Développement et validation du logiciel S4MPLE : application au docking moléculaire et à l'optimisation de fragments assistée par ordinateur dans le cadre du fragment-based drug design

Laurent Hoffer

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES

UMR 7140

THÈSE

présentée par

Laurent HOFFER

soutenue le : **03 juin 2013**

pour obtenir le grade de

Docteur de l'Université de Strasbourg

Discipline / Spécialité : Chimie / Chémoinformatique

Développement et validation du logiciel S4MPLE. Application au docking moléculaire et à l'optimisation de fragments assistée par ordinateur dans le cadre du Fragment-Based Drug Design.

THÈSE dirigée par :

Mr HORVATH Dragos	Directeur de recherche CNRS, Université de Strasbourg
Mr RENAUD Jean-Paul	Directeur de recherche CNRS en détachement, Sté NovAliX

RAPPORTEURS :

Mme IMBERTY Anne	Directrice de recherche CNRS, CERMAV
Mr VILLOUTREIX Bruno	Directeur de recherche INSERM, Université Paris-Diderot

---

MEMBRES DU JURY :

Mme IMBERTY Anne	Directrice de recherche CNRS, CERMAV
Mr VILLOUTREIX Bruno	Directeur de recherche INSERM, Université Paris-Diderot
Mr ROGNAN Didier	Directeur de recherche CNRS, Université de Strasbourg
Mr HORVATH Dragos	Directeur de recherche CNRS, Université de Strasbourg

Remerciements

Je souhaite ici rendre hommage et exprimer ma profonde reconnaissance à tous ceux qui, de près ou de loin, ont contribué à ce projet, et ce depuis sa définition jusqu'à son aboutissement.

En premier lieu, je tiens à exprimer ma gratitude à la société NovAliX pour m'avoir accueilli à la suite d'une candidature libre, et ce dès l'obtention de mon Master de Bioinformatique (Université de Strasbourg). A ce titre, je pense particulièrement au Docteur **Muller Pascal** qui fut mon premier contact interne au sein de l'entreprise, ainsi qu'aux responsables susnommés : le Docteur **Zeyer Denis** et M. **Jenn Stephan**.

Mes pensées vont aussi à l'Association Nationale de la Recherche Technique (ANRT), un acteur majeur visant à promouvoir des collaborations synergiques entre le Public et le Privé *via* le dispositif CIFRE, pour avoir donné l'opportunité de réaliser ce projet de recherche.

Je suis également reconnaissant envers le Professeur **Varnek Alexandre** pour m'avoir accueilli, tout au long de ces trois années, dans un excellent environnement scientifique au sein de son Laboratoire de Chémoinformatique de l'Université de Strasbourg.

Je voudrais ensuite remercier mes deux co-directeurs de thèse, les Docteurs **Horvath Dragos** et **Renaud Jean-Paul**, pour leur rigueur scientifique, leur disponibilité et bien entendu pour avoir constamment partagé leurs connaissances et expériences, le tout dans la bonne humeur et en me laissant une grande autonomie.

Bien évidemment, je voudrais adresser mes sincères remerciements aux Docteurs **Imberty Anne**, **Villoutreix Bruno** et **Rognan Didier** pour avoir accepté de faire partie de mon jury de thèse.

Je tiens également à saluer chaleureusement le Docteur **Marcou Gilles** pour ses nombreux conseils et les discussions diverses, autant scientifiques qu'informelles, que nous avons régulièrement eues.

Je souhaite aussi remercier tous les membres et collègues du Laboratoire de Chémoinformatique : **Muller Christophe** (un "heureux ex-thésard fraîchement diplômé" ayant toujours le mot pour rire), **De Luca Aurélie** (je te souhaite de valider prochainement ta thèse en plus d'un bon rétablissement), **Ruggiu Fiorella**, **Gaspard Héléna**, **Klimshuk Olga**, **Oprisiu Ioana**, **Bonachera Fanny**, **Ngoc Lam Nguyen**, **Khristova Tetiana**, **Kireeva Natalia**, **Chupakhin Vladimir**, **Solov'ev Vitaly**, **Baskin Igor** ; sans oublier les (ex-)doctorants du Laboratoire d'Innovation Thérapeutique, à savoir **Ray Anne-Marie**, **Meslamani Jamel**, **Desaphy Jérémy** et **Sturm Noé**.

Mes pensées vont aussi aux Docteurs en chimie **Ciapetti Paola** et **Klein Emmanuel**, au Docteur en biophysique structurale **Uhring Muriel**, et **Pocachard Nathalie** pour leur travail conséquent du côté expérimental au sein de l'entreprise NovAliX. Le projet interne n'a pu se terminer dans le délai imparti, mais merci pour vos conseils et vos efforts dans cette dernière ligne droite.

Pour finir, merci de tout cœur à **ma famille** bien-aimée qui m'a toujours soutenu, encouragé à poursuivre mes études (l'IUT semble bien loin maintenant), et sans qui je n'en serais pas là aujourd'hui.

Liste des abréviations

AG	Algorithme Génétique
AUC	Area Under the Curve (Aire sous la Courbe)
CO	Crossing Over
DDL	Degrés De Liberté
DM	Dynamique Moléculaire
DND	De Novo Design
EC	Echantillonnage Conformationnel
E_{pot}	Energie potentielle
ESI-MS-Nat	Spectrométrie de masse en conditions natives (ionisation de type électro-nébulisation)
FBDD	Fragment-Based Drug Design
FF	Force Field (Champ de Force)
FS	Fonction de Score / Fonction d'énergie
GC	Conjugate Gradient (Méthode du Gradient Conjugué)
HSP90	Heat Shock Protein 90
LE	Ligand Efficiency
LH	Liaison Hydrogène
MC	Monte Carlo
MM	Mécanique Moléculaire
MS	Mass Spectrometry (Spectrométrie de Masse)
PDB	Protein Data Bank
PIF	Pairwise Interaction Fingerprint (Fingerprint d'Interaction)
RMN	Résonance Magnétique Nucléaire
RMSD	Root Mean Square Deviation
ROC	Receiver Operating Characteristic
S4MPLE	Sampler for Multiple Protein-Ligand Entities
SBDD	Structure-Based Drug Design
SD	Steepest Descent (Méthode de la plus grande Pente)
SDF	Structure Data File

Table des matières

Introduction

Dans la recherche de molécules actives, la problématique d'optimisation des touches (“hits”) issues d'un criblage est complexe mais nécessaire dans l'optique du développement de ligands de haute affinité. Elle l'est d'autant plus pour le Fragment-Based Drug Design (FBDD) puisque le processus d'optimisation commence à partir de hits de petite taille (appelés fragments) possédant généralement une très faible affinité. Dans ce contexte, la modélisation aurait idéalement pour but de faciliter ce processus, entre autre *via* la prédiction du mode de liaison des différents hits et la suggestion de modifications potentiellement intéressantes au regard d'un ou de plusieurs critères objectifs (affinité, sélectivité, *etc.*).

Tout au long de ces dernières années, l'entreprise NovAliX a développé ses capacités dans le domaine innovant du FBDD. Son expertise dans des technologies essentielles pour ce nouveau paradigme de mise au point de molécules actives, entre autre la cristallographie et la spectrométrie de masse en conditions natives, en sont les meilleurs exemples.

La collaboration entre le Laboratoire de Chémoinformatique de l'Université de Strasbourg et la société NovAliX a pour but le développement de méthodes de modélisation représentant le pendant *in silico* des étapes clés du “FBDD expérimental”, d'où l'expression récurrente de “FBDD *in silico*”. La question sous-jacente est la suivante : “*Peut-on développer une stratégie virtuelle mais rationnelle d'optimisation, utilisable dans le contexte du FBDD, et reposant sur les concepts usuels de chémoinformatique et de modélisation moléculaire ?*”.

Ceci est d'autant plus audacieux que la difficulté à manipuler les ligands de très faible complexité comme les fragments, notamment au niveau de la qualité de prédiction de leur mode de liaison, est reconnue dans la littérature [1]. De plus, il est également connu que les fonctions de score calibrées sur des jeux d'entraînement composés essentiellement de complexes “ligand drug-like - récepteur” peuvent aboutir à de mauvaises prédictions / estimations pour d'autres classes de composés, comme par exemple les fragments. L'une des raisons est l'inadéquation vis-à-vis du domaine d'applicabilité desdites fonctions. Une nouvelle problématique apparaît alors : “*Une même fonction de score ou d'énergie est-elle compatible avec différentes classes de ligands ?*” ou “*Des taux de succès comparables, par exemple au niveau de la capacité à prédire le bon mode de liaison en docking,*

peuvent-ils être obtenus avec la même fonction pour des entités organiques de tailles variées (ligands drug-like vs. fragments) ?".

En plus de devoir s'atteler à ces questions fondamentales, ce travail s'inscrit également dans un projet plus large du Laboratoire de Chémoinformatique lié au développement interne de l'outil de modélisation moléculaire S4MPLE reposant sur un algorithme génétique hybride. Son objectif est de pouvoir gérer des problèmes difficiles et variés (gestion simultanée de plusieurs ligands, docking avec un site partiellement flexible, repliement d'oligopeptide), tout en restant le plus générique possible et en permettant une sélection fine des degrés de liberté (DDL) du système. La plupart des outils de docking sont limités au regard de l'ensemble des problèmes évoqués. De manière similaire, les outils classiques de mécanique moléculaire n'ont pas été conçus pour figer une partie arbitraire du système et emploient plus volontiers le formalisme des équations du mouvement de Newton que les algorithmes génétiques pour l'exploration conformationnelle du système.

Dans un but de synergie entre ces deux axes de travail, il est évidemment souhaitable d'intégrer S4MPLE dans la mesure du possible au sein du contexte FBDD *in silico*, notamment dans sa composante liée à l'échantillonnage conformationnel.

Après une première étape impliquant des développements et la validation de ce programme selon les canons en vigueur dans la communauté scientifique, un recentrage autour de la notion de FBDD *in silico* est entrepris afin de répondre au cahier des charges initial illustré par le titre de cette thèse, notamment *via* le docking puis l'optimisation virtuelle de composés de type fragment.

Le premier chapitre décrit le contexte général du projet, ainsi que des concepts fondamentaux régulièrement abordés tout au long de ce manuscrit, tant des points de vue théorique, expérimental et modélisation.

Le second chapitre, intitulé "matériel et méthodes", décrit de manière plus approfondie les outils développés et utilisés dans le cadre de ce travail. Un focus particulier est réalisé sur le programme S4MPLE, ainsi que sur les outils satellites relevant du protocole FBDD *in silico*.

Les chapitres suivants de ce manuscrit correspondent aux différents résultats et étapes clés du projet :

- le troisième chapitre aborde le développement et la validation de S4MPLE en tant que tel, et porte également sur la calibration de sa fonction d'énergie basée sur un champ de force
- le recentrage sur le FBDD commence lors du quatrième chapitre, et a pour but de démontrer la capacité de S4MPLE (et de sa nouvelle fonction d'énergie) à gérer des molécules de type fragment au cours de simulations de docking moléculaire
- l'avant-dernier chapitre concerne la partie FBDD *in silico* à proprement parler, et est dédié à l'optimisation virtuelle de fragment(s), *via* la suggestion dans un premier temps de molécules raisonnables pouvant être considérées comme une amélioration possible de ces composés initiaux, suivi de leur échantillonnage explicite dans le site de liaison à l'aide de S4MPLE. Ce protocole est validé à travers plusieurs études rétrospectives et variées portant sur des données issues de la littérature scientifique
- étant donné que cette thèse est financée par une bourse CIFRE, elle doit permettre des développements exploitables tant dans le milieu académique que dans un cadre industriel. Le dernier chapitre est consacré à l'application du protocole mis au point à un projet concret interne à l'entreprise NovAliX

1 Contexte général et notions fondamentales

Tout projet de développement de méthodes de modélisation, ayant pour but la conception de molécules actives, se situe à l'interface entre plusieurs disciplines comme la biologie, la chimie (théorique et expérimentale) et l'informatique. Ce premier chapitre vise à résumer le contexte biologique/pharmaceutique et les notions théoriques nécessaires à la compréhension des méthodes utilisées dans le cadre de ce travail. Par exemple, seront entre autre abordés le processus de développement d'un médicament, les principes de reconnaissance moléculaire, le FBDD, la modélisation moléculaire, l'échantillonnage conformationnel, le docking et le *de novo* design.

1.1 Le processus de développement d'un médicament

Un médicament est défini, selon l'article L5111-1 du code de santé public français [2], de cette manière : "*On entend par médicament toute substance ou composition présentée comme possédant des propriétés curatives ou préventives à l'égard des maladies humaines ou animales, ainsi que toute substance ou composition pouvant être utilisée chez l'homme ou chez l'animal ou pouvant leur être administrée, en vue d'établir un diagnostic médical ou de restaurer, corriger ou modifier leurs fonctions physiologiques en exerçant une action pharmacologique, immunologique ou métabolique*". Il est constitué d'un ou plusieurs principe(s) actif(s) ainsi que d'un ou plusieurs excipient(s) :

- les principes actifs présents dans les médicaments usuels appartiennent à la classe des petites molécules organiques. Ces dernières ont des utilités très variées dans l'organisme, allant du rôle de cofacteur qui est nécessaire pour réaliser certaines réactions enzymatiques, jusqu'à des rôles de signalisation endocrinienne ou synaptique. On parle d'inhibiteur lorsque le principe actif bloque le fonctionnement d'une enzyme, et les phénomènes d'activation ou d'inactivation d'un récepteur sont le résultat de principes actifs agissant respectivement comme agonistes, antagonistes ou agonistes inverses. Les principes actifs se lient à leurs cibles selon des principes de reconnaissance moléculaire qui sont décrits ci-dessous (voir le §1.2)
- les excipients peuvent être décrits comme une sorte de "substance d'emballage" du principe actif, sont responsables de l'aspect final du médicament (comprimé, gélule, *etc.*), et interviennent également au niveau de la libération du principe actif

Le développement d'un nouveau médicament est un processus complexe, long, très coûteux, et fortement encadré juridiquement à cause des risques de santé publique. En réalité, seules les plus

grandes entreprises de l'industrie pharmaceutique sont capables d'assurer le développement d'un médicament de A à Z. Celui-ci est divisé en plusieurs phases, chacune nécessitant la validation de la précédente pour pouvoir continuer. Il est à noter qu'à chaque étape, un grand nombre de molécules échoue à passer au stade suivant de développement. On estime qu'il faut 10-15 ans de recherche et d'essais cliniques pour mettre au point un médicament [3]. Les principales étapes, présentées dans la Figure 1, sont la phase de recherche, la phase préclinique et les différentes phases cliniques.

Figure 1: Les différentes étapes de développement d'un médicament.

Une brève description de chacune de ces phases est donnée ci-dessous.

La phase de recherche fondamentale

Un projet de développement de médicament commence généralement par une phase de recherche qui consiste en l'identification d'une cible impliquée dans une pathologie donnée. Les cibles thérapeutiques sont généralement des protéines, notamment des récepteurs ou des enzymes.

La phase préclinique

Une fois la cible validée, le but est de mettre au point un ligand modulant son activité, en l'activant ou en l'inhibant, afin d'obtenir l'effet curatif souhaité. L'identification de hits se fait par le criblage expérimental d'une chimiothèque. Des approches de modélisation, détaillées ultérieurement dans ce manuscrit, peuvent permettre de ne sélectionner qu'un sous-ensemble de molécules à tester en priorité. Les hits primaires confirmés sont ensuite optimisés dans le but d'améliorer leur affinité pour la cible, leur spécificité vis-à-vis d'autres macromolécules biologiques, et leurs propriétés pharmacocinétiques. Différents filtres, connus sous l'acronyme ADMET, sont appliqués pour éliminer les molécules indésirables le plut tôt possible dans le processus [4].

[A]bsorption

Cela consiste à analyser la faculté d'une molécule à pénétrer au sein de l'organisme après administration. Une faible solubilité (par exemple due à une trop grande hydrophobicité) ou une forte polarité ont un impact drastique sur l'absorption intestinale d'un composé.

[D]istribution

Ce critère mesure la capacité d'une molécule à diffuser, par exemple *via* le flux sanguin, à travers l'organisme. En effet, une molécule doit pouvoir passer d'un compartiment à un autre, afin de pouvoir arriver *in fine* à l'endroit où sa cible doit être atteinte. La forte liaison à des protéines plasmatiques a un impact négatif sur la distribution d'une molécule.

[M]étabolisme

Le filtre métabolisme vise à détecter (a) la stabilité de la molécule dans l'organisme qui impacte sur son temps d'action et (b) les métabolites de la molécule initiale, à savoir les composés résultant de sa dégradation ou de modifications enzymatiques ayant lieu au sein de l'organisme. Chez l'homme, les cytochromes P450 du foie sont les principales enzymes modifiant les xénobiotiques. Ces derniers sont notamment rendus plus hydrophiles par l'introduction d'atomes d'oxygène. Les métabolites peuvent être inactifs, plus actifs que le composé original et bien entendu potentiellement toxiques, d'où la nécessité de les caractériser et de les étudier.

[E]xcrétion

Afin d'éviter les phénomènes d'accumulation, souvent synonyme de toxicité, il faut veiller à ce que les composés administrés, ainsi que leurs métabolites, soient bien excrétés de l'organisme, par exemple *via* l'urine ou les selles.

[T]oxicité

Comme son nom l'indique, ce filtre sert à mesurer la toxicité d'un composé et de ses métabolites. Différents types de toxicité sont évalués, entre autre la cancérogénicité et la tératogénicité. Le test d'Ames est un exemple bien connu mesurant la toxicité d'un composé à travers son pouvoir mutagène / cancérogène [5].

Ces différents critères sont évalués à l'aide de tests *in vitro* (*ex*: Caco2 pour l'absorption intestinale) ou *in vivo* avec l'utilisation d'animaux modèles [4]. Des approches chémoinformatiques [6], telles les méthodes QS[A/P]R (Quantitative Structure–[Activity/Property] Relationship [7, 8]), permettent de créer des modèles mathématiques qui prédisent certaines propriétés/activités de molécules [9], en se basant sur des méthodes d'apprentissage [10] et des descripteurs calculés à partir des structures des composés [11]. Un bénéfice notable des méthodes QSAR est le fait de pouvoir identifier en amont les molécules les plus inappropriées du plan de synthèse et de test. Cependant, ils sont encore loin d'être assez précis afin de se substituer aux tests expérimentaux pour les candidats précliniques choisis. Sachant que de plus en plus de données structurales sont disponibles pour des cibles liées aux processus ADMET (cibles directes, entre autre les cytochromes, les protéines de transport comme HSA et des cibles à éviter comme le canal hERG ; cibles indirectes comme les récepteurs nucléaires régulant la

transcription de gènes codant pour des enzymes du métabolisme), des études de modélisation 3D basées sur ces structures peuvent également être entreprises en plus des approches QSAR [12]. Enfin, la chémogénomique [13, 14] peut également jouer un rôle essentiel *via* l'identification de cibles alternatives (essentiellement indésirables dans ce contexte ADMET à un stade précoce, voir les cibles évoquées ci-dessus) pour les ligands actifs sur la macromolécule biologique d'intérêt [12].
Les composés passant tous ces filtres avec succès deviennent des candidats médicament et entrent en phase clinique.

Les phases cliniques

Il s'agit des premières phases où le composé d'intérêt est administré à l'homme. Au cours du temps, plusieurs phases ont été définies, chacune répondant à une problématique bien spécifique. L'enjeu est de décider si le candidat peut devenir un médicament en tant que tel, en analysant sa balance bénéfices / risques (efficacité thérapeutique *vs.* effets secondaires).

La phase I

La phase I consiste à déterminer sur des volontaires sains l'innocuité du composé, les doses maximales tolérées, ainsi que l'apparition éventuelle d'effets secondaires encore inconnus à ce stade.

La phase II

La phase II est la première étape où l'on administre le candidat médicament à des patients, dont l'ordre de grandeur est de quelques centaines. Le but est d'évaluer la réelle efficacité de la molécule ainsi que les doses minimales efficaces - et donc la marge thérapeutique vis-à-vis des résultats obtenus à la phase précédente - tout en répertoriant l'apparition éventuelle de nouveaux effets secondaires.

La phase III

La phase III est conceptuellement similaire à la phase précédente : il s'agit à nouveau d'administrer la molécule à un groupe de patients. Ici, le nombre de personnes est sensiblement plus important, de l'ordre de quelques milliers. Des malades ayant des profils ethniques variés est également souhaitable pour être le plus proche possible de la réalité. Cette phase permet de confirmer la sécurité et l’efficacité du candidat médicament sur la pathologie ciblée. Une comparaison par rapport à un panel de molécules concurrentes déjà présentes sur le marché est également entreprise, notamment pour pouvoir estimer le service médical rendu par le candidat.

La phase d'enregistrement et de commercialisation

A ce stade, un dossier récapitulant l'ensemble des données des phases précliniques et cliniques est remis aux autorités de santé compétentes, dans le but d'obtenir une autorisation de mise sur le marché (AMM). Ce sésame ouvre la voie à la commercialisation de la molécule.

La phase IV

Cette phase relativement récente (également appelée pharmacovigilance) consiste à surveiller, à travers le réseau de médecins et après la mise sur le marché de la molécule, l'apparition d'effets indésirables non détectés précédemment.

1.2 Les principes de reconnaissance moléculaire

Les processus de reconnaissance moléculaire sont au cœur de tout projet de développement de molécules actives. Ils sont donc tout naturellement abordés dans ce premier chapitre sous différents prismes étroitement liés : leur représentation schématique, l'introduction des constantes d'équilibre et leur relation avec l'énergie libre de liaison standard.

1.2.1 Représentations simplifiées des phénomènes de reconnaissance moléculaire

Les processus de reconnaissance moléculaire sont généralement représentés de manière simplifiée sous la forme du modèle clé-serrure (voir la Figure 2) qui a été proposé par Hermann Emil Fischer dès 1894. Ce modèle fait apparaître la notion de complémentarité de forme entre une molécule et sa cible, et le complexe résultant est stabilisé par des interactions favorables entre les deux entités. Ces dernières sont introduites au §1.2.4.

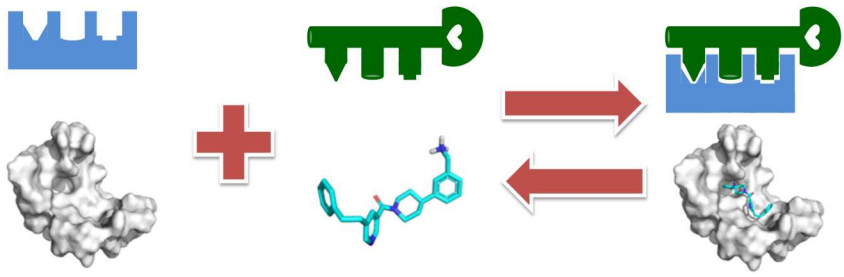


Figure 2: Le modèle clé-serrure pour modéliser la formation d'un complexe RL à partir d'un ligand L et d'un récepteur R.

Cependant, une serrure et une clé sont des entités rigides contrairement aux molécules qui peuvent souvent adopter un ensemble de conformations stables, d'où la limite de ce modèle, certes très intuitif

mais trop simpliste. La remise en cause de cette vision statique des choses a donné naissance à une forme alternative de représentation, à savoir le modèle main-gant ou d'ajustement induit (“induced fit”) proposé par Daniel Koshland en 1958. Dans ce dernier, la main représente le ligand qui peut à la fois changer de conformation et induire une modification de la forme du gant qui symbolise le récepteur.

1.2.2 La liaison d'un ligand à son récepteur

La liaison d'un ligand à sa cible est un évènement (généralement) réversible impliquant des interactions non liées. Soient R une cible libre, L un ligand libre et RL le complexe résultant de la liaison de L avec R (voir la Figure 2). On définit les constantes de dissociation K_d (en M) et d'association K_a (M^{-1}) à l'équilibre de la sorte :

$$K_d = \frac{[R] * [L]}{[RL]} = \frac{1}{K_a} \qquad \text{Eq 1}$$

En pratique, le K_d est la concentration en ligand qui reste libre alors que 50% des sites de la cible sont occupés à l'équilibre.

1.2.3 Aspects thermodynamiques des processus de reconnaissance moléculaire

Les processus de reconnaissance moléculaire sont régis par les lois de la thermodynamique [15, 16]. Puisque ce sous-chapitre n'a pas pour objectif d'être un résumé de cette discipline, seules les informations directement reliées aux processus de reconnaissance moléculaire seront brièvement introduites.

Lorsque l'on considère des transformations à pression P et température T constantes, l'enthalpie libre (ou énergie libre de Gibbs) G est le potentiel thermodynamique à considérer pour les étudier. Ce dernier est défini comme la différence de deux composantes, l'une enthalpique et l'autre entropique :

$$G = H - TS \qquad \text{Eq 2}$$

A température T constante, sa variation ΔG est égale à $\Delta G = \Delta H - T\Delta S$ car G, H et S sont des fonctions d'état qui ne dépendent que des états initial et final.

Cette variation est notamment utilisée pour prédire le sens d'évolution du système :

- $\Delta G < 0$: transformation spontanée thermodynamiquement possible
- $\Delta G = 0$: système à l'équilibre (aucune variation d'énergie libre)
- $\Delta G > 0$: transformation spontanée thermodynamiquement impossible

Lorsque l'on considère un système ligand-récepteur (comme défini dans la Figure 2),on s'intéresse à une énergie libre de liaison qui correspond à une différence d'énergie libre entre la forme complexée et les formes libres des entités :

$$\Delta G_{\text{liaison}} = G_{RL} - (G_R + G_L) \qquad \text{Eq 3}$$

Toutefois, ce type de grandeur dépend des conditions expérimentales (concentration des entités du système, température, *etc.*). Ainsi, seules les valeurs standards (par exemple G° et ΔG°) sont fixes et peuvent être tabulées. Ces constantes sont obtenues en réalisant les mesures dans des conditions dites standards (P = 1 atm, T = 298 °K, conditions sur la concentration des entités, *etc.*). En pratique, cette énergie libre standard de liaison ($\Delta G°_{liaison}$) est directement reliée aux constantes d'équilibre (K_a, K_d) définies précédemment :

$$\Delta G°_{\text{liaison}} = -RT\ \ln(K_a) = RT\ \ln(K_d) \qquad \text{Eq 4}$$

Le K_d, exprimé en M, représente donc aussi l'affinité d'un ligand pour sa cible. Enfin, les phénomènes de reconnaissance ligand-récepteur peuvent être dominés soit par la composante enthalpique soit par la composante entropique. Il est à noter qu'une compensation entropie-enthalpie change profondément le profil de liaison d'un composé à sa cible, mais ne change aucunement l'énergie libre. Ce type d'évènement (compensation entropie-enthalpie) est fréquent [17] et a été mis en évidence expérimentalement durant l'optimisation de séries de ligands [18].

1.2.4 Les principales classes d'interaction rencontrées en biologie

En dehors de la formation d'une liaison covalente entre le ligand et sa cible comme c'est le cas pour les inhibiteurs dits suicide, l'association d'un ligand à sa cible est réversible et résulte d'interactions favorables non liées / non covalentes entre eux. Etant donné que les principales classes d'interaction rencontrées en biologie et chimie médicinale seront régulièrement abordées dans ce manuscrit, elles

sont brièvement introduites dans ce sous-chapitre. Une interaction peut notamment impliquer un ligand et son récepteur (interactions intermoléculaires) ou concerner une seule entité (interactions intramoléculaires). On peut également se référer à la revue de Bissantz *et al* pour une description très approfondie et quasi-exhaustive des différents types d'interaction ligand-récepteur rencontrés en biologie [19].

La notion d'électronégativité

La notion d'électronégativité est fondamentale pour comprendre bon nombre d'interactions. On considère ici la définition de l'électronégativité au sens de Pauling. Concrètement, un atome électronégatif a la capacité d'attirer le nuage électronique vers lui-même. Lorsqu'il y a une différence d'électronégativité entre les atomes d'une liaison covalente, le doublet électronique de la liaison n'est plus partagé équitablement entre les atomes impliqués, et il y a l'apparition d'un dipôle (voir la Figure 3). On attribue une charge partielle δ^- à l'atome le plus électronégatif et une charge partielle δ^+ à l'atome le moins électronégatif. A l'opposé, lorsque la différence d'électronégativité est très faible ou nulle, le doublet électronique de la liaison reste globalement équitablement réparti entre les atomes.

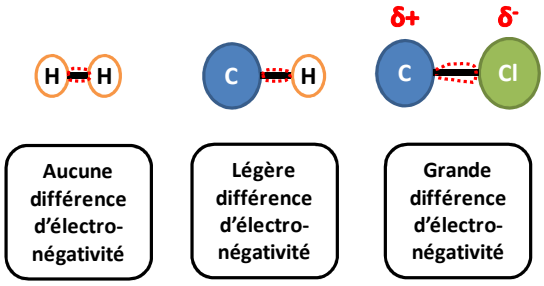


Figure 3: Impact sur une liaison covalente de la différence d'électronégativité entre les atomes la constituant.

Enfin, les particules chargées génèrent localement un champ électrique, ouvrant la voie à des interactions électrostatiques. Ces dernières sont attractives lorsque deux charges en interaction sont de signes opposés, et répulsives lorsque les charges sont de même signe.

Les liaisons ioniques

Les liaisons ioniques ou ponts salins concernent les fortes interactions électrostatiques attractives entre charges de signe opposé (voir la Figure 4 et la Figure 12A). La distance entre les centres chargés doit être inférieure à 4,0-4,5 Å environ pour que l'on considère l'interaction électrostatique comme une liaison ionique en tant que telle. Bien qu'elle soit plus faible que celle d'une liaison covalente, l'énergie

d'une liaison ionique reste très importante pour une interaction non covalente (de l'ordre de 10 kcal.mol^{-1}). Elles sont d'autant plus fortes lorsqu'elles se forment dans un environnement ayant une faible constante diélectrique (voir la loi de Coulomb décrite §1.4.1), comme c'est le cas dans le vide ou au cœur des protéines. Dans ce dernier cas, les liaisons ioniques stabilisent fortement leurs structures tridimensionnelles [20]. Par exemple, un pont salin est créé entre la chaîne latérale d'un résidu glutamate et celle d'un résidu lysine lorsque les charges sont situées à proximité dans l'espace.

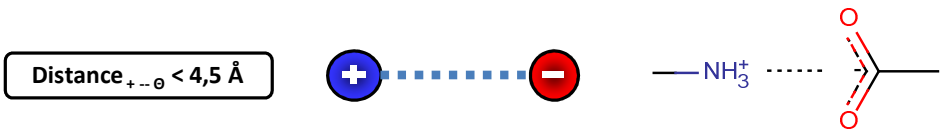


Figure 4: La liaison ionique.

Les liaisons hydrogène

Les liaisons hydrogène (LH, voir la Figure 5 et la Figure 12A/C) sont des interactions d'origine électrostatique entre un atome électronégatif dit "accepteur" (A porteur d'une charge partielle δ^-) et un atome d'hydrogène polaire (H porteur d'une charge partielle δ^+) qui est lié de manière covalente à un atome électronégatif dit "donneur" (D porteur d'une charge partielle δ^-) [21]. Les atomes d'oxygène et d'azote sont les donneurs et accepteurs les plus courants dans notre contexte d'intérêt. Une distance faible entre le donneur et l'accepteur est nécessaire pour qu'une LH soit significative, et son énergie est maximale (de l'ordre de 5 kcal.mol^{-1} pour celles rencontrées en biologie [21]) lorsque les atomes D-H..A sont alignés. Du fait de cette anisotropie, les LH sont considérées comme des interactions spécifiques.

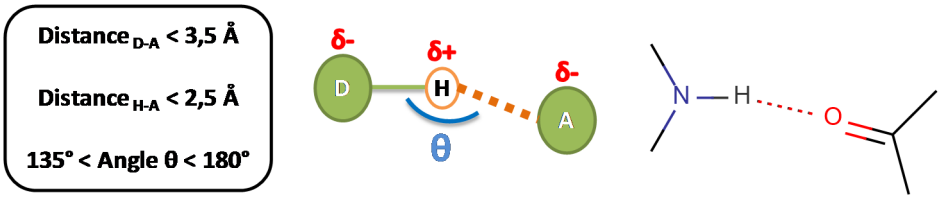


Figure 5: La liaison hydrogène.

Les interactions dipolaires impliquant un dipôle permanant

Dans les molécules globalement neutres, les interactions électrostatiques ne disparaissent pas pour autant. Même si les charges partielles se neutralisent, la séparation locale des charges due aux différences d'électronégativité génère néanmoins des dipôles. D'un point de vue pratique, ces interactions sont souvent décrites par la loi de Coulomb entre les charges partielles atomiques, plutôt que par le formalisme (plus complexe) régissant les interactions dipôle-charge ou dipôle-dipôle.

Elles incluent :

- les interactions de Keesom entre dipôles permanents (voir la Figure 6A). Ces derniers s'auto-organisent de manière à orienter face à face leurs charges de signes opposés, afin de maximiser les interactions électrostatiques. Au regard de cette définition, les liaisons hydrogène peuvent être vues comme une sous-famille des interactions de Keesom qui impliquent nécessairement un hydrogène polaire, mais de par leur caractère ubiquitaire et leur forte intensité, elles sont en pratique considérées comme une classe d'interaction à part entière
- les interactions de Debye entre dipôles permanents et dipôles induits (voir la Figure 6B). Par définition, une molécule apolaire possède un moment dipolaire nul. Cependant, sous l'effet d'un champ électrique, par exemple généré par un dipôle permanent proche telle une molécule polaire, il peut y avoir une légère polarisation de la molécule apolaire d'où l'apparition d'un dipôle induit capable de réaliser des interactions électrostatiques avec d'autres dipôles. Ce phénomène est également connu sous le nom de forces d'induction

Les interactions de Van der Waals

Les interactions de Van der Waals (VdW), également connues sous le nom d'interactions de London ou de forces de dispersion, impliquent des dipôles instantanés (voir la Figure 6C). Contrairement aux interactions dipolaires décrites auparavant, les interactions de VdW n'impliquent pas de dipôle permanent.

Lorsque les barycentres des charges positives et négatives coïncident, le moment dipolaire d'une molécule est nul, et celle-ci est considérée comme apolaire. Cependant, les électrons d'un atome ne sont pas figés : ils se déplacent dans des zones favorisées de l'espace appelées orbitales. Ainsi, il y a création en permanence de "micro-dipôles instantanés", car à chaque instant les barycentres des charges positives et négatives ne se superposent pas forcément. Ces dipôles dits instantanés vont également créer des interactions électrostatiques. Il est à noter que ce phénomène, produit par le mouvement des électrons autour du noyau, n'est bien entendu pas spécifique aux molécules apolaires et s'ajoute aux autres interactions dipolaires dans le cas de composés polaires.

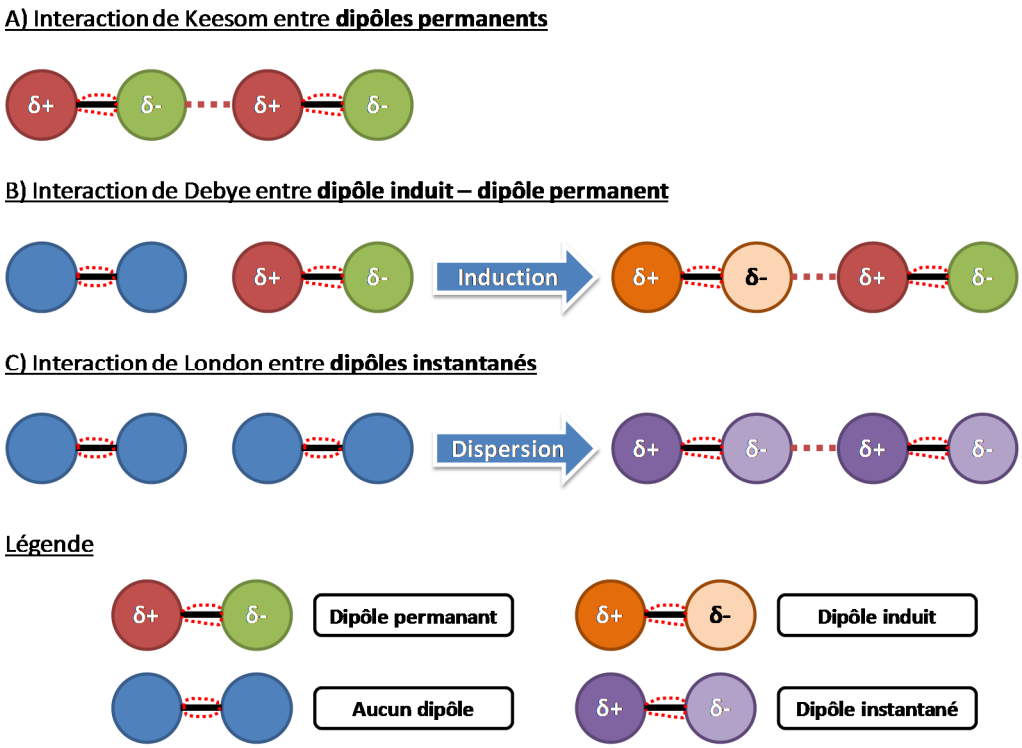


Figure 6: Illustration des principaux types d'interaction entre dipôles.

Les interactions de VdW sont de faible intensité (ordre de grandeur du kcal.mol^{-1}), mais elles sont néanmoins très importantes en pratique à cause de leur très grand nombre. Bien que cela puisse paraître assez simpliste, ce sont les interactions de VdW qui sont essentiellement responsables de la complémentarité stérique lors d'un phénomène de liaison. En effet, leur intensité augmente avec la diminution de la distance jusqu'à un seuil donné qui est fonction des éléments impliqués dans l'interaction. En dessous de cette distance limite, les interactions de VdW attractives deviennent négligeables au regard des interactions répulsives qui sont dues au trop grand rapprochement des nuages électroniques des atomes non liés considérés. C'est ce phénomène qui empêche d'avoir des interactions électrostatiques avec une distance extrêmement faible, ce cas étant même encouragé par la loi de Coulomb prise en dehors de toute autre considération.

L'effet hydrophobe et les contacts hydrophobes

En solution aqueuse, tout groupement accessible est entouré de molécules d'eau. Expérimentalement, on observe un réarrangement spontané des groupements apolaires les uns par rapport aux autres dans ce milieu polaire. C'est le cas dans la création des membranes biologiques (bicouche lipidique) qui sont constituées de molécules amphiphiles. Ce type de molécule est composé de deux pôles : l'un hydrophile, l'autre hydrophobe. On observe une auto-organisation de ces molécules dans l'eau, où les

queues hydrophobes se rassemblent entre elles au centre, et où seuls les groupements hydrophiles se retrouvent exposés au solvant tout en interagissant favorablement avec lui (voir la Figure 7).

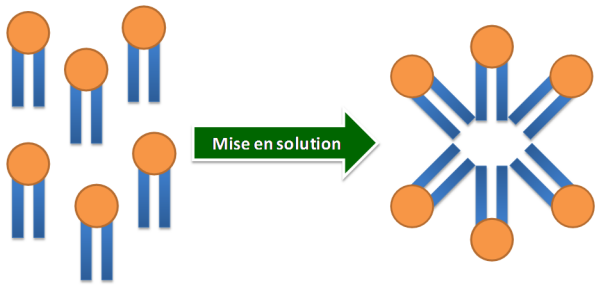


Figure 7: Auto-organisation de molécules amphiphiles dans un solvant aqueux.
Les zones orange représentent les parties hydrophiles et les queues bleues les chaînes hydrophobes.

Ce phénomène très complexe est appelé l'effet hydrophobe, et peut être résumé de cette façon : un soluté hydrophobe perturbe la structure tridimensionnelle usuelle de l'eau liquide car il se forme une sorte de "cage d'eau" dans la première couche de solvatation du soluté (il n'y a pas de liaison hydrogène entre le soluté hydrophobe et le solvant). Ces molécules d'eau sont considérées comme très ordonnées autour du soluté, d'où la notion de cage, et la conséquence directe est une perte d'entropie. Si les solutés hydrophobes (ou leurs groupements hydrophobes respectifs) se regroupent entre eux de manière à diminuer leur surface accessible à l'eau, alors il ne se forme qu'une seule, mais plus grande, cage de solvant autour de cet agrégat. Cependant, sa surface est plus petite que la somme des surfaces des cages autour de chaque soluté (considéré de manière individuelle). De fait, ce réarrangement permet de minimiser le nombre de molécules d'eau piégées autour des solutés et donc de maximiser le nombre de molécules d'eau libres, d'où une augmentation de l'entropie de l'univers (solutés + solvant dans son ensemble) qui explique le caractère spontané de cette réorganisation du système selon le second principe de la thermodynamique. Ainsi, les molécules apolaires n'interagissent pas favorablement avec l'eau, non pas parce qu'il y aurait une sorte de répulsion entre ces groupements et les molécules d'eau - il y a même des interactions de type Debye et London entre ces entités - mais parce que cette organisation spatiale impliquerait une augmentation défavorable de l'énergie libre *via* la diminution de sa composante entropique.

En solution aqueuse, le site de liaison et le ligand sont solvatés. Un évènement de liaison nécessite au préalable la désolvatation du site et du ligand. L'effet hydrophobe est responsable de la désolvatation spontanée du site et du ligand, dès lors qu'un contact intermoléculaire peut se créer entre zones hydrophobes de ces entités. Un enfouissement de ces zones est d'autant plus favorable qu'il diminue drastiquement la surface hydrophobe accessible au solvant (voir la Figure 12D). Outre l'aspect

entropique, ces contacts hydrophobes intermoléculaires augmentent également l'affinité *via* des interactions de VdW. A l'inverse, les interactions favorables de type LH entre groupes polaires du site et du ligand d'une part et le solvant d'autre part doivent être rompues avant la formation du complexe ligand-récepteur. Ce réarrangement-ci a un coût énergétique appelé "pénalité de désolvatation". La Figure 8 résume de manière schématique les aspects énergétiques liés aux interactions L-R en solution aqueuse en fonction de leur type (hydrophobe et hydrophile).

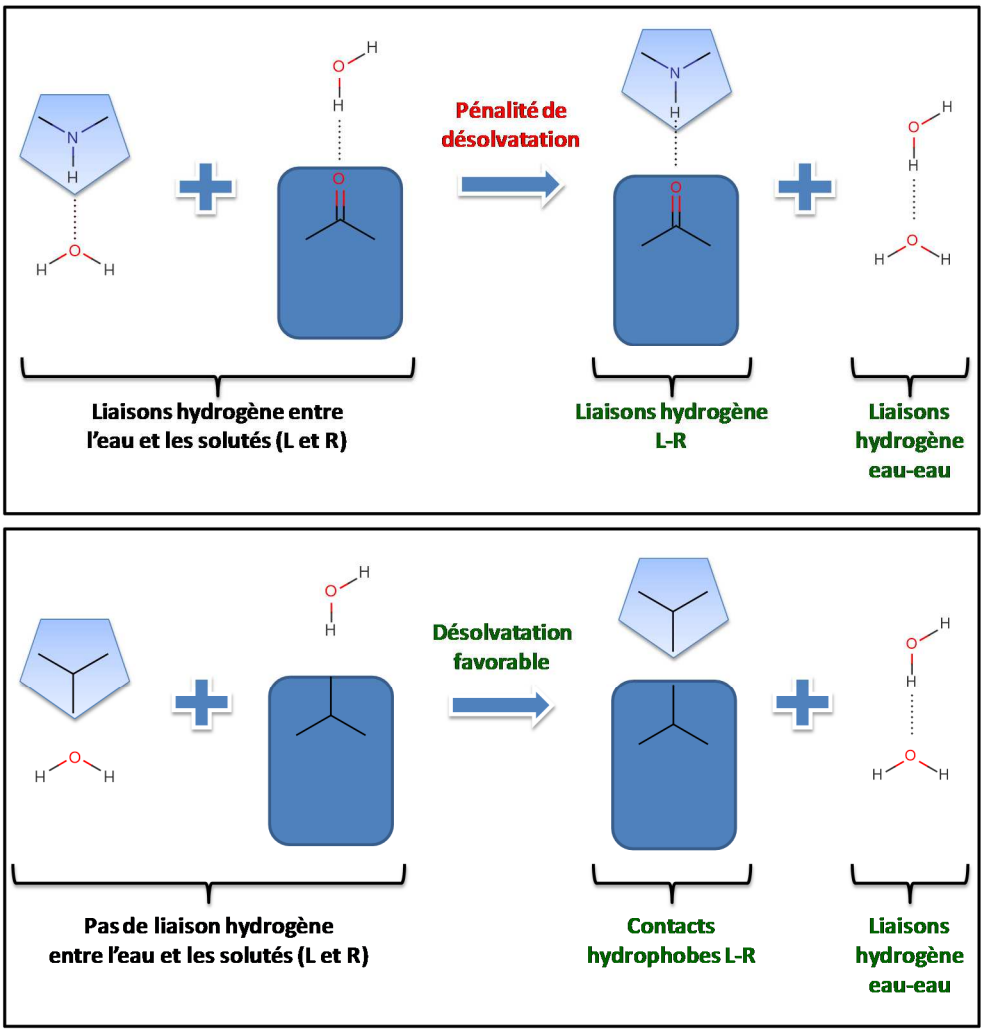


Figure 8: Impact de l'eau sur les aspects énergétiques des interactions ligand-récepteur (entre groupements polaires vs. entre groupements apolaires).

Le ligand est représenté sous la forme d'un pentagone et le récepteur sous la forme d'un rectangle aux angles arrondis. Les contributions énergétiquement défavorables sont représentées en rouge et celles favorables en vert.

Cet effet hydrophobe est fondamental pour favoriser puis maintenir le repliement tridimensionnel des peptides et protéines solubles, dont le cœur incorpore souvent de nombreux résidus à chaîne latérale

hydrophobe (contacts intramoléculaires) [22]. De même, l'incorporation de groupements hydrophobes est également très utilisée dans les phases d'optimisation pour augmenter l'affinité d'un ligand par "remplissage" de cavités hydrophobes [23], dans la mesure où sa solubilité et ses propriétés pharmacocinétiques de manière générale ne sont pas trop impactées. A titre d'exemple, l'ajout d'un simple méthyle (groupement considéré comme hydrophobe) en méta du cycle phényle d'un inhibiteur de l'enzyme dipeptidyl peptidase IV permet d'obtenir une IC_{50} environ 40 fois plus faible (5 nM vs 200 nM) [23]. Ce phénomène, où l'ajout d'un unique atome de carbone aliphatique permet un gain substantiel d'affinité, est appelé "méthyle magique", et un article récent lui est consacré [24]. Cette étude énumère des cas provenant de la littérature où un important gain d'affinité (d'un facteur 100 environ) est observé suite à l'ajout d'un simple groupe méthyle, et deux caractéristiques sont clairement mises en évidence :

- le méthyle se lie dans une cavité hydrophobe (il est entouré de chaînes latérales hydrophobes aliphatiques et/ou aromatiques)
- le méthyle favorise une conformation non liée proche de celle qu'adopte le ligand lorsqu'il s'associe à sa cible, d'où un bonus entropique aboutissant à un gain notable d'affinité. En effet, cette nouvelle contrainte sur la structure tridimensionnelle du ligand libre implique une diminution de la perte d'entropie conformationnelle à la suite du phénomène de liaison. Un exemple de ce type est l'incorporation d'un méthyle en ortho dans une structure bi-aryle

Les interactions entre systèmes π

Les interactions non liées π-π impliquent des systèmes π comme les cycles aromatiques. Ces derniers sont très souvent présents dans les ligands organiques et dans les macromolécules biologiques (H, F, W et Y dans le cas des protéines). Deux configurations sont principalement observées expérimentalement (voir la Figure 9) [25-27] :

- interaction π-π face à face - les deux cycles aromatiques sont parallèles et leurs centres géométriques sont distants de moins de 4 Å environ. Deux sous-catégories peuvent être créées selon la présence (conformation décalée) ou non (conformation en sandwich) d'un décalage entre les deux cycles parallèles
- interaction π-π en forme de T ("edge to face") - les deux cycles aromatiques forment un angle droit et leurs centres géométriques (centroïdes) sont distants de moins de 5,5 Å environ

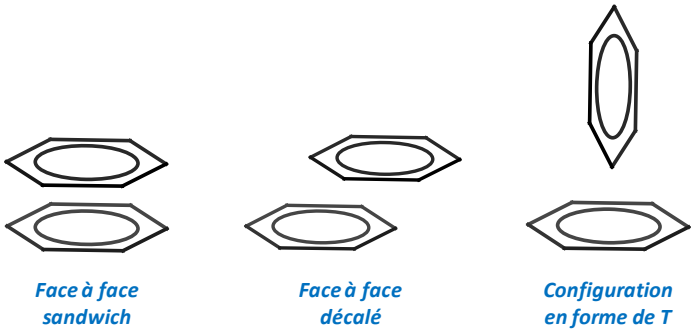


Figure 9: Les différentes configurations d'interactions non liées π-π.

Les limites de distance fournies ci-dessus sont celles utilisées dans les outils de modélisation développés par Schrödinger [28]. La Figure 12B montre un exemple d'interaction π-π du second type entre un ligand et un résidu aromatique du site de liaison. En biologie, on retrouve aussi ces interactions π-π stabilisatrices au sein de l'ADN et des protéines (interactions intramoléculaires) [29] et aux interfaces de complexes protéine-protéine et protéine-acide nucléique (interactions intermoléculaires) [30].

Les interactions cation-π

Comme leur nom le suggère (voir la Figure 10), ces interactions non liées de nature électrostatique impliquent un cation (un ion positif ou un groupe porteur d'une charge positive) et un système π (par exemple un cycle aromatique) [31], et sont relativement fréquentes en biologie et en chimie [32]. Comme pour d'autres classes d'interaction, un critère de distance permet d'identifier de manière géométrique les interactions cation-π : la distance entre le centre chargé et le centre géométrique du système π doit être inférieure à 5 Å environ [28].

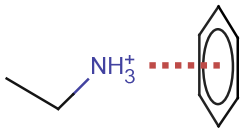


Figure 10: L'interaction non liée cation-π.

Au niveau protéique, les résidus R et K jouent le rôle du cation, tandis que F, Y et W représentent le système π [33]. Une fréquence plus importante est observée expérimentalement dans les protéines en faveur des interactions cation-π intramoléculaires impliquant R et W [34]. Ce type d'interaction est également fondamental au niveau épigénétique : des protéines spécifiques (“lecteurs épigénomes”) reconnaissent des lysines mono-, di- et/ou tri- méthylées par le biais de cages aromatiques [35, 36]. Enfin, le mode de liaison de nombreux ligands endogènes, parmi lesquels des neurotransmetteurs comme

l'acétylcholine et les catécholamines, implique ce type d'interaction [37]. La Figure 12C montre un exemple d'interaction cation-π entre un ligand et plusieurs chaînes latérales de résidus aromatiques du site de liaison.

Les interactions impliquant des métaux

Les groupements chimiques possédant un doublet d'électrons non appariés peuvent former une liaison de coordination avec un cation métallique, dès lors qu'un site de coordination est libre au niveau de cet ion (voir la Figure 11).

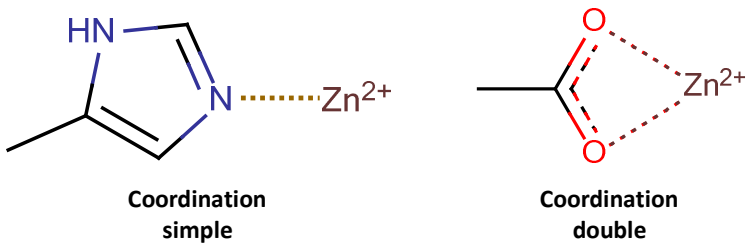


Figure 11: Exemples de liaison de coordination ligand-ion métallique.

Les résidus impliqués dans ce type d'interaction sont principalement l'aspartate, le glutamate, l'histidine et la cystéine, et les principaux ions métalliques rencontrés sont le zinc, fer, le calcium, le magnésium, le cuivre et le manganèse [38]. La distance d'une liaison de coordination est dépendante des groupements chimiques impliqués, et notamment du métal [39]. En pratique, une distance empirique maximale variant de 2,6 Å [28] à 2,8 Å [1] est généralement utilisée pour identifier ce type d'interaction.

De nombreuses protéines appartenant à la grande famille des métalloprotéines contiennent au moins un cation métallique dans leur structure. Les sites protéiques fixant ces cations ont divers rôles pouvant s'additionner, allant de la stabilisation de la structure tridimensionnelle de la protéine [40] jusqu'à une fonction catalytique dans la transformation substrat→produit pour la sous-famille des métalloenzymes. Dans ce dernier cas, le cation est lié à la protéine de manière à ce qu'au moins un site de coordination soit libre pour la fixation du substrat. Des cofacteurs peuvent également former des liaisons de coordination avec ces cations métalliques, comme par exemple les dérivés porphyriniques au sein de l'hémoglobine. Du fait de l'existence de plusieurs états d'oxydation pour certains cations métalliques, ces derniers sont fondamentaux en biologie, notamment pour réaliser les nombreuses réactions métaboliques de la classe oxydo-réduction.

Enfin, une stratégie efficace de design d'inhibiteurs de métalloprotéine consiste à concevoir un ligand qui forme une liaison de coordination avec un cation métallique du site de liaison [41] (voir la Figure

12B), entre autre *via* l'incorporation d'un groupement chélateur. Par exemple, de nombreux inhibiteurs de métalloenzymes à Zn^{2+} contiennent une fonction hydroxamate [42].

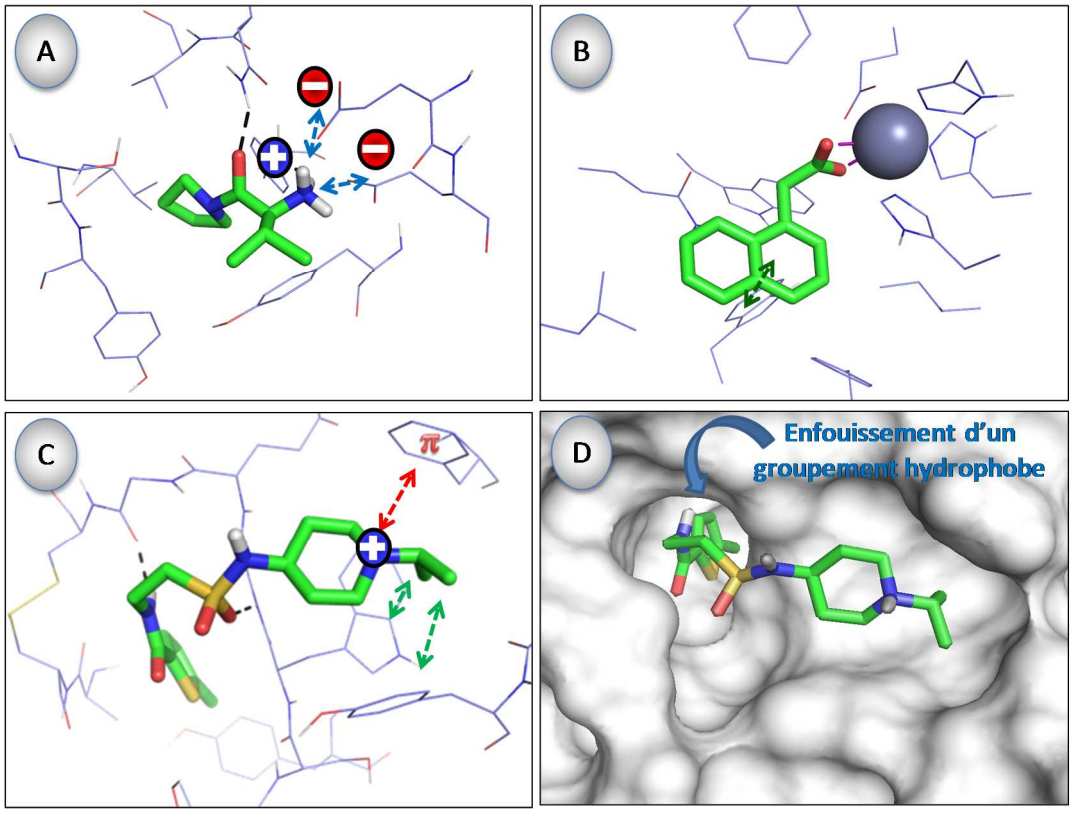


Figure 12: Illustration des principaux types d'interaction ligand-récepteur.

A) PDB 1N1M : mise en évidence d'interactions ioniques (flèches bleues) et de liaisons hydrogène (pointillés noirs). B) PDB 1LRH : illustration d'interactions ligand-métal (pointillés violets) et π-π (flèche vert foncé). C) PDB 4A7I : mise en évidence d'interactions cation-π (flèche rouge), de liaisons hydrogène (pointillés noirs) et d'interactions de VdW entre groupements hydrophobes (flèches vertes). D) PDB 4A7I : enfouissement d'un groupement hydrophobe dans une cavité hydrophobe du site.

1.3 Les notions de Structure-Based et Fragment-Based Drug Design

Les techniques de conception de molécules actives s'appuyant sur des données structurales (Structure-Based Drug Design - SBDD [43]) sont devenues fondamentales dans le monde pharmaceutique au sens large. Il s'agit d'un processus itératif et multidisciplinaire (voir la Figure 13).

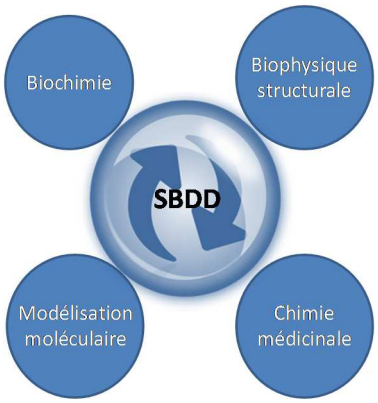


Figure 13 : Les disciplines impliquées dans le SBDD.

Diverses méthodes biophysiques, plus ou moins précises et possédant leurs propres avantages et inconvénients, sont utilisées pour avoir accès à une information structurale : de la microscopie électronique, à la résonance magnétique nucléaire (RMN), en passant par la résolution de structure par diffraction des rayons X sur des cristaux (X-ray). Par exemple, la RMN donne accès à une vision dynamique des molécules par l'analyse, entre autre, de phénomènes tels que le "Nuclear Overhauser Effect" (NOE). Celui-ci indique la proximité spatiale entre atomes du système, ce qui permet de construire des modèles structuraux de molécules en minimisant le nombre de violations de contraintes NOE. Cependant, il ne s'agit pas d'une photographie unique à un instant *t*, ce qui peut rendre plus difficile l'utilisation de telles structures en bioinformatique. A l'opposé, la X-ray donne accès à une structure unique mais très précise des molécules, dès lors que la résolution est suffisamment bonne.
Des banques de données, à l'échelle de la communauté scientifique mondiale, ont pour but de collecter toutes ces structures expérimentales. Les deux plus connues sont :

- la Cambridge Structural Database (CSD) qui se focalise sur les structures X-ray de petites molécules organiques et des complexes de faible taille comme les complexes organométalliques
- la Protein Data Bank (PDB) qui recense des structures X-ray et RMN de macromolécules biologiques (peptides, protéines et acides nucléiques) ainsi que leurs complexes

La PDB est une source d'information structurale fondamentale pour toutes les approches SBDD. Ces dernières ont entre autre pour but de déterminer le mode de liaison de ligands (potentiels ou connus) et de faciliter leur optimisation en suggérant des modifications sur la base de la structure 3D du complexe. De nombreuses méthodes de modélisation [44-46], parmi lesquelles le docking moléculaire et le *de novo* design (DND), reposent sur le concept de SBDD. Etant donné que ces deux dernières méthodes sont utilisées dans le cadre de cette thèse, elles seront également abordées dans ce large premier chapitre (voir les §1.6 et §1.7)

Le Fragment-Based Drug Design (FBDD), quant à lui, a bouleversé les méthodes de mise au point de nouvelles molécules actives dans le domaine de la recherche pharmaceutique. Il a pour but de produire des ligands de taille comparable aux candidats médicament usuels (“drug-like”) à partir de petites molécules organiques appelées fragments [47]. Une définition empirique, selon le même formalisme que les règles de Lipinski (Ro5) [48], énonce entre autre qu'un fragment possède une masse moléculaire inférieure à 300 Da [49]. Des seuils alternatifs, spécialement pour le critère masse moléculaire (ajout d'une borne inférieure et/ou diminution de la valeur supérieure), ont été proposés pour restreindre l'espace chimique des fragments [50, 51]. Ces derniers peuvent aussi être décrits comme des briques élémentaires, capables de se lier à une cible biologique, et qui sont incorporées dans des composés de taille et d'affinité plus importantes. Du fait de leur petite taille, ils possèdent généralement une faible affinité, située dans la gamme 10 µM - 10 mM, d'où la nécessité d'effectuer des criblages à forte concentration et d'utiliser des techniques biophysiques [52] très sensibles afin de pouvoir détecter la formation de complexes fragment-cible. Une cible manipulable par les méthodes biophysiques actuelles, ainsi qu'une librairie de fragments ayant une bonne solubilité et l'accès à des structures cristallographiques pour leur évolution sont néanmoins des prérequis pour pouvoir envisager l'utilisation d'une approche FBDD. Son principal défaut, implicitement lié à son concept même, réside dans la faible affinité de la plupart des hits. Ces derniers nécessitent donc une phase d'optimisation conséquente afin de déboucher sur un ligand de haute affinité. Certaines méthodes biophysiques (X-ray, RMN) donnent accès au mode de liaison du fragment à sa cible, et facilitent donc la phase d'optimisation en suggérant des modifications potentiellement intéressantes (voir SBDD ci-dessus).
Deux grands types d'optimisation permettent de transformer une touche (“hit”) en tête de série (“lead”) :

- optimisation par croissance (“growing”) : l'affinité du fragment initial est améliorée par ajout de groupements chimiques en vue d'améliorer les interactions avec la cible. Cette approche est similaire aux procédures classiques d'optimisation

- optimisation par liaison : il s'agit de lier deux fragments actifs et non compétitifs, de manière directe (“merging”) ou *via* l'introduction d'un groupement espaceur (“linking”). Cette démarche est plus complexe et nécessite deux fragments liés simultanément, mais peut aboutir à un gain d'affinité substantiel [53, 54] dans le cas idéal où le groupement espaceur permet de maintenir les fragments dans leur orientation initiale

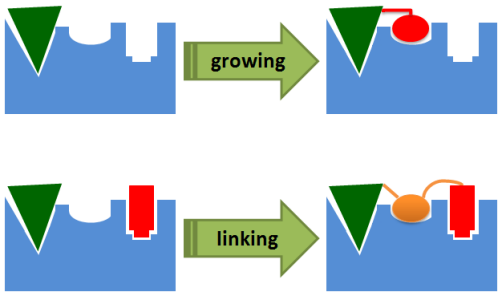


Figure 14: Exemples schématiques d'optimisation par growing et linking.

Le FBDD est devenu une alternative crédible au criblage à haut débit (High-Throughput Screening - HTS) dans l'identification de hits, tant dans le milieu académique qu'industriel. Il possède de nombreux atouts très intéressants, car complémentaires à ceux du HTS, parmi lesquels une meilleure exploration de l'espace chimique avec des chimiothèques de taille raisonnable (ordre de grandeur de 10^3-10^4 composés), une plus grande probabilité d'apparition d'un évènement de liaison, des interactions peu nombreuses mais de qualité (notion de “ligand efficiency” - LE [55]), et une plus large fenêtre d'optimisation tout en satisfaisant aux règles de Lipinski [48]. La notion de LE, définie comme l'énergie libre de liaison rapportée au nombre d’atomes lourds du ligand, est souvent plus importante pour des fragments que pour des ligands “drug-like” de référence, et ce malgré la relative faible affinité des fragments pour leur cible.

Les principales étapes du FBDD sont la conception de librairies de fragments, leur criblage expérimental, et leur nécessaire optimisation dans le but de générer des ligands de haute affinité (idéalement en conservant un bon LE tout au long du processus).

1.3.1 Revue dédiée au FBDD

Une revue [56] consacrée au FBDD, abordant à la fois les concepts clés, les approches expérimentales ainsi que les développements côté modélisation (chémo- et bioinformatique, *de novo* design), a été écrite dans le cadre de cette thèse. Elle est insérée dès la page suivante.

Fragment-Based Drug Design: Computational & Experimental State of the Art

Laurent Hoffer[1,2], Jean-Paul Renaud[*,2] and Dragos Horvath[*,1]

[1]*Laboratoire d'Infochimie, CNRS UMR7177, Université de Strasbourg, 4 rue Blaise Pascal, 67000 Strasbourg, France*

[2]*NovAliX, boulevard Sébastien Brant, 67400 Illkirch, France*

Abstract: Fragment-based screening is an emerging technology which is used as an alternative to high-throughput screening (HTS), and often in parallel. Fragment screening focuses on very small compounds. Because of their small size and simplicity, fragments exhibit a low to medium binding affinity (mM to μM) and must therefore be screened at high concentration in order to detect binding events. Since some issues are associated with high-concentration screening in biochemical assays, biophysical methods are generally employed in fragment screening campaigns. Moreover, these techniques are very sensitive and some of them can give precise information about the binding mode of fragments, which facilitates the mandatory hit-to-lead optimization. One of the main advantages of fragment-based screening is that fragment hits generally exhibit a strong binding with respect to their size, and their subsequent optimization should lead to compounds with better pharmacokinetic properties compared to molecules evolved from HTS hits. In other words, fragments are interesting starting points for drug discovery projects. Besides, the chemical space of low-complexity compounds is very limited in comparison to that of drug-like molecules, and thus easier to explore with a screening library of limited size. Furthermore, the "combinatorial explosion" effect ensures that the resulting combinations of interlinked binding fragments may cover a significant part of "drug-like" chemical space. In parallel to experimental screening, virtual screening techniques, dedicated to fragments or wider compounds, are gaining momentum in order to further reduce the number of compounds to test. This article is a review of the latest news in both experimental and *in silico* virtual screening in the fragment-based discovery field. Given the specificity of this journal, special attention will be given to fragment library design.

Keywords: Biophysical screening, computational chemistry, fragment-based drug design, fragment-based drug discovery, library design

INTRODUCTION

Nowadays, most drug discovery programs rely on a target-based approach. This kind of project begins with the identification and validation of a biological target (most often a receptor or enzyme) involved in a disease [1], the goal being to develop a ligand that is able to modulate the target activity, with acceptable pharmacokinetic properties and minimal side effects.

In large pharmaceutical companies, compound screening has progressively evolved from functional bioassays toward high-throughput screening (HTS) platforms. This evolution, initiated in the 90's, was the result of advances in combinatorial chemistry and improvements in both robotic automation and fast compound detection/analysis techniques. More recently, fragment-based drug discovery (FBDD) appeared as an alternative to HTS, and it has rapidly gained popularity in both academic and industrial worlds, fragment-based screening now being often conducted in parallel with HTS. Focusing on the detection of ligands of much smaller size than typical drugs, *i.e.* representing drug "fragments", which weakly bind to specific subsites of the target, FBDD aims at "recomposing" an eventually potent drug-like inhibitor by linking the various loosely binding fragments together, thus exploiting cooperative effects arising from the linkage. Besides, FBDD may provide starting points for challenging targets for which in-house libraries failed to produce any valuable hit.

Several biotechnology companies (Astex, Evotec, Plexxikon, SGX, Sunesis, Vernalis, Vertex), more or less dedicated to FBDD, emerged since the 1990s. Most molecule suppliers have enlarged their compound collections in response to the rise of FBDD; for instance, they now propose some "Rule of 3-compliant" (see below) collections of compounds.

What are "Fragments"?

Fragments, in the context of FBDD, are small compounds typically the size of the main moieties encountered in drugs, but some authors prefer a more specific definition, imposing constraints similar to Lipinski's rule defining "drug-likeness", also known as "Rule of 5" (Ro5) [2]. Thus the most widely accepted definition for fragments has been proposed by Congreve and co-workers [3]. According to their "Rule of 3" (Ro3), a fragment should have a molecular mass (MM) < 300 Da, a calculated logP ≤ 3, a number of hydrogen bond donors ≤ 3 and a number of hydrogen bond acceptors ≤ 3. They proposed some additional filters based on the number of rotatable bonds (NROT ≤ 3) and the polar surface area (PSA ≤ 60 Å^2). Some

*Address correspondence to these authors at the Laboratoire d'Infochimie, CNRS UMR7177, Université de Strasbourg, 4 rue Blaise Pascal, 67000 Strasbourg, France; Tel: +33 368 851 321; Fax: +33 368 851 589;
E-mail: horvath@chimie.u-strasbg.fr or NovAliX, boulevard Sébastien Brant, 67400 Illkirch, France; Tel: +33 368 330 200; Fax: +33 368 330 201;
E-mail : jpr@novalix-pharma.com

1386-2073/11 $58.00+.00

modifications of this rule have been suggested [4], particularly with regard to the molecular mass range [5]. For example, Card *et al*. screened a library of compounds with molecular mass between 125 and 350 Da to identify phosphodiesterase-4 hits [6]. This high upper limit has already been proposed by Teague *et al*. in the design of lead-like collections [7]. Another approach to roughly estimate the size of compounds is the use of a fuzzy metric like the number of heavy atoms. Because of their limited complexity, fragments generally bind to a target with low to medium affinity (from mM to µM). Thus they must be screened at high concentration in order to detect a binding event, and consequently they must have a high aqueous solubility. Therefore, measured or predicted solubility (logS) can be a useful filter during the building of fragment libraries. However, high-concentration screening often leads to false positives [8]. Sensitive biophysical methods are therefore often preferred to monitor binding events with fragment-like compounds. Among these methods, X-ray crystallography, NMR, surface plasmon resonance (SPR) and mass spectrometry (MS) are the most widely used. X-ray crystallography and NMR have been intensively used in FBDD projects because they are both very sensitive and can give access to structural information about the fragment binding mode.

Unlike the "drug-likeness" zone of chemical space, which is delimited by pharmacokinetically imposed constraints (solubility, permeability…) the "fragment" chemical space is an even fuzzier concept, for it is not the fragments *per se* but rather the resulting drug candidates that have to comply with these limitations. The choice of fragment size and nature – the compromise between size, solubility and binding detection limit – is rather dependent on the experimental methods used in FBDD. Various methods may allow for specific optimal scenarios: the same drug may be "built" from many small or from fewer large fragments. Having in mind that drug-likeness itself is a highly debatable topic, with many important drugs largely violating Lipinski's rule, fragment-likeness may as well be a tentative over-specification of the parameters of an open-ended research strategy.

The FBDD Strategy

FBDD relies on the screening of a small library of fragment size compounds in order to detect low-affinity binders. By contrast, in traditional high-throughput screening, as many molecules as possible are tested in order to find relatively potent compounds. The central idea in FBDD is to start the optimization process with low-complexity compounds. Fragment screening is more likely to identify compounds that can fit into the binding site. Binding fragments are then evolved to larger molecules with higher affinity/activity, this optimization step being facilitated when structural information is available. In 1981, Jencks claimed that complex compounds (e.g. molecule A-B) can be considered as the combination of several fragments (e.g. fragments A & B) which enable the binding event [9]. In this work, he considered the free energy of binding as the sum of contributions ("intrinsic binding energies") of fragments, modulated by a correction term which takes into account changes in translational and rotational entropy (due to the linking of moieties A and B). About 15 years later, Shuker and co-workers published the first practical example of FBDD called "SAR by NMR" [10].

There are three essential elements in the FBDD pipeline: a library of fragments, a method to screen them and to identify weak binders, and a strategy to evolve these low-affinity simple molecules into leads. And several challenges arise at each step, such as:

Designing the appropriate fragment library (e.g. in terms of diversity, chemical tractability, solubility...) for screening. HTS customarily relies on several types of compound libraries: standard "diverse" drug-like libraries for a wide range of targets, focused libraries for a specific target/family of targets, and combinatorial libraries, each type being used in a given context. For example, focused libraries are designed later in the discovery process, when an important amount of Structure-Activity Relationship (SAR) data is available for the target. These libraries often exhibit a higher hit rate compared to standard "diverse" libraries, but may not lead to "paradigm-breaking" discovery of new hits, because they were designed by extrapolation of known SAR information. By contrast, FBDD fragment libraries are created to be used with a wide range of targets, each putatively displaying a diverse set of sub-pockets to bind to. An efficient fragment library must, in any case, be able to cover a maximal chemical subspace, knowing that FBDD is typically used for "pioneering" exploration of the SAR landscape rather than for extrapolation of already known chemical space areas. Nevertheless, some focused fragment libraries have been built for specific use [11].

Choosing the right fragments to optimize when several hits are detected. The ligand efficiency (LE) concept, described later, tries to address this goal.

Finding the best way to evolve fragments into high-affinity ligands. Fragment optimization resembles traditional medicinal chemistry optimization, in which several moieties are replaced or added to a hit to improve both affinity and pharmacokinetic properties. There are three main strategies which can be used for evolving fragments into leads: merging, linking, and growing. Fragment merging and linking consist in the assembly of two non-competitive fragments, respectively directly or *via* a spacer called linker. The other approach, growing, consists in adding chemical groups to the initial fragment in order to increase its affinity by making additional interactions with the target. Growing has been the most successful strategy of fragment evolution since it is very intuitive and relatively reliable when a 3D structure of the [target/fragment] complex is available. The linking strategy is less used and is more challenging than growing, since it requires at least two non-competitive fragments, and both the orientation and the location of the fragments must be conserved in the final ligand after linking. Another approach, called "SAR by catalog" by Schulz and Hubbard in a review [12], consists in searching compounds from commercial databases which are similar to (or contain) the desired fragment, and to test these similar compounds. A requirement for FBDD success is that fragments act as anchors and are not forced to change their binding mode during the process of evolution towards the lead compound. Some FBDD success stories demonstrated that fragments can be optimized into potent ligands while maintaining the

binding mode of the initial hit. Counterexamples were also reported, such as the study of Babaoglu and co-workers [13]. They broke a beta-lactamase inhibitor into several fragments and determined their experimental binding modes by X-ray crystallography. The fragments did not bind to the target with the expected binding mode, as observed in the full ligand. Using fragments obtained by deconstructing a series of Bcl-x_L inhibitors and NMR as screening tool, Barelier *et al*. showed that these fragments could exhibit no binding to the target, or bind to the target either in a similar manner as in the starting inhibitor or at different sites [14].

FBDD has three main advantages that will be discussed in more detail later:

The chemical space can be more efficiently probed by screening small libraries of fragments (exploiting the "combinatorial advantage" – testing N fragments can generate as much information as testing the huge number of compounds resulting from possible combinations of these fragments). Besides, probing a target site with fragments may be intrinsically interesting for evaluating its "druggability" [15]. Based on 36 various drug discovery projects from AstraZeneca, Edfeld reported the use of fragment screening as an efficient way to estimate the druggability of a target [16]. A score (from high to low) is assigned based on the number of hits, affinity (if known) and diversity of hits. A good correlation was found between a high score and the probability of success for both HTS and hit-to-lead strategies. Finally this score, resulting only from a fragment screening, is of particular interest to prioritize novel targets in order to reduce the risk of failure (i.e. useless target from a drug discovery point of view).

Higher hit rates, since the probability of a binding event reduces dramatically with the complexity of the compound. Hann *et al*. studied with a simple mathematical model the probability of finding leads in relation with molecular complexity [17]. In their simplified model, the probability of actually measuring a binding event with an experimental method for a ligand that does bind increases with ligand complexity. But the probability of finding for a given compound a "useful event", defined as being both measurable and unique in its binding mode, reaches a maximum at intermediate complexity. Indeed small compounds may bind in several ways, as they may fit various subsites. On the other hand, it is more likely that a large molecule will have a unique binding mode since its global shape must fit the binding pocket. Conversely, the probability of binding goes down with increasing compound size, since bigger molecules contain more putative mismatch groups (steric, hydrogen-bond or ionic mismatches) with respect to the binding site. The low hit rate observed in HTS campaigns, where lead-like or drug-like compounds are screened, can be explained with this empirical model. Authors insist that a tradeoff must be found between small and complex compounds. In the context of FBDD, this tradeoff could be "screening fragment-like compounds with sensitive biophysical methods to maximize the binding event probability". In practice, a threshold can be applied to discard the smallest compounds (e.g. 100 or 150 Da) when building a fragment library in order to decrease the probability of multiple binding modes.

High binding efficiency per heavy atom. A theoretical advantage of FBDD is that compounds optimized from fragment hits should exhibit a high affinity while having a smaller size compared to compounds optimized from HTS hits. Indeed each part of the ligand should be close to optimal with respect to the binding site of the target. Thus, an emergent molecular mass window is available to perform subsequent optimization of other fundamental factors such as selectivity and pharmacokinetic properties. It should be noted that fragment-like compounds with an affinity comparable to or even better than that of drug-like compounds was already reported long time ago, for example the phenylethanolamine N-methyltransferase inhibitor SK&F64139 [18].

FBDD *vs* Classical HTS

Although HTS has identified many valuable hits for a wide range of targets [19], screening drug-sized compounds has shown its limitations. Indeed, the hit rate is often very low and many hits fail to reach optimization stages. Despite huge investments in HTS by large pharmaceutical companies, massive screening approaches did not yield the expected results [20] since the number of new drugs approved by the FDA has not increased these last years [21]. Optimization of HTS hits can be tricky because the simultaneous binding of several moieties can only be suboptimal; the direct consequence is that many HTS hits cannot be correctly optimized and lead to intractable problems. Typically, medicinal chemistry optimization generates larger and more hydrophobic compounds because lipophilic groups are often incorporated to increase affinity [2]. However poor pharmacokinetic properties are correlated with both high molecular mass and high hydrophobicity. Because of their larger size, HTS hits cannot be widely optimized while being "Ro5-compliant" (guidelines developed by Lipinski [2] to assess compound compatibility with oral administration). On the other hand, FBDD starts with low-complexity compounds, thus a larger "Lipinski-compliant" optimization window is available. Several examples in the literature demonstrate that FBDD can succeed where HTS has clearly failed - for example, in finding a non-peptidic lead for the aspartic protease BACE-1 [4].

However, HTS is still a method of choice for many targets of interest - notably membrane proteins like GPCRs and channels. Fragment binding events in these complex systems are intrinsically difficult to detect. Furthermore, membrane proteins cannot be produced in large quantities, and obtaining quality crystals that diffract enough to yield high-resolution structures remains challenging even though recent progress is encouraging [22].

FBDD *vs* Diversity-Oriented Synthesis (DOS)

A debate, between Hajduk on one side and Galloway and Spring on the other side, about the best strategy to use for drug discovery has been recently reported [23]. The former defends FBDD and the latter DOS. In one word, the goal of the DOS strategy is to synthesize lead-like/drug-like compounds (by definition larger than fragments) without the typical drawback (poor diversity) of combinatorial

chemistry-based libraries. Thus, one can expect that hits from DOS will exhibit a higher affinity than fragment hits. However, larger compounds lead to a much larger chemical space – and thus more difficult to cover – and a lower probability of binding event (see the description of the Hann model in the FBDD strategy part).

Exploiting Experimental Structures for Drug Design

X-ray crystallography and NMR provide experimental structures giving access to the detailed (atomic-scale) description of the binding mode(s) of compounds within the binding site. Through structure-based drug design (SBDD), one can rapidly progress from weak-affinity compounds toward lead-like molecules. Indeed, the structure of the binding site gives clues to pick up/build additional interactions in order to increase ligand affinity. In the context of FBDD, most published examples of fragment evolution have made use of SBDD approaches based on X-ray structures of target-ligand complexes. Murray and Blundell wrote a review dedicated to the crucial importance of structural biology in the FBDD field [24]. When the experimental structure of the target is available but not that of a target-ligand complex of interest, computational SBDD applications are used to model the ligand binding mode(s) by means of docking calculations, which will be reviewed in detail later.

Fragment Library Design

The first step in FBDD consists in creating fragment libraries. Several approaches have been proposed for that purpose. Since chemoinformatics methods are often used in this context, the design of fragment libraries will be described into the computational part. Most of these strategies can be used in the construction of both general and focused fragment libraries for a specific target or class of targets. The last type is often built by deconstructing a set of known binders for a given target [11b].

Chemical Space Navigation in FBDD

The number of potentially HTS-relevant synthesizable compounds of maximum 30 heavy atoms among most common elements such as C/N/O/S has been roughly estimated to more than 10^{60} [25]. The largest databases "only" contain several millions of compounds; so the chemical space cannot be properly sampled even in the case of a massive HTS screening. By contrast, Fink *et al*. computationally explored the FBDD-relevant chemical space of compounds up to 160 Da and concluded that it contains approximately $1.4x10^7$ compounds [26]. Clearly, a sample of 10^3 well-chosen fragments may stand a fair chance to be a statistically representative sample of these 10^7, whereas a classical HTS campaign of 10^6 molecules could never claim to be anywhere near to "represent" 10^{60}. The 10^{60} actually represent all the possible chemical combinations of the considered fragments, and actives amongst them would typically have at least some of their constituent fragments binding as they would as stand-alone molecules. Knowledge of the "stand-alone" binding modes may therefore significantly speed up the discovery of the rare actives, hence fragment-based screening, unlike HTS, implicitly provides a significant coverage of drug-like chemical space - or at least allows to "prune" large parts of uninteresting domains. There is a theoretical (but assumedly feeble) probability of existence of an active compound in which all its moieties lose their "native" most favorable conformation as stand-alone compounds in order to adopt the sub-optimal binding schemes ensuring a maximal overall binding free energy. Such a molecule could not be discovered by FBDD. This notwithstanding, the "combinatorial advantage" achieved by checking a limited number of fragments for their relative binding propensities, then postulating that the drug-like actives should be searched amongst the chemical combinations of strong fragment binders allows narrowing the virtually infinite drug-like chemical space by many orders of magnitude down to a tractable subset of compounds.

Thermodynamic Aspects of Ligand Binding

Molecular recognition obeys the laws of thermodynamics. A binding event involves crossing an entropic barrier due to, among others, the loss of rigid body rotational and translational entropy [27]. Murray *et al.* studied this phenomenon and concluded that the loss of rigid-body entropy barrier to binding is essentially independent of molecular mass [28]. Therefore, drug-like compounds which can pick up many interactions can more easily get over this unfavorable barrier compared to fragments, which are only able to participate in a limited number of interactions because of their size and must thus make optimal contacts to overcome this entropic penalty.

The knowledge of the free energy of binding (or the dissociation constant K_d of the complex) is of particular interest during the optimization stages. These experimental values can be approximated, among others, by several biophysical methods widely used in the FBDD field (mostly NMR and SPR), but only ITC can determine both components (enthalpy and entropy) of the free energy of binding. Freire proposed a method, called thermodynamic optimization plot, to represent both components (ΔH and $-T\Delta S$) on a two-dimensional chart [29]. Thermodynamic data resulting from different analogs can be displayed on this plot in order to determine, by an intuitive and visual analysis, the impact of each modification on enthalpy and entropy. He claimed that this approach should facilitate the understanding of thermodynamic data, and thus accelerate the affinity optimization of ligands. From a computational point of view, estimating the free energy of binding remains challenging although numerous methods have been developed to try to address this issue [30], but this subject is beyond the scope of this review.

Fragment Linking or Merging

The empirical, intuitive (albeit not very rigorous from a fundamental, statistical thermodynamics point of view) idea of a fragment-based decomposition of binding free energies has thus been validated: the optimization of ligand-receptor interactions for each fragment and the efficient incorporation of all these fragments into a single compound should produce a ligand with a higher affinity than the sum of affinities of the initial molecules. For a two-fragment A-B molecule, the "linking coefficient" E relates the dissociation

constant of A-B and the product of those of the separate entities A and B ($K_D^{AB} = E.K_D^A.K_D^B$) [27] ; in the case of an efficient linking, this coefficient can be highly favorable (i.e. << 1) (the linker keeps the fragments "frozen" in their relative positions when independently binding to the site - hence a purely beneficial entropic term). However, unfavorable linking coefficients are more likely: the entropic linking benefit can be easily lost because of constraints due to the linker (no longer allowing for an optimal relative positioning of the fragment), or suboptimal linker-binding site interactions. Nevertheless, several examples of efficient linking/merging have been published, such as the discovery by Borsi *et al.* of a highly potent MMP-12 ligand (affinity in nM range) resulting from the direct merging of two fragments of affinity in the mM range [27] ; Röhrig *et al.* studied the effect of both the nature and length of the linker on the affinity of modified ligands for FKBP [31] ; and Barker *et al.* synthesized a linked compound that was 1000 time more potent against Hsp90 than both starting fragments taken separately [32].

Ligand Efficiency

Kuntz and co-workers showed that the free energy of binding increases linearly with molecular mass up to roughly 15 heavy atoms, beyond which it grows only very little with mass [33]. Within this range, the free energy contribution per heavy atom is about -1.5 kcal/mol. This information is fundamental because it means that small compounds such as fragments can be more efficiently optimized in the first steps of evolution than larger hits. The ligand efficiency (LE) concept, which relates affinity to molecular size, has been proposed by Hopkins *et al.* [34] on the basis of this interesting discovery. More precisely, LE is defined as the free energy of binding divided by the number of non-hydrogen atoms. Strictly speaking, only the dissociation constant K_d (directly linked to the free energy of binding) should be used to calculate this criterion, but authors found that IC_{50} values can be used to estimate this coefficient. The ligand efficiency is widely used in the FBDD field since it allows not only to select the most interesting hits from a screening among a population of binders but also to monitor the quality of the optimization process.

The threshold of 0.3 kcal/mol/heavy atom is commonly used as the lower acceptable value for LE. This value represents the LE for a hypothetical reference compound with high affinity (10 nM) and a molecular mass of about 500 Da (upper limit in Ro5). Thus this threshold is used as a guideline to develop "Ro5-compliant" compounds.

The main drawback of LE is that it does not have the same behavior with ligands of different sizes. Indeed this coefficient is very sensitive for small compounds but relatively insensitive for larger compounds. To assess this problem, Reynolds *et al.* proposed a new metric called fit quality (FQ), computed by dividing LE by a scaling factor (called LE_Scale), which is determined by fitting the top ligand efficiency versus heavy atom count curve to an exponential function [35]. Other groups designed alternative criteria including the ligand-lipophilicity efficiency (LLE) [36] which subtracts a lipophilicity component (logP) to pIC_{50} values, the binding efficiency index (BEI) obtained by dividing the pIC_{50} (or pK_i, pK_d) by the MM in kDa [37], and the group efficiency (GE) [38] which estimates the binding energy contribution of an added functional group by dividing the change in binding energy by the change in the number of non-hydrogen atoms. The last criterion requires usual SAR data (binding affinity for a series of compounds differing each by one functional group) and has been applied to protein kinase B inhibitors [39]. By deconstructing a set of highly optimized inhibitors into minimal binding elements, Hajduk showed that an efficient optimization can strongly increase potency without sacrificing the BEI [40]. In other words, there could be a roughly linear correlation between affinity and molecular mass in an ideal optimization case.

Many reviews [8, 11b, 24, 41] and books [42] dedicated to FBDD have been published lately. An important number of FBDD success stories have been reported; several previously cited reviews analyze in detail some of them, notably when structural information and affinity/potency data are both available. The FBDD approach has been validated since several compounds discovered with this method have entered clinical trials, but to date there is no marketed drug derived from FBDD. Two reviews are dedicated to both potential and confirmed clinical candidates derived from fragment hits [41e, 43]. The present review will attempt to summarize both the computational and the experimental faces of FBDD, by addressing in more detail the key points that have been highlighted previously.

In Silico FBDD

Computational chemistry plays an important role at each step of FBDD projects. The creation of fragment libraries, the selection of fragments to be tested experimentally, and the evolution of fragments into potent compounds are stages where molecular modeling can be helpful. Some reviews summarize several uses of computational chemistry within the FBDD field [41f, 44]. Computational methods are often used at an early stage by performing "virtual screening" (VS). Their goal is to both decrease the size of the compound sets to be experimentally tested and increase the hit rate. Common virtual screening approaches used for drug-like molecules can be extended to fragment-like compounds. Finally, several tools dedicated to FBDD or adapted to this purpose have been developed these last years. The main computational approaches employed in FBDD will be briefly described, with several examples to illustrate these algorithms/strategies.

Fragment Library Design

There is globally a consensus about what constitutes a "good fragment library" and this topic has been extensively discussed [45]. Some of its properties are shared with standard screening libraries, while others are specific. The main factors are fragment physicochemical properties (molecular mass, logP, aqueous solubility), molecular diversity, synthetic accessibility for optimization, filtering of unwanted chemical groups, incorporation of recurrent scaffolds present in natural compounds or known drugs, and easy maintenance. A quick way to build a fragment library is to filter commercial databases of fragments with the above-mentioned filters.

It should be noted that the size of a fragment library is function of the biophysical technique used for screening [46]. Libraries that are screened by biophysical methods such as X-ray crystallography or NMR only contain 10^2-10^3 compounds compared to HTS libraries of 10^5-10^6 molecules. Recently, a large fragment library (10000 compounds) has been screened by NMR [47]. The largest fragment collection appears to be that of Graffinity (24000 compounds), which is screened by high-throughput SPR imaging [48] (see Experimental FBDD section).

Diversity-based libraries have the advantage that their range of applicability is much wider than that of biased libraries. In cases where there is sufficient information available about the target, biased/focused libraries are expected to yield a higher number of hits, however at the expense of the chemical diversity among the found hits.

The computational deconstruction of molecules into fragments by breaking bonds is widely used to generate fragment-like compounds [49]. These methods differ mainly in the dictionary of bonds to break.

Diversity-Driven Fragment Selection

As mentioned above, one major theoretical advantage of fragment-based screening with respect to HTS is the ability to cover a larger chemical space while screening a smaller number of compounds. To achieve this goal, the fragment library must be cleverly built such as to maximize the diversity of compounds. In other words, the relatively little number of compounds in the library must cover a significant part of the fragment chemical space.

Diversity enhancement algorithms can be applied to efficiently sample fragment chemical space while keeping only several hundreds or thousands compounds. These methods were developed outside the FBDD field but can nevertheless be applied to the design of fragment libraries – if appropriate care is taken to ensure that the originally employed molecular descriptors, tailored for drug-like molecules, continue to be relevant for fragment-sized compounds. For example, pharmacophore point triplets [50] of sizes typically encountered in drugs will generate very sparse signatures with fragment collections – they will have to be redefined on the basis of shorter inter-feature distance ranges, or skipped altogether in favor of less complex descriptors (pharmacophore pair counts).

Diversity methods typically rely on some compound clustering techniques, and then pick representatives at the core of each distinct cluster to form the subset of diverse molecules. Clustering may be typically similarity- or grid-based [51]. They will be briefly described below.

The first type, namely similarity-based clustering, uses a similarity metric to calculate "distances" (dissimilarity scores) between molecules represented as points in some N-dimensional Descriptor Space (DS) in which every axis represents a molecular descriptor. Clustering techniques and/or Kohonen neural networks [52] are then used to collect neighbor compounds into clusters. A representative subset of diverse molecules is obtained by choosing one compound from each cluster. Conventionally, molecules structures are encoded [53] as numeric or binary vectors (fingerprints), with an associated metric [54] adapted to the nature of these vectors.

Grid-based approaches somehow arbitrarily cut the DS into parallelepipedic cells, by making an arbitrary number of cuts along each of the N descriptor axes. A representative subset of diverse molecules is obtained by choosing one compound from each populated cell [51]. It offers the advantage of not requiring any calculation of pairwise distances, but does not formally guarantee that the resulting set is void of redundancies (very similar molecules may still be simultaneously selected if they happen to fall nevertheless into different cells).

Solubility-Assessed Fragment Libraries

Solubility is usually expressed as logS, where S is the saturating compound concentration in mol/L at thermodynamic equilibrium. The importance of this property has already been highlighted. Indeed, a high aqueous solubility of fragments is needed, since they need to be screened at high concentration. There are two ways to achieve this. The first is to experimentally measure the aqueous solubility of preselected compounds, and to subsequently filter them on the basis of the measured values of solubility. The determination of aqueous solubility in a high-throughput fashion has been reported [55]. A second approach consists in the prediction of the aqueous solubility by chemoinformatics methods, for example by building a QSPR model. This kind of technique will be described into a specific part dedicated to QSAR/QSPR modeling. Baurin *et al*. used a QSPR model as a pre-filter to select putative high aqueous solubility compounds, then experimentally measured this property for all selected compounds before integrating them into the final version of their fragment library [11c]. They confirmed the usefulness of their solubility model by comparing the predicted and measured solubility values from the initial fragment library: 88% fragments that passed the quality control were correctly predicted to have solubility $\geq$2mM and 88% fragments that failed this test were correctly predicted as insoluble at 2mM.

Synthetic Accessibility

The importance of the synthetic accessibility has also been previously highlighted. Since fragments are often weak binders, they must mandatorily be optimized into more potent compounds. Thus, all fragments present in a library must contain moieties that can be easily evolved. Several strategies to estimate or ensure the synthetic accessibility have been developed [56], but medicinal chemists are invited to validate the final selection anyway.

Lewell and coworkers developed the Retrosynthetic Combinatorial Analysis Procedure (RECAP) approach [49a]. It relies on the fragmentation of molecules around specific bond types, whose end atoms are flagged to take into account their previous chemical environment. These bond types result from common chemical reaction schemes; hence the retrosynthetic term. Based on atom flags and reaction schemes, a combinatorial library can then be created by merging the obtained fragments to build all synthetic accessible compounds. An alternative use of the method is to analyze common fragments present in a database. Since this method provides labeled fragments which can be evolved by

known chemistry, the RECAP approach can be used for both combinatorial chemistry and fragment library design. The methodology leads to the nowadays popular concept of "privileged structures", based on the observation that some fragments are, statistically speaking, over-represented for a given biological activity. Moreover, it is assumed that compounds that are similar to marketed drugs have a lower risk of failure in clinical tests ; thus fragments most often occurring within compounds in the WDI database [57] can be safely added to a fragment library.

Published Designs of Fragment Libraries

In 2004, Baurin *et al*. described the steps to create fragment libraries to be used in their SeeDs (Selection of experimentally exploitable Drug start points) protocol involving NMR screening [11c]. They built four distinct fragment libraries including a kinase-focused library by applying various filters (elements, molecular mass, wanted/unwanted functionalities) and using clustering and pharmacophore approaches. A QSPR model, estimating aqueous solubility of fragments, was used to discard potentially insoluble compounds. The solubility of each compound was nevertheless experimentally checked (threshold set to 2 mM) before incorporating fragments in the final collection. Chen and Hubbard recently reported some new aspects in their fragment library design protocol, and various analyses of their different versions of in-house fragment libraries [58]. They showed that the chemical complexity, measured by pharmacophoric triangles, has increased between two versions of their fragment library (2008 versus 2004) without increasing the average molecular mass.

Hajduk *et al.* proposed enriching fragment libraries with some scaffolds, among others biphenyl and diphenylmethane substructures [59]. For example, the biphenyl moiety, known to often bind to protein pockets of distinct classes [60], is rigid - hence entropically less hindered to bind. However, excessive use of aromatic fragments is not recommended, as these may lead to poorly soluble and/or DNA/RNA-intercalating compounds - two recent papers evidenced a correlation between the fraction of sp^3 carbons and protein-binding frequency and selectivity [61] as well as the likelihood that a compound makes a successful transition from lead to drug [62].

Scientists from Vertex developed the SHAPES strategy, which uses NMR as screening method [63]. By computationally breaking known drugs into rings, linkers and side chains, they showed that most drugs from the Comprehensive Medicinal Chemistry (CMC) database can be represented by a relatively small number of scaffolds. A substructure search was performed in the Available Chemicals Directory (ACD) with these scaffolds as queries to build a small fragment library (the SHAPES library), followed by a validation step on the found compounds focusing on some fundamental properties such as synthetic accessibility, aqueous solubility, diversity and a factor related to their screening approach (separation of NMR peaks).

Kolb *et al.* have developed the DAIM (Decomposition And Identification of Molecules) program, which breaks compounds into rigid fragments. Since this strategy is used in a broader workflow, the DAIM process will be described below in a part dedicated to the Caflisch software suite.

Gianti and Sartori proposed a workflow, integrated in the Pipeline Pilot package, that builds collections of "privileged fragments" [64]. They generate virtual privileged fragments by breaking compounds from drug or drug-like databases. Their breaking scheme leads to "decoration spheres" which are fuzzy substructures with variable length side chains. Several common filters, such as the Ro3 or a solubility threshold, are applied to the resulting fragments. They are then clustered to avoid redundancies, and are used as query inputs toward commercial database catalogs to build the final library.

Blomberg and co-workers described a series of tools developed at AstraZeneca several years ago that were used to build two distinct fragment libraries [65]. Among them, the Foyfi program generates hashed molecular fingerprints that are used to compute similarity between two compounds. The BigPicker program is dedicated to the design of compound cocktails in a way that tries to maximize internal diversity. Initially, compounds are randomly spread into groups and a score is computed for each subset. This score is the shortest Tanimoto distance between any two fingerprints of compounds in the group. Since the goal is to create cocktails with maximal diversity, the algorithm attempts to maximize the subset scores. A Monte Carlo strategy, based on the perturbation of randomly chosen subsets by swapping compounds that have a close neighbor in the same subset, is used to quickly find an acceptable solution. NMR- or X-ray-based screenings usually employ cocktails of compounds to speed up the process, and diverse mixtures are desired to avoid problems in hit identification ; thus BigPicker is particularly useful for these screening approaches.

Davies *et al.* created the "Fragments of Life" (FOL) library with an unusual but interesting approach (biology-oriented) [5]. They incorporated different types of fragments into their library. First, small compounds known to be produced by living organism were stored in the FOL-Nat subset. Known derivatives and bioisosteres of these molecules were stored in a second subset (FOL-NatD). Finally they added synthetic biaryl molecules, *a priori* closer to common fragments, to increase the diversity in a third subset (FOL-Biaryl). The merging of these three subsets lead to the full FOL library (1329 compounds), which is furthermore "Ro3-compliant". Besides, the methanol solubility has been checked for each fragment at a high threshold (≥50mM) to avoid solubility issues with their X-ray-based fragment screening strategy.

Finally, Schuffenhauer *et al*. from Novartis proposed an interesting idea for fragment library design [45c]. Their library is in reality composed of two subsets: a screening library containing compounds incorporating a linker group in a masked manner, and a synthesis library containing the corresponding unmasked compounds. Thus, hits can be optimized without loss of their initial key interactions. The synthesis library, composed of building blocks, facilitates the evolution of fragments into larger compounds.

Ligand-Based Drug Design (LBDD)

LBDD consists of modeling approaches that rely on the knowledge of a set of molecules that bind to the target of interest. LBDD approaches include similarity searching with

fingerprint-based or pharmacophore techniques, and QSAR models (quantitative structure-activity relationships). This kind of computational methods can be an alternative strategy for non-orphan targets for which structural information is not available. Thus LBDD is widely used with many pharmaceutically relevant targets, as for example the GPCRs [66]. The main LBDD strategies that were successfully employed in FBDD projects will be briefly described below.

Similarity-Based Methods

These approaches are based on the widely accepted similarity principle : "compounds with similar chemical structures usually possess similar physicochemical properties and biological activities". The similarity of chemical structures is computed with a chosen metric in the considered DS – as in above-mentioned dissimilarity selection techniques. However, the aim of the approach is now diametrically opposed : the search for analogs of validated binding fragments, keeping in mind that in the most interesting success stories the chemical "similarity" only concerns the key features controlling the binding, not the entire chemical structure. In particular, discovery of "bioisosteric" fragments featuring a conserved pharmacophore pattern – and hence similar affinities – but differing in terms of chemical composition and connectivity ("scaffold" hopping) is of paramount importance to the field. This kind of method is particularly useful in the "SAR by catalog" approach (briefly described above).

Pharmacophore Modeling

Central to the above-mentioned "scaffold hopping" goal, the "pharmacophore" is defined as an ensemble of steric and electronic features that is necessary to ensure the optimal interactions with a specific biological target ; the concept is widely discussed in a recent paper [67]. One way to use information from several active compounds is to build a consensus pharmacophore which contains favorable moieties for binding that can be classified into several families such as hydrophobic, aromatic, hydrogen donors, hydrogen acceptors, cationic and anionic. Some exclusion spheres can be added to roughly mimic binding site constraints. This ensemble of features is then used to screen databases looking for matching compounds. In the FBDD context, pharmacophoric approaches have been essentially used in the design of libraries, for example to enrich them with chemotypes of interest [68] or to analyze their complexity/diversity [11c, 58].

QSAR/QSPR

QSAR (quantitative structure-activity relationships) and QSPR (quantitative structure-property relationships) methods are widely used in chemoinformatics. They extract knowledge from already measured properties, put in relation to the structural elements (in particular, the associated molecular descriptors) of the respective compounds. The extracted information is "condensed" under the form of *in silico* models (equations estimating the property as a function of selected descriptors) which may thus predict the activity/property of new putative analogs, or simply distinguish likely active from probably inactive compounds (classification models). LogP [69] and aqueous solubility [70] are broadly predicted properties [71] because of their importance in the drug design field.

In FBDD, the most interesting use of this kind of approach is the prediction of aqueous solubility of molecules. Model applicability [72] is another key issue : must predictive models for fragment-sized compounds be trained on hand of specific "fragment" sets, or do "classical" models fitted against arbitrary compound collections cover the specific "fragment" zone of chemical space well enough ? Several reviews are dedicated to the *in silico* solubility predictions of molecules [73]. For example, Lamanna *et al*. employed a recursive partitioning decision tree learning method to build solubility models [74]. They used an in-house database of 3563 compounds with known solubility, which was subsequently split into a training set (2363 compounds) and a test set (1200 molecules). Only intuitive descriptors (understandable from a chemist's point of view) were used : molecular mass, the calculated octanol/water partition coefficient AlogP, polar surface area, number of rotatable bonds, number of hydrogen bond donors/acceptors and aromatic proportion (AP). The last one is defined as the number of aromatic atoms divided by to the total number of heavy atoms in the molecule. Their best model, showing an accuracy of 81% and a precision of 75% for the test set, was only based on two descriptors (MM & AP). Finally, their model was able to discard most of the known insoluble compounds from the test set.

An interesting attempt to bridge the gap between the fragment level and the properties of the final drug candidates is the Virtual Fragment Linking (VFL) method. It relies on an experimental fragment screening campaign, followed by the building of statistical Bayesian models to roughly predict activities of drug-like compounds containing one or more fragment-like binders [75]. The goal is to prioritize compounds from a HTS library by using information from a fragment screening. These approaches lead to interesting results for tricky targets such as GPCRs, for which structural data are globally unavailable.

Manoharan *et al*. employed QSAR methods in combination with a multi-objective strategy to perform the virtual optimization of an interesting scaffold [76]. They built QSAR models to predict both affinity (BACE-1 as target) and selectivity (BACE-1/BACE-2 and BACE-1/Cathepsin-D) of a set of compounds based on the same core. The descriptors were computed on the sidechains (R1 to R3 groups which are fragment-like compounds) of the corresponding Markush structure. The methodology integrates the fragment contributions, through their descriptors, into the QSAR model. Thus the putative positive effect of a fragment over another at a given position in the structure can be estimated. Then they identified bioisosteric fragments with an inverse-QSAR approach in order to seek for more diverse compounds. It should be noted that the new structures fitted into the applicability domain of the training set. Unfortunately, the affinity and selectivity of these promising compounds were not experimentally validated. Albeit scaffold-centric QSAR based on fragment contributions was one of the first methods to be employed [77], its explicit applications to FBDD are recent.

Structure-Based Drug Design (SBDD)

The knowledge of the 3D structure of the targeted macromolecule is of paramount importance in FBDD, for it greatly enhances the development of rational hypotheses for the optimal ways to interconnect the binding fragments into an active lead molecule. Therefore, SBDD is much more popular than LBDD in fragment-based approaches, being obviously the method of choice whenever X-ray screening is involved, or whenever the structural elucidation of the target structure can be achieved by NMR. Modeling site-ligand interactions is at the core of docking calculations, developed for drug-like molecules but which should, at least in principle, be much easier with fragments, which are relatively small and rigid. However, things are not that simple, as scoring functions calibrated on the basis of drug-like compounds (the weak point of docking protocols) may not automatically work with small fragments, arguably outside their applicability domains.

Docking & Post-Processing Strategies

Molecular docking has been relatively successfully used to predict experimental pose of drug-like compounds, and this technique can be applied to fragments too. It is usually used as a prescreen filter to select most promising fragments to be tested experimentally.

The low complexity of fragments has however highlighted the limitations of state-of-the-art docking programs. Several studies and reviews reported main success stories and pitfalls of fragment docking [78]. Scoring functions are often based on molecular mechanics and provide only a crude assessment of ligand-site recognition. By contrast, empirical (fitted) scoring functions are defined as a weighted sum of *ad hoc* ligand–receptor interaction contribution terms assumed to stand for important physico-chemical interactions (hydrogen bonding, Van der Waals contacts, entropic penalties, etc). These weights are obtained by a regression analysis on a reference set of receptor–ligand complexes with known binding constants, a constraint limiting the training to a pretty small dataset, implicitly carrying a risk of a bias towards specific types of molecules. It is common knowledge that scoring functions are still not reliable enough to systematically return the experimental pose as top-ranked pose for drug-like compounds. They are the Achilles' heel of docking approaches because of their simplicity. Solvation effects, including desolvation and interactions with water molecules, and electrostatics are recurrent weaknesses of scoring functions [79]. Post-processing approaches have been developed to increase the reliability of docking outputs [80]. The most intuitive way is to use scoring functions of different types to score the same poses. This method is called consensus scoring [79], and only compounds which are well scored by two or more functions are kept. Consensus scoring has been widely used for some years and lead to improved enrichment rates. Other post-processing approaches filter unwanted poses based on geometric rules or the knowledge of the binding mode. Marcou and Rognan developed the "Interaction FingerPrint" (IFP) [81] to address this issue : the IFP (which encodes ligand-target interactions on a pharmacophoric basis into a binary string) of a known ligand is used to select compounds sharing a similar binding mode. Besides, they showed that this metric is better than the widely used RMSD criterion for fragments. Competitor approaches as the "Protein Ligand Interaction Fingerprints" (PLIF) of MOE [82] can compute similar descriptors to filter poses based on the presence or absence of key interactions. Finally it has been shown – on a small dataset due to the method used – that rescoring GOLD poses using QM/MM (quantum mechanics/molecular mechanics) approaches may enhance the output accuracy [83].

Pharmacophoric constraints can be applied during the docking to increase the reliability of the results with respect to known SAR data. For example, the necessity to make a given hydrogen bond can be specified before the docking. Thus the sampling will be positively biased toward a known binding mode. The widely accepted assumption that poses close to the experimental binding mode are more probable is the basis of the approach. Although this technique prevents the discovery of new binding modes, it is efficient in the fragment context since usual scoring functions experience great difficulties in correctly ranking these small compounds.

For fragments, the scoring function issue is more crucial since, because of their size, they can be placed in many ways inside the binding site, contrarily to larger ligands where shape alone strongly limits the number of possibilities. Thus, in fragment docking, scoring functions must be able to distinguish the experimental pose from a huge number of not obviously impossible but nevertheless irrelevant poses. Also, again because of their size, fragments might simply fall outside the applicability domain of fitted scoring functions – which, like any QSAR approach, should in principle only be used with molecules that are similar enough to the training examples. Albeit, to our knowledge, the explicit applicability of scoring functions was never assessed, the emergence of interaction fingerprint-based filtering techniques to discriminate between "native-like" and "decoy" poses of drug ligands may implicitly act as an applicability check. Discarding, for a kinase inhibitor candidate, all the poses not showing the "classical" hydrogen bonding pattern with the hinge region amino acids (discarding the candidate itself as inactive if no such pose is found amongst the otherwise well-scored docking conformations) may well enhance the precision of docking calculations [84]. Unfortunately, fragments show limited interaction spots with binding sites and are weak binders, at best participating in one of the key interactions (in the kinase example above, the constraint in fragment docking would pick out only hinge region binding fragments, while leaving the rest of the catalytic site unexplored).

Despite these important issues, development of new tools for virtual screening of fragment-like compounds is aggressively pursued, especially since the recent expansion of experimental FBDD approaches. Indeed, if the binding mode of the starting fragment cannot be resolved with biophysical methods, it can be suggested by usual docking programs. In a similar manner, they may be used to screen virtual libraries of fragments. It should be noted that, even though current docking tools present real drawbacks concerning their ability to accurately predict fragment binding modes, encouraging results were obtained with Glide [85] in redocking studies [86] and virtual screening [87] focused on fragment-like compounds. The most widely

used softwares, such as DOCK [88], FlexX [89], GOLD [90], AutoDock [91] and Glide, were used to perform virtual screening of fragment libraries, and the subsequent optimization of validated virtual hits lead to high-affinity inhibitors for many relevant targets. Several examples thereof will be described below:

Agamennone *et al*. used several computational approaches, using among others GRID [92] and the GOLD docking software, to select fragments which were subsequently screened with NMR in order to develop inhibitors of the protein-protein S100B–p53 interaction [93].

Shoichet and Chen recently reported fragment docking studies using DOCK with the class A beta-lactamase CTX-M as target [94]. They screened two subsets of the ZINC database (http://zinc.docking.org): the fragment subset and the lead-like subset. None of the 37 lead-like tested compounds showed any inhibition of the hydrolysis of the beta-lactam substrate nitrocefin. By contrast, they identified ten hits with IC_{50} values in the mM range by testing 69 fragments. They subsequently solved the X-ray structure of seven fragment-target complexes and it should be noted that DOCK predicted the right binding mode with high accuracy. Finally, they evolved a tetrazole fragment into a more potent (K_i in the µM range) and specific, with respect to a class C beta-lactamase, inhibitor. They used a similar approach, involving both the Zinc database and the DOCK program, in a project dedicated to the class C beta-lactamase AmpC [95].

Zhang *et al*. reported the fragment-based development of malic enzyme inhibitors by using FlexX and a homology model of this enzyme [96]. Similarly, using an Aurora kinase A homology model relaxed by molecular dynamics, Warner and co-workers employed both LUDI [97] and Glide to facilitate the design of inhibitors for this enzyme [98].

Howard *et al*. developed highly potent thrombin inhibitors by using chemoinformatics and X-ray approaches [11a]. They first prepared a small, focused fragment library: they docked, using a modified version of GOLD, a fragment library and selected about 80 putative binders. By screening these fragments crystallographically, they found several low-affinity hits, the subsequent linking of which, using the X-ray structure of complexes, lead to a very interesting inhibitor (IC_{50} in the low nM range).

Englert *et al*. reported the discovery of inhibitors targeting the metalloproteinase thermolysin [99]. To achieve this goal, several filters were applied to create a focused fragment library. The most interesting one relates to the propensity to chelate the Zn^{2+} cation in the binding site. Secondly, they tested three different docking tools (AutoDock, FlexX and GOLD) with several scoring schemes, and used the best performer, AutoDock, for the virtual screening of fragments. The compounds to be tested were selected on the basis of a given pharmacophore. Finally, the assays reported several low-affinity hits, which were evolved into more potent compounds with the aid of fragment-target X-ray structures.

Röhrig *et al*. discovered indoleamine 2,3-dioxygenase (IDO) inhibitors with nanomolar potency in an enzymatic assay [100] using two *in silico* strategies involving their evolutionary docking algorithm EADock [101], a pharmacophore-based and a fragment-based approach, respectively. By docking known IDO inhibitors, they created a pharmacophore model which was subsequently used to design new ligands. In the context of the present review, the fragment-based strategy is more interesting. They built a fragment library from commercial databases and screened it with EADock to create maps of favorable binding modes for each fragment. By analyzing visually the docking results, they found several interesting fragments which could be directly linked (e.g. benzene and imidazole). SBDD optimization of these compounds designed from fragments and of those found *via* the pharmacophore method allowed the discovery of novel, highly potent inhibitor leads.

Caflisch *et al*. developed several tools dedicated to computational FBDD. Huang and Caflisch recently reviewed their tools [102]. Their software suite includes three programs (DAIM, SEED and FFLD) in their software suite, each designed to perform one step of the process. It should be noted that the second tool (SEED) needs several fragments, acting as anchors, which will be docked in the binding site. Thus one needs a method able to break compounds into fragments. The DAIM (Decomposition And Identification of Molecules) strategy has been developed to achieve this goal [103], and can be decomposed into four steps: 1) identification of rings within the compound, 2) identification of fragments defined as a set of atoms connected by unbreakable bonds, 3) merging of neighbors with previous fragments to create chemically relevant entities and 4) completion of the valences. A specific, fingerprint-based approach selects the three most appropriate fragments to be used in the subsequent fragment docking step. They showed that this number of fragments allows the correct positioning of a flexible ligand in a binding site. As mentioned above, the result of the DAIM process is used as an input for the SEED (Solvation Energy for Exhaustive Docking) tool [104], which searches optimal binding modes for fragment-like compounds in the binding site. By using a desolvation model, based on the continuum dielectric approximation, SEED is able to prioritize the right placement of hydrophobic fragments into a hydrophobic pocket. As SEED, FFLD (Fast Flexible Ligand Docking) will use the output of the previous step to start its own process [105]. This docking tool is based on a genetic algorithm to perform random modifications of ligand conformation, but the final placement of the full compound is biased toward the previously docked fragments. The full process leads to the docking of full ligands, but it is of course possible to stay in the fragment chemical space by using only the SEED tool and a fragment library. Their software suite was successfully used to discover new inhibitors targeting several relevant targets, among others beta-secretase [106], the West Nile Virus non-structural 3 protease (NS3pro) [107] and the *Plasmodium falciparum* plasmepsin II protease [108].

Dakshanamurthy reported a "SAR by catalog" study with *in silico* predictions as input [109]. They performed a virtual screening with the FlexX/FlexX-Pharm [110] approach to focus on compounds able to pick up common interactions between ligands and kinase binding sites. Indeed the FlexX-Pharm method allows including pharmacophore-like constraints in the docking calculation. The library was an in-house fragment library and the target was the vascular endothelial growth factor receptor-2 (VEGFR2) kinase, an

enzyme involved in the angiogenesis pathway. Docking under constraints provided putative fragment binders, which were used as queries to search larger compounds containing them. Experimental tests showed that several moderately potent compounds were able to inhibit angiogenesis without being cytotoxic. This docking strategy has also been successfully used for the discovery of dipeptidyl peptidase IV inhibitors which are potential anti-diabetic agents [111]. Machrouhi *et al*. designed new analogs of a non-specific reference inhibitor (compound C) of the 5'-AMP-activated protein kinase (AMPK) by using a modeling workflow [112]. They first built a homology model of this enzyme which was subsequently used to sample, with an in-house tool and a grand canonical Monte-Carlo strategy, three fragments obtained from the breaking of the reference ligand. The putative binding mode of compound C was obtained by assembling docked fragments located in the ATP-binding site. A protocol, involving minimization and a short molecular dynamics simulation, lead to a relaxed complex. This conformational state of the receptor was used to perform the screening of a fragment library. The result of this simulation lead to the selection and synthesis of several analogs, and some of them exhibited comparable IC_{50} (sub-μM range) and enhanced selectivity with respect to a panel of kinases.

De Novo Design

De novo design (DND) programs are able to automatically build, typically *via* a growth process based on a set of fragments and chemical interconnection rules, compounds fulfilling several criteria such as binding site or pharmacophoric constraints. Since the recent rise of experimental FBDD, computational techniques that facilitate the fragment evolution process, either through growing or by linking several non-overlapping fragments, are particularly hot research topics. In this sense, DND can be seen as "completely" virtual FBDD, bypassing the need for experimental check of putative affinities and binding modes of the initial fragments and the resulting concatenated products. It should be noted that using experimentally validated fragments as starting points for a DND run should increase the reliability of the predictions, and that the subsequent experimental test of newly designed compounds at each step (concept of interactive optimization which is not specific to DND evolution) may prune "dead" branches from the tree of all possible growth paths. Most DND methods try to identify energetically favorable positions within the binding site for a variety of probes, from chemical groups to fragment-like compounds, with the same type of scoring functions as those used for docking. In principle, DND algorithms cannot be deterministic because of the huge number of possible ways to grow a compound or link different entities. Besides, most molecules possess conformational degrees of freedom which increase the complexity of the search. This is why DND programs often use heuristics-based (randomness concept) approaches; thus different DND runs can lead to different results. Their theoretical main advantages are fast hit-to-lead optimization and a small number of compounds to be synthesized and subsequently tested. The success of DND strategies has been limited by both chemical synthesis problems and the low reliability of the predicted binding energies. Indeed, for a while, the main drawback of DND approaches has been to propose compounds impossible to synthesize. Newest approaches try to address this issue for example by incorporating rules to assess the feasibility of chemical synthesis, based on the fact that larger compounds can be produced by combining building blocks with common chemical reactions. These schemes are used to grow up compounds during the simulation. The RECAP reaction schemes [49a] are widely used, for instance in the TOPAS [113] and SQUIRREL [14] programs. DND has been reviewed some years ago by Schneider and Fechner [115].

DND tools are far from being perfect, but they were able to generate relevant predictions, such as active compounds of moderate activity (μM range) for a wide variety of targets. They are of particular interest in the early stages of fragment evolution. Some success stories, using a DND strategy in FBDD projects, will be briefly presented.

An important number of DND methods, developed before the rise of FBDD, are based on the fragment concept and growing/linking approaches [115], for example well-known tools as Multi-Copy Simultaneous Search (MCSS) [116] and SPROUT [117]. As for MCSS but by using another concept, called 3D Reference Interaction Site Model (3D-RISM), Imai *et al*. found favorable positions for fragments on the protein surface [118]. Ivetac and McCammon used the FT-MAP [119] approach, based on the fast Fourier transform correlation strategy, to map 16 chemically relevant fragment-like probes (e.g. isopropanol, cyclohexane, benzene, phenol…) on the whole surface of an ensemble of structures (experimental structures and snapshots from molecular dynamics simulations) of beta-1 and beta-2 adrenergic receptor [120]. They found several putative allosteric binding sites, more or less close to the orthosteric binding site of these class A GPCRs, which should be experimentally investigated. In the Fragment Screening by Replica Generation (FSRG), a fragment plus an ensemble of derivatives are processed [121]. Despite the fact that fundamental components such as coulombic terms and hydrogen-bond interactions are neglected, encouraging results were obtained with DUD datasets [122] to discard decoys from active fragments. The program LUDI [97], developed in 1992, also relies on the fragment concept and uses an empirical scoring function. It identifies favorable positions for fragments with respect to the target binding site. Subsequent merging of these fragments leads to the final compound. It was used in the design of several fragment-based inhibitors [123]. Takahashi and co-workers reported the design of moderate potency (IC_{50} in μM range) inhibitors of matrix metalloproteinases [123a]. The program LUDI was used to identify fragments that may interact with residues of the binding site. They selected fragments on the basis of binding modes of known inhibitors, and a hydroxamic moiety, well-known to coordinate zinc ions, was linked to their initial fragment. Chemical modifications of this compound lead to moderate potency inhibitors more or less specific toward several matrix metalloproteinases. Boehm *et al*. used LUDI to dock fragments (called "needles" in their paper) into the DNA-gyrase binding site in order to reduce the dataset (from 350K to 3K) while keeping potential actives that share conserved key interactions with respect to known inhibitors [123b]. The assays provided many hits and the SBDD optimization lead to an indazole

inhibitor ten times more potent than a reference inhibitor (novobiocin).

Zaliani *et al*. developed a software suite for fragment-based DND called NovoBench, including CoLibri, Recore, FragView, FragEnum, FlexNovo and Sylvia softwares [124]. One of these tools, Recore, has been developed to perform scaffold hopping [125]. It is able to quickly screen huge libraries of pre-processed fragments by using an index-based strategy. In the current version, pharmacophoric and binding site constraints are both available in order to guide the design process. More precisely, this method relies on "cut vectors" which are used to define moieties to replace in the scaffold hopping strategy. However these vectors can be used in other contexts, for example to optimize one or more fragments. In a linking strategy, the user has to place vectors in order to select ends to join, and the subsequent screening will return putative merged structures. In a growing strategy, the user has to place a vector on the extremity to evolve and another one in the binding site in order to guide the evolution in this direction. Resulting structures are ranked according their deviation from the initial geometry of the query. Thus, although developed for scaffold hopping purposes, Recore can be used in the FBDD field to evolve fragments into larger compounds.

Moriaud *et al*. reported the use of MED-SuMo and derived tools to create objects called MED-Portions with PDB files as input [126]. A MED-Portion is defined as an object encoding fragment-target sub-pocket pairs. By using a sub-pocket of interest as query, these MEDIT tools are able to search for similar sub-pockets. Since the MED-Portions include information about the fragment bound to the sub-pocket, potential small binders are suggested. In the case of several outputs, one for each sub-pocket, a new hybrid compound can be virtually built as in usual DND softwares. They applied their approach to distinct relevant target families such as GPCR and kinases. They also used this approach to study the allosteric binding site of the mitotic kinesin Eg5 [127].

Fragment linking tools for both experimentally validated and docked fragments are of particular interest in the FBDD context. Several recent programs are specialized in the linking of entities, as for example GANDI [128] and CONFIRM [129]. As mentioned in its name, GANDI (Genetic Algorithm-based de Novo Design of Inhibitors) uses an island-based genetic algorithm to sample linkers in order to join pre-docked fragments. These fragments were previously docked using the SEED tool (described above). An interesting feature is the ability to bias the sampling toward known binding modes or known inhibitors by using a similarity-based method. Finally, they showed that many built compounds are synthetically accessible or listed in the ZINC database and look like known kinase inhibitors. In the CONFIRM strategy, a "bridge" library is screened in order to find suitable linkers between fragments that have been selected by Glide used as the docking engine. Linked candidates are selected on the basis of their strain energy and the deviation of the original fragments between their initial position and their position in the linked compound.

EXPERIMENTAL FBDD

The 1990s have seen improvements of robotics and combinatorial chemistry [41j]. Both factors accelerated the rise of high-throughput screening (HTS) which consists in screening huge compound collections with fast biochemical-based assays to identify hits. In parallel, massive developments appeared in the biophysical field. Biophysical methods are able to detect weak binding events; this fact allowed the development of FBDD approaches. 1996 is a reference date in the FBDD field because it was the first time that a ligand was experimentally built from several fragments and a linker ; the authors called their technique SAR by NMR [10].

Nowadays, a wider range of techniques are used in FBDD campaigns, including NMR, X-ray crystallography, SPR and MS. In some cases, biochemical assays at high concentration of compounds can be employed.

The next paragraphs will be dedicated to these experimental screening techniques, and will emphasize their advantages and drawbacks. Several practical examples will be presented for each method.

Biochemical Assays/High-Concentration Screening (HCS)

In the FBDD field, biochemical assays are not the most used screening techniques in spite of their advantages (high throughput, sensitivity, low protein consumption, and wide applicability including membrane proteins). The weak affinity of most fragments requires screening at high concentration (up to 1 mM); however these conditions are known to produce false positives due to compound aggregation, chemical reactivity, interference with the assay, or protein denaturation, as well as false positives due to compound insolubility [130]. However, a recent comparative study with trypsin and MMP12 has shown that the false-negative and false-positive rates observed with 3 different biochemical assay methods were comparable to those obtained with SPR-based assays [131]. Besides, these experiments don't give any information about the binding mode of hits so they need to be complemented by X-ray crystallography. Another limitation is that the identification of second site binders is possible only if their binding allosterically modulates the primary binding site. Nevertheless, biochemical assays have been used for fragment screening, most of the time taking advantage of the high throughput and extreme sensitivity of fluorescence-based readout methods [132]. For example, interesting results were found for the Peroxisome Proliferator-Activated Receptor (PPAR) family since high-concentration screening (HCS) allowed to develop a PPAR pan-active ligand (indeglitazar, currently under clinical trials) [133]. Other success stories, involving HCS and high-throughput crystallography, include the design of phosphodiesterase 4 (PDE4) [6], B-RafV600E oncogenic mutant [134], beta-secretase (BACE-1) [135] and Hsp90 inhibitors [136]. Finally, it is worth noting that it is easy to perform fragment-based screening by HCS on existing HTS platforms, which groups already running such facilities can take advantage of.

X-Ray Crystallography

Crystallography-based screening consists in the soaking of preformed apo-target crystals with mixtures of several diverse compounds at high concentration, followed by acquisition of diffraction data [137]. The analysis of the electron density map allows to answer both questions "does a fragment in the mixture bind?" and "which one is the binder?". The initial criticism to crystallography was that it is not really compatible with the screening concept, which involves high throughput. With advances in robotics (e.g. automatic sample changers), instrumentation (X-ray sources and detectors), and automated density map interpretation, solving crystal structure of high resolution has become really fast [138]; thus the potential of crystallography-based screening has been achieved. An interesting development in this field is the program AutoSolve [139] which is able to answer both previous questions in an automated way.

X-ray crystallography yields an electron density map which provides a detailed description of ligand-target interactions at the atomic level; this information can be used in an SBDD approach to perform an efficient optimization. Another advantage of this method is its false positive rate equal to zero: molecules that bind to the surface or to another putative site can be easily discarded. The soaking mixture must be sufficiently diverse in shape to unambiguously identify the bound compounds in the electron density map. The selection of compounds to incorporate into each mixture can be done by using chemo-informatics methods to ensure the chemical diversity of each cocktail (see computational part).

Crystallography fragment screening was used to develop, among others, inhibitors of urokinase [137a], dihydroneopterin aldolase [140], p38α MAP kinase [141], *Trypanosoma brucei* nucleoside 2-deoxyribosyltransferase [142], thrombin [11a], BACE-1 [143], CDK2 [144], HCV RNA polymerase NS5b [145], JAK-2 [146], leukotriene A4 hydrolase [5], and Pim-1 [147].

Fragment screening by X-ray crystallography can identify low-affinity hits that would not have been found using functional assays due to their low potency. On the other hand, these compounds can exhibit quite high LE because of their low molecular mass. However, the need of important amounts of protein and the availability of highly diffracting crystals usable for soaking experiments limit the diversity of targets which can be screened with this approach. Besides, soaking may be unsuccessful when the protein-fragment interaction requires an induced fit but the protein is conformationally locked due to the crystal packing [41c]. Such restricted conformational changes may result on one hand in crystal cracking upon compound soaking, leading to diffraction loss, or on the other hand in absence of binding (false negatives). In the end, the relatively low success rate of primary crystallography-based screening has prompted many groups to screen fragment libraries using techniques such as NMR or SPR first (see following sections), and then to characterize in detail the binding mode of hits by X-ray crystallography using either soaking or co-crystallization studies or both. Whatever the primary screening technique used, there is a general consensus that X-ray crystallography is essential to elaborate fragment hits to improve their binding affinity and eventually turn them into leads.

Nuclear Magnetic Resonance (NMR)

NMR has long been used as a standard technique for macromolecular 3D structure determination [148]. More recently, it has also been recognized as a powerful tool for the detection and quantification of protein-ligand interactions [149]. Its versatility and utility at various stages of drug discovery, including screening, hit validation, hit-to-lead optimization, structural characterization of the target, and even target druggability, have been reviewed recently [150]. In particular, NMR is particularly well-suited for fragment screening thanks to its ability to detect weak affinity binders (with Kd as high as 5 mM), which is often the case with low molecular mass compounds. Indeed, NMR is perhaps the most widely used fragment screening technique. Screening by NMR was introduced in 1996 by the seminal "SAR by NMR" article [10], which constituted the first practical success of FBDD. In this study, fragments able to bind to FKBP were identified by monitoring chemical shift perturbations in two-dimensional $^{1}H/^{15}N$ HSQC (Heteronuclear Single-Quantum Coherence) spectra of the ^{15}N-labeled target – in other words, by observing protein amide chemical shift changes in 2D-NMR spectra (apo-protein *vs* protein+fragment). The NMR-based screening provided two interesting, non-competitive hits. In a second step, both fragments were individually optimized with the help of structural data. In a last step, the two resulting moderate affinity binders (respective Kd values of 2 and 100μM) were linked to yield a high affinity ligand (Kd value of 19 nM).

Briefly, there are two main approaches in NMR-based fragment screening: observation of the protein and observation of the ligand [150a, 151]. The most frequently used protein-observed method is chemical-shift perturbation, exemplified by $^{1}H/^{15}N$ or $^{1}H/^{13}C$ HSQC two-dimensional correlation spectra. Thanks to the ^{15}N or ^{13}C spectral editing, no signal from the ligand is observed and the screening can be performed at high ligand concentration; moreover, the technique provides precise information on the location of the binding site, allowing the screening of second or allosteric binding sites. Despite these advantages, the protein-observed approach suffers from several drawbacks : it requires important amounts of costly ^{15}N- or ^{13}C-labeled protein, long acquisition times, backbone ^{1}H-^{15}N resonance assignments (currently limited to targets with MM < 30-40 kDa), and prior knowledge of target structure. However, the development of cryogenic probes allowed to improve greatly the sensitivity and thus the data quality and the throughput of heteronuclear screening through the reduction in data acquisition times [152]. A higher throughput can be achieved by combining the use of a cryogenic probe with the screening of compound mixtures at low protein and compound concentrations [153].

To overcome these limitations, various ligand-observed techniques were developed. They require low concentrations of unlabeled protein, have a higher throughput thanks to shorter acquisition times, and are applicable to a wider range of targets (there is no limitation in the size of the protein). On the other hand, ligand-observed techniques do not provide structural

information about the binding sites (though specific binding can be assessed by competition experiments using reference ligands) and high affinity ligands cannot be directly detected. The most popular ligand-observed NMR screening techniques are STD (Saturation Transfer Difference) [154] and Water LOGSY (Water Ligand Optimized Gradient Spectroscopy) [155] because of their high sensitivity and ease of implementation and analysis. In STD NMR experiments, a protein signal is selectively irradiated, saturation propagates from the selected protein hydrogens to all protein hydrogens through spin diffusion (intramolecular ^{1}H-^{1}H cross-relaxation) and is also transferred from the protein to a bound ligand (intermolecular ^{1}H-^{1}H cross-relaxation) during its residence time at its binding site. The saturation signal is detected on the free ligand upon dissociation from the protein. A difference spectrum is recorded using as an "off-resonance" irradiation to remove the protein signals, resulting in a 1D ^{1}H spectrum where only resonances of binding compounds appear. With low affinity ligands, the fast exchange between the bound and the free states allows saturation transfer to many molecules, building up strong STD signals. Conversely, high affinity ligands typically have long residence times in their binding site and their STD signals are very weak; competition experiments have been designed to detect high affinity ligands [156]. In WaterLOGSY experiments, the intermolecular magnetization transfer occurs from bulk water to protein binding sites and on to bound ligands.

More recently, a competition-based ^{19}F NMR screening technique called FAXS (Fluorine chemical shift Anisotropy and eXchange for Screening) [157] was developed, taking advantage of ^{19}F detection (100% ^{19}F natural abundance, absence of signal overlap allowing to screen large compound mixtures). This technique, requiring small amounts of protein and short acquisition times, measures the displacement of a low-affinity fluorinated "spy" molecule, making it applicable to fragment screening in combination with *in silico* pre-screening [158].

An original screening method called TINS (Target Immobilized NMR Screening) was also developed, consisting in the immobilization of a target and a reference protein on a solid medium, and in the monitoring by NMR of fragment binding to the immobilized proteins [159]. An advantage of the method is the small amount of protein required since a single immobilized sample serves to screen the entire library. Very recently, TINS was compared with STD NMR, SPR and high-concentration screening for fragment hit identification [160]. The method could also be applied to a bacterial membrane protein solubilized in detergent micelles and lipid bilayer nanodiscs before immobilization [161]. A detailed comparison of the different NMR techniques with other biophysical techniques used for fragment screening was published recently [162]. Among others, initial protein-observed NMR screening was used in combination with other methodologies to develop inhibitors of stromelysin (MMP3) [163], protein tyrosine phosphatase 1B (PTP1B) [164], Bcl2 family proteins [165], Hsp90 [166], prostaglandin D2 synthase [167], X-linked inhibitor of apoptosis protein (XIAP) [168], and BACE-1 [47], and initial ligand-observed NMR screening was used in combination with other methodologies to develop BACE-1 inhibitors [169] and allosteric ligands of 3-phosphoinositide-dependent kinase-1 (PDK1) [170]; and a combination of ligand-observed and protein-observed NMR was used to identify ligands targeting the S100B–p53 interface [93]. STD NMR has also been used for the deconvolution of fragment pools selected from a biochemical assay [171].

Mass Spectrometry (MS)

Detection of protein-ligand interactions by electrospray ionization mass spectrometry (ESI-MS) is gaining popularity in the FBDD field owing to the small amounts of protein and ligands required [172]. Noncovalent electrospray ionization mass spectrometry (NC-ESI-MS or NC-MS), also called native mass spectrometry is a label-free, high-sensitivity technique than can detect non-covalent complexes with a K_d up to the millimolar range. The technology could be automated and miniaturized through the coupling of a nanoelectrospray time-of-flight mass spectrometer with a robotized nanochip-based injection system, resulting in a higher throughput and an increased sensitivity, and making it a valuable technique for fragment screening, as implemented in NovAliX [173]. Indeed, NC-MS proved to be a fast and efficient primary screening technique compared to X-ray crystallography [174]. Besides its usefulness as a screening technique, NC-MS also yields high-content information on protein-ligand interactions in terms of stoichiometry, specificity (through competition experiments), relative affinity, and strength of interaction [173]. NC-MS was involved in a "SAR by MS" study, where known inhibitors were deconstructed and the resulting set of fragments were analyzed for binding to the target by MS [175]. NC-MS was also used, in conjunction with NMR, SPR, and ITC, in a fragment-based approach for the development of protein-protein interaction inhibitors [176]. In a recent study [177], NC-MS served both to assess the binding of starting inhibitors and to detect the formation of improved inhibitors obtained by dynamic combinatorial chemistry through the *in situ* formation of a disulfide bond, reminiscent of the Tethering approach.

Tethering, a site-directed ligand discovery method, relies on the creation of a covalent bond between the ligand and the binding site [178]. The covalent bond is a disulfide bond involving a cysteine residue from the binding site and a thiol group from the ligand. Thus the presence of a cysteine is required in the binding site; it can be present in the wild-type protein or added by site-directed mutagenesis. To be compatible with the Tethering approach, the fragment library must be composed of compounds containing thiol groups. Compounds exhibiting some affinity for the targets that have a thiol group located near the cysteine will make a disulfide bridge with this residue provided the orientation is favorable. Once this covalent bond is created, a standard ESI-MS method is able to detect the adduct. This strategy has been used to develop interleukin 2 [179] and caspase-3 [180] inhibitors. Buck *et al.* used this approach to discover ligands for the C5a receptor, a member of the GPCR family, through the engineering of specific cysteine mutations [181].

Surface Plasmon Resonance (SPR)

SPR is a quantum phenomenon occurring when an incident beam of polarized light resonantly excites surface plasmons, which can be viewed as surface electromagnetic

waves propagating in a direction parallel to a metal/external medium interface. Practically, SPR is used to monitor changes in the refractive index of the medium adjacent to the metal surface (typically, a glass slide coated with gold), in particular when a ligand binds to a protein that has been attached to the glass surface, which is the principle of optical biosensors. Not only the binding event is detected, but also the kinetic parameters of binding (k_{on}, k_{off}) can be measured, giving access to the dissociation constant ($K_D = k_{off}/k_{on}$). SPR-based optical biosensors have long been used in drug discovery for compound screening and lead optimization [182]. Initially used for the validation and characterization of hits from a high-throughput screening, they are now able to detect the binding of weaker ligands thanks to sensitivity improvements, and are thus suitable for fragment screening [183]. Kinetic parameters k_{on} and k_{off}, and hence affinities can be derived from SPR experiments though for weak binders, compound solubility often limits the precision of the K_d determination. SPR-based techniques are used either as primary fragment screening techniques or for the estimation of the Kd of hits identified using other techniques. To quote but a few, recent examples of fragment-based SPR screening include the development of BACE-1 [184], MMP12 [185], X-linked inhibitor of apoptosis - caspase interaction [176] and chymase [186] inhibitors. New devices based on different detection principles: grating couplers (nanostructured optical sensors in a microplate format such as the Epic™ and BIND™ systems) and reflectometric interference spectroscopy (RIfS) (Octet™ system using a fiber-based approach) offer higher throughputs [183].

Finally, in standard optical biosensors, the target is immobilized on a surface and the binding of a small molecule modifies the surface's optical properties. In the reverse mode, Graffinity Pharmaceuticals has developed an original combination of chemical microarrays and SPR imaging able to generate simultaneously binding data for a protein target with more than 9,000 immobilized fragments per array [48]. In this unique set-up, the target protein is incubated with the chemical microarray, resulting in much higher response signals that are detected in a high-throughput fashion (for each campaign, 24,000 fragments and 86,000 lead-like compounds are screened). This approach is suitable for fragment screening and has been used in the discovery of thrombin [187] and MMP13 [188] inhibitors.

Isothermal Titration Calorimetry (ITC)

A method that is able to characterize all the fundamental thermodynamic parameters of binding, such as $\Delta H/T\Delta S/\Delta G$, is of particular interest, especially in the FBDD field since the choice of fragments to follow up often relies on their initial ligand efficiency (LE). Isothermal titration calorimetry (ITC) can provide the complete thermodynamic profile of a binding reaction by directly measuring the enthalpy change upon binding, and subsequently the binding affinity, and deriving from them the entropy change [189]. However, ITC is a very low-throughput technique that still requires important amounts of protein. Though not suitable for screening, it is used in combination with other techniques, bringing invaluable information during the hit-to-lead optimization [158, 168, 174, 190]. This has been recently illustrated by Edink *et al.* who reported a study about the optimization of a fragment bound to the acetylcholine binding protein (AchBP) target [191]. Since their fragment partially adopted a similar binding mode as lobeline, and since the latter compound is known to induce a conformational change of the binding site, they evolved their initial fragment in order to try to induce this conformational change. The docking tool GOLD was used to predict the binding mode of the designed compound, and it suggested an overlap between both moieties of interest. The predicted binding mode of this larger ligand was confirmed experimentally. Besides, binding of the compound lead to the expected conformational changes at the binding site. Thus, a compound resulting from the optimization of a fragment could induce the desired conformational changes in the binding site of AchBP. Finally, the X-ray structure of the complex with an analog showed that the ligand adopted an alternative binding mode. Using ITC, the authors observed differences in their thermodynamic binding signatures, and concluded that both structural biology and thermodynamic data bring essential information during evolution stages.

CONCLUSION

Fragment-based methods are now considered as a reliable approach in drug discovery and have been adopted by both biotechs and big pharmas. FBDD, which consists in screening fragments with sensitive biophysical techniques, has an important number of advantages as described in this review. The number and variety of targets demonstrate that FBDD has emerged has an alternative way to HTS to identify hits.

High affinity is a necessary property for drugs, but other parameters are fundamental in drug discovery. FBDD is able to generate compounds with both high affinity and good pharmacokinetic properties, as it proceeds from simple entities towards larger ligands and may be steered in a way to optimize all the relevant aspects simultaneously.

Each experimental fragment screening method has its advantages and drawbacks. Among them, X-ray crystallography and NMR can provide detailed (atomic-scale) binding mode information, of paramount importance in SBDD. Progresses about the production of well-diffracting GPCRs crystals should allow to extend the use of FBDD to this pharmaceutically important class of targets [22].

Computational chemistry applications in FBDD range from efficient fragment library design to selection of putative fragment binders and their subsequent evolution toward a lead. However, fragment-oriented approaches may have, due to the lower size/complexity of fragments, peculiarities that clearly distance them from “classical” chemoinformatics and docking tools, aimed at handling drug-like compounds. Yet, modeling fragment-sized ligands should be technically more readily feasible than modeling large ligands – albeit “classical” tools focused more on the latter. A typical illustration of the “Less is More” conundrum, FBDD may efficiently explore drug-like space by exploiting knowledge about the preferred binding behavior of fragments. It therefore avoids direct confrontation with the otherwise inextricable complexity of both the whole chemical space –

that cannot be covered by brute-force HTS – and the unconstrained modeling of large flexible ligands.

In the light of their intrinsic qualities, of recent developments, and of numerous success stories reported to date, fragment-based approaches (both computational and experimental) should clearly play a larger role in drug discovery in the future.

ABBREVIATIONS

DND = *De novo* design

DOS = Diversity-oriented synthesis

DS = Descriptor space

FBDD = Fragment-based drug discovery

GPCR = G protein-coupled receptor

HCS = High-concentration screening

HTS = High-throughput screening

LBDD = Ligand-based drug design

LE = Ligand efficiency

MM = Molecular mass

MS = Mass spectrometry

QSAR = Quantitative structure-activity relationship

QSPR = Quantitative structure-property relationship

SBDD = Structure-based drug design

SPR = Surface plasmon resonance

VS = Virtual screening

REFERENCES

[1] Sams-Dodd, F. Target-based drug discovery: is something wrong? *Drug Discov. Today,* **2005**, *10* (2), 139-147.

[2] Lipinski, C.; Lombardo, F.; Dominy, B.; Feeney, P. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.,* **2001**, *46* (1-3), 3-26.

[3] Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A rule of three for fragment-based lead discovery? *Drug Discov. Today,* **2003**, *8* (19), 876-877.

[4] Law, R.; Barker, O.; Barker, J. J.; Hesterkamp, T.; Godemann, R.; Andersen, O.; Fryatt, T.; Courtney, S.; Hallett, D.; Whittaker, M. The multiple roles of computational chemistry in fragment-based drug design. *J. Comput-Aided Mol. Design.,* **2009**, *23* (8), 459-473.

[5] Davies, D.; Mamat, B.; Magnusson, O.; Christensen, J.; Haraldsson, M.; Mishra, R.; Pease, B.; Hansen, E.; Singh, J.; Zembower, D.; Kim, H.; Kiselyov, A.; Burgin, A.; Gurney, M.; Stewart, L. Discovery of leukotriene A4 hydrolase inhibitors using metabolomics biased fragment crystallography. *J. Med. Chem.,* **2009**, *52* (15), 4694-4715.

[6] Card, G.; Blasdel, L.; England, B.; Zhang, C.; Suzuki, Y.; Gillette, S.; Fong, D.; Ibrahim, P.; Artis, D.; Bollag, G.; Milburn, M.; Kim, S.; Schlessinger, J.; Zhang, K. A family of phosphodiesterase inhibitors discovered by cocrystallography and scaffold-based drug design. *Nat. Biotechnol.,* **2005**, *23* (2), 201-207.

[7] Teague, S.; Davis, A.; Leeson, P.; Oprea, T. The design of leadlike combinatorial libraries. *Angew. Chem. Int. Ed. Engl.,* **1999**, *38* (24), 3743-3748.

[8] Erlanson, D. A.; McDowell, R. S.; O'Brien, T. Fragment-based drug discovery. *J. Med. Chem.,* **2004**, *47* (14), 3463-3482.

[9] Jencks, W. On the attribution and additivity of binding energies. *Proc. Natl. Acad. Sci. USA,* **1981**, *78* (7), 4046-4050.

[10] Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. Discovering high-affinity ligands for proteins: SAR by NMR. *Science,* **1996**, *274* (5292), 1531-1534.

[11] (a) Howard, N.; Abell, C.; Blakemore, W.; Chessari, G.; Congreve, M.; Howard, S.; Jhoti, H.; Murray, C. W.; Seavers, L. C. A.; van Montfort, R. L. M. Application of fragment screening and fragment linking to the discovery of novel thrombin inhibitors. *J. Med. Chem.,* **2006**, *49* (4), 1346-1355; (b) Congreve, M.; Chessari, G.; Tisi, D.; Woodhead, A. J. Recent developments in fragment-based drug discovery. *J. Med. Chem.,* **2008**, *51* (13), 3661-3680; (c) Baurin, N.; Aboul-Ela, F.; Barril, X.; Davis, B.; Drysdale, M.; Dymock, B.; Finch, H.; Fromont, C.; Richardson, C.; Simmonite, H.; Hubbard, R. E. Design and characterization of libraries of molecular fragments for use in NMR screening against protein targets. *J. Chem. Inf. Comput. Sci.,* **2004**, *44* (6), 2157-2166.

[12] Schulz, M. N.; Hubbard, R. E. Recent progress in fragment-based lead discovery. *Curr. Opin. Pharmacol.,* **2009**, *9* (5), 615-621.

[13] Babaoglu, K.; Shoichet, B. Deconstructing fragment-based inhibitor discovery. *Nat. Chem. Biol.,* **2006**, *2* (12), 720-723.

[14] Barelier, S.; Pons, J.; Marcillat, O.; Lancelin, J. M.; Krimm, I. Fragment-based deconstruction of Bcl-x(L) inhibitors. *J. Med. Chem.,* **2010**, *53* (6), 2577-2588.

[15] Hajduk, P.; Huth, J.; Fesik, S. Druggability indices for protein targets derived from NMR-based screening data. *J. Med. Chem.,* **2005**, *48* (7), 2518-2525.

[16] Edfeldt, F. N.; Breeze, A. L.; Folmer, R. H. Fragment screening to predict druggability (ligandability) and lead discovery success. *Drug Discov. Today,* **2011**, *16*(7-8), 284-287.

[17] Hann, M.; Leach, A.; Harper, G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.,* **2001**, *41* (3), 856-864.

[18] Pendleton, R.; Kaiser, C.; Gessner, G. Studies on adrenal phenylethanolamine N- methyltransferase (PNMT) with S K & F 64139, a selective inhibitor. *J. Pharmacol. Exp. Ther.,* **1976**, *197* (3), 623-632.

[19] Fox, S.; Wang, H.; Sopchak, L.; Khoury, R. High throughput screening: early successes indicate a promising future. *J. Biomol. Screen.,* **2001**, *6* (3), 137-140.

[20] Lahana, R. How many leads from HTS? *Drug Discov. Today* **1999**, 447-448.

[21] Munos, B. Lessons from 60 years of pharmaceutical innovation. *Nat. Rev. Drug Discovery,* **2009**, 959-968.

[22] Tate, C. G.; Stevens, R. C. Growth and excitement in membrane protein structural biology. *Curr. Opin. Struct. Biol.,* **2010**, *20* (4), 399-400.

[23] Hajduk, P. J.; Galloway, W. R.; Spring, D. R. Drug discovery: A question of library design. *Nature,* **2011**, *470* (7332), 42-43.

[24] Murray, C. W.; Blundell, T. L. Structural biology in fragment-based drug design. *Curr. Opin. Struct. Biol.,* **2010**, *20* (4), 497-507.

[25] Bohacek, R.; McMartin, C.; Guida, W. The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.,* **1996**, *16* (1), 3-50.

[26] Fink, T.; Bruggesser, H.; Reymond, J. Virtual exploration of the small-molecule chemical universe below 160 Daltons. *Angew. Chem. Int. Ed. Engl.,* **2005**, *44* (10), 1504-1508.

[27] Borsi, V.; Calderone, V.; Fragai, M.; Luchinat, C.; Sarti, N. Entropic contribution to the linking coefficient in fragment based drug design: A case study. *J. Med. Chem.,* **2010**, *53* (10), 4285-4289.

[28] Murray, C.; Verdonk, M. The consequences of translational and rotational entropy lost by small molecules on binding to proteins. *J. Comput. Aided Mol. Des.,* **2002**, *16* (10), 741-753.

[29] Freire, E. A thermodynamic approach to the affinity optimization of drug candidates. *Chemical Biology & Drug Design,* **2009**, 468-472.

[30] (a) Gilson, M. K.; Zhou, H. X. Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.,* **2007**, *36*, 21-42; (b) Ajay; Murcko, M. A. Computational methods to predict binding free energy in ligand-receptor complexes. *J. Med. Chem.,* **1995**, *38* (26), 4953-4967.

[31] Rohrig, C. H.; Loch, C.; Guan, J. Y.; Siegal, G.; Overhand, M. Fragment-based synthesis and SAR of modified FKBP ligands: Influence of different linking on binding affinity. *Chemmedchem,* **2007**, *2* (7), 1054-1070.

[32] Barker, J. J.; Barker, O.; Courtney, S. M.; Gardiner, M.; Hesterkamp, T.; Ichihara, O.; Mather, O.; Montalbetti, C. A.;

Müller, A.; Varasi, M.; Whittaker, M.; Yarnold, C. J. Discovery of a novel Hsp90 inhibitor by fragment linking. *ChemMedChem,* **2010**, *5* (10), 1697-700.

[33] Kuntz, I.; Chen, K.; Sharp, K.; Kollman, P. The maximal affinity of ligands. *Proc. Natl. Acad. Sci. USA,* **1999**, *96* (18), 9997-10002.

[34] Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand efficiency: a useful metric for lead selection. *Drug Discov. Today,* **2004**, *9* (10), 430-431.

[35] Reynolds, C.; Bembenek, S.; Tounge, B. The role of molecular size in ligand efficiency. *Bioorg. Med. Chem. Lett.,* **2007**, 4258-4261.

[36] Leeson, P.; Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discov.,* **2007**, *6* (11), 881-890.

[37] (a) Abad-Zapatero, C.; Metz, J. T. Ligand efficiency indices as guideposts for drug discovery. *Drug Discov. Today,* **2005**, *10* (7), 464-469; (b) Abad-Zapatero, C.; Perišić, O.; Wass, J.; Bento, A. P.; Overington, J.; Al-Lazikani, B.; Johnson, M. E. Ligand efficiency indices for an effective mapping of chemico-biological space: the concept of an atlas-like representation. *Drug Discov. Today,* **2010**, *15* (19-20), 804-811.

[38] Verdonk, M.; Rees, D. Group efficiency: a guideline for hits-to-leads chemistry. *ChemMedChem,* **2008**, *3* (8), 1179-1180.

[39] Saxty, G.; Woodhead, S. J.; Berdini, V.; Davies, T. G.; Verdonk, M. L.; Wyatt, P. G.; Boyle, R. G.; Barford, D.; Downham, R.; Garrett, M. D.; Carr, R. A. Identification of inhibitors of protein kinase B using fragment-based lead discovery. *J. Med. Chem.,* **2007**, *50* (10), 2293-2296.

[40] Hajduk, P. Fragment-based drug design: how big is too big? *J. Med. Chem.,* **2006**, *49* (24), 6972-6976.

[41] (a) Zartler, E. R.; Shapiro, M. J. Fragonomics: fragment-based drug discovery. *Curr. Opin. Chem. Biol.,* **2005**, *9* (4), 366-370; (b) Hajduk, P. J.; Greer, J. A decade of fragment-based drug design: strategic advances and lessons learned. *Nat. Rev. Drug Discovery,* **2007**, *6* (3), 211-219; (c) Hesterkamp, T.; Whittaker, M. Fragment-based activity space: smaller is better. *Curr. Opin. Chem. Biol.,* **2008**, *12* (3), 260-268; (d) Warr, W. A. Fragment-based drug discovery. *J. Comput-Aided Mol. Des.,* **2009**, *23* (8), 453-458; (e) Chessari, G.; Woodhead, A. J. From fragment to clinical candidate-a historical perspective. *Drug Discov. Today,* **2009**, *14* (13-14), 668-675; (f) Gozalbes, R.; Carbajo, R. J.; Pineda-Lucena, A. Contributions of computational chemistry and biophysical techniques to fragment-based drug discovery. *Curr. Med. Chem.,* **2010**, *17* (17), 1769-1794; (g) Ciulli, A.; Abell, C. Fragment-based approaches to enzyme inhibition. *Curr. Opin. Biotechnol.,* **2007**, *18* (6), 489-496; (h) Erlanson, D. A. Fragment-based lead discovery: a chemical update. *Curr. Opin. Biotechnol.,* **2006**, *17* (6), 643-652; (i) Leach, A.; Hann, M.; Burrows, J.; Griffen, E. Fragment screening: an introduction. *Mol. Biosyst.,* **2006**, *2* (9), 430-446; (j) Murray, C. W.; Rees, D. C. The rise of fragment-based drug discovery. *Nature Chemistry,* **2009**, *1* (3), 187-192; (k) Coyne, A. G.; Scott, D. E.; Abell, C. Drugging challenging targets using fragment-based approaches. *Curr. Opin. Chem. Biol.,* **2010**, *14* (3), 299-307.

[42] (a) Jahnke, W.; Erlanson, D., *Fragment-Based Approaches in Drug Discovery,* **2007**; (b) Jhoti, H.; Leach, A., Structure Based Drug DiscoVery. Springer, Ed. **2007**; (c) (Editor), H. R. E. Structure-Based Drug Discovery. *RSC Biomolecular Sciences* **2006**.

[43] de Kloe, G. E.; Bailey, D.; Leurs, R.; de Esch, I. J. P. Transforming fragments into candidates: small becomes big in medicinal chemistry. *Drug Discov. Today,* **2009**, *14* (13-14), 630-646.

[44] (a) Loving, K.; Alberts, I.; Sherman, W. Computational approaches for fragment-based and De novo design. *Curr. Top. Med. Chem.,* **2010**, *10* (1), 14-32; (b) Vangrevelinghe, E.; Rudisser, S. Computational approaches for fragment optimization. *Curr. Comput-Aided Drug Design.,* **2007**, *3* (1), 69-83.

[45] (a) Siegal, G.; Ab, E.; Schultz, J. Integration of fragment screening and library design. *Drug Discov. Today,* **2007**, *12* (23-24), 1032-1039; (b) Hubbard, R. E.; Chen, I.; Davis, B. Informatics and modeling challenges in fragment-based drug discovery. *Curr. Opin. Drug Discovery Dev.,* **2007**, *10* (3), 289-297; (c) Schuffenhauer, A.; Ruedisser, S.; Marzinzik, A.; Jahnke, W.; Blommers, M.; Selzer, P.; Jacoby, E. Library design for fragment based screening. *Curr. Top Med. Chem.,* **2005**, *5* (8), 751-762.

[46] Carr, R.; Congreve, M.; Murray, C. W.; Rees, D. C. Fragment-based lead discovery: leads by design. *Drug Discov. Today,* **2005**, *10* (14), 987-992.

[47] Wang, Y. S.; Strickland, C.; Voigt, J. H.; Kennedy, M. E.; Beyer, B. M.; Senior, M. M.; Smith, E. M.; Nechuta, T. L.; Madison, V. S.; Czarniecki, M.; McKittrick, B. A.; Stamford, A. W.; Parker, E. M.; Hunter, J. C.; Greenlee, W. J.; Wyss, D. F. Application of fragment-based NMR screening, X-ray crystallography, structure-based design, and focused chemical library design to identify novel mu M leads for the development of nM BACE-1 (beta-Site APP cleaving enzyme 1) inhibitors. *J. Med. Chem.,* **2010**, *53* (3), 942-950.

[48] Neumann, T.; Junker, H.; Schmidt, K.; Sekul, R. SPR-based fragment screening: advantages and applications. *Curr. Top. Med. Chem.,* **2007**, *7* (16), 1630-1642.

[49] (a) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP - Retrosynthetic combinatorial analysis procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.,* **1998**, *38* (3), 511-522; (b) Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M. On the art of compiling and using 'drug-like' chemical fragment spaces. *Chemmedchem,* **2008**, *3* (10), 1503-1507.

[50] Bonachéra, F.; Parent, B.; Barbosa, F.; Froloff, N.; Horvath, D. Fuzzy tricentric pharmacophore fingerprints. 1. Topological fuzzy pharmacophore triplets and adapted molecular similarity scoring schemes. *J. Chem. Inf. Model.*,**2006**, *46* (6), 2457-2477.

[51] Van Drie, J.; Lajiness, M., Approaches to virtual library design. *Drug Discov. Today,* **1998**, 274-283.

[52] Kangas, J. A.; Kohonen, T. K.; Laaksonen, J. T. Variants of self-organizing maps. *IEEE Trans. Neural Netw.,* **1990**, *1* (1), 93-99.

[53] Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. Virtual computational chemistry laboratory--design and description. *J. Comput. Aided Mol. Des.,* **2005**, *19* (6), 453-463.

[54] Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today,* **2006**, 1046-1053.

[55] Colclough, N.; Hunter, A.; Kenny, P.; Kittlety, R.; Lobedan, L.; Tam, K.; Timms, M. High throughput solubility determination with application to selection of compounds for fragment screening. *Bioorg. Med. Chem.,* **2008**, *16* (13), 6611-6616.

[56] Boda, K.; Seidel, T.; Gasteiger, J. Structure and reaction based evaluation of synthetic accessibility. *J. Comput. Aided Mol. Des.,* **2007**, *21* (6), 311-325.

[57] Derwent, World Drug Index (WDI), Thomson. *World Drug Index (WDI), Thomson*.

[58] Chen, I. J.; Hubbard, R. E. Lessons for fragment library design: analysis of output from multiple screening campaigns. *J. Comput-Aided Mol. Des.,* **2009**, *23* (8), 603-620.

[59] Hajduk, P.; Bures, M.; Praestgaard, J.; Fesik, S. Privileged molecules for protein binding identified from NMR-based screening. *J. Med. Chem.,* **2000**, *43* (18), 3443-3447.

[60] Barelier, S.; Pons, J.; Gehring, K.; Lancelin, J. M.; Krimm, I. Ligand specificity in fragment-based drug design. *J. Med. Chem.,* **2010**, *53* (14), 5256-5266.

[61] Clemons, P. A.; Bodycombe, N. E.; Carrinski, H. A.; Wilson, J. A.; Shamji, A. F.; Wagner, B. K.; Koehler, A. N.; Schreiber, S. L. Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. *Proc. Natl. Acad. Sci. USA,* **2010**, *107* (44), 18787-18792.

[62] Lovering, F.; Bikker, J.; Humblet, C. Escape from flatland: increasing saturation as an approach to improving clinical success. *J. Med. Chem.,* **2009**, *52* (21), 6752-6756.

[63] (a) Fejzo, J.; Lepre, C. A.; Peng, J. W.; Bemis, G. W.; Ajay; Murcko, M. A.; Moore, J. M. The SHAPES strategy: an NMR-based approach for lead generation in drug discovery. *Chem. Biol.,* **1999**, *6* (10), 755-769; (b) Lepre, C. Library design for NMR-based screening. *Drug Discov. Today,* **2001**, *6* (3), 133-140.

[64] Gianti, E.; Sartori, L. Identification and Selection of "Privileged Fragments" Suitable for Primary Screening. *J. Chem. Inf. Model.,* **2008**, *48* (11), 2129-2139.

[65] Blomberg, N.; Cosgrove, D. A.; Kenny, P. W.; Kolmodin, K. Design of compound libraries for fragment screening. *J. Comput-Aided Mol. Des.,* **2009**, *23* (8), 513-525.

[66] Smits, R. A.; Lim, H. D.; Hanzer, A.; Zuiderveld, O. P.; Guaita, E.; Adami, M.; Coruzzi, G.; Leurs, R.; de Esch, L. J. P. Fragment based design of new H-4 receptor-ligands with anti-inflammatory properties *in vivo*. *J. Med. Chem.,* **2008**, *51* (8), 2457-2467.

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

[68] Gozalbes, R.; Mosulén, S.; Carbajo, R.; Pineda-Lucena, A. Development and NMR validation of minimal pharmacophore hypotheses for the generation of fragment libraries enriched in heparanase inhibitors. *J. Comput. Aided Mol. Des.,* **2009**, *23*(8), 555-569.

[69] (a) Sun, H. A universal molecular descriptor system for prediction of LogP, LogS, LogBB, and absorption. *J. Chem. Inf. Comput. Sci.,* **2004**, 748-757; (b) Wildman, S.; Crippen, G. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.,* **1999**, 868-873; (c) Xing, L.; Glen, R. Novel methods for the prediction of logP, pK(a), and logD. *J. Chem. Inf. Comput. Sci.,* **2002**, 796-805.

[70] (a) Bruneau, P. Search for predictive generic model of aqueous solubility using Bayesian neural nets. *J. Chem. Inf. Comput. Sci.,* **2001**, *41* (6), 1605-1616; (b) Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. *J. Comput. Aided Mol. Des.,* **2005**, *19* (9-10), 693-703; (c) Butina, D.; Gola, J. Modeling aqueous solubility. *J. Chem. Inf. Comput. Sci.,* **2003**, *43* (3), 837-841; (d) Engkvist, O.; Wrede, P. High-throughput, *in silico* prediction of aqueous solubility based on one- and two-dimensional descriptors. *J. Chem. Inf. Comput. Sci.,* **2002**, *42* (5), 1247-1249; (e) Hou, T.; Xia, K.; Zhang, W.; Xu, X. ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach. *J. Chem. Inf. Comput. Sci.,* **2004**, *44* (1), 266-275; (f) Yan, A.; Gasteiger, J. Prediction of aqueous solubility of organic compounds based on a 3D structure representation. *J. Chem. Inf. Comput. Sci.,* **2003**, *43* (2), 429-434.

[71] Clark, M. Generalized fragment-substructure based property prediction method. *J. Chem. Inf. Model.,* **2005**, *45* (1), 30-38.

[72] Horvath, D.; Marcou, G.; Varnek, A. Predicting the predictability: a unified approach to the applicability domain problem of QSAR models. *J. Chem. Inf. Model.,* **2009**, *49* (7), 1762-1776.

[73] (a) Faller, B.; Ertl, P. Computational approaches to determine drug solubility. *Adv. Drug Deliv. Rev.,* **2007**, *59* (7), 533-545; (b) Delaney, J. Predicting aqueous solubility from structure. *Drug Discov. Today,* **2005**, *10* (4), 289-295.

[74] Lamanna, C.; Bellini, M.; Padova, A.; Westerberg, G.; Maccari, L. Straightforward recursive partitioning model for discarding insoluble compounds in the drug discovery process. *J. Med. Chem.,* **2008**, *51* (10), 2891-2897.

[75] Crisman, T.; Bender, A.; Milik, M.; Jenkins, J.; Scheiber, J.; Sukuru, S.; Fejzo, J.; Hommel, U.; Davies, J.; Glick, M. "Virtual fragment linking": an approach to identify potent binders from low affinity fragment hits. *J. Med. Chem.,* **2008**, *51* (8), 2481-2491.

[76] Manoharan, P.; Vijayan, R. S.; Ghoshal, N. Rationalizing fragment based drug discovery for BACE1: insights from FB-QSAR, FB-QSSR, multi objective (MO-QSPR) and MIF studies. *J. Comput. Aided Mol. Des.,* **2010**, *24* (10), 843-864.

[77] Kubinyi, H., *QSAR: Hansch Analysis and Related Approaches*. 1993.

[78] (a) Warren, G.; Andrews, C.; Capelli, A.; Clarke, B.; LaLonde, J.; Lambert, M.; Lindvall, M.; Nevins, N.; Semus, S.; Senger, S.; Tedesco, G.; Wall, I.; Woolven, J.; Peishoff, C.; Head, M. A critical assessment of docking programs and scoring functions. *J. Med. Chem.,* **2006**, *49* (20), 5912-5931; (b) Brooijmans, N.; Kuntz, I. D. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.,* **2003**, *32*, 335-373.

[79] Charifson, P.; Corkery, J.; Murcko, M.; Walters, W. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.,* **1999**, *42* (25), 5100-5109.

[80] Brewerton, S. C. The use of protein-ligand interaction fingerprints in docking. *Curr. Opin. Drug Discov. Devel.,* **2008**, *11* (3), 356-364.

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

[82] MOE, Molecular Operating Environment. http://www.chemcomp.com Chemical Computing Group Inc.

[83] Gleeson, M. P.; Gleeson, D., QM/MM As a Tool in Fragment Based Drug Discovery. A Cross-Docking, Rescoring Study of Kinase Inhibitors. *J. Chem. Inf. Model.,* **2009**, *49* (6), 1437-1448.

[84] Martin, E. J.; Sullivan, D. C. Surrogate AutoShim: predocking into a universal ensemble kinase receptor for three dimensional activity prediction, very quickly, without a crystal structure. *J. Chem. Inf. Model.,* **2008**, *48* (4), 873-881.

[85] Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.,* **2004**, *47* (7), 1739-1749.

[86] Loving, K.; Salam, N. K.; Sherman, W. Energetic analysis of fragment docking and application to structure-based pharmacophore hypothesis generation. *J. Comput-Aided Mol. Des.,* **2009**, *23* (8), 541-554.

[87] Kawatkar, S.; Wang, H. M.; Czerminski, R.; Joseph-McCarthy, D. Virtual fragment screening: an exploration of various docking and scoring protocols for fragments using Glide. *J. Comput-Aided Mol. Des.,* **2009**, *23* (8), 527-539.

[88] Kuntz, I.; Blaney, J.; Oatley, S.; Langridge, R.; Ferrin, T. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.,* **1982**, *161* (2), 269-288.

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

[90] Jones, G.; Willett, P.; Glen, R.; Leach, A.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.,* **1997**, *267* (3), 727-748.

[91] Goodsell, D. S.; Morris, G. M.; Olson, A. J. Automated docking of flexible ligands: Applications of AutoDock. *J. Mol. Recognit.,* **1996**, *9* (1), 1-5.

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

[93] Agamennone, M.; Cesari, L.; Lalli, D.; Turlizzi, E.; Del Conte, R.; Turano, P.; Mangani, S.; Padova, A. Fragmenting the S100B-p53 interaction: Combined virtual/biophysical screening approaches to identify ligands. *Chemmedchem,* **2010**, *5* (3), 428-435.

[94] Chen, Y.; Shoichet, B. K., Molecular docking and ligand specificity in fragment-based inhibitor discovery. *Nature Chemical Biology* **2009**, *5* (5), 358-364.

[95] Teotico, D.; Babaoglu, K.; Rocklin, G.; Ferreira, R.; Giannetti, A.; Shoichet, B. Docking for fragment inhibitors of AmpC beta-lactamase. *Proc. Natl. Acad. Sci. U SA,* **2009**, *106* (18), 7455-7460.

[96] Zhang, Y. J.; Wang, Z. L.; Sprous, D.; Nabioullin, R. *In silico* design and synthesis of piperazine-1-pyrrolidine-2,5-dione scaffold-based novel malic enzyme inhibitors. *Bioorg. Medicinal Chem. Lett.,* **2006**, *16* (3), 525-528.

[97] Böhm, H. The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *J. Comput. Aided Mol. Des.,* **1992**, *6* (1), 61-78.

[98] Warner, S. L.; Bashyam, S.; Vankayalapati, H.; Bearss, D. J.; Han, H. Y.; Von Hoff, D. D.; Hurley, L. H. Identification of a lead small-molecule inhibitor of the Aurora kinases using a structure-assisted, fragment-based approach. *Mol. Cancer Ther.,* **2006**, *5* (7), 1764-1773.

[99] Englert, L.; Silber, K.; Steuber, H.; Brass, S.; Over, B.; Gerber, H. D.; Heine, A.; Diederich, W. E.; Klebe, G. Fragment-based lead discovery: Screening and optimizing fragments for thermolysin inhibition. *Chemmedchem,* **2010**, *5* (6), 930-940.

[100] Rohrig, U. F.; Awad, L.; Grosdidier, A.; Larrieu, P.; Stroobant, V.; Colau, D.; Cerundolo, V.; Simpson, A. J. G.; Vogel, P.; Van den Eynde, B. J.; Zoete, V.; Michielin, O. Rational design of indoleamine 2,3-dioxygenase inhibitors. *J. Med. Chem.,* **2010**, *53* (3), 1172-1189.

[101] (a) Grosdidier, A.; Zoete, V.; Michielin, O. EADock: docking of small molecules into protein active sites with a multiobjective evolutionary optimization. *Proteins,***2007**, *67* (4), 1010-1025; (b) Grosdidier, A.; Zoete, V.; Michielin, O. Blind docking of 260 protein-ligand complexes with EADock 2.0. *J. Comput. Chem.,* **2009**, *30* (13), 2021-2030.

[102] Huang, D.; Caflisch, A. Library screening by fragment-based docking. *J. Mol. Recognit.,* **2010**, *23* (2), 183-93.

[103] Kolb, P.; Caflisch, A. Automatic and efficient decomposition of two-dimensional structures of small molecules for fragment-based high-throughput docking. *J. Med. Chem.,* **2006**, *49* (25), 7384-7392.

[104] (a) Majeux, N.; Scarsi, M.; Apostolakis, J.; Ehrhardt, C.; Caflisch, A. Exhaustive docking of molecular fragments with electrostatic solvation. *Proteins,* **1999**, *37* (1), 88-105; (b) Majeux, N.; Scarsi, M.; Caflisch, A. Efficient electrostatic solvation model for protein-fragment docking. *Proteins,* **2001**, *42* (2), 256-268.

[105] Budin, N.; Majeux, N.; Caflisch, A. Fragment-Based flexible ligand docking by evolutionary optimization. *Biol. Chem.,* **2001**, *382* (9), 1365-1372.

[106] (a) Huang, D.; Lüthi, U.; Kolb, P.; Edler, K.; Cecchini, M.; Audetat, S.; Barberis, A.; Caflisch, A. Discovery of cell-permeable non-peptide inhibitors of beta-secretase by high-throughput docking and continuum electrostatics calculations. *J. Med. Chem.,* **2005**, *48* (16), 5108-5111; (b) Huang, D. Z.; Luthi, U.; Kolb, P.; Cecchini, M.; Barberis, A.; Caflisch, A. *In silico* discovery of beta-secretase inhibitors. *J. Am. Chem. Soc.,* **2006**, *128* (16), 5436-5443.

[107] Ekonomiuk, D.; Su, X. C.; Ozawa, K.; Bodenreider, C.; Lim, S. P.; Otting, G.; Huang, D. Z.; Caflisch, A. Flaviviral protease inhibitors identified by fragment-based library docking into a structure generated by molecular dynamics. *J. Med. Chem.,* **2009**, *52* (15), 4860-4868.

[108] Friedman, R.; Caflisch, A. Discovery of plasmepsin inhibitors by fragment-based docking and consensus scoring. *Chemmedchem,* **2009**, *4* (8), 1317-1326.

[109] Dakshanamurthy, S.; Kim, M.; Brown, M. L.; Byers, S. W. *In-silico* fragment-based identification of novel angiogenesis inhibitors. *Bioorg. Medicinal Chem. Lett.,* **2007**, *17* (16), 4551-4556.

[110] Hindle, S.; Rarey, M.; Buning, C.; Lengaue, T. Flexible docking under pharmacophore type constraints. *J. Comput. Aided Mol. Des.,* **2002**, *16* (2), 129-149.

[111] Rummey, C.; Nordhoff, S.; Thiemann, M.; Metz, G. *In silico* fragment-based discovery of DPP-IV S1 pocket binders. *Bioorg. Medicinal Chem. Lett.,***2006**, *16* (5), 1405-1409.

[112] Machrouhi, F.; Ouhamou, N.; Laderoute, K.; Calaoagan, J.; Bukhtiyarova, M.; Ehrlich, P. J.; Klon, A. E. The rational design of a novel potent analogue of the 5'-AMP-activated protein kinase inhibitor compound C with improved selectivity and cellular activity. *Bioorg. Med. Chem. Lett.,* **2010**, *20* (22), 6394-6349.

[113] Schneider, G.; Lee, M.; Stahl, M.; Schneider, P. De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. *J. Comput. Aided Mol. Des.,* **2000**, *14* (5), 487-494.

[114] Proschak, E.; Sander, K.; Zettl, H.; Tanrikulu, Y.; Rau, O.; Schneider, P.; Schubert-Zsilavecz, M.; Stark, H.; Schneider, G., From Molecular Shape to Potent Bioactive Agents II: Fragment-Based de novo Design. *Chemmedchem,* **2009**, *4* (1), 45-48.

[115] Schneider, G.; Fechner, U., Computer-based de novo design of drug-like molecules. *Nature Reviews Drug Discovery* **2005**, *4* (8), 649-663.

[116] (a) Miranker, A.; Karplus, M. Functionality maps of binding sites: a multiple copy simultaneous search method. *Proteins,* **1991**, *11* (1), 29-34; (b) Schubert, C. R.; Stultz, C. M. The multi-copy simultaneous search methodology: a fundamental tool for structure-based drug design. *J. Comput-Aided Mol. Des.,* **2009**, *23* (8), 475-489.

[117] Mata, P.; Gillet, V.; Johnson, A.; Lampreia, J.; Myatt, G.; Sike, S.; Stebbings, A. SPROUT - 3D structure generation using templates. *J. Chem. Inf. Comput. Sci.,* **1995**, 479-493.

[118] Imai, T.; Oda, K.; Kovalenko, A.; Hirata, F.; Kidera, A. Ligand mapping on protein surfaces by the 3D-RISM theory: toward computational fragment-based drug design. *J. Am. Chem. Soc.,* **2009**, *131* (34), 12430-12440.

[119] Brenke, R.; Kozakov, D.; Chuang, G. Y.; Beglov, D.; Hall, D.; Landon, M. R.; Mattos, C.; Vajda, S. Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics,* **2009**, *25* (5), 621-627.

[120] Ivetac, A.; McCammon, J. A., Mapping the druggable allosteric space of G-protein coupled receptors: a fragment-based molecular dynamics approach. *Chem Biol Drug Des* **2010**, *76* (3), 201-17.

[121] Fukunishi, Y.; Mashimo, T.; Orita, M.; Ohno, K.; Nakamura, H. *In silico* fragment screening by replica generation (FSRG) method for fragment-based drug design. *J. Chem. Inf. Model.,* **2009**, *49* (4), 925-933.

[122] Huang, N.; Shoichet, B.; Irwin, J., Benchmarking sets for molecular docking. *J. Med. Chem.,* **2006**, 6789-6801.

[123] (a) Takahashi, K.; Ikura, M.; Habashita, H.; Nishizaki, M.; Sugiura, T.; Yamamoto, S.; Nakatani, S.; Ogawa, K.; Ohno, H.; Nakai, H.; Toda, M. Novel matrix metalloproteinase inhibitors: Generation of lead compounds by the in silico fragment-based approach. *Bioorg. Medicinal Chem.,* **2005**, *13* (14), 4527-4543; (b) Boehm, H.; Boehringer, M.; Bur, D.; Gmuender, H.; Huber, W.; Klaus, W.; Kostrewa, D.; Kuehne, H.; Luebbers, T.; Meunier-Keller, N.; Mueller, F. Novel inhibitors of DNA gyrase: 3D structure based biased needle screening, hit validation by biophysical methods, and 3D guided optimization. A promising alternative to random screening. *J. Med. Chem.,* **2000**, *43* (14), 2664-2674; (c) Oblak, M.; Grdadolnik, S. G.; Kotnik, M.; Jerala, R.; Filipic, M.; Solmajer, T. *In silico* fragment-based discovery of indolin-2-one analogues as potent DNA gyrase inhibitors. *Bioorg. Medicinal Chem. Lett.,* **2005**, *15* (23), 5207-5210.

[124] Zaliani, A.; Boda, K.; Seidel, T.; Herwig, A.; Schwab, C. H.; Gasteiger, J.; Claussen, H.; Lemmen, C.; Degen, J.; Parn, J.; Rarey, M. Second-generation de novo design: a view from a medicinal chemist perspective. *J. Comput-Aided Mol. Des.,* **2009**, *23* (8), 593-602.

[125] Maass, P.; Schulz-Gasch, T.; Stahl, M.; Rarey, M. Recore: a fast and versatile method for scaffold hopping based on small molecule crystal structure conformations. *J. Chem. Inf. Model.,* **2007**, *47* (2), 390-399.

[126] Moriaud, F.; Doppelt-Azeroual, O.; Martin, L.; Oguievetskaia, K.; Koch, K.; Vorotyntsev, A.; Adcock, S.; Delfaud, F., Computational fragment-based approach at PDB scale by protein local similarity. *J. Chem. Inf. Model.,* **2009**, *49*(2), 280-294.

[127] Oguievetskaia, K.; Martin-Chanas, L.; Vorotyntsev, A.; Doppelt-Azeroual, O.; Brotel, X.; Adcock, S.; de Brevern, A.; Delfaud, F.; Moriaud, F. Computational fragment-based drug design to explore the hydrophobic sub-pocket of the mitotic kinesin Eg5 allosteric binding site. *J. Comput. Aided Mol. Des.,* **2009**, *23*(8), 571-582.

[128] Dey, F.; Caflisch, A. Fragment-based de Novo ligand design by multiobjective evolutionary optimization. *J. Chem. Inf. Model.,* **2008**, *48* (3), 679-690.

[129] Thompson, D.; Denny, R.; Nilakantan, R.; Humblet, C.; Joseph-McCarthy, D.; Feyfant, E. CONFIRM: connecting fragments found in receptor molecules. *J. Comput. Aided Mol. Des.,* **2008**, *22* (10), 761-772.

[130] (a) McGovern, S.; Caselli, E.; Grigorieff, N.; Shoichet, B. A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J. Med. Chem.,* **2002**, *45* (8), 1712-1722; (b) Rishton, G. Nonleadlikeness and leadlikeness in biochemical screening. *Drug Discov. Today,* **2003**, *8* (2), 86-96; (c) Rishton, G. Reactive compounds and *in vitro* false positives in HTS. *Drug Discov. Today,* **1997**, 382-384.

[131] Boettcher, A.; Ruedisser, S.; Erbel, P.; Vinzenz, D.; Schiering, N.; Hassiepen, U.; Rigollier, P.; Mayr, L. M.; Woelcke, J. Fragment-Based Screening by Biochemical Assays. *Journal of Biomolecular Screening* **2010**, *15* (9), 1029-1041.

[132] Barker, J.; Courtney, S.; Hesterkamp, T.; Ullmann, D.; Whittaker, M. Fragment screening by biochemical assay. *Expert Opinion on Drug Discovery* **2006**, 225-236.

[133] Artis, D.; Lin, J.; Zhang, C.; Wang, W.; Mehra, U.; Perreault, M.; Erbe, D.; Krupka, H.; England, B.; Arnold, J.; Plotnikov, A.; Marimuthu, A.; Nguyen, H.; Will, S.; Signaevsky, M.; Kral, J.; Cantwell, J.; Settachatgull, C.; Yan, D.; Fong, D.; Oh, A.; Shi, S.; Womack, P.; Powell, B.; Habets, G.; West, B.; Zhang, K.; Milburn, M.; Vlasuk, G.; Hirth, K.; Nolop, K.; Bollag, G.; Ibrahim, P.; Tobin, J. Scaffold-based discovery of indeglitazar, a PPAR pan-active anti-diabetic agent. *Proc. Natl. Acad. Sci. USA,* **2009**, *106* (1), 262-267.

[134] Tsai, J.; Lee, J. T.; Wang, W.; Zhang, J.; Cho, H.; Mamo, S.; Bremer, R.; Gillette, S.; Kong, J.; Haass, N. K.; Sproesser, K.; Li, L.; Smalley, K. S.; Fong, D.; Zhu, Y. L.; Marimuthu, A.; Nguyen, H.; Lam, B.; Liu, J.; Cheung, I.; Rice, J.; Suzuki, Y.; Luu, C.; Settachatgul, C.; Shellooe, R.; Cantwell, J.; Kim, S. H.; Schlessinger, J.; Zhang, K. Y.; West, B. L.; Powell, B.; Habets, G.; Zhang, C.; Ibrahim, P. N.; Hirth, P.; Artis, D. R.; Herlyn, M.; Bollag, G. Discovery of a selective inhibitor of oncogenic B-Raf

kinase with potent antimelanoma activity. *Proc. Natl. Acad. Sci. USA,* **2008**, *105* (8), 3041-3046.

[135] Godemann, R.; Madden, J.; Krämer, J.; Smith, M.; Fritz, U.; Hesterkamp, T.; Barker, J.; Höppner, S.; Hallett, D.; Cesura, A.; Ebneth, A.; Kemp, J. Fragment-based discovery of BACE1 inhibitors using functional assays. *Biochemistry,* **2009**, *48* (45), 10743-10751.

[136] Barker, J. J.; Barker, O.; Boggio, R.; Chauhan, V.; Cheng, R. K. Y.; Corden, V.; Courtney, S. M.; Edwards, N.; Falque, V. M.; Fusar, F.; Gardiner, M.; Hamelin, E. M. N.; Hesterkamp, T.; Ichihara, O.; Jones, R. S.; Mather, O.; Mercurio, C.; Minucci, S.; Montalbetti, C.; Muller, A.; Patel, D.; Phillips, B. G.; Varasi, M.; Whittaker, M.; Winkler, D.; Yarnold, C. J. Fragment-based identification of Hsp90 inhibitors. *Chemmedchem.,* **2009**, *4* (6), 963-966.

[137] (a) Nienaber, V. L.; Richardson, P. L.; Klighofer, V.; Bouska, J. J.; Giranda, V. L.; Greer, J. Discovering novel ligands for macromolecules using X-ray crystallographic screening. *Nature Biotechnology* **2000**, *18* (10), 1105-1108; (b) Hartshorn, M. J.; Murray, C. W.; Cleasby, A.; Frederickson, M.; Tickle, I. J.; Jhoti, H. Fragment-based lead discovery using X-ray crystallography. *J. Med. Chem.,* **2005**, *48* (2), 403-413; (c) Verlinde, C.; Kim, H.; Bernstein, B.; Mande, S.; Hol, W. *Anti-trypanosomiasis drug development based on structures of glycolytic enzymes. In: Veerapandian, P., editor. Structure-based drug design. Marcel Dekker; Inc; New York: 1997. p. 365-394*.

[138] Blundell, T.; Jhoti, H.; Abell, C. High-throughput crystallography for lead discovery in drug design. *Nat. Rev. Drug Discov.,* **2002**, *1* (1), 45-54.

[139] (a) Mooij, W.; Hartshorn, M.; Tickle, I.; Sharff, A.; Verdonk, M.; Jhoti, H. Automated protein-ligand crystallography for structure-based drug design. *Chemmedchem.,* **2006**, 827-838; (b) Jhoti, H.; Cleasby, A.; Verdonk, M.; Williams, G. Fragment-based screening using X-ray crystallography and NMR spectroscopy. *Current Opinion in Chemical Biology* **2007**, *11* (5), 485-493.

[140] Sanders, W.; Nienaber, V.; Lerner, C.; McCall, J.; Merrick, S.; Swanson, S.; Harlan, J.; Stoll, V.; Stamper, G.; Betz, S.; Condroski, K.; Meadows, R.; Severin, J.; Walter, K.; Magdalinos, P.; Jakob, C.; Wagner, R.; Beutel, B. Discovery of potent inhibitors of dihydroneopterin aldolase using crystaLEAD high-throughput X-ray crystallographic screening and structure-directed lead optimization. *J. Med. Chem.,* **2004**, 1709-1718.

[141] Gill, A. L.; Frederickson, M.; Cleasby, A.; Woodhead, S. J.; Carr, M. G.; Woodhead, A. J.; Walker, M. T.; Congreve, M. S.; Devine, L. A.; Tisi, D.; O'Reilly, M.; Seavers, L. C. A.; Davis, D. J.; Curry, J.; Anthony, R.; Padova, A.; Murray, C. W.; Carr, R. A. E.; Jhoti, H. Identification of novel p38 alpha MAP kinase inhibitors using fragment-based lead generation. *J. Med. Chem.,* **2005**, *48* (2), 414-426.

[142] Bosch, J.; Robien, M.; Mehlin, C.; Boni, E.; Riechers, A.; Buckner, F.; Van Voorhis, W.; Myler, P.; Worthey, E.; DeTitta, G.; Luft, J.; Lauricella, A.; Gulde, S.; Anderson, L.; Kalyuzhniy, O.; Neely, H.; Ross, J.; Earnest, T.; Soltis, M.; Schoenfeld, L.; Zucker, F.; Merritt, E.; Fan, E.; Verlinde, C.; Hol, W. Using fragment cocktail crystallography to assist inhibitor design of Trypanosoma brucei nucleoside 2-deoxyribosyltransferase. *J. Med. Chem.,* **2006**, 5939-5946.

[143] Murray, C. W.; Callaghan, O.; Chessari, G.; Cleasby, A.; Congreve, M.; Frederickson, M.; Hartshorn, M. J.; McMenamin, R.; Patel, S.; Wallis, N. Application of fragment screening by X-ray crystallography to beta-secretase. *J. Med. Chem.,* **2007**, *50* (6), 1116-1123.

[144] Wyatt, P. G.; Woodhead, A. J.; Berdini, V.; Boulstridge, J. A.; Carr, M. G.; Cross, D. M.; Davis, D. J.; Devine, L. A.; Early, T. R.; Feltell, R. E.; Lewis, E. J.; McMenamin, R. L.; Navarro, E. F.; O'Brien, M. A.; O'Reilly, M.; Reule, M.; Saxty, G.; Seavers, L. C. A.; Smith, D. M.; Squires, M. S.; Trewartha, G.; Walker, M. T.; Woolford, A. J. A. Identification of N-(4-piperidinyl)-4-(2,6-dichlorobenzoylamino)-1H-pyrazole-3-carboxamide (AT7519), a novel cyclin dependent kinase inhibitor using fragment-based X-ray crystallography and structure based drug design. *J. Med. Chem.,* **2008**, *51* (16), 4986-4999.

[145] Antonysamy, S. S.; Aubol, B.; Blaney, J.; Browner, M. F.; Giannetti, A. M.; Harris, S. F.; Hebert, N.; Hendle, J.; Hopkins, S.; Jefferson, E.; Kissinger, C.; Leveque, V.; Marciano, D.; McGee, E.; Najera, I.; Nolan, B.; Tomimoto, M.; Torres, E.; Wright, T. Fragment-based discovery of hepatitis C virus NS5b RNA polymerase inhibitors. *Bioorg. Medicinal Chem. Lett.,* **2008**, *18* (9), 2990-2995.

[146] Antonysamy, S.; Hirst, G.; Park, F.; Sprengeler, P.; Stappenbeck, F.; Steensma, R.; Wilson, M.; Wong, M. Fragment-based discovery of JAK-2 inhibitors. *Bioorg. Medicinal Chem. Lett.,* **2009**, *19* (1), 279-282.

[147] Schulz, M. N.; Fanghanel, J.; Schafer, M.; Badock, V.; Briem, H.; Boemer, U.; Nguyen, D.; Husemann, M.; Hillig, R. C. A crystallographic fragment screen identifies cinnamic acid derivatives as starting points for potent Pim-1 inhibitors. *Acta Crystallographica Section D,* **2011**, *67* (3), 156-166.

[148] Wüthrich, K., *NMR of Proteins and Nucleic Acids (Wiley, 1986)*.

[149] Pellecchia, M.; Sem, D. S.; Wüthrich, K. NMR in drug discovery. *Nat. Rev. Drug Discov.,* **2002**, *1* (3), 211-219.

[150] (a) Pellecchia, M.; Bertini, I.; Cowburn, D.; Dalvit, C.; Giralt, E.; Jahnke, W.; James, T.; Homans, S.; Kessler, H.; Luchinat, C.; Meyer, B.; Oschkinat, H.; Peng, J.; Schwalbe, H.; Siegal, G. Perspectives on NMR in drug discovery: a technique comes of age. *Nat. Rev. Drug Discov.,* **2008**, *7* (9), 738-745; (b) Jahnke, W. Perspectives of biomolecular NMR in drug discovery: the blessing and curse of versatility. *J. Biomol. NMR,* **2007**, *39* (2), 87-90.

[151] Stockman, B.; Dalvit, C., NMR screening techniques in drug discovery and drug design. *Prog. N MR Spec.,* **2002**, 187-231.

[152] Russell, D.; Hadden, C.; Martin, G.; Gibson, A.; Zens, A.; Carolan, J. A comparison of inverse-detected heteronuclear NMR performance: conventional *vs* cryogenic microprobe performance. *J. Nat. Prod.,* **2000**, *63* (8), 1047-1049.

[153] Hajduk, P.; Gerfin, T.; Boehlen, J.; Häberli, M.; Marek, D.; Fesik, S. High-throughput nuclear magnetic resonance-based screening. *J. Med. Chem.,* **1999**, *42* (13), 2315-2317.

[154] Mayer, M.; Meyer, B. Characterization of ligand binding by saturation transfer difference NMR spectroscopy. *Angewandte Chemie-International Edition,* **1999**, 1784-1788.

[155] Dalvit, C.; Pevarello, P.; Tatò, M.; Veronesi, M.; Vulpetti, A.; Sundström, M. Identification of compounds with binding affinity to proteins *via* magnetization transfer from bulk water. *J. Biomol. NMR,* **2000**, *18* (1), 65-68.

[156] Wang, Y.; Liu, D.; Wyss, D. Competition STD NMR for the detection of high-affinity ligands and NMR-based screening. *Magnetic Resonance in Chem.,* **2004**, 485-489.

[157] Dalvit, C.; Fagerness, P.; Hadden, D.; Sarver, R.; Stockman, B. Fluorine-NMR experiments for high-throughput screening: theoretical aspects, practical considerations, and range of applicability. *J. Am. Chem. Soc.,* **2003**, *125* (25), 7696-703.

[158] Taylor, J. D.; Gilbert, P. J.; Williams, M. A.; Pitt, W. R.; Ladbury, J. E., Identification of novel fragment compounds targeted against the pY pocket of v-Src SH2 by computational and NMR screening and thermodynamic evaluation. *Proteins-Structure Function and Bioinformatics* **2007**, *67* (4), 981-990.

[159] Vanwetswinkel, S.; Heetebrij, R. J.; van Duynhoven, J.; Hollander, J. G.; Filippov, D. V.; Hajduk, P. J.; Siegal, G. TINS, target immobilized NMR screening: an efficient and sensitive method for ligand discovery. *Chem. Biol.,* **2005**, *12* (2), 207-216.

[160] Kobayashi, M.; Retra, K.; Figaroa, F.; Hollander, J.; Ab, E.; Heetebrij, R.; Irth, H.; Siegal, G. Target immobilization as a strategy for NMR-based fragment screening: comparison of TINS, STD, and SPR for fragment hit identification. *J. Biomol. Screen.,* **2010**, *15* (8), 978-989.

[161] Früh, V.; Zhou, Y.; Chen, D.; Loch, C.; Ab, E.; Grinkova, Y. N.; Verheij, H.; Sligar, S. G.; Bushweller, J. H.; Siegal, G. Application of fragment-based drug discovery to membrane proteins: identification of ligands of the integral membrane enzyme DsbB. *Chem. Biol.,* **2010**, *17* (8), 881-891.

[162] Dalvit, C. NMR methods in fragment screening: theory and a comparison with other biophysical techniques. *Drug Discov. Today,* **2009**, *14* (21-22), 1051-1057.

[163] Hajduk, P. J.; Sheppard, G.; Nettesheim, D. G.; Olejniczak, E. T.; Shuker, S. B.; Meadows, R. P.; Steinman, D. H.; Carrera, G. M.; Marcotte, P. A.; Severin, J.; Walter, K.; Smith, H.; Gubbins, E.; Simmer, R.; Holzman, T. F.; Morgan, D. W.; Davidsen, S. K.; Summers, J. B.; Fesik, S. W. Discovery of potent nonpeptide inhibitors of stromelysin using SAR by NMR. *J. Am. Chem. Soc.,* **1997**, *119* (25), 5818-5827.

[164] Szczepankiewicz, B.; Liu, G.; Hajduk, P.; Abad-Zapatero, C.; Pei, Z.; Xin, Z.; Lubben, T.; Trevillyan, J.; Stashko, M.; Ballaron, S.;

Liang, H.; Huang, F.; Hutchins, C.; Fesik, S.; Jirousek, M. Discovery of a potent, selective protein tyrosine phosphatase 1B inhibitor using a linked-fragment strategy. *J. Am. Chem. Soc.,* **2003**, *125* (14), 4087-4096.

[165] Oltersdorf, T.; Elmore, S.; Shoemaker, A.; Armstrong, R.; Augeri, D.; Belli, B.; Bruncko, M.; Deckwerth, T.; Dinges, J.; Hajduk, P.; Joseph, M.; Kitada, S.; Korsmeyer, S.; Kunzer, A.; Letai, A.; Li, C.; Mitten, M.; Nettesheim, D.; Ng, S.; Nimmer, P.; O'Connor, J.; Oleksijew, A.; Petros, A.; Reed, J.; Shen, W.; Tahir, S.; Thompson, C.; Tomaselli, K.; Wang, B.; Wendt, M.; Zhang, H.; Fesik, S.; Rosenberg, S. An inhibitor of Bcl-2 family proteins induces regression of solid tumours. *Nature,* **2005**, *435* (7042), 677-681.

[166] Huth, J. R.; Park, C.; Petros, A. M.; Kunzer, A. R.; Wendt, M. D.; Wang, X. L.; Lynch, C. L.; Mack, J. C.; Swift, K. M.; Judge, R. A.; Chen, J.; Richardson, P. L.; Jin, S.; Tahir, S. K.; Matayoshi, E. D.; Dorwin, S. A.; Ladror, U. S.; Severin, J. M.; Walter, K. A.; Bartley, D. M.; Fesik, S. W.; Elmore, S. W.; Hajduk, P. J. Discovery and design of novel HSP90 inhibitors using multiple fragment-based design strategies. *Chemical Biology & Drug Design* **2007**, *70*, 1-12.

[167] Hohwy, M.; Spadola, L.; Lundquist, B.; Hawtin, P.; Dahmen, J.; Groth-Clausen, I.; Nilsson, E.; Persdotter, S.; Von Wachenfeldt, K.; Folmer, R. H. A.; Edman, K. Novel prostaglandin D synthase inhibitors generated by fragment-based drug design. *J. Med. Chem.,* **2008**, *51* (7), 2178-2186.

[168] Huang, J. W.; Zhang, Z. M.; Wu, B. N.; Cellitti, J. F.; Zhang, X. Y.; Dahl, R.; Shiau, C. W.; Welsh, K.; Emdadi, A.; Stebbins, J. L.; Reed, J. C.; Pellecchia, M. Fragment-based design of small molecule X-linked inhibitor of apoptosis protein inhibitors. *J. Med. Chem.,* **2008**, *51* (22), 7111-7118.

[169] Geschwindner, S.; Olsson, L. L.; Albert, J. S.; Deinum, J.; Edwards, P. D.; de Beer, T.; Folmer, R. H. Discovery of a novel warhead against beta-secretase through fragment-based lead generation. *J. Med. Chem.,* **2007**, *50* (24), 5903-5911.

[170] Stockman, B.; Kothe, M.; Kohls, D.; Weibley, L.; Connolly, B.; Sheils, A.; Cao, Q.; Cheng, A.; Yang, L.; Kamath, A.; Ding, Y.; Charlton, M. Identification of allosteric PIF-pocket ligands for PDK1 using NMR-based fragment screening and 1H-15N TROSY experiments. *Chem. Biol. Drug Des.,* **2009**, *73* (2), 179-188.

[171] Mochalkin, I.; Miller, J. R.; Narasimhan, L.; Thanabal, V.; Erdman, P.; Cox, P. B.; Prasad, J.; Lightle, S.; Huband, M. D.; Stover, C. K. Discovery of antibacterial biotin carboxylase inhibitors by virtual screening and fragment-based approaches. *Acs Chemical Biology,* **2009**, *4* (6), 473-483.

[172] Hofstadler, S. A.; Sannes-Lowery, K. A. Applications of ESI-MS in drug discovery: interrogation of noncovalent complexes. *Nat. Rev. Drug Discov.,* **2006**, *5* (7), 585-595.

[173] Vivat Hannah, V.; Atmanene, C.; Zeyer, D.; Van Dorsselaer, A.; Sanglier-Cianferani, S. Native MS: an 'ESI' way to support structure- and fragment-based drug discovery. *Future Medicinal Chemistry,* **2010**, 35-50.

[174] Drinkwater, N.; Vu, H.; Lovell, K. M.; Criscione, K. R.; Collins, B. M.; Prisinzano, T. E.; Poulsen, S. A.; McLeish, M. J.; Grunewald, G. L.; Martin, J. L. Fragment-based screening by X-ray crystallography, MS and isothermal titration calorimetry to identify PNMT (phenylethanolamine N-methyltransferase) inhibitors. *Biochem. J.,* **2010**, *431* (1), 51-61.

[175] Ockey, D.; Dotson, J.; Struble, M.; Stults, J.; Bourell, J.; Clark, K.; Gadek, T., Structure-activity relationships by mass spectrometry: identification of novel MMP-3 inhibitors. *Bioorganic & Medicinal Chemistry* **2004**, 37-44.

[176] Moore, C. D.; Wu, H. H.; Bolanos, B.; Bergqvist, S.; Brooun, A.; Pauly, T.; Nowlin, D. Structural and biophysical characterization of XIAP BIR3 G306E mutant: Insights in protein dynamics and application for fragment-based drug design. *Chemical Biology & Drug Design.,* **2009**, *74* (3), 212-223.

[177] Rose, N. R.; Woon, E. C.; Kingham, G. L.; King, O. N.; Mecinović, J.; Clifton, I. J.; Ng, S. S.; Talib-Hardy, J.; Oppermann, U.; McDonough, M. A.; Schofield, C. J. Selective inhibitors of the JMJD2 histone demethylases: combined nondenaturing mass spectrometric screening and crystallographic approaches. *J. Med. Chem.,* **2010**, *53* (4), 1810-1818.

[178] (a) Erlanson, D.; Wells, J.; Braisted, A. Tethering: fragment-based drug discovery. *Annu. Rev. Biophys. Biomol. Struct.,* **2004**, *33*, 199-223; (b) Erlanson, D. A.; Hansen, S. K. Making drugs on proteins: site-directed ligand discovery for fragment-based lead assembly. *Curr. Opin. Chem. Biol.,* **2004**, *8* (4), 399-406.

[179] (a) Arkin, M.; Randal, M.; DeLano, W.; Hyde, J.; Luong, T.; Oslob, J.; Raphael, D.; Taylor, L.; Wang, J.; McDowell, R.; Wells, J.; Braisted, A. Binding of small molecules to an adaptive protein-protein interface. *Proc. Natl. Acad. Sci. USA,* **2003**, *100* (4), 1603-1608; (b) Braisted, A.; Oslob, J.; Delano, W.; Hyde, J.; McDowell, R.; Waal, N.; Yu, C.; Arkin, M.; Raimundo, B. Discovery of a potent small molecule IL-2 inhibitor through fragment assembly. *J. Am. Chem. Soc.,* **2003**, *125* (13), 3714-3715.

[180] Choong, I.; Lew, W.; Lee, D.; Pham, P.; Burdett, M.; Lam, J.; Wiesmann, C.; Luong, T.; Fahr, B.; DeLano, W.; McDowell, R.; Allen, D.; Erlanson, D.; Gordon, E.; O'Brien, T. Identification of potent and selective small-molecule inhibitors of caspase-3 through the use of extended tethering and structure-based drug design. *J. Med. Chem.,* **2002**, *45* (23), 5005-5022.

[181] Buck, E.; Wells, J. Disulfide trapping to localize small-molecule agonists and antagonists for a G protein-coupled receptor. *Proc. Natl. Acad. Sci. USA,* **2005**, *102* (8), 2719-2724.

[182] Cooper, M. A. Optical biosensors in drug discovery. *Nat. Rev. Drug Discov.,* **2002**, *1* (7), 515-528.

[183] Proll, F.; Fechner, P.; Proll, G. Direct optical detection in fragment-based screening. *Analytical and Bioanalytical Chemistry* **2009**, *393* (6-7), 1557-1562.

[184] Kuglstatter, A.; Stahl, M.; Peters, J.; Huber, W.; Stihle, M.; Schlatter, D.; Benz, J.; Ruf, A.; Roth, D.; Enderle, T.; Hennig, M. Tyramine fragment binding to BACE-1. *Bioorg. Med. Chem. Lett.,* **2008**, *18* (4), 1304-1307.

[185] Nordström, H.; Gossas, T.; Hämäläinen, M.; Källblad, P.; Nyström, S.; Wallberg, H.; Danielson, U. Identification of MMP-12 inhibitors by using biosensor-based screening of a fragment library. *J. Med. Chem.,* **2008**, *51* (12), 3449-3459.

[186] Perspicace, S.; Banner, D.; Benz, J.; Müller, F.; Schlatter, D.; Huber, W. Fragment-based screening using surface plasmon resonance technology. *J. Biomol. Screen.,* **2009**, *14* (4), 337-349.

[187] Neumann, T.; Junker, H.; Keil, O.; Burkert, K.; Ottleben, H.; Gamer, J.; Sekul, R.; Deppe, H.; Feurer, A.; Tomandl, D.; Metz, G. Discovery of thrombin inhibitor fragments from chemical microarray screening. *Letters in Drug Design & Discovery* **2005**, 590-594.

[188] Heim-Riether, A.; Taylor, S. J.; Liang, S.; Gao, D. A.; Xiong, Z.; Michael August, E.; Collins, B. K.; Farmer, B. T.; Haverty, K.; Hill-Drzewi, M.; Junker, H. D.; Mariana Margarit, S.; Moss, N.; Neumann, T.; Proudfoot, J. R.; Keenan, L. S.; Sekul, R.; Zhang, Q.; Li, J.; Farrow, N. A. Improving potency and selectivity of a new class of non-Zn-chelating MMP-13 inhibitors. *Bioorg. Med. Chem. Lett.,* **2009**, *19* (18), 5321-5324.

[189] (a) Holdgate, G.; Ward, W. Measurements of binding thermodynamics in drug discovery. *Drug Discov. Today,* **2005**, *10* (22), 1543-1550; (b) Chaires, J. B. Calorimetry and thermodynamics in drug design. *Annu. Rev. Biophys.,* **2008**, *37*, 135-151; (c) Leavitt, S.; Freire, E. Direct measurement of protein binding energetics by isothermal titration calorimetry. *Curr. Opin. Struct. Biol.,* **2001**, *11* (5), 560-566.

[190] Ciulli, A.; Williams, G.; Smith, A.; Blundell, T.; Abell, C. Probing hot spots at protein-ligand binding sites: a fragment-based approach using biophysical methods. *J. Med. Chem.,* **2006**, *49* (16), 4992-5000.

[191] Edink, E.; Rucktooa, P.; Retra, K.; Akdemir, A.; Nahar, T.; Zuiderveld, O.; van Elk, R.; Janssen, E.; van Nierop, P.; van Muijlwijk-Koezen, J.; Smit, A. B.; Sixma, T. K.; Leurs, R.; de Esch, I. J. Fragment growing induces conformational changes in acetylcholine-binding protein: A structural and thermodynamic analysis. *J. Am. Chem. Soc.,* **2011**, *133*(14), 5363-5371.

Received: December 16, 2010 **Revised: March 4, 2011** **Accepted: March 14, 2011**

1.4 La modélisation moléculaire

La modélisation, au sens général du terme, fait référence à une représentation plus ou moins simplifiée d'un processus. Comme son nom l'indique, la modélisation moléculaire s'inscrit dans un contexte chimique et/ou biologique afin de pouvoir simuler ce genre de système et prédire certaines propriétés d'intérêt. Pour ce faire, la modélisation moléculaire se base sur des formalismes mathématiques, plus ou moins proches de la réalité physique, et dont les deux principaux sont la chimie quantique et la mécanique moléculaire (MM).

La chimie quantique applique la théorie de la mécanique quantique aux molécules : elle vise à résoudre l'équation de Schrödinger. Elle a le désavantage d'être extrêmement complexe, en témoignent les ressources et les temps de calcul nécessaires. Ainsi, seuls de petits systèmes peuvent être étudiés avec ce formalisme mathématique qui permet une description précise d'un système en tenant compte à la fois des noyaux et des électrons. L'utilisation de légères approximations, telle celle de Born-Oppenheimer qui permet un découplage entre le mouvement des électrons et des noyaux du fait de leur grande différence de masse, ne permet pas non plus de s'atteler à des systèmes de taille sensiblement plus importante. Il est toutefois possible d'utiliser localement des approches *ab initio* au niveau d'un système ligand-récepteur [57], mais l'application de ce traitement à l'échelle d'une macromolécule biologique, dont l'ordre de grandeur est le millier d'atomes, reste actuellement inaccessible.

Dès lors, des approches alternatives, comme la mécanique moléculaire, se sont développées afin de pouvoir pallier ces limitations pratiques. D'importantes approximations sont réalisées en MM afin de rendre le formalisme mathématique bien plus abordable et moins coûteux en temps de calcul. Par exemple, la mécanique moléculaire repose sur la théorie de la mécanique classique Newtonienne pour décrire et simuler des systèmes moléculaires. Dans ce contexte, un atome est uniquement défini par une sphère rigide, sans tenir compte de ses sous-unités (noyau et électrons), et les liaisons sont simplement représentées par des ressorts. Afin de représenter la géométrie d'une molécule (voir la Figure 15), il suffit de tenir compte des longueurs de liaison, des angles de valence, et des angles dièdres décrivant la rotation autour d'une liaison, tout en incluant également des termes pour gérer les interactions non liées (interactions électrostatiques, liaisons hydrogène, interactions de Van der Waals, *etc.*, voir le §1.2.4).

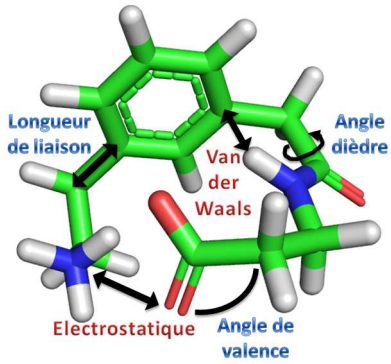


Figure 15: Illustration des principales composantes d'un champ de force (les termes liés sont représentés en bleu et les non liés en rouge).

Contrairement à la théorie quantique, les électrons ne sont pas explicitement simulés en MM. Par conséquent, cette dernière n'est pas à même de prédire des réarrangements électroniques du système comme lors de réactions chimiques (création / cassure de liaisons covalentes). Il s'agit là d'une des principales limites de la MM.

Les simulations de MM se basent sur la notion de champ de force (Force Field - FF). Ce dernier est composé de plusieurs potentiels évaluant différentes composantes énergétiques, le tout définissant l'énergie potentielle (E_{pot}) selon le principe d'additivité des termes. Un postulat fondamental en MM énonce que plus l'E_{pot} d'un système est faible, plus sa conformation est stable. Un FF repose sur un ensemble de paramètres empiriques qui décrivent le système et son état énergétique. Ils sont déterminés à partir de données expérimentales (structures obtenues par X-ray, spectroscopie infrarouge, interactions en phase gazeuse, énergie de solvatation, *etc.*) et/ou de calculs quantiques *ab initio*. Les paramètres ont pour but de reproduire le plus fidèlement ces données. AMBER [58, 59], CHARMM [60], OPLS [61], MMFF [62] et GROMOS [63] sont des FF très populaires, plus ou moins spécialisés pour certaines classes de molécules (protéines, acides nucléiques, carbohydrates, ligands, *etc.*). Il existe deux grandes classes de FF : la classe I dont font partie AMBER et CHARMM, et la classe II qui inclut des termes croisés entre les différents potentiels comme dans le champ de force MMFF.

En MM, un même élément du tableau périodique peut être représenté par plusieurs entités distinctes, appelées types atomiques, afin de tenir compte de l'environnement d'un atome donné. En effet, il est assez intuitif qu'un carbone impliqué dans une liaison amide ne peut pas être décrit par le même jeu de paramètres que l'atome de carbone sp^3 d'une chaîne aliphatique. L'hypothèse de transférabilité des paramètres est également émise, et stipule que ceux-ci peuvent être assignés à leur sous-structure de référence quel que soit l'environnement local. Le meilleur exemple concerne le dictionnaire de

paramètres des acides aminés. En effet, ce dictionnaire est utilisé pour simuler l'ensemble des protéines dès lors qu'elles ne contiennent que des résidus pour lesquels des paramètres sont connus. Les objectifs, les perspectives et les problèmes inhérents aux approches de modélisation moléculaire ont été discutés dans une revue de référence [64].

La MM permet d'évaluer l'énergie potentielle du système à partir des seules coordonnées de ses atomes, d'optimiser sa géométrie à l'aide de procédures de minimisation de l'E_{pot} (afin de trouver une conformation locale plus stable au regard des paramètres du FF), de réaliser des simulations de dynamique moléculaire et de tenter de prédire l'affinité de ligands pour leur cible. Ces différents points seront abordés ci-dessous, à la suite d'une description des termes usuels d'un FF.

1.4.1 La théorie des champs de force

Une focalisation sur le champ de force AMBER est faite car il s'agit du FF systématiquement utilisé dans les simulations réalisées dans le cadre de cette thèse. L'équation du FF AMBER définit l'E_{pot} du système à partir d'une somme de potentiels :

$$E_{pot} = E_{liaison} + E_{angle} + E_{dièdre} + E_{électrostatique} + E_{VdW} \quad \text{Eq 5}$$

avec :

$$E_{liaison} = \sum_{liaisons} K_r(r - r_0)^2 \quad \text{Eq 6}$$

$$E_{angle} = \sum_{angles} K_\theta(\theta - \theta_0)^2 \quad \text{Eq 7}$$

$$E_{dièdre} = \sum_{dièdres} \sum_{n=1}^{N} \frac{V_n}{2}[1 + \cos(m_n\phi - \delta_n)] \quad \text{Eq 8}$$

$$E_{électrostatique} = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(\frac{q_i q_j}{4\pi\varepsilon r_{ij}}\right) \quad \text{Eq 9}$$

$$E_{VdW} = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left[\varepsilon_{ij}\left(\frac{r_{0_{ij}}}{r_{ij}}\right)^{12} - 2\varepsilon_{ij}\left(\frac{r_{0_{ij}}}{r_{ij}}\right)^{6}\right] = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left[\frac{A_{ij}}{{r_{ij}}^{12}} - \frac{B_{ij}}{{r_{ij}}^{6}}\right] \quad \text{Eq 10}$$

Les trois premiers termes (liaison, angle et dièdre) appartiennent à la classe des interactions liées, tandis que les deux derniers (électrostatique et Van der Waals) font référence aux interactions non liées. Une description des différents paramètres et variables de l'équation de E_{pot} est disponible dans le Tableau 1.

Composante	Terme	Description
Liaison	$\mathbf{K_r}$	Constante de force
	$\mathbf{r_0}$	Distance à l'équilibre
	r	Distance courante
Angle de valence	$\mathbf{K_\theta}$	Constante de force
	$\boldsymbol{\theta_0}$	Valeur de l'angle à l'équilibre
	θ	Valeur courante de l'angle
Angle dièdre	$\mathbf{V_n}$	Hauteur de barrière du dièdre
	ϕ	Valeur courante du dièdre
	$\mathbf{m_n}$	Multiplicité
	$\boldsymbol{\delta_n}$	Phase
Electrostatique	**q**	Charge partielle
	ε	Constante diélectrique
	r_{ij}	Distance courante
Van der Waals	$\mathbf{r_{0_{ij}}}$	Distance optimale (fonction des rayons de VdW)
	r_{ij}	Distance courante
	$\boldsymbol{\varepsilon_{ij}}$	Profondeur du puits de potentiel

Tableau 1: Description des différents paramètres et variables de l'équation du FF.
Les données surlignées en gras correspondent à des paramètres du FF.

Dans les FF usuels comme AMBER, par opposition aux FF polarisables, les charges des atomes sont fixes durant l'intégralité de la simulation. Le terme même de champ de force fait référence à des forces F agissant sur le système, alors que l'on utilise des potentiels E dans l'équation du FF (voir l'Eq 5). Ces forces s'obtiennent en dérivant les différents potentiels. Le but est d'obtenir les forces agissant sur les atomes afin de pouvoir modifier leurs coordonnées en conséquence, mais le calcul formel direct de la dérivée correspondante est relativement complexe. Cependant la dérivée de chaque potentiel E par

rapport à sa coordonnée interne l, notée $\partial E / \partial l$, est relativement aisée à calculer du fait de l'utilisation de formules mathématiques simples pour chaque potentiel. Il en est de même pour la dérivée de chaque coordonnée interne par rapport à chaque coordonnée cartésienne *i* des atomes impliqués *n*, et qui est notée $\partial l / \partial x_{n_i}$. Le théorème de dérivation des fonctions composées permet de calculer la dérivée d'intérêt, notée $\partial E / \partial x_{n_i}$, et donc la force correspondante :

$$F = -\frac{\partial E}{\partial x_{n_i}} = -\frac{\partial E}{\partial l} * \frac{\partial l}{\partial x_{n_i}} \qquad \text{Eq 11}$$

Les expressions assez élémentaires des termes non liés (voir l'Eq 9 et l'Eq 10) proviennent aussi du fait que le nombre de paires à considérer est important : l'ordre de grandeur est en N^2, N étant le nombre d'atomes du système. Il est à noter que ces paires ne concernent que les atomes réellement non liés ou séparés d'au moins 3 liaisons covalentes.

Un potentiel supplémentaire modélisant les liaisons hydrogène est également inclus dans certains FF [65]. Dans le champ de force AMBER natif, la paramétrisation des termes électrostatique et VdW a permis de reproduire les liaisons hydrogène sans ajout de terme spécifique, entre autre en utilisant des charges partielles assez élevées pour les types atomiques impliqués dans les LH et un rayon de Van der Waals nul pour les hydrogènes polaires (HO, HW).

Pour contraindre la planarité de certains groupes comme les cycles aromatiques, des angles de torsion dits impropres, impliquant ici quatre atomes qui doivent rester dans un plan, sont également ajoutés. Dans le FF AMBER, les angles impropres sont traités *via* le même potentiel que les angles de torsion classiques.

Enfin, un FF peut considérer tous les atomes du système (“tout atomes”), traiter les groupements C-H comme un type atomique en soi (“atomes unis”), ou simuler une représentation simplifiée des éléments du système (“gros grain”). Du fait de ces approximations supplémentaires, les deux dernières approches - spécialement la dernière - permettent d'accélérer les simulations et donc d'en réaliser de plus longues.

Les paragraphes suivants sont consacrés à une description de chaque composante du FF AMBER.

La composante liaison covalente

Une liaison covalente, définie par les types atomiques des deux atomes impliqués dans celle-ci, possède une longueur optimale à l'équilibre (r_0). La Figure 16 illustre l'allure du terme de liaison (voir l'Eq 6) en fonction de la distance entre ses deux atomes.

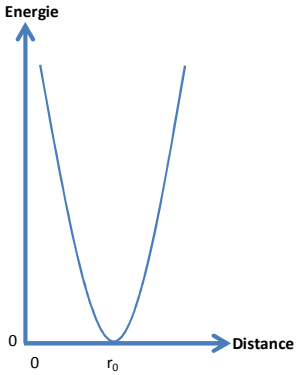


Figure 16: Illustration de l'allure du potentiel de liaison en fonction de la longueur de celle-ci.

Une liaison covalente est soumise à des mouvements de vibration qui diminuent ou augmentent sa longueur d'une faible amplitude, comme lors de la compression ou de l'élongation d'un ressort. Il est à noter que raccourcir ou allonger une liaison covalente d'une longueur donnée renvoie la même pénalité avec ce potentiel de liaison, ce qui est en contradiction avec les données expérimentales, où l'élongation est moins pénalisante énergétiquement que le raccourcissement. Le potentiel de Morse [66] permet de mieux modéliser cette asymétrie, mais est plus coûteux en temps de calcul, d'où l'utilisation usuelle du potentiel harmonique, dont les résultats restent toutefois très convenables.

La composante angle de valence

Un angle de valence, noté θ, est défini par trois atomes (voir la Figure 17).

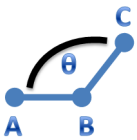


Figure 17: Représentation de l'angle de valence θ pour trois atomes A, B et C.

Il se calcule directement à partir de la formule du produit scalaire :

$$\theta = \widehat{ABC} = \cos^{-1}\left(\frac{\overrightarrow{BA} \cdot \overrightarrow{BC}}{\|\overrightarrow{BA}\| * \|\overrightarrow{BC}\|}\right) \qquad \text{Eq 12}$$

A l'instar des liaisons, les angles de valence possèdent également une valeur optimale à l'équilibre (θ_0). La composante angle du FF est modélisée selon le même formalisme que le terme de liaison, en substituant des angles aux distances (voir l'Eq 7).

La composante angle dièdre

Un angle dièdre, noté ϕ, est défini par quatre atomes (voir la Figure 18).

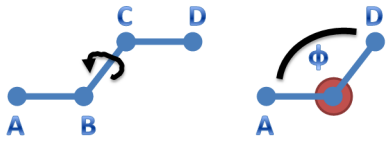


Figure 18: Représentation d'un angle dièdre ϕ défini par les atomes A, B, C et D autour de la liaison B-C (à gauche, vue de face ; à droite, vue de profil).

Une représentation de profil permet de mieux comprendre la stratégie, en quatre étapes, permettant de calculer la valeur d'un angle de torsion :

- soit un vecteur $\overrightarrow{V1}$, normal au plan défini par les atomes A, B et C, calculé à l'aide du produit vectoriel $\overrightarrow{BA} \wedge \overrightarrow{BC}$
- soit un vecteur $\overrightarrow{V2}$, normal au plan défini par les atomes B, C et D, calculé à l'aide du produit vectoriel $\overrightarrow{CB} \wedge \overrightarrow{CD}$
- la valeur de l'angle dièdre ϕ, appartenant à l'intervalle [0,π], est obtenue en calculant l'angle entre ces nouveaux vecteurs $\overrightarrow{V1}$ et $\overrightarrow{V2}$, comme précédemment pour l'angle de valence :

$$\phi = \cos^{-1}\left(\frac{\overrightarrow{V1} \cdot \overrightarrow{V2}}{\|\overrightarrow{V1}\| * \|\overrightarrow{V2}\|}\right) \qquad \text{Eq 13}$$

- le signe de ϕ, car un angle dièdre est conventionnellement défini sur l'intervalle [-π,π], est défini selon le signe du réel $v = \overrightarrow{BD} \cdot \left(\overrightarrow{BA} \wedge \overrightarrow{BC}\right) = \overrightarrow{BD} \cdot \overrightarrow{V1}$

Par opposition aux deux termes liés précédents, il n'y a pas systématiquement de valeur optimale unique pour un angle dièdre : un tel angle peut avoir plusieurs minima, plus ou moins énergétiquement favorables. Par conséquent, le formalisme mathématique précédent, utilisé pour les termes liaison et angle, ne peut s'appliquer à cette coordonnée interne. Un potentiel incluant une fonction trigonométrique est préférentiellement utilisé pour modéliser un angle de torsion. Dans le potentiel dièdre (Eq 8), également appelé potentiel de torsion, le terme V_n est la valeur de la barrière de rotation du dièdre considéré, tandis que m_n, ϕ et δ_n représentent respectivement sa multiplicité (ou périodicité de rotation), sa valeur courante et son angle de phase. Certains angles dièdres sont modélisés avec plusieurs jeux de paramètres, d'où l'indice *n* et la sommation sur l'ensemble des jeux de paramètres pour chaque angle de torsion (voir la Figure 19).

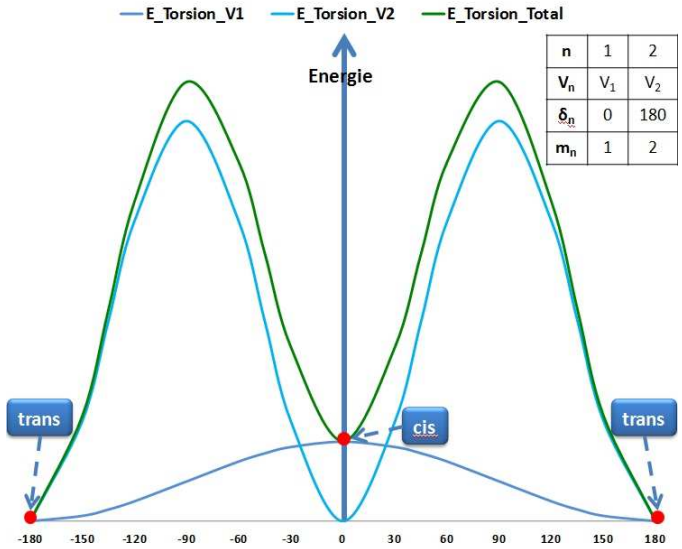


Figure 19: Illustration du potentiel de torsion usuel pour une liaison amide.
Les deux conformations expérimentales les plus favorables (''cis'' et ''trans'') sont bien représentées dans les puits de potentiel, avec une préférence pour la conformation ''trans''.

La composante électrostatique

La loi de Coulomb est utilisée pour calculer l'énergie électrostatique E entre deux atomes *i* et *j*, séparés d'une distance *r*, et respectivement porteurs d'une charge ponctuelle q_i et q_j (voir l'Eq 9). Pour une paire donnée, cette composante est énergétiquement favorable si les charges sont de signe opposé (interaction attractive avec E<0), mais devient une pénalité si elles sont de même signe (répulsion électrostatique avec E>0). Ce type d'interaction est maintenu à longue distance, par exemple aux alentours de 10 Å, car le potentiel coulombien diminue seulement en 1/r (voir la Figure 20).

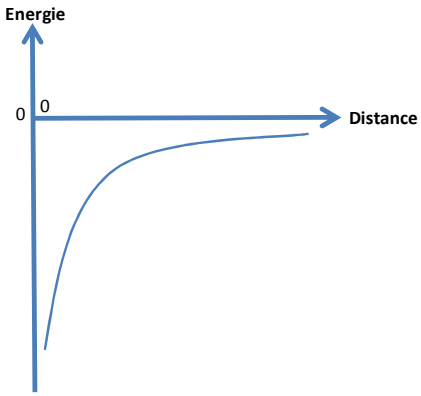


Figure 20: Représentation schématique de l'énergie calculée par le terme électrostatique du FF (loi de Coulomb) en fonction de la distance entre deux charges ponctuelles de signe opposé.

Un paramètre essentiel de ce potentiel, en dehors des charges partielles, est la valeur de la constante diélectrique ε du milieu environnant (vide, solvant polaire, solvant apolaire). La constante diélectrique représente la capacité du milieu environnant à maintenir une certaine distance entre deux charges ponctuelles, et ce grâce à un phénomène d'écrantage électrostatique. Par définition, on a $\varepsilon = \varepsilon_0 * \varepsilon_r$, avec ε_0 la permittivité du vide et ε_r la constante diélectrique relative (par rapport au vide) du milieu considéré. Dans le vide, $\varepsilon_r = 1$, alors que dans l'eau $\varepsilon_r \approx 80$, et il est généralement considéré que ε_r appartient à l'intervalle [2,4] au sein des protéines [67]. Par conséquent, les interactions électrostatiques sont d'une intensité beaucoup plus importante dans le vide que dans l'eau.

La composante de Van der Waals

Tout comme les interactions électrostatiques, cette composante se calcule également sur des paires d'atomes non liés. Soit une paire impliquant deux atomes de types atomiques *i* et *j* séparés d'une distance *r*. Le potentiel de Lennard-Jones est souvent utilisé pour modéliser les interactions de VdW (voir l'Eq 10). Le terme en $1/r^6$, ayant une base physique, modélise la composante attractive des interactions de VdW, d'où le signe négatif devant celle-ci. A l'inverse, le terme en $1/r^{12}$ modélise un phénomène de répulsion à trop faible distance, car il ne peut y avoir d'interpénétration des nuages électroniques entre deux atomes non liés. L'utilisation d'une puissance élevée est intuitivement logique et renforce le côté répulsif, mais la valeur 12 est totalement empirique : elle provient du fait qu'il est facile de calculer r^{12} connaissant r^6. Du fait des puissances utilisées, le terme attractif est dominant à distance moyenne tandis que le terme répulsif est dominant à courte distance (voir la Figure 21).

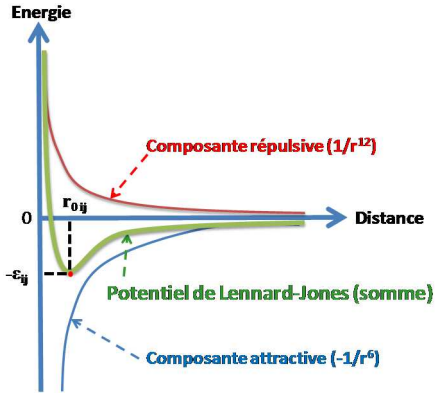


Figure 21: Représentation schématique du potentiel de Lennard-Jones, ainsi que de ses deux composantes, en fonction de la distance entre les deux atomes non liés de la paire considérée.

Les valeurs A_{ij} et B_{ij} présentes dans le potentiel (voir l'Eq 10) sont une combinaison des "vrais" paramètres de VdW que sont ε_{ij} et $r_{0_{ij}}$. Le paramètre ε_{ij} représente la profondeur de puits du potentiel de Van der Waals de la paire considérée. On peut l'interpréter comme une mesure de l'intensité de l'attraction de VdW entre les deux atomes de la paire. A la somme $r_{0_{ij}}$ des rayons de VdW, où les sphères atomiques se touchent sans s'interpénétrer, le potentiel de Lennard-Jones doit être minimal et égal à $-\varepsilon_{ij}$. Donc, dans un but d'efficacité (temps de calcul), on peut définir et pré-initialiser les termes A_{ij} et B_{ij} de cette manière :

$$A_{ij} = \varepsilon_{ij}(r_{0_{ij}})^{12}$$
$$B_{ij} = 2\,\varepsilon_{ij}(r_{0_{ij}})^{6} \qquad \text{Eq 14}$$

Les fichiers de paramètres d'un FF incluent uniquement les paramètres ε_i et r_{0_i} pour chaque type atomique *i*. Ces derniers ont été obtenus pour des paires homogènes (*i=j*) à partir de données expérimentales, entre autre *via* l'étude de gaz ou de structures X-ray. Pour une paire hétérogène (*i≠j*), on calcule les paramètres de Van der Waals ε_{ij} et $r_{0_{ij}}$ à partir des paramètres connus (ε_i ; ε_j ; r_{0_i} ; r_{0_j}) de cette manière :

$$r_{0_{ij}} = r_{0_i} + r_{0_j}$$
$$\varepsilon_{ij} = \sqrt{\varepsilon_i\,\varepsilon_j} \qquad \text{Eq 15}$$

1.4.2 La minimisation d'énergie

L'hypersurface d'énergie potentielle E_{pot} est une fonction des coordonnées des atomes du système. Cette dernière est très complexe car le nombre de variables est très important (3N coordonnées cartésiennes, avec N le nombre d'atomes du système). En MM, cette surface est également très accidentée : de faibles variations de coordonnées peuvent aboutir à de très fortes variations de l'E_{pot}. Du fait de la complexité de cette fonction et de sa surface associée, il est impossible en pratique de déterminer analytiquement les coordonnées correspondantes à son minimum global. Diverses méthodes d'exploration de la surface d'énergie potentielle, ayant pour but la recherche de minima (global et locaux), ont été développées, parmi lesquelles la minimisation d'énergie potentielle, la dynamique moléculaire, les simulations stochastiques de type Monte Carlo ou basées sur des algorithmes génétiques (voir aussi les §1.4.3 et §1.5.2).

Comme son nom l'indique, ce sous-chapitre est dédié à la partie minimisation d'énergie. La minimisation de l'E_{pot} est principalement utilisée dans un but d'optimisation de la géométrie courante d'un système jusqu'à atteindre le minimum énergétique le plus proche, généralement un minimum local (voir la Figure 22).

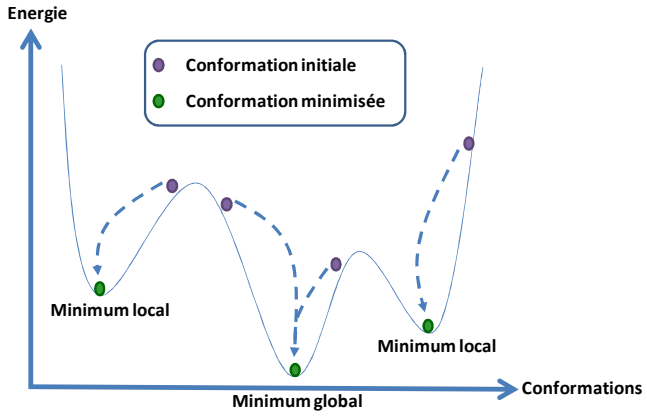


Figure 22: Minimisation dans un espace 1D partant de différentes conformations initiales, et annotation des différents minima obtenus (global et locaux).

Par exemple, la minimisation d'un système composé de deux atomes reliés par une liaison covalente aboutira *in fine* à une distance *r* séparant les deux atomes égale à r_0, à savoir la distance optimale d'après les paramètres du FF (voir l'Eq 6). Il est à noter que dans ce cas simpliste et très particulier, l'énergie finale est égale à 0 et correspond au minimum global. Plus précisément et partant d'une conformation donnée, l'objectif d'un algorithme de minimisation est de trouver la direction à suivre sur la surface d'énergie potentielle dans le but de diminuer petit à petit l'E_{pot} à travers le déplacement

des atomes du système. Cette recherche de direction peut reposer sur des concepts mathématiques usuels comme les dérivées, qui permettent entre autre de donner des informations sur les variations locales d'une fonction. Les algorithmes de minimisation les plus populaires reposent sur l'utilisation des dérivées :

- dérivées premières seules : méthodes de la plus grande pente et du gradient conjugué
- dérivées premières et secondes : méthodes de Newton-Raphson et "quasi-Newton"

Il est à noter que ce sont des processus itératifs et purement mathématiques : à chaque étape, la valeur de la fonction étudiée (ici, E_{pot}) diminue jusqu'à atteindre le minimum local le plus proche et cela sans pouvoir passer de barrières énergétiques (même très faibles), d'où le classement des algorithmes de minimisation dans les méthodes d'exploration relativement réduite de la surface d'énergie potentielle.

La méthode de la plus grande pente

La méthode de la plus grande pente ("Steepest Descent" - SD) est l'algorithme basé sur les dérivées premières le plus simple. En MM, la recherche du minimum à partir d'un point donné (initial ou intermédiaire) est faite dans la direction opposée au gradient ($-\nabla$) de la fonction E_{pot} par rapport aux coordonnées du système. Cette direction, correspondant à celle où la fonction décroît le plus vite, est nommée direction de la plus grande pente (ou descente), d'où le nom de l'algorithme.
Soit x_i un vecteur de dimension 3N représentant les coordonnées cartésiennes du système à l'étape *i* avec *i* partant de 0. A l'étape 0, le vecteur x_0 contient les coordonnées courantes du système avant minimisation. Chaque étape de minimisation *i* peut être décomposée en 3 sous-étapes :

1) identification de la direction de recherche D_i qui correspond à $-\nabla_i$, avec ∇_i le gradient de la fonction E_{pot} à l'étape *i*,
2) recherche d'un pas optimal k_i, rendant la fonction E_{pot} minimale, à l'aide d'une recherche linéaire le long de cette direction D_i,
3) calcul des nouvelles coordonnées x_{i+1}, composante par composante, à partir des coordonnées courantes x_i, du pas k_i et de la direction de recherche D_i :

$$x_{i+1} = x_i + k_i * D_i = x_i - k_i * \nabla_i \qquad \text{Eq 16}$$

La méthode SD est simple mais a comme inconvénient d'être peu efficace en terme de convergence, ce qui a donné lieu au développement d'algorithmes alternatifs plus compétitifs (voir ci-dessous).

La méthode du gradient conjugué

Cet algorithme ("Conjugate Gradient" - CG) est également uniquement basé sur les dérivées premières, et peut être vu comme une optimisation de la méthode de la plus grande pente.
Par définition, deux vecteurs D_1 et D_2 sont conjugués par rapport à une matrice M définie positive si :

$$D_1{}^T * M * D_2 = 0 \qquad \text{Eq 17}$$

Dans la méthode GC, les minimisations unidimensionnelles où l'on recherche le pas optimal sont réalisées non pas le long de la plus grande pente (-∇) comme dans l'approche SD, mais le long d'une direction dite conjuguée (Dconj) par rapport à celle de l'étape précédente. Le reste de l'algorithme CG est identique à celui de SD : il faut également réaliser une recherche linéaire pour trouver le pas optimal k_i à chaque étape *i*. Les nouvelles coordonnées x_{i+1} sont exprimées en fonction des coordonnées courantes x_i, du pas optimal k_i et de la direction de recherche conjuguée $Dconj_i$ selon la relation :

$$x_{i+1} = x_i + k_i * Dconj_i \qquad \text{Eq 18}$$

Bien qu'il ne repose que sur le calcul des dérivées premières, l'algorithme CG fait implicitement (donc sans devoir le calculer) référence au Hessien de la fonction E_{pot}. L'astuce de cet algorithme consiste à exprimer la direction de recherche conjuguée $Dconj_i$ en fonction de l'opposé du gradient courant $-\nabla_i$ (comme précédemment), tout en le modulant à l'aide de données connues (∇_i, ∇_{i-1}, $Dconj_{i-1}$) et en satisfaisant à la définition des vecteurs conjugués (voir l'Eq 17 avec le Hessien comme matrice M) :

$$Dconj_i = -\nabla_i + X_i * Dconj_{i-1} \qquad \text{Eq 19}$$

Les mathématiciens Fletcher et Reeves ont proposé le calcul du réel X_i de cette manière [68] :

$$X_i = \frac{\nabla_i{}^T * \nabla_i}{\nabla_{i-1}{}^T * \nabla_{i-1}} \qquad \text{Eq 20}$$

Il est à noter que l'expression de X_i n'a pas une forme unique : des formes alternatives, par exemple celle de Polak-Ribière, ont également été proposées.

Il a été montré que l'Eq 20 peut se réécrire sous cette forme en utilisant les normes des gradients :

$$X_i = \frac{\|\nabla_i\|^2}{\|\nabla_{i-1}\|^2} \qquad \text{Eq 21}$$

En pratique, c'est cette dernière expression qui est utilisée dans le calcul de la direction de recherche conjuguée $Dconj_i$. La formule de récurrence nécessite une direction de recherche conjuguée initiale : l'opposé du gradient, comme pour l'approche SD, est choisi par convention.
Du fait de l'utilisation de directions optimisées, l'algorithme CG converge plus rapidement que la méthode SD sans être réellement plus complexe, que ce soit au niveau du temps de calcul, de la mémoire vive utilisée ou de son implémentation, d'où sa forte popularité.

La méthode de Newton-Raphson

La méthode de Newton-Raphson repose sur l'hypothèse qu'une fonction *f* peut être approximée localement par une fonction quadratique autour d'un extremum (minimum ou maximum). De plus, cette fonction *f* peut être approximée par la troncature de son développement de Taylor (limité ici à l'ordre 2) autour d'un point donné x_0, avec $\Delta x = x - x_0$, et en utilisant le formalisme bra-ket car x est un vecteur 3N-dimensionnel :

$$f(x) = f(x_0) + \langle \Delta x | \nabla f_{x_0} \rangle + \frac{1}{2} \langle \Delta x | H_{x_0} | \Delta x \rangle \qquad \text{Eq 22}$$

Le gradient ∇f au point x s'approxime par rapport à celui du point x_0:

$$\nabla f = \nabla f_{x_0} + H_{x_0} | \Delta x \rangle \qquad \text{Eq 23}$$

Au minimum d'une fonction f, sa dérivée est nulle ($\nabla f = 0$), d'où le pas Δx à réaliser pour l'atteindre :

$$\Delta x = - {H_{x_0}}^{-1} | \nabla f_{x_0} \rangle \qquad \text{Eq 24}$$

En généralisant ce résultat, on détermine les nouvelles coordonnées (étape *i+1*) à partir de celles de l'étape courante (étape *i*) et du pas courant Δx_i :

$$x_{i+1} = x_i + \Delta x_i = x_i - {H_{x_i}}^{-1} | \nabla f_{x_i} \rangle \qquad \text{Eq 25}$$

Un avantage de la méthode Newton-Raphson est que le pas vers l'optimum est clairement défini, alors qu'il est déterminé à l'aide d'une recherche linéaire le long de la direction de la plus grande descente dans le cadre de l'algorithme SD. La méthode de Newton-Raphson est rigoureuse d'un point de vue mathématique et possède une bonne convergence (les dérivées secondes donnent en effet accès à la courbure de la fonction) : une seule itération est nécessaire pour atteindre l'optimum d'une fonction quadratique. Cependant, au niveau d'un point arbitraire d'une surface d'énergie, qui est tout sauf quadratique, le déplacement des x prôné par cette méthode n'a pas *a priori* plus de sens qu'un mouvement selon le gradient (tel quel ou conjugué). De plus, elle implique de devoir calculer l'inverse du Hessien (H^{-1}), ce qui est très coûteux à la fois en temps de calcul et en mémoire vive dans le cadre de la MM où de gros systèmes sont simulés.

Les algorithmes dits quasi-Newton sont basés sur la méthode de Newton-Raphson, décrite ci-dessus, tout en modifiant la façon dont est calculé le Hessien à chaque étape. Il peut être démontré qu'une approximation de l'inverse de la matrice Hessienne à l'étape *i+1* peut être obtenue à partir de celle de l'étape *i*. Dès lors, la matrice Hessienne ainsi que son inversion ne sont pas explicitement recalculées à chaque étape, ce qui rend les méthodes quasi-Newton plus compétitives par rapport à la méthode de Newton-Raphson classique. L'algorithme de Broyden-Fletcher-Goldfarb-Shanno (BFGS) figure parmi les méthodes quasi-Newton les plus populaires [69].

Les critères d'arrêt de la minimisation d'énergie

Les procédures de minimisation d'énergie intègrent généralement plusieurs critères d'arrêt, car le nombre d'étapes successives peut être très important jusqu'à l'obtention d'une totale convergence. Les critères d'arrêt courants sont :

- un nombre maximum d'étapes à réaliser
- un temps de calcul maximal alloué
- une norme de gradient en dessous d'un seuil prédéfini
- un gain d'énergie par étape inférieur à un seuil prédéfini

1.4.3 La dynamique moléculaire

La dynamique moléculaire (DM) est un type de simulation fondamental et très populaire en modélisation moléculaire, d'où la nécessité absolue de l'aborder ici. Cependant, étant donné que la DM n'est pas utilisée dans le cadre de ce travail, ce sous-chapitre n'a pour but que d'en être une brève introduction.

La dynamique moléculaire a pour objectif de produire une trajectoire représentative de l'évolution au cours du temps d'un système au niveau moléculaire [70, 71]. Cette technique est particulièrement utilisée pour étudier et simuler des macromolécules biologiques ou leurs complexes. Les trajectoires de DM ainsi que les conformations résultantes permettent ensuite de calculer certaines propriétés fondamentales comme des énergies libres [72], ou d'expliquer des mécanismes à l'échelle atomique comme la liaison d'un ligand à sa cible [73], le repliement d'un peptide [74] ou l'activation d'un récepteur [75]. De nombreuses suites logicielles comme AMBER [58, 59] et CHARMM [60] intègrent la possibilité de réaliser des simulations de DM. Ces dernières sont généralement constituées de trois phases successives :

1) le chauffage du système jusqu'à la température désirée de manière à donner une impulsion à ses atomes (notions de vitesse et d'énergie cinétique E_{cin})
2) l'équilibration du système à cette température pendant un certain temps
3) la production qui correspond à la trajectoire d'intérêt pour une durée donnée (nanosecondes jusqu'a millisecondes en fonction des phénomènes étudiés)

D'un point de vue théorique, la DM se base sur les équations de Newton de la mécanique classique régissant les mouvements. Finalement, on a [76]:

$$F_i = m_i\, a_i = m_i \frac{dv_i}{dt} = m_i \frac{d^2x_i}{dt} = -\frac{\partial E_{pot}(x_1, \dots, x_n)}{\partial x_i} \qquad \text{Eq 26}$$

avec F_i la force agissant sur l'atome *i*, m_i la masse de l'atome *i*, a_i l'accélération subie par l'atome *i*, v_i la vitesse instantanée de l'atome *i*, x_i le vecteur position courant de l'atome *i*, et les forces sont calculées comme précédemment (voir le §1.4.1).

Pour pouvoir obtenir une trajectoire, il va falloir déterminer les positions successives des atomes (coordonnées cartésiennes) du système *via* l'intégration dans le temps des équations du mouvement de Newton (voir l'Eq 26). Ceci est le rôle des algorithmes d'intégration comme par exemple l'algorithme de Verlet [77]. Tout comme pour certaines méthodes de minimisation (voir le §1.4.2), il repose sur le concept mathématique des développements de Taylor. Les développements tronqués de Taylor, limités à l'ordre 2, de la fonction position x(t) aux instants t+Δt et t-Δt s'écrivent :

$$x(t + \Delta t) = x(t) + \Delta t * v(t) + \frac{\Delta t^2}{2} * a(t) \qquad \text{Eq 27}$$

$$x(t-\Delta t) = x(t) - \Delta t * v(t) + \frac{\Delta t^2}{2} * a(t)$$

avec *v(t)* et *a(t)* étant respectivement la vitesse et l'accélération à l'instant *t*. A partir de la seconde équation de l'Eq 27, on peut isoler le terme impliquant la vitesse, d'où :

$$\Delta t * v(t) = x(t) - x(t-\Delta t) + \frac{\Delta t^2}{2} * a(t) \qquad \text{Eq 28}$$

L'injection de l'Eq 28 dans la première équation de l'Eq 27 permet d'aboutir à la forme finale de l'algorithme de Verlet (voir l'Eq 29), où les nouvelles coordonnées *x(t+Δt)* sont déterminées à partir des termes connus que sont les positions courantes *x(t)*, les positions précédentes *x(t-Δt)*, les accélérations courantes *a(t)*, et cela sans devoir explicitement calculer les vitesses *v(t)* :

$$x(t+\Delta t) = 2 * x(t) - x(t-\Delta t) + \Delta t^2 * a(t) \qquad \text{Eq 29}$$

Néanmoins, l'algorithme de Verlet classique est source d'imprécisions lors de son initialisation à $t=t_0$ car il requiert explicitement deux positions dont l'une est inconnue. Paradoxalement, le fait d'avoir pu éliminer le paramètre vitesse pose certains problèmes : bien qu'elles ne soient pas nécessaires pour produire la trajectoire, les vitesses sont requises pour calculer l'énergie cinétique E_{cin}, et vérifier la conservation de l'énergie totale du système pendant la simulation ($E_{totale} = E_{pot} + E_{cin} \approx$ constante). Afin d'outrepasser ces difficultés, des variantes de cet algorithme ont été développées, notamment celle dite "Verlet-vitesse" [78]. Le choix du pas d'intégration Δt est également fondamental pour obtenir une trajectoire pertinente : son ordre de grandeur usuel est la femto-seconde, car il doit être inférieur à la période des mouvements de plus hautes fréquences (de l'ordre de 10^{15} s^{-1} pour les vibrations des liaisons impliquant des atomes d'hydrogène). L'ajout de contraintes sur la longueur de ces liaisons permet d'augmenter légèrement ce pas d'intégration [79].

Comme indiqué auparavant, le nombre total de termes non liés devient vite prohibitif car il évolue en N^2 avec N le nombre d'atomes du système. En principe, chaque terme devrait être évalué, ce qui a une conséquence négative sur le temps de calcul nécessaire lors de chaque évaluation énergétique. Une astuce, consistant à ignorer les paires à longue distance, permet de diminuer cette contrainte. Ce seuil de distance, appelé seuil de coupure ("cut-off"), est en pratique lié au terme de Coulomb puisque c'est la composante non liée qui diminue le plus lentement avec la distance, et sa valeur est de l'ordre de 12 Å. Afin d'éviter des effets de seuil systématiques au niveau de la zone de coupure (l'erreur finale

pouvant être non négligeable sur l'ensemble des termes non liés), on utilise généralement une fonction d'amortissement (“switch”) dont le but est de moduler l'énergie calculée par le potentiel (voir la Figure 23). Ainsi, ce dernier renvoie la valeur usuelle en dessous d'un seuil donné de distance (d_{min}), 0 au delà d'un second seuil de distance (d_{max}) et une valeur modulée dans l'intervalle [d_{min},d_{max}].

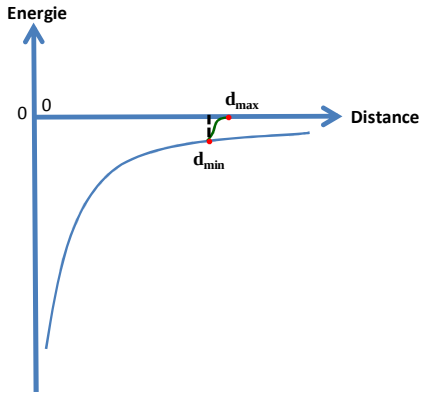


Figure 23: Illustration de l'emploi d'une fonction d'amortissement (cas du terme de Coulomb). Les deux paramètres d_{min} et d_{max} représentent l'intervalle où la fonction est réellement utilisée.

Les phénomènes de solvatation peuvent être traités de manière implicite ou explicite, ce dernier cas nécessitant la création d'une boîte d'eau autour des solutés d'intérêt. Lorsque l'on réalise des simulations en solvant explicite, le nombre de molécules d'eau est très important. Par conséquent, la remarque précédente sur le nombre de paires non liées est à nouveau d'actualité. De plus, les molécules d'eau à la périphérie de la boîte ne sont pas correctement modélisées car elles ne sont pas entourées de molécules d'eau au même titre que celles centrales. Afin d'éviter d'éventuels effets de bord, le système est représenté de manière périodique (souvent sous forme cubique) comme dans un cristal. Lorsqu'une molécule sort du cube par un côté, on la réintègre au sein de celui-ci par l'autre côté. Cette astuce permet de considérablement diminuer la taille du système à considérer tout en gommant les éventuels effets de bord au niveau de ses limites. Toutefois, il faut veiller à ce qu'une molécule ne soit pas en interaction avec l'une de ses images. Ceci est à nouveau réalisé *via* l'emploi de seuils de coupure dans le calcul des termes non liés.

La dynamique moléculaire est une autre manière d'explorer la surface d'énergie potentielle, en suivant cette fois les lois de Newton. Bien qu'elle permette de sauter des barrières énergétiques tout au long de la simulation, une simulation classique de DM ne permet pas de faire de grands sauts entre chaque pas d'intégration. De larges réarrangements conformationnels, comme lors du repliement d'un peptide, nécessitent une très longue simulation ou l'introduction de contraintes pour guider le processus. Ainsi, la DM permet une meilleure exploration de la surface d'énergie potentielle que la minimisation

d'énergie, cette dernière tombant systématiquement dans le puits d'énergie local. Toutefois, des méthodes alternatives ont été développées pour explorer cette surface complexe et accidentée de manière plus efficace / rapide (voir le sous-chapitre §1.5). La DM, qui possède certaines bases physiques, reste la méthode de choix lorsque l'on souhaite obtenir une trajectoire à l'issue de la simulation.

1.4.4 L'estimation de l'affinité d'un ligand

L'estimation précise de l'énergie libre de liaison d'un ligand pour une cible est d'une importance majeure en drug design. Malheureusement, cette donnée thermodynamique, elle-même composée de deux termes complexes (enthalpique et entropique), est très difficile à modéliser. Par conséquent, il est très compliqué de la prédire avec précision, et ce d'autant plus que la méthode utilisée renvoie rapidement une valeur (balance classique "précision *vs.* coût computationnel"). Ainsi, la capacité à prédire correctement l'affinité d'un composé reste encore aujourd'hui l'un des défis majeurs en modélisation moléculaire.

Des méthodes diverses, plus ou moins précises et rigoureuses d'un point de vue physique, ont été développées pour tenter de résoudre cette problématique, entre autre les fonctions de score empiriques (voir le §1.6.2), les modèles QSAR [80] (essentiellement limités à des séries congénériques) et les approches basées sur la MM. Au sein de cette dernière famille figurent les techniques d'intégration thermodynamique, de perturbation d'énergie libre et les méthodes MM-PBSA / MM-GBSA [81, 82]. La dernière, intitulée "Molecular Mechanical-Generalized Born Surface Area", est la plus compétitive au niveau de la balance "précision *vs.* coût computationnel" [67]. Elle permet d'approximer des énergies libres de liaison à partir de la différence entre les énergies libres des différentes espèces (complexe ligand-récepteur, ligand libre, récepteur libre, voir la Figure 24) solvatées de manière implicite.

$$\Delta G_{liaison} = G_{complexe} - (G_{prot\,éine} + G_{ligand})$$

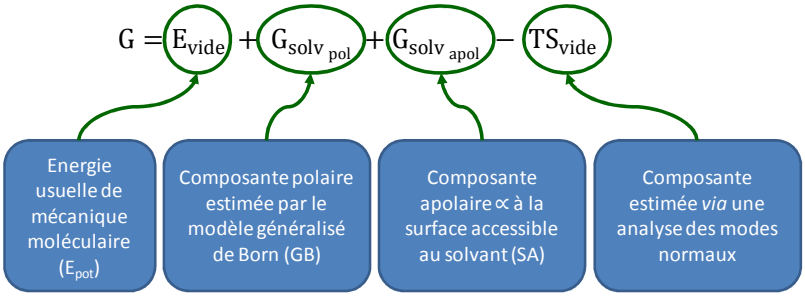


Figure 24: Décomposition du calcul de l'énergie libre de liaison selon la méthode MM/GBSA (adapté des références 83, 84).

L'énergie libre de chaque espèce est décomposée en un terme enthalpique et un terme entropique. La composante enthalpique est elle-même subdivisée en plusieurs composantes : une énergie en phase gazeuse estimée par le FF et une composante liée à la solvatation des espèces. Cette dernière combine une partie polaire estimée par le modèle généralisé de Born (la partie GB de GBSA), et une non polaire proportionnelle à la surface accessible au solvant (la partie SA de GBSA). La composante entropique reste très difficile à approximer. La pénalité entropique translationnelle/rotationnelle associée au phénomène de liaison est certes élevée (de l'ordre de 15-20 $kJ.mol^{-1}$), mais elle peut néanmoins être considérée comme quasi-constante car ne dépendant que très peu de la taille du ligand [85]. En revanche, la partie entropique conformationnelle, liée aux DDL internes du ligand, pose des problèmes quasi-insolubles en pratique, et les plus grandes approximations sont réalisées sur celle-ci [67]. La pénalité entropique conformationnelle associée au phénomène de liaison est souvent obtenue *via* une analyse des modes normaux sur un ensemble de conformations du système obtenues à l'aide de simulations de DM [86].

Dans le cadre de composés de type fragment, où l'éventail du nombre de torsions est très restreint, ce problème majeur est par conséquent moins pénalisant. A ce sujet, les fonctions de score empiriques (voir le §1.6.2) abordent généralement ce problème *via* un terme de pénalité proportionnel au nombre de torsions du ligand [87]. Afin d'obtenir des valeurs absolues d'énergie libre de liaison, les énergies relatives obtenues peuvent être calibrées par rapport à des valeurs expérimentales [67].

Les limites de la méthode MM-GBSA [67] sont clairement mises en évidence pour des ligands dont le phénomène de liaison est guidé principalement par le changement d'entropie, dans le cas de LH ligand-site médiées par des molécules d'eau et enfin pour distinguer des analogues dont l'un est structurellement contraint (version cyclisée d'un ligand, ou présence d'un ortho-méthyle comme discuté au §1.2.4) et l'autre non. Le premier peut avoir une affinité beaucoup plus importante pour une raison d'origine entropique, mais ce phénomène est très difficile à modéliser. Malgré ses limitations, l'approche MM-GBSA a été employée avec succès dans de nombreux projets [67].

1.5 Les autres méthodes d'échantillonnage conformationnel

L'échantillonnage conformationnel (EC) consiste à visiter les différentes conformations possibles d'un système, conçues comme des points dans l'espace des phases défini par tous les degrés de liberté de la molécule. Dans ce contexte, le système peut être illustré par une petite molécule organique (voir la Figure 25). Le but d'un EC consiste à générer un ensemble de conformères ou à rechercher la ou les conformations les plus stables. Le critère de stabilité est souvent une énergie, décrite par un Hamiltonien se basant sur un FF ou sur la théorie quantique.

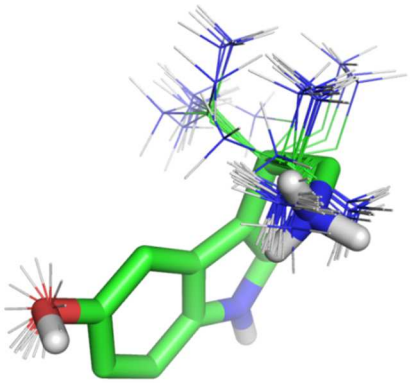


Figure 25: Exemple d'EC d'un ligand organique. La conformation initiale est représentée en stick.

Dans le contexte de l'EC d'une entité unique, les DDL explicitement échantillonnés se résument généralement aux angles dièdres du système. Dans le cas de plusieurs entités, telles les simulations de docking (voir le §1.6), des DDL supplémentaires (rotation et de translation inter-entités) sont ajoutés en plus de ceux internes à chaque entité. Pour des raisons de temps de calcul disponible, d'approximations souhaitées ou de connaissances biologiques sur la cible, certains DDL peuvent être modulés :

- totalement bloqués, par exemple lors d'une simulation de docking avec un site totalement rigide ou lors d'un EC d'une entité ayant des atomes fixes
- contraints, par exemple pour maintenir des atomes à proximité lors de la résolution d'une structure tridimensionnelle afin de satisfaire des contraintes NOE expérimentales [88]

Une multitude de méthodes, basées sur des concepts très variés, ont été développées pour répondre à cette problématique [76, 89]. Elles peuvent être classées en deux grandes familles : les méthodes déterministes et les méthodes stochastiques.

1.5.1 Approches déterministes

Les approches déterministes ont en commun de retourner le même résultat pour une simulation donnée qui serait exécutée plusieurs fois.

Scan systématique des degrés de liberté

Le scan systématique des DDL est la méthode la plus intuitive : on énumère toutes les conformations possibles en bouclant avec un pas donné sur l'ensemble des DDL (angles dièdres). Le pas doit être suffisamment fin pour pouvoir énumérer un ensemble pertinent, tout en étant suffisamment grand pour éviter un nombre excessif de conformations ou l'obtention de conformères quasi-identiques. Etant donné que chaque conformation de l'ensemble fini est évaluée énergétiquement, il est possible de trouver le minimum global au regard de la fonction énergétique utilisée dès lors que l'ensemble est représentatif. Un pas de 30° est usuel, mais cela aboutit tout de même à 12 possibilités pour chaque angle dièdre, d'où une explosion du nombre de conformations (N) avec l'augmentation du nombre de DDL (n) : $N = 12^n$. En pratique, ce type d'approche ne peut être appliqué qu'à des systèmes ayant un faible nombre de DDL. Le programme MacroModel permet entre autre de réaliser un scan systématique des angles de torsion d'une molécule [90].

Création de structures 3D basées sur des règles et des dictionnaires structuraux

D'autres approches déterministes se basent sur des règles d'assemblage et sur une bibliothèque préexistante de conformations énergétiquement favorables pour un ensemble de sous-structures. Ainsi, le passage d'une structure 2D (table de connectivité) à une ou plusieurs structure(s) 3D revient à assembler de manière combinatoire les différentes géométries précalculées de chaque bloc, puis de sélectionner le ou les conformères à conserver *via* leur évaluation énergétique. Une relaxation de type descente de gradient peut également être appliquée à l'ensemble généré. Bien que chacun possède des caractéristiques propres, les programmes CORINA [91, 92], CONCORD [92, 93], OMEGA [94] et DG-AMMOS [95] reposent globalement sur ce type de stratégie. La plupart de ces approches, très efficaces d'un point de vue du temps de calcul, sont focalisées sur la génération de conformères pour des petites structures comme des ligands organiques ou des complexes organométalliques de taille relativement réduite. Une stratégie, s'appuyant à la fois sur un alphabet structural (en l'occurrence, un dictionnaire de conformations concernant quatre résidus protéiques consécutifs) et un FF à gros grain, a permis de prédire la conformation expérimentale de molécules bien plus complexes comme des peptides [96-98].

1.5.2 Approches stochastiques

Par opposition aux méthodes déterministes, les approches stochastiques explorent l'espace conformationnel de manière aléatoire. Elles peuvent donc ne pas retourner systématiquement le même résultat pour une simulation identique exécutée plusieurs fois, dès lors que le temps de simulation est insuffisant au regard de la taille de l'espace de recherche. Il est à noter que le minimum global peut tout à fait être trouvé par ces approches ; la principale difficulté étant qu'il est difficile voire impossible de s'en assurer en pratique pour des problèmes de grande complexité.
Seules des méthodes stochastiques dont la fonction objective est de type énergie potentielle d'un FF sont considérées ici. Les principaux algorithmes stochastiques pour réaliser un EC sont de type Monte Carlo, évolutionnaires et relatifs au phénomène d'auto-organisation.

Les simulations Monte Carlo

Les méthodes Monte Carlo (MC) sont des méthodes stochastiques d'optimisation globale très générales qui reposent sur des modifications aléatoires des DDL. Ce genre d'algorithme, qui permet de sonder l'espace de recherche du problème et qui consiste en un certain nombre de cycles, est notamment utilisé pour résoudre des problèmes complexes. Appliqués à la modélisation moléculaire, les algorithmes MC permettent d'échantillonner un grand nombre de configurations diverses du système, et une évaluation énergétique permet d'identifier les configurations de plus basse énergie. Ce type d'algorithme est associé à un critère de Métropolis [99], dont le formalisme est très similaire à celui du facteur de Boltzmann ($e^{-\frac{E}{k_b T}}$), dans le but de pouvoir conserver, selon un tirage aléatoire, une conformation d'énergie moins favorable. En effet, un puits de potentiel encore plus intéressant peut être atteint en passant par une conformation intermédiaire d'énergie plus élevée. Ce dernier point fait référence à la notion de passage de barrières énergétiques, et ces dernières peuvent être très importantes du fait de la forte rugosité de la fonction E_{pot}. Ainsi, cette méthode peut permettre de sortir d'un éventuel minimum local dans lequel on serait piégé. Pour rester cohérent, il faut néanmoins rendre la probabilité d'acceptation d'une configuration d'énergie plus élevée d'autant plus faible que le surplus énergétique vis-à-vis de la meilleure conformation est grand. Un algorithme classique de MC associé au critère de Métropolis est illustré à la Figure 26.
Une approche de recuit simulé, où la température du système est considérée comme une variable [100], pilote généralement les simulations MC : la température, qui est élevée initialement, baisse régulièrement au fur et à mesure des cycles, afin de pouvoir rejeter plus facilement des configurations de plus haute énergie au fur et à mesure de ceux-ci, notamment pour pouvoir se focaliser sur l'identification des minimum locaux des configurations visitées. Il reste cependant à définir le

protocole du recuit simulé, à savoir l'intensité de la baisse de température et sa fréquence d'application.

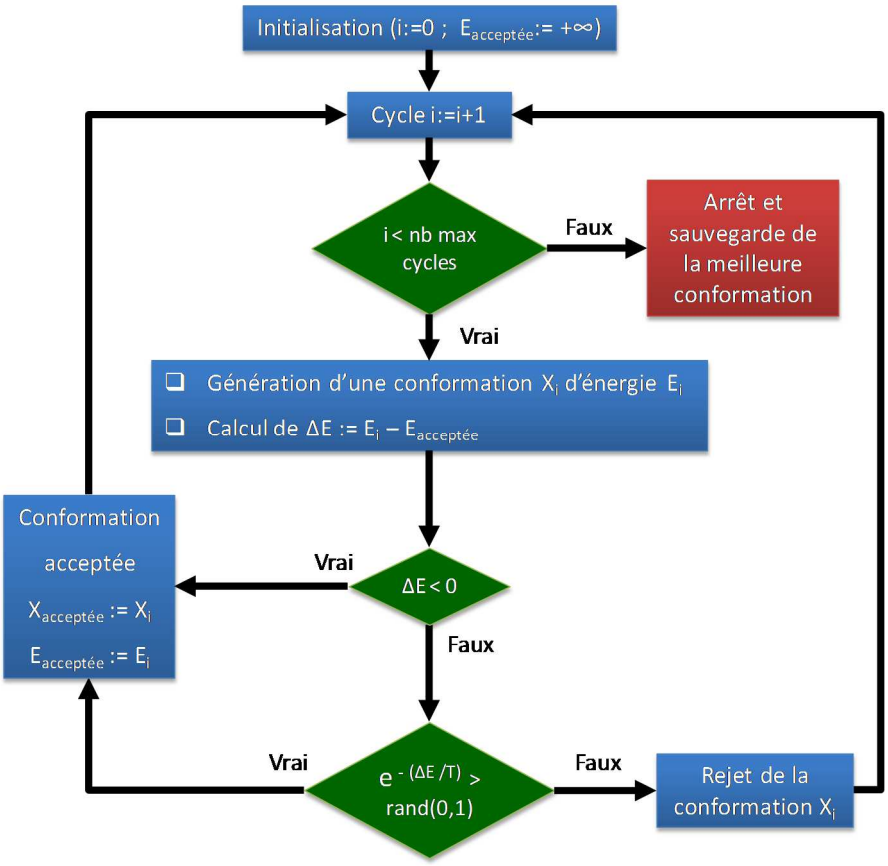


Figure 26: Illustration d'un algorithme Monte Carlo avec critère de Métropolis. La fonction rand(0,1) renvoie un réel appartenant à l'intervalle [0,1].

Les algorithmes évolutionnaires et génétiques

Un algorithme évolutionnaire est également une méthode stochastique d'optimisation globale, mais qui s'inspire de la théorie de Darwin. Le Darwinisme fait référence à la notion de sélection naturelle qui stipule que les individus les plus aptes à vivre dans un environnement donné auront une plus grande probabilité de survivre, et donc de se reproduire génération après génération. Un algorithme évolutionnaire repose également sur la notion de population et une fonction d'évaluation (“fitness”) mesure l'aptitude d'un individu par rapport à son environnement.

Au sein de cette classe d'algorithmes, les algorithmes génétiques (AG) sont les plus populaires, et miment l’évolution biologique en faisant l’analogie entre le vecteur des DDL et le génotype d'un individu représenté par un chromosome. Ce chromosome est soumis à des mutations ou des crossing-over en guise de stratégie d’exploration de l’espace des DDL (les deux classes d'opérateur sont

introduites ci-dessous). Par leur capacité à modifier ce génome, ces opérateurs génétiques sexués (cross-over, impliquant deux chromosomes parents) et asexués (mutation, produisant un descendant à partir d'un seul individu) influent donc sur le phénotype qui n'est autre que le résultat de l'expression des gènes. Dans notre contexte, il s'agit de la géométrie moléculaire résultant des dièdres courants listés dans le chromosome, et la viabilité de ce nouvel individu est bien dictée par son phénotype : stabilité=f(géométrie)=E_{pot}. Enfin, la probabilité de maintien d'un individu dans une population de taille fixe croît avec son score de viabilité ("fit"). Par conséquent, il y a une concentration des traits génétiques (structuraux) favorables dans la population au fil des cycles successifs de reproduction/sélection appelés générations.

Tout AG consiste en l'initialisation aléatoire d'une population d'un nombre donné d'individus, suivie par un nombre de générations où les différents opérateurs s'appliquent sur les chromosomes d'individus courants selon des fréquences prédéfinies. Un exemple d'AG standard est illustré à la Figure 27.

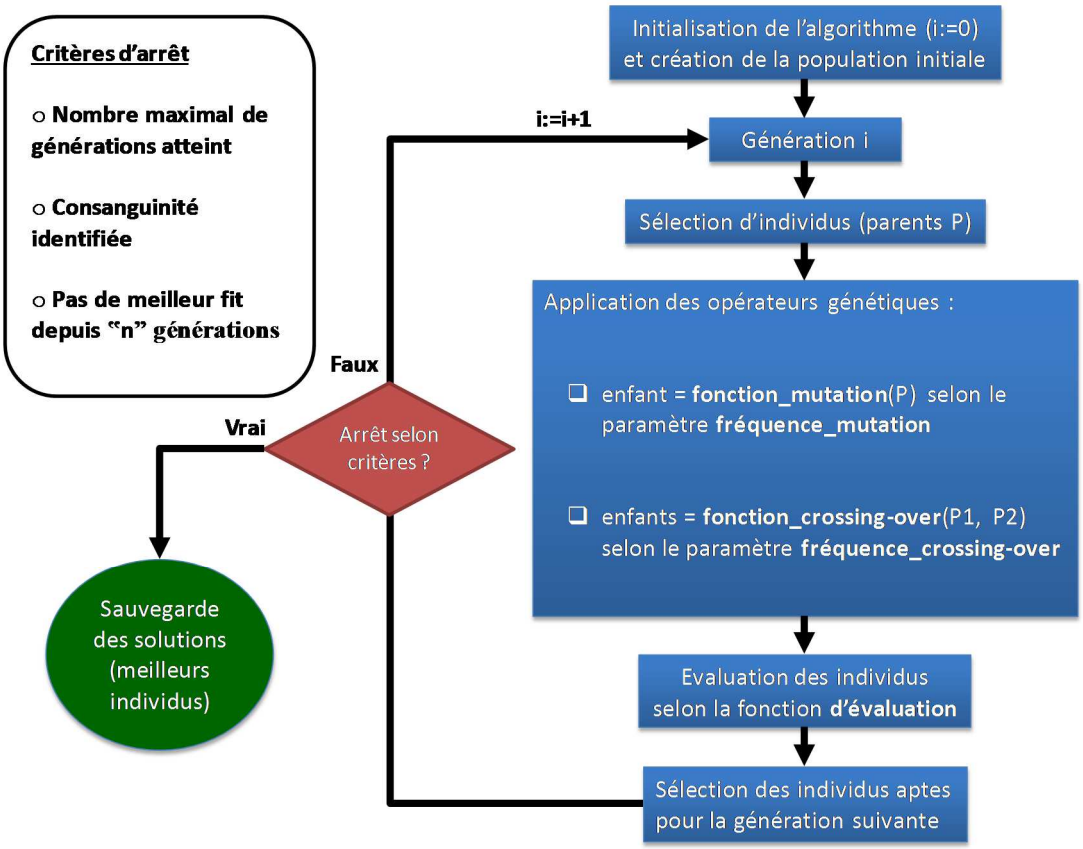


Figure 27: Illustration d'un algorithme génétique standard.

Il existe différentes stratégies pour sélectionner les individus aptes à survivre pour la génération suivante. L'élitisme est une stratégie courante qui consiste à ne garder que les N meilleurs individus

triés selon la fonction d'évaluation, avec N la taille de la population. De plus, le sous-ensemble des N' meilleurs individus de la population (l'élite, avec N' << N) ne peut pas être modifié ou remplacé.

Les critères d'arrêt usuels d'un AG sont un nombre maximal de générations, la convergence vers un même état des différents individus de la population (notion de consanguinité), et l'incapacité à trouver un meilleur individu après un certain nombre de générations. Il faut bien entendu être capable de détecter une éventuelle consanguinité au moyen de critères appropriés. Dans le cas d'un EC, une trop grande similitude conformationnelle au sein de la population, estimée par exemple par des RMSD faibles (après superposition) ou à travers l'utilisation d'empreintes d'interaction, est synonyme de consanguinité. Si la consanguinité apparaît trop tôt, alors l'algorithme peut retourner un optimum local, d'où l'emploi d'un opérateur de diversité génétique. Ce dernier vise à remplacer par de nouveaux individus ceux ayant la moins bonne énergie parmi les membres trop proches génétiquement.

Comme en biologie, une mutation est une modification ponctuelle d'un chromosome, et l'opérateur mutation le plus simple consiste à sélectionner un individu de manière aléatoire, et à remplacer la valeur d'un de ses DDL tiré au hasard sur son chromosome (notion de locus) par une valeur elle-même générée aléatoirement (voir la Figure 28). En plus d'impacter le génome des individus, et donc de favoriser l'exploration de l'espace de recherche, l'opérateur mutation possède la qualité de pouvoir introduire de nouveaux caractères non présents dans la population à un instant donné.

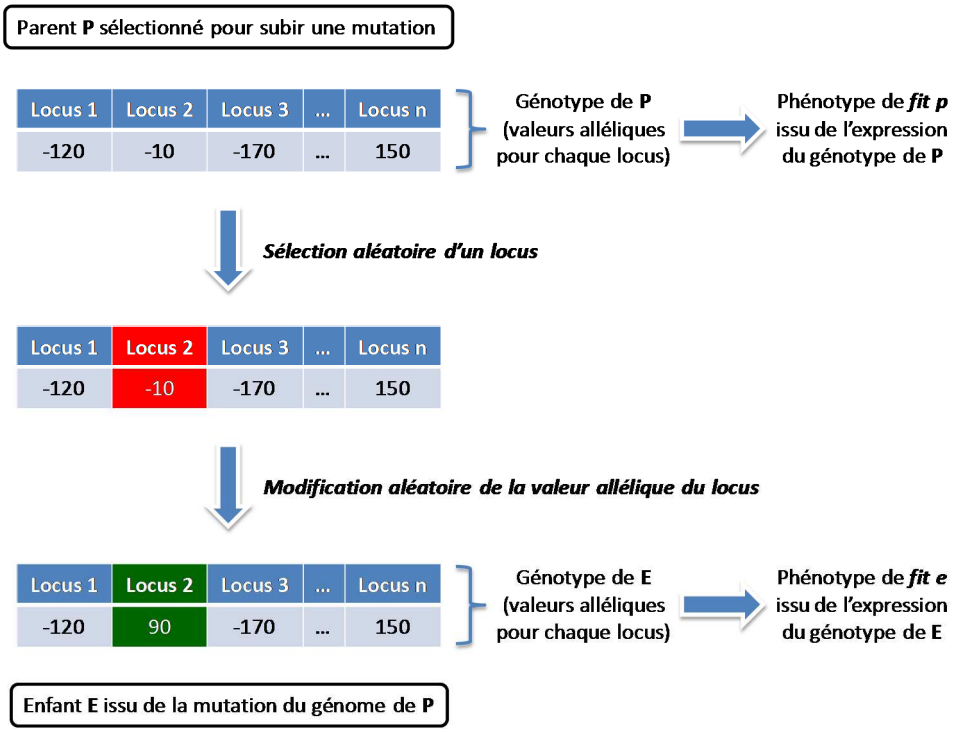


Figure 28: Représentation d'un opérateur mutation standard.

Le terme "fit" fait référence à l'évaluation de l'individu par la fonction de fitness.

En biologie, un crossing-over (CO) consiste en l'appariement de deux chromosomes homologues qui vont s'échanger des fragments homologues par recombinaison génétique. Selon le même principe, un opérateur CO classique va sélectionner deux individus tirés au sort parmi la population, puis va identifier un point de CO également de manière aléatoire, pour finir par créer deux enfants issus de la recombinaison des DDL des deux parents (voir la Figure 29). L'idée sous-jacente de cet opérateur est que si chacun des parents possèdent des caractères intéressants mais complémentaires, leur union peut aboutir à un enfant "exceptionnel" vis-à-vis de la fonction d'évaluation.

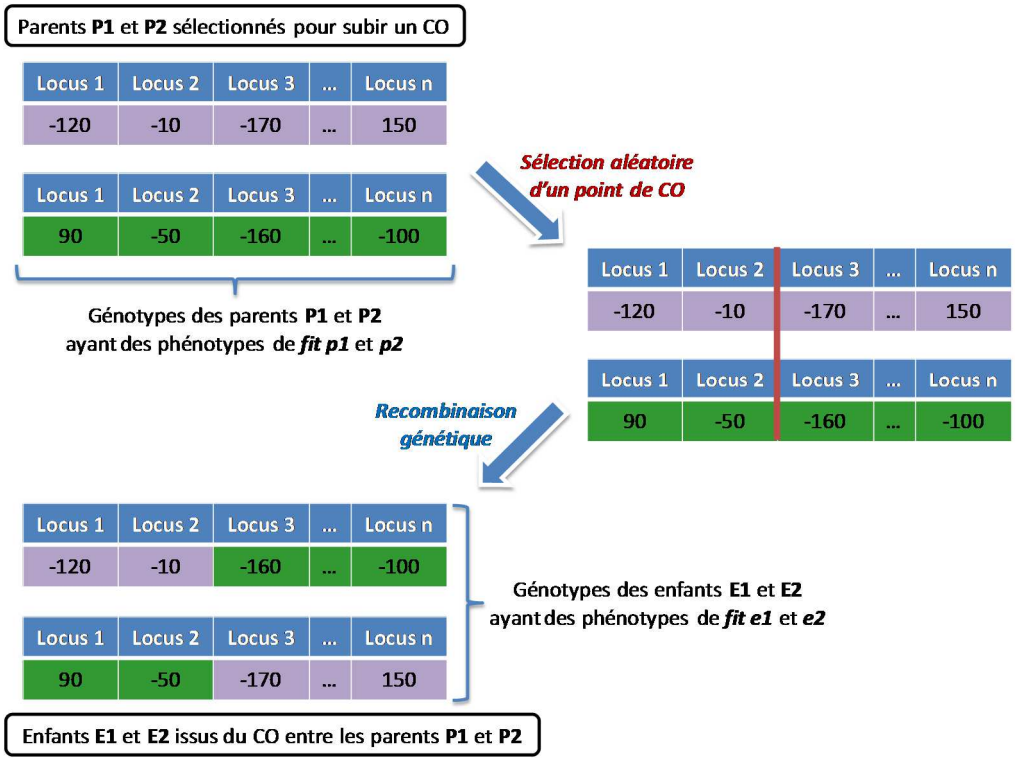


Figure 29: Représentation d'un opérateur crossing-over standard.
Le terme "fit" fait référence à l'évaluation de l'individu par la fonction de fitness.

Un avantage notable des AG est leur forte adaptation à la taille d'un problème donné : il suffit de jouer sur les paramètres majeurs (taille de la population, nombre de générations, fréquences des opérateurs) pour pouvoir le traiter en théorie.

Lorsque la fonction d'évaluation est basée sur un FF, l'optimisation locale d'un individu par descente de gradient devient possible. Ce genre d'algorithme génétique "hybride" est également connu sous le nom d'AG lamarckien [101]. En effet, une optimisation locale peut être bien plus efficace pour converger vers le minimum le plus proche que d'attendre par exemple la mutation visant à impacter de la même manière un DDL donné. Cependant, cette variante très populaire d'AG est confrontée au dilemme

"optimisation locale des individus *vs.* échantillonnage de l'espace de recherche" : il est à la fois nécessaire de réaliser des recherches locales d'optimum tout en permettant une exploration efficace (par définition non locale) de l'espace de recherche. Ceci explique les paramètres supplémentaires inhérents à tout AG de type lamarckien (nombre d'étapes, fréquences et critères de convergence des minimisations, *etc.*).

Enfin, les AG sont très populaires dans le cadre du docking (voir le §1.6) : de nombreux programmes, entre autre GOLD [102], Autodock [101], EADock [103], FlipDock [104], LGA [105] et FITTED [106], reposent sur eux pour réaliser la phase d'échantillonnage conformationnel. C'est également le cas de S4MPLE, et plus de précisions seront apportées dans la partie "Matériel et méthodes" dédiée (voir le §2.1).

Les algorithmes d'auto-organisation

Ces algorithmes sont également des méthodes stochastiques d'optimisation globale très générales qui peuvent être adaptées pour identifier des minimums énergétiques dans le cadre d'un EC. Ils partagent le concept d'auto-organisation qui fait référence à la capacité d'un groupe d'individus à s'auto-organiser de manière performante, alors que chaque individu considéré séparément ne possède que des capacités cognitives très limitées. Il y a deux sous-types principaux (colonies de fourmis et essaims) qui sont rapidement introduits au regard de leur popularité actuelle. Tout comme les algorithmes évolutionnaires, les algorithmes d'auto-organisation reposent sur une population.

Les algorithmes de colonies de fourmis ("Ant Colony Optimization" - ACO) [107, 108] s'inspirent du comportement d'une colonie de fourmis à la recherche de nourriture. De nombreux chemins coexistent entre la nourriture et la fourmilière, et le chemin initialement emprunté pour trouver la nourriture est totalement aléatoire. Cependant, au bout d'un certain temps, les fourmis utilisent préférentiellement le chemin optimal (le plus court) lorsqu'elles ramènent la nourriture au niveau de la fourmilière. Cette aptitude vient d'un processus indirect de communication entre les différentes fourmis faisant intervenir un dépôt volatil d'hormones (phéromones) qui les attire. Ainsi, un chemin non optimal contiendra peu de phéromones et la piste s'estompera petit à petit à cause de leur caractère volatil, alors qu'un chemin très emprunté sera renforcé petit à petit, notamment à cause des allers et retours fréquents des fourmis du fait de la longueur optimale (minimale) du trajet.

Dans le contexte de l'EC, chaque fourmi tente de trouver des zones de l'espace des phases favorables d'un point de vue énergétique et dépose des phéromones virtuelles afin que les autres membres de la population se focalisent sur ces zones intéressantes au fur et à mesure des itérations. Par ailleurs, les zones explorées les moins favorables disparaissent petit à petit à cause de l'évaporation des

phéromones virtuelles. Un algorithme ACO a notamment été utilisé avec succès en docking à travers l'outil PLANTS [109].

Les algorithmes à essaim (“Particle Swarm Optimization” - PSO) [110] s'inspirent initialement du comportement d'un groupe d'oiseaux en vol. Dans ce contexte, chaque oiseau serait considéré comme une particule présente au sein de l'essaim. Chaque particule du système va se déplacer à travers l'espace de recherche en fonction d'une équation de mouvement intégrant une vitesse courante aléatoire et un biais vers les solutions optimales obtenues à ce stade (locales et sur l'ensemble de l'essaim) dans le but d'identifier les zones de l'espace de recherche les plus favorables énergétiquement. Comme dans les autres algorithmes stochastiques, le processus identifie itération après itération des minimums de plus en plus profonds jusqu'à convergence. Plusieurs outils de docking, entre autre SODOCK [111], pso@autodock [112], ParaDockS [113] et FIPSDock [114], reposent sur un algorithme PSO.

1.6 Le docking et les fonctions de score / d'énergie

Le docking, qui est aussi connu sous le nom d'amarrage moléculaire, peut être défini comme la prédiction de la structure d'un complexe "n-aire" à l'échelle atomique [15]. Généralement, le système considéré est binaire et implique deux entités qui sont un site de liaison et une petite molécule organique. C'est ce cas de figure classique, le docking "mono ligand-récepteur", qui sera implicitement traité dans ce chapitre, même si des systèmes alternatifs (protéine-protéine, protéine-acide nucléique) ont également fait l'objet de recherches [115, 116]. De nombreuses méthodes de docking ont été développées, mais elles partagent toutes deux étapes clés :

1) l'exploration de l'espace conformationnel des DDL sélectionnés (intra-entités comme en "EC classique", mais aussi inter-entités pour pouvoir positionner une entité par rapport à une autre) *via* une stratégie de recherche, comme celles évoquées au §1.5. Cette stratégie doit être suffisamment efficace pour être capable de visiter le mode de liaison connu d'une molécule active
2) l'estimation énergétique des différentes configurations visitées du système à l'aide d'une fonction de score / d'énergie. Par convention, le terme de fonction de score (FS) sera préférentiellement utilisé dans ce sous-chapitre

Les simulations de docking permettent de s'atteler à plusieurs problématiques fondamentales dans le domaine de la recherche pharmaceutique :

- la prédiction du mode de liaison d'un actif connu pour lequel il n'existe pas encore de données structurales. Cette information, lorsqu'elle est valide, permet de rationaliser la phase d'optimisation de la molécule (SBDD)
- la sélection d'un sous-ensemble d'une chimiothèque à cribler expérimentalement de manière prioritaire (voir le criblage virtuel et la notion d'enrichissement au §1.6.3). Le docking agit comme un filtre dans le contexte du criblage virtuel

1.6.1 Docking et DDL considérés

Il y a trois types d'approches de docking en fonction des DDL explicitement considérés pendant la simulation : blocage des DDL internes de l'ensemble des entités (docking rigide), activation des DDL internes du ligand (docking semi-flexible), voire extension de la flexibilité à une partie de la cible (docking flexible). Ces différents types d'approches sont décrits ci-dessous, ainsi que leurs avantages et inconvénients respectifs.

Le docking rigide

L'algorithme ne s'intéresse qu'aux DDL de rotation et de translation inter-entités. En d'autres termes, il n'explore que les différentes positions de la conformation rigide du ligand au sein du site également rigide. Etant donné que les DDL internes à chaque entité sont totalement bloqués, c'est le type de simulation le plus proche du modèle clé-serrure décrit préalablement. Pour des raisons évidentes, les premières tentatives de docking étaient de ce genre [117]. Les algorithmes de docking rigide peuvent être très rapides, mais sont souvent peu utiles car la conformation bioactive du ligand est rarement connue *a priori*. Cependant, en combinaison avec un générateur de conformères, cette stratégie redevient intéressante car le principal point négatif est théoriquement réglé. Cette synergie est employée par les logiciels de docking FRED [118, 119] et MS-DOCK [120]. Néanmoins, cette approche risque de rejeter certains actifs si :

1. le générateur de conformères n'a pas produit la géométrie bioactive
2. la géométrie bioactive a été générée mais son positionnement dans le site échoue à cause de la rugosité du potentiel
    a. il y a une légère imprécision géométrique, comme par exemple une distance entre points d'ancrage du ligand surestimée d'une fraction d'ångström au regard de la distance maximale tolérée par la géométrie figée du site
    b. un réarrangement conformationnel, comme un mouvement de chaîne latérale, est nécessaire pour accommoder le conformère bioactif

Le docking semi-flexible

Dans ce cadre, le ligand est flexible mais le site reste rigide. La conformation du ligand est échantillonnée selon des techniques variées, incluant les algorithmes génétiques, les méthodes géométriques comme la construction incrémentale [121], les algorithmes d'auto-organisation et les simulations Monte Carlo. Une approche de ce type est donc en principe capable de répondre favorablement aux situations 1 et 2.a mentionnées auparavant. Les cas 2.a sont vite résolus par une étape de minimisation (relaxation des DDL du ligand dans le site), alors que l'EC du ligand en présence du site permettra même d'aboutir à des géométries bioactives tendues, non générées spontanément par l'EC du ligand libre (cas 1). Le docking semi-flexible est la stratégie implémentée dans la plupart des outils de docking actuels (entre autre FlexX [121], Glide [122], DOCK [123], Surflex [124] et PLANTS [109]) car elle est un bon compromis entre rapidité des simulations du fait du nombre limité de DDL et pertinence des résultats, dès lors qu'il n'y a que très peu de réarrangements conformationnels du site à la suite de la fixation du ligand. Le cas 2.b évoqué ci-dessus reste donc problématique pour le

docking semi-flexible. Bien que le récepteur reste rigide au sens où il n'y a aucun mouvement d'atome lourd, certains outils permettent néanmoins d'optimiser le réseau de LH potentielles entre le ligand et le site [125].

Le docking flexible

Un docking flexible implique que certains DDL du site sont également débloqués. C'est la stratégie la plus proche du modèle main-gant décrit au §1.3.3, et donc la plus pertinente pour modéliser les phénomènes de reconnaissance moléculaire. Malheureusement, le déblocage des DDL du site augmente de manière considérable le nombre de possibilités à explorer, d'autant plus lorsque des mouvements de la chaîne principale sont rendus possibles. Dans ce dernier cas, la géométrie du site peut être sensiblement bouleversée, d'où la difficulté pour estimer la pertinence des réarrangements obtenus à l'issue du processus. Enfin, l'estimation énergétique est également plus complexe et plus lente car des termes supplémentaires liés aux contacts intra-récepteur doivent également être ajoutés, alors qu'ils étaient négligés auparavant. Le déblocage d'une partie du site met aussi en évidence des carences au niveau des FS, notamment dans leur capacité à prédire correctement les réarrangements conformationnels attendus.

Il existe plusieurs alternatives, plus ou moins réalistes, pour prendre en compte la flexibilité du site. La plus intuitive, qui est aussi celle demandant le moins de développements techniques, consiste à docker un ligand donné dans un ensemble de conformations du site de liaison qui resteront rigides pendant la simulation. Cela revient à réaliser n simulations de docking semi-flexible pour chaque ligand, où n est le nombre de structures du site. Cet ensemble peut être construit à l'aide de structures expérimentales provenant de la PDB et/ou de conformations relaxées issues d'une trajectoire de dynamique moléculaire. La principale difficulté consiste à créer un ensemble minimal représentatif des différentes conformations possibles du site. L'analyse de modes normaux de vibration a été employée avec succès pour identifier des conformations pertinentes à inclure dans l'ensemble de structures [126]. Etant donné que ce docking flexible est très proche dans l'esprit du type semi-flexible, il est relativement rapide tant que n reste petit. Dans la stratégie de FlexE [127], un modèle unique du site est généré à partir de la superposition d'un ensemble de structures de référence. Les parties non conservées sont notamment décrites sous la forme de zones alternatives (par exemple, différents rotamères d'un résidu sont explicitement inclus dans le modèle).

Les autres approches dites flexibles échantillonnent explicitement certains DDL internes du site au même titre que ceux du ligand. Les chaînes latérales sont les DDL les plus faciles à considérer car elles n'ont qu'un impact limité sur l'aspect global du site de liaison. Malgré cela, le nombre de possibilités augmente de manière exponentielle avec le nombre et la complexité des chaînes latérales

débloquées. Certaines versions d'Autodock [128] et de GOLD [102] permettent de réaliser des simulations incluant des chaînes latérales flexibles.
D'autres programmes comme RosettaLigand [129, 130], FlipDock [104], FITTED [106] et IFD [131] intègrent la possibilité de pouvoir échantillonner une partie de la chaîne principale, ouvrant la voie à de plus larges réarrangements conformationnels du site de liaison. Contrairement à la dynamique moléculaire, le docking flexible ne considère qu'une partie du site à échantillonner explicitement : selon la stratégie de recherche utilisée, les mouvements peuvent être très larges (MC, AG) alors qu'ils restent locaux après chaque pas d'intégration en dynamique moléculaire, d'où la nécessaire limitation des DDL en docking flexible. En revanche, il n'est pas exclu que le système soit relaxé dans son intégralité par des minimisations d'énergie pendant la phase d'EC et/ou dans une phase de post-traitement.

Selon les approches de docking, la structure tridimensionnelle du site est utilisée ou encodée de différentes manières : le site peut être utilisé comme tel (“tout atome”), sous une représentation simplifiée (“gros grain”) dans le but d'accélérer le processus par rapport à la forme précédente, voire sous la forme de grilles précalculées pour divers potentiels. Cette dernière stratégie permet de fortement simplifier le coût des évaluations énergétiques, mais pose problème lors du passage à un site partiellement flexible car une modification de sa structure a un impact direct sur lesdites grilles.
A l'heure actuelle, le défi principal reste toujours le développement d'algorithmes de docking flexible, prédisant les éventuels changements conformationnels avec précision et en un temps raisonnable (de l’ordre de la dizaine de minutes) afin de pouvoir cribler des librairies dont l'ordre de grandeur est de plusieurs milliers de composés. Néanmoins, dans des cas ponctuels où seuls quelques actifs de mode de liaison inconnu sont considérés, des temps de calculs plus importants (de l'ordre de plusieurs heures par ligand) restent tout à fait acceptables puisque le seul critère décisif est la précision des géométries prédites [132].

1.6.2 La notion de fonction de score

Une fonction de score, dans le sens le plus large, est une méthode mathématique qui permet d'estimer le degré d'interaction entre deux entités (généralement un ligand et un site de liaison), et cela de manière très rapide *via* l'utilisation de fortes approximations concernant les phénomènes de reconnaissance moléculaire (voir le §1.2) [15]. Dans ce sens, l'énergie d'un FF, pilotant le positionnement optimal d’un ligand dans le site actif, peut aussi être considérée comme une fonction de score. Une énergie d’interaction est alors obtenue en considérant les composantes intermoléculaires, mais on peut également ajouter une composante intramoléculaire qui permet par

exemple de modéliser la différence d'énergie entre les états lié et non lié du ligand. Cette énergie permet de prédire la géométrie de la pose optimale, mais aussi de classer des ligands les uns par rapport aux autres (“ranking”). Bien qu'elle ne soit pas une estimation de l’énergie libre de liaison, elle pourra quand même servir comme hypothèse de sélection de ligands lors d'un criblage virtuel (potentiellement avec succès si les effets entropiques associés aux ligands ne sont pas prépondérants ou ne varient pas trop).

D’autres fonctions de score retournent cette fois une estimation de l'énergie libre de liaison à partir des coordonnées d'une unique pose. On utilise les FS pour réévaluer *a posteriori* les différentes poses identifiées lors de la phase d'EC. Une FS repose sur un pari particulièrement osé : pouvoir prédire très rapidement une énergie libre (une propriété d’un ensemble de géométries couvrant une zone de l’espace des phases) à partir d’une géométrie unique représentative de cet ensemble. Sous cette hypothèse, la FS serait capable d'identifier le bon mode de liaison d'un composé donné parmi un ensemble de poses variées, et de classer les différentes molécules selon leur ordre de grandeur en terme d'affinité dans le contexte d'un criblage virtuel. La FS apparaît donc comme la clé de voûte de toute approche de docking, et son but très ambitieux et paradoxal par nature en fait également le point le plus critique [133]. En plus du docking à proprement parler, les FS sont également utilisées pour estimer l'affinité de composés dérivés d'un hit ou issus de simulations de type *de novo* design (voir le §1.7).

Enfin, les fonctions de score peuvent être divisées en trois grandes classes : les FS basées sur un FF, les FS empiriques et celles de type potentiels statistiques (“knowledge-based”).

Les fonctions de score basées sur un champ de force

Comme son nom l'indique, ce type de fonction repose sur le formalisme des FF introduit au §1.4.1. Une énergie est obtenue en considérant les composantes intermoléculaires (Van der Waals et électrostatique), auxquelles on ajoute une composante intramoléculaire (énergie de déformation / tension) qui permet par exemple de modéliser la différence d'énergie entre les états lié et non lié du ligand. Enfin, la modélisation de phénomènes de désolvatation du ligand et du site est généralement incluse *via* l'utilisation d'un modèle de solvant implicite. Les principales qualités de cette classe de fonction proviennent de sa capacité à réaliser des optimisations locales par descente de gradient et de pouvoir modéliser d'un point de vue énergétique les réarrangements éventuels du site de liaison. En revanche, ces fonctions sont relativement complexes, et bien qu'il puisse y avoir une corrélation entre l'énergie brute calculée et l'énergie libre de liaison, cette première en est rarement une bonne estimation au kcal/mol près. Le Tableau 2 liste des exemples de FS de ce type.

Les fonctions de score empiriques

Il s'agit d'une somme pondérée d'interactions ligand-récepteur (liaisons hydrogène, électrostatique, contribution hydrophobe, *etc.*) auxquelles des termes de pénalité peuvent être adjoints (clashes L-R, pénalité entropique conformationnelle liée au phénomène de liaison du ligand, *etc.*) [134]. Les poids *w* associés aux différentes composantes (voir l'Eq 30) sont obtenus à l'aide d'une méthode de régression linéaire multiple sur un jeu d'entraînement.

$$\Delta G_{liaison} \propto SCORE$$

où SCORE peut être exprimé de cette manière Eq 30

$$SCORE = w_{LH} * E_{LH} + w_{VdW} * E_{VdW} + w_{Nombre\_de\_DDL} * Nombre\_de\_DDL$$

Ce jeu est composé de complexes ligand-récepteur pour lesquels sont connues à la fois les constantes d'affinité (K_d, K_i, voire IC_{50}) et les structures tridimensionnelles correspondantes. Une homogénéité des types de constante est fortement recommandée pour l'obtention de résultats pertinents et robustes. La phase de paramétrisation va permettre un apprentissage à partir de chaque mode de liaison expérimental et va déterminer les poids consensus aboutissant à la meilleure corrélation entre l'énergie calculée et les constantes expérimentales. L'équation finale (voir l'Eq 30), qui est similaire à un modèle QSAR, permet donc de retourner à partir d'une pose unique une estimation de l'affinité expérimentale. Malgré leur popularité, ces FS ont néanmoins des défauts. Comme pour toute approche de ce type, le degré de confiance envers les résultats est tributaire de l'appartenance au domaine d'applicabilité du complexe évalué [135]. Cependant, cette condition n'est pas vérifiée en pratique. Du fait des contraintes pesant sur l'ajout de complexes dans le jeu d'entraînement, la taille de ce dernier est relativement modeste par rapport au nombre de structures de complexes L-R déposées à la PDB. De plus, certains jeux d'entraînement peuvent être plus ou moins biaisés vers certaines cibles ou classes de ligands. Cela a une conséquence sur la précision de l'estimation énergétique qui n'est pas homogène. Enfin, l'apprentissage ne se fait que sur les modes de liaison expérimentaux, donc sans ajout de leurres telles des mauvaises poses qu'il faudrait réussir à discriminer. Le Tableau 2 liste des FS appartenant à cette classe.

Les fonctions de score de type potentiels statistiques

Ces fonctions intègrent des potentiels statistiques *A(i,j,d)* sur des paires non liées de classes d'atomes (*i*,*j*) séparés d'une distance *d*. La fonction de score PMF est un membre notable de cette classe, et le score retourné par celle-ci est tout simplement la somme des potentiels sur l'ensemble des paires considérées [136]. La paramétrisation des potentiels est issue de l'analyse statistique d'un grand nombre de structures expérimentales (provenant par exemple de la PDB). L'idée sous-jacente est que les interactions entre paires d'atomes qui apparaissent plus fréquemment que selon une distribution de référence sont énergétiquement favorables puisque l'analyse porte sur des structures représentatives de minima énergétiques. Par exemple, une liaison hydrogène entre un accepteur de type *i* et un donneur de type *j* à une distance *d* faible est une interaction très fréquente, d'où une contribution *A(i,j,d)* favorable. A l'inverse, une distance très faible *d'* aboutissant à un clash entre deux atomes de types *i'* et *j'* est rare, d'où une valeur du potentiel *A(i',j',d')* défavorable. Par opposition aux fonctions empiriques, les FS de type potentiels statistiques ne nécessitent que de l'information structurale, ce qui permet de construire des jeux d'entraînement de taille bien plus importante, et donc statistiquement plus robustes et variés [137]. Mais puisque la seule donnée structurale ne permet pas d'y associer une affinité, les estimations énergétiques ne peuvent s'apparenter à une réelle estimation de l'énergie libre de liaison. Enfin, il est à noter que cette approche est purement statistique et implicite dans la mesure où elle ne décompose pas le score final en une somme de composantes physiques avec les problèmes inhérents à l'estimation de chaque composante [137]. Des exemples de FS de type potentiels statistiques sont également listés dans le Tableau 2.

Enfin, il existe également des FS "hybrides" dans le sens où elles mélangent au moins deux classes de FS. Bien que basée sur le FF AMBER, la FS d'Autodock [138] est également empirique dans le sens où elle a aussi été paramétrée sur un jeu d'entraînement afin de retourner une estimation de l'énergie libre de liaison. Cette FS se présente sous la forme d'une combinaison linéaire de termes pondérés, et dont certains sont typiques d'un FF (VdW, électrostatique). La procédure de calibration est comparable à celle décrite dans le paragraphe dédié aux FS dites empiriques.

Malheureusement, à l'heure actuelle, il n'y a pas de FS idéale : les évaluations énergétiques ne sont pas toujours fiables, et aucune FS n'est systématiquement supérieure aux autres quel que soit le système étudié [133, 139].

Classe de la fonction de score	Nom de la fonction de score
Empirique	ChemScore [140]
	LUDI [134]
Champ de force	DOCK [123]
	GoldScore [102]
Potentiels statistiques	DrugScoreX [141]
	PMF [137]
Hybride (Champ de force / Empirique)	AutoDock [138]

Tableau 2: Liste non exhaustive de fonctions de score par classe.
Lorsque la FS n'a pas de nom propre, celui du logiciel l'intégrant est signalé.

1.6.3 Les différents types de simulation

Différents types de simulation de docking (redocking, cross-docking et criblage virtuel) existent et sont brièvement décrits ci-dessous. Ils sont également complémentaires lors des phases de validation d'une nouvelle méthode.

Le redocking

Le redocking consiste à tenter de repositionner un ligand dans son site de liaison à partir du complexe L-R correspondant, duquel le composé a été préalablement extrait. En d'autres termes, un site apte à lier le ligand que l'on souhaite docker est préparé à partir du complexe L-R d'intérêt, d'où l'appellation "redocking". Si la conformation bioactive du ligand est utilisée, un docking rigide peut suffire à reproduire la géométrie du complexe, alors que si l'on part d'une conformation aléatoire, un docking semi-flexible s'avère souvent nécessaire. Bien entendu, il s'agit théoriquement de simulations relativement aisées, mais en pratique les choses ne sont pas si simples, notamment à cause de la grande flexibilité de certains ligands et des carences de la FS qui peut favoriser un mauvais mode de liaison. Le redocking reste cependant l'étape fondamentale pour valider un nouvel algorithme de docking. Des jeux d'entraînement spécifiques [142, 143], contenant des complexes L-R de haute résolution, des cibles diverses et des ligands de taille et châssis moléculaire variés, ont également été construits pour faciliter la validation de nouvelles méthodes et la comparaison d'algorithmes différents.

Le taux de succès d'un algorithme de docking intègre à la fois la qualité d'implémentation de sa stratégie d'EC et la précision de sa FS pour l'évaluation énergétique des poses. La partie relative à l'EC

est essentiellement validée par la capacité à reproduire, sans aucune considération énergétique, le bon mode de liaison du ligand pour la quasi-totalité des complexes du jeu. Le critère usuel de succès est le RMSD ("Root-Mean-Square Deviation", voir l'Eq 31) qui estime globalement la distance moyenne entre atomes homologues de deux conformations différentes du même système, comme par exemple une pose de docking par rapport au mode de liaison expérimental.

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\left(X_{i_{Pred}} - X_{i_{Exp}}\right)^2 + \left(Y_{i_{Pred}} - Y_{i_{Exp}}\right)^2 + \left(Z_{i_{Pred}} - Z_{i_{Exp}}\right)^2\right]} \qquad \text{Eq 31}$$

avec N le nombre d'atomes, les étiquettes Pred/Exp qui caractérisent respectivement la pose prédite et la pose expérimentale, et enfin X/Y/Z qui représentent les coordonnées cartésiennes dans un espace 3D. Un seuil maximal de 2 Å est souvent utilisé pour discriminer une bonne pose d'une mauvaise pose [144]. Le RMSD est une métrique simple à calculer et intuitive en ce qui concerne son interprétation, d'où sa forte popularité. Néanmoins, le RMSD est aussi sensible à la taille des entités : un même seuil de RMSD est difficilement utilisable pour du docking de fragments, de ligands drug-like et de protéines (docking protéine-protéine). De plus, le RMSD usuel souffre aussi de la non prise en compte de la symétrie ou pseudo-symétrie de certaines molécules, d'où une valeur qui peut être très élevée alors qu'une grande partie des contacts est satisfaite. Cet inconvénient majeur peut être résolu avec un RMSD insensible aux éléments de symétrie. Un RMSD élevé, synonyme d'échec, peut également être obtenu dans le cas d'un ligand possédant un groupement flexible exposé au solvant et ce malgré une parfaite reproduction de sa partie enfouie. Des métriques alternatives, comme les empreintes d'interaction [1, 145] ou RSR [146], ont été développées pour pallier ces aspects négatifs (voir le §1.6.4). En redocking, la qualité de la fonction de score se mesure avec le même critère que pour la qualité d'EC, mais en ne prenant en compte pour chaque complexe que la ou les meilleures poses au regard de ladite fonction.

Le cross-docking

Le cross-docking s'apparente au redocking dans le sens où il n'est toujours question que de prédire la géométrie d'un complexe pour lequel le composé à docker est connu pour se lier à la cible. La principale différence, qui est cependant de taille, réside dans le fait que le site utilisé n'est pas forcément la conformation capable d'accommoder le ligand d'intérêt. En pratique, une validation de type cross-docking consiste à sélectionner n complexes et à docker chaque ligand dans chaque site, d'où un mélange de simulations (redocking et cross-docking). Il est à noter que le cross-docking est

plus proche de la réalité que le redocking, comme par exemple lorsque l'on cherche à prédire le mode de liaison d'un actif connu en utilisant une structure du site récupérée à partir de la PDB. Le fait de tenter de docker un ligand dans un site potentiellement non apte à l'accueillir, car nécessitant des réarrangements conformationnels, explique en lui-même la plus grande difficulté de ce type de simulation et la baisse des taux de succès [132, 147]. Seules des approches impliquant au moins une flexibilité partielle du site sont réellement aptes à donner de bons résultats en cross-docking de manière générale.

Les critères de succès du cross-docking sont globalement similaires à ceux du redocking, à la différence qu'une superposition des différents sites par rapport à une référence commune s'avère nécessaire pour calculer des RMSD corrects.

Le criblage virtuel

Ce dernier type de simulation ajoute une difficulté supplémentaire : non seulement les molécules ne sont pas dockées dans leur site natif, mais de plus la méthode doit être capable de discriminer les molécules inactives ou faiblement actives au profit des ligands réels, d'où l'importance d'avoir la fonction de score la plus fiable possible. Conceptuellement, le criblage virtuel s'apparente à n simulations de docking, n étant la taille de la chimiothèque à cribler [148]. Selon la taille de cette dernière (de quelques molécules jusqu'à plusieurs millions de composés), l'utilisateur devra choisir l'outil le plus adapté à ses besoins en fonction de la qualité souhaitée des résultats et des DDL considérés d'une part, et de la rapidité d'exécution vis-à-vis de la puissance machine disponible d'autre part. Enfin, le criblage virtuel est la stratégie utilisée dans le cadre de projets de drug design prospectifs pour identifier de nouveaux hits potentiels à confirmer de manière expérimentale.

Etant donné que le criblage virtuel implique des systèmes différents (un site de liaison unique mais divers composés organiques), il apparaît évident que le rôle de la fonction de score est prépondérant ne serait-ce qu'à travers le classement des différents composés les uns par rapport aux autres. Bien que fondamentale dans une optique d'optimisation, la capacité à reproduire le bon mode de liaison en criblage virtuel est assez peu évaluée, et est remplacée par deux autres critères qui sont l'enrichissement et l'aire sous la courbe ROC [149]. Le facteur d'enrichissement (“enrichment factor”) EF(n%) mesure la capacité de la méthode à concentrer les composés actifs en tête de liste dans les n premiers % de la chimiothèque criblée, et s'exprime de cette manière :

$$\mathrm{EF(n\%)} = \frac{\left(\frac{\text{Nb actifs dans le top n\%}}{\text{Nb molécules dans le top n\%}}\right)}{\left(\frac{\text{Nb actifs dans la chimiothèque}}{\text{Nb molécules dans la chimiothèque}}\right)} = \frac{\left(\frac{\mathrm{TP}}{\mathrm{TP+FP}}\right)}{\left(\frac{\mathrm{TP+FN}}{\mathrm{TP+TN+FP+FN}}\right)} \qquad \text{Eq 32}$$

où TP est le nombre de vrais positifs, FP le nombre de faux positifs, TN le nombre de vrais négatifs et FN le nombre de faux négatifs. Une méthode de criblage efficace retournera donc un EF élevé (EF>>1), en supposant qu'il existe des actifs dans la chimiothèque. Il est à noter qu'en pratique, seuls les EF(n%) avec n petit (en pratique de 0,01% à 10% environ) sont réellement importants, puisque ce sont les molécules en tête de liste qui seront préférentiellement testées expérimentalement. L'autre critère, appelé courbe ROC ("Receiver Operating Characteristic", voir la Figure 30) [149], est une courbe paramétrique représentant le taux de vrais positifs TPR ("True Positive Rate", également connu sous le nom de Sensibilité) en fonction du taux de faux positifs FPR ("false positive rate", équivalent à 1 - Spécificité). Le paramètre est souvent une probabilité, mais dans le contexte du docking, le paramètre considéré est le score attribué à chaque ligand après un tri préalable (du plus favorable au moins favorable). Bien que différente de l'EF dans sa définition, la courbe ROC a également pour but d'estimer les performances d'une méthode de classification, notamment à travers sa capacité à discriminer molécules actives et inactives.

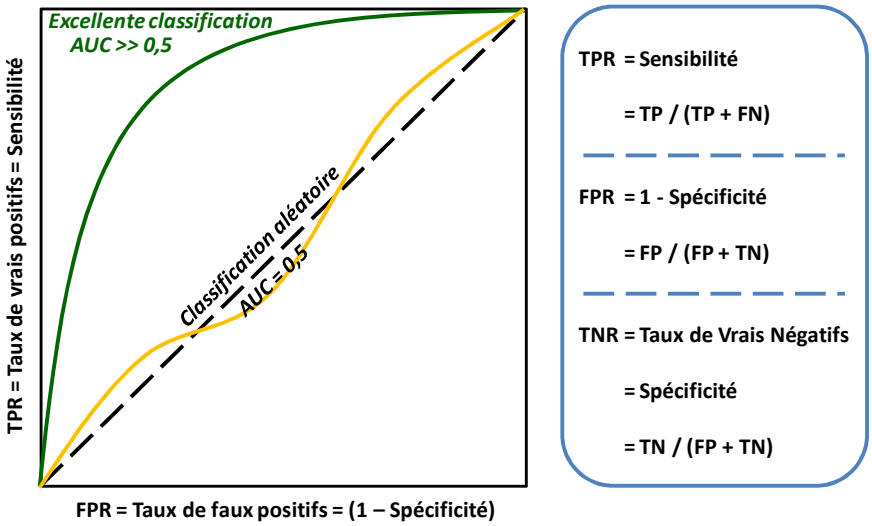


Figure 30: Superposition de deux courbes ROC représentatives d'une excellente classification (courbe verte) et d'une classification moyenne (proche d'un tirage aléatoire, courbe orange).

Contrairement au facteur d'enrichissement, les deux variables d'une courbe ROC sont bornées et appartiennent à l'intervalle [0;1]. Ainsi, il est possible de calculer une aire bornée sous la courbe ROC, notée AUC ("Area Under the Curve"), à des fins de comparaison entre diverses méthodes. Pour résumer, la courbe illustre de manière détaillée les performances de la méthode, tandis que la valeur AUC résume à elle seule ses performances globales.

1.6.4 Les stratégies post-traitement

Du fait de la très relative fiabilité des fonctions de score [139], diverses stratégies de post-traitement ont été développées pour corriger ce défaut majeur. L'approche la plus intuitive, appelée "consensus scoring" consiste à combiner l'information de plusieurs fonctions de score - préférentiellement de natures différentes [150] - afin de ne sélectionner que les composés ayant une bonne évaluation sur plusieurs d'entre elles [151]. Cette méthode permet d'atténuer les erreurs (surestimation et/ou sous-estimation) inhérentes à chaque fonction considérée séparément. Une alternative consiste à réévaluer énergétiquement les différentes poses avec des méthodes plus rigoureuses, limitant de fait leur large application. L'utilisation de modèles complexes de solvant implicite comme MM-GBSA [152] ou d'un Hamiltonien de nature quantique [153] a été décrite dans la littérature.

Afin de diminuer les éventuelles limitations de chaque programme de docking en terme d'EC, il a été proposé d'utiliser une démarche dite "consensus docking" qui analyse les poses générées par des outils distincts lors d'une phase de post-traitement [154].

Enfin, ce type d'analyse peut être purement géométrique *via* l'utilisation d'empreintes d'interaction qui encodent diverses classes d'interactions L-R de manière facilement manipulable par un ordinateur (par exemple, sous forme d'une chaîne de bits). A partir d'une empreinte de référence comme le mode de liaison d'un actif connu, il est aisé d'extraire les composés interagissant avec la cible d'une manière similaire à celle de l'actif [1, 145, 155], avec l'idée sous-jacente que les molécules partageant un mode de liaison commun avec une molécule active ont une probabilité plus importante d'être elles-mêmes actives par rapport à des molécules choisies uniquement sur la base de leur score.

1.7 Le de novo design

Le terme *de novo* design (DND) fait référence à des approches de modélisation moléculaire ayant pour but de construire une molécule active à partir de la seule structure d'un site de liaison. Dans le cas où aucun actif n'a été mis en évidence et que seule la structure 3D de la cible est connue, le DND peut apparaître comme une alternative au docking. Ce cadre minimal peut généralement être complété par des contraintes pharmacophoriques, ciblant des résidus connus pour interagir avec des actifs de référence, afin de guider le processus de construction du ligand. De nombreux algorithmes de DND reposent dans un premier temps sur l'identification de zones énergétiquement favorables au sein du site de liaison pour un panel de sondes organiques (cyclohexane, phénol, méthanol, urée, *etc.*). Vient ensuite l'étape de construction du ligand à proprement parler *via* la liaison de différentes sondes entre elles au moyen de règles de connexion. A titre d'exemple, LUDI [156], MCSS [157], SPROUT [158], LigBuilder [159], SkelGen [160] et TOPAS [161] sont des programmes de DND plutôt connus dans le domaine. En soi, le DND est une stratégie de construction de molécules et se base sur le docking/scoring pour évaluer la pertinence des composés générés pendant la phase de construction (ajout/suppression de groupements, modification de la nature d'un atome, *etc.*) dans le but d'optimiser l'affinité calculée pour la cible. Cependant, puisque le DND repose sur le docking/scoring, la faible fiabilité de prédiction des affinités y est implicitement héritée. Néanmoins, quelques réussites notables ont été publiées au fil des ans, et elles sont décrites dans une revue dédiée au *de novo* design [162].

Etant donné qu'il y a une construction virtuelle de composés, l'autre point critique concerne bien évidemment l'accessibilité chimique des molécules issues du processus (la faisabilité chimique et la facilité de synthèse). Ce dernier aspect a été en partie amélioré *via* l'utilisation de règles chimiques rétro-synthétiques (par exemple les règles RECAP qui sont construites à partir de réactions chimiques usuelles [163]) pour combiner virtuellement les différents éléments d'une manière raisonnable. De récentes approches, plus complexes mais également plus réalistes comme l'encodage de certaines réactions chimiques usuelles (réactifs→produits) [164, 165], ont permis d'atteindre un haut niveau d'accessibilité chimique pour les molécules construites *de novo* [166-168].

La phase de liaison des sondes moléculaires en DND est conceptuellement très similaire au type d'optimisation "linking" en FBDD (voir le §1.3), d'où ces quelques lignes l'introduisant. Il est à noter que de nombreuses méthodes ont été développées bien avant la récente expansion du FBDD expérimental, et un certain regain d'intérêt est observé depuis pour le DND.

1.8 Les objectifs de cette thèse

Malgré l'expansion rapide du FBDD expérimental, peu de méthodes *in silico* dédiées à ce nouveau paradigme avaient été développées au commencement de cette thèse. Depuis, un certain nombre d'approches ont été publiées. Elles ont d'ailleurs fait l'objet d'une description dans notre revue consacrée au FBDD [56] et dans l'article dédié à l'optimisation virtuelle de fragments (voir le §5.5).

Ce travail de thèse porte sur le développement, la validation et l'application de techniques de modélisation adaptées au FBDD. Autrement dit, son objectif principal est la mise au point d'un protocole *in silico* dont la finalité serait d'optimiser chacune des étapes clés et communes à toute stratégie de FBDD (voir la Figure 31) :

1) la conception de chimiothèques de fragments à l'aide d'outils chémoinformatiques, afin de ne sélectionner que des composés présentant entre eux une certaine diversité chimique tout en contenant des groupements fonctionnels pour la nécessaire phase d'optimisation
2) la prédiction du mode de liaison de molécules de type fragment par des approches de docking moléculaire, et sa généralisation dans un contexte de criblage virtuel en vue d'extraire un sous-ensemble de la chimiothèque à cribler expérimentalement de manière prioritaire
3) la suggestion d'optimisations potentiellement intéressantes du ou des fragments initiaux (hits) en se basant sur les notions de "growing" et "linking" introduites précédemment (voir le §1.3)

Au démarrage de ce projet, il a été convenu entre les parties prenantes de ne se focaliser que sur les objectifs 2) et 3).

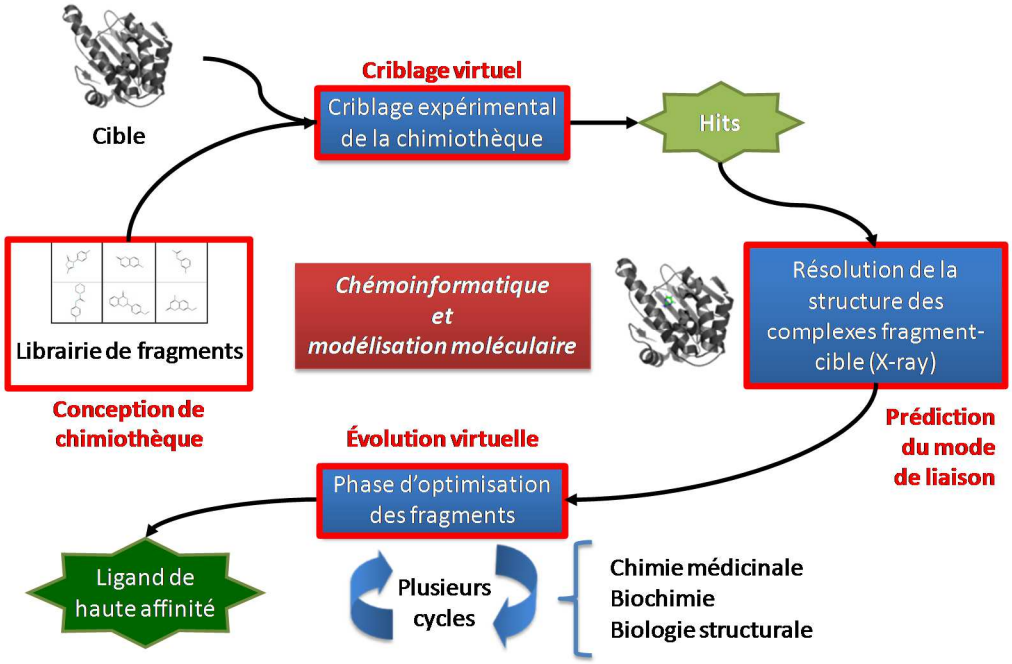


Figure 31: Exemple de démarche classique de FBDD.

Les rectangles rouges indiquent les étapes clés dans lesquelles la chémoinformatique et la modélisation moléculaire peuvent être impliquées.

2 Matériel et méthodes

Ce chapitre contient une description des principaux outils utilisés et/ou développés dans le cadre de ce travail.

2.1 Présentation de S4MPLE

2.1.1 Caractéristiques générales

S4MPLE (Sampler for Multiple Protein-Ligand Entities) est un outil développé au Laboratoire de Chémoinformatique de l'Université de Strasbourg sous la responsabilité du Dr. Horvath, et dont le projet a été initié en amont de ma thèse. Il s'agit d'un programme de modélisation moléculaire utilisant une fonction d'énergie basée sur le formalisme des FF, et dont l'échantillonnage conformationnel repose sur un algorithme génétique. La combinaison d'un algorithme évolutif avec un FF permet de réaliser des optimisations locales, comme par exemple une minimisation de l'énergie potentielle du système. Ce genre d'algorithme hybride est également connu sous le nom d'algorithme génétique de type Lamarckien [101].

L'algorithme d'échantillonnage de S4MPLE est complètement général vis-à-vis des DDL du système, ces derniers pouvant être spécifiés précisément. Par défaut, tous les atomes sont considérés comme libres, et un fichier de configuration (“fixed_atoms”) permet de préciser les atomes fixes du système. Dès lors, il est possible d'envisager des simulations très variées (voir la Figure 32) : échantillonnage conformationnel d'un ligand, tentative de repliement *ab initio* d'un oligopeptide, docking d'un ou de plusieurs ligand(s) dans un site totalement rigide ou partiellement flexible.

Des informations complémentaires, de nature technique et/ou algorithmique, sont disponibles dans le guide d'utilisateur (voir le §2.1.3) et l'article I (voir le §3.4).

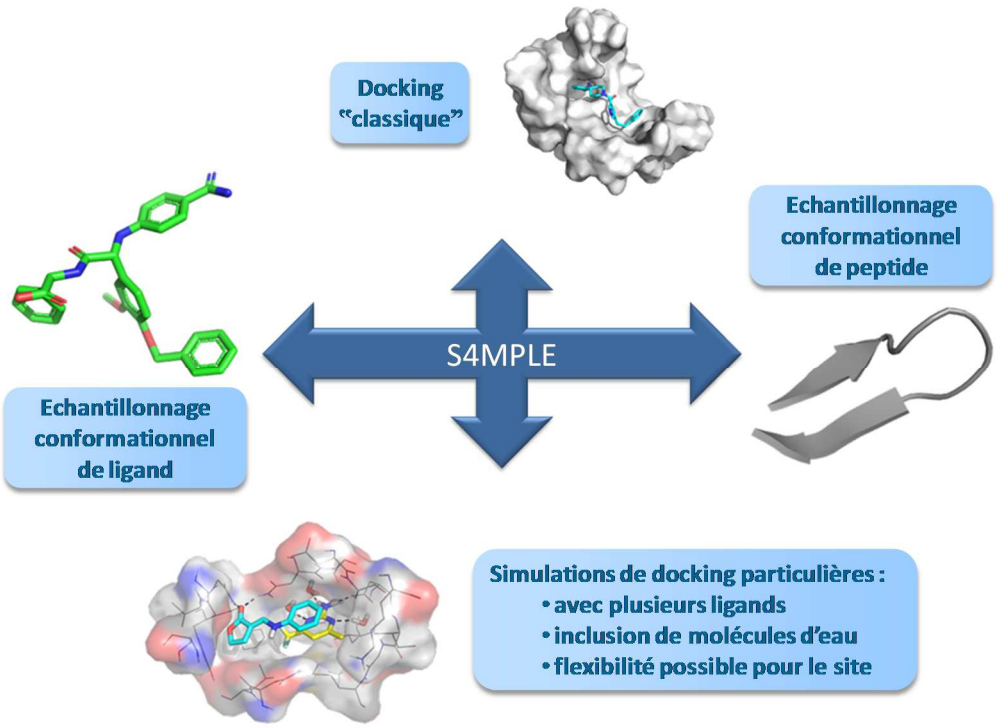


Figure 32: Les principaux types de simulations théoriquement réalisables avec S4MPLE.

S4MPLE ne fait pas de distinction conceptuelle entre le site de liaison et le ligand à docker, comme c'est généralement le cas pour les outils classiques de docking qui attendent explicitement ces deux types de fichier en entrée. Ainsi durant l'initialisation du programme, les fichiers moléculaires présents (ligand | peptide | ligand + site de liaison) sont simplement interprétés afin de définir et de paramétrer le système considéré. En fonction du statut (fixe ou libre) des atomes, les DDL explicites à échantillonner (comme les angles de torsion) sont automatiquement détectés sur l'ensemble du système. Il y a cependant deux limitations inhérentes à toute approche de MM, la première étant le temps de calcul alloué, et la deuxième la disponibilité de paramètres de FF pour l'intégralité du système étudié.

2.1.2 Architecture de S4MPLE

D'un point de vue technique, le programme est écrit en Pascal Objet et la compilation se fait avec le compilateur FreePascal [169] *via* l'environnement de développement Lazarus [170]. Le Pascal Objet est un langage de programmation orientée objet (POO) fortement typé, insensible à la casse, et dont la déclaration des variables se fait dans un bloc préliminaire. Cette dernière propriété confère à ce langage une très bonne lisibilité au détriment d'une certaine souplesse, comme la possibilité de déclarer une nouvelle variable de portée locale au sein d'une boucle. Un fichier source Pascal commence par un mot clé spécifiant le type de celui-ci, à savoir un programme ou un module/unité.

Une unité, contrairement à un programme, n'est pas indépendante et peut être utilisée par un programme ou une autre unité.
L'architecture de S4MPLE se compose du programme principal et de plusieurs unités. Les différentes unités contiennent soit des fonctions de base (unités dites utilitaires), soit la description d'une ou plusieurs classes. Par convention, le nom de toutes les classes commence par la lettre T. Une description concise de chaque unité est donnée ci-dessous.

Programme S4MPLE
Il s'agit du programme principal qui contient la définition d'une classe héritant de *TCustomApplication*. Parmi les éléments majeurs, la redéfinition de la méthode *DoRun*() de la classe *TCustomApplication* permet de spécifier ce qui doit être exécuté au lancement de l'application. Le traitement des arguments, spécifiant les actions à mener (lecture des fichiers moléculaires et des paramètres du FF, minimisation du système ou appel à l'algorithme génétique), est réalisé dans cette méthode fondamentale. Les objets issus des classes *TMolecule* et *TForceField* sont systématiquement crées, alors que ceux instanciés à partir des classes *TEFunc* et *TEvolAlgorithm* le sont en fonction des arguments donnés lors du lancement du programme.

Unité ForceField
Cette unité sert uniquement à lire les fichiers de paramètres du FF sélectionné, ainsi que de précalculer un certain nombre de termes (voir unité *FFEngine*). L'objet, instancié à partir de la classe *TForceField*, est détruit après la phase d'initialisation du programme.

Unité Geometry
Cette unité contient les classes représentant les briques élémentaires en MM, ainsi que des méthodes de base concernant le maniement de vecteurs (produits scalaire et vectoriel, normalisation, *etc.*) dans un espace tridimensionnel. Comme leurs noms l'indiquent, les classes *TAtom*, *TPair*, *TAngle* et *TTorsion* contiennent respectivement les définitions d'un atome, d'une paire d'atomes, d'un angle de valence et d'un angle de torsion. Par exemple, un objet, instance de la classe *TAngle*, contient notamment la valeur courante de l'angle de valence, les références vers les trois objets *TAtom* définissant cet angle, la référence du terme du FF associé aux types atomiques considérés, en plus de méthodes permettant de calculer la valeur de cet angle et les dérivées correspondantes à partir de coordonnées cartésiennes.

Unités MolConstants et SamplingTypes

MolConstants est une unité utilitaire contenant des constantes chimiques issues du tableau périodique, ainsi que des paramètres internes considérés comme immuables, par opposition aux paramètres spécifiés dynamiquement lors de l'exécution du programme. *SamplingTypes* contient simplement la définition de toutes les structures de données. Par exemple, une entité *TContrib* est définie pour stocker l'énergie potentielle du système ainsi que ses différentes composantes.

Unité Fragments

Un objet *TFragment* est un objet sophistiqué focalisé sur un DDL qui peut être intra- ou intermoléculaire. Le premier type correspond aux DDL explicites usuels en EC que sont les angles dièdres. Le second type implique des entités distinctes, comme par exemple un ligand qui sera positionné par rapport à son site au moyen d'un contact favorable. Les fonctions mutation et crossing-over de l'algorithme génétique opèrent sur des objets instanciés à partir de la classe *TFragment* quel que soit leur type : le choix d'implémentation initial permet donc un traitement unifié des différents DDL. Un objet *TFragment* contient également la liste des atomes pilotés par le mouvement de ladite torsion (cas d'un DDL intramoléculaire), et une méthode permettant d'imposer une valeur donnée à l'angle de torsion *via* une rotation autour de cet axe arbitraire dans un repère 3D. Il est à noter que cette dénomination "fragment" a été employée lors du développement initial de cet outil (en amont de cette thèse) et a été maintenu par la suite. Ainsi, dans le seul contexte de l'architecture de S4MPLE, le terme fragment ne fait aucunement référence à un ligand de faible masse.

Unité FFEngine

Il s'agit de la classe qui sert au stockage en mémoire des paramètres du FF pour le système d'intérêt. Pour chaque composante de l'équation de calcul de l'E_{pot}, il y a un objet associé qui contient les paramètres du FF nécessaires pour simuler le système : *TBondEngine* (liaison covalente), *TNBEngine* (interactions non liées), *TAngEngine* (angle), *TTorsEngine* (torsion), *TChirOopEngine* (maintien d'une chiralité initiale et contrainte sur les angles impropres), *TContEngine* (contact favorable), *TConstraintEngine* (contraintes). L'objet *TFFEngine* est simplement un objet contenant les références de l'ensemble de ces composantes, et ces dernières sont toutes des fonctions de coordonnées internes ou de distances interatomiques.

Unité Fingerprint

Le fingerprint d'interaction, dont le but est d'encoder la configuration du système, est défini par la classe *TFingerprint*. Chaque élément/case du fingerprint, instance de la classe *TInvolvedPairs*,

contient les paires équivalentes afin de produire un fingerprint insensible à d'éventuelles symétries présentes dans le système (par exemple, une fonction acide carboxylique déprotonée).

Unité MolUnit

L'unité *MolUnit* définit la classe *TMolecule* qui contient tous les éléments nécessaires pour modéliser un système moléculaire (liste des atomes, liaisons, angles, torsions, le nombre d'entités, *etc.*). Cette classe recèle également des objets complexes comme le *FFEngine* et le fingerprint d'interaction. Le choix d'intégrer au sein de cette classe des méthodes qui vont au-delà de la simple gestion d'une molécule a été fait lors de la phase de développement initiale. La classe *TMolecule* comporte des méthodes permettant de lire/écrire des fichiers moléculaires (sdf, car, mol2), de mettre à jour les coordonnées internes, de calculer l'E_{pot} et les dérivées en communiquant avec d'autres objets instanciés comme le *FFEngine*. L'objet unique, issu de l'instanciation de cette classe, est donc la clé de voûte du programme. Lorsque l'on souhaite minimiser l'E_{pot} du système (voir unité *EFunctions*) ou échantillonner celui-ci (voir unité *EvolAlgorithm*), la référence d'un objet *TMolecule* est systématiquement passée aux constructeurs correspondants.

Unités EFunctions et OptimTools

OptimTools est également une unité utilitaire. Elle contient une implémentation, adaptée aux structures de données définies initialement dans l'unité *SamplingTypes*, des différents algorithmes de minimisation disponibles dans la librairie *DMath* [171] :

- algorithme de la plus forte pente (Steepest Descent - SD)
- algorithme du gradient conjugué (Conjugate gradient - CG)
- algorithme de Broyden-Fletcher-Goldfarb-Shanno (BFGS)

L'unité *EFunctions* définit la classe *TEfunc*, dont les principales fonctions reposent sur les algorithmes d'optimisation implémentés dans l'unité *OptimTools*. L'opérateur de minimisation locale exhaustive, alternant des descentes de gradient sur une surface perturbée ou non d'E_{pot}, est implémenté dans cette unité.

Unité Chromosome

Comme son nom l'indique, cette unité définit un objet chromosome (classe *TChromosome*). Il contient entre autre l'énergie courante, les valeurs des DDL d'un individu donné et l'implémentation des différents opérateurs génétiques.

Unité EvolAlgorithm

Les différentes stratégies d'EC basées sur le formalisme des AG sont décrites dans la classe *TEvolAlgorithm* de cette unité.

2.1.3 Guide d'utilisateur

Le guide d'utilisateur est inséré dès la page suivante. Il répertorie de nombreuses informations techniques relatives à l'installation de S4MPLE, aux fichiers moléculaires supportés, aux différents fichiers de configuration, aux options disponibles lors du lancement d'une simulation, et enfin à la préparation des fichiers moléculaires (macromolécules et ligands).

S4MPLE

User-guide

&

Technical Reference

Chém Informatique
STRASBOURG

12-2012

Overview

1 Introduction

S4MPLE (Sampler For Multiple Protein-Ligand Entities) is a flexible molecular modeling tool, supporting empirical force field-driven conformational sampling and geometry optimization heuristics using a hybrid genetic algorithm (GA).

Allowing full control of the considered degrees of freedom (DoF), S4MPLE is a completely general approach to visit the conformational space of arbitrary molecules or molecular complexes. In theory, S4MPLE can be used in:

- conformational sampling of small organic compounds (1 entity),
- oligo-peptide folding (1 entity),
- protein loop repositioning in protein homology models (1 entity),
- ligand docking using various scenarios, from full rigid to partially flexible binding site (2 entities),
- simultaneous docking of several ligands into a same binding site (at least 3 entities).

By default, all atoms are mobile unless their rank numbers (in the order in which they are input) are listed in the "fixed_atoms" file (see §3.2.1). The list of explicit DoF, internally called fragments, to be used by the genetic algorithm is built automatically, based on chemical common sense (double bonds will not serve in torsion-drive mutations). Rings will be broken into fragments and subjected to intensive sampling only if a ring bond is listed in the "broken_bonds" file (see §3.2.3). Otherwise, ring geometry may only change due to external forces acting thereon during regular gradient-based minimizations

S4MPLE, written in object-Pascal, is used in command-line mode. As all molecular modeling approach, its main limitations are:

- the studied system size *vs.* available computational resources,
- the availability of force field (FF) parameters for the studied system.

This document facilitates, for the end-user, the installation and a first use of S4MPLE.

2 Installation procedure

S4MPLE has been compiled and tested on a linux x86_64 distribution (Mandriva 2010.2), and a shell is recommended to run it.

The latest version of S4MPLE (as binary file) is available from the laboratory website: http://infochim.u-strasbg.fr

After the extraction of the downloaded tar.gz file, you should have the following files and directories:

- Files
    - S4MPLE: binary file of the program
    - this user-guide in PDF format
- Directories:
    - parm_ff/: it contains the whole set of parameters for both available native FF (CVFF [1] or AMBER/GAFF [2, 3]). The latter FF is systematically used in any reported simulations.
    - parm_ff_fit/: it contains some additional parameters.
    - scripts_prepare_lig/: scripts needed to parameterize ligand molecule(s)
    - scripts_tuto/: scripts piloting the listed tutorials
    - tutorial_sampling/: an example to do conformational sampling of one ligand
    - tutorial_docking/: an example to do standard ligand docking
    - sup_data/: Supplementary material concerning redocking/benchmarking studies described in publications

2.1 Default setup

This default setup is the easiest way to install and run S4MPLE. It will works for projects which do not need the preparation of new ligands. Obviously, this default setup is sufficient to play with S4MPLE using the provided tutorials.

The default setup, step by step:

1) Move both parameter folders parm_ff and parm_ff_fit to some dedicated location on your file system..

2) Set (in your default .<whatever shell you use>rc setup file) the environment variable **SETUP_DIR** to point to the directory in which the parameter folders reside (so that 'ls –l $SETUP_DIR/parm_ff' will display the parameter folder contents)

3) Execute the following installation commands
    - cd $SETUP_DIR/parm_ff/
    - ./**set_amber99_gaff_core_ff.sh**
    - cd $SETUP_DIR/parm_ff_fit/
    - ./**set_amber99_gaff_fit_ff.sh**

4) For convenience, add the S4MPLE binary file to some directory in your **PATH**.

Now you should be able to launch both provided tutorials:

- Conformational sampling of an organic compound (see §5.1),

- Docking of a ligand into its binding site (see §5.2).

2.2 Full setup (involving AMBERTools)

In practice, the full setup is recommended for the end-user, since this procedure is needed in all cases where a new ligand is involved.

The AMBERTools software suite [4] is needed to prepare any new ligand. The release 1.3 of AMBERTools was initially used, and all tests were performed on this version.

AMBERTools can be downloaded, free of charges, at http://ambermd.org.

The full setup, step by step:

1) do the default setup (as described in §2.1)

2) install **JAVA**, available in any Linux distribution or from http://www.java.com

3) install the **ChemAxon** API, available from http://www.chemaxon.com

4) install a **Tcl** interpreter, available in any Linux distribution or from http://www.tcl.tk

5) install the **AMBERTools** software suite (see the provided installation file)

6) set the absolute path of the AMBERTools directory as the environment variable **AMBERHOME**
    - ls –l **$AMBERHOME/bin/antechamber $AMBERHOME/bin/parmchk** must not return any errors

7) For convenience, add the "scripts_prepare_lig" directory to your PATH

You should now be able to prepare new ligands according to the workflow described in details in §4.2

3 Technical reference

This chapter describes the technical points which are necessary for the end-user. It includes obviously file formats and available options. For more information about the genetic algorithm implementation or the FF-based energy function, the reader can refer to the associated papers[5, 6].

3.1 The input files and their formats

S4MPLE is able to use input files of various formats:

- .mol2: file in the SYBYL MOL2 format.
- .car: file in the CAR format originally introduced by BioSym InsightII (PDB-like with additional information like partial charges).
- .sdf: file in the MDL format.
- .mol: file in the MDL format.
- .out: this "chromosome file" contains poses or conformers encoded as a linear string. This kind of file is generated by S4MPLE and can be re-used to generate full-blown molecular files (one conformer per line).
- .fp: binary interaction fingerprint file. This kind of file is generated by S4MPLE and can be re-used to compute the fingerprint difference between that of the current conformer and that of a reference (alternative to RMSD criterion), or to guide conformational searching towards (attractor) or away (tabu) from conformations with interaction patterns similar to the one described in the .fp file

S4MPLE will seek in the working directory for input molecular structure files with the prefix "ref", in that order:

1) ref.mol2, ref.car: target file or protein-ligand complex file previously saved with S4MPLE,
2) ref.sdf, ref.mol: ligand file (mono or multi-structures).

It should be noted that at least one ref.xxx file must be present, and symbolic links can be used in practice to avoid renaming the input files.
Molecular output files will be, by default, generated using the input format. Their automatically generated names will include the current energy level and the pose rank, like in "best-30.8_0.mol2" – the pose of energy=-30.8 kcal/mol is ranked 0 (best pose found). Furthermore, additional information is stored in these files (in comment fields): the energy and its components, RMSD values with respect to input geometry of the ref file, *etc.*

3.2 Main configuration files

Various configuration files can be used, in addition to molecular files, to control the simulation. The most important ones are "fixed_atoms" and "hot_spots".

3.2.1 File fixed_atoms

This file is used to set the status (free or flexible) of all atoms in the system. It lists the fixed atoms in the system with 1 atom ID per line. The atom "ID" simply refers to the rank number of atom in the final list managed by the program. This follows the order in which atoms are read from input files, with assumed protein site files (of type .mol2 or .car) read first. Therefore, the bad news is that you need to perform some arithmetic if, for example, you wish to declare the second atom of the LIGAND as fixed, you need to add the atom ID N_{site}+2 in the fixed_atoms file, where N_{site} is the number of atoms in the receptor file. The good news is that you may block some ligand atoms, and free some protein moieties if you wish – classical docking/sampling tools will typically not allow this. By default (no fixed_atoms file), all atoms are considered flexible.
To perform a semi-flexible docking (free ligand but rigid binding site), just copy the atom lines from the site (mol2 file) in the fixed_atoms file.
Albeit S4MPLE does not need to make any assumption about the nature of the submitted molecules, in practice some of the docking options (hot spot definitions, use of geometric centers to direct ligand pose - see further on) do need to know which entity is the site, and which is the ligand. Please note that, by default, the site file is expected to contain at least one fixed atom – otherwise, S4MPLE will consider all the present entities as equivalent partners.

3.2.2 File hot_spots

This configuration file lists the hot spots for the initial ligand placement in the binding site. As for the fixed_atoms file, 1 atom ID per line must be provided. This gives a list of preferential anchoring points in the "active site" of the protein – it basically tells the tool where the "active site" is to be expected. Anchoring points must be either Hydrogen bond donors, acceptors or hydrophobic groups, which will be paired with potential partners from the ligand (in the ligand, all the H-bonding partners and hydrophobes count as "hot").

This file is useless in the context of a single entity: hot spot pairing serves to enable genetic operators to implicitly deal with inter-molecular degrees of freedom, implicitly controlling the position of the ligand with respect to the site. When a single molecular graph is subjected to sampling, covalent bonds are used to modify geometry.

If no hot_spots file are provided (not knowing where the active site is would be a good excuse), all qualifying (hydrogen bond acceptor/donor or carbons) site atoms are considered "hot". The latter case leads to full blind docking, but obviously needs a larger number of generations. This file can be automatically generated, if desired (see §3.3.5). Note that you need not specify ALL the potential anchoring points of the active site as hot_spots. This selection has no impact on force field energy calculations, it just pilots the initial attempt to randomly position the ligand with respect to the site. The algorithm will, for example, randomly draw hat spot #5, a hydrophobic carbon of the site. Then it will randomly pick one of the hydrophobic carbons of the ligand (if there are none, the tentative is aborted and a new site hot spot is picked). Say that the drawn lottery ball is carbon #7 of the ligand – then, the software will try to place the ligand in such a way as to establish a hydrophobic contact between hotspot 5 and carbon 7 (all while avoiding clashes, as far as possible). If, while trying to do this, the software also happens to bring together hydrogen bond donors and acceptors that are not listed as hot spots – very well, that will definitely count in energy evaluation. Therefore, it is recommended to list only the *most deeply buried* anchoring points of the site in hot_spots: the initial positioning of the ligand close to those will certainly cause other contacts to form spontaneously (if the correct pose does not imply interaction with the hot spots – that is not a problem, either: subsequent energy minimization may push the ligand away if needed: pushing the ligand out of the site is easy – dragging it in is difficult).

3.2.3 File broken_bonds

This file contains the bonds to break during the creation of the explicit DoF. One broken bond is defined by listing both involved atom IDs, separated by free spaces, in the bond (one definition per line).

Rings will be broken into DoF and subjected to intensive sampling only if a ring bond is listed in the "broken_bonds" file. Otherwise, ring geometry may only change due to external forces acting thereon during regular gradient-based minimizations. Automated recognition of rigid rings (that do not require any sampling) *vs* flexible rings that should be "opened", *i.e.* automatic generation of the "broken_bonds" file, is envisaged but not yet undertaken.

This option can be used to enhance the sampling of a flexible protein loop too.

3.2.4 File minfragsize

The minfragsize file is used to change the default behavior during the creation of the DoF. By default, one explicit "fragment" is created if one end contains a minimal number of atoms. Creating a "fragment" means that the program includes the explicit rotational degree of freedom associated to the fragment (the bond connecting it to the rest of the molecule) on the list of axes eligible for mutations/cross-overs. However, performing a cross-over swapping methyl groups is basically a waste of time. If an axis is not explicitly used for genetic operators because the fragment it carries is not big enough, this does not imply that its geometry is fixed – it may well change, following "Lamarckian" gradient-directed moves. It simply means that the program will never focus to explicitly alter the geometry of that specific moiety (also see below, with respect to the –n option of S4MPLE). It is possible to specify distinct behaviors between entities, as for example for ligands and flexible binding sites. Parameters are defined according to this convention (one per line): "<keyword> <value>". The available keywords are:

- mfs: set the same threshold for both ligand and target
- mfs_lig: set the threshold for ligand
- mfs_ target: set the threshold for target

For example, it is possible to use a low value (mfs_lig 3) for ligand in order to explicitly sample most DoF (moieties larger than a methyl group) while performing a fuzzier flexible site docking by using a larger threshold (mfs_target 6).

3.2.5 File runparams

This configuration file allows the user to change several critical parameters of a launched evolutionary-based simulation *after* the simulation has been launched. This file is revisited at each iteration, and internal updates are made on the fly. Parameters are defined according to this convention (one per line): "<keyword> <value>". The available keywords are:

- ngen: change the number of desired generations
- runtime: change the current time limit

3.2.6 File active_pairs.lst

This file lists the atoms pairs to consider in the simulation. This file is usually created by the "prune" strategy of the evolutionary algorithm (see there), and can be useful in cases where a large binding site is defined with only specific flexible moieties. The latter strategy performs a preliminary simulation in order to prune some pairs from the whole non-bonded list (pairs involving atoms which are never close from each other are discarded). Using a pruning run before the actual simulation may significantly alleviate the pair list, but the user has to use it carefully.

3.2.7 File frozen_bonds

The file allows to constraint dihedral angles to their native values by listing involved atoms IDs as in the broken_bonds file (under testing at the moment).

3.2.8 File bonds

The final configuration file is only used in the context of the CAR format for the target. It lists explicitly the bonds, if they cannot be automatically detected (as in PDB file, there is no connectivity table for the usual protein residues in CAR file).

3.2.9 Chromosome I/O files

The term "conformer chromosome refers as a pose encoded as a linear string. This string contains all coordinates (which are converted to integers by multiplying by 1000, then rounding up) of the free atoms of the system. Fixed atom coordinates will always be equal to

the initial values read from data files. Note that in a simulation in which you'd wish to gradually unlock increasingly large parts of the molecule (or, on the contrary, to fix moieties) you cannot use chromosomes obtained with a different fixed_atoms list. You will need to edit the chromosome file, and insert the input file coordinates for the freshly unlocked atoms in the string (at the correct positions), or, on the contrary, delete the coordinates of freshly fixed atoms.

Chromosome files (see pop_in and pop_out options below) are generated during the sampling process, in order to save the so-far near-optimal visited geometries. The first column of these plain text files will contain the force field energy value, followed by the chromosome string as defined above. They are reread for post-processing purposes and conversion to full-blown output molecular files of the best sampled geometries, for visualization in your favorite graphical interface.

3.3 Available options in S4MPLE

The command-line help is detailed here.

3.3.1 Main options

-h: use this option to print the command-line help.

-i <directory>: use this option to set the working directory. The latter must contain all molecular and control files.

-f <directory>: use this option to specify the force field directory relative to a reference location specified by the environment variable $SETUP_DIR (see §2.1). Thus, use –f parm_ff to use the native Core Amber/GAff, and –f parm_ff_fit to use the solvation-extended version.

The minimization of the system is called with the option -m, while the genetic algorithm is launched with the option -e (see below).

3.3.2 Specific management of DoF

-n <binary value>: use this option to specify the amide mode.
Available values are:

- 0 = amide bonds are not considered as explicit DoF
- 1 = amide bonds are explicitly sampled

The default value is 0 (FALSE).

-q <value>: use this option to automatically unlock polar hydrogens whatever their status from fixed_atoms file.
Supported values are:

- 0 = do nothing
- 1 = unlock hydrogens from hydroxyl groups
- 2 = unlock all polar hydrogens

The default value is 0.

3.3.3 Energy minimization switches

-a <value>: use this option to set the convergence threshold for the gradient.
The default value is $1.0e^{-7}$.

-b <value>: use this option to set the required gain in energy to be achieved by a minimization step.
The default value is 0.10 kcal.

Within the full minimization procedure, a step not managing to lower energy by more than that counts as a "failure", and minimization undertakes a bond softening/retightening cycle to escape the local minimum.

-m <value>: use this option to specify the maximal number of allowed "failures" in the above sense, before stopping the full minimization procedure.

3.3.4 Sampling and post-processing

-e <colon-separated options string>: use this option to specify the task to perform (evolutionary algorithm, post-processing) and pass the corresponding control parameters (see detailed discussion below).

-j <value>: use this option to specify the energy window width with respect to the so-far best energy. The default value is +30.0 kcal/mol.

-s <binary value>: use this option to set the output status.
Available values are:

- 0 = only save chromosomes (DoF of the system encoded as linear strings)
- 1= save molecules as both molecular files (one per instance) and chromosome

The default value is 1 (TRUE).

3.3.5 Miscellaneous controls

-y <binary value>: use this option to use geometric centers of entities to help the placement in the binding site when docking.
Available values are:

- 0 = Off
- 1 = On

The default value is 0 (FALSE). This assumes that the receptor file (ref.car or ref.mol2, in which at least one atom is declared fixed) only includes the active site neighborhood – i.e. represents a "sphere" of residues centered on the active site cleft (typically, we use MOE to cut a sphere of residues of 10 A around the active site), in order to make sure that the geometric mean of the site atom coordinates actually pinpoints the active site. Turning the geometric center option on triggers the program to double-check that the initial pose not only features some randomly picked site-ligand favorable contact, but also that the ligand is well located into the site (sometimes, if hot spots at the outer edge of the binding site are defined, S4MPLE may generate initial poses in which the ligand interacts with such a hot spot, all while pending outside the active site cleft).

-r <value>: use this option to set the RMSD mode.

Available values are:

- 0 = do not compute any RMSD
- 1 = compute standard RMSD
- 2 = same as 1 but using a symmetry-compliant RMSD for ligands

The default value is 2.

-o <value>: use this option to automatically create the hot_spots file.

If the geometric centers mode is on (see -y switch), then all putative hot spots (hydrogen bond donors, hydrogen bond acceptors and carbons) around the geometric center of the site are saved (maximal number = value).

If the geometric centers mode is off, then only the putative hot spots around the loaded ligand are extracted (maximal number = no limit).

-z <filename>: use this option to specify a custom reference binary fingerprint from an external file.

By default, the binary fingerprint of the starting geometry is employed.

-t <binary value>: use this option to set the superimposition mode for single-entity sampling.

Available values are:

- 0 = Off
- 1 = On

The default value is 0 (FALSE).

-w <value>: use this option to set the water contact mode.

Available values are:

- 1 = standard case, the water contact mode is off.
- number > 1 = the contact strength between ligand and water molecules is scaled by the value.

By default, the water contact mode is off (FALSE) and the scaling value is 1.0.

-k <value>: use this option to set the force constant used in site constrained minimization.

The latter are performed with respect to their native coordinates.

3.3.6 Sampler Switches

This sub-chapter sums up the currently supported task control options with the -e flag. The options have to be concatenated as option1=value1:option2=value2:... using colons as separators. The key option is the strategy, to be passed as "strat=chosen strategy keyword".

Preliminary run

The prune strategy can be used to perform a preliminary simulation in order to prune some pairs from the whole non-bonded list (pairs involving atoms which are never close from each other are discarded). Using a pruning run before the actual simulation may significantly alleviate the pair list.

This will produce, locally, a list of non-bonded pairs that were found to be relevant. The file will be implicitly used by the following actual simulation, restricting the list of actually monitored NB pairs to a subset of these preselected pairs. The idea is that very far atoms may never directly meet, whereas they may indirectly interact by jointly exercising forces on intermediate geometric elements.

At the moment, this approach has not been extensively tested.

Running a GA-driven sampling

To actually perform a GA-driven sampling run, set option "strat" to one of the supported heuristics below (the default strategy is strat=evol):

- strat=evol: default evolutionary strategy,
- strat=base: basic GA,
- strat=elit: elitist GA,
- strat=tour: tournament-based GA,
- strat=hc: hill climbing GA.

The "evol", "base" and "hc" strategies have been extensively tested, the others less so. Note that evol requires many more generations, as it features a single individual which is modified by the operators per generation.

Further strategy-related options include:

- npop=<population size>
  This switch is used to set the population size. The default value is 50.

- ngen=<generation-number>
  This switch is used to set the number of generations. The default value is 500.
  Some few hundred will do for simple docking, but a full deployment on a grid will be needed for highly complex systems (work in progress).

- water=[true/false]
  This switch tries to perform an optimization of free waters around each kept pose using the GA during the post-processing stage. The default value is FALSE.

- minfpdiff=<value>
  This switch is used to set the minimal difference for interaction fingerprints (FP) of two non-redundant conformers (related to fingerprint size). The default value is 0.01.

- pop_out=<output filename>
  This switch is used to set the output filename in which conformer chromosomes are output.

- search=local
  Just add this option in the GA options string to force a sampling in the neighborhood of the input structure.

- seed=[true/false]

This switch is used to reset the random seed at the beginning of the simulation. Distinct simulations can be performed using this option. By default, there is no modification of the random seed.

- runtime=<value>
  This switch is used to set maximal allowed run time in hours for the simulation. The time limit is disabled (default behavior) when this parameter is equals to 0.

Post-processing stage

To post-process a brute conformer chromosome list, obtained from a previous simulation, use:

- strat=filter
  This strategy enables the post-processing stage. After reading the input conformer chromosomes file, poses are filtered according two criterions:
  - energy width,
  - redundancy using internal interaction fingerprints.

  All kept poses are directly saved or subjected to post-processing stage according to specified options.

- minfpdiff=<value>
  This switch is used to set the minimal difference for interaction fingerprints (FP) of two non-redundant conformers (related to fingerprint size). The default value is 0.01.

- pop_out=<output filename>
  This switch is used to set the output filename in which kept conformer chromosomes are output.

- pop_in=<input filename>
  This switch is used to set the input filename which contain the conformer chromosomes (poses as linear strings). This file is typically the output filename of the simulation.

- maxSaved=<value>

This switch is used to set maximal number of poses to save during the filtering stage. By default, this number is 200.

- optimize=<value>

  This switch is used to optimize the kept poses read from the input chromosomes file. It has the same behavior as with the -m option above, except that minimization is systematically applied to any conformer before saving. A value of 0 leads to the usual gradient-based procedure without softening/retightening cycles, while larger values enable the full minimization procedure (see -m option). The returned geometry will be confronted again with more stable ones, and if still not redundant it will be marked for saving.

- optoh=<value>

  This switch is used to enable the pre-optimization of hydroxyl groups.

  Available values are:

  - 0 = do not pre-optimize OH groups
  - 1 = pre-optimize hydroxyl groups from ligands
  - 2 = pre-optimize hydroxyl groups from full system

  The default value is 0.

- dark=<value>

  This switch is used to set the intrinsic bonus for darker fingerprints in kcal/mol per percent darkness (default=0) in order to preferentially select conformers with many realized interactions.

3.4 RMSD Calculation

S4MPLE can compute distinct RMSDs (Root-mean-square deviation). They are computed over all heavy atoms without superimposition (by default). Native coordinates are used as reference coordinates for RMSD calculation, but custom reference coordinates can be used using a specific field (<INIT>) in SDF/MOL molecular files (see §4.2.1)

Several RMSD are computed and stored in molecular files:

- All: standard RMSD over the full system

- Calpha: standard RMSD over C-alpha atoms of target
- Target: standard RMSD for the full target
- TargetFlex: standard RMSD over free atoms of target
- Lig_X: standard or symmetrical-compliant RMSD for the ligand X

4 Preparation of the molecular input files

This chapter is dedicated to the preparation of molecular input files, including organic compounds or biological entities.

4.1 Preparation of biomolecule files

These may be oligo-peptides or protein binding sites. They need to be prepared and saved as mol2 file, from their PDB entries with common graphical molecular modeling soft. Preparation includes reading the PDB file, fixing erroneous bond orders, assigning protonation status of acido-basic groups, checking whether force field parameter assignment is successful within your favorite modeling software (a necessary, but not always sufficient coherence test). All tests were performed with mol2 generated by the program MOE [7].

We recommend to cut out a sphere of residues within some 10 Å around the active site (in general – within the neighborhood of flexible moieties). Save this molecular subset to the .mol2 file. Make sure your modeling software does not try to overzealously fill in the empty valences at the cut points with hydrogens – that will artificially create $–NH_2$ and, perhaps, aldehyde groups in the final structure, confusing the AMBER FF parameterization process [which expects –NH-C-C(=O)- backbones].

When using the AMBER FF, the partial charges and atomic types are assigned, on the fly, from the customized topology file using topological indexes. Briefly, each atom from a given residue has an unique flag, and all usual residues have been processed in order to compute these reference flags.

The full list of allowed residue names with the available AMBER FF are:

- standard residues: ALA ARG ASN ASP CYS GLN GLU GLY HIS ILE LEU LYS MET PHE PRO SER THR TRP TYR VAL
- usual patches: ACE NHE NME AMN CXL
- water residue: HOH

- special histidine: HID HIE HIP
- special protonation states: ASZ CYX GLZ LYZ
- cysteine in disulfide bond: CYM
- N-terminal residues (all): NALA NARG NASN NASP NCYS NCYX NGLN NGLU NGLY NHID NHIE NHIP NILE NLEU NLYS NMET NPHE NPRO NSER NTHR NTRP NTYR NVAL
- C-terminal residues (all): CALA CARG CASN CASP CCYS CCYX CGLN CGLU CGLY CHID CHIE CHIP CILE CLEU CLYS CMET CPHE CPRO CSER CTHR CTRP CTYR CVAL

Any other residues, such as co-factors, have to be prepared as ligands. At the moment, nucleic bases were not investigated. It should be noted that standard names can be systematically used for the 20 common residues. During the initialization of the program, special residues are discovered from their common name (e.g. HIP from a HIS in the site file, CALA from ALA, CYM from CYS ...) and the relevant partial charges and atomic types are assigned.

4.2 Preparation of a ligand

There are no AMBER dictionary files for ligands: GAFF must be used to obtain specific parameters, and partial charges also need to be provided by the user. The following ligand preparation workflow, consisting of several steps, is used in docking simulations:

- computing Gasteiger partial charges [8] using ChemAxon libraries [9] (property <CHG>),
- adding GAFF atomic types using the Antechamber tool [4] (property <TYP>),
- using the Parmchk tool [3] to check whether there are missing parameters (*e.g.* bonds, angles or torsions). In that case, Parmchk tries to compute the missing parameters using empirical rules, and the latter are added to their respective FF parameters files,
- generating a single conformer using ChemAxon libraries (avoiding starting from the expected solution) and saving initial coordinates from heavy atoms in property <INIT>.

4.2.1 Computing the partial charges

At the beginning of the project, it has been decided to use Gasteiger charges for ligands. Although the mol/sdf format is a good way to encode ligand information (connectivity table, information about formal charges and stereochemistry ...), there are no specific location for the partial charges. A special field, named CHG, is used to store these pre-computed charges.

The provided JAVA program **PrepLigFull**, based on the ChemAxon API (version 5.7), is routinely used to compute and store the partial charges, as desired in the ligand input files. Besides, this tool can be employed to predict other properties, among other the major microspecie at a given pH.

The command-line manual from PrepLigFull is displayed below:

```
Unix Shell > java -jar PrepLigFull.jar

PrepLig : currently supported command-line switches:

I/O
 -f : Input filename
 -o : Output filename
 -c : standardization string or file

Microspecies
 -fix_protons : do not modify protonation state of compounds
 -pH : pH for microspecies calculation (default=7.4)
 -min_ms_pop : Minimal threshold for microspecies (1 = 100%)
 -explicit_microspecies : Enumerate microspecies explicitly (see min_ms_pop option)

Tautomers
 -explicit_tautomers : Enumerate tautomers explicitly

Conformers
```

```
-single : Only save the best conformer (lowest energy)
-maxconfs : Maximal number of conformers
-noH : Don't prehydrogenize the Conformer plugin
-tlim : Time limit for Conformers calculation (in seconds for each molecule)

Other
-save_coords : save initial coordinates in <INIT> property
-h : Print this help text

Gasteiger charges are stored in the <CHG> property

Several options need valid ChemAxon licenses !
```

The simplest way to do this first preparation step is:

```
Unix Shell > java -jar PrepLigFull.jar -f input.sdf -o input_chg.sdf -fix_protons
```

If you are interested about microspecies prediction at usual pH:

```
Unix Shell > java -jar PrepLigFull.jar -f input.sdf -o input_chg.sdf -pH 7.4
```

If you want to save initial coordinates for subsequent use (e.g. computing RMSD in redocking experiment, but starting from a non-native conformer):

```
Unix Shell > java -jar PrepLigFull.jar -f input.sdf -o input_chg.sdf -save_coords -fix_protons
```

All these commands will add the property CHG into the output files. Mono or multi-structures can be used as input file, and a new conformer is generated from the starting structure.

4.2.2 Assigning the AMBER/GAFF atomic types

This step involves two key tools, developed by GAFF authors, namely Antechamber [4] and Parmchk. Several scripts were developed to perform the remaining steps into one single command.

- for a single ligand:

```
Unix Shell > gaff_00_mono.sh <input sdf file>
```

- for batch processing of a given directory containing ligand molecular files:

```
Unix Shell > gaff_00_all input <input directory>
```

5 Tutorials

Two distinct tutorials are provided to allow a quick test of S4MPLE. All simulation involves at least two main steps:

1) sampling or docking itself,
2) post-processing of previously saved conformers using gradient-based optimizations.

The ligand docking run includes a preliminary relaxation of the X-ray conformation of the ligand within the binding pocket.

5.1 Conformational Sampling

In this example, the small ligand serotonin is briefly sampled.

Unix shell: > **./ scripts_tuto/sampling_ligands.sh tutorial_sampling parm_ff_fit**

The results are provided in the results.tar.gz files too.

All configuration files and options were previously described, respectively in §3.2 and §3.3.

5.2 Ligand docking

In this example, a small organic ligand is docked into its cognate binding site.

Unix shell: > **./scripts_tuto/docking_filtering_opt.sh tutorial_docking parm_ff_fit 1N1M**

The results are provided in the results.tar.gz files too.

All configuration files and options were previously described, respectively in §3.2 and §3.3.

6 Bibliography

1. Hagler, A. t.; E, H.; S, L., Energy functions for peptides and proteins: I, Derivation of a consistent force field including the hydrogen bond from amide crystals. *J. Am. Chem. Soc.* **1974,** *96*, 5319-5327.
2. Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; Debolt, S.; Ferguson, D.; Seibel, G.; Kollman, P., AMBER a package of computer-programs for applying molecular mechanics, normal-mode analysis, molecular-dynamics and free-energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications* **1995,** *91* (1-3), 1-41.
3. Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *Journal of Computational Chemistry* **2004,** *25* (9), 1157-1174.
4. Wang, J. M.; Wang, W.; Kollman, P. A.; Case, D. A., Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics & Modelling* **2006,** *25* (2), 247-260.
5. Hoffer, L.; Horvath, D., S4MPLE - Sampler For Multiple Protein-Ligand Entities: Simultaneous docking of several entities. *J Chem Inf Model* **2012,** *submitted*.
6. Hoffer, L.; Chira, C.; Marcou, G.; A., V.; Horvath, D., S4MPLE - Sampler For Multiple Protein-Ligand Entities: Methodology & Rigid-Site Docking Benchmarking. *J. Mol. Graph. Model.* **2012,** *submitted*.
7. MOE, Molecular Operating Environment. *http://www.chemcomp.com* **Chemical Computing Group Inc**.
8. Gasteiger, J.; Marsilli, M., A New Model for Calculating Atomic Charges in Molecules. *Tetrahedron Lett.* **1978**, 3181-3184.
9. ChemAxon Calculation of Partial Charge Distributions. http://www.chemaxon.com/marvin/help/calculations/charge.html (accessed Feb. 2009).

2.1.4 Champs de force et fonction d'énergie

La fonction d'énergie de S4MPLE repose sur le formalisme des FF préalablement décrit au §1.4.1. Les différents points abordés dans ce sous-chapitre correspondent à une description des FF disponibles, ainsi que des modifications qui y ont été apportées.

Les champs de force natifs

Deux FF distincts sont actuellement disponibles :

- CVFF (Consistent Valence Force Field) [172], disponible au sein du programme Discover [173]
- AMBER / GAFF (Assisted Model Building with Energy Refinement [58, 59] / General AMBER Force Field [174])

Le champ de force CVFF a été importé d'une implémentation antérieure du programme opérant uniquement dans un espace torsionnel [175, 176]. Le champ de force CVFF est actuellement assez peu utilisé dans la communauté scientifique au regard du nombre d'articles s'y référant. Cependant il a l'avantage d'être compatible à la fois avec les macromolécules biologiques et les petites molécules organiques, ce qui permet une modélisation de complexes ligand-récepteur.
Une recherche bibliographique a mis en évidence le développement relativement récent de GAFF qui est une extension de AMBER pour simuler les petites molécules [174]. L'utilisation conjointe de ces deux FF permet donc également de simuler ce genre de complexes.
Le premier développement notable réalisé fut d'ailleurs d'intégrer les paramètres d'AMBER / GAFF tout en gardant la possibilité de revenir au FF initial (CVFF) en cas de besoin. Etant donné que l'équation de calcul de l'énergie potentielle (E_{pot}) est très similaire, voire identique pour la plupart des termes, cet ajout a pu se faire avec des modifications mineures du code source et un préformatage identique des différents fichiers de paramètres. Ceux d'AMBER / GAFF sont disponibles gratuitement et sont récupérés sur un site internet dédié [177]. Il en est de même pour la suite logicielle AMBERTools [59] (version 1.3 sous licence GPL). Plusieurs composants d'AMBERTools, dont Antechamber [178] et Parmchk, servent à la préparation des ligands. Il est à noter que l'ensemble des simulations décrites dans le cadre cette thèse repose sur l'utilisation des champs de force AMBER / GAFF.

Les composantes énergétiques additionnelles

Des composantes énergétiques additionnelles [179], également importées de la version précédente de l'outil et d'une complexité similaire aux termes usuels d'un FF, sont intégrées à l'équation *in vacuo* de calcul de l'E_{pot} du système. Ces termes représentent un modèle simple de solvatation continu, ainsi

qu'une composante récompensant certaines interactions favorables. Les effets du solvant et l'enfouissement énergétiquement favorable de zones hydrophobes, absents du FF *in vacuo* initial, sont régulièrement montrés du doigt comme le talon d'Achille de la plupart des fonctions d'énergie / de score actuelles [15]. Par convention, le FF AMBER / GAFF natif est appelé "Core FF", tandis que le terme "Fit FF" fait référence à la version modifiée intégrant les termes additionnels. Ces composantes énergétiques additionnelles sont introduites ci-dessous et sont également décrites dans la partie "Matériel et Méthodes" de l'article I (inséré au niveau du §3.4 à la page 107 de ce manuscrit).

Le modèle de solvant implicite

L'utilisation d'un solvant explicite, modélisé par l'ajout de nombreuses molécules d'eau dans le système, est globalement non adapté avec un AG. En effet, le fait de déplacer une molécule d'eau d'un endroit de la boîte d'eau à un autre est essentiellement une perte de temps. Pour cette raison, un modèle de solvant implicite est préférentiellement utilisé avec un AG.

Un modèle de solvant implicite a notamment pour but d'approximer l'énergie associée au processus de solvatation d'un soluté chargé ; ce dernier étant modélisé par un diélectrique faible, et le solvant par un diélectrique fort [180]. Le modèle de solvant implicite inclus dans S4MPLE est constitué de deux termes de faible complexité : une constante diélectrique relative dépendante de la distance et un terme de désolvatation basé sur des paires d'atomes non liés [179] (voir ci-dessous).

La "constante diélectrique relative dépendante de la distance" $\varepsilon_r(d)$ correspond à la constante diélectrique relative qui est multipliée par la distance d entre les deux charges ponctuelles d'une paire. Ce procédé [181] permet de modéliser très simplement le phénomène d'écrantage électrostatique qui aboutit à une diminution des interactions électrostatiques lorsque des molécules d'eau ($\varepsilon_{r\_eau} \approx 80$) sont intercalées entre deux charges ponctuelles. De fait, le terme de Coulomb (voir l'Eq 9) devient fonction de la distance au carré ($1/d^2$).

En toute rigueur, la composante électrostatique du processus de solvatation devrait être estimée en utilisant une méthode se basant sur l'équation de Poisson. Mais sa complexité, tant des points de vue théorique que pratique (précision numérique des calculs), limite son utilisation. Ainsi, des approximations plus ou moins fortes sont réalisées pour rendre plus facile la modélisation de cette composante. Par exemple, les modèles de solvant continu reposent sur l'approximation que la composante électrostatique du processus de solvatation peut être réduite à une interaction entre la distribution de charges du soluté et un continuum diélectrique représentant le solvant [182].

Gilson et Honig ont proposé un terme de champ de force modélisant la composante électrostatique du processus de solvatation à travers le déplacement d'un diélectrique fort (par exemple, le solvant H_2O) par un diélectrique faible s'approchant (soluté) [183]. La variation d'énergie associée à ce phénomène, impliquant un calcul intégral de volume, s'écrit :

$$\Delta E_{déplacement} = \left(\frac{1}{\varepsilon_{int}} - \frac{1}{\varepsilon_{solv}}\right) \sum_i \frac{{q^2}_i}{32\pi^2} \int_{V_{int}} \frac{dV}{d_i^4} \qquad \text{Eq 33}$$

avec l'indice *i* qui concerne les atomes du système, q_i la charge partielle de l'atome *i* et ε_{int} / ε_{solv} les constantes diélectriques respectives des milieux considérés (intérieur du soluté et solvant). Ce formalisme est insuffisant pour estimer des énergies de transfert (par exemple vide→eau), mais est néanmoins suffisant pour approximer des différences d'énergie de solvatation à la suite d'un réarrangement conformationnel ou de la liaison d'un ligand à son récepteur [183]. Ils ont également proposé de remplacer le lourd calcul intégral du volume par une constante $\overline{V}$ représentant un volume moyen des atomes et une somme sur l'ensemble de ses atomes voisins *j* () [182, 183], d'où :

$$\Delta E_{déplacement} = \left(\frac{1}{\varepsilon_{int}} - \frac{1}{\varepsilon_{solv}}\right) \overline{V} \sum_i \frac{{q^2}_i}{32\pi^2} \sum_j \frac{1}{d_{ij}^4} \qquad \text{Eq 34}$$

La seule variable restante est la distance interatomique d_{ij}. Cette expression en $1/d^4$ et fonction des charges partielles atomiques a inspiré la définition arbitraire du terme non lié de désolvatation [179] :

$$E_{desolv_{ij}} = d?solv\_factor \frac{{q_i}^2 + {q_j}^2}{d_{ij}^4} \qquad \text{Eq 35}$$

Ce terme de désolvatation inclut une constante nommée "desolv_factor", nécessitant une paramétrisation, afin de lui donner un poids relatif par rapport aux autres composantes du FF. Il est à noter que cette composante de désolvatation est bien évidemment nulle dans le cas du Core FF. Cette expression du terme de désolvatation possède certaines propriétés intéressantes :

- lorsque deux atomes polaires sont à proximité (d_{ij} faible), ils ont dû subir au préalable une désolvatation qui a un coût énergétique ($E_{desolv_{ij}} > 0$)
- lorsqu'un atome polaire n'a aucun voisin à proximité, il est considéré comme solvaté et ne perçoit pas de pénalité particulière car toutes les distances d_{ij} sont élevées

Enfin, les atomes apolaires, porteurs de charges partielles faibles par définition, subissent une pénalité de désolvatation très faible avec ce formalisme. Cependant, ils sont majoritaires dans les (bio)molécules organiques. Afin de ne pas laisser la somme de nombreuses contributions peu

significatives peser trop fort par rapport aux rares contributions importantes, un seuil ajustable "minq_to_desolv" est finalement considéré pour éventuellement permettre d'ignorer ces premières. Des détails supplémentaires vis-à-vis de ce facteur correctif sont disponibles au §3.2.

La composante "contact favorable"

Le terme de contact, également basé sur des paires d'atomes, se focalise sur deux types d'interaction : les liaisons hydrogène et la proximité de carbones dans l'espace (C-C). Ce dernier point permet notamment de récompenser l'enfouissement de groupements hydrophobes dans des cavités hydrophobes, phénomène connu pour être très bénéfique d'un point de vue énergétique (voir le §1.2.4) et cependant assez peu récompensé par le terme attractif de Van der Waals par rapport au même enfouissement dans une cavité polaire. Les deux types d'interaction sont calculés selon le même formalisme :

$$E_{contact_{ij}} = k_{ij} C_{ij} \qquad \text{Eq 36}$$

La constante de force k_{ij} est fonction de la nature de l'interaction (LH, contact C-C). Il s'agit d'une constante (paramètre "hbond_bonus") dans le cas d'une LH, mais elle est fonction des types atomiques impliqués pour un contact C-C.

La variable C_{ij} représente la qualité de celle-ci (présente/absente/partielle) en fonction de la distance. Etant donné que seule la distance intervient dans le calcul de cette composante, il apparaît plus judicieux de définir une liaison hydrogène D-H..A par sa distance H-A plutôt que par sa distance D-A. Le paramètre C_{ij} vaut 0 lorsque la distance d_{ij} est supérieure à la limite fixée d_{max}, 1 lorsqu'elle est inférieure à un seuil minimal d_{min} et une valeur intermédiaire dans l'intervalle [d_{min} ; d_{max}]. Par conséquent, ce terme inclut une fonction d'amortissement ("switch") comme évoqué précédemment au §1.4.3. Comme pour le terme de désolvatation, cette composante additionnelle est bien évidemment nulle dans le cas de Core FF.

Tous les termes non liés sont donc maintenant des fonctions en d^{2n}, ce qui a l'avantage d'éviter de devoir systématiquement calculer la distance d à partir de d^2 sur les nombreuses paires non liées. De plus, l'obtention de d^{2n} avec n=(1,2,3,6) à partir de d^2 est également optimale d'un point de vue informatique *via* l'exponentiation rapide.

Après une calibration des différentes constantes au sein du champ de force CVFF modifié, les termes additionnels avaient amélioré la capacité à simuler le repliement *ab initio* d'oligopeptides (au moins un conformère proche de la structure native en tête de liste) [175, 184]. Le passage de CVFF à AMBER / GAFF change potentiellement la donne : une nouvelle calibration des différents paramètres s'avère nécessaire. De plus, il faut également vérifier la pertinence de ces termes additionnels après calibration par rapport au FF natif (Core FF). Le détail du protocole de calibration, à savoir ses différentes étapes et sa validation, sera présenté et discuté ultérieurement dans le §3.2.

2.1.5 Les principales caractéristiques de l'algorithme génétique

Les principales caractéristiques de l'AG implémenté dans S4MPLE sont décrites dans un article :

- la gestion des DDL du système
- les différents opérateurs génétiques crossing-over et mutation
- les principales stratégies d'évolution (sous la forme de pseudo-code)
- le mécanisme de contrôle de la diversité. Il est basé sur des empreintes d'interaction représentant de manière simplifiée une conformation donnée du système à travers l'identification des contacts favorables présents. L'acronyme utilisé pour identifier le fingerprint d'interaction est PIF (“Pairwise Interaction Fingerprint”)

Par conséquent, il recommandé à ce stade de lire les sous-chapitres §2.4 à §2.6 de la partie “Matériel et Méthodes” de l'article I [185] (voir le §3.4 à la page 107 de ce manuscrit).

Il est à noter que les mêmes opérateurs génétiques sont capables de traiter les DDL intra- et intermoléculaires de manière totalement transparente et unifiée, ce qui ouvre la voie à des simulations avec un nombre N variable d'entités qui est fixé initialement :

- N=1 → EC classique
- N=2 → docking d'une entité
- N=3 → docking de deux entités
- N>3 → docking de plus de deux entités

Cette faculté concernant l'EC de plusieurs entités différentes est peu fréquente en docking [186]. Bien qu'un nombre plus élevé est techniquement possible, “seul” un docking limité à deux ligands réels (hors molécules d'eau libres) a été réalisé lors de cette thèse (voir le §4.2).

2.2 Présentation des outils du protocole d'évolution

Deux programmes GenLinkersDB et JMolEvolve, basés sur l'API JAVA de ChemAxon [187], ont été développés dans le cadre de cette thèse pour répondre à la problématique de création d'une chimiothèque focalisée sur un ou deux fragment(s) de référence. Ils sont décrits ci-dessous.

2.2.1 L'outil GenLinkersDB

Généralités

Le programme GenLinkersDB génère des banques de "linkers" en fragmentant une librairie classique de molécules, comme par exemple la ZINC [188], selon les règles RECAP (Retrosynthetic Combinatorial Analysis Procedure) [163] de rétro-synthèse virtuelle. L'algorithme RECAP identifie certaines liaisons chimiques particulières, les clive, et ajoute une étiquette ("flag") relative à la nature de la liaison chimique à chaque nouvelle extrémité (voir la Figure 33). L'hypothèse fondamentale selon laquelle des étiquettes complémentaires peuvent être reconnectées à l'aide d'une liaison covalente est émise. Etant donné qu'un objet de l'API ChemAxon reprend cet algorithme, ce premier outil s'avère finalement relativement simple au niveau de son implémentation.

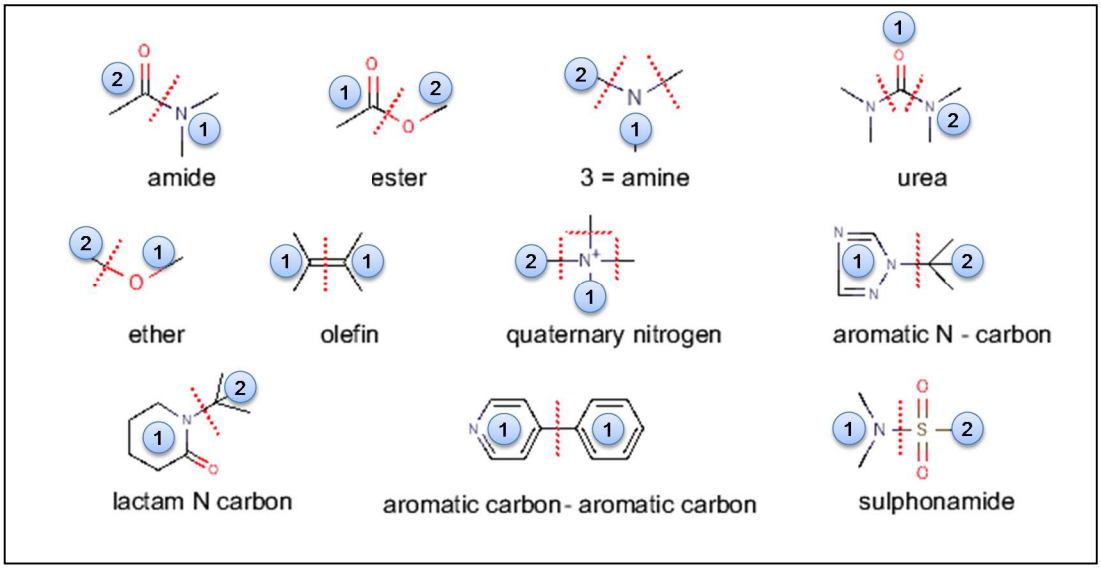


Figure 33: Dictionnaire des liaisons chimiques recherchées par l'approche RECAP usuelle, et illustration des étiquettes obtenues après fragmentation.

Par convention et pour éviter toute confusion avec les fragments du FBDD définis au §1.3, les groupements obtenus après fragmentation selon le principe RECAP seront appelés "linkers". Par souci de simplification, ce dernier terme sera employé quel que soit le type de simulation : optimisation par

croissance (“growing”) ou liaison (“linking”). La Figure 34 montre l'exemple de la fragmentation exhaustive d'une molécule avec l'algorithme RECAP tel qu'il est implémenté dans l'API ChemAxon (voir l'Annexe 1 pour le fichier de configuration incluant toutes les options utilisées).

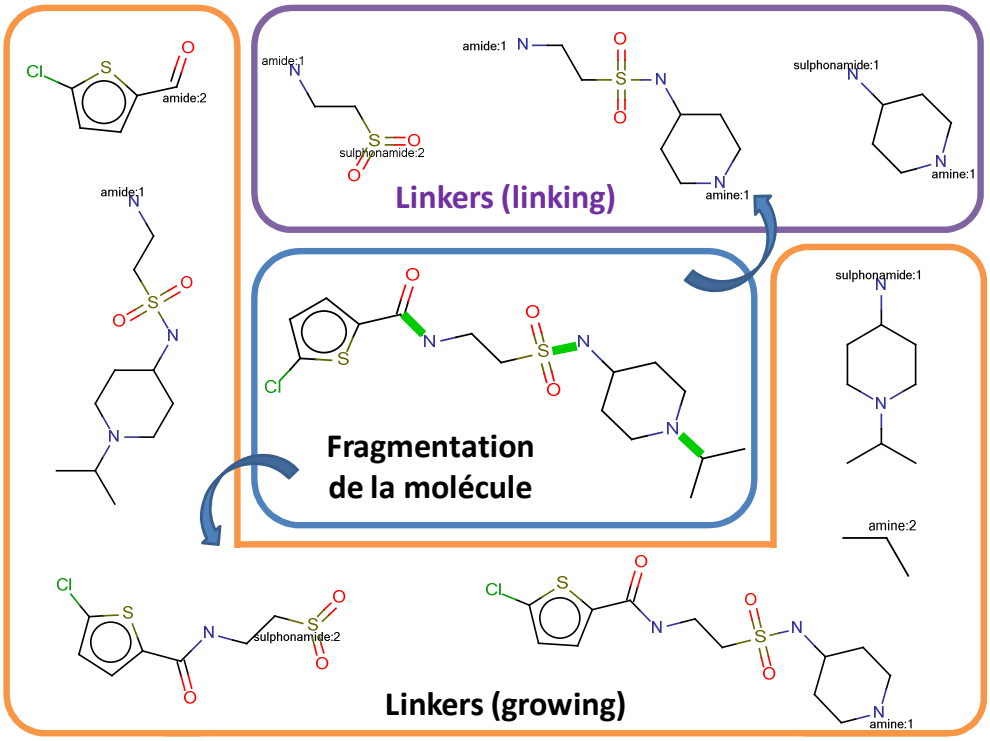


Figure 34: Exemple de la fragmentation exhaustive d'une molécule.
Les traits en vert représentent les sites de coupure identifiés selon le dictionnaire RECAP prédéfini.

Ce procédé illustré à la Figure 34 peut évidemment être généralisé à une chimiothèque de molécules, et le résultat de la fragmentation est une banque de linkers. Dans notre contexte d'étude, seuls les linkers à une (growing) ou deux (linking) étiquettes sont conservés ; les autres étant tout simplement éliminés. Une fois la chimiothèque de linkers générée, il est possible de [163] :

1) créer des chimiothèques combinatoires de molécules en reconnectant de manière exhaustive tous les linkers compatibles entre eux. Par exemple, la concaténation de deux linkers issus de la fragmentation précédente reconduit à la molécule originale (voir la Figure 35)
2) construire des chimiothèques focalisées (procédé détaillé ultérieurement au §2.2.2) sur une ou plusieurs référence(s). C'est cette stratégie qui sera utilisée ultérieurement dans la première étape du protocole FBDD *in silico*, en prenant comme référence un ou deux fragment(s) connu(s) pour se lier au site de liaison

3) analyser les linkers obtenus de manière à identifier les châssis ou groupements les plus fréquents au sein d'une banque de molécules

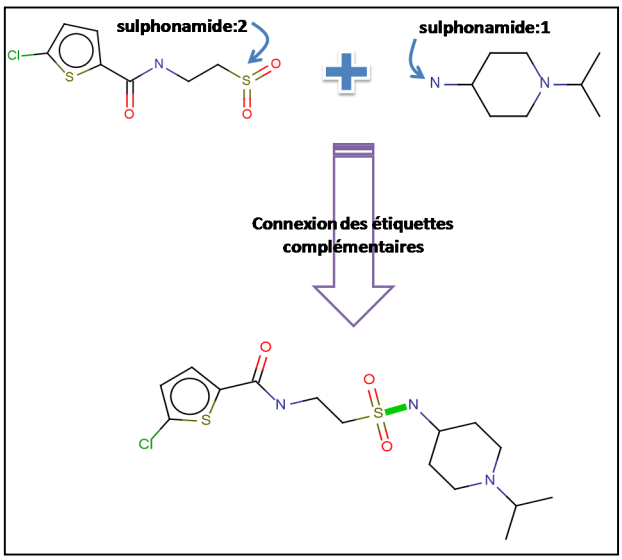


Figure 35: Exemple de création d'un nouveau composé à partir de linkers compatibles (ces derniers possèdent des étiquettes complémentaires).

Les différentes étapes de construction des banques de linkers sont illustrées à la Figure 36, et ne sont bien évidemment réalisées qu'une seule fois. La version dite "clean" de la ZINC a été utilisée comme source de molécules pour le processus de fragmentation. Ce sous-ensemble d'environ 9,5 millions de composés "drug-like" a été filtré de manière à éliminer les molécules contenant des groupements indésirables ou réactifs. Une limite à 300 Da pour les linkers est arbitrairement choisie puisque l'objectif final est d'optimiser un fragment tout en restant globalement dans l'espace chimique des Ro5.

Il est à noter que cette technique de chémoinformatique n'est pas parfaite, dans le sens où elle ne garantit pas la possibilité de synthèse d'une molécule. En effet, bien qu'elle ait l'avantage d'être intuitive, rapide et génératrice de molécules raisonnables (si la banque initialement fragmentée est chimiquement cohérente, par exemple *via* l'absence de groupements indésirables), cette approche ne tient pas compte des voies de synthèse organique, de l'ajout ou non de groupement protecteur et des synthons directement ou commercialement disponibles. De plus, le contexte des atomes n'est pas non plus pris en compte : une amine aliphatique ou une autre liée à un cycle aromatique n'ont pas le même comportement, mais elles sont néanmoins considérées sous la même étiquette. Cette stratégie RECAP a été néanmoins poursuivie à la fois pour des contraintes de temps et sachant que les points principaux de ce travail sont l'EC et le docking de composés.

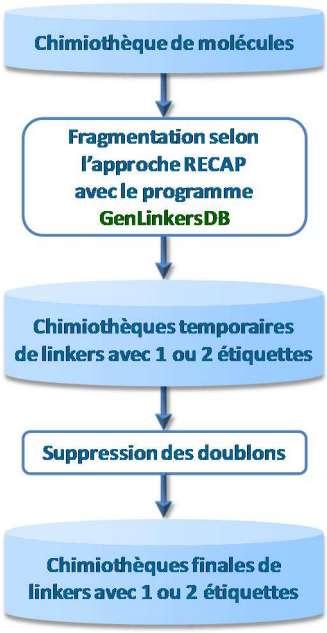


Figure 36: Processus de création des chimiothèques de linkers à partir d'une source de molécules.

Suppression des linkers doublons

La fragmentation exhaustive d'une grande banque de molécules diverses est nécessaire pour pouvoir couvrir une part non négligeable de l'espace chimique des linkers (graphes moléculaires et étiquettes). Naturellement, cette fragmentation aboutit à une énorme chimiothèque temporaire où il y a de nombreux doublons. Contrairement au cas usuel, un doublon ne se résume pas ici à un même graphe moléculaire. En effet, deux graphes moléculaires identiques mais possédant des étiquettes différentes, ou des étiquettes identiques mais placées sur un autre nœud (atome) doivent être considérés comme différents. La Figure 37 résume de manière schématique les différents cas de figure rencontrés lors de l'élimination des linkers redondants. Par exemple, le graphe moléculaire du linker 5 est unique, il est donc conservé. Bien que les linkers 1, 2, 3, 4 et 6 aient le même graphe, ils ne sont pas tous des doublons. Les linkers 1 et 3 ont la même étiquette mais celle-ci n'est pas située sur le même nœud du graphe, ils sont donc également conservés. Il en est de même pour les linkers 2 et 6. En revanche, les linkers 1 et 4 ont en commun leur graphe moléculaire, leur étiquette et le nœud la portant : ils sont donc identiques, et seul l'un des deux est conservé dans la banque finale.

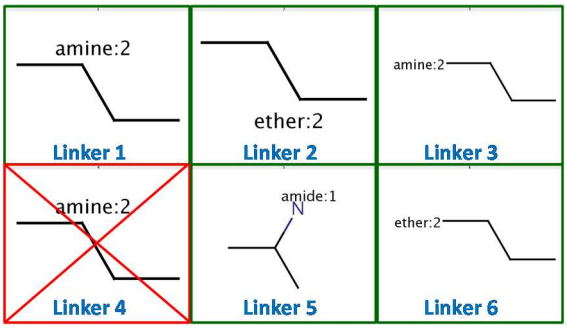


Figure 37: Exemple de linkers qui sont considérés comme uniques (vert) ou redondants (rouge).

Une première tentative d'identification des doublons, reposant sur des SMILES canoniques et leur comparaison en tant que chaînes de caractères, n'a pas été concluante : certains graphes moléculaires identiques avaient néanmoins des SMILES canoniques différents. Cette approche intuitive d'identification des doublons a donc été rapidement abandonnée au profit d'une méthode basée sur la recherche d'isomorphismes de graphes moléculaires (voir la Figure 38). Une implémentation de ce type d'algorithme est d'ailleurs disponible dans l'API ChemAxon. Pour pouvoir utiliser directement ces fonctions dans le contexte de l'étude, une astuce a consisté à ajouter un élément chimique absent de la banque au niveau de l'étiquette des linkers. Dans le cas d'une banque de linkers avec deux étiquettes, deux éléments chimiques différents sont utilisés, et l'étiquette la plus petite (au sens d'une comparaison de chaînes de caractères) est systématiquement associée au même élément. Dès lors, il devient possible d'utiliser tels quels les algorithmes d'isomorphisme de graphes, et ce en dernier ressort lorsque les étiquettes sont identiques et les SMILES canoniques différents entre le linker courant à tester et un linker préalablement identifié comme unique. Par exemple, le traitement du sous-ensemble incluant les 6 linkers de la Figure 37 retourne le même résultat, à savoir 5 linkers conservés sur les 6. Cependant, cette méthode rigoureuse aboutit rapidement à une impasse avec l'augmentation du nombre de molécules à comparer. Pour pallier ce problème, la gigantesque chimiothèque temporaire (plusieurs dizaines de millions de linkers), issue du processus de fragmentation de la ZINC, est automatiquement scindée en sous-chimiothèques selon le critère masse moléculaire avec un pas de 1 Dalton. L'idée sous-jacente est bien entendu que deux molécules de masses différentes ne peuvent pas être identiques. Ce procédé très simple a permis d'utiliser l'approche rigoureuse décrite ci-dessus. De plus, le fait de morceler les chimiothèques a l'avantage notable de pouvoir construire une banque finale dont les éléments sont classés par masse moléculaire croissante. Ainsi, l'arrêt du processus de création d'une chimiothèque focalisée sur une référence n'a aucune incidence sur l'espace chimique de complexité plus faible. Dans le cas général, à savoir une banque non triée de linkers, l'arrêt du traitement ferait en sorte que des linkers de plus faible complexité seraient potentiellement manqués.

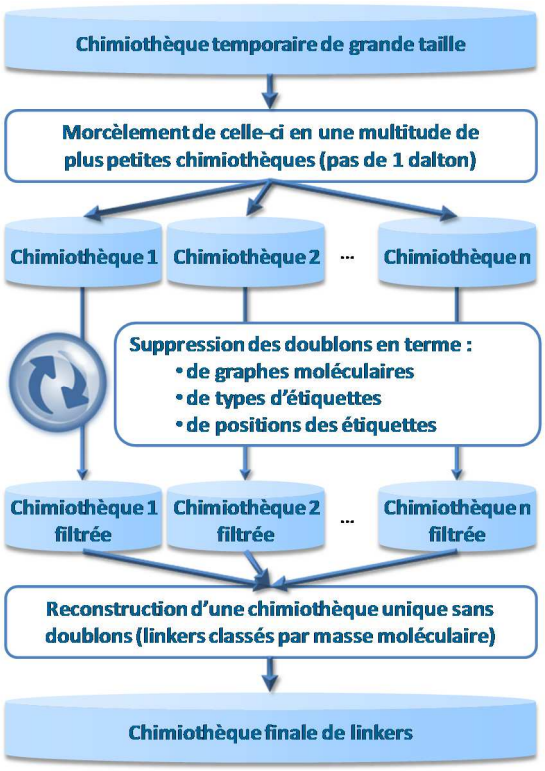


Figure 38: Illustration de la stratégie employée pour supprimer les linkers en double.

Les banques de linkers

Différents types de banques sont créés. Un aperçu des deux grandes classes (linking et growing) est donné à la

Figure 39, et le Tableau 3 recense le nombre d'entités uniques pour plusieurs exemples de banques qui ont été générées.

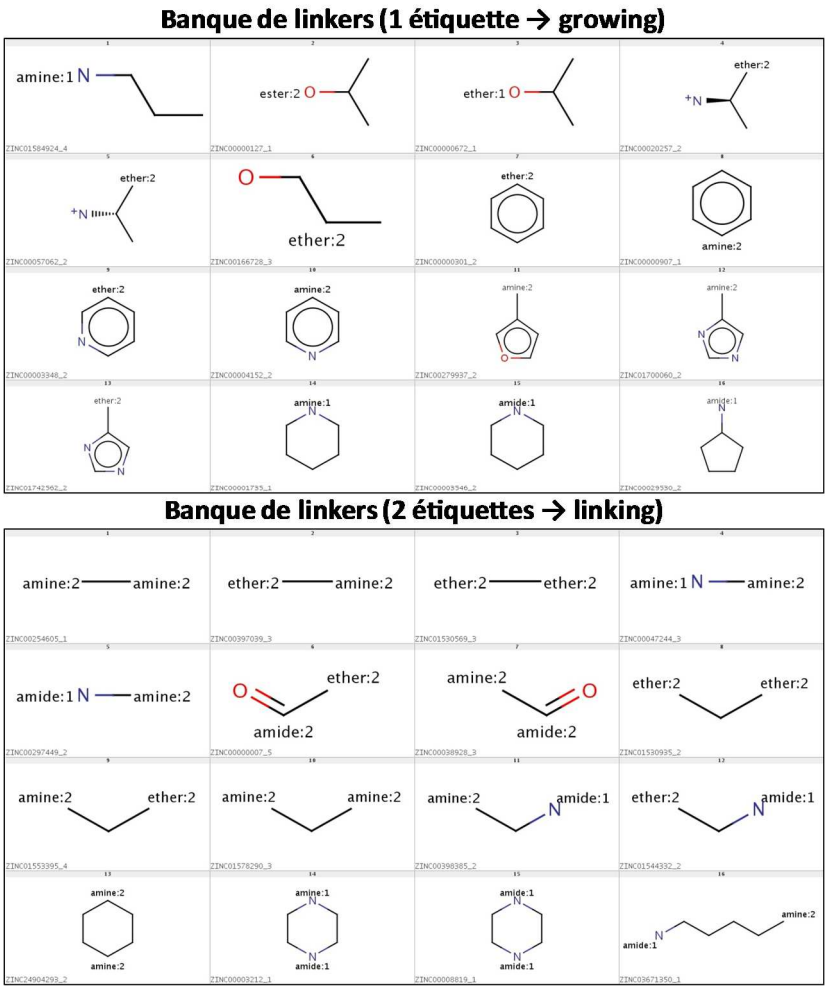


Figure 39: Aperçu à l'aide de MarvinView [187] d'un sous-ensemble d'une banque de linker.

Banque	Nombre de linkers en milliers (K)
Growing | Masse < 300 Da	3611 K
Growing | Masse < 300 Da | Eléments autorisés [HCNOSP]	2951 K
Linking | Masse < 300 Da	2164 K
Linking | Masse < 300 Da | Eléments autorisés [HCNOSP]	1898 K
Growing | Nombre d'atomes lourds [1-6] | Nombre de cycle [0]	2,3 K
Growing | Masse < 300 Da | Occurrence min du linker [100]	40 K
Linking | Masse < 300 Da | Occurrence min du linker [100]	26 K

Tableau 3: Nombre d'entités pour un sous-ensemble des banques de linkers.

Le filtre sur les éléments (“éléments autorisés [HCNOSP]”) permet essentiellement de ne pas tenir compte des linkers porteurs d'halogènes. Ceci permet par exemple de favoriser une certaine diversité au regard du nombre considéré de molécules. Par exemple, seul le linker “phényl” sera considéré au lieu de tous ses dérivés (F/Cl/Br/I en ortho/méta/para) potentiellement présents dans la banque généraliste. Cet effet est encore amplifié avec une banque composée uniquement de linkers populaires / fréquents : seuls les linkers dont le nombre d'occurrences est supérieur à un seuil prédéfini sont inclus dans celle-ci. Une alternative est de partir d'une version de la ZINC préalablement traitée avec l'algorithme de Murcko [189] : la chimiothèque initiale à fragmenter n'est constituée que de châssis moléculaires épurés. Dans ce dernier cas, une grande variété de châssis est obtenue avec une banque de linkers relativement petite. Cette dernière peut par exemple être utilisée en première approche afin de trouver une structure cyclique particulièrement adaptée à une cavité donnée. Par la suite, les versions “décorées” des cycles mis en évidence seront finalement extraites et utilisées pour générer des analogues. A l'inverse, une banque de petits linkers sans cycle à une étiquette est plutôt adaptée pour générer des analogues d'un composé donné dans une optique d'optimisation plus fine d'un lead déjà connu. Toutes ces dernières banques (fonction du nombre d'occurrences, sans cycles ou “murcko-isées”) n'ont toutefois pas été utilisées dans le cadre de ce travail puisqu'elles ont été générées après la fin de la quasi-totalité des simulations.

2.2.2 L'outil JMolEvolve

Généralités

Le programme JMolEvolve construit de nouveaux composés en connectant les linkers compatibles, provenant d'une chimiothèque, avec le ou les fragment(s) de référence (voir la Figure 40).

Les deux types usuels d'optimisation de fragments dans le cadre du FBDD expérimental sont modélisés ici : une stratégie de type growing est utilisée lorsqu'un fragment est donné en entrée, tandis qu'une stratégie linking est employée lorsque deux fragments sont soumis par l'utilisateur. Dans un but de simplification, l'acronyme G&L sera utilisé pour décrire ces deux stratégies d'optimisation à la fois.

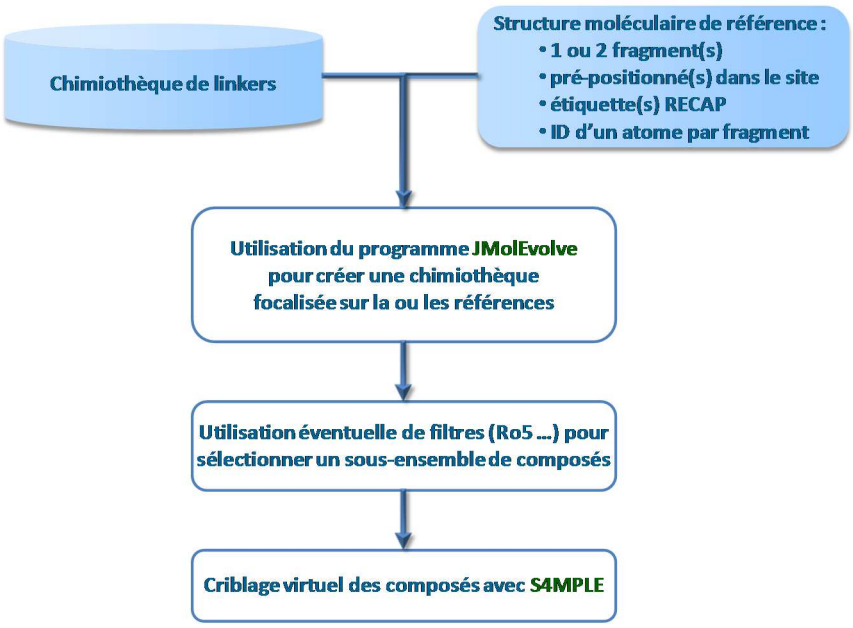


Figure 40: Principe général du programme JMolEvolve et perspective d'utilisation.

Le processus aboutit *in fine* à la création d'une chimiothèque focalisée. Les structures chimères "fragment-linker" ou "fragment1-linker-fragment2" (selon le type d'évolution) sont exportées dans le format SDF qui est très populaire pour représenter des petites molécules organiques. Bien que conçu pour ce type de molécules, le format SDF n'intègre pas la possibilité de stocker de manière native des charges partielles (seulement des charges formelles), d'où l'utilisation de champs/propriétés dédiés. Il est à noter que les charges partielles sont systématiquement traitées de cette manière (hors contexte du G&L) lors de la préparation des ligands en vue d'une simulation avec S4MPLE. Il apparaît donc pertinent de traiter les propriétés additionnelles du contexte G&L de cette manière.
Pour chaque composé chimère nouvellement créé, JMolEvolve :

- initialise un certain nombre de propriétés, notamment
  - <CHG> qui stocke les charges partielles atomiques de type Gasteiger [190] (plug-in ChemAxon) de la molécule finale
  - <INIT> qui contient les coordonnées initiales du fragment, avant la génération d'un conformère dans le but d'obtenir une structure 3D réaliste en sortie
  - <LINK> qui permet de différencier la partie fragment de celle du linker au sein de la molécule nouvellement formée
- identifie la micro-espèce majoritaire à pH physiologique (plug-in ChemAxon, si cela est souhaité par l'utilisateur). Cette option est fondamentale car la liaison de deux fragments peut aboutir à un état de protonation inadéquat (voir la Figure 41)

- génère un conformère de basse énergie de la molécule (plug-in ChemAxon)
- calcule plusieurs propriétés physico-chimiques et topologiques (plug-in ChemAxon)

Les informations stockées dans les propriétés <LINK>, <INIT> et <CHG> seront lues par S4MPLE lors de la phase d'initialisation de la simulation au même titre que les paramètres du FF.
Il est à noter que des filtres usuels (Ro5, nombre maximal de liaisons à rotation libre, *etc.*) peuvent être utilisés en sortie pour diminuer la taille de la chimiothèque à cribler virtuellement.

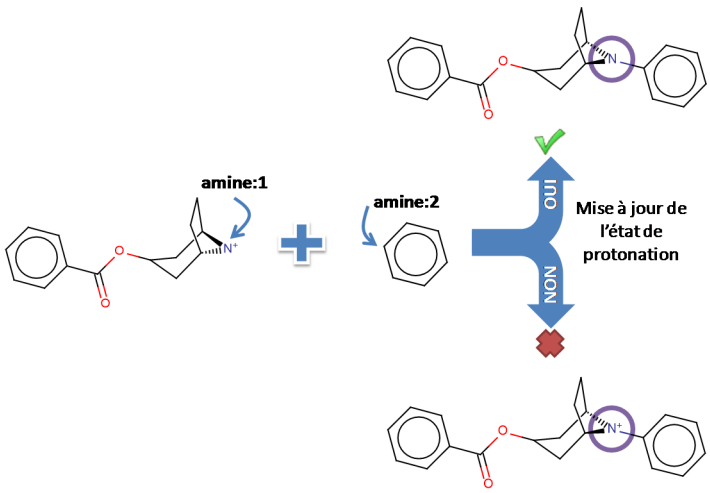


Figure 41: Intérêt de l'option de mise à jour de l'état de protonation des molécules générées.

Architecture de JMolEvolve

Le programme JMolEvolve a nécessité un plus grand développement (voir le Tableau 4), notamment pour pouvoir gérer les stratégies G&L au sein d'un même programme et de manière totalement transparente pour l'utilisateur.

paquetage jme	paquetage jme.gui
JMolEvolve	**JMEFrame**
MolecEvolve	JMEPanelGenDB
MolecGrowing	JMEPanelEvol
MolecLinking	JMEPanelInfo
RecapCompl	JConsole
GenLinkersDB	JConsolePanel

Tableau 4: Liste des différents paquetages et classes de JMolEvolve.
Les classes en gras sont exécutables (shell ou mode graphique) et celle en italique est abstraite.

Seules les classes non abordées jusqu'alors du paquetage *jme* seront décrites ci-dessous ; celles du paquetage *jme.gui* n'étant que la surcouche graphique basée sur l'API Swing du langage de programmation JAVA. A ce propos, l'interface graphique inclut également la possibilité de piloter le programme GenLinkersDB décrit préalablement, d'où son inclusion dans le paquetage *jme*. Le programme JMolEvolve repose fortement sur les concepts de PPO d'héritage (voir la Figure 42) et de polymorphisme.

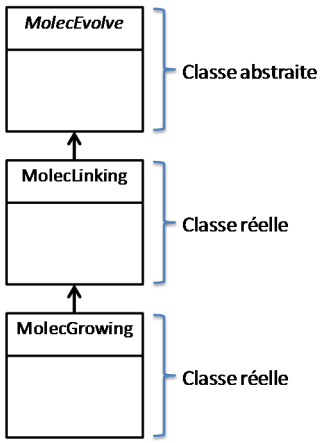


Figure 42: Hiérarchisation des classes MolecEvolve, MolecLinking et MolecGrowing. Les flèches indiquent une relation d'héritage.

La classe abstraite *MolecEvolve* n'introduit qu'un comportement général minimal puisqu'elle ne contient que les signatures des différentes méthodes clés. A l'opposé, les classes *MolecLinking* et *MolecGrowing* sont réelles et sont les descendantes de *MolecEvolve* : elles doivent contenir une implémentation des méthodes introduites dans la classe abstraite *MolecEvolve*. Dans ce contexte, un objet *MolecLinking* contient principalement :

- une molécule sous forme d'objet *chemaxon.struc.Molecule*
- les diverses propriétés relatives à l'étiquette d'attache et à l'atome impliqué
- les méthodes (incluant les accesseurs et mutateurs) nécessaires pour initialiser, modifier, valider et récupérer les propriétés

Un objet *MolecLinking* fait référence à deux atomes et deux étiquettes. Ce choix de conception permet de définir la classe *MolecGrowing* comme une sous-classe de *MolecLinking*, où seuls les éléments liés à la première étiquette sont considérés. Par conséquent, il y a une ré-implémentation minimale de la plupart des fonctions.

La classe *JMolEvolve* est la classe principale puisqu'exécutable du paquetage *jme*. Elle permet de réaliser le processus de création d'une chimiothèque focalisée en manipulant les différents types

d'objets préalablement décrits dans ce sous-chapitre, *via* la gestion des différents arguments et les entrées/sorties (structures initiales, banque de linkers, composés nouvellement crées). Lors de ce processus, toutes les entités moléculaires sont manipulées sous la forme d'objets *MolecEvolve* (la classe abstraite), mais elles sont instanciées à l'aide de constructeurs différents selon le type d'évolution (voir la Figure 43) :

- mode growing - toutes les entités moléculaires (fragment initial et linkers) sont réellement des objets de type *MolecGrowing*
- mode linking - les fragments initiaux restent de type *MolecGrowing*, tandis que les linkers provenant de la banque éponyme sont des objets instanciés à partir de la classe *MolecLinking*

Ces spécificités sont automatiquement réalisées dès l'initialisation du programme, puis grâce au polymorphisme, les étapes suivantes du processus deviennent absolument identiques pour le G&L.

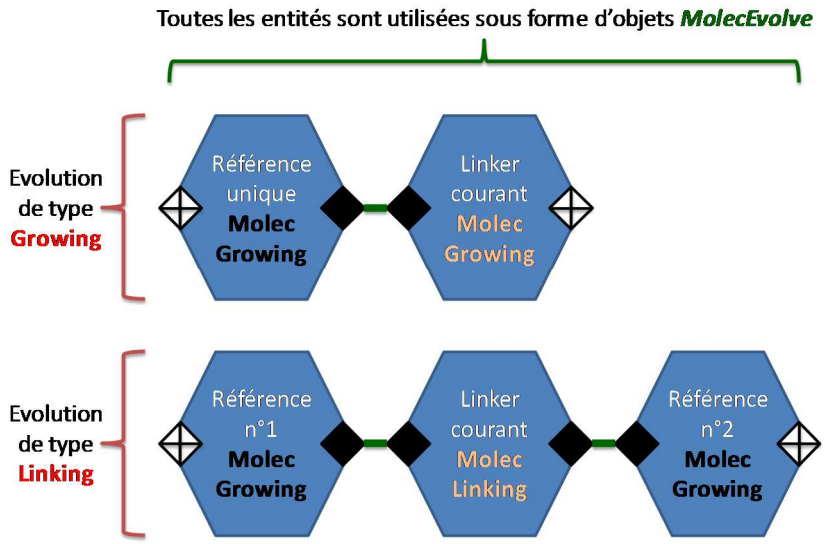


Figure 43: Illustration des différents objets en mémoire en fonction du mode d'évolution.

Enfin, un objet instancié à partir de la classe *RecapCompl* a pour but principal, à partir d'un fichier de configuration ("flags.txt"), de retourner l'étiquette complémentaire de celle donnée en paramètre (par exemple celle du fragment de référence). Une fois cette étiquette connue, il est facile de n'extraire que les linkers compatible lors du parcours de la banque de linkers.

2.2.3 Guide rapide d'utilisation de GenLinkersDB et JMolEvolve

Les deux outils peuvent s'utiliser tant en ligne de commande qu'en mode graphique à partir d'un unique fichier JAR (JMolEvolve.jar). L'avantage du premier mode est évident pour pouvoir

automatiser des processus, tandis que le second est plus convivial pour l'utilisateur, et ce notamment dans la gestion des différents arguments. De plus, le mode graphique permet d'éditer directement la structure chargée afin de respecter les conventions des différents sites RECAP de coupure.

GenLinkersDB

Les arguments du programme GenLinkersDB sont les suivants :

argument 1	banque de molécules à fragmenter
argument 2	préfixe pour le nom des banques de linkers

Tableau 5: Liste des arguments du programme GenLinkersDB.

Le fichier de configuration du fragmenteur RECAP de ChemAxon (voir l'Annexe 1), nommé "recap.xml", doit se situer dans le répertoire pointé par la variable d'environnement JMEDIR.

JMolEvolve

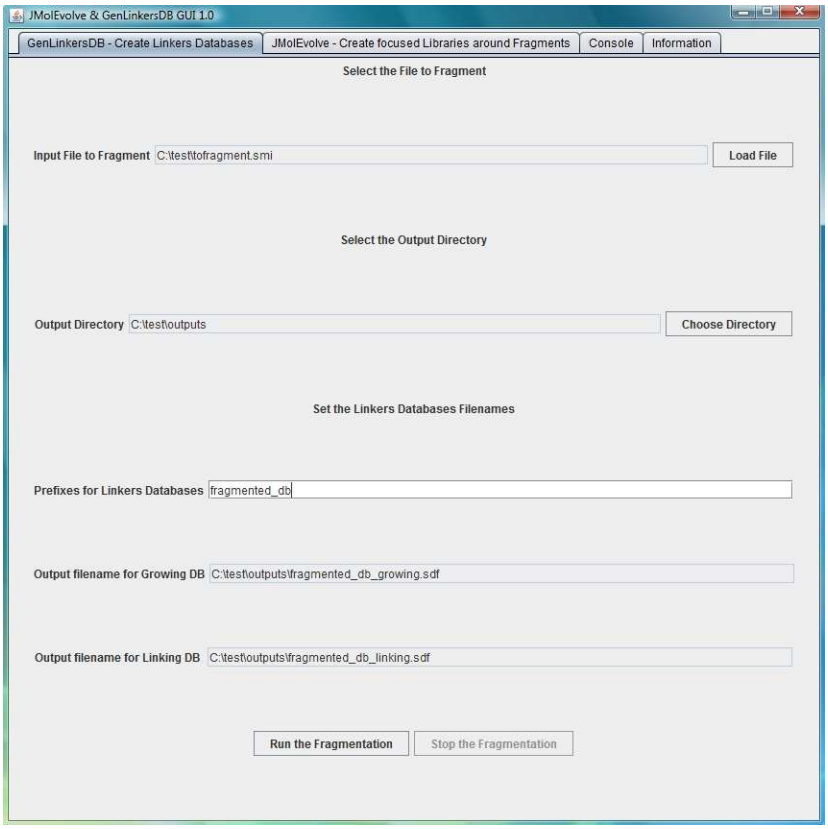
Les arguments du programme JMolEvolve sont les suivants :

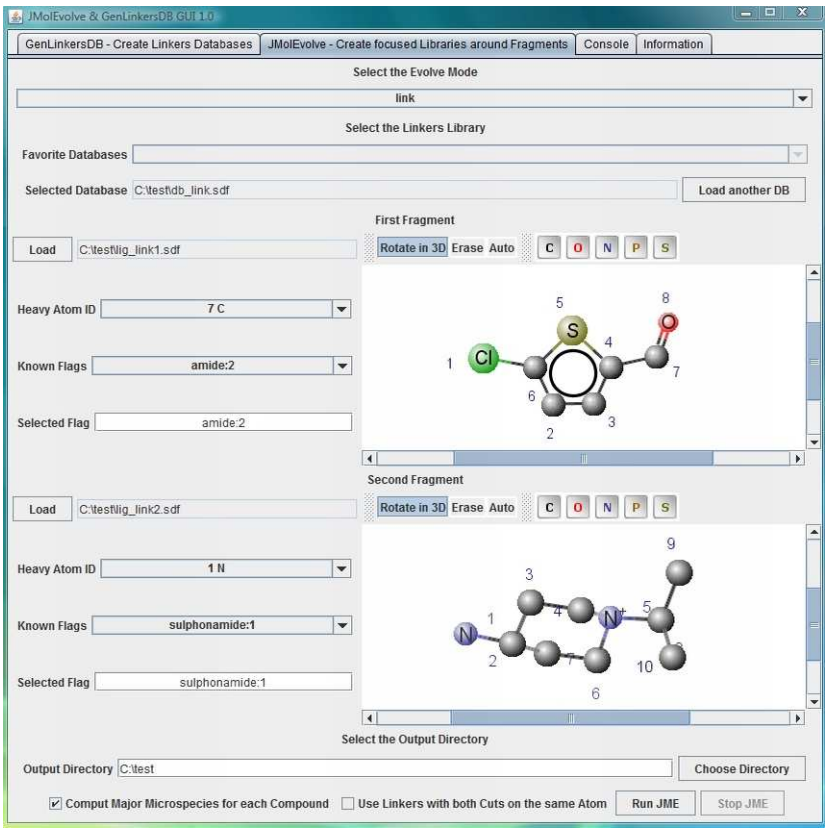
Growing + Linking	arg. 1	répertoire de travail pour la sauvegarde de la chimiothèque focalisée
	arg.2	fichier correspondant à une banque de linkers (de type growing ou linking)
	arg.3	booléen (true/false) pour la prédiction ou non de l'espèce majoritaire à la suite de la fusion du précurseur et d'un linker compatible
	arg.4	fichier moléculaire de référence (fragment/précurseur I)
	arg.5	ID de l'atome qui subit le processus d'évolution (fragment/précurseur I)
	arg.6	étiquette RECAP correspondante (fragment/précurseur I)
Linking	arg.7	fichier moléculaire de référence (fragment/précurseur II)
	arg.8	ID de l'atome qui subit le processus d'évolution (fragment/précurseur II)
	arg.9	étiquette RECAP correspondante (fragment/précurseur II)
	arg.10	booléen (true/false) pour spécifier ou non l'incorporation de linkers particuliers

Tableau 6: Liste des arguments du programme JMolEvolve.
Le terme de "linkers particuliers" fait référence à ceux où les deux étiquettes se trouvent sur le même atome (ils peuvent ne pas être désirables en pratique).

Le fichier de configuration fourni "flags.txt", contenant la liste des étiquettes usuelles et leur membre complémentaire, doit également se situer dans le répertoire pointé par la variable d'environnement JMEDIR. Un fichier de configuration facultatif ("db.txt") localisé dans ce même répertoire de référence, comprenant les chemins absolus des banques de linkers favorites, peut être utilisé avec la partie graphique afin de pouvoir sélectionner rapidement l'une de celles-ci.

Enfin, des captures d'écran de chaque onglet de la surcouche graphique de JMolEvolve sont insérées dans le Tableau 7.


JMolEvolve & GenLinkersDB GUI 1.0
GenLinkersDB - Create Linkers Databases
JMolEvolve - Create focused Libraries around Fragments
Console
Information
Select the File to Fragment
Input File to Fragment
C:\test\tofragment.smi
Load File
Select the Output Directory
Output Directory
C:\test\outputs
Choose Directory
Set the Linkers Databases Filenames
Prefixes for Linkers Databases
fragmented_db
Output filename for Growing DB
C:\test\outputs\fragmented_db_growing.sdf
Output filename for Linking DB
C:\test\outputs\fragmented_db_linking.sdf
Run the Fragmentation
Stop the Fragmentation


JMolEvolve & GenLinkersDB GUI 1.0
GenLinkersDB - Create Linkers Databases
JMolEvolve - Create focused Libraries around Fragments
Console
Information
Select the Evolve Mode
link
Select the Linkers Library
Favorite Databases
Selected Database
C:\test\db_link.sdf
Load another DB
First Fragment
Load
C:\test\lig_link1.sdf
Rotate in 3D
Erase
Auto
C
O
N
P
S
Heavy Atom ID
7 C
Known Flags
amide:2
Selected Flag
amide:2
Second Fragment
Load
C:\test\lig_link2.sdf
Heavy Atom ID
1 N
Known Flags
sulphonamide:1
Selected Flag
sulphonamide:1
Select the Output Directory
Output Directory
C:\test
Choose Directory
Comput Major Microspecies for each Compound
Use Linkers with both Cuts on the same Atom
Run JME
Stop JME

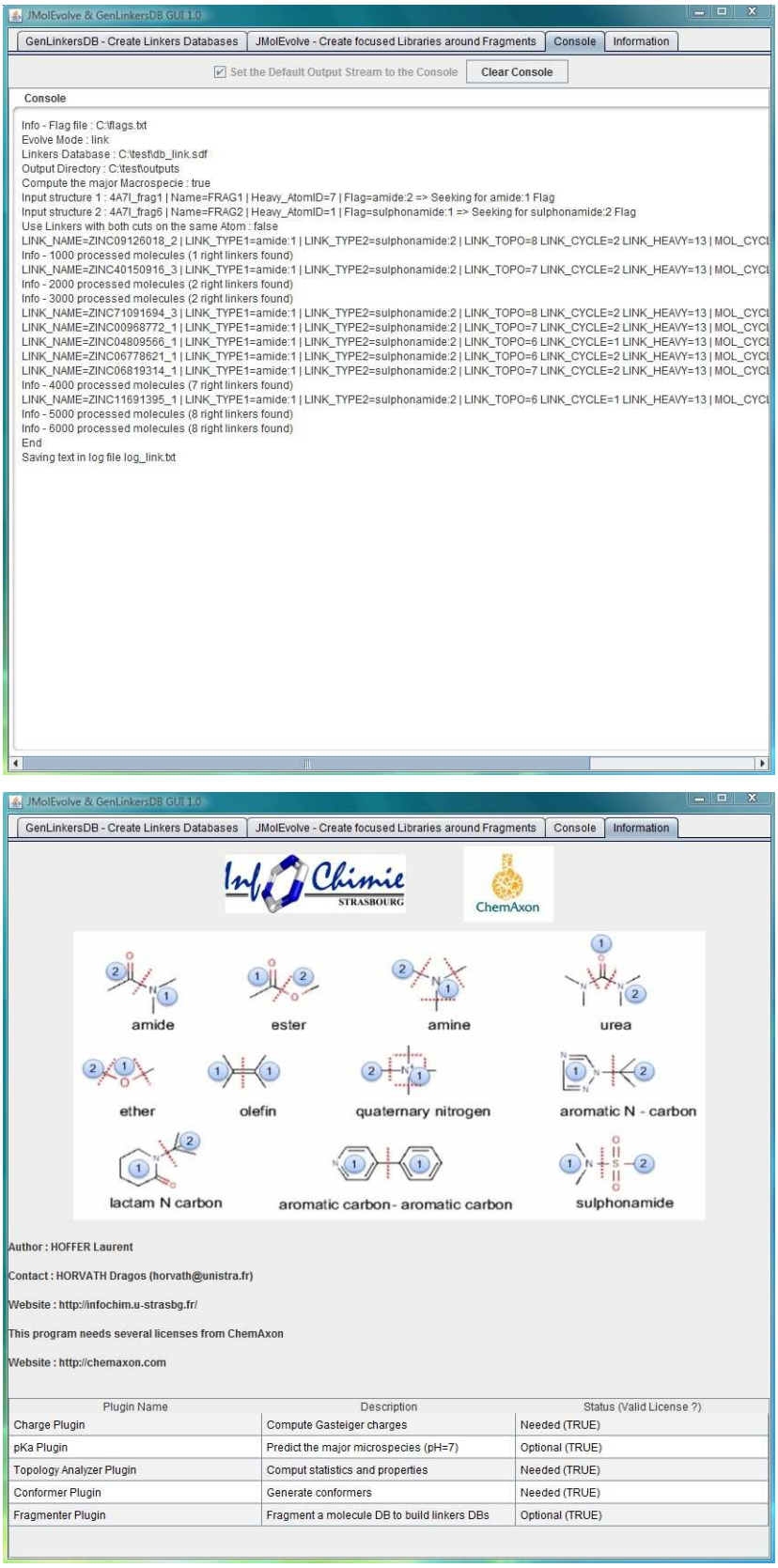


Tableau 7 : Captures d'écran des différents onglets de la surcouche graphique de JMolEvolve.

3 Développements et validation du programme S4MPLE

Ce chapitre décrit les principaux développements informatiques réalisés durant cette thèse, la procédure de calibration de la fonction d'énergie et la validation de l'outil selon les canons en vigueur dans la communauté scientifique.

3.1 La phase de développement informatique

Le développement de cet outil a commencé en amont de cette thèse et a été initié par le Dr. Horvath. Néanmoins, des développements constants, allant de la simple résolution de bugs à l'ajout de nouvelles fonctionnalités, ont été entrepris tout au long de cette thèse.

L'un des premiers travaux a consisté à rendre possible l'utilisation d'AMBER / GAFF tout en gardant la possibilité de revenir au FF initial. Cette première étape est réalisée avec des modifications mineures du code grâce à un préformatage donné des fichiers de paramètres; ce dernier étant facilité par la très grande similarité des différentes composantes énergétiques des deux FF.

De nombreuses améliorations et de nombreux développements ont été apportés tout au long de la thèse, parmi lesquels :

- la possibilité d'utiliser (lecture/écriture/relecture) des fichiers au format MOL2, qui est un format très répandu pour stocker des molécules (ligands organiques et macromolécules biologiques), à la place du format CAR qui est une sorte de fichier PDB modifié. Il en est de même quant à la capacité à pouvoir directement réutiliser, avec une assignation des différents paramètres du FF, des fichiers MOL2 de complexes ligand-récepteur préalablement sauvegardés *via* S4MPLE
- l'assignation automatique des paramètres du FF (charges partielles et types atomiques) des résidus protéiques usuels, *via* l'utilisation d'index topologiques (similaires à ceux décrits pour le PIF). Le FF AMBER repose sur un fichier de topologie qui décrit les types atomiques et les charges associés à chaque atome de chaque résidu, tel un dictionnaire. Un index topologique est précalculé pour chaque atome de chaque résidu considéré. Une clé est ensuite obtenue en concaténant l'élément chimique et l'index. Les différentes clés sont ensuite intégrées dans une table de hachage, et les valeurs associées aux clés sont les paramètres du FF (charges et types atomiques). Pour être cohérent, il a été vérifié qu'au sein d'un résidu donné, seuls les atomes réellement équivalents possèdent des index topologiques identiques. Les résidus considérés incluent les résidus protéiques usuels et leurs différents états de protonation, les variantes N_{ter} et C_{ter} des entités précitées et divers autres résidus (patch N_{ter}, C_{ter}, molécule d'eau, *etc.*). Lors

de la phase d'assignation des paramètres d'un résidu donné, une boucle est réalisée sur l'ensemble de ses atomes pour vérifier l'exactitude des différents index. Si l'ensemble des index est retrouvé alors il y a une assignation des différents paramètres à la volée, sinon un avertissement est lancé. Malheureusement, ce procédé nécessite de devoir spécifier manuellement le type explicite de chaque résidu dans les fichiers moléculaires (par exemple HID au lieu de HIS). Afin de pallier ce défaut, une liste des résidus apparentés est ajoutée dans l'un des fichiers de configuration de S4MPLE. Par exemple, l'alanine peut être représentée par les résidus ALA (alanine standard), NALA (alanine en N_{ter}) et CALA (alanine en C_{ter}). Ainsi, pour paramétrer une alanine localisée en N_{ter} mais nommée ALA dans le fichier moléculaire, un premier essai est systématiquement fait avec ce type par défaut qui échoue ici, puis un second essai est réalisé avec le type associé NALA qui aboutit *in fine* à la bonne paramétrisation. Ce procédé est utilisé pour tous les résidus, en incluant également les nombreux tautomères et états de protonation de l'histidine

- la création de scripts permettant de paramétrer de nouveaux ligands de manière automatique *via* l'utilisation des outils fournis dans AMBERTools [59] (Antechamber [178], ParmChk), et avec une mise à jour des paramètres préformatés lus par S4MPLE si cela s'avère nécessaire
- l'ajout d'une procédure de calcul de RMSD prenant en compte des éléments de symétrie (identifiés selon le même principe que pour le PIF). Le but est de retourner une valeur pertinente (non surestimée) pour des ligands usuels possédant une certaine symétrie. Les diverses combinaisons à tester pour le calcul du RMSD sont obtenues *via* la permutation d'atomes équivalents en prenant garde que chaque ID soit bien entendu présent une fois et une seule pour ne pas biaiser le calcul. Elles sont générées durant la phase d'initialisation, et le RMSD retourné est tout simplement la valeur minimale sur l'ensemble des combinaisons précalculées
- l'amélioration, si souhaité, du placement initial du ligand dans le site durant la phase d'initialisation de l'AG en se basant sur les centres géométriques du ligand et du site. Le ligand subit un nouveau placement lorsque les deux centres sont trop éloignés au regard d'une distance donnée (6 Å par défaut)
- la possibilité de réaliser des minimisations sous contraintes (terme harmonique) temporaires des atomes flexibles de l'entité biologique par rapport à leurs coordonnées initiales. Cette option peut être utile dans le cas du post-traitement d'un docking ultra-flexible de manière à favoriser une localisation proche de la structure native pour les différents résidus flexibles dans la mesure où ce rapprochement n'est pas incompatible avec le positionnement du ligand. Ainsi,

seules les régions où des réarrangements conformationnels s'avèrent réellement nécessaires possèderont une conformation alternative, ce qui se révèle plus cohérent et facilite d'autant plus la phase d'analyse des résultats

- l'optimisation du positionnement des groupements hydroxyles (ligand seul ou ligand+site) des différentes poses lors de la phase de post-traitement dans le but d'optimiser le réseau de liaisons hydrogène. Il y a une rotation selon un pas défini (par exemple 30°) autour de chaque torsion pilotant la position de l'hydrogène polaire d'un groupe hydroxyle, suivie d'une estimation énergétique. La position finale retenue est celle qui minimise l'énergie. Cette possibilité est intéressante dans le cas d'un cross-docking ou d'un criblage virtuel, puisque le réseau de LH du site peut être modifié à la suite de la fixation du ligand

Bien que cela ne fasse pas à proprement parler d'un développement informatique, une stratégie d'optimisation des principaux paramètres de l'algorithme génétique est également entreprise. Ce type d'algorithme incluant de nombreux paramètres plus ou moins empiriques, il est bien connu que leur détermination est terriblement complexe et nécessite de très nombreuses simulations. Pour des raisons objectives de calendrier, cette partie fondamentale a été réalisée en aval des différents benchmarks de redocking mentionnés dans les chapitres 3 et 4. Pour chaque paramètre, un nombre restreint de valeurs possibles est arbitrairement choisi. Les principaux paramètres d'intérêt sont :

- la stratégie d'évolution (**evol** / base / elit)
- le nombre de générations (200 / 300 / **500**)
- l'utilisation ou non d'une stratégie "tabu" (oui / **non**)
- les critères d'arrêt pour les minimisations
    - seuil pour la norme du gradient (**1e-7** / 0,01 / 0,1 / 1)
    - gain d'énergie minimal entre chaque étape (**0,01** / 0,1 / 1 / 10)

Les valeurs initialement par défaut sont en gras, et sont très strictes pour les critères d'arrêt des minimisations. Toutes les combinaisons de paramètres sont testées en docking sur un sous-ensemble du Astex Diverse Set [143], avec 8 simulations indépendantes pour chaque complexe. Ce sous-ensemble comprend 8 complexes PDB (1HWW, 1K3U, 1KZK, 1LPZ, 1S19, 1T46, 1Y6B, 2BM2) comprenant des ligands de complexités variées (masse moléculaire, nombre de liaisons à rotation libre, *etc.*) pour lesquels une erreur systématique de FF n'a pas été mise en évidence. Pour chaque combinaison et chaque complexe, la médiane de l'énergie minimale sur les 8 simulations indépendantes est calculée. Le but est ensuite de trouver un compromis "qualité *vs.* temps de simulation", notamment par rapport

aux paramètres par défaut utilisés dans les benchmarks. Ici, le terme de qualité fait référence à la capacité à trouver une énergie (E_{pot}) la plus faible possible.

Les stratégies "base" et "evol" ont donné de meilleurs résultats que la stratégie "elit". L'une des conclusions est que le nombre de générations doit être maintenu pour des ligands "drug-like" (500), mais il peut être logiquement baissé pour des composés moins complexes comme les fragments (200). Le fait d'être moins strict dans les critères d'arrêt des minimisations pendant la phase d'EC a un impact notable et logique sur le temps de calcul sans pour autant détériorer la qualité des résultats. Ainsi, le temps de calcul est diminué de 30% environ par rapport au temps moyen signalé dans le premier article (de l'ordre de plusieurs heures, voir le §3.4) rien qu'en passant le gain minimal d'énergie entre deux étapes de minimisation de 0,01 à 1. Enfin, la stratégie "tabu" actuelle, dont le but est de pénaliser des régions de l'espace de recherche déjà explorées, n'a pas un impact suffisamment intéressant sur l'EC lorsque l'on met dans la balance la détérioration logique du temps de simulation.

3.2 La procédure de calibration de la fonction d'énergie

Ce sous-chapitre décrit la stratégie utilisée pour paramétrer les différentes constantes d'intérêt (la liste exhaustive est donnée dans le Tableau 8). Ces paramètres sont notamment liés aux composantes énergétiques additionnelles du "Fit FF" qui ont été introduites préalablement au §2.1.4.

Le terme de contact reposant sur une constante de force pour chaque type atomique de l'élément carbone (N>30), ces derniers sont répartis en trois catégories (carbone aliphatique, carbone aromatique et carbone dans un environnement polaire). Cette mutualisation des types atomiques, dont la classification finale est donnée dans le tableau 1 de l'article I (voir le §3.4), a permis de fortement diminuer le nombre de paramètres à déterminer.

Les FF AMBER et GAFF étant respectivement spécialisés sur les macromolécules biologiques et les petites molécules organiques, il a été convenu de diviser la procédure de calibration en deux étapes, chacune reposant sur une réévaluation énergétique ("rescoring") d'ensembles de conformères/poses appartenant à ces deux mondes :

1) la première étape se focalise sur un jeu de peptides de structure variée, et vise à déterminer la plupart des paramètres décrits dans le Tableau 8, à l'exception des constantes de force **Kc_type** (Kc_polarized, Kc_arom et Kc_aliph) du terme de contact qui restent fixes à la valeur 0,1 (voir ci-dessous)
2) la seconde étape utilise des complexes ligand-récepteur de la PDB dans le but d'optimiser les trois constantes **Kc_type**

Cette dissociation vient également du fait qu'une première tentative a été effectuée avec l'intégralité des types atomiques de l'élément carbone comme paramètres à part entière, d'où un nombre de variables justifiant cette dichotomie. De plus, un certain consensus autour du poids 0,1 avait alors émergé pour de nombreux types atomiques de l'élément carbone. Cette stratégie en deux étapes a été maintenue après la mutualisation des types atomiques C en trois classes distinctes.

Paramètres	Descriptions et valeurs utilisées par défaut pour Core FF
epsilon	Ce paramètre représente la constante diélectrique relative (défaut=2).
desolv_factor	Il s'agit du poids relatif du terme de désolvatation (défaut=0).
minq_to_desolv	Ce paramètre permet de ne considérer que certaines paires d'atomes pour le terme de désolvatation. Seules les paires dont l'un des atomes possède une charge partielle supérieure (en valeur absolue) à ce seuil sont considérées. Autrement dit, $E_{desolv_{ij}}$ est nulle pour les paires filtrées (défaut=0, toutes les paires sont considérées).
hbond_bonus	Ce paramètre modélise la constante de force du terme contact favorable pour les interactions de type liaison hydrogène (défaut=0).
repulsive_factor	C'est la constante de proportionnalité ajoutée devant la composante répulsive du terme de VdW dans le but de moduler celle-ci (défaut=1).
vicinal_weight	Ce paramètre module le poids de la composante répulsive du terme de VdW des paires non liées dites "vicinales" des angles de torsion (défaut=0,5).
Kc_polarized	Ce paramètre représente la constante de force du terme contact favorable pour un carbone dans un environnement polaire (défaut=0).
Kc_arom	Ce paramètre représente la constante de force du terme contact favorable pour un carbone aromatique (défaut=0).
Kc_aliph	Ce paramètre représente la constante de force du terme contact favorable pour un carbone aliphatique (défaut=0).
desolv_scale_ion	C'est un facteur de proportionnalité, spécifique aux paires non liées impliquant un ion, pour la composante désolvatation (défaut=1, voir le §3.2.2).
desolv_scale_hb	C'est un facteur de proportionnalité, spécifique aux paires non liées de type liaison hydrogène, pour la composante désolvatation (défaut=1, voir le §3.2.2).

Tableau 8: Liste des différents paramètres d'intérêt dans le cadre de la procédure de calibration de la fonction d'énergie.

La phase finale de validation, dont le but est de vérifier la pertinence des termes additionnels du champ de force "Fit FF" par rapport au natif "Core FF", consiste à réaliser une étude de redocking sur un jeu de référence externe également composé de complexes PDB (voir le §3.3).

3.2.1 Etape 1) rescoring de conformères de peptides

Etant donné que cette première partie est peu décrite dans l'article I disponible au §3.4, elle est intégralement décrite et discutée ici. Elle repose sur un jeu de peptides de structures secondaires variées (voir le Tableau 9) pour attribuer essentiellement les paramètres du modèle de solvant.

PDB	Nombre de résidus	Méthode expérimentale	Type de structure secondaire
1L2Y	20	RMN	hélice α
1LE1	13	RMN	brin β
1UAO	10	RMN	petit brin β
1VII	36	RMN	hélice α
2KEF	25	RMN	brin β (+4 ponts disulfures)

Tableau 9: Caractéristiques générales des peptides d'intérêt.

Les sous-étapes de cette première partie sont détaillées ci-dessous et illustrées à la Figure 44 :

I. Un grand nombre de conformères, de l'ordre de plusieurs centaines de milliers pour chaque peptide, est généré à l'aide de S4MPLE en utilisant le FF natif AMBER / GAFF. Une stratégie non biaisée (32 simulations indépendantes avec un EC par défaut, 100 individus dans la population, 10000 générations) et une biaisée (1 simulation avec la stratégie de recherche locale à partir de la structure native) sont utilisées pour générer un ensemble de conformères comportant à la fois des structures proches du repliement natif et des structures plus ou moins éloignées de celui-ci. Ce mélange de conformères, comportant beaucoup plus de structures non natives, est nécessaire puisque le but ici n'est pas d'évaluer la capacité de S4MPLE à réaliser du repliement *ab initio* de peptides, mais bien de se focaliser uniquement sur l'aspect énergétique, d'où la nécessité de structures proches du repliement natif.
II. La phase suivante consiste à déterminer des jeux de paramètres, à l'aide d'un scan aléatoire mais borné, qui permettent de classer en tête de liste des conformères similaires au repliement

expérimental pour le plus grand nombre possible de peptides du jeu. On entend par tête de liste les 30 meilleurs conformères d'un point de vue énergétique (E_{pot}) et non redondants à un seuil de *minfpdiff* donné. Ces meilleurs conformères sont simplement extraits, pour chaque peptide du jeu, à l'aide d'une réévaluation énergétique sur l'ensemble préalablement généré à l'étape I.

III. Les quelques jeux de paramètres prometteurs sont ensuite utilisés pour réaliser un nouvel EC avec S4MPLE (non biaisé et biaisé comme précédemment à l'étape I) afin de confirmer leur qualité potentielle, à savoir permettre de classer des conformères natifs en tête de liste (même critères qu'à l'étape II). Si cet EC, utilisant les nouveaux paramètres, découvre de nouvelles géométries non natives et plus stables que la meilleure géométrie correctement repliée, ce jeu est écarté et les nouvelles mauvaises géométries sont ajoutées à l'ensemble de conformères de référence.

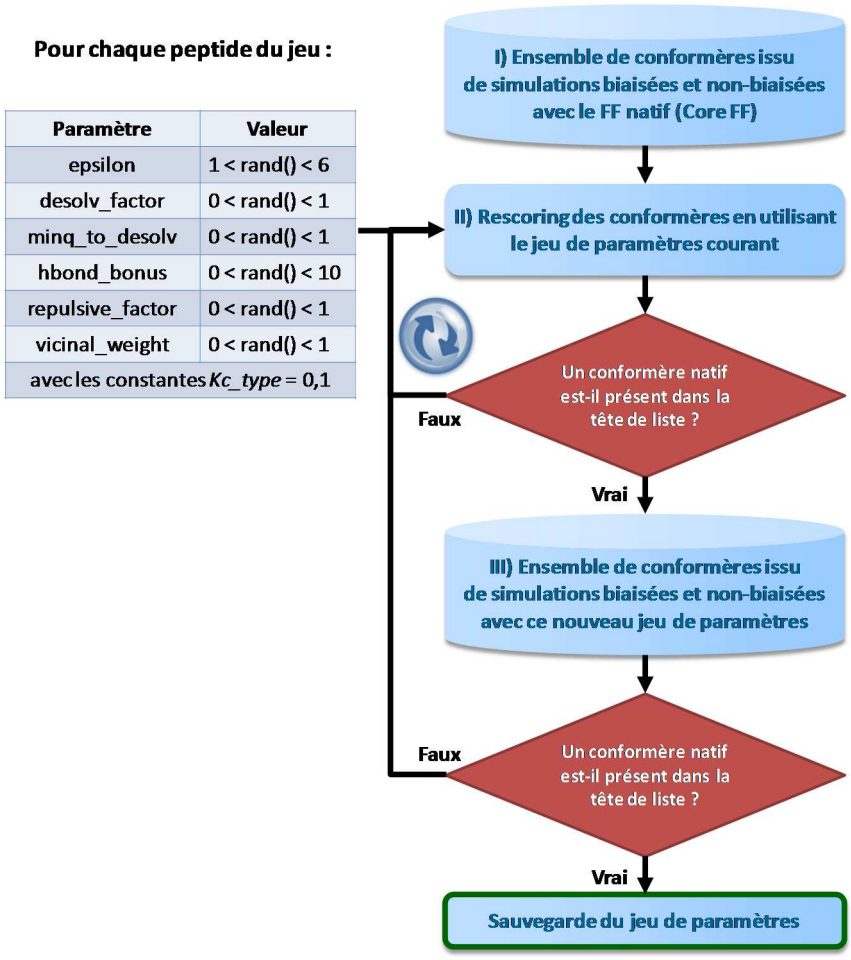


Figure 44: Description de la première étape du protocole de calibration.

Le critère de similarité (conformère natif ou non) repose sur le PIF (empreinte d'interaction) avec comme référence celui de la structure expérimentale. L'expression "tête de liste" fait référence aux 30 meilleurs conformères (selon E_{pot}) et non redondants (également avec le PIF).

Un jeu de paramètres, satisfaisant au mieux aux critères énoncés ci-dessus après plusieurs dizaines d'itérations, est donné dans le Tableau 10. Ce "FF préliminaire" a permis d'obtenir en tête de liste (top 30, voire à la première place pour 1L2Y et 1VII) au moins un conformère proche de la structure native pour tous les peptides à l'exception de 1LE1. Ces résultats sont toutefois encourageants, notamment par rapport à ceux du FF natif car un conformère non natif est systématiquement retrouvé au rang 1 avec ce dernier (par conséquent, le peptide 1LE1 pose également problème). Ce type de résultats, à savoir le fait de trouver des conformères non natifs de plus basse énergie que celle de la structure expérimentale relaxée en utilisant un FF brut (sans modèle de solvant implicite ou explicite), n'est pas si étonnant et a déjà été évoqué dans la littérature [191].

Paramètre	Valeur
epsilon	4
desolv_factor	0,1
minq_to_desolv	0,125
hbond_bonus	2
repulsive_factor	0,75
vicinal_weight	0,033

Tableau 10: Détail du meilleur jeu de paramètres ("FF préliminaire") obtenu à l'issue de cette première étape de la procédure de calibration.

Certaines valeurs, en plus de permettre une relative extraction des conformères natifs, apparaissent assez intuitives :

- une constante diélectrique relative égale à 4, ce qui est dans la gamme usuelle
- un facteur de proportionnalité pour le terme répulsif de VdW légèrement inférieur à 1 (une valeur extrêmement faible n'aurait aucun sens par exemple)
- une valeur pour le paramètre "minq_to_desolv" permettant de différencier les paires polaires (par exemple O..O) ou partiellement polaires (par exemple O..C) des paires impliquant deux atomes non polaires (C..C), ce qui était le but souhaité lors de l'ajout de ce paramètre dans la composante pénalité de désolvatation (voir le §2.1.4)

Etant donné que cette étude complète (EC biaisé et non biaisé, étape III de la Figure 44) est à nouveau intégralement réalisée pour la version finale du "Fit FF", le détail des résultats pour chaque peptide

selon les deux FF considérés (Core FF et Fit FF) sera donné et discuté ultérieurement au §3.5. Enfin, ces résultats ouvrent la voie à la seconde étape impliquant des complexes PDB ligand-récepteur.

3.2.2 Etape 2) rescoring de poses ligand-récepteur

Cette seconde étape, reposant sur les résultats issus de l'étape 1 (voir le Tableau 10), a pour but d'optimiser les constantes **Kc_type** du terme de contact à partir du rescoring d'un important nombre de poses sur un jeu de complexes PDB ligand-récepteur (Astex/CCDC-Clean [142], N=191, voir le Tableau 11). Etant donné que cette étape est intégralement décrite dans l'article I (inséré au niveau du §3.4), seuls les grandes lignes et les résultats sont décrits et discutés ici.

1A28	1BBP	1CX2	1EOC	1HDC	1LST	1POC	1TNL	2CPP	3TPI
1A42	1BL7	1D0L	1EPB	1HFC	1LYB	1PPC	1TRK	2CTC	4AAH
1A4G	1BMA	1D3H	1EPO	1HIV	1LYL	1PPH	1TYL	2DBL	4COX
1A4Q	1BYB	1D4P	1ETR	1HPV	1MBI	1PPI	1UKZ	2FOX	4CTS
1A6W	1BYG	1DBB	1ETS	1HRI	1MCQ	1PSO	1ULB	2GBP	4DFR
1A9U	1C12	1DBJ	1ETT	1HSB	1MDR	1PTV	1UVS	2H4N	4FBP
1AAQ	1C1E	1DD7	1F0R	1HSL	1MLD	1QCF	1UVT	2IFB	4LBD
1ABE	1C5C	1DG5	1F0S	1HYT	1MMQ	1QPE	1WAP	2LGS	5ABP
1ABF	1C5X	1DHF	1F3D	1IBG	1MRG	1QPQ	1XID	2MCP	5CPP
1ACJ	1C83	1DID	1FAX	1IDA	1MRK	1RNE	1XIE	2PCP	5ER1
1ACL	1CBS	1DOG	1FEN	1IMB	1MTS	1RNT	1YDR	2PHH	7TIM
1ACM	1CBX	1DR1	1FGI	1IVB	1MUP	1ROB	1YDT	2QWK	
1AI5	1CIL	1DWB	1FKG	1IVQ	1NCO	1RT2	1YEE	2R07	
1AOE	1CKP	1DWC	1FKI	1JAP	1NGP	1SNC	25C8	2TMN	
1AQW	1CLE	1DWD	1FL3	1KEL	1OKL	1SRJ	2AAD	2TSC	
1ATL	1COM	1EAP	1FLR	1LAH	1OKM	1TDB	2ACK	2YPI	
1AZM	1COY	1EBG	1FRP	1LCP	1PBD	1TMN	2ADA	3CPA	
1B58	1CPS	1EED	1GLP	1LDM	1PDZ	1TNG	2AK3	3ERD	
1B9V	1CQP	1EI1	1GLQ	1LIC	1PHD	1TNH	2CHT	3ERT	
1BAF	1CVU	1EJN	1HAK	1LNA	1PHG	1TNI	2CMD	3HVT	

Tableau 11: Liste des différents complexes PDB utilisés dans le jeu d'entraînement (sous-ensemble du jeu Astex/CCDC-Clean).

Le FF préliminaire a été utilisé tel quel en redocking sur un sous-ensemble de ce jeu de complexes PDB (N=30, ligands variés, présence/absence d'ions métalliques dans le site).

Sur la base de l'analyse visuelle de la meilleure pose énergétique en sortie de docking, il est rapidement apparu que le FF préliminaire est globalement pertinent, à l'exception de certains cas très spécifiques :

- la coordination entre le ligand et le cation métallique est souvent approximative bien que le RMSD soit faible sur l'ensemble du ligand. Par exemple, une coordination simple est préférée à la coordination double qui est observée expérimentalement dans le complexe 1CBX
- des liaisons hydrogène de qualité très moyenne au niveau de l'angle D-H..A, aux alentours de 140°, pour certains complexes impliquant des oses. Une grande densité de liaisons hydrogène (via les groupements hydroxyles dont l'hydrogène polaire est flexible) est observée dans ce type de complexe, entre autre 1ABE et 1ABF

L'hypothèse selon laquelle le paramètre **desolv_factor** du terme de désolvatation (voir l'Eq 35) est responsable de cette situation est rapidement émise : la valeur obtenue est vraisemblablement trop élevée et la pénalité annule l'attraction électrostatique à très faible distance. De plus, ce terme en Q^2 (avec Q la charge partielle) semble également très élevé avec une entité très chargée comme un cation métallique. La diminution de la valeur de ce paramètre a permis de restaurer des interactions de qualité, validant l'hypothèse émise au début du paragraphe. A ce stade, trois stratégies distinctes sont envisagées :

- retourner à l'étape précédente (rescoring de conformères de peptides)
- diminuer la valeur du paramètre **desolv_factor**
- mettre à l'échelle certaines paires non liées spécifiques

La première, qui est également la plus radicale, est mise de côté à cause des résultats encourageants obtenus à l'étape précédente avec le jeu de peptides. La seconde voie aurait également comme conséquence de baisser la pénalité liée à des interactions non favorables à faible distance comme les paires accepteur-accepteur et donneur-donneur. Ainsi, le choix de se focaliser sur la mise à l'échelle de certaines paires non liées spécifiques comme les LH ou les paires impliquant un ion est réalisé, d'autant plus que des entités fortement chargées (ions métalliques) étaient absentes du jeu de peptides.
Deux paramètres indépendants de mise à l'échelle (**desolv_scale_ion** et **desolv_scale_hb** déjà introduits dans le Tableau 8) sont donc définis, et peuvent varier de 0 à 1 par incrément de 0,1. La recherche de leur valeur optimale est obtenue empiriquement (approche par essais et erreurs en docking) sur le sous-ensemble des 30 complexes ayant mis en évidence ces difficultés ponctuelles. Le meilleur compromis obtenu est **desolv_scale_ion**=**desolv_scale_hb**=0,1. Ce simple réglage

supplémentaire a permis de corriger le tir par rapport aux imprécisions du FF préliminaire (notamment *via* l'amélioration des critères géométriques des interactions incriminées) sans pour autant détériorer le docking des autres complexes. Dès lors, il a été convenu de continuer la procédure de calibration de la fonction d'énergie.

A l'instar de la phase focalisée sur des peptides, un grand nombre de poses de docking est généré pour chaque complexe à l'aide de S4MPLE. La figure 4 de l'article I (voir le §3.4) détaille les étapes de cette seconde partie de la calibration de la fonction d'énergie. Pour aboutir à un ensemble de poses variées et représentatives, chaque ligand a été docké en utilisant plusieurs variantes de FF :

- le champ de force natif Core FF
- le champ de force dit "FF préliminaire", ainsi que certaines variantes avec des valeurs **Kc_type** représentatives de l'espace de recherche borné (voir ci-dessous), entre autre :
    - des poids nuls pour l'ensemble des classes (**Kc_type**=0,000)
    - des poids élevés pour l'ensemble des classes (**Kc_type**=0,300)
    - une asymétrie entre les trois classes considérées, comme un poids plus important pour la classe des carbones polaires (idem pour un poids moins important)

Une simulation de docking de 400 générations a été réalisée pour chaque complexe et pour chaque variante de FF. Les poses expérimentales minimisées dans le site de liaison avec les différentes variantes du FF y sont également incluses. Un scan systématique, sur un total de 10 valeurs discrètes (0,000; 0,010; 0,025; 0,050; 0,075; 0,100; 0,150; 0,200; 0,250; 0,300) pour chaque paramètre, est utilisé pour déterminer le triplet de constantes **Kc_type** permettant de classer le plus souvent possible un mode de liaison expérimental au premier rang selon l'E_{pot} du système. Le critère pour distinguer une pose de docking proche du mode de liaison expérimental est le RMSD avec un seuil limite à 2 Å. Ainsi, pour chaque complexe, la meilleure pose extraite de l'ensemble correspondant est sauvegardée pour chacune des 1000 combinaisons de l'espace de recherche.
Sur l'ensemble de ces combinaisons, le meilleur résultat obtenu est 154/191 complexes avec la meilleure pose considérée comme similaire au mode de liaison expérimental, soit un taux de succès de 80,6%. Les 11 triplets ayant abouti à ce même seuil sont détaillés dans le tableau 2 de l'article I (voir le §3.4). A l'exception d'un triplet unique, tous les autres se situent dans une même zone de l'espace de recherche, à savoir un poids très faible pour les carbones dans un environnement polaire ($0 \leq$ **Kc_polarized** $\leq 0,025$) et des poids bien plus élevés pour les carbones aromatiques ($0,075 \leq$ **Kc_arom** $\leq 0,2$) et aliphatiques ($0,15 \leq$ **Kc_aliph** $\leq 0,2$). Ce regroupement dans une zone privilégiée

de l'espace de recherche rend ces résultats bien plus robustes par rapport à une éventuelle dissémination des points au sein de cet espace. De plus, cette discrimination vis-à-vis de la classe **Kc_polarized** a intuitivement plus de sens physique dans le contexte d'un terme conceptuellement ajouté pour renforcer les interactions entre zones hydrophobes. Le triplet final retenu est la médiane sur l'ensemble de ces résultats, soit **Kc_polarized**=0,01 et **Kc_arom**=**Kc_aliph**=0,15. Enfin, il est à noter que :

- ce triplet est déjà naturellement présent dans les 11 meilleurs
- l'utilisation d'un seuil de RMSD plus strict (1,5 Å) renvoie à nouveau le triplet final parmi les meilleures combinaisons, et avec un taux de succès qui reste tout à fait acceptable (74%)

Le jeu final de paramètres, nommé Fit FF (voir le tableau 3 de l'article I, §3.4), est ensuite validé au moyen de simulations de redocking sur un jeu externe à la procédure de calibration (voir ci-dessous).

3.3 Validation à l'aide de simulations de redocking

L'utilisation d'un jeu externe reconnu au sein de la communauté scientifique permet de valider la pertinence des jeux de paramètres obtenus. Ce jeu, nommé Astex Diverse Set [143], ne comporte que des complexes de haute résolution où il n'y a aucune ambiguïté sur la densité électronique du ligand. Comme son nom l'indique, il inclut de nombreuses cibles et des ligands aux châssis moléculaires très variés. Le taux de succès est à nouveau défini comme le pourcentage de complexes pour lesquels le RMSD de la meilleure pose selon E_{pot} est inférieur à 2 Å par rapport à la position expérimentale du ligand. Le taux de succès relatif à l'EC est le même critère mais sans distinction de rang : par exemple, il y a un succès dès qu'une pose correcte est retrouvée dans les 30 poses sauvegardées.

Le détail relatif à la préparation des différents complexes (ligands, sites de liaison, fichiers de configuration) est disponible dans l'article associé (voir le §3.4). Il est à noter que pour éviter tout biais, les points d'ancrage (“hot spots”) sont systématiquement définis avec le mode automatique pour toutes les simulations de redocking.

La statistique relative à la qualité d'EC valide S4MPLE en tant que programme de docking puisque celle-ci se situe autour de 95% en moyenne sur 10 simulations indépendantes sur l'ensemble du jeu de validation externe. De plus, les champs de force Core FF et Fit FF ont obtenu de très bons taux de succès, respectivement 76% et 85% (également sur 10 simulations indépendantes). L'avantage sensible de Fit FF par rapport à Core FF valide l'ajout des termes additionnels et leur calibration. Il est à noter que ce taux reste relativement élevé en utilisant des seuils de RMSD bien plus stricts : il passe par exemple pour le Fit FF de 85% à 78% et 69%, respectivement pour des seuils de 2,0 Å, 1,5 Å et

1,0 Å. Le détail de ces résultats ainsi qu'une discussion à propos des échecs systématiques sont disponibles dans l'article associé (voir le §3.4).

La première étape du projet, à savoir la validation usuelle de l'outil d'EC S4MPLE ainsi que sa fonction d'énergie Fit FF basée sur les champs de force AMBER / GAFF, est validée par ces résultats. Il devient dès lors possible d'envisager d'autres types de simulations, notamment celles se focalisant sur des molécules organiques de type fragment. Ce dernier cas permet de se recentrer par rapport au cahier des charges initial, et fera l'objet du chapitre 4.

3.4 Article I

L'article I est inséré dès la page suivante (en cours de soumission) [185].

S4MPLE – Sampler For Multiple Protein-Ligand Entities: Methodology & Rigid-Site Docking Benchmarking

Laurent HOFFER[‡§], Camelia CHIRA[†], Gilles MARCOU[‡], Alexandre VARNEK[‡] & Dragos HORVATH[‡*]

‡ Université de Strasbourg, 1 rue B. Pascal, Strasbourg 67000, France,

§ Novalix, BioParc, bld Sébastien Brant, BP 30170, Illkirch 67405 Cedex, France,

† Centre for the Study of Complexity at Babes-Bolyai University, Cluj-Napoca, Romania

* Corresponding author: dhorvath@unistra.fr

Abstract

This paper describes the development of the program S4MPLE (Sampler for Multiple Protein-Ligand Entities), which is able to perform extensive conformational sampling of one or more molecules, using an original, solvent-effect-enabled extension of the AMBER/GAFF force field. It is designed as a basic, unified approach, treating docking as a particular case of conformational sampling, involving several molecules. It makes no fundamental distinction between binding site and ligand, and allows full control of the considered *vs.* fixed degrees of freedom, putatively allowing docking simulations with sidechain and backbone flexibility of the protein. By contrast to a previous version operating on torsional degrees of freedom [1, 2], the herein presented tool supports complete flexibility, being based on a Lamarckian Genetic Algorithm using specially designed genetic operators, transparently addressing both intra- and intermolecular degrees of freedom. It is designed to eventually operate in massively distributed computing environments.

Since S4MPLE is not meant to be yet another rigid docking program, dedicated studies highlighting some of its specific abilities – docking of fragment-like compounds, cross-docking with flexible protein sites, simultaneous docking of multiple ligands, including free crystallographic waters – were or are being undertaken and published [3] . The present paper, however, discusses (a) the methodology, (b) the set-up of the force field energy functions and (c) their validation in classical redocking tests, using the single-CPU workstation version of the tool. S4MPLE uses the AMBER force field [4] to simulate macromolecules, and GAFF [5] to manage small compounds such as ligands. Additional terms, such as a favorable contact bonus (for hydrophobic contacts and hydrogen bonds) and a pair-based desolvation term [6], were added. Calibration and validation of these additional terms in the herewith resulting augmented AMBER/GAFF force field was performed on reference sets of PDB complexes (Astex-CCDC [7] and Astex Diverse Set [8] for the calibration and validation, respectively). More than 80% success in redocking was achieved (RMSD of top-ranked pose < 2.0 Å).

1 Introduction

The knowledge of the experimental structure of the target is precious information, enabling the use of structure-based drug design (SBDD) methods. Through SBDD, one can rapidly progress from weak-affinity compounds toward lead-like molecules. One of the main SBDD approaches is docking, which can be defined as the prediction of ligand-target complex structures at the atomic scale. It can be used for various purposes:

- filtering a database in order to select and test alleged binders to a target
- solving the binding mode of known binders in order to rationalize their subsequent optimizations

A lot of docking algorithms have been developed. They all include two key steps: the sampling of the degrees of freedom (DoF) of the system, and the scoring of potential complexes (ranking of poses). The search algorithm must be sufficiently efficient to perform a relevant sampling of partners. There are three main docking simulations types:

- rigid docking (target and ligand are considered rigid)
- semi-flexible docking (flexible ligand and rigid target)
- flexible docking (flexible ligand and partly flexible target)

The first developed docking approaches belonged to first category. These programs are very fast but inaccurate for significant number of problems since the bioactive conformation of ligand is rarely known. The program FRED [9] behaves as a fully rigid tool during the sampling step, but it uses a pool of conformers for each ligand, so it avoids the main disadvantage of the fully rigid approach. Most other tools explicitly sample ligand conformations during the docking process. Simultaneously, the search engine explores the different positions of the ligand within the rigid binding site, using translational and rotational degrees of freedom of the ligand. FlexX [10], Glide [11], DOCK [12] and Surflex [13] are widely used software of the semi-flexible category. Still, the hypothesis of a rigid target is often penalizing, since ligand binding often implies structural rearrangements of the target site, ranging from simple side chain displacements to whole loop/subdomain rearrangements. There are different ways to perform flexible target docking. The most intuitive approach is to use an ensemble of different conformations of the receptor, and to perform semi-flexible docking on each target structure. Various conformations of the binding site obtained by X-ray or snapshots from molecular dynamics simulation can be used to create the collection of inputs. An alternative way is to generate a single

united protein model, as in the FlexE strategy [14], from an ensemble of superimposed structures. In this model, the non-conserved regions are treated as alternative locations. Another way to simulate a flexible target is to explicitly add more or less flexibility for the binding site. As for example, Autodock [15, 16] and Gold [17, 18] can add local flexibility : from free polar hydrogens to flexible sidechains. The programs RosettaLigand [19], FlipDock [20], FITTED [21-23] and IFD [24] allow for both backbone and sidechain flexibility. Conversely to molecular dynamics, only a substructure of the target is considered as flexible in most flexible docking runs (e.g. some sidechains, or part of backbone and sidechains). The real challenge is to develop algorithms able to screen a database with flexibility in order to simulate conformational changes of the binding site (phenomenon known as "induced fit effects").

1.1 Sampling strategies

In practice, an exhaustive exploration of the degrees of freedom of the whole system is not possible, especially with large ligands and/or flexible binding sites. Molecular dynamics-based approaches have a low propensity to bypass energy barriers, and are very slow to move the ligand into the constraining active site. However, they have a real physical basis, and are very useful during local optimization stages in order to find the closest local minima. Alternative search heuristics are able to quickly (relatively) converge toward an acceptable solution but not necessarily the best one. The most popular stochastic heuristics are Monte Carlo and evolutionary-based strategies. These optimization algorithms are very general, so they can be adapted for many purposes, among them docking. Monte Carlo (MC) strategies [25] consist in several cycles of random modifications of a system coupled with a thermal bath, so that the acceptance probability of a random step is dictated by the Metropolis criterion (the Boltzmann term of the associated energy variation) at a given temperature. In Simulated Annealing (SA) [26], the temperature typically decreases throughout the simulation.

Evolutionary/Genetic Algorithms (EA/GA) [27-30] rely on simulations of Darwinian evolution: "individuals" or "chromosomes" (vectors encoding points in the problem space to visit – here, coordinates of a molecule or of a molecular complex) are more likely to be selected into the next generation population if their fitness is higher – here, the conformational stability. In practice, the sampling is realized on a population of chromosomes. Chromosomes undergo random modifications with biological operators such as mutation and crossing over. Several docking tools, based on EA/GA, have been developed, among others GOLD, Autodock, FlipDock, EAdock [28, 31], LGA [32] and FITTED. Since S4MPLE is GA-driven, the technical details thereof will be discussed in the Methods section.

1.2 Scoring functions

In-depth conformational sampling should, in principle, enumerate enough possible states to allow determination of the free energies of various conformers (understood as ensembles of geometries within a potential energy well). If so, then the free energy differences between bound and dissociated site-ligand states should directly lead to the binding affinity of that protein-ligand pair. In practice, this is not possible – both because sampling is never exhaustive enough, but mainly because the energy function of geometry is insufficiently accurate. Therefore, most programs include an additional scoring function – basically a QSAR equation trying to predict binding free energy as a function of a (typically, the lowest-energy) site-ligand pose. Such empirical scoring functions [33-35] are a weighted sum of ligand–receptor interaction terms. These often include hydrogen bonds, ionic bonds and hydrophobic contributions, clashes or even entropic penalties. Weights of each term of this QSAR equation are obtained by a regression analysis on a reference set of receptor–ligand complexes with known binding affinities, or from relative occurrence likelihoods in experimental structural databases.

At this point, S4MPLE does not include any scoring function, the only criterion used to rank sampled geometries being their force field-based energies (potential energy). The present work focuses on the ability of S4MPLE to reproduce experimentally relevant geometries on the basis of the simple force field energy criterion. Further efforts will be dedicated to quantitative ligand affinity predictions.

1.3 Overview of this Article

S4MPLE has been conceived as a completely general conformational sampling program, specifically targeted at flexible docking problems – which are nothing but conformational sampling of two or more interacting species. Full control over the considered and respectively frozen DoF allows S4MPLE to be used in any of the three main docking scenarios cited in the introduction, and beyond (small peptide folding, protein loop repositioning in protein homology models, docking in presence of mobile protein loops, simultaneous or concurrent docking of several ligands into a same site, *etc.*). Previous publications concerning this development were of technical nature, and considered the deployment strategies of the GA-driven conformational sampling algorithm on computer grids [27, 36-39]. As mentioned before, the specific difficult sampling and docking problems for which S4MPLE was originally designed were or will be subject of independent publications. This work focuses on molecular modeling results, specifically addresses three key aspects that need to be taken into account when designing and validating a new sampling and docking tool:

- The conformational sampling algorithm and its novel elements (genetic operators for fully flexible modeling, differentiable interaction fingerprints to monitor conformer diversity)
- The set-up and fitting of the force field (FF) engine used for energy calculations, based on the classical AMBER [4, 40] /Generalized AMBER (GAFF) [5, 40], but augmented with chirality control terms, a continuum solvent model and favorable contact (hydrophobic, H bonding) bonus terms. The latter, together with classical customizable FF terms – such as the chosen dielectric constant, and a herein introduced repulsive Van der Waals coefficient weighing term – need fitting.
- S4MPLE proficiency in "classical" docking benchmarks.

2 Methods

2.1 S4MPLE

S4MPLE is a flexible, modular molecular modeling tool, based on a hybrid Genetic Algorithm (GA), combining molecular modeling-specific optimization in addition to classical evolutionary sampling strategies. Allowing full control of the considered degrees of freedom, S4MPLE is a completely general approach to visit the conformational space of arbitrary molecules or molecular complexes. Its focus is on thorough geometry sampling, without need to rely on compound class-specific working hypotheses (amino acid rotamer libraries, *etc.*), potentially restraining its applicability domain. As such, it may be equally well used for conformational sampling and docking – which is nothing but sampling of a ligand in presence of a binding site. The "site" does however not need to be a protein, which may eventually render S4MPLE useful for simulations of arbitrary molecule self-assembly processes. Being conceived in view of large-scale deployment on computer grids, the only limitations of its applicability are (a) the studied system size *vs.* available computational resources, and (b) the availability of force field parameters for the studied molecules. S4MPLE is written in object-Pascal, and used in command-line mode.

2.2 Force field

Two force fields are currently implemented in S4MPLE:

- CVFF (Consistent Valence Force Field) [41], which has been imported from previous work on a torsion-driven sampling tool [37, 38] (XXX 2*2 different ref XXX) and,

- AMBER/GAFF, recently implemented and exclusively used in the herein reported results. AMBER [4] is a widely used force field for decades to simulate biological macromolecules such as proteins and nucleic acids. Recently, its authors published an extension thereof, to handle small compounds such as ligands [5]. This new force field is called GAFF and its accuracy level is comparable to other small-compounds force fields such as MMFF94 [42] or Tripos FF [43].

2.2.1 Continuum solvent model and Contact terms

In addition to the classical *in vacuo* force field terms, S4MPLE includes a specific continuum solvation term of maximal simplicity, in order to minimize its computational cost. Explicit solvent boxes, a wide-used option in molecular dynamics, are not compatible with evolutionary sampling procedures supporting large-scale random jumps in conformational space.

S4MPLE solvent model terms include simple functions of inter-atomic distances, of the same complexity as usual force field terms:

- A pair-based desolvation term (see Eq. 1) [6], function of partial charges Q and distance d between two atoms i and j, and scaled by a generic constant σ (more about this in the force field fitting paragraph §2.8)

$$Edesolv_{ij} = \sigma \frac{{Q_i}^2 + {Q_j}^2}{d_{ij}^4} \qquad \text{Eq. 1}$$

- A linearly distance-dependent relative dielectric constant (ε_r) is used in the Coulomb term [44], which *de facto* makes it a function of $1/d^2$, with a practical advantage: avoiding the need to draw the square root of the sum of squares of coordinate differences in the Euclidean formula for non-bonded pairs. Non-bonded terms, including desolvation and a 12-6 Lennard-Jones function, are now functions of d^{2n}

$$ECoulomb_{ij} = \frac{332\ Q_i Q_j}{\varepsilon_r d_{ij}} \qquad \text{Eq. 2}$$

- Contact terms (see Eq. 3) [6], rewarding favorable interactions such as hydrophobic contacts and hydrogen bonds were added to the force field equation. In this context, hydrophobic contacts

refer to close carbons in space, and hydrogen bond donors are hydrogens linked to hetero atoms. Since the same formalism, based on distance only, is used to monitor both these kinds of interactions, it appeared more judicious to select polar hydrogens instead of their porting heteroatoms in the definition of hydrogen bonds. Constant κ_{ij} is a function of the nature of the contact, and the types of involved atoms *i* and *j*, see parameter fitting below. C_{ij} encodes contact strength: full contact $C_{ij}=1$ is assumed at $d_{ij}<d_{min}$. Contact ceases completely at $d_{ij}>d_{max}$, and its strength varies smoothly within the switching range $[d_{min},d_{max}]$. This range is contact type dependent: for hydrophobic contacts, a range of [4.5, 5.5] Å is used. For hydrogen bonds, the range is atom pair specific: [*(sum of Van der Waals Radii)-0.5, (sum of Van der Waals Radii)+0.1*]:

$$Econtact_{ij} = \kappa_{ij} C_{ij} = \kappa_{ij} \left[0.5 + 0.5 * cos\left(\pi \frac{d_{ij}^2 - d_{min}^2}{d_{max}^2 - d_{min}^2} \right) \right] \qquad \text{Eq. 3}$$

Since these terms are not included into the native FF, the additional parameters needed calibration, as outlined in the dedicated chapter (§2.8) below.

2.2.2 Context-specific Termination Function

For each atom pair, featuring a Coulomb term in $1/d^2$, repulsive and attractive Lennard-Jones, desolvation in $1/d^4$, and perhaps a contact term, the distance-dependent contribution asymptotically drops to zero. S4MPLE uses pair-specific cut-off values: cut_{ij} is determined by backwards scanning the distance range, starting from *maxcut=15 Å*, towards ever shorter distances, until the calculated pairwise energy contribution, all terms confounded, exceeds *minPairContrib=0.01 kcal/mol*. This typically happens within the 10…12 Å range for pairs featuring strong Coulomb contribution, down to ~6 Å for Van der Waals-dominated interactions. In order to avoid cut-off artifacts, a termination function $\omega_{ij} = 1 - 2\, d_{ij}^2/cut_{ij}^2 + d_{ij}^4/cut_{ij}^4$ is employed. Therefore, the total pairwise non-bonded contributions can be written as:

$$Enonbonded_{ij} = \omega_{ij}\left(ECoulomb_{ij} + EvdW_{ij} + Edesolv_{ij} + Econtact_{ij}\right) \qquad \text{Eq. 4}$$

2.2.3 Out-of-plane and Chiral Constraint terms

Random jumps in problem space may lead to highly distorted geometries, which may relax by chiral center inversion. Chiral constraints are defined with respect to the initial configuration of chiral carbons, and are always strictly equal to zero unless the chiral carbon geometry closely approaches the highly distorted planar state required for inversion. Since in the current implementation, this is a relative term aimed to preserve the initial configuration, the priorities of the four substituents (***L***owest, ***l***ow, ***h***igh, ***H***ighest) of the chiral center may be arbitrary – they are taken according to the internal atom numbering (see Figure 1). An arbitrary face of the tetrahedron (say *l, h, H*) is selected, and a chiral direction vector, orthogonal to the *(l,h,H)* plane is defined as $\overrightarrow{Hl} \times \overrightarrow{hl}$, for example. Here, the vectors represent relative position vectors of corners *H* and *h* with respect to *l*. Eventually, the position vector of the chiral center c with respect to *l*, $\overrightarrow{cl}$, is computed, and projection of $\overrightarrow{cl}$ along the direction $\overrightarrow{Hl} \times \overrightarrow{hl}$ is estimated as the dot product of these normed vectors:

$$p = \overrightarrow{cl} \cdot \left(\overrightarrow{Hl} \times \overrightarrow{hl}\right) / \left\|\overrightarrow{Hl} \times \overrightarrow{hl}\right\| \qquad \text{Eq. 5}$$

The actual sign of p would reflect the absolute stereochemistry, if priorities *L, l, h, H* would have been assigned according to the Cahn-Ingold-Prelog (CIP) rules [45]. This is irrelevant here, and the absolute value does not matter. The chirality penalty is defined with respect to the original value of p_0 calculated in the initial geometry. It equals zero if $p \cdot p_0 > 0$ (in other words, the sign has not changed, then there is no chirality inversion), but otherwise linearly increases with respect to $|p|$, with a default chirality violation proportionality constant of $k_{chir}=100$.

The same formalism can be used to force planarity around trigonal substituents. In this case, the *L* substituent is missing (no problem, since not explicitly entering the chirality index), and p should ideally equal zero. In the current implementation, no penalty is considered if $|p| < p_{oop}=0.01$, the latter being an empirically chosen allowable out-of-plane deformation threshold. Otherwise, the penalty increases like $k_{oop}(|p|-p_0)$. At this point, a generic $k_{oop}=200$ is in use for all planar centers (including amide nitrogens).

2.3 Potential Energy Surface: Avoiding Singularities

In classical Molecular Dynamic (MD) simulations, singularities of the energy terms in $1/d^n$ at zero inter-atomic are off bounds, because the simulations comply with the energy conservation principle. This is not the case with more aggressive problem space sampling heuristics, such as genetic

algorithms. Therefore, in order to avoid systematic checking for zero distances, the non-bonded squares of distances is implicitly augmented by a small increment *d2offset=0.01*. This has no impact on the precision of these relatively long-range interactions but effectively prevents division by zero errors.

2.4 Hybrid Evolutionary Operators

S4MPLE represents the further development of an evolutionary sampling tool deployed on computer grids, which was however limited to torsional degrees of freedom. Unlike the previously developed rigid rotamer approach, where the vector of torsional angle values assigned to each considered rotational axis naturally formed the "chromosome" of the problem, herein considered full flexibility requires working with Cartesian coordinates of each atom. Or, the matrix x_{ij}, with i being the atom index and $j=1..3$ the coordinate index ($j=1$ means "x", 2 is "y" and 3 is "z") is a poor support for direct crossovers and mutations. Single-point mutations make little sense in this context (changing one coordinate of one atom at a time would lead to a locally distorted geometry of extremely high energy, yet similar to its parent). Plain crossing-over of two sets of Cartesian coordinates is bound to lead to physically impossible geometries, because, unlike torsional angle vectors, these are not invariant to roto-translation. Therefore, adapted genetic operators had to be developed.

2.4.1 Atom Flexibility Status.

By default, all atoms are considered flexible. An explicit list of fixed atoms – for example, binding site atoms – needs to be provided. S4MPLE starts by checking for rotatable (singe exocyclic) bonds, in order to break up the molecule into fragments, which constitute the "operands" on which the generic operators (see §2.4.4 and §2.4.5) will apply. By default, a minimal size of five atoms is required per fragment: $–CH_3$ will not be accepted as fragment, *i.e.* the program will not attempt exchanges of $–CH_3$ groups between different conformers, members of the population. Any exocyclic bond (out of which only single bonds, except amide by default, are considered) divides a molecule into two moieties. The smaller (amongst moieties not containing any fixed atoms) will preferentially become the fragment associated to that bond. When both moieties happen to contain fixed atoms, there will be no fragment associated to that bond. Therefore, some atoms may be not part of any fragment, without being fixed. Let us refer to these as "passive" moieties. Passive moieties are not fixed: albeit their internal degrees of freedom are not being explicitly sampled, their geometry may nevertheless change – during gradient optimization, for example.

A ring system will typically count as a single fragment, and intra-cyclic bonds are not considered as recombination points either. This behavior can be changed by formally declaring one of the ring bonds as “broken”. Using this trick, S4MPLE will specifically ignore a given bond during the fragment list build-up, but its harmonic FF contribution is not modified. The consequence is the ability to explicitly sample different ring conformations.

Disjoined molecular graphs void of any fixed atoms are considered as stand-alone molecules (e.g. ligands). They will be rendered as “non-covalently connected” fragments, assuming one of the putative favorable inter-fragment contacts to play the role of connector instead of the covalent bond – see further on.

2.4.2 “Lamarckian” Local Optimization

Occasional gradient-based local optimization (a.k.a “Lamarckian” local optimization), during the evolutionary sampling procedure, is mandatory for molecules. This is due to the extreme ruggedness of the energy function. An atom misplaced by as little as 0.1 Å may cause an energy excess of hundreds of kcal/mol, causing that otherwise “perfect” geometry to fail in Darwinian selection. It is unwise to wait for a very long time until an appropriate mutation alleviates the bad contact, when few steps of gradient-based optimization may instantly solve the problem. All genetic operators will include, as a last step, a random number (between 20 and 50) of conjugate gradient (CG) relaxation iterations. These procedures will be generically referred to as LO (“Local” – or “Lamarckian” – Optimization).

2.4.3 Exhaustive Energy Minimization

This operator is a succession of various descent methods (SD, CG, BFGS), alternatively applied to (a) the actual energy and (b) to a modified energy landscape, with the weight of the bond stretching contributions downscaled from the default 1.0 to some random value within (0.0,0.2). Softened bonds allow for a temporarily less rugged potential, with respect to which gradient-based methods might bypass some minor energy barriers, thus escaping some irrelevant local minimum. The following minimization cycle is performed after restoring nominal bond strength, until it is observed that bond softening/retightening did no longer allow a move towards any better local minimum. This approach is not routinely used as a genetic operator, but typically employed to refine so-far best sampled geometries before output, or at post-processing stages.

2.4.4 Cross-Over Operators

In the following, molecule crossover, producing a chromosome child:=**CROSS(**$parent_1$,$parent_2$**),** shall refer to a generic procedure taking two mating individuals as arguments, and randomly calling one of the two recombination operators designed here: *fragment recombination*, preferentially used in 80% of cases, and the alternative *uniform torsional crossover*. Eventually, it performs a gradient relaxation as mentioned in §2.4.2.

- ***Fragment recombination*** randomly picks a pair of complementary molecular fragments (**F,f**) defined at the beginning of the simulation. Such fragments may correspond either to two radicals (**F** containing more atoms than **f**), covalently connected to each other by a rotatable bond (see
- Figure 2), or a loose ligand **f** moving freely with respect to the *receptor* moiety **F**. In this way, folding and docking, the two key applications typically dealt with by different software suites, are here unified. The fragment recombination takes place in several stages. First (step a in Figure 2) the geometry of **F** (x_{ij}, atom $i \in$**F,** coordinate $j=1..3$) is taken from the first mate, whereas the geometry of **f** is imported from the second, as **f'**. Next (step b), **f'** will be reconnected to **F,** *i.e.* the imported coordinates (x_{ij}, $i \in$**f'**) are submitted to a roto-translation meant to replace them in a chemically meaningful way with respect to **F**. For bound fragments this means restoring valence bond length and angles around the cut bond to chemically acceptable values.

  For loose ones, the role of the chemical bond to be restored is formally assigned to a randomly picked hydrophobic contact or hydrogen bond involving one atom from **F** and another from **f'**. By default, S4MPLE would randomly try to realize any of the putative favorable contacts between them. Practically, however, when **F** is a large biomolecule, it will offer very many putative contact partners. Unbiased browsing through all these possibilities to suggest possible ligand placements (all around the protein surface) – would however represent a huge waste of time. Therefore, a **hot_spots** file enumerating the eligible contact partners of the active site (preferentially at the cleft bottom) is used to proactively orient the ligand towards the binding cleft. All carbons, acceptors and donors in the smaller "ligand" moiety count as putative contact atoms. First, a putative pair of matching contact atoms **a**$\in$**F** and **a'**$\in$**f'** (hydrophobe-hydrophobe or acceptor-donor) is selected. Next, for both **a** and **a'** S4MPLE tries to find a random solvent-accessible point on each of the contact spheres surrounding these atoms at Van der Waals plus solvent probe radius (here 1.1 Å, slightly less than the water probe radius of 1.4 Å: slight clashes may be tolerated in initial poses). If detection of such an accessible point fails, for either **a** or **a'**, then the pair (**a**, **a'**) is dropped and the algorithm then tries to pick another – until all the 50 pair picking trials are exhausted. Suppose

the selected accessible points were **P** and **p'** respectively. Partners are then brought within contact distance, *i.e.* the atom **a'** is placed at **P**. Next, **f'** is rotated in order to render vectors **a'-p'** and **a-P** antiparallel – meaning that **f'** is being tentatively kept, as much as possible, outside of the exclusion volume of **F**.

The above-mentioned constraints for rearrangement of **f'**, for both fragment junction types (covalent and loose), do not completely define the new geometry of the child structure. In particular, **f'** is free to rotate around the newly formed **F-f'** bond (or contact axis). This torsional degree of freedom may be fixed according to two alternative fragment fitting procedures:

- *DockC* randomly rotates **f'** around the **F-f'** axis, either until a clash-free arrangement is obtained, or a maximal number of attempts *(50)* is reached.
- *DockE* pursues random rotations while attempting to minimize the inter-fragment interaction energies. It stops and returns the most stable pose found so far if *20* successive attempts failed to discover any better one.

DockE, the more resource-consuming of the two approaches, is randomly called in 30% of cases. Eventually, the geometry returned by *DockC/DockE* is submitted to LO. LO may dramatically modify geometry, and therefore the genetic material of parents is not necessarily preserved in the child.

- ***Uniform torsional crossovers*** browse through the list of rotatable bonds, randomly pick one of the two parents as "donor" and set the associated torsion angle value in the child geometry equal to the one of the donor. This is a usual type of crossing-over operator. This approach does not operate on valence angle/bond lengths (these are inherited exclusively from the first mate), nor on the relative positioning of loose fragments. A LO step completes the procedure.

2.4.5 Mutations

The Mutation operator child:=**MUT**(parent) first randomly picks a pair of fragments (**F,f**), then checks whether these are bound or loose fragments. In the first case, a torsional angle associated to the inter-fragment bond is forced to change, either randomly or by means of a temporary constraint term in the energy function, followed by gradient optimization of geometry on this perturbed energy surface ("driven" mutations as reported previously [2]). Otherwise (loose fragment case), mutation is performed as a crossover of the molecule with itself. In other words, there is a repositioning of **f** with respect to **F**, allowing new inter-fragment contacts to be established, and leading to novel geometries.

2.4.6 Initialization

Random Initialization **RI** consists in scrambling the current geometry of the molecule object in memory by random rotations around torsional axes. Next, each fragment is tentatively rearranged in a clash-free manner (*DockC, previously described in §2.4.4*). A LO step completes the procedure.

2.5 Population Diversity Control: Interaction Fingerprints

Any population-based heuristics is strongly tributary to a population diversity control mechanism. In absence of such, the risk of premature convergence is very important (via accumulation of minor mutants of a dominant, relatively stable individual corresponding to some local energy minimum). S4MPLE adopts the postulate that two geometries may be considered redundant if they share a same set of contacts. This postulate is embodied by novel, fuzzy and differentiable Pairwise Interaction Fingerprints (PIF). Unlike classical ligand-protein IF used in docking [46-48], PIFs are

- general, regrouping both intra- and intermolecular favorable contacts: hydrogen bonds and hydrophobic contacts
- symmetry-compliant, *i.e.* invariant to swapping of the contact status of topologically equivalent atoms: rotation of 180° of a carboxylate group having one of its oxygens acting as acceptor in a hydrogen bond will not change the fingerprint, as the hydrogen bond now involves a different, yet topologically equivalent oxygen.
- smooth and differentiable, rather than binary: contact status varies smoothly between "absent" (0) and "fully established" (1.0) as the corresponding contact distance scans the switching range $[d_{min},d_{max}]$.

Building this fingerprint, for a given molecular geometry, relies on preliminary atom typing work (done at the molecule input stage). This includes, first, detection of hydrophobes (carbons), hydrogen bond donors and acceptors. Assignment into the latter two categories is based on the AMBER/GAFF force field types:

- h-bond donors : H, HO, HW, hn, ho, hw
- h-bond acceptors : O, O2, OW, OH, OS, NB, NC, NY, o, oh, os, ow, n1, n2, n3, na, nb, nc, nd, ne, nf

Atoms of types others than listed above or amongst the considered classes of hydrophobes (detailed in Table 1) are not accounted for in PIFs.

Next, "symmetry sets" S_k of topologically equivalent atoms are built – let their number be $k=1..N_S$, while N denotes the total number of atoms. Atoms are topologically equivalent if they are of the same

element (same mass M_i), were assigned the same partial charge Q_i and have identical values of their below-given topological indexes. These indexes (I_i^1, I_i^2) capture the environment of the atom *i* in the molecular graph (topological distances, *i.e.* the shortest path-based number of bonds between, *i* and *j*, are labeled T_{ij}). For generality, atoms from disjoined moieties (*i.e.* i in ligand, j in the protein) are set to "infinity" (practically, $T_{ij}=999$).

Note that these symmetry classes are also employed by S4MPLE to calculate symmetry-compliant root-mean-square-deviation (RMSD) values of atomic coordinates (not further detailed here – additional information available upon request). See Appendix 1 for the pseudo-code describing the symmetry set build-up.

Figure 3 exemplifies the mapping of monitored contacts in the fingerprint.

Note that any member atom *i* of a set *S* is equally distant, in terms of topological distance, from any other atom *j* member of a set *s*. Therefore, the topological distance *T* between symmetry sets can be defined as:

$$T(S,s) = T_{ij} \quad \forall\, i \in S \; and \; \forall\, j \in s \qquad \text{Eq. 6}$$

The contact fingerprint is then built with respect to all pairs of sets (*s,S*) where either both *s* and *S* are sets of hydrophobic carbons, or *s* are acceptors while *S* are donors. Only pairs of sets with at least one free atom, and not topologically too close (*i.e.* always close in space), are selected: $T(s,S)>5$. Therefore, the fingerprint dimension N_F equals to the total number of set pairs above. The PIF is thus defined as a vector of $k=1..N_F$ real values, in which every element PIF_k monitors the contact strength associated to a pair (s_k,S_k) of symmetry sets fulfilling the above condition.

Element PIF_k is thus a function of coordinates of a variable number of atoms $i \in S_k$ and $j \in s_k$. Pairwise contact intensity C_{ij} for each atom pair *(i,j)* can be inferred from Eq. 3. Then, PIF_k could be chosen to equal the *maximal* of all possible pairwise C_{ij}. Unfortunately, this would not be a differentiable function with respect to the coordinates of involved atoms. Therefore – except for the trivial case when $PIF_k = 0$ because all $C_{ij}=0$ – the "self-weighed" average of contacts is used to define *PIF*:

$$PIF_k = \frac{\sum_{i \in S_k} \sum_{j \in s_k} C_{ij}^2}{\sum_{i \in S_k} \sum_{j \in s_k} C_{ij}} \qquad \text{Eq. 7}$$

A geometry can thus be characterized by its interaction fingerprint, and fingerprint dissimilarity scores can be used to assess geometry redundancy.

The main application of PIFs is the geometry redundancy check. Used by the default evolutionary strategies, this basically amounts to calculating the fingerprint size-relative block distance $\delta(g,G) = \sum_{k=1}^{N_F} \left| PIF_k^G - PIF_k^g \right| / N_F$ between two geometries *g* and *G*, each encoded by a chromosome in the population. This block distance represents a generic fraction of contacts having different status in *G* and *g* respectively. A user-defined *minfpdiff* cutoff (typically 0.01 for rigid site docking) is used to define geometry redundancy: *g* and *G* are redundant if the status of more than 99% of the theoretically possible contacts monitored in the PIF is unchanged.

The degree of originality $\omega(G)$ of an individual *G* is defined as the smallest $\delta(g,G)$ between *G* and any other *fitter* individual *g*. In other words, if *G* and *g* encode near identical geometries, but geometry g is energetically more stable, then G alone will be assigned a low originality score and potentially replaced by the diversification strategy. Two distinct routines called **ReplaceMostRedundant** and **RandomizeIfRedundant** are in charge of population diversification, controlled by *minfpdiff*.

ReplaceMostRedundant allows an offspring *O* with a high degree of originality to potentially replace one of the redundant individuals in the population. The eligibility of *O* as a potential replacement increases with its originality: if $\omega(O)>minfpdiff$, the child *O* is certain to participate, rather than strictly compete against its parents. The most redundant member (*MR*) of the current population is determined as follows:

- first, $\omega_{min}=min[\omega(M)]$ – the minimal degree of originality over all members *M* of the population – is found. *MR* is taken to be the least fit individual among the least original members M (those with $\omega(M) \leq 1.1\,\omega_{min}$).
- An empirical trade-off between diversification and acceptation of less fit individuals is achieved: individual *O* replaces *MR* if its excess energy *E(O)-E(MR)* in kcal/mol is "compensated" by the gain in diversity, which is empirically converted into energy units by means on an user-defined *div2E* parameters. If $E(O)-E(MR) \leq div2E[\omega(O)-\omega(MR)]$ and $\omega(O)>\omega(MR)$, then the individual *O* replaces *MR*, and an offspring is accepted in the population and coexist with its parents. By default, children only replace their parents.

RandomizeIfRedundant can be applied to each member of the population, once per generation – after mutations, crossovers and replacements have updated the population. Depending on the degree of originality $\omega(I)$ of the current individual *I*, of energy *E(I)*, this will be considered for randomization if $rand()<\omega(I)/minfpdiff$. Attempts to randomly generate a new individual *M* of energy *E(M)* and originality $\omega(M)$ are made, until

- *M* represents an acceptable trade-off between fitness loss *vs.* originality gain with respect to *I*: $E(M)-E(I)<div2E[\omega(M)-\omega(I)]$, or
- the allotted number of attempts (5) is reached.

2.6 Evolutionary strategies

Several evolutionary strategies were built on the basis of the above operators – three of which are described here. They are based on a set of standard procedure calls (described above as genetic operators). The pseudo code from Appendix 2 depicts a standard genetic algorithm (SGA), while the code from Appendix 3 depicts a simple evolutionary procedure ("evol"). The term **MightReplace** stands for the standard challenge of the parents by the offspring. If an offspring is not original enough to enter the population by replacing a redundant individual (see **ReplaceMostRedundant** in §2.5), it may challenge its parents on basis of its fitness score alone: if fitter, it will replace its parent with the worst fit.

Alternatively, an elitist strategy ensures the survival of the most fitted individuals in the next generation. The considered size of the elite subpopulation is 20% of the entire population. At each generation, the population is ranked based on the energy function and the best fitted chromosomes are declared elite, and cannot be modified or replaced. The general scheme of the considered Elitist Genetic Algorithm (*EGA*) is the same with that of SGA with the main difference that an elite individual will not be replaced by any of the *ReplaceMostRedundant* or *RandomizeIfRedundant* procedures. The elite subpopulation is recomputed after the diversification stage, as the last stage of the generation loop.

2.7 Datasets

2.7.1 Astex/CCDC clean subset

Herein, a further subset of the so-called "clean" Astex/CCDC subset [7], containing 191 complexes has been considered (see Supplementary Material) according to several additional filters:

- No covalently bound ligand
- No complexes where the symmetry-related units are mandatory to explain the binding mode (as seems to be the case, among other, for 1ETA or 3CLA – check performed using the "symexp" command in Pymol [49]).

2.7.2 Astex Diverse set

This [8] is a compilation of reliable and diverse complexes for the modeler community. The 85 selected PDB [50] complexes concern targets of agro/pharmaceutical interest. The X-ray complexes possess high resolution, and feature no clashes. A particular attention has been given to the quality of the electron density around the ligand, not including complexes with high nominal resolution but a lack of density for the ligand. Unlike in the previous set, all the ligands are drug-like.

2.8 Force Field Parameter Tuning Protocol

The calibration of the additional force-field terms outlined below started from a reasonable initial estimation thereof. These were not obtained by systematic scanning, but seemed to be reasonable choices in previously peptide folding experiments mentioned in [1, 39].

As described in the introduction, S4MPLE is based on an extension of the classical AMBER/GAFF force field. Additional terms, introduced above, involve several new tunable parameters:

- *epsilon*: proportionality constant of the distance-dependent relative dielectric constant $\varepsilon_r = epsilon \times d$.
- *repulsive_factor*: the weight of the Van der Waals repulsive term. Therefore, $EvdW_{ij} = repulsive\_factor \times A_{ij}/d_{ij}^{12} - B_{ij}/d_{ij}^{6}$, where A_{ij} and B_{ij} represent the default AMBER/GAFF Van der Waals pairwise interaction coefficients. At a default value of 1.0, original AMBER/GAFF Van der Waals contributions are returned (at $d^2 >> d2offset$, see §2.3)
- *vicinal_weight*: for vicinal (three bonds apart) pairs *i,j*, the entire Van der Waals term $EvdW_{ij}$ is scaled down by *vicinal_weight*, a "trick" already employed in the original AMBER in order to correct the systematic errors due to using a same Van der Waals functional form for both short and long-range interactions.
- *desolv_factor* : the value to be assigned to the desolvation parameter σ from Eq. 1, for all the pairs *i,j* except the special cases outlined below:
- *minq_to_desolv*: the minimal charge threshold for the desolvation term (*i.e.* if $max(|Q_i|,|Q_j|) < minq\_to\_desolv$ then $\sigma = 0$).
- *desolv_scale_ion* and *desolv_scale_hb*: the value of scaling factors for some specific desolvation terms (pairs involving ion or hydrogen-bonds). These extra parameters were required in order to achieve successful docking predictions – a preliminary attempt to fit a set of parameters including only *desolv_factor* to control desolvation failed (results not shown). They are not *a priori* introduced terms, but were added in order to escape a fitting bottleneck.

- *hbond_bonus*: the common value for all κ_{ij} in Eq. 3 corresponding to hydrogen bond interactions.
- Hydrophobic constants K_c for each hydrophobic carbon class (see Table 1), introduced in order to define hydrophobic contact constants κ_{ij} in Eq. 3 as the *average* of constants of the involved carbons.

Therefore, three different force field setups can be defined (Table 3). The first, termed "Core" FF, simply represents the default vacuum AMBER/GAFF terms. The second, "Preliminary" FF assumes above-mentioned estimates for the parameters of the additional terms, with a single weight ($K_{polarized}$=K_{arom}=K_{aliph}=0.1) for all carbons types, and no scaling of the desolvation term (*desolv_scale_ion*=*desolv_scale_hb*=1.0). Eventually, the "Fit" FF is the result of a two-stage fine tuning of some of these parameters, as detailed below:

- First, tuning of the desolvation parameters (*desolv_scale_ion* and *desolv_scale_hb*), based on the observations that the "Preliminary" set-up leads to both (a) bad coordination between ligand and metal ion for several metallo-protein complexes (e.g. mono-dentate coordination preferred to bi-dentate coordination for the acid group in 1CBX), and (b) low quality hydrogen bonds (e.g. angle D-H..A) for complexes involving sugars. Thus, the idea to specifically rescale desolvation term of the above-cited classes of interactions. Associated weights were fixed by trial-and-error docking simulations of concerned protein-ligand complexes.

  Last, a systematic scan of hydrophobicity parameters associated to considered classes of carbons was performed. The rather numerous carbon force field types (33 = 16 from AMBER and 17 from GAFF), were regrouped into 3 different "hydrophobicity" classes: aliphatic, aromatic and polarized, each being attributed a hydrophobicity constant K_c (see Table 1). A total of 10 discrete values (0; 0.010; 0.025; 0.050; 0.075; 0.100; 0.150; 0.200; 0.250; 0.300) are explicitly considered as possible K_c choices, allowing for 1000 combinations to scan. The scan was bound to highlight a triplet which systematically ranks, for a large majority of Astex/CCDC-clean complexes, native-like poses as lowest-energy states, out of large and decoy-rich sets of poses. Pose sets were generated by S4MPLE, in iteratively repeated runs employing various FF parameterization schemes (see Figure 4). In practice and for each complex, the pool of poses is rescored using the fitted parameters and one of the 1000 combinations. The top-ranked pose is saved for each complex and all tested combinations, thus it is possible to extract triplets which most frequently favor the expected binding mode.

Validation of the herein obtained "Fit" FF configuration was done by docking of Astex Diverse ligands, and assessment of redocking success.

2.9 Redocking protocol using S4MPLE

The preparation of the ligands consists of several steps:

- computing partial charges (Gasteiger type) using ChemAxon libraries [51]
- adding GAFF atomic types using the Antechamber tool [40, 52]
- using the Parmchk tool [5, 40] to check whether there are missing parameters (*e.g.* bonds, angles or torsions). In that case, Parmchk computes the missing parameters using empirical rules, and these new parameters are added to their respective force field parameters files
- generating a single conformer using ChemAxon libraries (avoiding to start from the expected solution, when the docking accuracy may be artificially enhanced)

Binding sites are prepared using MOE and its Protonate3D protocol [53]. Partial charges and atom types are assigned from the specific AMBER topology file during the initialization of the program.

As mentioned before, S4MPLE does not request any formal definition of the binding pocket. Local protein subdomains, including only residues with at least one atom at 10 Å or less with respect to known ligand(s), are used for docking. These are cut out of the PDB structures, after the hydrogen assignment step in MOE. Usual redocking protocols often use a value of 6.5 Å [54-56]. Here, larger binding domains are employed, in order to maintain compatibility with the classical FF cut-off magnitudes (9…12 Å). S4MPLE relies on a list of protein atoms to be preferentially used to anchor the ligand, by randomly positioning it such as to establish favorable contacts with these listed "hot spots". Hot spots were automatically picked, by detecting the putative hydrophobic contact and hydrogen bonding centers in the close neighborhood of the binding subdomain center (no biased choice of protein groups involved in actually observed contacts). Explicit removal of remote protein moieties is however a potential source of artifacts, given the otherwise unrestricted ligand mobility: this may be pushed out of the site, and led to form "favorable" fake contacts with – in practice – inaccessible protein atoms. Therefore, ligand poses having their geometric centers at more than 8 Å away from the center of the binding subdomain are systematically discarded in order to facilitate *a posteriori* analysis of results.

Co-factors (prepared/parameterized like ligands, but maintaining their experimental geometry) and ions are included in the binding site, and all waters are removed. During the redocking benchmarks, all binding site atoms are considered as fixed.

Evolutionary algorithms involve numerous parameters: size of the population, number of generations and specific parameters related to genetic operators (including the frequencies of their call). In addition, the similarity threshold (*minfpdiff*) defining conformational redundancy needs to be set. The redocking simulations consisted in 10-fold runs of Astex Diverse Set complexes. Each run took 500 generations of 50 individuals using the "evol" strategy, mutation probability of 1/10, crossing-over probability 1/10, and *minfpdiff*=0.01. The number of generations has been set up on the basis of preliminary redocking runs involving both small and large ligands. Sampling success (ability to save a native-like pose using the RMSD metric) has been monitored in function of the number of generations (see

Figure 5) and seen to reach a plateau around 400. Thus, the default number of generations has been set up to a slightly larger value (500). After actual docking, so-far visited poses are subjected to a filtering step, involving selection of a non-redundant poses and further optimization (using the exhaustive energy minimization strategy defined in §2.4.3) of the best poses (potential energy within +30 kcal/mol with respect to sampled top conformer).

Success rate (percentage of correctly predicted complex geometries, at given RMSD level) is then reported as the *average* of success rates of each independent run. In other words, all the 10 independent simulations are monitored independently: if, for a same ligand, 6 out of 10 succeeded and 4 failed, six successes and four failures will be counted. The final success average equals to the sum of successes/(10 x number of benchmarked complexes).

2.10 Redocking protocol using FlexX

This was carried out for benchmarking purposes, in order to compare the behavior of S4MPLE with respect to an available, state-of-the-art docking tool. The program LeadIT (version 2.0.2), developed by BioSolveIT, is used and includes a version of the popular docking tool FlexX [10]. Protein site files prepared by MOE as mentioned beforehand lead however to unexpectedly low results in FlexX simulations (results not shown). In particular, the relatively large definition of the binding site seemed to prevent FlexX from efficiently discovering native poses, as no docking constraints were imposed (e.g. avoiding poses near remote residues). Therefore, a FlexX-specific site preparation protocol was adopted. As with MOE, cofactors and metals located near ligand of interest were preserved, and all water molecules were removed. Hydrogens were however added within LeadIT, and adjustments of residues protonation states and orientation of polar hydrogens were manually corrected if necessary.

FlexX binding sites include all complete residues with at least one atom within a distance of up to 6.5 Å with respect to the reference ligand. Default values for all options relative to ligand preparation are used, except for the auto-assignment of protonation states. These are the same as those used in S4MPLE runs. Pharmacophoric restraints are disabled for metals in order to avoid putative biases during the docking process. Default FlexX docking options are used.

3 Results & Discussion

3.1 Force Field Fitting

3.1.1 Tuning of Desolvation Contributions

The Preliminary FF configuration was problematic in modeling of some complexes featuring strong polar interactions, where the Core FF successfully converged toward the expected solution. Bad coordination between ligand and metal ion for several metallo-protein complexes (e.g. mono-dentate coordination preferred to bi-dentate coordination for the acid group of the ligand in 1CBX complex), and low quality hydrogen bonds (e.g. bad D-H..A angle values of ~120° in hydrogen bonds) in complexes involving sugars (1ABE, 1ABF) are examples of such failures. Obviously, the desolvation term at preliminary setup (σ=0.1) is too high, cancelling out the benefic electrostatic terms at very low distance (hydrogen-bond case). Besides, dielectric solvent models are notoriously challenged by heavily charged ions, known to cause dielectric saturation phenomena [57]. Therefore, a specific treatment of the desolvation terms involving favorably interacting highly charged partners appears as necessary. Two independent scaling constants *desolv_scale_ion* and *desolv_scale_hb* – for pairs including a metal cation and for hydrogen bonding partners (specifically for the polar H/acceptor pair) – were introduced. Their values were allowed to scan the range from 1 to 0, and a consensus over selected complexes emerged for values of 0.1 (data not shown). Scaling down, by a factor 10, desolvation contributions of metal ions restored the expected bi-dentate coordination in 1CBX. The same holds for hydrogen bonds: scaling down the desolvation term of hydrogen-bond pairs improved predicted geometries. Note that acceptor-acceptor or polar H-polar H desolvation terms are not being scaled down.

3.1.2 Optimizing Hydrophobic Contact Strengths

Over the 1000 combinations ($K_{polarized}$, K_{arom}, K_{aliph}) of hydrophobic factors, the best result at training stage was to obtain 154 correct (top-ranked pose at RMSD<2.0 Å) out of 191 Astex/CCDC complexes (80.6% success rate). This top result was independently achieved with 11 different combinations. All except one coherently come from a same zone of the combination space: negligible weights for polarized carbons ($K_{polarized} \approx 0$), contrasting with high aromatic and hydrophobic weights ($K_{arom} \approx K_{aliph} \approx 0.15$). The outlier (0.2, 0.15, 0.15) is atypical because it considers polarized carbons as the most hydrophobic – also, the 154 complexes being well-scored by it significantly differ from the consensual ones retrieved at all other, physically more meaningful, setups. Note that the scanned K_C range properly encompasses the optimal range: at extremes (0, 0, 0) and (0.3, 0.3, 0.3) success rates plummeted to 145 and 142/191 respectively. The absolute worst combination (0.2, 0.1, 0.3) only leads to 138 well-predicted complexes. Eventually, the kept combination used in the "Fit" force field was taken at the core of the region: taking the median over listed values for each parameter in Table 2 highlights the selected triplet (0.01, 0.15, 0.15). Finally this combination appears consistent since it is also among the best ones with respect to the stricter success criterion RMSD≤1.5 Å (74% success rate). The Table 3 lists all parameters and their final values for the Fit FF setup.

3.2 Redocking - Astex Diverse set

Both Core and Fit FF setups are challenged to dock validation set complexes (Astex Diverse Set [8]). The Fit FF obtained an excellent 85% success rate at RMSD<2.0 Å for the top ranked pose (average over 10-fold independent runs). The results of redocking simulations are summarized in both Table 4 and

Figure 6. This accuracy is equivalent to those of "state of art" docking tools [8, 32, 58-60]. The lower success rate of RosettaLigand may be due to therein considered protein flexibility.

It is clear that force field tuning was important: there is a significant increase in accuracy of the Fit FF (85%) with respect to the standard vacuum AMBER/GAFF Core (76%) FF. Another important point is the ability to maintain an excellent/good accuracy with even more stringent RMSD thresholds. Thus, for Fit FF, success rate goes down by only 7% at 1.5 Å (78%) and 16% at 1.0 Å (69%) with respect to the usual 2.0 Å threshold. The ability of S4MPLE to almost systematically retrieve a pose close to the experimental one within the top 30 poses is shown too (see Table 4). Details of each individual run/complexes are available in the supplementary data (e.g. RMSD of top-ranked pose,

lower RMSD within the top 30 poses). Comments about global failure cases, and differences between both force field setups are discussed below in more detail. For clarity reasons, only complexes where Core FF and Fit FF-driven docking runs had different success status are shown. In these cases, there is at least one native-like pose (in complexes with systematic errors, the overlay of the native pose and two different "wrong" poses is too crowded to be readable).

3.2.1 Insights on failure cases

Three different main scenarios may describe docking failures:

1. All-atom RMSD is high, albeit the top-ranked pose is chemically meaningful, reproducing all the key ligand-site contacts. This may happen :

   (a) if the ligand includes a moiety dangling out in the solvent, or

   (b) the ligand is "pseudo-symmetric", in the sense that it features chemically similar groups which could, in principle, compete for a same binding spot.

   In situation (a), the "native" structure of the dangling ligand moiety is likely an over-interpretation of electron density data – or an artifact where crystal packing constrains this otherwise moving moiety. In case (b), say a chlorophenyl-(symmetric linker)-tolyl, the ligand may probably allow for two distinct, comparably populated binding modes (with Cl-Phe and Me-Phe switching binding pockets), whereas electron density fitting would most likely highlight only one of those. Note that S4MPLE pairwise interaction fingerprint (PIF) similarity would, unlike RMSD scores, not consider cases 1.a as docking failures. Indeed, calculated fingerprint would be approximately the same as the native fingerprint, whatever happens to the dangling moiety. Scenario 1.b is a real challenge, and can only be evidenced by visual inspection.
2. Neither the top-ranked pose, nor any other of sampled poses, match the native one. This may be due to :

   (a) insufficient sampling,

   (b) inappropriate definition of the binding site (in redocking, this includes ignoring key waters mediating ligand-site interactions – in cross-docking, the binding site must undergo mandatory conformational changes to accommodate the ligand), or

   (c) a highly unrealistic force field setup, causing native-like structures to be of high energy, thus effectively preventing them from being sampled and saved.

3. The top-ranked pose differs from the native, but the latter is being sampled and listed among less stable ones. This is clearly :

   (a) imputable to the force field setup, counting as a failure to identify the expected binding mode as the absolute energy minimum, or

   (b) inappropriate definition of the binding site (like in 2.b – however, of more limited impact on results since the expected binding mode has been saved), or

   (c) an entropic effect: minima deeper than the native one exist, as predicted by the FF, but are narrow and therefore not significantly populated at room temperature.

   It is not clear at this point how the narrowness of a minimum impacts on its probability to be visited by an EA. According to common sense, hypothesis (c) is unlikely: if such narrow minima exist, their discovery by the algorithm should be an expectedly rare stochastic event. EA sampling does not furnish positive proof for a 3.c scenario, so this explanation should be invoked with great caution.

This being said, convergence towards the native structure as so-far lowest energy pose is not an absolute warranty of success – the FF energy landscape may nevertheless feature deeper fake minima, which were not visited, given the limited sampling effort per individual redocking run. However, the native structure is, at least, an attractor in problem space – an important local minimum, if not the absolute one.

There are some general failures to converge towards low-RMSD poses, irrespectively of the FF setup: 1G9V, 1GM8, 1GPK, 1HP0, 1HVY, 1JJE, 1MEH, 1TZ8, 1XM6, 1YGC and 2BR1. These will be detailed below, and assigned to the scenarios mentioned above:

- 1G9V, 1GM8 & 1HVY ligands have limited direct contacts with the binding site: most of the ligand/site hydrogen bonds are bridged with waters. Worse, these ligands feature solvated carboxylate groups, not directly interacting with the target. In 1HVY, the dihydroquinazoline moiety from drug Tomudex is perfectly docked, whereas the two loose carboxylates are, in absence of crystallographic waters, attracted towards neighboring basic residues. Therefore, direct site-ligand contacts are all well predicted – this fits scenario 1.a. Same holds for 1GM8, the hydrophobic phenylacetamide moiety is correctly positioned, unlike the anionic β-lactam moiety. 1G9V, however, is totally dependent on specific water-mediated interactions, therefore classifying as failure 3.b (the correct pose is being sampled, but not top ranked).
- 2BR1 & 1XM6 are further examples of (3.b) complexes where water-mediated interactions play an important role. By contrast to the examples above, these water-mediated contacts occur deeply

within the binding site. In the top-ranked pose of 2BR1, the ligand is bound "head-to-tail", in burying the methoxyphenyl moiety into a deep pocket filled with water in the X-ray structure. In 1XM6, the location of the propoxyphenyl group is perfect but the oxazolidone moiety is erroneously predicted to directly interact with a Zn^{2+} ion (in the experimental structure, a water molecule occupies this area).

- 1GPK & 1MEH are classical FF failure cases (3.a), native conformations being sampled but not top-ranked. The best poses favor an ionic bond instead of two hydrogen bonds, observed in the PDB file, showing that fine tuning of ionic/polar interactions and desolvation penalty is not perfect. In 1GPK the top-ranked pose is inverted, but nevertheless fulfills expected hydrophobic contacts with the site. In 1MEH, the solvent-exposed carboxylate unsurprisingly prefers the neighborhood of R414, instead of the crystallographic hydrogen bonds to S262. This has a very limited impact on the correctly predicted pose of the buried aromatic moiety.
- 1HP0 is an interesting case, where a non-native conformer with flipped deaza-adenine group is nevertheless almost correctly docked – fulfilling all of the experimentally observed contacts. The ligand adopts a correct binding mode around the calcium ion, and reproduces the stacking interactions of the flipped aromatic moiety. Non-bonded energies are similar, thus the preference for the native conformer should have been played out in terms of intra-ligand strain contributions. This is not happening (the native conformer is visited, but ranked slightly less well). At this point, it is difficult to formally rank this as a force field failure 3.a, rather than a genuine multiple binding mode example 1.b. 1JJE is in a similar situation, featuring a pseudo-symmetric ligand, derivative of 2,3-dibenzylsuccinic acid in which one of the Phe groups ports a -O-CH_2-O- bridge (benzodioxole). The latter makes, however, no additional strong contact to the protein, whereas the central, succinyl moiety coordinating both Zn^{2+} ions is perfectly predicted. It seems that the benzodioxole group may be equally well accommodated on each side of the binding site. Docking does not favor the one expected. It is not clear, however, whether the experimentally observed electron density is witness of a single binding mode which S4MPLE fails to reproduce, or whether it is a statistical expression of a slight preference among two almost equally populated modes. The "head-to-tail" binding of this pseudo-symmetric ligand results in an impressive RMSD of 8 Å.
- In 1TZ8, the native PDB pose surprisingly exhibits several slight ligand-site clashes (e.g. distance L17_CD1--Ligand_C=3.0 Å). Expectedly, even the Fit FF with slightly downscaled repulsive Van der Waals terms, leads to a wrong top-ranked pose – albeit it enumerates the native one. It should be noted that this complex is often reported as a failure [59, 60]. Furthermore, authors of the Astex

compilation dataset reported this complex as borderline in term of quality of the ligand's electronic density [8]. Hence, this case should be most likely classified as 3.b, or less likely as a force field problem (3.a).

- 1YGC: the most solvent-exposed moiety (arylsulphonamide) adopts a wrong alternative conformation (direct hydrogen-bond with the C58 backbone instead of two water-mediated hydrogen bonds with Y94 and T98), whereas the rest of this large ligand is properly docked.

Most of these cases are often described as common failures in other docking papers [8, 59, 60]. However, all had the native pose visited by S4MPLE, albeit not top-ranked – yet, only few amongst the above cases are obviously due to FF errors.

For some other complexes (1KZK, 1MZC, 1UML, 1R58, 1Y6B and 2BM2), the native pose was not systematically visited during each of the 10 runs. However, once found, the native poses turned out to be top-ranked. Since these ligands are large compounds and the best energy over all runs is close to the X-ray binding mode, these failures are the result of a lack of sampling. Hence, these complexes occasionally exemplify the 2.a scenario: 500 evolutionary generations at population size of 50 may not guarantee convergence (albeit the 10-fold repeated runs are eventually sufficient). This fact is not surprising for heuristics-based algorithms.

3.2.2 FF-based scoring: Core Vs Fit

As previously described, a real improvement is observed using the Fit FF. There are 8 complexes for which the fitted scheme converged towards the expected solution, while the Core FF failed, *vs.* only one exemplifying the opposite situation. The below mentioned cases (see

Figure 7 and Table 5) are, unless otherwise noted, representatives of the 3.a-type failures (native pose sampled, but not ranked as most stable) with Fit FF (1OQ5) or Core FF (all the others) :

- 1OQ5: Fit FF has the second-ranked pose matching the expected binding mode, but the most stable one appears clearly non-native whereas Core FF selects the native one. In this Fit FF failure, the ligand is rotated around the pyrazole--benzenesulphonamide axis, with the tolyl group occupying an alternative sub-pocket.
- 1N2J: this small ligand (pantoate) makes four direct hydrogen bonds with the binding site. Unexpectedly, using the Core force field, the top-ranked pose does not make any of these, and places the carboxylate group in the hydrophobic area where the two methyl groups of the rather small ligand are expected. Favorable contact terms (hydrogen-bond and hydrophobic enclosure),

and desolvation (penalizing the burial of a $–COO^-$ in a hydrophobic pocket) in Fit FF successfully restore the correct binding mode at the top of the list.

- 1NAV: the compound is buried and makes several hydrogen bonds with the binding site. The Fit force field finds the native binding mode with a great precision (RMSD≈0.5 Å), whereas the Core FF selects a translated pose (several Å away from the experimental binding mode: RMSD≈6 Å) in order to create an ionic bond between the carboxylic group of the ligand and the solvent-exposed R266. This is another example of overestimated polar contributions in absence of scaled down electrostatics and desolvation.
- 1R58: in that case, both scoring schemes lead to an acceptable solution for buried and metal coordination groups of the ligand, but the chlorophenyl moiety is wrongly docked by the Core FF, which has no particular incentive to flip this group in order to bring in within hydrophobic contact range with Y444, as seen in the experimental structure.
- 1UML: this ligand is located near a Zn^{2+} ion but does not directly interact with it in the X-ray structure. Fit FF leads to a perfect pose, while the Core FF completely misplaces the ligand, in order to allow for an interaction with that cation.
- 1V0P: Fit FF returns a perfect top-ranked solution, in which the carboxylate group of the ligand is solvent-exposed and doesn't make any direct contact with the protein. At the opposite, the Core FF favors a pose where the ligand is less buried, allowing the carboxylate group to make several fake hydrogen bonds with the binding site.
- 1W1P: this complex contains a small cyclic dipeptide (Gly-L/Pro). Although both hydrogen bonds with the site are conserved using the Core FF, the RMSD is high because the ligand is flipped with respect to the experimental binding mode. This is rather a 1.a scenario, in which ligand flipping does not affect observed interaction patterns. Conversely, the Fit FF top-ranked pose matches the exact binding mode.
- 1XOQ: Although two cations are present in the binding site (Zn^{2+} and Mg^{2+}), the drug (Roflumilast) does not directly interact with them in the crystal structure. The Core FF forces a binding mode where the dichloropyridine group of the ligand is close to the magnesium ion, whereas the Fit FF returns the expected binding mode with two hydrogen-bonds with the same sidechain (Q369).
- 1YVF: this case is similar to 1V0P, featuring a solvent-exposed carboxylate, forced by Core FF into several non-native interactions with R394. The best pose from the Fit FF reproduces almost exactly the experimental binding mode.

Generally speaking, Core FF binding modes favored less buried ligand poses, with fake polar contacts. This is coherent with a FF setup overestimating Coulomb terms, ignoring desolvation penalties and not rewarding hydrophobic enclosure.

4 Conclusion

S4MPLE is a general conformational sampling tool, based on evolutionary algorithms. It was designed to support full atom flexibility, and based on a set of powerful general generic operators, including an original conformer population diversity control mechanism, based on differentiable pair-wise interaction fingerprints. As a consequence, S4MPLE sees no fundamental difference between actual sampling of a single molecule, or docking of one – or several – ligand(s) into a target site. This is due to the fact that its genetic operators automatically detect the context in which they are called – recombination of covalently bound or non-bonded fragments – and seamlessly adapt to it. Potential degrees of freedom are automatically detected, but can be controlled by the user: fixing of molecular moieties allows the approach to concentrate on relevant degrees of freedom, be they intra- or intermolecular. Thus, the theoretical applicability range of S4MPLE is broad: from small peptide folding and ligand sampling, to rigid site docking, to classical flexible docking (with moving side chains), to extreme flexible docking (allowing larger movements of protein loops, backbone included), to multiple ligands docking.

Care was taken to adapt the genetic operators to the specific nature of the energy landscape. Herein used "driven" mutations or fragment recombinations are much more likely to produce relevant geometries, by contrast to the typical random tinkering with torsional angle values, bound to lead to clash-rich, impossible geometries.

Pairwise monitoring of the status of all putative contacts (including, unlike in classical IF [46, 48], intra-ligand and intra-site terms – if site flexibility is enabled) fuzzily represents the molecular geometry of the whole system. PIFs are not meant to extract similar binding modes shared by chemically different compounds, albeit this information can be easily mined for, by analyzing the populated PIF elements.

S4MPLE can be employed, in response to the complexity of the considered problem, either as stand-alone single CPU process, or as computer-grid deployed, massively parallel simulation (work in progress).

This first paper, in a series dedicated to S4MPLE, addresses the following main issues:

- The key aspects of its algorithm and underlying force field model – the latter being an extension, including among others a continuum solvent model, of the popular AMBER/GAFF force field.
- Fitting the parameters of the additional force field terms. S4MPLE does not so far feature any scoring function aimed at returning conformational free energies: at this point, energetic stability (potential energy) is the only criterion used to rank the sampled geometries. Based on the typical rigid-docking benchmark sets, parameter configurations, maximizing the rescoring success rate (*i.e.* number of Astex/CCDC clean protein-ligand complexes in which the energetically top-ranked pose was native-like), were sought.
- Validation of above fitted parameters, and in-depth study of rigid-receptor redocking results, on an external set of complexes (Astex Diverse).

It was shown that outfitting the already quite well performing AMBER/GAFF force field with very simple desolvation and hydrophobic contact terms clearly improved rigid-receptor redocking success scores. At a strict count of redocking successes, at a customarily employed criterion of RMSD<2 Å, within the widely used Astex Diverse data set, S4MPLE ranked better or similar to state-of-the-art docking tools. Furthermore, many of the cases counting as failures were shown to be situations in which multiple ligand binding modes cannot be easily excluded (e.g. almost symmetric ligands). Furthermore, all the non-native top-ranked poses correspond (if a sufficient number of generations is allowed) to deeper energy wells compared to the level reached when relaxing the native structure in our enhanced AMBER/GAFF force field. These failure, when not attributable to ignored water-mediated interactions due to deletion of intervening crystallographic waters, outline (expectable) limitations of the force field-based potential energy model. In no instance S4MPLE systematically failed to reach the baseline energy level of the relaxed native structures, showing that its sampling procedures are powerful, successfully avoiding entrapment in local optima, notably using regularly approaches avoiding consanguinity within the population. In respect to these results, the herein developed, solvation-enabled AMBER/GAFF extension should be preferentially used for drug-like ligand docking. It is yet unclear whether the herein introduced FF terms scale up conveniently for use in large-scale conformational sampling, such as peptide folding.

At the moment, S4MPLE is still a prototype under development. The runs were quite time-consuming (several hours/single CPU for the average rigid-site redocking simulation), but

(a) a plethora of technical but important parameters – such as, for example, stopping criteria of local search procedures, *etc.* – were not yet optimized, being set by default to rather strict values.

(b) the main goal behind this development is not to add one more rigid-site docking program to an already long list. So far, the purpose of this development was to assess in how far experimental docking poses can be correctly predicted on the basis of the herein defined energy function (no re-ranking based on fitted free energy scoring functions) if unbiased sampling is allowed.

The use of S4MPLE for classical rigid-site docking could be easily enhanced – for example, by a manual, knowledge-based selection of site hot spots, by contrast to the automatic approach used here. Also, the availability of smooth, differentiable interaction fingerprints allows for straightforward inclusion of problem-specific knowledge (imposing specific contacts to occur on the protein side). However, having passed these first classical tests, further work will address problems of larger complexity – fragment-like compounds docking, multi-ligand docking and binding site flexibility.

5 Acknowledgements

The authors wish to thank the staff of the two computer centers which hosted the simulations: HPC (High-performance computing) of the University of Strasbourg and HPC of the chemistry faculty of Cluj-Napoca. All pictures depicting ligand-protein structures have been created using Pymol [49]

6 Supporting Material

The following:

- x86_64 executable of S4MPLE,
- User guide,
- various ligand preparation tools and
- force field parameter distributions
- spreadsheet with the list of the training set compounds used to calibrate Fit FF, and detailed docking results for each of the repeated docking runs are available for download on http://infochim.u-strasbg.fr, DOWNLOADS section (tar.gz). Docked poses are available upon request (mol2 format)

7 Figures

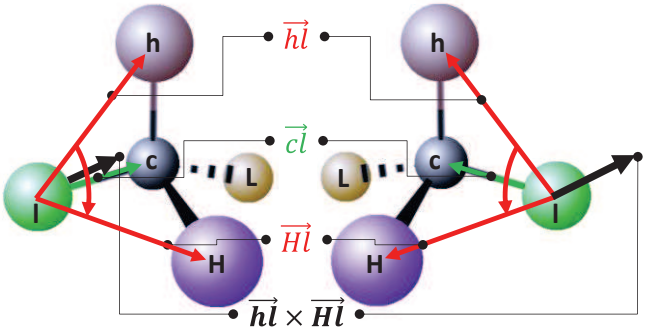


Figure 1: Calculation of the chirality index used to preserve the configuration of asymmetric carbons.

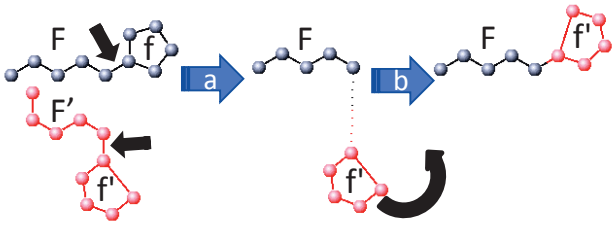
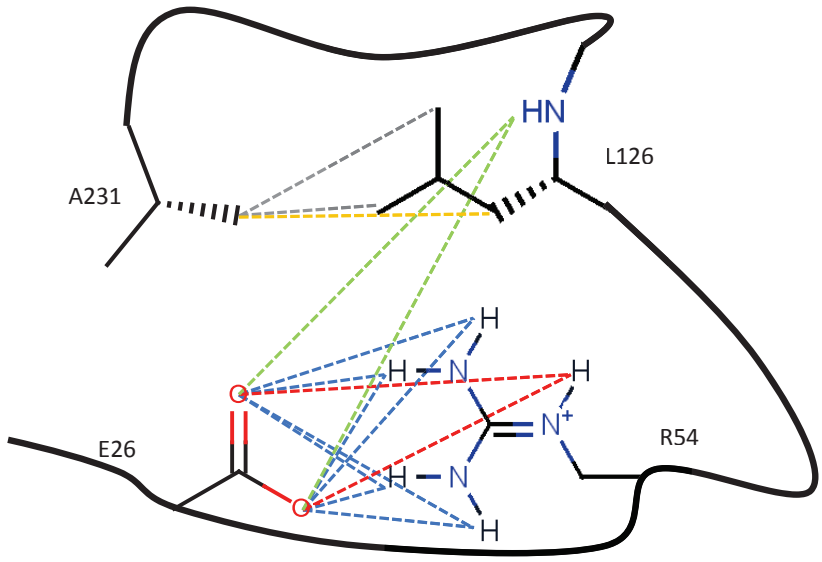


Figure 2: Principle of fragment recombination.



Fingerprint:	1.0	0.1	0.0	…	1.0	0.5
Type:	Hbond	Hbond	Hbond	…	C-C	C-C
Contributing distances:	E26:OE* - R54:HZ*	E26:OE* - R54:HE	E26:OE* - L126:HN	…	A231:CB – L126:CD*	A231:CB – L126:CB

Figure 3: Principle of Interaction Fingerprints: each element represents a unique putatively favorable interaction, which may be embodied by different topologically equivalent atoms. For example, the E26-R54 ionic contact above is considered as “established” (associated PIF element set to 1.0) as soon as either of equivalent R54 HZ are within contact distance of either of E26 OE. Practically, for all distances between equivalent atoms (dotted lines of a same color) the corresponding contact strength are calculated and eventually averaged according to Eq 7. This biased average (favoring strong contact contributions) is reported in the associated PIF cell (color matching distance lines).

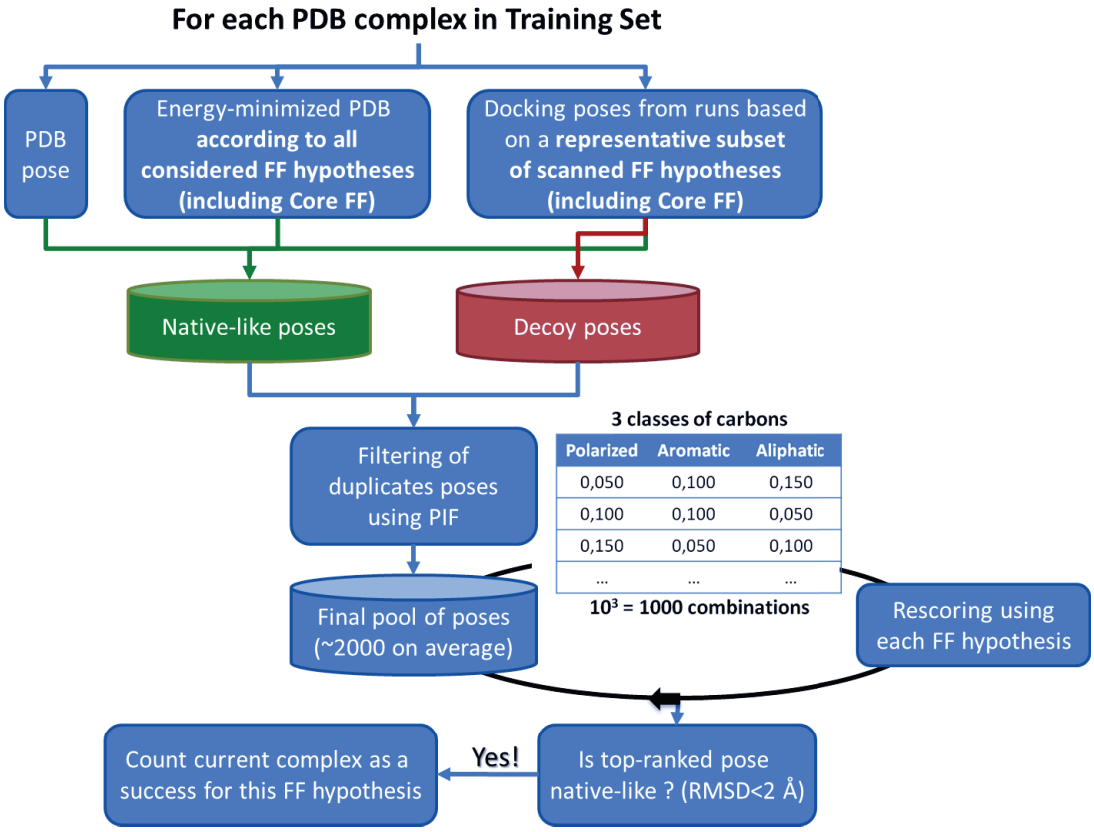


Figure 4: Strategy used in the force field parameter tuning protocol. "Decoy poses" are non-native-like, to be distinguished from the native-like by a proper choice of FF parameters. These three parameters were subjected to a systematic scan, formally covering a "cube" of possibilities in parameter space. Within this parameter space volume, the "representative" subset of possible FF configurations, includes the Core FF and Preliminary FF setups, plus setups at the extremes (corners) and at the center of the scanned parameter "cube". This "representative" subset is used to generate docking poses (one 400-generation run per FF configuration, per complex).

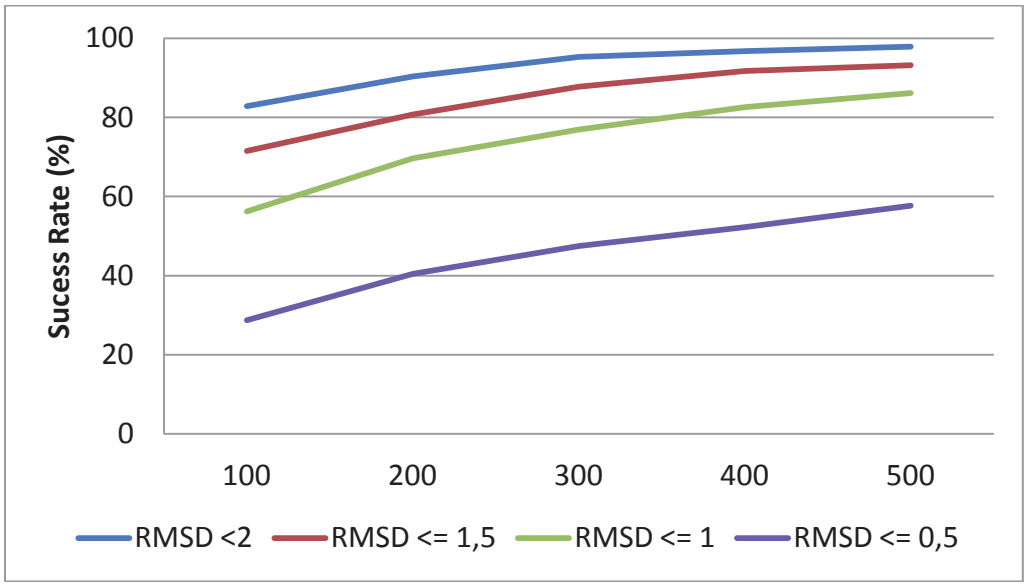


Figure 5: Ability to sample X-ray binding modes (irrespective of their ranking in terms of energy), in function of the number of generations at different RMSD thresholds (dataset : Astex Diverse Set).

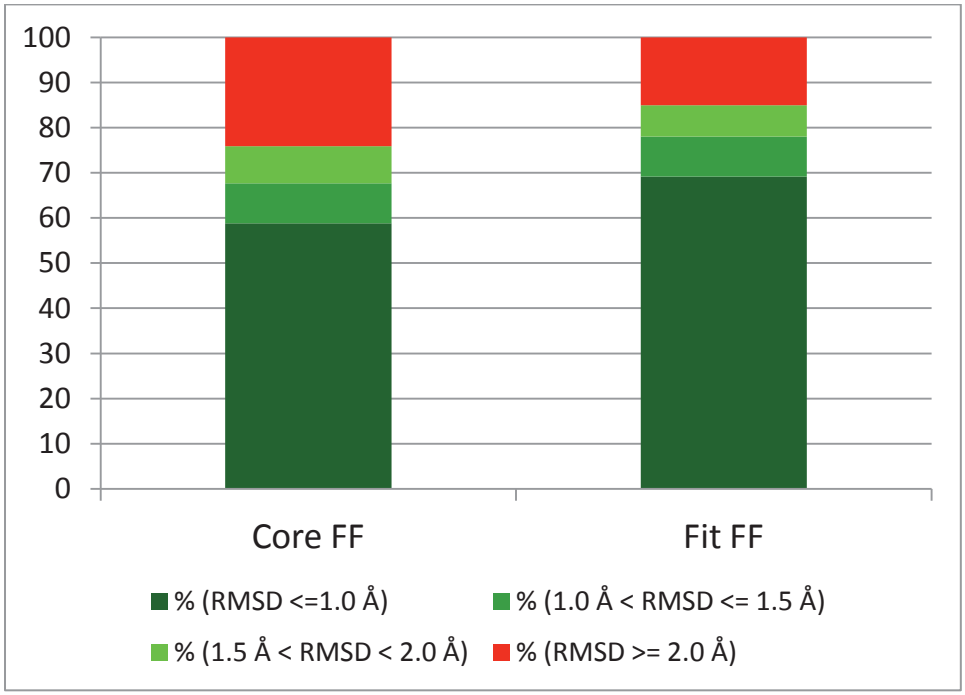


Figure 6: Docking performances on the Astex Diverse Set for both Core FF and Fit FF.

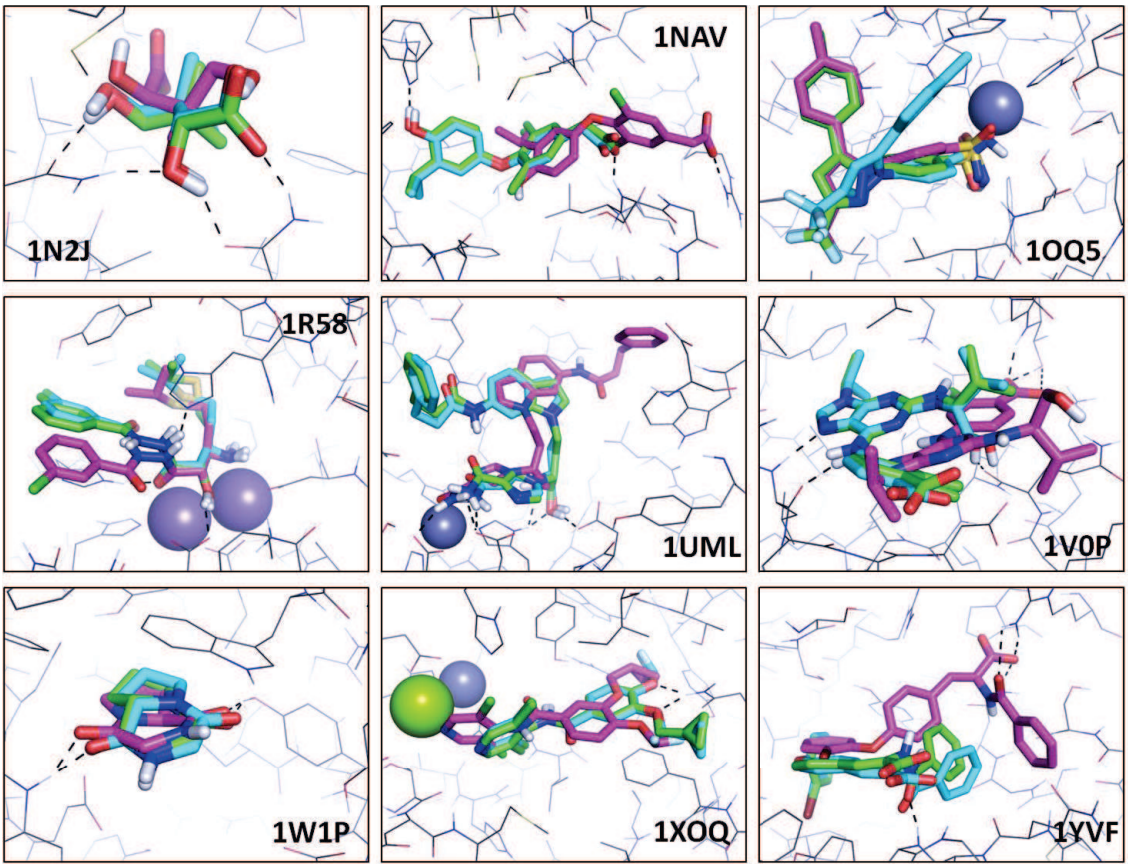


Figure 7: X-ray and top-ranked poses obtained using Fit FF and Core FF from the 9 PDB complexes discussed in the chapter §3.2.2 (Core Vs Fit). Native poses are shown in green, whereas Fit FF poses and Core FF are displayed in blue and purple respectively. Hydrogen-bonds are shown as black dotted line and ions are represented as spheres.

8 Tables

Table 1: Description of the AMBER/GAFF carbon types and their classes.

Category	Force Field	Atomic Type	Description (from AMBER/GAFF parameters file)
Polarized	AMBER	C	sp2 C carbonyl group
	GAFF	c	sp2 C carbonyl group
Aromatic	AMBER	C*	sp2 arom. 5 memb.ring w/1 subst. (TRP)
	AMBER	CA	sp2 C pure aromatic (benzene)
	AMBER	CB	sp2 aromatic C, 5&6 membered ring junction
	AMBER	CC	sp2 aromatic C, 5 memb. ring HIS
	AMBER	CN	sp2 C aromatic 5&6 memb.ring junct.(TRP)
	AMBER	CR	sp2 arom as CQ but in HIS
	AMBER	CV	sp2 arom. 5 memb.ring w/1 N and 1 H (HIS)
	AMBER	CW	sp2 arom. 5 memb.ring w/1 N-H and 1 H (HIS)
	GAFF	ca	sp2 C in pure aromatic systems
	GAFF	cc	sp2 carbons in non-pure aromatic systems
	GAFF	cd	sp2 carbons in non-pure aromatic systems
	GAFF	cp	Head sp2 carbons connecting rings in bi-phenyls
	GAFF	cq	Head sp2 carbons connecting rings in bi-phenyls
Aliphatic	AMBER	CT	sp3 aliphatic C
	GAFF	c1	sp C
	GAFF	c2	sp2 C
	GAFF	c3	sp3 C
	GAFF	ce	Inner sp2 carbons in conjugated systems
	GAFF	cf	Inner sp2 carbons in conjugated systems
	GAFF	cg	Inner sp carbons in conjugated systems
	GAFF	ch	Inner sp carbons in conjugated systems
	GAFF	cu	sp2 carbons in triangle systems
	GAFF	cv	sp2 carbons in square systems
	GAFF	cx	sp3 carbons in triangle systems
	GAFF	cy	sp3 carbons in square systems

Table 2: Triplets of weights which lead to the best rescoring results (the selected triplet is shown in bold).

Number of successfully predicted complexes	Weights (K)		
	$K_{polarized}$	K_{arom}	K_{aliph}
154/191	0.000	0.075	0.150
	0.000	0.100	0.150
	0.000	0.150	0.150
	0.000	0.150	0.200
	0.000	0.200	0.200
	0.010	0.075	0.150
	0.010	0.100	0.150
	0.010	**0.150**	**0.150**
	0.010	0.200	0.200
	0.025	0.150	0.150
	0.200	0.150	0.150

Table 3: Considered FF parameter schemes.

Parameter	Core FF	Preliminary FF	Fit FF
epsilon	2	4	4
desolv_factor	0.0	0.1	0.1
minq_to_desolv	0	0.125	0.125
hbond_bonus	0	2	2
repulsive_factor	1.00	0.75	0.75
vicinal_weight	0.5	0.033	0.033
desolv_scale_ion	-	1.0	0.1
desolv_scale_hb	-	1.0	0.1
$K_{polarized}$	0.0	0.1	0.01
K_{arom}	0.0	0.1	0.15
K_{aliph}	0.0	0.1	0.15

Table 4: Docking performance of several tools on the Astex Diverse Set (* statistics from closest protocols with respect to those presented here). With S4MPLE, Saved poses include the top 30 non-redundant (at *minfpdiff=0.01*) most stable geometries.

Docking Tools (Scoring)	Success Rate (%) Top ranked-pose	Success Rate (%) Saved poses
S4MPLE (Core FF)	76	93
S4MPLE (Fit FF)	85	96
GOLD (Goldscore) * [17]	75-81	Unavailable
FlexX	71	91
Plants (ChemPLP) [58]	87	97
Plants (PLP) [58]	84	
LGA (LargeAll) [32]	63	Unavailable
RosettaLig [59]	58	92
SKATE [60]	87	98

Table 5: RMSD of top-ranked poses for Core FF and Fit FF from complexes discussed in the dedicated chapter §3.2.2 (Core Vs Fit).

PDB	Core FF	Fit FF
1N2J	3,85	1,42
1NAV	6,26	0,35
1OQ5	0,92	2,83
1R58	3,03	0,74
1UML	7,59	0,66
1V0P	7,60	0,41
1W1P	2,94	0,39
1XOQ	4,04	0,30
1YVF	6,11	0,89

9 Appendices

$N_S := 0$

foreach *atom i from* 1 *to N* ***do***

if *(i is hydrophobic, donor or acceptor)* ***then***

calculate atom ID record $R_i := \left\{ M_i, Q_i, I_i^1 = \sum_{j \neq i}^{N} \frac{M_j Q_j}{T_{ij}} , I_i^2 = \sum_{j \neq i}^{N} \frac{M_j Q_j}{T_{ij}^2} \right\}$

if *(∃ S_k associated to record R_i)* ***then***

add i as a new member of S_k

else

inc(N_S)

create new set S_{N_S}, *associated to record* R_i

add i as a new member of S_k

end if *(∃ …)*

end if *(i is hydrophobic, donor or acceptor)*

end *atom loop*

Appendix 1: Pseudo-code of the symmetry set build-up.

```
Population initialization using RI      #Random initialization as in §2.4.6
while (stop condition not true) do
    for (each individual c_i in current population) do
        if (crossover probability met) then
            Select mate c_j using tournament selection
            c_o:=CROSS(c_i, c_j)    #CROSS as in §2.4.4
            MightReplace(c_o, c_i)
        end if
        if (mutation probability met) then
            c_m:=MUT(c_i)       #MUT as in §2.4.5
            if (not ReplaceMostRedundant(c_m)) then MightReplace(c_o, c_i)
        end if
    end for
    RandomizeIfRedundant    #Method as in §2.5
end while
```

Appendix 2: Pseudo-code of the SGA procedure.

```
Population initialization using RI      #Random initialization as in §2.4.6
while (stop condition not true) do
    pick one individual c_i from the current population
    RandomizeIfRedundant (c_i)   #Method as in §2.5
    while (attempts < allowed genetic operations) do
      if (mutation probability met) then
        c_c:=MUT(c_i)        #MUT as in §2.4.5
      else if (crossover probability met) then
         Select mate c_j randomly from the current population
        c_c:=CROSS(c_i, c_j)    #CROSS as in §2.4.4
      end if
      if (c_c is diverse wrt minfpdiff) then
        if (not ReplaceMostRedundant(c_c)) then MightReplace(c_c, c_i)
      else MightReplace(c_c, c_i)
      end if
      save_best_individual_if_new_top_found
    end while
end while
```

Appendix 3: Pseudo-code of the evolutionary ("evol") procedure.

10 References

1. Horvath, D.; Brillet, L.; Roy, S.; Conilleau, S.; Tantar, A.-A.; Boisson, J.-C.; Melab, N.; Talbi, E.-G. Local vs. global search strategies in evolutionary GRID-based conformational sampling & docking. In *IEEE Congress on Evolutionary Computation CEC 09*, IEEE: Trondheim, Norway, 2009; pp 247-254.
2. Parent, B., Kökösy, A., Horvath, D. Optimized Evolutionary Strategies in Conformational Sampling. *Soft Computing* **2007,** 11, 63-79.
3. Hoffer, L.; Horvath, D. S4MPLE - Sampler For Multiple Protein-Ligand Entities: Simultaneous docking of several entities. *J Chem Inf Model* **2012,** submitted.
4. Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; Debolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. AMBER a package of computer-programs for applying molecular mechanics, normal-mode analysis, molecular-dynamics and free-energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications* **1995,** 91, 1-41.
5. Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *Journal of Computational Chemistry* **2004,** 25, 1157-1174.
6. Horvath, D. A virtual screening approach applied to the search for trypanothione reductase inhibitors. *J Med Chem* **1997,** 40, 2412-23.
7. Nissink, J. W.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A new test set for validating predictions of protein-ligand interaction. *Proteins* **2002,** 49, 457-71.
8. Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J Med Chem* **2007,** 50, 726-41.
9. McGann, M. FRED pose prediction and virtual screening accuracy. *J Chem Inf Model* **2011,** 51, 578-96.
10. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology* **1996,** 261, 470-489.
11. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of Medicinal Chemistry* **2004,** 47, 1739-1749.
12. Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J Comput Aided Mol Des* **2001,** 15, 411-28.
13. Jain, A. N. Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *Journal of Medicinal Chemistry* **2003,** 46, 499-511.
14. Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. FlexE: Efficient molecular docking considering protein structure variations. *Journal of Molecular Biology* **2001,** 308, 377-395.
15. Goodsell, D. S.; Morris, G. M.; Olson, A. J. Automated docking of flexible ligands: Applications of AutoDock. *Journal of Molecular Recognition* **1996,** 9, 1-5.
16. Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. Distributed automated docking of flexible ligands to proteins: Parallel applications of AutoDock 2.4. *Journal of Computer-Aided Molecular Design* **1996,** 10, 293-304.
17. Jones, G.; Willett, P.; Glen, R.; Leach, A.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* **1997,** 267, 727-48.

18. Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins-Structure Function and Genetics* **2003,** 52, 609-623.
19. Meiler, J.; Baker, D. ROSETTALIGAND: Protein-small molecule docking with full side-chain flexibility. *Proteins-Structure Function and Bioinformatics* **2006,** 65, 538-548.
20. Zhao, Y.; Sanner, M. F. FLIPDock: Docking flexible ligands into flexible receptors. *Proteins-Structure Function and Bioinformatics* **2007,** 68, 726-737.
21. Corbeil, C. R.; Moitessier, N. Docking ligands into flexible and solvated macromolecules. 3. Impact of input ligand conformation, protein flexibility, and water molecules on the accuracy of docking programs. *J Chem Inf Model* **2009,** 49, 997-1009.
22. Corbeil, C. R.; Englebienne, P.; Yannopoulos, C. G.; Chan, L.; Das, S. K.; Bilimoria, D.; L'Heureux, L.; Moitessier, N. Docking Ligands into flexible and solvated macromolecules. 2. Development and application of FITTED 1.5 to the virtual screening of potential HCV polymerase inhibitors. *Journal of Chemical Information and Modeling* **2008,** 48, 902-909.
23. Corbeil, C. R.; Englebienne, P.; Moitessier, N. Docking ligands into flexible and solvated macromolecules. 1. Development and validation of FITTED 1.0. *J Chem Inf Model* **2007,** 47, 435-49.
24. Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. Novel procedure for modeling ligand/receptor induced fit effects. *J Med Chem* **2006,** 49, 534-53.
25. Hsu, H.-P.; Grassberger, P. A Review of Monte Carlo Simulations of Polymers with PERM. *J. Stat. Phys.* **2011,** 144, 597-637.
26. Ingber, L. Simulated annealing: Practice versus theory. *Journal of Mathematical Computation Modelling* **1993,** 18, 29-57.
27. Tantar, A.-A., Conilleau, S., Parent, B., Melab, N., Brillet, L., Roy, S., Talbi, E.-G., Horvath, D. Docking and Biomolecular Simulations on Computer Grids: Status and Trends. *Curr. Comput.-Aided Drug Des.* **2008,** 4, 235-249.
28. Grosdidier, A.; Zoete, V.; Michielin, O. EADock: Docking of small molecules into protein active sites with a multiobjective evolutionary optimization. *Proteins-Structure Function and Bioinformatics* **2007,** 67, 1010-1025.
29. Schefzick, S.; Bradley, M. Comparison of Commercially Available Genetic Algorithms: GAs as Variable Selection Tool. *J. Comput. Aided Mol. Des.* **2004,** 18, 511-521.
30. Thomsen, R. Flexible ligand docking using evolutionary algorithms: investigating the effects of variation operators and local search hybrids. *Biosystems* **2003,** 72, 57-73.
31. Zoete, V.; Grosdidier, A.; Cuendet, M.; Michielin, O. Use of the FACTS solvation model for protein–ligand docking calculations. Application to EADock. *J. Mol. Recognit.* 23, 457-461.
32. Fuhrmann, J.; Rurainski, A.; Lenhof, H. P.; Neumann, D. A new Lamarckian genetic algorithm for flexible ligand-receptor docking. *J Comput Chem* **2010,** 31, 1911-8.
33. Shen, Q. C.; Xiong, B.; Zheng, M. Y.; Luo, X. M.; Luo, C.; Liu, X. A.; Du, Y.; Li, J.; Zhu, W. L.; Shen, J. K.; Jiang, H. L. Knowledge-Based Scoring Functions in Drug Design: 2. Can the Knowledge Base Be Enriched? *Journal of Chemical Information and Modeling* **2011,** 51, 386-397.
34. Neudert, G.; Klebe, G. DSX: A Knowledge-Based Scoring Function for the Assessment of Protein-Ligand Complexes. *Journal of Chemical Information and Modeling* **2011,** 51, 2731-2745.
35. Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *Journal of Medicinal Chemistry* **2000,** 43, 4759-4767.
36. Horvath, D.; Brillet, L.; Roy, S.; Conilleau, S.; Tantar, A. A.; Boisson, J. C.; Melab, N.; Talbi, E. G.; Ieee. Local vs. Global Search Strategies in Evolutionary GRID-based Conformational Sampling & Docking. In *2009 IEEE Congress on Evolutionary Computation, Vols 1-5*, IEEE: New York, 2009; pp 247-254.

37. Tantar, A.-A.; Melab, N.; Talbi, E.-G.; Parent, B.; Horvath, D. A parallel hybrid genetic algorithm for protein structure prediction on the computational grid. *Future Generation Computer Systems* **2007,** 23, 398-409.
38. Parent, B.; Koekoesy, A.; Horvath, D. Optimized evolutionary strategies in conformational sampling. *Soft Computing* **2007,** 11, 63-79.
39. Parent, B., Tantar, A., Melab, N., Talbi, E.-G., Horvath, D. Grid-based Evolutionary Strategies Applied to the Conformational Sampling Problem. In *IEEE Congress on Evolutionary Computation, CEC 2007*, Singapore, 2007; pp 291-296.
40. Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Swails, J.; Goetz, A. W.; Kolossvary, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wolf, R. M.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Salomon-Ferrer, R.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. AMBER 12. *University of California, San Francisco* **2012**.
41. Dauberosguthorpe, P.; Roberts, V. A.; Osguthorpe, D. J.; Wolff, J.; Genest, M.; Hagler, A. T. Structure and energetics of ligand-binding to proteins - escherichia-coli dihydrofolate reductase trimethoprim, a drug-receptor system. *Proteins-Structure Function and Genetics* **1988,** 4, 31-47.
42. Halgren, T. A. MMFF VII. Characterization of MMFF94, MMFF94s, and other widely available force fields for conformational energies and for intermolecular-interaction energies and geometries. *Journal of Computational Chemistry* **1999,** 20, 730-748.
43. Tripos, I. *Sybyl*, 8.0; St. Louis, MO, 2007.
44. Mazur, J.; Jernigan, R. Distance-dependent dielectric-constants and their application to double-helical DNA. *Biopolymers* **1991,** 31, 1615-1629.
45. Mcnaught, A. D. Recent Iupac Nomenclature Recommendations. *Journal of Pharmacy and Pharmacology* **1984,** 36, 643-643.
46. Marcou, G.; Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J Chem Inf Model* **2007,** 47, 195-207.
47. Brewerton, S. C. The use of protein-ligand interaction fingerprints in docking. *Curr Opin Drug Discov Devel* **2008,** 11, 356-64.
48. Deng, Z.; Chuaqui, C.; Singh, J. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J Med Chem* **2004,** 47, 337-44.
49. DeLano, W. L. *The PyMOL Molecular Graphics System*, DeLano Scientific: San Carlos, CA, USA, 2002.
50. RCSB Protein Data Bank. http://www.rcsb.org/pdb/
51. ChemAxon. Calculation of Partial Charge Distributions. http://www.chemaxon.com/marvin/help/calculations/charge.html (Feb. 2009).
52. Wang, J. M.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics & Modelling* **2006,** 25, 247-260.
53. *MOE (Molecular Operating Environment)*, 2005.06; Chemical Computing Group, Inc.: Montreal, 2005.
54. Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **2004,** 57, 225-42.
55. Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. *Proteins* **1999,** 37, 228-41.
56. Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J Med Chem* **2000,** 43, 4759-67.

57. Horvath, D.; Lippens, G.; vanBelle, D. Development and parametrization of continuum solvent models .2. A unified approach to the solvation problem. *Journal of Chemical Physics* **1996,** 105, 4197-4210.
58. Korb, O.; Stützle, T.; Exner, T. E. Empirical scoring functions for advanced protein-ligand docking with PLANTS. *J Chem Inf Model* **2009,** 49, 84-96.
59. Davis, I. W.; Baker, D. RosettaLigand docking with full ligand and receptor flexibility. *J Mol Biol* **2009,** 385, 381-92.
60. Feng, J. A.; Marshall, G. R. SKATE: a docking program that decouples systematic sampling from scoring. *J Comput Chem* **2010,** 31, 2540-54.

3.5 Rescoring de conformères de peptides avec la fonction Fit FF

Comme souligné préalablement, le jeu de peptides a fait l'objet de nouvelles simulations (identiques à celles de l'étape 1.I du §3.2.1 avec un EC biaisé et non biaisé) en utilisant la version finale de la fonction d'énergie (Fit FF). Ici, une phase de post-traitement supplémentaire est réalisée sur l'ensemble des conformères générés. Son but est d'optimiser et de sauvegarder pour chaque peptide les meilleures géométries, à savoir les conformères non redondants ayant l'E_{pot} la plus faible (un nombre maximal de 500 conformères est toléré).

Comme toujours dans le cadre d'une approche heuristique et spécialement pour des problèmes très complexes à l'instar de l'EC de peptides, le conformère de plus basse énergie (E_{pot}) obtenu à l'issue des simulations n'est pas forcément le minimum global. Mais cela donne néanmoins un aperçu de la pertinence des paramètres du FF considéré ; en effet, des paramètres inadéquats ont une très faible probabilité de faire ressortir des structures natives en tête de liste. Dans ce contexte, un conformère natif est une géométrie qui possède :

- une distance faible entre son fingerprint et celui de la structure expérimentale. Pour une meilleure compréhension car plus intuitif, le critère RMSD est donné dans les différents tableaux. Le RMSD, calculé uniquement sur la chaîne principale (“main chain”), est obtenu après superposition des chaînes des deux géométries. Cette action est réalisée avec la fonction “align” du logiciel Pymol [192], et tous les RMSD relatifs aux peptides sont calculés de cette façon
- un repliement globalement conservé, notamment en ce qui concerne les structures secondaires de type hélice ou brin (critère visuel sous Pymol)

Les résultats détaillés sont donnés ci-dessous dans le Tableau 12 (Core FF) et le Tableau 13 (Fit FF).

Pour tous les peptides du jeu, un conformère non natif possède l'E_{pot} la plus faible dans le cas de Core FF. Les peptides 1L2Y, 1UAO et 2KEF ont chacun un conformère natif avec un rang très correct (bien que non égal à 1). A l'opposé, les peptides 1LE1 et 1VII ont des résultats plus mitigés : la différence d'énergie entre le meilleur conformère natif et le conformère de plus basse énergie est plus élevée, ce qui a une conséquence négative directe sur le rang du meilleur conformère natif.

Core FF					
PDB	Structure exp. minimisée	Conformère le plus stable (meilleure énergie)			
	Energie FF	Rang	Energie FF	RMSD	Différence d'énergie avec la pose non native la plus stable
1L2Y	-126,9	1	-190,7	**3,4**	/
1LE1	-90,0	1	-130,4	**5,2**	/
1UAO	-65,3	1	-99,0	**4,8**	/
1VII	-227,2	1	-412,0	**6,4**	/
2KEF	-18,1	1	-157,4	**6,0**	/
PDB	Structure exp. minimisée	Conformère le plus proche de la structure native (distance minimale entre les fingerprints)			
	Energie FF	Rang	Energie FF	RMSD	Différence d'énergie avec la pose non native la plus stable
1L2Y	-126,9	3	-182,3	**0,9**	8,4
1LE1	-90,0	26	-114,1	**0,7**	16,3
1UAO	-65,3	12	-93,5	**0,4**	5,5
1VII	-227,2	18	-378,6	**2,3**	33,4
2KEF	-18,1	2	-155,9	**0,9**	1,5

Tableau 12: Résumé des simulations avec la fonction d'énergie Core FF.

Fit FF					
PDB	Structure exp. minimisée	Conformère le plus stable (meilleure énergie)			
	Energie FF	Rang	Energie FF	RMSD	Différence d'énergie avec la pose non native la plus stable
1L2Y	-144,5	1	-192,5	**0,4**	-1,0
1LE1	-108,5	1	-167,4	**4,0**	/
1UAO	-84,0	1	-109,6	**4,0**	/
1VII	-324,5	1	-423,1	**1,8**	-0,6
2KEF	-81,4	1	-190,6	**1,3**	-1,2
PDB	Structure exp. minimisée	Conformère le plus proche de la structure native (distance minimale entre les fingerprints)			
	Energie FF	Rang	Energie FF	RMSD	Différence d'énergie avec la pose non native la plus stable
1L2Y	-144,5	1	-192,5	**0,4**	0,0
1LE1	-108,5	60	-153,9	**0,9**	13,5
1UAO	-84,0	3	-107,7	**0,7**	1,9
1VII	-324,5	1	-423,1	**1,8**	0,0
2KEF	-81,4	1	-190,6	**1,3**	0,0

Tableau 13: Résumé des simulations avec la fonction d'énergie Fit FF.

Sur les 5 peptides du jeu, la fonction d'énergie Fit FF a convergé 3 fois (1L2Y, 2KEF, 1VII) vers un repliement très similaire à la structure expérimentale. En ce qui concerne les 2 échecs au rang 1, un conformère natif est très rapidement retrouvé pour 1UAO (rang 3), mais le problème persiste pour le cas du peptide 1LE1.

La Figure 45 montre l'exemple du peptide 1L2Y (son nom usuel est "cage du tryptophane") où un succès est obtenu avec le Fit FF, alors que Core FF aboutit à un échec. A l'exception des extrémités N_{ter} et C_{ter} qui se déplacent de quelques ångströms dans le but de "fermer la structure" à l'aide d'une liaison ionique, le reste de celle-ci se superpose parfaitement à la structure expérimentale RMN. De plus, ces extrémités sont également les parties les moins statiques en réalité, comme le montre le fichier PDB qui contient plusieurs dizaines de structures. A l'opposé dans le cas du Core FF, le conformère de plus basse énergie est totalement déplié : le tryptophane central est relégué en périphérie et l'on observe un réseau de liaisons ioniques / hydrogène qui ne décrivent absolument pas le repliement natif.

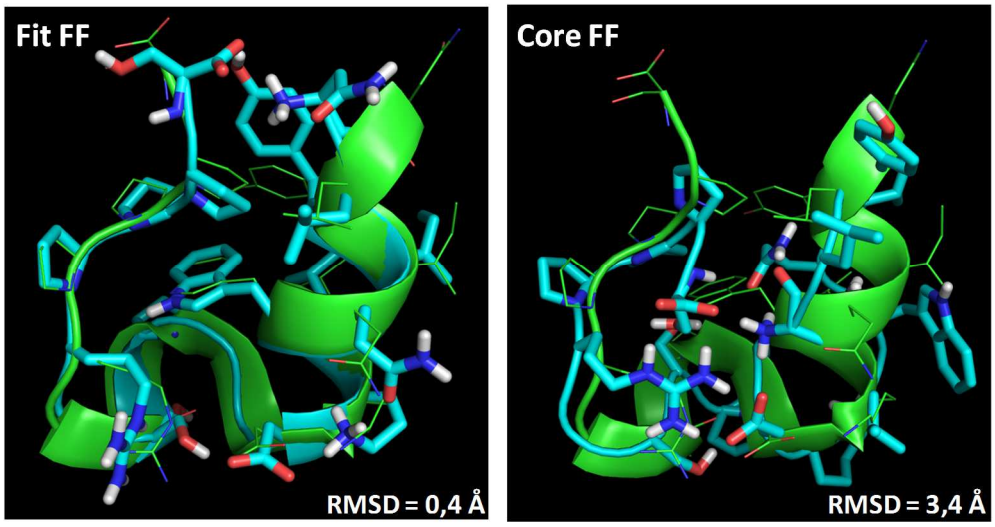


Figure 45: Superposition de la structure expérimentale 1L2Y (vert) et du meilleur conformère obtenu (bleu) selon E_{pot} avec les deux FF considérés.

Bien qu'objectivement non parfaits, ces derniers résultats de la fonction d'énergie Fit FF (qui repose sur les champs de force AMBER / GAFF tout en incluant entre autre un modèle de solvant implicite) confirment la tendance observée lors des simulations de validation de docking ligand-récepteur (voir le §3.3).

3.6 Corrélation entre énergie du FF et affinité expérimentale

Jusqu'à présent, seule la capacité à favoriser le mode de liaison expérimental par rapport à un important nombre de poses leurres a été activement recherchée, tout en ayant ajouté des composantes modélisant certains aspects énergétiques favorables et défavorables des phénomènes de liaison. Dans le cas idéal (paramètres natifs et additionnels optimaux), il devrait néanmoins exister une certaine corrélation entre l'énergie d'interaction du FF et des constantes expérimentales d'affinité (K_d, K_i, *etc.*). On ne parle pas ici de l'estimation d'énergie libre de liaison relative ou absolue, mais bien d'énergie directement calculée à partir d'un FF. Dans ce contexte de physique Newtonienne, l'E_{pot} est d'autant plus basse que la confirmation du système associée est stable. De plus, leur ordre de grandeur n'est généralement pas comparable avec ce type de constante thermodynamique : un ligand ayant un K_d picomolaire n'a qu'une énergie libre de liaison environ égale à -15 kcal.mol^{-1}.
La corrélation (ou l'absence de corrélation) entre énergie du FF et affinité expérimentale est analysée dans ce sous-chapitre à partir de structures X-ray et des données expérimentales associées.

3.6.1 Méthode et données expérimentales

Une constante expérimentale de liaison ligand-récepteur est répertoriée pour la plupart des complexes du jeu de validation (Astex Diverse Set). Malheureusement, ces données ne sont pas homogènes car étant de natures différentes (K_d, K_i, EC_{50}, IC_{50}, K_m, *etc.*). L'étude porte sur le sous-ensemble des K_i car ils sont à la fois fortement liés à la notion d'affinité et relativement nombreux sur l'ensemble du jeu (29 données distinctes avec une grande amplitude : pK_i variant de -4,1 à -10,5). De plus, le K_i est une mesure expérimentale plus rigoureuse que les IC_{50} ou EC_{50} car moins dépendante de certaines conditions expérimentales utilisées lors de la mesure [193].

Pour chaque complexe, le ligand flexible est minimisé dans le site rigide non solvaté en partant de la géométrie expérimentale du fichier PDB (le fichier moléculaire du site est celui créé pour les simulations de redocking). L'énergie potentielle obtenue à l'issue de la relaxation du ligand dans le site est nommée $E_{pot\,complexe}$, et inclut entre autre les composantes intermoléculaires ligand-récepteur.
Jusqu'à ce stade, on s'est intéressé à différents systèmes (peptides, complexes ligand-récepteur) mais toujours de manière indépendante. Par conséquent, l'E_{pot} pouvait être directement utilisée pour classer les différents conformères ou poses selon leur énergie propre. Dans ce contexte-ci, où des systèmes variés (différents ligands dans différents sites) sont regroupés, il faut passer à une mesure de l'énergie incorporant une composante fonction de la conformation relaxée du ligand libre. A ce titre, la

bioconformation du ligand est minimisée seule à l'aide d'un algorithme adéquat, et l'énergie potentielle obtenue est appelée $E_{pot\,ligand\_seul}$. Il est à noter que le site reste figé ici ; par conséquent, il n'y a pas besoin d'inclure une composante liée à la conformation relaxée du site libre. L'énergie finalement considérée, nommée $E_{interaction}$, est la soustraction de ces deux énergies : $E_{pot\,complexe} - E_{pot\,ligand\_seul}$.
Contrairement à ce cas d'étude, un site unique est considéré dans le cadre d'un criblage virtuel : les différents systèmes ne diffèrent plus que par le ligand criblé, à la seule condition que les DDL du site restent inchangés durant l'intégralité du criblage. De fait, la composante intra-site est naturellement prise en compte par l'énergie potentielle du complexe $E_{pot\,complexe}$ (avec ou sans déblocage de DDL du site). Le terme spécifique $E_{pot\,protéine\_seule}$, représentant la conformation relaxée du site selon les DDL considérés, reste constant. Par conséquent, ce terme $E_{pot\,protéine\_seule}$ peut donc ne pas être retranché, dans la mesure où des énergies relatives sont suffisantes dans l'optique d'un classement des différents ligands. On peut alors utiliser la différence d'énergie potentielle, définie ci-dessus, pour classer les différents ligands d'un criblage virtuel.

Les données expérimentales sont finalement confrontées avec les énergies potentielles (E_{pot}) et d'interaction ($E_{interaction}$) calculées selon les deux schémas énergétiques habituels (Core FF et Fit FF). Le score attribué par le logiciel de docking FlexX [121] est également inclus dans cette étude à des fins de comparaison. Pour ce dernier cas de figure où les résultats de docking sont utilisés puisqu'il n'est pas possible d'optimiser directement la conformation expérimentale, les seuls complexes considérés sont ceux pour lesquels la meilleure pose est correcte (RMSD < 2 Å). En effet, le mauvais docking d'un ligand (par exemple, 1KZK qui est un ligand complexe de haute affinité) peut aboutir à un mauvais score pour une raison d'échec d'EC totalement indépendant de l'évaluation énergétique. 21 complexes sont finalement retenus au regard des 29 complexes initialement considérés (1GPK, 1HVY, 1KZK, 1MMV, 1T9B, 1UML, 1YGC et 1YQY sont exclus).
Une régression linéaire simple est réalisée sous Excel pour évaluer la corrélation entre les énergies et les données expérimentales. Le coefficient de corrélation R^2 correspondant est également calculé à l'aide de cet outil. Ce coefficient est un indicateur permettant de juger de la qualité d'une régression car il évalue l'adéquation entre le modèle et les données observées.

3.6.2 Résultats et discussion

Les résultats de cette étude, à savoir les différentes énergies calculées, sont donnés dans le Tableau 14 pour chaque complexe d'intérêt. Les données brutes (K_i en µM) [143] sont transformées en pKi = -log(K_i), après une conversion préalable des K_i en M.

PDB	pK_i	Core FF			Fit FF		
		Epot complexe	Epot ligand_seul	E interaction	Epot complexe	Epot ligand_seul	E interaction
1gpk	-5,4	-26,9	24,3	-51,2	-46,6	10,2	-56,8
1hnn	-6,2	-35,9	33,0	-68,9	-36,2	24,8	-61,0
1hp0	-6,7	-5,8	60,0	-65,8	-23,7	46,7	-70,4
1hvy	-6,2	-54,6	28,6	-83,2	-69,3	16,7	-85,9
1j3j	-8,0	-20,0	35,8	-55,8	-39,9	20,2	-60,1
1jd0	-8,2	-37,2	43,7	-80,9	-12,5	43,3	-55,8
1kzk	-10,4	-51,8	28,5	-80,3	-93,4	0,1	-93,5
1l2s	-4,6	-11,8	32,0	-43,8	-26,7	23,4	-50,1
1lpz	-7,6	-18,7	46,6	-65,3	-44,5	29,8	-74,3
1mmv	-7,2	-77,0	2,1	-79,0	-70,3	4,6	-74,9
1n1m	-5,7	-47,7	16,3	-64,0	-42,1	8,5	-50,7
1n2v	-4,1	6,9	47,1	-40,2	-2,1	41,5	-43,6
1n46	-10,5	-45,9	32,8	-78,7	-84,4	9,2	-93,6
1of6	-6,0	-72,8	-6,7	-66,1	-70,0	-5,9	-64,0
1owe	-6,2	-29,7	26,0	-55,7	-45,1	8,0	-53,0
1oyt	-7,2	-30,5	49,0	-79,5	-50,3	31,6	-81,9
1q4g	-6,9	-42,3	15,3	-57,6	-61,2	4,6	-65,8
1r1h	-8,9	-103,7	-7,6	-96,1	-96,7	-11,4	-85,3
1r55	-6,8	-62,9	10,6	-73,6	-65,8	-0,5	-65,3
1t9b	-6,9	-21,4	32,6	-54,0	-44,6	15,5	-60,0
1tt1	-4,2	-78,2	4,8	-83,1	-64,0	5,0	-69,0
1u1c	-5,4	-41,8	15,8	-57,5	-57,1	6,7	-63,9
1uml	-7,5	-12,0	55,0	-67,0	-42,4	39,9	-82,3
1v48	-8,2	-55,7	38,5	-94,2	-53,9	27,9	-81,8
1w2g	-4,6	-23,3	22,1	-45,4	-30,9	17,4	-48,3
1ygc	-9,5	-46,7	45,6	-92,3	-75,5	19,7	-95,2
1yqy	-7,6	-30,1	33,7	-63,7	-40,0	24,3	-64,3
1z95	-7,1	-40,0	19,4	-59,3	-73,0	5,8	-78,8
2bm2	-7,8	-29,7	43,0	-72,7	-55,3	17,7	-73,0

Tableau 14: Données expérimentales et énergies calculées pour chaque complexe PDB considéré.

Les droites de corrélation sont illustrées à la Figure 46 et les différentes valeurs des coefficients de corrélation sont répertoriées dans le Tableau 15.

FF	Energie considérée	Nombre de points	Coefficient de corrélation
Core FF	$E_{pot\ complexe}$	29	$R^2 \approx 0,06$
Fit FF	$E_{pot\ complexe}$	29	$R^2 \approx 0,29$
Core FF	$E_{interaction}$	29	$R^2 \approx 0,39$
Fit FF	$E_{interaction}$	29	$R^2 \approx 0,58$
Score FlexX		21	$R^2 \approx 0,14$

Tableau 15: Coefficients de corrélation obtenus selon la fonction d'énergie considérée.

Bien que reposant sur un nombre limité de points, en l'occurrence 29, plusieurs commentaires émergent de ces résultats :

- une corrélation faible voire quasi-inexistante est observée en tenant uniquement compte de l'E_{pot} du système. Ce résultat était attendu dans la mesure où cette énergie peut très fortement différer de l'énergie d'interaction car elle inclut sans aucune modulation les composantes intramoléculaires du ligand. Cet effet peut aussi être amplifié avec GAFF étant donné que les constantes de force associées au terme "angle de valence" sont très élevées [174] : l'énergie basale d'un ligand, à savoir l'énergie potentielle du ligand seul après relaxation, peut être maintenue à une valeur relativement haute si les angles de valence diffèrent légèrement de leur valeur théorique.
- le passage à une énergie d'interaction améliore sensiblement les résultats par rapport à l'énergie potentielle : le coefficient de corrélation atteint 0,58 et 0,39, respectivement pour les fonctions d'énergie Fit FF et Core FF. Précédemment (voir le §3.3), on a montré au moyen de simulations de redocking que le Fit FF permettait de converger plus souvent vers le mode de liaison expérimental que le Core FF. Cette même tendance se retrouve ici avec une meilleure corrélation entre les énergies et les affinités, en soulignant toutefois que les valeurs R^2 restent relativement moyennes. Plusieurs valeurs aberrantes ("outliers"), discutées ci-dessous, sont clairement mises en évidence (voir la Figure 46)
    - 1JD0) ce fragment de 221 Da, qui interagit avec un cation métallique Zn^{2+}, est sous-estimé avec le Fit FF alors qu'il est dans la norme avec le Core FF. Une hypothèse est

que cet artefact provient du terme de désolvatation qui reste trop pénalisant malgré la mise à l'échelle des paires non liées impliquant un cation. Il est à noter que l'estimation de l'affinité de ligands se liant à des métalloprotéines est connu pour être problématique, et certaines études mettent tout simplement de côté ce type de complexe [67]

  - 1HVY) l'énergie est surestimée alors qu'il s'agit d'un ligand drug-like (456 Da) qui possède une affinité somme toute moyenne (de l'ordre du µM). La relaxation du ligand dans le site aboutit à la formation d'une liaison ionique avec LYS77 qui est en réalité entourée de trois molécules d'eau d'après la structure X-ray. Par conséquent, une surestimation de la composante électrostatique et l'échec du modèle de solvant continu à se substituer à des interactions médiées par des molécules d'eau sont des hypothèses crédibles pour expliquer cette énergie trop favorable au regard de l'affinité réelle du ligand. Celui-ci est aussi un outlier pour le Core FF
  - 1TT1) ce fragment est surestimé à la fois avec Fit FF et Core FF. Son mode de liaison, impliquant deux liaisons ioniques et plusieurs liaisons hydrogène, rend aisément compréhensible l'obtention d'une énergie très favorable. Malgré le terme de pénalité lié à la désolvatation, ce complexe reste dans la catégorie des outliers bien qu'il y ait une réduction de l'erreur de prédiction par rapport à Core FF (l'énergie avec le FF natif est comparable à celle obtenue avec des ligands de $pK_i \approx -10$)
- à l'instar de l'E_{pot} seule, on observe une corrélation quasi-inexistante ($R^2 \approx 0,14$) entre le score de docking de FlexX et les 21 valeurs de pK_i considérées. Les deux plus grands outliers sont à nouveau 1TT1 (surestimation) et 1JD0 (sous-estimation) comme pour le Fit FF

Nos résultats très convenables ($R^2 \approx 0,6$) sont issus d'une stratégie où il n'y a eu aucun entraînement explicite avec des constantes d'affinité. De plus, une corrélation relative ("tendance") entre énergies calculées et affinités expérimentales est un fait récurrent dans le domaine de la modélisation moléculaire [67]. A ce titre, les meilleures fonctions de score empiriques (voir le §1.6.2) explicitement entrainées sur des affinités expérimentales ou les approches de type MM-GBSA (voir le §1.4.4) ne font pas beaucoup mieux ($R^2 \approx 0,63$ pour MM-GBSA) [67, 194]. Toutefois, il faut souligner que la taille des jeux (entraînement et/ou validation) est beaucoup plus importante dans les études de Sotriffer *et al* [194] et Greenidge *et al* [67], ces dernières étant dédiées à cette problématique fondamentale d'estimation des affinités. Enfin, une étude a montré que l'utilisation d'une constante diélectrique relative fonction de la distance est très compétitive vis-à-vis de méthodes plus rigoureuses comme MM-GBSA [180].

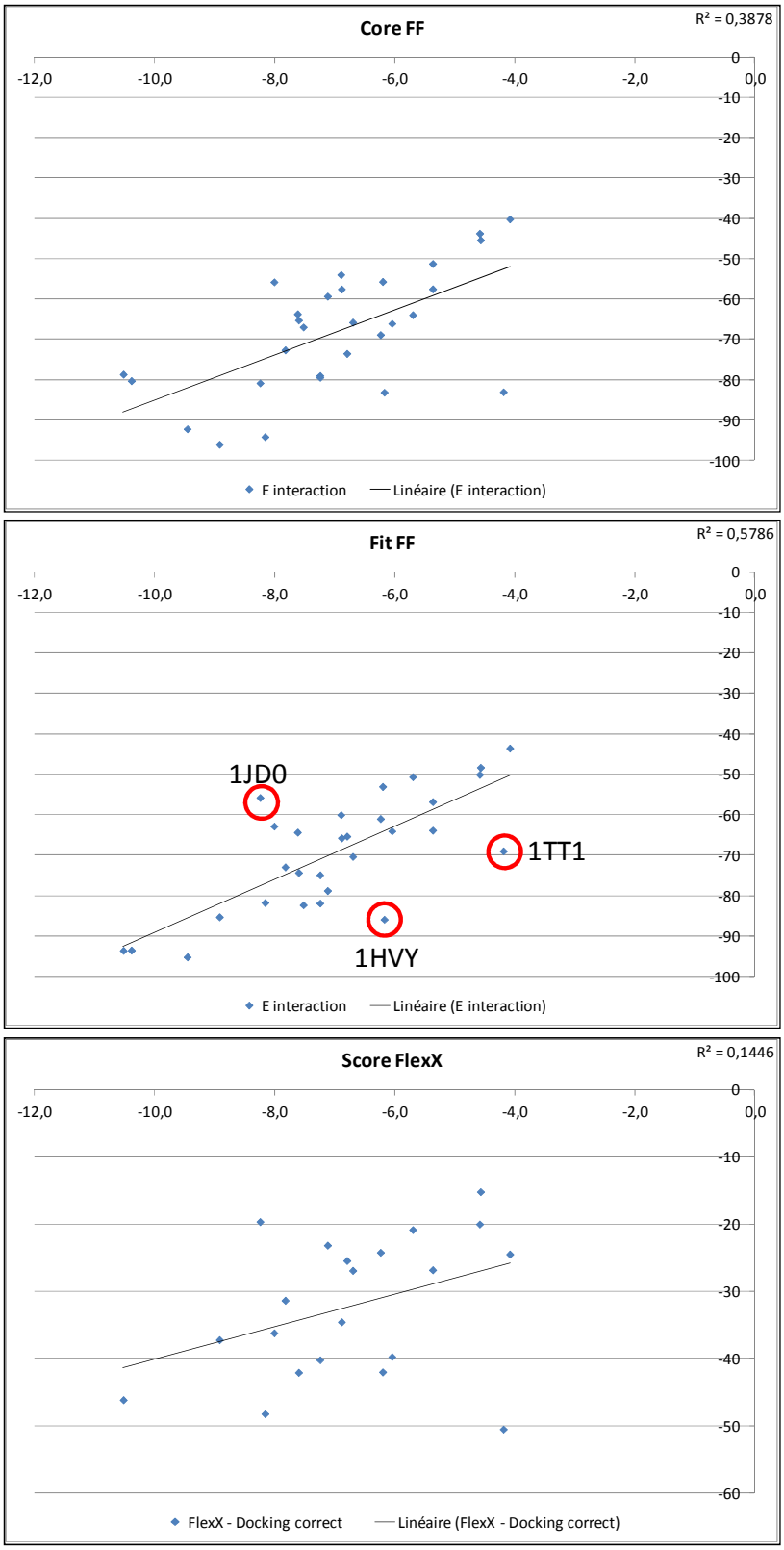


Figure 46: Corrélation entre l'énergie d'interaction ou le score (axe Y) et les pKi (axe X). Haut : Core FF ; Milieu : Fit FF ; Bas : FlexX. Les cercles indiquent les outliers pour le Fit FF.

4 Docking de fragments

Le cahier des charges impose que S4MPLE soit capable de docker correctement des ligands de type fragment. Or ces molécules organiques sont connues pour être assez problématiques en docking étant donné que la fonction d'énergie / de score doit souvent être capable de discriminer le bon mode de liaison au sein d'une multitude de possibilités (du fait de leur petite taille), le tout dans une fenêtre énergétique assez restreinte. Une étude a aussi montré que la similarité de mode de liaison, évaluée à l'aide d'empreintes d'interaction (IFP), s'est révélée plus efficace pour identifier la bonne pose que l'utilisation d'une fonction de score, spécialement pour des composés de très faible complexité comme les fragments [1]. Néanmoins, un travail récent [195] a conclu qu'il n'y a pas de différences notables au niveau des taux de succès de docking entre des ligands de type fragment et des ligands de plus grande complexité. Dans le cas où un échec est observé en docking, les premiers posent problème pour les raisons évoquées ci-dessus, alors que les seconds mettent plutôt une insuffisance d'EC en évidence.

S4MPLE est théoriquement capable de jongler avec plusieurs entités mobiles, mais encore faut-il le prouver à l'aide de cas concrets. A ce titre, certains résultats expérimentaux obtenus dans le cadre de projets FBDD sont une formidable opportunité de valider cette capacité. Des complexes ternaires, incluant deux fragments distincts liés simultanément dans leur site de liaison, ont été déposés à la PDB. Ce type d'information structurale est notamment le point de départ d'une stratégie d'optimisation de type linking. De plus, certaines méthodes expérimentales, comme le criblage de composés par spectrométrie de masse en conditions natives [52], permettent de détecter ce type d'évènement (liaison simultanée de deux ligands), mais sans l'accès aux modes de liaison à l'échelle atomique. Une approche de prédiction de la géométrie de complexes ternaires peut s'avérer utile, au minimum dans l'attente de la résolution expérimentale de la structure (X-ray). Enfin, il a été montré que le mode de liaison d'un fragment donné peut changer lors de la liaison simultanée de deux composés dans le site [196], d'où l'intérêt de tenter de prédire directement la géométrie du complexe ternaire fragment1-fragment2-récepteur. C'est pourquoi des simulations ayant pour but de reproduire les modes de liaison de deux fragments au sein d'un complexe ternaire ont été entreprises.

Au préalable, la capacité à docker correctement des fragments est évaluée à l'aide de simulations de redocking et selon les mêmes critères que précédemment (voir le §3.3).

Ces résultats ont fait l'objet d'un article publié [197] qui est également inséré dans ce document (voir le §4.3). Par conséquent, seules les grandes lignes sont décrites et discutées ici. Enfin, la publication fait référence à des simulations multi-entités variées (ligand+molécules d'eau ou fragment1+fragment2). Au regard du titre de ce mémoire, seule la partie dédiée aux ligands de type fragment sera mentionnée dans ce chapitre.

4.1 Docking classique appliqué à des fragments

Cette étape de validation préliminaire repose sur un ensemble de 142 complexes PDB fragment-récepteur. Il est constitué à partir de trois sources : Astex Diverse Set [143], Astex/CCDC-Clean [142] et les complexes décrits dans une revue de référence du FBDD [47]. Un complexe fragment-récepteur est simplement identifié lorsque le ligand principal possède une masse inférieure à 300 Da, ce seuil étant celui intégré dans les Ro3 [49]. Les protocoles de préparation des fichiers moléculaires et les paramètres des simulations sont conservés. Dans un but de benchmark, les deux fonctions d'énergie (Core FF et Fit FF) sont à nouveau employées comme précédemment (voir le §3.3).

Un succès est à nouveau défini selon le critère RMSD de la meilleure pose selon E_{pot} qui doit être inférieur à 2 Å par rapport à la conformation cristallographique du ligand. Les taux de succès donnés sont obtenus en moyennant ceux obtenus sur 5 simulations indépendantes sur l'ensemble du jeu considéré. On utilise également la nomenclature usuelle où les poses sont classées par énergie, et le rang 1 correspond à la plus favorable.

Comme l'ont récemment publié Verdonk *et al* [195], les taux de succès de docking de fragments sont comparables à ceux obtenus sur un jeu composé à plus de 50% de composés drug-like (voir le §3.3). A nouveau, on observe un avantage sensible de la fonction d'énergie Fit FF (86%), en terme de taux de succès, par rapport à Core FF (77%). De plus, l'analyse de la capacité à reproduire une pose correcte dans les 30 meilleures révèle qu'un succès est quasi-systématiquement observé (98% des cas pour les deux fonctions considérées). Ce dernier point était toutefois attendu dans la mesure où ces molécules de faible complexité ne doivent pas poser de problème d'EC outre mesure. Les résultats détaillés sont disponibles dans le tableau 3 et les figures 1 et 2 de la publication (voir le §4.3).

Un seuil critique de 2 Å pour le RMSD peut apparaître fort généreux pour des molécules de si faible complexité (faible masse, nombre de DDL limité). En pratique, un RMSD légèrement inférieur se traduit souvent par une translation du fragment, ce qui peut aboutir à la perte de la plupart des interactions clés. Par contraste, les ligands drug-like, pour lesquels le seuil de 2 Å a été défini, sont généralement plus flexibles et les interactions clés peuvent être parfaitement reproduites malgré un RMSD se situant dans cette zone. Dès lors, un seuil plus strict de RMSD peut être envisagé et une valeur de 1,5 Å a été récemment employée [195]. Aux seuils de 1,5 et 1 Å, le taux de succès s'établit respectivement à 79% et 70% pour la fonction d'énergie Fit FF. Cette baisse relativement faible des résultats, malgré l'emploi de seuils beaucoup plus stricts, confirme la tendance observée au seuil par défaut de 2 Å. En d'autres termes, soit il y a un échec du docking au niveau de la meilleure pose, soit il y a un succès impliquant une très bonne reproduction du mode de liaison du fragment. Il est à noter qu'un seuil encore plus strict, de l'ordre de 0,5 Å, peut perdre tout son sens : le positionnement du

ligand n'est pas forcément connu avec une précision d'orfèvre, car celui-ci découle de l'interprétation d'une carte de densité électronique obtenue avec les limites inhérentes à l'utilisation des techniques de cristallographie sur des entités biologiques (limite de résolution moyenne, *etc.*).

La capacité à docker des composés de type fragment étant validée, il devient possible de passer à l'étape suivante, à savoir le docking simultané de plusieurs entités mobiles.

4.2 Docking simultané de plusieurs fragments

Cette partie se focalise sur la prédiction de la géométrie de complexes ternaires fragment1-fragment2-récepteur. Le fait de simuler plusieurs entités mobiles au sein d'un site de liaison augmente de manière très importante le nombre de combinaisons possibles.
Deux approches différentes sont développées pour retourner *in fine* des géométries potentielles pour un complexe ternaire donné :

1) un docking multi-entités explicite. Les deux fragments sont simulés en même temps dans le site de liaison. Durant la phase d'EC, l'AG tente ici de déterminer la configuration du système fragment1-fragment2-site qui aboutit à l'E_{pot} la plus faible. La phase de post-traitement usuelle, à savoir la minimisation des poses non redondantes dans une fenêtre d'énergie donnée, complète la simulation
2) un docking dit séquentiel. Comme son nom l'indique, cette méthode vise dans un premier temps à simuler séparément chaque fragment dans le site de liaison. Puis les meilleures poses non redondantes sont extraites, et chaque pose du fragment1 est ensuite associée à chaque pose du fragment2. Ce processus est très facile à automatiser car les poses sont également sauvegardées sous forme de lignes de coordonnées (voir le guide d'utilisateur de S4MPLE au niveau du §2.1.3). Enfin, la phase de post-traitement usuelle est lancée sur ces nouvelles poses non redondantes avec une tolérance plus forte sur la fenêtre énergétique afin de tolérer initialement la présence de légers clashes. En plus de supprimer les éventuels clashes, la relaxation peut également permettre d'optimiser les contacts favorables inter-fragments. Du fait du découplage de l'EC des DDL, cette seconde approche est en théorie bien plus aisée que la première

Une recherche dans la banque de structures PDB a permis d'extraire des complexes ternaires d'intérêt. Seuls des fragments non cofacteur avec au moins un cycle sont considérés. De plus, la publication associée au code PDB doit faire allusion au FBDD. Au final, seuls quelques complexes sont

récupérés. Vu que la plupart des structures appartiennent à la même cible (“Heat Shock Protein 90” - HSP90), il a été décidé de ne conserver que les complexes de celle-ci (2QFO, 2XDU, 2YEI, 2YEJ et 3HZ1, voir le tableau 2 de la publication). La phase d'EC du docking est composée de 8 simulations indépendantes de 2000 générations, et les résultats sont ensuite rassemblés avant la phase de post-traitement. Le nombre maximal de poses à sauvegarder est 500.

Les deux protocoles ont parfaitement fonctionné sur le cas 2QFO où il y a à la fois une forte interaction inter-fragment entre systèmes π et des affinités non négligeables pour des composés si simples (K_d<<1mM). La meilleure pose selon E_{pot} contient au minimum un fragment correctement placé, quelle que soit l'approche, pour trois autres complexes (2XDU, 2YEJ et 3HZ1). Toutefois, aucun des deux fragments n'est correctement prédit au niveau de la meilleure pose pour le complexe 2YEI avec la méthode multi-entités. Néanmoins, une configuration que l'on peut qualifier de proche de l'expérimentale (RMSD $_{fragment1}$ < 1,5 Å et RMSD $_{fragment2}$ < 1,5 Å) est systématiquement reproduite pour chaque complexe et ce pour les deux approches, avec une tendance plus prononcée en ce qui concerne le docking séquentiel dans la capacité à produire des RMSD et des rangs très faibles. Les résultats détaillés du docking simultané de plusieurs fragments sont disponibles dans le tableau 7 et la figure 5 de la publication (voir le §4.3).

En pratique, il est souhaitable d'employer les deux stratégies pour la prédiction prospective de la structure ternaire d'un complexe, car chacune possède ses propres avantages et inconvénients. Par exemple, la méthode multi-entités est plus rigoureuse mais également plus complexe (nombre plus élevé de DDL), mais a l'avantage d'avoir une phase post-traitement beaucoup plus légère que celle de la stratégie séquentielle qui souffre d'un problème combinatoire. Etant donné que le même système est simulé avec les mêmes DDL (au final), les résultats des deux approches sont directement fusionnables, ce qui est un dernier argument en faveur du lancement parallèle des deux stratégies.

Bien que reposant sur un faible échantillon, une tendance se dégage entre l'affinité d'un fragment et la capacité à bien prédire son positionnement. Cette observation, assez intuitive, a déjà été évoquée dans l'étude de Verdonk *et al* portant sur des simulations de docking “mono-fragment” [195].

Ce bilan convaincant valide la capacité de S4MPLE à gérer plusieurs entités mobiles de type fragment, et ouvre des perspectives intéressantes quant à son utilisation au sein de projets FBDD. Enfin, ces résultats permettent de passer à la dernière phase du projet, à savoir le développement d'une stratégie d'optimisation virtuelle de fragments (voir le chapitre 5).

4.3 Article II

L'article II [197] est inséré dès la page suivante.

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

S4MPLE – Sampler For Multiple Protein–Ligand Entities: Simultaneous Docking of Several Entities

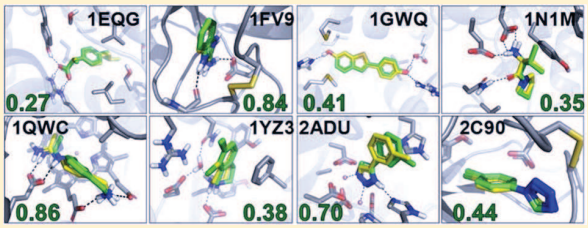
Laurent Hoffer[‡,§] and Dragos Horvath*[,‡]

[‡]Université de Strasbourg, 1 rue B. Pascal, Strasbourg 67000, France

[§]Novalix, BioParc, bld Sébastien Brant, BP 30170, Illkirch 67405 Cedex, France

Supporting Information

ABSTRACT: S4MPLE is a conformational sampling tool, based on a hybrid genetic algorithm, simulating one (conformer enumeration) or more molecules (docking). Energy calculations are based on the AMBER force field [Cornell et al. *J. Am. Chem. Soc.* **1995**, *117*, 5179.] for biological macromolecules and its generalized version GAFF [Wang et al. *J. Comput. Chem.* **2004**, *25*, 1157.] for ligands. This paper describes more advanced, specific applications of S4MPLE to problems more complex than classical redocking of drug-like compounds [Hoffer et al. *J. Mol. Graphics Modell.* **2012**, submitted for publication.]. Here, simultaneous docking of multiple entities is addressed in two different important contexts. First, simultaneous docking of two fragment-like ligands was attempted, as such ternary complexes are the basis of fragment-based drug design by linkage of the independent binders. As a preliminary, the capacity of S4MPLE to dock fragment-like compounds has been assessed, since this class of small probes used in fragment-based drug design covers a different chemical space than drug-like molecules. Herein reported success rates from fragments redocking are as good as classical benchmarking results on drug-like compounds (Astex Diverse Set [Hartshorn et al. *J. Med. Chem.* **2007**, *50*, 726.]). Then, S4MPLE is successfully challenged to predict locations of fragments involved in ternary complexes by means of multientity docking. Second, the key problem of predicting water-mediated interaction is addressed by considering explicit water molecules as additional entities to be docked in the presence of the "main" ligand. Blind prediction of solvent molecule positions, reproducing relevant ligand-water-site mediated interactions, is achieved in 76% cases over saved poses. S4MPLE was also successful to predict crystallographic water displacement by a therefore tailored functional group in the optimized ligand. However, water localization is an extremely delicate issue in terms of weighing of electrostatic and desolvation terms and also introduces a significant increase of required sampling efforts. Yet, the herein reported results – not making use of massively parallel deployment of the software – are very encouraging.



1. INTRODUCTION

Docking can be defined as the structural prediction of a complex (e.g., ligand-protein) at the atomic resolution scale. In principle, this is just a particular instance of the more general problem of conformational sampling: docking consists in the sampling of ligand geometry, of the intermolecular roto-translational degrees of freedom (DoF), plus optionally of some internal DoF of the binding site. In practice, however, usual docking simulations make an explicit distinction between the two different molecules classes (binding site and ligands). A lot of methods, using various algorithms, have been developed to address the docking problem.[5] Most of them expect two distinct molecular input files – the binding site, in some biomolecule-compatible format (.pdb, .mol2, *etc.*) and a ligand, in a file explicitly providing bond order information (.mol2, .sdf, *etc.*). Exceptions to this classical scheme (protein–protein docking, self-assembly of small organic molecules) are usually not supported by main-stream docking programs and are often handled by dedicated soft.[6,7] Furthermore, classical docking tools all include two key steps: the (typically force-field driven) sampling of the DoF of the system and the scoring/ranking of obtained poses. The latter step implicitly assumes the receptor to be a protein, binding one single ligand – typically drug-like for the scoring functions were trained under these assumptions. Therefore, such scoring functions are often misleading when applied to small, fragment-like molecules[8] or in docking scenarios conserving explicit site waters.

In fact, molecular recognition phenomena are never a two-body problem: involved partners are solvated. Solvent effects are known to be very complex – they are extremely hard to model – and have a major role in the free energy of binding.[9] Specific terms, accounting among other for the desolvation phenomenon of ligand and binding site, have been developed to enhance the accuracy of scoring functions.[10,11] While solvation is increasingly taken into account, the critical aspect of handling displaceable crystallographic water molecules in the binding site, putatively involved in water-mediated hydrogen bonds between the ligand and the target,[12,13] is not properly handled by classical docking tools, which cannot cope with

Received: October 15, 2012
Published: December 10, 2012

multiple mobile entities in the protein site. It has been shown that conserving tightly bound waters in simulations significantly enhances the accuracy of cross-docking results[14] – yet, the nontrivial question is which waters to conserve. Huang et al. investigated the effect of including waters in massive virtual screening.[15] Different approaches, for example based on energy computations[16] or classifier algorithms,[17] tried to address this fundamental question by flagging waters as conserved or displaceable. Besides, some ligands may gain in binding energy by explicitly replacing poorly anchored crystallographic H_2O in order to directly interact with the target.

Specific approaches typically deal with these issues by introducing dedicated DoF piloting the status of water molecules. Much effort has been invested to determine the preferred location of water probes in protein sites.[18] Alternatively, docking approaches incorporate crystallographic water molecules in the usual docking process where, at least the oxygen atom is considered rigid, but the water particles can be turned ON/OFF during the simulation.[19,20] As for example, FlexX is able to both rotate waters in order to optimize the hydrogen bonds network and turn them ON/OFF if necessary. However, an alternative approach allowing explicit particles, modeling water molecules, to be introduced during the incremental ligand reconstruction step has been published by FlexX authors,[20] but it was not implemented in the commercial releases of the program.

Recently, Lie et al. reported[21] an interesting approach, implemented in the Molegro Virtual Docker (MVD) program, involving waters in docking. The ligand is surrounded by waters at every moment, and the solvent molecules can be toggled on/off in order to make a favorable direct contact with the site. Thus, the number of active waters can change during the process, and a simple ″entropic″ penalty is added to somehow compensate for the ″appearance″ of activated water molecules (i.e., addition of their interaction terms from the force field/scoring function). By contrast to most of the previously described algorithms, waters are not included into the binding site, they are not static since they move with the ligand, and fair results were obtained using a data set composed of 12 PDB complexes.

The sampling and docking tool S4MPLE (Sampler For Multiple Protein–Ligand Entities) advocates a completely generalized approach, treating docking as a particular case of sampling of multiple entities. It may search for the lowest-energy arrangements of arbitrary molecular and supramolecular systems, starting from a single molecule (ligand geometry sampling/small peptide folding) to site-ligand complexes (classical docking), to ternary (or even higher-order) systems, addressing both inter- and (user-controlled) intramolecular DoF. S4MPLE[3] is a molecular modeling tool based on a hybrid "Lamarckian" Genetic Algorithm (GA), combining local optimizations in addition to classical evolutionary sampling strategies. Allowing full control of the considered DoF, S4MPLE is a completely general approach to explore the conformational space of the flexible parts of the studied system. By default, all atoms of the system are considered flexible. An explicit list of fixed atoms – for example, binding site atoms – needs to be provided.

The energy function is based on the AMBER/GAFF force-field (FF) formalism. AMBER[22] is a popular FF used to simulate biological macromolecules. Recently, an extension to handle small compounds such as ligands has been reported: General AMBER Force Field (GAFF).[2] In addition to the classical *in vacuo* FF terms, S4MPLE includes a specific continuum solvation term of maximal simplicity, in order to minimize the computational overhead associated with it. The solvent model terms include simple functions (linearly distance-dependent relative dielectric constant and pair-based desolvation term) of interatomic distances of the same complexity as usual FF terms. Explicit solvent boxes,[23] widely used in molecular dynamics, are not compatible with evolutionary-based strategies involving large-scale random jumps in conformational space. Besides, a term rewarding favorable interactions such as hydrogen bonds and hydrophobic enclosure is included too. All these additional terms, their calibration/validation, and the genetic operators are described elsewhere.[3] This "Fit FF" leads to enhancement with respect to the "Core FF" in redocking experiments using the most popular docking data set (Astex Diverse Set[4]).

Occasional local optimization is mandatory for molecules during the evolutionary-based sampling procedure. This is due to the extreme ruggedness of the FF-based energy function. All genetic operators will include, as a last step, a small local optimization. A population diversity control mechanism, in the form of interaction fingerprints, is included in order to avoid the risk of premature convergence.

This work continues to explore the performances and putative applications of the single CPU workstation version of S4MPLE, focusing on multiple entity docking, and addressing two main key issues:

- the problem of water-mediated ligand binding. The implicit solvent terms model the shielding effect of water on electrostatics but not the water-mediated hydrogen bonds. Here, waters are treated as additional ligands competing for anchorage to the active site, and
- the issue of simultaneous docking of multiple fragment-like ligands in Fragment-Based Drug Discovery/Design (FBDD). FBDD emerged this past decade as a powerful way to identify fragment-hits which are subsequently evolved into high affinity lead compounds.[24] Any rigorous *in silico* FBDD strategy must include two main steps: prediction of fragment binding modes and their subsequent optimization/evolution into more potent ligands. Prior to simultaneous docking of multiple fragments, a single fragment redocking study was performed to estimate the ability of S4MPLE to cope with fragment-size ligands.

Poses are solely evaluated according to their AMBER/GAFF energies, without any additional scoring function based reranking (not straightforwardly applicable to multiple entity contexts, anyway). These multiple entity simulations proved to be more complex and time-consuming that the standard single-ligand benchmark previously reported. Nevertheless, S4MPLE successfully managed to predict both water-mediated interactions and multiple fragment binding modes, within the classical limitations due to FF accuracy.

2. METHODS

2.1. Brief Overview of S4MPLE. As already mentioned, a detailed technical description of the GA-driven sampling procedure and benchmarking studies with respect to classical docking problems were submitted for publication elsewhere.[3] Here, the original features of S4MPLE will be briefly mentioned in order to allow a better understanding of the current results.

S4MPLE is a completely general approach to visit the conformational space of arbitrary molecules or complexes. As such, it may be equally well used for conformational sampling and docking – which is nothing but sampling of a ligand in the

dx.doi.org/10.1021/ci300495r | J. Chem. Inf. Model. 2013, 53, 88–102

presence of a binding site. The 'site' does not need to be a protein, which may eventually render S4MPLE useful for simulations of arbitrary molecule self-assembly processes. It was conceived in view of large-scale deployment on computer grids, but the current paper describes the workstation version of the tool. Its applicability is only limited by (a) the studied system size vs available computational resources and (b) the availability of force field parameters for the studied molecules. The program is written in object-Pascal and used in command-line mode.

2.1.1. Force Field. The empirical force field (FF) used by S4MPLE in the current work is an adapted AMBER[22]/GAFF[2] parameter set, including a continuum solvent model, based on simple functions of interatomic distances:

- A linearly distance-dependent relative dielectric constant is used in the Coulomb term, which *de facto* makes it a function of $1/d^2$.
- A pair-based desolvation term[25] in $1/d^4$, function of the squares of partial charges (with tunable proportionality constant).
- Contact terms,[25] rewarding favorable interactions such as hydrophobic contacts and hydrogen bonds.

Since these terms are not "native" to AMBER/GAFF, the additional parameters needed calibration, as described in ref 3. The resulting "Fit FF" was used in the present work. Benchmarking of fragment-like ligand docking was also performed, for comparative reasons, with the plain AMBER/GAFF formalism (without desolvation, without contact terms), herein referred to as "Core FF".

2.1.2. General, Chemically Meaningful Genetic Operators. The ability of S4MPLE to indiscriminately handle intra- and intermolecular degrees of freedom is the key to sample ensembles including an arbitrary number of independent species and thus be applicable as classical conformational sampler, docking program, and multiple ligand docking tool. This ability is achieved through the appropriate design of the genetic operators, which in typical docking approaches focus on torsional angles to control intramolecular and on roto-translational DoF. In S4MPLE, a genetic operator works on some randomly picked molecular substructure, which may be either covalently connected or not. If connected, the operator will perform movements all while respecting the covalent constraints (bond length, valence angles). If not − the substructure in question being, for example, one of the ligands competing for a binding pocket of the active site − the guidance role of the absent covalent bond is taken over by a putatively favorable contact axis, randomly chosen as a pair of atoms (one in the substructure, one being an external partner) that shall be brought together in order to form a hydrophobic or hydrogen bonding contact. When constrained by covalent bonding distance, operators would basically perform some rotation around the covalent bond (directed mutation) or swapping of substructures between parents (crossing-overs). With nonbonded substructures, operators will follow the optimal contact distance, returning basically random (and yet as clash-free as possible) poses featuring this randomly chosen contact. Please refer to the technical article previously mentioned for detailed information.

2.1.3. Population Diversity Control: Interaction Fingerprints. S4MPLE adopts the postulate that two geometries may be considered redundant if they share a same set of contacts. This postulate is embodied by novel, fuzzy, and differentiable Pairwise Interaction Fingerprints (PIF). Unlike classical ligand-protein IF used in docking,[8,26,27] PIFs are

- general, regrouping both intra- and intermolecular favorable contacts: hydrogen bonds and hydrophobic contacts
- symmetry-compliant, i.e. invariant to swapping of the contact status of topologically equivalent atoms.
- smooth and differentiable, rather than binary: contact status varies smoothly between "absent" (0) and "fully established" (1.0) as a function of contact distance.

A geometry is characterized by its interaction fingerprint, and fingerprint dissimilarity scores are used to assess geometry redundancy, by means of a user-defined dissimilarity threshold.

2.2. Data Sets. *2.2.1. Fragments Set.* All complexes featuring fragment-like ligands from the Astex Diverse set,[4] Astex/CCDC clean subset (no covalent ligand),[28] and from Congreve's review[29] are merged into the fragment data set. A compound is defined as a fragment if its mass is below 300 Da. This mass threshold is used in the rules of three, a set of empirical rules to define fragment-like compounds.[30] The total number of complexes, after removing one PDB duplicate (1N1M), is 142. This complex is only included into the smaller set (Congreve subset) before computing the statistics for each complexes sources (subsets).

2.2.2. Ligand-Water Set. A subset of Astex Diverse set, composed of 16 complexes in which water-mediated interactions play a key role, is used to test the multientities docking ability of S4MPLE. In this run, the ligand plus one or several explicit water molecules (considered as additional ligands) are docked, in an attempt to correctly position both ligand and waters as the lowest-energy configuration. These examples include, unsurprisingly, many complexes for which standard water-free redocking failed[3] (RMSD of top-ranked pose > 2 Å). All investigated PDB complexes are listed in Table 1. Additionally, S4MPLE has been challenged to correctly predict the position of a crystallographic water in the poly(ADP-ribose) polymerase (PARP) structure, which is alternatively seen to either mediate a ligand-site interaction

Table 1. 16 PDB Complexes Selected for the Simulations with Free Explicit Waters[b]

PDB	number of free waters	water IDs
1G9V	3	9596 9209 9269
1GM8	2	13899 13905
1GPK	1	9154
1HVY	6	20311 20320 20344 20383 20461 20689
1LPZ	1	5166
1MEH	1	5560
1N2J	2	8735 8978
1N2V	4	6547 6550 6556 6559
1OPK	2	7964 8009
1P2Y	1	6511
1Q41	1	11312
1SQ5	3	5012 5123 5126
1T9B	2	18426 21222
1XM6	1	5563 5488[a]
2BR1	1	5381
2BSM	3	3687 4044 4047

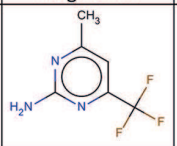
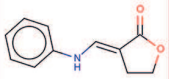
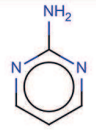
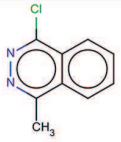
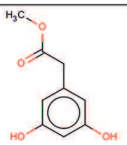
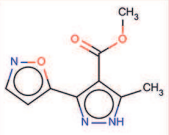
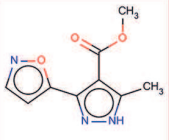
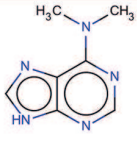
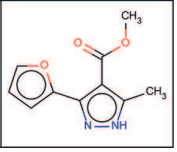
[a]An additional water (fixed oxygen) is included since it coordinates a metal ion in the vicinity of the ligand. [b]Misdocked complexes in classical water-free redocking simulations are displayed in bold.

dx.doi.org/10.1021/ci300495r | J. Chem. Inf. Model. 2013, 53, 88−102

(3PAX complex) or is displaced by a larger ligand making a direct interaction with the site (2PAX complex).

2.2.3. Multifragments Set. The target Heat Shock Protein 90 (Hsp90) has been intensively studied within FBDD projects,[31−36] and a large number of fragment-protein complexes have been deposited in the PDB. Among them, there are 5 ternary complexes (2QFO,[35] 2XDU,[32] 2YEI/2YEJ,[36] and 3HZ1,[31] see Table 2) which include two

Table 2. 5 Ternary Hsp90 Complexes Selected for the Multifragments Docking Simulations

PDB	Fragment 1	Fragment 2
2QFO		
2XDU		
2YEI		
2YEJ		
3HZ1		

noncompetitive fragments, other than cocrystallized solvent and small organic compounds such as glycerol, *etc.* S4MPLE has been challenged to reproduce binding modes of the simultaneous binders for these five ternary complexes.

2.3. Redocking Protocol Using S4MPLE. The preparation of the ligands consists of several steps: computing partial charges (Gasteiger type) and generating a single conformer using ChemAxon libraries,[37] adding GAFF atomic types with the "antechamber" tool,[38] and employing the "parmchk" tool[38] to check whether there are missing parameters (e.g., bonds, angles, or torsions). The conformer generation step avoids starting from the expected solution, thus the docking accuracy is not artificially enhanced. Binding sites are prepared using MOE[39] and its Protonate3D protocol directly from PDB files. Partial charges and atom types from protein atoms are assigned from the AMBER topology file during the initialization of the program. Binding sites are defined as subdomains including only residues with at least one atom at 10 Å or less with respect to reference compound(s). For fragment set redocking benchmarks, hot spots atoms, preferentially used to anchor the ligand, are automatically picked, by detecting the putative hydrophobic and hydrogen bonding centers in the close neighborhood of the binding site center. Cofactors and ions are included in the binding site, and all waters are removed. During the fragment set redocking benchmarks, all binding site atoms are fixed, by contrast to all other simulations, where polar hydrogens of hydroxyl groups are flexible in order to dynamically optimize the hydrogen bond network.

The redocking simulations of fragment set complexes consisted in 5-fold runs. Each run took 500 generations of 50 individuals, with default mutation and crossing-over probabilities, and the conformational redundancy parameter *minfpdiff* = 0.01. Each simulation can be described by two stages: the first consists in the docking process itself, where poses with the lowest potential energy are sought and saved. The second stage belongs to postprocessing category: there is an optimization cycle of selected poses (nonredundant configurations within an energy window +30 kcal/mol with respect to best one) from the previous stage, and the best poses are saved (maximum 30). Success rate (percentage of correctly predicted complex geometries, at given RMSD level) is then reported as the *average* of success rates of each independent run. Unless mentioned, this preparation workflow and the associated parameters are used for the other simulations too.

2.4. Redocking Protocol Using FlexX. FlexX is a very popular docking tool.[40] At the moment, it is included within the LeadIT (version 2.0.2) platform software developed by BioSolveIT. The redocking of the fragment set is performed with FlexX too, for benchmarking purposes. LeadIT is used to prepare a binding site from the PDB files: hydrogens are added, cofactors and metals located near fragment of interest are preserved, and all water molecules are removed. Residue protonation states and orientation of polar hydrogens are manually fixed if necessary. FlexX binding sites include all full residues with at least one atom within a distance of up to 6.5 Å with respect to the reference fragment.

Default values for the entire preparation workflow and docking process are used, except for fragment protonation states and metals where pharmacophoric restraints are disabled. Concerning the protonation states, those defined for S4MPLE runs are employed. For FlexX docking, success rates rely on RMSD listed within LeadIT.

2.5. Multientity Docking Simulations. Due to the completely general and unified treatment of the DoF, S4MPLE does not theoretically limit the number and the status of entities. This has several advantages and allows modeling of competitive or noncompetitive binders with respect to the same site. Unlike in multiple-copy approaches like MCSS,[41] each compound interacts with all others as in usual FF implementations. Technically, S4MPLE reads a multimolecule SDF file and produces relative arrangements of various conformers of all the herein present compounds. Stoichiometry may be controlled by explicitly dupli(multipli)-cating concerned molecules in the input file. Since storage of a protein site in SDF format is cumbersome, alternative formats (MOL2 or CAR) are used for reading/writing biological macromolecules and their complexes.

Two main types of multiple entity simulations are performed:

- simultaneously docking of one ligand and one (or more) explicit waters
- simultaneously docking of two noncompetitive fragments

2.5.1. Docking of One Ligand and Free Explicit Waters. In the herein work, waters can be allowed to move freely, being modeled like ligands (but not appear or disappear). The only specific treatment reserved to H_2O is the use of TIP3P charges

dx.doi.org/10.1021/ci300495r | J. Chem. Inf. Model. 2013, 53, 88−102

instead of the Gasteiger charges customarily employed with ligands. The challenge here consisted of checking whether addition of water molecules, in the absence of any information of their crystallographic positions, would allow it to predict the ligand-water-site interaction network, in correctly locating both ligand and at least some of the added water molecules. The explicit waters are also embedded into the continuum solvent model − their presence is necessary to highlight specific interactions that cannot be reduced to simple dielectric shielding. By default, S4MPLE treats all the DoF equally and pays no special attention to the "real" ligand vs accompanying water molecules. However, docking of a ligand plus few H_2O molecules in an active site, which is completely stripped of crystallographic waters, may not necessarily find the water-mediated ligand-site interactions. The explicit waters might well be captured in significant site-water-site interactions. This makes it difficult to predict the number of explicit waters that must be added in order to solve the problem. Since the goal here is to try to reproduce experimentally observed binding modes, the experimental structure was used to answer this question. Various scenarios, featuring from 1 to 6 explicit waters, have been tested depending on the complex of interest. An addition of more water particles amounts to an increase in DoF − a significant one, if the site is frozen and the ligand itself is relatively small. Therefore, some general and some specific strategies have been developed for ligand+water docking:

- protocol A: usual settings, 2000 generations,
- protocol B: usual settings, 5000 generations,
- protocol C: protocol A, with a temporary bias toward water-ligand interactions,
- protocol D: protocol C, followed by a specific postprocessing refinement of water positions around the kept ligand poses.

The bias in protocol C is introduced in order to artificially enhance the chance of discovering poses with ligand-water-site interactions over site-water-site interactions, as hinted above. Recall that genetic operators will tentatively place waters in the neighborhood of randomly picked hydrogen bond donor and acceptor "hot spots", which may belong either to the protein site or to the ligand. Site hot spots are automatically defined according to the same procedure used in the default redocking study,[3] such as to pick acceptors/donors situated close to the bottom of the site. Any putative donor/acceptor of the ligand automatically counts as "hot spot". By default, hot spots are identically treated by the genetic operators, irrespectively of their origin. The bias consists in

- automatically removing putative site-water hydrogen bonds from the list of favorable contacts used by the crossover procedures, and
- specifically scaling up hydrogen bond interactions between waters and the actual ligand by a factor of 2 (with respect to default FF settings). After the sampling stage, this energy bias is removed, during a postprocessing stage reranking poses according to their unbiased Fit FF energy levels.

Protocol D includes a postprocessing step meant to "shake up" the suboptimally positioned waters at the docking step. Applied to each of the kept poses, this protocol allows the docked water configurations to be challenged by different water poses, evolved during 100 generations, from a population including the docked pose as an initial individual. During this procedure, the ligand atoms are assigned passive status, i.e. they are ignored by all the genetic operators except local gradient-based optimizations. As in default docking, all these protocols (A−D) are followed by a final relaxation of all so-far kept binding modes, after the pruning of worst and redundant poses.

2.5.2. Docking of One Ligand and Displacing a Water Molecule. The analysis of X-ray complexes often highlights water-mediated hydrogen bonds between the ligand and the binding site. A subsequent ligand optimization can be a modification in order to displace the mediating water (especially if it appears poorly anchored), allowing a direct interaction between the ligand and the target. The ability of S4MPLE to predict this kind of effect is assessed, based on experimental data of this nature: two PDB complexes of the poly(ADP-ribose) polymerase, including distinct bound fragments, are used (2PAX/3PAX).[42] In the 3PAX structure, the methoxybenzamide compound makes three direct hydrogen bonds with the binding site (two with G863 and one with S904) and one water-mediated with E988. In the 2PAX structure, the amino-naphthalimide molecule adopts a similar binding mode but with a direct hydrogen bond with E988. The naphthalimide compound can be described as a slight optimization of the methoxybenzamide fragment in order to maintain the amide group into its optimal configuration through cyclization and to displace the water while maintaining the polar interaction with the E988 side chain.

In a first step, S4MPLE is challenged to predict the location of studied fragments into their own binding site (redocking) with one free water in both cases. In the ideal case, it will find the correct binding mode of fragments and the good location of the free water in the 3PAX structure. Since the water does not mediate the hydrogen-bond between the fragment and the target in the 2PAX structure, it should be docked in another area of the site (e.g., site-water-site interaction). Simulations were performed according to both above-mentioned protocols C and D, in order to verify whether the "aggressive" favoring of water-ligand interactions would not prevent the eviction of the water mediated interaction in 2PAX.

Additionally, cross-docking simulations were also performed, in order to ensure that redocking success was not conditioned by the (very small) changes in active site geometries.

2.5.3. Docking of Two Organic fragments. The aim of this study is to evaluate the ability of S4MPLE to predict the simultaneous locations of two compounds within a given binding site. By contrast to previous multientities simulations involving one ligand and explicit solvent molecules, two "real" ligands are docked here. The investigated protein is the molecular chaperone Heat Shock Protein 90 through several ternary PDB complexes containing each two bound fragments. For convenient reasons, the first fragment of a given PDB code (XXXX) will be referred as XXXX_1 and the second as XXXX_2.

Two distinct multientity docking protocols (simultaneous and sequential) have been investigated. The simultaneous approach may be more rigorous than docking of individual known fragment binders, because it may implicitly account for any potential competition of different compounds for the same binding pocket and for favorable interfragment interactions. Nevertheless, the number of DoF quickly increases when simulating more than one compound; therefore, a higher number of generations must be allowed. In practice, 8 independent runs are performed and subsequently followed by the merging of all poses before the final relaxation stage. The number of generations is set to 2000. A maximum of 500 poses and a *minfpdiff* value of 0.004 are employed. All the other parameters use previously described values (see §2.3). Binding

sites are prepared as usual, including all residues within a radius of 10 Å around the considered reference (here two distinct fragments). Conserved waters, over a large number of X-ray structures and mediating ligand-site interactions around the hot spot D93 (4 H_2O in complexes 2QFO, 2XDU, 2YEJ, 3HZ1 and 3 H_2O in 2YEI), are included into the binding site too.

Alternatively, sequential docking decouples the exploration of DoF of involved ligands, i.e. breaks the general simultaneous sampling down to two sequential analyses of each ligand alone, and then tries to find combinations of poses that are mutually compatible. In the first step of the sequential docking, fragments are individually docked into their binding site with the same parameters as in fragment redocking benchmark (see §2.3). In a second step, each kept pose of fragment A is automatically associated with anyone of fragment B in order to generate all the combinations. The last step consists of the usual poses filtering, with a larger energy window (*excess_energy* = 20000 kcal/mol), allowing several slight overlaps – but not competitive poses for the same subpocket, in which both fragments significantly overlap. Selected combinations are eventually energy-minimized, putatively allowing the initially tolerated bad contacts to rearrange into favorable interactions between the binding fragments.

3. RESULTS AND DISCUSSION

3.1. Redocking of Fragments. The ability of S4MPLE to dock fragment-like compounds is evaluated using usual redocking simulations. It should be noted that these molecules can be very challenging since they can adopt a huge number of binding modes, because of their small size. Thus the scoring function must be able to highlight the good binding mode out of a lot of wrong poses, generally with scores/energies in a pretty small window. At the opposite, the selection of the correct binding mode of larger ligands is paradoxically easier: albeit the problem space volume may be much larger, the total number of clash-free poses within this volume is relatively small, as there are not so many ways to make the relatively large ligand fit into the defined binding site. As far as the issue of sampling these clash-free poses is solved, the scoring function must only be able to discriminate the correct binding mode among a smaller number of decoy conformations. However, a recent study[43] concluded that there is no significant difference in docking accuracies of "usual" ligands vs fragments, but authors highlighted that the observed failures stemmed from different causes as outlined above (scoring vs sampling issues). The authors also demonstrated that high Ligand-Efficiency (LE, defined as the affinity divided by the number of heavy atoms[44]) fragments are also docked with more accuracy – this can be understood intuitively, but a rigorous validation is still required. In S4MPLE, scoring is consistently performed in terms of the force field energy function used at the sampling stage, so artifacts due to fragment-like binders being outside the applicability domain of drug-like compound specific scoring functions are not expected.

The fragments redocking results with S4MPLE are summarized in Figure 1, Figure 2, and Table 3. Full details about selected complexes and each run are also available in the Supporting Information. Success rates are quite high, and the Fit FF significantly outperforms the Core FF energy function (respectively 86% and 77% at 2 Å for the top-ranked pose), as in previous benchmark studies.[3] By contrast, the redocking using FlexX only led to 72% for the same criteria. However, by definition, these compounds are pretty small and do not make a

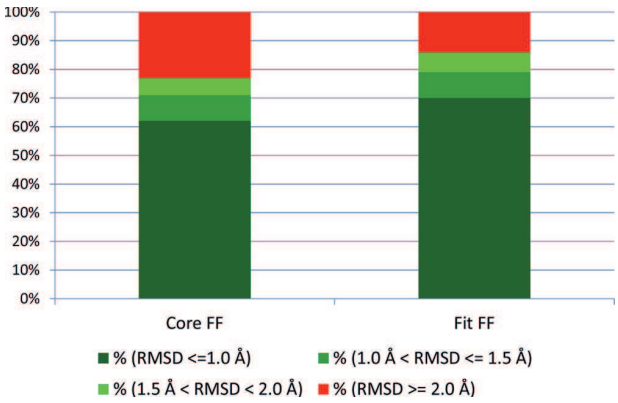


Figure 1. Docking performances (% complexes with RMSD of the top-ranked pose in a given range) over the full fragment set, for both Core FF and Fit FF energy schemes.

lot of interactions with the binding site compared to larger ligands. The usual RMSD threshold (2 Å) can be misleading: using an alternative metric[8] or more stringent thresholds (e.g., 1.5 Å or even 1 Å)[43] should be more appropriate to discard failures from successes. A noteworthy point is that the success rate only decreases of 7% by lowering the RMSD threshold from 2 Å to 1.5 Å for the Fit FF mode (79% accuracy). Besides, at the stringent 1 Å limit, the accuracy is still a quite high 70%. Expectedly, there are no sampling problems with these smaller compounds: the expected binding mode is quasi-systematically found among the saved poses (around 98%). Many (93) of these complexes (source: Astex/CCDC) were present in the calibration protocol of the Fit FF:[3] the consequence could be an artificially bias toward the experimental solution. However, the other complexes not used for FF fitting (Astex Diverse Set and complexes from the Congreve's review) show equivalent or even better success rates. Indeed, most failure cases come from the Astex/CCDC-clean-subset itself for both Core FF and Fit FF.

Several papers focusing on fragments docking were published these last few years.[8,43,45−47] Among them, complexes from Congreve's review[29] are often used as a reference data set.[45,46] Although small (12 PDB complexes – it was not designed for benchmarking), the latter covers diverse targets, a variety of interactions (purely van der Waals, hydrogen bonds, ionic bonds, or even metal coordination), and a wide affinity range (from sub-nM to high-mM). A comparative table (Table 4) summarizes the redocking results of S4MPLE and those found in the literature for these 12 complexes. 1QWC did not systematically converge toward the expected solution using Fit FF, hence the slightly lower average success rate. Figure 2 shows the superimposition of both experimental and predicted binding modes (using the Fit FF scoring scheme) and the polar interactions between fragments and the binding pocket. Haider et al. used two popular tools (GOLD[48] and MCSS[41,49,50]) to dock these fragments into their targets.[46] Besides, they investigated the impact of a rigorous rescoring scheme based on a Generalized Born Surface Area (GBSA) solvation model. Their results show a global success rate of 67% and 75% in the strategy based on GOLD, respectively without/with the GBSA rescoring scheme. At the opposite, Loving et al. described a perfect accuracy (100%) over this data set,[45] using Glide as docking engine.[51] However it should be noted that all waters around fragments are systematically included in the binding site

dx.doi.org/10.1021/ci300495r | J. Chem. Inf. Model. 2013, 53, 88−102

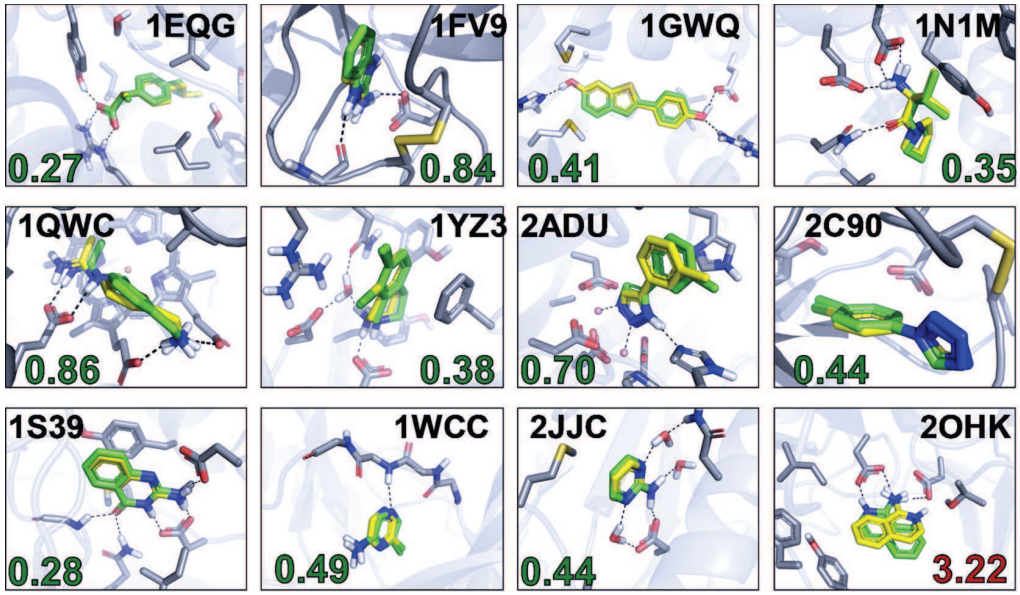


Figure 2. Top-ranked poses from redocking simulations of Congreve's data set, using the Fit FF energy function (RMSD with respect to X-ray coordinates are provided for each complex). Experimental and docked locations are represented in green and yellow, respectively. Main polar interactions are depicted as dash lines.

Table 3. Docking Performance of Fragment-Like Compounds (Full Fragment Set and Its Subsets) Using S4MPLE[a]

data set (size)	energy function	success rate (%) top-ranked pose	success rate (%) top 30 poses
full fragment set (142)	Core FF	77	98
	Fit FF	86	98
Congreve subset (12)	Core FF	92	100
	Fit FF	88	95
Astex Diverse subset (37)	Core FF	84	97
	Fit FF	90	99
CCDC clean subset (93)	Core FF	72	97
	Fit FF	84	98

[a]Success rates, defined as the percentage of correctly predicted geometries with RMSD < 2 Å for the top-ranked pose, are reported as the average success rates of the 5 independent runs.

Table 4. Docking Performance of Several Docking Tools on the Congreve Data Set[a]

docking program	inclusion of crystallographic waters	success rate for the top-ranked pose (%)
S4MPLE (Core FF)	2JJC	92
S4MPLE (Fit FF)	2JJC	88
GLIDE[45]	all sites	100
MCSS[46]	2JJC/1YZ3	67
MCSS (GBSA rescoring)[46]	2JJC/1YZ3	67
GOLD[46]	2JJC/1YZ3	67
GOLD (GBSA rescoring)[46]	2JJC/1YZ3	75

[a]SAMPLE success rates are reported as the average of success rates of the 5 independent runs, while other values are directly extracted from the literature. Binding sites in which crystallographic waters were not removed during the docking simulations are listed.

(distance threshold of 5 Å around the reference fragment) and certainly channel the ligand toward its expected position. Our results are good (one failure out of the twelve cases) using water-free binding sites, except for the Hsp90 2JJC complex, where waters around the D93 residue are known to be conserved over a huge number of PDB complexes and absolutely needed to explain the binding of a fragment mostly interacting with the site through water-mediated bridges. Surprisingly, the failure is not the same for both FF setup schemes: 1WCC for Core FF and 2OHK for Fit FF, see Figure 2. In the last case, the main interactions are globally conserved, but there is a flip of the isoquinolin cycle on itself.

As a final remark, it is interesting to point out that success rates obtained with fragment-like compounds are similar to those concerning a data set mainly composed of drug-like molecules,[3] as very recently described by Verdonk et al.[43]

3.2. Docking of One Ligand and Free Explicit Waters.

The results of these multiligands simulations, depicted in Table 5, are analyzed according to several criteria:

- "top-ranked pose" and "over saved poses": ability to accurately reproduce the experimental binding mode, for the top-ranked pose and, respectively, within one of the saved ones (not necessarily the most stable) and
- "over all entities": ability to visit at least once the expected location for every entity (ligand and water) over the whole set of saved poses

The heavy-atom RMSD is employed as metric to discard successes from failures, and two different thresholds are monitored (1 Å and 2 Å). By default, 2 Å is employed. However the more stringent value has its importance too, since a stricter limit should be more consistent with waters (only one heavy atom). Indeed a RMSD of 2 Å can be accepted for drug-like ligand poses, but not H_2O poses, for it does not guarantee the preservation of hydrogen bonds.

The first two result columns "top-ranked pose" in Table 5 refer to the best-energy configuration retrieved by each protocol. Column "lig" reports the percentage of complexes

dx.doi.org/10.1021/ci300495r | J. Chem. Inf. Model. 2013, 53, 88−102

Table 5. Results for Each Protocol (A–D) from Ligand-Water Docking Simulations[a]

	% (RMSD < 2 Å)					
	top-ranked pose		over saved poses		over all entities	
protocols	lig	waters	lig	waters	lig	waters
protocol A	56	15	94	29	94	74
protocol B	56	9	100	32	100	74
protocol C	56	26	100	41	100	71
protocol D	69	41	94	76	94	94
	% (RMSD < 1 Å)					
	top-ranked pose		over saved poses		over all entities	
protocols	lig	waters	lig	waters	lig	waters
protocol A	31	9	81	18	88	56
protocol B	25	6	75	21	88	62
protocol C	38	15	81	32	81	62
protocol D	56	38	81	62	88	94

[a]For each considered criteria, values are divided into two subcategories (respectively for ligands and waters). Statistics for two different RMSD thresholds (1 and 2 Å) are compiled.

in which the ligand entity has been placed within the given RMSD threshold with respect to experiment: 56% means that in 9 complexes out of 16, protocol A (see §2.5.1) managed to place the ligand within 2 Å in the most stable pose. Column "waters" must be understood in the context of a variable number of added H_2O entities in each complex. If in a complex with W water molecules only w are within 2 Å from expected position, this complex will count only as a partial success: the success counter is incremented by w/W. The reported percentage represents thus the sum of all w/W indices, over the 16 considered complexes and may be alternatively read as the percentage of correctly placed water molecules. The next two columns "over saved poses" refer no longer to the top-ranked pose but to the geometrically best saved pose, featuring a correctly positioned ligand and as many well located waters as possible. This means that, for example, protocol A managed to visit at least one state with correctly positioned ligand for 15 out of 16 complexes (94%), at RMSD < 2 Å. However, over these states, only 29% of added water molecules also converged toward experimental positions. Eventually, the last two result columns "over all entities" refer to the overall ability to visit at least once each of the experimentally relevant positions of ligands and waters. Within the low-energy poses that were saved in protocol A, 74% of water molecules are at least once placed where they are expected – however, not simultaneously (some conformers witness a correct position of water 1, others of water 2, *etc.*).

It must be noted that this data set contains an important number of complexes which failed in water-free redocking simulations (7/16, see Table 1 and the associate study).[3] These failures should now be fixed, in as far as they were actually caused by the lack of water molecules. However, if different sources of errors affect those compounds (typical FF inaccuracies), then any errors in ligands placement will obviously prevent waters to be correctly located. Figure 3 shows the main scenarios which can be obtained in this kind of simulation (see Supplementary Information for details about all complexes):

a) successful placement of all entities (ligand and H_2O) for the top-ranked pose,

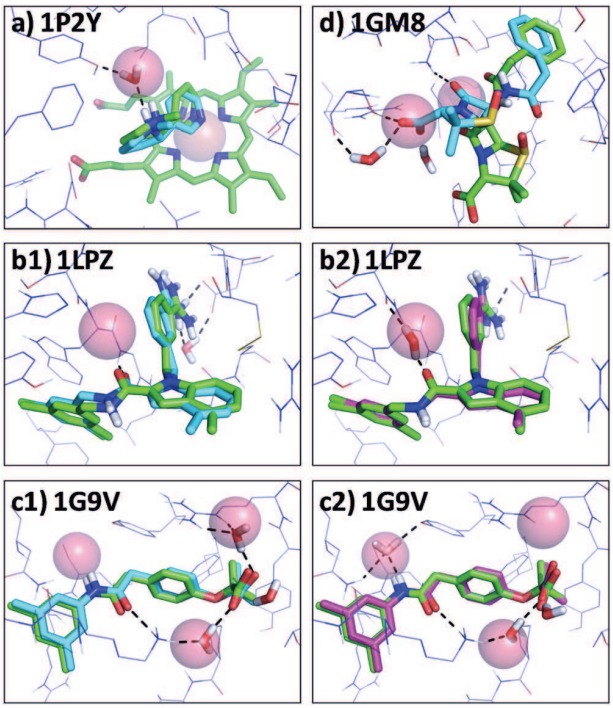


Figure 3. Examples of scenarios obtained in multientities docking simulations with free waters using protocol D: a.) success in prediction of all locations of ligand and water(s) for top-ranked (1P2Y); b.) experimental geometry found among saved poses – other than first (1LPZ); c.) partial success (correct positioning of several entities, including ligand, and at least one water, 1G9V); and d.) no acceptable solution (1GM8). Experimental ligand conformer is in green, while expected water locations are red spheres. Predicted poses are represented in blue (if top-ranked) or purple (other rank) for ligand, and waters are represented as stick models. Main interactions are depicted as dash lines.

b) successful placement of all entities, but for a higher energy pose,

c) partial success: ligand and some but not all waters simultaneously and correctly located (e.g., over all poses),

d) full failure: there is no acceptable solution for the ligand and at least one water molecule over the whole set of poses.

Results of each docking protocol (see §2.5.1) will be briefly analyzed in the following:

3.2.1. Protocols A and B: Usual Docking Settings with 2000/5000 Generations. Although the number of generations is critical (shorter simulations lead to expectedly bad results, data not shown), going for 5000 generations, instead of 2000, does not improve the poor success rates with respect to the three RMSD criteria. This is particularly true for the ability to simultaneously reproduce the expected binding mode of all entities within a same pose – preferentially the most stable one. The top-ranked statistics are equal to previous water-free redocking results for ligands (56% success rate), while the water molecules are poorly positioned (15%). Within the entire set of poses, the binding mode of ligands is well retrieved (94% and 100%), but the prediction rate of water-mediated interaction patterns remains low (≈30%). However, the actual hot spots involved in water-mediated interactions are indeed attractors in which water molecules are often found: about 3/4 of these locations are visited by waters for these two protocols. Logically, complexes with a lot of free waters (e.g., 1HVY,

dx.doi.org/10.1021/ci300495r | J. Chem. Inf. Model. 2013, 53, 88–102

1G9V, 1GM8) are, for pure probabilistic reasons, much more difficult to predict and lower the overall statistics, whereas simpler systems with a single additional water are well predicted (e.g., 1LPZ, 1P2Y). These results show that the sampling procedure is quite effective, in the sense that independent moieties do indeed visit their expected positions − however, the consensual event (all moieties being at their correct position at the same time) is too rare using default settings to reproducibly occur during a span of 5000 generations only.

In as far as the (energy-minimized) experimental pose is not the most stable in terms of its computed FF energy, failing to find it cannot be held against the sampling procedure − in this case, the expected geometry would be perceived as some arbitrary intermediately stable state. In this context, such failure may not necessarily be interpreted as a straightforward FF parametrization problem. It is not clear upfront that an explicit water molecule, in the presence of protein site, ligand and implicit solvent, would spontaneously favor the ligand-water-site mediator spot, for many other crystallographic waters have been removed during site preparations, and only a few have been reinstated during these simulations. The FF model relies on a fine balance between implicit desolvation and interactions with explicit water molecules. Since continuum desolvation is included in the energy function, the energy bonus due to water-mediated bridges is small, for it includes the desolvation penalty of the polar groups. Two water molecules embedded in a perfect solvent model should, on the average, not seek to make contact to each other, for the favorable Coulomb terms should be canceled out by the desolvation contribution. Unilateral site-water or ligand-water contacts may be only slightly beneficial, if the corresponding hydrogen bonding strength exceeds typical water−water interactions, or if water binding at a given hot spot is entropically favored. It is unclear to what extent FF calculations are accurate enough to reproduce the subtle balance of conflicting energy terms (desolvation penalty, Coulomb interaction, and hydrogen bond bonus) involved in the process.

On the opposite, if the expected configurations are the most stable ones, or energetically very close to the absolute minimum and were nevertheless not sampled, this means that the herein envisaged sampling problems were much harder than standard flexible ligand docking − and this, albeit the number of additional DoF does not significantly increase. It is the nature of the landscape, however, which may change. Water-mediated interaction contributions in the presence of an implicit desolvation term being small, the landscape may feature many possible local minima of rather equivalent depths, and not separated by any significant barriers − flat and shallow energy landscapes are notoriously difficult to sample. Furthermore, GAs are effective sampling algorithms due to their ability to inherit favorable "traits" of the ancestors. In flexible ligand docking, if a correct ligand-site anchoring point is found, this interaction is a favorable "trait" because it provides local stabilization. Individuals containing it will have, statistically speaking, larger chances of survival, and the well-formed contact will entropically enhance other hot spots, topologically close on the ligand, to approach their corresponding anchoring points in the site. Docking is locally cooperative − very much like folding of a protein α-helix. By contrast, in multiple docking with several interacting species, it does not pay off to correctly position a water molecule unless the ligand is also in place. In other words, an energy bonus due to the site-water-ligand bridge will only emerge if both water and ligand are well positioned. Otherwise, an individual featuring the correct site-water contact will not be preferentially selected with respect to other site-water contact featuring competitors, which cannot lead to the proper ternary configuration. Likewise,[52] folding of β-sheets is more difficult than folding of α-helices: a state with one strand in β-sheet configuration is not energetically favored before the complementary half is formed and zipped into its final position. There is no local cooperativity in ligand-water-site docking either. Note that simultaneous docking of multiple fragments binding to different subpockets, i.e. not (significantly) interacting with each other in the lowest-energy poses would not suffer from such an entropic bottleneck (§3.4). The correct placement of one fragment in its subpocket does pay off, irrespectively whether the other fragments are well-located or not (in as far as they are not clashing).

In order to understand whether the poor success rates above are due to (a) suboptimal binding energies, inclusively due to FF inaccuracies or rather due to (b) difficult sampling cause by lacking local cooperativity, two additional docking protocols were envisaged:

3.2.2. Protocol C: Protocol A + Energetic Bias in Favor of Water-Ligand Interactions. In this strategy, a (temporary, sampling stage-only) bias toward water-ligand hydrogen bonds is automatically added in order to decrease the frequencies of site-water-site interactions, which are not interesting in the context of this study. This strategy improves water position predictions for the top-ranked pose (26% vs 15%) and over the saved poses (41% vs 29%). Note that the above "top-ranked" refers to the unbiased energy, after removal of the temporary strengthening of ligand-water hydrogen bonds. The improvement is thus mainly due to the sampling protocol spending more time in the relevant problem space, featuring compulsory ligand-water contacts. As in protocols A and B, low complexity systems are obviously more efficiently reproduced with respect to high complexity system. Even so, the final success rate is judged insufficient − yet, it suggests anyway that adapting the sampling strategy to the specific nature of this problem may pay off, hence the development of the strategy D with final water refinement.

3.2.3. Protocol D: Protocol C + Water Refinements around Selected Poses. This approach involves a refinement of all waters around each of the poses kept for the ligand. All new low-energy configurations are stored before the final relaxing step, where the considered DoF are minimized. Compared to previous protocols, there is a great improvement of results in all monitored categories. Thus, the accuracy of top-ranked pose increases from 56% to 69% for ligands and from 26% to 41% concerning waters. The improvement appears more even flagrant over the entire set of saved poses: regarding water location prediction, the accuracy jumps from 41% to 76%. Besides, most of these waters are particularly well predicted: 62% at 1 Å with respect to 76% at 2 Å. Regarding the ability to visit the expected location of all entities, there is an enhancement with respect to previous schemes, especially for waters (from 62% to 94% at 1 Å). These high accuracies reflect the ability to generate poses with both ligand and waters close to their expected locations, even with high complexity systems (e.g., 1G9V, 1HVY, 1N2V, 1SQ5, 2BSM which all contain at least 3 free waters).

Clearly, undertaking an exhaustive search of the optimal locations of water molecules around each of the putative ligand positions does pay off. This is consistent with the fact that the energy landscape is rather flat with respect to the DoF of

waters. Furthermore, the interaction fingerprint changes little with respect to water positions, since these are responsible for few (typically one or two) contacts − poses with identical ligand placements, differing only with respect to water positioning are at risk of being discarded as "redundant" during the evolutionary process. Default simulations would spend most of the effort in trying to explore alternative poses for the main ligand but do not spend enough time in trying to rearrange waters around each pose. If the latter aspect is artificially enhanced, like in the current protocol, significant improvement can be obtained: if the correct ligand pose is among the considered (true in 94% of the cases), the specific optimization of the explicit solvent surrounding it would eventually enumerate the wanted configuration as well. It is also clear − and not really surprising − that the desired water-mediated interaction configurations are not systematically the lowest energy configuration. This is partly an intrinsic weakness of the approach: if water-binding pockets of higher affinity that the ligand-water-site key points exist, they should be filled with explicit water before trying to position bridging water molecules, which are otherwise at risk of being captured by such energetically more rewarding sites. As for example, a water molecule is often located between two negatively charged side chains. The other part of responsibility pertains to the intrinsic FF inaccuracies. Unfortunately, it is very difficult to precisely delimit the role of each of these failure scenarios. In either case, these inaccuracies are however not large enough in order to completely disqualify the wanted configurations from the final list of kept poses.

The challenging complex 1GM8 is a major source of errors: the binding mode of the penicillin-G derivative involves several water-mediated interactions and the carboxylate group, totally solvent exposed, does not make any polar interaction (even mediated) with the binding site. The FF greatly favors (by a computed −24 kcal/mol) a wrong binding mode involving this anionic group in both direct and water-mediated hydrogen bonds with the site. This example accounts for the 6% of cases in which a correct ligand pose never made it into the subset of saved geometries ("over saved poses" criterion decreasing from 100% to 94%). In previous protocols, the correct ligand pose was nevertheless saved − here, the aggressive optimization of water positions provides an additional stabilization for the wrong ligand pose, which eventually leads to discarding of the correct.

Like 1GM8, 1G9V is a recurrent failure in both classical (implicit solvent-only) and previous docking protocols. However, this protocol led to several poses, including the top-ranked one, very close to the expected configuration ($RMSD_{ligand}$ < 0.5 Å and 2 waters out of 3 make the crystallographic ligand-water-site interactions - see Figure 3.c1 and c2). Inclusion of waters allowed generating a good top-ranked solution for the 1G9V complex, when water-free simulations, whatever the considered energy scheme, failed to converge toward the experimental binding mode.

To our knowledge, S4MPLE is the only tool not dealing with crystallographic waters in terms of on/off toggles but as actual physical entities. Empirical simulations with variable numbers of atoms are notoriously difficult to interpret, as the (free) energy functions piloting these were designed to account for conformational changes, not for variations of the included interaction lists. The ″entropic″ penalty should be rather understood as an empirical compensation for the varying number of terms in the scoring function. This is a weak point of all attempts to deal with solvent by inserting or deleting waters, and does not concern S4MPLE, where explicit waters are displaced to alternative locations, not deleted. The weakness of our water treatment strategy is that the initial number of explicit waters was chosen as an empirical parameter - in perspective, it might be interesting to determine this number by ″saturating″ all the hydrogen bonding options of the ligand, like in the MVD strategy.[21]

Albeit the so-far reported results are encouraging, the coexistence of explicit water molecules and implicit solvent may eventually require some dedicated fine-tuning of force field parameters. For example, the hydrophobic term, implicitly accounted for by the number of hydrophobic contacts, stands for the entropy gain associated with water molecules leaving the geometrically constrained orientations at the interface and rejoining the bulk solvent. It is unclear how the behavior of explicit waters (the "entropy" of which could, at least in principle, be illustrated by the ensemble of visited stable poses) would interplay with the implicit solvent terms.

3.3. Docking of One Ligand and Displacing a Water Molecule. This simulation could be classified as a special case of the previous runs involving one ligand and explicit water molecules, and the results are summarized in Table 6.

Table 6. RMSD from Cross-Docking Experiments on the PARP Target (2PAX/3PAX Complexes) from the Displacing Water Study

		binding site 2PAX	binding site 3PAX
entities 2PAX	fragment 2PAX	0.19	0.29
	water (no bridged HB)	10.67	10.43
entities 3PAX	fragment 3PAX	0.31	0.30 (#1)
			0.42 (#2)
	water (bridged HB)	0.80	2.78 (#1)
			0.82 (#2)

Concerning the 3PAX complex, both first and second poses from redocking contain the correct binding mode for the ligand (RMSD of 0.30 Å and 0.42 Å respectively). While preserving the expected water-mediated hydrogen bond in both geometries, only the second one correctly locates the water from a RMSD point of view (0.82 Å vs 2.78 Å). Indeed, the water is slightly translated in the top-ranked pose in order to make an additional (non-native) interaction with a tyrosine (Y907 - flexible hydroxyl group considered) in the vicinity of the ligand. The second pose, from an energetically point of view, perfectly reproduces the experimental data. Redocking of 2PAX leads to the expected solution with a very low RMSD (0.19 Å) for the ligand, which is larger than 3PAX and features no water-mediated hydrogen bond with the E988 side chain: the water molecule, as expected, moves to some arbitrary hydrophilic subpocket of the site, not far (∼1.5 Å) from the location of an original crystallographic water (HOH 1018 in 3PAX). The amine group of the 2PAX ligand takes the place of water in 3PAX. The cross-docking runs converged toward the experimental binding modes for both ligands (RMSD of 0.29 Å and 0.31 Å) too, and the location of water is correctly predicted when needed (mediated interaction with the 3PAX fragment, but no bridged interaction with the larger 2PAX fragment). Figure 4 displays the obtained binding modes in the cross-docking runs. This last example, in addition to previous simulations on the ligand-water set, highlights the ability of

dx.doi.org/10.1021/ci300495r | J. Chem. Inf. Model. 2013, 53, 88−102

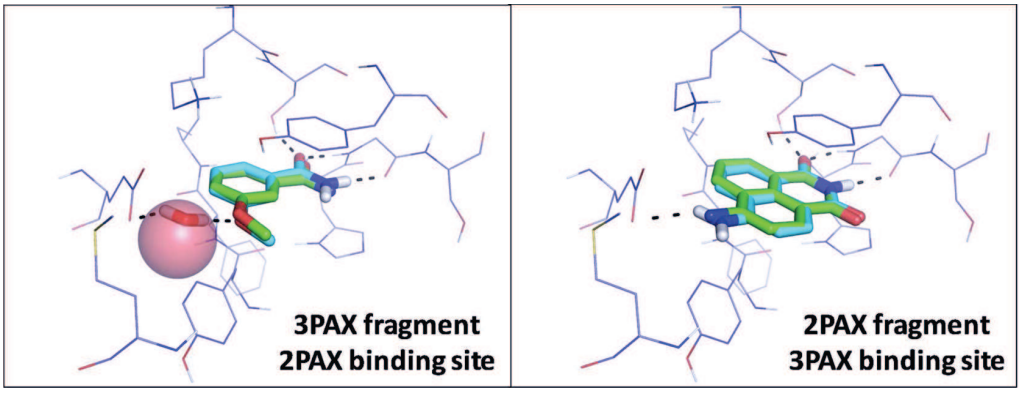


Figure 4. Structures from cross-docking run for the PARP target. Experimental ligand conformer is in green, while expected water locations are red spheres. Top-ranked poses are represented in blue for ligand and waters as stick. Main interactions are depicted as dash lines.

S4MPLE to deal with explicit water particles. Besides, the locations of entities, when correctly found, are generally reported with high accuracy. This holds irrespectively of the employed water docking protocol C or D, showing that statistically favoring ligand-water contacts did not spuriously lead to a failure to evince the water molecule.

3.4. Docking of Two Organic Ligands. Prediction of simultaneous binding modes of fragment-like compounds is an excellent benchmark for testing the multiligand sampling capacity of S4MPLE. This ability is of particular interest in the FBDD field.[24] Indeed, the probability of simultaneous binding in the same active site is higher than for larger molecules, and predicted binding modes are of direct interest, as this constitutes the starting point for drug design by fragment linking. Fragment screening based on native mass spectrometry[53] is able to detect a multibinding event. Although useful, this information is insufficient for a rational optimization since the binding modes remains unknown. In theory, X-ray crystallography can easily solve this problem, but it can be long and expensive in practice. Therefore, docking algorithms, able to deal with several entities, can be used to predict the geometry of these ternary complexes "fragment1+fragment2+protein". The docking results, if judged credible (e.g., interactions with known hot spots), enable the subsequent use of SBDD for a more rational optimization or evolution by means of a linking strategy. Moreover, it has been experimentally demonstrated that fragments can adopt different binding modes when bound alone or simultaneously with another compound (see PDB codes 3HYY, 3HYZ, and 3HZ1 and the associated article[31]). Thus, it appears more consistent and reliable to try to get access to the geometry of the ternary complex itself. Table 7 sums up all the results relative to the multifragments docking performed on selected Hsp90 complexes.

3.4.1. Simultaneous Protocol. A successful docking is achieved for the complex 2QFO, in terms of top-ranked pose and accuracy (RMSD ≈ 0.5 Å for both compounds). It should be noted that large conformational differences of the site are observed between 2QFO and other structures. In the latter, an additional subpocket, systematically occupied by one fragment, emerges as a result of a restructuration as α-helix of the sequence K100 to E120. The consequence is that the 2QFO binding site is half as large as in other investigated structures (note that, given the definitions of active sites as sets of residues within a 10 Å sphere around any atom of the ligand, current simulations include about 70 residues in 2QFO). As in the X-ray structure, 2QFO_1 interacts with the hot spot D93 (directly and through close waters), while 2QFO_2 makes one hydrogen-bond with N51 and an interfragment Π-stacking interaction.

A pose close to the experimental solution is also retrieved for the 4 other cases but unfortunately not as the lowest FF-based energy configuration of the system. In practice, several best poses (e.g., top 30) are checked with thoroughness in prospective docking. Accordingly, correct predictions were generated for 4 out of the 5 investigated ternary complexes (2QFO, 2XDU, 2YEJ, and 3HZ1) within the 12 best poses. The expected locations of the two fragments are not simultaneously retrieved as the top-ranked pose. Among them, there are 3 complexes with one good fragment prediction in the top-ranked pose (2XDU_2, 2YEJ_1, 3HZ1_1). Only the 2YEI docking appears as a complete failure, since the best pose is totally wrong (two large RMSDs for the considered fragments). Besides, the best-ranked couple of acceptable poses for 2YEI_1 and 2YEI_2 lags behind (#187).

Table 7. Results from Multifragments Simulations for Both Investigated Protocols (Simultaneous and Sequential)[a]

	simultaneous docking					
	top-ranked pose			all saved poses		
PDB	rank	RMSD fragment1	RMSD fragment2	rank	RMSD fragment1	RMSD fragment2
2QFO	1	0.40	0.61	1	0.40	0.61
2XDU	1	13.47	0.62	12	1.17	0.69
2YEI	1	3.05	4.36	187	1.47	1.21
2YEJ	1	0.47	4.12	5	0.34	1.26
3HZ1	1	0.37	4.37	2	0.29	0.68
	sequential docking					
	top-ranked pose			all saved poses		
PDB	rank	RMSD fragment1	RMSD fragment2	rank	RMSD fragment1	RMSD fragment2
2QFO	1	0.26	0.46	1	0.26	0.46
2XDU	1	0.34	4.23	5	0.37	0.71
2YEI	1	0.98	4.31	3	0.98	1.25
2YEJ	1	0.59	4.24	10	0.53	1.34
3HZ1	1	0.66	4.44	4	0.54	0.51

[a]RMSD of individual fragments and corresponding ranks for the best pose and the closest native-like configuration (rank >1) are provided.

dx.doi.org/10.1021/ci300495r | J. Chem. Inf. Model. 2013, 53, 88−102

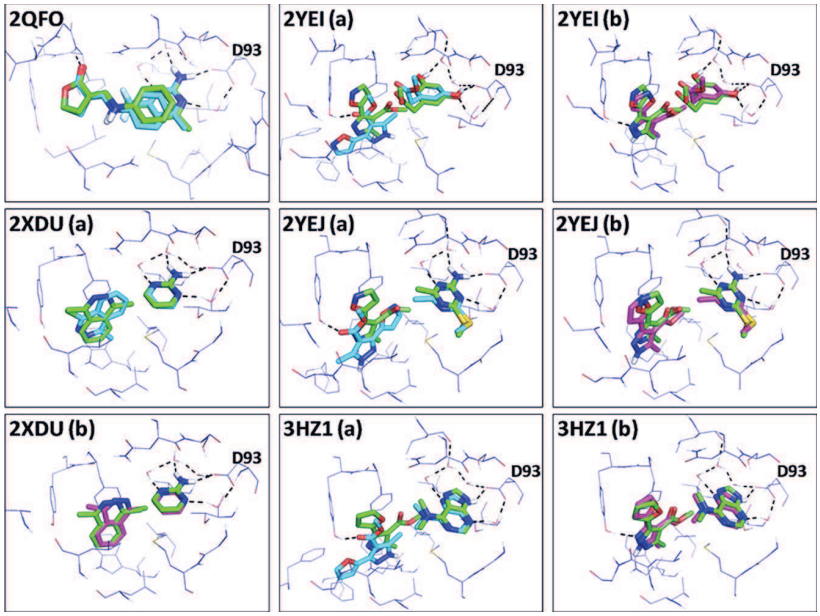


Figure 5. Multifragment simulations (sequential protocol). Experimental locations are displayed in green for fragments. Predicted poses are represented in blue (top-ranked) or purple (other rank) for ligand. Main polar interactions are depicted as dash lines. The hot spot D93 is indicated.

Since the native conformations are quite accurately sampled with this "brute-force" approach, but not ranked as #1, the classical FF failure hypothesis is the preferred explanation. It should be noted that the scaffolds of misdocked fragments are highly similar: a methyl-pyrazole linked to a heterocycle (a furan for 2YEI_2 and 2YEJ_2 or an isoxazole for 3HZ1_2) with a masked acidic group. The latter is enclosed by several hydrophobic side chains such as F138, L107, M98, and V150; therefore, it does not make any polar interaction with the site (this "quite useless" chemical function was later removed in the 3HZ1 fragment optimization using a linking strategy[31]). Finally, both 2YEI fragments contain an ester function which poorly interacts with the target, which seems to give more weight to the hypothesis of a less than optimal FF parametrization of esters.

3.4.2. Sequential Protocol. This alternative strategy, less greedy in terms of computing resources (e.g., smaller number of generations), has been performed in parallel to the simultaneous protocol. Although the fragments are docked separately in a first step, the final relaxation is performed over reasonable putative configurations of the system, including both investigated compounds within the binding site. Therefore, putative interfragment contacts are taken into account during this final stage. Figure 5 shows, for all investigated structures, the closest pose to X-ray data in addition to the top-ranked pose, each superimposed to the experimental one. As before, the 2QFO geometry is perfectly reproduced, even at the first rank, while pyrazole derivatives adopt wrong binding modes. By contrast to the simultaneous strategy, the top-ranked pose of the 2YEI complex contains one successfully predicted fragment (2YEI_1). This time, the correctly predicted 2XDU fragment from the lowest-energy pose is not the same: using the sequential scheme, the potential energy is slightly better, and the location of the amino-pyrimidine fragment (2XDU_1) is perfectly reproduced in the vicinity of D93 (RMSD < 0.5 Å) to the detriment of the less well anchored phthalazine fragment (2XDU_2).

Generally speaking, the same trend emerges from the ability to reproduce the structure of these ternary complexes using this sequential strategy, with a slight improvement toward lower RMSD and with better acceptable ranks (when the top-ranked pose does not match the experimental data). This is coherent with the nature of this docking problem, where fragments are not strong competitors (they would not target a same subpocket). Interfragment interactions do exist, but they are less decisive with respect to much stronger site-fragment anchoring contributions (unlike in explicit water placement, which is highly dependent on the ligand position). Actually, the 2QFO system which displays the strongest interfragment contact (a Π-stacking of respective phenyl rings) also happens to be the best predicted.

The analysis of the whole set unsurprisingly highlights that fragments with most polar groups involved in hydrogen bonds (e.g., 2QFO_1, 2QFO_2, 2XDU_1, 2YEI_1, 2YEJ_1, 3HZ1_1) are more often successfully docked, by contrast to others (2XDU_2, 2YEI_2, 2YEJ_2, 3HZ1_2). Besides, fragments from the 2QFO structure exhibit pretty high affinity with respect to their small size ($K_{d\_2QFO\_1}$ = 20 μM, $K_{d\_2QFO\_2}$ = 150 μM, see ref 35), compared to some of the failed ones (e.g., $K_{d\_2XDU\_2}$ > 1 mM, $IC_{50\_3HZ1\_2}$ = 1 mM, see refs 32 and 31 respectively). This goes in the same direction as one main conclusion from the study of Verdonk et al. about fragment docking, in which they demonstrated that an important factor affecting docking accuracy of fragment-like compounds is ligand efficiency (LE).

Finally, it should be noted that top-ranked poses from docking simulations are systematically lower or equal in energy than geometries obtained by simple energy minimization of X-ray poses, accrediting the efficiency of the sampling engine. Although not perfect, these results, in addition to previous ones relative to the fragment set, demonstrate that S4MPLE is able to reproduce expected binding modes for fragment-like compounds, and even in the tricky multifragments case.

dx.doi.org/10.1021/ci300495r | J. Chem. Inf. Model. 2013, 53, 88−102

Therefore, a usage of S4MPLE within FBDD projects becomes possible and promising.

4. CONCLUSION

S4MPLE, a conformational sampling tool based on a hybrid evolutionary algorithm, has been implemented to allow extensive atom flexibility, all the while being general and transcending the classical "one site, one ligand" docking scheme. It includes an original population diversity check relying on pairwise interaction fingerprints and generic operators acting in 3D Cartesian space on user-defined and general DoF. As a consequence, S4MPLE provides an equivalent treatment of intra- and intermolecular DoF, making no distinction between conformational sampling of a single compound or docking. Thus, the theoretical applicability range of S4MPLE is only limited by the set of available FF parameters and accessible computing time. This program is meant to address more difficult sampling and docking problems by contrast to high throughput docking at few seconds/ligand. While designed for deployment on grids, the current work relies on the few-CPU "workstation" version of S4MPLE. This second paper addressed in more details the following main issues:

- Docking of fragment-like compounds, in order to assess the applicability of the AMBER/GAFF force field to this class of organic ligands,
- Simultaneous docking of several entities, including two noncompetitive fragments, or one ligand and explicit free waters.

These results demonstrate the ability of S4MPLE to deal with several entities: experimental configurations are often generated but not systematically top-ranked. AMBER/GAFF displayed, for docking of single fragment-like ligands, success rates comparable to state-of-the-art drug-like molecule docking benchmarks. The previously reported[3] fitting of additional FF terms (continuum solvent model, contact bonus terms) proved beneficial for fragment-like docking as well. So far, we do not see any systematic FF bias due to lower ligand sizes although specific FF deficiencies remain – for example, a tendency to force ester groups into fake contacts with the protein site.

Beyond such "classical" FF problems, docking of explicit waters within a continuum solvent, in attempts to predict ligand-water-site interactions, is a very difficult problem because the targeted configurations are not necessarily the energetically most rewarding ones (the mobile explicit water may well be trapped into site subpockets binding crystallographic waters even tighter). Therefore, if the experimental ligand-water-site-featuring configurations were not ranked as first among sampled states, this is not (necessarily) a FF failure. The ability to enumerate such states within the typical top 30-saved geometries should count as success. Success rates are, expectedly, higher in simulations with fewer explicit waters. In general, adding explicit waters does not trigger a huge increase of the number of DoF but alters the energy landscape of the problem. The additional DoF create a quite flat energy profile (explicit water Coulomb interactions being counter-balanced by the continuum desolvation term), or the difficulty of a sampling problem is known to increase significantly in such cases (flat landscapes, with many shallow local minima). Alternative strategies enhancing the sampling of the inter-molecular DoF of waters were introduced in order to deal with this problem.

Eventually, multiple docking of fragment-like ligands successfully managed to return experimental structures of ternary complexes, albeit – again – not always at the top of saved conformer lists. This is due to the above-cited FF problems.

The strength of S4MPLE is its flexible control of DoF and its adaptability to the computational resources one possesses: on a workstation, extensive fixing of estimated low-mobility moieties is mandatory. The key plus of S4MPLE is that, with extensive computer power, it is possible to easily unlock more DoF and go for in-depth sampling using massively parallel deployment. At the moment, S4MPLE is still a prototype. However, having successfully passed these tests about fragments docking, interesting perspectives should be reached in the FBDD, once a strategy relative to the *in silico* optimization (growing and linking) of fragments will be developed. Future work will focus in that direction.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Detailed docking results with free water molecules, for each employed protocol (spreadsheets in ci300495r_si_001.xlsx) and detailed fragment redocking results, per individual simulation (ci300495r_si_002.xlsx). This material is available free of charge via the Internet at http://pubs.acs.org. The following are available for download on http://infochim.u-strasbg.fr, DOWNLOADS section (tar.gz): x86_64 executable of S4MPLE; User guide; various ligand preparation tools and force field parameter distributions; and a spreadsheet with the list of the training set compounds used to calibrate Fit FF and detailed docking results for each of the repeated docking runs. Docked poses are available upon request (PyMol .pse format).

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: dhorvath@unistra.fr.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors wish to thank the staff of the two computer centers which hosted the simulations: HPC (high-performance computing) of the University of Strasbourg and HPC of the chemistry faculty of Cluj-Napoca. All pictures depicting ligand-protein structures have been created using Pymol.[54]

■ REFERENCES

(1) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules. *J. Am. Chem. Soc.* **1995**, *117* (19), 5179–5197.

(2) Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25* (9), 1157–1174.

(3) Hoffer, L.; Chira, C.; Marcou, G.; Varnek, V.; Horvath, D. S4MPLE - sampler for multiple protein-ligand entities: methodology & rigid-site docking benchmarking. *J. Mol. Graphics Modell.* **2012**, submitted.

(4) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **2007**, *50* (4), 726–41.

(5) Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. A review of protein-small molecule docking methods. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 151−166.
(6) Janin, J.; Henrick, K.; Moult, J.; Eyck, L. T.; Sternberg, M. J.; Vajda, S.; Vakser, I.; Wodak, S. J. CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins* **2003**, *52* (1), 2−9.
(7) Janin, J. Protein-protein docking tested in blind predictions: the CAPRI experiment. *Mol. BioSyst.* **2010**, *6* (12), 2351−62.
(8) Marcou, G.; Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.* **2007**, *47* (1), 195−207.
(9) Brooijmans, N.; Kuntz, I. D. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335−373.
(10) Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S. A semiempirical free energy force field with charge-based desolvation. *J. Comput. Chem.* **2007**, *28* (6), 1145−1152.
(11) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.* **2006**, *49* (21), 6177−96.
(12) Lu, Y.; Wang, R.; Yang, C.; Wang, S. Analysis of ligand-bound water molecules in high-resolution crystal structures of protein-ligand complexes. *J. Chem. Inf. Model.* **2007**, *47* (2), 668−675.
(13) Poornima, C.; Dean, P. Hydration in drug design. 1. Multiple hydrogen-bonding features of water molecules in mediating protein-ligand interactions. *J. Comput.-Aided Mol. Des.* **1995**, *9* (6), 500−512.
(14) Thilagavathi, R.; Mancera, R. L. Ligand-protein cross-docking with water molecules. *J. Chem. Inf. Model.* **2010**, *50* (3), 415−21.
(15) Huang, N.; Shoichet, B. K. Exploiting ordered waters in molecular docking. *J. Med. Chem.* **2008**, *51* (16), 4862−5.
(16) Barillari, C.; Taylor, J.; Viner, R.; Essex, J. W. Classification of water molecules in protein binding sites. *J. Am. Chem. Soc.* **2007**, *129* (9), 2577−87.
(17) Raymer, M. L.; Sanschagrin, P. C.; Punch, W. F.; Venkataraman, S.; Goodman, E. D.; Kuhn, L. A. Predicting conserved water-mediated and polar ligand interactions in proteins using a K-nearest-neighbors genetic algorithm. *J. Mol. Biol.* **1997**, *265* (4), 445−64.
(18) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857.
(19) Verdonk, M. L.; Chessari, G.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Nissink, J. W.; Taylor, R. D.; Taylor, R. Modeling water molecules in protein-ligand docking using GOLD. *J. Med. Chem.* **2005**, *48* (20), 6504−15.
(20) Rarey, M.; Kramer, B.; Lengauer, T. The particle concept: placing discrete water molecules during protein-ligand docking predictions. *Proteins* **1999**, *34* (1), 17−28.
(21) Lie, M. A.; Thomsen, R.; Pedersen, C. N. S.; Schiøtt, B.; Christensen, M. H. Molecular docking with ligand attached water molecules. *J. Chem Inf. Model.* **2011**, *51* (4), 909−917.
(22) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; Debolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. AMBER a package of computer-programs for applying molecular mechanics, normal-mode analysis, molecular-dynamics and free-energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* **1995**, *91* (1−3), 1−41.
(23) Marrone, T. J.; Gilson, M. K.; Mccammon, J. A. Comparison of continuum and explicit models of solvation: potentials of mean force for alanine dipeptide. *J. Phys. Chem.* **1996**, *100* (5), 1439−1441.
(24) Hoffer, L.; Renaud, J. P.; Horvath, D. Fragment-based drug design: computational and experimental state of the art. *Comb. Chem. High Throughput Screening* **2011**, *14* (6), 500−520.
(25) Horvath, D. A virtual screening approach applied to the search for trypanothione reductase inhibitors. *J. Med. Chem.* **1997**, *40* (15), 2412−23.
(26) Brewerton, S. C. The use of protein-ligand interaction fingerprints in docking. *Curr. Opin. Drug Discovery Dev.* **2008**, *11* (3), 356−64.
(27) Deng, Z.; Chuaqui, C.; Singh, J. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem.* **2004**, *47* (2), 337−44.
(28) Nissink, J. W.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A new test set for validating predictions of protein-ligand interaction. *Proteins* **2002**, *49* (4), 457−71.
(29) Congreve, M.; Chessari, G.; Tisi, D.; Woodhead, A. J. Recent developments in fragment-based drug discovery. *J. Med. Chem.* **2008**, *51* (13), 3661−3680.
(30) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A rule of three for fragment-based lead discovery? *Drug Discovery Today* **2003**, *8* (19), 876−877.
(31) Barker, J. J.; Barker, O.; Courtney, S. M.; Gardiner, M.; Hesterkamp, T.; Ichihara, O.; Mather, O.; Montalbetti, C. A.; Müller, A.; Varasi, M.; Whittaker, M.; Yarnold, C. J. Discovery of a novel Hsp90 inhibitor by fragment linking. *ChemMedChem* **2010**, *5* (10), 1697−700.
(32) Murray, C. W.; Carr, M. G.; Callaghan, O.; Chessari, G.; Congreve, M.; Cowan, S.; Coyle, J. E.; Downham, R.; Figueroa, E.; Frederickson, M.; Graham, B.; McMenamin, R.; O'Brien, M. A.; Patel, S.; Phillips, T. R.; Williams, G.; Woodhead, A. J.; Woolford, A. J. Fragment-based drug discovery applied to Hsp90. Discovery of two lead series with high ligand efficiency. *J. Med. Chem.* **2010**, *53* (16), 5942−55.
(33) Woodhead, A. J.; Angove, H.; Carr, M. G.; Chessari, G.; Congreve, M.; Coyle, J. E.; Cosme, J.; Graham, B.; Day, P. J.; Downham, R.; Fazal, L.; Feltell, R.; Figueroa, E.; Frederickson, M.; Lewis, J.; McMenamin, R.; Murray, C. W.; O'Brien, M. A.; Parra, L.; Patel, S.; Phillips, T.; Rees, D. C.; Rich, S.; Smith, D. M.; Trewartha, G.; Vinkovic, M.; Williams, B.; Woolford, A. J. Discovery of (2,4-dihydroxy-5-isopropylphenyl)-[5-(4-methylpiperazin-1-ylmethyl)-1,3-dihydroisoindol-2-yl]methanone (AT13387), a novel inhibitor of the molecular chaperone Hsp90 by fragment based drug design. *J. Med. Chem.* **2010**, *53* (16), 5956−69.
(34) Barker, J. J.; Barker, O.; Boggio, R.; Chauhan, V.; Cheng, R. K. Y.; Corden, V.; Courtney, S. M.; Edwards, N.; Falque, V. M.; Fusar, F.; Gardiner, M.; Hamelin, E. M. N.; Hesterkamp, T.; Ichihara, O.; Jones, R. S.; Mather, O.; Mercurio, C.; Minucci, S.; Montalbetti, C.; Muller, A.; Patel, D.; Phillips, B. G.; Varasi, M.; Whittaker, M.; Winkler, D.; Yarnold, C. J. Fragment-based identification of Hsp90 inhibitors. *ChemMedChem* **2009**, *4* (6), 963−966.
(35) Huth, J. R.; Park, C.; Petros, A. M.; Kunzer, A. R.; Wendt, M. D.; Wang, X. L.; Lynch, C. L.; Mack, J. C.; Swift, K. M.; Judge, R. A.; Chen, J.; Richardson, P. L.; Jin, S.; Tahir, S. K.; Matayoshi, E. D.; Dorwin, S. A.; Ladror, U. S.; Severin, J. M.; Walter, K. A.; Bartley, D. M.; Fesik, S. W.; Elmore, S. W.; Hajduk, P. J. Discovery and design of novel HSP90 inhibitors using multiple fragment-based design strategies. *Chem. Biol. Drug Des.* **2007**, *70*, 1−12.
(36) Roughley, S.; Hubbard, R. How well can fragments explore accessed chemical space? A case study from heat shock protein 90. *J. Med. Chem.* **2011**, *54* (12), 3989−4005.
(37) http://www.chemaxon.com.
(38) Wang, J. M.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graphics Modell.* **2006**, *25* (2), 247−260.
(39) *MOE (Molecular Operating Environment)*, 2005.06; Chemical Computing Group, Inc.: Montreal, 2005.
(40) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261* (3), 470−489.
(41) Schubert, C. R.; Stultz, C. M. The multi-copy simultaneous search methodology: a fundamental tool for structure-based drug design. *J. Comput.-Aided Mol. Des.* **2009**, *23* (8), 475−489.
(42) Ruf, A.; de Murcia, G.; Schulz, G. E. Inhibitor and NAD+ binding to poly(ADP-ribose) polymerase as derived from crystal structures and homology modeling. *Biochemistry* **1998**, *37* (11), 3893−900.

(43) Verdonk, M. L.; Giangreco, I.; Hall, R. J.; Korb, O.; Mortenson, P. N.; Murray, C. W. Docking performance of fragments and druglike compounds. *J. Med. Chem.* **2011**, *54* (15), 5422−31.
(44) Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today* **2004**, *9* (10), 430−431.
(45) Loving, K.; Salam, N. K.; Sherman, W. Energetic analysis of fragment docking and application to structure-based pharmacophore hypothesis generation. *J. Comput.-Aided Mol. Des.* **2009**, *23* (8), 541−554.
(46) Haider, M. K.; Bertrand, H. O.; Hubbard, R. E. Predicting fragment binding poses using a combined MCSS MM-GBSA approach. *J. Chem. Inf. Model.* **2011**, *51* (5), 1092−1105.
(47) Sándor, M.; Kiss, R.; Keseru, G. M. Virtual fragment docking by Glide: a validation study on 190 protein-fragment complexes. *J. Chem. Inf. Model.* **2010**, *50* (6), 1165−72.
(48) Jones, G.; Willett, P.; Glen, R.; Leach, A.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267* (3), 727−48.
(49) Stultz, C. M.; Karplus, M. MCSS functionality maps for a flexible protein. *Proteins* **1999**, *37* (4), 512−29.
(50) Caflisch, A.; Miranker, A.; Karplus, M. Multiple copy simultaneous search and construction of ligands in binding sites: application to inhibitors of HIV-1 aspartic proteinase. *J. Med. Chem.* **1993**, *36* (15), 2142−67.
(51) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47* (7), 1739−1749.
(52) Dill, K. A.; Fiebig, K. M.; Chan, H. S. Cooperativity in protein-folding kinetics. *Proc. Natl. Acad. Sci.* **1993**, *90* (5), 1942−1946.
(53) Hannah, V.; Atmanene, C.; Zeyer, D.; Van Dorsselaer, A.; Sanglier-Cianferani, S. Native MS: an 'ESI' way to support structure- and fragment-based drug discovery. *Future Med. Chem.* **2010**, 35−50.
(54) DeLano, W. L. *The PyMOL Molecular Graphics System*; DeLano Scientific: San Carlos, CA, USA, 2002.

4.4 Criblage virtuel d'une chimiothèque de fragments

Le criblage expérimental d'une chimiothèque propriétaire (321 composés), uniquement constituée de fragments, a été réalisé en amont de cette thèse au sein de l'entreprise NovAliX. La cible est la protéine chaperonne HSP90. Plus d'informations à propos de la cible et du criblage expérimental sont données dans le chapitre 6 étant donné que l'étude prospective est le prolongement direct de ces résultats expérimentaux.

4.4.1 Matériel et méthodes

Des données d'intérêt étant disponibles (liste des fragments actifs, nombre total d'actifs), le criblage virtuel de cette chimiothèque est également réalisé afin de quantifier le pouvoir prédictif de S4MPLE et des différentes fonctions d'énergie. A cette chimiothèque sont également ajoutés les composés de type fragment (masse < 300 Da) employés dans une étude de modélisation en partie focalisée sur ladite cible [152] : 1QYE, 1ZWH, 2CCS, 2QFO (deux fragments distincts), 2WI1, 2WI2, 3EKO et 3FT5. Au final, il y a environ 12% d'actifs (40/330) dans la chimiothèque.

Le site de liaison est préparé à partir de la structure X-ray 3D0B avec inclusion des molécules d'eau conservées sur un grand nombre de complexes de HSP90. Il s'agit de HOH235, HOH236 et HOH353 pour 3D0B. Le criblage virtuel est réalisé avec S4MPLE selon les deux fonctions d'énergie usuelles (Core FF et Fit FF) et avec le logiciel de docking FlexX.

Avec S4MPLE, les hydrogènes polaires sont débloqués au niveau site de liaison et les paramètres par défaut de l'AG sont utilisés, notamment 500 générations et une population constituée de 50 individus. Après la phase d'EC (docking), les différentes poses non redondantes sont soumises à une phase d'optimisation comme décrit préalablement (voir le §3.4). Pour chaque fragment, il y a un docking de celui-ci dans le site, suivi d'un rapide EC du fragment seul. Ces deux simulations aboutissent chacune à un minimum énergétique donné, et l'énergie finale considérée est la soustraction de ces deux valeurs comme indiqué précédemment (voir le §3.6). Les différents fragments sont ensuite triés par énergie croissante et annotés en fonction de leur statut connu (actif / inactif).

4.4.2 Résultats et discussion

Les différents criblages virtuels sont ensuite analysés selon les critères introduits préalablement au §1.6.3, à savoir le facteur d'enrichissement et l'aire sous la courbe ROC. Les résultats bruts sont analysés à l'aide de l'outil KNIME [198] pour calculer les facteurs d'enrichissement à différents seuils

précoces (voir le Tableau 16), tracer les courbes ROC et déterminer les AUC correspondantes (voir la Figure 47).

Protocole de criblage	Seuil 1%	Seuil 3%	Seuil 10%
S4MPLE (Core FF)	7,5 (EF=7,5)	7,5 (EF=2,5)	14,0 (EF=1,4)
S4MPLE (Fit FF)	7,5 (EF=7,5)	12,5 (EF=4,2)	22,5 (EF=2,25)
FlexX	0,0 (EF=0)	5,0 (EF=1,67)	12,5 (EF=1,25)

Tableau 16: Pourcentages d'actifs retrouvés et facteurs d'enrichissement pour divers seuils et protocoles de criblage virtuel.

Comme lors de l'étude de la corrélation vis-à-vis des constantes d'affinité (voir le §3.6), les résultats issus de FlexX sont les moins intéressants, suivi par ceux de Core FF pour finir par ceux de Fit FF. Pour ces deux derniers protocoles, les facteurs d'enrichissement précoces sont tout à fait corrects sur ce cas : des actifs sont retrouvés immédiatement en tête de liste. Par exemple, dans les 33 meilleurs fragments (10% de la base) selon le protocole impliquant la fonction d'énergie Fit FF, il y a 9 actifs connus (22,5% des actifs) alors qu'un choix aléatoire n'aurait permis d'en récolter statiquement que 4 environ, d'où le facteur d'enrichissement de 2,25.

La même tendance ressort également des courbes ROC, bien qu'objectivement les AUC obtenues restent au mieux assez moyennes. Avec FlexX, le démarrage est plus lent et l'AUC est égale à 0,56 soit légèrement mieux qu'une distribution aléatoire (AUC=0,5). Bien que trouvant des actifs plus rapidement, la courbe ROC s'essouffle par la suite dans le cas de Core FF pour finir avec une AUC égale à 0,5. A l'opposé, la fonction d'énergie Fit FF permet d'obtenir à la fois un meilleur démarrage selon le critère EF et une meilleure AUC (AUC=0,69).

Enfin, le fragment principal, à savoir celui qui a été considéré par les chimistes et les biologistes sur la base des résultats du criblage par spectrométrie de masse, est également retrouvé dans le top 8% de la chimiothèque (position 24 sur 330). C'est la structure cristallographique du complexe impliquant ce composé qui servira de base de travail pour la phase d'optimisation décrite dans le chapitre 6. Avant de passer à ce cas d'étude prospectif, il est toutefois nécessaire de développer et valider la stratégie d'optimisation virtuelle de fragments sur des cas rétrospectifs issus de la littérature scientifique. Ces différentes étapes font l'objet du chapitre suivant.

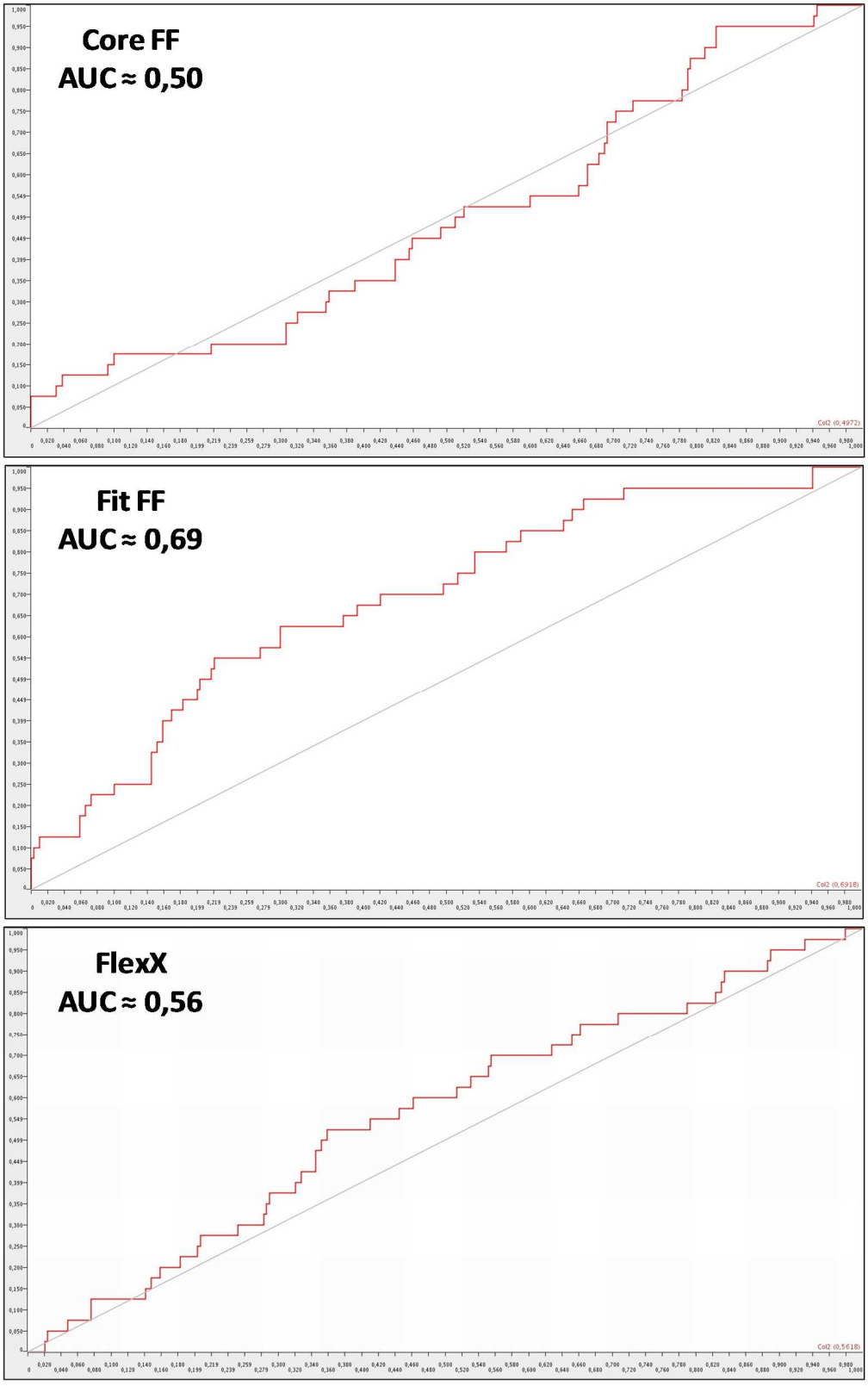


Figure 47: Courbes ROC et AUC pour chaque protocole de criblage virtuel.

Haut : S4MPLE avec Core FF ; Milieu : S4MPLE avec Fit FF ; Bas : score FlexX.

5 Stratégie d'optimisation virtuelle de fragments

Le prérequis concernant la capacité de S4MPLE à simuler des fragments étant validé à ce stade, un protocole virtuel d'optimisation fragment→ligand, *via* les stratégies de growing et linking introduites précédemment au §1.3, est développé. Il a pour but de suggérer des optimisations potentielles de hits d'une manière rationnelle tout en tenant compte de la structure 3D du site de liaison. Une focalisation sur des hits de type fragment est réalisée dans ce contexte, mais il n'est pas conceptuellement limité à cette classe de hits. On rappelle que dans un souci de simplification, le terme linker sera employé quel que soit le type de simulation (growing ou linking).

Outre le fait de devoir répondre à cette problématique spécifique, cette stratégie d'optimisation / d'évolution doit également reposer sur le travail réalisé en amont, notamment à travers l'utilisation du programme S4MPLE. La fonction d'énergie Fit FF doit également être employée puisque c'est elle qui a obtenu les meilleure performances vis-à-vis de ces molécules de très faible complexité.

La fine gestion des DDL et la conception des opérateurs génétiques permettent d'employer S4MPLE, avec une mise à jour minime, comme outil dans le cadre du FBDD *in silico*. Sa capacité intrinsèque à réaliser l'EC d'une sous-structure spécifique, liée de manière covalente à un groupement donné (une chaîne latérale flexible reliée à la chaîne principale fixe, mais également un linker connecté au fragment initial qui resterait fixe), en est un parfait exemple. En d'autres termes, S4MPLE peut être employé pour optimiser la conformation interne d'un linker tout en intégrant les contraintes stériques du site. De plus, certaines de ses caractéristiques (multi-entités [197] et flexibilité du site) peuvent aussi être directement incorporées au protocole virtuel d'optimisation. Ainsi, les principaux scénarios rencontrés au niveau du FBDD devraient pouvoir être théoriquement abordés : optimisation par growing ou linking, avec la possibilité d'incorporer une certaine flexibilité au niveau du site et de gérer la médiation de molécules d'eau *via* des liaisons hydrogène (ligand - - H_2O - - site, voire ligand - - H_2O - - H_2O - - site).

Cette stratégie d'optimisation *in silico* et sa validation rétrospective sur des données essentiellement issues de la littérature du FBDD ont fait l'objet d'un troisième article [199] (inséré au niveau du §5.5). Par conséquent, seules les grandes lignes seront décrites et discutées dans ce chapitre.

5.1 Choix de la stratégie d'optimisation

Cette problématique d'optimisation de hits peut être abordée de diverses manières, notamment *via* le criblage virtuel d'une chimiothèque focalisée sur le(s) fragment(s) actif(s) de départ ou par ajouts successifs de groupements chimiques comme dans une simulation de DND (voir le §1.7). Dans ce

dernier cas, le processus commencerait toutefois à partir d'un composé de référence. Bien que l'approche DND soit potentiellement plus rapide car reposant sur une simulation unique, c'est celle basée sur le criblage virtuel d'une chimiothèque focalisée qui est choisie, et ce pour plusieurs raisons théoriques et/ou purement techniques qui sont discutées ci-dessous.

L'ajout de groupements chimiques nécessite une assignation des types atomiques du FF à la volée et une mise à jour des différents termes énergétiques du FF précalculés au moment de l'initialisation. De plus, la détection des types atomiques est réalisée avec l'outil externe Antechamber [178] fourni par les auteurs de GAFF, ce qui complique la mise à jour en temps réel. Enfin, les approches de modélisation moléculaire reposent généralement sur un système défini initialement (nombre fixe d'entités, chacune ayant une topologie donnée), et le but de la simulation consiste à échantillonner les différents objets les uns par rapport aux autres.

Les charges atomiques sont partielles et non formelles en MM. Par conséquent, l'ajout de groupements supplémentaires entraîne la nécessité de régulièrement mettre à jour les charges partielles tout au long de la simulation. Cette remarque rejoint celle des types atomiques sur l'aspect de modification des paramètres de base du système considéré. De plus, il faut également inclure la gestion de l'état de protonation du ligand qui évolue au fur et à mesure, ce qui nécessite soit l'incorporation d'un modèle de pK_a soit l'encodage de règles empiriques.

Une optimisation en temps réel implique aussi un développement informatique beaucoup plus important sur l'outil d'EC lui-même, alors qu'un criblage virtuel ne consiste qu'à lancer n simulations, n étant le nombre de composés de la chimiothèque à cribler.

Enfin, le fait de passer par une librairie externe de molécules permet d'avoir la main sur ce qui est réellement criblé. Par exemple, des filtres classiques (masse, nombre de liaisons à rotation libre, *etc.*) voire des ensembles de règles [48, 200, 201] peuvent permettre de réduire la zone de recherche à une portion souhaitée de l'espace chimique.

Néanmoins, l'approche par criblage virtuel nécessite la création au préalable d'une librairie focalisée sur le ou les hits de départ, d'où le développement des outils JMolEvolve et GenLinkersDB (introduits précédemment au §2.2) afin de répondre à cette problématique.

5.2 Description du protocole d'évolution

Le protocole d'optimisation de fragment est constitué de trois phases principales :

1) la création d'une chimiothèque focalisée sur un (growing) ou deux (linking) composé(s). Leur mode de liaison peut être expérimental ou déterminé à l'aide d'une simulation préliminaire de docking. A ce stade, on part du principe que la chimiothèque a été générée avec succès comme décrit au §2.2
2) la phase de criblage virtuel qui consiste en un EC en deux étapes pour chaque composé de la banque criblée
    a. S4MPLE est utilisé pour réaliser un EC biaisé des composés au sein du site de liaison en utilisant des contraintes sur les atomes provenant du fragment initial. Une vision concrète de cette phase, selon le type d'évolution, est donnée au niveau de la Figure 48. Cette première étape opérant sur un système ligand-récepteur est suivie d'une optimisation des différentes poses non redondantes avec un ligand totalement libre (l'ensemble des contraintes pesant sur lui sont préalablement supprimées)
    b. S4MPLE est ensuite utilisé pour faire un EC classique (sans aucune contrainte) du ligand seul (hors site). Les différents conformères non redondants sont finalement optimisés, comme il est d'usage à l'aide de procédures de minimisation
3) le classement final des composés selon des critères énergétiques ou géométriques. Les informations relatives à ces critères sont sauvegardées dans les fichiers moléculaires en vue de la phase d'analyse ultérieure réalisée à l'aide de scripts bash
    a. le critère principal est évidemment énergétique puisque l'on se place dans une optique de MM basée sur un FF. Une différence d'énergie potentielle est à nouveau considérée, selon le même principe qu'au §3.6, entre celle minimale issue de la phase 2.a (complexe) et celle minimale provenant de la phase 2.b (ligand seul)
    b. le second indicateur concerne la déviation (estimée *via* le critère RMSD) des atomes lourds provenant du hit initial par rapport à leurs coordonnées finales dans la molécule chimère. Un seuil est défini par l'utilisateur, et les composés dont le RMSD est supérieur seront tout simplement écartés

Ce principe général est illustré à la figure 2 de l'article (voir le §5.5).

De plus, le protocole repose directement sur certaines spécificités techniques de S4MPLE. Par conséquent, les modifications au niveau du code source sont donc mineures, les principales étant :

- l'utilisation d'un statut fixe pour les atomes provenant du hit initial permet de réaliser un EC contraint lorsque le mode growing ou linking est activé. En effet, seuls les composés évolués partageant le mode de liaison du fragment initial sont réellement d'intérêt dans le contexte du FBDD. Techniquement, il ne reste qu'à automatiser le statut fixe pour les atomes provenant du fragment initial. Une étiquette représentant cette information peut être stockée dans le fichier du ligand simulé, comme c'est déjà le cas pour d'autres paramètres fondamentaux du FF (charges partielles et types atomiques)
- dans le seul contexte du linking, les deux extrémités sont initialement fixes (étape 2.a). Par conséquent, l'identification des DDL explicites (tels les dièdres) n'est pas directement réalisable, dans la mesure où aucun atome fixe ne doit être présent des deux côtés de la liaison centrale du dièdre. Toutefois, l'utilisation du statut de liaison dit "broken" permet d'outrepasser cette difficulté et d'identifier automatiquement les DDL. Cette astuce fut initialement implémentée pour réaliser l'EC de structures cycliques (voir le guide d'utilisateur de S4MPLE, disponible au niveau du §2.1.3) ; elle trouve donc ici une voie alternative d'emploi. Il ne reste qu'à automatiser l'activation de ce statut particulier au niveau d'une des deux liaisons qui servent de "jonction fragment-linker" lorsque le mode linking est activé

Conceptuellement, le linker à échantillonner est semblable à la chaîne latérale d'un résidu dont la chaîne principale serait fixe. Par conséquent, S4MPLE ne voit pas de ligand à part entière lors de la phase 2.a d'EC sous contraintes : le fichier de configuration usuel "hot_spots", contenant les points d'ancrage préférentiels pour le docking des ligands, est donc absolument inutile dans le contexte du growing/linking.

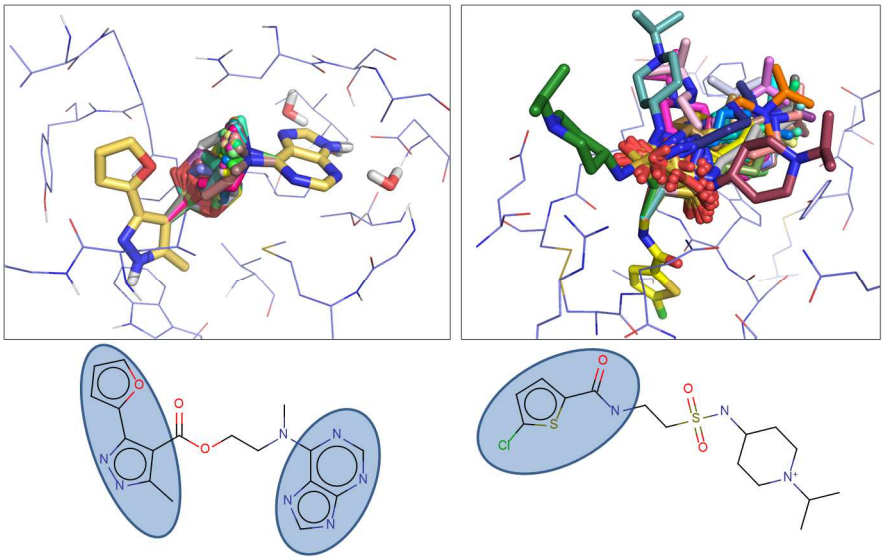


Figure 48: Illustration de la phase 2.a d'EC sous contraintes pour des simulations de type growing (à droite) ou linking (à gauche).

Les atomes entourés sont considérés ici comme fixes et proviennent des composés de départ.

Un axiome du FBDD stipule que le mode de liaison initial du fragment doit être conservé lors de la phase d'évolution. Par exemple dans le contexte du linking, un linker approprié doit en principe conserver le mode de liaison initial de chacun des fragments du complexe ternaire. Par conséquent, une trop grande déviation des atomes lors de la phase finale de relaxation de l'étape 2.a) est :

- soit le signe d'un linker inapproprié (cas où l'énergie d'interaction devrait aussi être assez peu favorable)
- soit le résultat d'une optimisation qui serait en dehors du domaine d'applicabilité du FBDD (voir l'axiome ci-dessus). Un composé chimère ayant une énergie favorable, mais dont les fragments ont été forcés à quitter leur position lorsqu'ils sont liés seuls, peut tout à fait être actif. Néanmoins, si le choix se présente, des analogues d'énergie tout aussi favorable mais maintenant de surcroît les fragments dans leur position initiale sont à préférer. Cette stratégie conservatrice recoupe le principe d'utilisation des empreintes d'interaction [1], ces dernières favorisant les modes d'interaction connus par rapport au seul critère de la fonction de score

Les critères énergie et RMSD sont ainsi complémentaires (voir la Figure 49).

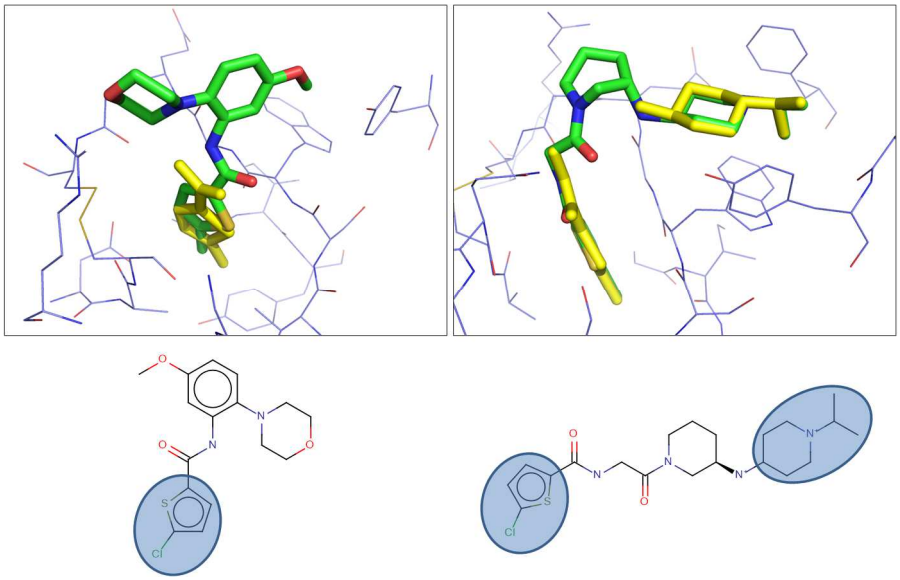


Figure 49: Complémentarité entre les critères énergie et RMSD.

A gauche : simulation de type growing où le critère RMSD permet d'écarter le composé lorsque le seuil défini est strict. A droite : simulation de type linking où le composé est relativement mal classé à cause du critère énergétique (conformation contrainte). Les atomes entourés sont initialement fixes mais redeviennent libres pendant la phase finale de relaxation, d'où le calcul d'un RMSD. Ce dernier est calculé sur les atomes lourds représentés en jaune, à savoir ceux provenant du hit de départ.

Néanmoins, des résultats préliminaires ont mis en évidence des composés ayant une énergie relativement favorable avec au moins une fonction amide de conformation peu probable (cis voire beaucoup plus rarement "ni cis ni trans"). Dans le contexte du growing/linking, ce type de liaison est systématiquement considéré comme un DDL explicite, d'où la possibilité d'obtenir des conformations d'amide non trans. En effet, la pénalité énergétique associée à ces "mauvaises" conformations peut être largement compensée par des interactions favorables entre le site et le linker, d'où l'introduction de ce filtre géométrique supplémentaire : les composés, dont la meilleure pose intègre une fonction amide de conformation indésirable, sont mis de côté. On entend par conformation indésirable une déviation de plus de 10° par rapport à la conformation trans (180°). L'utilisation conjointe de ces différents critères doit permettre d'écarter les linkers indésirables et d'enrichir la tête de liste avec des linkers appropriés.

Les contraintes initiales, bien qu'automatiques par essence, peuvent néanmoins être modulées par l'utilisateur. En effet, il peut être souhaitable de débloquer par exemple un DDL en amont de la jonction "fragment-linker". Ce cas est observable au niveau de la Figure 49 (à gauche) : le fragment initial intègre la fonction amide, mais la torsion correspondant à la liaison thiophène--amide a

néanmoins été débloquée lors de la simulation. En pratique, il suffit de déclarer les numéros (ID) des atomes lourds à débloquer (les hydrogènes associés le seront également), et les DDL supplémentaires seront automatiquement identifiés. Dans un souci de simplicité et de cohérence, ces numéros doivent rester les mêmes quel que soit le ligand criblé. Ceci est facilement obtenu avec une définition des atomes lourds issus du hit de départ avant ceux des parties variables (linkers) au sein des fichiers moléculaires. Un fichier unique de configuration, nommé "unlock", est donc suffisant pour l'ensemble du criblage virtuel.

Enfin, un aspect notable de S4MPLE concerne la possibilité d'incorporer une certaine flexibilité au niveau du site. Celle-ci est fondamentale pour réaliser des simulations plus ou moins proches du modèle main-gant évoqué au §1.2.1 :

- la liaison d'un ligand peut nécessiter des réarrangements conformationnels du site de liaison
- un mouvement de chaîne latérale peut ouvrir l'accès à une poche initialement inaccessible

De plus, il peut être intéressant de débloquer les atomes du site à proximité du fragment initial et des linkers pendant la phase finale de relaxation. Cela favorise grandement l'élimination d'éventuels clashes et l'optimisation des contacts favorables ligand-site
Cette capacité est mise en évidence à travers un exemple concret (décrit ultérieurement au §5.4.3) où le déblocage de certains DDL du site est absolument nécessaire.

5.3 Les simulations de validation du protocole

Ce protocole est appliqué de manière rétrospective à divers cas d'étude, extraits de la littérature du FBDD, dans un but de validation de la méthode. Une recherche bibliographique a mis en évidence plusieurs travaux incluant à la fois un hit primaire de type fragment, un lead optimisé à partir de celui-ci, et des données structurales. Les fragments incluant au moins une étiquette RECAP sont sélectionnés dans la mesure où le but de ces études rétrospectives est de valider de manière rigoureuse la stratégie décrite précédemment.
Trois cibles émergent de cette recherche : le facteur X activé (FXa), la protéine chaperonne HSP90 et la protéine de capture de l'acétylcholine ("Acetylcholine Binding Protein" - AChBP). Les deux premières sont des cibles d'intérêt thérapeutique évident étant donné les graves pathologies qui peuvent y être associées (respectivement, thrombose [202] et cancer [203]). L'AChBP est une protéine modulant l'influx nerveux *via* la capture du neurotransmetteur acétylcholine au niveau de la fente synaptique [204]. Mais contrairement aux deux cibles précédentes, l'AChBP n'est pas d'un intérêt

pharmaceutique direct puisqu'elle n'est produite que chez certains gastéropodes. Cependant, il s'agit d'une protéine soluble qui possède une forte homologie de séquence avec le domaine extracellulaire du récepteur nicotinique de l'acétylcholine (nAChR). Ce dernier est un récepteur membranaire ionotrope sensible à la fixation d'un ligand, et l'accès à sa structure 3D (X-ray) reste très difficile. Ainsi, l'AChBP sert en pratique de modèle expérimental pour l'étude du récepteur nAChR [204, 205].
A travers ces différentes cibles et les données expérimentales associées, tous les scénarios de growing/linking évoqués ci-dessus sont clairement mis en évidence :

- la cible FXa est associée à trois études indépendantes (deux simulations de type growing et un cas de linking) et le site de liaison est créé à partir de la structure 4A7I [206]
- une optimisation de type linking est réalisée pour la cible HSP90 et le site est généré à partir de la structure 3HZ1 [196]
- une optimisation de type growing est effectuée pour la cible AChBP dans un site partiellement flexible (obtenu à partir de la structure 2Y54 [207]). Toutes les chaînes latérales qui pointent vers le ligand et situées à 6 Å de l'ammonium du fragment initial sont débloquées ; en pratique cela correspond à 5 résidus qui sont Y53, Y91, W145, Y186 et Y193

Etant donné qu'une partie du ligand est contrainte dans ce type de simulation, un nombre plus restreint de cycles est nécessaire par rapport à un docking classique. Concrètement, 200 générations sont réalisées pour toutes les simulations, à l'exception de celle impliquant un site flexible où un nombre de 400 générations est employé.

Dans le second article [197], on a illustré la capacité de S4MPLE à gérer quelques molécules d'eau explicites comme des ligands additionnels afin de prendre en compte des LH indirectes (ligand -- H_2O -- site). On s'est posé la question de savoir si cette capacité pouvait être intégrée telle quelle au protocole FBDD, et surtout si elle pouvait avoir un impact positif (par exemple, dans le classement des linkers par ordre de priorité).
Une recherche bibliographique supplémentaire est réalisée afin d'identifier des exemples concrets permettant de mettre en évidence des situations où des molécules d'eau sont fondamentales dans le processus d'optimisation d'un fragment initial. Ce type de données ouvre la voie à des simulations, difficiles mais ambitieuses, de growing/linking incluant des molécules d'eau libres. Malheureusement, cette recherche n'a pas permis d'extraire un cas d'étude lié au FBDD, contrairement aux cibles antérieures. L'analyse du jeu préalablement utilisé lors de la validation de la capacité multi-entités de S4MPLE (ligand + molécules d'eau) [197] a néanmoins mis en évidence plusieurs cas, notamment celui

de l'Acétolactate Synthase (AlSynth), qui pourraient être utilisés dans ce sens, à condition que certaines hypothèses soient émises puisqu'il n'y a pas à proprement parler de fragment initial et de lead évolué à partir de celui-ci.

Une optimisation de type growing avec deux molécules d'eau libres est effectuée pour la cible AlSynth, et le site de liaison est généré à partir de la structure 1T9B [208]. Du fait de l'absence du fichier de configuration hot_spots (cas de toutes les simulations de growing/linking), les molécules d'eau libres, qui sont considérées comme des ligands à part entière, sont spécifiquement échantillonnées autour des points d'ancrage automatiques (fragment + linker + H_2O). Dans ce cas d'étude AlSynth, une simulation "classique" sans aucune molécule d'eau est effectuée en plus de celle en incluant afin de voir leur impact. L'ajout des deux H_2O se fait simplement *via* une concaténation de fichiers moléculaires, et les paramètres TIP3P [209] sont utilisés pour les modéliser.

Un complément d'information est disponible dans l'article associé (voir le §5.5) pour l'ensemble des données expérimentales de référence et des simulations. Les principaux points sont résumés dans le tableau 1.

5.4 Résultats et discussion

Plusieurs critères objectifs sont utilisés pour juger de la pertinence des résultats des simulations rétrospectives. Certains sont liés à la phase préliminaire de création de la chimiothèque focalisée et d'autres au criblage virtuel à proprement parler :

- capacité à générer au sein de la chimiothèque le ligand de référence et/ou des analogues très proches à partir du fragment initial
- capacité à reproduire le mode de liaison, expérimentalement connu, du ligand de référence
- capacité à classer ce ligand et/ou des analogues en tête de liste par rapport aux autres molécules générées qui sont considérées dans leur grande majorité comme des leurres ("decoys"). Le terme analogue fait essentiellement allusion à des molécules très similaires au ligand connu (par exemple, un tolyl à la place d'un phényl). Bien entendu, des molécules considérées comme des analogues peuvent néanmoins être inactives sur la cible. De même que certains leurres pourraient aussi être actifs. Dans ce contexte rétrospectif, il y donc a une focalisation sur les ligands validés expérimentalement puisque c'est la seule information réellement digne de confiance

- capacité à reproduire certains changements conformationnels connus lorsque des DDL du site sont débloqués
- capacité à favoriser des composés dont le linker est impliqué dans des interactions (LH) médiées avec le site lors de l'ajout de molécules d'eau explicites et libres

Les résultats relatifs aux différentes cibles sont brièvement décrits et discutés ci-dessous, sachant que le détail est systématiquement disponible dans l'article associé (voir le §5.5).

5.4.1 Cible FXa

L'intégralité de ces simulations repose sur les données de l'article de Nazaré *et al* [206]. Il discute notamment de la décomposition d'un petit inhibiteur de FXa de haute affinité (LE=0,49, ce qui est très élevé) en plusieurs fragments, et les K_i sont mesurés pour chaque composé. Cette étude, comme d'autres auparavant [53], met en évidence le bonus énergétique obtenu avec l'utilisation d'un linker approprié (voire avec la liaison directe de deux fragments dans le contexte du merging).
Dans le cadre de ce travail, le but est de pouvoir générer l'inhibiteur connu à partir de différentes sources et de le classer en tête de liste pour chaque criblage virtuel. Malheureusement, les structures des complexes avec les différents fragments n'ont pas été résolues (ce point a été confirmé par les auteurs de l'étude). Dès lors, deux hypothèses de travail doivent être émises :

- hypothèse 1) le fragment "5-chloro-2-thiophène-carboxamide" se lie dans la poche S1. D'autres ligands porteurs de ce groupement le situent également dans cette cavité (codes PDB : 2P95, 2VVC, 2VVU, 2VVV, 2VWL, 2VWM, 2VWN, 2VWO, 2W26 et 3TK6). De plus, le docking de ce fragment avec S4MPLE converge aussi vers ce mode de liaison (RMSD < 0,5 Å, voir la Figure 50). Cette première hypothèse semble donc très raisonnable
- hypothèse 2) le fragment "4-amino-1-isopropyl-pipéridine" se lie dans la poche S4. Tout comme pour le fragment chlorothiophène, un groupement pipéridine se retrouve incorporé dans d'autres inhibiteurs de FXa, et il est localisé dans la même cage aromatique (codes PDB : 2BOH, 2BQ6, 2BQ7 et 2BQW). De plus, le docking de ce second fragment avec S4MPLE converge aussi vers ce mode de liaison (RMSD ≈ 0,5 Å, voir la Figure 50). Cette seconde hypothèse semble également très raisonnable

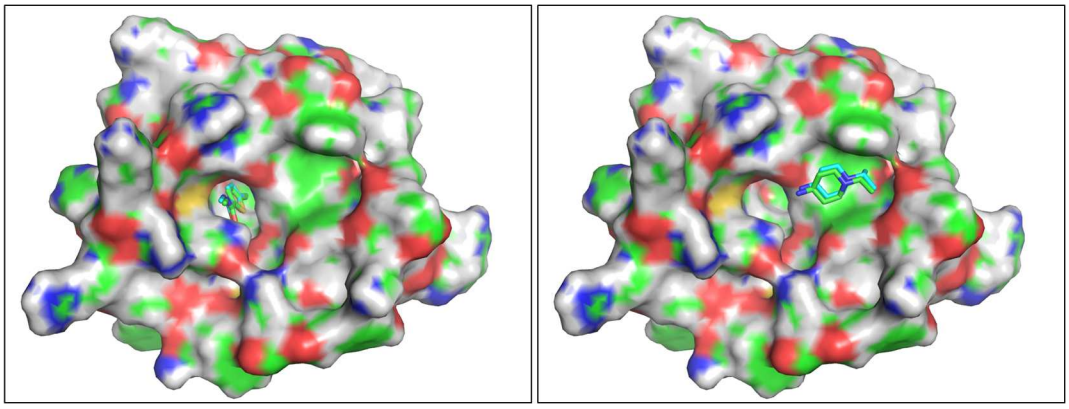


Figure 50: Superposition du mode de liaison expérimental (atomes de carbone en vert) et de la meilleure pose en docking (atomes de carbone en bleu) pour les deux fragments considérés. A gauche : cas du fragment se liant dans la poche S1. A droite : cas du fragment se liant dans S4.

Growing I

Ce premier cas d'étude nommé "growing I", partant d'un fragment de LE moyen (0,36), est relativement simple puisqu'il suffit de trouver un groupement de taille modérée (isopropyl-pipéridine) pour recréer l'inhibiteur de référence. Mais il a néanmoins toute sa place dans le contexte d'une première étape de validation de méthode, d'autant plus qu'un bon rang n'est pas acquis par avance.
Dans le cas où tous les atomes du fragment initial seraient fixes, seule la poche S4 serait réellement échantillonnée. Plusieurs DDL du fragment initial sont donc débloqués afin de réaliser un EC plus réaliste du ligand dans le site. Au final, seule la partie 5-chloro-2-thiophène-carboxamide reste contrainte. Comme attendu, l'inhibiteur est automatiquement généré. Il figure également en tête de liste (rang #3, voir la figure 3 de l'article) du criblage virtuel portant sur une banque de plus de 4000 composés, et son mode de liaison est fidèlement reproduit. Par conséquent, cette première étape est validée au regard des critères énoncés au début de ce sous-chapitre.

Growing II

Ce second cas d'étude nommé "growing II", qui part d'un micro-fragment de LE très élevé (0,58), est beaucoup plus ambitieux dans la mesure où une large portion doit être ajoutée afin de reproduire fidèlement l'inhibiteur de référence. Malheureusement, ce dernier n'est pas exactement reproduit à l'issue de la création de la chimiothèque focalisée. Ce large groupement d'environ 15 atomes lourds n'est pas présent dans la banque de linkers, d'où l'impossibilité de générer le ligand connu à l'identique. Il s'agit d'une des limites de cette approche empirique.

Le criblage virtuel a malgré tout été entrepris car certaines molécules générées peuvent être considérées comme assez proches de la référence. Par exemple, le linker inclut un groupement éthylsulfonamide suivi d'un large groupement hydrophobe terminal (toutefois, sans cation).
Sur les 8000 molécules criblées, une molécule du top 10 est similaire d'un point de vue pharmacophorique au ligand d'intérêt (voir les figures 4 et 5 de l'article). D'autres composés potentiellement intéressants, mais réellement plus éloignés en terme de points pharmacophoriques, sont également présents dans le top 100 final (voir la figure 6 de l'article). Enfin, les énergies issues du criblage sont moins favorables que celle du ligand de référence simulé dans les mêmes conditions.
Bien qu'imparfaits car la molécule exacte n'a pu être générée automatiquement (problème essentiellement combinatoire vu le nombre d'atomes lourds à ajouter pour passer du fragment au lead), les résultats de cette seconde étude plus ambitieuse sont toutefois encourageants, dans la mesure où plusieurs molécules partageant des points pharmacophoriques avec l'inhibiteur de référence ont été générées et plutôt bien classées (top 10 et/ou top 100) au regard des 8000 composés criblés.

Linking

La chimiothèque générée à partir des deux fragments contient 650 composés et l'inhibiteur de référence y est inclus. Le composé d'intérêt est classé en tête de liste comme espéré (rang #6, voir la figure 7 de l'article), et le linker adopte bel et bien le mode de liaison expérimental. Selon les critères évoqués auparavant et en tenant compte des deux hypothèses concernant les modes de liaison des fragments, ce premier cas d'étude de type linking apparaît comme un succès.

5.4.2 Cible HSP90

Cette simulation de linking repose à nouveau sur un article lié à un projet expérimental basé sur des fragments [196]. Les deux principales différences par rapport à celle de FXa sont :

- une structure X-ray à la fois pour les fragments (complexe ternaire 3HZ1) et pour le ligand optimisé (3HZ5). Il n'y a donc aucune hypothèse vis-à-vis des modes de liaison des fragments
- un inhibiteur final de plus grande affinité mais n'incorporant pas exactement les deux fragments initiaux (un groupement ester est remplacé par une chaîne aliphatique), ce qui a une conséquence négative directe sur la capacité à pouvoir générer exactement le composé d'intérêt

Sachant qu'il y a assez peu de données structurales disponibles dans une optique de linking (structure du complexe ternaire initial et structure du complexe avec le ligand optimisé), il a toutefois été convenu d'entreprendre cette étude.

Des composés relativement similaires au ligand connu, mais intégrant systématiquement la fonction ester, sont bien entendu présents dans la banque à cribler. Au final, ils sont retrouvés très rapidement lors du parcours de la liste triée par énergie (voir la figure 8 de l'article).
Malgré l'incapacité à reproduire à l'identique le ligand de référence pour la raison évoquée ci-dessus, ce second criblage virtuel de type linking a logiquement favorisé des composés équivalents à l'inhibiteur connu selon le critère distance topologique (défini comme le nombre minimal de liaisons covalentes entre deux atomes) entre les hétérocycles pyrazole et purine.

5.4.3 Cible AChBP

Ce cas d'étude a pour but de démontrer la capacité de S4MPLE à réaliser une optimisation *in silico* tout en ayant un site de liaison partiellement flexible. Les données expérimentales provenant de l'article de Edink *et al* [207] sont sans aucune ambiguïté : les pK_i du fragment initial et de deux molécules évoluées à partir de celui-ci sont connus, tout comme les structures X-ray des complexes associés. L'analyse de ces dernières met en évidence un réarrangement conformationnel du site (essentiellement la Y91) entre la liaison du fragment et celle des ligands optimisés. Par conséquent, ces données expérimentales sont tout à fait pertinentes dans l'optique du déblocage de certains DDL du site pendant le criblage virtuel.
L'optimisation du fragment initial se résumant globalement à l'ajout d'un groupement éthyl-benzene, il n'est pas étonnant de pouvoir créer automatiquement les deux inhibiteurs d'intérêt. Au final, plus de 4500 composés sont criblés virtuellement dans le site flexible d'AChBP.
Contrairement aux précédents criblages où les ligands de référence (si générés) étaient toujours présents dans les meilleures molécules selon les critères de classement choisis, les deux inhibiteurs d'intérêt se retrouvent ici un peu plus loin dans la liste, respectivement dans le top 4% et le top 10%. Toutefois, un analogue très proche (un tolyle remplaçant le groupement phényle, voir la figure 11 de l'article) est situé dans le top 3%, et le méthyle additionnel se retrouve à environ 3,5 Å de plusieurs atomes de carbone des chaînes latérales de Y91 et W145. Enfin, le réarrangement conformationnel attendu, à savoir la rotation de la chaîne latérale de Y91 tout en maintenant les autres dans leur état initial, est bien reproduit (voir la figure 12 de l'article).
Malgré un classement légèrement moins performant, sous l'hypothèse pessimiste qu'aucune molécule de meilleur rang ne soit active, ce criblage virtuel plus réaliste (car impliquant un site flexible) a quand même permis de répondre favorablement aux principaux critères de succès évoqués au début de ce sous-chapitre.

5.4.4 Cible AlSynth

Ce dernier cas d'étude a pour enjeu de mettre en évidence la capacité de S4MPLE à faire une optimisation *in silico* tout en incluant des molécules d'eau libres. En effet, il est connu depuis longtemps que celles-ci jouent un rôle essentiel dans les phases d'optimisation.

La seule donnée structurale disponible est la géométrie du complexe AlSynth-inhibiteur (PDB 1T9B). Par conséquent, une hypothèse est réalisée quant à la liaison du fragment initial ; ce dernier étant défini comme une sous-structure de l'inhibiteur de référence en le coupant au niveau de sa jonction urée-sulfonamide. Le fragment considéré est le (4-méthoxy-6-méthyl-1,3,5-triazin-2-yl)-urée, et l'on suppose son mode de liaison identique à celui observé dans l'inhibiteur final.

Plus précisément, le but est de montrer que l'ajout des H_2O a un impact positif au niveau du rang pour le composé de référence. Son mode de liaison implique une interaction médiée par une molécule d'eau au niveau de la partie considérée comme linker. Parallèlement, on s'attend à un classement moins efficace pour la simulation classique, puisque la complémentarité ligand/site est moins probante au sein d'un environnement privé de la possibilité de réaliser des LH médiées (ligand - - H_2O - - site). Il est à noter que l'une des molécules d'eau sert de contrôle dans la mesure où elle devrait systématiquement réaliser deux LH médiées entre le groupement urée du fragment initial et le site. La seconde molécule d'eau est ajoutée afin d'être potentiellement impliquée dans une interaction médiée impliquant le linker.

Dans le cadre de la simulation dite classique, le ligand d'intérêt obtient, comme attendu, un rang acceptable mais au-delà du top 1% : il est classé au rang #62 sur un total de 1550 composés criblés (top 4%). A l'inverse, ce même composé obtient un rang bien plus favorable (rang #6, voir la figure 14 de l'article) lorsque les molécules d'eau sont incorporées dans la simulation. Les molécules d'eau étant échantillonnées autour du ligand et d'elles-mêmes, elles peuvent occuper une place de choix lorsque le linker possède un groupement pouvant être impliqué dans une LH médiée. A l'inverse, lorsque la structure du linker est dépourvue de ce type de point d'ancrage, la seconde H_2O reste ici majoritairement à la sortie du site.

D'un point de vue structural (voir la figure 13 de l'article), le chlorophényle de la référence occupe la cavité attendue, et les deux molécules d'eau sont également prédites avec précision (RMSD ≈ 0,5 Å). L'interaction supplémentaire entre l'un des oxygène du groupement sulfonamide et le résidu K251 explique le saut favorable, au niveau du critère rang, de l'inhibiteur de référence entre les deux types de simulation de growing.

Enfin, et c'est à la fois rassurant et somme toute logique, il existe un chevauchement non négligeable au niveau des 50 meilleurs composés de chaque simulation : 30 molécules sont communes aux deux

têtes de liste. En effet, ces molécules, dont le linker comporte des groupements hydrophobes interagissant favorablement avec V191, P192, A195, A200, F201 et K251, n'ont aucune raison d'être mal classées dans le contexte d'un EC impliquant des molécules d'eau.
Bien que reposant sur des hypothèses de travail, nécessitées par un manque de données expérimentales très spécifiques, cette dernière étude plutôt ambitieuse a permis de mettre en évidence une capacité singulière de l'outil S4MPLE dans le contexte d'une optimisation *in silico* de type growing.

Pour conclure, cette stratégie d'optimisation virtuelle de fragments, reposant sur plusieurs outils (GenLinkersDB, JMolEvolve et S4MPLE), est validée sur la base de ces différentes études rétrospectives. De plus, des aspects singuliers ont été mis en évidence à travers ces simulations d'optimisation *in silico*, notamment l'incorporation d'une certaine flexibilité au niveau du site et la gestion de molécules d'eau libres. Ces résultats ouvrent des perspectives intéressantes pour des projets prospectifs. A ce titre, le dernier chapitre décrit l'emploi de cette stratégie dans le cadre d'un projet interne de l'entreprise NovAliX.

5.5 Article III

L'article III est inséré dès la page suivante dans sa version actuelle (“galley proof”) [199].

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

In Silico Fragment-Based Drug Discovery: Setup and Validation of a Fragment-to-Lead Computational Protocol Using S4MPLE

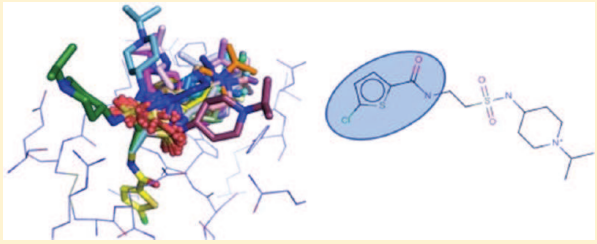
Laurent Hoffer,[†,‡] Jean-Paul Renaud,[‡] and Dragos Horvath[†,*]

[†]Université de Strasbourg, 1 rue B. Pascal, Strasbourg 67000, France

[‡]NovAliX, BioParc, bld Sébastien Brant, BP 30170, Illkirch 67405 Cedex, France

Supporting Information

ABSTRACT: This paper describes the use and validation of S4MPLE in Fragment-Based Drug Design (FBDD)—a strategy to build drug-like ligands starting from small compounds called fragments. S4MPLE is a conformational sampling tool based on a hybrid genetic algorithm that is able to simulate one (conformer enumeration) or more molecules (docking). The goal of the current paper is to show that due to the judicious design of genetic operators, S4MPLE may be used without any specific adaptation as an in silico FBDD tool. Such fragment-to-lead evolution involves either growing of one or linking of several fragment-like binder(s). The native ability to specifically "dock" a substructure that is covalently anchored to its target (here, some prepositioned fragment formally part of the binding site) enables it to act like dedicated de novo builders and differentiates it from most classical docking tools, which may only cope with non-covalent interactions. Besides, S4MPLE may address growing/linking scenarios involving protein site flexibility, and it might also suggest "growth" moves by bridging the ligand to the site via water-mediated interactions if H_2O molecules are simply appended to the input files. Therefore, the only development overhead required to build a virtual fragment→ligand growing/linking strategy based on S4MPLE were two chemoinformatics programs meant to provide a minimalistic management of the linker library. The first creates a duplicate-free library by fragmenting a compound database, whereas the second builds new compounds, attaching chemically compatible linkers to the starting fragments. S4MPLE is subsequently used to probe the optimal placement of the linkers within the binding site, with initial restraints on atoms from initial fragments, followed by an optimization of all kept poses after restraint removal. Ranking is mainly based on two criteria: force-field potential energy and RMSD shifts of the original fragment moieties. This strategy was applied to several examples from the FBDD literature with good results over several monitored criteria: ability to generate the optimized ligand (or close analogs), good ranking of analogs among decoy compounds, and accurate predictions of expected binding modes of reference ligands. Simulations included "classical" covalent growing/linking, more challenging ones involving binding site conformational changes, and growth with optional recognition of putatively favorable water-mediated interactions.



1. INTRODUCTION

1.1. Focus and Outline of this Work. This paper describes the use of S4MPLE for the in silico fragment-to-lead evolution step in Fragment-Based Drug Discovery (FBDD). This approach typically implies fragment concatenation operations, starting from experimentally validated bound fragment-like hits in the target site and has been addressed by several dedicated tools. It includes either linking of several independently binding fragments or growing of one fragment-like compound to lead-like size. For clarity purposes, the single term "linker" will refer to added atoms/groups with respect to starting hit(s) in both linking and growing contexts. In silico FBDD is basically a (structure-based) subdomain of de novo drug design (DND, see section 1.3), requiring a starting moiety of validated affinity and of known binding mode to the target.

The focus of this paper does therefore not lie on the generic stumbling blocks of FBDD/DND fragment-to-lead evolution: fragment library design, chemical space coverage, assessment of generated compound feasibility, etc. (see section 1.3). It is to show that, with a minimalistic set of tools for linker management and focused library generation, both classical and more challenging in silico fragment-to-lead evolution can be successfully carried out with a generic conformational sampling program without further adaptation. In silico FBDD emerges, if one may say so, as a collateral benefit of the versatility of the operators designed within S4MPLE.

S4MPLE, a Lamarckian genetic algorithm for generic conformational sampling, has the particularity of full control over all the degrees of freedom (DoF). This is due to a deceptively simple and yet fundamental choice of generic operators recombining the local geometries of various molecular moieties. Rather than distinguishing, like many classical sampling/docking programs, between intra-molecular DoF associated to covalent bonds/torsional axes and intermolecular Euler angles/quaternions, S4MPLE completely ignores this arbitrary classification and proposes a moiety recombination scheme driven by favorable contact formations. The latter may be covalent bonds as much as

Received: January 10, 2013
Published: March 28, 2013

hydrogen bonds or hydrophobic contacts. It has been shown[1] how this choice allows S4MPLE to (A) simultaneously dock several entities, including free crystallographic waters added—or removed, when replaced by ligand groups—in order to predict multi-molecular complex geometries. Reversely, S4MPLE may (B) dock a piece of a molecule like a linker into a formal active site of the protein including the initial fragment-like hit(s) to which the linker is covalently connected. In addition, it may, if needed, also unlock DoF of specific protein moieties (linking/growing in flexible active sites). Eventually, it flawlessly supports a combined (A+B) scenario, trying to simultaneously include waters that might indirectly anchor the evolved lead to the site, i.e., formally "grow" the fragment by adding both covalent substituents and non-covalent partners (solvent molecules) mediating favorable interactions.

1.2. Short Overview of Fragment-Based Drug Design. FBDD is an emerging technology that builds ligands from low complexity compounds called fragments.[2−4] Because of their small size, they often have low potency. Therefore, very sensitive methods, such as biophysical techniques (NMR, SPR, MS, etc.), are employed in the screening campaigns.[5−8] Besides, X-ray crystallography and NMR are both able to resolve the binding modes of fragments within the binding site—this information greatly facilitates the fragment-to-lead optimization using Structure-Based Drug Design (SBDD) strategies.[9] A main advantage of FBDD is that fragment hits often have strong affinity with respect to their pretty small size (i.e., good ligand efficiency LE[10]). Also, hit rates in fragment collections are higher than in drug-like ones.[11] Screening an appropriate diverse library (e.g., 10^4−10^5 fragments[12]) is in principle equivalent to the exploration of a much larger drug-like space domain than accessible by traditional high-throughput screening.

There are three main strategies to optimize fragments into high affinity drug-like ligands: merging, linking, and growing.[13] Growing looks like usual hit2lead optimization because chemical groups are simply added to one single fragment. Both merging and linking strategies assemble two fragments that bind simultaneously. However, only linking adds additional atoms by means of a spacer. Ideally, binding modes of original fragments must be conserved in the final compound after merging/linking without adding undesirable strain energy, enabling the bonus gain in affinity associated with successful linking.[14] Linking is more challenging than growing, and there are some very interesting success stories where two low affinity fragments are linked into one high affinity ligand using a spacer.[15−17] However, it is not surprising that a lot of success stories in FBDD used growing optimization, with the support of X-ray data.[2]

1.3. In Silico Fragment-Based Drug Design. In parallel to the growing interest in experimental FBDD, computer-based strategies, targeting main bottlenecks of FBDD (library design, fragment hits identification and their subsequent optimization into leads), are being developed.[18,19]

Chemoinformatics[20] can be used to design a fragment library with high scaffold diversity, while only incorporating compounds containing chemical functions known to be easily alterable in order to facilitate their subsequent growth.[21] When the 3D structure of the target is available, a prioritized fragment subset for experimental testing can be built by means of virtual screening approaches, such as docking or pharmacophore screening. Docking has been more or less successfully used to predict the binding mode of drug-like ligands,[22] and in practice, this technique can be applied to fragment-like compounds too. Small size implies low numbers of putative favorable contacts, thus narrow binding energy ranges. While this challenge of scoring may become critical with very small compounds, a recent study[23] concluded that there is no significant difference in docking accuracies of drug-like ligands (sampling issues) vs fragment-like compounds (scoring issues). Good results were also reported with Glide[24] in both fragment redocking studies[25] and virtual screening.[26]

Finally, fragment→lead optimizations can be suggested by de novo design (DND) algorithms or similar in silico growing/linking strategies. The main goal of DND is to generate chemically sound and original structures predicted to have desired physicochemical and biological properties. The process can often be guided by using both structure-based docking, pharmacophore constraints,[27,28] or any other 2D/3D property prediction model.[29] Compound growth is based on chemical linkage rules, such as the RECAP[30] or BRICS[31] reaction schemes, in order to increase the synthetic accessibility of suggested compounds. LUDI,[32] Legend,[33] SPROUT,[34] SkelGen,[35] GroupBuild,[36] and LigBuilder[37] are common DND programs. TOPAS[38] and Flux[39] are RECAP-based DND software. For more details, see Schneider and Fechner.[40] Albeit a marginal topic in this work, it is important to highlight the importance of synthesis tractability of virtually generated compounds.[41−45] For example, generic reaction schemes[46] (reactants→products, with −R groups) are encoded in order to mimic real chemistry, as in the DOGS procedure.[42]

Various DND algorithms have been either explicitly developed to address some of FBDD issues or can be used in this context. The search for energetically favorable locations for various chemical entities is already known as probe mapping, and MCSS,[47,48] FTMap,[49−51] or GRID[52] are popular programs of that kind. After a search step that includes various organic probes (e.g., isopropanol, cyclohexane, benzene, etc,), FTMap performs a cluster analysis in order to identify putative hot spots. The latter are defined as consensus areas over several probes and obviously with favorable energies. Using the small but diverse "Congreve data set",[2] Hall et al. showed that their FTMap strategy was able to identify the main hot spots (defined as the subpocket that binds the studied fragment hit) as the top-ranked cluster, and some other clusters matched lead moieties evolved from these starting fragments.[50] Recore has been developed to perform scaffold hopping,[53] and relies on "exit vectors" defining moieties to replace. However, such vectors can be used to define the direction in which a compound must be optimized, hence the possible use of Recore within FBDD projects. Besides, pharmacophore and binding site constraints can be added to guide the evolution process. In the FBDD context, fragments linking algorithms are of particular interest when binding modes of simultaneously fragment binders are known. Two recent tools are specialized in the linking of predocked fragments: CONFIRM[54] and GANDI.[55] In CONFIRM, a pre-prepared library of "bridges" is screened to extract appropriate linkers in order to join predocked fragments. In this study, the Glide docking tool[24] was used to locate the starting fragments, and the validation of the method was performed on retrospective cases. The program GANDI uses fragments poses (e.g., generated by the SEED program,[56] developed in the same laboratory) as inputs too. It relies on a genetic algorithm involving an island paradigm and a tabu search during the sampling stage, and the energy function is based on the force field formalism.

To conclude, the three main issues of FBDD can be addressed, with more or less success, by computer-based approaches. Several reviews described both sides (experimental and computational) of fragment-based drug discovery.[4,57]

1.4. Introduction of the Virtual Screening Protocol. In order to apply S4MPLE as a virtual screening engine for suited

dx.doi.org/10.1021/ci4000163 | J. Chem. Inf. Model. 2013, 53, 836−851

linkers, the only additional developments required by this virtual fragment→ligand growing/linking procedure were two Java programs based on ChemAxon API:[58] a linker generator and a focused library builder, automatically connecting compatible linkers to the prepositioned reference fragments.

In a nutshell, S4MPLE may be employed to specifically optimize positions of linker atoms only, in search of the most stable linker conformation strainlessly connecting the prepositioned fragment(s)—one in growing and two in linking. Thus, the valuable information of the initial positioning of the starting fragment(s) can be exploited. S4MPLE will first optimize fragment-linker and intra-linker covalent structures under constraints given by the fixed ends of prepositioned fragments and the excluded volume of the (fully rigid or partially flexible) target site. While initial fragment(s) atoms do not move, but the site is likely to adapt its geometry to the linker, site-specific DoF can be enabled. Eventually, constraints are lifted and the so-far best poses of the now fully flexible ligand are allowed to freely relax. In order to identify putatively favorable water-mediated interactions during growing or linking, the user merely has to concatenate the desired number of water molecules at the end of the ligand candidate files produced by the library management tools.

This strategy (see Methods) is retrospectively applied to case studies from the FBDD literature. Its success is measured according to several criteria:

- ability to create the known reference ligands (and/or close analogs) from starting fragment hits
- ability to prioritize known actives relatively early within the ranked list of generated compounds (supposed inactive, although some of them can be real binders)
- ability to accurately reproduce the experimental binding mode of reference ligands
- ability to predict binding site conformational changes (when site flexibility is enabled)
- ability to produce a better ranking/prioritization of known actives involved in water-mediated interactions, upon explicit adding of water molecules to the simulation. Comparatively, the water-free growing procedure is expected to penalize the experimental ligand (due to unaccounted water-mediated hydrogen bonds).

2. METHODS

The virtual fragment→ligand growing/linking protocol is explained below. Eventually, data sets used in the retrospectives studies are introduced.

2.1. S4MPLE. S4MPLE (Sampler for Multiple Protein−Ligand Entities), a molecular modeling program based on a Lamarckian genetic algorithm, has been described previously.[1,59] This conformational tool, allowing the selection of considered degrees of freedom of the system, can be employed for a wide variety of simulation types: conformational sampling of ligands or small peptides and docking of both fragment-sized and drug-sized compounds. There is no explicit limit with respect to the number of considered entities—simultaneous docking of multiple ligands is supported. The energy function relies on the force field (FF) formalism and uses AMBER[60] and GAFF[61] to respectively simulate peptide and small organic moieties of the considered system. Here, all simulations are performed with the "Fit FF" energy scheme.[59] The control of conformational similarity is performed by a symmetry-compliant pair-based interaction fingerprint (PIF) that monitors two interaction types: close carbons contacts (based on C−C distance) and hydrogen bonds. Two configurations of the system are considered equivalent if the Hamming[62] distance between their fingerprints is lower than a user-defined threshold. The program is written in object-Pascal and used in command-line mode.

2.2. Fragment-to-Lead Development Tools. Two Java programs, GenLinkersDB and JMolEvolve, based on the ChemAxon API,[58] serve to virtually evolve starting fragment(s) into lead-like or drug-like molecules. In other words, they are used in order to create a focalized library around one (growing) or two (linking) starting hit(s).

2.2.1. GenLinkersDB. GenLinkersDB creates linker libraries by fragmenting a database following RECAP cleavage rules. The RECAP algorithm[30] runs through molecular structures, detects bonds of interest using a predefined dictionary, and cleaves them while assigning specific flags to the resulting "loose" valences of atoms. These flags encode the chemical context or neighborhood of the atom in its molecule of provenience, much like FF atomic types. Once a molecule has been fragmented, it is possible to build it back by connecting fragments containing complementary flags. As for example, when an amide bond is broken, two fragments are created with two distinct but complementary amide flags (e.g., amide:1 for N−* and amide:2 for C(=O)−*). An extensive fragmentation (including single atom fragments, cuts between heteroatoms and cycles) is performed to maximize the number of generated linkers. Fragments with one or two flags are respectively saved in "growing" or "linking" databases. As already mentioned and for convenience, the term "linker" will refer to the added atoms/groups in both cases (growing or linking). A mass threshold of 300 Da is defined because the goal consists in optimizing precursor compounds into Ro5-compliant[63] molecules. Raw fragment databases, immediately after the fragmentation step, are huge and contain many duplicates. These are removed in a two stages procedure:

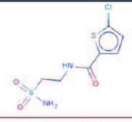
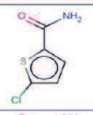
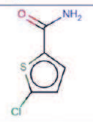
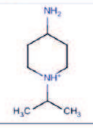
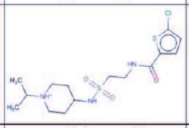
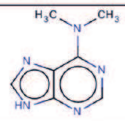
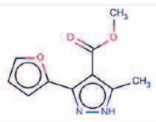
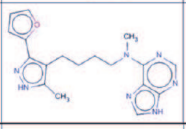
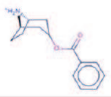
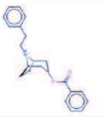
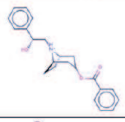
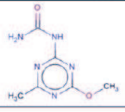
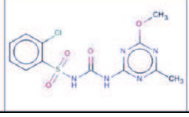
- Linkers are stored into different files, indexed by their mass (storage file M contains ligands of mass [M, M+1 (Da)].
- Specific functions (see below), based on ChemAxon graph isomorphism algorithms, are subsequently called to remove redundant entries within these smaller and more tractable files.

Indeed, linker uniqueness cannot be defined on the naked molecular graph but can be defined on the RECAP-flag property-labeled graph. A same skeleton may occur in conjunction to different or differently located RECAP flags. It should be noted that mass-splitting allows not only the acceleration of duplicate searches but also results in building complexity-ranked linker databases. The virtual evolution process can be started with the shortest linkers and stopped as soon as results show that desired linker size may have been passed.

The clean subset of the ZINC[64] version 12, containing more than 9 million molecules, is used as input file for the full fragmentation protocol (described above). The final numbers of linkers are about 2.2 and 3.6 million, respectively, for linking and growing databases. Subsets with lower mass limits or element constraints (e.g., only H, C, N, O, S, and P elements) are easily extractable therefrom.

2.2.2. JMolEvolve. JMolEvolve builds new molecules by merging one original precursor fragment and chemically compatible linkers based on the complementarity of the RECAP flags. Precursor fragments must be input as 3D molecule files with the coordinates corresponding to their docked or experimentally determined position relative to the active site and linkage points (e.g., fragment atom ID(s) to be used in linking/growing). This growing mode can be switched to linking mode by simply providing two precursors. In practice, the user must

Table 1. Description of All Growing/Linking Protocols[a]

Target	Protocol	Molecule ID	Structure	Binding data	RECAP flags	Linkers database	Filters	Generated & screened compounds	Binding site flexibility
FXa	Growing I (Small)	2		Ki ≈ 60 μM LE=0.36	sulphonamide	Growing Mass 250 Da HCNOSP	HAC < 12	≈ 4000 compounds Reference : yes Analogs : yes	Polar hydrogens
	Growing II (Large)	1		Ki ≈ 60 μM LE=0.58	amide	Growing Mass 250 Da HCNOSP	12 ≤ HAC ≤ 15 + various filters	≈ 8000 compounds Reference : no Analogs : yes	Polar hydrogens
	Linking	1		Ki ≈ 60 μM LE=0.58	amide	Linking Mass 150 Da	Canonical amide "C(=O)[NH]" preserved	≈ 650 compounds Reference : yes Analogs : yes	Polar hydrogens
		3		LE=0.27	amine or amide or sulphonamide				
	Targeted Reference Lead	**4**		Ki ≈ 0.002 μM LE=0.49					
Hsp90	Linking	11		IC50 ≈ 1500 μM	amine	Linking Mass 250 Da	/	≈ 300 compounds Reference : no Analogs : yes	Polar hydrogens
		12		IC50 ≈ 1000 μM	ester				
	Targeted Reference Lead	**13**		IC50 = 1.5 μM					
AChBP	Growing	21		pKi = 5.3 LE=0.43	amine	Growing Mass 200 Da HCNOSP	HAC < 14 + charged amine + various filters	≈ 4500 compounds References : yes Analogs : yes	Several sidechains + polar hydrogens
	Targeted Reference Leads	**22**		pKi = 7.5 LE=0.41					
		23		pKi = 7.0 LE=0.37					
AlSynth	Growing	31		pKi = N/A	urea	Growing Mass 200 Da	/	≈ 1550 compounds References : yes Analogs : yes	/
	Targeted Reference Lead	**32**		pKi = 7					

[a] It includes a 2D depiction of fragments with their binding data. Information about used linker database, employed RECAP flags, generated and screened ligands, filters, and flexible binding site moieties are provided, too.

specify a linker database, the considered RECAP flag(s), above-mentioned structure file(s) for precursor(s), and linkage points. Initially, a "chimera" ligand is built by plugging in the linker group into the reference molecular graph and submitting it to the

dx.doi.org/10.1021/ci4000163 | J. Chem. Inf. Model. 2013, 53, 836−851

ChemAxon conformer plug-in in order to obtain one relevant conformer thereof. It however stores the initial fragment coordinates as data fields to be used by S4MPLE in order to override the new coordinates produced by the ChemAxon tool. The user may specify if the major microspecies based on the ChemAxon pK_a plug-in should be computed. Eventually, hydrogens are added, and atomic partial charges (Gasteiger[65]) and various properties (heavy atom count, mass, number of rings, etc.) are computed and saved as an annotated MDL .sd file. The last ligand preparation step consists in the assignment of GAFF atomic types using antechamber[66] and their check with parmchk from the same software suite (AMBERTools[67]).

Both linking and growing strategies may generate a huge number of putative compounds. Common filters (e.g., heavy atom count HAC, rotatable bond count, ring count, acceptor count, donor count, chiral center count—see cutoff values in Table 1) are used to reduce the size of the final library to screen and to discard non-drug-like molecules.

2.3. Virtual Screening Protocol. S4MPLE is subsequently used to simulate the compounds within the binding site, ignoring the DoF of precursor fragments. These have by default a "fixed atom" status, like the majority—or totality—of site atoms. It is a good strategy, however, to unlock the immediate neighborhood of the linkage point(s) in precursor fragments, as local geometric rearrangements in the bound fragments may occur upon linking. In growing (Figure 1, top), S4MPLE will break down the unlocked part of the ligand into pieces on which genetic operators will apply. This is an unambiguous operation because for each rotatable bond the molecular moiety not including the fixed atoms will be taken. If a linker, however, is spanned between two fixed fragments (Figure 1, bottom), S4MPLE cannot detect explicit DoF to the linker part—either moiety around a linker bond contains fixed atoms and cannot qualify as mobile pieces associated to an explicit DoF. This issue is addressed similarly to the intra-cyclic conformer sampling problem: a bond at the interface fragment-linker is automatically selected and formally "broken". This allows S4MPLE to detect mobile pieces ranging from the other fixed fragment atoms to this "broken" bond. There will be no explicit DoF associated to the "broken" bond, but its end is implicitly subject to all the moves of the detected pieces. Because the bond is formally broken but not removed from the molecular energy expression, its harmonic terms (bond stretching and angle bending) will be minimized in linker poses respecting these covalent constraints. Because S4MPLE includes a chirality inversion penalty term, the "broken" bond may include chiral atoms as well.

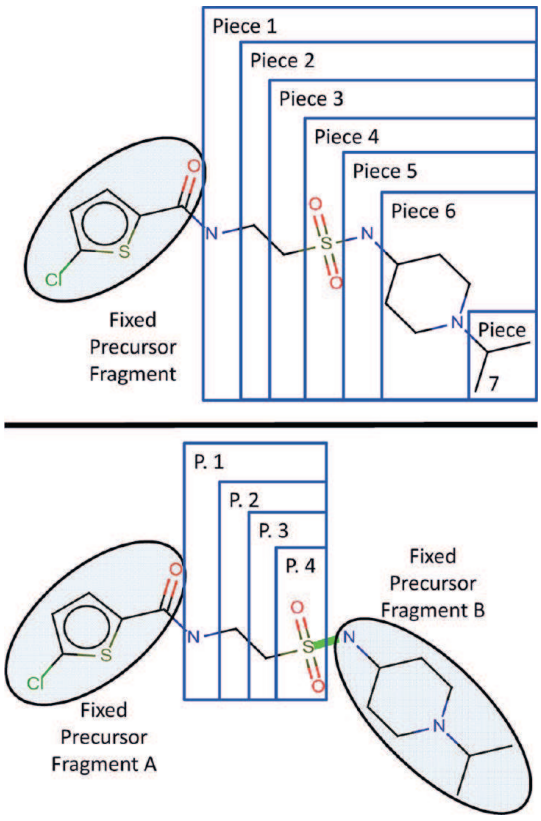


Figure 1. Automatic detection of moving pieces used in the evolutionary sampling routine. Top: Growing scenario (fixed precursor fragment). Pieces are hierarchically defined with respect to rotatable bonds, such as to include only moving atoms (here, amide bonds are considered rotatable as well). Bottom: Linking scenario, with the formally (green) "broken" bond. Moving pieces are all anchored to fragment A and end at the "broken" bond.

Genetic Algorithm-driven sampling consisted, for all simulations with rigid binding sites, in 200 generations of 20 individuals with all other parameters set to their default values.

In order to probe for putative water-mediated interactions (Acetolactate Synthase study case), no change of protocol is needed: it suffices to append the desired number of explicit water molecules (two, in this case) at the end of the multimolecule .sdf file. Albeit in the previous work we had reported specific protocols aimed to enhance the sampling of ligand-water-site bridges,[1] these were not used here (by default, waters tend to be positioned such as to preferentially form hydrogen bonds with ligand atoms or other waters, the natural "hot spots". No explicit hot spot list was provided in any reported simulations. Please refer to the cited publication and the S4MPLE user guide for more information). However, in order to account for the increased problem space, the 200 generation simulation was run in duplicate. Also, a duplicate 200 generation run without waters was performed in parallel, the goal being to compare the net impact of including explicit waters on the pertinence of the virtual screening.

When binding site flexibility was enabled (Acetylcholine Binding Protein study case), the sampling effort was extended to 400 generations. Like previously outlined,[1,59] a pretty large binding site is defined from the reference X-ray structure using the same workflow.

The sampling stage is followed by an optimization of all kept poses (all non-redundant poses at given interaction fingerprint dissimilarity threshold *minfpdiff*, within a given energy window +30 kcal/mol), now unlocking all the ligand DoF. Growing runs employed a *minfpdiff* = 0.005, while linking used a smaller *minfpdiff* = 0.001 because both ends are restrained. The fine minimization consisted in a systematic optimization of hydrogens from hydroxyl groups of the considered system (ligand and binding site, when unlocked), followed by the "exhaustive energy minimization" operator.[59] Figure 2 describes the full virtual screening protocol.

Given the initial fixing of precursor atoms, the final ligand conformers may retain an important intra-molecular strain component—all subsequent optimization notwithstanding. Therefore, scoring will not consider only ligand-site interaction terms but also include an estimate of this intra-molecular strain. The energetic criterion considered here thus equals the potential energy of the complex minus the potential energy of the best

dx.doi.org/10.1021/ci4000163 | J. Chem. Inf. Model. 2013, 53, 836−851

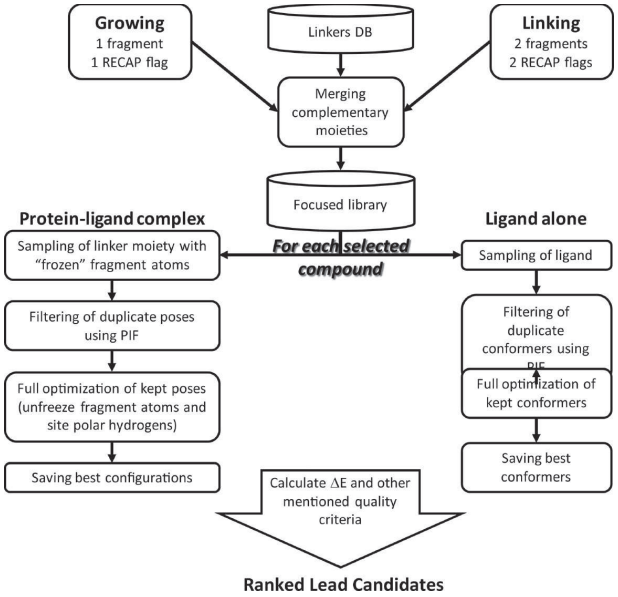


Figure 2. Workflow of the virtual screening protocol used in all studies.

conformer of the free ligand (the latter being obtained by a rapid run, consisting in 100 generations, in absence of the protein site). This would also include intra-site strain energies if site DoFs are enabled. With explicitly added water molecules, the basis energy level of waters in the free state is constant, so needs not be subtracted from the complex energy in as far as this score only serves for ranking/prioritization of candidates.

The final prioritization of the candidates is based on multiple criteria, the main being the force-field energy difference defined above. In the ideal case, this FF-based energy should discard bad linkers (constrained configurations, systematic clashes, etc.). Furthermore, ligands in which the initial fragments quit their initial poses, as soon as their DoF are unlocked at the final minimization stage, are unlikely to represent potent binders—or, at least they are outside the applicability domain of FBDD. A RMSD criterion is employed in addition to the energy: molecules with RMSD shifts over initially fixed precursor atoms beyond a user-defined threshold (respectively 0.5 and 1.0 Å in growing and linking procedures) are automatically discarded. Eventually, poses with obvious local structural flaws (that are but slightly penalized in terms of energy) are also discarded: here, this concerns molecules for which the best conformation within the binding site contains unlikely amide configurations (deviating by >10° from trans). Combining all these criteria is expected to rank interesting linkers in the top hit-list.

2.4. Data Sets. Several lead optimization strategies (starting from a fragment-like ligand) or explicit growing/linking approaches are extracted from literature, the goal being to virtually reproduce these success stories. Various targets emerged from this selection: activated Factor X (FXa), Heat Shock Protein 90 (Hsp90), and Acetylcholine Binding Protein (AChBP). Acetolactate Synthase (AlSynth) was used to exemplify growing runs with explicit waters. Table 1 sums up, for each data set, the main information about inputs, filters, and used options.

2.4.1. Factor Xa (FXa). This serine protease, involved in the coagulation cascade, has been intensively studied for decades, and several tens of complexes are deposited in the Protein Data Bank[68] (PDB). FXa is the activated form of the enzyme, and its inhibitors are putative anticoagulants drugs. A large variety of molecular scaffolds are present in these active compounds. Nazare et al. recently exemplified the affinity gain due to successful linking[14] from the deconstruction of small FXa inhibitors of high affinity.[69] The reference drug-like compound **4**, directly obtained from the merging of two low affinity fragments, has an impressive LE of 0.49.

The present is a retrospective study, where both in silico growing and linking strategies are applied to "precursors" from the fragmented reference ligand. Canonical RECAP flags amine, amide, and sulphonamide being present, the herein envisaged virtual evolution strategy to recreate the potent inhibitor **4** makes perfect chemical sense. The starting materials are one X-ray structure (4A7I) and binding data. Two growing and one linking protocols are carried out:

- Growing I (small): optimization from fragment **2** (LE = 0.36) with the sulphonamide extension. The chlorothiophene-amide moiety is fixed; other atoms of fragment **2** are mobile.
- Growing II (large): evolution from fragment **1** (LE = 0.58) with the amide flag. Chlorothiophene is the fixed moiety of this fragment.
- Linking: connect fragment **1** (flag amide), a complementary spacer and fragment **3** (flags amine, amide, or sulphonamide). Both original rings (chlorothiophene and isopropyl-piperidine) are kept fixed at the first stage.

2.4.2. Heat Shock Protein 90 (Hsp90). The Heat Shock Protein 90 (Hsp90) is a molecular chaperone involved in the refolding of several client proteins related to cancer.[70,71] Hsp90 is overexpressed into some cancerous cells and makes them more resistant to treatments, so its inhibition should reduce their life expectancy. The popular Hsp90 target has been intensively studied within FBDD projects,[72−76] and numerous complexes have been deposited in the PDB. Among them, several structures contain two fragments (concomitant binders) in the same binding site, enabling a linking strategy. The X-ray structure 3HZ5 contains a larger ligand (**13**), which results from the linking of N-dimethyl-adenine (**11**) and a pyrazole-derivative (**12**) fragment, after removal of the ester group of the latter (because not involved in favorable site contacts). These two non-competitive fragment-hits are present in the structure 3HZ1 and are linkable because the reactive centers are not sterically hindered. This technique was employed by Barker et al.[73] to combine the fragments **11** and **12** into ligand **13**. Each individual fragment exhibits very low potency (IC_{50} in low mM range), whereas the linked inhibitor gained 3 orders of magnitude (IC_{50} = 1.5 μM).

A virtual linking simulation is performed for the same purpose. The binding site is taken from the 3HZ1 structure. The set of anchor atoms to keep fix at the first sampling stage comprises the purine and pyrazole-derivative rings.

2.4.3. Acetylcholine Binding Protein (AChBP). The Acetylcholine Binding Protein is released into the synaptic cleft in order to capture the acetylcholine neurotransmitter. The direct consequence is the modulation of the synaptic transmission. In practice, this protein serves as model to study related membrane proteins such as ligand-gated ion channels.[77] Using AChBP as the target, Edink et al. described a fragment optimization (growing) resulting in conformational changes in the binding site.[78] As in the Hsp90 study, the starting materials are binding values and structural data for both fragment and evolved ligands. The 2Y54 structure includes the fragment **21** that corresponds to the Murcko scaffold[79] of cocaine. The visual analysis of the complex shows that the small

dx.doi.org/10.1021/ci4000163 | J. Chem. Inf. Model. 2013, 53, 836−851

organic compound binds to the bottom of cavity, inside an aromatic cage perfectly suitable for binding hydrophobic cations such as choline or tropine. The comparison of different X-ray structures, including either fragment **21** (pK_i = 5.3) or a reference potent inhibitor (α-lobeline, pK_i = 8.6), suggested one optimization way, while highlighting that conformational changes must be mandatory to accommodate ligands evolved from fragment **21**. As expected, the designed molecules bind to AChBP and the X-ray structures of these complexes (2Y56/2Y57) showed one main modification in the binding site conformation: another Y91 rotamer giving access to a new subpocket, occupied by the added chemical groups. Compounds **22** and **23**, optimized from fragment **21**, have respective pK_i values of 7.5 and 7. In addition to a thermodynamic analysis of binding profiles, the authors demonstrated that fragment evolution can (obviously) impact the pocket conformation.

From an in silico point of view, a retrospective growing analysis based on these data appears very interesting and challenging at the same time because flexibility of the binding site must be enabled for relevant reasons (experimentally confirmed binding site rearrangements). Although there is only one large shift between the 2Y54 and 2Y56/2Y57 structures (Y91 rotamer), all side chains around the ligand are unlocked (Y53, Y91, W145, Y186, and Y193) in order to make more challenging and reliable simulations. In this growing case, based on the amine RECAP flag, no additional unlocked atoms are needed because the evolution process directly starts from the bridged ring of fragment **21**. The binding site is taken from the 2Y54 X-ray structure. The resulting database of grown analogs is huge; thus different filters based on molecular properties are used to create a more tractable set, as in the large FXa growing II protocol. Because the key role of the positive charge carried by the ammonium group is well acknowledged, analogs in which the starting nitrogen loses its basic character by coupling to the linkers are discarded before docking/sampling in order to save time. This filtering is straightforward based on the predicted protonation state in the major microspecies returned by the ChemAxon pK_a plugin.

2.4.4. Acetolactate Synthase (AlSynth). This study case has been chosen because the targeted reference ligand **32** shows two key water-mediated interactions with the site (PDB 1T9B[80]), all while allowing a straightforward retrosynthetic cut of the species (at the level of the urea group). There is no experimental knowledge of the binding status and position of the starting fragment **31** (4-methoxy-6- methyl-1,3,5-triazin-2-yl)-urea; its location seen in the final lead is therefore assumed.

The carbonyl-sulphonamide group is expected to have a pK_a value of about 4 (according to the ChemAxon pK_a prediction tool[81]), and the ligand is indeed considered under its ionized anionic form in docking benchmarks (Astex data set[82]). However, the position of the first key crystallographic water—perfectly symmetrically located at the convergence point of a bidentate hydrogen bond with both the N−H groups of the urea—suggests that within the active site the effect of the protein has modified the local pK_a value as to prefer the neutral form. It was known from our previous results[1] that S4MPLE was able to dock the entire ligand (at that time taken as the anionic species $C(=O)[N^-]SO_2$) together with its two crystallographic waters, leading to a predicted geometry within the top 10 best ranked ones in very close agreement with the PDB structure. However, the placement of the urea-binding water was no longer in a symmetric position between the urea nitrogens. For simplicity, the neutral form of the ligands was imposed in these simulations.

The first above-mentioned crystallographic water bridge involves atoms of the prepositioned root fragment and is expected to occupy that position regardless of the nature of the growing moiety. It could have been fixed and integrated to the site, but was nevertheless allowed to move freely, serving as an implicit control of sampling quality.

The second water mediates the interaction between the sulphonyl oxygen and the cation of K251. This is the water molecule that may eventually participate (if possible, as in the case of the chlorophenylsulphonyl moiety) in non-covalent "growing", completing the connected linker and granting an extra stabilization for the binding. Otherwise, it is allowed to adopt another favorably interacting position with the site and/or the ligand.

The goal here is to retrospectively find the chlorophenylsulphonyl moiety, out of a large collection of alternatives, and to show that its finding is significantly enhanced when accounting for explicit waters.

3. RESULTS AND DISCUSSION

All results relative to the different growing and linking protocols are reported and discussed below. These should answer the previously issued questions (see the end of section 1.4). The initial fixing of precursor fragment atoms is useful for several reasons. First, the goal consists in the optimization of an original hit already present in the binding pocket (experimental location, or obtained by docking). Therefore, only binding modes sharing the initial interactions are really interesting. Besides, in the context of linking, a suitable spacer must freeze original fragments into their relative positions when independently bound to the site. Finally, these constraints have the advantage to speed up the process compared to a full blind docking.

In prospective projects, it may be interesting to redock all selected compounds without any particular bias. If unbiased docking leads to a same optimal pose as the growing/linking strategy, this may be interpreted as an additional argument augmenting the confidence in the success of that particular compound. However, if unbiased docking reveals lower energy poses with original fragments in unexpected locations, this may signal a force field artifact (fake minimum of the energy landscape, not accessible under growing/linking constraints). Alternatively, it may simply reveal a better overall pose with fragments choosing only slightly suboptimal alternative binding pockets. However, practice shows that observation of known key interactions in a ligand-site complex is often a more reliable success indicator than favorable scoring function values—this is why interaction fingerprints were first introduced.[83,84] Because in the herein addressed retrospective studies the experimental poses of final ligands are known, no final unbiased docking was undertaken to illustrate the expected direct anchoring of the ligand its site. Having the native or native-like linkers or growth fragments highlighted as preferred choices out of the large set of decoy substructures counts here as success. The question whether these best options are good enough to yield a potent ligand is already answered.

3.1. Factor Xa. This target is widely studied here in three protocols (two growing and one linking). Different fragments are used as inputs, but the ultimate goal is always the same: generating the high LE reference **4** and/or close analogs and recognizing them as putative "top hits" among all the other generated growth/linking products. Departure fragments **1** and **2** have low but measurable K_i, therefore they should be discovered during a typical fragment screening campaign. Growing protocol II

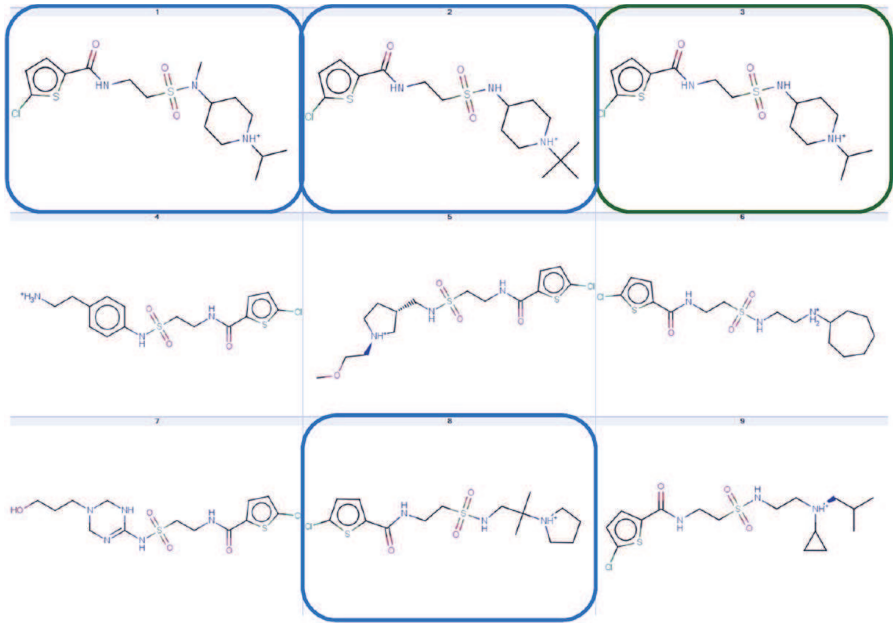


Figure 3. Top nine molecules from the FXa growing I virtual screening. Reference compound and closest analogs are respectively highlighted in green and blue.

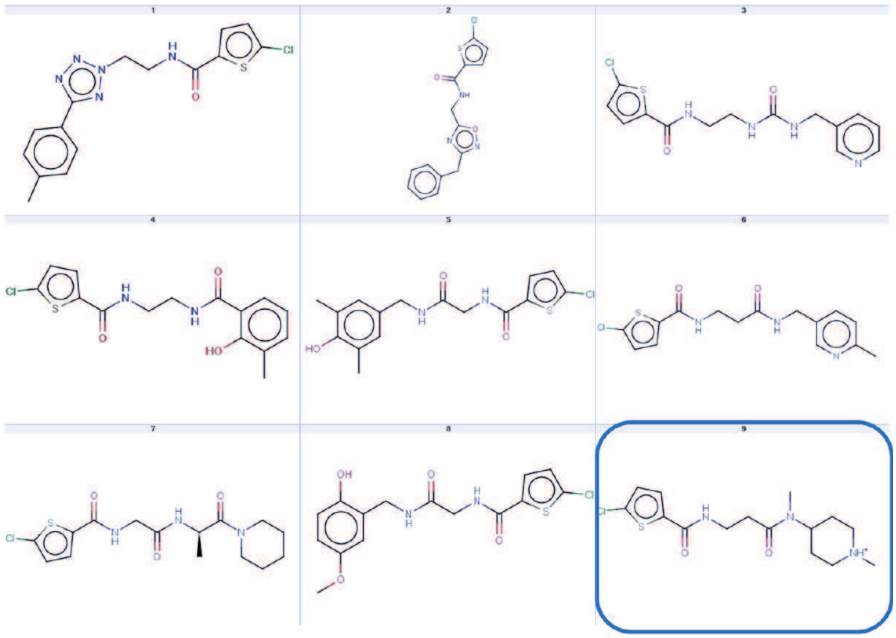


Figure 4. Top nine molecules from the FXa growing II virtual screening. Closest analog is highlighted in blue.

mimics a real situation, where only highest LE fragments are selected for the subsequent optimization stages. While the protocol II has stronger arguments, it is really more challenging because a larger moiety must be designed via the growing strategy. Conversely, strategy I is merely required to add an isopropyl-piperidine group to the original fragment. The virtual linking approach, realized here for the POC, would probably not have been undertaken in real conditions because fragment **3** has a very low affinity/efficiency (LE = 0.27). Unfortunately, the binding modes of fragments **1**, **2**, and **3** alone were not experimentally solved, contrary to that of reference product **4**. This information has been confirmed by the first author of the study of interest.[69] However, various known inhibitors contain the chlorothiophene substructure (**1**), and the latter always binds the S1 pocket as for ligand **4**. Besides, docking (with S4MPLE) of fragments **1** and **3** always led to the supposed binding mode each time with a large energy difference with respect to the second pose. Thus, the assumption that individual fragments have the same binding mode as in the grown/linked potent inhibitor appears as a reasonable hypothesis for these retrospective in silico studies, although several examples from FBDD projects highlighted the opposite case.[73]

3.1.1. Growing Protocol I (Small). This growing protocol, simpler than the others, is the first step of the POC workflow:

dx.doi.org/10.1021/ci4000163 | *J. Chem. Inf. Model.* 2013, 53, 836−851

is the strategy able to prioritize the chemical group seen in molecule **4** matching the S4 pocket of FXa? The fact of unlocking some significant DoF in the precursor fragment **2** (the "sulphonamide spacer") opens the theoretical possibility to reach various alternative pockets with the grown chain (a correctly prepositioned rigid sulphonamide would have obviously oriented the substituents toward S4 making this challenge too easy). Over the 4000 grown ligands, the top nine compounds, selected after pruning those with unacceptable RMSD shifts and distorted amide conformations, contain the known inhibitor (compound **4** ranked as #3) and several very close analogs (ranked as #1, #2, and #8). These top nine molecules are depicted in the Figure 3. Although only moderate-sized chemical groups were sought here (cyclic aliphatic ammonium), the best binding mode perfectly matches the experimental one (RMSD < 0.5 Å). On the basis of these encouraging results, this first growing protocol is validated.

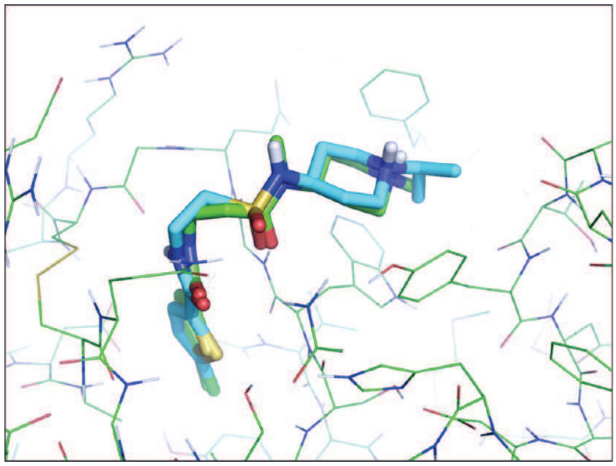


Figure 5. Superimposition of compound #9 (methyl-amide) to the high LE inhibitor **4** (sulphonamide) from the FXa growing II virtual screening.

3.1.2. Growing Protocol II (Large). In this case study, the preliminary in silico growing stage is expected to connect the ethyl-sulphonamide-piperidine-isopropyl group (15 heavy atoms) to the starting fragment **1** to exactly reproduce the reference **4**. Unfortunately, this inhibitor is not automatically generated. An analysis of the large growing database revealed that the desired group ($_{\text{amide-flag}}$-ethyl-sulphonamide-piperidine-isopropyl) was simply not among the options retained when fragmenting the ZINC-clean collection. This is a clear limit of the knowledge-based approach used here: although containing more than 9 million compounds, the ZINC database only represents a fraction of the huge drug-like chemical space. Despite the absence of desired ligand (**4**), the virtual screening was undertaken nevertheless because potentially interesting compounds including an ethyl-sulphonamide spacer or a terminal hydrophobic cation were generated. Because contributions of non-specific site-ligand interactions to the binding energy scale up with putative ligand size, this growing was confined to fragments within a size window of 12 to 15 heavy atoms. Within these specifications, about 8000 molecules were finally screened, and several putative interesting compounds emerged among the top 100 hits. As before, the top nine compounds are reported (Figure 4). In particular, the ninth hit appears pretty close to the active inhibitor **4** with respect to two criteria:

- Chemical structure: Amide linkage substitutes the sulphonamide spacer and a terminal N-methyl replaces the N-isopropyl one.
- Binding mode: Superimposition of experimental binding mode of the reference and of predicted hit 9 is depicted in Figure 5.

Figure 6 shows some more hits within the top 100, all of which may be considered rather similar to **4** because they all include either an ethyl-sulphonamide spacer followed by a terminal hydrophobic group or a terminal charged hydrophobic structure like the isopropyl-piperidine.

Finally, from an energy point of view, the top ranked ligand exhibits a higher FF energy difference (less favorable) than the

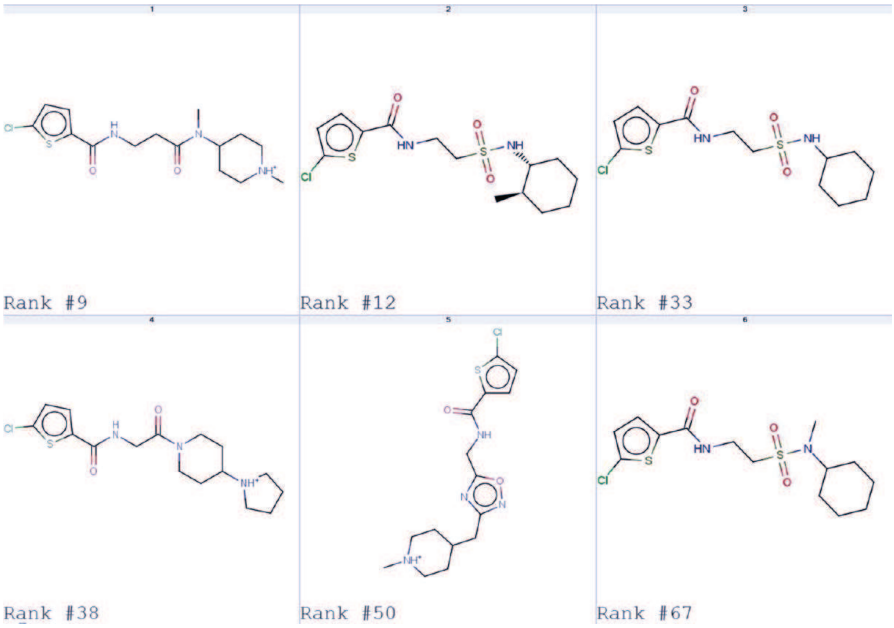


Figure 6. Most similar compounds to reference **4** within the top 100 hits from the FXa growing II virtual screening.

dx.doi.org/10.1021/ci4000163 | *J. Chem. Inf. Model.* 2013, 53, 836−851

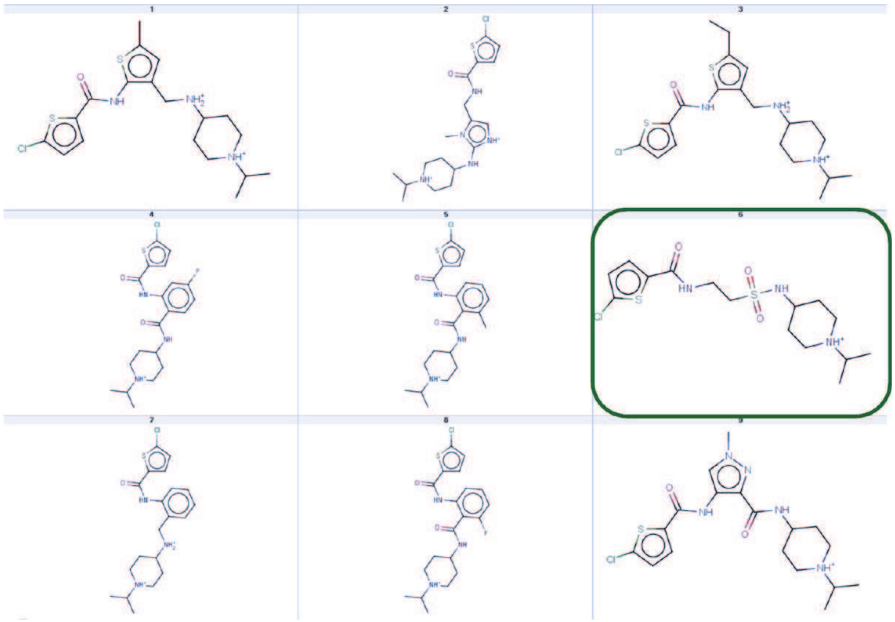


Figure 7. Top nine molecules from the FXa linking virtual screening. Reference compound is highlighted in green.

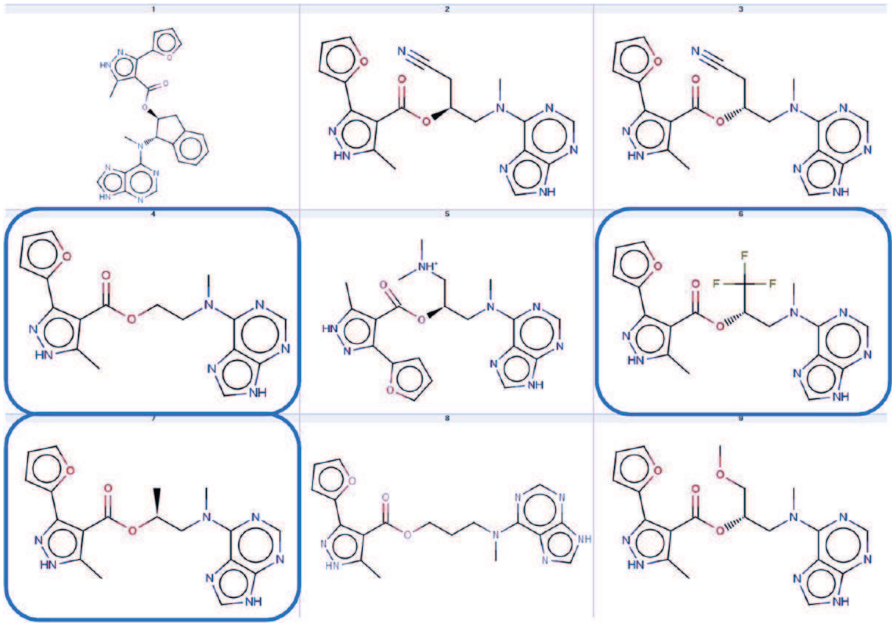


Figure 8. Top nine molecules from the Hsp90 linking virtual screening. Closest analogs are highlighted in blue.

reference compound **4** simulated in the same conditions. If the needed moiety would have been present at growth stage, **4** would have been top ranked with respect to all herein retained candidates. Although not perfect, the results of this larger virtual screening highlight valuable perspectives for prospective growing studies.

3.1.3. Linking Strategy. This first linking approach, consisting in the enumeration and assessment of about 650 compounds, generated and retrieved the expected ligand in the top nine ligands (reference compound **4** ranked as #6). No specific filters were employed here to reduce the set of compounds to screen, except for the preservation of a canonical amide C(=O)[NH] (with one polar H, at least). The sulphonamide spacer occupies its expected location. However, comments about the accuracy of retrieved binding mode are less relevant in this situation because both ends are constrained in the linker sampling stage (before the relaxation of the full ligand). Compared to the growing protocol I, only few analogs are retrieved here—expectedly, because only the chemical space of the linker moiety is subject to exploration. One close analog, with an additional carbon in the linker, is ranked #90. Visually, this new carbon displaces the sulfonamide group that loses its favorable interactions (hydrogen bonds) with the site, hence a logically worse ranking. Figure 7 displays the top nine results of this linking simulation. Surprisingly, there are several molecules including two ammonium ions according to the ChemAxon major microspecies plug-in. This may be correct

dx.doi.org/10.1021/ci4000163 | *J. Chem. Inf. Model.* 2013, 53, 836−851

for the unbound ligands but may change due to pK_a shifts upon binding. However, such subtle effects are beyond the reach of FF-based modeling. Also, the relative gain in binding energy upon presence of a second cation may be an artifact due to ignoring counterion effects. All in all, chemical common sense recommends caution in accepting these species as putative actives.

To conclude, the linking procedure automatically generated and top ranked the desired inhibitor **4** from fragments **1** and **3**, thus representing a success.

3.2. Heat Shock Protein 90. Conversely to the FXa studies, where hypothesis are made about the fragments binding modes, X-ray data here give access to simultaneous binding modes of non-competitive fragments (3HZ1) in addition to the evolved compound (3HZ5). However, the final compound does not fully include the starting fragments: the masked acidic group present in the pyrazole-derivative **12** disappears from the optimized ligand. Authors highlighted the fact that the ester group does not make tight interactions with the site. Besides, this chemical function is known to be frequently cleaved in vivo. Therefore, this group was removed in the linking process and was replaced by a simple alkyl spacer of length suggested by molecular modeling. The direct consequence is that our procedure can only generate more or less close analogs including the ester function but not the exact known reference **13**. Given that only few data (including X-ray support for individual fragments and evolved compound) relative to linking optimization are available, this retrospective linking study has been undertaken anyway.

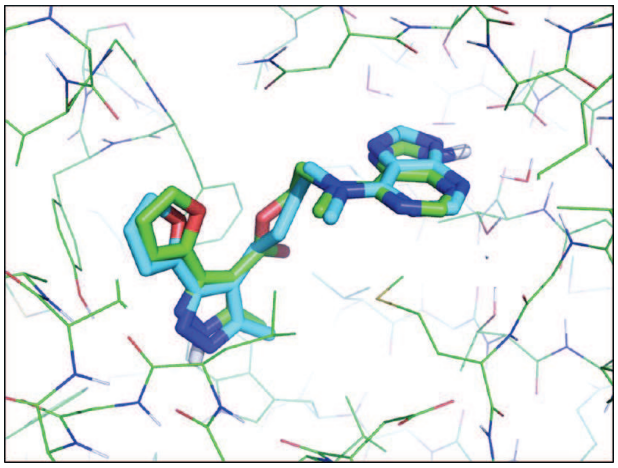


Figure 9. Superimposition of the reference inhibitor **13** (green) to its closest screened analog (ranked as #4 and displayed in blue).

Ester-containing analogs are retrieved at ranks #4, #6, and #7 (Figure 8). Ligand #4 looks like the closest analog of the reference because its spacer is very simple (not ramified) and does not contain any stereocenter, conversely to linkers of #6 and #7. Empirically, a good superimposition between the active (**13**) and these suggested molecules can be observed, preserving the hydrogen bonds network between the purine ring and the binding site (Figure 9). Interestingly, the protocol favored molecules (eight among the top nine) with a topological distance of 6 between their heterocycles (pyrazole and purine), as in the known inhibitor **13**.

Despite the inability to generate the reference inhibitor **13**, which does not exactly incorporate original fragment hits, this in silico linking procedure generated compounds with rather similar linkers (same length as in the active ligand, but including an ester function) within the top hit list.

3.3. Acetylcholine Binding Protein. A strong feature in S4MPLE, exemplified by this case, is that growth/linking may be performed all while considering binding site flexibility (during the sampling and/or the final refinement), whereas many state-of-the-art tools still operate today on rigid sites only. Most issues mentioned in the introduction are addressed by this simulation: design of known inhibitors, sampling of generated molecules (e.g., picking up the expected interactions between the site and known ligands), ranking of references in the top hit list, and prediction of binding site conformational changes. The optimization of fragment **21** into actives **22** or **23** consists in adding a pretty simple group (e.g., phenyl with two spacer

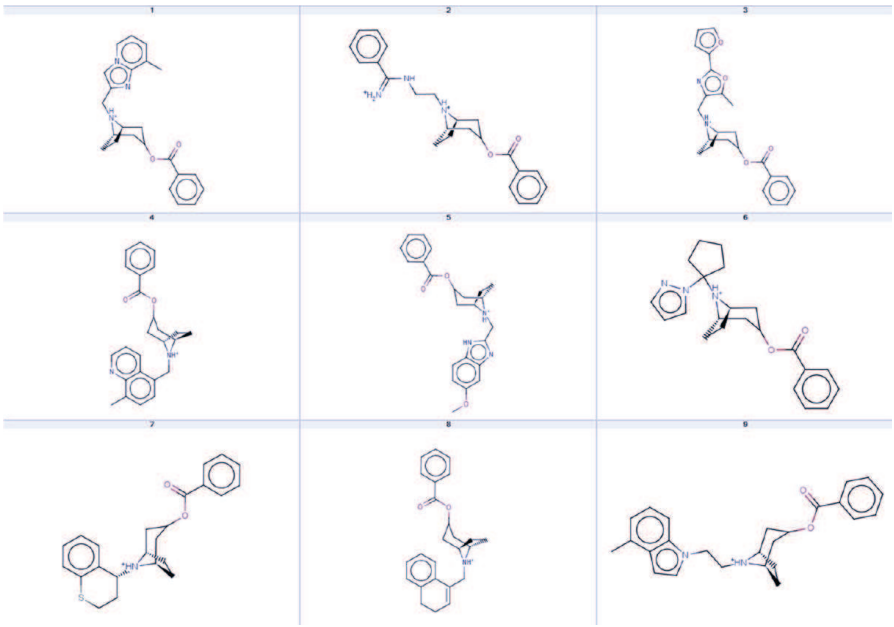


Figure 10. Top nine molecules from the AChBP growing virtual screening.

dx.doi.org/10.1021/ci4000163 | J. Chem. Inf. Model. 2013, 53, 836−851

carbons for molecule **22**). As expected, these molecules are automatically generated from the fragment **21** using the amine RECAP flag. The top nine retrieved molecules are displayed in the Figure 10. By contrast to previous successful cases, molecules **22** and **23** are not ranked in the top hit list (e.g., top 1%) in this more challenging study, involving about 4500 screened entities. However, ligand **23** is ranked among the top 4%, whereas ligand **22** appears in top 10%, and one close analog of **22** (tolyl group substitutes the phenyl ring) is ranked in the top 3%. Figure 11

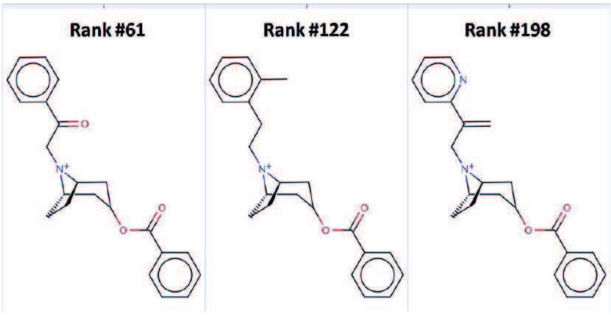


Figure 11. Most similar compounds to references **22** or **23** within the top 5% hits from the AChBP growing virtual screening.

shows some hits within the top 5%, all rather similar to references **22** or **23**. Finally, the expected new rotamer of the tyrosine (Y91) is successfully predicted for sampled reference molecules. The additional hydrogen-bond between compound **23** and binding site backbone is reproduced, while the other four flexible side chains correctly fall back into their original conformations. Figure 12 illustrates starting materials (experimental binding modes of fragment **21** and ligand **23** within their respective binding sites) and sampled geometries of reference **23** into the partly flexible 2Y54 binding site.

Although the ranking appears a bit less efficient in that specific case (pessimistically assuming that all compounds ranked above the expected references are false positives), most previously introduced issues are favorably addressed by this in silico growing protocol with partial binding site flexibility.

3.4. Acetolactate Synthase. If this simulation is run as a classical water-free growing process, like in examples above, the targeted reference ligand **32** would have been ranked as the 62nd most promising moiety to connect to the departure triazinyl-urea fragment **31**, out of about 1550 candidates (i.e., within top 4%). With explicit waters, however, its ranking is boosted to position 6 (i.e., within top 0.4%). Therefore, inclusion of explicit-free waters appears as an important factor in order to highlight the suitability of the chlorophenylsulphonyl moiety as a correct choice for growing. Furthermore, the waters are both found at expected locations, all while the geometry of the bound ligand is perfectly reproduced (Figure 13). The best nine ligands selected by the explicit-water simulation are shown in Figure 14. Intriguingly, the overall best ranked compound, having N-bound phenylalanine as selected growing moiety, is common to both explicit-water and water-free runs. The negative carboxylate therein forms a strong salt bridge with the lysine (albeit rather solvent exposed), and waters are both located in the neighborhood of the carboxylate. It is an illustration of a possible compound in which the water-mediated interaction would be replaced by a direct contact, showing that if such alternatives were present among the possibilities opened by the growing moiety database they could be found as well.

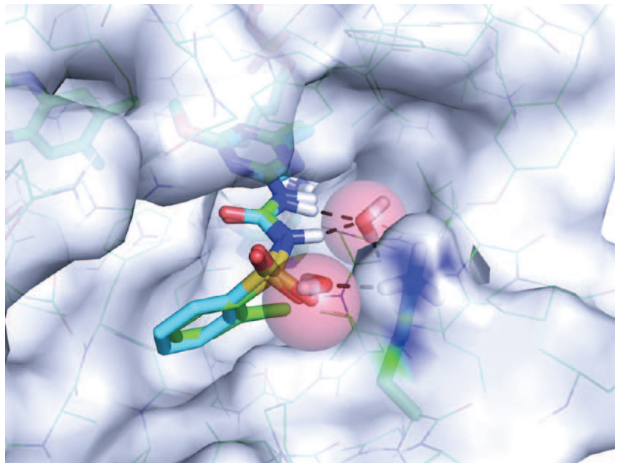


Figure 13. Superimposition of experimental and predicted geometries for the reference compound **32** from the AlSynth growing virtual screening. Predicted waters locations are shown as sticks, and experimental ones are displayed as spherical.

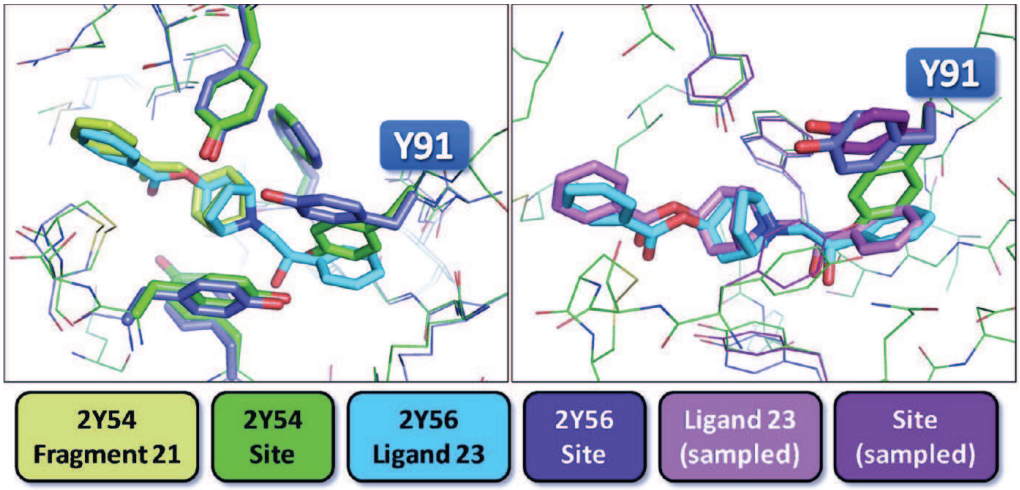


Figure 12. Superimposition of experimental structure (2Y54 vs 2Y56, left) and resulting sampled configurations (right). Albeit the initial site geometry was taken from 2Y54, the sampled geometry matches closely the expected 2Y56. Conformational changes of Y91 are highlighted.

dx.doi.org/10.1021/ci4000163 | *J. Chem. Inf. Model.* 2013, 53, 836−851

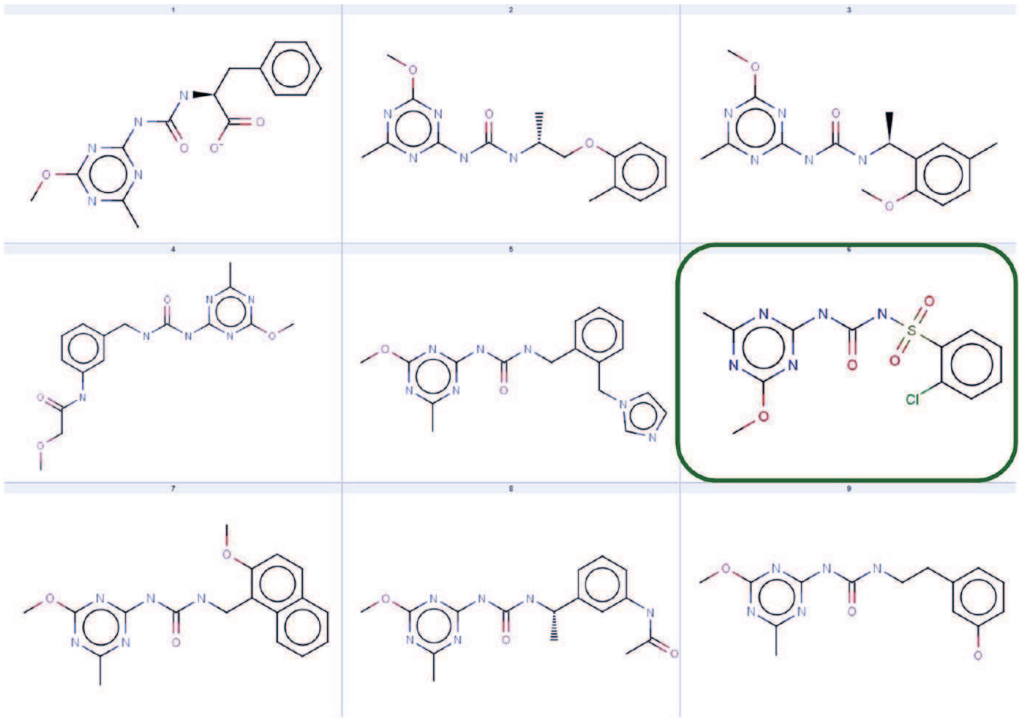


Figure 14. Top nine molecules from the AlSynth growing virtual screening. Reference compound is highlighted in green.

Unfortunately, we have no knowledge of the activity level of this interesting derivative. Finally, the top 50 selected ligands in both runs have nevertheless significant overlap: common species are molecules with hydrophobic groups (mainly interacting with V191, P192, A195, A200, F201, and K251), obviously well ranked irrespective of the use of explicit waters.

4. CONCLUSION AND PERSPECTIVES

On the basis of several retrospective studies, this in silico fragment-to-lead protocol involving JMolEvolve and S4MPLE is validated by its ability to automatically generate expected ligands (and/or close analogs) from starting fragments, to rank them with some effectiveness among a set of decoy compounds, to reproduce their experimental binding mode, and to predict slight site conformational changes when binding site flexibility enabled. Furthermore, S4MPLE—initially designed in view of conformational sampling and docking—is not only well able to take charge of problems typically addressed by specific FBDD/DND computer programs but allows going beyond classical applications. It may easily include site flexibility, but foremost, it is able to consider non-covalent alternatives to classical growing (i.e., including non-covalent partners such as water molecules to mediate favorable contacts with the site). This allows priority retrieval of growing moieties that would not be top-ranked under default conditions because their complementarity to the site is less apparent if bridging water molecules are not present. All this is achieved without any prerequisite knowledge of the expected water positions and without any specific alterations of the protocol: the only user-defined parameter is the number of explicit water molecules to consider (where adding too many should not impact on the found binding modes, albeit it would impact on the convergence rate).

This optimization strategy is currently being applied in prospective FBDD projects (work in progress at NovAliX, Strasbourg). Even the herein used minimalistic but robust linking/growing strategy may certainly produce novel compounds if provided with original starting fragments and linkers. Assessment of synthetic accessibility was not the subject of this structure-based study, although the RECAP-based procedure did not generate odd structures. Nevertheless, future work will focus on that important point in order to increase the usefulness of our in silico FBDD protocol.

S4MPLE-driven docking will at least ensure that retained linking/growing options match well the shape of the active site and are rather void of intra-molecular strain. Will they be active? This is the key question. Depending on the accuracy of the herein used AMBER/GAFF force field, of the additional desolvation and contact terms, and on entropic aspects no one really knows how to accurately simulate, etc. This paper evidenced the ability of the generic tool S4MPLE to address atypical problems in the sense that "docking" of linkers is driven by both covalent and non-covalent constraints, all while supporting side chain flexibility. The versatility of S4MPLE allows it to easily adapt to the "philosophy" of FBDD: specifically search for docked solution in which reference fragments remain in the positions they would spontaneously adopt when unconnected. Most important is the ability of S4MPLE to prioritize groups that benefit from water-mediated interactions with the site. This formally exemplifies (as far as we can tell) an original growth strategy of a (ligand + mediating water) non-covalent complex by contrast to classical growth through covalent bonds only. In most cases, S4MPLE managed to highlight (i.e., to prioritize in the energy-ranked list) known binders out of many decoys and to discover native-like poses (including the last example with freely roaming water molecules). The good news—from an algorithm development point of view—is that the tool did not get lost in phase space and did not get blocked in local high-energy minima (no failure to "drag" growing/linking moieties of initially random coordinates into a constrained active site) but actually found the relevant minima. That, furthermore, the minima fortunately coincided with experimental input is

dx.doi.org/10.1021/ci4000163 | J. Chem. Inf. Model. 2013, 53, 836−851

mainly the merit of AMBER/GAFF, partially enhanced by specific contributions.

Because of the full control over considered DoF, S4MPLE is a versatile tool covering from massively parallel computing of large ligands and small proteins (work in progress) to the medium throughput construction of reasonable growth/linker entities (at less than a CPU hour per ligand under a preliminary protocol employing a fixed number of evolutionary generations chosen such as to suffice for the most flexible among the encountered candidate compounds). Optimizing the termination conditions (rendering them dependent on the effective number of DoF and/or replacing the total generation number criterion by a maximal number of "stagnant" generations having produced no better offspring) must still be undertaken.

■ ASSOCIATED CONTENT

Supporting Information

S4MPLE (x86_64) version can be uploaded from http://infochim.u-strasbg.fr (see Downloads). Both compound files used in this work and the ChemAxon-driven growth/linking tools are available upon request (ChemAxon licenses needed by the end users). This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: dhorvath@unistra.fr.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank the staff of the two computer centers that hosted the simulations: HPC (High-performance computing) of the University of Strasbourg and HPC of the chemistry faculty of Cluj-Napoca. All images of ligand−protein structures were created using Pymol.[85]

■ REFERENCES

(1) Hoffer, L.; Horvath, D. S4MPLE - Sampler for multiple protein−ligand entities: Simultaneous docking of several entities. *J. Chem. Inf. Model.* **2012**, *53* (1), 88−102.

(2) Congreve, M.; Chessari, G.; Tisi, D.; Woodhead, A. J. Recent developments in fragment-based drug discovery. *J. Med. Chem.* **2008**, *51* (13), 3661−3680.

(3) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A rule of three for fragment-based lead discovery? *Drug Discovery Today* **2003**, *8* (19), 876−877.

(4) Hoffer, L.; Renaud, J. P.; Horvath, D. Fragment-based drug design: Computational and experimental state of the art. *Comb. Chem. High Throughput Screening* **2011**, *14* (6), 500−520.

(5) Neumann, T.; Junker, H.; Schmidt, K.; Sekul, R. SPR-based fragment screening: advantages and applications. *Curr. Top. Med. Chem.* **2007**, *7* (16), 1630−1642.

(6) Perspicace, S.; Banner, D.; Benz, J.; Müller, F.; Schlatter, D.; Huber, W. Fragment-based screening using surface plasmon resonance technology. *J. Biomol. Screeing* **2009**, *14* (4), 337−349.

(7) Vivat Hannah, V.; Atmanene, C.; Zeyer, D.; Van Dorsselaer, A.; Sanglier-Cianférani, S. Native MS: An 'ESI' way to support structure- and fragment-based drug discovery. *Future Med Chem* **2010**, *2* (1), 35−50.

(8) Orita, M.; Warizaya, M.; Amano, Y.; Ohno, K.; Niimi, T. Advances in fragment-based drug discovery platforms. *Exp. Opin. Drug Discovery* **2009**, *4* (11), 1125−1144.

(9) Murray, C. W.; Blundell, T. L. Structural biology in fragment-based drug design. *Curr. Opin. Struct. Biol.* **2010**, *20* (4), 497−507.

(10) Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today* **2004**, *9* (10), 430−431.

(11) Hann, M.; Leach, A.; Harper, G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (3), 856−864.

(12) Murray, C. W.; Rees, D. C. The rise of fragment-based drug discovery. *Nature Chem.* **2009**, *1* (3), 187−192.

(13) Erlanson, D. A. Fragment-based lead discovery: A chemical update. *Curr. Opin. Biotechnol.* **2006**, *17* (6), 643−652.

(14) Borsi, V.; Calderone, V.; Fragai, M.; Luchinat, C.; Sarti, N. Entropic contribution to the linking coefficient in fragment based drug design: A case study. *J. Med. Chem.* **2010**, *53* (10), 4285−4289.

(15) Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. Discovering high-affinity ligands for proteins: SAR by NMR. *Science* **1996**, *274* (5292), 1531−1534.

(16) Hajduk, P. J.; Sheppard, G.; Nettesheim, D. G.; Olejniczak, E. T.; Shuker, S. B.; Meadows, R. P.; Steinman, D. H.; Carrera, G. M.; Marcotte, P. A.; Severin, J.; Walter, K.; Smith, H.; Gubbins, E.; Simmer, R.; Holzman, T. F.; Morgan, D. W.; Davidsen, S. K.; Summers, J. B.; Fesik, S. W. Discovery of potent nonpeptide inhibitors of stromelysin using SAR by NMR. *J. Am. Chem. Soc.* **1997**, *119* (25), 5818−5827.

(17) Szczepankiewicz, B.; Liu, G.; Hajduk, P.; Abad-Zapatero, C.; Pei, Z.; Xin, Z.; Lubben, T.; Trevillyan, J.; Stashko, M.; Ballaron, S.; Liang, H.; Huang, F.; Hutchins, C.; Fesik, S.; Jirousek, M. Discovery of a potent, selective protein tyrosine phosphatase 1B inhibitor using a linked-fragment strategy. *J. Am. Chem. Soc.* **2003**, *125* (14), 4087−4096.

(18) Law, R.; Barker, O.; Barker, J. J.; Hesterkamp, T.; Godemann, R.; Andersen, O.; Fryatt, T.; Courtney, S.; Hallett, D.; Whittaker, M. The multiple roles of computational chemistry in fragment-based drug design. *J. Comput.-Aided Mol. Des.* **2009**, *23* (8), 459−473.

(19) Hubbard, R. E.; Chen, I.; Davis, B. Informatics and modeling challenges in fragment-based drug discovery. *Curr. Opin. Drug Discovery Dev.* **2007**, *10* (3), 289−297.

(20) Hann, M.; Green, R. Chemoinformatics - A new name for an old problem? *Curr. Opin. Chem. Biol.* **1999**, *3* (4), 379−383.

(21) Schuffenhauer, A.; Ruedisser, S.; Marzinzik, A.; Jahnke, W.; Blommers, M.; Selzer, P.; Jacoby, E. Library design for fragment based screening. *Curr. Top. Med. Chem.* **2005**, *5* (8), 751−762.

(22) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432* (7019), 862−865.

(23) Verdonk, M. L.; Giangreco, I.; Hall, R. J.; Korb, O.; Mortenson, P. N.; Murray, C. W. Docking performance of fragments and druglike compounds. *J. Med. Chem.* **2011**, *54* (15), 5422−5431.

(24) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47* (7), 1739−1749.

(25) Loving, K.; Salam, N. K.; Sherman, W. Energetic analysis of fragment docking and application to structure-based pharmacophore hypothesis generation. *J. Comput.-Aided Mol. Des.* **2009**, *23* (8), 541−554.

(26) Kawatkar, S.; Wang, H. M.; Czerminski, R.; Joseph-McCarthy, D. Virtual fragment screening: an exploration of various docking and scoring protocols for fragments using Glide. *J. Comput.-Aided Mol. Des.* **2009**, *23* (8), 527−539.

(27) Huang, Q.; Li, L. L.; Yang, S. Y. PhDD: A new pharmacophore-based de novo design method of drug-like molecules combined with assessment of synthetic accessibility. *J. Mol. Graph. Model.* **2010**, *28* (8), 775−787.

(28) Lippert, T.; Schulz-Gasch, T.; Roche, O.; Guba, W.; Rarey, M. De novo design by pharmacophore-based searches in fragment spaces. *J. Comput-Aided Mol. Des.* **2011**, *25* (10), 931−945.

(29) Besnard, J.; Ruda, G. F.; Setola, V.; Abecassis, K.; Rodriguiz, R. M.; Huang, X. P.; Norval, S.; Sassano, M. F.; Shin, A. I.; Webster, L. A.; Simeons, F. R.; Stojanovski, L.; Prat, A.; Seidah, N. G.; Constam, D. B.; Bickerton, G. R.; Read, K. D.; Wetsel, W. C.; Gilbert, I. H.; Roth, B. L.; Hopkins, A. L. Automated design of ligands to polypharmacological profiles. *Nature* **2012**, *492* (7428), 215−220.

(30) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP - Retrosynthetic combinatorial analysis procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (3), 511−522.
(31) Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem* **2008**, *3* (10), 1503−1507.
(32) Böhm, H. The computer program LUDI: A new method for the de novo design of enzyme inhibitors. *J. Comput.-Aided Mol. Des.* **1992**, *6* (1), 61−78.
(33) Nishibata, Y.; Itai, A. Automatic creation of drug candidate structures based on receptor structure - Starting point for artificial lead generation. *Tetrahedron* **1991**, *47* (43), 8985−8990.
(34) Mata, P.; Gillet, V.; Johnson, A.; Lampreia, J.; Myatt, G.; Sike, S.; Stebbings, A. SPROUT - 3D structure generation using templates. *J. Chem. Inf. Comput. Sci.* **1995**, 479−493.
(35) Stahl, M.; Todorov, N. P.; James, T.; Mauser, H.; Boehm, H. J.; Dean, P. M. A validation study on the practical use of automated de novo design. *J. Comput.-Aided Mol. Des.* **2002**, *16* (7), 459−478.
(36) Rotstein, S.; Murcko, M. Groupbuild - A fragment-based method for denovo drug design. *J. Med. Chem.* **1993**, *36* (12), 1700−1710.
(37) Wang, R.; Gao, Y.; Lai, L. LigBuilder: A multi-purpose program for structure-based drug design. *J. Mol. Model.* **2000**, 498−516.
(38) Schneider, G.; Lee, M.; Stahl, M.; Schneider, P. De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. *J. Comput.-Aided Mol. Des.* **2000**, *14* (5), 487−494.
(39) Fechner, U.; Schneider, G. Flux (1): A virtual synthesis scheme for fragment-based de novo design. *J. Chem. Inf. Model.* **2006**, *46* (2), 699−707.
(40) Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discovery* **2005**, *4* (8), 649−663.
(41) Vinkers, H.; de Jonge, M.; Daeyaert, F.; Heeres, J.; Koymans, L.; van Lenthe, J.; Lewi, P.; Timmerman, H.; Van Aken, K.; Janssen, P. SYNOPSIS: SYNthesize and OPtimize system in silico. *J. Med. Chem.* **2003**, *46* (13), 2765−2773.
(42) Hartenfeller, M.; Zettl, H.; Walter, M.; Rupp, M.; Reisen, F.; Proschak, E.; Weggen, S.; Stark, H.; Schneider, G. DOGS: Reaction-driven de novo design of bioactive compounds. *PLoS Comput Biol* **2012**, *8* (2), e1002380.
(43) Gillet, V.; Myatt, G.; Zsoldos, Z.; Johnson, A. SPROUT, HIPPO and CAESA: Tools for de novo structure generation and estimation of synthetic accessibility. *Perspect. Drug Discovery Des.* **1995**, *3*, 34−50.
(44) Boda, K.; Seidel, T.; Gasteiger, J. Structure and reaction based evaluation of synthetic accessibility. *J. Comput.-Aided Mol. Des.* **2007**, *21* (6), 311−325.
(45) Zaliani, A.; Boda, K.; Seidel, T.; Herwig, A.; Schwab, C. H.; Gasteiger, J.; Claussen, H.; Lemmen, C.; Degen, J.; Parn, J.; Rarey, M. Second-generation de novo design: A view from a medicinal chemist perspective. *J. Comput.-Aided Mol. Des.* **2009**, *23* (8), 593−602.
(46) Hartenfeller, M.; Eberle, M.; Meier, P.; Nieto-Oberhuber, C.; Altmann, K.; Schneider, G.; Jacoby, E.; Renner, S. A collection of robust organic synthesis reactions for in silico molecule design. *J. Chem. Inf. Model.* **2011**, *51* (12), 3093−3098.
(47) Miranker, A.; Karplus, M. Functionality maps of binding sites: A multiple copy simultaneous search method. *Proteins* **1991**, *11* (1), 29−34.
(48) Schubert, C. R.; Stultz, C. M. The multi-copy simultaneous search methodology: A fundamental tool for structure-based drug design. *J. Comput.-Aided Mol. Des.* **2009**, *23* (8), 475−489.
(49) Brenke, R.; Kozakov, D.; Chuang, G. Y.; Beglov, D.; Hall, D.; Landon, M. R.; Mattos, C.; Vajda, S. Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics* **2009**, *25* (5), 621−627.
(50) Hall, D. R.; Ngan, C. H.; Zerbe, B. S.; Kozakov, D.; Vajda, S. Hot spot analysis for driving the development of hits into leads in fragment-based drug discovery. *J. Chem. Inf. Model.* **2012**, *52* (1), 199−209.
(51) Ngan, C. H.; Bohnuud, T.; Mottarella, S. E.; Beglov, D.; Villar, E. A.; Hall, D. R.; Kozakov, D.; Vajda, S. FTMAP: extended protein mapping with user-selected probe molecules. *Nucleic Acids Res.* **2012**, *40* (Web Server issue), W271−W275.
(52) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857.
(53) Maass, P.; Schulz-Gasch, T.; Stahl, M.; Rarey, M. Recore: A fast and versatile method for scaffold hopping based on small molecule crystal structure conformations. *J. Chem. Inf. Model.* **2007**, *47* (2), 390−399.
(54) Thompson, D.; Denny, R.; Nilakantan, R.; Humblet, C.; Joseph-McCarthy, D.; Feyfant, E. CONFIRM: Connecting fragments found in receptor molecules. *J. Comput.-Aided Mol. Des.* **2008**, *22* (10), 761−772.
(55) Dey, F.; Caflisch, A. Fragment-based de novo ligand design by multiobjective evolutionary optimization. *J. Chem. Inf. Model.* **2008**, *48* (3), 679−690.
(56) Majeux, N.; Scarsi, M.; Apostolakis, J.; Ehrhardt, C.; Caflisch, A. Exhaustive docking of molecular fragments with electrostatic solvation. *Proteins* **1999**, *37* (1), 88−105.
(57) Gozalbes, R.; Carbajo, R. J.; Pineda-Lucena, A. Contributions of computational chemistry and biophysical techniques to fragment-based drug discovery. *Curr. Med. Chem.* **2010**, *17* (17), 1769−1794.
(58) ChemAxon. http://www.chemaxon.com (accessed April 4, 2013).
(59) Hoffer, L.; Chira, C.; Marcou, G.; Varnek, A.; Horvath, D. 2013, publication in preparation.
(60) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; Debolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. AMBER a package of computer-programs for applying molecular mechanics, normal-mode analysis, molecular-dynamics and free-energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* **1995**, *91* (1−3), 1−41.
(61) Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25* (9), 1157−1174.
(62) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Model.* **1998**, *38*, 983−996.
(63) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3−26.
(64) Irwin, J. J.; Shoichet, B. K. ZINC - A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45* (1), 177−182.
(65) Gasteiger, J.; Marsilli, M. A new model for calculating atomic charges in molecules. *Tetrahedron Lett.* **1978**, 3181−3184.
(66) Wang, J. M.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graphics Modell.* **2006**, *25* (2), 247−260.
(67) Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Swails, J.; Goetz, A. W.; Kolossvry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wolf, R. M.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Salomon-Ferrer, R.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. *AMBER 12*; University of California, 2012.
(68) RCSB Protein Data Bank. http://www.rcsb.org/pdb/home/home.do (accessed April 4, 2013).
(69) Nazaré, M.; Matter, H.; Will, D. W.; Wagner, M.; Urmann, M.; Czech, J.; Schreuder, H.; Bauer, A.; Ritter, K.; Wehner, V. Fragment deconstruction of small, potent factor Xa inhibitors: exploring the superadditivity energetics of fragment linking in protein−ligand complexes. *Angew. Chem., Int. Ed. Engl.* **2012**, *51* (4), 905−911.
(70) Janin, Y. L. Heat shock protein 90 inhibitors. A text book example of medicinal chemistry? *J. Med. Chem.* **2005**, *48* (24), 7503−7512.
(71) Janin, Y. L. ATPase inhibitors of heat-shock protein 90, second season. *Drug Discovery Today* **2010**, *15* (9−10), 342−353.

(72) Barker, J. J.; Barker, O.; Boggio, R.; Chauhan, V.; Cheng, R. K. Y.; Corden, V.; Courtney, S. M.; Edwards, N.; Falque, V. M.; Fusar, F.; Gardiner, M.; Hamelin, E. M. N.; Hesterkamp, T.; Ichihara, O.; Jones, R. S.; Mather, O.; Mercurio, C.; Minucci, S.; Montalbetti, C.; Muller, A.; Patel, D.; Phillips, B. G.; Varasi, M.; Whittaker, M.; Winkler, D.; Yarnold, C. J. Fragment-based Identification of Hsp90 Inhibitors. *ChemMedChem* **2009**, *4* (6), 963−966.
(73) Barker, J. J.; Barker, O.; Courtney, S. M.; Gardiner, M.; Hesterkamp, T.; Ichihara, O.; Mather, O.; Montalbetti, C. A.; Müller, A.; Varasi, M.; Whittaker, M.; Yarnold, C. J. Discovery of a novel Hsp90 inhibitor by fragment linking. *ChemMedChem* **2010**, *5* (10), 1697−1700.
(74) Huth, J. R.; Park, C.; Petros, A. M.; Kunzer, A. R.; Wendt, M. D.; Wang, X. L.; Lynch, C. L.; Mack, J. C.; Swift, K. M.; Judge, R. A.; Chen, J.; Richardson, P. L.; Jin, S.; Tahir, S. K.; Matayoshi, E. D.; Dorwin, S. A.; Ladror, U. S.; Severin, J. M.; Walter, K. A.; Bartley, D. M.; Fesik, S. W.; Elmore, S. W.; Hajduk, P. J. Discovery and design of novel HSP90 inhibitors using multiple fragment-based design strategies. *Chem. Biol. Drug Des.* **2007**, *70*, 1−12.
(75) Murray, C. W.; Carr, M. G.; Callaghan, O.; Chessari, G.; Congreve, M.; Cowan, S.; Coyle, J. E.; Downham, R.; Figueroa, E.; Frederickson, M.; Graham, B.; McMenamin, R.; O'Brien, M. A.; Patel, S.; Phillips, T. R.; Williams, G.; Woodhead, A. J.; Woolford, A. J. Fragment-based drug discovery applied to Hsp90. Discovery of two lead series with high ligand efficiency. *J. Med. Chem.* **2010**, *53* (16), 5942−5955.
(76) Woodhead, A. J.; Angove, H.; Carr, M. G.; Chessari, G.; Congreve, M.; Coyle, J. E.; Cosme, J.; Graham, B.; Day, P. J.; Downham, R.; Fazal, L.; Feltell, R.; Figueroa, E.; Frederickson, M.; Lewis, J.; McMenamin, R.; Murray, C. W.; O'Brien, M. A.; Parra, L.; Patel, S.; Phillips, T.; Rees, D. C.; Rich, S.; Smith, D. M.; Trewartha, G.; Vinkovic, M.; Williams, B.; Woolford, A. J. Discovery of (2,4-dihydroxy-5-isopropylphenyl)-[5-(4-methylpiperazin-1-ylmethyl)-1,3-dihydroisoindol-2-yl]methanone (AT13387), a novel inhibitor of the molecular chaperone Hsp90 by fragment based drug design. *J. Med. Chem.* **2010**, *53* (16), 5956−5969.
(77) Sixma, T. K.; Smit, A. B. Acetylcholine binding protein (AChBP): A secreted glial protein that provides a high-resolution model for the extracellular domain of pentameric ligand-gated ion channels. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32* (1), 311−334.
(78) Edink, E.; Rucktooa, P.; Retra, K.; Akdemir, A.; Nahar, T.; Zuiderveld, O.; van Elk, R.; Janssen, E.; van Nierop, P.; van Muijlwijk-Koezen, J.; Smit, A. B.; Sixma, T. K.; Leurs, R.; de Esch, I. J. Fragment growing induces conformational changes in acetylcholine-binding protein: A structural and thermodynamic analysis. *J. Am. Chem. Soc.* **2011**, *133*, 5363−5371.
(79) Bemis, G.; Murcko, M. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887−2893.
(80) McCourt, J. A.; Pang, S. S.; Guddat, L. W.; Duggleby, R. G. Elucidating the specificity of binding of sulfonylurea herbicides to acetohydroxyacid synthase. *Biochemistry* **2005**, *44* (7), 2330−2238.
(81) ChemAxon pKa Calculator Plugin. https://www.chemaxon.com/products/calculator-plugins/property-predictors/ (accessed February 2013).
(82) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein−ligand docking performance. *J. Med. Chem.* **2007**, *50* (4), 726−741.
(83) Martin, E. J.; Sullivan, D. C. Surrogate AutoShim: Predocking into a universal ensemble kinase receptor for three dimensional activity prediction, very quickly, without a crystal structure. *J. Chem. Inf. Model.* **2008**, *48* (4), 873−881.
(84) Marcou, G.; Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.* **2007**, *47* (1), 195−207.
(85) DeLano, W. L. *The PyMOL Molecular Graphics System*; DeLano Scientific: San Carlos, CA, 2002.

6 Application à un cas d'étude prospectif

La stratégie virtuelle d'optimisation de fragment, décrite et validée au chapitre 5, est finalement appliquée à un projet prospectif interne à l'entreprise. Le hit primaire de type fragment fut initialement identifié, en amont de cette thèse, à l'aide d'un criblage expérimental par spectrométrie de masse.

6.1 Description de la cible

La cible, Heat Shock Protein 90 (HSP90), appartient à la famille des chaperonnes moléculaires ATP-dépendantes [210]. La protéine a une masse d'environ 90 kDa d'où son nom, et la forme humaine possède 732 résidus (code Uniprot [211] : P07900). Ce type de protéine a pour fonction d'assister d'autres protéines dites "clientes" dans le but de favoriser leur bon repliement et donc leur fonctionnement correct [212]. Au sein de cellules saines, HSP90 est spécifiquement exprimée en réponse à un choc thermique afin de tenter de résoudre les problèmes de dénaturation / altération des structures 3D de ses protéines clientes. Parmi ces dernières [213] figurent des kinases (Src, Raf), des facteurs de transcription (récepteurs nucléaires des hormones stéroïdes) et même des télomerases. HSP90 joue donc un rôle positif dans la survie d'une cellule soumise à un stress, *via* la protection de protéines fondamentales pour assurer son bon fonctionnement. L'inhibition de HSP90 aboutit donc à une très forte diminution de son activité chaperonne. Dans ces conditions, les tâches accomplies par ses protéines clientes sont très perturbées lorsque la cellule fait face à un stress. Au regard du rôle clé de certaines de ses protéines clientes (surexprimées et/ou mutées) dans certains cancers, HSP90 est devenue une cible majeure et centrale pour lutter contre cette maladie [214, 215]. Des inhibiteurs de HSP90 sont en cours d'étude (au stade clinique) dans le cadre d'un traitement contre le cancer [203].
D'un point de vue structural, HSP90 est constituée de trois domaines (N-terminal, intermédiaire et C-terminal) [216]. Le domaine C-terminal est lié au processus d'homodimérisation et celui dit intermédiaire recrute les protéines clientes. Le domaine N-terminal, dit domaine ATPase, est celui qui lie l'ATP, et l'hydrolyse de ce composé fournit l'énergie nécessaire pour assurer la fonction chaperonne. L'analyse de la structure PDB 1BYQ montre que la partie adénine de la molécule d'ADP interagit avec le résidu clé ("hot spot") D93 au moyen d'une LH directe et de plusieurs LH médiées par des molécules d'eau hautement conservées (sur la base de l'observation de nombreuses structures de HSP90), tandis que ses groupements phosphates sont impliqués dans des liaisons de coordination avec un ion Mg^{2+} et des LH avec N51 et F138 (voir la Figure 51). De nombreux inhibiteurs compétitifs de l'ATP ont été développés à l'aide d'approches de type SBDD [217, 218] et/ou FBDD [196, 215, 219-221].

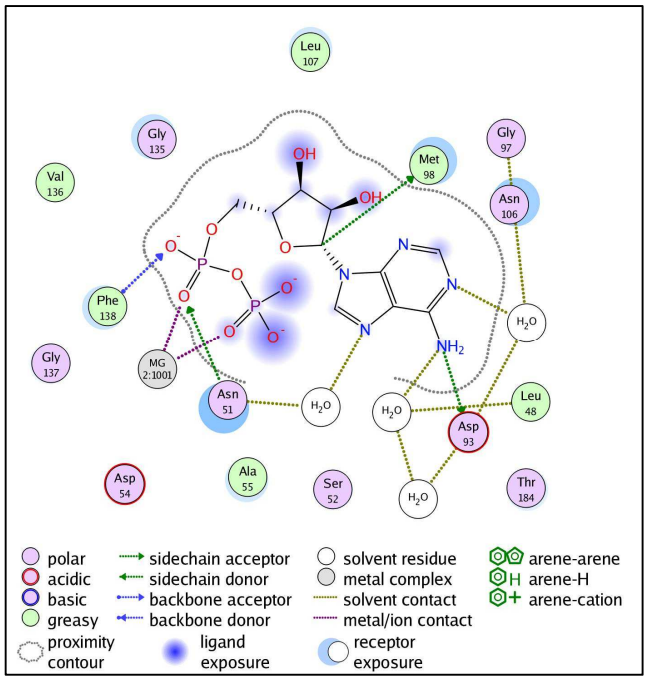


Figure 51: Diagramme 2D d'interaction [222] entre la molécule d'ADP et le site de liaison de HSP90. Pour plus de clarté, seules les molécules d'eau à proximité du groupement adénine sont représentées.

Enfin, une partie du domaine N-terminal de HSP90 est présent sous plusieurs conformations alternatives (voir la partie gauche de la Figure 52) selon qu'il y ait ou non un ligand et selon le type de ligand. La zone autour de D93 est quant à elle très conservée d'un point de vue conformationnel.

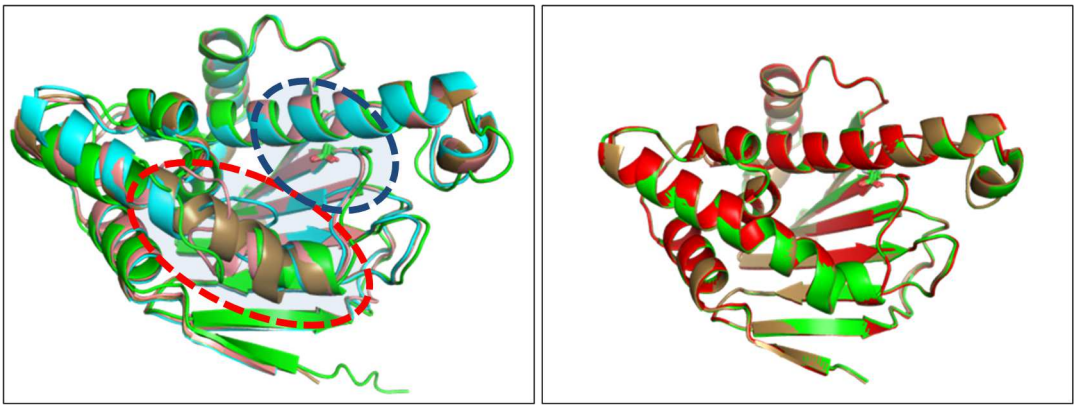


*Figure 52: Superposition 3D de plusieurs structures de HSP90 représentées sous forme de cartoon. A gauche : illustration de diverses conformations du domaine N-terminal (codes PDB : 1BYQ en vert, 2JJC en bleu, 2QFO en rose et 3D0B en marron). A droite : superposition de la structure 3D0B avec les deux structures cristallographiques "3D0B-like" contenant les fragments flx337 et flx338 (voir le §**Erreur ! Source du renvoi introuvable.**). La chaîne latérale de D93 est affichée en stick, et les zones en pointillées illustrent des régions variables (rouge) ou très conservées (bleu) en terme de conformation.*

<u>Remarque</u> : ce projet prospectif n'a pu se terminer dans les délais impartis (fin de la thèse), et est de fait encore en cours au moment où ces lignes sont écrites. Par conséquent, le reste de ce chapitre (à savoir les données expérimentales initiales, la phase de modélisation et une partie de celles expérimentales ; 14 pages) a été supprimé dans cette version dite de diffusion du manuscrit de thèse.

Conclusion générale et perspectives

Cette thèse avait comme objectif principal le développement, la validation et l'application d'approches de chémoinformatique et de modélisation moléculaire appliquées au FBDD. Plus spécifiquement, il s'agissait de travailler sur deux étapes critiques du FBDD qui sont :

- la prédiction des modes de liaison de composés organiques de type fragment notamment à l'aide d'approches de docking moléculaire
- la capacité à faciliter la phase d'optimisation d'un fragment de départ *via* la suggestion de structures à synthétiser selon une approche SBDD et en réutilisant les notions fondamentales de "growing" et "linking"

De plus, ce travail devait également porter sur des développements informatiques au niveau du logiciel S4MPLE. Il s'agit d'un outil d'EC, interne au Laboratoire de Chémoinformatique, utilisant un AG hybride et dont la fonction objective repose sur un FF.

Dans un but de synergie entre ces deux axes de travail, il était souhaité d'intégrer S4MPLE dans la mesure du possible au sein du contexte FBDD *in silico*, notamment dans sa composante SBDD.

Les principales conclusions de ce travail, dont la thématique principale aborde des domaines et sujets variés, sont présentées ci-dessous.

Tout d'abord, les différents aspects du FBDD, tant expérimentaux que *in silico*, ont fait l'objet d'une revue qui a été écrite pendant cette thèse.

La première phase du projet a consisté en un travail sur la fonction d'énergie de S4MPLE, et plus spécifiquement sur la calibration des termes additionnels, entre autres ceux liés au modèle de solvant implicite. Etant donné qu'ils sont de nature non liée et donc en nombre important, ces termes sont définis de manière à ce qu'ils soient d'une complexité comparable aux termes usuels des FF. L'approche sélectionnée, basée non pas sur une calibration impliquant des données d'affinité et des poses (uniques) X-ray, a consisté à pouvoir préférentiellement distinguer le bon conformère (ou mode de liaison) d'une multitude de mauvaises solutions.
Idéalement, le fruit de la première étape de calibration fonctionnerait pour des peptides variés, et serait directement transférable à d'autres systèmes comme par exemple des systèmes ligand-récepteur.

Cependant, aucun des schémas énergétiques testés et/ou sélectionnés (FF natif mais également tous les autres jeux de paramètres) n'a permis de gérer correctement l'un des peptides du jeu (1LE1) dont le repliement implique un brin β et 2 interactions π-π en forme de T. Bien qu'imparfaits, les résultats préliminaires sur le jeu de peptides étaient toutefois meilleurs qu'avec le FF natif au regard des critères utilisés, d'où la poursuite dans cette voie.

Le passage à des systèmes ligand-récepteur a ensuite révélé d'autres difficultés au niveau des paramètres, et plus particulièrement vis-à-vis des cations et des LH. Trois alternatives, déjà discutées précédemment, étaient disponibles à ce stade. Etant donné les contraintes de temps et le souhait de ne pas diminuer globalement la valeur de l'un des paramètres mis en cause (*desolv_factor*), la seule alternative qui restait consistait à traiter ces paires non liées de manière spécifique au moyen d'un facteur correctif, au risque d'un "overfitting" éventuel. Leur calibration a reposé sur une approche par essais et erreurs à l'aide de simulations de docking.

Ensuite, la recherche des poids optimaux du terme de contact, basée sur un scan systématique et une réévaluation d'ensembles de poses, a permis d'identifier un triplet améliorant sensiblement les résultats en docking (Fit FF *vs.* Core FF et FF préliminaire) en utilisant un jeu de validation externe reconnu dans la communauté.

Etant donné certains tâtonnements et un overfitting éventuel au niveau d'une des étapes de calibration, le jeu final de paramètres (Fit FF) fut encore testé selon deux voies : l'EC biaisé et non biaisé de peptides (comme précédemment), et la corrélation entre les énergies FF et des affinités expérimentales. Le fait de réobtenir des résultats comparables pour le Fit FF par rapport au FF préliminaire est à noter. Le jeu final de paramètres a même permis d'améliorer le cas du peptide 2KEF, puisque maintenant le meilleur conformère évalué possède bien un repliement natif. Ainsi, la mise à l'échelle de certaines paires spécifiques et le fait de donner des poids différents aux différentes classes de carbone n'ont ni chamboulé ni dénaturé le FF lorsque l'on retourne à des systèmes "100% peptide" (ces modifications étaient toutefois nécessaires pour les complexes ligand-récepteur). Enfin, une relative corrélation, tout à fait acceptable au regard de l'état de l'art, est observée entre les énergies calculées selon Fit FF et les affinités expérimentales (K_i). Les études dédiées à cette problématique incorporent bien évidemment un plus grand nombre de points, mais une absence totale de corrélation aurait certainement été mise en évidence sous l'hypothèse d'un jeu de paramètres incohérents.

Cette première phase, chronophage, laborieuse et décevante dans ses premiers essais, a fini par aboutir à des résultats concrets et exploitables. Toutefois, elle a bien mis en évidence les difficultés extrêmement connues pour obtenir des paramètres pertinents et universels, afin de pouvoir les transférer directement d'un système à un autre.

Après certains développements informatiques, le programme S4MPLE a été validé selon les règles en vigueur dans la communauté scientifique à l'aide de simulations de redocking. Des résultats comparables à ceux d'outils de référence ont été obtenus lors de ces simulations. De plus, les taux de succès entre des jeux de validation divers (Astex Diverse Set essentiellement composé de complexes "ligand drug-like - récepteur" *vs.* un ensemble uniquement composé de complexes "fragment-récepteur") sont également comparables, ce qui permet de répondre favorablement - au moins partiellement - à l'une des problématiques initiales qui était "*Une même fonction de score ou d'énergie est-elle compatible avec différentes classes de ligands ?*". Toutefois, une étude plus systématique, portant sur un plus grand nombre de complexes et intégrant à la fois des simulations de redocking et de cross-docking, permettrait de répondre de manière plus rigoureuse à cette question. Cette dernière remarque est également valable pour la plupart des études sur le sujet.

Parallèlement, certaines caractéristiques originales, entre autre le fait de pouvoir simuler simultanément plusieurs entités, ont été mises en évidence. La prise en compte dynamique de molécules d'eau en docking est l'un des sujets actuels majeurs au même titre que la gestion de la flexibilité. Ainsi, il a été montré que S4MPLE est capable de docker un ligand tout en incluant des molécules d'eau explicites libres afin de pouvoir détecter des interactions médiées. Dans le même esprit, la gestion de deux ligands, en l'occurrence deux fragments, est également validée *via* la capacité de S4MPLE à prédire la géométrie de complexes ternaires qui sont bien plus fréquents dans le contexte du FBDD.

Ces résultats encourageants sur la bonne gestion de composés de type fragment ont ouvert la voie à la dernière phase, à savoir la problématique dédiée à l'optimisation virtuelle de composés reposant sur une approche 3D. Un axiome du FBDD stipule que le mode de liaison du fragment de départ doit être conservé lors de la phase d'évolution. Par exemple dans le contexte du linking, un espaceur approprié doit en principe conserver le mode de liaison de chacun des fragments du complexe ternaire initial.
Une stratégie en deux étapes principales a été définie pour répondre à cette problématique : 1) il y a création d'une chimiothèque focalisée sur le ou les fragment(s) de départ, et 2) un criblage virtuel des différents composés générés est réalisé avec S4MPLE. Bien que cet EC soit contraint dans une première phase, ce qui n'est pas en contradiction vis-à-vis de l'axiome énoncé ci-dessus, l'intégralité du ligand est finalement relaxée, et c'est cette énergie non biaisée qui sert à classer les composés. Ce protocole a été validé avec succès à l'aide de simulations rétrospectives diverses (growing et linking) et plus ou moins ambitieuses (ajout d'une certaine flexibilité au niveau du site ou de molécules d'eau libres) reposant sur des données provenant essentiellement de la littérature scientifique du FBDD.

L'intégralité de ces résultats permet finalement de répondre favorablement au cahier des charges illustré par une autre problématique définie dans l'introduction : "*Peut-on développer une stratégie virtuelle mais rationnelle d'optimisation, utilisable dans ce contexte du FBDD, et reposant sur les concepts usuels de chémoinformatique et de modélisation moléculaire ?*".
Cette stratégie *in silico* est actuellement appliquée sur un cas d'étude concret au sein de l'entreprise NovAliX. Malheureusement, ce projet interne n'a pu se terminer dans les délais impartis (fin de cette thèse), et il n'est pas possible à ce stade de conclure favorablement ou défavorablement sur la base des résultats expérimentaux trop peu nombreux.

En termes de perspectives, ce travail pourra être complété dans l'avenir par plusieurs développements visant à corriger / améliorer certains points qui ont émergé lors de ce travail. Les principaux sont introduits ci-dessous.

Tout d'abord, on pourrait imaginer une amélioration de la stratégie de création de la population initiale. Cet aspect est très important pour un AG : plus la population initiale est de qualité (bonne diversité et conformations de basse énergie) et plus l'AG est censé converger rapidement vers l'une des solutions optimales. Actuellement, la minimisation que subit chaque nouvel individu peut potentiellement aboutir à d'importantes distorsions afin de supprimer les mauvais contacts de VdW (clashes). Le positionnement actuel, basé sur un contact favorable pour amener le ligand dans le site, pourrait ainsi être amélioré *via* l'utilisation d'un plus grand nombre de points (3 par exemple). Toutefois, la stratégie actuelle est très efficace pour pouvoir "tapisser" le site dans le cas du docking de fragments qui font par nature peu d'interactions avec lui. Ainsi, tout comme au sein du logiciel FlexX, on pourrait envisager une approche variable d'ancrage du ligand (aléatoire, ou fonction de sa taille). Enfin, l'aspect "temps de calcul" n'est pas non plus négligeable afin de rendre S4MPLE bien plus compétitif sur ce point. A ce titre, un projet ANR, incluant un Laboratoire d'informatique, a été déposé afin de travailler sur le déploiement de S4MPLE et sur des aspects d'actualité très techniques (GPU).

L'approche actuellement utilisée pour construire la chimiothèque focalisée est basée sur une complémentarité d'étiquettes RECAP. Elle est certes simple et intuitive, mais elle ne tient pas compte des voies de synthèse organique et des synthons potentiellement disponibles. Par conséquent, il faut améliorer ce point afin de rendre la méthode plus efficace au plan expérimental.
L'encodage et la gestion de réactions chimiques (réactifs→produits), en lieu et place de la complémentarité d'étiquettes, est une voie d'amélioration possible [164]. Cela devrait permettre d'être à

la fois plus proche de l'accessibilité chimique tout en intégrant directement les synthons réellement disponibles, d'où un meilleur rendement attendu au niveau de la phase de synthèse organique.

Enfin, il a été convenu au début de ce projet de ne pas s'atteler à l'aspect "création de chimiothèques de fragments" qui est pourtant fondamental, et qui a fait l'objet d'une certaine littérature. Ainsi, un travail dans cette voie apparaît comme une perspective légitime, et ce d'autant plus qu'il permettrait de compléter l'arsenal actuel (docking, criblage virtuel, évolution virtuelle) avec le dernier point restant vacant vis-à-vis des étapes clés du FBDD dans lesquelles la chémoinformatique et la modélisation moléculaire peuvent être impliquées.

Communications scientifiques

Ce chapitre répertorie les différentes communications scientifiques réalisées pendant le déroulement de cette thèse. Comme le veut l'usage, le terme communication regroupe les publications dans un journal à comité de lecture et les communications orales et par affiches présentées lors de congrès.

Publications dans un journal à comité de lecture

Un article a été publié ou soumis (en attende de validation au moment de l'écriture) pour chacun des principaux chapitres (3, 4 et 5) de cette thèse :

Article I (voir le §3.4)

Auteurs	Hoffer Laurent, Chira Camelia, Marcou Gilles, Varnek Alexandre et Horvath Dragos
Titre	S4MPLE – Sampler For Multiple Protein-Ligand Entities. Methodology & Rigid-Site Docking Benchmarking
Journal	Journal of Cheminformatics
Editeur	Chemistry Central
Pagination	N/A (article soumis)
DOI	N/A (article soumis)

Article II (voir le §4.3)

Auteurs	Hoffer Laurent et Horvath Dragos
Titre	S4MPLE – Sampler For Multiple Protein-Ligand Entities. Simultaneous docking of several entities
Journal	Journal of Chemical Information and Modeling (JCIM)
Editeur	ACS Publications
Année	2012
Volume	53
Pagination	88-102
DOI	http://dx.doi.org/10.1021/ci300495r

Article III (voir le §5.5)

Auteurs	Hoffer Laurent, Renaud Jean-Paul et Horvath Dragos
Titre	In Silico Fragment-Based Drug Discovery: Setup and Validation of a Fragment-to-Lead Computational Protocol using S4MPLE
Journal	Journal of Chemical Information and Modeling (JCIM)
Editeur	ACS Publications
Année	2013
Volume	N/A (article accepté)
Pagination	N/A (article accepté)
DOI	http://dx.doi.org/10.1021/ci4000163

Une revue, consacrée au FBDD de manière générale, a également été écrite lors de ce travail de thèse :

Revue (voir le §1.3.1)

Auteurs	Hoffer Laurent, Renaud Jean-Paul et Horvath Dragos
Titre	Fragment-Based Drug Design: Computational & Experimental State of the Art
Journal	Combinatorial Chemistry & High Throughput Screening (CCHTS)
Editeur	Bentham
Année	2011
Volume	14
Pagination	500-520
DOI	http://dx.doi.org/10.2174/138620711795767884

Communications par affiche

Une première affiche scientifique

Auteurs : Hoffer Laurent, Varnek Alexandre et Horvath Dragos

Titre : Sampling, Docking & Fragment-Based Drug Design

a été présentée lors :

- des 5èmes Journées de la Société Française de Chémoinformatique (2011 - Cabourg)

URL : http://www.chemoinformatique.fr/modules/smartsection/item.php?itemid=38

- des 8èmes Rencontres des Chimistes Théoriciens du Grand Est (2012 - Dijon)

URL : http://www.icb.cnrs.fr/RCTGE/

Une seconde affiche scientifique

Auteurs : Hoffer Laurent, Varnek Alexandre et Horvath Dragos

Titre : Fragment-Based Drug Design using S4MPLE

a été présentée lors :

- de la 3rd Summer School on Chemoinformatics (2012 - Strasbourg)

URL : http://infochim.u-strasbg.fr/spip.php?rubrique132

- du congrès international Suprachem (2012 - Strasbourg)

URL : http://suprachem2012.u-strasbg.fr/

Communication orale

Une communication orale a été présentée lors du congrès Groupe de Graphisme et Modélisation Moléculaire (GGMM) qui s''est déroulé à La Rochelle en juin 2011 :

Auteurs : Hoffer Laurent, Varnek Alexandre et Horvath Dragos

Titre : Sampling, Docking & Fragment-Based Drug Design

URL : http://ggmm2011.wordpress.com

Bibliographie

1. Marcou, G.; Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J Chem Inf Model* **2007,** 47, 195-207.
2. http://www.legifrance.gouv.fr. In.
3. Leeson, P.; Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat Rev Drug Discov* **2007,** 6, 881-90.
4. Hughes, J. P.; Rees, S.; Kalindjian, S. B.; Philpott, K. L. Principles of early drug discovery. *Br J Pharmacol* **2011,** 162, 1239-49.
5. Josephy, P. D. New developments in the ames assay: High-sensitivity detection of mutagenic arylamines. *BioEssays* **1989,** 11, 108-112.
6. Varnek, A.; Baskin, I. Chemoinformatics as a Theoretical Chemistry Discipline. *Molecular Informatics* **2011,** 30, 20-32.
7. Katritzky, A.; Kuanar, M.; Fara, D.; Karelson, M.; Acree, W.; Solov'ev, V.; Varnek, A. QSAR modeling of blood : air and tissue : air partition coefficients using theoretical descriptors. *Bioorganic & Medicinal Chemistry* **2005,** 13, 6450-6463.
8. Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics* **2010,** 29, 476-488.
9. Muratov, E.; Fourches, D.; Artemenko, A.; Kuz'min, V.; Zhao, G. Y.; Golbraikh, A.; Polischuk, P.; Varlamova, E.; Baskin, I.; Palyulin, V.; Zefirov, N.; Li, J. Z.; Gramatica, P.; Martin, T.; Hormozdiari, F.; Dao, P.; Sahinalp, C.; Cherkasov, A.; Oberg, T.; Todeschini, R.; Poroikov, V.; Zaharov, A.; Lagunin, A.; Filimonov, D.; Varnek, A.; Horvath, D.; Marcou, G.; Muller, C.; Xi, L. L.; Liu, H. X.; Yao, X. J.; Hansen, K.; Schroeter, T.; Muller, K. R.; Tetko, I.; Sushko, I.; Novotarskyi, S.; Baker, N.; Reed, J.; Barnes, J.; Tropsha, A. Collaborative QSAR analysis of Ames mutagenicity. *Abstracts of Papers of the American Chemical Society* **2011,** 241, 1.
10. Varnek, A.; Baskin, I. Machine Learning Methods for Property Prediction in Chemoinformatics: Quo Vadis? *Journal of Chemical Information and Modeling* **2012,** 52, 1413-1437.
11. Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I. V.; Marcou, G. ISIDA - Platform for virtual screening based on fragment and pharmacophoric descriptors. *Curr. Comput.-Aided Drug Des.* **2008,** 4, 191-198.
12. Moroy, G.; Martiny, V. Y.; Vayer, P.; Villoutreix, B. O.; Miteva, M. A. Toward in silico structure-based ADMET prediction in drug discovery. *Drug Discov Today* **2012,** 17, 44-55.
13. Rognan, D. Chemogenomic approaches to rational drug design. *Br J Pharmacol* **2007,** 152, 38-52.
14. van Westen, G.; Wegner, J.; IJzerman, A.; van Vlijmen, H.; Bender, A. Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *Medchemcomm* **2011,** 2, 16-30.
15. Brooijmans, N.; Kuntz, I. D. Molecular recognition and docking algorithms. *Annual Review of Biophysics and Biomolecular Structure* **2003,** 32, 335-373.
16. Holdgate, G.; Ward, W. Measurements of binding thermodynamics in drug discovery. *Drug Discov Today* **2005,** 10, 1543-50.
17. Dunitz, J. D. Win some, lose some: enthalpy-entropy compensation in weak intermolecular interactions. *Chem Biol* **1995,** 2, 709-12.
18. Freire, E. A Thermodynamic Approach to the Affinity Optimization of Drug Candidates. *Chemical Biology & Drug Design* **2009**, 468-472.
19. Bissantz, C.; Kuhn, B.; Stahl, M. A medicinal chemist's guide to molecular interactions. *J Med Chem* **2010,** 53, 5061-84.
20. Bosshard, H.; Marti, D.; Jelesarov, I. Protein stabilization by salt bridges: concepts, experimental approaches and clarification of some misunderstandings. *Journal of Molecular Recognition* **2004,** 17, 1-16.

21. Steiner, T. The hydrogen bond in the solid state. *Angewandte Chemie-International Edition* **2002,** 41, 48-76.
22. Pace, C. N.; Shirley, B. A.; McNutt, M.; Gajiwala, K. Forces contributing to the conformational stability of proteins. *FASEB J* **1996,** 10, 75-83.
23. Boehringer, M.; Fischer, H.; Hennig, M.; Hunziker, D.; Huwyler, J.; Kuhn, B.; Loeffler, B. M.; Luebbers, T.; Mattei, P.; Narquizian, R.; Sebokova, E.; Sprecher, U.; Wessel, H. P. Aryl- and heteroaryl-substituted aminobenzo[a]quinolizines as dipeptidyl peptidase IV inhibitors. *Bioorganic & Medicinal Chemistry Letters* **2010,** 20, 1106-1108.
24. Leung, C. S.; Leung, S. S.; Tirado-Rives, J.; Jorgensen, W. L. Methyl effects on protein-ligand binding. *J Med Chem* **2012,** 55, 4489-500.
25. McGaughey, G. B.; Gagné, M.; Rappé, A. K. pi-Stacking interactions. Alive and well in proteins. *J Biol Chem* **1998,** 273, 15458-63.
26. Sinnokrot, M. O.; Valeev, E. F.; Sherrill, C. D. Estimates of the ab initio limit for pi-pi interactions: the benzene dimer. *J Am Chem Soc* **2002,** 124, 10887-93.
27. Hunter, C.; Sanders, J. The nature of pi-pi interactions. *Journal of the American Chemical Society* **1990,** 112, 5525-5534.
28. http://www.schrodinger.com/kb/1556. In.
29. Burley, S. K.; Petsko, G. A. Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science* **1985,** 229, 23-8.
30. Kumar, N. V.; Govil, G. Theoretical studies on protein-nucleic acid interactions. III. Stacking of aromatic amino acids with bases and base pairs of nucleic acids. *Biopolymers* **1984,** 23, 2009-24.
31. Ma, J.; Dougherty, D. The cation-pi interaction. *Chemical Reviews* **1997,** 97, 1303-1324.
32. Dougherty, D. A. Cation-pi interactions in chemistry and biology: a new view of benzene, Phe, Tyr, and Trp. *Science* **1996,** 271, 163-8.
33. Dougherty, D. A. Cation-pi interactions involving aromatic amino acids. *J Nutr* **2007,** 137, 1504S-1508S; discussion 1516S-1517S.
34. Gallivan, J. P.; Dougherty, D. A. Cation-pi interactions in structural biology. *Proc Natl Acad Sci U S A* **1999,** 96, 9459-64.
35. Yun, M.; Wu, J.; Workman, J. L.; Li, B. Readers of histone modifications. *Cell Res* **2011,** 21, 564-78.
36. Musselman, C. A.; Lalonde, M. E.; Côté, J.; Kutateladze, T. G. Perceiving the epigenetic landscape through histone readers. *Nat Struct Mol Biol* **2012,** 19, 1218-27.
37. Zacharias, N.; Dougherty, D. A. Cation-pi interactions in ligand recognition and catalysis. *Trends Pharmacol Sci* **2002,** 23, 281-7.
38. Harding, M. Geometry of metal-ligand interactions in proteins. *Acta Crystallographica Section D* **2001,** 57, 401-411.
39. Harding, M. Small revisions to predicted distances around metal sites in proteins. *Acta Crystallographica Section D* **2006,** 62, 678-682.
40. Salgado, E. N.; Radford, R. J.; Tezcan, F. A. Metal-directed protein self-assembly. *Acc Chem Res* **2010,** 43, 661-72.
41. Holmquist, B.; Vallee, B. L. Metal-coordinating substrate analogs as inhibitors of metalloenzymes. *Proc Natl Acad Sci U S A* **1979,** 76, 6216-20.
42. Rouffet, M.; Cohen, S. M. Emerging trends in metalloprotein inhibition. *Dalton Trans* **2011,** 40, 3445-54.
43. Hubbard. Structure-Based Drug Discovery. *RSC Biomolecular Sciences* **2006**.
44. Bohacek, R.; McMartin, C.; Guida, W. The art and practice of structure-based drug design: a molecular modeling perspective. *Med Res Rev* **1996,** 16, 3-50.
45. Marrone, T. J.; Briggs, J. M.; McCammon, J. A. Structure-based drug design: computational advances. *Annu Rev Pharmacol Toxicol* **1997,** 37, 71-90.

46. Wolber, G.; Langer, T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Mod.* **2005,** 45, 160-169.
47. Congreve, M.; Chessari, G.; Tisi, D.; Woodhead, A. J. Recent developments in fragment-based drug discovery. *Journal of Medicinal Chemistry* **2008,** 51, 3661-3680.
48. Lipinski, C.; Lombardo, F.; Dominy, B.; Feeney, P. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* **2001,** 46, 3-26.
49. Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A rule of three for fragment-based lead discovery? *Drug Discovery Today* **2003,** 8, 876-877.
50. Law, R.; Barker, O.; Barker, J. J.; Hesterkamp, T.; Godemann, R.; Andersen, O.; Fryatt, T.; Courtney, S.; Hallett, D.; Whittaker, M. The multiple roles of computational chemistry in fragment-based drug design. *Journal of Computer-Aided Molecular Design* **2009,** 23, 459-473.
51. Davies, D.; Mamat, B.; Magnusson, O.; Christensen, J.; Haraldsson, M.; Mishra, R.; Pease, B.; Hansen, E.; Singh, J.; Zembower, D.; Kim, H.; Kiselyov, A.; Burgin, A.; Gurney, M.; Stewart, L. Discovery of leukotriene A4 hydrolase inhibitors using metabolomics biased fragment crystallography. *J Med Chem* **2009,** 52, 4694-715.
52. Renaud, J. P.; Delsuc, M. A. Biophysical techniques for ligand screening and drug design. *Current Opinion in Pharmacology* **2009,** 9, 622-628.
53. Borsi, V.; Calderone, V.; Fragai, M.; Luchinat, C.; Sarti, N. Entropic Contribution to the Linking Coefficient in Fragment Based Drug Design: A Case Study. *Journal of Medicinal Chemistry* **2010,** 53, 4285-4289.
54. Ichihara, O.; Barker, J.; Law, R.; Whittaker, M. Compound Design by Fragment-Linking. *Molecular Informatics* **2011,** 30, 298-306.
55. Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today* **2004,** 9, 430-431.
56. Hoffer, L.; Renaud, J. P.; Horvath, D. Fragment-Based Drug Design: Computational and Experimental State of the Art. *Combinatorial Chemistry & High Throughput Screening* **2011,** 14, 500-520.
57. Hitaoka, S.; Matoba, H.; Harada, M.; Yoshida, T.; Tsuji, D.; Hirokawa, T.; Itoh, K.; Chuman, H. Correlation Analyses on Binding Affinity of Sialic Acid Analogues and Anti-Influenza Drugs with Human Neuraminidase Using ab Initio MO Calculations on Their Complex Structures – LERE-QSAR Analysis (IV). *Journal of Chemical Information and Modeling* **2011,** 51, 2706-2716.
58. Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; Debolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. AMBER a package of computer-programs for applying molecular mechanics, normal-mode analysis, molecular-dynamics and free-energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications* **1995,** 91, 1-41.
59. Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Swails, J.; Goetz, A. W.; Kolossvary, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wolf, R. M.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Salomon-Ferrer, R.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. AMBER 12. *University of California, San Francisco* **2012**.
60. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM - A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry* **1983,** 4, 187-217.
61. Jorgensen, W.; Maxwell, D.; TiradoRives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society* **1996,** 118, 11225-11236.

62. Halgren, T. A. MMFF VII. Characterization of MMFF94, MMFF94s, and other widely available force fields for conformational energies and for intermolecular-interaction energies and geometries. *Journal of Computational Chemistry* **1999,** 20, 730-748.
63. Gunsteren, W. f. v. *Biomolecular simulation: the Gromos96 manual and user guide*. vdf Hochschulverlag Eth Zurich: 1996.
64. van Gunsteren, W. F.; Bakowies, D.; Baron, R.; Chandrasekhar, I.; Christen, M.; Daura, X.; Gee, P.; Geerke, D. P.; Glättli, A.; Hünenberger, P. H.; Kastenholz, M. A.; Oostenbrink, C.; Schenk, M.; Trzesniak, D.; van der Vegt, N. F.; Yu, H. B. Biomolecular modeling: Goals, problems, perspectives. *Angew Chem Int Ed Engl* **2006,** 45, 4064-92.
65. Vedani, A. YETI: An interactive molecular mechanics program for small-molecule protein complexes. *Journal of Computational Chemistry* **1988,** 9, 269-280.
66. Morse, P. M. Diatomic Molecules According to the Wave Mechanics. II. Vibrational Levels. *Physical Review* **1929,** 34, 57-64.
67. Greenidge, P. A.; Kramer, C.; Mozziconacci, J. C.; Wolf, R. M. MM/GBSA Binding Energy Prediction on the PDBbind Data Set: Successes, Failures, and Directions for Further Improvement. *J Chem Inf Model* **2012**.
68. Fletcher, R.; Reeves, C. M. Function minimization by conjugate gradients. *Computer Journal* **1964,** 7, 149-154.
69. Liu, D. C.; Nocedal, J. On the Limited Memory BFGS Method for Large Scale Optimization. *Mathematical Programming* **1989,** 45, 503-528.
70. Karplus, M.; McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat Struct Biol* **2002,** 9, 646-52.
71. Hansson, T.; Oostenbrink, C.; van Gunsteren, W. Molecular dynamics simulations. *Curr Opin Struct Biol* **2002,** 12, 190-6.
72. Kollman, P. Free-energy calculations : applications to chemical and biochemical phenomena. *Chemical Reviews* **1993,** 93, 2395-2417.
73. Durrant, J. D.; McCammon, J. A. Molecular dynamics simulations and drug discovery. *BMC Biol* **2011,** 9, 71.
74. Brooks, C.; Case, D. Simulations of peptide conformational dynamics and thermodynamics. *Chemical Reviews* **1993,** 93, 2487-2502.
75. Grossfield, A. Recent progress in the study of G protein-coupled receptors with molecular dynamics computer simulations. *Biochim Biophys Acta* **2011,** 1808, 1868-78.
76. Christen, M.; van Gunsteren, W. F. On searching in, sampling of, and dynamically moving through conformational space of biomolecular systems: A review. *J Comput Chem* **2008,** 29, 157-66.
77. Verlet, L. Computer Experiments on Classical Fluids. I. Thermodynamical Properties of Lennard-jones Molecules. *Phys. Rev.* **1967,** 159, 98-103.
78. Swope, W.; Andersen, H.; Berens, P.; Wilson, K. A computer-simulation method for the calculation of equilibrium-constants for the formation of physical clusters of molecules - application to small water clusters. *Journal of Chemical Physics* **1982,** 76, 637-649.
79. van Gunsteren, W. F. Constrained dynamics of flexible molecules. *Molecular Physics* **1980,** 40, 1015-1019.
80. Lill, M. A. Multi-dimensional QSAR in drug discovery. *Drug Discov Today* **2007,** 12, 1013-7.
81. Gilson, M. K.; Zhou, H. X. Calculation of protein-ligand binding affinities. *Annu Rev Biophys Biomol Struct* **2007,** 36, 21-42.
82. Michel, J.; Foloppe, N.; Essex, J. Rigorous Free Energy Calculations in Structure-Based Drug Design. *Molecular Informatics* **2010,** 29, 570-578.
83. Hu, R.; Barbault, F.; Maurel, F.; Delamar, M.; Zhang, R. Molecular dynamics simulations of 2-amino-6-arylsulphonylbenzonitriles analogues as HIV inhibitors: interaction modes and binding free energies. *Chem Biol Drug Des* **2010,** 76, 518-26.

84. Barbault, F.; Maurel, F. Is inhibition process better described with MD(QM/MM) simulations? The case of urokinase type plasminogen activator inhibitors. *J Comput Chem* **2012,** 33, 607-16.
85. Murray, C.; Verdonk, M. The consequences of translational and rotational entropy lost by small molecules on binding to proteins. *J Comput Aided Mol Des* **2002,** 16, 741-53.
86. Hou, T.; Wang, J.; Li, Y.; Wang, W. Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J Chem Inf Model* **2011,** 51, 69-82.
87. Bohm, H.-J. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *Journal of Computer-Aided Molecular Design* **1994,** 8, 243-256.
88. de Bakker, P. I.; Furnham, N.; Blundell, T. L.; DePristo, M. A. Conformer generation under restraints. *Curr Opin Struct Biol* **2006,** 16, 160-5.
89. Liwo, A.; Czaplewski, C.; Ołdziej, S.; Scheraga, H. A. Computational techniques for efficient conformational sampling of proteins. *Curr Opin Struct Biol* **2008,** 18, 134-9.
90. Mohamadi, F.; Richards, N.; Guida, W.; Liskamp, R.; Lipton, M.; Caufield, C.; Chang, G.; Hendrickson, T.; Still, W. MACROMODEL - An integrated software system for modeling organic and bioorganic molecules using molecular mechanics. *Journal of Computational Chemistry* **1990,** 11, 440-467.
91. Gasteiger, J.; Hiller, C.; Rudolph, C.; Sadowski, J. Automatic-generation of 3D-atomic coordinates for organic-molecules. *Abstracts of Papers of the American Chemical Society* **1991,** 202, 36-CINF.
92. Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of automatic 3-dimensional model builders using 639 X-ray structures. *Journal of Chemical Information and Computer Sciences* **1994,** 34, 1000-1008.
93. Rusinko, A.; Skell, J.; Balducci, R.; Pearlman, R. CONCORD - rapid generation of high-quality approximate 3-dimensional molecular coordinates. *Abstracts of Papers of the American Chemical Society* **1986,** 192, 12-COMP.
94. Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *Journal of Chemical Information and Modeling* 50, 572-584.
95. Lagorce, D.; Pencheva, T.; Villoutreix, B. O.; Miteva, M. A. DG-AMMOS: a new tool to generate 3d conformation of small molecules using distance geometry and automated molecular mechanics optimization for in silico screening. *BMC Chem Biol* **2009,** 9, 6.
96. Camproux, A.; Gautier, R.; Tuffery, P. A hidden Markov model derived structural alphabet for proteins. *Journal of Molecular Biology* **2004,** 339, 591-605.
97. Maupetit, J.; Derreumaux, P.; Tuffery, P. PEP-FOLD: an online resource for de novo peptide structure prediction. *Nucleic Acids Research* **2009,** 37, W498-W503.
98. Maupetit, J.; Derreumaux, P.; Tufféry, P. A fast method for large-scale de novo peptide and miniprotein structure prediction. *J Comput Chem* **2010,** 31, 726-38.
99. Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *The Journal of chemical physics* **1953,** 21, 1087-1092.
100. Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983,** 220, 671-680.
101. Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry* **1998,** 19, 1639-1662.
102. Jones, G.; Willett, P.; Glen, R.; Leach, A.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* **1997,** 267, 727-48.

103. Grosdidier, A.; Zoete, V.; Michielin, O. EADock: Docking of small molecules into protein active sites with a multiobjective evolutionary optimization. *Proteins-Structure Function and Bioinformatics* **2007,** 67, 1010-1025.
104. Zhao, Y.; Sanner, M. F. FLIPDock: Docking flexible ligands into flexible receptors. *Proteins-Structure Function and Bioinformatics* **2007,** 68, 726-737.
105. Fuhrmann, J.; Rurainski, A.; Lenhof, H. P.; Neumann, D. A new Lamarckian genetic algorithm for flexible ligand-receptor docking. *J Comput Chem* **2010,** 31, 1911-8.
106. Corbeil, C. R.; Englebienne, P.; Moitessier, N. Docking ligands into flexible and solvated macromolecules. 1. Development and validation of FITTED 1.0. *J Chem Inf Model* **2007,** 47, 435-49.
107. Dorigo, M.; Stützle, T. *Ant Colony Optimization*. MIT Press: 2004.
108. Bonabeau, E.; Dorigo, M.; Theraulaz, G. Inspiration for optimization from social insect behaviour. *Nature* **2000,** 406, 39-42.
109. Korb, O.; Stützle, T.; Exner, T. E. Empirical scoring functions for advanced protein-ligand docking with PLANTS. *J Chem Inf Model* **2009,** 49, 84-96.
110. Kennedy, J.; Eberhart, R. In *Particle swarm optimization*, Neural Networks, 1995. Proceedings., IEEE International Conference on, Nov/Dec 1995, 1995; 1995; pp 1942-1948 vol.4.
111. Chen, H. M.; Liu, B. F.; Huang, H. L.; Hwang, S. F.; Ho, S. Y. SODOCK: Swarm optimization for highly flexible protein-ligand docking. *Journal of Computational Chemistry* **2007,** 28, 612-623.
112. Namasivayam, V.; Günther, R. pso@autodock: a fast flexible molecular docking program based on Swarm intelligence. *Chem Biol Drug Des* **2007,** 70, 475-84.
113. Meier, R.; Pippel, M.; Brandt, F.; Sippl, W.; Baldauf, C. ParaDockS: a framework for molecular docking with population-based metaheuristics. *J Chem Inf Model* **2010,** 50, 879-89.
114. Liu, Y.; Zhao, L.; Li, W.; Zhao, D.; Song, M.; Yang, Y. FIPSDock: a new molecular docking technique driven by fully informed swarm optimization algorithm. *J Comput Chem* **2013,** 34, 67-75.
115. Janin, J.; Henrick, K.; Moult, J.; Eyck, L. T.; Sternberg, M. J.; Vajda, S.; Vakser, I.; Wodak, S. J.; Interactions, C. A. o. P. CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins* **2003,** 52, 2-9.
116. Janin, J. Protein-protein docking tested in blind predictions: the CAPRI experiment. *Mol Biosyst* **2010,** 6, 2351-62.
117. Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology* **1982,** 161, 269 - 288.
118. McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian docking functions. *Biopolymers* **2003,** 68, 76-90.
119. McGann, M. FRED pose prediction and virtual screening accuracy. *J Chem Inf Model* **2011,** 51, 578-96.
120. Sauton, N.; Lagorce, D.; Villoutreix, B. O.; Miteva, M. A. MS-DOCK: accurate multiple conformation generator and rigid docking protocol for multi-step virtual ligand screening. *BMC Bioinformatics* **2008,** 9, 184.
121. Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. *Proteins* **1999,** 37, 228-41.
122. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of Medicinal Chemistry* **2004,** 47, 1739-1749.
123. Ewing, T. J.; Makino, S.; Skillman, G. A.; Kuntz, I. D. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *Journal of Computer-Aided Molecular Design* **2001,** 15, 411-428.
124. Jain, A. N. Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *Journal of Medicinal Chemistry* **2003,** 46, 499-511.

125. Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins-Structure Function and Genetics* **2003,** 52, 609-623.
126. Sperandio, O.; Mouawad, L.; Pinto, E.; Villoutreix, B. O.; Perahia, D.; Miteva, M. A. How to choose relevant multiple receptor conformations for virtual screening: a test case of Cdk2 and normal mode analysis. *Eur Biophys J* **2010,** 39, 1365-72.
127. Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. FlexE: Efficient molecular docking considering protein structure variations. *Journal of Molecular Biology* **2001,** 308, 377-395.
128. Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem* **2009,** 30, 2785-91.
129. Meiler, J.; Baker, D. ROSETTALIGAND: Protein-small molecule docking with full side-chain flexibility. *Proteins-Structure Function and Bioinformatics* **2006,** 65, 538-548.
130. Davis, I. W.; Baker, D. RosettaLigand docking with full ligand and receptor flexibility. *J Mol Biol* **2009,** 385, 381-92.
131. Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. Novel procedure for modeling ligand/receptor induced fit effects. *J Med Chem* **2006,** 49, 534-53.
132. Sutherland, J. J.; Nandigam, R. K.; Erickson, J. A.; Vieth, M. Lessons in molecular recognition. 2. Assessing and improving cross-docking accuracy. *J Chem Inf Model* **2007,** 47, 2293-302.
133. Plewczynski, D.; Łaźniewski, M.; Augustyniak, R.; Ginalski, K. Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *J Comput Chem* **2011,** 32, 742-55.
134. Bohm, H. Prediction of binding constants of protein ligands: A fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *Journal of Computer-Aided Molecular Design* **1998,** 12, 309-323.
135. Horvath, D.; Marcou, G.; Varnek, A. Predicting the predictability: a unified approach to the applicability domain problem of QSAR models. *J Chem Inf Model* **2009,** 49, 1762-76.
136. Muegge, I.; Martin, Y. C.; Hajduk, P. J.; Fesik, S. W. Evaluation of PMF scoring in docking weak ligands to the FK506 binding protein. *J Med Chem* **1999,** 42, 2498-503.
137. Muegge, I. PMF scoring revisited. *J Med Chem* **2006,** 49, 5895-902.
138. Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S. A semiempirical free energy force field with charge-based desolvation. *Journal of Computational Chemistry* **2007,** 28, 1145-1152.
139. Plewczynski, D.; Łaźniewski, M.; von Grotthuss, M.; Rychlewski, L.; Ginalski, K. VoteDock: consensus docking method for prediction of protein-ligand interactions. *J Comput Chem* **2011,** 32, 568-81.
140. Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J Comput Aided Mol Des* **1997,** 11, 425-45.
141. Neudert, G.; Klebe, G. DSX: A Knowledge-Based Scoring Function for the Assessment of Protein-Ligand Complexes. *Journal of Chemical Information and Modeling* **2011,** 51, 2731-2745.
142. Nissink, J. W.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A new test set for validating predictions of protein-ligand interaction. *Proteins* **2002,** 49, 457-71.
143. Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J Med Chem* **2007,** 50, 726-41.
144. Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **2004,** 57, 225-42.
145. Deng, Z.; Chuaqui, C.; Singh, J. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J Med Chem* **2004,** 47, 337-44.

146. Yusuf, D.; Davis, A. M.; Kleywegt, G. J.; Schmitt, S. An alternative method for the evaluation of docking performance: RSR vs RMSD. *J Chem Inf Model* **2008,** 48, 1411-22.
147. Erickson, J. A.; Jalaie, M.; Robertson, D. H.; Lewis, R. A.; Vieth, M. Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *J Med Chem* **2004,** 47, 45-55.
148. Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004,** 432, 862-5.
149. Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters* **2006,** 27, 861-874.
150. Oda, A.; Tsuchida, K.; Takakura, T.; Yamaotsu, N.; Hirono, S. Comparison of consensus scoring strategies for evaluating computational models of protein-ligand complexes. *Journal of Chemical Information and Modeling* **2006,** 46, 380-391.
151. Charifson, P.; Corkery, J.; Murcko, M.; Walters, W. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J Med Chem* **1999,** 42, 5100-9.
152. Haider, M. K.; Bertrand, H. O.; Hubbard, R. E. Predicting Fragment Binding Poses Using a Combined MCSS MM-GBSA Approach. *J Chem Inf Model* **2011**.
153. Gleeson, M. P.; Gleeson, D. QM/MM As a Tool in Fragment Based Drug Discovery. A Cross-Docking, Rescoring Study of Kinase Inhibitors. *Journal of Chemical Information and Modeling* **2009,** 49, 1437-1448.
154. Paul, N.; Rognan, D. ConsDock: A new program for the consensus analysis of protein-ligand interactions. *Proteins* **2002,** 47, 521-33.
155. Brewerton, S. C. The use of protein-ligand interaction fingerprints in docking. *Current Opinion in Drug Discovery & Development* **2008,** 11, 356-364.
156. Böhm, H. The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *J Comput Aided Mol Des* **1992,** 6, 61-78.
157. Caflisch, A.; Miranker, A.; Karplus, M. Multiple copy simultaneous search and construction of ligands in binding sites: application to inhibitors of HIV-1 aspartic proteinase. *J Med Chem* **1993,** 36, 2142-67.
158. Mata, p.; Gillet, v.; Johnson, a.; Lampreia, j.; Myatt, g.; Sike, s.; Stebbings, a. SPROUT - 3D structure generation using templates. *Journal of Chemical Information and Computer Sciences* **1995**, 479-493.
159. Wang, R.; Gao, Y.; Lai, L. LigBuilder: A multi-purpose program for structure-based drug design. *Journal of Molecular Modeling* **2000**, 498-516.
160. Stahl, M.; Todorov, N. P.; James, T.; Mauser, H.; Boehm, H. J.; Dean, P. M. A validation study on the practical use of automated de novo design. *J Comput Aided Mol Des* **2002,** 16, 459-78.
161. Schneider, G.; Lee, M.; Stahl, M.; Schneider, P. De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. *J Comput Aided Mol Des* **2000,** 14, 487-94.
162. Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nature Reviews Drug Discovery* **2005,** 4, 649-663.
163. Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP - Retrosynthetic combinatorial analysis procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *Journal of Chemical Information and Computer Sciences* **1998,** 38, 511-522.
164. Hartenfeller, M.; Eberle, M.; Meier, P.; Nieto-Oberhuber, C.; Altmann, K.; Schneider, G.; Jacoby, E.; Renner, S. A Collection of Robust Organic Synthesis Reactions for In Silico Molecule Design. *Journal of Chemical Information and Modeling* **2011,** 51, 3093-3098.
165. Bode, J. Reactor. ChemAxon Ltd., Maramaros koz 2/a, Budapest, 1037 Hungary. www.chemaxon.com. Contact ChemAxon for pricing information. *Journal of the American Chemical Society* **2004,** 126, 15317-15317.

166. Vinkers, H.; de Jonge, M.; Daeyaert, F.; Heeres, J.; Koymans, L.; van Lenthe, J.; Lewi, P.; Timmerman, H.; Van Aken, K.; Janssen, P. SYNOPSIS: SYNthesize and OPtimize System in Silico. *J Med Chem* **2003,** 46, 2765-73.
167. Hartenfeller, M.; Zettl, H.; Walter, M.; Rupp, M.; Reisen, F.; Proschak, E.; Weggen, S.; Stark, H.; Schneider, G. DOGS: reaction-driven de novo design of bioactive compounds. *PLoS Comput Biol* **2012,** 8, e1002380.
168. Zaliani, A.; Boda, K.; Seidel, T.; Herwig, A.; Schwab, C. H.; Gasteiger, J.; Claussen, H.; Lemmen, C.; Degen, J.; Parn, J.; Rarey, M. Second-generation de novo design: a view from a medicinal chemist perspective. *Journal of Computer-Aided Molecular Design* **2009,** 23, 593-602.
169. http://freepascal.org. In.
170. http://lazarus.freepascal.org. In.
171. http://sourceforge.net/projects/dmath. In.
172. Dauber-Osguthorpe, P.; Roberts, V. A.; Osguthorpe, D. J.; Wolff, J.; Genest, M.; Hagler, A. T. Structure and energetics of ligand binding to proteins: Escherichia coli dihydrofolate reductase-trimethoprim, a drug-receptor system. *Proteins: Structure, Function, and Genetics* **1988,** 4, 31-47.
173. http://accelrys.com. In.
174. Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *Journal of Computational Chemistry* **2004,** 25, 1157-1174.
175. Horvath, D.; Brillet, L.; Roy, S.; Conilleau, S.; Tantar, A. A.; Boisson, J. C.; Melab, N.; Talbi, E. G.; Ieee. Local vs. Global Search Strategies in Evolutionary GRID-based Conformational Sampling & Docking. In *2009 IEEE Congress on Evolutionary Computation, Vols 1-5*, IEEE: New York, 2009; pp 247-254.
176. Parent, B.; Koekoesy, A.; Horvath, D. Optimized evolutionary strategies in conformational sampling. *Soft Computing* **2007,** 11, 63-79.
177. http://ambermd.org. In.
178. Wang, J. M.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics & Modelling* **2006,** 25, 247-260.
179. Horvath, D. A virtual screening approach applied to the search for trypanothione reductase inhibitors. *Journal of Medicinal Chemistry* **1997,** 40, 2412-2423.
180. Ferrara, P.; Gohlke, H.; Price, D.; Klebe, G.; Brooks, C. r. Assessing scoring functions for protein-ligand interactions. *J Med Chem* **2004,** 47, 3032-47.
181. Mazur, J.; Jernigan, R. Distance-dependent dielectric-constants and their application to double-helical DNA. *Biopolymers* **1991,** 31, 1615-1629.
182. Horvath, D.; Lippens, G.; vanBelle, D. Development and parametrization of continuum solvent models .2. A unified approach to the solvation problem. *Journal of Chemical Physics* **1996,** 105, 4197-4210.
183. Gilson, M. K.; Honig, B. The inclusion of electrostatic hydration energies in molecular mechanics calculations. *Journal of Computer-Aided Molecular Design* **1991,** 5, 5-20.
184. Horvath, D.; Chira, C. Simplified Chain Folding Models as Metaheuristic Benchmark for Tuning Real Protein Folding Algorithms? In *2010 IEEE Congress on Evolutionary Computation*, IEEE: Barcelona, 2010; pp 1-8.
185. Hoffer, L.; Chira, C.; Marcou, G.; Varnek, A.; Horvath, D. S4MPLE - Sampler For Multiple Protein-Ligand Entities: Methodology & Rigid-Site Docking Benchmarking. *Journal of Cheminformatics* **2013,** submitted for publication.
186. Li, H.; Li, C. Multiple ligand simultaneous docking: orchestrated dancing of ligands in binding sites of protein. *J Comput Chem* **2010,** 31, 2014-22.
187. http://www.chemaxon.com. In.
188. Irwin, J. J.; Shoichet, B. K. ZINC - A free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling* **2005,** 45, 177-182.

189. Bemis, G.; Murcko, M. The properties of known drugs. 1. Molecular frameworks. *J Med Chem* **1996,** 39, 2887-93.
190. Gasteiger, J.; Marsilli, M. A New Model for Calculating Atomic Charges in Molecules. *Tetrahedron Lett.* **1978**, 3181-3184.
191. Legrand, S.; Merz, K. The application of the genetic algorithm to the minimization of potential-energy functions. *Journal of Global Optimization* **1993,** 3, 49-66.
192. DeLano, W. L. *The PyMOL Molecular Graphics System*, DeLano Scientific: San Carlos, CA, USA, 2002.
193. Cer, R.; Mudunuri, U.; Stephens, R.; Lebeda, F. IC50-to-K-i: a web-based tool for converting IC50 to K-i values for inhibitors of enzyme activity and ligand binding. *Nucleic Acids Research* **2009,** 37, W441-W445.
194. Sotriffer, C. A.; Sanschagrin, P.; Matter, H.; Klebe, G. SFCscore: scoring functions for affinity prediction of protein-ligand complexes. *Proteins* **2008,** 73, 395-419.
195. Verdonk, M. L.; Giangreco, I.; Hall, R. J.; Korb, O.; Mortenson, P. N.; Murray, C. W. Docking performance of fragments and druglike compounds. *J Med Chem* **2011,** 54, 5422-31.
196. Barker, J. J.; Barker, O.; Courtney, S. M.; Gardiner, M.; Hesterkamp, T.; Ichihara, O.; Mather, O.; Montalbetti, C. A.; Müller, A.; Varasi, M.; Whittaker, M.; Yarnold, C. J. Discovery of a novel Hsp90 inhibitor by fragment linking. *ChemMedChem* **2010,** 5, 1697-700.
197. Hoffer, L.; Horvath, D. S4MPLE - Sampler For Multiple Protein-Ligand Entities: Simultaneous Docking of Several Entities. *J Chem Inf Model* **2012,** 53, 88-102.
198. Berthold, M.; Cebron, N.; Dill, F.; Gabriel, T.; Kotter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B.; Preisach, C.; Burkhardt, H.; SchmidtThieme, L.; Decker, R. KNIME: The Konstanz Information Miner. *Data Analysis, Machine Learning and Applications* **2008**, 319-326.
199. Hoffer, L.; Renaud, J.; Horvath, D. In silico Fragment-Based Drug Discovery: setup and validation of a fragment-to-lead computational protocol using S4MPLE. *J Chem Inf Model* **2013,** accepted for publication.
200. Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is there a difference between leads and drugs? A historical perspective. *J Chem Inf Comput Sci* **2001,** 41, 1308-15.
201. Lagorce, D.; Sperandio, O.; Galons, H.; Miteva, M. A.; Villoutreix, B. O. FAF-Drugs2: free ADME/tox filtering tool to assist drug discovery and chemical biology projects. *BMC Bioinformatics* **2008,** 9, 396.
202. Pinto, D. J.; Smallheer, J. M.; Cheney, D. L.; Knabb, R. M.; Wexler, R. R. Factor Xa inhibitors: next-generation antithrombotic agents. *J Med Chem* **2010,** 53, 6243-74.
203. Dai, C.; Whitesell, L. HSP90: a rising star on the horizon of anticancer targets. *Future Oncol* **2005,** 1, 529-40.
204. Sixma, T. K.; Smit, A. B. Acetylcholine binding protein (AChBP): a secreted glial protein that provides a high-resolution model for the extracellular domain of pentameric ligand-gated ion channels. *Annu Rev Biophys Biomol Struct* **2003,** 32, 311-34.
205. Grimster, N. P.; Stump, B.; Fotsing, J. R.; Weide, T.; Talley, T. T.; Yamauchi, J. G.; Nemecz, Á.; Kim, C.; Ho, K. Y.; Sharpless, K. B.; Taylor, P.; Fokin, V. V. Generation of candidate ligands for nicotinic acetylcholine receptors via in situ click chemistry with a soluble acetylcholine binding protein template. *J Am Chem Soc* **2012,** 134, 6732-40.
206. Nazaré, M.; Matter, H.; Will, D. W.; Wagner, M.; Urmann, M.; Czech, J.; Schreuder, H.; Bauer, A.; Ritter, K.; Wehner, V. Fragment deconstruction of small, potent factor Xa inhibitors: exploring the superadditivity energetics of fragment linking in protein-ligand complexes. *Angew Chem Int Ed Engl* **2012,** 51, 905-11.
207. Edink, E.; Rucktooa, P.; Retra, K.; Akdemir, A.; Nahar, T.; Zuiderveld, O.; van Elk, R.; Janssen, E.; van Nierop, P.; van Muijlwijk-Koezen, J.; Smit, A. B.; Sixma, T. K.; Leurs, R.; de Esch, I. J. Fragment Growing Induces Conformational Changes in Acetylcholine-Binding Protein: A Structural and Thermodynamic Analysis. *J Am Chem Soc* **2011**.

208. McCourt, J. A.; Pang, S. S.; Guddat, L. W.; Duggleby, R. G. Elucidating the specificity of binding of sulfonylurea herbicides to acetohydroxyacid synthase. *Biochemistry* **2005,** 44, 2330-8.
209. Jorgensen, W.; Chandrasekhar, J.; Madura, J.; Impey, R.; Klein, M. Comparison Of Simple Potential Functions For Simulating Liquid Water. *Journal of Chemical Physics* **1983,** 79, 926-935.
210. Janin, Y. L. Heat shock protein 90 inhibitors. A text book example of medicinal chemistry? *J Med Chem* **2005,** 48, 7503-12.
211. Apweiler, R.; Bairoch, A.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Natale, D. A.; O'Donovan, C.; Redaschi, N.; Yeh, L. S. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **2004,** 32, D115-9.
212. Hendrick, J. P.; Hartl, F. U. Molecular chaperone functions of heat-shock proteins. *Annu Rev Biochem* **1993,** 62, 349-84.
213. Richter, K.; Buchner, J. Hsp90: chaperoning signal transduction. *J Cell Physiol* **2001,** 188, 281-90.
214. Xu, W.; Neckers, L. Targeting the molecular chaperone heat shock protein 90 provides a multifaceted effect on diverse cell signaling pathways of cancer cells. *Clin Cancer Res* **2007,** 13, 1625-9.
215. Barker, J. J.; Barker, O.; Boggio, R.; Chauhan, V.; Cheng, R. K. Y.; Corden, V.; Courtney, S. M.; Edwards, N.; Falque, V. M.; Fusar, F.; Gardiner, M.; Hamelin, E. M. N.; Hesterkamp, T.; Ichihara, O.; Jones, R. S.; Mather, O.; Mercurio, C.; Minucci, S.; Montalbetti, C.; Muller, A.; Patel, D.; Phillips, B. G.; Varasi, M.; Whittaker, M.; Winkler, D.; Yarnold, C. J. Fragment-based Identification of Hsp90 Inhibitors. *Chemmedchem* **2009,** 4, 963-966.
216. Pearl, L. H.; Prodromou, C. Structure and mechanism of the Hsp90 molecular chaperone machinery. *Annu Rev Biochem* **2006,** 75, 271-94.
217. Dymock, B.; Barril, X.; Brough, P.; Cansfield, J.; Massey, A.; McDonald, E.; Hubbard, R.; Surgenor, A.; Roughley, S.; Webb, P.; Workman, P.; Wright, L.; Drysdale, M. Novel, potent small-molecule inhibitors of the molecular chaperone Hsp90 discovered through structure-based design. *J Med Chem* **2005,** 48, 4212-5.
218. van Montfort, R. L. M.; Workman, P. Structure-based design of molecular cancer therapeutics. *Trends in Biotechnology* **2009,** 27, 315-328.
219. Huth, J. R.; Park, C.; Petros, A. M.; Kunzer, A. R.; Wendt, M. D.; Wang, X. L.; Lynch, C. L.; Mack, J. C.; Swift, K. M.; Judge, R. A.; Chen, J.; Richardson, P. L.; Jin, S.; Tahir, S. K.; Matayoshi, E. D.; Dorwin, S. A.; Ladror, U. S.; Severin, J. M.; Walter, K. A.; Bartley, D. M.; Fesik, S. W.; Elmore, S. W.; Hajduk, P. J. Discovery and design of novel HSP90 inhibitors using multiple fragment-based design strategies. *Chemical Biology & Drug Design* **2007,** 70, 1-12.
220. Murray, C. W.; Carr, M. G.; Callaghan, O.; Chessari, G.; Congreve, M.; Cowan, S.; Coyle, J. E.; Downham, R.; Figueroa, E.; Frederickson, M.; Graham, B.; McMenamin, R.; O'Brien, M. A.; Patel, S.; Phillips, T. R.; Williams, G.; Woodhead, A. J.; Woolford, A. J. Fragment-based drug discovery applied to Hsp90. Discovery of two lead series with high ligand efficiency. *J Med Chem* **2010,** 53, 5942-55.
221. Woodhead, A. J.; Angove, H.; Carr, M. G.; Chessari, G.; Congreve, M.; Coyle, J. E.; Cosme, J.; Graham, B.; Day, P. J.; Downham, R.; Fazal, L.; Feltell, R.; Figueroa, E.; Frederickson, M.; Lewis, J.; McMenamin, R.; Murray, C. W.; O'Brien, M. A.; Parra, L.; Patel, S.; Phillips, T.; Rees, D. C.; Rich, S.; Smith, D. M.; Trewartha, G.; Vinkovic, M.; Williams, B.; Woolford, A. J. Discovery of (2,4-dihydroxy-5-isopropylphenyl)-[5-(4-methylpiperazin-1-ylmethyl)-1,3-dihydroisoindol-2-yl]methanone (AT13387), a novel inhibitor of the molecular chaperone Hsp90 by fragment based drug design. *J Med Chem* **2010,** 53, 5956-69.
222. MOE. Molecular Operating Environment. *http://www.chemcomp.com* **Chemical Computing Group Inc**.

Annexes

Annexe 1 : Fichier de configuration XML utilisé pour la fragmentation RECAP (API ChemAxon).

```
<?xml version="1.0" encoding="UTF-8"?>

<FragmenterConfiguration Version ="0.1" schemaLocation="fragment_schema.xsd">

 <Fragmenter>
  <Actions>
   <Action ID="amide" Structure="[O:3]=[C!$(C([#7])(=O)[!#1!#6]):2]-[#7!$([#7][!#1!#6]):1]>>[O:3]=[C:2].[#7:1]"/>
   <Action ID="ester" Structure="[#6!$([#6](O)~[!#1!#6])][O:2][C:1]=O>>[C:1]=O.[#6][O:2]"/>
   <Action ID="amine" Structure="[#6:2]-[N!$(N[#6]=[!#6])!$(N~[!#1!#6])!X4:1]>>[N:1].[#6:2]"/>
   <Action ID="urea" Structure="N[C:1]([N:2])=O>>N[C:1]=O.[N:2]"/>
   <Action ID="ether" Structure="[#6]-[O!$(O[#6]~[!#1!#6]):1]-[#6:2]>>[#6:2].[O:1]-[#6]"/>
   <Action ID="olefin" Structure="[C:1]=[C:1]>>[C:1].[C:1]"/>
   <Action ID="quatN" Structure="[#6:1]-[N$(N([#6])([#6])([#6])[#6])!$(NC=[!#6]):2]>>[#6:1].[N:2]"/>
   <Action ID="aromN-carbon" Structure="[n:1]-[#6!$([#6]=[!#6]):2]>>[n:1].[#6:2]"/>
   <Action ID="lactamN-carbon" Structure="[C:3](=[O:4])@-[N:1]!@-[#6!$([#6]=[!#6]):2]>>[C:3](=[O:4])[N:1].[#6:2]"/>
   <Action ID="aromc-aromc" Structure="[c:1]-[c:1]>>[c:1].[c:1]"/>
   <Action ID="sulphonamide" Structure="[#7:1][S:2](=O)=O>>[#7:1].[S:2](=O)=O"/>
  </Actions>
  <Params>
   <SDFTags Uid="UID" CutIds="REACTIONS" CutCounts="COUNTS" CutSum="SUM" Count="COUNT"
          FragmentSets="FRAGMENTSETS"/>
   <Fragmentation Extensive="true"/>
  </Params>
 </Fragmenter>

 <Reviser>
  <Recap Class="chemaxon.fragmenter.Recap">
   <Params>
    <Limits MaxCutCount="2" MinAtomCount="1"/>
    <Options CutRingCHetero="true"/>
   </Params>
  </Recap>
 </Reviser>

</FragmenterConfiguration>
```

Iconographie

Sauf mention contraire, les représentations des molécules en 2D et 3D ont été réalisées respectivement à l'aide des programmes MarvinView [187] et Pymol [192].

Table des figures

Table des tableaux

Table des annexes

Laurent HOFFER

Développement et validation du logiciel S4MPLE. Application au docking moléculaire et à l'optimisation de fragments assistée par ordinateur dans le cadre du FBDD.

Ce travail de thèse a pour but de développer le pendant *in silico* des étapes clés du Fragment-Based Drug Design (FBDD), et ce dans le cadre plus général du développement de l'outil de modélisation moléculaire S4MPLE. En un mot, le FBDD génère des ligands "drug-like" à partir de petites molécules organiques (fragments). Après une étape de validation de S4MPLE et de sa fonction d'énergie, un recentrage autour du FBDD est réalisé, à travers le docking moléculaire puis l'optimisation virtuelle de fragments en mimant les méthodes d'optimisation de type growing et linking. Cette stratégie repose entre autre sur 1) la création d'une chimiothèque focalisée en connectant un ou deux fragment(s) avec des linkers pré-générés, et 2) l'échantillonnage avec S4MPLE des composés chimères dans le site avec certaines contraintes. Des simulations de growing et de linking plus ou moins ambitieuses (site flexible, ajout de H2O libres) permettent de valider cette approche avec plusieurs études rétrospectives basées sur des données expérimentales de la littérature. La dernière phase de la thèse a consisté à appliquer ce protocole *in silico* à un projet interne de l'entreprise.

Mots clés: FBDD, échantillonnage conformationnel, docking, algorithme génétique, champ de force.

This work aims to develop *in silico* methods targeting the key stages of Fragment-Based Drug Design (FBDD), participating to the development of the molecular modeling tool S4MPLE. Briefly, FBDD generates "drug-like" ligands from small organic molecules called fragments. After a validation step of S4MPLE and its energy function, the work focused on FBDD: molecular docking of fragments and their subsequent virtual optimization. The latter mimics standard evolution strategies in FBDD (growing and linking). This *in silico* approach involves among other two key stages 1) building of a focused library by plugging in pre-generated linkers into reference fragments using rules and 2) sampling of these new compounds under atomic and binding site constraints. Validation simulations, relying on known experimental data, included "classical" growing/linking and more challenging ones (site flexibility, free waters). Finally, this strategy is applied to one project of the company.

Keywords: FBDD, conformational sampling, docking, genetic algorithm, force field.