

A MODEL TO PREDICT PROTEIN-LIGAND COMPLEXES WITH KNOWN  
STRUCTURE

January 17, 2011

GLORIA ANGÉLICA SANDOVAL PÉREZ  
ID: 598211

Thesis presented as a partial requirement for the degree of  
MASTER OF SCIENCES  
MASTER IN BIOMEDICAL ENGINEERING

Supervisor:  
LUIS FERNANDO NIÑO V. Ph.D.  
Associated Professor

NATIONAL UNIVERSITY OF COLOMBIA  
SCHOOL OF MEDICINE  
BOGOTÁ D.C.  
2010

Approved by the School of Medicine in fulfillment of the requirements to grant the title of Master in Biomedical Engineering.

---

LUIS FERNANDO NIÑO V Ph.D.  
Supervisor of the Thesis

---

FABIO A. GONZALEZ O. Ph.D.  
Member of the Jury

---

EDGAR EDUARDO ROMERO C. Ph.D.  
Member of the Jury

National University of Colombia  
Bogotá D.C., November 2010

*To my parents  
Gloria Cristina and Carmen Julio*

# ACKNOWLEDGEMENT

I used to believe that anything could be accomplished by just one person –no matter how big it was–, that one person alone could finish something up to the very last detail if enough effort and dedication is put to the task. In that way, at the end of the story, that "something" would belong to this one person and to no one else. However, things that are created by one person tend to be unreal or incomplete and so does the person who creates them.

Nowadays I realize that real things are made up of small pieces, each of which is important and unique. If one piece is missing or fails to work properly, then reality would feel, smell, and hear unpleasant.

So, I would like to thank to each of the pieces that made this work real and prevent it from decaying. To my parents for their constant support; to my LISI-mates, especially to David Becerra for sharing his ideas with me, for his patience and enlightening discussions. To my thesis director and to all the people who were involved in some part of this work and are not mentioned here: without your help this work would not have been possible.

# RESUMEN

Este trabajo se enfoca en el problema de acoplamiento molecular (docking), el cual es de gran interés para las áreas de quimioinformática y diseño racional de fármacos. Básicamente, el problema de docking consiste en predecir e identificar los complejos más estables formados entre una macromolécula como DNA, RNA o una proteína (denominada receptor) y una molécula orgánica pequeña (denominada ligando). Aunque, desde inicios de los 80 hasta la actualidad se han dedicado grandes esfuerzos para resolver el problema de docking, este sigue siendo un problema desafiante y varios aspectos aún no han sido satisfactoriamente resueltos.

Esta tesis propone un modelo de optimización multi-objetivo de predicción de complejos proteína–ligando, basado en las contribuciones energéticas debidas a la interacción molecular. El modelo se evaluó con un set de complejos de referencia, que han sido ampliamente utilizados en la validación y evaluación de otros modelos de docking; estos complejos se encuentran en la base de datos del protein data bank (PDB), identificados como: 1ABE, 1CDG, 1ACM y 1BAF. Los resultados obtenidos muestran que el modelo es flexible en cuanto al uso de algoritmos de optimización y funciones de energía; adicionalmente, predice las ubicaciones y conformaciones del ligando en el sitio de unión del receptor de forma tal que los complejos obtenidos son energéticamente estables.

# ABSTRACT

This work focuses on the docking problem, which is one of the most important problems in chemoinformatics and drug design. Basically, it deals with the prediction and identification of the most stable complexes formed between a macromolecule such as DNA, RNA or a protein (namely the receptor) and a small organic molecule (called the ligand). Although a great deal of work has been carried out since the late eighties to solve the docking problem, it is still very challenging and several issues have not yet been solved.

This thesis proposes a multi-objective model for predicting protein–ligand complexes based on the energetic contributions that occur in molecular interactions. The performance of the model was evaluated over a set of benchmark complexes that have been widely used in the validation and evaluation of other docking models; these complexes are reported in the protein data bank (PDB) as: 1ABE, 1CDG, 1ACM, and 1BAF. The results show that the proposed model is flexible to the use of different search algorithms and energy functions; it predicts ligand localizations on the receptor binding site and the ligand conformations that form adequate complexes with the receptor.

# Contents

<b>ACKNOWLEDGEMENT</b>	<b>ii</b>
<b>RESUMEN</b>	<b>iii</b>
<b>ABSTRACT</b>	<b>iv</b>
<b>1 INTRODUCTION</b>	<b>1</b>
<b>2 BACKGROUND: THE DRUG DESIGN CYCLE</b>	<b>5</b>
2.1 <i>IN SILICO</i> METHODS . . . . .	9
2.1.1 DRUG DESIGN WITHOUT KNOWN RECEPTOR . . . . .	11
2.1.2 DRUG DESIGN WITH KNOWN RECEPTOR STRUCTURE . . . . .	11
<b>3 BACKGROUND: DOCKING</b>	<b>12</b>
3.1 DEFINITION OF THE BINDING SITE . . . . .	13
3.2 MOLECULAR REPRESENTATION . . . . .	16
3.3 EXPLORATION OF THE SEARCH SPACE . . . . .	18
3.3.1 SYSTEMATIC APPROACHES . . . . .	18
3.3.2 DETERMINISTIC APPROACHES . . . . .	19
3.3.3 STOCHASTIC METHODS . . . . .	20
3.4 SCORING FUNCTIONS . . . . .	22

<i>CONTENTS</i>	vi
3.4.1 FORCE FIELDS . . . . .	23
3.4.2 EMPIRICAL . . . . .	24
3.4.3 KNOWLEDGE-BASED . . . . .	24
<b>4 AN EVOLUTIONARY APPROACH TO THE DOCKING PROBLEM</b>	<b>26</b>
4.1 DEFINITION OF THE BINDING SITE . . . . .	26
4.2 SELECTION OF THE MOLECULAR REPRESENTATION METHOD	30
4.2.1 MOLECULAR REPRESENTATIONS USED BY THE PROPOSED METHOD . . . . .	32
4.2.2 ALTERNATING BETWEEN MOLECULAR REPRESENTATIONS . . . . .	34
4.2.2.1 FROM THE ALGEBRAIC TO THE TRIGONOMETRIC REPRESENTATION . . . . .	35
4.2.2.2 FROM THE TRIGONOMETRIC TO THE ALGEBRAIC REPRESENTATION . . . . .	36
4.3 IMPLEMENTATION OF THE SCORING FUNCTION . . . . .	38
4.4 EXPLORATION OF THE SEARCH SPACE . . . . .	41
4.4.1 MULTI-OBJECTIVE OPTIMIZATION . . . . .	42
4.4.1.1 THE MULTI-OBJECTIVE EVOLUTIONARY ALGORITHM . . . . .	45
4.4.1.2 THE MULTI-OBJECTIVE FORMULATION . . . . .	46
4.4.1.3 DECISION-MAKING PHASE . . . . .	48
4.5 EVALUATION OF THE OBTAINED RESULTS . . . . .	48
4.6 FRAMEWORK OF EXPERIMENTATION . . . . .	50
4.6.1 MOLECULAR COMPLEXES . . . . .	50
4.7 ASPECTS TO EVALUATE . . . . .	52

<i>CONTENTS</i>	vii
<b>5 RESULTS AND DISCUSSION</b>	<b>54</b>
5.1 1ABE . . . . .	54
5.2 1ACM . . . . .	55
5.3 1BAF . . . . .	62
5.4 1CDG . . . . .	63
<b>6 CONCLUSIONS AND PERSPECTIVES</b>	<b>70</b>
<b>7 APPENDIX</b>	<b>73</b>
<b>Bibliography</b>	<b>75</b>

# List of Figures

2.1	Drug Design Cycle . . . . .	6
2.2	Structure of Enalapril . . . . .	7
2.3	Enalapril. Simplified structure . . . . .	7
2.4	<i>In silico</i> methods . . . . .	10
3.1	Docking . . . . .	14
3.2	Cavity identification methods . . . . .	15
3.3	Contour defined by physicochemical interactions . . . . .	16
3.4	General diagram representation of a genetic algorithm cycle	21
4.1	Receptor Surface. . . . .	27
4.2	Surface of the Human Immunodeficiency Virus-1 Protease (PDB ID: 1aaq). . . . .	28
4.3	2D Example of sphere generation over a surface . . . . .	29
4.4	Cavity selected on the Human immunodeficiency virus-1 Pro- tease . . . . .	30
4.5	Data representation of molecular structures . . . . .	32
4.6	Comparison between the trigonometric and algebraic repre- sentations of a molecule . . . . .	34
4.7	Atom types, bonds, rotatable and non-rotatable bonds. . . . .	35

4.8	Dihedral angles in a molecule . . . . .	37
4.9	Planes, vectors and angles used for the calculation of di- dral angles. . . . .	37
4.10	Atoms affected by changes on dihedral angles of the rotatable bonds . . . . .	38
4.11	The $n$ -dimensional parameter space maps to the $m$ -dimensional objective space. . . . .	43
4.12	An ideal multi-objective optimization procedure. . . . .	44
4.13	NSGA II procedure. . . . .	46
4.14	Chromosome representation used for the genetic algorithm. . . . .	47
4.15	Points to define angles $\alpha$ , $\beta$ , $\gamma$ or $\delta$ . . . . .	49
5.1	Complex 1ABE. Changes in the Pareto Front as the number of evaluations increases. . . . .	55
5.2	Complex 1ABE. RMDS and energy values according to the number of evaluations. . . . .	56
5.3	Complex 1ABE. Changes in RMSD <sub>1</sub> and RMSD <sub>3</sub> values as the number of evaluations increases. . . . .	57
5.4	Complex 1ABE. Dispersion in RMSD values according to the number of evaluations. . . . .	58
5.5	Solutions established by the proposed method for the 1ABE complex. . . . .	58
5.6	Complex 1ACM. Changes in the Pareto Front as the number of evaluations increases. . . . .	59
5.7	Complex 1ACM. RMDS and Energy values according to the number of evaluations. . . . .	60
5.8	Solutions established by the proposed method for the 1ACM complex. . . . .	61

5.9	Complex 1ACM. Changes in RMSD1 and RMSD3 values as the number of evaluations increases. . . . .	61
5.10	1ACM Complex. Dispersion in RMSD values according to the increase in the number of evaluations. . . . .	62
5.11	Solution established by the proposed method for the 1BAF complex. . . . .	63
5.12	Complex 1BAF. Changes in the Pareto Front as the number of evaluations increases. . . . .	64
5.13	Complex BAF. RMDS and Energy values according to the number of evaluations. . . . .	65
5.14	Complex BAF. Changes in RMSD1 and RMSD3 values as the number of evaluations increases. . . . .	66
5.15	Complex BAF. Dispersion in RMSD values according to the increase in the number of evaluations. . . . .	66
5.16	Complex CDG. Changes in the Pareto Front as the number of evaluations increases. . . . .	67
5.17	Complex CDG. RMDS and Energy values according to the number of evaluations. . . . .	67
5.18	Solutions established by the proposed method for the 1CDG complex. . . . .	68
5.19	Complex CDG. Changes in $\text{RMSD}_1$ and $\text{RMSD}_3$ values as the number of evaluations increases . . . . .	68
5.20	Complex CDG. Dispersion in RMSD values according to the increase in the number of evaluations. . . . .	69

# List of Tables

4.1	Data set of complexes . . . . .	51
5.1	Complex 1ABE. Decision Maker selection at different number of evaluations of the method. . . . .	59
5.2	1CDG Complex. Decision Maker selection at different number of evaluations of the method. . . . .	69

# Chapter 1

## INTRODUCTION

Molecular interactions that take place within living organisms are complex and involve many different molecules. A simplified way of seeing them is to consider a receptor–ligand system, where the receptor is a macromolecule involved in a specific biochemical pathway of a target organism, and the ligand is a molecule that binds to such receptor, affecting its functioning. Different types of macromolecules such as DNA, RNA and proteins, are considered receptors. Among them, proteins account for the largest percentage of receptors that are of pharmacological interest mainly due to their abundance and importance in biochemical pathways. In turn, many different types of molecules can act as ligands, including proteins, peptides (molecules made up by a few amino acids) and small organic molecules of exogenous or endogenous origin. Both the receptor and the ligand are embedded in a particular environment, which consists of water molecules, ions, ligands and several other different molecules that can affect the interaction between both molecules. In summary, molecular interactions can be represented as dynamic systems where all molecules are continuously moving.

Predicting the structures that may result from the interaction of a receptor and a ligand (known as the docking problem) is of special interest in drug design given that the ligand can affect the behavior of the receptor. In other words, if the malfunctioning of a receptor that is involved in a specific pathway and is associated to a particular disease, is “restored” by binding a particular ligand, then such ligand can be considered a potential drug.

Thanks to the development of techniques to elucidate the structural conformation of a molecule inside living organisms and to the wide accessibility to these data, a rational drug development process has become possible by driving drug research toward the identification of the structures that need to be attacked in a specific illness; this prevents wasting time and resources in molecules of improbable pharmacological activity. In consequence, the inclusion of structural information has transformed drug development from a trial-and-error experimentation process into a rational drug design process to discover new drugs [8, 49].

Several computational tools have been developed in an attempt to solve the docking problem and hence reduce the number of ligands to be evaluated *in vivo* by focusing on those with the ability to bind to the receptor involved in the disease of interest. Nevertheless, modeling biological systems is not a trivial task. It demands making different simplifications and assumptions either to simulate the binding process or to select the molecules that will bind to the receptor. Among such simplifications, those that have had important implications on the results of the predictions are the ones made over the dynamics of the system and the interaction energies that are considered in the simulation of the binding process [86, 13].

One of the strongest simplifications regarding the system dynamics consists in considering structures as rigid bodies. Under this simplification, the ability of the model to predict the protein–ligand structure is drastically reduced, as mentioned by Trovov *et al.* in their 2008 study where they report that “docking algorithms predict an incorrect binding pose for about 50 to 70% of all ligands when only a single fixed receptor conformation is considered” [89]. When the model considers both the receptor and the ligand as flexible structures, the effectiveness of the prediction increases to 71% [46, 27].

On the other hand, the energetic implications of the interaction are simplified when the docking model uses scoring functions to evaluate energy changes in the binding interface. One type of scoring functions are force field (FF) scoring functions. These functions must assign atom types to evaluate the interactions that occur between the two molecules. Atom types are assigned based on the nature of the atom itself, the number of covalent bonds and its neighbor atoms [91]; however, since it is impossible to identify the amount of atom types in all organic molecules, certain

degree of generalization is needed in order to assign the atom types; however, such overgeneralization introduces an error to the obtained energy value. There are other examples of this kind of simplifications or assumptions; but they are not described here as they are beyond the scope of this work. As a result of energetic simplifications, there is low correlation between experimentally obtained energy values and computed energy values. The scoring functions have been subjected to continuous optimization and novel approaches have been proposed to improve the results, but still predictions are not good enough [68, 93].

In general terms, although the docking problem has been widely explored and studied over the years, it still is an open problem. This work aims to propose a computational model to face both of the issues mentioned above, namely, the system dynamics and energy measures.

The system dynamics is determined by the flexibility of the ligand during the docking process. The flexibility of the molecule is defined by the rotatable bonds in the ligand such that changes occurring in those bonds produce all possible molecular conformations for each ligand.

Although docking models usually evaluate the energy by using a force field scoring function as a single objective to optimize the interaction between the molecules, the model proposed in this work is based on the idea that “non-covalent bonds are critical for maintaining the three-dimensional (3D) structure of large molecules such as proteins and nucleic acids” [57] and that non-covalent bonds can stabilize unusual conformations in small ligands when they are bound to a receptor; causing a detriment in the internal energy. Therefore, as a natural consequence, a multi-objective optimization approach was chosen to predict protein–ligand structures (here called complexes), where the objective functions to be optimized were the non-bond and bond energy terms.

This document is organized in the following chapters:

Chapter 2 explains the cycle of drug design and gives an overview of *in silico* methods used in drug design (or computational drug design tools).

Chapter 3 provides a general explanation of the methods that have been proposed to solve the docking problem.

Chapter 5 describes each of the five steps that were followed to develop the model: the definition of the binding site, selection of the molecular representation methods, application of the scoring function, exploration of the search space and evaluation of the results.

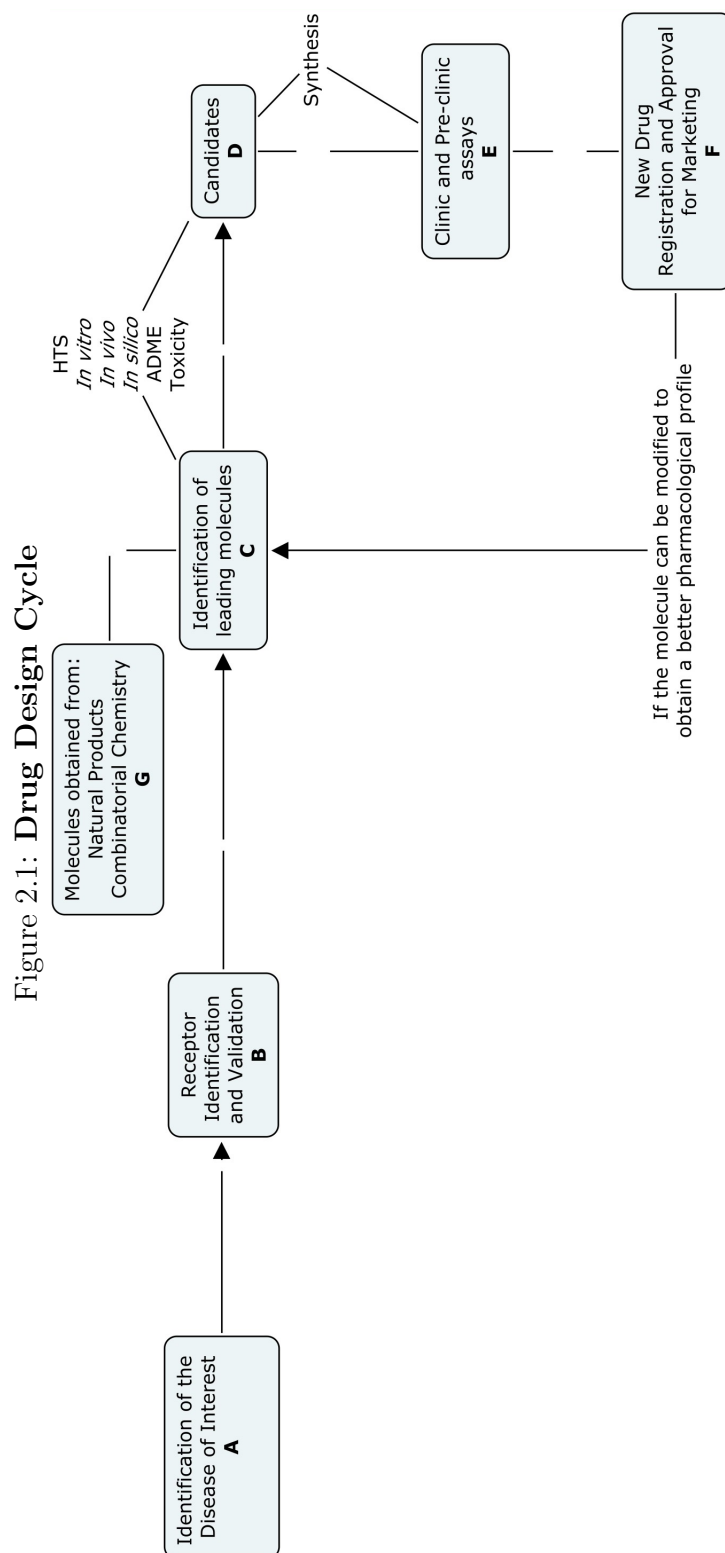
Chapter 5 contains the experimental framework, which includes the complexes, experiments and measures used to evaluate the proposed method. Additionally, it presents the results obtained with the proposed method. Finally, the document presents the conclusions and perspectives for future works in Chapter 6.

# Chapter 2

## BACKGROUND: THE DRUG DESIGN CYCLE

The development of new molecules for the treatment of known diseases is of special interest for life sciences. In general, the task of discovering a new drug molecule is considered as a cycle where several experiments at different levels (*in vitro*, *in vivo*, *in silico*, etc.) are carried out to select pharmacologically active molecules that have reduced side effects or can be administered in smaller dosages. The cycle begins with the identification of an illness and its pathophysiological study (steps A and B in Figure 2.1); it includes the study of the molecular processes underlying the development of the disease and, ideally, the characterization of the most important receptors involved in such illness (stage C in Figure 2.1).

This work will focus on developing tools for evaluating candidate drug molecules when the receptor is known. The receptor will be a structure inside or outside the cell, whose interaction with exogenous or endogenous molecules (ligands) can alter biochemical cascades associated with the disease of interest. Receptors can include DNA, RNA or proteins; but the drug design cycle focuses largely on proteins because they are the most common type of receptors found in living organisms. Besides, some proteic receptors are specific for many biochemical pathways, hence, if the function of the proteic receptor is altered by binding of an exogenous ligand, it is possible to find molecules with specific pharmacological activity on such receptor (drugs) and therefore reduce some side effects, which is why the study of proteins involved in pathological processes is essential for the development of new drugs.



Once the target receptor involved in the biochemical pathway to be affected is characterized, a new drug begins to be searched within a large amount of molecules (stage G in Figure 2.1). These initial molecules could be of diverse origins, for example, they could be compounds extracted from natural sources, synthetic products or compounds resulting from combinatorial chemistry processes.

To have a more clear idea about the large size of the search space, let's consider an example proposed by Oprea [59] with the molecules explored for the discovery of Enalapril, a prominent pro-drug angiotensin converter enzyme (ACE) inhibitor. Given the parameters as outlined in the patent covering Enalapril (see Figure 2.2), the author estimated the total number of compounds included in the generic claim for Enalaprilat, its active ingredient. The compound in the patent is described in Figure 2.3. “where the substituents R2 and R7 are hydrogens and R6 is a hydroxyl group; R and R3 are described as lower alkoxy groups, for example methyl, ethyl, isopentyl, and similar radicals; R1 is described as a substituted lower alkyl, which corresponds to a phenyl group; R4 and R5 are described as being lower alkyl groups, which may be linked to form a cyclic 4- to 6-membered ring in this position. According to these parameters, the number of compounds included in the patent gets close to  $1.72 \times 10^{13}$  (more than 17 trillion compounds)” (*Taken from Tudor Oprea, 2005, Chemoinformatics in Drug Discovery*) [59].

Figure 2.2: **Structure of Enalapril**

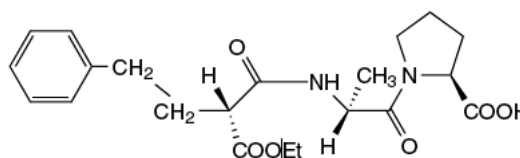
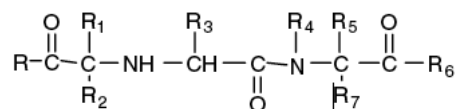


Figure 2.3: **Enalapril**. Simplified structure



In this kind of search spaces, it is necessary to answer some questions in order to reduce the number of compounds, for example: What is the binding affinity between the different ligands and the studied receptor? Which chemical and physical features of the ligand will determine the biological activity? How will the absorption, distribution, metabolism and elimination of the drug be? Is it possible to predict the toxicity of a specific molecule? Finding the answers to these questions is important as they allow to narrow the initial number of lead molecules to about 10 compounds with the desired characteristics, called drug candidates. In Figure 2.1, this process lies between the identification of lead molecules and drug candidates (stage C and D, respectively). Assays to select molecules at the first stages of the cycle are expensive in terms of money and time, therefore drug developers have introduced software tools and simulations of biological systems to reduce the number of *in vivo* and *in vitro* assays and so enhance the amount of resources invested to obtain a new drug. In fact, some new drugs such as  $\beta$ -inhibitors[5] and hypocholesterol drugs have been developed thanks to *in silico* methods and therefore their use and development have been adapted as another useful tool in the cycle of drug design.

Once the drug candidates are selected and synthesized, they enter pre-clinical and clinical trials to select the one that has the best pharmacological profile (stage E in Figure 2.1); this molecule is then registered and approved for use and commercialization. An important aspect in the development and research for new drugs is that the process does not end once a molecule is released to the public; on the contrary, it is still kept under study in order to establish adverse drug effects, undesirable unknown interactions with food and other drugs, safety profile in specific ethnic groups and other possible pharmacological uses. The results obtained at this stage of the cycle most of the times lead to the conclusion that the molecule could be modified by increasing its selectivity for the receptor in order to obtain a better safety profile or decrease the dosage. Also, molecules derived from other commercially available drugs can be used as lead molecules in a new drug design cycle and follow the same process described above.

## 2.1 *IN SILICO* METHODS

As mentioned in the above section, *in silico* drug design methods have been adapted as an important tool in the drug design cycle. Since this work focuses on the development of a new *in silico* drug design method, it is important to consider their advantages and disadvantages compared to other methods:

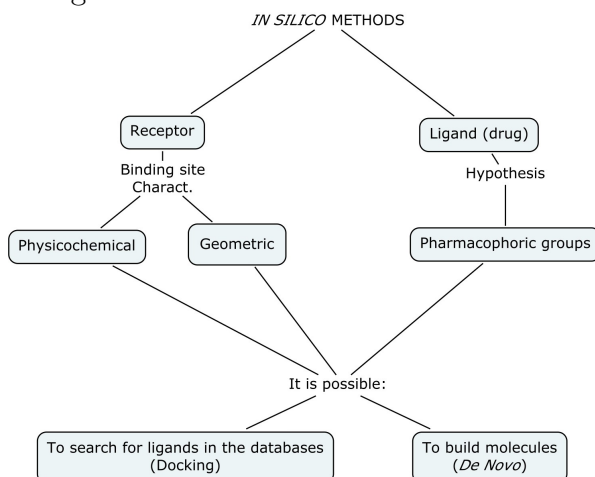
- Saving time in the process of obtaining new drug candidates.
- Decreasing the cost of *in vivo* and *in vitro* experimentation.
- Decreasing the use of animals in experimentation.
- Predicting unknown ligands that could interact with the receptor under study.

On the other hand, some of the disadvantages are:

- Inability to predict the toxic effects of a ligand.
- Impossibility to predict undesirable interactions with concomitantly administered drugs. Some studies have been conducted to establish whether drugs and food are metabolized through the same pathway; however, it is not possible to determine all the possible interactions of a specific molecule [11, 26, 76, 6].
- Although it is possible to find a ligand with high binding affinity for a receptor, this does not guarantee that such ligand would have any type of biological activity because there are still other important issues, such as its absorption, distribution, metabolism and elimination, to be taken into account.

The application of *in silico* methods to drug design can be summarized in two types of models (see Figure 2.4):

- **Models of biological systems when information about the structure of the receptor is not available.** In this type of methods, although information about the receptor is not available, there is a set of molecules known to have the desired pharmacological activity and based on which molecules with similar activity can be hypothesized. A research group at the University of Washington has dedicated some efforts to extract information and develop tools to produce molecules when structural information about the receptor is not available [17]. This approach is called the Pharmacophore-Based Virtual Screening
- **Models of biological systems when information about the structure of the receptor is available.** In the early eighties, two groups headed by Goodford [34] and Kuntz [52], respectively, lead the development of techniques when the 3D structure of the receptor was known. Goodford and colleagues not only developed the use of probe atoms and chemical groups to map the binding site, but also identified the optimal binding sub-sites; this was a prelude for positioning and assembling fragments to build a new molecule [73, 74]. On the other hand, the studies by Kuntz and colleagues have proposed the major innovations for exploring the formation of possible complexes, even including approaches under solvation conditions [82].

Figure 2.4: *In silico* methods

### 2.1.1 DRUG DESIGN WITHOUT KNOWN RECEPTOR

#### PHARMACOPHORE

A pharmacophore is a subset of pharmaceutically relevant molecular descriptors, which are potential key points for the interaction of the ligand with the receptor [15]. The success of studies on pharmacophores have motivated efforts to extend the domain to non-congeneric series where the structural similarity between active molecules in the same bio-assay is not obvious. Certainly, the studies by Beckett and Casey on opioids to define those parts of the active molecules (pharmacophoric groups) that are essential for the pharmacological activity were of seminal importance for this type of models [7]. Kier and Aldrich [50] further developed the concept of pharmacophore and applied it to rationalize the Structure Activity Relation (SAR) of several systems. These methods are undergoing continuous development. In general, they are intended to generate a pharmacophore hypothesis by finding molecular characteristics associated with a biological effect. These characteristics are found by considering several drug molecules with similar known biological activity and overlapping their structures, seeking to detect similarities between their chemical groups, such as for example in hydrogen bond donors and acceptors, aromatic rings, etc.

### 2.1.2 DRUG DESIGN WITH KNOWN RECEPTOR STRUCTURE

If the 3D structure of the receptor involved in the disease of interest is known and the binding site(s) has been identified and characterized, both geometrically and physicochemically, it is possible to tackle drug design from two points of view: by constructing the structure of a new molecule with high binding affinity for the receptor (*de novo* design) or by searching in some databases for molecules with ability to bind to the receptor (docking) (See Figure 2.4); this latter approach is the one explored in this work and will be explained in detail in the next chapter.

# Chapter 3

## BACKGROUND: DOCKING

Molecular docking aims to predict the structure of receptor–ligand complexes, being the receptor usually a protein and the ligand a small organic molecule [13]. Since the development of combinatorial chemistry, molecular docking has been applied to aid in the design of combinatorial compound libraries (large collections of chemical compounds obtained by combinatorial chemistry) and their *in silico* pre-screen to identify potential new drugs[13]. In the earlier studies, docking was successfully implemented in drug design as a tool to search drug compound databases and optimize lead candidates; nowadays, docking programs are used to prioritize molecules for *in vivo* or *in vitro* assays.

The first docking programs allowed automating methods to search for molecular complexes with geometric and chemical complementarities. To simplify the search and reduce the degrees of freedom, protein and ligand structures were considered as rigid bodies, except with respect to translation and rotations. Also, the atoms of both the receptor and the ligand were explicitly expressed and simple scoring functions were applied. Since the prominent works by DesJarlais [24], who used a simple approach with some of the ligand’s degrees of freedom, several algorithms have been developed to explore molecular flexibility. When more degrees of freedom are added to the system, a larger number of potential protein–ligand complexes can be explored, thus increasing the likelihood of finding compounds with high binding affinity for the receptor. This is reflected as an increase in the effectiveness of the proposed methods.

Any method proposed for solving the docking problem has to deal with the following aspects [13] (Figure 3.1):

1. Definition of the binding site.
2. Selection of the molecular representation method.
3. Implementation of scoring functions.
4. Exploration of the search space.
5. Evaluation of the similarities between the predicted structures.

Each one of these aspects will be explained briefly below.

### 3.1 DEFINITION OF THE BINDING SITE

As shown by the first box in Figure 3.1, the binding site can be characterized by its geometric and physicochemical properties.

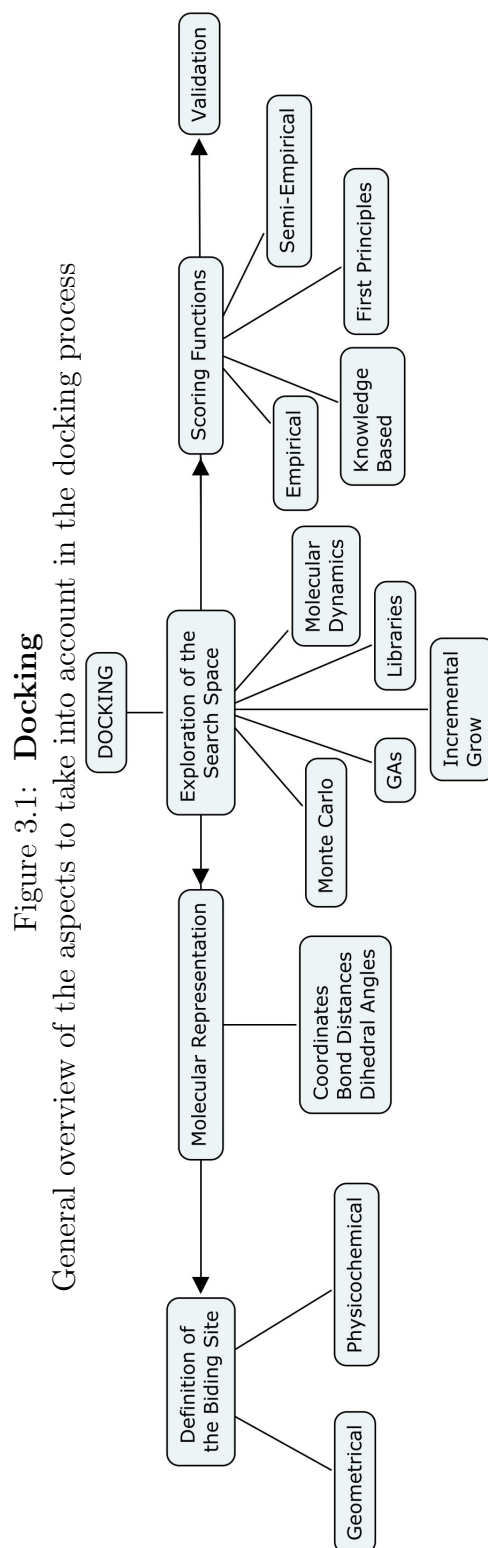
#### Geometric Aspects

Several approaches have been developed to define the geometric features of the binding site; CAST, PASS, POCKET and SURFNET are some of such methods [56, 39, 55, 53]. These methods are shown in the Figure3.2.

CAST (See figure 3.2A) represents atoms in the protein's binding pocket as spheres; centers in the spheres are linked with the other centers to obtain triangles. If three contiguous triangles do not pass through the atoms and one of them is obtuse, it is possible to say that a binding site is defined [56].

In PASS, the protein surface is covered by virtual spheres. Additional layers are located over the surface in an iterative process, until the spaces are filled in. In Figure 3.2 D, the binding sites are depicted as black points.

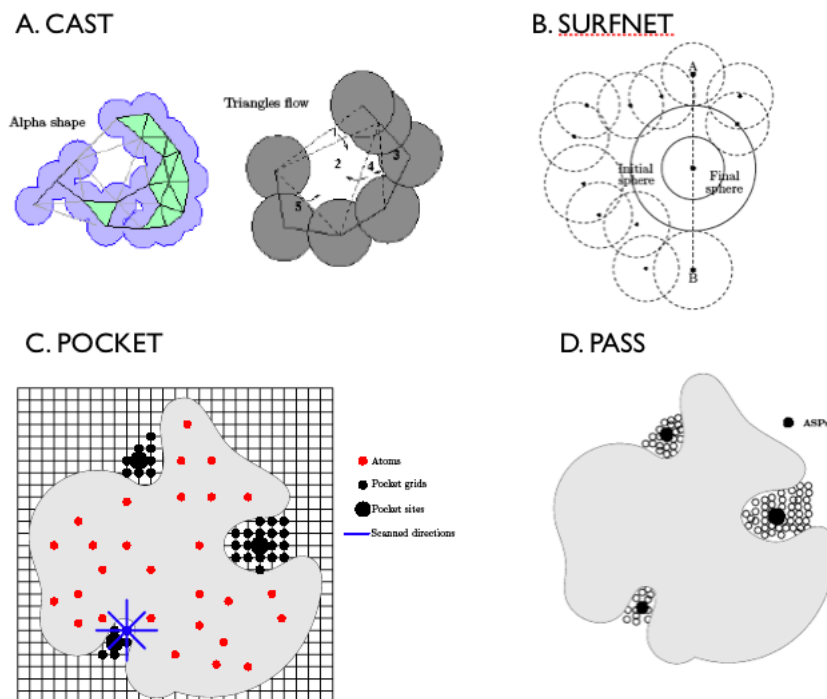
POCKET and LIGSITE are methods that identify pockets by using a series of simple operations on a cubic grid. Both libraries start by generating a regular Cartesian



grid. To start, all grid points are labeled as solvent and set to a value of 0; grid points that are inaccessible to the solvent are assigned a value of  $-1$ . As a straightforward method for this steric validation, distance checks are conducted to see whether a solvent molecule centered at a grid point overlaps with any atom of the protein [39, 55].

In the SURFNET approach, which is shown in Figure 3.2B, a sphere is placed between the van der Waals surfaces of a pair of atoms. The radius of this gap sphere is then reduced until it is not penetrated by any of the neighboring atoms. All the resulting final gap spheres are stored and compute to describe the shapes and sizes of the protein cavities [53]. Similar approaches can be found in [41] and [69].

Figure 3.2: Cavity identification methods



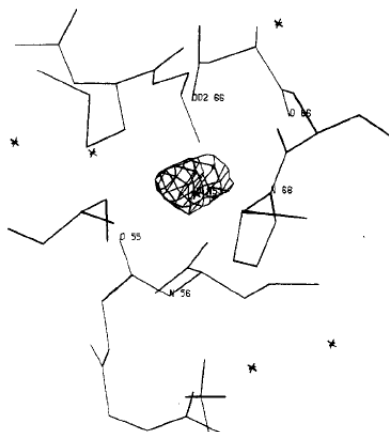
These images were taken from Huang et al. 2006 [42]

### Physicochemical Aspects

In addition to the need of defining the shape of the binding site, it is also necessary to characterize its physicochemical environment in order to determine the binding

affinity between both molecules. One of the first approaches to this task was proposed by Goodford in 1985. Figure 3.3 shows the result of applying this approach to phospholipase A2 (protein) and a water molecule (probe molecule). This approach computes the interaction energy of both molecules and gives an array of energy values, which define a contour where the interaction between the probe and the protein would be most likely to take place (dark region in the middle of Figure 3.3). Some other approaches are closely related to this idea [36, 23, 51, 77, 90, 84, 18, 9, 32]. One of the most interesting ones is the approach proposed by Trovov and Abagyan [4], who also used the idea of forming contours in order to recreate the volume and shape of the binding site.

Figure 3.3: **Contour defined by physicochemical interactions**



The structure used in the calculation was the phospholipase-A2 macromolecule in interaction with water molecules in a region that should have an interaction energy of -9 kcal/mol. The image was taken from the work reported by Goodford in 1985 [34].

## 3.2 MOLECULAR REPRESENTATION

An important problem for developing computational tools for drug design is the molecular representation. A molecular representation should allow introducing and manipulating data of molecular structures on a computer so as to obtain useful

results for the drug design process. Several methods have been proposed to represent a molecule, but in general terms, they are all closely related to the problem to be solved. For example, problems related to searching for conserved regions in a protein can be solved by using sequence information, which is a molecular representation with low level of detail, but if the problem involves calculating the binding energy, a more detailed molecular representation is necessary.

Taking into account the level of detail, molecular representations could be seen as implicit (compact) and explicit (detailed). Compact representations such as the SMILES format, which represent the molecule as a string of characters by specifying atoms and bonds, have the advantage of reducing the quantity of data needed to represent a molecule. However, information about atom localizations in the 3D space and the presence of hydrogens are lost in this type of representations. On the contrary, explicit representations do include atom types and their respective bonds with the other atoms. They have different levels of detail depending on whether atoms and bond types, presence or absence of hydrogen atoms are specified, so they can be adjusted to improve the exploration of the conformational space. However, such improvement causes an increase in the quantity of data needed per molecule. Since both aspects, detail and quantity of the data, are always in conflict, a wide spectrum of approaches ranging from very compact to highly detailed has been proposed.

Specifically for the docking problem, different methods have been formulated, among which the ones proposed by Kuntz et al., DesJarlais et al. and Rarey et al. [24, 54, 71] are some of the most outstanding ones. These methods represent the ligand as a composition of rigid parts in which each part is separately matched within the docking site without assuming dependence on the order in which the ligand fragments are placed.

The representation method proposed by Smillie et al. is based on clique finding, which is an NP-complete problem where each atom is considered as a vertice and covalent bonds as edges of the graph[83].

In the approach proposed by Mitzutani et al., the molecular reconstruction is done based on energetically favorable docking models by considering only the formation of hydrogen bonds between heteroatoms in the ligand and the protein receptor. The

position, orientation and conformation of the ligand are defined by the distances between hydrogen atoms in the ligand and hydrogen atoms in the protein receptor with ability to form hydrogen bonds [62].

The method proposed by Sandack et al. is based on an extension and generalization of the Hough transform and the geometric hashing paradigms for rigid object recognition. In this method, both molecules (the ligand and the receptor) are represented as sets of outstanding points, which represent their molecular surfaces. With these surfaces it is possible to screen a database of known ligands (models) for all ligands showing substantial partial surface complementarity to the receptor surface, without colliding with the receptor [78].

Goodsell and Olson employed the Metropolis method to search for ligand conformations and combine it with energy evaluations; this approach considers that molecular representations should allow changing each degree of freedom of the rigid body (the receptor), such as the translation and rotation, and also include the degree of freedom of each torsion angle of the ligand [35].

### **3.3 EXPLORATION OF THE SEARCH SPACE**

Once the appropriate molecular representation method(s) has been defined, it is necessary to explore the 3D space defined by the binding site and the conformations that this ligand could adopt in the search space. The exploration of the search space is important because the structures of some drug candidate molecules are unknown and the stable conformations that they would adopt in the binding site of a specific receptor are different from the ones they appear to have when they bind to other structures in vacuum.

According to a classification of the different approaches to explore ligand conformations in the binding site proposed by Kuntz et al. [13], the algorithms employed for conformational searches can be classified into one of the following three categories: systematic, stochastic and deterministic.

#### **3.3.1 SYSTEMATIC APPROACHES**

In systematic searches, the exploration is done based on the values associated to each

degree of freedom and each of these values is explored in a combinatorial fashion. As the number of degrees of freedom increases, the number of evaluations increases too. To deal with this problem, some constraints are introduced to prevent the algorithm from exploring zones that can give wrong solutions.

### **Incremental Construction or Fragment-based**

Generally, these methods divide the ligand into rigid and flexible pieces; rigid sections are defined between rotatable bonds. One rigid part is chosen by the method; particularly the fragment with less possible anchor positions. The number of anchor possibilities is denoted as  $n$  and it will determine the number of branches of the search tree. At the next level, the method selects the next fragment with the lowest anchor position possibilities and so on, until all the fragments have been included. Some of these methods take into account chemical interactions in order to reduce the number of anchor possibilities [71].

Some programs and approaches have introduced incremental construction (fragment-based) methods as an alternative to explore the conformational space. Examples of such methods are: DOCK 4.0[27], FlexX[71], LUDI [10], ADAM [63], Hammerhead [95], SLIDE [79] and SPECITOPE [80].

### **3.3.2 DETERMINISTIC APPROACHES**

These methods look for molecular conformations with low energy in such a way that the current states are closely related to the immediate previous conformations in a deterministic way. One important problem of deterministic methods is that they are prone to get trapped in local minima.

#### **Molecular Dynamics (MD)**

Explorations based on molecular dynamics work basically by setting a trajectory to explore the protein surface; however, one of the biggest problem with this kind of methods is that an MD trajectory will become often trapped in a local minimum and will not be able to step over high energy conformational barriers [13][88]. Moreover, they are extremely dependent on the initial state of the system, which makes it necessary to execute the process many times changing the initial conditions in order

to explore the whole protein surface and all the molecular conformational options. This increases the time demanded to explore the conformational space, making it very expensive in terms of time and computational resources. Examples of this kind of methods are AMBER [19] and CHARMM [14].

### 3.3.3 STOCHASTIC METHODS

Search algorithms based on stochastic methods make random changes, usually by changing a degree of freedom at once, guiding these changes by a scoring function that will allow the search to converge to good solutions. This kind of methods overcome the drawbacks of deterministic methods; however, they may have some convergence problems.

#### Monte Carlo Methods

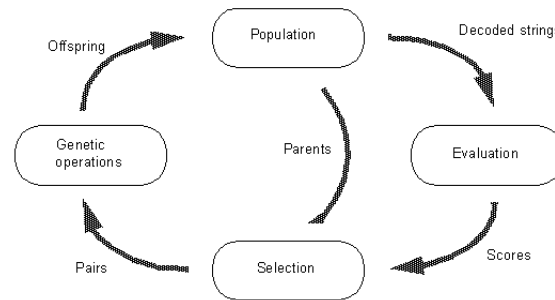
Standard Monte Carlo (MC) methods or Metropolis MC [61] consider the ligand as a whole. The position of the molecule is changed by random translation or rotation movements. Some implementations also include changes in rotatable angles, making it possible to explore and evaluate different molecular conformations. In these methods, ligands are usually placed randomly at the binding site, and changes are made randomly. The resulting ligands are evaluated and those with the lowest energy are chosen for the next cycle [13]. Early implementations of AutoDock [31] [33] used MC simulations; MC methods are also implemented in the program ICM [1] [2] and the MCDOCK package, which makes part of the DOCK program developed by Kuntz's group. Others examples are QXP [3, 45], GLIDE [47, 28], PRO\_LEADS and prodock [52].

#### Genetic Algorithms

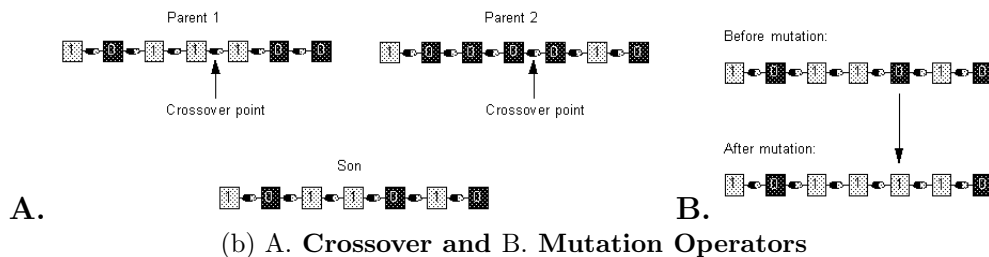
Generic Algorithms are also probabilistic search procedures, which are designed to work on large spaces involving states that can be represented as strings. The basic idea of genetic algorithms was developed in 1960 by Holland [40]; he presented the genetic algorithm concept as an emulation of the biological evolution and gave a theoretical framework for adaptation under the genetic algorithm. Holland's Genetic Algorithm evolves a population of "chromosomes", which are strings that represent solution samples distributed on the search space, driven by a kind of natural selection

together with genetic-inspired operators of crossover, mutation and inversion. Figure 3.4 illustrates the model. The operators used in this kind of algorithms can be described as follows:

Figure 3.4: **General diagram representation of a genetic algorithm cycle**



(a) **Reproduction cycle**



(b) **A. Crossover and B. Mutation Operators**

Figures were taken from Dilvan de Abreu Moreira, 1995 Agents: A Distributed Client/Server System for Leaf Cell Generation

- The *selection operator* tends to choose the most fitted chromosomes such that the selected chromosomes will produce in average more offspring than the less fitted ones.
- The *crossover operator* exchanges parts of two chromosomes in a similar way as the biological recombination between two single chromosomes does.

- The *mutation operator* makes random changes on the values at some locations in the chromosomes.
- The *inversion operator* reverses the order in which genes are arranged; however, this operator is not commonly used.

In summary, Genetic Algorithm methods should have at least the following elements

- A population of chromosomes.
- Evaluation and selection according to fitness.
- Crossover to produce new offspring.
- Random mutation.

Genetic algorithms have been used on several docking programs such as GOLD [46] and AutoDock [65], just to mention a few examples.

### 3.4 SCORING FUNCTIONS

Scoring functions are used as the objective functions to guide the docking process and to help evaluate the different kinds of conformations and locations that could be acquired by the ligand in the binding site; this kind of functions should allow selecting the most promising structures from the whole set of possibilities.

Ideally, scoring functions should estimate the binding energy between macromolecular receptors and small organic molecules in a fast way; a task that many of them have achieved. However, they should also calculate accurate binding free energies, but no scoring function has come even close to that [81].

There are important disadvantages in all scoring functions. All of them calculate the binding affinity as a sum of independent terms; hence, their complexity depends on the number of terms. Many of them disregard entropic effects because they evaluate interactions between a rigid receptor and a ligand in a single, frozen binding mode, and they also ignore specific solvation and desolvation effects.

In spite of the problems mentioned above, many scoring functions have been widely used in docking models with successful results. In general, docking models use different kinds of scoring functions and make it possible to obtain a consensus result[93]. Typically, they attempt to optimize the prediction results yielded by the docking programs, but they do not provide a better understanding of the protein–ligand interaction. Therefore, scoring functions are selected based on the specific conditions of the problem at hand. Also, discussions about which scoring function is most suitable for a particular system are often difficult when the 3D structure of the complex has not been experimentally observed [68]. In consequence, a blind combination of some arbitrary scoring functions would not necessarily lead to better results [93] , which is why the optimization of energy predictions is still an important field of work in docking models.

Scoring functions used in docking could be classified into three categories: force fields, empirical scoring functions, and knowledge-based. Each of these categories will be explained below.

### 3.4.1 FORCE FIELDS

Force field scoring functions have been developed using physical principles. Since all possible molecular interactions between a receptor and biologically interesting small organic molecules could be represented and optimized with different levels of detail and information, a wide range of force fields functions has been developed.

The first simulations of molecular interactions did not include explicit information on the molecules (such as atomic data) or excluded parameters such as the solvent, which is where the molecular interaction takes place, mainly because computer power was limited. Weiner et al. proposed one of the earliest force field scoring functions[94], even before the interaction of complex molecules could be simulated in an explicit solvent [20]. As computer power increased, it was possible to obtain

more realistic simulations, such as the one proposed by Jorgensen and co-workers, who proposed explicit solvent representations [48].

In a simplistic model, force field functions could be represented as shown in Equation 2, where (a) terms of bond and angles, which are represented by a simple diagonal harmonics (these terms are generally taken from reported experimental structure determinations), (b) Torsion parameters, which define the dihedral angles between atoms, (c) van der Waals terms, which are represented by 6–12 potentials and (d) Electrostatic interactions modeled by Coulombic interactions, which are taken from quantum calculations or empirical determinations[19]. (a) and (b) are energy terms related to bond atoms, while (c) and (d) are energy terms related with non-bond atoms.

$$E_{pair} = \underbrace{\sum_{bond} k_r (r - r_{eq})^2 + \sum_{angles} k_\theta (\theta - \theta_{eq})^2}_{(a)} + \underbrace{\sum_{dihedral} \frac{v_n}{2} \times [1 + \cos(n\phi - \gamma)]}_{(b)} + \sum_{i < j} \left[ \underbrace{\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6}}_{(c)} + \underbrace{\frac{q_i q_j}{\epsilon R_{ij}}}_{(d)} \right]$$

### 3.4.2 EMPIRICAL

Empirical scoring functions calculate the binding energy as a weighted sum of explicit hydrogen bonding and hydrophobic contact terms. Hydrogen bonds are described by distant terms and angles, and the binding affinity is defined by the deviation of distances and angles from idealized hydrogen bonding geometries. Additionally, these terms are weighted by the  $K_i$  (dissociation constants), typically obtained from the experimental assays reporting the protein–ligand structure.

Some of the empirical scoring functions used in docking models are DrugScore [84] and SuperStar [90].

### 3.4.3 KNOWLEDGE-BASED

This kind of energy functions were proposed as an answer to the massive increase

of structures in the protein data bank (PDB)<sup>1</sup>. These functions are built exclusively from statistical analysis of the complex structures that have been experimentally defined to date. They are based on the assumption that shorter inter-atomic distances between atoms with favorable interactions occur more often than average, whereas unfavorable interactions occur less frequently.

First, atom types must be defined on the ligand and the protein receptor; then when this problem has been resolved, the distances and frequencies between the different atom types can be established. Knowledge-based functions assume that repulsions are interactions that will not be present in the reported structures.

Thus, score values typically consist of hundreds of small contributions, which makes it difficult to interpret the results. An example of knowledge-based scoring functions is the one developed by Gohlke et al. [33] and Muegge [66].

---

<sup>1</sup> <http://www.pdb.org>

# Chapter 4

## AN EVOLUTIONARY APPROACH TO THE DOCKING PROBLEM

As explained in Chapter 2, the following are the the tasks that a solution to the docking problem must consider:

1. Definition of the binding site.
2. Selection of the molecular representation method.
3. Exploration of the search space.
4. Implementation of the scoring function.

The way in which the proposed method tackled each task will be explained in detail in this chapter.

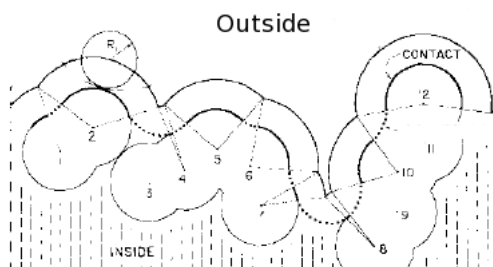
### 4.1 DEFINITION OF THE BINDING SITE

The characterization of the binding site allows defining the environment where the ligand will accomodate and its conformation in such space defined in the receptor. In the proposed method, the definition of the binding site will focus on the geometrical aspects, while the physicochemical properties (what anchor the ligands bind

to the binding site) will be evaluated by the scoring function. The geometric characterization begins by identifying the receptor surface; this reduces the number of possible places where to find cavities in the receptor molecule, since it is shown as a compact structure and not as a cluster of points (atoms) in the space. Once the receptor surface has been determined, it is possible to define the number of cavities in the protein, as well as their size and volume.

In the method postulated in this work the molecular surface of the receptor is calculated taking into account the approach proposed by Richards in 1977 [72]; who defined the molecular surface using van der Waals radii. According to this approach, each atom in the molecule is represented by its  $x$ ,  $y$  and  $z$  coordinates, which represent the atom's center and van der Waals radius. These values are used to generate potential spheres and the sum of these spheres yield a molecular surface. Figure 4.1 illustrates Richards' approach for calculating a molecular surface.

Figure 4.1: **Receptor Surface.**

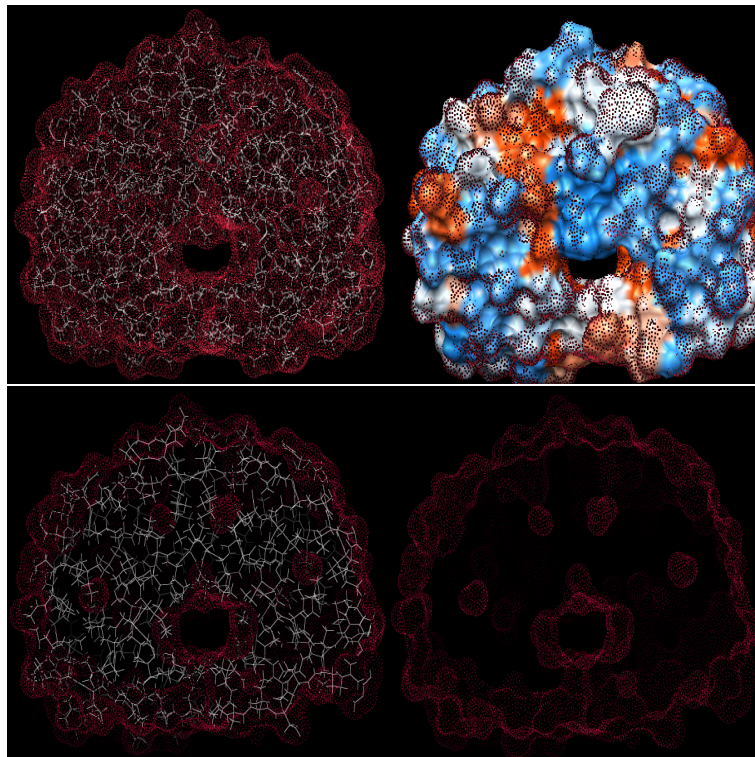


In this figure, the first line that binds the circles (which are the receptor atoms represented in two dimensions), defines the van der Waals surface. Image taken from Richards, 1977 [72].

Richards' proposal allows to define all the cavities in the receptor, including those to which solvents have no access to. This could be seen as an advantage because it defines cavities that are not available to the solvent when the molecule is crystallized, but that could be available at some moments during the normal functioning of the protein. An example of a protein surface obtained according to Richards' method

is shown in Figure 4.2. The surface was calculated using the DMS program<sup>1</sup>, and the graphics were generated using Chimera<sup>2</sup>; both of which were developed by the University of California, San Francisco (UCSF).

Figure 4.2: **Surface of the Human Immunodeficiency Virus-1 Protease** (PDB ID: 1aaq).



**A.** The protein is represented by the bonds shown in white while the calculated surface is shown as a red contour.

**B.** The calculated surface is the result of the sum of the van der Waals radii.

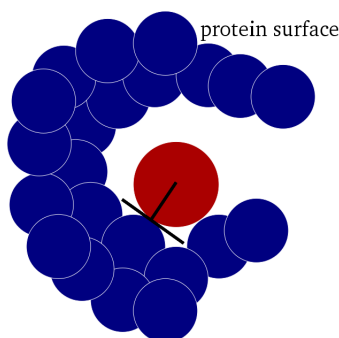
**C and D.** Even cavities located inside of the protein, which are not accessible to the solvent, are identified by Richards' method.

---

<sup>1</sup> <http://www.cgl.ucsf.edu/chimera/docs/UsersGuide/midas/dms1.html>

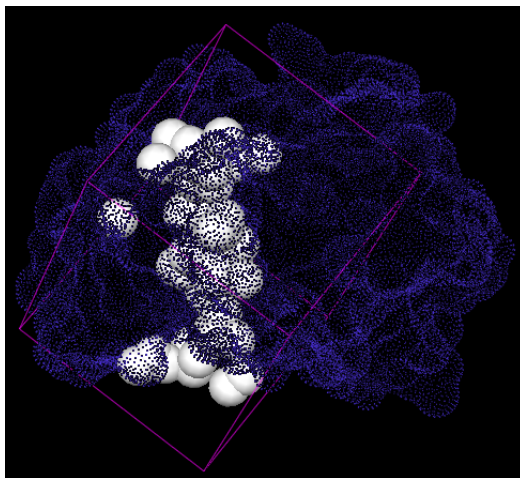
<sup>2</sup> <http://www.cgl.ucsf.edu/chimera/>

Figure 4.3: 2D Example of sphere generation over a surface



It should be noted that even though this method identifies cavities within the protein, it does not provide information about the shape and size of such cavities, nor it chooses in which binding site to evaluate the small organic molecules during the docking process. To overcome this issue, geometric complementarity tools have been implemented, among which one of tools that has shown better results is the `sphgen_cpp` tool used by the DOCK program [52]. This program generates a negative image of the receptor surface by placing spheres over the surface. Each sphere touches the molecular surface at two points and has its radius along the surface normal vector, as shown in Figure 4.3. The spheres are calculated over the entire surface, so that there is approximately one sphere per surface point. The representation obtained with the spheres is then filtered to keep only the largest spheres associated with the receptor atoms. A clustering algorithm is then applied to the filtered set to generate the negative image of the receptor surface and clusters are then ordered according to the number of spheres, which will correspond to the size of the cavity. Since the clusters could be located in different places over the receptor, the binding site where the ligand will be docked is selected according to the localization of the ligand that was reported in the elucidation of the protein–ligand complex structure. The cluster that defines this binding site is also used to define the receptor zone to be explored. Figure 4.4 shows the cavity selected as a binding site for the example shown in Figure 4.2.

Figure 4.4: Cavity selected on the Human immunodeficiency virus-1 Protease



## 4.2 SELECTION OF THE MOLECULAR REPRESENTATION METHOD

In docking, molecular representations must deal with two important aspects: the localization of the ligand with regards to the protein and the conformations that it can adopt during the binding process. The localization of the ligand with respect to the protein is solved by representing both in the same coordinate system and limiting its location to the area defined in the identification of the binding site; this helps to reduce considerably the vast search space to a specific area in the protein. Conversely, the ligand conformations could be seen as combinations of rotatable bond angles. Therefore, the degrees of freedom depend on translations and rotations over the  $x$ ,  $y$  and  $z$  axes and the dihedral angles defined by the number of rotatable bonds in the molecule.

The localization of the ligand in the binding site is computed using the same coordinate system and explicitly stating the locations of the atoms by their  $x$ ,  $y$  and  $z$  coordinates. Changes resulting from translations and rotations of the molecule are reflected as changes in the coordinates of all the molecule's atoms. However, exploring the different conformations of the molecule is not an easy task, the first available algorithms considered fragments of molecules or molecular groups with ability to establish specific bonds, as opposed to considering only dihedral angles,

which would make the exploration of the search space easier. Besides, algorithms that use representations based on information about rotatable bonds and dihedral angles are most commonly applied to proteins in which information about bond distances, bond angles and torsion angles can be limited to the twenty amino acids that conform them. In this work, an algorithm was implemented based on this idea so that different data representation methods could be alternated to simplify the exploration of the search space.

Some of the molecular representations that can be easily interchanged are shown in Figure 4.5. Each has its own advantages and disadvantages, as explain below.

### **Algebraic Representations**

In algebraic representations, such as the one shown in Figure 4.5(a), three coordinates ( $x$ ,  $y$  and  $z$ ) are computed for each atom, which is the most straightforward method of representing a molecular structure. With this method the potential energy function can be expressed, evaluated, or minimized when a structure has been defined in terms of Cartesian coordinates. Moreover, structural or dynamics changes are described in terms of atomic coordinates as well [75].

### **Geometric Representations**

An example of a ggeometric representation is presented in Figure 4.5(b). These kind of representations are based on the inter-atomic distances between pairs of atoms in the molecule. Its main advantage is that some inter-atomic distances, such as for example distances between covalently bonded atoms, are reported by experimental structure determinations. Therefore, it is possible to reconstruct a molecule when all the distances between the atoms have been reported [75].

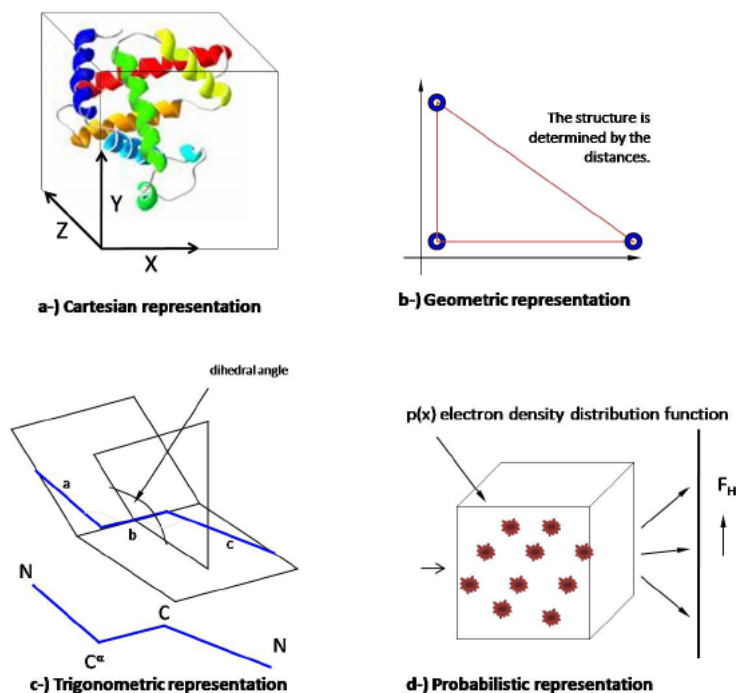
### **Trigonometric Representations**

Trigonometric representations rely on the idea that molecules can be represented by angles and distances. The main advantage of this type of representation is that few data is needed given that bond distances and bond angles are constant, and therefore it is only needed to represent dihedral angles to re-construct a molecule (see Figure 4.5c) [75].

## Probabilistic Representations

Probabilistic representations use entropy methods as a tool to reconstruct the electron density function based on X-ray crystallographic techniques [75].

Figure 4.5: Data representation of molecular structures



### 4.2.1 MOLECULAR REPRESENTATIONS USED BY THE PROPOSED METHOD

Of the four types of interchangeable representation methods described above, the docking model proposed in this work uses an algebraic and a trigonometric representation. Both representations have their own advantages and therefore are conveniently alternated during the docking process according to the specific needs of a particular stage.

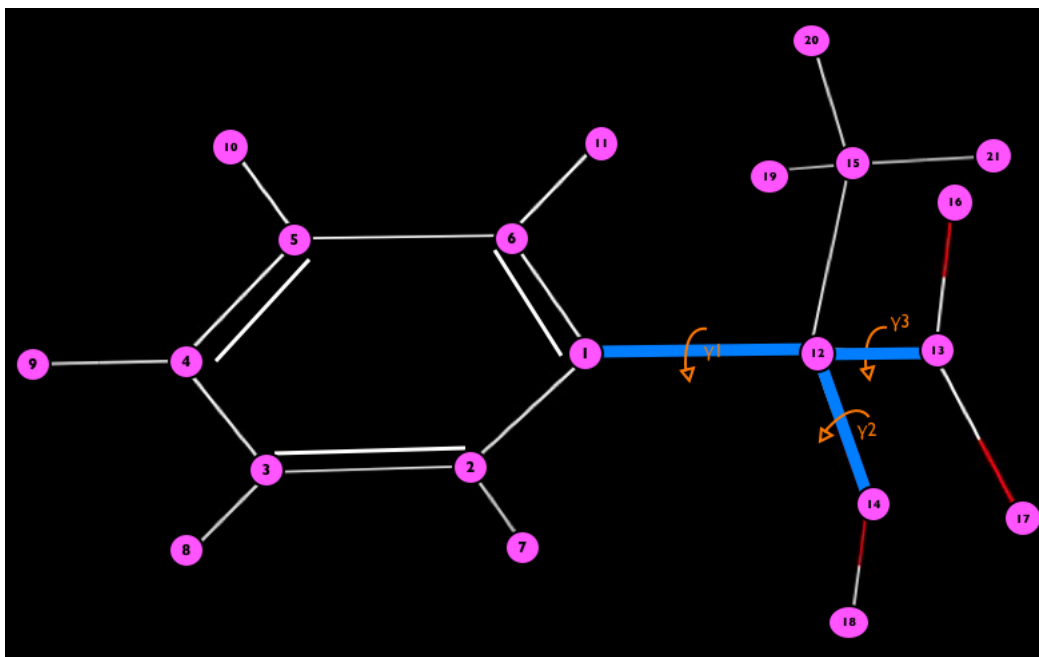
The *trigonometric representation* allows expressing the molecule in terms of torsion angles, which causes a significant reduction in the amount of information needed to represent a molecule. With this type of representation, it is easier to make changes in the ligand's molecular conformations so as to identify which ones bind more stably to the receptor. In addition, this representation can be more easily manipulated by the genetic algorithm used in the proposed approach, as it will be explained in Section 4.4 (Exploration of the search space) .

To represent a molecule, the relevant torsion angles needed are those formed between the four atoms that are involved in a rotatable bond. For example, to obtain a trigonometric representation of the molecule shown in Figure 4.6, the amount of necessary data is equal to the number of rotatable bonds, and their corresponding dihedral angles  $\gamma_1$ ,  $\gamma_2$ , and  $\gamma_3$ . The worst case occurs when the molecule is exclusively composed by atoms with single bonds; however, in this case the number of rotatable bonds is always significantly lower than the total number of atoms.

This highlights an important aspect about the trigonometric representation: it depends on the number of rotatable bonds in the molecule, which is no cause for concern in the case of proteins and nucleic acids (RNA and DNA) because their structures are limited to a few, relatively small, known amino acids or nucleotide *building blocks* given that the number of possibilities are vast and difficult to delimit.

In an *algebraic representation*, atoms are represented in an explicit way and the necessary data is larger than the data of the trigonometric representation. For example, the number of data needed to obtain an algebraic representation of the molecule shown in Figure 4.6, is equal to the  $x$ ,  $y$  and  $z$  coordinates of each of its 21 atoms and the explicit description of each covalent bond. The level of detail of an algebraic representation can increase, for example, by discriminating between atom types, not only according to their hybridization state but also by the atoms that they are bond to, as well as by specifying the type of covalent bond formed between them, not only depending on whether it is single, double or triple, but also specifying the chemical group in which it is established (e.g. amido, amino, carboxyl and aromatic groups). This was an important aspect to consider for the purpose of this work given that the higher the level of detail, the better the evaluation of the receptor–ligand complex made by the scoring function would be.

Figure 4.6: Comparison between the trigonometric and algebraic representations of a molecule



The violet circles represent the 21 atoms in the molecule. The blue lines represent the rotatable bonds and the white lines are non-rotatable bonds. In the algebraic representation, it is necessary to make explicit all atoms and bonds, while in the trigonometric representation only dihedral angles are required.

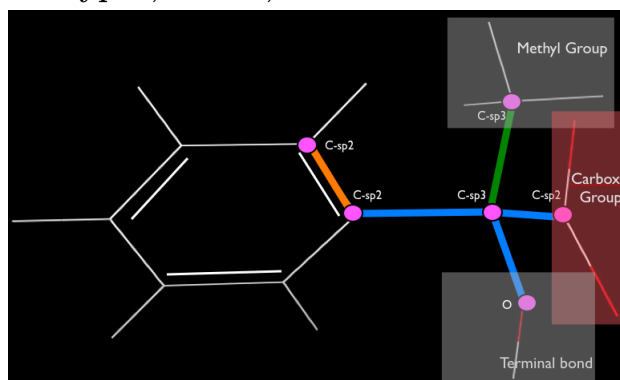
Although data manipulation at the computational level becomes more complex with an algebraic representation, the evaluation stage of the proposed docking method becomes easier and better results are obtained with this representation.

#### 4.2.2 ALTERNATING BETWEEN MOLECULAR REPRESENTATIONS

As explained above, a trigonometric or an algebraic representations are necessary at specific stages of the proposed docking method. For this reason, the alternation between the two representations was an important aspect to consider in order to perform an efficient search and evaluation of the molecules.

Both the ligand and the receptor are input to the program as algebraic representa-

Figure 4.7: Atom types, bonds, rotatable and non-rotatable bonds.



According to the conditions established for defining a rotatable bond, the molecule in the figure has three rotatable bonds, which are represented in blue. Although the bond represented in green is a single bond, it is one of the exceptions described in the text since it belongs to a methyl group. The orange bond is an example of a non-rotatable bond, as it is established between carbons with  $sp^2$  hybridization and belongs to a ring. The violet circles correspond to the atoms that are involved in such bonds, their hybridizations are noted together with the atom symbol. For example, a carbon with  $sp^2$  hybridization is denoted as C-sp2.

tions, but at different stages of the algorithm the ligand's algebraic representation is replaced by its trigonometric representation. For this purpose, the proposed method implemented an algorithm to perform an automatic and relatively fast alternation between the algebraic and trigonometric representations. This algorithm is based on the studies performed by Dong and Wu [25] on proteins, but applied to small organic molecules.

#### 4.2.2.1 FROM THE ALGEBRAIC TO THE TRIGONOMETRIC REPRESENTATION

To alternate between the algebraic and the trigonometric representations, the algorithm first defines the number of rotatable bonds in the small organic molecule. These bonds are defined using the tool developed by Kairys<sup>3</sup>, which makes it possible to distinguish between rotatable and non-rotatable bonds based on the defi-

<sup>3</sup> [http://www.ccl.net/cca/software/PERL/Find\\_Rotatable\\_Bonds/index.shtml](http://www.ccl.net/cca/software/PERL/Find_Rotatable_Bonds/index.shtml)

nitions made by Makino, Kuntz [58] and Hanser et al. [38]. Rotatable bonds are mainly defined by the atom types involved in the bond. Basically, rotatable bonds could be single bonds; those formed by atoms with  $sp^3-sp^3$  and  $sp^3-sp^2$  hybridizations. Some other bonds, such as those established between  $sp^2-sp^2$  atoms (double bonds), and  $sp^3-sp^3$  atoms present in a ring, could be considered as rotatable, but they are excluded since the flexibility of these bonds is limited. If it is necessary to explore the different conformations resulting from rotating some of the bonds that were excluded, the conformations must be previously generated and each molecule must be treated as unique. Additionally, terminal bonds, such as bonds between carbon-hydrogen atoms and methyl groups, are not considered as rotatable bonds by the proposed method. Some examples of atom types, bonds, and rotatable and non-rotatable bonds are illustrated in the molecule shown in Figure 4.7.

Once both of the atoms involved in a rotatable bonds are identified, two contiguous atoms have to be considered to calculate the dihedral angle. For example in Figure 4.8, the dihedral  $\gamma_1$  angle is computed using the information on the four contiguous atoms depicted in violet. In the same way, the dihedral angles  $\gamma_1$ ,  $\gamma_2$ , and  $\gamma_3$  are computed for each rotatable bond in the molecule to obtain a trigonometric representation of the molecule.

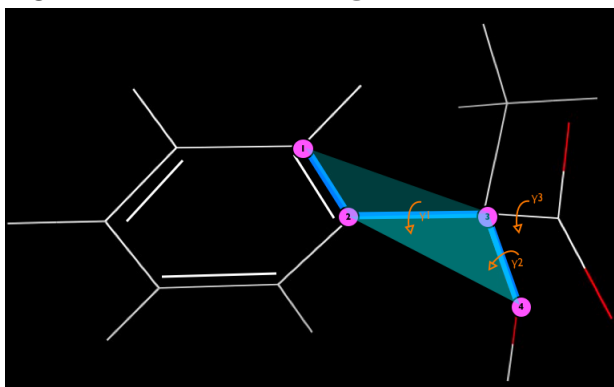
Accordingly, dihedral angles are computed using Equation 4.1; Figure 4.9 shows the vectors and angles needed for such computation.

$$\cos\alpha_{ijkl} = \frac{(axb)(axc) - (axc)(bxb)}{\|a\| \|b\|^2 \|c\| \sin\alpha_{ijk} \sin\alpha_{jkl}} \quad (4.1)$$

#### 4.2.2.2 FROM THE TRIGONOMETRIC TO THE ALGEBRAIC REPRESENTATION

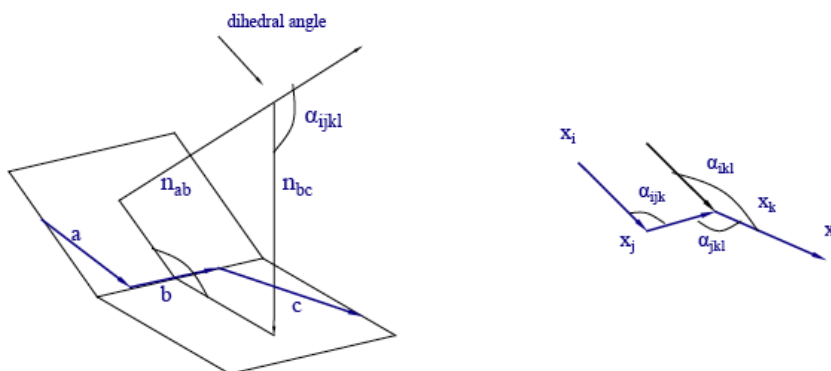
The transformation from the trigonometric to the algebraic representation is done according to the method proposed by Dong and Wu [25]. The molecular reconstruction is performed by computing the localization of a fourth atom, based on the information about the coordinates, distances, bond angles and dihedral angles of the three previous atoms. An example of this is explained in Appendix 1.

Figure 4.8: Dihedral angles in a molecule



Atoms 1, 2, 3 and 4 are defining the planes (represented in blue) for calculating the dihedral angle. The rotatable bonds in that molecule are expressed by  $\gamma_1$ ,  $\gamma_2$ , and  $\gamma_3$ .

Figure 4.9: Planes, vectors and angles used for the calculation of dihedral angles.



The four atoms involved in the rotatable bond are represented by  $x_i$ ,  $x_j$ ,  $x_k$ ,  $x_l$ ; each bond is noted by vectors  $a$ ,  $b$  and  $c$ ; the angle formed between vectors  $a$  and  $b$  is named  $\alpha_{ijl}$ , the angle formed between  $b$  and  $c$  is named  $\alpha_{jkl}$ . Finally, the calculated angle is the one defined by the normals to the planes formed by the four atoms, which is named  $\alpha_{ijkl}$ .

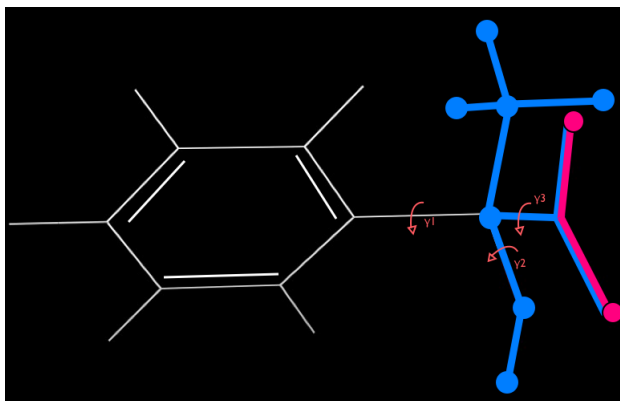
Information regarding distances, bond angles and non-rotatable dihedral angles are considered constants and therefore are stored based on the algebraic representation of the molecule. On the contrary, variable data such as those regarding dihedral angles from rotatable bonds are expressed in the trigonometric representation. Since changes in the dihedral angles of rotatable bonds are propagated in the molecule to atoms that are directly or indirectly involved in the bond; these changes are regis-

tered in a new algebraic representation. An example of this situation is presented in Figure 4.10. If  $\gamma_3$  is rotated, the atoms highlighted in pink will adopt a new orientation based on the angle specified in the trigonometric representation. However, if the change is produced in  $\gamma_1$ , all the atoms highlighted in blue will have a new orientation. Although in this example only changes in one direction are considered, it is possible to introduce changes in any direction.

Therefore, it is possible to represent any conformation of a molecule by computing the localization of the fourth atom based on the information of the three previous atoms (see Appendix 1 ).

Since the proposed method can produce stable and unstable molecules, the scoring function is used to evaluate the viability of such predicted structures.

Figure 4.10: **Atoms affected by changes on dihedral angles of the rotatable bonds**



Atoms highlighted in pink are affected by changes in  $\gamma_3$ , while changes  $\gamma_1$  are reflected in atoms highlighted in blue.

### 4.3 IMPLEMENTATION OF THE SCORING FUNCTION

Energy functions are used to guide the search for molecules with the ability of bind to protein receptors based on the evaluations and assignment of scores that will allow to differentiate between stable and unstable protein–ligand complexes.

An energy function based on the force field was chosen as the scoring function. This function includes data from both quantum chemistry calculations and experimentation, in the specific case of this work, it allows to evaluate energy of both large molecules (DNA, RNA and proteins) as well as of small organic molecules. It is worth noting that computations based on quantum chemistry are more accurate than the experimental ones, but the former are computationally more expensive. However, evaluations performed by force field scoring functions are not necessarily closely related to the experimental results and the same occurs even with other types of scoring functions. Some successful docking programs have used scoring functions based on the force field to evaluate molecular complexes [27, 46].

In general, energy functions based on the force field assign scores to the evaluated molecules by adding energetic terms from the molecules involved in the system [91]. The molecules can be of different kinds, including small organic molecules, proteins, DNA, RNA or solvent, just to mention few. Each term in Equation 4.2 is computed by considering all the atoms involved in the system.

We can describe the energy as a sum of bond and non-bond energetic contributions (see Equation 4.2). Bond terms refer to the molecule's internal energy and they include covalent bonds (bonds), bond angles (angles) and torsion angles (dihedrals), while non-bond terms are related to interactions where no covalent bonds are established but electrostatic attractions or repulsion forces are presented: they include van der Waals forces, which are represented by the term  $(\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6})$ , and Coulomb forces represented by the term  $(\frac{q_i q_j}{\epsilon R_{ij}})$ .

$$E_{pair} = \sum_{bonds} k_r (r - r_{eq})^2 + \sum_{angles} k_\theta (\theta - \theta_{eq})^2 + \sum_{dihedral} \frac{v_n}{2} \times [1 + \cos(n\phi - \gamma)] + \sum_{i < j} \left[ \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\epsilon R_{ij}} \right] \quad (4.2)$$

**Bond terms** could be seen as hard and soft parameters. Hard parameters are bonds and bond angles, and their values are obtained from experimental data of the reported structures. They generally consist of a reference value and a range where it can fluctuate; deviations from the reference value are penalized by an unfavorable score. In contrast, torsion angles are considered as the soft parameter

because they could be influenced by non-bond terms; however, they are also obtained from experimental data and are modeled by a harmonic function were the rotatable angle could vary between 0 and 360 degrees, with minimizations made by quantum chemistry calculations.

Reference values determined from experimentation change according to the type of atoms involved in the bonding, such that the distance between atoms of carbons with  $sp^2$  hybridization (C=C) is not the same as the distance between carbon atoms with  $sp^3$  hybridization (C-C). The same phenomenon occurs with bond angles where three atoms are involved, or torsion angles where four atoms are interconnected. A higher discrimination of atom types allows a better approximation of the energy values predicted by the Force Field Scoring Function to energy values obtained by the experimental studies. As in the molecular representation, in small organic molecules atoms are not limited to just a few types, compared to macromolecules such as proteins and nucleic acids (RNA and DNA), and therefore they cannot be easily discriminated.

We used the atom types defined by the Amber Force Field developed by UCSF University [19] and its expansion to small organic molecules [91] was used. Atom types were assigned by the antechamber program [92], while those that were not assigned by the program were determined by taking into account: the atom symbol (for example C, O, H, N, P, etc.), hybridization, and neighbor atoms.

The reference values to assign the atom types of the protein receptor were taken from AMBER [19], while reference values for the ligand were obtained from GAFF [91]. However, it may happen that some reference values were not present, especially, in the case of ligands. In such cases, the reference values were taken from those which had the nearest atom type array. For example, if the reference value for the dihedral angle formed by the atoms X1-X2-X3-X4 has not been determined experimentally and therefore reported in GAFF [85], the values for X1-X2-X3-X5, where X5 has the same atom symbol and hybridization of X4, would be considered instead.

**Non-bonding terms** referred to any kind of interactions occurring without the existence of covalent bonds. These interactions could be established by electric charges of ions, partial atom charges, and atomic radii that produce repulsion when two atoms collide. The calculation of the attraction or repulsion due to electric atom

charges is done according to the Coulomb law, while attraction or repulsion forces due to the atoms' radii are calculated according to the Lennard–Jones potential function [34].

The radii of the atoms were taken from either the AMBER or GAFF force fields depending on whether the evaluated molecule was a protein or small organic molecule. These values were used to calculate the score of attraction or repulsion due to the van der Waal terms considered in Equation 4.2.

The partial atomic charges, which are employed to calculate electrostatic attractions or repulsions, as expressed in Equation 4.2, can be computed according to different methods; however, there is no consensus about which is the best method to obtain them. In this work, a widely used method based on atomic connectivities was adopted. This method was introduced by Gasteiger [29] and its calculation is relatively fast compared to the calculations based on quantum chemistry; such as with the AM1-BCC method [43]. In small organic molecules a partial charge values must be calculated for each atom of a new input molecule, whereas in organic macromolecules partial charge values are initially assigned for each of the atoms of the 21 amino acids (in the case of proteins), and therefore do not have to be calculated everytime a new proteins is introduced, which greatly simplifies calculations.

Generally, the scoring functions that are used to evaluate complexes protein–ligand give unique energy values; However, as explained before, it is possible to consider those values as terms that evaluate different aspects of the molecule: the internal energy, which is determined by the bonding terms, and the external energy, which is defined by the non-bonding terms. These two terms could be in conflict when there are two or more molecules interacting in the same system; for example, in vacuum a molecule takes a particular conformation that is stable under such conditions; here bond and non-bond terms have different weights compared to when the molecule is interacting with other molecules and therefore could adopt different conformations, which are stabilized by non-bonding interactions.

## 4.4 EXPLORATION OF THE SEARCH SPACE

The docking problem will thus be considered as a multi-objective optimization prob-

lem, unlike the typical search approaches which focus on the optimization of a single objective function and look for the molecule with the best energy score. Such energy values are obtained from a scoring function, which evaluates the molecular complexes. Only recently, multi-objective optimization began to be applied to the solution of bioinformatics or computational biology problems [37].

The multi-objective approach is implemented in the proposed method to deal with the conflicting terms in the energy function. It does not consider the energy function as a unique value, but instead it optimizes both energy contributions from the bonding and non-bonding terms. Although the docking problem can have other objectives that could be optimized, such as geometric complementarity, the proposed model will only consider the bonding and non-bonding terms.

#### 4.4.1 MULTI-OBJECTIVE OPTIMIZATION

A multi-objective optimization is an appropriate approach to the solution of problems where one or more optimal solutions may arise when several objectives are considered. Typically, the objectives are estimated very differently and are often conflicting aspects of the solutions [75].

In general, a multi-objective optimization problem (MOOP) can be defined as follows:

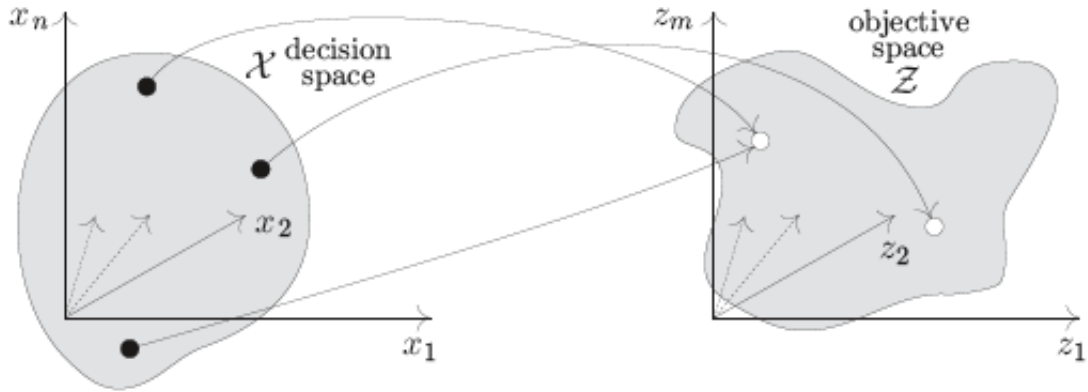
**Definition 3.1** Multi-Objective Optimization Problem:

Finding a vector  $x = [x_1, x_2, \dots, x_n]^T$  which:

- i*) Satisfies the  $r$  equality constraints, such that  $h_i(x) = 0$ ,  $1 \leq i \leq r$ ;
- ii*) Is subject to the  $s$  inequality constraints  $g_i(x) \geq 0$ ,  $1 \leq i \leq s$ ;
- iii*) and optimizes the vector function  $z = f(x) = [f_1(x); f_2(x), \dots, f_m(x)]^T$

According to this definition, the MOOP focuses on finding the optimal values for the vector  $x$  (in the space of variable decisions  $X$ ), that minimizes/maximizes the function  $f(x)$  (in the objective space  $Z$ ). This optimization process is carried out by taking in to account that the constraints  $h$  and  $g$  need to be satisfied. The vector  $x$  is an  $n$ -dimensional decision vector in the  $X$  space, and  $z = f(x)$  is an objective vector that maps  $X$  into  $R_m$ , where  $m$  is the number of objectives to be optimized.

Figure 4.11: The  $n$ -dimensional parameter space maps to the  $m$ -dimensional objective space.



The image of  $X$  in the objective space is the set of all attainable points  $Z$  (See Figure 4.11)

The optimal set of solutions is defined using the concept of dominance, which is used to compare two solutions.

**Definition 3.2** Dominance. A vector solution  $x$  is said to dominate a solution  $x'$ , denoted by  $x \prec x'$  if  $z_i \leq z'_i, \wedge, \exists_i : z_i < z'_i$  for  $1 \leq i \leq m$

In other words,  $x \prec x'$ ; if the solution  $x$  is no worse than  $x'$  in all the objectives, and  $x$  is strictly better than  $x'$  in at least one objective. Note that if solution  $x$  does not dominate solution  $x'$ , this does not imply that  $x'$  dominates  $x$ . Furthermore, there are three possible outcomes of this relation:  $x$  dominates  $x'$ ;  $x$  is dominated by  $x'$ ; or  $x$  and  $x'$  do not dominate each other.

A Pareto optimal set of solutions  $X$  consists of all those vectors that do not dominate each other, and it is impossible to improve it in any objective without getting worse results in another objective. Thus a Pareto optimal front can be defined as follows:

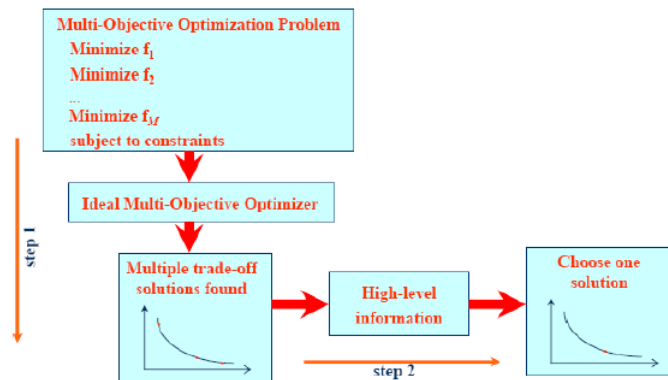
**Definition 3.3** Pareto Optimum Solution. A solution  $x \in X$  is said to be a Pareto optimal solution if and only if there is no  $x' \in X$  such that  $x \prec x'$ .

There are two aspects that are preserved in the set of solutions that belong to a

Pareto front: 1) any two solutions non-dominated in the Pareto front must not be dominated by each other. 2) Any solution that does not belong to the set of non-dominated solutions is dominated by at least one of the solutions in the set.

It is clear that the multiple solutions obtained in the Pareto front consist of the best results for the evaluated objectives. However, it is also necessary to have a high-level decision maker in order to choose the Pareto optimal solution. Figure 4.12 shows the principles involved in an ideal optimization procedure [22]. Step 1 produces multiple trade-off solutions, but in the next step, a higher level of information is necessary in order to choose a final solution.

Figure 4.12: **An ideal multi-objective optimization procedure.**



**Step 1**, which includes the optimization process, is illustrated in the first three boxes. A Pareto front with more than one solution is obtained at the end of this step; however, in **Step 2** the main goal is to obtain one solution from the Pareto front, which is done by introducing high level information.

Typically, a Pareto front includes a high number of solutions. Therefore, stochastic methods such as evolutionary algorithms allow to get in one run a set of solutions instead of a single one, unlike classic search algorithms which need of several runs in order to obtain different solutions. Genetic algorithms mimic natural evolution to perform search and optimization processes to find the best solutions for a problem. Evolutionary algorithms that deal with multiple objectives are called Multi-Objective Evolutionary Algorithms (MOEA).

In this work, a MOEA called Elitist Non-Dominated Sorting Genetic Algorithm

(NSGA-II) was used [22]. Other MOEA are also available in the proposed model, but NSGA-II is most widely implemented MOEA in diverse kinds of problems. This is specially useful since the use of MOEAs in biological problems is relatively new and therefore is not known which MOEA is better for this kind of problems.

#### 4.4.1.1 THE MULTI-OBJECTIVE EVOLUTIONARY ALGORITHM

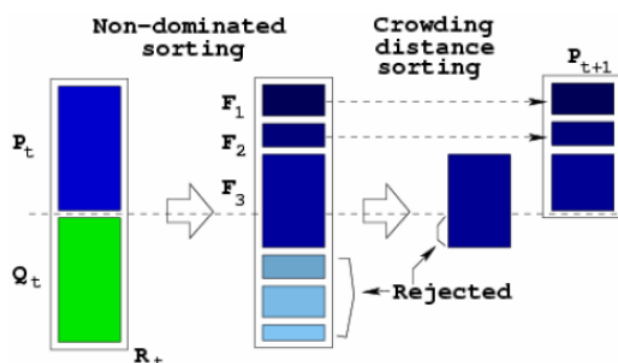
The NSGA-II algorithm was introduced by Deb. Et al in the year 2000[22]. This algorithm is based on an elite-preserving strategy and an explicit diversity-preserving mechanism [75].

Figure 4.13 depicts the NSGA II algorithm. In general, it operates as follows [21]:

- Step 1: Combine parent ( $P_t$ ) and offspring ( $Q_t$ ) population and create  $R_t = P_t \cup Q_t$ .  
Perform a non-dominated sorting of  $R_t$  and identify different fronts:  $F_i$ ,  $i = 1, 2, \dots, etc.$
- Step 2: Set new population  $P_{t+1} = \emptyset$ . Set a counter  $i = 1$ .  
Until  $|P_{t+1}| + |F_i| > N$ , perform  $P_{t+1} = P_{t+1} \cup F_i$  and  $i = i + 1$
- Step 3: Perform a crowding-sorting procedure and include the most widely spread ( $N - |P_{t+1}|$ ) solutions by using the crowding distance values in the sorted  $F_i$  to  $P_{t+1}$ .
- Step 4: Create offspring population  $Q_{t+1}$  from  $P_{t+1}$  by using the crowded tournament selection, crossover and mutation operators.

One of the most interesting aspects of the NSGA-II algorithm is the use of the crowding distance among the non-dominated solutions, which evaluates the diversity

Figure 4.13: NSGA II procedure.



of the population. The crowding distance can be calculated either in the objective or in the variable space. The NSGA-II algorithm was used to find whether the tested ligand bind to the receptor's binding site, specifically, to evolve conformations and locations of the ligand in the binding site of the receptor.

A Java-based framework called jMetal<sup>4</sup> was used in the study and experimentation of the multi-objective algorithm to solve the docking problem.

#### 4.4.1.2 THE MULTI-OBJECTIVE FORMULATION

In multi-objective problems, it is necessary to define two spaces: the decision space that represents the set where the solutions would be, and the objective space, where the set of solutions is represented as a function of the objectives to be optimized. Also, a set of constraints need to be defined, which will determine the set of feasible solutions.

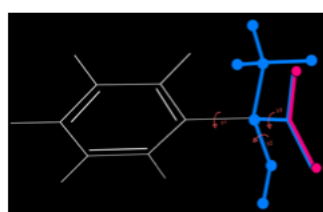
**Decision space** (decision variables). In genetic algorithms, the decision variables are represented as chromosomes that contain the information of the solutions to the problem.

In this work, the chromosomes represent conformations and locations of small organic molecules in the binding site. Particularly, a molecule is represented as an array of real values that represent the location of the ligand (ID) and the molecule's

<sup>4</sup><http://jmetal.sourceforge.net/>

conformation through the dihedral angles of rotatable bonds. Figure 4.14 depicts a representation of the possible solutions the molecule that appears next to the chromosome representation. It is possible to see in this example that the length of the chromosome depends on the number of rotatable bonds in the ligand. The operators over individuals of the algorithm are defined in a similar way to those explained in Chapter 2, except by mutation in the ID variables. Particularly, in ID, six variables can be mutated: three for the translation and three for the rotation (over the axis  $x$ ,  $y$  and  $z$ ). However, the number of changes was limited to three when a mutation in the ID is performed. Moreover, these changes are randomly chosen.

Figure 4.14: **Chromosome representation used for the genetic algorithm.**



$\gamma_1$	$\gamma_2$	$\gamma_3$	ID
------------	------------	------------	----

$\gamma_1$  = Rotatable angles

ID = Identification related with  
the location

**Objective Space** (objective functions). In the proposed method, different terms of the scoring functions are selected to formulate the multi-objective docking problem. Particularly, two objectives were considered in this work.

According to the energy contributions in the scoring function, the two objectives were defined as follows: The first objective corresponds to the energy contributions from the covalent bonds between atoms (bonding terms), such as bonds, bond angles and torsion angles. The second objective is related to the interactions where a covalent bond does not exist, such as electrostatic attraction and repulsion forces, and van der Waals terms. The energy terms for these two terms will be denoted as:

$$f_1 = E_{bond}$$

$$f_2 = E_{non-bond}$$

**Feasible regions** (Equality/inequality constraints). The constraints that determine the feasible solutions in the proposed method are of two kinds: steric constraints

and limitations to the location of the ligand outside the binding site. The steric constraints are represented by the collisions between atoms from the same molecule and with atoms of the receptor. These collisions are penalized by the scoring function, but are not included within the optimization constraints. On the other hand, the second type of constraints is related to infeasible locations of the ligand. These are included by delimiting the binding site, and solutions out of this physical limit will not be generated.

#### 4.4.1.3 DECISION-MAKING PHASE

The NSGA-II multi-objective algorithm evolves a set of ligands that will bind to a receptor and yields a set of good solutions based on the selected objectives. However, it is necessary to include a high level analysis in order to choose one of the solutions in the final Pareto front. In general, this is a difficult task, specially when the number of objectives is high and the set of solutions is large.

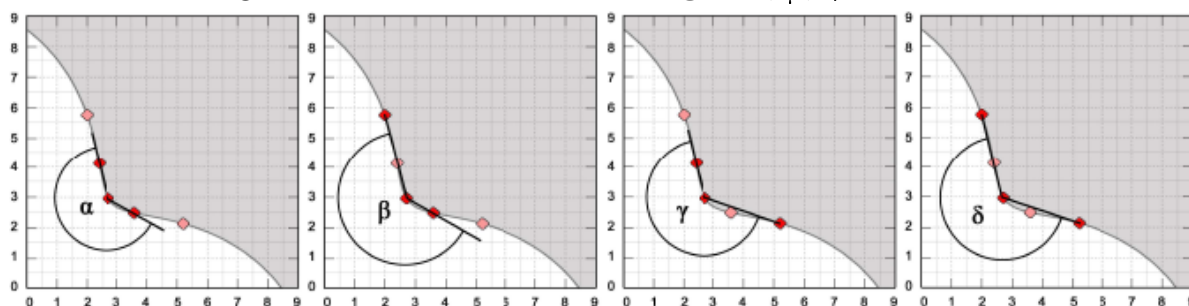
A proposal made by Branke et al. [12] focuses the selection of the solutions from the Pareto front on the identification of the knees, i.e., regions in the Pareto front where small displacements produce a big detriment on at least one of the objectives. Two methods to find the knees were considered: one based on angles and the other one based on the utilities.

The angles are defined by two lines that are traced between four points in the Pareto Front (see Figure 4.15). These angles indicate the presence or absence of a knee. The larger the angles  $\alpha$ ,  $\beta$ ,  $\gamma$  or  $\delta$  are, the higher the probability of classifying the solution as a knee.

The decision-making process based on the utilities focuses on the cost of choosing the next best solution. It can be defined as the additional cost that must be accepted if the best solution is not available and the second best solution is chosen. In this work, both decision makers (based on angles and utility) were implemented.

## 4.5 EVALUATION OF THE OBTAINED RESULTS

The proposed approach was tested with molecular complexes reported in PDB . This

Figure 4.15: Points to define angles  $\alpha$ ,  $\beta$ ,  $\gamma$  or  $\delta$ .

The figure shows the four angles ( $\alpha$ ,  $\beta$ ,  $\gamma$  or  $\delta$ ) computed to define the presence or absence of kees in the Pareto front. Each box shows the points (or the solutions) taken into account to define each angle.

method was evaluated using a metric widely used in the prediction of 3D molecular structures: the root mean square deviation (RMSD).

The RMSD measures the distance between the predicted conformation and the reported structure, such that a value of zero indicates an exact coincidence between both structures. The RMSD is defined by the following equation 2:

$$RMSD_{(a,b)} = \sqrt{\frac{\sum_{i=1}^n |r_{ai} - r_{bi}|^2}{n}}$$

Though the molecular complexes are superimposed, only the distance between the atoms of the ligand are taken into account to compare the molecular complexes, because the receptor is kept fixed and no atom in the receptor changes. Accordingly,  $r_{ai}$  and  $r_{bi}$  are the locations of the  $i$ -th atom of the ligands  $a$  and  $b$ , respectively.

The RMSD does not depend exclusively on the conformation and location of the molecule; it is also determined by the size of the molecule and the protein structure. Therefore, it is not always true that the best alignment has the best RMSD.

In this work, the Superimpose tool included in the set of programs TINKER<sup>5</sup>, developed by the Ponder laboratory, was used. Superimpose calculates the distance between the atoms of a molecule. It allows to superimpose specific atoms or the complete structure.

<sup>5</sup> <http://dasher.wustl.edu/tinker/>

This tool also allows to translate and rotate the molecule, thus minimizing the RMSD value. Particularly, it produces three RMSD values:  $\text{RMSD}_1$ , which measures the distances between atom of both ligand structures without inducing changes in translation or rotation. This gives an idea about the ability of the method to find the ideal location in the binding site.  $\text{RMSD}_2$  corresponds to the distance between atoms after being translated to the same position of the reported atom; this measure does not give much information about the goodness of method itself. The third value  $\text{RMSD}_3$ , is the result of translating and rotating the ligand in order to obtain the best possible superposition between the structures, and allows to know whether the molecular conformations are the same or not.

## 4.6 FRAMEWORK OF EXPERIMENTATION

This chapter describes the experimental framework followed to evaluate the behavior of the proposed model. In brief, experiments were carried out on four protein–ligand complexes (PDB IDs: 1ABE, 1ACM, 1BAF, 1CDG). The performance of the model was evaluated by considering the divergence between predicted and reported structures and the results obtained from such comparisons are described and discussed.

### 4.6.1 MOLECULAR COMPLEXES

The complexes used to evaluate the proposed method were chosen based on the following characteristics: (1) proteins involve in the complexes have a remarkable importance in pharmacology, (2) the complex structure is defined with sufficiently high resolution so as to be suitable for the docking evaluation, and (3) complexes have been widely used in the evaluation and validation of several of the methods available to approach the docking problem. The pharmacological relevance of the complexes is explained in the following paragraphs, while the structural characteristics in regards to number of atoms, rotatable bonds and resolution of the structure determination method are presented in Table 4.1.

Complex 1ABE has been important for understanding the role of periplasmic proteins in active transport [70]. It is formed by the L-arabinose-binding protein from *Escherichia coli* and both  $\alpha$ - and  $\beta$ - L-arabinose monomers [70].

Table 4.1: Data set of complexes

PDB ID	Protein	Ligand	L.A.	R.B.	Method	Resolution (Å)
1ABE	L-Arabinose-Binding Protein	L-Arabinose	20	4	X-Ray Diffraction	1.70
1ACM	Aspartate Carbamoyltransferase	<i>N</i> -(Phosphonacetyl)-L-Aspartic acid	26	6	X-Ray Diffraction	2.80
1BAF	Igg1-Kappa AN02 Fab	<i>N</i> -(2-aminoethyl)-4,6-dinitro- <i>N'</i> -(2,2,6,6-tetramethyl-1-oxy-piperidin-4-yl) -benzene-1,3-diamine	35	4	X-Ray Diffraction	2.90
1CDG	Cyclo dextrin glycosyl-transferase	Maltose	45	12	X-Ray Diffraction	2.00

L.A.: Number of ligand atoms.

R.B.: Number of rotatable bonds.

The protein in the 1ACM complex, the aspartate transcarbamoylase catalyzes the reaction between Carbamoyl Phosphate and L-aspartate to form *N*-carbamoyl-L-aspartate and inorganic phosphate. The carbamoyl aspartate thus formed proceeds through the pyrimidine biosynthetic pathway ultimately leading to the formation of pyrimidine nucleotides [85]. The ligand has been used as an inhibitor of this enzyme in the treatment of cancer.

Complex 1BAF corresponds to the crystal structure of the Fab fragment of the murine monoclonal anti-dinitrophenyl-spin-label antibody AN02 forming a complex with its hapten, which has been solved at a resolution of 2.9 Å [16].

According to Moriwaki et al. [64], the protein involved in the 1CDG complex is the Cyclodextrin Glycosyl-Transferase (CGTase); a monomeric enzyme with a molecular weight of about 74,500 Da, containing a sequence of amino acids structurally similar to the  $\alpha$ -amylase enzyme. Besides the cyclization reaction, which forms the Cyclodextrins (CDs), CGTase catalyzes the coupling reaction and the disproportionation of linear maltodextrins (see [67] and [60]). This reaction opens the ring structure and exchanges segments of maltodextrin linear chains. Since the discovery of the CGTase from *Bacillus macerans* in 1903, the production of this enzyme has been studied in several bacterial lineages, such as *Bacillus elaterium*, *Bacillus macerans*, *Klebsiella pneumoniae*, and *Bacillus stearothermophilus* (see [87] and [44]).

*Bacillus* species are the major producers of industrial enzymes. Various applications (e.g., in detergents, pulp and paper processing industry) have prompted the isolation of strains from a variety of alkaline environments as source of enzymes with suitable activities. Enzymes capable of producing predominantly a particular type of CD can decrease the costs of downstream purification and hence are commercially valuable [30].

## 4.7 ASPECTS TO EVALUATE

The proposed approach was evaluated on the chosen molecular complexes by considering the following aspects:

- The behavior of the Pareto fronts at a different number of evaluations in order to monitor the minimization process of both objectives (bond and non-bond terms).
- The  $\text{RMSD}_1$  values of the predicted molecular complexes in different generations of the genetic algorithm. This was meant to provide a general idea about the algorithm convergence to molecular complexes similar to the ones that are experimentally available.
- Comparison of  $\text{RMSD}_1$  and  $\text{RMSD}_3$  values along the iterative process. This gives an idea about the convergence to locations and conformations of the ligand molecules.
- Frequency histograms of RMSD values at different generations of the genetic algorithm as a measure of the dispersion and convergence of the proposed method.

# Chapter 5

## RESULTS AND DISCUSSION

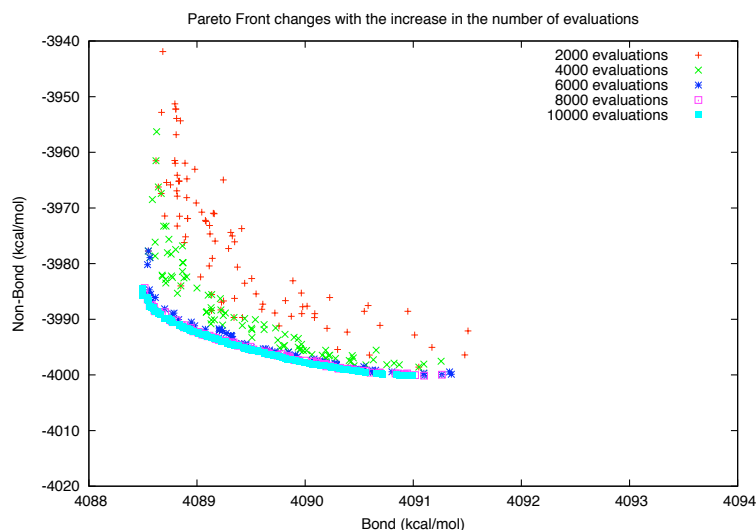
### 5.1 1ABE

The search is guided by the energy values to minimize the bond and non-bond terms. The behavior of this minimization process can be observed in Figure 5.1 showing the convergence of the non-dominant solutions to minimum energy values. In the specific case of the 1ABE complex, the values obtained at 10,000 evaluations are highlighted in blue. The figure shows the convergence of both the bond energy objective and the non-bond objective, which has a higher dispersion in the early evaluations of the algorithm.

The dispersion of the solutions and their subsequent convergence to precise conformations and locations in the binding site of the experimentally obtained complexes are measured by the  $\text{RMSD}_1$  of the structure superposition, and the correlation between the energy terms at different number of evaluations. This is shown in Figure 5.2. Notice that as the number of evaluations increases, RMSD values tend to zero at low energy values.

Figure 5.3 shows the minimum, maximum and average  $\text{RMSD}_1$  and  $\text{RMSD}_3$  values after a certain number of evaluations. The average RMSD decreases significantly as the number of evaluations increases, until finally stabilizing near zero. As shown in Figure 5.3, the RMSD remains constant after several evaluations; however, the

Figure 5.1: **Complex 1ABE. Changes in the Pareto Front as the number of evaluations increases.**



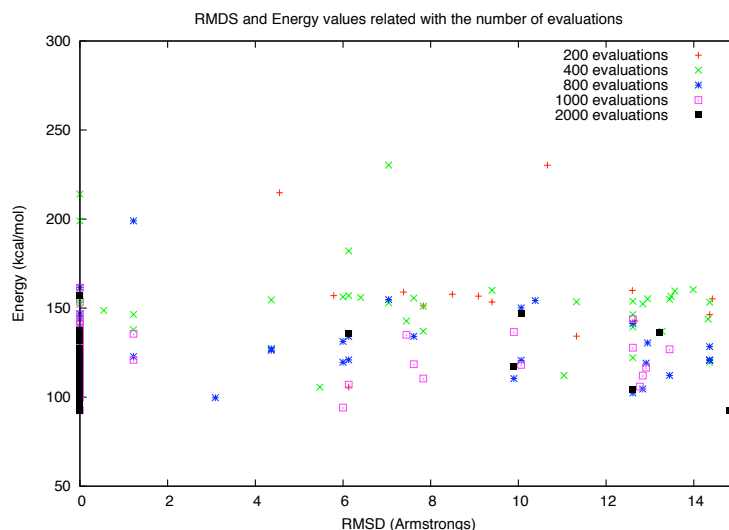
number of solutions in this RMSD range decreases with the time and number of generations, as shown in Figure 5.4.

Table 5.1 shows the  $\text{RMSD}_1$  values of the solutions selected by the decision maker. In general, the best solutions were found by the method based on utility for the 1ABE complex. Figure 5.5 shows some of the results, from which it can be observed that the best location for the ligand was in the binding site because the rest of the space was occupied by the receptor and so collisions between the atoms would otherwise had occurred.

## 5.2 1ACM

The changes in the Pareto front due to the minimization of both objectives are shown in Figure 5.6. Same as with the 1ABE complex, a faster convergence occurred in the objective related to the bond terms. In this sense, energy values for the first evaluations ranged between 4,630 and 4,638 (Kcal/mol) and such variation remained

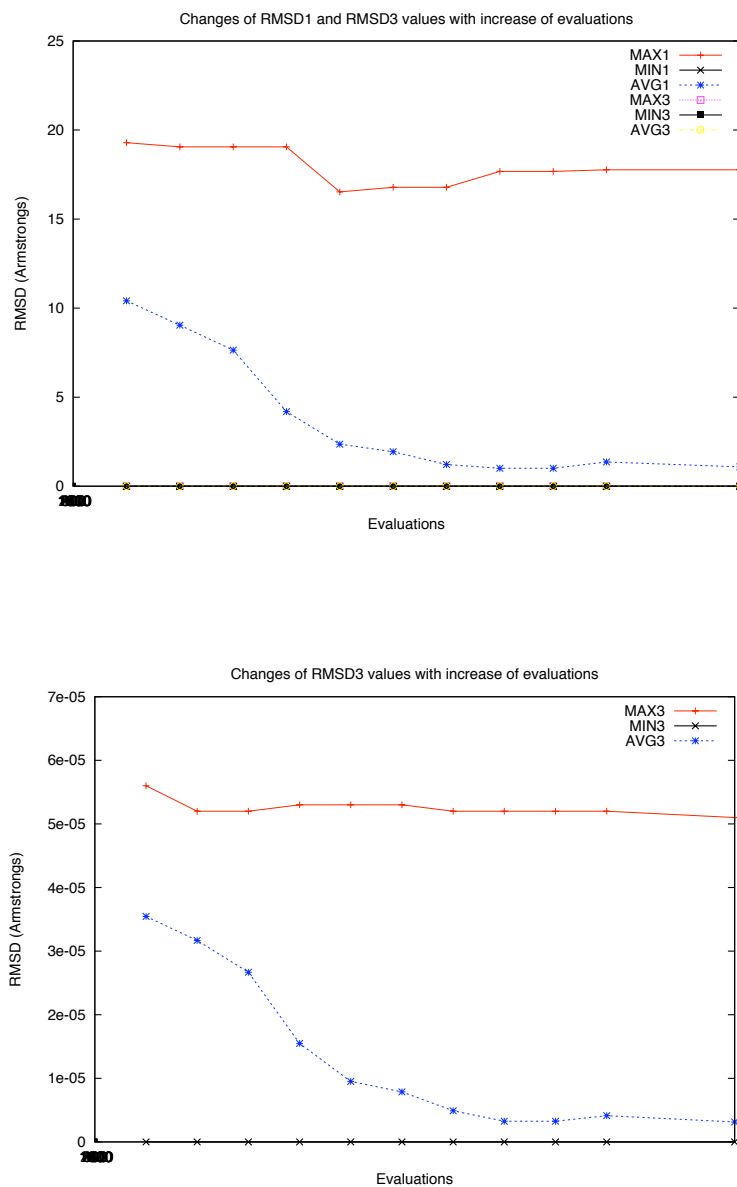
Figure 5.2: **Complex 1ABE. RMDS and energy values according to the number of evaluations.**



stable for the last evaluation (between 4,630 to 4,634 Kcal/mol), which means that the energy value was optimized with less than 20,000 evaluations. In contrast, non-bond terms showed a larger range of variation in the first evaluations (-9,100 and -8,200 (Kcal/mol), which decreased considerably for the last evaluation (-9100 to -8800 Kcal/mol). In conclusion, the convergence was faster for bond terms than for non-bond terms. .

In spite of the apparent convergence of the Pareto front after 20,000 evaluations, it is possible to observe in Figure 5.7 that two groups of individuals are formed at 20,000 evaluations, which are shown as black points between RMSD values of 7–10 Å and 12–15 Å, and energy values around -4,400 and -4,200 Kcal/mol. In contrast, the desired behavior would be a simultaneous decreasing of RMSD and energy values, in which a group of complexes with low RMSD and energy values is to be obtained at the end of the process. The difference between the desired results and the results obtained with the model could be related to the gap in the energy values between the reported structure and the structures yielded by the model.

Figure 5.3: Complex 1ABE. Changes in  $\text{RMSD}_1$  and  $\text{RMSD}_3$  values as the number of evaluations increases.



The energy value reported for 1ACM is -4,665.67 Kcal/mol, contrary to the range obtained in the optimization process (see Figure 5.7). Therefore, the algorithm needs

Figure 5.4: Complex 1ABE. Dispersion in RMSD values according to the number of evaluations.

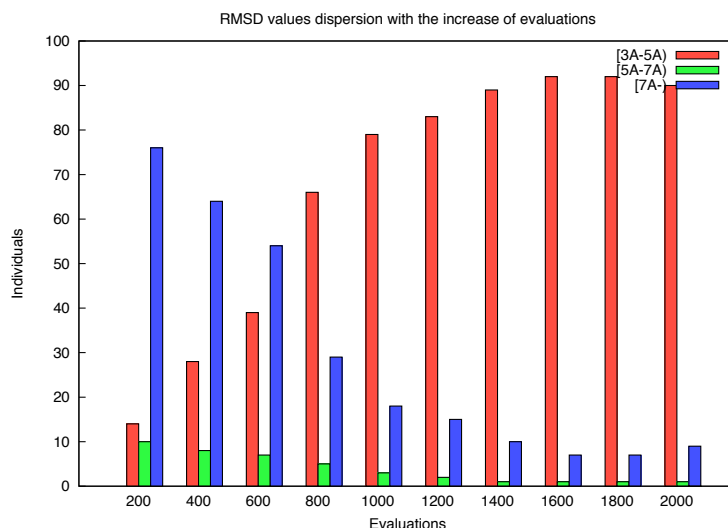
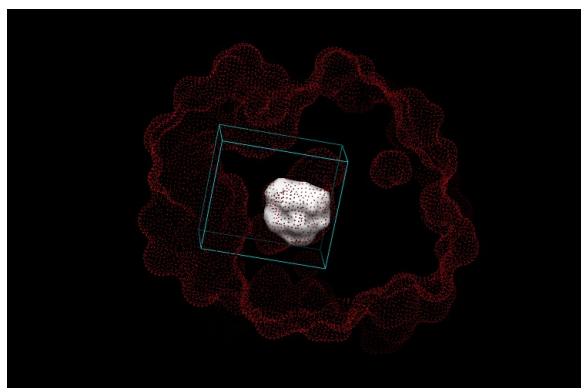


Figure 5.5: Solutions established by the proposed method for the 1ABE complex.



The receptor surface is represented by the red points and the ligand surface is presented as a solid white body. The defined binding site is represented by the blue box limiting the area where the ligand could be located.

a higher number of evaluations than those proposed initially to reach the ideal value of energy, which is clearly lower than the one achieved at 20,000 evaluations.

Table 5.1: Complex 1ABE. Decision Maker selection at different number of evaluations of the method.

Evaluations	Decision Maker	RMSD <sub>1</sub> (Å)
200	Utility	6.1296
	Angles	16.1812
500	Utility	0.0000
	Angles	0.0000
1,000	Utility	0.0000
	Angles	0.0000
1,500	Utility	0.0000
	Angles	13.4469
2,000	Utility	14.8077
	Angles	0.0000

Figure 5.6: Complex 1ACM. Changes in the Pareto Front as the number of evaluations increases.

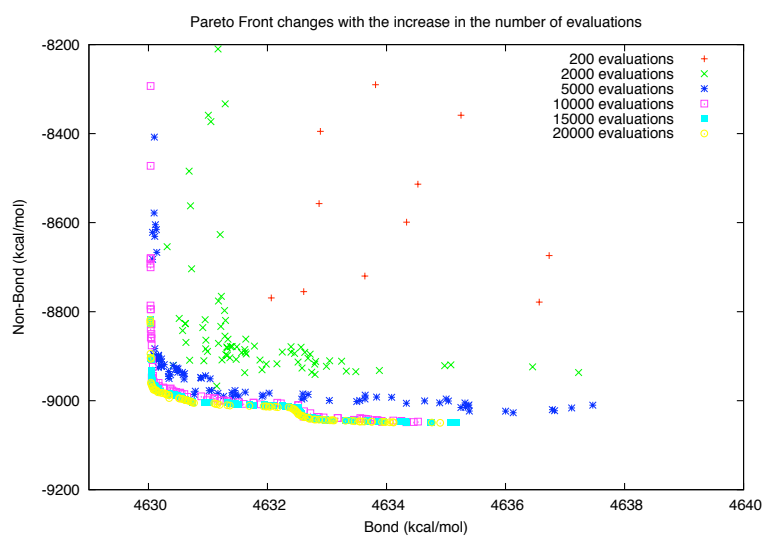
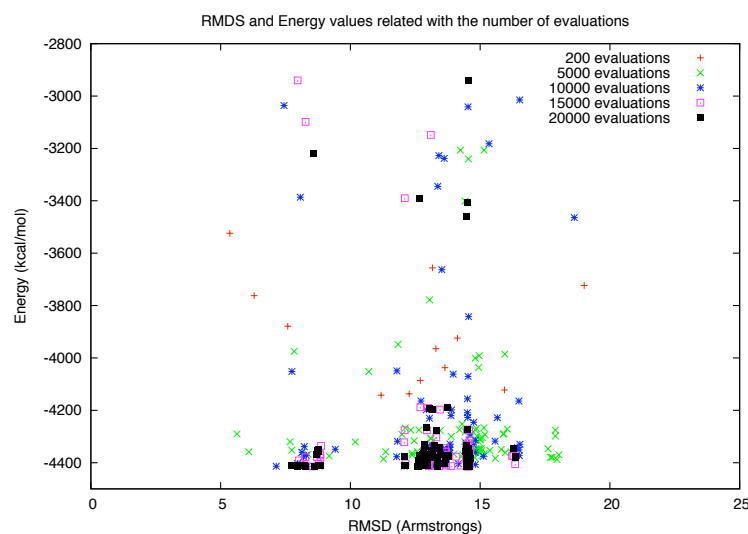


Figure 5.8 shows that some solutions can get into the binding site but the atoms may collide, therefore they are penalized with high energy values and the ligands are located outside the surface where energy values are better. Therefore, it is possible

that the complex needs of additional constraints in its geometric parameters related to the definition of the binding site, or modifications in the parameters of the scoring function that makes it possible for the ligand to get inside closed cavities despite colliding with the protein. This task would be approach in a future work.

Figure 5.7: **Complex 1ACM. RMDS and Energy values according to the number of evaluations.**



The apparent convergence to high RMSD values when the molecule has not reached its better conformation, as well as locations expressed as low energy values, could produce the effect observed in Figures 5.9 and 5.10, where the molecule has found an apparently stable state, but such stability is not necessarily true, as mentioned above.

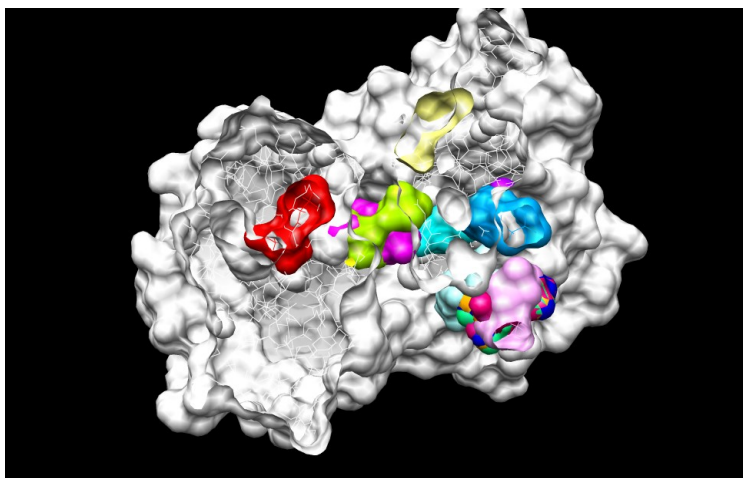
The solutions yielded by the model are not the best ones from an energetic point of view, as evidenced by comparing RMSD values to those obtained with programs such as GOLD<sup>1</sup> (1.23 Å) and DOCK<sup>2</sup> (1.11 Å) .

The decision maker was not considered in this example because all the molecules

<sup>1</sup>[http://www.ccdc.cam.ac.uk/products/life\\_sciences/gold/validation/original\\_gold\\_test\\_set/](http://www.ccdc.cam.ac.uk/products/life_sciences/gold/validation/original_gold_test_set/)

<sup>2</sup>[http://dock.compbio.ucsf.edu/Test\\_Sets/v5.4\\_protein\\_summary.html](http://dock.compbio.ucsf.edu/Test_Sets/v5.4_protein_summary.html)

Figure 5.8: Solutions established by the proposed method for the 1ACM complex.



The white surface corresponds to the receptor, the reported ligand is represented by the pink molecular surface in the middle of the receptor, and the other colored surfaces to solutions found by the model.

Figure 5.9: Complex 1ACM. Changes in RMSD1 and RMSD3 values as the number of evaluations increases.

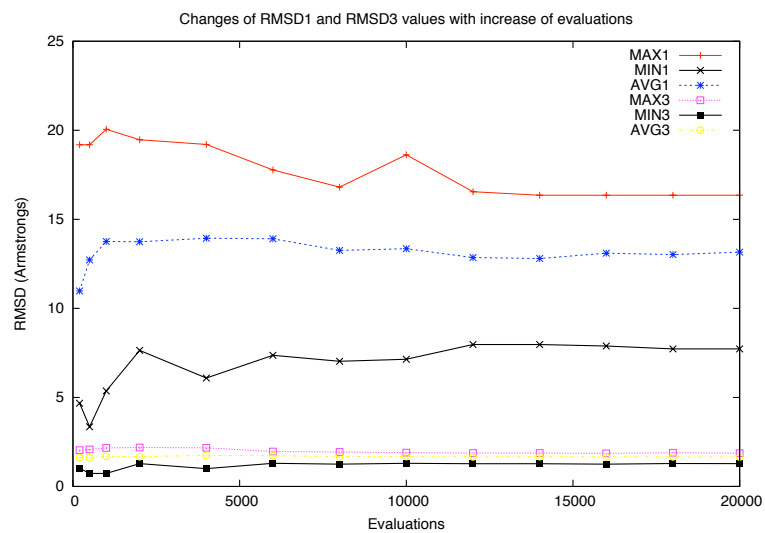
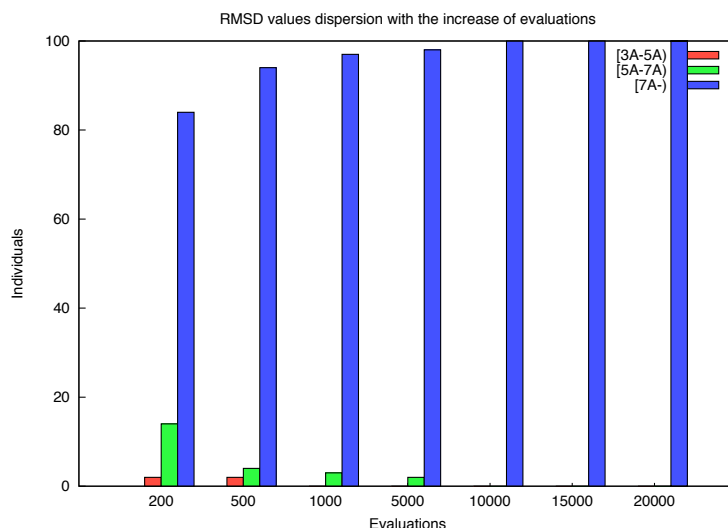


Figure 5.10: 1ACM Complex. Dispersion in RMSD values according to the increase in the number of evaluations.



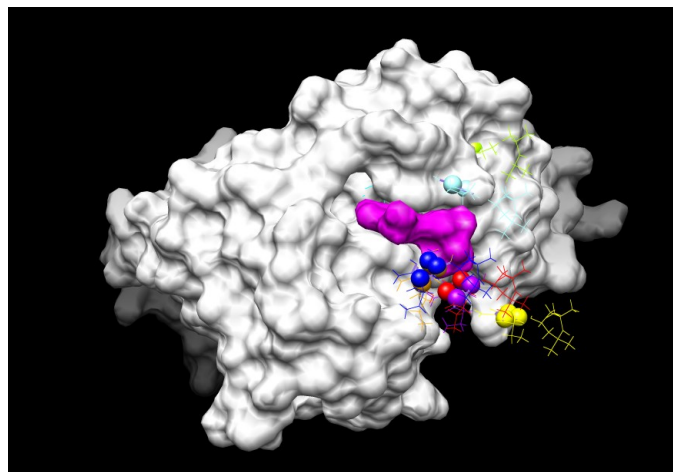
were located outside the protein cavity. The chemical environment on the site where the ligands were located by the proposed method is completely different from the real binding site, which affects the objective related to the non-bond terms as well as the decision maker, as this latter works with information on Pareto front and therefore would not be useful to select the best molecule.

### 5.3 1BAF

The behavior of the 1BAF complex was similar to the one of the 1ACM complex. The Pareto front converged to low energy values, as shown in Figure 5.12; however, there was not much correlation between the  $\text{RMSD}_1$  and the energy (see Figure 5.13). This shows that the number of evaluations was probably insufficient and, similar to the case of the 1ACM complex, several atom collisions could have occurred in the binding site where the molecule is supposed to accommodate, as observed in Figure 5.11, thus reducing the probability of finding molecules close to the binding site, as

it would be locating the ligands on the protein's surface regions with low energy interaction, but far away from the binding site.

Figure 5.11: **Solution established by the proposed method for the 1BAF complex.**



The white surface corresponds to the receptor and the magenta surface to the ligand. Other ligand conformations and locations are represented by atoms and sticks. It is possible to find molecules near the surface and the binding site, but not inside the binding site. Same as in the 1ACM complex, there are collisions between ligand and protein atoms due to the disposition of the binding cavity.

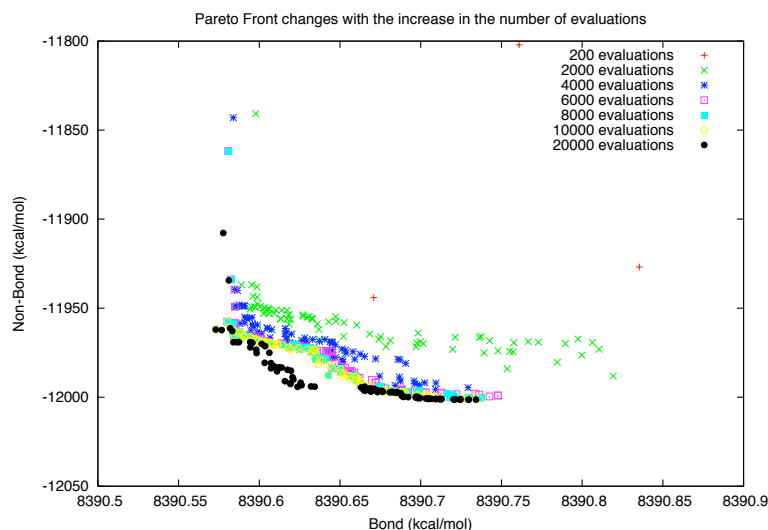
The apparent convergence to high RMSD values when the molecule has not reached its best conformation, as well as low energy locations, could produce the effect observed in Figures 5.14 and 5.15, where the molecule has reached an apparently stable state but has not necessarily reached the best one, same as it occurred with the 1ACM complex.

## 5.4 1CDG

As mentioned in the previous complexes, the Pareto front changes with the number of evaluations to search for the minimum energy values, as shown in Figure 5.16.

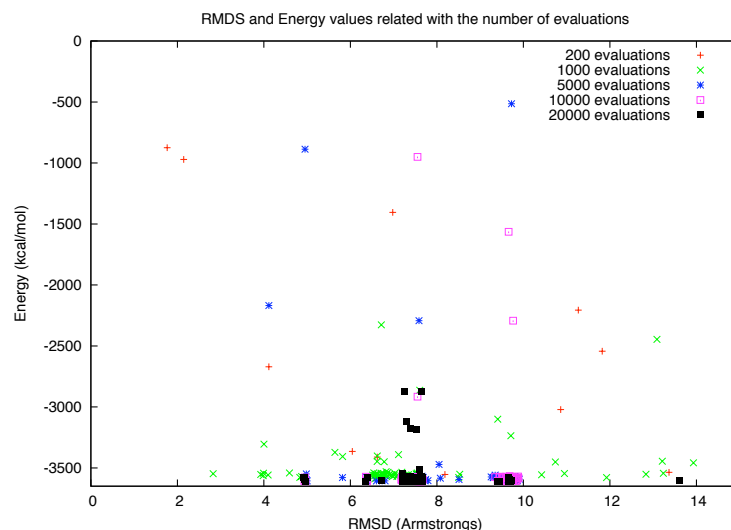
In this experiment, the convergence to bond energy terms occurred faster than for non-bond terms, but less faster than for the previous ligands because the number

Figure 5.12: Complex 1BAF. Changes in the Pareto Front as the number of evaluations increases.



of rotatable bonds was 12, which is considerably higher than the number of rotatable bonds in other molecules shown in Table ???. The high number of rotatable bonds makes it difficult to predict the molecules location and conformation in the binding site for most of the available methods. However, as seen in Figure 5.17, the proposed method showed a good performance for predicting both the location and conformation reported for the 1CDG ligand, which has a large number of rotatable bonds. Such molecules are represented by a group of low energy solutions and low RMSD values observed in Figure 5.17. Besides, Figure 5.18 shows that the cavity in this protein, unlike in 1ACM and 1BAF, does not produce collisions between atoms of the molecules, making it easier and precise to predict the complex. The decrease in RMSD values depicted in Figure 5.19 shows that average values decrease significantly as the number of evaluations becomes larger, until reaching values close to zero. Although the maximum RMSD values remains constant, the number of solutions with lower RMSD values increases, as it can be observed in Figure 5.20. The ability of the algorithm to find adequate conformations and locations for ligand molecules with a considerable number of rotatable bonds is specially attractive and

Figure 5.13: Complex BAF. RMDS and Energy values according to the number of evaluations.



useful for predictions about protein–ligand complexes, especially when dealing with ligands that cannot be easily handled by using other approaches.

Table 5.2 shows the  $\text{RMSD}_1$  values for the solutions chosen by the implemented decision maker algorithm for different number of evaluations. The table allows to compare which of the criteria (angles or utility) is the most appropriated one to select the complex that will be reported to the user. In this specific case, the best results were selected by the approach that takes into account the angles. This means that the lowest RMSD structure, at different number of evaluations, would be selected having into account this criteria; however, more experiments with different complexes are needed in order to justify whether to use angles or utility as the only criteria to choose the final solution.

Figure 5.14: Complex BAF. Changes in RMSD1 and RMSD3 values as the number of evaluations increases.

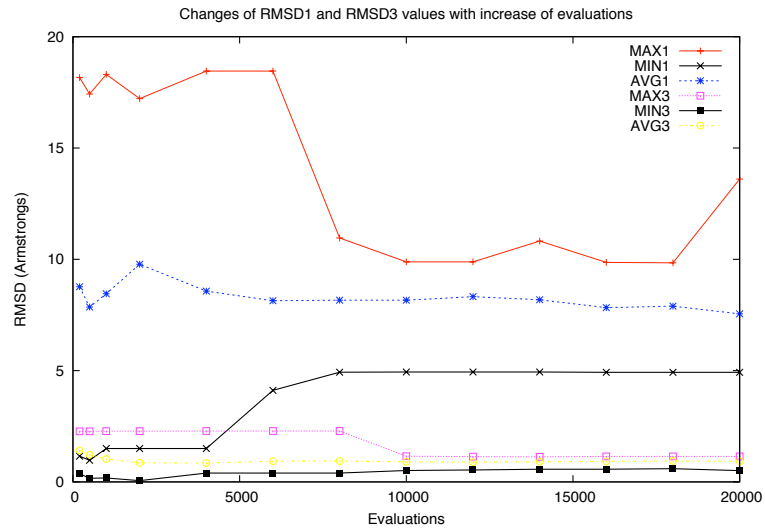


Figure 5.15: Complex BAF. Dispersion in RMSD values according to the increase in the number of evaluations.

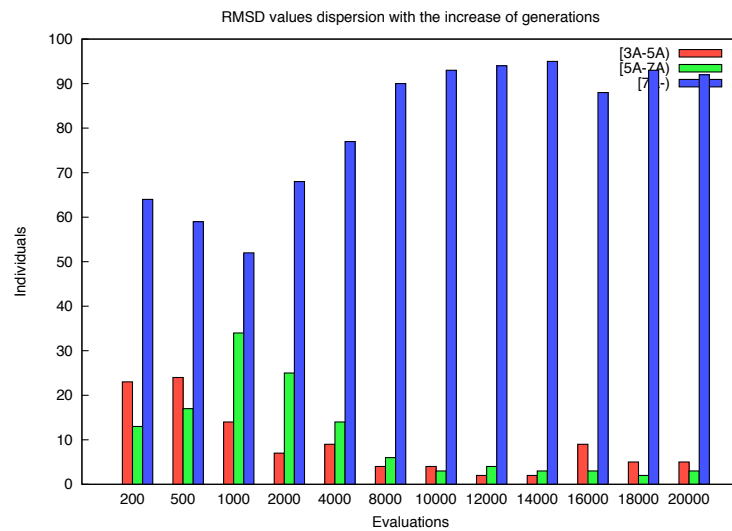


Figure 5.16: Complex CDG. Changes in the Pareto Front as the number of evaluations increases.

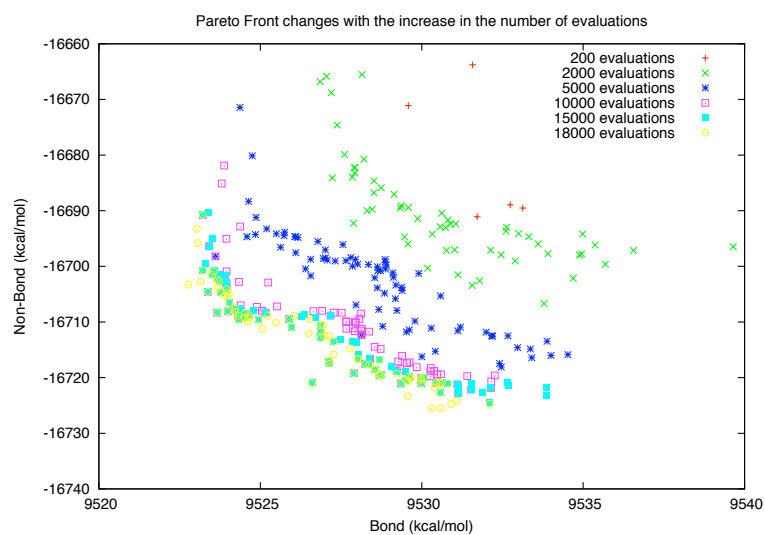


Figure 5.17: Complex CDG. RMDS and Energy values according to the number of evaluations.

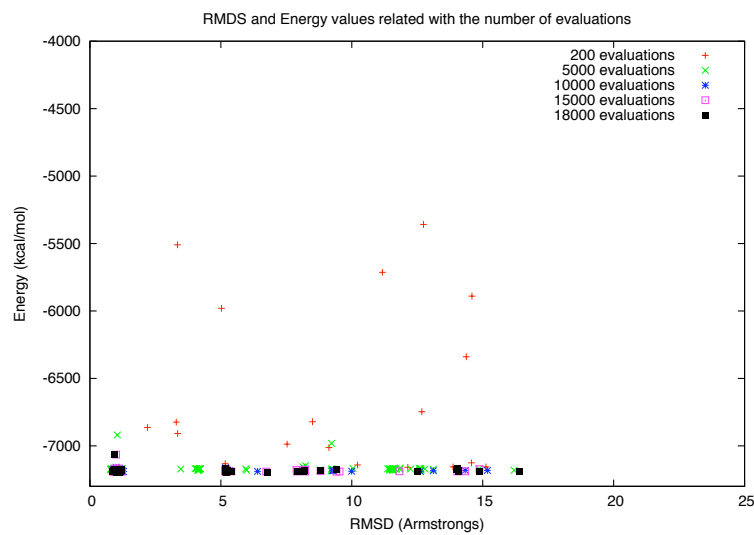
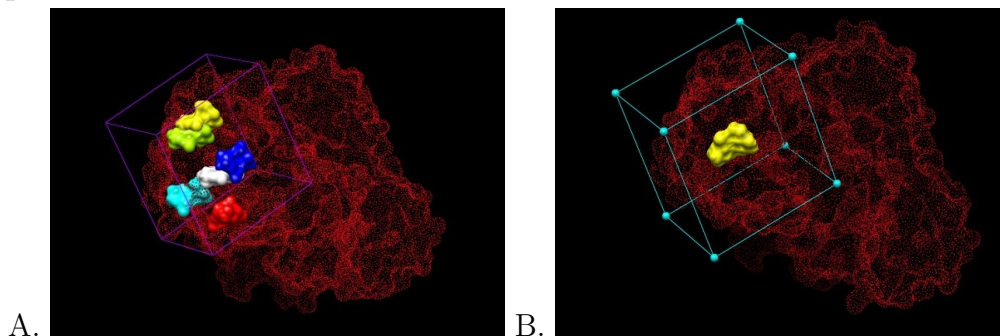


Figure 5.18: Solutions established by the proposed method for the 1CDG complex.



Panel A shows the possible solutions that were found by the algorithm. Panel B presents the complex reported in PDB.

Figure 5.19: Complex CDG. Changes in  $\text{RMSD}_1$  and  $\text{RMSD}_3$  values as the number of evaluations increases

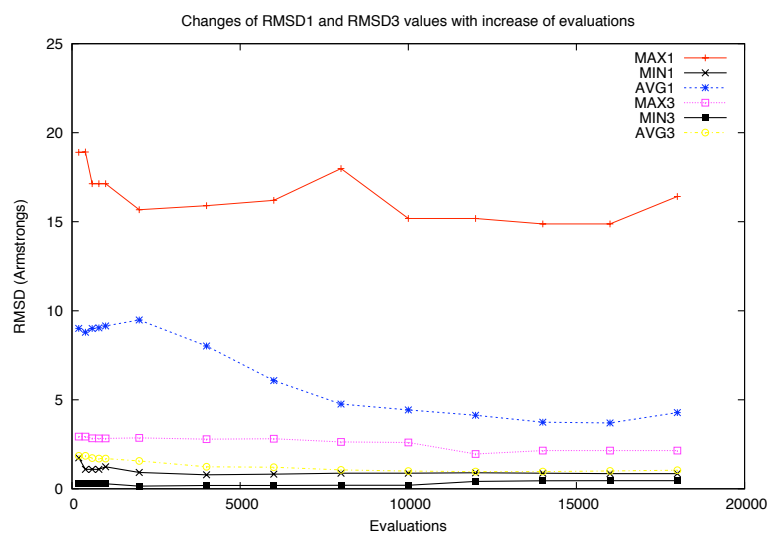


Figure 5.20: Complex CDG. Dispersion in RMSD values according to the increase in the number of evaluations.

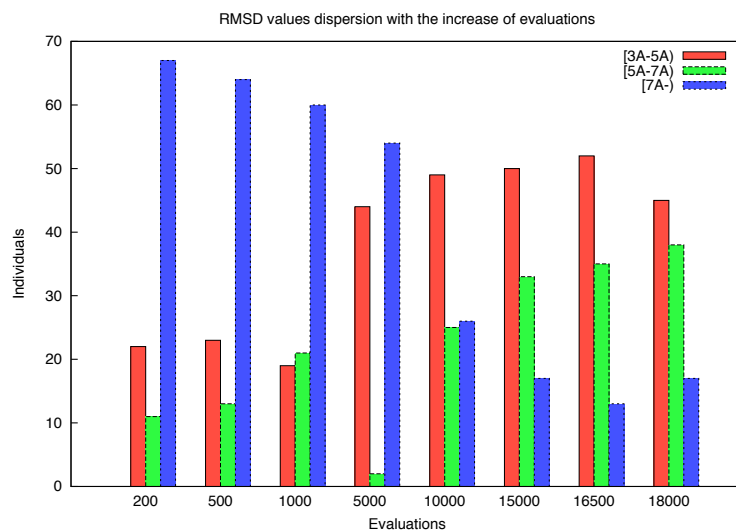


Table 5.2: 1CDG Complex. Decision Maker selection at different number of evaluations of the method.

Evaluations	Decision Maker	RMSD <sub>1</sub> (Å)
200	Utility	12.6537
	Angles	11.1741
1,000	Utility	13.7963
	Angles	7.7924
5,000	Utility	1.0764
	Angles	4.1159
10,000	Utility	5.2105
	Angles	1.2507
14,000	Utility	1.2241
	Angles	5.1728
18,000	Utility	6.7644
	Angles	1.0861

# Chapter 6

## CONCLUSIONS AND PERSPECTIVES

This research contributes to the solution of docking problem by introducing a novel method for predicting molecular complexes based on a multi-objective approach at an atomic conformational level.

In fact, this thesis proposes a methodology to tackle the docking problem. It can be seen as a valuable bioinformatical tool for predicting the formation of molecular complexes between proteins and ligands. It is important to notice that the implementation of this methodology was done using some publicly available third-party software tools as well as some software tools developed as part of this thesis. The latter software tools allow the exploration of conformational spaces of small organic molecules as well as of different energy functions (such as CHARMM19 and 27, AMBER 94, 96, 98 and 99) and different MOEA algorithms (such as NSGA-II, SPEA2, PAES, PESAI and IBEA) to evolve different protein conformations.

After analyzing the experimental framework, it is clear that the proposed method is good in specific cases, but to make it suitable for many kind of complexes, further work would be necessary in order to consider the possibility of introducing heuristic information to restrict the search space and enhance the efficacy and efficiency of the algorithm. With respect to this method it can be concluded that:

The method proposed for representing small organic molecules was adequate for the search of the conformational space because it reduced the number of variables that needed to be managed by the evolutionary algorithm. The subsequent alternation

from strings of characters to 3D representations in an efficient way (being able to obtain molecules in different formats such as xyz, Sybyl or pdb) allowed evaluating the complexes satisfactorily with the scoring function. Changes in the molecular representations were possible by the development and implementation of an algorithm during the development of this work, which allowed generating any molecular conformation. The algorithm to alternate between the molecular representations was of crucial importance for the different stages of the docking problem.

The scoring function measures the energy of a complex once the binding between protein and ligand has been established. The approach proposed for small organic molecules was appropriate for this work since the scoring function found conformations and locations that were stable, although in some specific cases such conformations and localization were not ideal. It is important to emphasize that the correlation between structures and energy was measured as a relationship between RMSD errors (in Angstroms) and energy (Kcal/mol). Nevertheless, as explained in Chapter 4, the correlation between both is not clear, specially when the binding site is a cavity where several collisions between protein and ligand atoms could occur. This led to the conclusion that further work is needed to consider additional scoring functions that evaluate the solutions and penalize these collisions in a soft way and/or the inclusion of geometric constraints to avoid ignoring possible good locations by allowing the exploration of cavities that are not accessible to the solvent or whose shapes and sizes produce several collisions between both the ligand and the receptor.

A multi-objective evolutionary algorithm enable the method to find molecular complexes with 3D structures relatively close to the ones reported for the benchmark structures. The ability of finding good complexes when the ligand has a high number of rotatable bonds is of remarkable importance and interest in this kind of approaches, considering that some of the other available methods have problems predicting complexes when a ligand with high number of rotatable bonds is considered. The proposed method found non-dominated solutions near to or inside the binding site.

Since the objectives to be optimized by the MOEA were based on non-bonding and bonding interactions, and taking into account that some of the best results were

found for molecules with a significantly large number of rotatable bonds, it could be concluded that this kind of interactions are suitable to model the formation of complexes when the ligands are molecules that can take different configurations due to the influence of non-bonding forces.

This work presented a novel method to explore the formation of protein–ligand complexes, which has great impact in the general understanding of the mechanisms underlying molecular interactions. It would be especially useful in the discovery and selection of molecules with possible biological activity in drug development based on *in silico* techniques. However, additional work is necessary in order to enhance the performance of the proposed method.

Future work to extend and improve this study may include: enhancing the method’s ability to predict molecular complexes, evaluating and validating the proposed method, and exploring different areas where the method could be applied. The proposed method could be improved by introducing additional aspects such as additional geometric constraints, other energy functions and different multi-objective evolutionary algorithms.

Due to the high computational cost of the proposed method, it would be useful to explore the use of some parallelization techniques in order to reduce the time required to obtain the molecular predictions.

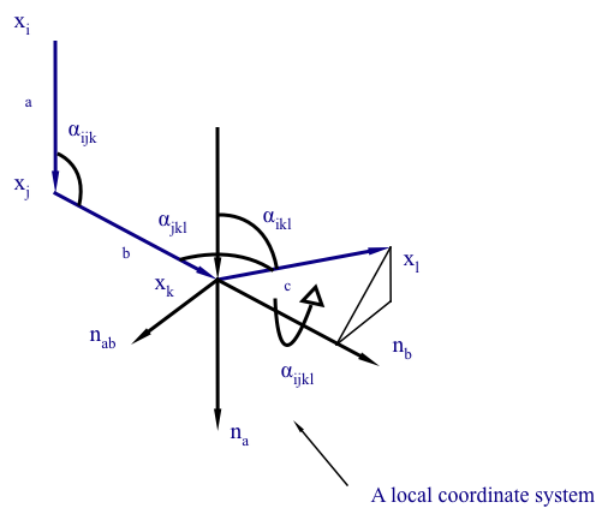
Also, other compounds such as the solvent and other biological ligands can be included in the model in order to make the simulation process closer to reality.

Finally, the ultimate goal would be to include the proposed method in a drug design cycle in order to evaluate its performance for evaluating drug candidates.

# Chapter 7

## APPENDIX

ANNEXE 1: Calculation of atomic coordinates based on distances, bond angles and the dihedral angles.



If the atoms correspond to:

$$x_i = (u_i, v_i, w_i)$$

$$x_j = (u_j, v_j, w_j)$$

$$x_k = (u_k, v_k, w_k)$$

$$x_l = ?$$

**And  $x_l$  can be seen as:**

$$x_l = x_k + c \text{ where } c = c_1 + c_2 + c_3$$

$$c_1 = -\|c\| \cos \alpha_{ikl} n_a$$

$$c_2 = -\|c\| \cos \alpha_{jkl} n_b$$

$$c_3 = \|c\| \sin \alpha_{jkl} \sin \alpha_{ijk} n_{ab}$$

**$n_a$ ,  $n_b$  and  $n_{ab}$  are vectors which define a new coordinate system based on the known vectors, which are found according to the following expressions:**

$$n_a = a / \|a\|$$

$$n_b = b / \|b\|$$

$$n_{ab} = a \times b / (\|a\| \|b\| \sin \alpha_{ijk})$$

# Bibliography

- [1] R. Abagyan and M. Totrov. Biased probability monte carlo conformational searches and electrostatic calculations for peptides and proteins. *J Mol Biol*, 235(3):983–1002, January 1994.
- [2] Ruben Abagyan, Maxim Totrov, and Dmitry Kuznetsov. Icm: A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *Journal of Computational Chemistry*, 15(5):488–506, September 2004.
- [3] Laleh Alisaraie, Lars A. Haller, and Gregor Fels. A qxp-based multistep docking procedure for accurate prediction of protein-ligand complexes. *Journal of Chemical Information and Modeling*, 46(3):1174–1187, 2006. PMID: 16711737.
- [4] J. An, M. Totrov, and R. Abagyan. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol Cell Proteomics*, 4(6):752–761, June 2005.
- [5] PR Andrews, JM Carson, A Caselli, MJ Spark, and R Woods. Conformational analysis and active site modelling of angiotensin-converting enzyme inhibitors. *Journal of medicinal chemistry*, 28(3):393–399, 1985.
- [6] John H. Ansede and Dhiren R. Thakker. High-throughput screening for stability and inhibitory activity of compounds toward cytochrome p450-mediated metabolism. *Journal of Pharmaceutical Sciences*, 93(2):239–255, 2004.
- [7] A. H. Beckett and A.F. Casey. Synthetic analgesics: stereochemical considerations. *J. Pharm. Pharmacol.*, 6(12):986–1001, 1954.

- [8] N Beerenwinkel, T Sing, T Lengauer, J Rahnenführer, K Roomp, I Savenkov, R Fischer, D Hoffmann, J Selbig, K Korn, H Walter, T Berg, P Braun, G Fätkenheuer, M Oette, J Rockstroh, B Kupfer, R Kaiser, and M Däumer. Bioinformatics; computational methods for the design of effective therapies against drug resistant hiv strains. *21(21):3943–3950*, 2005.
- [9] Andrey A. Bliznyuk and Jill E. Gready. Simple method for locating possible ligand binding sites on protein surfaces. *Journal of Computational Chemistry*, 20(9):983–988, 1999.
- [10] H. J. Bohm. Ludi: rule-based automatic design of new substituents for enzyme inhibitor leads. *Journal of computer-aided molecular design*, 6(6):593–606, December 1992.
- [11] P. Bonnabry, J. Sievering, T. Leemann, and P. Dayer. Quantitative drug interactions prediction system (q-dips): a computer-based prediction and management support system for drug metabolism interactions. *European Journal of Clinical Pharmacology*, 55:341–347, 1999.
- [12] Jürgen Branke, Kalyanmoy Deb, Henning Dierolf, and Matthias Osswald. Finding knees in multi-objective optimization. In *In the Eighth Conference on Parallel Problem Solving from Nature (PPSN VIII). Lecture Notes in Computer Science*, volume 3242, pages 722–731, 2004.
- [13] Natasja Brooijmans and Irwin D. Kuntz. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.*, 32:335–373, 2003.
- [14] Bernard R. Brooks, Robert E. Bruccoleri, Barry D. Olafson, David J. States, S. Swaminathan, and Martin Karplus. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2):187–217, FebruaryFebruary 1983.
- [15] Nathan Brown. Chemoinformatics—an introduction for computer scientists. *ACM Comput. Surv.*, 41(2):38, 2009.
- [16] A. T. Brünger, D. J. Leahy, T. R. Hynes, and R. O. Fox. 2.9 a resolution structure of an anti-dinitrophenyl-spin-label monoclonal antibody fab fragment

- with bound hapten. *Journal of molecular biology*, 221(1):239–256, September 1991.
- [17] Olson Edward C. and Christoffersen Ralph E., editors. *Computer-Assisted Drug Design*. AMERICAN CHEMICAL SOCIETY, WASHINGTON, D. C., 11 1979.
- [18] Stephen Campbell, Nicola Gold, Richard Jackson, and David Westhead. Ligand binding: functional site location, similarity and docking. *Current Opinion in Structural Biology*, 13(3):389–395, 2003.
- [19] D.A. Case, T.A. Darden, T.E. Cheatham, J. Wang C.L. Simmerling, R.E. Duke, R. Luo, M. Crowley, Ross C. Walker, W. Zhang, K.M. Merz, B.Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossváry, K.F.Wong, F. Paesani, J. Vanicek, X.Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, D.H. Mathews, M.G. Seetin, C. Sagui, V. Babin, and P.A. Kollman. Amber 10. *University of California, San Francisco*, 2008.
- [20] Wendy D. Cornell, Piotr Cieplak, Christopher I. Bayly, Ian R. Gould, Kenneth M. Merz, David M. Ferguson, David C. Spellmeyer, Thomas Fox, James W. Caldwell, and Peter A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, May 1995.
- [21] Kalyanmoy Deb, Samir Agrawal, Amrit Pratap, and T. Meyarivan. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: Nsga-ii, 2000.
- [22] Kalyanmoy Deb and Deb Kalyanmoy. *Multi-Objective Optimization Using Evolutionary Algorithms*. Wiley, 1 edition, June 2001.
- [23] S. Dennis, T. Kortvelyesi, and S. Vajda. Computational mapping identifies the binding sites of organic solvents on proteins. *Proc Natl Acad Sci U S A*, 99(7):4290–4295, April 2002.
- [24] R. L. DesJarlais, R. P. Sheridan, J. S. Dixon, I. D. Kuntz, and R. Venkataraghavan. Docking flexible ligands to macromolecular receptors by molecular shape. *J Med Chem*, 29(11):2149–2153, November 1986.

- [25] Qunfeng Dong and Zhijun Wu. A linear-time algorithm for solving the molecular distance geometry problem with exact inter-atomic distances. *Journal of Global Optimization*, 22(1):365–375, 01 2002.
- [26] Sean Ekins. Predicting undesirable drug interactions with promiscuous proteins in silico. *Drug Discovery Today*, 9(4):276–285, 2004.
- [27] T. J. A. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz. Dock 4.0: Search strategies for automated molecular docking of flexible molecule databases. *Journal of Computer-Aided Molecular Design*, pages 411–428, May 2001.
- [28] Richard A. Friesner, Jay L. Banks, Robert B. Murphy, Thomas A. Halgren, Jasna J. Klicic, Daniel T. Mainz, Matthew P. Repasky, Eric H. Knoll, Mee Shelley, Jason K. Perry, David E. Shaw, Perry Francis, and Peter S. Shenkin. Glide: A new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal of Medicinal Chemistry*, 47(7):1739–1749, March 2004.
- [29] Johann Gasteiger and Mario Marsili. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron*, 36(22):3219 – 3228, 1980.
- [30] B. N. Gawande and A. Y. Patkar. Purification and properties of a novel raw starch degrading-cyclodextrin glycosyltransferase from klebsiella pneumoniae as- 22. *Enzyme and microbial technology*, 28(9-10):735–743, June 2001.
- [31] Valerie J. Gillet, Peter Willett, and John Bradshaw. The effectiveness of reactant pools for generating structurally-diverse combinatorial libraries. *Journal of Chemical Information and Computer Sciences*, 37(4):731–740, July 1997.
- [32] Meir Glick, Daniel D. Robinson, Guy H. Grant, and W. Graham Richards. Identification of ligand binding sites on proteins using a multi-scale approach. *Journal of the American Chemical Society*, 124(10):2337–2344, March 2002.
- [33] H. Gohlke, M. Hendlich, and G. Klebe. Knowledge-based scoring function to predict protein-ligand interactions. *Journal of molecular biology*, 295(2):337–356, January 2000.

- [34] P. J. Goodford. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.*, 28:849–857, 1985.
- [35] D. S. Goodsell and A. J. Olson. Automated docking of substrates to proteins by simulated annealing. *Proteins*, 8(3):195–202, 1990.
- [36] Olgun Guvench and Alexander D. MacKerell. Computational fragment-based binding site identification by ligand competitive saturation. *PLoS computational biology*, 5(7):e1000435+, July 2009.
- [37] Julia Handl, Douglas B. Kell, and Joshua Knowles. Multiobjective optimization in bioinformatics and computational biology. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, 4(2):279–292, April 2007.
- [38] Th. Hanser, Ph. Jauffret, and G. Kaufmann. A new algorithm for exhaustive ring perception in a molecular graph. *Journal of Chemical Information and Computer Sciences*, 36(6):1146–1152, 1996.
- [39] M. Hendlich, F. Rippmann, and G. Barnickel. Ligsite: automatic and efficient detection of potential small molecule-binding sites in proteins. *J Mol Graph Model*, 15(6), December 1997.
- [40] John H. Holland. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. The MIT Press, April 1992.
- [41] Bingding Huang and Michael Schroeder. Ligsitecsc: predicting ligand binding sites using the connolly surface and degree of conservation. *BMC Structural Biology*, 6(1):19, 2006.
- [42] Bingding Huang and Michael Schroeder. Ligsitecsc: predicting ligand binding sites using the connolly surface and degree of conservation. *BMC Structural Biology*, 6(1):19, 2006.

- [43] Araz Jakalian, Bruce L. Bush, David B. Jack, and Christopher I. Bayly. Fast, efficient generation of high-quality atomic charges. am1-bcc model: I. method. *Journal of Computational Chemistry*, 21(2):132–146, 2000.
- [44] R. Jamuna, N. Saswathi, R. Sheela, and S. V. Ramakrishna. Synthesis of cyclodextrin glucosyl transferase by bacillus cereus for the production of cyclodextrins. *Applied biochemistry and biotechnology*, 43(3):163–176, December 1993.
- [45] Lin Jiang, Ying Gao, Fenglou Mao, Zhijie Liu, and Luhua Lai. Proteins structure, function, and genetics. potential of mean force for protein-protein interaction studies. 46(2):190–196, 2002.
- [46] G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol*, 267(3):727–748, April 1997.
- [47] William L. Jorgensen and Corky Jenson. Temperature dependence of tip3p, spc, and tip4p water from npt monte carlo simulations: Seeking temperatures of maximum density. *Journal of Computational Chemistry*, 19(10):1179–1186, 1998.
- [48] William L. Jorgensen and Julianto Pranata. Importance of secondary interactions in triply hydrogen bonded complexes: guanine-cytosine vs uracil-2,6-diaminopyridine. *Journal of the American Chemical Society*, 112(5):2008–2010, February 1990.
- [49] IM Kapetanovic. Chemico-biological interactions; computer-aided drug discovery and development (caddd): in silico-chemico-biological approach. 171(2):165–176, 2008.
- [50] L. B. Kier and H. S. Aldrich. A theoretical study of receptor site models for trimethylammonium group interaction. *Journal of theoretical biology*, 46(2):529–541, August 1974.

- [51] Tamas Kortvelyesi, Michael Silberstein, Sheldon Dennis, and Sandor Vajda. Improved mapping of protein binding sites. *Journal of Computer-Aided Molecular Design*, 17(2):173–186, February 2003.
- [52] Irwin D. Kuntz, Jeffrey M. Blaney, Stuart J. Oatley, Robert Langridge, and Thomas E. Ferrin. A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology*, 161(2):269–288, 1982.
- [53] R. Laskowski. Surfnet: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of Molecular Graphics*, 13(5):323–330, October 1995.
- [54] Andrew R. Leach and Irwin D. Kuntz. Conformational analysis of flexible ligands in macromolecular receptor sites. *J. Comput. Chem.*, 13(6):730–748, 1992.
- [55] D. G. Levitt and L. J. Banaszak. Pocket: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J Mol Graph*, 10(4):229–234, December 1992.
- [56] J. Liang, H. Edelsbrunner, and C. Woodward. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci*, 7(9):1884–1897, September 1998.
- [57] Harvey F. Lodish. *Molecular cell biology*. W.H. Freeman and Company, 5rev ed edition, August 2003.
- [58] Shingo Makino and Irwin D. Kuntz. Automated flexible ligand docking method and its application for database search. *Journal of Computational Chemistry*, 18(14):1812–1825, November 1997.
- [59] Garland R. Marshall. *Introduction to Chemoinformatics in Drug Discovery- A personal View In Chemoinformatics in Drug Discovery.*, volume 23, chapter 1, pages 1–19. WILEY-VCH Verlag GmbH and Co. KGaA, 2005.
- [60] G. Matioli, G. M. Zanin, and F. F. De Moraes. Characterization of cyclodextrin glycosyltransferase from bacillus firmus strain no. 37. *Applied biochemistry and biotechnology*, 91-93:643–654, 2001.

- [61] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, June 1953.
- [62] M. Y. Mizutani and A. Itai. Efficient method for high-throughput virtual screening based on flexible docking: discovery of novel acetylcholinesterase inhibitors. *J Med Chem*, 47(20):4818–4828, September 2004.
- [63] Miho Yamada Mizutani, Nobuo Tomioka, and Akiko Itai. Rational automatic search method for stable docking models of protein and ligand. *Journal of Molecular Biology*, 243(2):310 – 326, 1994.
- [64] Cristiane Moriwaki, Glauciane L. Costa, Rubia Pazzetto, Gisella M. Zanin, Flávio F. Moraes, Márcia Portilho, and Graciette Matioli. Production and characterization of a new cyclodextrin glycosyltransferase from bacillus firmus isolated from brazilian soil. *Process Biochemistry*, 42(10):1384 – 1390, 2007.
- [65] Garrett M. Morris, David S. Goodsell, Ruth Huey, and Arthur J. Olson. Distributed automated docking of flexible ligands to proteins: Parallel applications of autodock 2.4. *Journal of Computer-Aided Molecular Design*, 10(4):293–304, 08 1996.
- [66] Ingo Muegge. A knowledge-based scoring function for protein-ligand interactions: Probing the reference state. *Perspectives in Drug Discovery and Design*, 20(1):99–114, 12 2000.
- [67] Akira Nakamura, Keiko Haga, and Kunio Yamane. The transglycosylation reaction of cyclodextrin glucanotransferase is operated by a ping-pong mechanism. *FEBS Letters*, 337(1):66 – 70, 1994.
- [68] A. Oda, K. Tsuchida, T. Takakura, N. Yamaotsu, and S. Hirono. Comparison of consensus scoring strategies for evaluating computational models of protein-ligand complexes. *J. Chem. Inf. Model.*, 46(1):380–391, January 2006.
- [69] Klaus P. Peters, Jana Fauck, and Cornelius Frömmel. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J. Mol. Biol.*, 256:201–213, 1996.

- [70] F. A. Quioco and N. K. Vyas. Novel stereospecificity of the l-arabinose-binding protein. *Nature*, 310(5976):381–386, 1984.
- [71] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol*, 261(3):470–489, August 1996.
- [72] F. M. Richards. Areas, volumes, packing, and protein structure. *Annual Review of Biophysics and Bioengineering*, 6(1):151–176, 1977.
- [73] D. Ringe. Structure-aided drug design: crystallography and computational approaches. *J. Nucl. Med.*, 36(6 Suppl):28S–30S, 1995.
- [74] Dagmar Ringe and Carla Mattos. Analysis of the binding surfaces of proteins. *Medicinal Research Reviews*, 19(4):321–331, 1999.
- [75] David Camilo Becerra Romero. A parallel multi-objective ab-initio model for protein folding prediction at an atomic conformation level, November 2009.
- [76] Amin Rostami-Hodjegan and Geoff Tucker. 'in silico' simulations to assess the 'in vivo' consequences of 'in vitro' metabolic drug-drug interactions. *Drug Discovery Today: Technologies*, 1(4), 2004.
- [77] J. Ruppert, W. Welch, and A. N. Jain. Automatic identification and representation of protein binding sites for molecular docking. *Protein Sci*, 6(3):524–533, March 1997.
- [78] Bilha Sandak, Ruth Nussinov, and Haim J. Wolfson. A method for biomolecular structural recognition and docking allowing conformational flexibility, 1997.
- [79] Volker Schneck and Leslie Kuhn. Virtual screening with solvation and ligand-induced complementarity. *Perspectives in Drug Discovery and Design*, 20(1):171–190, 12 2000.
- [80] Volker Schneck, Craig A. Swanson, Elizabeth D. Getzoff, John A. Tainer, and Leslie A. Kuhn. Screening a peptidyl database for potential ligands to proteins with side-chain flexibility. *Proteins: Structure, Function, and Genetics*, 33(1):74–87, 1998.

- [81] Tanja Schulz-Gasch and Martin Stahl. Scoring functions for protein-ligand interactions: a critical perspective. *Drug Discovery Today: Technologies*, 1(3):231–239, 2004.
- [82] B. K. Shoichet, A. R. Leach, and I. D. Kuntz. Ligand solvation in molecular docking. *Proteins*, 34(1):4–16, January 1999.
- [83] A. S. Smellie, G. M. Crippen, and W. G. Richards. Fast drug-receptor mapping by site-directed distances: A novel method of predicting new pharmacological leads. *ChemInform*, 22(44):306, November 1991.
- [84] C. A. Sotriffer, H. Gohlke, and G. Klebe. Docking into knowledge-based potential fields: a comparative evaluation of drugscore. *J Med Chem*, 45(10):1967–1970, May 2002.
- [85] J. W. Stebbins, D. E. Robertson, M. F. Roberts, R. C. Stevens, W. N. Lipscomb, and E. R. Kantrowitz. Arginine 54 in the active site of escherichia coli aspartate transcarbamoylase is critical for catalysis: a site-specific mutagenesis, nmr, and x-ray crystallographic study. *Protein science : a publication of the Protein Society*, 1(11):1435–1446, November 1992.
- [86] Erk Subasi and Cagatay Basdogan. A new haptic interaction and visualization approach for rigid molecular docking in virtual environments. *Presence: Teleoper. Virtual Environ.*, 17(1):73–90, 2008.
- [87] J. Szejtli. *Cyclodextrin technology*. Dordrecht, (1988).
- [88] R. D. Taylor, P. J. Jewsbury, and J. W. Essex. A review of protein-small molecule docking methods. *J Comput Aided Mol Des*, 16(3):151–166, March 2002.
- [89] M Totrov and R Abagyan. Flexible ligand docking to multiple receptor conformations: a practical alternative. *Current opinion in structural biology*, 18(2):178–184, 2008.
- [90] M. L. Verdonk, J. C. Cole, P. Watson, V. Gillet, and P. Willett. Superstar: improved knowledge-based interaction fields for protein binding sites. *J Mol Biol*, 307(3):841–859, March 2001.

- [91] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development and testing of a general amber force field. *J Comput Chem*, 25(9):1157–1174, July 2004.
- [92] Junmei Wang, Wei Wang, Peter A. A. Kollman, and David A. A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graph Model*, February 2006.
- [93] Renxiao Wang, Yipin Lu, and Shaomeng Wang. Comparative evaluation of 11 scoring functions for molecular docking. *Journal of Medicinal Chemistry*, 46(12):2287–2303, June 2003.
- [94] Scott J. Weiner, Peter A. Kollman, David A. Case, Chandra U. Singh, Caterina Ghio, Guliano Alagona, Salvatore Profeta, and Paul Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*, 106(3):765–784, February 1984.
- [95] W. Welch, J. Ruppert, and A. N. Jain. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chem Biol*, 3(6):449–462, June 1996.