



UNIL | Université de Lausanne

Faculté de biologie  
et de médecine

# **EADock: design of a new molecular docking algorithm and some of its applications.**

**Thèse de doctorat ès science de la vie (PhD)**

présentée à la

Faculté de biologie et de médecine  
de l'Université de Lausanne

par

**Aurélien Grosdidier**

Diplômé de l'UFR de Pharmacie, Université Joseph Fourier, Grenoble, France

**Jury**

Prof. Yves Poirier, Rapporteur  
Prof. Olivier Michielin, Directeur de thèse  
Dr. Vincent Zoete, Expert Interne  
Dr. Roland Stote, Expert Externe

Lausanne, 2007





UNIL | Université de Lausanne

Faculté de biologie  
et de médecine

Ecole Doctorale

# Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

<i>Président</i>	Monsieur Prof.	Yves <b>Poirier</b>
<i>Directeur de thèse</i>	Monsieur Prof.	Olivier <b>Michielin</b>
<i>Rapporteur</i>	Monsieur Prof.	Yves <b>Poirier</b>
<i>Experts</i>	Monsieur Dr	Vincent <b>Zoete</b>
	Monsieur Dr	Roland <b>Stote</b>

le Conseil de Faculté autorise l'impression de la thèse de

**Monsieur Aurélien Grosdidier**

titulaire d'un DEA en bioinformatique de l'Université de Genève

intitulée

**EADock : design of a new molecular docking algorithm  
and some of its applications**

Lausanne, le 10 mai 2007

pour Le Doyen  
de la Faculté de Biologie et de Médecine

Prof. Yves Poirier





*À Celui qui,  
alors que nous ne pouvons qu'espérer les prolonger,  
offre un sens à nos vies.*



# 1 Table of content

1 Table of content.....	7
2 Acknowledgments.....	13
3 Summary.....	15
3.1 English.....	15
3.2 French.....	16
4 Abbreviation.....	19
5 Introduction.....	21
5.1 Toward new drugs.....	21
5.2 The docking problem.....	24
5.2.1 Statement.....	24
5.2.2 Theoretical problems and methodological answers.....	26
5.2.2.1 Sampling heuristics.....	27
5.2.2.2 Scoring functions.....	30
5.2.3 Modeling molecular interactions.....	31
5.3 Molecular docking: state of the art.....	34
5.3.1 Common implementations.....	34
5.3.1.1 AutoDock.....	34
5.3.1.2 Gold.....	34
5.3.1.3 FlexX.....	35
5.3.1.4 DOCK and ICM.....	35
5.3.2 Benchmarking benchmarks.....	35
5.3.2.1 Experts.....	36
5.3.2.2 Test sets.....	36
5.3.2.3 Comparable and large search space.....	36
5.3.2.4 Comparable seeding.....	37
5.3.2.5 Number of evaluated poses and CPU time.....	37
5.3.2.6 Outcome.....	37
5.3.3 Performance Benchmarks.....	37
5.4 Docking challenges.....	38

5.4.1 Scoring functions.....	39
5.4.1.1 Usual methods evaluating the binding free energy.....	39
5.4.1.2 Accounting for the entropy.....	40
5.4.1.3 Role of water molecules.....	40
5.4.1.4 Flexibility of the protein.....	40
5.4.1.5 Reproducibility issue.....	41
5.4.2 Sampling heuristics.....	41
5.5 Two applications of docking in drug design.....	41
5.5.1 Two impacts of docking softwares.....	42
5.5.2 Virtual screening.....	43
5.5.2.1 Overview.....	43
5.5.2.2 State of the art.....	44
5.5.3 Fragment-based rational drug design.....	45
5.5.3.1 Overview.....	45
5.5.3.2 State of the art.....	49
5.5.3.3 Conclusion.....	52
5.6 General conclusion.....	52
6 Presentation and benchmark of EADock.....	55
6.1 Docking algorithm.....	56
6.1.1 Seeding.....	58
6.1.2 Selection.....	58
6.1.3 Diversity.....	61
6.1.4 Postprocessing.....	63
6.2 Dataset.....	63
6.3 Algorithm assessment and benchmark.....	65
6.4 Algorithm performance.....	67
6.4.1 Algorithm assessment.....	67
6.4.2 Benchmarks.....	76
6.5 Discussion.....	79
6.5.1 Benchmarking docking algorithms.....	79
6.5.2 Our approach.....	80

7 Applications.....	85
7.1 Overview.....	85
7.2 Understanding molecular principles.....	87
7.2.1 Regulation of the Na,K-ATPase.....	87
7.2.1.1 Biological context.....	87
7.2.1.2 Modeling approach.....	88
7.2.1.3 Results.....	89
7.2.1.4 Conclusion.....	91
7.2.2 Regulation of the nuclear hormone receptor PPAR $\alpha$ .....	91
7.2.2.1 Biological context.....	91
7.2.2.2 Modeling approach.....	93
7.2.2.3 Results.....	95
7.2.2.4 Conclusion.....	96
7.3 Understanding the action of known compounds.....	97
7.3.1 Docking of DEHP/MEHP on the nuclear hormone receptor PPAR $\gamma$ .....	97
7.3.1.1 Biological context.....	97
7.3.1.2 Modeling approach.....	99
7.3.1.3 Results.....	99
7.3.1.4 Conclusion.....	99
7.3.2 Impact of the biotransformation of the Imatinib on its binding mode.....	101
7.3.2.1 Biological context.....	101
7.3.2.2 Modeling approach.....	102
7.3.2.3 Results.....	104
7.4 Lead discovery and optimization.....	107
7.4.1 Material and Methods.....	107
7.4.1.1 Overview.....	107
7.4.1.2 EADock.....	109
7.4.1.3 Forcefield.....	109
7.4.2 Targeting the integrin.....	110
7.4.2.1 Biological context.....	110
7.4.2.2 Assessment of EADock.....	111

7.4.2.3 Design of peptide inhibitors of $\alpha 5\beta 1$ .....	112
7.4.2.4 Conclusion.....	116
7.4.3 Design of peptidic PPAR $\alpha$ ligands.....	116
7.4.3.1 Biological context.....	117
7.4.3.2 Assessment of EADock.....	117
7.4.3.3 Design of peptide ligands of hPPAR $\alpha$ .....	117
7.4.3.4 Conclusion.....	120
7.4.4 Targeting the indoleamine deoxygenase.....	121
7.4.4.1 Biological context.....	121
7.4.4.2 Modeling approach.....	124
7.4.4.3 Results.....	125
7.4.4.4 Design of new inhibitors.....	131
7.4.4.5 Conclusion.....	131
8 Perspectives.....	135
8.1 Improvements.....	136
8.1.1 Performance improvement.....	136
8.1.1.1 Scoring.....	136
8.1.1.2 Sampling.....	139
8.1.2 Usability.....	146
8.1.2.1 Input.....	147
8.1.2.2 Speed.....	149
8.1.2.3 Output.....	151
8.2 Conclusion.....	154
9 Conclusion.....	157
10 Bibliography.....	159







## 2 Acknowledgments

I would like to warmly thank my supervisor Prof. Olivier Michielin for giving me the opportunity to join his group, for his support and unbreakable trust whatever the scientific and technical challenges. A big special thank to my co-supervisor and friend Vincent Zoete, whose contribution to this PhD thesis is enormous. It has been a real pleasure to work with such a talented, inexhaustible and upright person. A whole lot of thanks to all the present and past group members, by order of appearance: Theres Fagerberg, Pierre Chodanowski, Antoine Leimgruber, Michel Cuendet, Hamid Hussain-Kahn, Simon Bernèche, Ute Röhrig, Justyna Iwaszkiewicz, Mathias Ferber and Thierry Schüpbach. Thanks for all the discussions, support, great moments, and never-ending preparation of manuscripts. It has been a great pleasure working with you all.

This work would not have been possible without the VITAL-IT team of the Swiss Institute of Bioinformatics, and the Cluster versus Cancer Project for providing the computational resources. EADock is a great tool, although a bit hungry sometimes. Thank you all.

Online communities have been very helpful during these four years, and I would like to thank the DLFP team for its extraordinary responsiveness to anything, including real questions, trolls, and FUDs.



## 3 Summary

### 3.1 English

The pharmaceutical industry has been facing several challenges during the last years, and the optimization of their drug discovery pipeline is believed to be the only viable solution. High-throughput techniques do participate actively to this optimization, especially when complemented by computational approaches aiming at rationalizing the enormous amount of information that they can produce. *In silico* techniques, such as virtual screening or rational drug design, are now routinely used to guide drug discovery. Both heavily rely on the prediction of the molecular interaction (docking) occurring between drug-like molecules and a therapeutically relevant target. Several softwares are available to this end, but despite the very promising picture drawn in most benchmarks, they still hold several hidden weaknesses. As pointed out in several recent reviews, the docking problem is far from being solved, and there is now a need for methods able to identify binding modes with a high accuracy, which is essential to reliably compute the binding free energy of the ligand. This quantity is directly linked to its affinity and can be related to its biological activity. Accurate docking algorithms are thus critical for both the discovery and the rational optimization of new drugs.

In this thesis, a new docking software aiming at this goal is presented, EADock. It uses a hybrid evolutionary algorithm with two fitness functions, in combination with a sophisticated management of the diversity. EADock is interfaced with the CHARMM package for energy calculations and coordinate handling. A validation was carried out on 37 crystallized protein-ligand complexes featuring 11 different proteins. The search space was defined as a sphere of 15 Å around the center of mass of the ligand position in the crystal structure, and conversely to other benchmarks, our algorithm was fed with optimized ligand positions up to 10 Å root mean square deviation (RMSD) from the crystal structure. This validation illustrates the efficiency of our sampling heuristic, as correct binding modes, defined by a RMSD to the crystal structure lower than 2 Å, were identified and ranked first for 68% of the complexes. The success rate increases to 78% when

considering the five best-ranked clusters, and 92% when all clusters present in the last generation are taken into account. Most failures in this benchmark could be explained by the presence of crystal contacts in the experimental structure.

EADock has been used to understand molecular interactions involved in the regulation of the Na,K-ATPase, and in the activation of the nuclear hormone peroxisome proliferator-activated receptors  $\alpha$  (PPAR $\alpha$ ). It also helped to understand the action of common pollutants (phthalates) on PPAR $\gamma$ , and the impact of biotransformations of the anticancer drug Imatinib (Gleevec®) on its binding mode to the Bcr-Abl tyrosine kinase. Finally, a fragment-based rational drug design approach using EADock was developed, and led to the successful design of new peptidic ligands for the  $\alpha 5\beta 1$  integrin, and for the human PPAR $\alpha$ . In both cases, the designed peptides presented activities comparable to that of well-established ligands such as the anticancer drug Cilengitide and Wy14,643, respectively.

### 3.2 French

Les récentes difficultés de l'industrie pharmaceutique ne semblent pouvoir se résoudre que par l'optimisation de leur processus de développement de médicaments. Cette dernière implique de plus en plus de techniques dites "haut-débit", particulièrement efficaces lorsqu'elles sont couplées aux outils informatiques permettant de gérer la masse de données produite. Désormais, les approches *in silico* telles que le criblage virtuel ou la conception rationnelle de nouvelles molécules sont utilisées couramment. Toutes deux reposent sur la capacité à prédire les détails de l'interaction moléculaire entre une molécule ressemblant à un principe actif (PA) et une protéine cible ayant un intérêt thérapeutique. Les comparatifs de logiciels s'attaquant à cette prédiction sont flatteurs, mais plusieurs problèmes subsistent. La littérature récente tend à remettre en cause leur fiabilité, affirmant l'émergence d'un besoin pour des approches plus précises du mode d'interaction. Cette précision est essentielle au calcul de l'énergie libre de liaison, qui est directement liée à l'affinité du PA potentiel pour la protéine cible, et indirectement liée à son activité biologique. Une prédiction précise est d'une importance toute particulière pour la découverte et l'optimisation de nouvelles molécules actives.

Cette thèse présente un nouveau logiciel, EADock, mettant en avant une telle précision. Cet algorithme évolutionnaire hybride utilise deux pressions de sélections, combinées à une gestion de la diversité sophistiquée. EADock repose sur CHARMM pour les calculs d'énergie et la gestion des coordonnées atomiques. Sa validation a été effectuée sur 37 complexes protéine-ligand cristallisés, incluant 11 protéines différentes. L'espace de recherche a été étendu à une sphère de 15 Å de rayon autour du centre de masse du ligand cristallisé, et contrairement aux comparatifs habituels, l'algorithme est parti de solutions optimisées présentant un RMSD jusqu'à 10 Å par rapport à la structure cristalline. Cette validation a permis de mettre en évidence l'efficacité de notre heuristique de recherche car des modes d'interactions présentant un RMSD inférieur à 2 Å par rapport à la structure cristalline ont été classés premier pour 68% des complexes. Lorsque les cinq meilleures solutions sont prises en compte, le taux de succès grimpe à 78%, et 92% lorsque la totalité de la dernière génération est prise en compte. La plupart des erreurs de prédiction sont imputables à la présence de contacts cristallins.

Depuis, EADock a été utilisé pour comprendre les mécanismes moléculaires impliqués dans la régulation de la Na,K-ATPase et dans l'activation du peroxisome proliferator-activated receptor  $\alpha$  (PPAR $\alpha$ ). Il a également permis de décrire l'interaction de polluants couramment rencontrés sur PPAR $\gamma$ , ainsi que l'influence de la métabolisation de l'Imatinib (PA anticancéreux) sur la fixation à la kinase Bcr-Abl. Une approche basée sur la prédiction des interactions de fragments moléculaires avec protéine cible est également proposée. Elle a permis la découverte de nouveaux ligands peptidiques de PPAR $\alpha$  et de l'intégrine  $\alpha 5\beta 1$ . Dans les deux cas, l'activité de ces nouveaux peptides est comparable à celles de ligands bien établis, comme le Wy14,643 pour le premier, et le Cilengitide (PA anticancéreux) pour la seconde.



## 4 Abbreviation

3D	Three Dimensional
ABNR	Adopted Basis Newton-Raphson
API	Application Programming Interface
CAPRI	Critical Assessment of Prediction of Interactions
CHARMM	Chemistry at HARvard Molecular Mechanics
CPU	Central Processing Unit
DoF	Degree of Freedom
EA	Evolutionary Algorithm
EADock	Evolutionary Algorithm for Docking
FB-RDD	Fragment-Based Rational Drug Design
FDA	Food and Drug Administration
GB-MV2	Generalized Born using Molecular Volume, analytical method 2
GBSA	Generalized Born Surface Area
HTS	High Throughput Screening
kcal	Kilocalorie
LE	Ligand Efficiency
MD	Molecular Dynamics
MM	Molecular Mechanics
MMFF	Merck Molecular Force Field
MW	Molecular Weight
NMR	Nuclear Magnetic Resonance
PB	Poisson-Boltzmann
PDB	Protein Data Bank
RDD	Rational Drug Design
RMSD	Root Mean Square Deviation
ROI	Region Of Interest
SASA	Solvent Accessible Surface Area
SD	Steepest Descent
vdW	van der Waals
VS	Virtual Screening





## 5 Introduction

While the questions about the origin and purpose of life are very personal and have no clear social answers, the pragmatic consequences of life, among which are the ineluctable competition with other living forms and the fight against diseases and death, were always addressed by human societies in order to survive. Science led to historical successes resulting in a jump of life expectancy, for instance with the discovery of antibiotics. Many now think that the worst enemy of human beings are human themselves. It is certainly true that our personal behaviors can impact our health badly. This effect can be direct, e.g. people with a sedentary lifestyle are more at risk than others to develop cardiovascular diseases. It can also be indirect, through toxic compounds that were released around us, such as asbestos fibers causing lung cancers. While it is certainly worth focusing on our own behavior, such a human-centric point of view should not hide the dynamic of life itself, in which death is unavoidable. Even the benefit of a longer life expectancy comes at the price, for instance, of an increase of the incidence of neurodegenerative diseases. Similarly, each fight won with antibiotics remains fragile, as resistant bacterial strains are already a major concern, especially in hospitals.

More efficient medicines are certainly required to face these challenges, as well as possible yet still unknown threats in the coming years, keeping in mind that the armament race against death is already lost, focusing on diseases with humility, with the hope to make life better and possibly longer.

### 5.1 *Toward new drugs*

The development of a new drug, from the identification of a biological target to the patient, is a multi-step process that can be roughly split into pre-clinical development and clinical trials (see Table 1).

The goal pursued during the pre-clinical development is to identify and optimize a lead compound regarding to a targeted biological activity. This requires the investigation of its pharmacodynamics and pharmacokinetics properties, and often starts with *in silico* assay,

	Preclinical development	Clinical trial			Pharmacovigilance
		Phase I	Phase II	Phase III	Phase IV
Years	3.5	1	2	3	Additional post marketing testing required by FDA
Test Population	Laboratory and animal studies	20 to 80 healthy volunteers	100 to 300 patient volunteers	1000 to 3000 patient volunteers	
Purpose	Assess safety and biological activity	Determine safety and dosage	Evaluate effectiveness, look for side effects	Verify effectiveness, monitor adverse reactions from long-term use	
Success Rate	5,000 compounds evaluated	5 enter trials			

Table 1: developing drugs: a multi step process. Adapted from Dale E. Wierenga and C. Rovert Eaton, Office of Research and Development, [http://www.allp.com/drug\\_dev.htm](http://www.allp.com/drug_dev.htm)

which are then pushed *in vitro* and ultimately *in vivo* in cells and model organisms.

Once the lead compound reaches a satisfying activity and safety, and once its galenic formulation is defined, the three phases of clinical trials successively starts. Their first goal is to check that the compound is not harmful for humans (phase 1). The dose-effect relation is then deeply investigated (phase II), before the definitive assessment of the drug on 1000 to 3000 volunteers patients (phase III). Once on the market, the pharmacovigilance stage (phase IV) starts, aiming at the detection, assessment, understanding and prevention of adverse effects, particularly long term and short-term side effect.

Taking about twelve years, the development of a new drug is increasingly costly, and was estimated to USD \$359 in 1993 as reported by the Congressional Office of Technology Assessment<sup>1</sup>, USD \$897 millions in the late 90s [1]. Last year, a study of the Tufts Center for the Study of Drug Development even mentioned USD \$1.2 billion dollars to develop a new biotechnology product<sup>2</sup>. This study also reported that the costs of clinical trials and pre-clinical development are similar (USD \$625 millions and USD \$615 millions, respectively).

While these costs are increasing, the pharmaceutical industry has been facing several challenges during the last two years [2], resulting in numerous merging/acquiring guided

1 [http://www.allp.com/drug\\_dev.htm](http://www.allp.com/drug_dev.htm)

2 <http://csdd.tufts.edu/NewsEvents/NewsArticle.asp?newsid=69>

by short-term considerations [3]. This should not hide that the key for finding long-term solutions is likely to be the optimization of the drug development pipeline [4].

This optimization should aim at driving along the shortest path toward drug delivery to the patient [5], filtering out unlikely directions and focusing on the most promising projects [6]. Such a speedup of the drug development process would also lead to a better reactivity against new threats, cheaper drugs, and hopefully the development of new drugs for orphan diseases.

One of the methods to implement such an optimization is called translational research<sup>3</sup>, where the usual “bench-to-bed” one-way processing of biological/medical knowledge is replaced by a two-way communication. In such a multidisciplinary research environment, biological and medical knowledge goes back from *in vivo* to *in vitro* to *in silico*. A general trend is to allow an early clinical phase to take place for infinitesimal doses of a drug under development, in order to examine with care and accuracy its pharmacokinetics and pharmacodynamics properties and identify potential issues [7]. While communication by itself is not the magic bullet, it is likely to help rationalizing the enormous amount of information (and noise) that can be generated by the ever-increasing and highly-recommended usage of high-throughput methods at all stages [5].

During the last ten years, such methods were increasingly used in two directions. High-throughput techniques are now used to identify new biologically relevant targets [8], particularly with the development of microarray techniques in genomics and proteomics projects. The resulting explosion of biological targets comes with the development of other high-throughput methods aiming at designing active compounds more efficiently, involving pharmaco- and toxicogenomics, experimental medicine [8], and of course computational chemistry. The many roles of computational chemistry in drug design are reviewed in [9] and encompass virtual screening, *de novo* design, evaluation of drug-likeness and the determination of protein-ligand interaction.

---

3 <http://nihroadmap.nih.gov/clinicalresearch/overview-translational.asp>

## 5.2 The docking problem

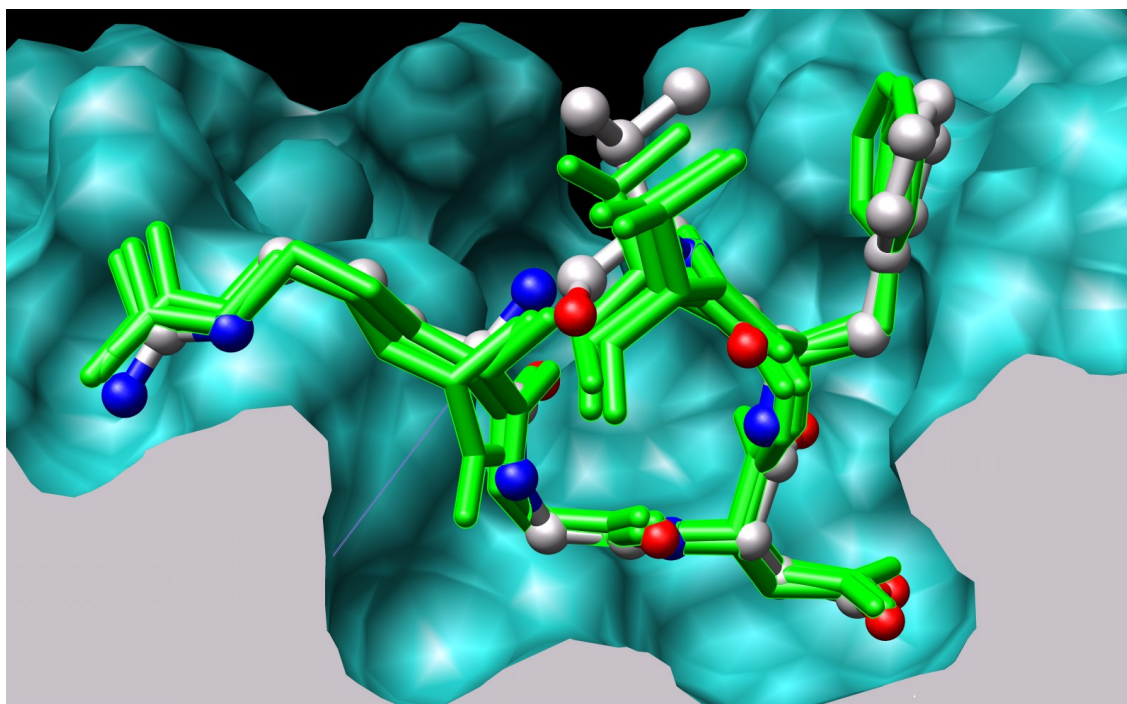
### 5.2.1 Statement

The interaction between molecules is the key of most, if not all, biological events. Molecular details of such interactions are of major interest and can be characterized experimentally using X-ray crystallography or NMR. Unfortunately, the overwhelming number of different molecules in a single cell makes an exhaustive experimental determination of all these interactions by far out of reach. Starting from the structures of unbound molecules, *in silico* molecular docking is an attempt to predict the structure of the corresponding complex. Such a computational approach is considerably easier to set up, cheaper and faster than the experimental methods mentioned above. Docking softwares are valuable tools in pharmacy and medicine, as most drugs are small molecules (ligands) designed to interact with biologically relevant target proteins (receptors) in order to act on the biological pathway they are involved in. This thesis focuses on this particular aspect of molecular docking.

In this introduction, the docking problem is stated and some common answers are summarized. The five most cited softwares are briefly introduced, together with the performance one can expect from such tools. Several challenges for the years to come are then presented, as well as two applications of docking softwares in drug design.

An overview of a typical docking procedure is shown in Figure 2. The first step is to obtain a structure for the receptor, by X-ray crystallography, NMR, or modeling techniques. The more accurate the physical description of this structure, the more relevant, accurate and useful the predicted binding mode (see Figure 1).

Therefore, it should be checked carefully regarding to two aspects. First, the structure should correspond to a biological conformation that is relevant to the targeted biological mechanism. For instance, the presence of crystal contacts in X-ray structures should be verified, as well as the impact of the presence/absence of other interacting partners such as cofactors. Second, the quality of the structure should be verified at an atomic level. For instance, the docking of a ligand is likely to fail if the region encompassing its native



*Figure 1: Side view of the experimentally determined binding mode of the anticancer drug Cilengitide (ball and stick) on the surface of its biological target, the  $\alpha V\beta 3$  integrin. Binding modes predicted with the algorithm presented in this thesis are shown in green sticks. The accuracy of such predictions opens the field of structure-based rational optimizations of the active compound.*

binding mode includes unresolved atoms, has a poor sequence identity with the template structure (if created by homology modeling), or encompasses flexible residues (reflected by a high B-factor if the structure has been determined by X-ray, or multiple conformations if determined by NMR).

If such issues are identified, they have to be addressed during the preparation of the structures for the docking. The latter also includes the resolution of steric clashes, or the assignment of protonation state. The conformation of the ligand is usually optimized by the docking software, and is thus usually not critical. Once both structures are prepared, the docking software can be used with ad hoc parameters to propose one or several putative binding modes, which can be further investigated.

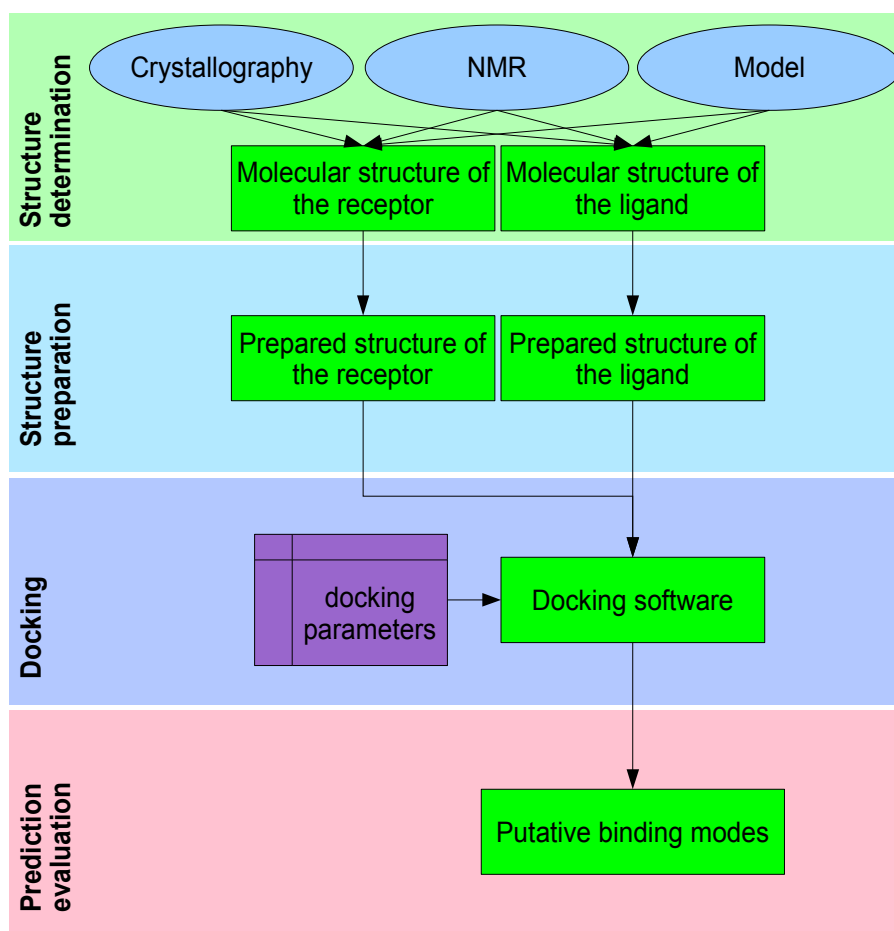


Figure 2: Typical docking pipeline. See text for details.

### 5.2.2 Theoretical problems and methodological answers

To be successful, docking softwares must be able to generate several binding modes and, among them, recognize the native one. Docking can be thus considered as the optimization of structural and energetic criteria described by a scoring function given a set of degrees of freedom corresponding to the ligand and the receptor conformations and their relative positions. This simple formulation should not hide the two challenges it contains. The first challenge comes from the size of the search space, which grows exponentially with the number of degrees of freedom of the system. Its exhaustive exploration is thus not feasible, and all methods primarily rely on heuristic sampling techniques to generate binding modes, which must be carefully designed. The second challenge is to define a scoring function able to pinpoint the native binding mode among

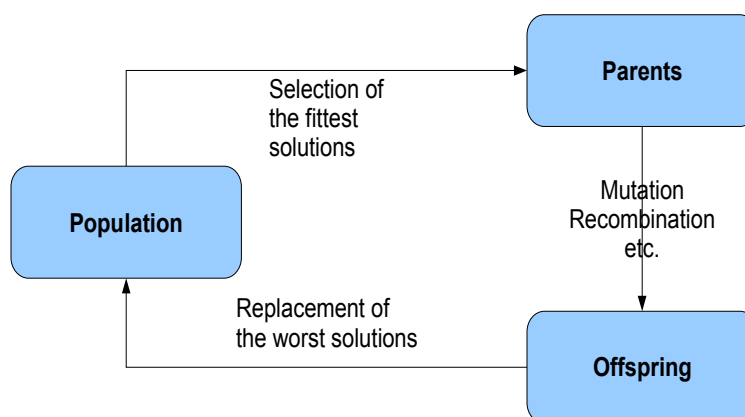
all the ones that are generated.

#### **5.2.2.1 Sampling heuristics**

As mentioned above, several sampling heuristics are used to face the complexity of the search space [10]. They can be classified in three families: approaches derived from a so-called systematic search, molecular dynamics simulation techniques, and stochastic methods. Systematic search methods would be too costly to be applied directly, but several approaches use them in combination with either filtering techniques [11] or with an incremental reconstruction of the ligand. The general principle of the latter is to split the ligand in rigid and flexible fragments: one [12] [13] [14] or several [15] [16] rigid fragments are docked on the surface of the protein and the ligand is reconstructed. Standard molecular mechanics simulation techniques (molecular dynamics and minimization) are appealing because of their physical foundations, but are time consuming and not effective at crossing high free energy barriers within accessible simulation time [17]. However, the reduction of van der Waals and electrostatic repulsions was found to improve the sampling by lowering conformational transition energy barriers [18] [19]. Stochastic methods (Monte Carlo, genetic algorithms, and tabu search) are general optimization techniques with a limited physical basis, and are able to explore the search space ignoring energy barriers.

Evolutionary algorithms (EA) are generic iterative stochastic optimization procedures mimicking the adaptive process of natural evolution, classified as artificial intelligence techniques (this latter also encompass neurocomputing and fuzzy system). On the contrary to most optimization techniques, they focus on several putative solutions at the same time, in a so-called population (see Figure 3).

Briefly, this population is subjected to a selection process, implemented as an objective (or fitness) function describing the problem to optimize, usually involving many degrees of freedom (DoF). The collection of values of each DoF of a solution defines its “genes”.



*Figure 3: Schematic representation of a typical evolutionary cycle taking place in EA. See text for details.*

The selection of the fittest solutions (called parents) according to this objective function is counterbalanced by the generation of new solutions (called children) in order to maintain diversity in the population. Such an offspring is generated by modifying the parental genes thanks to so-called operators, which can be classified depending on the number of parents needed to create a new solution. Operators modifying a single parent to create a child are called mutations, while operators combining two or more solutions into a single child are called recombinations. In Lamarckian genetic algorithms, a local search is performed around children that are created by operators, enhancing the efficiency of the search [20]. This modification of the solution can be viewed as a phenotype modification that is then introduced back in the genome.

Once offspring are generated, it replaces the worst solutions of the population. The latter is then exposed again to the fitness function, and the evolution goes on for the next iteration (generation). As the number of generations increases, the average fitness of the population of solutions is supposed to increase, and several highly fit solutions are expected to appear.

Interestingly, no problem-dependent algorithmic adaptations are needed, even when the fitness function is discontinuous and noisy. Such versatility made EA suitable for various problems, where they can be used as black box optimization procedure. However, these



naïve approaches can also be combined with problem-specific knowledge to drive the search more efficiently, for instance by using semi-stochastic operators or a local search adapted to the underlying energy landscape, if the latter can be at least roughly described. Such approaches are called hybrid evolutionary algorithms, and have proven to be much more efficient [21].

When applied to the docking problem, the fitness function describes the interactions between the ligand and the receptor. This optimization is performed by varying the degrees of freedom related to the ligand and receptor positions, orientations and conformations. During the evolutionary cycle, worst solutions are likely to be replaced by children, created from parents selected among the fittest solutions. This process is repeated until a convergence has been reached in the population, or after a fixed number of generations. Evolutionary algorithms require a balance between diversity and selection, controlling the distribution of solutions in the search space, so that they can efficiently speculate on new solutions with expected improved fitness. A high diversity combined with a slow renewal rate of the population leads to a robust and slow algorithm, roughly similar to a Monte-Carlo search. Conversely, a low diversity with a fast solution turnover is likely to cause a premature convergence [22]. This sampling bias, which can be controlled by evolutionary parameters, is a very powerful aspect of evolutionary algorithms, as they can be tuned according to the complexity of the problem to solve. Two limits can be pointed out. First, this biased sampling does not follow a Boltzmann statistic, and thus does not provide direct insights into the thermodynamical properties of a system, such as its binding free energy. Secondly, evolutionary algorithms extensively use stochastic elements, and consequently, finding the optimal solution is not guaranteed within a finite period of time. More efficient hybrid approaches [21], in which problem-specific knowledge is used to drive evolution, are now widely used for docking [20] [23] [24]. Most of them use the efficient stochastic search of evolutionary algorithms to cross energy barriers and obtain rough minima, which are subsequently refined by a local search like energy minimization [20].

Several recent publications [25] [26] [27] introduced a two-step approach reducing the complexity of the docking problem. First, putative binding pockets are identified, in which

the ligand conformation is optimized. These approaches were reported to be very efficient for virtual screening (VS) [28], but this work has been heavily criticized recently [29]. Moreover, it is anyway not relevant for rational drug design (RDD), which aims at designing a ligand in order to achieve a high specificity for a predetermined binding region.

#### **5.2.2.2 Scoring functions**

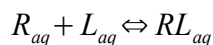
In addition to sampling issues, a common bottleneck of docking programs is the scoring function, responsible for driving the search, discriminating and ranking the generated binding modes [10]. The ideal score would be the free energy of the ligand-receptor association [30] [31], as it would make the ranking of different ligands possible, which is of major importance for VS. Since we are interested in ranking different binding modes of one given ligand on the surface of a given protein, relative free energies of association would be sufficient. Unfortunately, such calculations are currently too computationally demanding. As VS usually implies several thousand dockings to enrich a database, there is a need for fast scoring functions and necessary approximations. Conversely, RDD requires a very accurate but computationally expensive scoring to reliably predict the binding mode for a few tens of complexes. A scoring function must be both efficient and selective [19]: it must be able to drive the search as smoothly as possible (i.e. provide an energy gradient), as well as able to identify the correct binding mode in a set of decoys (i.e. this correct binding mode should have an energy lower than those fake binding modes). Scoring functions can be classified into three families: empirical scoring functions, knowledge-based, and force field based scoring functions. Empirical scoring functions are expressed as a weighted sum of terms arising from given molecular interactions, such as hydrogen bonds, ionic and van der Waals interactions [32] [33]. The weighting factors are fitted on a database of complexes with known structures and binding free energies. Their transferability to complexes outside the training database is thought to be more limited compared to force field-based scoring functions. The second family is based on potentials of mean force that are derived from large datasets of experimental 3D structures [34] [35] [36] [37] [38]. The third family of scoring functions is based on molecular mechanics force

fields, summing the interaction energy and the internal energies of both partners, and ideally taking into account the solvent effect. If the protein is kept rigid, its internal energy does not change and can be ignored, speeding up the evaluation of a binding mode. These scoring functions are usually sensitive to atomic coordinates, limiting their applications in cross-docking experiments [39]. Softened van der Waals potentials, in which the contribution of the repulsive term is limited to allow some steric clashes without penalizing too much the corresponding binding mode, have the advantage of being less sensitive to atomic coordinates in these cases, but also suffer from being less selective [40]. As force field-based scoring functions are not trained on a set of complexes, a good transferability to real world applications can be expected. As an example of this third family, two docking approaches using the CHARMM [41] package were published previously: DARWIN [24] and CDOCKER [18].

No perfect scoring functions has been found yet (see below), and it has been shown that a consensus scoring, filtering docking results using several scoring functions, lead to a considerable reduction of unrealistic docking modes that may have a favorable score according to a given score [42].

### 5.2.3 *Modeling molecular interactions*

The noncovalent, reversible association of receptor (R) and ligand (L) to form a receptor-ligand complex (RL) generally occurs in an aqueous, electrolyte-containing solution (Equation 1):



*Equation 1*

Under equilibrium, this reaction is determined by the standard Gibb's free energy of binding  $\Delta G^\circ$ . This quantity is related to the experimentally determined association constant  $K_A$  and  $K_D$  defined in Equation 2 by Equation 3, and is composed of an enthalpic ( $\Delta H^\circ$ ) and an entropic ( $T\Delta S^\circ$ ) contributions (T refers to the absolute temperature).

$$\Delta G^o = K_A = K_D^{-1} = \frac{[RL]}{[R][L]} T \Delta S^o$$

Equation 2

The CHARMM molecular mechanics force field [43] is an all-atom empirical potential energy function for molecular modeling and dynamics studies of proteins. This energy function includes bounded and non-bounded contributions, as described in Table 2.


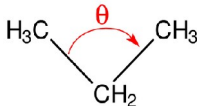
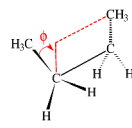
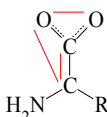

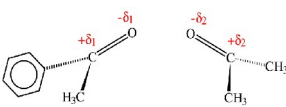
Energy term	Illustration	Equation
Bond length		$E_{bond} = \sum_{bonds} k_r (r - r_0)^2$
Bond angle		$E_{angle} = \sum_{angles} k_\theta (\theta - \theta_0)^2$
Dihedral angle		$E_{torsion} = \sum_{dihedrals} k_\phi [1 + \cos(n\phi - \delta)]$
Improper angle		$E_{impropers} = \sum_{angles} k_\omega (\omega - \omega_0)^2$
van der Waals		$E_{vdW} = \sum_{i,j} 4\epsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right]$
Electrostatic		$E_{elec} = \sum_{i,j} \frac{q_i q_j}{\epsilon r_{ij}}$

Table 2: Energy terms used in CHARMM.

Bond length stretching is modeled by an harmonic potential, where  $r$  is the bond length,  $r_0$  the equilibrium distance and  $k_r$  the bond-stretching force constant. Bond angle bending is also modeled by an harmonic potential, where  $\theta$  is the bond angle,  $\theta_0$  the equilibrium value and  $k_\theta$  the angle bending force constant. The torsion of the dihedral angles is modeled by a cosine expansion, where  $\phi$  is the dihedral angle,  $k_\phi$  its force constant,  $n$  its multiplicity and  $\delta$  its phase. The non-bonded interactions between two

atoms (here named  $i$  and  $j$ ) are modeled using a Coulomb potential term for electrostatic interactions and a Lennard-Jones potential for van der Waals interactions, where  $q_i$  and  $q_j$  are the atomic charges of atoms  $i$  and  $j$ ,  $\epsilon_{ij}$  the dispersion well depth,  $\sigma_{ij}$  the Lennard-Jones diameter,  $r_{ij}$  the non-bonded distance and  $\epsilon$  the dielectric constant.

The sum of those terms gives the complete functional form of the total CHARMM enthalpy, shown in Equation 4.

$$E = \sum_{bonds} k_r (r - r_0)^2 + \sum_{angles} k_\theta (\theta - \theta_0)^2 + \sum_{impropers} k_\omega (\omega - \omega_0)^2 + \sum_{dihedrals} k_\phi [1 + \cos(n\phi - \delta)] + \sum_{i,j} 4\epsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \sum_{i,j} \frac{q_i q_j}{\epsilon r_{ij}}$$

Equation 4

The analytical Generalized Born model (GB-MV2) [44] [45] can be used to calculate the electrostatic solvation energy,  $\Delta G_{elec,solv}$ . This model was found to reproduce the solvation free energies calculated by solving the Poisson equation with 1 % accuracy. The Poisson method for obtaining solvation energies is generally considered a benchmark for implicit solvation calculations. However, GB-MV2 is much faster than solving the Poisson equation (by a factor of about 20) and is therefore very useful to calculate  $\Delta G_{elec,solv}$  for a large number of structures or conformations, and was used in this study. The Generalized Born equation has the following functional form, where  $\epsilon$  is the relative permittivity of the medium (solvent), and  $q_i$  and  $q_j$  are the atomic charges of atoms  $i$  and  $j$ ,  $r_{ij}$  is the interparticle distance,  $\alpha_i$  and  $\alpha_j$  the atoms Born radii and

$$D_{ij} = \frac{r_{ij}^2}{(K_s \alpha_i \alpha_j)} \text{ with } K_s \text{ being a constant set to 8.}$$

$$\Delta G_{(elec,solv)} = \frac{-1}{2} \left(1 - \frac{1}{\epsilon}\right) + \sum_i \sum_j \frac{q_i q_j}{(r_{ij}^2 + \alpha_i \alpha_j^{-D_{ij}})^{0.5}}$$

Equation 5

## 5.3 Molecular docking: state of the art

### 5.3.1 Common implementations

More than 30 programs are currently available [46] and most of them are dedicated to VS. The five most frequently cited ones represent 65% of the citations found in the literature: AutoDock (27%) [20], GOLD (15%) [23], FlexX (11%) [47], DOCK (6%)[12] and ICM (6%) [48]. Interestingly, the two most cited docking softwares AutoDock and GOLD are based on evolutionary algorithms, providing a convenient way to implement separately the sampling heuristic and the scoring function, and in both, the latter is based on a force field. This direction thus seems to be promising.

#### 5.3.1.1 AutoDock

AutoDock is by far the most cited implementation, with 27% of the citations gathered. Conversely to the four others most cited softwares, its citation share has increased regularly between 2001 and 2005. It relies on a Lamarckian genetic algorithm combined with a scoring function based on the AMBER force field [49], and is known for its robustness and accuracy [46] [50]. This flexible software is available for free for academic usage, and is thus often used to investigate new aspects of docking and implement new ideas [51] [52] [53] [54].

#### 5.3.1.2 Gold

The second most cited implementation is GOLD (15%). Interestingly, it also combines a genetic algorithm and a force field based scoring function. It reaches a very good success rate on one of the most comprehensive available test set of complexes [55], but systematic problems are reported for polar ligands and the docking inside large cavities [56].

### **5.3.1.3 FlexX**

The program FlexX comes third, with 11% of the citations. Its sampling heuristic is based on an incremental reconstruction of the ligand using docked fragments as anchors, and its empirical scoring function takes into account entropic, hydrogen-bonding, ionic, aromatic and lipophilic terms. The four latter are scaled by heuristic distance and angle-dependent penalties functions [34] [57]. Its speed comes at the price of a limited performance when docking very flexible ligands [56].

### **5.3.1.4 DOCK and ICM**

DOCK and ICM both represent 6% of the citations each. The former performs a sampling similar to FlexX, which is very fast. Its scoring function does not contain explicit hydrogen-bonding, nor solvation/desolvation, nor hydrophobicity terms, and its accuracy is limited [46]. Several extensions are available, e.g. the docking to multiple receptor structures to account for the flexibility of the protein [58] or a scoring with the GBSA solvation model [59].

ICM is based on Monte Carlo minimization in the internal coordinates space. Its scoring function is based on ECEPP/3 [60], which estimates the entropy of the side chains, and contains an approximated electrostatic solvation term [61]. It was found satisfying at docking and enriching [28] [57].

## **5.3.2 Benchmarking benchmarks**

Docking programs are regularly benchmarked [62] [63]. A Critical Assessment of Prediction of Interactions (CAPRI, <http://capri.ebi.ac.uk/>) was launched in 2001, but only addresses protein-protein interactions. Setting up a fair comparison between different docking softwares is not trivial [64] although the docking of small molecule into protein docking might be significantly helped by such a contest [65]. A fair benchmark should address the points below.

### **5.3.2.1 Experts**

With a few notable exceptions [62] [63], small molecules-protein docking benchmarks are carried out by experts who are often involved in one of the benchmarked algorithms [64]. Since experience is required to get out the most of a docking software [63], their superior knowledge about their own tool may bias comparisons. It also raises a few concerns about their independence, as companies are very active in this field [66].

### **5.3.2.2 Test sets**

According to [64], the ideal test set should not be biased toward a given protein family. Instead, it should be large enough to span representative high-resolution complexes, curated to remove or repair errors such as steric clashes or crystal contacts. Complexes with missing residues or covalent bonds should be avoided. Binding data (such as  $K_D$  or  $IC_{50}$ ) should be available for each complex. The benchmark of programs using fitted objective functions should not be performed on the same test set used for its training. Instead, the database should be partitioned in a training set and test set.

Four databases meeting these criteria can be suggested: LPDB [67], CCDC/Astex [55], PDBbind [68] and BindingMOAD [69]. They contain approximately 260, 300, 290 and 470 curated complexes, respectively.

### **5.3.2.3 Comparable and large search space**

For a fair comparison, all programs should have comparable definition of the search space. The definitions of the region of interest explored by the different programs must be similar. If too small, no clear conclusions can be drawn about the performance of the sampling heuristic [64] [70], and scoring failure may also be hidden, as a more thorough sampling might have led to binding modes with a more favorable score, as recently discussed in [70]. All programs should face a similar set of degrees of freedom, such as rotations and translations, dihedral angles, bond lengths and bond angles. The same structural information such as the presence of water molecules or specific protonation states, if any, should be used by all programs when they take account of it.



#### **5.3.2.4 Comparable seeding**

All programs should be fed with similar conformations. It was recently suggested that the randomization of the degrees of freedom should be biased toward energetically favorable solutions, with a minimal distance to the crystal structure, to demonstrate the ability of the scoring function to discriminate between different energetically favorable minima [70] and highlight the efficiency of the sampling heuristic [64], respectively.

#### **5.3.2.5 Number of evaluated poses and CPU time**

All programs should be compared in their latest available version, and the number of evaluated poses and CPU time should be considered.

#### **5.3.2.6 Outcome**

Even if a fair comparison is set up, recognizing how successful a prediction is not trivial. A maximum RMSD to the experimental structure is usually used to define a correct prediction, but it does not reflect how realistically the molecular interactions are reproduced [71]. This observation led to the IBAC classification of success and failure, which is based on the presence of key interactions between the ligand and the receptor [71]. Unfortunately, it requires a tedious manual inspection of the experimental structure of each complex. The RMSD is thus widely used, a success being reported if the predicted binding mode fall within 2 Å RMSD to the native structure. The reader should also not forget that the experimental structure used as a reference is only an average structure.

Two different reasons can explain docking failures. When the native binding mode is not even sampled, a sampling failure is reported. When the native binding mode is generated, but not correctly discriminated among the numerous decoys also sampled, a scoring failure is reported.

### **5.3.3 Performance Benchmarks**

A recent critical assessment of 37 scoring functions with 10 docking programs (each of

them with several docking protocols), was carried out on eight pharmaceutically relevant targets [62], each of them with up to 200 ligands. When considering the most efficient protocol and all binding modes proposed by docking programs, at least one of them was able to dock more than 40% of the known and crystallized ligands investigated within 2 Å RMSD to the crystal structure for 7 out of their 8 targets. It also appears that for several targets, this success rate increases up to 90% and 100% when considering RMSD to the crystal structure lower than 2 Å or 4 Å, respectively. This promising result was obtained in reasonably fair yet optimal conditions, involving computational chemists with either expertise in each particular protein target or expertise in a particular docking algorithms. Authors concluded that docking algorithms were able to explore conformational space sufficiently well to generate correctly docked poses, despite big variations depending on the target protein. They also observed that scoring functions were less successful at distinguishing the crystallographic conformation from the set of docked poses, as the performance observed was much lower when considering only the top ranked binding mode.

All in all, it seems that there are no universally efficient docking program yet [62] [72], although some of them seem to be consistently better than others, such as GOLD [73]. In a real world prediction, several of them should be used and their predictions should be compared. The reader should also keep in mind that a key aspect of performance that has not been addressed yet is the statistical significance of the different accuracies of benchmarked softwares [64].

## **5.4 Docking challenges**

“Despite the very promising picture drawn, molecular docking still holds several hidden weaknesses, and the so-called docking problem is far from being solved” [46]. The general impression is one of inconsistent performance in combination with a trend toward improvement [74]. This is supported by the recent benchmark mentioned above [62], where docking seems to have reached a plateau and is waiting for an important breakthrough [65]. Both scoring functions and sampling heuristics are facing big challenges in the years to come.

### **5.4.1 Scoring functions**

Strangely, a recent study reported the success of a docking program able to predict complexes correctly even though the ligand protonation state does not correspond to the experimental structure [75], whereas it is known that protonation has a considerable influence on the orientation of a docked ligand [76]. Similarly, the docking of ligands into rigid receptors that have been crystallized with another ligand was reported to be successful [57] even though the atomic description of the binding site was known to be wrong. Such suspicious performances open the question about scoring functions [76]: how could they discriminate between a good and a bad binding mode if a binding mode known to be wrong is recognized as the good one ? In fact, most docking failures can be attributed to scoring functions [62]. An explanation is that many docking tools are used for VS (see below). Their scoring functions must be fast and require a detrimental level of approximations. The different scoring functions available today are at best weakly correlated with the binding free energy [77], which would be the ideal scoring function especially for VS, and most of the time, no correlation can be observed [62] [65].

#### **5.4.1.1 Usual methods evaluating the binding free energy**

Several exact methods calculating the binding free energy are available, among which thermodynamic integration (TI) and free energy perturbation (FEP). They are by far too slow to be combined routinely with the sampling heuristic performed by docking softwares. The binding free energy can also be estimated by approximate methods based on the sampling of several conformations (such as LIE, MM-PBSA, MM-GBSA). Unfortunately, calculating trajectories for each putative binding mode to average thermodynamical quantities is also too slow.

To cope with the throughput required by VS, most scoring functions available today evaluate a single binding mode without further sampling. This relies on the assumption that only this one is significantly occupied [77]. This particularly impacts force field based scoring functions, which are sensitive to small variations of atomic coordinates [39], and there is a great interest in new and more accurate scoring functions bridging the gap

between the costly binding free energy calculations/estimations and current scoring functions [62] [65].

#### **5.4.1.2 Accounting for the entropy**

Even in the most sophisticated scoring functions, the entropy contribution to the binding free energy is poorly modeled or ignored, despite its significant impact [74]. It is widely assumed that the entropic contribution of bond stretches and angle bends are negligible. The assumption is often done for rotational and translational entropy, although they are known to vary not only from one complex to another, but also among alternative bound conformations of a single complex [78]. A recent study considering these two terms was reported to be successful [52].

The loss of flexibility of the ligand upon binding, resulting from a reduction of the energetically accessible rotamers, has an average impact on the binding free energy ranging between 4 and 5 kcal/mol [72]. The overall uncertainty was reported to be of around 5-10 kcal/mol, spanning several logs of affinity [74]. An ideal scoring function should account for this reduction of the accessible conformational space, but this penalty is usually calculated as a function of the number of frozen dihedral angles [74]. This crude approximation makes the consistent and successful ranking of diverse compounds inconceivable when using docking score to estimate the binding free energy [74].

#### **5.4.1.3 Role of water molecules**

Another challenge is the modeling of the role of water molecules in solvation, desolvation and ligand binding. Despite some approaches have shown to be successful [70] [79], this field seems to be still waiting for more accurate calculations [74] [80] [81].

#### **5.4.1.4 Flexibility of the protein**

A convenient way to represent protein flexibility implicitly is to use softened van der Waals potentials. An explicit flexibility can be incorporated by methods combining several conformations of the structure of the receptor into a single map of interaction energy [82],

or into a pharmacophore that weights more the consensus regions of a protein than the flexible regions of the active site [83], or into a specific knowledge-based pair potentials [84]. The docking can also be directly combined with MD simulations where parts of the protein are free to move [85].

The flexibility of the protein is clearly a challenge for the years to come [65], both for the sampling heuristic and the scoring function, as the search space can be dramatically increased and a lot of noise can be generated by remote atomic movements not related to the binding mode evaluated.

#### **5.4.1.5 Reproducibility issue**

Ultimately, the research toward more effective and selective scoring functions [19] is difficult as the re-implementations of published scoring functions often perform differently from the original [86]. As stated in [64], this suggests that “while authors may be meticulous in documenting the exact variants of the scoring functions they use, we fear that the subtleties are often lost when their original work is cited”.

#### **5.4.2 Sampling heuristics**

While the need for more accurate scoring functions is highlighted by most benchmarks, their improvement is likely to require the generation of more relevant and refined binding modes [63]. Sampling is thus still part of the problem, as a perfect sampling heuristic would be impaired by a bad scoring function, so would be a perfect scoring function combined with a bad sampling heuristic. To improve docking accuracy, the balance historically found between the performance of the sampling heuristics and the accuracy of the scoring functions will thus have to be reconsidered [63].

Nonetheless, docking softwares are already used to design new drugs.

### **5.5 Two applications of docking in drug design**

As stated above, docking softwares are valuable tools in pharmacy and medicine, as most

drugs are small molecules (ligands) designed to interact with biologically relevant target proteins (receptors) in order to act on the biological pathway they are involved in. The identification of an efficient ligand (lead compound) is part of a process taking place between the need for a biological activity and drugs delivered to the patients. This process is briefly presented below to highlight the role of docking softwares, and two of their most common applications are then presented.

### **5.5.1 Two impacts of docking softwares**

The number of experimentally resolved protein structures is growing exponentially thanks to huge efforts and improvements in crystallographic techniques. Several of these protein structures are potential targets for the pharmaceutical industry, and the importance of structure-based drug design has thus increased during the past few years. Several computational approaches based on these structures aim at rationalizing experiments by focusing on compounds more likely to have the desired activity, bioavailability and toxicity. Such an early filtering is very helpful regarding to the 5000 compounds required to obtain a single drug (see Table 1). Marketed drugs are usually close to the first active compounds identified (called hits) [87]. Their characteristics are thus very important, and several hits belonging to different chemical families should ideally be identified [5] [62]. Therefore, hit discovery relies heavily on the sampling of the chemical space, for which computational methods are very efficient, especially when considering VS or RDD [5]. These *in silico* methods are now widely used [5] and generate substantial profits [66]. Docking tools have a central role among these methods. Their most common application, VS, intends to rank several thousands of small molecules (usually taken from a database) according to a few properties related to the binding to a pharmaceutically relevant target. Complementary to VS, RDD, such as fragment-based approaches, suggests structure-based modifications of a lead compound, or even lead compound themselves. These two typical applications are described hereafter, as well as a deeper insight into fragment-based rational drug design (FB-RDD).

## 5.5.2 Virtual screening

### 5.5.2.1 Overview

Drugs are usually small organic molecules with less than 30 heavy atoms. The size of the corresponding chemical space is estimated to  $10^{60}$  [88]. Drug-like molecules can be characterized by the “rule of five” [89]: they should have less than 5 hydrogen-bond donors and 10 acceptors, a molecular weight lower than 500 g/mol, and a ClogP lower than 5, but a representative subset of such compounds is still by far out of reach of high-throughput methods. However, once compounds obeying this rule of five are carefully chosen, taking into account their drug-likeness [9] and diversity, docking softwares can be used to identify the ones most likely to be active [65] (see Figure 4). This filtering aims at focusing the numerous affinity assays that need to be performed on compounds favorably ranked. Compared to traditional experimental high-throughput screening (HTS), VS is faster, cheaper, and provides a better sampling of the chemical space [81]. Nevertheless, VS should be used in combination with traditional HTS, as a biological assay is always needed to validate computational methods (see Figure 4).

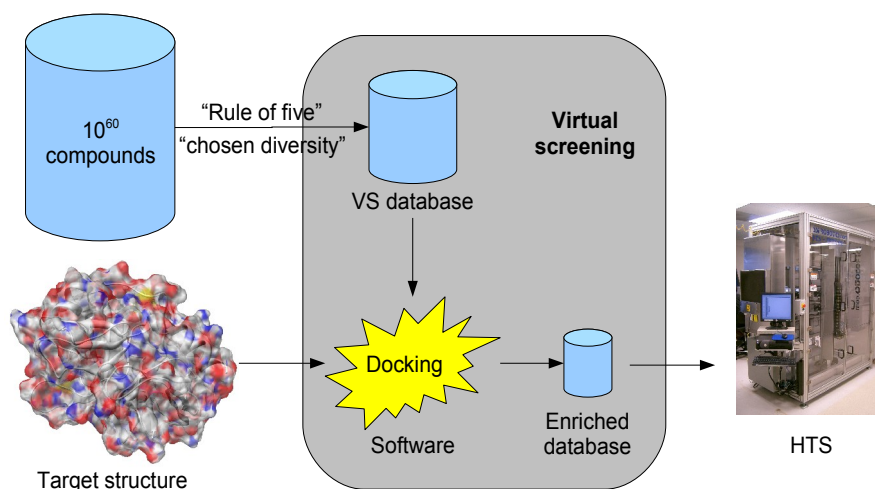


Figure 4: Overview of virtual screening . See text for details.

### 5.5.2.2 State of the art

A recent benchmark of 10 docking programs and 37 scoring functions on 8 biological targets concluded that VS is able to identify active molecules representing several potential leads [62]. As expected, similar yet experimentally inactive molecules are identified as such, and for all but one targets, at least one combination of docking program/scoring function lead to a significant enrichment.

While promising, the results should not hide the inconsistent performance of VS [62] [77]. Another benchmark of 10 scoring functions and 4 docking engines showed that the discriminatory power of today's scoring functions requires better poses, more accurate bioactive conformations, and also more accurate binding poses [63]. As mentioned above, this underlines that the sampling problem is not resolved yet.

A recent study [90] compared the enrichments resulting from the VS of five targets with a known structure, either using the five structures available, or five homology models derived from templates, or even the template structure themselves. Unexpectedly, the enrichments resulting from the VS of homology models was not correlated with the quality of the template structure, and were equal or even greater than the enrichment obtained from the VS of the five crystal structures [90]. Even more surprising, the screening of the structures used as templates also led to a similar enrichment. Because of the molecular



structure differences between the targets and the templates protein, such a result was not expected, and is not significantly rationalized when some degree of flexibility of the protein is allowed. This supports the idea that there is at best a poor correlation between the accuracy of a binding mode and the enrichment, as mentioned by [62] [65] [91]. This statement might come as a surprise regarding to the hype around VS during the last decade. However it is consistent with the fact that the docking problem itself is not resolved yet, and that there are big issues that have to be addressed in the prediction and recognition of a correct binding mode, as well as in the reasonable evaluation of its binding free energy. Recently, some even suggested that the enrichment provided by VS might be more due to filtering out bad solutions than selecting the good ones [65] [74]. From a scientific point of view, this approach is only marginally satisfying although it has been shown to be somewhat effective [62].

### **5.5.3 *Fragment-based rational drug design***

#### **5.5.3.1 *Overview***

The inconsistent performance of VS is likely to be due to the level of approximation required to reach a docking speed compatible with a screening of large databases of drug-like compounds. Interestingly, most known drugs are made of a relatively rigid regions combined with more flexible linkers (Figure 5), defining the so-called molecular fragments.

One way to limit the size of these databases is to search only for smaller molecules, or even fragments of molecules, and then optimize, grow or link these fragments to produce a lead, and (in the best cases) a drug (Figure 6, Figure 8).

Designing drugs from pieces can reduce the dimensionality of the search and dramatically improve the chances of finding good starting points for the drug discovery process against novel drug targets [94]. It also corresponds to a general trend to screen with low MW compounds, also called fragments, which are then optimized [95]. The goal is to identify millimolar to micromolar hits [96] [97] that can be subjected to an optimization.

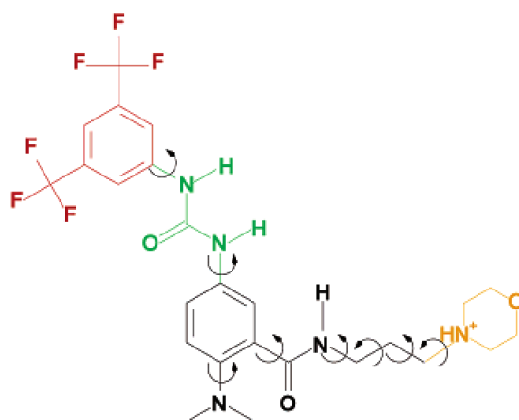


Figure 5: Decomposition of a low micromolar  $\beta$ -Secretase inhibitor into fragments, which are shown in different colors [92] [93].

An interesting consequence of fragment-based rational drug design (FB-RDD) is that it is likely to lead to drugs with a better binding energy per atom, also called “ligand efficiency” [95] [96] [98], which is believed to improve the yield of drug discovery (Figure 7, Figure 8).

#### 5.5.3.1.1 Definition of fragments

Several methods are available to identify interesting fragments [93] [99] [100] [101] [102] [103] [104]. The general trend observed is that Lipinski's “rule of five” for drugs can be rephrased as a “rule of three” for fragments [97]. Fragments typically have a molecular weight lower than 300 Da, a maximum of three hydrogen bond donors and acceptors, and a maximum ClogP of 3.

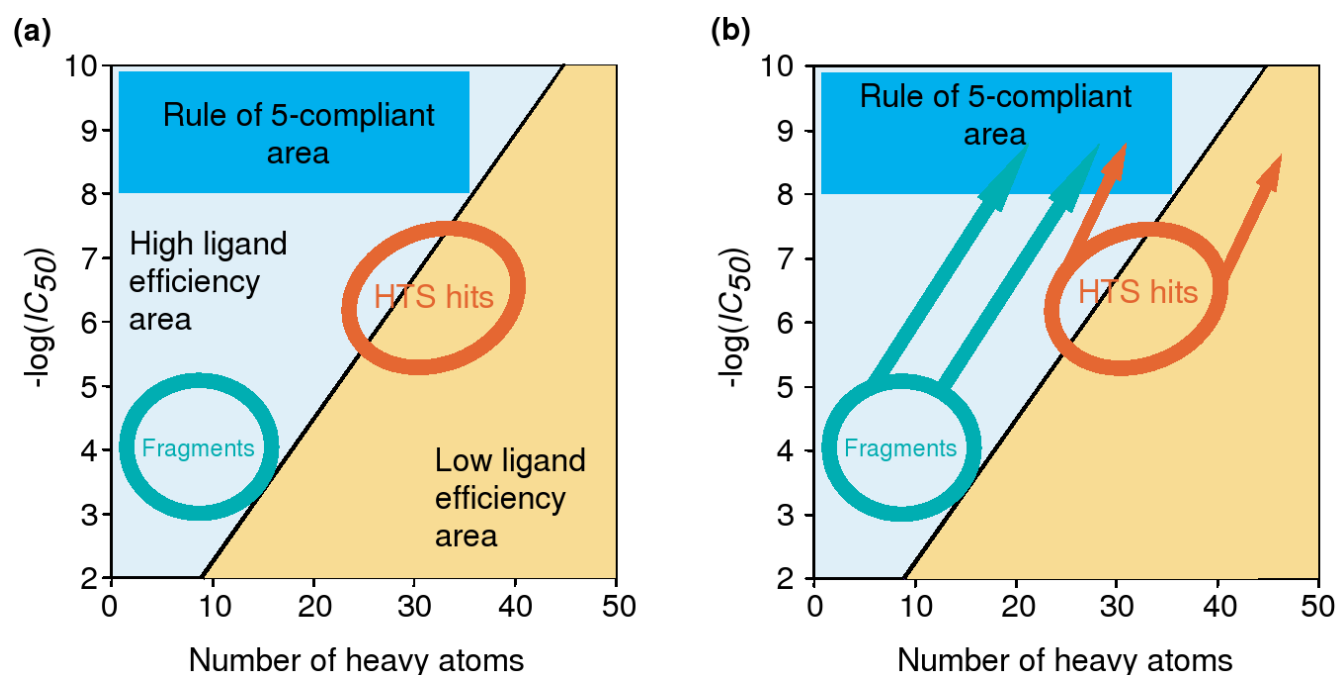


Figure 7: Ligand efficiency and chemical tractability of a hit. (a) The concept of ligand efficiency (LE) can be used to assess the quality of initial screening hits and also to monitor the quality of the leads as they are optimized. The darker blue area is the region where compounds would be at least 10nM in potency and also obey Lipinski's molecular weight guide. Low-affinity and/or low-molecular weight fragments are shown in the green circle. The red oval depicts the broad cross section of assay-detectable hits from HTS, and includes low molecular weight lead-like compounds that would be seen as chemically tractable as well as many more less-attractive compounds with poor ligand efficiency. Potency optimization of fragments or HTS hits will tend to be linked to an increase in complexity and molecular weight. An efficient optimization will be one in which potency is increased without a reduction in ligand efficiency, as illustrated by the green and red arrows. Due to the low complexity and molecular weight of fragments, they are more likely to lead to a rule-of-five compliant lead compound ([96]).

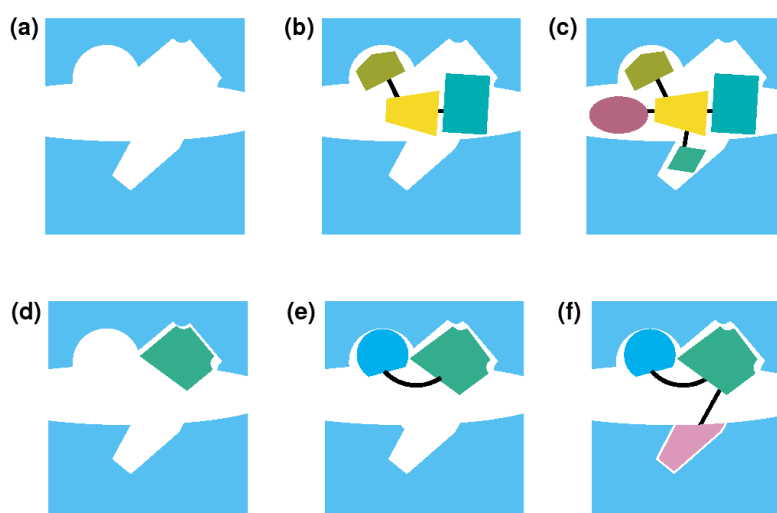


Figure 8: Schematic representation of 'drug-like' HTS hits and fragments as start points for drug discovery. **(a)** Cartoon representation of the active site of a protein, in which there are three pockets that are likely to be "hot spots" for inhibitors to bind. **(b)** Representation of a typical drug-sized HTS hit binding to the active site. The HTS hit is functionally complex and makes numerous but low-quality interactions around the key pockets. The affinity is the summation of interactions spread across the whole molecule. **(c)** representation of a potency-optimized compound derived from the HTS hit in (b), which now better fills the binding site, but at the expense of increased molecular weight and significant complexity. Such optimized ligands are likely to have poor drug-like properties (e.g. size, synthetic complexity, low solubility, multiple functional groups that might be metabolized). **(d)** representation of a "ligand efficient" fragment making a small number of high-efficiency interactions to one of the "hot spots" within the active site. Due to their small size, such fragments would usually not have good activity in a biological assay (typically in the millimolar or high micromolar range). **(e)** representation of a ligand efficient hit compound making good quality interactions in the active site based around a small "template". Such a compound might be expected to be active in a biological assay. This lead-like compound might be rapidly identified from the fragment hit in (d) using structural information of how the fragment binds to the receptor. **(f)** representation of an advanced lead, derived from the hit in (e), making further high-efficiency interactions in the active site, while retaining the key interactions from the original fragment in the "hot spot" of the binding site. This lead has been "evolved" into neighboring binding pockets to produce a compact, ligand efficient and potent lead. ([96])

### **5.5.3.2 State of the art**

#### **5.5.3.2.1 Wet-lab methods**

##### **5.5.3.2.1.1 Fragments positioning**

###### **5.5.3.2.1.1.1 Functional assays**

Inhibiting fragments can be identified using functional assays [88]. No structure is needed and the resulting optimized fragments are likely to lead to a functional inhibitor, not only to a good binder. However, no structural information can be collected, and the linking of fragments is thus difficult and requires that the binding modes of two fragments are in the neighborhood of each other. Notable successes were reported [88] [95] [105]. For instance, Maly and co-workers [106] screened a set of small fragments in a functional assay to identify inhibitors of the kinase c-Src, an important oncology target. Fragments with significant activities were then joined using different linkers. The constructed molecules were then re screened to identify the most potent inhibitors. Several nanomolar inhibitors of c-Src were identified and displayed a greater specificity against related enzymes.

###### **5.5.3.2.1.1.2 NMR and X-ray screening**

NMR and X-ray screening allows the identification of fragments binding to a biological target [107]. Their binding mode can then be identified, and the optimization of fragments is thus made easier. Several successful examples can be found in [88] [95] [105]. The two main concerns of NMR screening are the need for a high field NMR spectrometer and a big amount of radio labeled  $^{15}\text{N}$  protein to carry out experiments.

###### **5.5.3.2.1.1.3 MS-based approaches**

Mass spectrometry can also be used to identify interesting fragments [88]. Two variants exists depending on whether the fragment is covalently linked to the target [94] or not [108] [109] [110]. The binding modes are resolved by X-ray crystallography.

#### 5.5.3.2.1.2 Toward lead compounds

Several methods are proposed in the literature to identify a lead compound starting from one or more interesting fragments. Combinations of these methods are usually used [88].

When a single fragment has been identified, it can be optimized by various substitutions or expansions in order to improve affinity and other properties [88]. When two or more fragments are available, they can be merged or linked. This approach can be used to reconstruct known leads [88] once they have been split in fragments. A third way encompassing areas such as dynamic combinatorial chemistry uses the target as a template for the synthesis of inhibitors from fragments [88]: the target protein is used both to select and to combine pairs of fragments *in situ*. It assembles its own inhibitor by selecting fragments that can cross-link to each other when brought into mutual proximity.

#### 5.5.3.2.1.3 Noteworthy successes

Several successes are presented in a recent review [95] containing references for active compounds that have been found for several targets, such as JNK1 ( $IC_{50} = 0.024 \mu M$ ), PDE4D ( $IC_{50} = 0.019 \mu M$ ), capthepsin S ( $IC_{50} = 0.009 \mu M$ ), DNA gyrase B ( $IC_{50} = 25 \mu M$ ), lactate dehydrogenase ( $IC_{50} = 0.042 \mu M$ ), anthrax lethal factor ( $IC_{50} = 0.032 \mu M$ ), HCV-IRES IIA ( $IC_{50} = 0.72 \mu M$ ), IMPDH ( $IC_{50} = 0.076 \mu M$ ), PDE4 ( $IC_{50} = 0.0009 \mu M$ ), hNK<sub>2</sub> receptor ( $IC_{50} = 0.016 \mu M$ ), DPP-IV ( $IC_{50} = 0.023 \mu M$ ), DHNA ( $IC_{50} = 1.5 \mu M$ ), CDK2 ( $IC_{50} = 350 \mu M$ ), thrombin ( $IC_{50} = 400 \mu M$ ), or PTP1B ( $IC_{50} = 86 \mu M$ ). Other noteworthy successes can be found in [88] [105] [111].

#### 5.5.3.2.2 In silico methods

The *in silico* docking of fragments share the same advantages than VS over traditional experimental HTS. It is faster, cheaper, easier to set up, but requires an experimental validation during or at the end of the optimization process. While better suited for small or rigid fragments [88], FB-RDD provides a better sampling [5]. As it is able to find the most probable binding mode of a fragment within a given region of interest, it can also probably identify binders that are too weak to be detected by the methods described above, even though they would be interesting. Another advantage is that several

overlapping binding modes can be identified whereas they would lead to ambiguity in X-ray crystallography.

#### **5.5.3.2.2.1 State of the art**

##### **5.5.3.2.2.1.1 Skeleton and atom types**

This approach is a two steps procedure: first, initial genetic skeleton of carbon atoms is generated [112] [113] [114] [115] [116] [117] [118] [119]. Atom types are then chosen to optimize the electrostatic or hydrophobic complementarity of the protein [120] [121] [122] [123] [124] [125] [126]. Depending on how fragments are chosen, the synthesis of interesting compound might not be feasible.

##### **5.5.3.2.2.1.2 Fragment docking**

Several programs were designed to dock fragments in favorable conformations, such as Ludi [127], GRID [128], X-CITE [129], SEED [130] or MCSS [131].

##### **5.5.3.2.2.1.3 Fragment linking**

Several approaches can be found in the literature such as HOOK [132], DLD [133], and others [134]. The latter links small functional groups that have been either experimentally determined or determined with Multiple Copy Simultaneous Search (MCSS, [131]). Its scoring function has been designed to select fragments interacting favorably with the receptor and recognize molecules with satisfactory bond lengths, angles and dihedral angles. The refinement of these molecules is performed using a Monte-Carlo sampling and a simulated annealing protocol.

An interesting approach has been implemented by the Caflisch group, where fragments (identified with DAIM [93]) are docked with the program SEED [130], and used as anchor points by FFLD [16] to propose docking modes for one or several compounds.

#### **5.5.3.2.2.2 Applications**

An approach combining DAIM, FFLD and SEED was found to be very successful at

identifying micromolar inhibitors for the  $\beta$ -Secretase [92], an interesting target regarding to the Alzheimer disease.

Another noteworthy success was reported in [135], where an *in silico* screening for potential low molecular weight inhibitors of the DNA gyrase is combined with a biased high-throughput screening, and a 3D guided optimization process. A new inhibitor was found, ten times more potent than the reference inhibitor novobiocin.

### **5.5.3.3 Conclusion**

Most studies around FB-RDD were published during the last five years, and were carried out in the industry [66] [88] by companies such as Vertex, Sareum, Astex Technology or Sunesis Pharmaceutical. The docking, linking and optimization of fragments *in silico* is very likely to grow in the coming years, with docking programs able to predict accurately the positions and orientation of fragments, as both are required for the linking/optimization to be successful. The docking problem is not solved yet [65], but the relatively low chemical complexity compared to VS leaves more space for slower and more accurate scoring functions, and more efficient sampling heuristics.

## **5.6 General conclusion**

As stated in [46]: “Despite the very promising picture drawn, molecular docking still holds several hidden weaknesses, and the so-called docking problem is far from being solved. The lack of a suitable scoring function, able to efficiently combine both accuracy and speed, is perhaps the most detrimental weakness. The results of a docking experiment should therefore not be taken as the final result of a structural study, but rather as a good starting point for a deeper and more accurate analysis. In this sense, docking must be necessarily fast, enabling large quantities of data to be considered, and reasonable and coherent solutions to be generated. However, the final result (geometry, binding free energy) should always be determined by a more accurate and precise methodology, naturally slower”.



The main goal pursued during this PhD thesis was to end up with a docking algorithm that is able to deal with real world applications, precise enough for RDD, and versatile enough to be used as a toolbox to investigate new answers and development that may lead to important breakthrough [65]. This docking algorithm is called EADock.



## 6 Presentation and benchmark of EADock

Most published docking algorithms were designed for VS. The resulting time constraint implies that either a very fast (and thus less accurate) scoring scheme is used, or that the sampling is limited around the supposed binding pocket, or both. The docking algorithm for RDD proposed in this work, Evolutionary Algorithm for Docking (EADock) provides a unique combination of four methods, which have been presented separately [18] [24] [136] [137] [138].

First, the thorough sampling heuristic of EADock is inspired by evolutionary algorithms, and uses a combination of two fitness functions. The first one, which neglects solvent effects, is used to drive the search toward local minima because of its efficiency and speed. These minima are then exposed to a more selective and computationally demanding fitness function, which includes the solvation free energy. This approach thus relies on the assumption that minima of the second fitness are also minima of the first, though their rank may be different [139].

Secondly, a mechanism inspired by tabu search restricts the search space as the evolution proceeds, by storing a list of previously visited unfavorable docking poses and preventing the search from revisiting these poses, thus facilitating the exploration of new conformational space. This continuous update of the search space also ensures that the evolution does not converge to complexes that do not correspond to a minimum of the second and more selective fitness.

Third, the sampling is performed with operators that combine a broad and a local search of the conformational space. Some of these operators are semi-stochastic, dealing with rotations and translations. Other operators, called “smart operators” aim at crossing energy barriers by transiently modifying the fitness landscape, in a physical and deterministic manner.

Fourth, aside from this flexible sampling framework, coordinates handling and energy calculations are delegated to the CHARMM package, for which a Java API was developed.

EADock is thus able to use the latest improvements available in CHARMM, especially sophisticated solvation models such as GB-MV2 [44] [45].

The predictive ability of EADock was benchmarked on a previously used set of 37 protein complexes [57]. A successful prediction was defined by a RMSD between the predicted binding mode and the crystal structure, calculated for heavy atoms of the ligand (referred to as RMSD), lower than 2 Å. Despite challenging starting conformations, a huge search space and a very short evolution, such complexes were identified and proposed for 92 % of the test complexes, and ranked first for 68 % of them. Some failures may be explained by the existence of a bond between the ligand and its receptor that is out of the scope of our scoring functions. For all remaining failures except for one, a significant interaction was found between the ligand and a neighboring complex of the crystal unit cell.

## **6.1 Docking algorithm**

An overview of the algorithm is outlined in Figure 9.

EADock is initialized with parameters relative to the docking (such as reference coordinates for the targeted protein and the ligand, and a list of free dihedral angles) and for the evolutionary process itself (such as the population size and the number of generations). Of course, the crystal structure, if any, is never used in any way to introduce a bias or a driving force in the algorithm. A region of interest (ROI) is defined as a sphere. Depending on its position and its radius, this sphere can be focused around the binding site, or encompass the whole protein surface.

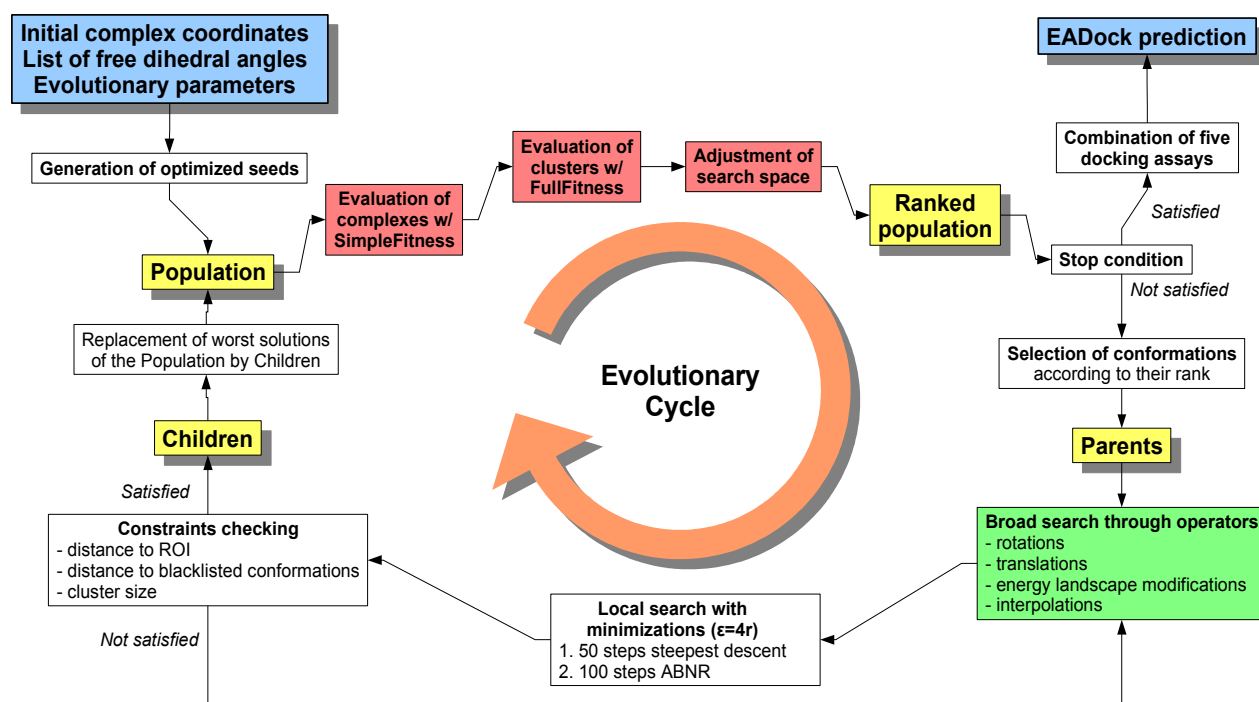


Figure 9: Main steps of the docking algorithm implemented in EADock. Typical parameters are a population size of 250 complexes, 400 generations, a clustering cutoff of  $2 \text{ \AA}$ , and a maximum cluster size of 8 elements.

From a technical point of view, EADock is a Java program. It relies on a generic evolutionary engine called Jeep and on a docking-specific code which is interfaced with a molecular mechanics engine. For this study, we chose the CHARMM package (Figure 10).

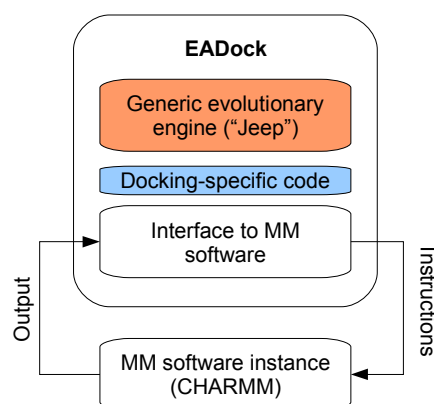


Figure 10: Software design

### 6.1.1 Seeding

The first population to be evolved, generation zero, is filled with decoys from the reference coordinates of the ligand. These decoys are referred to as seeds. Each seed is generated by random translation and rotation starting from the reference coordinates, followed by a sequential optimization of every user-defined dihedral angle. Resulting complexes are optimized by a routine called *SmartAttractor*. This procedure translates the ligand to a minimum of the interaction energy, close to the protein surface, along an axis defined by the two closest atoms. This minimum is identified by iteratively translating the ligand toward the protein with a step of 0.1 Å and minimizing its energy.

Starting from the conformations obtained from the translations/rotations, each dihedral angle specified by the user is optimized sequentially by the *OptRot* operator (see below). After this optimization, the ligand is further minimized using 50 steps of steepest descent (SD) followed by 100 steps of Adopted Basis Newton-Raphson (ABNR). Such a search guarantees that energetically unfavorable rotamers are not retained in the first generation, and that the free dihedral angles, bond lengths and valence angles of the ligand are optimized for each initial binding mode.

### 6.1.2 Selection

Once a population has been created, two fitness functions are successively applied, providing the only two driving forces in EADock. First, complexes are ranked according to a fast and simple scoring function, the *SimpleFitness*. Secondly, *clusters of complexes* are formed and confronted to an accurate and slower scoring function, the *FullFitness*, and the ranks of their centers are updated.

The *SimpleFitness* is equal to the total energy of the system calculated with the CHARMM22 molecular mechanics force field, with a dielectric constant of 1 and no cutoff:

$$SimpleFitness = E_{intra}^{ligand} + E_{intra}^{recept.} + E_{inter}$$

$E_{intra}^{ligand}$  and  $E_{intra}^{recept.}$  are the internal energy of the ligand and the receptor, respectively.

They are equal to the sum of the internal bonded (bonds, angles, etc.) and non-bonded (electrostatic and van der Waals interactions) terms. When the receptor is fixed, its internal energy,  $E_{intra}^{recept.}$ , is constant.  $E_{inter}$  is the interaction energy between the ligand and the receptor, and is equal to the sum of the van der Waals and electrostatic interaction energies. The *SimpleFitness* is fast, but neglects the effect of solvent known to have an important contribution to the binding free energy. It is nevertheless likely to focus on reasonable solutions [139].

The *FullFitness* is subsequently used to evaluate clusters that are identified using the RMSD matrix between all complexes in the population. The most favorably ranked complex is chosen as center for the first cluster. Its neighboring complexes in the population, defined with a RMSD threshold, are assigned to this first cluster. The next most favorably ranked complex is chosen as the center for the second cluster, and its neighbors are assigned to this second cluster. This procedure continues until all complexes of the population have been assigned to a cluster. When at least three clusters have reached their maximum number of members (typically 8), their *FullFitness* is computed.

The *FullFitness* of a cluster is calculated by averaging the 30 % most favorable effective energies of its elements, in order to limit the risk of a few complexes penalizing the whole cluster. This effective energy is written as the sum of the total energy of the system and a solvation term. Neglecting the solute entropic contribution, we can write:

$$G_{eff} = E_{intra}^{ligand} + E_{intra}^{recept.} + E_{inter} + \Delta G_{elec, solv} + \sigma \times SASA$$

where  $E_{intra}^{ligand}$ ,  $E_{intra}^{recept.}$ , and  $E_{inter}$  are calculated as described above. The solvation energy is composed of the electrostatic,  $\Delta G_{elec, solv}$ , and the non-polar contributions. The latter can be considered as the sum of a cavity term and a solute-solvent van der Waals term, and is assumed to be proportional to the solvent accessible surface area, SASA [140] [141]. We use a value of 0.0072 kcal/(mol Å<sup>2</sup>) for the parameter  $\sigma$  [142] [143] [144] and the SASA was calculated analytically in CHARMM.

$\Delta G_{elec,solv}$  is calculated using the analytical Generalized Born Molecular Volume (GB-MV2) model implemented in CHARMM that is about 20 times faster than solving the Poisson equation. Recent results showed that the deviation between the desolvation energies calculated with the GB-MV2 model and with the Poisson-Boltzmann (PB) model is constant for a series of different conformations of a given complex, which means that the use of GB-MV2 does not alter the ranking of binding modes [145].

Each fitness function modifies the rank of complexes in the population, and this rank is used to select parents for the next generation; the centers of clusters with the most favorable *FullFitness* are ranked on top of all other complexes in the current population. Conversely, centers corresponding to clusters with a less favorable *FullFitness* are removed from the population and added to the tabu list.

Evolutionary algorithms require a balance between selection and generation of diversity, the latter being embodied in new complexes created from parent complexes. In EADock, parents are chosen to refine and fill identified clusters, so that these clusters can compete with respect to their *FullFitness* as frequently as possible, limiting the risk of discarding interesting complexes due to the poor selectivity of the *SimpleFitness*. Parent complexes are selected from the top ranked half of the population, according to their internal rank in the cluster to which they belong, then by the rank of this cluster among other clusters. Isolated binding modes are considered as mono-element clusters. The best member of each cluster is selected first, then the second best member of each cluster, and so on. This selection continues with elements of the following ranks until enough parents have been collected. Members of small clusters can be selected several times.

In summary, EADock uses two fitness functions on two different levels: complexes are ranked according to the *SimpleFitness* (fast, efficient) to guarantee reasonable electrostatic and van der Waals interactions. Clusters of complexes, corresponding to binding modes, are then evaluated by the *FullFitness* (slow, selective) taking into account the solvent effect, and the search space is adjusted consequently. Both fitness functions modify the rank of complexes in the population, as this rank is used to select parents for the next generation. After a user-defined number of generations, the evolution is stopped.



### 6.1.3 Diversity

To generate a child, one or two parent complexes are selected according to their rank in the collection described above, and modified by an operator. Operators combine a broad search procedure (see below) followed by a local search through energy minimization. The latter was shown to speed up convergence and improve prediction quality [20] by resolving simple steric problems that might be introduced by the former, as well as adjusting valence angles and bond lengths. Once a child belonging to the search space has been generated and confronted to the tabu list, it is included in the population. If it belongs to a cluster that is already full, it replaces the member with the least favorable *SimpleFitness*. Otherwise, the worst ranked solution of the whole population is replaced. This prevents the premature convergence of the whole population to a single minimum.

Several operators are available to generate new complexes. Four operators optimize the position and orientation of the ligand relative to the protein: two consist of random rotations around a random axis and two of random translations along a random axis. For each kind of movement, two sets of parameters are used, either focusing on a local (rotations up to  $40^\circ$  and translations up to 2.5 Å) or on a long-range exploration (free rotations and translations up to 10 Å). The latter are referred to as long-range operators, the former and all other operators being short-range operators. As long-range operators are likely to deeply modify the ligand pose, the *SmartAttractor* procedure described previously is applied to the newly generated pose in order to avoid steric clashes or complexes with little or no interaction between the ligand and its receptor.

The *OptRot* operator optimizes the free dihedral angles of the ligand. One of these dihedral angles is randomly chosen and optimized as follows; two groups of atoms are identified, one on each side of the bond defining the rotation axis. The first group of atoms is held fixed while the second is rotated by  $60^\circ$  steps. Scanned poses are minimized using 50 steps of SD followed by 100 steps of ABNR, and assigned a score with the *SimpleFitness*. The conformation of the second group of atoms with the lowest score is

kept with a Metropolis-like criterion. The angle scan is then repeated swapping the two groups of atoms in order to rotate the first one, and the rotamer with the most favorable score is retained.

Three operators, *ElectrostaticOptimizer*, *VanDerWaalsOptimizer* and *SoftLigand*, were designed to modify the ligand binding mode of the parent thanks to a minimization in a transiently altered force field. To some extent, this can be related to the limitation of the repulsive van der Waals term which has been described in [18] [19] that also smooths the energy landscape. They all follow the same principle: first, the relative contribution of a specific energy term is artificially increased or decreased. Secondly, an energy minimization is performed in this altered force field. Third, initial contributions of energetic terms are restored, and fourth, an additional energy minimization is performed to relax the ligand in the original force field.

*ElectrostaticOptimizer* and *VanDerWaalsOptimizer* transiently increase by a factor of five the electrostatic or the van der Waals interaction energy, respectively. Both the sampling and relaxation minimizations follow the same scheme consisting of 50 steps of SD, followed by 100 steps of ABNR. *SoftLigand* transiently decreases the self-energy of the ligand by a factor of four. This alteration of the force field allows the ligand to be transiently distorted in order to cross energy barriers and to improve its interaction with the surface of the protein during 150 steps of ABNR minimization. The relaxation minimization consists of 500 steps of ABNR.

The last operator, *Interpolator*, uses two parent complexes if the RMSD between them ranges from 0.2 Å to 5 Å. A set of interpolated conformations are generated and optimized by the *SmartAttractor* procedure (see Seeding), and the conformation with the lowest interaction energy is retained.

Once a parent has been randomly chosen, an operator is selected based on its probability to be applied, which is increased automatically according to its contribution to the fitness improvement over the last five generations. This procedure, called automatic operator scheduling, has been described previously [21]. In brief, each time a child is created, the

operator that was applied is credited with the *SimpleFitness* difference between this child and its parent. Every fifth generations, the probabilities of operators are adjusted according to this credit. In addition, a bias is introduced depending on the size of the cluster to which the parent belongs. If this cluster has reached its maximum allowed size, long-range operators are more likely to be selected in order to escape from this identified minimum. Conversely, short-range operators are more likely to be applied if the cluster is almost empty, to refine it and increase its number of members, in order to evaluate it as soon as possible with the *FullFitness*.

#### **6.1.4 Postprocessing**

The reliability of each docking experiment is enhanced by combining several (typically five) independent evolutions. Complexes for which the effective energy has been calculated are merged into a single optimized population, which is then reclustered. These new clusters are ranked according to their *FullFitness*, which is calculated as described previously.

A 2 Å RMSD threshold between the top-ranked cluster and the crystal structure defines an successful prediction. If no such conformation was ever sampled, a sampling failure is reported. If a successful complex was generated but lost before its evaluation by the *FullFitness*, a *SimpleFitness* failure is reported. If an acceptable cluster was evaluated by the *FullFitness*, but lost afterwards because of its poor score, a *FullFitness* failure is reported.

### **6.2 Dataset**

To ensure unbiased benchmarks [64], complexes used for the validation of EADock were taken from a previous study [57]. They are presented in Table 3.

Experiment	Binding site accessibility	Complex	q	DoF	Hb A.	Hb D.	Mass	% B. Sur.
28 test cases: algorithm assessment and benchmark	Accessible	<i>Carbonic anhydrase</i>						
		1cil	-1	3	6	2	323.4	85.1
		1okl	0	2	4	1	249.3	87.7
		1cnx	0	10	6	3	331.4	74.2
		<i>Neuraminidase</i>						
		1nsc	-1	4	9	6	308.3	92.0
		1nsd	-1	4	8	5	290.3	92.6
		1nnb	-1	4	8	5	290.3	89.7
		<i>Ribonuclease</i>						
		1gsp	0	2	9	3	360.3	80.2
		1rhl	-2	3	10	4	361.2	78.1
		1rls	-2	3	10	4	361.2	79.2
		<i>Trypsin</i>						
		3ptb	1	1	0	2	121.2	94.6
		1tnq	1	1	0	1	114.2	91.6
		1tnj	1	2	0	1	122.2	92.4
		1tnk	1	3	0	1	136.2	91.0
		1tni	1	4	0	1	150.2	85.6
		1tnl	1	1	0	1	134.2	92.7
		1tpp	0	2	3	2	206.2	86.9
		1pph	1	7	3	3	429.6	69.9
	Poorly accessible	<i>Carboxypeptidase</i>						
		1cbx	-1	3	4	1	207.2	98.2
		3cpa	0	4	4	3	238.2	97.7
		6cpa	-1	9	8	2	477.4	82.3
		<i>Penicillopepsin</i>						
		1apt	1	17	6	5	501.7	85.9
		1apu	0	15	6	4	485.7	85.0
		<i>Thermolysin</i>						
		3tmn	0	5	3	3	303.4	73.0
		5tln	-1	7	5	3	320.3	79.8
		6tmn	-1	11	8	3	471.5	73.2
		<i><math>\epsilon</math>-Thrombin</i>						
		1etr	0	7	6	4	504.6	87.9
		1ets	1	7	4	4	522.7	88.3
		1ett	1	5	3	3	429.6	88.2
9 test cases: algorithm benchmark only	Buried	<i>Cytochrome P-450cam</i>						
		1phf	0	1	1	1	144.2	100.0
		1phg	0	3	3	0	226.3	100.0
		2cpp	0	0	1	0	152.2	100.0
		<i>Intestinal FABP</i>						
		1icm	-1	11	2	0	227.4	95.6
		1icn	0	14	2	1	282.5	96.0
		2ifb	-1	13	2	0	255.4	96.9
		<i>L-Arabinose</i>						
		1abe	0	0	5	4	150.1	100.0
		1abf	0	0	5	4	164.2	100.0
		5abp	0	1	6	5	180.2	100.0

Table 3: Our approach was tested on 37 complexes. This table shows complexes that have been used for the sampling heuristic assessment and for the benchmark of the algorithm. This distinction is based on the accessibility of the binding site. This table lists the PDB code for each complex, as well as the charge of the ligand ( $q$ ), the number of internal degrees of freedom of the ligand (DoF), the number of hydrogen-bond donors and acceptors (Hb. A. and Hb. D.), the mass of the ligand (Mass) and the percentage of surface of the ligand buried upon complexation (% B. Sur.)

As can be seen, featured ligands are diverse in terms of charge, molecular weight, number of internal degrees of freedom, number of hydrogen-bond donors and acceptors, and logP.

The binding sites can be classified according to their accessibility. Binding sites are considered buried if the fraction of the ligand surface buried upon binding is greater than 95% for all ligands. This is the case for cytochrome, L-arabinose and intestinal FABP. Non-buried binding sites can be either easily (trypsin, neuraminidase, ribonuclease, carbonic anhydrase), or poorly accessible ( $\epsilon$ -thrombin, thermolysin, penicillopepsin, carbocypeptidase), depending on the shape of the binding pocket.

Titration groups were considered to be in their standard protonation state at neutral pH. The protonation state of histidine residues was defined based on inspection of their environment. The proteins and ions were modeled using the all-atom CHARMM22 [43] force field. Missing hydrogens in the crystal structure were added using the HBUILD [146] procedure of CHARMM. Missing parameters for the ligand, for use in conjunction with CHARMM22, were derived from the Merck Molecular Force Field (MMFF) [147] [148] [149] [150] [151] by taking the dihedral angle term as is, and the quadratic part of the bond and angle energy terms. The partial charges and van der Waals parameters of the ligand atoms were taken from the MMFF. The ligands were modeled with all hydrogens.

Before starting the docking process, the crystal structures were minimized using 100 steps of SD with the GB-MV2 solvation model. No cutoff was used. This short minimization was used to remove clashes arising from the crystal structure and hydrogen atoms placement without affecting the protein conformation. The RMSD between the starting and final conformations, calculated for all heavy atoms, was always lower than 0.15 Å. The ligand was removed before starting the docking process.

### **6.3 Algorithm assessment and benchmark**

First, the scoring strategy and the sampling heuristic were assessed. Then, the predictive ability of EADock was benchmarked in conditions similar to a real application.

For each test case, a hundred decoys were generated using the seeding procedure described above. The *SimpleFitness* and the effective energy of these decoys were calculated, and plotted against the RMSD to the crystal structure. The convergence of the evolutionary process relies on the correspondence between minima of the two fitness functions. Such a correspondence implies that if a complex has been first minimized in the *SimpleFitness* force field, a subsequent minimization in the *FullFitness* force field has only a limited impact on its coordinates. This was confirmed for each test case, using 250 decoys generated with the seeding procedure described above, except that no minimizations were performed (conformations A). These conformations were minimized in the *SimpleFitness* force field (50 steps of SD followed by 100 steps of ABNR) giving conformations B. The latter were further minimized in the *FullFitness* force field (100 steps of SD) giving conformations C. The RMSDs between corresponding poses in B and C were calculated ( $\text{RMSD}_{\text{BC}}$ ). In view of the hypothesis that the two fitness share the same minima, the distribution of  $\text{RMSD}_{\text{BC}}$  is expected to be close to zero. To provide a reference, conformations A were also minimized directly in the *FullFitness* force field (100 steps of SD) giving conformations D. The RMSDs between corresponding poses in A and D were also calculated ( $\text{RMSD}_{\text{AD}}$ ). The  $\text{RMSD}_{\text{BC}}$  and  $\text{RMSD}_{\text{AD}}$  distributions were compared, in order to verify that  $\text{RMSD}_{\text{BC}}$  is closer to zero than  $\text{RMSD}_{\text{AD}}$ .

The quality of the sampling of a docking algorithm can be measured by its ability to converge on the crystal structure when starting from remote seeds. Docking assays were thus performed with increasingly challenging seedings for the 28 test cases with an accessible or poorly accessible binding pocket. Test cases with buried binding pockets were excluded; since their binding pockets are *a priori* unambiguously identified, they would have led to an overestimation of the efficiency of our sampling heuristic. A total of 140 docking assays were performed: for each of the 28 test cases, five groups of seeds were generated as described above, with a RMSD to the crystal structure varying between 0-3 Å (easiest), 2-5 Å, 4-7 Å, 6-9 Å and 8-11 Å (most difficult). For all docking runs, 25 out of the 250 docking poses in the population were renewed at each generation. Children were generated from parent conformations selected out of the top-ranked half of the population. All operators shared the same base probability (0.2), and their maximum

adaptive probability was set to 0.1. A clustering distance cutoff of 2 Å was used together with a maximum of 8 conformations per cluster. Identified clusters were evaluated by averaging the three most favorable effective energies of their members.

Finally, EADock was benchmarked in real conditions on the same test set: a docking assay was realized for the 37 test cases, using the parameters described above, combined with a less restricted seeding. Seeds were generated between 3 Å and 10 Å RMSD to the crystal structure for test cases with accessible or poorly accessible binding pockets (seeds too close to the crystal structure were explicitly excluded, see discussion). The ROI was limited to a radial distance of 15 Å around the center of mass of the crystal structure. For buried test cases, seeds were generated between 3 Å and 5 Å RMSD while the ROI was limited to 5 Å to prevent sampling outside of the binding pocket.

## 6.4 Algorithm performance

First, the scoring function and sampling heuristic of the algorithm were assessed by investigating the relationship between the *SimpleFitness* or the *FullFitness* with the RMSD to the crystal structure using a set of decoys, and by starting the evolution from unfavorably biased seeds.

Second, the predictive ability of the algorithm was benchmarked in realistic conditions, excluding seeds with a RMSD to the crystal structure lower than 3 Å RMSD, in order not to introduce a favorable bias in our results (see Discussion). Both evaluations were performed on the same data set of 37 ligands [57].

### 6.4.1 Algorithm assessment

Four representative examples of the relationships that we observed between the RMSD and both the *SimpleFitness* and *FullFitness* are shown in Figure 11: a successful identification of an acceptable cluster by both fitness (A), a *SimpleFitness* failure (B), a *FullFitness* failure (C), and a failure of both fitness (D). Compared to the *SimpleFitness*, the *FullFitness* is generally noisier, and its driving force is present only near the global

minimum.

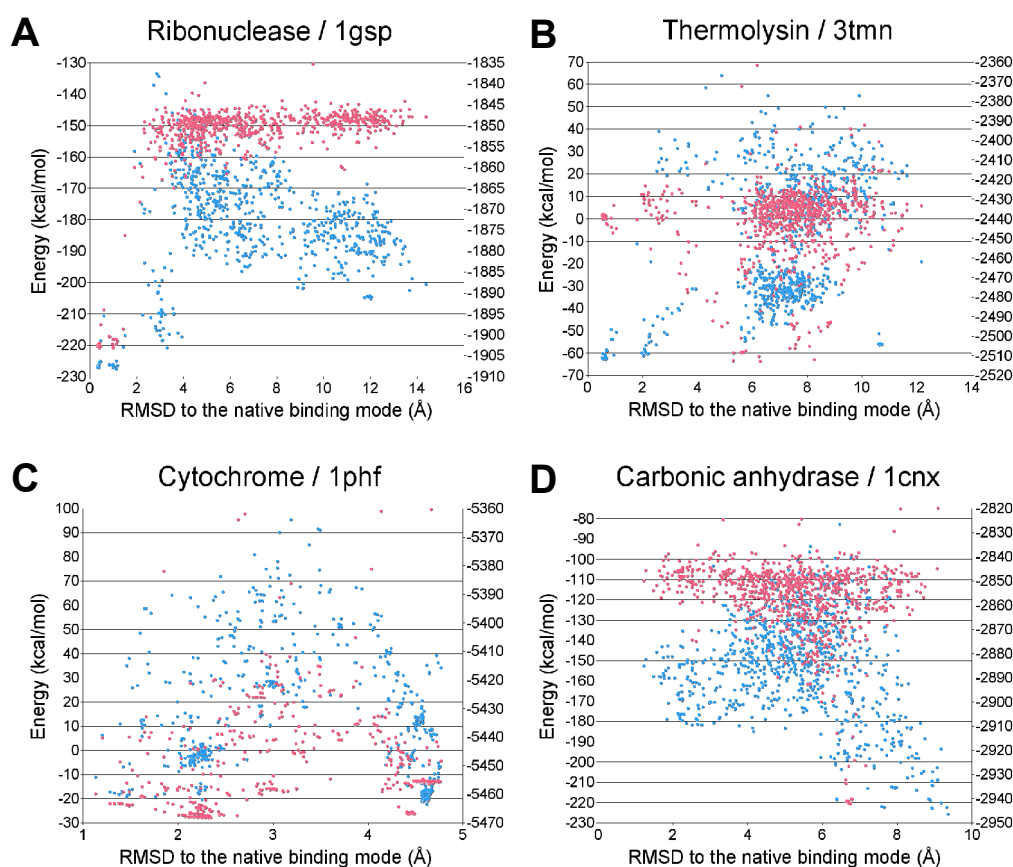


Figure 11: Correlation between the RMSD and the SimpleFitness (pink, left Y-axis) and the FullFitness (blue, right Y-axis). The four plots correspond to four representative test cases. For each of them, a set of 1000 decoys were generated using the seeding procedure (see Material and Method). A. both fitness are able to identify a cluster very close to the crystal structure. B. the SimpleFitness fails at ranking the cluster corresponding to the crystal structure correctly, but point it out as a local minimum which is ranked correctly by the FullFitness. C, a correct cluster was ranked first by the SimpleFitness, but not by the FullFitness. In this case, a bond exists between the ligand and the receptor. D, both fitness fail at identifying an acceptable cluster.



The correspondence between minima of both fitness functions was assessed by plotting histograms for  $\text{RMSD}_{\text{BC}}$  and  $\text{RMSD}_{\text{AD}}$  (see Material and Methods). A representative example, ribonuclease/1gsp, is shown in Figure 12.

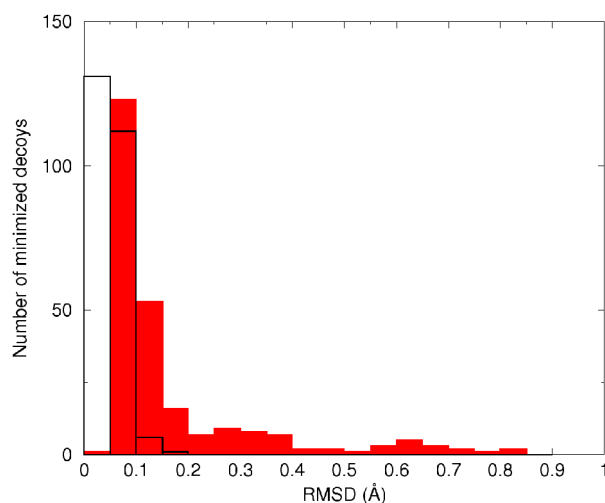


Figure 12: Histograms of  $\text{RMSD}_{\text{BC}}$  and  $\text{RMSD}_{\text{AD}}$  (see Material and Method) for a representative test case (ribonuclease/1gsp). As can be seen, the distribution of  $\text{RMSD}_{\text{BC}}$  (black lines) is tighter than the distribution of  $\text{RMSD}_{\text{AD}}$  (red bars), and corresponds to much lower RMSD. This supports the hypothesis that the minima of the *SimpleFitness* are also minima of the *FullFitness*.

As can be seen, the  $\text{RMSD}_{\text{BC}}$  is lower and its distribution is much tighter. This supports the hypothesis that minima of the *SimpleFitness* are also minima of the *FullFitness*. As their ranking may be different, the ROI is dynamically updated in order to prevent the sampling to be focused around minima of the *SimpleFitness* that are not relevant according to the *FullFitness*.

Each seeding was assessed according to the randomization of the ligand position and to its conformation prior to starting the docking procedure, since both have a significant impact on the results [64]. To estimate the conformational diversity among seeds, each one was fitted to the ligand crystal structure conformation (used as a reference), and the

corresponding *fitted* RMSD ( $\text{RMSD}_{\text{fit}}$ ) was calculated. It is important to note that the RMSD between seeds and the crystal structure reflect a combination of the randomization of the ligand position and conformation, while the  $\text{RMSD}_{\text{fit}}$  between seeds and the crystal structures reflect the randomization of the ligand conformation alone. For the former, the distinction was made between test cases with an accessible or poorly accessible binding site, and test cases with a buried binding site (Figure 13A). For the latter, a representative example, trypsin/1pph, is shown in Figure 13B.

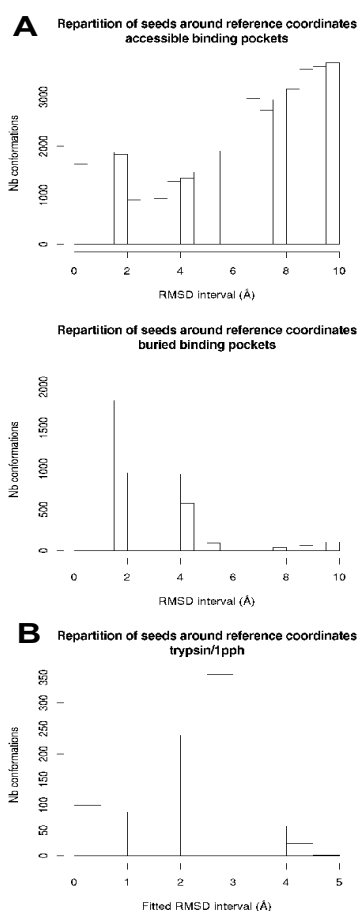


Figure 13: A. Assessment of the distribution of seeding conformations around the crystal structure: histogram of the RMSD between all seeds and the crystal structure for the 28 accessible and the 9 buried test cases. B. Fitted RMSD between 1250 seeds and the crystal structure for trypsin/1pph. The reference conformation ( $\text{RMSD}=0$ ) corresponds to that found in the crystal structure, which is poorly represented.

As can be seen, the native ligand conformation was poorly represented, showing the good randomization of seeds dihedral angles. This thorough randomization makes the docking assay more realistic and difficult [152].

In our approach, clusters are used to establish the connection between the *SimpleFitness* and the *FullFitness*. An example of the structural variability within clusters is shown in Figure 14A. All members of the depicted clusters correspond to a well-defined binding mode.

During evolution, clusters are evaluated by the *FullFitness*, by averaging over the 30% most favorable effective energies of their members. This implies that for each cluster, the distribution of the effective energies of these members is tight enough for their average to be relevant. The standard deviation of these distributions,  $\sigma_{\text{eff}}$ , was measured for all clusters that reached their maximum size (Figure 14B). For 50 %, 80 % and 90 % of the clusters,  $\sigma_{\text{eff}}$  is below 1.29 kcal/mol, 3.93 kcal/mol and 6.36 kcal/mol, respectively. This indicates that the distribution of the effective energies of elements belonging to a cluster are narrow, and that the *FullFitness* calculated for clusters are relevant.

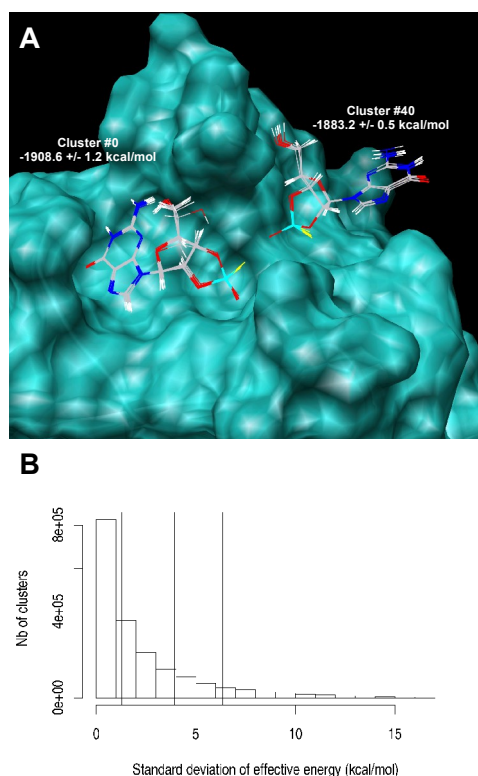


Figure 14; A. variability of coordinates and energies within two top ranked clusters for ribonuclease/1gsp. All molecular graphics images were produced using the UCSF Chimera package [153]. B. Distribution of the standard deviation of the effective energy inside the 1.8 million clusters encountered during 140 docking runs. The vertical lines at  $\sigma_{eff} = 1.29$  kcal/mol,  $\sigma_{eff} = 3.93$  kcal/mol and  $\sigma_{eff} = 6.36$  kcal/mol corresponds to the percentiles 50, 80 and 90, respectively (the 1 % highest standard deviations are not represented)

The intrinsic ability of “smart operators” to generate low energy conformations was assessed. Due to the automatic operator scheduling policy implemented in EADock, this ability is reflected by the probability of an operator to be applied. As described in Methods, the algorithm can adjust the probability of operators from 0.2 to 0.3 depending on the fitness of the children they produced. These probabilities were averaged over the 140 dockings performed. The adaptive probabilities of the smart operators are larger than those of the stochastic operators, illustrating their competitive advantage over standard operators (Figure 15).

In order to measure the impact of smart operators on convergence, two docking assays

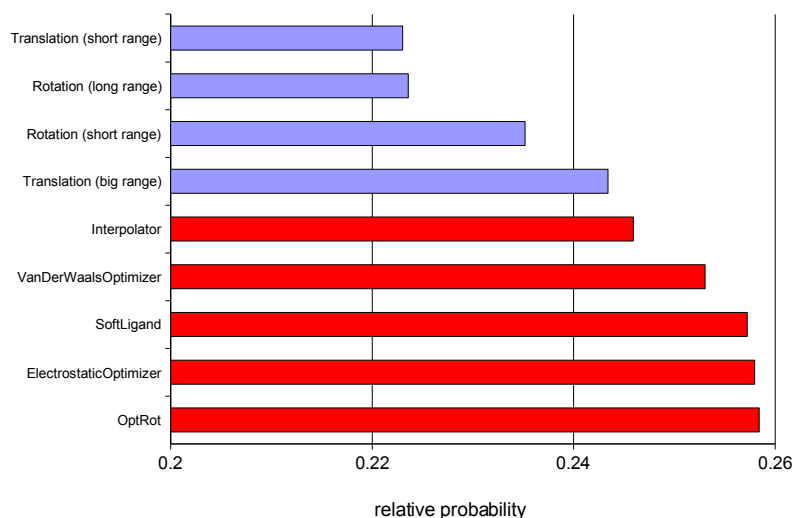
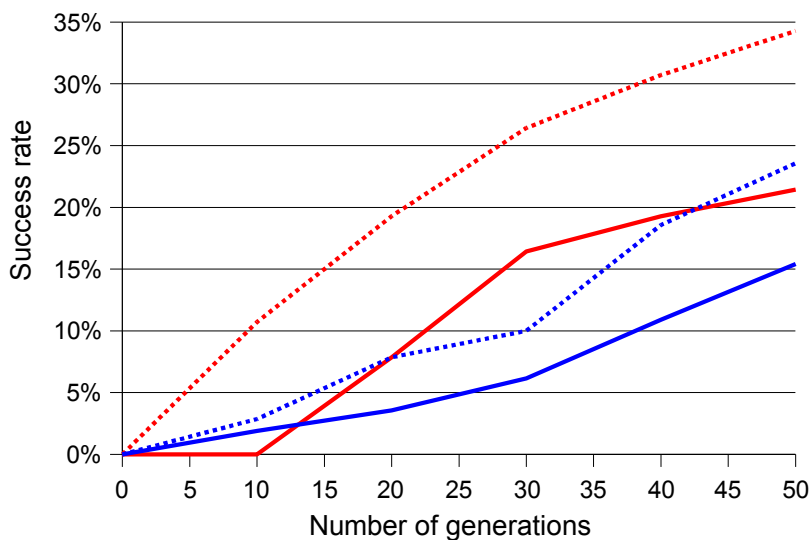


Figure 15: Relative probabilities of operators averaged over the 140 runs. Smart operators (red) significantly outperform classical operators (blue).

were realized with seeds ranging from 8 Å and 11 Å RMSD to the crystal structure, one with classical operators, the other one with both classical and smart operators. The speed of convergence increases with the use of smart operators, especially during the early generations (Figure 16), because of their more realistic physical description of the system. This allows a thorough exploration of the search space that could never been achieved by stochastic operators.

In order to assess the performance of the sampling heuristic, the cluster from the optimized population (combined last generations of the five independent runs, see Methods) with the lowest average RMSD to the crystal structure was retained, whatever its rank and *FullFitness*. The success rate remained high regardless of the seed distribution (Figure 17).



*Figure 16: Success rate considering either the top-ranked cluster (full line) or the five top-ranked clusters (dashed line), with and without smart operators (red and blue curves, respectively), as a function of the number of generations. During the early generations, the success rate is higher with smart operators.*

At least one acceptable binding mode was found within the five top ranked clusters for 89 % of the test cases, except when seeds are generated between 8-11 Å RMSD to the crystal structure (79 %). If all clusters present in the last generation are considered (between 30 and 60 depending on the test case), this increases up to 100 %, 96 %, 93 %, 93 % and 86 % when seeds are within 0-3 Å, 2-5 Å, 4-7 Å, 6-9 Å and 8-11 Å RMSD to the crystal structure, respectively. Noteworthy, RMSD less than 1 Å are reported for 70 % of the cases, and up to 90 % when the RMSD between seeds and the crystal structure ranges from 0 to 3 Å (data not shown).

The algorithm was able to generate at least one conformation within a 2 Å RMSD threshold to the ligand in the crystal structure for 97 % of the 140 docking assays. Only four sampling failures were reported for penicillopepsin/1apt and 1apu, thermolysin/6tmn

and  $\epsilon$ -thrombin/1etr, when starting from the most remote seeds. These four sampling failures reflect a difficulty for the sampling heuristic to generate a reasonable conformation inside a poorly accessible binding pocket (Figure 18). The remaining failures can be attributed to our scoring function, although for most of them, a significant interaction was found between the ligand and a neighboring complex in the crystal (see below).

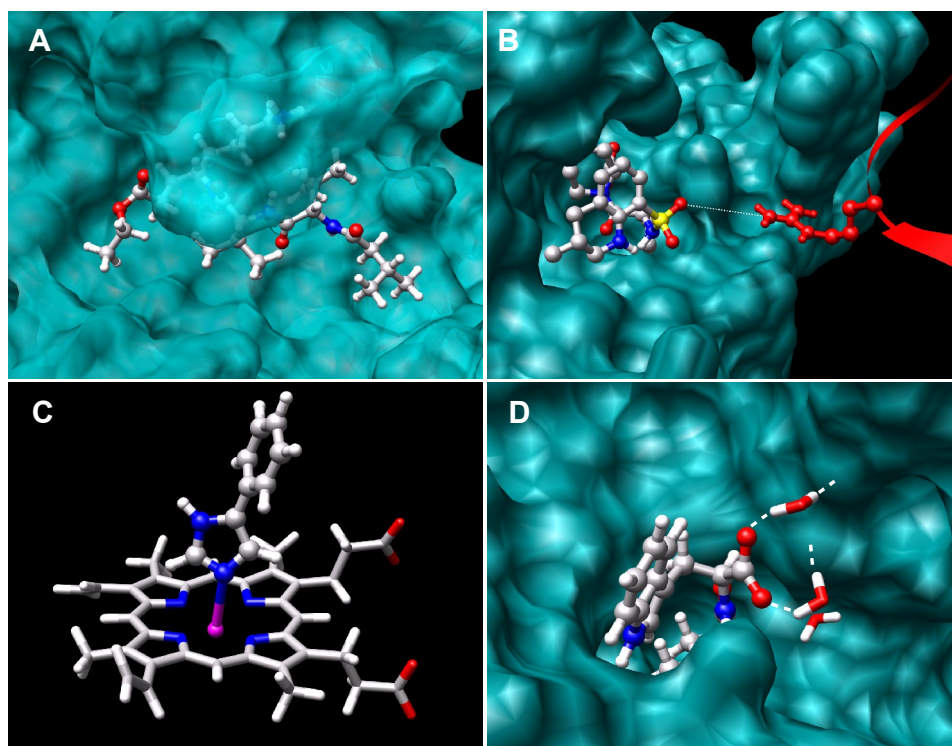


Figure 18: A. view of the single sampling failure reported, corresponding to a poorly accessible binding site (penicillopepsin/1apt), B. example of a typical contact between the ligand and a neighboring complex in the crystal (thermolysin/5tln, 5.2 Å), misleading our scoring functions, C. example of the limits of the FullFitness: bond between the ligand and the heme of the receptor (cytochrome/1phf), misleading the FullFitness. D. example of the limits of the SimpleFitness, with protein/ligand interactions mediated by crystal water molecules (thermolysin/3tmn).

### 6.4.2 Benchmarks

Aside from the algorithm assessment presented above, the predictive performance of EADock was benchmarked for the 37 test cases using a realistic seeding (see Material and Methods) ranging from 3 Å to 10 Å RMSD to the crystal structure. Again, the crystal conformation was excluded from the seeds. All predictions are shown in Table 4.

A cluster with an average RMSD to the crystal structure lower than 2 Å was ranked first for 68 % of the test cases. When considering the five top-ranked solutions or all solutions surviving through the evolution, the success rate increased up to 78 %, and 92 %, respectively. In order to compare the accuracy of the predicted poses to other programs benchmarked in [57], we focused on the 11 test cases corresponding to native docking experiments in both studies. Despite the exclusion of the crystal structure from the seeds, the average RMSD between the best clusters predicted by EADock and crystal structures is 0.75 Å. This is significantly better than what was reported for ICM (1.04 Å), AutoDock (2.46 Å), GOLD (3.31 Å), FlexX (3.85 Å) and DOCK (3.87 Å).



Seeding	Binding site accessibility	PDB codes of complexes	%Rec.Acc.Surf in the ROI	First acceptable cluster			Possible explanation
				Rank	Average RMSD (Å)	$\Delta$ FullFitness (kcal/mol)	
3-10 Å RMSD between seeds and the native binding mode	Accessible	<i>Carbonic anhydrase</i>					
		1cil	10.0%	49	1.08	26.78	Poor description of the interaction between the ligand and a Zn atom
		1okl	9.7%	58	1.57	34.53	
		1cnx	12.3%	46	1.65	55.28	
		<i>Neuraminidase</i>					
		1nsc	8.1%	1	0.52	-	-
		1nsd	8.1%	1	0.69	-	-
		1nnb	8.8%	1	0.74	-	-
		<i>Ribonuclease</i>					
		1gsp	31.1%	1	0.85	-	-
		1rhl	32.0%	1	1.12	-	-
		1rls	32.4%	1	0.98	-	-
		<i>Trypsin</i>					
		3ptb	15.9%	1	0.49	-	-
		1tng	16.1%	1	0.21	-	-
		1tnj	17.5%	1	0.86	-	-
		1tnk	17.2%	1	1.19	-	-
		1tni	17.1%	2	1.98	2.96	Crystal contact
		1tnl	17.0%	1	0.99	-	-
		1tpp	15.0%	1	0.35	-	-
		1pph	15.4%	1	0.49	-	-
	Poorly accessible	<i>Carboxypeptidase</i>					
		1cbx	9.7%	1	0.58	-	-
		3cpa	9.6%	1	0.85	-	-
		6cpa	10.8%	2	0.98	0.88	Crystal contact
		<i>Penicillopepsin</i>					
		1apt	13.3%		Sampling failure		-
		1apu	12.9%	6	0.68	12.03	-
		<i>Thermolysin</i>					
		3tmn	11.5%	1	0.57	-	-
		5tln	11.1%	7	1.86	28.98	Crystal contact
		6tmn	10.1%		Fitness failure		Water mediated interactions
		<i><math>\epsilon</math>-Thrombin</i>					
		1etr	12.6%		Fitness failure		Crystal contact
		1ets	12.0%	1	1.16	-	-
		1ett	12.4%	1	0.79	-	-
3-5 Å RMSD between seeds and the native binding mode	Buried	<i>Cytochrome P-450cam</i>					
		1phf	0.3%	2	1.83	1.62	Bond between the ligand and the receptor
		1phg	0.2%	3	1.65	0.97	
		2cpp	0.2%	1	0.19	-	
		<i>Intestinal FABP</i>					
		1icm	1.4%	1	0.66	-	-
		1icn	1.0%	1	1.87	-	-
		2ifb	0.8%	1	0.76	-	-
		<i>L-Arabinose</i>					
		1abe	0.3%	1	0.18	-	-
		1abf	0.3%	1	0.68	-	-
		5abp	0.3%	1	0.64	-	-

Table 4: Summary of predictions for each test case, presenting the distance between the seeds and the crystal structure, the fraction of the receptor surface included in the ROI, the rank of the first acceptable cluster and the average RMSD between its members and the crystal structure. If the best ranked acceptable cluster is not ranked first, the FullFitness difference with the top ranked cluster is shown. If the first acceptable cluster is not ranked first, a possible explanation is given.

Failures are illustrated in Figure 18. A single sampling failure was observed for penicillopepsin/1apt that has a poorly accessible binding site (Figure 18A). However, when the complex corresponding to the crystal structure is generated, it is successfully identified by the two fitness functions (data not shown).

For trypsin/1tni, the cluster with the most favorable energy is rejected as the RMSD to the crystal structure is 2.1 Å, hardly above the 2 Å threshold defining a successful prediction. This is due to the different conformations of the phenyl ring. Interestingly, the B-factors reported for the corresponding atoms in the original PDB file are high (47.1 Å<sup>2</sup> on average, to be compared with an average of 17.6 Å<sup>2</sup> over all other atoms), suggesting that this part of the ligand is more flexible. In addition, a contact was observed between this phenyl ring and a neighboring complex of the crystal unit cell.

Similar crystal contacts were identified for ε-thrombin/1etr, carbocypeptidase/6cpa and thermolysin/5tln. For the latter, a ionic interaction was found between the ligand and a neighboring complex in the unit cell (Figure 18B).

The ligands of cytochrome/1phf and cytochrome/1phg, are linked to the heme of the receptor, while our molecular mechanics force field-based scoring function does not account for bonds between interacting partners (Figure 18C). These bonds were not reported by Bursulaya et al [57], although discussed in the original article [154].

Other benchmarked programs reported in [57], except ICM, systematically failed at identifying the correct binding modes for all test cases belonging to the carbonic anhydrase family. This could be due to a limitation of the force fields in describing the interaction between the ligand and the zinc ion in the binding pocket [39]. It is worth mentioning that the ROI used with ICM was not described, and it is not possible to exclude that it encompasses only a restricted number of putative binding modes.

Our validation test set was too small to significantly quantify the improvement provided by using GB-MV2 (*FullFitness*) over  $\epsilon=1$  (*SimpleFitness*). However, using this sophisticated implicit solvation model led to successes for 3tmn, 1ets, 1etr and 3cpa. For example, three water molecules mediate interactions between the ligand and the protein

in the 3tmn complex (Figure 18). According to the *SimpleFitness* (i.e. *in vacuo*), at least two of these water molecules are needed in order to rank the crystal structure first (Figure 16B). Without these structural water molecules, the crystal structure corresponds only to a local minimum of the *SimpleFitness*. However, this local minimum is deep enough to be represented by a cluster, which is ranked first according to the *FullFitness* (Figure 16B). This is a strong argument in favor of realistic solvent models such as GB-MV2. Noteworthy, the docking of the 3tmn test case was not successful when the tabu list (see Material and Methods) was not used. This highlights the need to limit the sampling of uninteresting regions of the search space, according to the *FullFitness*, when discrepancies are found between the two fitness functions. In a previous study, only ICM was able to identify this binding mode correctly but, again, its ROI was not described and its sampling might have been limited inside the binding pocket [57].

## 6.5 Discussion

### 6.5.1 Benchmarking docking algorithms

Two questions must be addressed to benchmark a docking algorithm: 1) its ability to generate the good solution through sampling, and 2) its ability to recognize this solution as the correct one by its scoring function. Unfortunately, evaluations of sampling heuristic and scoring function are tightly coupled, as scoring failures cannot be identified if the good solutions are not even generated. Also, the crystal structure is believed to be at least a local minimum of the scoring function. Therefore, if the seeds are too close to the crystal structure or if the region of interest is too small, an algorithm with a weak sampling heuristic is likely to succeed even if its scoring function is deficient, since it will be unable to sample a remote (physically irrelevant) global minimum, and will converge toward the closest local one (not far from the crystal structure). The quality of the algorithm is thus not assessed and the benchmark is not relevant, because successful predictions can either be attributed to a good scoring function or to a lack of sampling [64]. In addition, using seeds too close to the crystal structure also implies that the latter is *a priori* identified, which does not correspond to a real prediction (see list in [64]).

### 6.5.2 Our approach

In order to avoid such problems during the benchmark of EADock, a minimal distance of 3 Å RMSD between the seeds and the crystal structure was enforced after random rotations and translations of the ligand, as well as modifications of its internal degrees of freedom. As a result, to be successful, the sampling heuristic must be able to identify minima deeper than those provided by this high quality seeding.

Such challenging conditions make a direct comparison between EADock and the programs previously benchmarked in [57] on the same validation set difficult for two reasons. First, in the previous study, the crystal structure was kept in the seeds. Second, the ROI used with EADock in our study encompass 14 % of the receptor surface, whereas the previous study defined ROI encompassing only 10 % (AutoDock), 4.6 % (Gold) and 1.7 % (DOCK and FlexX). While such limited ROI are not suitable for the benchmark of docking programs (see above), they are relevant if the binding pocket is known prior to the docking study.

Despite unfavorable conditions, the search realized by EADock converges within only 50000 *SimpleFitness* and 15000 *FullFitness* evaluations on average, depending on the evolutionary path explored. This highlights the efficiency of our sampling heuristic, as it can be compared to the 2500000 poses evaluated by AutoDock in [57]. Another recent study based on a different test set, reported between 200000 and 400000 evaluations to converge [138], depending on the software used. This also corresponds to the number of fitness evaluations usually observed [155]. This good performance of the sampling heuristic of EADock is also supported by the successful docking of all ligands if the RMSD is used as a fitness, no matter the distance between the seeds and the crystal structure (data not shown). This points out that the success rate of EADock is limited by our scoring functions. Interestingly, most scoring failures may be explained by crystal contacts or bonds between the ligand and the receptor.

The estimation of the binding free energy requires that the predicted and the native binding mode are as close as possible. When considering successful predictions, the average RMSD between crystal structures and binding modes proposed by EADock is only

0.75 Å. This gain over other approaches is expected to play a key role for rational lead optimization.

To demonstrate the efficiency of our approach under extreme conditions on a real application, the RGD cyclic pentapeptide was docked on the  $\alpha V\beta 3$  integrin. The same docking parameters were used, except that the ROI was set to a sphere with a radius of 25 Å centered on the binding site, encompassing 65000 Å<sup>3</sup>. Seeding conformations were generated far away from the binding pocket, between 15 Å and 25 Å RMSD to the crystal structure. Despite these difficult conditions, EADock is able to identify the crystal structure, with a RMSD of only 1.17 Å (Figure 19).

This points out the efficiency of our sampling heuristic and scoring function in a real application, and opens the field of a rational design of active compounds derived from Cilengitide, which is of major interest given its clinical impact [156].

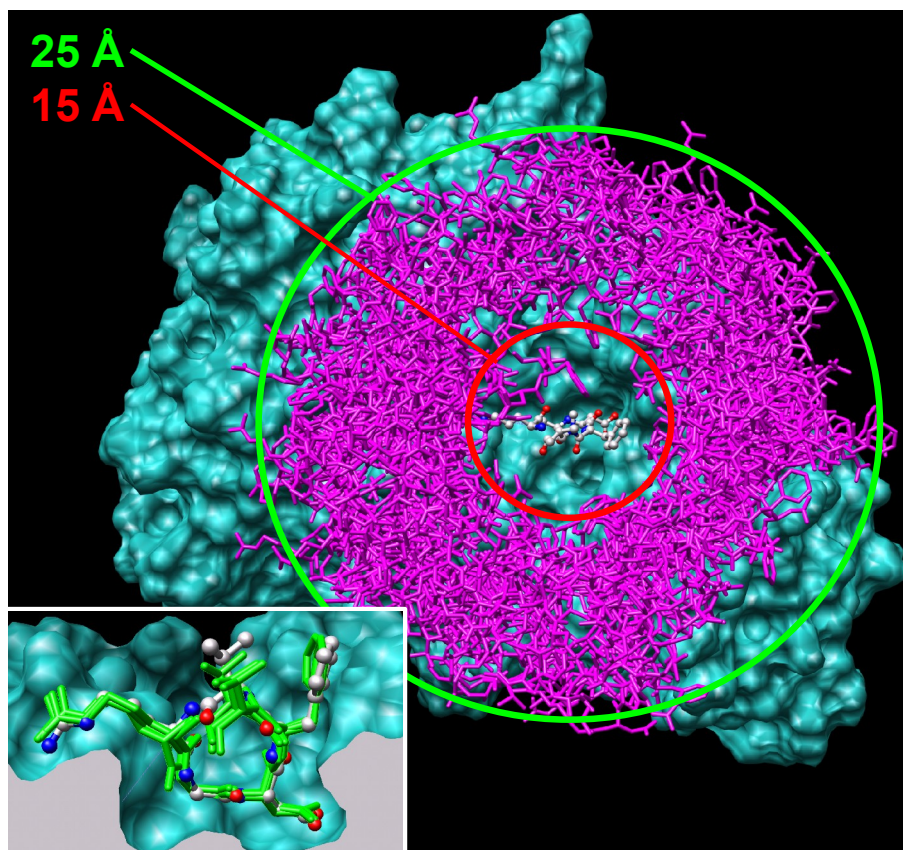


Figure 19: Docking of a RGD cyclic pentapeptide on the  $\alpha V\beta 3$  integrin. Seeds (pink sticks) were generated between 15 Å and 25 Å RMSD to the crystal structure (red and green circles, respectively). The ROI was defined by a 25 Å radius sphere encompassing 65449 Å<sup>3</sup>. The binding mode of the ligand found in the crystal is shown in ball and stick (inset). A standard docking was performed, and the top ranked cluster was found to have a binding mode identical to that of the crystal structure, as shown by the three top scoring conformations (green sticks). The average RMSD to the crystal structure for these three conformations is 1.17 Å.

All our results are obtained by the combination of a multi-objective optimization taking into account the solvation free energy of the complex, and an efficient sampling heuristic able to converge with a very limited number of docking modes evaluation. Taking into account the solvation free energy with a force-field based scoring function has three drawbacks: it is too sensitive to small variation in the atomic coordinates, it does not provide a clear driving force, and it is computationally demanding. Our docking algorithm proposes convenient solutions for each of these three limitations. First, variations of the

effective energy according to the coordinates are smoothed by an averaging over several docking modes within the same cluster. Second, a driving force is introduced into the evolutionary process: the *SimpleFitness* is used as a filter for the identification of a reasonable docking pose. Third, as the *SimpleFitness* is fast to compute, it is a first approximation to the expensive *FullFitness*. These three improvements allow EADock to use a fitness function based on the solvation free energy calculated with the CHARMM package. This provides several advantages over other approaches such as universality (i.e. it is not limited to docking on proteins) and its average description of the solvent effect. The use of this universal scoring function is believed to increase the transferability of our results to other receptor/ligand families (e.g. ligand/DNA) as well as other docking strategies like, for example, molecular fragments. The latter opens the door to a unified approach toward FB-RDD.





## 7 Applications

### 7.1 Overview

The ability to identify the most favorable binding mode of small molecules to a biologically relevant target opens the door to *in silico* structure-based drug design. Aside from the methodological development described in the previous chapter, collaborations with experimentalists were carried out in various fields and are described in this chapter. These collaborations fall into the following categories:

1. understanding the molecular principles governing the regulation of proteins. Two examples of major importance, the Na,K-ATPase and the Peroxisome Proliferator-Activated Receptor  $\alpha$  (PPAR $\alpha$ ) were studied. Among other roles (see below), the main task of the Na,K-ATPase is to keep the Na<sup>+</sup> and K<sup>+</sup> gradients that permit the maintenance of the cell volume and the membrane potential. The latter is a nuclear hormone receptor regulating the transcription of genes with consequences on the metabolism of lipids.
2. understanding the action of known compounds. The impact of biotransformation on the binding mode of the anticancer drug Imatinib to the c-Abl kinase, and the molecular mechanism underlying the interaction of two common pollutants with the PPAR $\gamma$  were studied.
3. discovering and optimizing new active compounds. Three studies targeting the human PPAR $\alpha$ , the indoleamine deoxygenase (IDO) and the integrin  $\alpha 5\beta 1$  and  $\alpha V\beta 3$  were performed. The first is the target for the lipid lowering fibrates, and the two latter are recent and interesting targets regarding to cancer therapy [157]. A FB-RDD approach (see Chapter 1 “Introduction”) was developed with Vincent Zoete. This approach is built around EADock, used in combination with several other software pieces, and is briefly presented in this chapter.

The use of EADock in real applications had a major impact on its development, leading to

progressive improvements in terms of functionality, stability and usability. On the technical side, because of the different aims and biological systems investigated during these collaborations, the implementation of EADock had to be revisited several times in order to gain the required versatility. For instance, specific operators were developed for a docking study of the FXYP7 transmembrane helix that is involved in the regulation of the Na,K-ATPase pump; pseudo Morse potentials were added to both fitness to mimic the creation of a covalent bond between the IDO and its ligands (this improvement is still under investigation and not presented here, see Chapter 4 “Perspectives”); and the flexibility of the receptor was taken into account successfully when docking into the *Xenopus* PPAR $\alpha$  its known regulator Wy-14,643.

## **7.2 Understanding molecular principles**

### **7.2.1 Regulation of the Na,K-ATPase**

This work has been published in the Journal of Biological Chemistry [158].

#### **7.2.1.1 Biological context**

Na,K-ATPase transports three Na<sup>+</sup> against two K<sup>+</sup> across the plasma membrane of animal cells by using the energy of the hydrolysis of ATP. Its main task is the maintenance of the transmembrane Na<sup>+</sup> and K<sup>+</sup> gradients that permit the maintenance of the cell volume and the membrane potential. Moreover, Na<sup>+</sup> gradients provide the energy for many secondary transport systems of vital importance. In addition to these basic functions, Na,K-ATPase is involved in many specialized tissue functions such as transepithelial Na<sup>+</sup> transport, and muscle and neuronal excitability. Due to its important physiological role, Na,K-ATPase is finely regulated. Established regulatory mechanisms include changes in the intracellular Na<sup>+</sup> concentration that produce short term regulation of the Na,K-ATPase transport rate, phosphorylation of the subunit by cAMP-dependent protein kinase and protein kinase C that influences the distribution of Na,K-ATPase between the plasma membrane and intracellular stores, and long term regulation that increases the number of Na,K-ATPase units [159]. Recently, a novel regulatory mechanism has been identified that involves tissue- and isozyme-specific interactions between Na,K-ATPase and small membrane proteins of the FXYD protein family [160]. The FXYD protein family contains seven members that are characterized by one transmembrane domain and a signature sequence that contains the FXYD motif and 3 other conserved amino acids [161]. FXYD2 or the subunit of Na,K-ATPase was the first FXYD protein that was identified as a specific modulator of renal Na,K-ATPase [162] [163] [164] [165] [166]. It is now well established that also FXYD1 [167], a phospholemman-like protein from shark [168], FXYD4 (corticosteroid hormone-induced factor) [164] [169], and FXYD7 [170] also play a tissue-specific role in Na,K-ATPase regulation. Significantly, each of these auxiliary subunits produces a distinct functional effect on Na,K-ATPase that is adapted to the physiological

needs of the tissues in which they are expressed. The functional effects of FXYD proteins on Na,K-ATPase have been studied extensively, but the molecular basis of these effects is unknown, and very little is known on the interaction sites in the Na,K-ATPase and the FXYD proteins that mediate the efficient association between these two proteins and transmit the functional effects of FXYD proteins on Na,K-ATPase. Experiments based on thermal denaturation suggest that the association of FXYD2 occurs with transmembrane (TM) domains 8–10 (see Figure 21) [171]. Moreover, a recent model [172], deduced from an electron crystallographic analysis at 9.5 Å resolution of renal Na,K-ATPase and by taking as a basis the high resolution structure of the Ca-ATPase [173], predicts that FXYD2 is located in a pocket made up of TM9, TM6, TM4, and TM2 of the Na,K-ATPase subunit. In this study, we investigated the role of TM9 of the Na,K-ATPase subunit in the structural and functional interaction with FXYD proteins. For this purpose, we produced a model of the Na,K-ATPase subunit to determine amino acids in the TM9 helix, which point to TM2 and which could potentially interact with FXYD proteins.

#### **7.2.1.2 Modeling approach**

Based on our previous homology model of the Bufo Na,K-ATPase [174] in the E1 conformation, a model of the FXYD7/Na,K-ATPase complex was built using an early version of EADock [70]. In brief, starting from an arbitrary conformation of the FXYD7 helix in the vicinity of TM2 and TM9 of the Bufo Na,K-ATPase subunit, the FXYD7 coordinates were refined using two operators translating or rotating the FXYD7 helix around a random axis. Two other operators were designed to rationalize the search, performing rotations around or translations along the axis of the modeled fragment of FXYD7 (Figure 20A and Figure 20B, respectively). The fifth and last operator was a semistochastic *Interpolator* combining two high scoring complexes to generate a new position and orientation of the FXYD7 fragment (Figure 20C).

After each operator was applied, a short energy minimization of the FXYD7 helix as well as TM9 and TM2 residues was performed using the CHARMM program. The minimization consisted of 30 steps of steepest descents followed by 50 steps of Adopted Basis Newton-Raphson. The fitness of a complex was defined as its total enthalpic energy calculated by

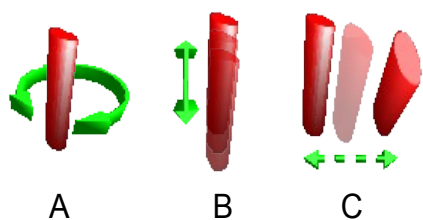


Figure 20: Operators designed for the docking of the FXYP7 into the Na,K-ATPase. See text for details

CHARMM using the CHARMM22 [43] parameters with a soft van der Waals potential. An implicit lipid layer was added by means of a dielectric switch in the Generalized Born-Simple Switch solvation model [175] [176] with its default parameter set. No mutation data were used in the conformational space sampling nor in the ranking of the complexes. The best scoring complexes generated during the evolutionary search were clustered, based on heavy atom root mean square difference values, and a contact list was generated for each cluster.

### 7.2.1.3 Results

Two clusters were identified: the lowest energy cluster A (83% of the conformers), and cluster B (17% of the conformers) with a mean energy of +63 kcal/mol compared to cluster A. Both clusters share the same anchoring of FXYP7 in the cleft near the extracellular space, with contacts involving residues Gln26 (cluster A) and Thr27 (cluster B) of FXYP7, with Phe967 from TM9 (Figure 21). FXYP7 interacts with residues Ile953, Phe956, Glu960, Leu964, and Phe967 and Leu968 from TM9 and Tyr149, Ile143, Arg156, and Ile157 from TM2. Two residues of FXYP7 (Met30 and Phe37) are overrepresented in the contact list. Met30 is close to both Leu964 and Leu968 (TM9) and Ile142 (TM2). Two alternative rotamers for Phe37 were isolated in each cluster with favorable interactions with either Phe956 (TM9) or Tyr149 (TM2). Next to Phe37, Val38 also interacts with Tyr149 (TM2). Two hydrophobic residues of FXYP7, Ile44 and Leu45, fill the widest part of the TM9 –TM2 groove, stabilized by several contacts with Ile953 and Ile157 (structures from cluster B) or Arg156 (aliphatic part of the side chain, structures from cluster A).

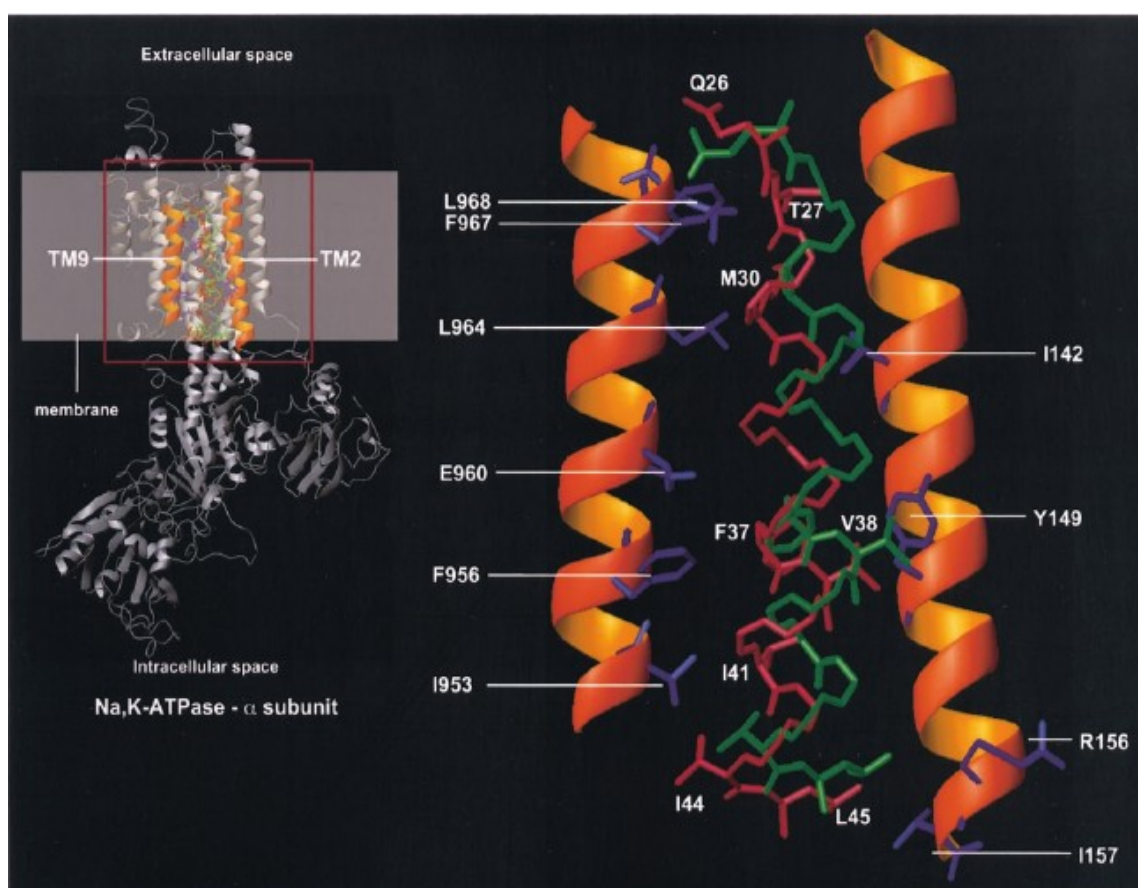


Figure 21: Docking results. Left, overview of the Na,K-ATPase subunit, showing TM2 and TM9, and two average structures from cluster A (green) and B (red). The location of the membrane is represented by a transparent rectangle. Right, detailed view of the interaction between TM9 (left) and TM2 (right) and two average structures of FXYD7, from cluster A (green) and B (red). Heavy atoms of residues involved in the interaction are shown in blue for the Na,K-ATPase subunit and in green or in red for cluster A and B.

FXYD7 residues	Interaction with Na,K-ATPase subunit	
	Cluster A	Cluster B
Gln26	Phe967	Phe967
Thr27		Phe967
Met30	Ile142/Leu964	Ile142/Leu964/Leu968
Phe37	Tyr149/Phe956	Tyr149/Phe956/Glu960
Val38	Tyr149	Tyr149
Ile41		Ile953
Ile44	Ile953	Ile953
Leu45	Arg156	Ile157

Table 5: Contact list derived from the docking calculations. See text for details.

#### **7.2.1.4 Conclusion**

The model of FXYD7 docked with the Na,K-ATPase subunit in the E1 conformation is in excellent agreement with the experimental data [158]. All residues in TM9, Ile953, Phe956, Glu960, Leu964, and Phe967, identified by the mutagenesis analysis are involved in contacts between FXYD7 and the Na,K-ATPase subunit in the model. In addition, the model predicts Ile142, Tyr149, Arg156, and Ile157 in TM2 and Gln26, Met30, Val38, Phe37, Ile41, Ile44, and Leu45 in FXYD7 as interaction sites.

Interestingly, the stabilizing interactions involving Leu964 and Phe967 in TM9 are similar in both clusters. Gly29 and Gly40 in FXYD7, the substitution of which has previously been shown experimentally to significantly affect the association efficiency with Na,K-ATPase [177], were not predicted to form favorable contacts by the docking. It remains to be shown whether substitution of these glycine residues could perturb correct folding of FXYD proteins, thus limiting complex formation. Alternatively, it remains to be investigated whether there might be a correlation between the potential implication of Gly40 in oligomer formation [178] and the association efficiency of Gly40 mutants in FXYD7 as well as in FXYD2 [179]. Finally, the residues suggested by the docking study only, such as, Tyr149, Ile157, Arg156, Ile142, and Leu968 in the Na,K-ATPase subunit, and the predicted amino acids in FXYD7, are good candidates for further mutagenesis analysis.

### **7.2.2 Regulation of the nuclear hormone receptor PPAR $\alpha$**

This work was carried out by Vincent Zoete, Aurélien Grosdidier and Pierre Chodanowski during a collaboration with the groups of Liliane Michalik and Walter Wahli (UNIL-Center for Integrative Genomics). It has been published recently [180].

#### **7.2.2.1 Biological context**

Peroxisome proliferator-activated receptors (PPARs) are members of the nuclear receptor family. As ligand-dependent receptors, PPARs form heterodimers with the RXR and adopt an active conformation in the presence of an agonist [181]. Natural ligands of PPARs

include fatty acids (FA) and eicosanoids. Additional co-activator proteins are recruited to create a complex that binds to peroxisome proliferator response elements (PPRE) in target genes and stimulates their expression. In addition to this canonical mechanism, it has been recently found that PPARs can also function independently, in the absence of a hetero-partner [182]. Three isotypes of human PPAR, called  $\alpha$ ,  $\gamma$  and  $\beta/\delta$ , have been characterized [183], showing distinct tissue distributions, physiological roles and ligand specificity. These aspects have been intensively reviewed in recent years [183] [184] [185] [186] [187] [188]. In brief, PPAR $\alpha$  is found in liver, kidney, heart, and muscle. It is important for the uptake and oxidation of FA and lipoprotein metabolism. PPAR $\alpha$  is the target for the lipid lowering fibrates. PPAR $\gamma$  is localized in fat, large intestine, and macrophages. It plays an important role in adipocyte differentiation and is the receptor for a well-known class of antidiabetic insulin sensitizers drugs, the thiazolidinediones (TZD), such as rosiglitazone. PPAR $\delta$  is expressed in most cell types. PPAR $\delta$  agonists play important roles in dyslipidemia, cancer treatment, and cell differentiation within the central nervous system.

A “mouse trap” model of the ligand-dependent transcription factors activation has been proposed by D. Moras and coworkers and is now widely accepted [189]. In this model, the AF-2 helix H12 closes on the ligand binding site in response to ligand binding and the resulting active form of the receptor can bind a co-activator.

The study of ligand binding domains (LBDs) suggest that the receptors can adopt the active conformation even in the absence of agonists [190] [191], in agreement with the existence of a basal activity in absence of ligand [192]. All these data indicate that PPAR LBD acts as a dimmer switch and does not adopt a well-defined structure in the absence of ligand, but rather shows an equilibrium of conformations [193] [194]. Ligand binding would shift this conformational equilibrium to a state that favors co-activator recruitment, through direct contacts between the ligand and the AF-2 domain and a global stabilization of the LBD (see Figure 22). The dynamic properties of helix 12 are a major determinant of AF-2 domain activity, and were investigated using *in silico* approaches, in presence and in absence of the ligand, in the Xenopus PPAR $\alpha$  (xPPAR $\alpha$ ).



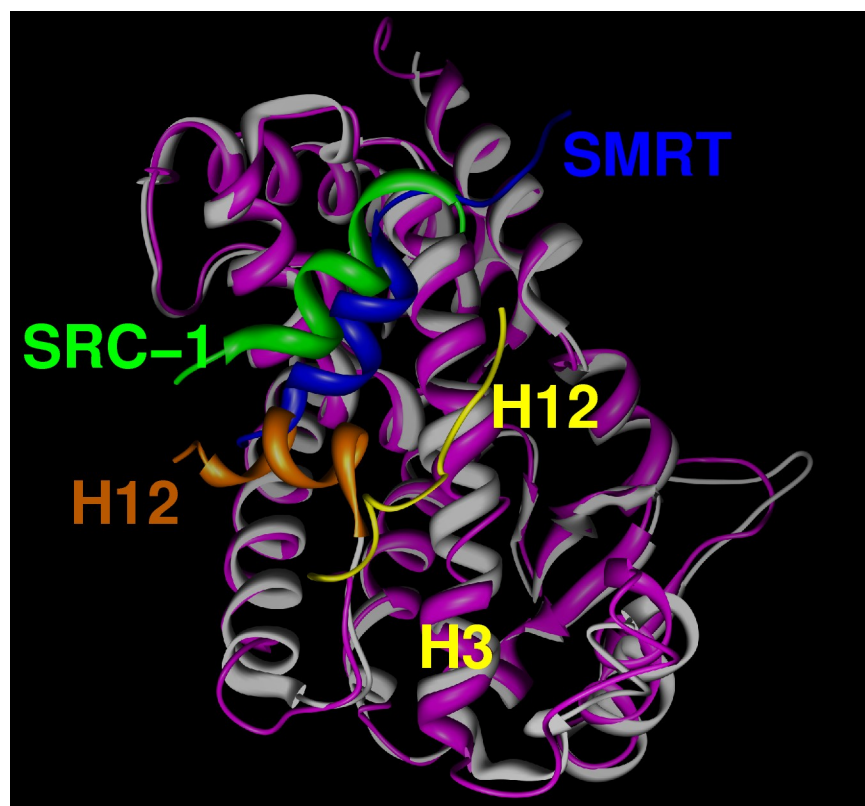


Figure 22: Structural superposition of PPAR $\alpha$ /GW409544/SRC-1 [195] (white) and PPAR $\alpha$ /GW6471/SMRT [196] (magenta). In the PPAR $\alpha$ /GW409544/SRC-1 structure, the AF-2 helix H12 in its active conformation and the SRC-1 co-activator are orange and green, respectively. In the PPAR $\alpha$ /GW6471/SMR structure, the AF-2 helix H12 in an inactive conformation and the SMRT co-repressor are yellow and blue, respectively. The co-repressor occupies partly the position of the active conformation of AF-2.

A recent review about PPAR structure in relation with ligand specificity, molecular switch and interactions with regulators is available in [197].

### 7.2.2.2 Modeling approach

#### 7.2.2.2.1 Deriving the apo xPPAR $\alpha$ model

Since the experimental structure of xPPAR $\alpha$  in complex with Wy-14,643 is not available, an homology model of the xPPAR $\alpha$  ligand binding domain (SwissProt accession number P37232) was built using MODELLER v6.2 [198], using the human PPAR $\alpha$  (PDB code 1K7L) as a template. In this structure, helix 12 is in the closed and active conformation.

CLUSTALW was used to perform a pairwise sequence alignment of the *Xenopus laevis* and human sequences. The sequence identity between the two molecules is 90%. Default parameters of the homology modeling routine were used. The energy of the model was then minimized using the CHARMM program and the CHARMM19 force field, with a dielectric constant of 1 and a 20 Å cutoff. The minimization consisted of 30 steps of Steepest Descent (SD) followed by 30 steps of Adopted Basis Newton-Raphson (ABNR). The positions of the C $\alpha$  atoms were constrained using a mass weighted harmonic force constant of 10 kcal/(mol Å<sup>2</sup>).

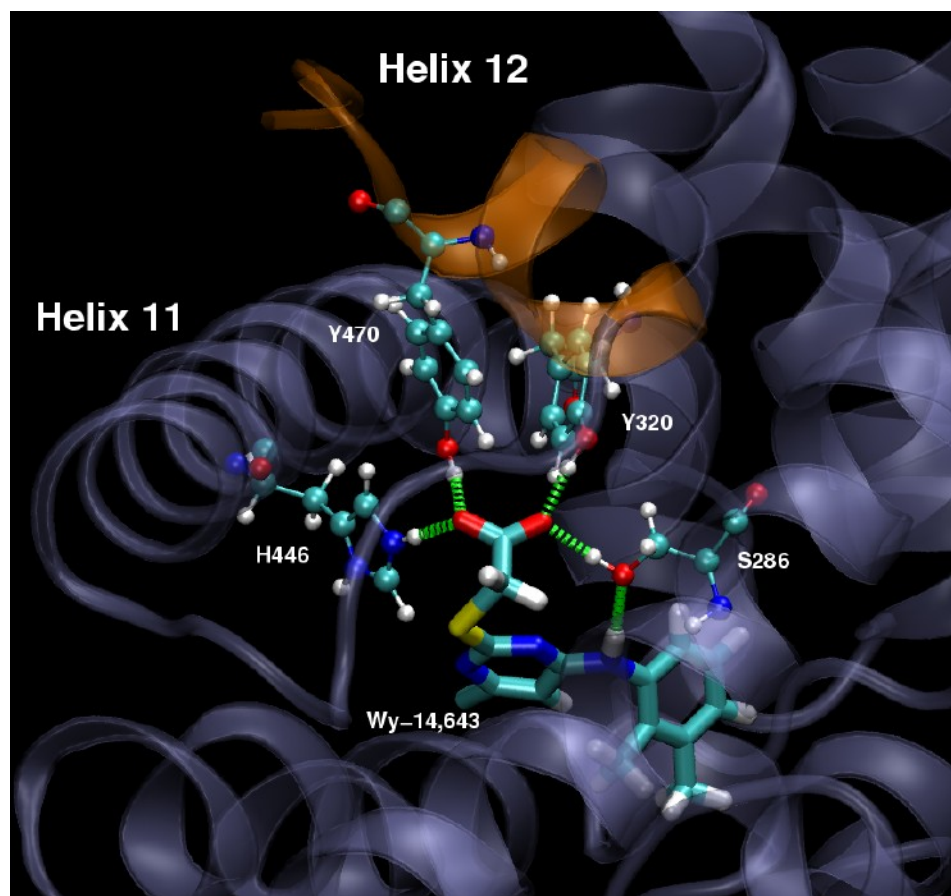
#### **7.2.2.2.2 Docking of Wy-14,643**

Missing parameters for the ligand, for use in conjunction with the CHARMM22 all atoms molecular mechanics force field, were derived from the Merck Molecular Force Field (MMFF), by taking the dihedral angle term as is, but only the quadratic part of the bond and angle energy terms. The partial charges and van der Waals parameters of the ligand atoms were taken from the MMFF. The ligands were modeled with all hydrogens.

Based on the xPPAR $\alpha$  model mentioned above, a model of the xPPAR $\alpha$ /Wy-14,643 complex was built using EADock [70]. Details of the calculations are presented in chapter 2. In brief, starting from a set of 250 randomly generated initial conformations, positions and orientations of Wy-14,643 inside the known binding site of xPPAR $\alpha$ , the coordinates of the ligand were refined using several operators, renewing 10% of the population at each generation. Thorough exploration of the accessible conformational space of the ligand relative to the protein surface was submitted to the evolutionary pressure of a scoring function that takes account of the solvent effect thanks to the GB-MV2 implicit solvent model. Residues of the binding site were flexible during the docking to account for the inherent inaccuracy of coordinates in the xPPAR $\alpha$  model, and of the induced fit of the protein in the presence of the ligand. These include residues 247, 253, 257, 278-279, 281-283, 285-286, 320, 323-324, 327, 336, 338, 345, 360-361, 446, 450 and 470. After 400 generations, the conformations with the lowest energy were further minimized by 100 steps of SD using the GB-MV2 generalized Born model. The lowest energy conformation was used for further molecular dynamics simulation.

### 7.2.2.3 Results

A schematic representation of the hydrogen bonds and the van der Waals interactions between xPPAR $\alpha$  and Wy-14,643 in the calculated binding mode is given below in Figure 23.



*Figure 23: 3D representation of hydrogen bonds between xPPAR $\alpha$  and Wy-14,643, in the binding mode proposed by EADock. The figure was done with the VMD program.[199]*

The polar carboxylate function of Wy-14,643 makes hydrogen bonds with the side chains of residues Ser286, Tyr320, His446 and residue Tyr470 of Helix 12. Such a hydrogen bond pattern has already been found experimentally, for example between the carboxylate head of the AZ242 [200], of the TZD head group of the rosiglitazone (PDB code 1FM6) and GW409544 [195] ligands and the corresponding polar residues of the human PPAR protein: Ser280, Tyr314, His440 and Tyr464. Other studies aiming at docking theoretically some PPAR ligands with the FlexX [47] program also found similar results

[201] [202]. An additional hydrogen bond takes place between the NH aniline function of Wy-14,643 and the side chain OG atom of Ser286 (Figure 23). The hydrophobic tail of Wy-14,643 extends in the hydrophobic pocket of the LBD, where it makes van der Waals contacts with Phe279, Cys282, Thr285, Thr289, Met323, Phe324, Leu325, Val340, Met336, Val338, Met361 and Val450.

#### **7.2.2.4 Conclusion**

Starting from this complex model, the role of the ligand binding on the protein stability, has been investigated using the approach developed by V. Zoete and M. Meuwly [203]. The method is based on the notion that the binding free energy corresponding to the alchemical complexation of a given side chain (considered as a “pseudo-ligand”) into the rest of the protein (considered as a “pseudo-receptor”) reflects the importance of this side chain to the thermodynamic stability of the protein [203]. This method could explained experimentally determined variations in PPAR activity upon mutation of some helix 12 residues, and reversely, it pointed out important additional residues that were confirmed experimentally.

In order to reveal the key principle behind helix 12 regulation, we combined the functional study of PPAR $\alpha$  mutant transcriptional activity and MD simulations. These two approaches used very different observation time windows, the former analyzing the steady state receptor activity, and the latter local changes at the nanosecond scale. The findings of the molecular modeling simulations have triggered functional validation experiments and, reciprocally, intriguing experimental results suggested new molecular modeling simulations that helped in their interpretation. The complementation of the two approaches and the concordance in the results enabled us to draw robust conclusions regarding the molecular mechanisms governing PPAR $\alpha$  helix 12 regulation. See [180] for further details.

### **7.3 Understanding the action of known compounds**

#### **7.3.1 Docking of DEHP/MEHP on the nuclear hormone receptor PPAR $\gamma$**

This work has been carried out by Vincent Zoete and Aurélien Grosdidier, during a collaboration with the group of Béatrice Desvergne (UNIL-Center for Integrative Genomics). It has been submitted recently.

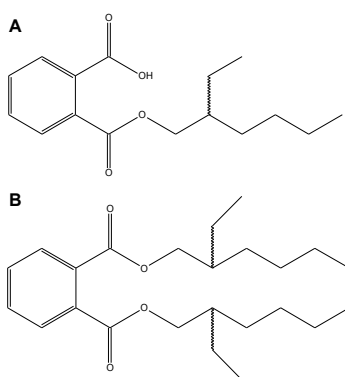
##### **7.3.1.1 Biological context**

Exposure to endocrine disrupting chemicals (EDCs) can lead to detrimental effects in human and animal populations by interfering with the synthesis, the elimination and the mechanisms of action of hormones. At the molecular level, these compounds act by activating or inhibiting enzymatic activities of hormone biosynthesis and by targeting nuclear receptors (NRs), a class of transcription factors that regulate gene expression programs in response to lipophilic hormones and mediators. NRs constitute a wide family of receptors regulating diverse physiological actions, out of which most members share the capacity to regulate gene expression in response to ligand binding.

Metabolism constitutes an aspect of the maintenance of homeostasis that requires the cooperation of various organs in order to control the balance between energy storage and utilization according to the nutritional status and the needs of the organism. This balance is regulated at the transcriptional level by the integrated action of nuclear receptors as well as other transcription factors. The prevalence of the metabolic syndrome has dramatically increased during the past decades and it has been suggested that predisposition to obesity could be acquired during fetal development, both through nutritional supply and exposition to environmental factors.

Given their central role in metabolic regulations, PPARs potentially constitute important targets for environmental factors. The large PPAR ligand-binding pocket that can accommodate a wide variety of ligands [197] raises the question of whether PPAR activity and PPAR-regulated pathways could be affected by an exposure to EDCs. We have focused

the present study on the interference of phthalate esters with PPAR regulated processes. Phthalates are widely used industrial chemicals which primarily serve as plasticizers to soften PVC but are also found in cosmetics, perfumes and certain drugs as well as in industrial paints and solvents. Di-Ethyl-Hexyl-Phthalate (DEHP, Figure 24) is among the most abundantly used phthalate esters with an annual worldwide production estimated around 2 million tons according to Swiss authorities (Federal Office of Public Health<sup>4</sup>). DEHP is incorporated non-covalently into flexible plastics used for manufacturing a wide variety of daily products including medical devices and food packaging and its propensity to leach can lead to high levels of human exposure. The biological effects of DEHP are hence of major concern but so far elusive. Upon ingestion, pancreatic lipases present in the intestine convert DEHP to its monoester equivalent Mono-Ethyl-Hexyl-Phthalate (MEHPP, Figure 24) which is preferentially absorbed. In addition, MEHP can also be produced by plasmatic and hepatic lipases, which transform DEHP directly reaching the blood through absorption or medical contamination. This metabolite activates the three PPAR isotypes and mediates the action of DEHP on hepatic peroxisome proliferation via PPAR $\alpha$ . In this study, we focused on the mechanisms through which MEHP interferes with PPAR $\gamma$  signaling as mentionned in 7.2.2.1. PPAR $\gamma$  is localized in fat, large intestine, and macrophages. It plays an important role in adipocyte differentiation and is the receptor for a well-known class of antidiabetic insulin sensitizers drugs, the thiazolidinediones (TZD).



*Figure 24: chemical structures of the Mono-Ethyl-Hexyl-Phthalate (MEHP, A) and the Di-Ethyl-Hexyl-Phthalate (DEHP, B)*

<sup>4</sup> <http://www.bag.admin.ch/themen/chemikalien/00228/01378>

#### **7.3.1.2 Modeling approach**

The binding of R- and S-MEHP was modeled using EADock, based on two structures of hPPAR $\gamma$  complexed to AZ242 (PDB reference 1I7I) and to an  $\alpha$ -aryloxyphenylacetic acid partial agonist (1ZEO) in order to take into account a possible induced fit of the protein by ligand complexation. In brief, starting from a set of 250 randomly generated initial positions of MEHP in the PPAR $\gamma$  binding site, the coordinates of the ligand were refined using several operators, renewing 10% of the population at each generation. The docking was stopped after 400 generations and the conformation with the lowest energy was retained.

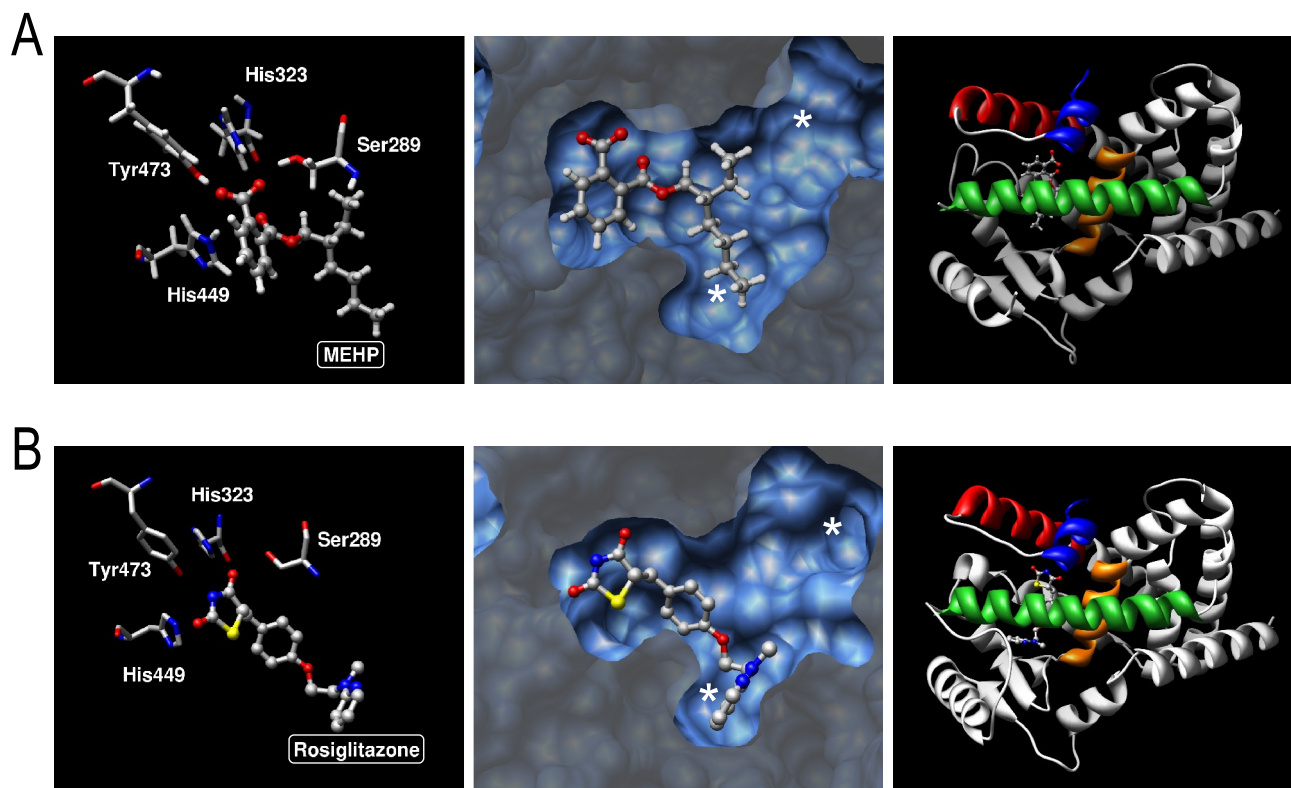
#### **7.3.1.3 Results**

In order to characterize whether differences in binding mode could potentially account for the different activation properties of MEHP and Rosiglitazone, we modeled the binding of MEHP to two structures of the PPAR $\gamma$  ligand binding pocket (LBD) and compared it to the binding of Rosiglitazone available from the crystal structure of the PPAR $\gamma$  LBD in complex with this agonist. The results obtained for the docking of MEHP in the 1I7I and 1ZEO structures of hPPAR $\gamma$  are very similar and both the R- and the S-enantiomer of MEHP could fit in the PPAR $\gamma$  LBD (data not shown). Although explored by the ligand during the docking process, the additional pocket of 1ZEO was not used in the proposed binding mode (see supplementary information). MEHP contacted S289, H323, H367 and Y473 (A), a set of residues important for the stabilization of the interaction between Rosiglitazone and the receptor (B). Furthermore, the contact between the carboxylate function of the phthalic acid ring and Y473, a residue from helix 12 important for transcriptional activation, suggests that the activity of MEHP relies on the stabilization of helix 12.

#### **7.3.1.4 Conclusion**

These observations support the absence of activity of the DEHP parent compound where this carboxylate is esterified by a bulky and hydrophobic chain. MEHP and

Rosiglitazone bind to the PPAR $\gamma$  LBD in similar configurations where only one side of the T-shaped binding pocket is occupied and where similar residues are contacted. Thus, the difference in affinity and in efficacy between MEHP and Rosiglitazone most likely reflects subtle variations in the binding mode, which lead to less productive conformational



*Figure 25: MEHP and Rosiglitazone bind similarly to the PPAR $\gamma$  ligand binding domain. The binding of the R enantiomer of MEHP to the human PPAR $\gamma$  LBD (structure 1I7I) was modeled as described in the material and methods section (A) and compared to the reported structure of the hPPAR $\gamma$  LBD complex with Rosiglitazone (B). Left panels represent interactions with key residues of the LBD. Middle panels describe the positioning in the LBD cavity where asterisks represent the two parts of the T-shaped ligand binding pocket. The right panels show the position of the ligand in the secondary structure of the receptor. Helices contacting the ligand are colored as follows: H3, green; H5, orange; H11, red and H12, blue. Note that hydrogen atoms of Rosiglitazone are not represented because not present in the PDB structure.*

changes upon MEHP binding. However, the full-characterization of the differential



changes in the three-dimensional structure of the LBD would require the crystallization of the PPAR $\gamma$  LBD in complex with MEHP.

### **7.3.2 Impact of the biotransformation of the Imatinib on its binding mode**

This work has been carried out by Vincent Zoete and Aurélien Grosdidier, during a collaboration with Bertrand Rochat, responsible for the quantitative Mass Spectrometry Facility shared by the Faculté de Biologie et Médecine of the Université de Lausanne. It has been submitted recently.

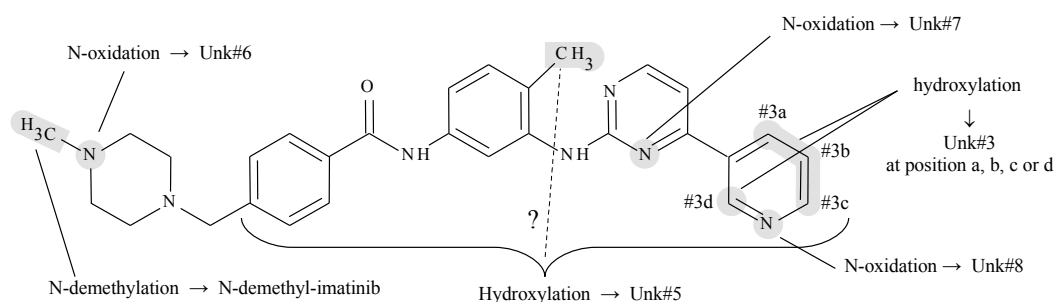
#### **7.3.2.1 Biological context**

Protein kinases are enzymes transferring phosphate from adenosine triphosphate to specific amino acids of substrate proteins, activating signal-transduction pathways involved in variety of biologic processes. Several of these protein kinases are deregulated and over expressed in human cancers, such as the Bcr-Abl tyrosine kinase involved in chronic myelogenous leukemia (CML), which have been extremely studied [204] [205] [206] [207] [208]. The phosphorylations mediated by Bcr-Abl are impaired by the drug Imatinib (Gleevec®, Figure 26), which is extremely efficient against CML [209] [210].

Besides the variability of systemic pharmacokinetics mediated by hepatic drug metabolizing enzymes (DME), DME activity in the targeted cells may be relevant for drug disposition in tumors [211] [212]. Indeed, in addition to the activity of influx and efflux systems through the membrane of cancer cells, it appears crucial to study local activity of DME to better understand bioactivation or degradation inside the target cells. Data underscoring that specific drug metabolism in cancer cells may play a role in cancer resistance are accumulating. Recent results [213] have shown that DME are able to biotransform Imatinib in six different metabolites that were identified and elucidated using liquid chromatography coupled with tandem mass spectrometers: one N-demethyl, 2 hydroxy (identified as Unk#3 and Unk#5) and 3 N-oxide (identified as Unk#6, Unk#7 and Unk#8) (Figure 26).

To investigate qualitatively the effect of these biotransformations on the ligand affinity

relative to Imatinib, these 6 metabolites have been docked to Abl in its inactive conformation using EADock.



*Figure 26: Chemical structure of Imatinib (Gleevec®;  $C_{29}H_{31}N_7O$ ; monoisotopic mass = 493.3 Da) with the structure proposal of Imatinib metabolites formed in the microsomal incubations and studied in this work. The interrogation mark indicates that, according to published spectral data, the structure of Unk#5 could be the hydroxy benzylic metabolite. The position of the hydroxy group of Unk#3 can take place at 4 carbons numbered a, b, c and d. Therefore, the Unk#3 possible chemical structures were defined as Unk#3a, Unk#3b, Unk#3c or Unk#3d.*

### 7.3.2.2 Modeling approach

#### 7.3.2.2.1 Parameters and coordinates handling.

Missing parameters for Imatinib and its metabolites, for use in conjunction with the CHARMM22 [43] all atoms molecular mechanics forcefield were derived from the Merck Molecular Force Field (MMFF [147] [148] [149] [150] [151]), by taking the dihedral angle term as is, but considering only the quadratic part of the bond and angle energy terms. The partial charges and van der Waals parameters of the ligand atoms were taken also from the MMFF. The ligands were modeled with all hydrogens.

The simulations were performed starting from the X-ray structure of the c-Abl kinase in complex with Imatinib resolved at 2.1 Å: (entry 1IEP in the Protein Data Bank). Titratable groups were taken in their standard protonation state at neutral pH. The protonation states of the histidine residues were set based on visual inspection of their environment.

The isolated complex was minimized using 100 steps of Steepest Descent (SD) algorithm using the GB-MV2 Generalized Born model to remove sterical clashes in the X-ray structure. The heavy atom RMSD between the X-ray structure and the minimized molecule is only 0.1 Å. All calculations were performed using the CHARMM program (version c31b1) and the CHARMM22 forcefield. The ligand was removed from the binding site before performing the docking.

#### **7.3.2.2.2 Docking of Imatinib and its metabolites to c-Abl kinase.**

The crystal structure of the catalytic region of human Abl kinase in complex with Imatinib has indicated that the interaction takes place at the kinase domain in the inactive and unphosphorylated conformation of Abl and engage 21 amino acid residues [214]. The docking of Imatinib and its metabolites to c-Abl kinase were performed using the EADock program [70]. The details of the calculations are presented in Chapter 2 “Material and Methods”. In brief, starting from a set of 250 randomly generated initial conformations, positions and orientations of the ligand in the region of the c-Abl kinase binding site, the coordinates of the ligand were refined using an evolutionary algorithm, renewing 10% of the population at each generation. The thorough exploration of the accessible conformational space of the ligand relative to the protein surface was submitted to the evolutionary pressure of a scoring function that takes account of the solvent effect thanks to the GB-MV2 implicit solvent model. The maximum allowed distance between the explored putative binding modes and the center of the binding site was 10 Å. This defined a search space encompassing the exterior of the binding site, in case a given metabolic modification of Imatinib would prevent a docking into the protein binding site. Protein residues were fixed during the docking. After 400 generations, the conformations with the lowest energy were further minimized by 100 steps of SD using the GB-MV2 generalized Born model. The lowest energy conformation was retained.

#### **7.3.2.2.3 Calculation of void regions between Imatinib and c-Abl kinase.**

The calculation of void regions between Imatinib and the c-Abl kinase was performed using the SURFNET program [215] implemented in UCSF Chimera molecular graphics software. Default parameters were used.

#### **7.3.2.3 Results**

The docking of Imatinib to c-Abl kinase was performed to critically assess the approach, thanks to the availability of the complex X-ray structure. The RMSD between the experimental and calculated binding modes of Imatinib is only 0.17 Å. All the important interactions between the ligand and the protein determined experimentally are reproduced in the modeled complex, illustrating the relevance of our method.

The different Imatinib metabolites were docked to c-Abl kinase, and are shown in the Figure 27 and in Table 6. The proposed binding modes for all Imatinib metabolites are very similar to that of the native molecule; i.e. none of these metabolic modifications was found to prevent the docking. However, several meaningful differences were found.

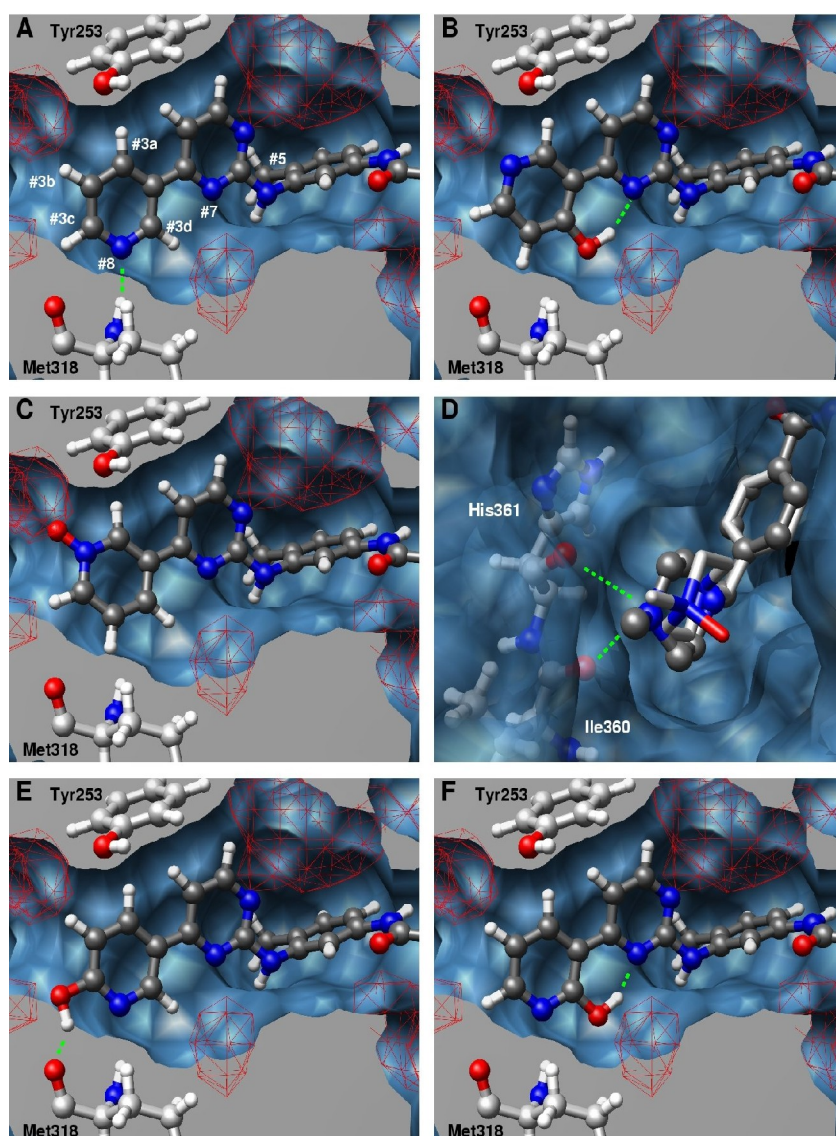


Figure 27: Docking of Imatinib metabolites. See text for details. (A) Imatinim, (B) Unk#3a, (C) Unk#8, (D) Unk#6, Unk#3c, Unk#3d

In the case of metabolites #3a and #8, the metabolic modification implies the addition of a bulky hydroxyl group or oxygen atom in a position where no space is found by the SURFNET program in the native Imatinib/c-Abl kinase complex (Figure 27A). Position #3a of the native Imatinib is in contact with the side chain of Tyr253, while position #8 (i.e. the nitrogen atom of the pyridine cycle) makes a hydrogen bond with the backbone NH atoms of Met318. Consequently, in the binding modes calculated by EADock for these two metabolites, the pyridine cycle is flipped so that the additional atoms point toward a void region in the complex (Figure 27B and Figure 27C). In both cases, the hydrogen bond

between the nitrogen atom of the pyridine cycle and the NH backbone atoms of Met138 is lost. In the case of metabolite #3a, the flipped conformation allows an internal hydrogen bond between the additional hydroxyl group and a nitrogen atom of the ligand pyrimidine cycle (Figure 27B). According to EADock, the oxidation of the N-methyl group (metabolite #6) of the piperidine cycle also has an impact on the ligand binding mode. In this case, the added oxygen atom is shown to prevent the hydrogen bond that is taking place between the protonated N-methyl group of the Imatinib piperidine cycle and the Ile360 and His361 backbone carbonyls. This leads to a modified position of the oxidized piperidine cycle according to the native Imatinib (Figure 27D). In the three cases, the binding mode of the rest of these metabolites is similar to that of the native Imatinib. The binding mode modifications described above reveal the sterical incompatibility of the metabolites for the exact binding mode of the native Imatinib without protein or ligand conformational rearrangements. Since no compensatory favorable interaction between the ligand and the protein is added by the modifications, this may be expected to cause a significant decrease of the ligand affinity for the protein.

All other metabolites (#3b, #3c, #3d, #5, #7 and #9) were shown theoretically to dock into the c-Abl kinase similarly to Imatinib. This result is in agreement with the fact that these metabolic modifications imply an addition in a void region of the c-Abl kinase/Imatinib complex according to SURFNET. However, some of these modifications change the interaction scheme between the ligand and the protein. In the case of metabolites #3c, the added hydroxyl group makes a hydrogen bond with the backbone carbonyl of Met318 (Figure 27E). In metabolite #3d, the added hydroxyl group makes an internal hydrogen bond with a nitrogen atom of the pyrimidine cycle of the ligand (Figure 27F). These additional interactions between the ligand and c-Abl kinase, or within the ligand, might increase the ligand affinity for the protein. This later statement should warrant further investigations, both *in silico* and *in vitro*.

## 7.4 Lead discovery and optimization

### 7.4.1 Material and Methods

#### 7.4.1.1 Overview

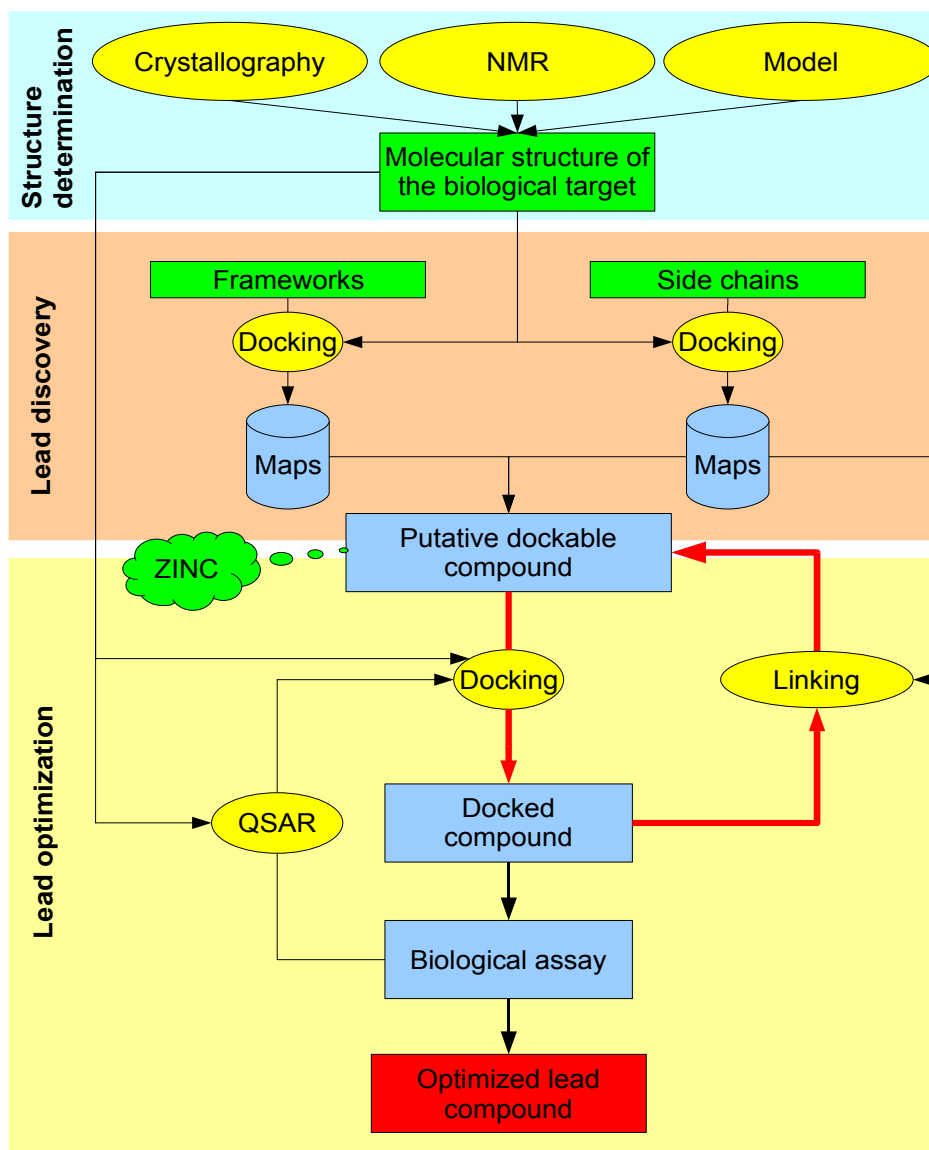


Figure 28: Overview of our generic rational drug design pipeline. Data are shown in blue, except starting and final data in green and red, respectively. Methods are shown in yellow. The optimization cycle is materialized by red arrows. See text for details.

A FB-RDD approach (see Chapter 1 “Introduction”), essentially using EADock, was used to design new peptidic ligands of the  $\alpha 5 \beta 1$  integrin and of the human PPAR $\alpha$  nuclear hormone receptor. An outline of this approach, which is believed to lead to an optimized lead compound starting from the molecular structure of a biological target, is presented in Figure 28.

In this approach, maps of all most favorable positions and orientations of small molecular fragments on the protein surface are calculated using the EADock program (see Figure 28). These fragments can be classified in frameworks and side chains [99] [100]. In our approach, frameworks are typically fragments of known lead compounds or virtual lead compounds (i.e. compounds designed *ab initio* and presenting a binding mode calculated by EADock compatible with the targeted binding site, which is well defined and reproducible). Side chains are small chemical fragments that could be linked to frameworks. In the case of peptide design, they correspond to all natural side chains, except glycine, which has no side chain, and proline. Once their most favorable positions in the binding site of the protein have been calculated with EADock, the possible linkings between frameworks and side chains are investigated by checking some geometrical and chemical rules (see Figure 29).

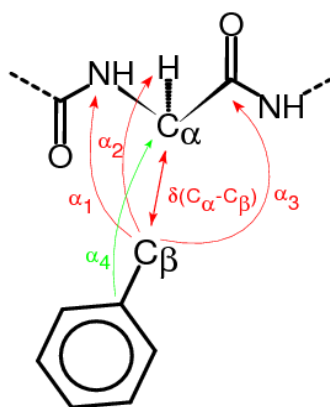


Figure 29: Distance and angles that are checked to verify the possibility of linking the backbone of a peptidic lead compound (top) with a putative side chain (bottom).



In the case of peptide design, the distance between the C $\alpha$  backbone atom and the C $\beta$  putative side chain atom should be close to 1.54 Å, while the different angles C $\alpha$ -C $\beta$ -C $\gamma$ , N-C $\alpha$ -C $\beta$ , H-C $\alpha$ -C $\beta$  and C-C $\alpha$ -C $\beta$  should be close to 109-110°. Given the corresponding virtual mutations, a list of putative peptide ligands is constructed in a combinatorial way. Then, each of these putative ligand is docked *in silico* in the protein active site using EADock. The most promising ones, in terms of interactions with the targeted protein, are retained for further modifications using the same FB-RDD approach. This new cycle of optimization might lead to new putative mutations in response to an eventual limited repositioning of the peptide backbone during the docking step. This procedure might be repeated several times (see Figure 28, “optimization cycle”), finally leading to a list of molecules to be tested experimentally. The measured experimental affinity might be used to generate a QSAR model that could be used in subsequent dockings (see Chapter 4 “Perspectives”).

#### **7.4.1.2 EADock**

EADock is used for the *in silico* docking experiments, i.e. finding the most favorable binding mode of a given ligand [70]. It is also used to calculate maps of the most favorable positions of frameworks and side chains, using the following modified parameters. The cluster size was reduced to 1.5 Å, with a maximum of 5 members, and the total population size was increased to 500 members. The blacklisting procedure was switched off to preserve all the most favorable binding modes generated during the evolutionary process. A radius of 15 Å was used to define the region of interest, encompassing the targeted binding site.

#### **7.4.1.3 Forcefield**

All molecular mechanics calculations were performed using the CHARMM program (version c31b1) and the CHARMM22 forcefield. Titratable groups were taken in their standard protonation state at neutral pH. The protonation state of the histidine residues were set based on visual inspection of their environment.

## **7.4.2 Targeting the integrin**

This work is being carried out by Vincent Zoete and Aurélien Grosdidier, in collaboration with Curzio Rüegg and Gian Carlo Alghisi from the Laboratory of tumor angiogenesis and melanoma research of the Centre Hospitalier Universitaire Vaudois. the corresponding manuscript is in preparation.

### **7.4.2.1 Biological context**

The survival of cells depends on many factors, among which the attachment to extracellular matrix (ECM) components, mediated by cell adhesion molecules such as integrins. Integrins are transmembrane  $\alpha\beta$ -heterodimeric proteins. 18  $\alpha$  and 8  $\beta$  subunits have been identified, forming 24 known dimers whose expression depends on cell type and cellular function. They are reviewed in [216] [217] [218]. Each integrin subunit has a large extracellular, a short transmembrane and small intracellular domains. They are the main receptors for ECM proteins like collagen, fibronectin and laminin. Cell-matrix interaction via integrins is essential for embryonic development, proliferation, survival, adhesion, differentiation, and the migration of cells [218] [219] [220] [221] [222] (see Figure 30).

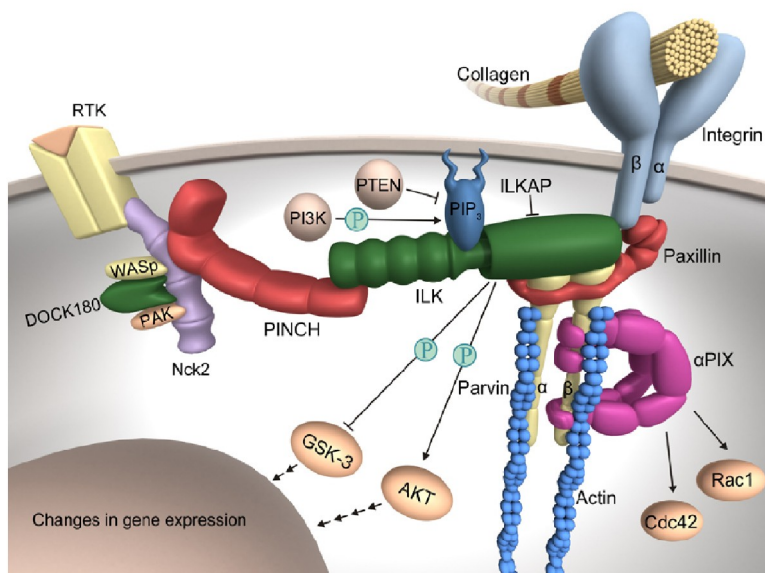


Figure 30: Overview of the integrin pathway [223].

Tumor cells have developed effective mechanisms to escape contact inhibition, which is mediated by such cell adhesion molecules. The over expression or loss of integrin contributes to several tumors, and changes in integrin expression patterns differentially affect tumor invasion and metastasis [224] [225] [226] [227] [228].

At the crossroad of such a major survival pathway, integrins are particularly interesting targets for cancer therapy, and a number of ligands have already been designed [157]. The Cilengitide is one of the most promising one. This cyclic pentapeptide c(RGDf(NMe)V) contains the RGD motif also found in fibronectin, known to bind to most integrins. It was found to be an efficient  $\alpha V\beta 3$  and  $\alpha 5\beta 1$  (to a lower extent) integrin inhibitors that limits tumor growth and angiogenesis, and is currently studied in clinical trials (phase I or II depending on the indication, see Chapter 1 "Introduction"). Our FB-RDD approach described above was used to propose new peptide inhibitors using the RGD motif as a template.

#### 7.4.2.2 Assessment of EADock

The ability of EADock to properly dock peptides on the surface of integrin molecule was assessed using the Cilengitide/ $\alpha V\beta 3$  integrin complex for which an X-ray structure is

available (1L5G in the PDB). Rapidly, EADock was found to reproduce the experimental binding mode of Cilengitide, with a RMSD of 1.17 Å, even when the algorithm was initiated with a population distributed far from the binding site: from 15 to 25 Å RMSD from the experimental binding mode (see Chapter 2 “Material and Methods”). Then, EADock was assessed for its ability to calculate meaningful maps of favorable positions for molecular fragments on the surface of the  $\alpha V\beta 3$  integrin. For this purpose, maps were calculated for the aspartate and arginine side chain fragments using the approach described in 7.4.1.2, and compared to the actual position of the aspartate and arginine side chains of Cilengitide. These two side chains were chosen since they are part of the RGD motif of Cilengitide, which is present in nearly all integrin natural ligands and has been shown to be important for the interaction with the receptor. Figure 31 shows the RGD motif of Cilengitide bound to the  $\alpha V\beta 3$  integrin, along with the map calculated for the aspartate side chain. As can be seen, the most favorable position for the aspartate side chain fragment (acetate) according to the EADock scoring function, which reflects the binding affinity between the fragment and the protein, is nearly superimposed to the actual position of the aspartate side chain of the RGD motif. Similarly, the most favorable positions calculated for the arginine side chain fragment make ionic interactions with the Asp 150 and 218 of chain  $\alpha$ , similarly to the arginine residue of Cilengitide. The sixth most favorable position is in fact superimposed to the arginine residue of the inhibitor (see Figure 31). These results illustrate the efficiency of EADock for both ligand docking and map calculation on the integrin receptor.

#### **7.4.2.3 Design of peptide inhibitors of $\alpha 5\beta 1$ .**

No structure was available for the  $\alpha 5\beta 1$  integrin. A homology model has been realized based on the structure of the  $\alpha V\beta 3$  integrin bound to the Cilengitide (1L5G in PDB, 52% identity), using MODELLER. The resulting structure was minimized by 30 steps of steepest descent and 30 steps of Adopted Newton-Raphson, then by 100 steps of steepest descent with the GB-MV2 solvent model. Based on the fibronectin sequence and structure (1TTG in PDB), we investigated the RGDSP linear pentapeptide as a possible virtual lead compound to initiate a cycle of sequence optimization. First, the RGDSP peptide was

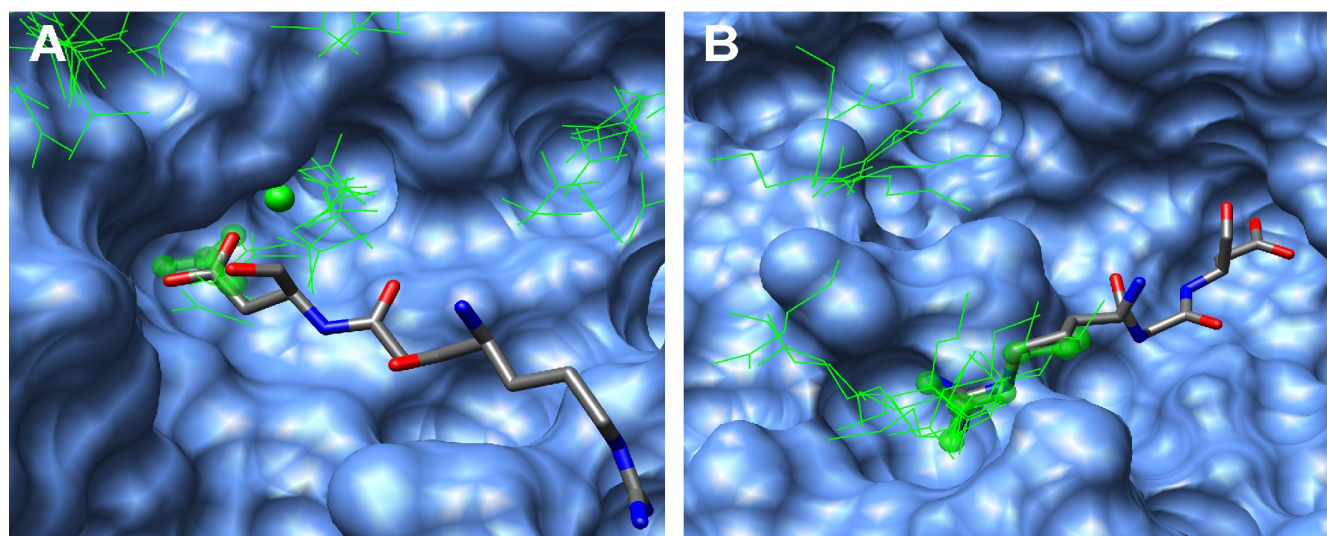


Figure 31: (A) The RGD motif of Cilengitide is shown in thick lines, colored according to the atom type. The rest of the molecule is not represented for the clarity of the figure. The surface of the  $\alpha V\beta 3$  integrin receptor is shown. Favorable positions for the acetate fragment are shown in green thin lines, except the most favorable one, which is shown in transparent ball and stick representation. (B) Same as (A) but for the arginine side chain fragment. The sixth most favorable position is shown in transparent ball and stick representation.

docked on the  $\alpha 5\beta 1$  integrin using EADock. Figure 32 shows the most favorable calculated binding mode.

As can be seen, the position of the RGD motif of the peptide is very similar to that seen in the experimental binding mode of Cilengitide on  $\alpha V\beta 3$ . The arginine residue makes ionic interactions with the Asp227 side chain of chain  $\alpha$ , and the aspartate residue with the backbones of Asn553 and Tyr462 (chain  $\beta$ ) and with an  $Mg^{2+}$  structural ion, respectively. Additional hydrogen bonds also take place between the NH backbone atoms and the hydroxyl side chain function of the peptide serine residue on the one hand, and the backbone of carbonyl of the integrin Asn553 (chain  $\beta$ ) on the other hand. Similarly to the N-methylValine residue of Cilengitide, the proline residue does not make any interaction with the integrin receptor. The binding mode obtained by EADock was reproducible and stable enough to be used as a virtual lead compound for further sequence modifications.

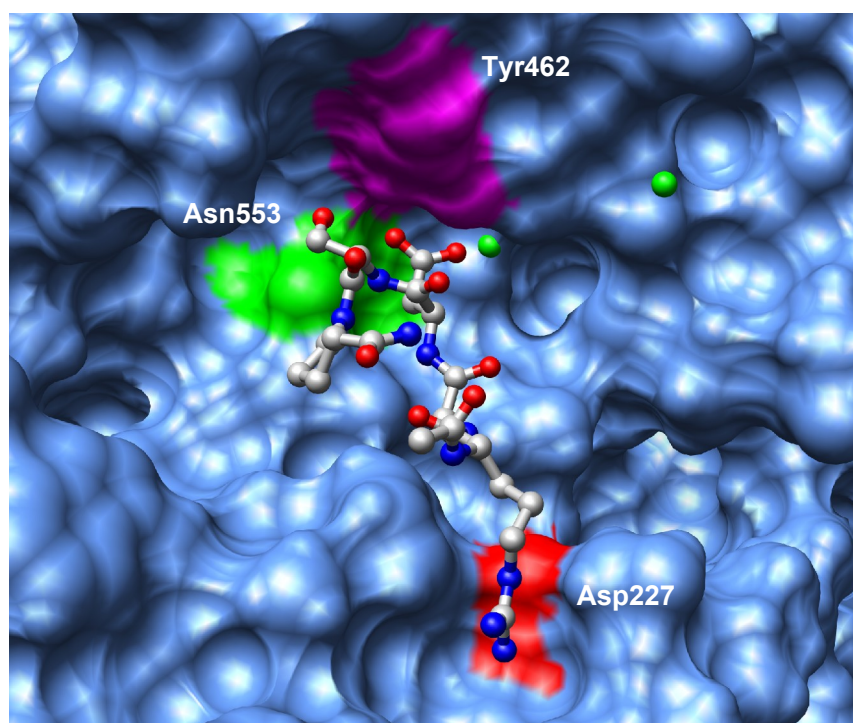


Figure 32: Binding mode of RGDSP

The RGDSP linear peptide was submitted to a cycle of sequence optimization, following the approach described earlier. Several pentapeptide sequences were suggested by the approach and were retained for experimental analysis. They are listed in Table 7.

Peptide	IC <sub>50</sub> ( $\alpha 5\beta 1$ )
Cilengitide, i.e. c(RGDf[Met]V)	5.5 $\mu$ M
RGD	650 $\mu$ M
RGDSP	45 $\mu$ M
RGDLP	175 $\mu$ M
RGDFP	17.5 $\mu$ M
RGDWP	12 $\mu$ M

Table 7: Sequence and IC<sub>50</sub> of the peptides experimentally tested on the  $\alpha 5\beta 1$  integrin.

For example, Figure 33A shows the map of favorable positions for the tryptophan side chain fragment calculated by EADock. The most favorable position, in terms of binding free energy, makes several van der Waals interactions with Leu512, Pro515 and Cys516, and a hydrogen bond with the Ser518 side chain (chain  $\beta$ ). It is well positioned to replace the fourth side chain (Ser) of the RGDSP lead compound. This led to the automatic design



of the RGDWP sequence and docking of the corresponding peptide. As can be seen in Figure 33B, the tryptophan residue of RGDWP, in its calculated binding mode, has a position similar to that displayed by the parent molecular fragment that was used to derive this putative compound. This illustrates that the basis of the approach does make sense: selecting a molecular fragment displaying numerous favorable interactions with the targeted protein and well positioned to be linked to the lead compound actually led to a new peptide whose calculated binding mode conserved the positioning of this particular side chain, as well as that of the rest of the molecule.

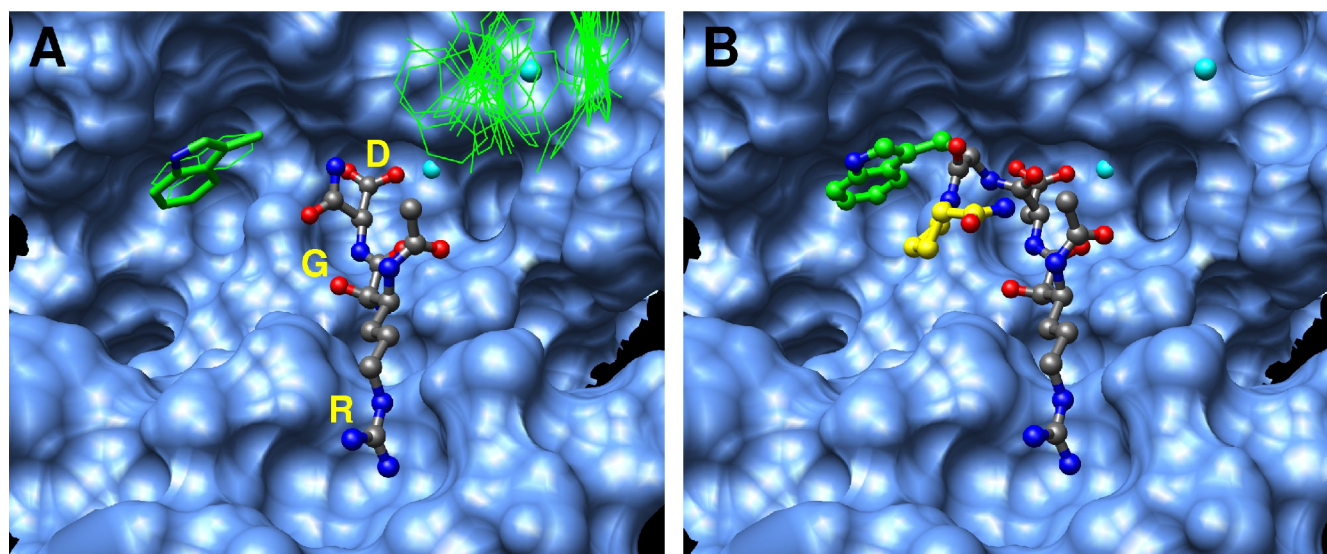


Figure 33: (A) The RGDSP peptide docked on the surface of the  $\alpha 5 \beta 1$  integrin is shown in ball and stick representation. The S and P residues are not shown for the clarity of the figure. The surface of the  $\alpha 5 \beta 1$  integrin receptor is shown. Favorable positions for the tryptophan fragment are shown in green thin lines, except the most favorable one, which is shown in transparent thick lines. (B) Binding mode calculated by EADock for the linear RGDWP peptide.

As could be expected from the fact that the proline residue does not make any interaction with the targeted protein, no possible sequence modifications was found for this residue. On the contrary, several sequence modifications were suggested by the approach for the serine residue, including mutations to leucine, phenylalanine and tryptophan. The four corresponding linear pentapeptides were synthesized and tested experimentally, along

with the tripeptide RGD that was taken as a reference. As can be seen in Table 7, the RGD tripeptide shows a low affinity for the  $\alpha 5\beta 1$  receptor, with an  $IC_{50}$  of only 650  $\mu M$ . The RGDSP, whose sequence was extracted from fibronectin, shows a much better  $IC_{50}$ , i.e. 45  $\mu M$ . Among the three peptides whose sequence were derived from the *in silico* approach, two display an increased activity compared to that of RGDSP: RGDFP and RGDWP. As already reported for  $\alpha V\beta 3$ , the fourth residue shows therefore a preference for aromatic side chains. The most active peptide, RGDWP, exhibits an activity very similar to that of the cyclic peptide Cilengitide.

#### **7.4.2.4 Conclusion**

The FB-RDD procedure used in this study led to the design of several linear pentapeptides showing significant experimental affinity for the  $\alpha 5\beta 1$  receptor. The measured activities were higher than that of the RGD motif and the RGDSP lead compound, and similar to that of the well known Cilengitide cyclic pentapeptide. This study thus illustrates the ability of this *in silico* approach to design new ligands with increased affinity for the targeted protein.

The results are only a first preliminary step in the design of  $\alpha 5\beta 1$  inhibitors. Further rounds of sequence optimization will be performed to potentially increase the affinity of these peptides. Also, several modifications will be intended to increase the peptide resistance to metabolism, such as the introduction of D residues in the sequence, or cyclization of the molecule. In addition, the latter could also lead to an increase in affinity for the protein, since it reduces drastically the conformational space of the peptide, compared to the linear one, and therefore limits the entropic penalty upon complexation. As can be seen in Figure 33B, the proximity of the C- and N-termini of the peptide in its predicted bioactive conformation suggests that the latter could be conserved during such a cyclization.

#### **7.4.3 Design of peptidic PPAR $\alpha$ ligands**

This work was carried out by Vincent Zoete, Lina Yip-Sondernegger and Aurélien



Grosdidier, in collaboration with the groups of Liliane Michalik and Walter Wahli (UNIL-Center for Integrative Genomics). The manuscript is in preparation.

#### **7.4.3.1 Biological context**

PPAR $\alpha$  is found in liver, kidney, heart and muscle and activates genes responsible for maintaining the homeostasis of the fatty acids and lipoprotein metabolism through their uptake and oxidation. PPAR $\alpha$  is the target of the lipid lowering fibrates.

#### **7.4.3.2 Assessment of EADock**

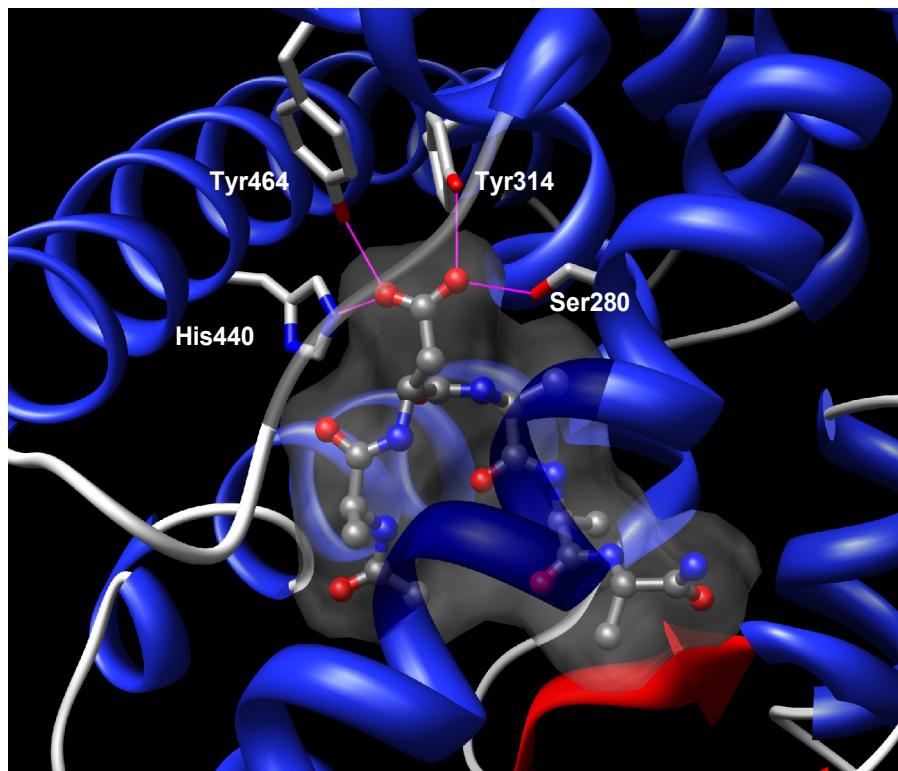
All calculations were done using the X-ray structure of the human PPAR $\alpha$  molecule bound to the GW409544 agonist (1K7L in PDB). This ligand was removed prior to all calculations.

The ability of EADock to properly dock molecules within the binding site of PPAR molecules has been assessed previously [180]. Also, the ability of EADock to calculate reliable maps of favorable positions for small molecular fragments has been tested during a previous study aiming at designing peptide inhibitors for integrin receptors (see previous application).

#### **7.4.3.3 Design of peptide ligands of hPPAR $\alpha$ .**

Since no large peptide ligand is available for PPAR $\alpha$ , the first step of the approach was to construct an initial peptide virtual lead compound. The latter is a peptide that has not been tested experimentally, thus we don't know whether it is active or not. However, it is characterized by the fact that it exhibits a well defined and reproducible binding mode according to *in silico* approaches, so that it can be used as a first template for further sequence optimization. Analysis of all organic synthetic ligands show that they share the same structural property as the natural ligands: they have a polar head (generally a carboxylate function or a thiazolidinedione cycle), responsible for the interactions with the polar part of the binding site (Ser280, Tyr314, His440 and Tyr464), and a hydrophobic tail making van der Waals interactions with the rest of the binding site, which is itself mainly

hydrophobic [197]. Based on this, we chose the ADAAA peptide as a virtual lead compound to initiate the sequence optimization. This peptide was docked in the PPAR $\alpha$  using the EADock program. The estimated binding mode shows that, as expected, the aspartate side chain makes a network of four hydrogen bonds with the polar part of the binding site, while the alanine residues fill part of the rest of the hydrophobic large binding site, providing a first positioning of the backbone (see Figure 34).



*Figure 34: Calculated binding mode of the virtual lead compound ADAAA. The carboxylate function of the Asp residue makes a hydrogen bonds network with Ser280, Tyr314, His440 and Tyr464.*

The virtual lead compound was submitted to two rounds of sequence optimization. No interesting substitution was found for the N-terminal alanine residue. According to the FB-RDD, the aspartate residue in position 2 could be replaced by a glutamate side chain, while a mutation of the alanine in position 3 to serine was suggested for its ability to make an additional hydrogen bond with the Ser280 side chain. Little room is present around the

alanine in position 4, and consequently, no mutation was suggested by the approach for this residue. Finally, it was suggested that a mutation of the C-terminal alanine to tryptophan could take place. As can be seen in Figure 35, a favorable position for a tryptophan side chain fragment is situated adequately to be linked to the C-terminal residue and makes favorable van der Waals interactions with the surrounding hydrophobic residues (Cys275, Val332 and Ala333), as well as a possible orthogonal  $\pi$ - $\pi$  interaction with the Tyr334 side chain (T-stacking).

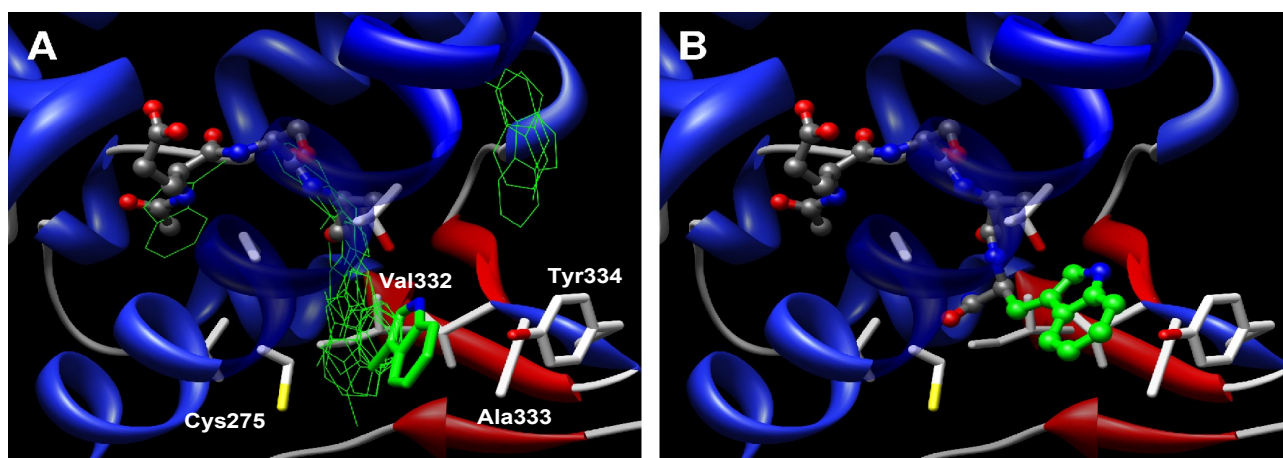


Figure 35: (A) The calculated binding mode of the AESAA peptide is shown in ball and stick representation. The N- and C-terminal residues are not given for the clarity of the figure. The favorable positions of the tryptophan side chain fragment are shown in green thin lines, except the one that was found by the reconnection procedure to be a putative side chain for the C-terminal residue. (B) Calculated binding mode for the ESAW peptide. The actual position of the Trp side chain is colored in green.

Several peptides were retained for experimental analysis, i.e. ADAAA (as a reference, Figure 34), AESAW, ADSAW, ESAW (its binding mode is shown in Figure 36) and Fmoc-ESAW. ADAAA was found to be inactive experimentally.

This result is however interesting, because it shows that a virtual lead compound does not require a detectable activity to be used for sequence optimization, as long as it exhibits a

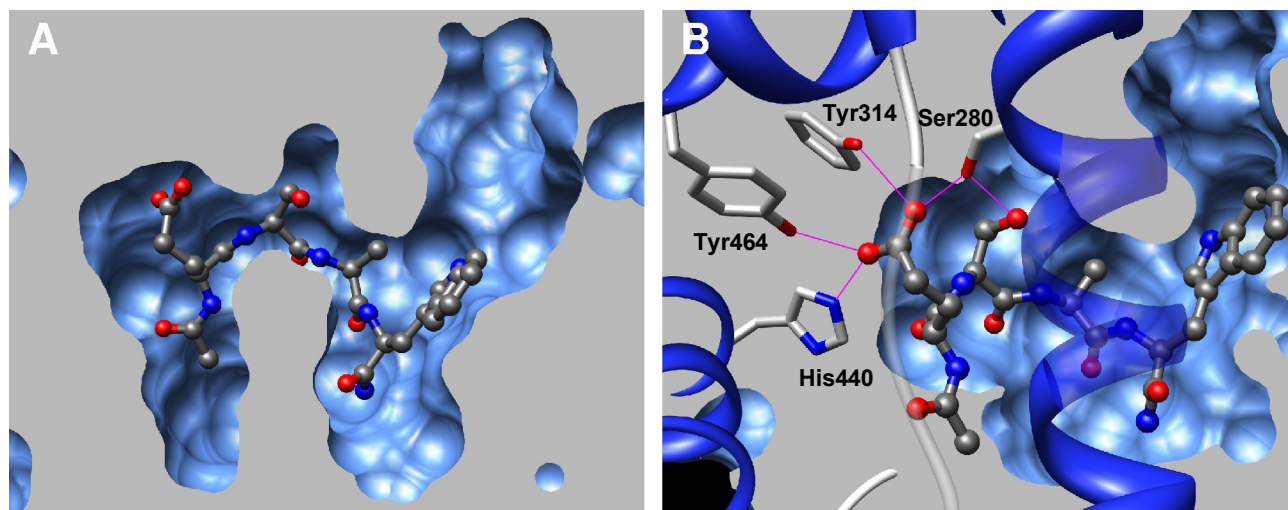


Figure 36: (A) The calculated binding mode for the ESAW peptide (ball and stick) in PPAR $\alpha$ . The protein surface is clipped to show the space filling of the binding site realized by the peptide. (B) Hydrogen bonds network between the glutamate side chain of ESAW and Ser280, Tyr314, His440 and Tyr464.

well defined *in silico* binding mode. Preliminary results showed that AESAW does have a limited but detectable activity, while ESAW exhibits a low micromolar activity, similar to that of the well established PPAR $\alpha$  organic ligand Wy 14,643. These first results are very encouraging in view of the absence of experimental lead compound, and of the fact that this sequence was derived using only *in silico* approaches. This highlights the efficiency of our *in silico* FB-RDD approach.

#### 7.4.3.4 Conclusion

The FB-RDD procedure used in this study led to the design of several linear tetrapeptides and pentapeptides, some of them showing significant experimental affinity against the PPAR $\alpha$ . The measured activity of the ESAW tetrapeptide is similar to that of Wy14,643, a well established organic ligand of PPAR $\alpha$ . This study thus illustrates the ability of this *in silico* approach to design new ligand with high affinity for the targeted protein.

Further rounds of sequence optimization will be performed to potentially increase the affinity of these peptides for PPAR $\alpha$ . Also, several modifications will be intended to increase the peptide resistance to metabolism, such as the introduction of D residues

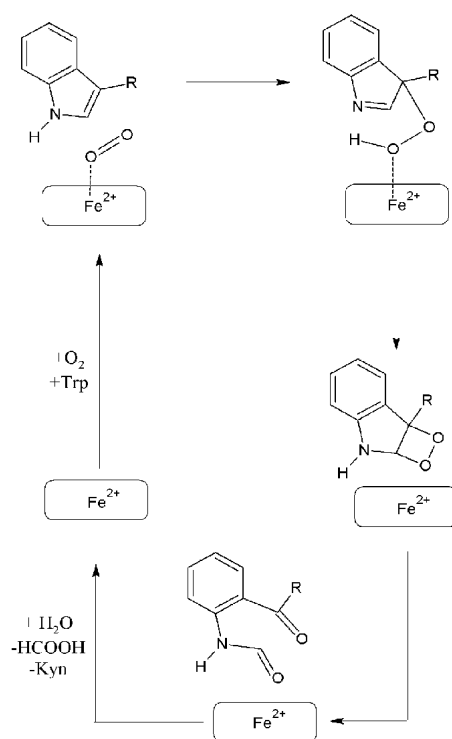
in the sequence.

#### 7.4.4 Targeting the indoleamine deoxygenase

This work is being carried out by Üte Röhrig, Vincent Zoete and Aurélien Grosdidier. The manuscript is in preparation.

##### 7.4.4.1 Biological context

The heme-containing enzyme indoleamine 2,3-dioxygenase (IDO, EC 1.13.11.52) has recently been implicated in the establishment of pathological immune tolerance by tumors. IDO catalyzes the initial and rate-limiting step in the catabolism of tryptophan (Trp) along the kynurenine pathway (see Figure 37) [229].



*Figure 37: IDO catalytic cycle. In the first step of the catalytic cycle, IDO binds both the substrate and molecular oxygen in the distal heme site. The enzyme catalyzes the cleavage of the pyrrole ring of the substrate and incorporates both oxygen atoms. It releases N-formyl kynurenine, which is subsequently hydrolyzed to kynurenine.*

By depleting Trp locally, IDO blocks the proliferation of T lymphocytes, which are extremely sensitive to Trp shortage [230]. The observation that many human tumors constitutively express IDO introduced the hypothesis that its inhibition could enhance the effectiveness of cancer immunotherapy. In fact, results from *in vitro* and *in vivo* studies suggest that the efficacy of therapeutic vaccination of cancer patients might be improved by concomitant administration of an IDO inhibitor [231]. Up to date, the best known IDO inhibitors display affinities in the micromolar range (see Table 8).

Inhibitor	Stereochemistry	K [ $\mu$ M]	Reference
4-Phenyl-Imidazole (PIM)		8	[Sono1989]
1-Methyl-Trp (1MT)	L	34	[Peterson1994]
1-Methyl-Trp (1MT)	L,D	34.2	[Muller2005]
MTH-Trp (MTH)	L,D	11.6	[Muller2005]
Trp (Trp)	L,L	147	[Peterson1994]
4-F,7-F-Trp (47FF)	L	40	[Sono1996]
5-F,7-F-Trp (57FF)	L	24	[Sono1996]
7-F-Trp (7F)	L	37	[Sono1996]
3-benzofuranyl-Trp (OIN)	L,D	25	[Cady1991]
3-benzothienyl-Trp (SIN)	L,D	70	[Cady1991]
2,5-Dihydro-Phe (DHP)	L	230	[Watanabe1978]
BR1		97.7	[Gaspari2006]
BR2		82.5	[Gaspari2006]
BR3		41	[Gaspari2006]
BR4		34	[Gaspari2006]
BR5		42.1	[Gaspari2006]
BR6		179.6	[Gaspari2006]
BR7		47.6	[Gaspari2006]
BR8		72.4	[Gaspari2006]
BR9		62.4	[Gaspari2006]
BR10		149.4	[Gaspari2006]
BR11		1267	[Gaspari2006]
BR12		37	[Gaspari2006]
BR13		13.2	[Gaspari2006]
BR14		363.6	[Gaspari2006]
BR15		17.2	[Gaspari2006]
BR16		11.6	[Gaspari2006]
BR17		28.4	[Gaspari2006]
BR18		20.5	[Gaspari2006]
BR19		NI	[Gaspari2006]
BR20		342.3	[Gaspari2006]
BR21		NI	[Gaspari2006]
BR22		202	[Gaspari2006]
BR23		1292	[Gaspari2006]
BR24		328.7	[Gaspari2006]

Table 8: IDO inhibitors

Both competitive and noncompetitive inhibitors have been identified, the latter being

mostly  $\beta$ -carboline derivatives [232]. Competitive inhibitors are mainly derived from Trp or from the natural compound brassinin, many of them incorporating an indole ring. In *in vivo* studies, mainly the Trp analog 1-methyl-Trp (1MT) has been used ( $K = 34 \mu\text{M}$ ) so far.

The recently resolved crystal structure of human IDO [233] can serve as a scaffold for the design of new IDO inhibitors. To this aim, we focus first on elucidating the binding modes and affinities of all competitive inhibitors, where a binding constant value has been experimentally measured (Table 8, Figure 38). Those include Trp derivatives with indole ring substitutions (1MT, 47FF, 57FF, 7F), Trp derivatives with indole ring modifications (SIN, OIN, DHP), Trp derivatives with side chain modifications (MTH), TRP<sub>2</sub>, brassinin derivatives (BR1-BR24), and PIM. The second goal of this study is to propose rational modifications of these compounds according to the FB-RDD pipeline described above.

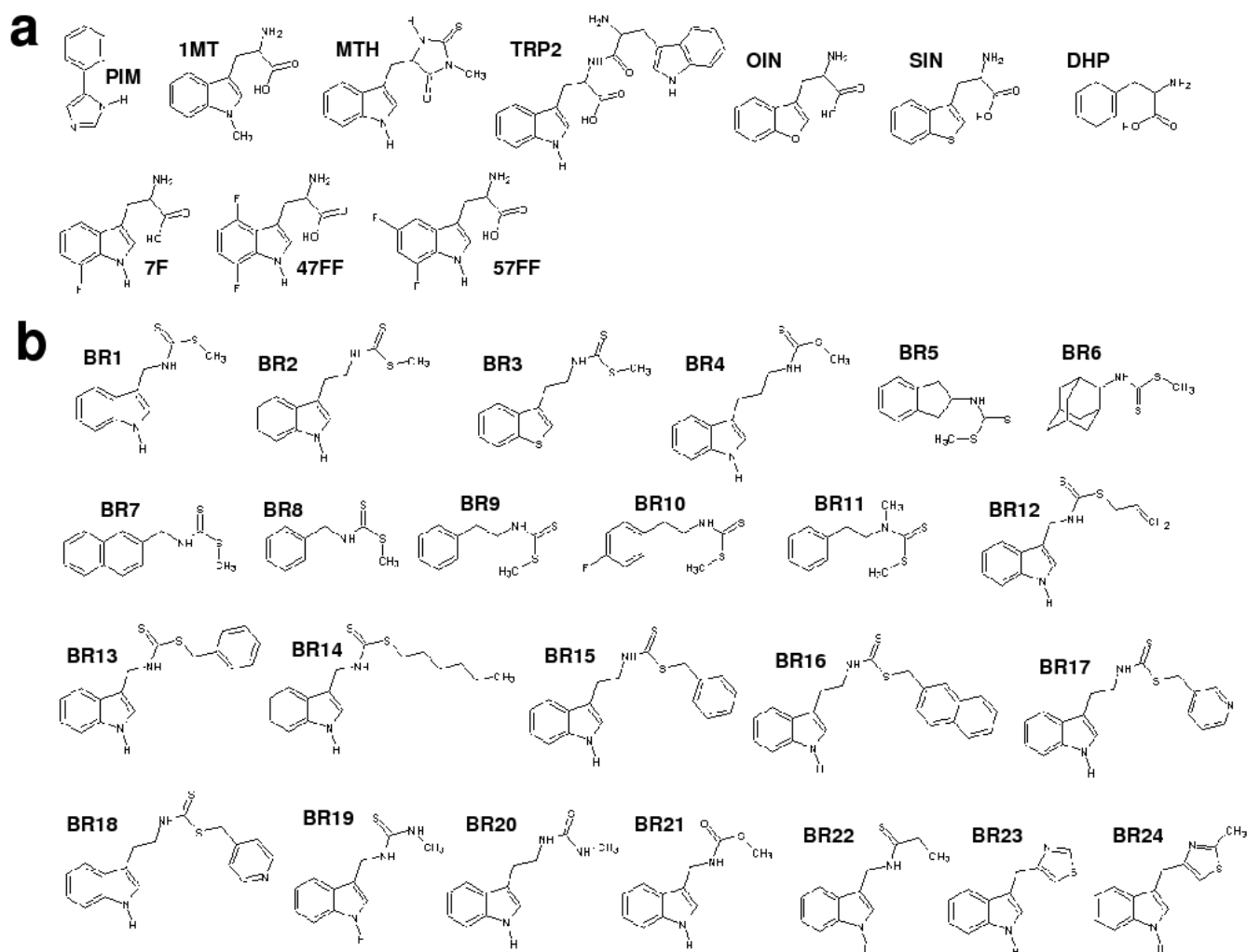


Figure 38: All docked inhibitors

#### 7.4.4.2 Modeling approach

In the X-ray structure, PIM is bound in a deep binding site, with its phenyl ring inside a large hydrophobic pocket (Pocket A, Figure 39). The imidazole nitrogen is coordinated to the heme iron with a distance of 2.1 Å. The PIM binding site is made up of residues Tyr126, Cys129, Val130, Phe163, Phe164, Ser167, Leu234, Gly262, Ser263, Ala264, and the heme ring. Possible hydrogen-bonding sites are the SH group of Cys129, the OH group of Ser167, the backbone CO group of Gly262, the backbone NH group of Ala164, and the heme 7-propionate group. Larger ligands than PIM may also interact with Phe226, Arg231, Ser235, Phe291, Ile354, and Leu384, which are located at the binding site entrance. Here, additional hydrogen bonds are possible with the sidechain of Arg231. A hydrophobic pocket in this region is provided by Phe163, Phe226, Arg231, Leu234,



Ile354, and the heme ring (Pocket B, Figure 39).

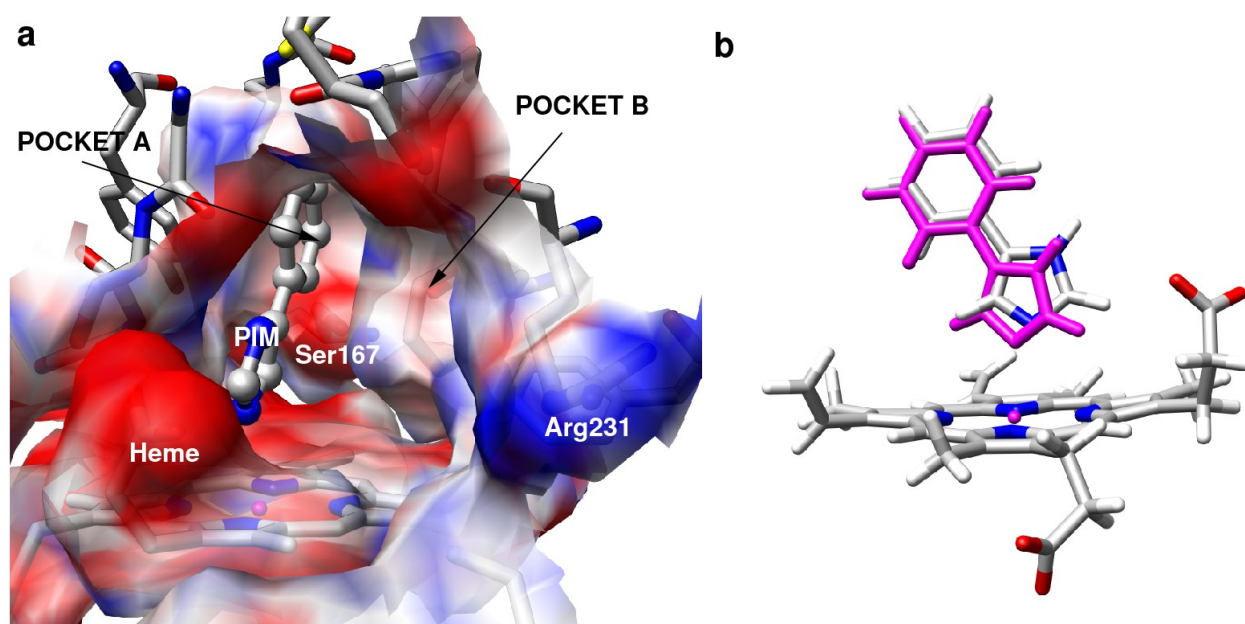


Figure 39: *a. Crystal structure of IDO: Binding site with bound PIM ligand. Two hydrophobic pockets and some important residues are labeled. The surface is colored by its electrostatic potential (red=negative, blue=positive). b. Superposition of PIM crystal structure (magenta) and best predicted PIM structure.*

### 7.4.4.3 Results

#### 7.4.4.3.1 Docking of frameworks

##### 7.4.4.3.1.1 Docking of the co-crystallized inhibitors

As a first assessment of the ability of EADock to identify relevant ligand binding modes in IDO, we docked 4-phenylimidazole (PIM), which has been co-crystallized in one of the resolved crystal structures (2D0T). The best docking solutions are very close to the crystal structure (ligand heavy atom RMSD 0.7 Å, Figure 39). The distance between the imidazole nitrogen and the iron is somewhat larger (2.8 Å) than in the X-ray structure (2.1 Å). This is expected as our empirical force field neglects any covalent contribution to the Fe-N bond. PIM does not form any hydrogen bond with the protein. Its NH group is pointing towards the solvent. The result demonstrates that EADock is capable of identifying the most favorable binding site in agreement with experimental data.

#### 7.4.4.3.1.2 Docking of the substrate tryptophan

Human IDO degrades both D and L-Trp. In order to explore the complex formed between Trp and the enzyme, we docked both isomers. Studies were carried out in absence of O<sub>2</sub>.

For L-Trp, an energetically favorable binding mode is found inside the binding site, with the indole nitrogen pointing towards the heme iron, the phenyl ring filling the hydrophobic pocket, and the  $\alpha$ -amino group forming a salt bridge with the heme 7-propionate (Figure 40).

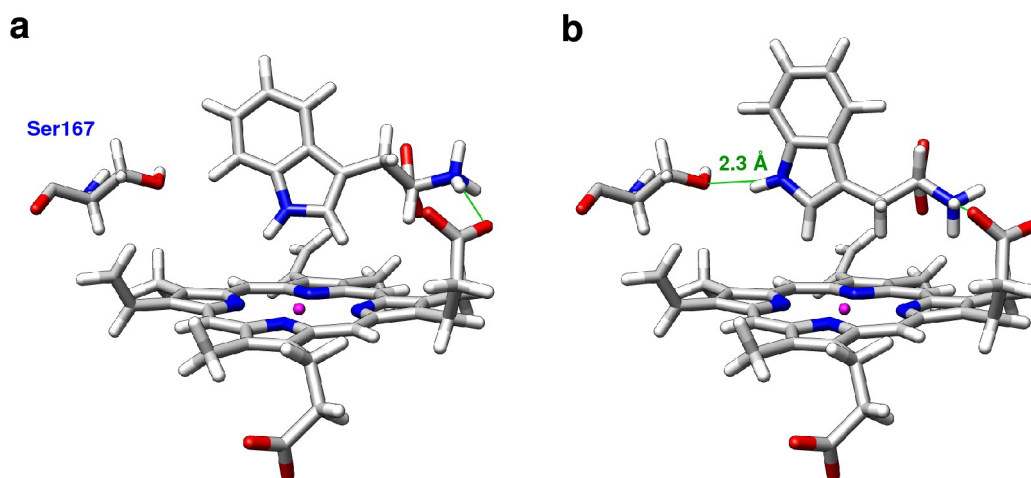


Figure 40: Binding mode of Trp. a. L-Trp. b. D-Trp.

A similar binding mode is found for D-Trp, but a difference consists in the location of the indole NH group, which points further away from the heme iron and forms an additional hydrogen bond with the sidechain of Ser167 (Figure 40). Despite this additional favorable interaction, the energy of this pose is about 5 kcal/mol higher than that for L-Trp.

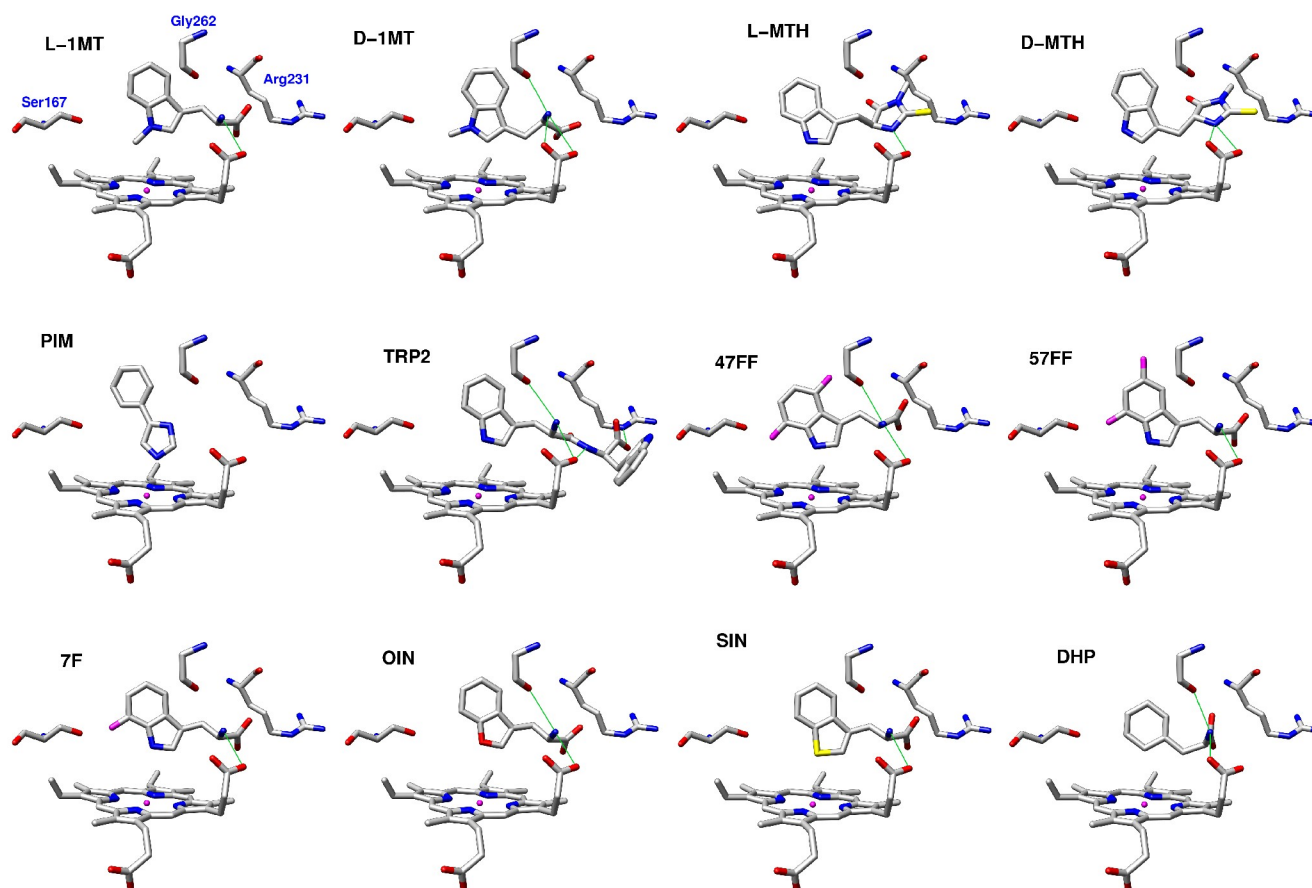
It is conceivable that a binding mode similar to the ones found here could be possible also in the presence of oxygen, which would allow for the proposed catalytic cycle [233] to take place. While the proposed binding mode displays many favorable interactions between the receptor and the ligand, the Trp carboxylate group is located in a hydrophobic environment. However, preliminary results from a molecular dynamics simulation of Trp-bound IDO in explicit water show that the positively charged sidechain of Arg231 can move to a position where it forms a direct hydrogen bond with the

carboxylate group of the Trp ligand.

### **Docking of other known inhibitors**

#### **7.4.4.3.1.2.1 Docking of tryptophan derivatives**

For most Trp derivatives, we find a conserved binding mode similar to the binding mode of Trp (Figure 41).



*Figure 41: Docked other known inhibitors.*

While the plane of the indole ring is very similar in all ligands and roughly perpendicular to the heme plane, the “tilt” of the indole ring and its distance with respect to the heme plane shows some variation. This flexibility might be connected to the poor substrate specificity of IDO. The positively charged  $\alpha$ -amino group is always hydrogen bonded to the heme propionate and additionally sometimes to the backbone oxygen of Gly262. The L isomers of the oxygen and of the sulfur analogues of Trp (OIN, SIN) bind exactly in the

same position as the parent compound, while the D isomers seem to be less easily accommodated within the binding site – here, all ligands are found to dock outside of the binding site.

Trp<sub>2</sub> binds with the first Trp inside the binding site, while the second Trp displays hydrogen bonds with the heme propionate and the sidechain of Arg231.

The most commonly used IDO inhibitor in cellular and in *in vivo* assays is 1MT. We find a favorable docking mode for both isomers of 1MT inside the binding site, similar to the binding mode of Trp. The indole ring is pushed further away from the heme by the presence of the methyl group.

#### **7.4.4.3.1.2.2 Docking of brassinin derivatives**

We docked all 24 experimentally investigated brassinin derivatives. Except for one compound (BR6), we obtain a very conserved binding mode for all ligands (Figure 42), suggesting that the rigid pocket approximation is quite appropriate.

In the parent compound (BR1), the indole ring binds in a similar fashion than in D-Trp. The NH group of the dithiocarbamate hydrogen bonds to the backbone oxygen of Gly262. Most other compounds follow this mode, with optional hydrogen bonds of the indole NH group to Ser167 and of the dithiocarbamate NH group to the heme propionate.

In analogy to Trp, the brassinin compounds do not bind directly to the heme iron. The low affinity of BR11 can be explained by the fact that the hydrogen bond to heme propionate is eliminated by the presence of the aminomethyl group. In BR13 and BR15, two of the most active inhibitors, the additional phenyl ring interacts mainly with the sidechain of Arg231, possibly forming some cation- $\pi$  interaction. In BR23 and BR24 the thiazole ring can neither interact favorably with the heme propionate group nor with the backbone oxygen of Gly262 because it does not include a polar hydrogen atom, thus explaining the low affinity of these ligands. However, from the obtained geometries it is not obvious why BR20 has a low affinity.

For all compounds, the sulfur atoms of the dithiocarbamate group are located in a highly hydrophobic and polarizable side pocket of the binding site (Pocket B, Figure 39). This environment seems to be well suited to accommodate the large, polarizable sulfur atoms.

Compounds, where one or two of the sulfur atoms are replaced by nitrogen or oxygen (BR19-BR22), do not generally dock in the same position as their sulfur counterparts. Therefore, it is difficult to determine what exactly is the effect of the substitution. We are currently working on developing a QSAR model that can explain the different activities of these compounds. In BR6, at variance with all other compounds, the indole ring is replaced by a highly non-planar ring system, which cannot be accommodated in the binding site of the crystal structure. However, since BR6 shows a similar activity to some planar compounds, it might be assumed that the binding site is able to adapt also to this compound.

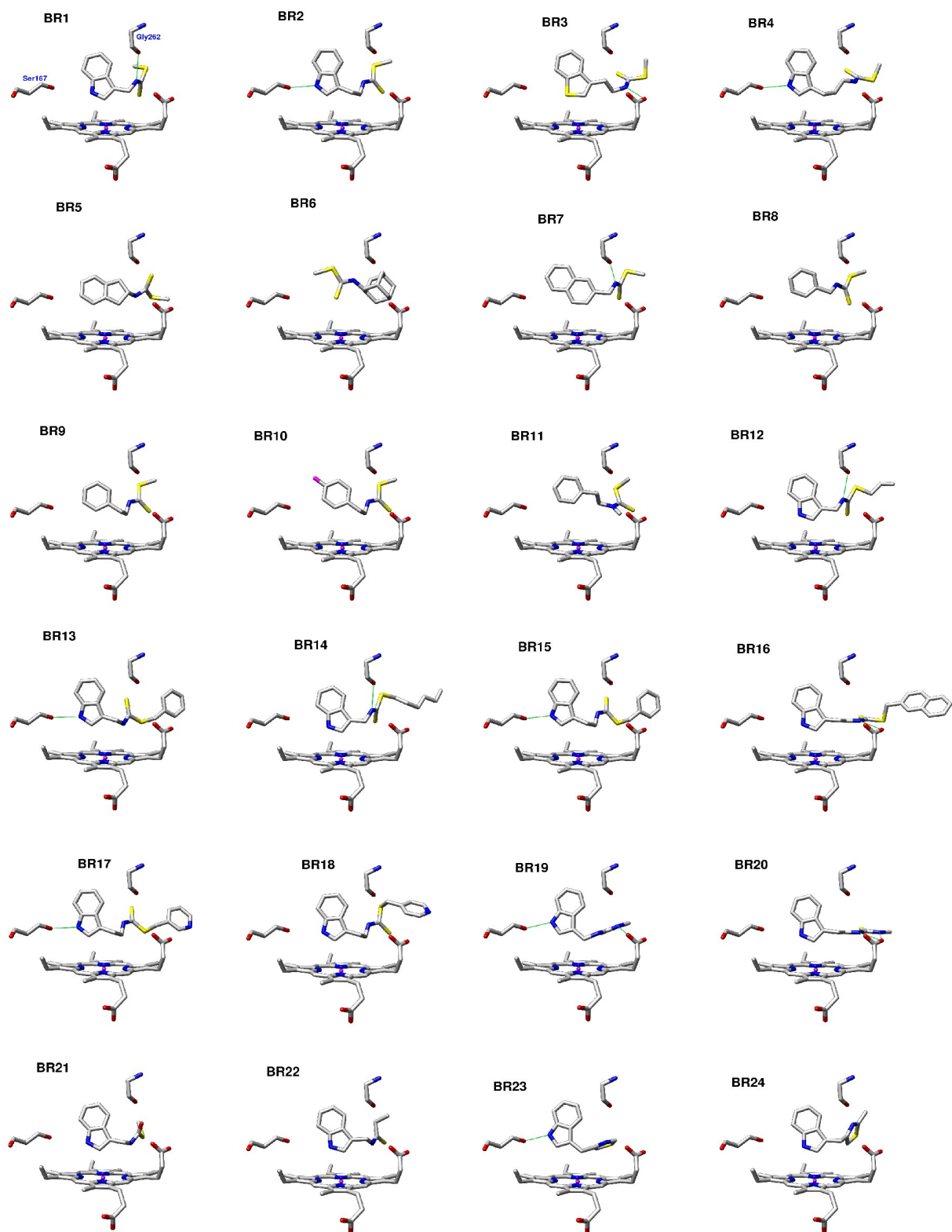


Figure 42: Docked brassinin derivatives

#### **7.4.4.4 Design of new inhibitors**

Despite the challenging interaction of the compounds with the iron of the heme, relevant binding modes were observed. They can be combined with fragment maps generated by EADock (see above).

#### **7.4.4.5 Conclusion**

The results from the docking of PIM into the IDO binding site suggests that our docking algorithm is capable of finding and correctly ranking the conformation of IDO ligands despite the difficulties stemming from the presence of a transition metal in the active site. We are now addressing this issue by using an extra potential to the two scoring functions of EADock in order to obtain a better description of the interactions with the iron atom (see Chapter 4 “Perspectives”).

Docking of different ligands into the IDO binding site shows that the latter is large enough and able to accommodate both L and D isomers of Trp and its derivatives, in agreement with experimental data. The fact that all known IDO inhibitors (except for one compound) can be docked inside the active site of the 2D0T X-ray structure illustrates that induced fit does probably not play a major role in the binding process. However, e.g. in case of Trp, we would expect a sidechain movement of Arg231 in order to optimally bind the ligand. We will investigate this point further by using a flexible protein during the docking procedure.

Based on the observed geometries of the bound ligands, we conclude that a good ligand should display some or all of the features summarized by the pharmacophore in Figure 41. A QSAR model, which will serve for rationalizing the observed activities and for predicting the activities of new ligands, is currently under development.

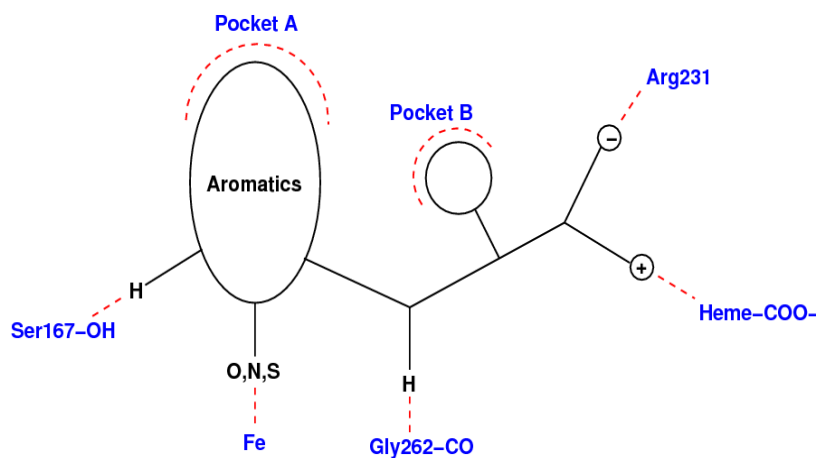


Figure 43: Suggestion for a pharmacophore: (i) a large hydrophobic fragment to fill pocket A in the binding site; (ii) an atom that can coordinate to the heme iron such as oxygen, nitrogen, sulfur (even if not necessary for binding, see e.g. the case for Trp, this should increase the affinity); (iii) a positively charged aminogroup that can form a saltbridge with the heme 7-propionate; (iv) a negatively charged group that can form a saltbridge with Arg231; (v) a hydrophobic group that can form van der Waals interactions with pocket B; and (v) groups that can hydrogen bond to Ser167 and to Gly262.







## 8 Perspectives

In a prospective way, this chapter gives some insights about several aspects that could significantly improve both the performance and the usability of our docking approach. Interestingly, the two most cited programs to date (AutoDock and GOLD) are evolutionary algorithms, illustrating their ability to provide an appropriate way of implementing separately the sampling heuristic used to generate poses and the scoring function used to rank them. The research around evolutionary computation is a very active field with major improvements being reported [234]. This theoretical flexibility can be combined by the practical flexibility resulting from the use of up to date software development techniques such as object-oriented design. The resulting software inherently allows a convenient bridge to be established between research-grade and production-grade softwares. The former should be a toolbox allowing an easy investigation and complete reorganization of the methodological basements of the algorithm. This usually comes at the price of usability, with several and somewhat confusing options and features made available, together with the unavoidable corresponding software flaws. On the contrary, production-grade software must be stable, reliable, easy to use and efficient. This comes at the price of a reduction of the number of features, keeping only the most efficient and useful ones.

Thanks to the use of an object oriented language and of an up to date and rapidly evolving modeling engine, EADock is designed to fulfill both requirements, leaving the room for the usual bug fixes and progressive methodological refinements as well as for big methodological jumps that are believed to be required to improve docking accuracy [65].

Several of these improvements are described in this chapter. Some of them are already implemented but not tested, others are to be implemented in the coming months. The first category of improvements addresses the performance issues noticed during the validation (see Chapter 2 “Material and Methods”), regarding both the scoring function and the sampling heuristic. The second category addresses the usability of the software.

## **8.1 Improvements**

### **8.1.1 Performance improvement**

#### **8.1.1.1 Scoring**

##### **8.1.1.1.1 Challenges**

As stated by [77] “Today’s understanding of the physical chemistry of molecular recognition may prove to be incomplete. Specific issues likely to be of continuing interest include changes in ligand, protein, and solvent entropy upon binding; the suspicious tendency of many binding models to yield affinities that correlate strongly with the molecular weight of the ligand; electronic polarization; the modeling of metal-containing binding sites; and the thermodynamic implications of water binding sites at the ligand-protein interface.”

As mentioned in the Chapter 1 “Introduction” and reflected by our benchmark, no ideal scoring function has been found yet. Several general trends can be identified, and will be investigated using EADock.

##### **8.1.1.1.2 A need for decoys**

The development of scoring functions may become tedious if a large number of different and *a priori* equally interesting ideas are formulated. The evaluation of the corresponding scoring functions must be performed and take into account its ability to impulse a driving force and to discriminate between the right solution and a set of decoys. Such requirements are often incompatible, as the former implies a focus on the general trend while the latter requires a careful inspection of small variations. An interesting discussion about this problem focusing on the way the van der Waals interactions are taken into account can be found in [40].

A convenient way to optimize a scoring function regarding to both needs is to use decoys. Its ability to drive the search can be trained and estimated by using rough decoys.

Conversely, highly refined decoys can be used to train and estimate the ability of a scoring function to recognize the right solution.

One of the databases mentioned in the Chapter 1 “Introduction”, the LPDB, provides both rough and refined decoys for each test complex, and is well adapted to CHARMM as the corresponding PSF and CRD files are available. We are planning to use it to test the ability of new scoring functions to drive the search. To assess the selectivity of new scoring functions, highly refined decoys will also be generated by EADock.

#### **8.1.1.1.3 Refine the scoring**

Once a clear benchmark has been set up, several ideas are worth investigating. The first is an extension of the current scoring function of EADock to be able to deal with ligands that may bind covalently to their receptor. The second is to account for the entropy contribution to the binding free energy. The third is the use of polarizable force field. The forth approach could be used to merge the current scoring strategy with a QSAR relationship when several compounds with known activities are docked on the same receptor.

#### ***User defined extra potential***

CHARMM gives the opportunity for the user to define its own distance restraints. A combination of two such potentials can be fitted on quantum calculations to mimic the formation of a covalent bond between the ligand and its receptor (Figure 44). Such potential can be added to both fitness functions of EADock, allowing the docking of ligand interacting with metalloenzymes, such as the iron of the heme group of the IDO enzyme (see Chapter 3 “Applications”). This was implemented in EADock and is currently being tested. Preliminary results showed a better agreement between the predicted and the experimental binding modes (Figure 45).

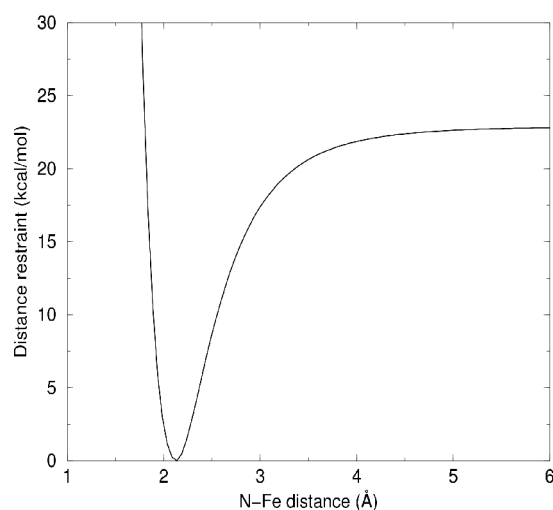


Figure 44: Example potential that can mimic the formation of a covalent bond. See text for details.

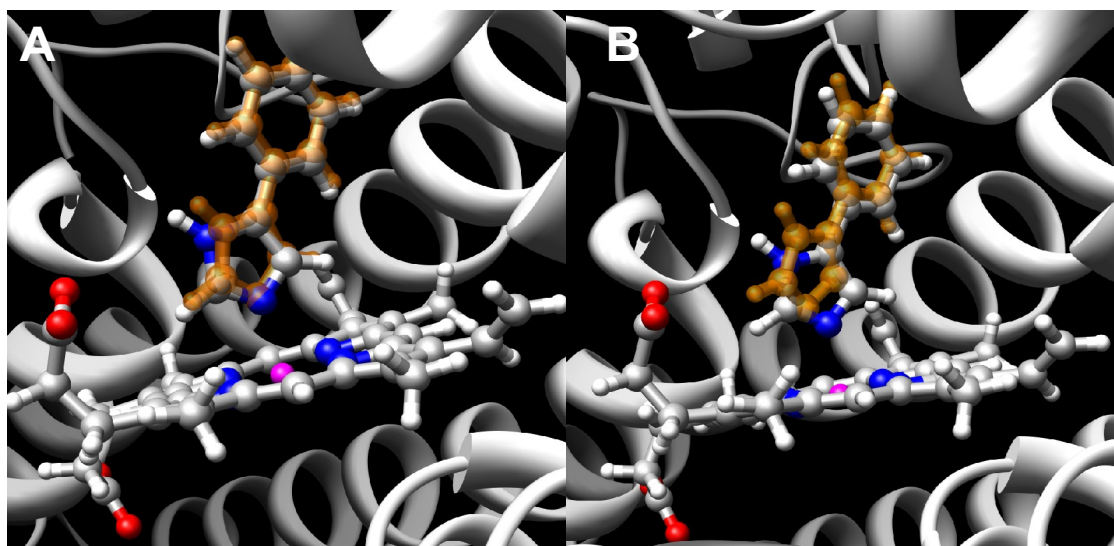


Figure 45: (A) Docking of the phenylimidazole (PIM) with an extra potential mimicking the formation of a covalent bound. The predicted binding mode is shown in transparent orange, the atoms of the ligand in the native binding mode are colored according to their type. The distance between PIM and the iron is 2 Å, and the RMSD to the native binding mode is 0.5 Å. (B) Docking of PIM without the extra potential. The distance between PIM and iron is 2.8 Å. Although the RMSD to the native binding mode is 0.85 Å, this binding mode is visually less convincing.

### ***Iterative integration with QSAR or LIE/LIECE approach***

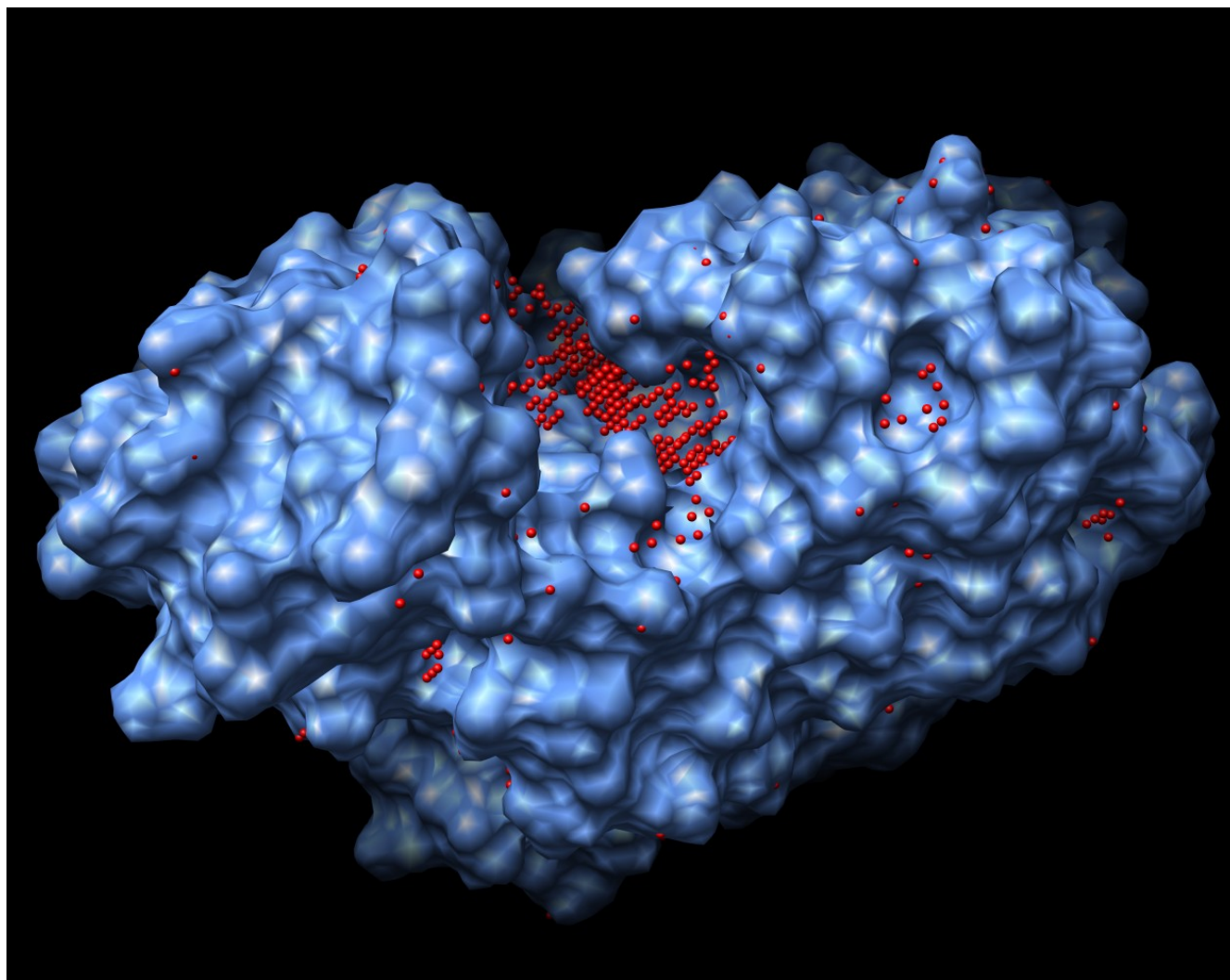
When a set of compounds with known activities has been docked on the same target, they can be used to generate an initial QSAR model [235] or optimize the  $\alpha$  and  $\beta$  parameters of a rapid scoring functions like LIECE [236]. Such models can be combined with the scoring function of EADock to drive the search more efficiently. Other compounds can then be docked more accurately, and the corresponding binding modes can be used to iteratively refine the scoring model. Such an iterative approach might be worth investigating for the design of active compounds (see Chapter 3 “Applications”).

#### **8.1.1.2 Sampling**

As mentioned in [63] and explained in Chapter 1 “Introduction”, the objective function must be consistent with the sampling heuristic, and vice versa. A scoring function aiming at being very accurate cannot be combined with a rough sampling heuristic. Although the performance of EADock (as published in [70]) is good regarding to its sampling, we have several ideas to improve it. Three directions are currently being explored: the rational reduction of the search space, a more realistic description of molecular interactions, and a better efficiency of the evolutionary process itself.

##### **8.1.1.2.1 Search space reduction**

The search space can be reduced and its exploration made easier by several tricks. First, when generated, if ligand poses are not close enough to the receptor, the former should be attracted by the cavities of the latter. Second, the split of the ligand into flexible and rigid parts would be a convenient way to describe at the same time highly flexible ligands (such as peptides) and rigid ligands (such as cyclic molecules), simplifying the corresponding search space. Third, the refinement of dihedral angles should not be random but biased toward the most likely conformations [237]. Fourth, the list of binding modes that are blacklisted should be periodically reevaluated and cleaned up to define a smarter constraint on the sampling heuristic.



*Figure 46: cavities detected on the penicillopepsin/1apt test case, grid points are shown in red.*

Several problematic binding modes with a very limited contact between the ligand and the receptor were generated and observed during the validation of EADock and some applications described in the previous chapters. Although these binding modes are likely to be removed from the evolution because of their poor FullFitness, the current SmartAttractor procedure (see Chapter 2 “Material and Methods”) can probably be improved by attracting a remote ligand into protein cavities rather than to the closest part of the surface. Such cavities can be identified prior to the docking, by our own permissive variant of the PocketFinder algorithm [238], which is based on a grid (see Figure 46).



Depending on its accessibility, each point of the grid can be assigned a mean field potential attracting the heavy atoms of the ligand.

Then, if the sampling heuristic generates a binding mode in which the distance criteria between the ligand and the receptor is not met (see Chapter 2 “Material and Methods”), the potentials assigned to the closest grid points are used to attract the ligand in the corresponding cavity by 500 steps of ABNR minimization. To make the path to the cavity easier to follow for the ligand, it is made softer (see the SoftLigand operator in Material and Methods). The system is then relaxed in the unbiased force field by 500 steps of ABNR minimization. This approach is implemented in EADock and currently being assessed.

The sampling bias toward cavities can be finely tuned by tweaking the maximum distance criteria allowed between the ligand and the receptor, the number of attracting points taken into account, and by the potentials themselves. Such an approach is believed to be less restrictive than the docking of the ligand inside a binding pocket identified prior to the docking (see Chapter 1 “Introduction”). This latter approach allows an impressive reduction of the ROI but will fail if the binding pocket is not identified correctly. The method we propose here does not significantly slow down a docking run, and is likely to be more robust because it only adds a tunable bias to the sampling.

#### ***Adaptation to the complexity of the ligand***

Another way to limit the search space comes from the observation that while some ligands such as peptides are highly flexible, some are very rigid. For instance, most drug frameworks are made of one or several aromatic cycles. The conformational space of such ligands could be reduced by limiting the conformational exploration to their flexible parts, as their rigid parts are very unlikely to be distorted in a low-energy binding mode.

#### ***Rotamer libraries***

The current exploration of the dihedral angle space is performed by a discrete sampling of conformations generated by 60 degrees rotations. Such a rough exploration guarantees

that the dihedral space is more or less uniformly explored. However, this is not satisfying since some conformations are known to be more favorable than others [239]. Instead of a fixed step of 60 degrees, the sampling of dihedral angles should be biased toward these most favorable values.

Such an approach was recently successfully validated [237] for proteins, and can be easily implemented by considering existing rotamer libraries. A similar sampling bias could also be introduced for organic ligands. It could either be defined manually, or derived from MD simulations prior to the docking.

#### ***Blacklisted conformations***

The search space would be described more efficiently if the list of blacklisted binding modes was cleaned up regularly to filter out its redundancy.

#### ***8.1.1.2.2 More realistic description of the interaction***

The rationalization of the search space that would be allowed by the improvements described above can either lead to shorter run times, or it can be reinvested to open a way toward a better description of the molecular interaction taking place between the ligand and the receptor.

#### ***Conformational change of the receptor***

The first obvious step in this direction would be to take the flexibility of the receptor into account. A convenient way to do this implicitly is to use softened van der Waals potentials. This drives the search efficiently, but it comes at the price of selectivity [40]. This issue can be addressed by starting the docking procedure using a softening of the potential, which is progressively decreased, keeping both the efficiency and the selectivity of these interactions [40]. Such a trick can be combined with the exploration of the conformational space of the receptor, either by explicitly modifying the position of a group of atoms, or by allowing a passive induced fit by minimization. Such approaches are implemented in EADock, but not critically assessed yet even though it already led to very successful

results, as described in the previous chapter.

#### ***Free water molecules***

The interaction between a ligand and its receptor may involve water molecules participating in a network of hydrogen bonds stabilizing the complex. Such water mediated interactions were found in the thermolysin family of our test set (see Chapter 2 “Material and Methods”). Although the complex was stabilized enough by the GB-MV2 implicit solvation model used in the FullFitness, such a statistical description can not reproduce the interactions made by key water molecules. The introduction of water molecules, explicitly moved during the docking, might help docking correctly more difficult test cases.

#### **8.1.1.2.3 *Evolution should never rest***

All along the evolution, the population of binding modes is clustered many times. This clustering allows the identification of local minima and the coupling between the two fitness functions, and impacts directly the management of the diversity (see Chapter 2 “Introduction”).

Despite its critical role, the clustering algorithm used in [70] heavily depends on the ranking of the binding modes in the population. This obvious lack of reliability is a major concern for future evolutions. It could be replaced by an UPGMA-like clustering algorithm, as follows. The clustering of the two closest elements of the distance matrix between binding modes is done first, instead of taking them according to their order in the population. Second, the center of a cluster is not defined by its rank in the population to cluster. Instead, the binding mode with the lowest RMSD to the others within the same cluster is chosen as the center. This is likely to lead to a more reliable identification of local minima.

The FullFitness distribution of the elements within two distinct clusters may partly overlap. However, with the implementation published in [70], one of the two compared clusters will be discarded and its center will be blacklisted anyway. The fate of these

clusters can thus be very different despite very similar FullFitness, and the wrong one might be blacklisted because it has not been refined enough compared to the others. While this mechanism is a key feature of EADock, the probability of such a mistake can be evaluated by comparing the FullFitness distribution of the two clusters with a non-parametric Wilcoxon statistical test. Depending on the statistical significance of the energy difference, the cluster with the less favorable FullFitness may be saved or discarded. This rational statistical assessment, together with the clustering algorithm described above, was implemented in EADock, but the corresponding expected performance improvement was not assessed yet.

Another potentially interesting option would be to cluster the population according to the RMSD and to the SimpleFitness or FullFitness. Such a two-way clustering could perform even better by identifying large unfavorable regions of the search space, for instance corresponding to positions of a ligand floating around a charged sidechain of the receptor but not making additional interactions with it. Conversely, favorable regions would be more carefully inspected for energetic details. Unfortunately, the number of clustering happening during the evolution is huge (typically more than 10000), and the speed of the clustering algorithm may have a major impact on the docking speed and may be problematic from a software optimization point of view. Nevertheless, such two-way clustering is now under investigation.

Complementary to the safer ranking of more reliably identified local minima, the yield and robustness of the evolutionary process itself can be improved when facing a broad and more complex energy landscape. Once enough local minima are identified, newly generated binding modes are likely to be the one eliminated between generations. This lead to a stable population over several generations. Once a disruption happens, for instance coming from the blacklisting procedure, the identified local minima massively disappear and a significant part of the population is renewed (see Figure 47).

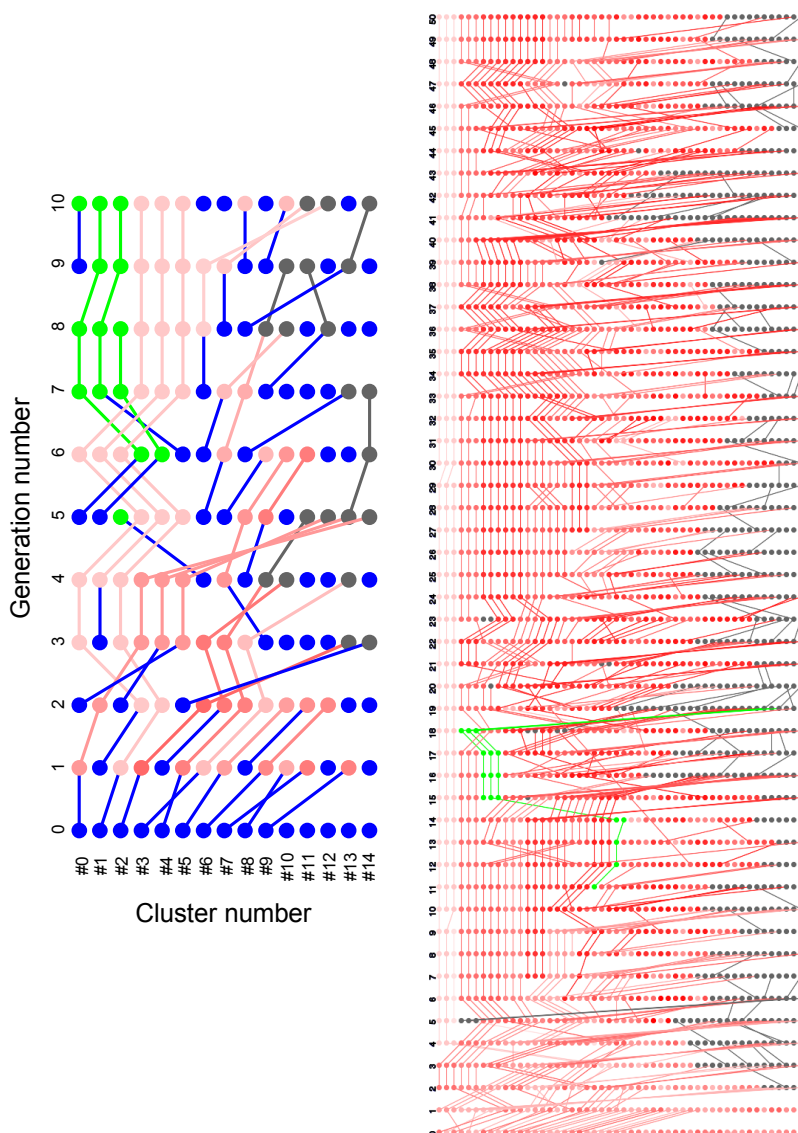


Figure 47: Left evolution of clusters ranking as a function of the generation. Segments are used to join the positions of a given cluster along the docking process. Green: binding modes < 2 Å RMSD to the experimental structure, gray: discarded solutions, blue: new solutions, pink: solutions above 2 Å RMSD (the closer from 2 Å, the lighter). Right, overview of a more realistic evolution: green: binding modes < 2 Å RMSD to the experimental structure, gray: discarded solutions, pink: solutions above 2 Å RMSD (the closer from 2 Å, the lighter). Relatively stable populations can be seen between generations 6 and 15, or 23 to 33, separated by mass extinction happening for instance at generation 16 or 40.

While a stabilization of the population during several generations may indicate a convergence of the evolutionary process, the corresponding generations are useless until another interesting region of the search space is identified or refined. Such identification or refinement implies that a sufficient number of binding modes in the population are available from an evolutionary point of view to carry out the sampling.

A convenient way to challenge a stable population without using a sharp blacklisting would be to temporary overweight one or several energetic terms in the two fitness functions, such as the van der Waals interaction energy, self energy of the ligand, the electrostatic interaction energy, or surface buried upon binding. The resulting disruption would result in the elimination of clusters having a reasonable overall FullFitness coming from compensated yet highly suspicious unfavorable interactions. Interestingly, the overweighting of van der Waals interactions would select the clusters with the most favorable ligand efficiency [98] (see Chapter 1 “introduction”).

The different criteria and the corresponding threshold could certainly be prioritized by looking at decoys in a database such as the LPDB (see Chapter 1 “introduction”). Alternatively, the most disruptive energy term and its ideal threshold value could be identified from the evolving population of binding mode itself.

### **8.1.2 Usability**

During the past years, EADock has been used as a docking toolbox, able to test and validate new ideas trying to reach an interesting performance level. As described above this process will go on in the future. On the other side, what makes a software interesting is its ability to go successfully from the world of benchmarks to real world applications. Such a transition requires a favorable balance between the performance of the program and its usability. As discussed above, the former can be improved by several means. This is also the case for the latter, by improving the input, the output, and the speed of EADock.

### **8.1.2.1 Input**

A first step toward a better usability is to limit the amount of information that EADock has to feed with, and by using a comprehensive user interface.

#### **8.1.2.1.1 Smarter user input**

Both docking-specific and evolutionary parameters should be at least suggested, if not completely automated. The former include the definition of the rotatable dihedral angles, symmetries, an easy management of tautomeric states, and an integration with already-existing scripts to generate PSF and CRD files from PDB files easily

An automatic determination of reasonable evolutionary parameters would allow a better management of docking jobs. Such an automatic determination routine could take into account the size of the ROI, the number of putative binding pockets of the protein (for instance depending on its accessible surface and number of identified binding pockets, see above), and the number of different conformations accessible for the ligand (which depends on the type and number of the corresponding degrees of freedom).

To avoid a tedious resubmission of jobs that crashed or had been stopped, restart files will have to be generated.

#### **8.1.2.1.2 Better user interface**

While the command line is probably the most suitable interface for intensive docking studies running on clusters, end users might also be interested in more friendly interfaces.

##### **8.1.2.1.2.1 Graphical user interface**

As EADock itself is implemented in Java, the implementation of a Java-based graphical user interface would be straightforward, but would require a significant software development.

Alternatively, existing programs such as UCSF Chimera can be extended by plug-ins. The

integration of the docking software DOCK into Chimera is already on its way<sup>5</sup>, and a similar plug-in could be rapidly developed for EADock.

#### 8.1.2.1.2.2 Web service

During the last years, the biological community has been getting used to web-services for common bioinformatics tasks. Such an interface for EADock would allow researchers to submit their docking jobs easily (Figure 48), while the docking itself would be running on a dedicated back-end cluster (Figure 52). Such a web service is currently being developed.

The screenshot shows a web browser window titled "SwissDock webservice - konqueror". The address bar shows a URL starting with "http://". The page has a menu bar with "Location", "Edit", "View", "Bookmarks", "Tools", "Settings", and "Help". Below the menu bar is a search bar with "Google Search". The main content area is titled "SwissDock webservice" and contains the following sections:

- Initial set of coordinates**
  - CHARMM topology file (PSF) :
  - CHARMM coordinate file (CRD) :
  - Ligand topology file (RTF) :
  - Ligand parameter file (PAR) :
  - These files can be generated from a PDB file from [here](#)
- Definition of the search space**
  - Selection for the ligand : SELE SEGID  END
  - Region of interest radius :  Å
  - Coming features:
    - Free dihedrals of the ligand
    - flexible part of the receptor
    - ...
- Evolutionary parameters**
  - Population size :  agents
  - Renewal rate :
  - Number of generation :
- Seeding**
  - Generate seeds between  Å RMSD to the ligand in the initial set of coordinates. Coming features:
    - upload your own seeding file

A "Submit" button is located at the bottom of the form.

Figure 48: Prototype of a web service for EADock.

5 <http://www.cgl.ucsf.edu/chimera/docs/ContributedSoftware/dockprep/dockprep.html>



### 8.1.2.2 Speed

Compared to typical run times reported for other docking softwares (hours for EADock vs minutes for VS dedicated softwares), the docking accuracy of EADock comes at the price of speed. However, due to the need for human expertise when designing new drugs *in silico*, the docking itself is not necessarily a bottleneck. Nonetheless, the faster EADock, the better the user experience. Profiling results showed that depending on the size of the system, the CPU time used by CHARMM ranges from 95% to 99% of the total docking run.

Optimizing the Java code thus appears to be useless, leaving us with two options: either CHARMM itself should be accelerated, or the algorithm itself should be tweaked.

#### 8.1.2.2.1 CHARMM optimization

The compilation of CHARMM was found to have a deep impact on its speed, especially for some hardware architectures (IA64, PPC). For common platform (x86, AMD64), the GNU compiler, was found to perform much better although it is still significantly slower than proprietary compilers such as ICC. Choosing an optimal compiler is the first and easy way to accelerate CHARMM.

CHARMM was not developed with high performance computing in mind, but physical relevance and accuracy. The code readability was thus favored against raw performance. There is certainly room for improvement in the routines called by EADock, and such optimizations are currently being carried out in the SIB. Although very efficient, such optimizations should be implemented very carefully as they often lead to unmaintainable softwares when the optimization/readability balance becomes unfavorable.

A general trend for today's heavy computational tasks is to use hardware acceleration. In this case, a dedicated processor is designed for a specific task such as non bonded interactions calculations. Several hardware accelerators providers are available today to speed up molecular dynamics programs, such as ClearSpeed<sup>6</sup> or IBM's MD GRAPE<sup>7</sup>. Interestingly, the most widely distributed processors dedicated to hardware accelerations

<sup>6</sup> <http://www.clearspeed.com>

<sup>7</sup> <http://www.research.ibm.com/grape>

are the Graphic Processor Units (GPU) of general purpose graphics card. These highly specialized processors are much cheaper and progressively adapted to more general tasks<sup>8</sup> thanks to the active participation of the two manufacturers delivering the fastest solutions, NVIDIA and ATI. The MD simulation software GROMACS has been ported on such architecture, and is used by the Folding@home project<sup>9</sup>.

Such optimizations by hardware acceleration are currently being carried out in the SIB and are likely to have interesting consequences on the speed of CHARMM, and thus on the speed of EADock.

#### **8.1.2.2.2 Parallelization**

The single-CPU performance of CHARMM is a clear limitation of EADock today, and the optimization of an existing software as mentioned above requires expert knowledge and may takes time. Intrinsically, evolutionary algorithms are embarrassingly parallel problems that can be dramatically accelerated by spreading the calculations among several processors or machines.

The first approach is to split the generation of children and the fitness evaluation among several machines. Such a fine granularity is likely to be efficient, but the evolutionary process will have to wait until the slowest job has been completed, and the queuing system must be much faster than the individual jobs.

The second and more interesting approach is to have an evolutionary process on each CPU core, all of them being organized in so-called island models [22]. Such a distribution allows the preservation of the diversity in the different populations, and was found to be more reliable than a single evolutionary process [22]. The scaling is expected to be extremely good even on slow networks because the required bandwidth is very low: only the most interesting binding modes are exchanged between the populations (Figure 49)

Island parallelization is technically easy to deploy in a research environment, where several computers are usually available. Modern CPUs are so fast that they are waiting

<sup>8</sup> <http://www.gpgpu.org>

<sup>9</sup> <http://folding.stanford.edu>

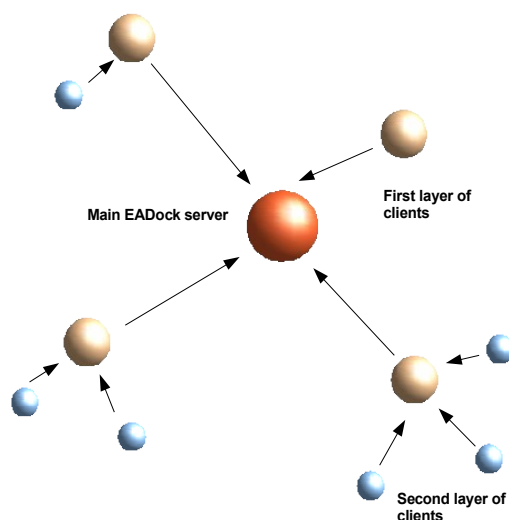


Figure 49: Example of a topology achievable with EADock. The migrations of interesting binding modes is represented by arrows.

most of the time and a parallelization on workstations allows a more rational usage of the computing resources. Nowadays, there is a great interest in such grid computing approaches like Swiss Bio Grid (<http://www.swissbiogrid.org>). A flexible implementation of the parallelization through such island models is available in EADock.

Alternatively, the growing online community using broadband Internet access inspired the BOINC project [<http://boinc.berkeley.edu>]. The high number of CPU available should not hide that their reliability decreases as the time required by a job increases. Several issues can be pointed out with these projects, such as the most commonly found operating system, Microsoft® Windows® (which only has a limited uptime), the license of the programs distributed and the legal issues that may be raised if an interesting results comes out.

### 8.1.2.3 Output

Once a docking run is launched, its status can be monitored in a text file. Unfortunately, this output has not been assigned to a single object responsible for the interaction with the user, but is spread among many objects. This should be addressed in the coming

releases. An exception framework catching CHARMM crashes should also be implemented to help understanding what has gone wrong.

Once complete, the binding modes predicted by EADock are dumped in CRD format, so that they can be easily read by CHARMM or translated into PDB and visualized. A visualization tool for the docking outputs is available in UCSF Chimera. Although it has been implemented for DOCK, the docking results from EADock can be loaded and nicely represented/filtered (see Figure 50).

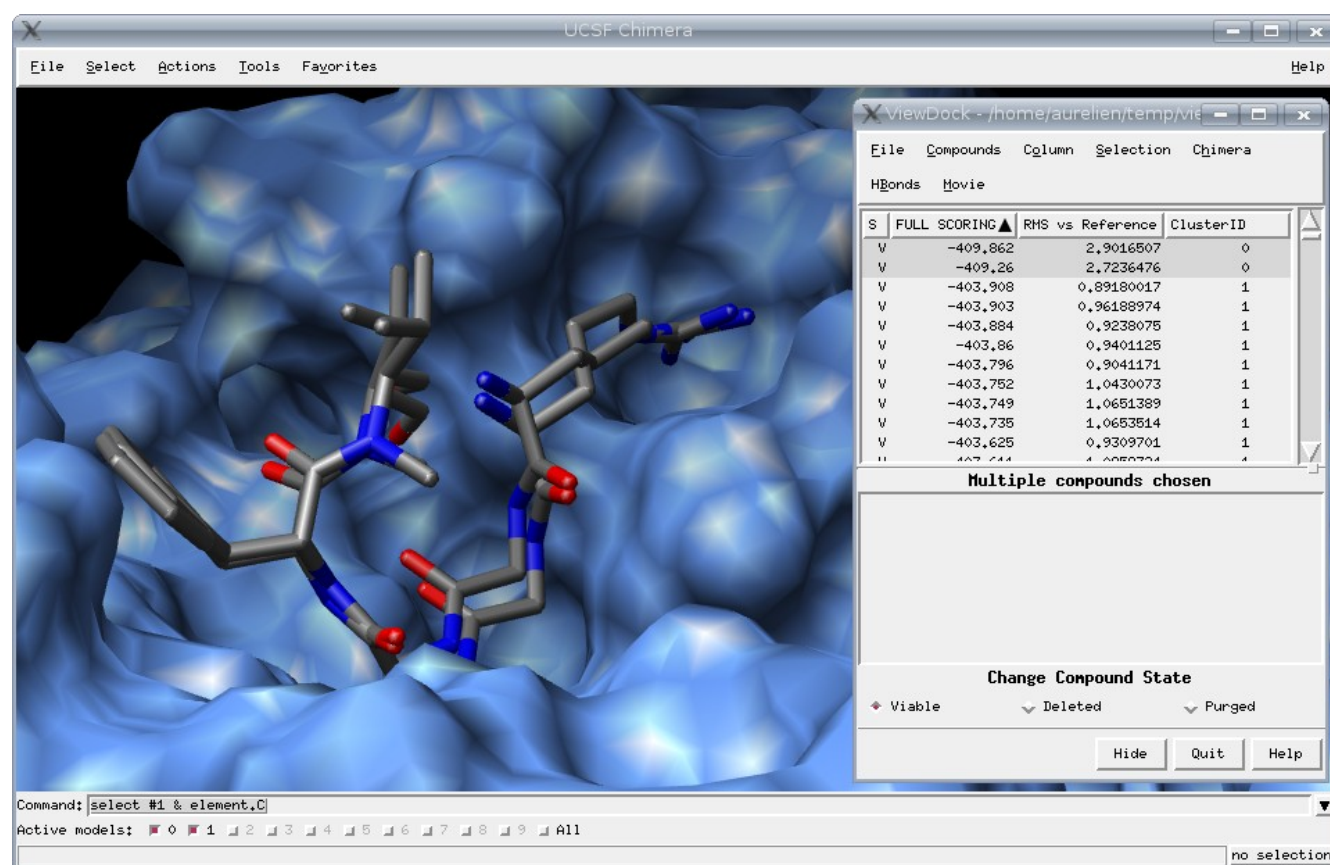


Figure 50: EADock predictions visualized with the UCSF Chimera/ViewDock plugin.

An extension to ViewDock would be interesting to implement in order to take into account the specific features of EADock, for instance to select and represent clusters of binding modes easily, or to map energy differences into color space.

Finally, due to the huge amount of information processed during a docking run, a bridge between EADock and a database should be implemented, to allow an easy mining of the

evolutionary process and an easier access to decoys. Such a database would be much more convenient than a flat file storage to centralize several tens of dockings for several tens of projects. Such a database is currently being developed (see Figure 51) and a putative information pipeline is depicted in Figure 52.

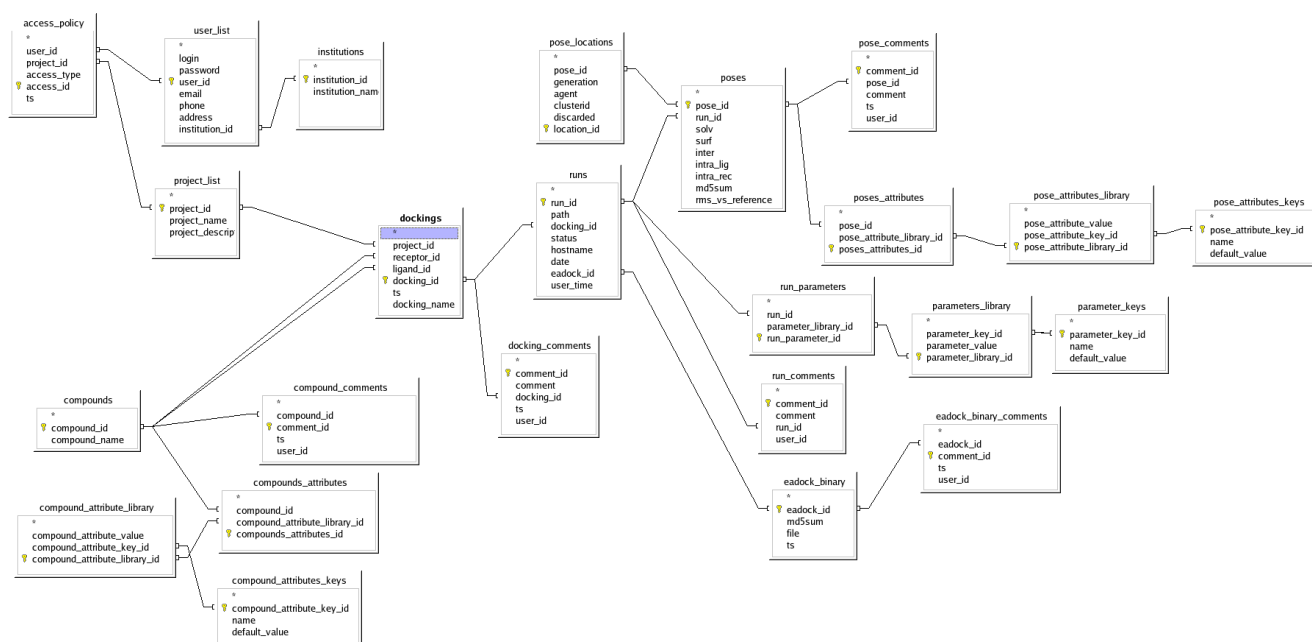


Figure 51: EADockDB overview

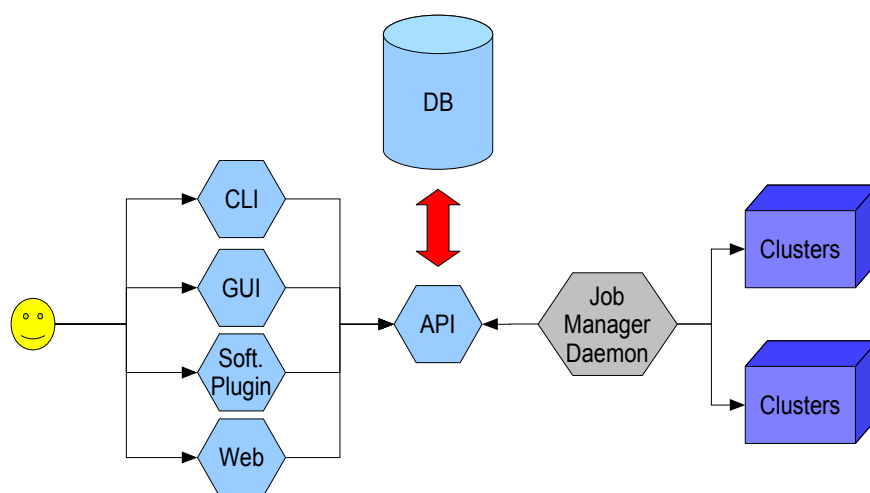


Figure 52: Information flowchart

## **8.2 Conclusion**

While the docking accuracy has been found to be relevant (see Chapter 2 “Material and Methods” and [70], and while the software usability led to satisfying applications (see Chapter 3 “Applications”), the methodological and software developments are not stopped. On the contrary, 32 new versions of EADock were released between June 2006 and January 2007, implementing most of the features described in this chapter. Today, the bottleneck is certainly the lack of a clean database to benchmark these improvements. To this aim, a deep inspection and manual curation of the LPDB is currently being performed in our group by Vincent Zoete.







## 9 Conclusion

Theoretical molecular docking approaches are complementary to *in vitro* and *in vivo* experiments, and help the interpretation of biological observations. They can help understanding the key molecular interactions between a ligand and its receptor, and provide information for efficient structure-based molecular design of new active compounds. Moreover, they can help reducing the number of biological assays by providing criteria to focus on a subset of a large collections of molecules. On the other hand, experimental data is key to successful development of theoretical tools. This synergy between theoretical and experimental approaches is crucial for a satisfactory evolution of biology, pharmacology and medicine. Such a so-called “translational research”, where the usual “bench-to-bed” one-way processing of biological/medical knowledge is replaced by a two-way communication, is believed to shorten the path toward drug delivery to the patient.



## 10 Bibliography

- 1: M. Moran, Cost of Bringing New Drugs To Market Rising Rapidly, Psychiatr News (38) 25-, 2003
- 2: S. Frantz, Pharma faces major challenges after a year of failures and heated battles, Nat Rev Drug Discov (6) 5--7, 2007
- 3: P. M. Danzon and A. Epstein and S. Nicholson, Mergers and Acquisitions in the Pharmaceutical and Biotech Industries, NBER Working Paper (10536) , 2004
- 4: S. Frantz, Pipeline problems are increasing the urge to merge., Nat Rev Drug Discov (5) 977--979, 2006
- 5: G. M. Keseru and G. M. Makara, Hit discovery and hit-to-lead approaches., Drug Discov Today (11) 741--748, 2006
- 6: S. Frantz, Study reveals secrets to faster drug development., Nat Rev Drug Discov (5) 883, 2006
- 7: S. Kummar and R. Kinders and L. Rubinstein and R. E. Parchment . and A. J. Murgo and J. Collins and O. Pickeral and J. Low and S. M. Steinberg and M. Gutierrez and S. Yang and L. Helman and R. Wiltout and J. E. Tomaszewski and J. H. Doroshow, Compressing drug development timelines in oncology using phase '0' trials, Nat Rev Cancer (7) 131--139, 2007
- 8 R. S. Larson, Bioinformatics And Drug Discovery, 2005
- 9: W. L. Jorgensen, The many roles of computation in drug discovery., Science (303) 1813--1818, 2004
- 10: D. B. Kitchen and H. Decornez and J. R. Furr and J. Bajorath, Docking and scoring in virtual screening for drug discovery: methods and applications., Nat. Rev. Drug. Discov. (3) 935-49, 2004
- 11: R. A. Friesner and J. L. Banks and R. B. Murphy and T. A. Halgren and J. J. Klicic and D.

- T. Mainz and M. P. Repasky and E. H. Knoll and M. Shelley and J. K. Perry and D. E. Shaw and P. Francis and P. S. Shenkin, Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy., *J. Med. Chem.* (47) 1739-49, 2004
- 12: T. J. Ewing and S. Makino and A. G. Skillman and I. D. Kuntz, DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases., *J. Comput. Aided. Mol. Des.* (15) 411--428, 2001
- 13: H. Claussen and C. Buning and M. Rarey and T. Lengauer, FlexE: efficient molecular docking considering protein structure variations., *J. Mol. Biol.* (308) 377-95, 2001
- 14: M. D. Miller and S. K. Kearsley and D. J. Underwood and R. P. Sheridan, FLOG: a system to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure., *J. Comput. Aided. Mol. Des.* (8) 153--174, 1994
- 15: W. Welch and J. Ruppert and A. N. Jain, Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites., *Chem. Biol.* (3) 449--462, 1996
- 16: N. Budin and N. Majeux and A. Caflisch, Fragment-Based flexible ligand docking by evolutionary optimization., *Biol. Chem.* (382) 1365-72, 2001
- 17: A. E. Cho and J. A. Wendel and N. Vaidehi and P. M. Kekenyes-Huskey and W. B. Floriano and P. K. Maiti and W. A. Goddard, The MPSim-Dock hierarchical docking algorithm: application to the eight trypsin inhibitor cocrystals., *J. Comput. Chem.* (26) 48-71, 2005
- 18: G. Wu and D. H. Robertson and C. L. Brooks and M. Vieth, Detailed analysis of grid-based molecular docking: A case study of CDOCKER-A CHARMM-based MD docking algorithm., *J. Comput. Chem.* (24) 1549-62, 2003
- 19: M. Vieth and J. D. Hirst and A. Kolinski and C. L. Brooks, Assessing energy functions for flexible docking., *J. Comput. Chem.* (19) 1612-1622, 1998
- 20: G. M. Morris and D. S. Goodsell and R. S. Halliday and R. Huey and W. E. Hart and R. K. Belew and A. J. Olson, Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function., *J. Comput. Chem.* (19) 1639-1662, 1998

- 21 L. Davis, Handbook of Genetic Algorithms, 1991
- 22 D. E. Goldberg, Genetic Algorithms in Search Optimization and Machine Learning, 1989
- 23: G. Jones and P. Willett and R. C. Glen and A. R. Leach and R. Taylor, Development and validation of a genetic algorithm for flexible docking., J. Mol. Biol. (267) 727-48, 1997
- 24: J. S. Taylor and R. M. Burnett, DARWIN: a program for docking flexible molecules., Proteins (41) 173-91, 2000
- 25: F. Glaser and R. J. Morris and R. J. Najmanovich and R. A. Laskowski and J. M. Thornton, A method for localizing ligand binding pockets in protein structures., Proteins (62) 479--488, 2006
- 26: M. Nayal and B. Honig, On the nature of cavities on protein surfaces: Application to the identification of drug-binding sites., Proteins () , 2006
- 27: J. An and M. Totrov and R. Abagyan, Pocketome via comprehensive identification and classification of ligand binding envelopes., Mol. Cell. Proteomics (4) 752--761, 2005
- 28: H. Chen and P. D. Lyne and F. G. and T. Lovell and J. Li, On evaluating molecular-docking methods for pose prediction and enrichment factors., J. Chem. Inf. Model. (46) 401--415, 2006
- 29: E. Perola and W. P. Walters and P. J. Charifson, Comments on the Article "On Evaluating Molecular-Docking Methods for Pose Prediction and Enrichment Factors", J. Chem. Inf. Model. () --, 2007
- 30: T. Simonson and G. Archontis and M. Karplus, Free energy simulations come of age: protein-ligand recognition., Acc. Chem. Res. (35) 430--437, 2002
- 31: P. Kollman, Free Energy Calculations: Applications to Chemical and Biochemical Phenomena, Chemical Reviews (93) 2395--2417, 1993
- 32: M. K. Holloway, A priori Prediction of Activity for HIV-1 Protease Inhibitors Employing Energy Minimization in the Active Site, J. Med. Chem. (38) 305-317, 1995

- 33: C. Perez and M. Pastor and A. R. Ortiz and F. Gago, Comparative binding energy analysis of HIV-1 Protease inhibitors: incorporation of solvent effects and validation as a powerful tool in receptor-based drug design, *J. Med. Chem.* (41) 836-852, 1998
- 34: H. J. Boehm, The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure, *J. Comput. Aided Mol. Des.* (8) 243-256, 1994
- 35: H. J. Boehm, 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, *J. Comput. Aided Mol. Des.* (12) 309-323, 1998
- 36: R. D. Head and M. L. Smythe and T. I. Oprea and C. L. Waller and S. M. Green and G. R. Marshall, VALIDATE: a new method for the Receptor-Based prediction of binding affinities of Novel ligands, *J. Am. Chem. Soc.* (118) 3959-3969, 1996
- 37: I. Muegge and Y. Martin, A general and fast scoring function for protein-ligand interactions: a simplified potential approach, *J. Med. Chem.* (42) 791-804, 1999
- 38: I. Muegge, A knowledge-based scoring function for protein-ligand interactions: Probing the reference state, *Perspectives in Drug Discovery and Design* (20) 99-114, 2000
- 39: P. Ferrara and H. Gohlke and D. J. Price and G. Klebe and C. L. Brooks, Assessing scoring functions for protein-ligand interactions., *J. Med. Chem.* (47) 3032-47, 2004
- 40: C. J. Camacho and S. Vajda, Protein docking along smooth association pathways., *Proc. Natl. Acad. Sci. U S A* (98) 10636--10641, 2001
- 41: B. R. Brooks and R. E. Bruccoleri and B. D. Olafson and D. J. States and S. Swaminathan and M. Karplus, CHARMM: a program for macromolecular energy minimization, and dynamics calculations., *J. Comput. Chem.* (4) 187-217, 1983
- 42: P. S. Charifson and J. J. Corkery and M. A. Murcko and W. P. Walters, Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins., *J Med Chem* (42) 5100--5109, 1999
- 43: A. D. MacKerell and D. Bashford and M. Bellott and R. L. Dunbrack and J. D. Evanseck

and M. J. Field and S. Fischer and J. Gao and H. Guo and S. Ha and D. Joseph-McCarthy and L. Kuchnir and K. Kuczera and F. T. K. Lau and C. Mattos and S. Michnick and T. Ngo and D. T. Nguyen and B. Prodhom and W. E. Reiher and B. Roux and M. Schlenkrich and J. C. Smith and R. Stote and J. Straub and M. Watanabe and J. Wiorkiewicz-Kuczera and D. Yin and M. Karplus, All-atom empirical potential for molecular modeling and dynamics studies of proteins, *Journal of Physical Chemistry B* (102) 3586-3616, 1998

44: M. S. Lee and F. R. Salsbury and C. L. Brooks, Novel generalized Born methods, *J. Chem. Phys.* (116) 10606-10614, 2002

45: M. S. Lee and M. Feig and F. R. Salsbury and C. L. Brooks, New analytic approximation to the standard molecular volume definition and its application to generalized born calculations, *J. Comput. Chem.* (24) 1348-1356, 2003

46: S. F. Sousa and P. A. Fernandes and M. J. Ramos, Protein-ligand docking: current status and future challenges., *Proteins* (65) 15--26, 2006

47: M. Rarey and B. Kramer and T. Lengauer and G. Klebe, A fast flexible docking method using an incremental construction algorithm., *J Mol Biol* (261) 470--489, 1996

48: R. Abagyan and M. Totrov and D. Kuznetsov, ICM—a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation, *J. Comput. Chem.* (15) 488--506, 1994

49: Y. Duan and C. Wu and S. Chowdhury and M. C. Lee and G. Xiong and W. Zhang and R. Yang and P. Cieplak and R. Luo and T. Lee and J. Caldwell and J. Wang and P. Kollman, A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations., *J Comput Chem* (24) 1999--2012, 2003

50: H. Park and J. Lee and S. Lee, Critical assessment of the automated AutoDock as a new docking tool for virtual screening., *Proteins* (65) 549--554, 2006

51: C. Hetényi and D. vanderSpoel, Efficient docking of peptides to proteins without prior knowledge of the binding site., *Protein Sci.* (11) 1729-37, 2002

- 52: A. M. Ruvinsky and A. V. Kozintsev, New and fast statistical-thermodynamic method for computation of protein-ligand binding entropy substantially improves docking accuracy., J. Comput. Chem. (26) 1089-95, 2005
- 53: C. Hetényi and G. Paragi and U. Maran and Z. Timár and M. Karelson and B. Penke, Combination of a Modified Scoring Function with Two-Dimensional Descriptors for Calculation of Binding Affinities of Bulky, Flexible Ligands to Proteins., J. Am. Chem. Soc. (128) 1233-1239, 2006
- 54: N. Moitessier and E. Westhof and S. Hanessian, Docking of Aminoglycosides to Hydrated and Flexible RNA., J. Med. Chem. (49) 1023-1033, 2006
- 55: J. W. M. Nissink and C. Murray and M. Hartshorn and M. L. Verdonk and J. C. Cole and R. Taylor, A new test set for validating predictions of protein-ligand interaction., Proteins (49) 457-471, 2002
- 56: E. Kellenberger and J. Rodrigo and P. Muller and D. Rognan, Comparative evaluation of eight docking tools for docking and virtual screening accuracy., Proteins (57) 225-242, 2004
- 57: B. D. Bursulaya and M. Totrov and R. Abagyan and C. L. Brooks, Comparative study of several algorithms for flexible ligand docking., J. Comput. Aided. Mol. Des. (17) 755-63, 2003
- 58: R. M. Knegtel and I. D. Kuntz and C. M. Oshiro, Molecular docking to ensembles of protein structures., J Mol Biol (266) 424-440, 1997
- 59: X. Zou and S. Yaxiong and I. D. Kuntz, Inclusion of Solvation in Ligand Binding Free Energy Calculations Using the Generalized-Born Model, Journal of the American Chemical Society (121) 8033-8043, 1999
- 60: JY Trosset and HA Scheraga, Reaching the global minimum in docking simulations: a Monte Carlo energy minimization approach using Bezier splines., Proc. Natl. Acad. Sci. U S A (95) 8011-5, 1998
- 61: R. Abagyan and M. Totrov, Biased probability Monte Carlo conformational searches



- and electrostatic calculations for peptides and proteins., J Mol Biol (235) 983--1002, 1994
- 62: G. Warren and C. Andrews and A.-M. Capelli and B. Clarke and J. Lalonde and M. Lambert and M. Lindvall and N. Nevins and S. Semus and S. Senger and G. Tedesco and I. Wall and J. Woolven and C. Peishoff and M. Head, A Critical Assessment of Docking Programs and Scoring Functions., J Med Chem (49) 5912--5931, 2006
- 63: M. Kontoyianni and G. S. Sokol and L. M. McClellan, Evaluation of library ranking efficacy in virtual screening., J Comput Chem (26) 11--22, 2005
- 64: J. C. Cole and C. W. Murray and J. Willem and M. Nissink and R. D. Taylor and R. Taylor, Comparing protein-ligand docking programs is difficult., Proteins ( ) , 2005
- 65: A. R. Leach and B. K. Shoichet and C. Peishoff, Prediction of Protein-Ligand Interactions. Docking and Scoring: Successes and Gaps., J Med Chem (49) 5851--5855, 2006
- 66: K. A. Thiel, Structure-aided drug design's next generation., Nat Biotechnol (22) 513--519, 2004
- 67: O. Roche and R. Kiyama and C. L. Brooks, Ligand-protein database: linking protein-ligand complex structures to binding data., J. Med. Chem. (44) 3592-8, 2001
- 68: R. Wang and X. Fang and Y. Lu and C.-Y. Yang and S. Wang, The PDBbind database: methodologies and updates., J. Med. Chem. (48) 4111-9, 2005
- 69: L. Hu and M. L. Benson and R. D. Smith and M. G. Lerner and H. A. Carlson, Binding MOAD (Mother Of All Databases)., Proteins (60) 333--340, 2005
- 70: A. Grosdidier and V. Zoete and O. Michielin, EADock: docking small molecules into protein active sites with a multiobjective evolutionary optimization, Proteins: structure, function and bioinformatics (in press) , 2007
- 71: R. T. Kroemer and A. Vulpetti and J. J. McDonald and D. C. Rohrer and J.-Y. Trosset and F. Giordanetto and S. Cotesta and C. McMartin and M. Kihlén and P. F. W. Stouten, Assessment of docking poses: interactions-based accuracy classification (IBAC) versus crystal structure deviations., J. Chem. Inf. Comput. Sci. (44) 871-81, 2004

- 72: E. Perola and P. S. Charifson, Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding., J. Med. Chem. (47) 2499-510, 2004
- 73: M. Kontoyianni and L. M. McClellan and G. S. Sokol, Evaluation of docking performance: comparative data on docking algorithms., J Med Chem (47) 558--565, 2004
- 74: J. Tirado-Rives and W. L. Jorgensen, Contribution of conformer focusing to the uncertainty in predicting free energies for protein-ligand binding., J Med Chem (49) 5880--5884, 2006
- 75: N. P. Todorov and P. H. Monthoux and I. L. Alberts, The influence of variations of ligand protonation and tautomerism on protein-ligand recognition and binding energy landscape., J Chem Inf Model (46) 1134--1142, 2006
- 76: A. J. S. Knox and M. J. Meegan and G. Carta and D. G. Lloyd, Considerations in compound database preparation--"hidden" impact on virtual screening results., J Chem Inf Model (45) 1908--1919, 2005
- 77: M. K. Gilson and H.-X. Zhou, Calculation of Protein-Ligand Binding Affinities., Annu Rev Biophys Biomol Struct ( ) , 2007
- 78: A.J. Bordner and C.N. Cavasotto and R.A. Abagyan, Accurate Transferable Model for Water, *n*-Octanol, and *n*-Hexadecane Solvation Free Energies, Journal of Physical Chemistry B (106) 11009-11015, 2002
- 79: M. L. Verdonk and G. Chessari and J. C. Cole and M. J H. and C. W. Murray and J. W. M. Nissink and R. D. Taylor and R. Taylor, Modeling water molecules in protein-ligand docking using GOLD., J Med Chem (48) 6504--6515, 2005
- 80: T. Young and R. Abel and B. Kim and B. J. Berne and R. A. Friesner, Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding., Proc Natl Acad Sci U S A (104) 808--813, 2007
- 81: G. Klebe, Virtual ligand screening: strategies, perspectives and limitations., Drug Discov Today (11) 580--594, 2006

- 82: F. Osterberg and G. M. Morris and M. F. Sanner and A. J. Olson and D. S. Goodsell, Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock., *Proteins* (46) 34--40, 2002
- 83: K. L. Meagher and H. A. Carlson, Incorporating protein flexibility in structure-based drug discovery: using HIV-1 protease as a test case., *J. Am. Chem. Soc.* (126) 13276-81, 2004
- 84: S. Radestock and M. Boehm and H. Gohlke, Improving binding mode predictions by docking into protein-specifically adapted potential fields., *J Med Chem* (48) 5466--5479, 2005
- 85: P. H. J. Keizers and C. deGraaf and F. J. J. deKanter and C. Oostenbrink and K. A. Feenstra and J. N. M. Commandeur and N. P. E. Vermeulen, Metabolic regio- and stereoselectivity of cytochrome P450 2D6 towards 3,4-methylenedioxy-N-alkylamphetamines: in silico predictions and experimental validation., *J Med Chem* (48) 6117--6127, 2005
- 86: R. Wang and Y. Lu and S. Wang, Comparative evaluation of 11 scoring functions for molecular docking., *J. Med. Chem.* (46) 2287-303, 2003
- 87: J. R. Proudfoot, Drugs, leads, and drug-likeness: an analysis of some recently launched drugs., *Bioorg Med Chem Lett* (12) 1647--1650, 2002
- 88: D. A. Erlanson and R. S. McDowell and T. O'Brien, Fragment-based drug discovery., *J Med Chem* (47) 3463--3482, 2004
- 89: C. A. Lipinski and F. Lombardo and B. W. Dominy and P. J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, *Advanced Drug Delivery Reviews* (46) 3--26, 2001
- 90: V. Kairys and M. X. Fernandes and M. K. Gilson, Screening drug-like compounds by docking to homology models: a systematic study., *J. Chem. Inf. Model.* (46) 365--379, 2006
- 91: C. Bissantz and G. Folkers and D. Rognan, Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations., *J Med Chem* (43)

4759--4767, 2000

92: D. Huang and U. Luethi and P. Kolb and M. Cecchini and A. Barberis and A. Caflisch, In silico discovery of beta-secretase inhibitors., J Am Chem Soc (128) 5436--5443, 2006

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

94: D. A. Erlanson and J. A. Wells and A. C. Braisted, Tethering: fragment-based drug discovery., Annu Rev Biophys Biomol Struct (33) 199--223, 2004

95: D. A. Erlanson, Fragment-based lead discovery: a chemical update., Curr Opin Biotechnol (17) 643--652, 2006

96: R. Carr and M. Congreve and C. W. Murray and D. C. Rees, Fragment-based lead discovery: leads by design., Drug Discov Today (10) 987--992, 2005

97: M. Congreve and R. Carr and C. Murray and H. Jhoti, A 'rule of three' for fragment-based lead discovery?, Drug Discov Today (8) 876--877, 2003

98: A. L. Hopkins and C. R. Groom and A. Alex, Ligand efficiency: a useful metric for lead selection., Drug Discov Today (9) 430--431, 2004

99: G. W. Bemis and M. A. Murcko, The properties of known drugs. 1. Molecular frameworks., J. Med. Chem. (39) 2887--2893, 1996

100: G. W. Bemis and M. A. Murcko, Properties of known drugs. 2. Side chains., J. Med. Chem. (42) 5095--5099, 1999

101: X. Q. Lewell and D. B. Judd and S. P. Watson and M. M. Hann, 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 (38) 511--522, 1998

102: P. J. Hajduk and M. Bures and J. Praestgaard and S. W. Fesik, Privileged molecules for protein binding identified from NMR-based screening., J Med Chem (43) 3443--3447, 2000

- 103: E. Jacoby and J. D. and M. J. J. Blommers, Design of small molecule libraries for NMR screening and other applications in drug discovery., *Curr Top Med Chem* (3) 11--23, 2003
- 104: M. Vieth and M. G. Siegel and R. E. Higgs and I. A. Watson and D. H. Robertson and K. A. Savin and G. L. Durst and P. A. Hipkind, Characteristic physical properties and structural fragments of marketed oral drugs., *J Med Chem* (47) 224--232, 2004
- 105: D. C. Rees and M. Congreve and C. W Murray and R. Carr, Fragment-based lead discovery., *Nat Rev Drug Discov* (3) 660--672, 2004
- 106: D. J. Maly and I. C. Choong and J. A. Ellman, Combinatorial target-guided ligand assembly: identification of potent subtype-selective c-Src inhibitors., *Proc Natl Acad Sci U S A* (97) 2419--2424, 2000
- 107: S. B. Shuker and P. J. Hajduk and R. P. Meadows and S. W. Fesik, Discovering high-affinity ligands for proteins: SAR by NMR., *Science* (274) 1531--1534, 1996
- 108: S. A. Hofstadler and R. H. Griffey, Analysis of noncovalent complexes of DNA and RNA by mass spectrometry., *Chem Rev* (101) 377--390, 2001
- 109: E. E. Swayze and E. A. Jefferson and K. A. Sannes-Lowery and L. B. Blyn and L. M. Risen and S. Arakawa and S. A. Osgood and S. A. Hofstadler and R. H. Griffey, SAR by MS: a ligand based technique for drug lead discovery against structured RNA targets., *J Med Chem* (45) 3816--3819, 2002
- 110: Y. He and J. Yang and B. Wu and D. Robinson and K. Sprankle and P.-P. Kung and K. Lowery and V. Mohan and S. Hofstadler and E. E. Swayze and R. Griffey, Synthesis and evaluation of novel bacterial rRNA-binding benzimidazoles by mass spectrometry., *Bioorg Med Chem Lett* (14) 695--699, 2004
- 111: M. Krier and J. X. deAraújo-Júnior and M. Schmitt and J. Durantón and H. Justiano-Basaran and C. Lugnier and J.-J. Bourguignon and D. Rognan, Design of small-sized libraries by combinatorial assembly of linkers and functional groups to a given scaffold: application to the structure-based optimization of a phosphodiesterase 4 inhibitor., *J Med Chem* (48) 3816--3822, 2005

- 112: R. A. Lewis and P. M. Dean, Automated site-directed drug design: the formation of molecular templates in primary structure generation., *Proc R Soc Lond B Biol Sci* (236) 141--162, 1989
- 113: R. A. Lewis and P. M. Dean, Automated site-directed drug design: the concept of spacer skeletons for primary structure generation., *Proc R Soc Lond B Biol Sci* (236) 125--140, 1989
- 114: R. A. Lewis, Automated site-directed drug design: approaches to the formation of 3D molecular graphs., *J Comput Aided Mol Des* (4) 205--210, 1990
- 115: R. A. Lewis, Automated site-directed drug design: a method for the generation of general three-dimensional molecular graphs., *J Mol Graph* (10) 131--143, 1992
- 116: R. A. Lewis and D. C. Roe and C. Huang and T. E. Ferrin and R. Langridge and I. D. Kuntz, Automated site-directed drug design using molecular lattices., *J Mol Graph* (10) 66--78, 106, 1992
- 117: A. R. Leach and S. R. Kilvington, Automated molecular design: a new fragment-joining algorithm., *J Comput Aided Mol Des* (8) 283--298, 1994
- 118: N. P. Todorov and P. M. Dean, Evaluation of a method for controlling molecular scaffold diversity in de novo ligand design., *J Comput Aided Mol Des* (11) 175--192, 1997
- 119: V. J. Gillet and W. Newell and P. Mata and G. Myatt and S. Sike and Z. Zsoldos and A. P. Johnson, SPROUT: recent developments in the de novo design of molecules., *J Chem Inf Comput Sci* (34) 207--217, 1994
- 120: N. P. Todorov and P. M. Dean, A branch-and-bound method for optimal atom-type assignment in de novo ligand design., *J Comput Aided Mol Des* (12) 335--349, 1998
- 121: S. L. Chan and P. L. Chau and J. M. Goodman, Ligand atom partial charged assignment for complementary electrostatic potentials, *Journal of computer-aided molecular design* (6) 461--474, 1992
- 122: M. T. Barakat and P. M. Dean, The atom assignment problem in automated de novo drug design. 5. Tests for envelope-directed fragment placement based on molecular

similarity., J Comput Aided Mol Des (9) 457--462, 1995

123: M. T. Barakat and P. M. Dean, The atom assignment problem in automated de novo drug design. 4. Tests for site-directed fragment placement based on molecular complementarity., J Comput Aided Mol Des (9) 448--456, 1995

124: M. T. Barakat and P. M. Dean, The atom assignment problem in automated de novo drug design. 3. Algorithms for optimization of fragment placement onto 3D molecular graphs., J Comput Aided Mol Des (9) 359--372, 1995

125: M. T. Barakat and P. M. Dean, The atom assignment problem in automated de novo drug design. 2. A method for molecular graph and fragment perception., J Comput Aided Mol Des (9) 351--358, 1995

126: M. T. Barakat and P. M. Dean, The atom assignment problem in automated de novo drug design. 1. Transferability of molecular fragment properties., J Comput Aided Mol Des (9) 341--350, 1995

127: H. J. Boehm, LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads., J Comput Aided Mol Des (6) 593--606, 1992

128: P. J. Goodford, A computational procedure for determining energetically favorable binding sites on biologically important macromolecules., J Med Chem (28) 849--857, 1985

129: R. A. Laskowski and J. M. Thornton and C. Humblet and J. Singh, X-SITE: use of empirically derived atomic packing preferences to identify favourable interaction regions in the binding sites of proteins., J Mol Biol (259) 175--201, 1996

130: N. Majeux and M. Scarsi and J. Apostolakis and C. Ehrhardt and A. Caflisch, Exhaustive docking of molecular fragments with electrostatic solvation., Proteins (37) 88--105, 1999

131: A. Miranker and M. Karplus, Functionality maps of binding sites: a multiple copy simultaneous search method., Proteins (11) 29--34, 1991

132: M. B. Eisen and D. C. Wiley and M. Karplus and R. E. Hubbard, HOOK: a program for finding novel molecular architectures that satisfy the chemical and steric requirements of

a macromolecule binding site., *Proteins* (19) 199--221, 1994

133: C. M. Stultz and M. Karplus, Dynamic ligand design and combinatorial optimization: designing inhibitors to endothiapepsin., *Proteins* (40) 258--289, 2000

134: A. R. Leach and R. A. Bryce and A. J. Robinson, Synergy between combinatorial chemistry and de novo design., *J Mol Graph Model* (18) 358--67, 526, 2000

135: H. J. Boehm and M. Boehringer and D. Bur and H. Gmuender and W. Huber and W. Klaus and D. Kostrewa and H. Kuehne and T. Luebbers and N. Meunier-Keller and F. Mueller, 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* (43) 2664--2674, 2000

136: D. Huang and U. Luethi and P. Kolb and K. Edler and M. Cecchini and S. Audetat and A. Barberis and A. Caflisch, Discovery of cell-permeable non-peptide inhibitors of beta-secretase by high-throughput docking and continuum electrostatics calculations., *J Med Chem* (48) 5108--5111, 2005

137: C. A. Baxter and C. W. Murray and D. E. Clark and D. R. Westhead and M. D. Eldridge, Flexible docking using Tabu search and an empirical estimate of binding affinity., *Proteins* (33) 367--382, 1998

138: J. Pei and Q. Wang and Z. Liu and Q. Li and K. Yang and L. Lai, PSI-DOCK: Towards highly efficient and accurate flexible ligand docking., *Proteins* ( ) , 2006

139: F. Sirockin and C. Sich and S. Improta and M. Schaefer and V. Saudek and N. Froloff and M. Karplus and A. Dejaegere, Structure activity relationship by NMR and by computer: a comparative study., *J. Am. Chem. Soc.* (124) 11073--11084, 2002

140: R. B. Hermann, Theory of hydrophobic bonding. II. Correlation of hydrocarbon solubility in water with solvent cavity surface area, *J. Phys. Chem.* (76) 2754 - 2759, 1972

141: G. L. Amidon and S. H. Yalkowsky and S. T. Anik and S. C. Valvani, Solubility of nonelectrolytes in polar solvents. V. Estimation of the solubility of aliphatic monofunctional compounds in water using a molecular surface area approach, *J. Phys.*



Chem. (79) 2239 - 2246, 1975

142: H. Gohlke and C. Kiel and D. A. Case, Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes, J. Mol. Biol. (330) 891-913, 2003

143: W. Hasel and T. F. Hendrikson and W. C. Still, A rapid approximation to the solvent accessible surface areas of atoms, Tetrahedron Comput. Methodol. (1) 103-116, 1988

144: W. C. Stil and A. Tempczyk and R. C. Hawley and T. Hendrickson, Semianalytical treatment of solvation for molecular mechanics and dynamics, J. Am. Chem. Soc. (112) 6127 - 6129, 1990

145: V. Zoete and M. Meuwly and M. Karplus, Study of the insulin dimerization from binding free energy calculations and per-residue free energy decomposition, Proteins (61) 79-93, 2005

146: A. T. Brunger and M. Karplus, Polar hydrogen positions in proteins: empirical energy placement and neutron diffraction comparison., Proteins (4) 148-156, 1988

147: T. H. Halgren, Merck Molecular Force Field. V. Extension of MMFF94 Using Experimental Data, Additonal Computational Data, and Empirical Rules, J. Comput. Chem (17) 616-641, 1996

148: T. H. Halgren, Merck Molecular Force Field. IV. Conformational Energies and Geometries for MMFF94, J. Comput. Chem (17) 587-615, 1996

149: T. H. Halgren, Merck Molecular Force Field. III. Molecular Geometries and Vibrational Frequencies for MMFF94, J. Comput. Chem (17) 553-586, 1996

150: T. H. Halgren, Merck Molecular Force Field. II. MMFF94 van der Waals and Electrostatic Parameters for Intermolecular Interactions, J. Comput. Chem (17) 520-552, 1996

151: T. H. Halgren, Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94, J. Comput. Chem (17) 490-519, 1996

- 152: G. R. Stockwell and J. M. Thornton, Conformational diversity of ligands bound to proteins., *J. Mol. Biol.* (356) 928--944, 2006
- 153: E. F. Pettersen and T. D. Goddard and C. C. Huang and G. S. Couch and D. M. Greenblatt and E. C. Meng and T. E. Ferrin, UCSF Chimera--a visualization system for exploratory research and analysis., *J. Comput. Chem.* (25) 1605--1612, 2004
- 154: T.L. Poulos and A.J. Howard, Crystal structures of metyrapone- and phenylimidazole-inhibited complexes of cytochrome P-450cam., *Biochemistry* (26) 8165--8174, 1987
- 155: E. Perola and W. P. Walters and P. S. Charifson, A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance., *Proteins* (56) 235-49, 2004
- 156: Patricia A Burke and Sally J DeNardo and Laird A Miers and Kathleen R Lamborn and Siegfried Matzku and Gerald L DeNardo, Cilengitide targeting of  $\alpha(v)\beta(3)$  integrin receptor synergizes with radioimmunotherapy to increase efficacy and apoptosis in breast cancer xenografts., *Cancer Res* (62) 4263--4272, 2002
- 157: S. Hehlhans and M. Haase and N. Cordes, Signalling via integrins: implications for cell survival and anticancer strategies., *Biochim Biophys Acta* (1775) 163--180, 2007
- 158: C. Li and A. Grosdidier and G. Crambert and J.-D. Horisberger and O. Michielin and K. Geering, Structural and functional interaction sites between Na,K-ATPase and FXYD proteins., *J. Biol. Chem.* (279) 38895--38902, 2004
- 159: A. G. Therien and R. Blostein, Mechanisms of sodium pump regulation., *Am J Physiol Cell Physiol* (279) C541--C566, 2000
- 160: G. Crambert and K. Geering, FXYD proteins: new tissue-specific regulators of the ubiquitous Na,K-ATPase., *Sci STKE* (2003) RE1, 2003
- 161: K. J. Sweadner and E. Rael, The FXYD gene family of small ion transport regulators or channels: cDNA sequence, protein signature sequence, and expression., *Genomics* (68) 41--56, 2000
- 162: A. G. Therien and R. Goldshleger and S. J. Karlish and R. Blostein, Tissue-specific

distribution and modulatory role of the gamma subunit of the Na,K-ATPase., J Biol Chem (272) 32628--32634, 1997

163: P. Béguin and X. Wang and D. Firsov and A. Puoti and D. Claeys and J. D. Horisberger and K. Geering, The gamma subunit is a specific component of the Na,K-ATPase and modulates its transport function., EMBO J (16) 4250--4260, 1997

164: E. Arystarkhova and R. K. Wetzel and N. K. Asinovski and K. J. Sweadner, The gamma subunit modulates Na(+) and K(+) affinity of the renal Na,K-ATPase., J Biol Chem (274) 33183--33185, 1999

165: P. Béguin and G. Crambert and S. Guennoun and H. Garty and J. D. Horisberger and K. Geering, CHIF, a member of the FXYD protein family, is a regulator of Na,K-ATPase distinct from the gamma-subunit., EMBO J (20) 3993--4002, 2001

166: H. X. Pu and F. Cluzeaud and R. Goldshleger and S. J. Karlish and N. Farman and R. Blostein, Functional role and immunocytochemical localization of the gamma a and gamma b forms of the Na,K-ATPase gamma subunit., J Biol Chem (276) 20370--20378, 2001

167: G. Crambert and M. Fuzesi and H. Garty and S. Karlish and K. Geering, Phospholemman (FXYD1) associates with Na,K-ATPase and regulates its transport properties., Proc Natl Acad Sci U S A (99) 11476--11481, 2002

168: Y. A. Mahmmoud and G. Cramb and A. B. Maunsbach and C. P. Cutler and L. Meischke and F. Cornelius, Regulation of Na,K-ATPase by PLMS, the phospholemman-like protein from shark: molecular cloning, sequence, expression, cellular distribution, and functional effects of PLMS., J Biol Chem (278) 37427--37438, 2003

169: H. Garty and M. Lindzen and R. Scanzano and R. Aizman and M. Fuezesi and R. Goldshleger and N. Farman and R. Blostein and S. J. D. Karlish, A functional interaction between CHIF and Na-K-ATPase: implication for regulation by FXYD proteins., Am J Physiol Renal Physiol (283) F607--F615, 2002

170: P. Béguin and G. Crambert and F. Monnet-Tschudi and M. Uldry and J.-D. Horisberger and H. Garty and K. Geering, FXYD7 is a brain-specific regulator of Na,K-ATPase alpha 1-

beta isozymes., EMBO J (21) 3264--3273, 2002

171: C. Donnet and E. Arystarkhova and K. J. Sweadner, Thermal denaturation of the Na,K-ATPase provides evidence for alpha-alpha oligomeric interaction and gamma subunit association with the C-terminal domain., J Biol Chem (276) 7357--7365, 2001

172: H. Hebert and P. Purhonen and H. Vorum and K. Thomsen and A. B. Maunsbach, Three-dimensional structure of renal Na,K-ATPase from cryo-electron microscopy of two-dimensional crystals., J Mol Biol (314) 479--494, 2001

173: C. Toyoshima and M. Nakasako and H. Nomura and H. Ogawa, Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 Å resolution., Nature (405) 647--655, 2000

174: J.-D. Horisberger and S. Kharoubi-Hess and S. Guennoun and O. Michielin, The fourth transmembrane segment of the Na,K-ATPase alpha subunit: a systematic mutagenesis study., J Biol Chem (279) 29542--29550, 2004

175: W. Im and M. Feig and C. L. Brooks, An implicit membrane generalized born theory for the study of structure, stability, and interactions of membrane proteins., Biophys J (85) 2900--2918, 2003

176W. Im and M. S. Lee and C. L. Brooks, Generalized born model with a simple smoothing function, 2003

177: G. Crambert and C. Li and L. K. Swee and K. Geering, FXD7, mapping of functional sites involved in endoplasmic reticulum export, association with and regulation of Na,K-ATPase., J Biol Chem (279) 30888--30895, 2004

178: A. G. Therien and C. M. Deber, Interhelical packing in detergent micelles. Folding of a cystic fibrosis transmembrane conductance regulator construct., J Biol Chem (277) 6067--6072, 2002

179: H. X. Pu and R. Scanzano and R. Blostein, Distinct regulatory effects of the Na,K-ATPase gamma subunit., J Biol Chem (277) 20270--20276, 2002

180: . Michalik and V. Zoete and G. Krey and A. Grosdidier and L. Gelman and P.

Chodanowski and J. N. Feige and B. Desvergne and W. Wahli and O. Michielin, Combined simulation and mutagenesis analyses reveal the involvement of key residues for PPARalpha helix 12 dynamic behavior., *J Biol Chem* () , 2007

181: B. Desvergne and W. Wahli, Peroxisome proliferator-activated receptors: nuclear control of metabolism., *Endocr Rev* (20) 649--688, 1999

182: M. J. Leaver and E. Boukouvala and E. Antonopoulou and A. Diez and L. Favre-Krey and M. T. Ezaz and J. M. Bautista and D. R. Tocher and G. Krey, Three peroxisome proliferator-activated receptor isotypes from each of two species of marine fish., *Endocrinology* (146) 3150--3162, 2005

183: J. N. Feige and L. Gelman and L. Michalik and B. Desvergne and W. Wahli, From molecular action to physiological outputs: peroxisome proliferator-activated receptors are nuclear receptors at the crossroads of key cellular functions., *Prog Lipid Res* (45) 120--159, 2006

184: L. Michalik and B. Desvergne and W. Wahli, Peroxisome-proliferator-activated receptors and cancers: complex stories., *Nat Rev Cancer* (4) 61--70, 2004

185: B. Desvergne and L. Michalik and W. Wahli, Be fit or be sick: peroxisome proliferator-activated receptors are down the road., *Mol Endocrinol* (18) 1321--1332, 2004

186: L. Gelman and L. Michalik and B. Desvergne and W. Wahli, Kinase signaling cascades that modulate peroxisome proliferator-activated receptors., *Curr Opin Cell Biol* (17) 216--222, 2005

187: C. Diradourian and J. Girard and J.-P. Pégrier, Phosphorylation of PPARs: from molecular characterization to physiological relevance., *Biochimie* (87) 33--38, 2005

188: C. E. Juge-Aubry and E. Hammar and C. Siegrist-Kaiser and A. Pernin and A. Takeshita and W. W. Chin and A. G. Burger and C. A. Meier, Regulation of the transcriptional activity of the peroxisome proliferator-activated receptor alpha by phosphorylation of a ligand-independent trans-activating domain., *J Biol Chem* (274) 10505--10510, 1999

- 189: J. P. Renaud and N. Rochel and M. Ruff and V. Vivat and P. Chambon and H. Gronemeyer and D. Moras, Crystal structure of the RAR-gamma ligand-binding domain bound to all-trans retinoic acid., *Nature* (378) 681--689, 1995
- 190: R. T. Nolte and G. B. Wisely and S. Westin and J. E. Cobb and M. H. Lambert and R. Kurokawa and M. G. Rosenfeld and T. M. Willson and C. K. Glass and M. V. Milburn, Ligand binding and co-activator assembly of the peroxisome proliferator-activated receptor-gamma., *Nature* (395) 137--143, 1998
- 191: J. Uppenberg and C. Svensson and M. Jaki and G. Bertilsson and L. Jendeborg and A. Berkenstam, Crystal structure of the ligand binding domain of the human nuclear receptor PPARgamma., *J Biol Chem* (273) 31108--31112, 1998
- 192: H. Keller and P. R. Devchand and M. Perroud and W. Wahli, PPAR alpha structure-function relationships derived from species-specific differences in responsiveness to hypolipidemic agents., *Biol Chem* (378) 651--655, 1997
- 193: B. A. Johnson and E. M. Wilson and Y. Li and D. E. Moller and R. G. Smith and G. Zhou, Ligand-induced stabilization of PPARgamma monitored by NMR spectroscopy: implications for nuclear receptor activation., *J Mol Biol* (298) 187--194, 2000
- 194: L. Nagy and J. W. R. Schwabe, Mechanism of the nuclear receptor molecular switch., *Trends Biochem Sci* (29) 317--324, 2004
- 195: H. E. Xu and M. H. Lambert and V. G. Montana and K. D. Plunket and L. B. Moore and J. L. Collins and J. A. Oplinger and S. A. Kliewer and R. T. Gampe and D. D. McKee and J. T. Moore and T. M. Willson, Structural determinants of ligand binding selectivity between the peroxisome proliferator-activated receptors., *Proc Natl Acad Sci U S A* (98) 13919--13924, 2001
- 196: H. E. Xu and T. B. Stanley and V. G. Montana and M. H. Lambert and B. G. Shearer and J. E. Cobb and D. D. McKee and C. M. Galardi and K. D. Plunket and R. T. Nolte and D. J. Parks and J. T. Moore and S. A. Kliewer and T. M. Willson and J. B. Stimmel, Structural basis for antagonist-mediated recruitment of nuclear co-repressors by PPARalpha., *Nature* (415) 813--817, 2002

- 197: V. Zoete and A. Grosdidier and O. Michielin, Peroxisome proliferator-activated receptor structures: Ligand specificity, molecular switch and interactions with regulators., *Biochim Biophys Acta* ( ) , 2007
- 198: A. Fiser and R. K. Do and A. Sali, Modeling of loops in protein structures., *Protein Sci* (9) 1753--1773, 2000
- 199: W. Humphrey and A. Dalke and K. Schulten, VMD -- Visual Molecular Dynamics, *Journal of Molecular Graphics* (14) 33-38, 1996
- 200: P. Cronet and J. F. Petersen and R. Folmer and N. Blomberg and K. Sjoebloom and U. Karlsson and E. L. Lindstedt and K. Bamberg, Structure of the PPARalpha and -gamma ligand binding domain in complex with AZ 242; ligand selectivity and agonist activation in the PPAR family., *Structure* (9) 699--706, 2001
- 201X. Y. Xu and F.Cheng and J. Hua and S. Xiao and M. Luo and L. Li and C. Li and D. Yue and Y. Du and F. Ye and S. Hao and J. Da and Y. Zhu and H. Liang and J. Kai and X. Chen, Agonist-PPAR? interactions: Molecular modeling study with docking approach, 2003
- 202: C. Yu and L. Chen and H. Luo and J. Chen and F. Cheng and C. Gui and R. Zhang and J. Shen and K. Chen and H. Jiang and X. Shen, Binding analyses between Human PPARgamma-LBD and ligands., *Eur J Biochem* (271) 386--397, 2004
- 203: V. Zoete and M. Meuwly, Importance of individual side chains for the stability of a protein fold: computational alanine scanning of the insulin monomer., *J Comput Chem* (27) 1843--1857, 2006
- 204: S. Faderl and M. Talpaz and Z. Estrov and S. O'Brien and R. Kurzrock and H. M. Kantarjian, The biology of chronic myeloid leukemia., *N Engl J Med* (341) 164--172, 1999
- 205: C. L. Sawyers, Chronic myeloid leukemia., *N Engl J Med* (340) 1330--1340, 1999
- 206: M. W. Deininger and J. M. Goldman and J. V. Melo, The molecular biology of chronic myeloid leukemia., *Blood* (96) 3343--3356, 2000
- 207: J. M. Goldman, Tyrosine-kinase inhibition in treatment of chronic myeloid leukaemia., *Lancet* (355) 1031--1032, 2000

- 208: D. T. Holyoake, Recent advances in the molecular and cellular biology of chronic myeloid leukaemia: lessons to be learned from the laboratory., *Br J Haematol* (113) 11--23, 2001
- 209: B. J. Druker and C. L. Sawyers and H. Kantarjian and D. J. Resta and S. F. Reese and J. M. Ford and R. Capdeville and M. Talpaz, Activity of a specific inhibitor of the BCR-ABL tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the Philadelphia chromosome., *N Engl J Med* (344) 1038--1042, 2001
- 210: B. J. Druker and M. Talpaz and D. J. Resta and B. Peng and E. Buchdunger and J. M. Ford and N. B. Lydon and H. Kantarjian and R. Capdeville and S. Ohno-Jones and C. L. Sawyers, Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia., *N Engl J Med* (344) 1031--1037, 2001
- 211: M. Michael and M. M. Doherty, Tumoral drug metabolism: overview and its implications for cancer therapy., *J Clin Oncol* (23) 205--229, 2005
- 212: B. Rochat, Role of cytochrome P450 activity in the fate of anticancer agents and in drug resistance: focus on tamoxifen, paclitaxel and imatinib metabolism., *Clin Pharmacokinet* (44) 349--366, 2005
- 213: M. Marull and B. Rochat, Fragmentation study of imatinib and characterization of new imatinib metabolites by liquid chromatography-triple-quadrupole and linear ion trap mass spectrometers., *J Mass Spectrom* (41) 390--404, 2006
- 214: B. Nagar and W. G. Bornmann and P. Pellicena and T. Schindler and D. R. Veach and W. T. Miller and B. Clarkson and J. Kuriyan, Crystal structures of the kinase domain of c-Abl in complex with the small molecule inhibitors PD173955 and imatinib (STI-571)., *Cancer Res* (62) 4236--4243, 2002
- 215: R. A. Laskowski, SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions., *J Mol Graph* (13) 323--30, 307-8, 1995
- 216: R. O. Hynes, Integrins: bidirectional, allosteric signaling machines., *Cell* (110) 673--687, 2002



- 217: M. A. Schwartz and M. H. Ginsberg, Networks and crosstalk: integrin signalling spreads., *Nat Cell Biol* (4) E65--E68, 2002
- 218: F. M. Watt, Role of integrins in regulating epidermal adhesion, growth and differentiation., *EMBO J* (21) 3919--3926, 2002
- 219: C. Brakebusch and R. Faessler, The integrin-actin connection, an eternal love affair., *EMBO J* (22) 2324--2333, 2003
- 220: M. A. Schwartz and R. K. Assoian, Integrins and cell proliferation: regulation of cyclin-dependent kinases via cytoplasmic signaling pathways., *J Cell Sci* (114) 2553--2560, 2001
- 221: M. A. Schwartz, Integrin signaling revisited., *Trends Cell Biol* (11) 466--470, 2001
- 222: N. Zahir and V. M. Weaver, Death in the third dimension: apoptosis regulation and tissue architecture., *Curr Opin Genet Dev* (14) 71--80, 2004
- 223R. Faessler, *Molekulare Medizin*, 2004
- 224: A. J. D'Ardenne and P. I. Richman and M. A. Horton and A. E. Mcaulay and S. Jordan, Co-ordinate expression of the alpha-6 integrin laminin receptor sub-unit and laminin in breast cancer., *J Pathol* (165) 213--220, 1991
- 225: B. Felding-Habermann, Integrin adhesion receptors in tumor metastasis., *Clin Exp Metastasis* (20) 203--213, 2003
- 226: L. V. Parise and J. Lee and R. L. Juliano, New aspects of integrin signaling in cancer., *Semin Cancer Biol* (10) 407--414, 2000
- 227: O. W. Petersen and L. Rønnov-Jessen and A. R. Howlett and M. J. Bissell, Interaction with basement membrane serves to rapidly distinguish growth and differentiation pattern of normal and malignant human breast epithelial cells., *Proc Natl Acad Sci U S A* (89) 9064--9068, 1992
- 228: K. Wilhelmsen and S. H. M. Litjens and A. Sonnenberg, Multiple functions of the integrin alpha6beta4 in epidermal homeostasis and tumorigenesis., *Mol Cell Biol* (26)

2877--2886, 2006

229: S. Yamamoto and O. Hayaishi, Tryptophan pyrrolase of rabbit intestine. D- and L-tryptophan-cleaving enzyme or enzymes., J Biol Chem (242) 5260--5266, 1967

230: P. Hwu and M. X. Du and R. Lapointe and M. Do and M. W. Taylor and H. A. Young, Indoleamine 2,3-dioxygenase production by human dendritic cells results in the inhibition of T cell proliferation., J Immunol (164) 3596--3599, 2000

231: C. Uyttenhove and L. Pilotte and I. Theate and V. Stroobant and D. Colau and N. Parmentier and T. Boon and B. J. vandenEynde, Evidence for a tumoral immune resistance mechanism based on tryptophan degradation by indoleamine 2,3-dioxygenase., Nat Med (9) 1269--1274, 2003

232: N. Eguchi and Y. Watanabe and K. Kawanishi and Y. Hashimoto and O. Hayaishi, Inhibition of indoleamine 2,3-dioxygenase and tryptophan 2,3-dioxygenase by beta-carboline and indole derivatives., Arch Biochem Biophys (232) 602--609, 1984

233: H. Sugimoto and S.-I. Oda and T. Otsuki and T. Hino and T. Yoshida and Y. Shiro, Crystal structure of human indoleamine 2,3-dioxygenase: catalytic mechanism of O<sub>2</sub> incorporation by a heme-containing dioxygenase., Proc Natl Acad Sci U S A (103) 2611--2616, 2006

234: J. A. Vrugt and B. A. Robinson, Improved evolutionary optimization from genetically adaptive multimethod search., Proc Natl Acad Sci U S A () , 2007

235: A. Cherkasov and F. Ban and Y. Li and M. Fallahi and G. L. Hammond, Progressive docking: a hybrid QSAR/docking approach for accelerating in silico high throughput screening., J Med Chem (49) 7466--7478, 2006

236: D. Huang and A. Caflisch, Efficient evaluation of binding free energy using continuum electrostatics solvation., J Med Chem (47) 5791--5797, 2004

237: A. Strizhev and E.J. Abrahamian and S. Choi and J.M. Leonard and P.R.N. Wolohan and R.D. Clark, The Effects of Biasing Torsional Mutations in a Conformational GA, Journal of Chemical Information and Modeling () , 2006

- 238: M. Hendlich and F. Rippmann and G. Barnickel, LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins., J Mol Graph Model (15) 359--63, 389, 1997
- 239: A. A. Canutescu and A. A. Shelenkov and R. L. Dunbrack, A graph-theory algorithm for rapid protein side-chain prediction., Protein Sci (12) 2001--2014, 2003