PhD program in Biomedical Engineering

ENRICHMENT OF VIRTUAL SCREENING RESULTS USING INDUCED-FIT TECHNIQUES

Doctoral Thesis by:

Jelisa Iglesias

Thesis supervisor:

Victor Guallar Tasies

Thesis co-supervisor:

Jorge Estrada Collado

Thesis ponent:

Alexandre Perera i Lluna

Universitat Politècnica de Catalunya

Life Science Department at Barcelona Supercomputing Center

Barcelona, September 2019

"*Every step is a first step if it's a step in the right direction*"

Tyffany Aching in *I shall wear midnight* by sir Terry Pratchett

Abstract

This thesis explains the design, development, test on the benchmarking dataset DUD-e and the application to an industrial VS project of the PELE VS platform.

The most common and quick tools used on VS campaigns do not take into account the induced-fit (IF) effect, although there are some methodologies capable of reproducing this effect they are either time consuming or very limited on which transformations the protein may undergo. In this thesis with the development of the PELE VS platform we aim at using the simulation software Protein Energy Landscape Exploration (PELE) to account for the IF effect.

The PELE software uses a Monte Carlo (MC) algorithm coupled with an energy minimization step to explore the ligand conformations and model the protein. This approach allows the program to account for both big conformational changes and small local changes of the protein and to perform a good conformational search of the ligand, which coupled can account for the if effect with only a quick simulation.

PELE has been traditionally, and successfully, used in the enzyme engineering field where only a few compounds per protein are usually tested and studied. In order to apply the program to the VS field, where thousands of compounds are tested *in silico*, the first step was to automatize the whole procedure of preparing, launching and analysing the simulations. Thus, during this thesis the PELE VS platform has been developed altogether with other auxiliary tools.

Once the platform was developed, we wanted to test the behaviour of PELE on a well known benchmarking dataset, thus we tried our methodology on the DUD-e dataset. Since this dataset is formed by more than 100 proteins, we chose a few proteins for each of the families present in the dataset, reducing the number of proteins to 21 systems. Then, we tried to use a general protocol for all the chosen proteins in order to improve the results of currently used scoring function (SF) in the field. After studying the simulations and trying several protocols on this subset we observed that every protein (or at least family) that we want to study needs an specific simulation protocol in order to correctly reproduce the IF effect and improve the results of the most used SF Glide from Schrödinger.

Finally, we applied the platform and our previous hypothesis to an industrial VS campaign, as part of the collaborative Retos project: Silicoderm. In this case we worked with only one protein

and several compounds and we confirmed the need for a tailored simulation protocol for the receptor in order to improve results.

Preface

The drug discovery field is responsible for the apparition of new drugs and therapies against diseases. Behind each new drug that reaches the market there are usually from 15 to 20 years of work from an interdisciplinary team of people, from chemists to doctors. Although the clinical assays seem to have improved their success rates, investing into a drug development project is still a high risk investment, due to the high amounts of compounds that never reach the market.

The work developed in this thesis wants to improve the ratio of compounds that can reach the later phases of the drug discovery process, and eventually reach the market, by taking into account protein flexibility and the IF effects produced upon ligand binding when performing the first steps of the drug discovery process, the Virtual Screening (VS). In the traditional methodologies used in the VS process the flexible behaviour of the proteins has been often neglected or treated minimally, and we hope that introducing it will improve the success rates of the selected compounds.

In order to introduce the flexible behaviour of the proteins into the process we will use the simulation software PELE, which has been shown to be good at predicting the right binding modes of molecules.

One of the main objectives of this thesis was to develop and test a platform that allows the use of PELE within the scope of VS projects.

While testing the platform we also wanted to develop a general methodology capable of improving the VS results of a gold-standard methodology.

Finally, we aimed at applying this new methodology to an active VS project in collaboration with a pharmaceutical company.

As the reader will be able to read all these objectives have been accomplished and they will be explained through the different chapters of this thesis; which is structured into five chapters plus 2 annexes.

The first chapter is an introduction explaining the basic concepts needed to understand this thesis, the state of the art of the VS field methodologies, the problems of this field, the motivation for this thesis and the thesis main objectives.

The second chapter called PELE VS contains the explanation about the software developed during this thesis. It explains the problems each program solves, how the programs are structured

and the reasons for their design.

The third chapter is called DUD-e Study. It contains the results of applying the VS framework (which is explained in chapter 2) combined with PELE simulations to the DUD-e dataset[66], and an explanation of the results obtained.

The fourth chapter is called "Retos project: SilicoDerm". It explains the methodology used and the results obtained during my participation in the collaborative SilicoDerm project. This project is carried in collaboration with the company Almirall thanks to a Retos grant provided by the MINECO

The fifth chapter are the conclusions; it summarizes the level of accomplishment of the objectives and it provides the general and specific conclusions of this thesis as well as some further work.

Acknowledgments

I would like to thank my director Victor Guallar for giving me the opportunity to join his lab and guiding me through this thesis; my co-director Jorge Estrada for the long discussions about the thesis and his guidance, and my tutor Alexandre Perera for his guidance through all last four years. I also would like to thank Suwipa Saen-Oon not only for her valuable help with the projects but also for all her encouragement through all four years.

I also would like to thank all the other PhD students and master students with whom Ive shared these last four years. Thanks for all those fun conversations over lunch, for the beers after work and dinners, thanks for being there and sharing this journey together.

Finally, to my family and friends whose support has helped me to keep pushing through the thesis, thank you, without your encouragement this work probably wouldn't have seen the light.

David when I started this thesis you were my boyfriend, now that Im finishing it you're my husband I cannot thank you enough. You have been my light house shining over the storm of emotions this thesis has brought me, thanks for always being there and for all your help.

Contents

List of Tables

List of Figures

Acronyms

aa amino acid. 1, 16, 17, 22, 25, 32, 33

aces Acetylcholinesterase. 1, 48, 50, 52, 53, 57, 61, 64, 69

adrb1 Androgen receptor 1. 1, 46, 50, 52, 53, 57, 64, 69

adrb2 Androgen receptor 2. 1, 46, 48, 50, 52, 57, 64, 69

andr1 Androgen Receptor 1. 1, 47, 48, 50, 52, 55, 57, 61, 64, 69

BE Binding Energy. 1, 37

BP binding pocket. xiii, 1, 4, 9, 10, 17, 55, 57, 61, 64, 79, 85, 92, 95, 96

BSC Barcelona Supercomputing Center. 1, 13

cdk2 Cyclin-dependent kinase 2. 1, 47, 50, 52, 64, 69–71, 73, 75

drd3 Dopamine D3 receptor. 1, 46, 50, 52, 53, 57, 61, 64, 69

EF enrichment factor. 1, 49, 50, 52, 53, 61, 69, 71, 73, 75, 105

esr1 Estrogen Receptor 1. 1, 47, 48, 50, 52, 55, 57, 61, 64, 69

gcr Glucocorticoid Receptor. 1, 47, 48, 50, 52, 53, 55, 57, 61, 64, 69–71, 73, 75

GPCR G-Protein Coupled Receptor. 1, 13, 46

h-bond hydrogen bond. 1, 46, 83, 85–87, 89, 90, 92–96, 101

hiv HIV protease. 1, 13, 47, 48, 55

hivw HIV protease containing one water molecule, which is responsible for the most common conformation of the binding pocket. 1, 47, 48, 50, 52, 57, 64, 69

HPC High Performance Computer. 1, 11, 13, 33

hs90 Heat shock protein HSP 90-alpha. 1, 48, 50, 52, 53, 55, 57, 64, 69

HTS high-throughput screening. 1, 7, 8

IF induced-fit. iii, v, 1, 6, 10, 15, 23, 49, 106

jak2 Tyrosine-protein kinase JAK2. 1, 46, 48, 50, 52, 53, 57, 61, 64, 69–71, 73, 75

LVS ligand-based VS. 1, 8

MC Monte Carlo. iii, 1, 11, 12

mcr Mineralocorticoid Receptor. 1, 47

mcrin Mineralocorticoid Receptor with the MET-852 pointing towards the inside of the protein. 1, 48, 50, 52, 53, 57, 64, 69–71, 73, 75

mcrout Mineralocorticoid Receptor with the MET-852 pointing towards the outside of the protein. 1, 48, 50, 52, 53, 57, 61, 64, 69

MD molecular dynamics. 1, 9, 10

NHR Nuclear Hormone Receptor. 1, 13, 46–48, 50, 52, 57, 69–71, 73, 75, 76

nram Neuraminidase. 1, 48, 50, 52, 53, 57, 61, 64, 69

PELE Protein Energy Landscape Exploration. iii, v, vi, xiii, 1, 6, 11–17, 19, 21–25, 27–36, 44–49, 51, 55, 69, 70, 75, 105

PELE BE PELE Binding Energy. xiv, 1, 48, 49, 55, 59, 79, 85–87, 90, 92–94, 97–101, 104

ppar Peroxisome proliferator-activated receptor gamma. 1, 47, 48, 50, 52, 55, 57, 61, 64, 69–71, 73, 75

prgr Progesterone Receptor. 1, 47, 48, 50, 52, 55, 57, 61, 64, 69

rxra Retinoid X Receptor Alpha. 1, 47, 48, 50, 52, 55, 57, 61, 64, 69

SASA Solvent Accessible Surface. 1, 75

SF scoring function. iii, 1, 15, 16, 23, 24, 37, 40, 41, 48, 106

SI Supplementary Information. 1, 52, 79

SVS structure-based VS. 1, 8–10

thb Thyroid Hormone Receptor Beta-1. 1, 47, 48, 50, 52, 53, 55, 57, 61, 64, 69

thrb Thrombin. 1, 47, 48, 50, 52, 57, 64

try1 Trypsin 1. 1, 47, 48, 50, 52, 64, 69

VS Virtual Screening. v, xiv, 1, 7, 8, 14, 15, 19, 23, 25, 44, 46, 77, 83, 86, 92, 101, 105, 106

wee1 Serine/threonine-protein kinase WEE1. 1, 47, 48, 61

Glossary

amino acids An Organic compound consistent on a backbone formed by an amine and a carboxyl functional groups, bonded by the peptide bond, and a side chain. 1

enzyme A protein that catalyses a chemical reaction. 1, 4, 13

GPCR The G-protein coupled receptors or GPCRs is a family of proteins that act as cell membrane receptors, reading the signals from the environment, by binding extracellular molecules, and starting a signal cascade inside the cell by activating G-proteins. 1

kinase A family of proteins that adds phosphates to proteins, mainly, as it performs a chemical reaction it can be considered an enzyme. 1, 46, 70, 71, 73, 75

near-native pose A protein's structure derived from a simulation that closely resembles to the experimental structure. 1, 77, 87, 90, 92, 101

NHR A family of proteins placed on the cell nucleus membrane, that act as receptors of the intracellular signals, by binding the molecules present in the cytoplasm. 1

protease A family of proteins in charge of breaking other proteins. 1, 13, 46, 47

protein A molecule with a biological function that is formed by amino acids bonded by a peptide bond. 1

Chapter 1

Introduction

1.1 Introductory concepts

The aim of this thesis is the development of a new methodology to improve the Virtual Screening (VS) procedure with the inclusion of the induced-fit (IF) effect by simulating the protein-ligand complex using the Protein Energy Landscape Exploration (PELE) program and a re-score process of the new structure.

These three lines involve a quite complex process depicted in Figure 1.1 which will be explained through all this thesis. In order to understand the process, we first need to understand some key concepts such as VS, drug, target/protein and IF.

1.1.1 Drugs and proteins

The first concept in our list is "drug". A drug is any molecule capable of causing an effect in the human body. This effect can go from calming a stomach-ache to combating cancer, and from being localized (anesthetize a tooth) to global (control the sugar level in blood).

Most of currently used drugs accomplish their effect by targeting one protein, that is, the drugs interact with the protein by binding to it. Logically, we now need to know what a protein is, our second concept.

A protein is a sequence of bonded by peptide bonds. The amino acids (aas) are molecules with the structure in Figure 1.2a where R represents a variable group called sidechain. In nature there are 20 standard side chains used by all forms of life from microorganisms to humans (Figure 1.2b. The sequence of the aas that form the protein is encoded by the DNA.

Proteins are present in every part of the cells and oversee most of cell's functions, from capturing the environmental stimuli to the replication of DNA. Thus, the malfunction of a protein most likely will cause a disease. Some examples include: (i) when the amyloid precursor protein isn't

(a) Thesis graphic abstract

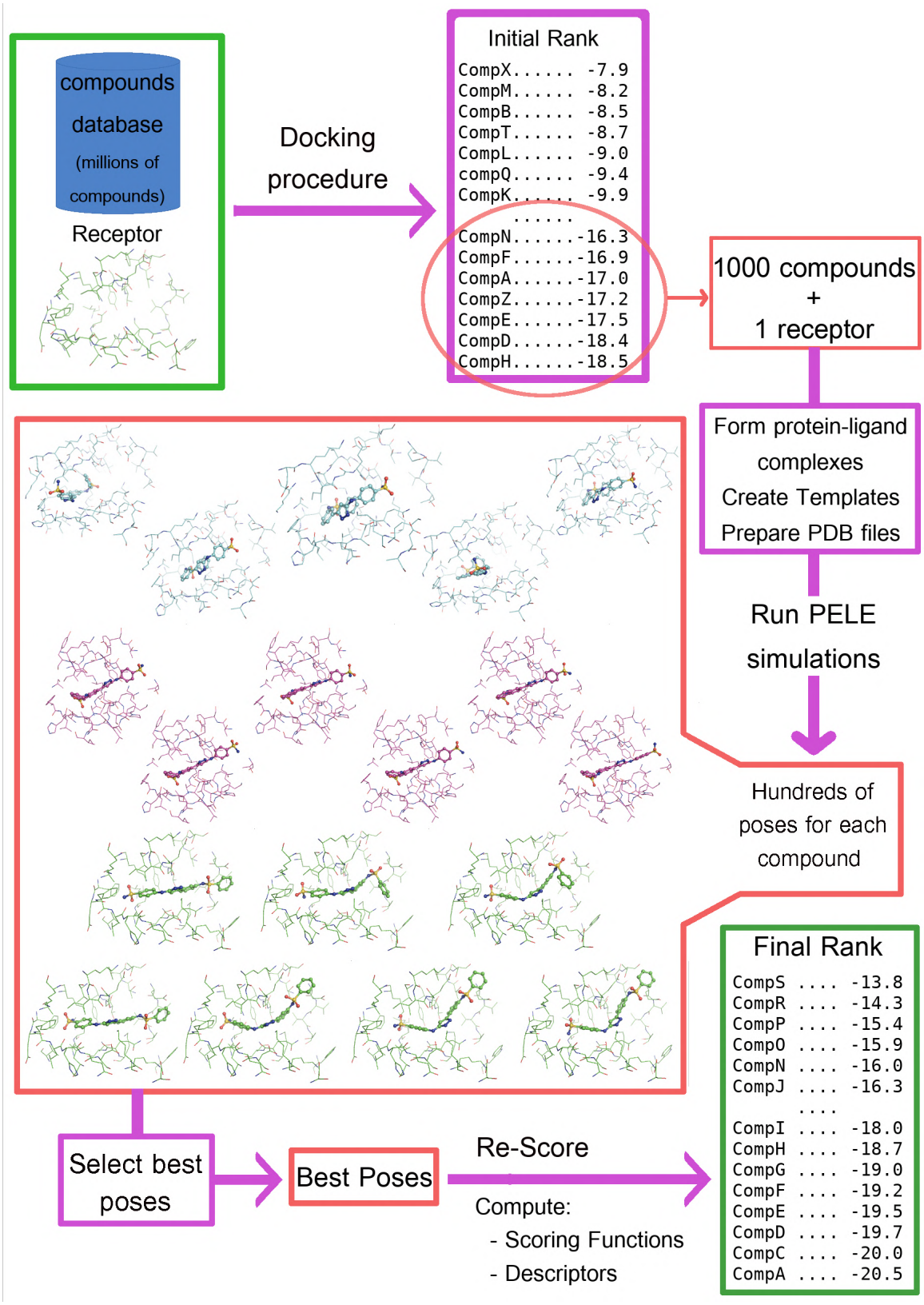
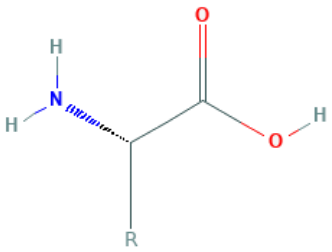


Figure 1.1: This thesis graphic abstract

(a) general structure.



(b) 20 standard structures.

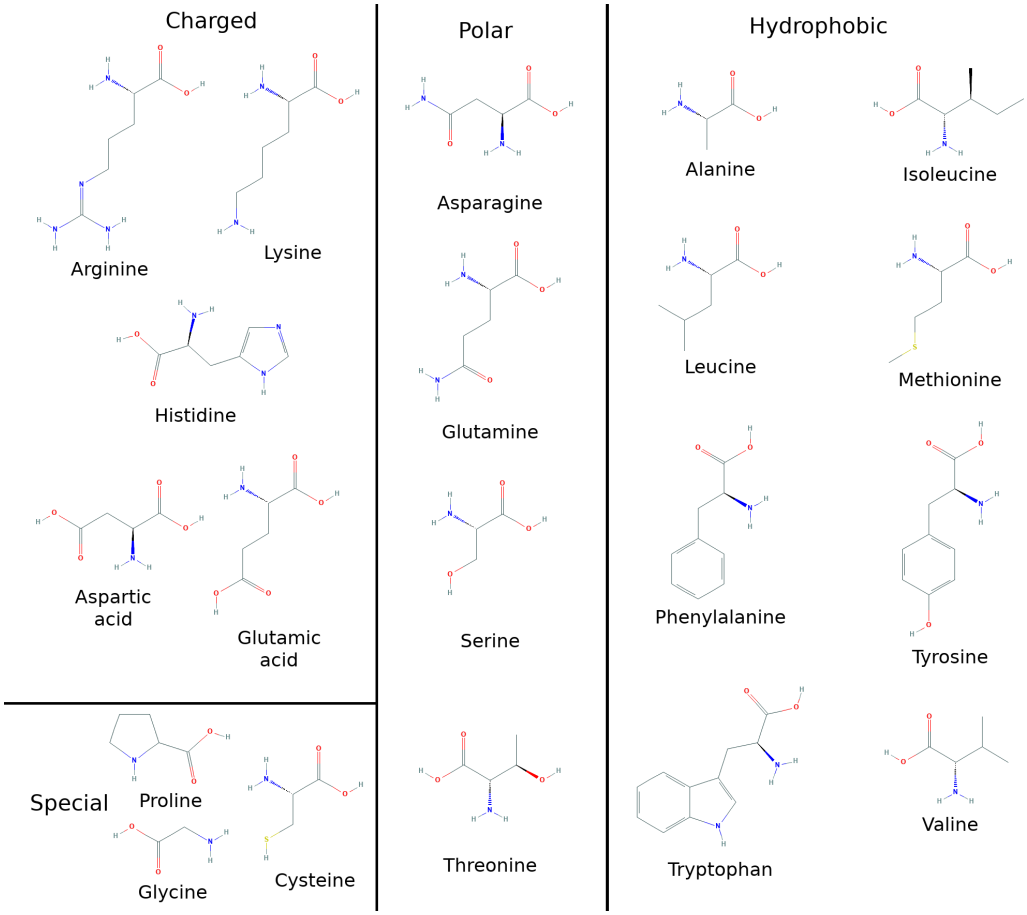


Figure 1.2: Standard grouped by their electrostatic properties.

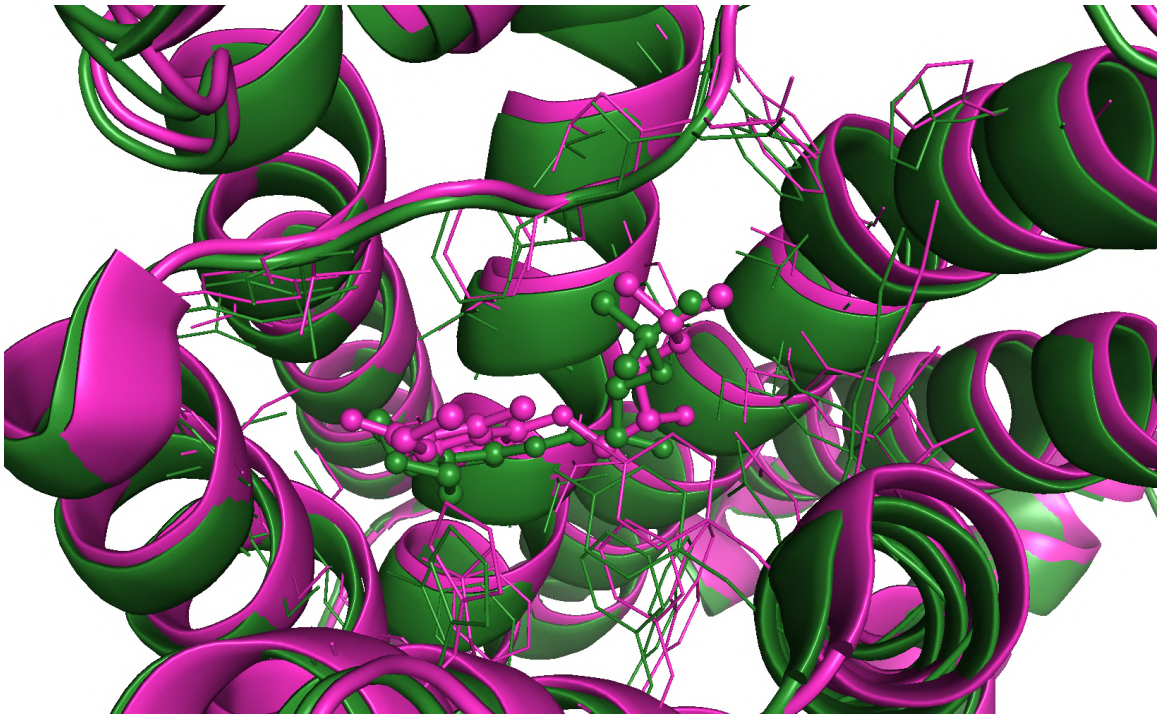


Figure 1.3: Two structures of the human beta2 adrenergic receptor in complex bound to two different ligands to show the conformational differences due to binding. The protein's structure is shown using the cartoon (backbone) and lines (BP) representations, while the ligands are shown using the ball and stick representation.

correctly processed the β-amyloid protein generated misfolds and aggregates causing alzheimer in the long term [29], (ii) when the p-53 protein malfunctions it causes cancer [61, 36], (iii) when the DiAminoOxidasa (DAO) enzyme doesn't work correctly it causes the DAO deficiency syndrome [60], (iv) the malfunction of insulin or its receptors causes diabetes.

The malfunction of a protein can be given (i) by an overactivity, or (ii) by a reduced or lack of activity of the protein. The first case can be produced by an overexpression of the protein or by an undesired activation of the protein, due to an excess of the native compound or to the interaction with non-native compounds caused by mutations of the DNA that modifies the binding site (BS) or binding pocket (BP), the region of the protein that interacts with the compounds. The second case can be caused by the lack of production of the protein itself, by a misfolding of the protein or by a mutation of the DNA that affects the protein's structure and/or its dynamics, or a small change in the BP that prevents the binding of compounds, from now on called ligands, and/or the chemical reaction that should take place

The proteins fold into dynamic complex 3D structures, these can be resolved by X-ray crystallography, by cryo-EM or by NMR. The X-ray crystallography requires the protein's crystal; thus,

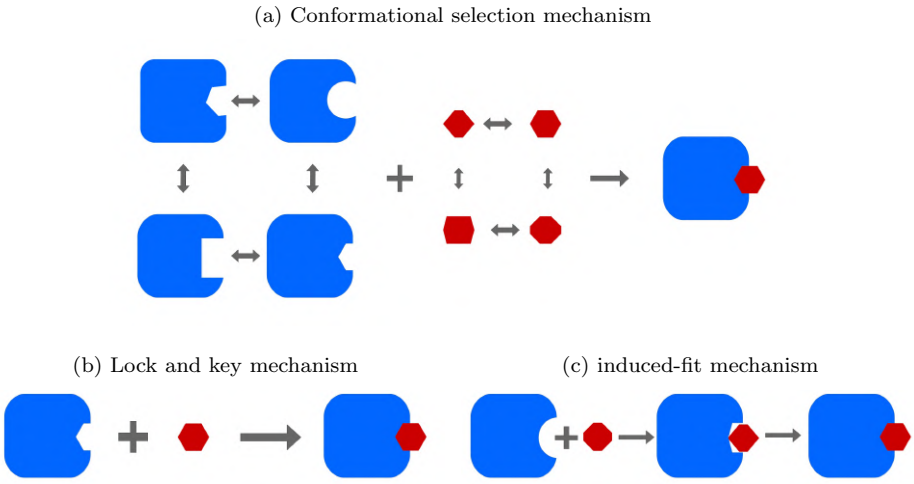


Figure 1.4: Schematics of the different Binding Mechanism theories.

it only provides a snapshot of the protein under non-cellular conditions. We can find crystallographic structures of a protein without ligands (apo-structure) and with different ligands bound (holo-structure), where we might appreciate changes in the protein structure. One example is the adrb2 protein shown in Figure 1.3. In this Figure we can appreciate how the beta andro-receptor 2 (a membrane receptor pertaining to the G-protein coupled receptor 1 family) changes it's structure upon binding of different ligands. These changes in conformation occur naturally in cell conditions because a protein is a flexible entity that can accommodate different ligands.

Proteins oversee many different functions of the cells: they are in charge of receiving signals, contract the cell, build new proteins, read and duplicate the DNA, produce energy, etc. We can categorize them according to the function they perform, their sequence similarity and/or their structural similarity; each of these categories of proteins is called a protein family.

That proteins are capable of binding different ligands is something proven by the experimental data, but the exact mechanism of the binding process is still unclear. There are three main theories about how binding occurs: (i) the lock and key theory, (ii) the conformational selection and (iii) the induced-fit theory. Figure 1.4 shows a schematic of each theory

The first theory that tried to explain the ligand binding process is the lock and key [20] proposed by Fischer in 1894 (Figure 1.4b). According to this theory the ligand and the protein are rigid entities that fit perfectly with one another. The binding occurs when the protein and ligand meet.

The second theory is the conformational selection [22] proposed by Frauenfelder, Parak, & Young in 1988 (Figure 1.4a). According to this theory the protein and the ligand are flexible entities that sample a given number of conformations, the bound conformation among them. The binding occurs when both systems happen to have the right conformation.

Lastly in 1958 Khosland, proposed the induced-fit theory [48] (Figure 1.4c) which states that

the protein undergoes the conformational changes needed to bind the ligand when this is in close proximity, and both elements change their conformation to fit perfectly.

Nowadays, with all the experimental data that we have at our disposal the lock and key theory has been discarded, and the binding process is described using the conformational sampling theory, the induced-fit theory or a mix of both theories. These theories treat the protein as a flexible entity, making the protein flexibility a key point, even if traditionally neglected, of the drug discovery process.

We have now introduced the third important concept of the thesis: the IF; we'll try to model it using the PELE software.

So far, we've seen what is a drug that targets a protein, what is this protein and how the drug binds to the protein, which can be explained by a mix of the conformational selection and the IF mechanisms. The last basic thing we need to know is how to differentiate between the compounds that will become good drugs and those that won't.

The ideal measure of a compound's potency, or binding affinity, is the Free Energy of Gibbs or ΔG; which tells us how good a binder is. Usually, experimental assays, however, measure half maximal inhibitory concentration (IC50), the dissociation constant (K_d) or the inhibition constant (K_i).

The IC50 value is the concentration at which the compound inhibits the biological function of the protein by half. K_d is an equilibrium constant measuring the dissociation/association of a given ligand. In addition, it equals the concentration of the free ligand at which half of it is associated with the receptor. K_i measures the equilibrium constant in competitive inhibition studies, and it's equivalent to K_d in single ligand cases.

$$K_{eq} = \frac{IC50}{1 + \frac{[L]}{K_m}} \tag{1.1}$$

$$\Delta G = -RTln(K_{eq}) \tag{1.2}$$

The IC50 value can change depending on the experiment used to obtain it, thus it cannot be compared among different targets or even the same target if the experimental conditions have changed. But it can be converted to the equilibrium constant (either K_i or K_d) using the formula in equation 1.1 where K_{eq} is the equilibrium constant, IC50 is the value obtained from the experiment, [L] is the maximum concentration of the ligand used for the experiment and K_m is the concentration of ligand at which the target's activity is at half maximal.

In turn the equilibrium constants (K_i or K_d) can be converted into ΔG value using the equation 1.2. Where the R is the gas constant, T is the temperature and K_{eq} is the K_i or K_d values.

1.1.2 Drug discovery process

The drug discovery process is a long and expensive process with really high attrition rates. This process can be divided in the following stages: (i) target discovery, (ii) target validation, (iii) hit finding, (iv) hit to lead, (v) lead optimization, (vi) pre-clinical stage and (vii) clinical stages.

The exact number of compounds that make it from the VS campaign to the market is really difficult to obtain, since most of the research is kept confidential until the compound reaches the clinical phases. The success rate of the clinical phases can be computed from the public information available, but it's a complex calculation highly influenced by the data used [14, 35, 77]. The latest analysis show that the probability of success of the clinical phases also varies greatly depending on the study's target [77]. Roughly, only 10% of the compounds reaching clinical phase I eventually reach the market.

We can say there are two main reasons for this low rate of success: (i) the apparition of undesired secondary effects on later phases (toxicity issues and lack of efficacy are the main reasons for the low success rates in clinical phases) and (ii) the low ratio of true positives derived from the initial VS that can move to the next phases.

The apparition of undesired secondary effects is a hard to asses problem, because sometimes the effects cannot be detected until the compound is experimentally tested in a complex model such as animals or humans, due to the lack of different tissues in the *in vitro* models and the simplistic model of the *in silico* screening where only one protein is studied instead of the hundreds of thousands of proteins that are acting at any moment in a human body. The prediction of adverse effects is very challenging as it can be linked to population polymorphisms that are not tested in phases (i) to (iii).

The low ratio of true positives derived from an initial screening of compounds is an equally complex issue. The most successful high-throughput screening (HTS) can present a 5% of active compounds, usually it renders between 2% and 3% of active compounds. 90% of the compounds obtained in this phase will present undesired secondary effects such as toxicity, low solubility or low activity, which will discard them from becoming drugs. From this percentages it is easy to imagine how many thousands of compounds have to be tested to obtain a few that will become prospective drugs.

If we combine these two problems with the fact that the chemical space is almost infinite, we end up with a combinatorial problem of huge proportions; which cannot be solved experimentally due to the amount of time and money required to solve it. Thanks to the advances of the computational methods and technology nowadays we can tackle this problem by using computational methodologies called VS.

The VS methodologies include any and all the computational methods used during the drug discovery process. These methodologies allow us to test more compounds, in a more inexpensive and quicker way [28]. These methodologies do not replace the experiments completely but are used as a complement to the experimental HTS assays.

In a HTS assay from 1000 compounds to 100000 compounds are tested, depending on the resources available and it can take from months to years [43]. While with the VS methodologies we can test millions of compounds in a few days.

1.2 VS state of the art

Thanks to the general improvement of the computational field, both the resources at our disposal and the computational methods have evolved rapidly over the last decades, and thus the VS process has evolved from an experimental methodology that may or may not improve the results, to an integral step in this process. The VS process allows the researcher to improve the quality and increase the chemical variability of the compounds chosen to undergo further experiments, at the same time reducing the cost and time spent searching for compounds that may become prospective drugs [28].

The VS process is nowadays an important step of the drug discovery field, that encompasses different stages in itself starting from filtering compounds from databases to the structural studies of how the compounds may bind the receptor; thus, even providing information on the mechanism of action of the compounds. There are many good reviews that explain in more detail the complexity of this process and the possible steps in which it can be divided. [28, 51, 53, 21, 8, 57, 12, 31]

For the scope of this thesis what we need to know is that the VS methodologies can be divided into two categories: the ligand-based VS (LVS) and the structure-based VS (SVS). The main difference between these two types lays in what they base the search of new compounds on [5, 46, 21].

The classification of the VS methodologies into LVS and SVS doesn't translate into a separation of techniques. Some techniques such as molecular fingerprints or pharmacophores [52, 3] can be included into both categories depending on the structure they use to build the models.

This categorization of the methods doesn't make them incompatible, in fact in many cases the process involves the use of both types of screenings in a hierarchical manner. First, the LVS techniques are used to filter the millions of compounds available. Once filtered the SVS is used to further filter the compounds and to estimate their affinity to the target.

1.2.1 Ligand based VS

The LVS uses the known ligands to look for new compounds based on their properties; this type of methodologies are usually faster than the SVS and present similar results [65, 50]. These methods are dependant on the existence of known ligands of the receptor.

The LVS can be divided into 2-D methodologies and 3-D methodologies [21], depending on whether they take into account only the type of atoms and bonds present in the molecule while performing a similarity search [59] or they take into account the conformation of the molecule.

In this category we can also include the filtering criteria used to choose compounds based on their physico-chemical properties such as the Lipinski rules [56], the number of rotatables bonds present in the compounds, since it has been shown their influence in the oral availability of drugs [75, 2] or their logD, which has been proven to influence the drugs intestinal permeability [19].

The main issue with this type of methodologies is that they require the knowledge about known ligands, in most cases this information will be available, but for the most novel targets we may not have it. The other issue that may arise is that this type of methodology tends to produce similar compounds to those already known; which is a problem if we are searching for novel compounds to patent.

1.2.2 Structure-based method

The SVS uses the structural information of the receptor, or when available the information about known ligands-receptor complexes. This information can be obtained from a structure derived from experimental methodologies or from modelling processes [11].

In this category are included plenty of techniques such as pharmacophores derived from the receptor's structure, the fingertips describing the ligand-complex interactions [13], docking and even molecular dynamics (MD).

In order to use these methodologies, we need to know the target's structure, or at least the structure of proteins with a similar structure, in order to create a homology model. Since proteins from the same family are assumed to present similar structures, due to the similarity in their sequence and function, they are used to create models of the target's structure.

For those targets where we have structural information about how they bind to known compounds, or known ligands, we can extract the BP characteristics. This information can be added into the search for new compounds in order to better filter them.

However, if this information is not available, there are methods that allow us to estimate the BP such as homology modelling or druggability studies. Also, some of these methods, like the pharmacophores or docking ones, can be used to make exploratory searches over compound databases that bind to the expected BP.

One of the most common techniques used to search for new prospective ligands, or hits, is the docking methodology. The common step among all the docking techniques is the modelling of the protein and the ligand as two grids that need to interlock with the minimum clashes possible. In order to do so they split the process into two steps: the sampling and the scoring. During the sampling phase the ligand explores the 3D space looking for the empty gaps inside the BP, thus, optimizing the contacts and minimizing the clashes between the molecules. In the scoring phase the poses are scored, and the program selects the best one based on this score.

When these technologies appeared both grids were considered rigid and the ligand was simply moved in the 3D space in order to match the BP during the sampling phase. As the computational

resources available improved, these technologies started to take consider the ligand flexibility, by using multiple conformations of the molecule which undergo both phases. The main differences between the different docking techniques lay in how they asses the sampling phase and the score they use.

1.2.3 Incorporation of flexibility to the docking

As mentioned before, nowadays most docking methodologies take into account the flexibility of the ligand but still neglect protein flexibility due to the computational cost required to account for it. Nonetheless, in the last years several methods to incorporate this flexibility have appeared. These new methods have been shown to be able to obtain prospective ligands in several cases [41].

Several of the methodologies used have been based on the ensemble docking methodologies [40], where the docking is performed on multiple structures of the same target. In this category we can find programs ready to use multiple structures for the docking such as HYBRID and FRED programs [64, 63] owned by OpenEye, or the 4D docking developed by ICM [73]; and programs that provide tools to generate the multiple structures needed and/or analyse the results of a traditional docking over multiple targets of the same protein. The methods used to obtain the structures needed for the ensemble docking can include: the generation of multiple structures using experimental methodologies such as X-ray (crystals with different ligands but same receptor) or NMR experiments; the use of different snapshots obtained from a simulation [18] or from modelling approaches.

There are also methods capable of performing structural re-arrangements by means of adding a step where the sidechains present in the BP explore several possible conformations in order to somehow emulate the IF effect. Some of the methodologies included in this category are the IFD developed by Schrödinger [72] or the SCARE methodology from ICM[7], which mainly perform a rigid docking on a modified BP to minimize steric clashes and then samples sidechains of the residues that clash with the best poses, but these methods do not sample the backbone. Although these methods are almost as quick as the traditional docking methodologies and are capable to reproduce some of the IF effects, due to the lack of backbone sampling, these methods aren't capable of reproducing big conformational changes affecting the protein's backbone.

The other way to account for target flexibility when performing SVS is to use MD or other simulation methodologies such as MCPRO from Jorgensen[44] or ProtoMS[78] from Essex's lab. As we have seen, the simulations can be used to obtain multiple structures of the receptor (or target) in order to perform the docking. It can also be used in order to perform mechanistic studies of the binding procedure or to introduce the IF effect and to estimate the binding affinity or ΔG of a compound. [41]. These methods allow to reproduce any kind of IF effect produced upon ligand binding., Nevertheless they are also extremely time-consuming: while a traditional docking can asses thousands of compounds in just a few hours, these methodologies require several hours per compound to asses.

1.3 Introduction to PELE and examples of use

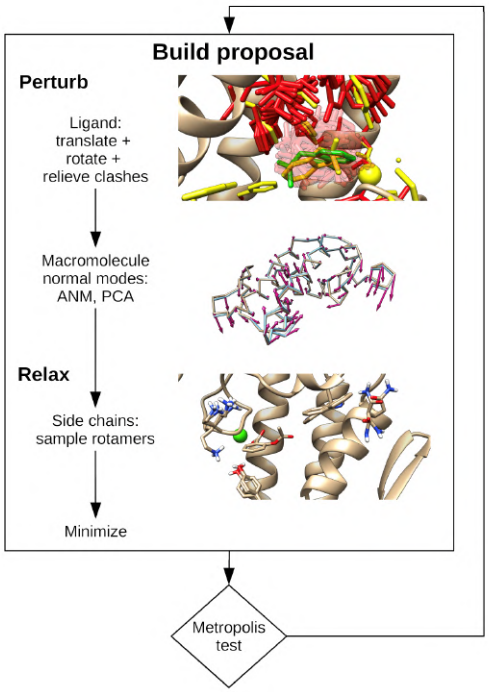


Figure 1.5: This image represents the steps that the PELE simulation performs. It has been extracted from the book: Monte Carlo Techniques for Drug Design: The Success Case of PELE [27]

The PELE Method

PELE [45, 58, 6] introduces an unconventional Monte Carlo (MC) procedure where a stochastic perturbation is followed by a relaxation step using protein structure prediction algorithms. The code was intended to map the energy landscape of a protein-ligand complex and expanded later to model protein-protein and protein-DNA interactions. It was designed to use massively parallel High Performance Computer (HPC) , where a computing core runs an individual trajectory; typically, simulations involve from tens to hundreds of processors for thousands of MC steps.

More in detail, the heuristic MC procedure of PELE involves the following steps (Figure 1.5):

Ligand Perturbation. The ligand, which is built by a rigid core and a set of rotatable fragments, is initially perturbed by forcing a translation and a rotation. From a list of different perturbation poses (typically between 1 and 20) PELE chooses the one with the lowest total energy. Each pose is randomly generated, where several moves are tried until a free-clash combination is found. Based on the different goals, translation and rotations take different values: a binding

site search uses large translations (around 3-6 Å), while a local exploration of the active site is typically restricted to values ¡ 1 Å. To enhance the sampling of rare events, the translational vector may be kept for a given number of MC steps.

Protein Perturbation. As a second step, the backbone is perturbed following normal modes calculated using the Anisotropic Network Model (ANM) or a user-given vector(s), for example built from a Principal Component Analysis (PCA) derived from X-ray structures, molecular dynamics, etc. It is possible to use a single mode, or to mix several ones; if using ANM (the most common approach) modes are selected from (typically) the 6 lowest frequency modes, since they are the ones more closely related to conformational movements. To apply the movement harmonic constraints are placed on the CAs target positions, and an all-atom energy minimization is performed using the multi-scale Truncated Newton algorithm.

Side Chain Sampling. Side chains readjustment is done for a selected list of residues, such as those surrounding the ligand. The side chains that increased their energy the most along the previous perturbation step, and for the ligand itself. The sampling first considers only rotamers as possible conformations and places the best rotamer (after clustering) in a residue by residue way. This is iterated until two rounds do not significantly change the prediction.

Minimization. The sampling procedure ends by a multi-scale Truncated Newton minimization where atoms representing the nodes in the ANM phase might be weakly constrained (so we do not undo the perturbation).

As in other MC procedures, the resulting structure is then accepted or rejected based on a Metropolis test. Overall, each MC step takes between 20 seconds and a minute, depending on the system size, number of perturbation tries, number of side chains to predict, etc. PELE uses the OPLS 2005 or the AMBER99sbBSC0 energy function and parameters along with one of the three different implicit solvents implemented: the OBC model [70] with a non-polar term following the ACE model [25], the Surface Generalized Born (SGB) model [24] and the Variable Dielectric Generalized Born model [62]. PELE allows placing specific discrete water molecules and has recently incorporated an additional perturbation step performing a quick MC on selected explicit waters (following the ligand perturbation). In addition, an enhanced sampling strategy using adaptive techniques and reward functions has been recently introduced, AdaptivePELE [55]. In this approach, several iterations of standard PELE simulation are run for a reduced number of MC steps (4-20): the epochs. After each epoch, all conformations are clustered and each cluster is assigned a reward value that favours (by default) clusters less explored; additional rewards might be defined from user defined properties such as reaction coordinates, ligand exposure to solvent, etc. Then, the next epoch is started using selected conformations based on the reward function, as new seeds.

Examples of PELEs applications

PELE was designed to map ligand migration pathways. In its initial application, with only few hours of a small computational cluster (4 cores) it mapped how a palmitate fatty acid escaped the fatty acid binding protein, without any bias [55]. The fast performance of PELE , for example, allowed the first ligand migration study on the (tetramer) human hemoglobin, identifying differences between the α and β subunits and between their T and R states [71]. Using larger HPC resources, PELE could perform a complete non-biased exploration, mapping the entire active site search and binding on significantly challenging systems, such as prolyl oligopeptidase [49], the binding of porphyrin into Gun4 [47], or in cisplatin non-covalent binding into a DNA receptor, etc. [10]. Such type of analysis was further enhanced with the development of AdaptivePELE [4]. This procedure improved approximately one order of magnitude the exploration performed by PELE , as seen in its application in GPCR or in Nuclear Hormone Receptors . PELE's modelling capabilities drove significant biomedical/pharmaceutical studies, attracting, in addition, the interest from some pharmaceutical companies. Application studies included research on cancer targets such as mTOR [17] and BCL-2 [39], on the glycosylation disorder through the study of the human phosphomannomutase2 receptor [1], and on diverse NHRs , in collaboration with AstraZeneca [16, 30]. Additional biomedical studies in this line, for example, include the prediction of drug resistance in the HIV protease (hiv)-1 protease receptor, where PELE was able to (blindly) identify high resistance patients using their viruses sequence data [37].

Besides biomedical applications, PELE has been successfully used in enzyme engineering biotechnological applications. Mutational designs have introduced improved enzymes on several systems, including oxidation of secondary alcohols in flavin systems, 2-phenyethanol oxidation in toluene 4-Monooxygenase [38], laccases [26], etc.

We've chosen to use this methodology for the thesis because, although it is a simulation methodology, it allows us to perform both local and global samplings in a short time. The user can choose to perform a local sampling, consuming as little as half an hour, or to perform big conformational changes, in several hours. The usual simulation is in the middle of these two types of exploration, optimizing the sidechains while allowing for some small changes in the backbone, and it takes a few hours. PELE has been developed by the EAPM group in the Life Science department at the Barcelona Supercomputing Center (BSC), where this thesis has taken place.

1.4 Objectives

In the last years the protein's flexibility has been proven to be a key point for the VS procedure. As such several methods and procedures have been developed to take it into account.

The focus of this thesis lays on the development of methodologies that allows us to use PELE on VS campaigns.

1. The development of a platform to allow the use of PELE in VS campaigns where thousands of compounds need to be tested *in silico*.

2. The development of a general simulation protocol capable of improving the VS results compared to a gold-standard methodology

3. The use of the developed platform and simulation protocol in a real-life industrial VS campaign.

Chapter 2

PELE VS

This chapter explains the design and implementation of the PELE VS platform, which I've developed. The platform is capable of generating a ranking of the compounds provided based on their estimated activity, taking into account the IF effect they produce on the protein, by performing a PELE simulation and computing several scoring functions (SFs).

This platform automatically prepares and analyses PELE simulations, selecting the best structures from each of the simulations, and re-scores them using multiple scoring functions to generate a ranking of the compounds.

The platform is also capable of introducing mutations into the proteins structures and placing missing atoms into the protein residues. This allows the user to use incomplete structures or introduce mutations into known crystals in order to observe the possible changes in activity.

It can also compute several public and proprietary scoring functions and generate a .csv file to simplify analysis. This way the user can study how well the different SF perform, combine them into one, or just pick the best one according to the user's criteria.

In this chapter I'll explain the PELE VS platform developed during this thesis. The main purpose of this platform is the automation of our procedure due to the amount of data to be treated and the repetitive nature of the processes involved.

2.1 General Problem description

During a VS campaign up to a few millions of compounds are assessed in a short period of time using computational techniques. The result of the VS is a ranking of the assessed compounds, according to their estimated affinities. The top compounds from this ranking will be experimentally tested to ascertain their binding properties and their fitness as a drug candidate.

We hope that by introducing a short simulation, done by the group's simulation program PELE, we can improve the protein-ligand complexes structures, in order to create a new ranking with any

of the supported scoring function (SF). We expect that the new ranking has either a better ratio of true binders in the top positions or that the potency of these compounds is better, ideally both will be attained. Due to the computational cost of our approach, we don't expect to asses millions of compounds, but thousands of them.

In order to process thousands of compounds the procedure needs to be automated. If we were to do it manually, the initial preparation, launch, and analysis of the simulations of approximately 200 compounds can take up to 2 months of work.

A second step would be to compute all the SF supported by the platform, when done manually one by one it requires up to another 2 months. The last step is to extract the data from the outputs, which amounts to another month.

If we want to apply our procedure to thousands of compounds, doing so manually would take years, which is unsustainable. Thus, we need to automate the procedure in order to make it efficient and competitive in a VS campaign.

2.2 Protein preparation

This section explains the mut_prep4pele module. It explains what problem is solved by this module, the workflow designed to solve the problem and finally how the workflow has been implemented. Each of these explanations corresponds with one of the following subsections: 2.2.1, 2.2.2 and 2.2.3

2.2.1 Problem description

The first requirement of our simulation program PELE, is a 3D structure of the protein with no missing atoms. It also requires a non-standard nomenclature for certain residues; mainly those with multiple protonation states. Finally, like many other simulation programs, it requires that the atoms of the protein are named in a specific way.

In order to obtain structures that meet the aforementioned requisites I've developed a program called mut-prep4pele which is available at `https://github.com/Jelisa/mut_prep4pele`. This program is capable of placing hydrogens, complete residues with missing atoms in their sidechains and produce mutated structures for the 20 standard aa.

The program will change residue names and atom names to match the naming requirements of PELE. If the structure is complete and it has the protonation states already set, the program won't modify them; otherwise the program will set the standard protonation states.

The standard protonation states will render: the arginine and lysine positively charged, the glutamate and aspartate negatively charged and the histidine neutral with the hydrogen placed on the carbon$_E$ atom.

Another requirement of PELE is that the ligand should be placed in its own chain and if it is not of proteic nature its atoms should present no repetition of names; in consequence, the program

has the option to change the atom names to make them unique inside a specified chain.

PELE is also used on the field of enzymology, where the research is focused on improving the protein's activity by modifying its BP, which means introducing mutations onto the protein. Thus, the program has the option to introduce mutations into the proteins.

2.2.2 Workflow

Figure 2.1 is the graphical representation of the main program's workflow, which will be explained in more detail in the following lines.

The program input is the complete path to a 3D structure of a protein or protein-ligand complex and the options chosen by the user.

The program parses the options chosen, which involves making sure the options provided are valid, and if a mutation is required, check whether it's valid or not.

Then the program tries to read the 3D structure provided by the user. If it's unreadable the program will generate an error and crash, otherwise it will read the file and find the initial and last residues present.

If the structure presents insertion codes, the program will renumber the structure to eliminate them. The insertion codes are used when two residues have the same residue number in order to differentiate them but PELE is unable to work with them correctly. Thus, when present, the program will renumber the structure to eliminate them.

Then, the program will check for the presence of gaps in the structure using a distance criterion. With the crystallization process sometimes the more flexible parts of the proteins cannot be obtained thus generating an incomplete structure with gaps. The program will detect a gap if the C atom of a residue and the N atom of the following one are at more than the user-defined distance, the pdb_resolution option, or a default value of 1.55 Å. The resolution of a pdb file sets the minimum distance at which the bonded atoms can be placed without error, that means a pdb file with a resolution of 3 Å can present this distance between the C atom of a residue and the N atom of the following residue without presenting any gaps, thus this bigger distance should be used to avoid detecting inexistent gaps.

The following step that the program performs is to fix the residues' names to match the PELE nomenclature. In this step the program checks whether the residues' names match the PELE nomenclature or not. If they don't match the names will be changed to match the nomenclature. It will look for the hydrogens present for the histidine, lysine, glutamic and aspartic aa and depending on them it will adopt the residue name respective to the protonation state.

Next, the program checks and fixes the atoms' names. During this step the program will ensure that the atoms of each residue match the atomic names of the residue with the PELE nomenclature.

Then, the program will check the structure looking for any missing atoms, any extra atoms, any metal to coordinate and the residues without a PELE template. The presence of the two first cases

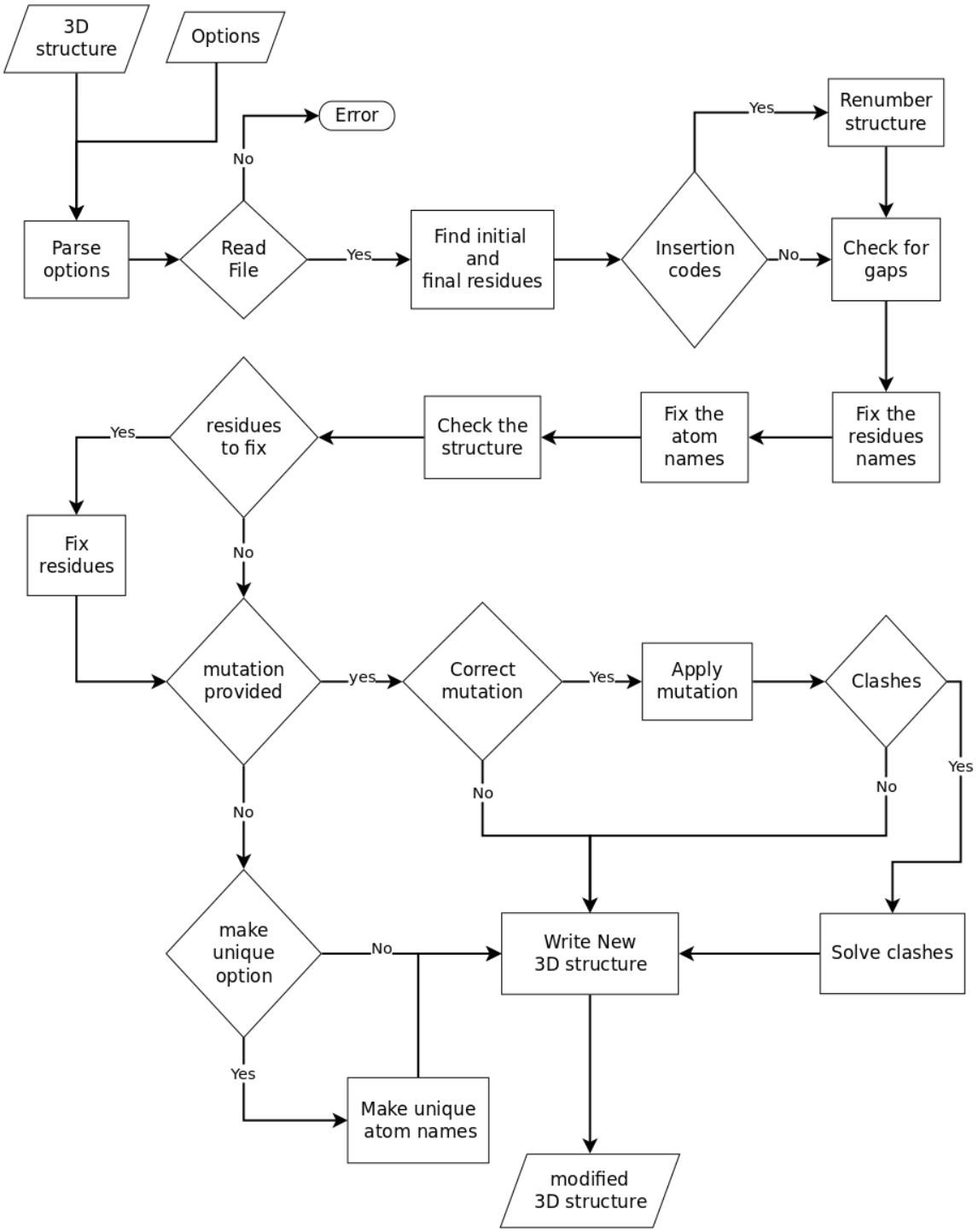


Figure 2.1: mut_prep4pele.py flowchart.

requires that the structure is fixed; in case of metals to coordinate the program will provide this information to the user so they can take them into account when preparing the PELE simulation. The user should be aware that mutprep will only correct those residues with a template.

If there are any extra or missing atoms the program will eliminate or add the needed atoms or transform those that need to be changed. The atoms will need to be changed only if the charge_terminals option is chosen, in which case the program will add one extra hydrogen atom to the N-terminal residue so it's positively charged, and transform one hydrogen atom into an oxygen atom and eliminate one hydrogen from the C-terminal residue.

Following the check and correction of the structure to meet the PELE requirements the program reaches a point where its dual nature shows. On one hand, this program is designed to introduce mutations to the protein structure provided while ensuring their "correctness" to be used by PELE. On the other hand, it's been adapted to meet the needs of VS campaigns.

If no mutation is provided the program will check the option make_unique. If it has been used, the program will change the atoms' names of the chain provided by the option in a way that no two atoms present the same name.

Otherwise, for each mutation provided, the program will check whether the mutation is valid or not. If it's valid it will apply the mutation, then check for clashes and try to solve them.

In any case, at the end of both paths, the program will write the new structure obtained into a .pdb file.

2.2.3 Implementation

In this subsection we'll present how the previous workflow has been implemented into several packages and classes, and how do they interact. We'll also present the module's dependencies and its inputs, outputs and options as well as an example of use.

Dependencies

The script has the following dependencies:

ProDy 1.8.X A library to read 3D structures of biomolecules in pdb format and manipulate them.

Numpy A mathematical library.

Modules

Figure 2.2 shows each of the modules and their relationships. This program is quite complex and as such its composed by one main module and 11 auxiliary modules.

There are three basic modules that provide information to the rest of the program: the `enviroment_parameters`, the `parameters_help` and the `global_variables` modules. The modules on this level do not present dependencies to other modules, while being extensively used by

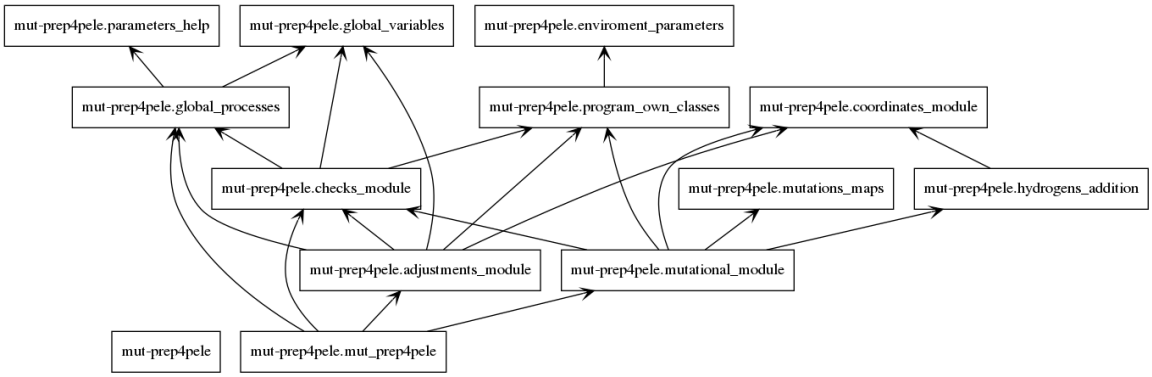


Figure 2.2: Module's relationship for the mut_prep4pele package.

the modules in the next three levels of the tree. The `enviroment_parameters` module contains information about where the program is installed, and its only variable should be modified upon download of the program. This variable allows the user to call the program from wherever he wants. The `parameters_help` module contains all the information about the program options. Lastly, the `global_variables` modules contains all the constant variables the program may need.

On the next level of the module's tree in Figure 2.2, we find the modules: `global_processes`, `program_own_classes` and `coordinates_modules`.

The `global_processes` module contains the general use functions. As such it contains the functions that process the options provided to the program and the functions that affect the whole protein at the same time. This module depends on the `parameters_help` and the `global_variables` MODULES.

The `program_own_classes` contains the implementation of the two classes developed for this program. The `coordinates_module` contains all the functions used to modify a residue's position, or the atom's positions. This module depends on the `enviroment_parameters`. This dependency is due to the classes' dependency on the information contained in the Data folder; this folder is provided with the program thus the dependency at the module's level.

The `coordinates_modules` contains the functions in charge of computing and modifying the atom's coordinates, together with the functions implementing the mathematical functions needed. This module doesn't depend on other modules.

On the middle level of the tree we have the `checks_module`, the `mutations_map` and the `hydrogens_addition` modules.

The `checks_module` module contains the functions that perform any kind of check. The checks the program has to perform go from checking the requested mutation, to check the protein's structure. This module depends on the `global_processes`, the `program_own_classes` and the `global_variables` modules. The module uses just one of the implemented functions in the `global_processes` module one time inside one of the function's implemented. While the other

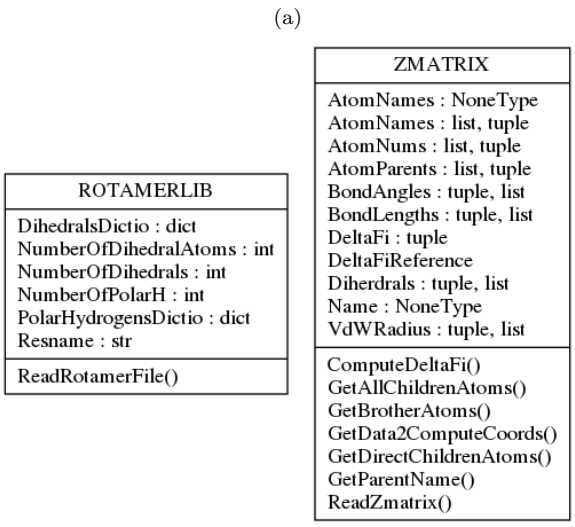


Figure 2.3: UML representation of the classes used by mut_prep4pele.

two modules needed are more extensively used throughout all the functions implemented in this module.

The `mutations_map` contains the information about how the program should perform the mutations. This module has no functions and it has no dependencies on any other modules. Although the information in this module is constant, it's only used by one other module, thus I decided to incorporate it into a separate module from the `global_variables` module.

The `hydrogens_addition` module contains the functions needed to place hydrogen atoms into the protein. This module depends on the `coordinates_modules`.

At the fourth level of the tree there are only two modules: the `adjustments_module` and the `mutational_module`.

The `adjustments_module` contains functions that perform corrections on the protein to ensure that its format and its structure won't cause a malfunction of the simulation program PELE. This module depends of the following modules: `global_processes`, `checks_module`, `global_variables`, `program_own_classes` and `coordinates_modules`.

The `mutational_module` implements the functions in charge of performing the mutations and/or modifying the protein's structure. It depends on the modules: `checks_module`, `program_own_classes`, `coordinates_modules`, `mutations_map` and `hydrogens_addition`.

The last level contains the main module, the `mut_prep4pele` module. This is the module the user should invoke in order to use the program. The programs depend on the modules: `global_processes`, `checks_module`, `adjustments_module` and the `mutational_module`.

Classes

Figure 2.3 shows the UML representation of the classes and their relationship. As we can see, the program has two independent classes: the `ROTAMERLIB` class and the `ZMATRIX` class.

The `ROTAMERLIB` is used to parse and store the information contained in the rotamer libraries files used by PELE. These files contain the information about how the aa sidechains can be positioned. This information is used to minimize the number of sidechain's clashes of a mutated residue with the rest of the original protein.

The `ZMATRIX` class stores an extended Zmatrix information extracted from the PELE Templates. The Zmatrix of a molecule gives information on the canonical bonding distances, angles and dihedral angles present. The PELE Templates also provide information on how the molecule's atoms are connected, and energetic information, this information is also saved into the class, thus the extended Zmatrix. This class is used whenever a modification of the protein has to be done. It's also used to check the correct form of the residues.

Input and Options

The script has only one mandatory argument, and then it has several optional arguments.

Its only mandatory argument is called *input_pdb* or *ipdb*. This parameter specifies the path to the .pdb file(s) with the structure(s) to prepare for PELE or mutate.

Additional arguments include:

output_pdb or *opdb* takes the path to write the final pdb. If it isn't specified, the program will generate a pdb with the same name as the initial pdb but with the suffix _processed.

mutation or *mut* specifies the desired mutation(s). It should be a string with the desired mutation(s) in the following format: 'residue XXX N to YYY' where: XXX is the original residue, N is the number of the residue and YYY is the desired aa. The names of the residues should be the three letters code for the aa, with the exception of the histidine residue for which one of the following names should be used: HID, HIP, HIE, depending on the desired protonation state.

mutants_from_file or *mut_file* specifies a path to read the output names and the mutations from a file. The file should have one mutation per line and no blank lines. Each line should have the following format "output_name→XXX␣N␣C␣YYY ", where: → indicates a tab space, ␣ marks a space character, XXX is the initial aa in 3 letters code, N the residue number, C the chain to modify and YYY the desired aa in 3 letters code.

mutant_multiple argument is a flag to indicate that all the given mutations should be placed on the same 3D structure. When present the program will create one structure with all the mutations specified without checking for steric clashes.

charge_terminals option is a flag to indicate the program that it should charge any terminal residue, those at the beginning and ending of the sequence and those involved in a gap.

no_gaps_ter option is a flag. When this option is present the program won't add a TER mark whenever it finds a gap in the sequence.

pdb_resolution takes as argument a float indicating the pdb resolution. This number will be used as the maximum allowed between the N atom of a residue and the C atom of the previous residue when checking for gaps. In this check the minimum distance that will be used is 1.55A.

make_unique option takes a single argument: a letter from the alphabet. It should be used to specify the name of a chain present in the structure for which the program will set unique atom's names.

remove_terminal_missing option is a flag. when chosen the program will remove the terminal residues in case they are missing heavy atoms in the backbone of the protein(that is: N, CA, C, O).

The `mut_prep4pele.py` script is the main module in the mut_preplibrary package

2.3 Virtual Screening Framework

This section explains how the VS platform general workflow, its structure and the reasons behind the platform's design.

The problem this framework wants to solve has been previously explained in Section 2.1. In summary, this framework aims to rank the provided compounds according to their estimated ΔG, taking into account the IF effect. In order to account for the IF effect we'll perform PELE simulations; to estimate the compounds ΔG, we'll use a SF.

2.3.1 Workflow

The platform workflow has five parts: (i) simulation preparation, (ii) PELE simulation, (iii) pose selection, (iv) re-score procedure, (v) and (vi) other analysis tools. Each and every ONE of these parts is automatized, and most of the code was developed during this thesis, with the exception of the part (ii) which is manual and only uses the group's software PELE.

The workflow depicted in Figure 2.4 represents the aforementioned parts; although the names that appear in the image agree with the names of each of the packages that compose the framework instead of the parts names. Each of the packages implements one of the parts mentioned at the beginning of this subsection.

The reasons for this division of processes is a combination of the amount of computational resources needed for certain steps and the licensing of the external programs used, which prevent the

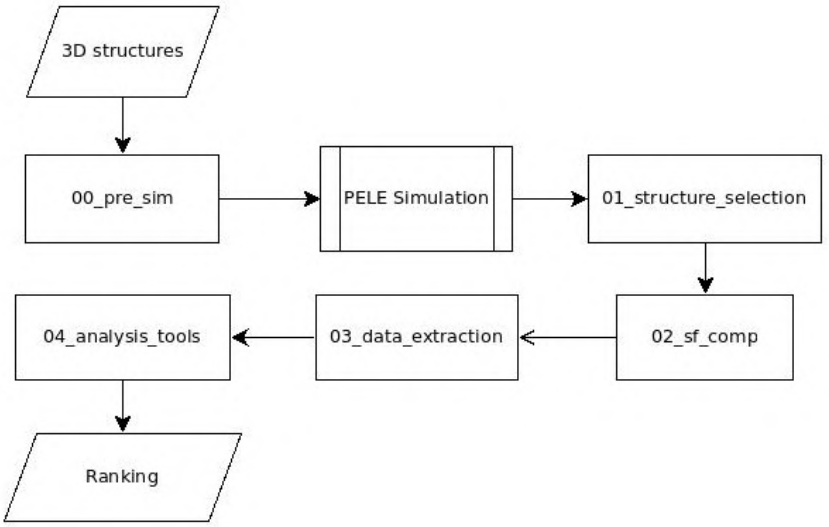


Figure 2.4: The VS framework general workflow. The names agree with the module's names that can be found at the GitHub repository.

installation of the software in certain types of machines, thus, the different steps are run in different machines rendering the complete automatization of the platform too complex for the scope of this thesis. Nevertheless, the development of a wrapper script that allows the complete automatization of the whole process should be an easy task if all the programs are present in one place.

The platform input is a series of files containing 3D structural biomolecule coordinates, which can either be protein-ligand complexes (with the same or different proteins), or a series of ligands and one receptor with which the ligand-protein complex will be formed. Independently of the type of input provided, the first step of the platform will generate a folder containing all the files and data structure needed to perform PELE simulation for each of the files in the input.

The PELE simulation has to be launched by the user manually, wherever they have installed the software. During this simulation, from hundreds to thousands of new structures are generated depending on the parameters, for each of the files provided; the new structures generated represent the same biomolecule as the initial one but with some changes in their conformation (structural rearrangement). Thus, from now on, they'll be called poses.

Then, the platform selects the best pose(s) according to a metric selected by the user. These poses will be extracted into a folder selected by the user, and in this folder there will be one sub-folder per input containing the extracted pose(s) of that system.

The next package will compute the selected SF for each of the inputs provided, generating one output file per selected score and input file.

Lastly, another package will extract the scores values into a csv file, which will rank the different ligands or complexes provided as the initial input.

Each package is composed of several modules that work together for one purpose, with the exception of the package 04_analysis_tools, which is a compilation of simple stand-alone scripts used to perform different types of analysis or tasks, such as plot energy profiles or merge all the reports produced by PELE.

In order to better understand what each part of the framework is responsible of, each of the packages is explained in more detail within the following sections with the exception of the package called 04_analysis_tools. Due to its toolbox nature; all the other packages present several options and can be used in many forms, but always with the same goal.

2.4 Simulations preparation

2.4.1 Problem description

This package meets the need to automatically generate all the files and folders needed to run a PELE simulation.

In order to perform the simulation, PELE requires a pdb file containing the biological system to simulate and the Data and Documents folders. The Documents folder contains information about the format of the control file for the PELE simulation, while the Data folder contains the Templates, RotamerLibraries and solvent information for the common aa that form the proteins, and for the most common molecules present like waters, and some ions.

In order to store these files for our specific system like: the ligand, non-natural aa, heme groups, etc. we can either modify the Data folder and add the needed files or use the DataLocal folder. Due to the size of the Data folder and the fact it's needed for all PELE simulations the recommendation is to have only one general copy of Data, and to use the DataLocal folder to store the data of our specific system.

As we've seen, we need a few files to run PELE simulations; to launch this process manually for one system is not a problem, it can take from 2 to 10 minutes. However, when we need to prepare the files for a thousand systems, the 2-10 minutes become 33-167 hours. Since in a VS campaign we'll work with thousands of compounds the amount of time needed to prepare them manually becomes absurd. Thus, we need to automatize this process.

That is this module's main function: to automatically prepare all the files needed to run thousands of PELE simulations. The script will generate one folder for each input file containing all the data needed to launch a PELE simulation. In order to generate this data structure the user should call the `sims_preparation.py` script inside the `00_pre_sim` folder of the VS_framework repository, which uses a couple other supporting scripts inside the folder and calls a couple external scripts developed by other members of the group.

2.4.2 Workflow

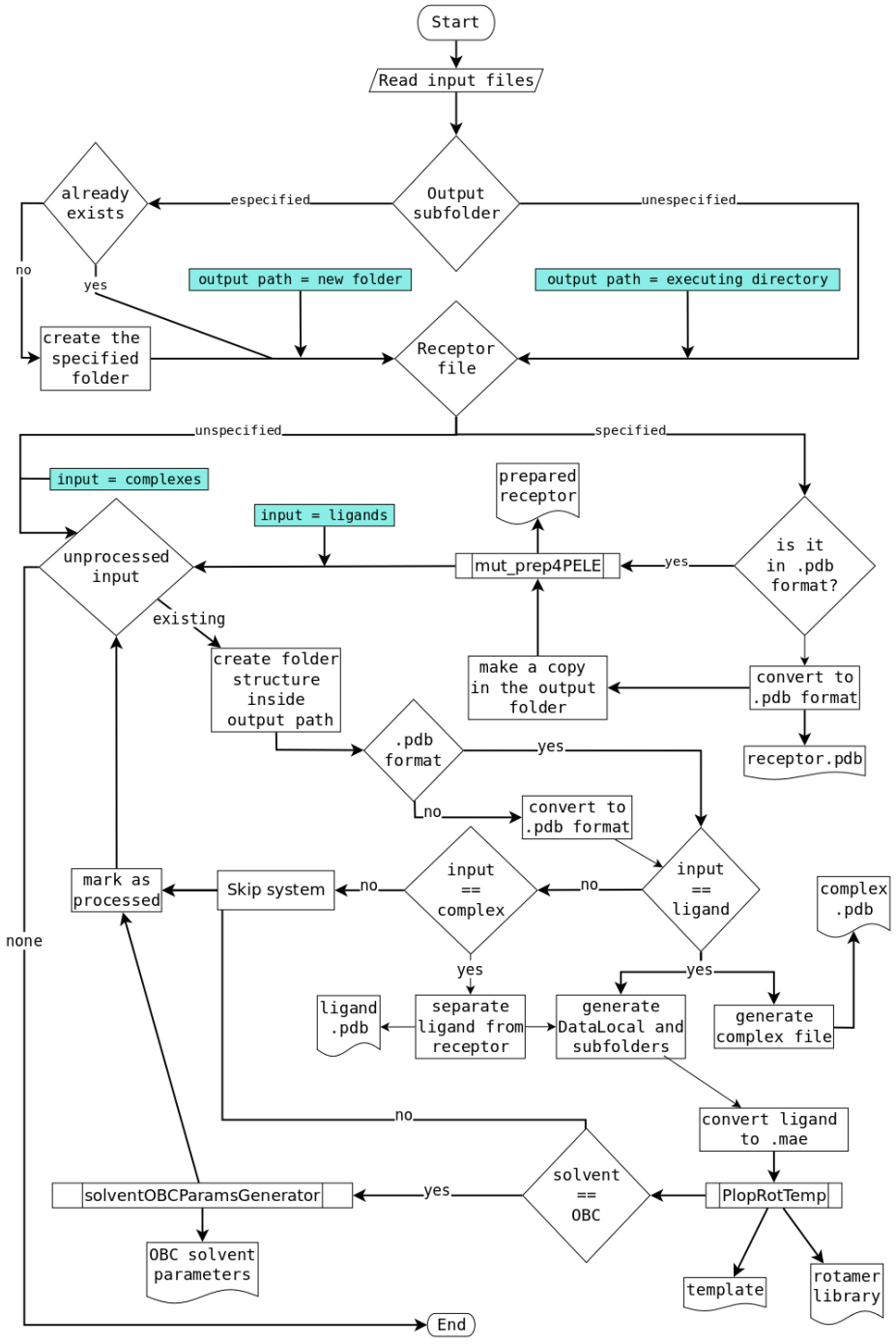


Figure 2.5: Sims_preparation.py flowchart.

In Figure 2.5 we can observe the flowchart describing the program's workflow. The main function of the program is to create one folder containing all the information needed to run a PELE simulation, for each file in the input. In order to achieve this, the program takes several steps and may skip those systems that, for any given reason, cannot be simulated.

The first step is to read all the input files into a list. Next it checks whether the output folder where the data should be generated is created or no. If it doesn't exist it asks the user if it should be created, if so, the program will create the folder, otherwise it will stop.

The second step is to understand which kind of data the program is dealing with as input: either simple ligands or complexes protein-ligand. In order to do so it checks whether the receptor option has been selected or not.

If the receptor option hasn't been selected the program will treat the input files as protein-ligand complexes files and process them. On the other hand, if it has been set, the program will use the structure provided by this option as the protein receptor and it will treat the input files as small molecules.

This means the program will check whether the file provided as receptor is in pdb format, or not, and if it isn't it will convert it to .pdb format. Then, it will launch the external program `mut_prep4pele.py` to ensure that the receptor has the right format to enter the PELE simulation software, and that the chain chosen to contain the ligand (provided as input) isn't already in use on the receptor structure.

The next step is to iterate over all the input files provided until all the files have been processed. For each of the files provided the first step is to pre-process the file, in order to do so two sub-actions are taken: (i) create the basic data structure and (ii) check the format.

To perform the first of these steps, the program will extract a general name out of the complete path provided for the file by using only the name of the file without extension, and then it will create a folder named with the general name inside the output folder. The second step consist in checking whether the file is in one of the supported formats or not, if it isn't in .pdb format it will be converted to it.

After this pre-process of the provided input file, the program will perform different operations: if it's a ligand, the program will generate the ligand-protein complex; and if it's a ligand-protein complex, it will extract the ligand.

In case the input files contain only ligand structures, the program will first call the mut_prep4pele with the ligand to make sure it complies with the naming requisites of the PELE simulation program, such as having unique atom names in case of a non-peptidic ligand, and makes sure the ligand chain matches the one specified by the ligand_chain option. Then the program generates the ligand-protein complex with the receptor provided by merging the processed structures into one. In the case of ligand-protein complexes, the program first runs the mut_prep4pele for the complex, and then it extracts the ligand from the .pdb file using the chain provided by the ligand_chain option.

On the next step, if everything has proceeded correctly the new folder created should contain at least two files: the ligand-protein complex in .pdb format, which will be used for the PELE simulations, and the ligand in .pdb format, that will be used to generate the ligand templates for the PELE simulation. After this, the ligand file is converted to .mae format using the Schrödinger converter in order to be able to launch the PlopRotTemp program and generate the PELE templates, unless the option no_templates is used, in which case the file won't be converted and the templates won't be generated.

Next, the program now generates the DataLocal folder inside the new subfolder and the subfolder structure needed by PELE. Afterwards it calls the PlopRotTemp external program in order to generate the PELE templates needed for the simulation in a temporary folder; then these templates are moved into the DataLocal subfolders.

Finally, if the simulation uses the OBC solvent the corresponding subfolder and templates will be generated also in the DataLocal folder, by means of the `solventOBCParamsGenerator.py` script.

2.4.3 Implementation

Dependencies

The script has the following dependencies:

- The ProDy 1.8.X library to read and manipulate pdb files.
- The module `mut_prep4pele.py` see section 2.2
- The `PlopRotTemp_S_2017` module
- Schrödinger Software: Maestro Academic version or private version
- The `obc_param_generator.py` script developed by the group.

Modules

This package is composed of the module `sims_preparation.py` as the main module and other 4 auxiliary modules. It also has a dependency on the external module mut_prep4pele. The module's relationships are depicted in Figure 2.6.

The module enviroment_parameters contains all the information about the default values to use for the program. It includes the path to the external software required by the program and the default values of the optional arguments.

The constant_values module contains the values of the internal constant variables that won't change unless the program is modified.

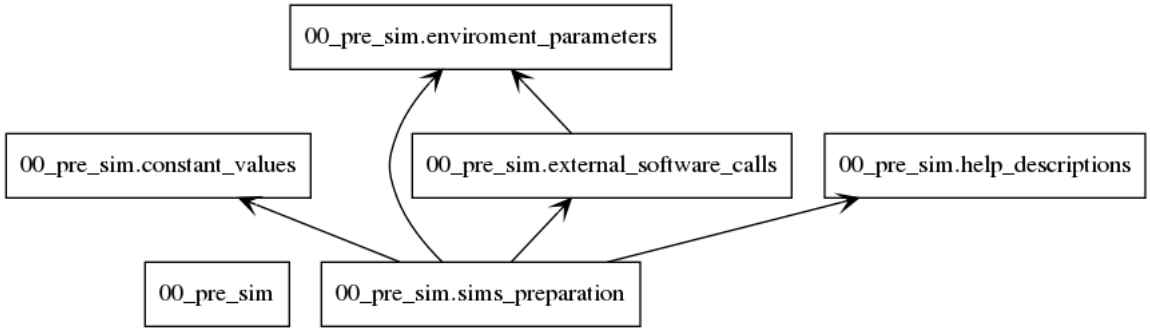


Figure 2.6: Modules relationship for the 00_pre_sim package.

The external_software_calls contains all the variables used to call the external programs. These variables contain the general format of the commands to launch the external programs; they're completed with the system specific information by the main module.

The help_descriptions contains the descriptions of each of the options in the program and it's used by the main module to generate the help messages.

Input and options

The program has 4 groups of arguments as input. A mandatory argument group consists on the input files; an optional arguments group encompassing all the arguments related to the behaviour of the program options; a PELE options group, which covers all arguments related to PELE simulation characteristics; and an External Software path which involves the options used to specify the complete paths to the external software called by the program.

Mandatory arguments

The script's only mandatory argument is the *input_files*, or *input*, which should consist in a list of 3D structures to process. If the structures provided are protein-ligand complexes with the ligand in the default chain to be used, it doesn't require anything else. Otherwise, if the ligand is present in a different chain, the user should use the ligand_chain to specify the correct chain. Contrary, if the provided structures are formed only by ligands, the user should provide a protein to form the protein-ligand complexes using the receptor option. The program will consider all the input complexes or single ligands depending on whether the receptor option is specified or not, and it isn't able to work with both types of inputs simultaneously.

Optional arguments

All the parameters in this group are optional arguments and, if the user doesn't use them, the program will use the default values. These parameters are used to define the type of input provided, where the output should be written, and which steps of the process can be omitted. Summarizing, the parameters in this group are used to set up the behaviour of the program.

receptor takes the complete path to a 3D structure to use as the protein receptor for the protein-ligand complex during the PELE simulations. The structure should be of proteic nature, even though it may contain waters and/or ions, and it should be encoded in any free format such as pdb or mol2 or in the Schrödinger's proprietary format .mae, and it should be correctly aligned with ligands provided as input. This option should be used only when the files given with the input_files option contain 3D structures of only ligands. Its default value is an empty string meaning the input files will be treated as protein-ligand complexes.

subfolders_path is used to specify the path where the files should be created. This path should point to a pre-existing path or to a new folder inside a pre-existing one. Its default value is `./`, which means the program will generate the folders for each input file into the execution directory.

conf_template or *conf_file_template* takes one single argument consisting in the complete path to a PELE control file template. This template is going to be used by the program to generate as many PELE control files as needed in order to launch one PELE simulation for each input file. The default value should point to the template provided with the platform, and it should be modified upon installation by changing the variable conformational_template inside the `enviroment_parameters.py` file.

conf_file_suffix takes as argument a string containing any desired suffix to add to the PELE's control file's names. This option is useful when the user wants to create new control files inside folders with a previous PELE simulation without rewriting nor re-using the previous control files and in order to reuse the templates. If this option isn't used the control file will be systemID_solvent.conf where the systemID derives from the input file name, if it's a single core simulation and systemID_pele_adaptive_sampling_.conf in case of an adaptive PELE simulation.

adaptive_sampling takes the complete path to an adaptive PELE's control file's template as single argument. This template will be used to generate the control file required to launch an adaptive PELE simulation.

ligand_chain takes one argument: a letter from the alphabet. This letter should specify the chain where the ligand is (if the input consists of complexes) or should be (if the input are ligands) in the protein-ligand complex 3D structure needed as the PELE simulations starting point. The specified chain should contain only the ligand and unless it is from peptidic nature, it should be formed by one single residue. The default value (specified by the ligand_chain inside `enviroment_parameters.py`) is Z, in order to ensure that the ligand is the last molecule on the 3D file; a requirement of external programs used in later steps.

no_templates is a flag used to tell the program not to generate the templates needed by the PELE simulation program, which by default are generated. This option is of interest in order to save computational time (this is the most time-consuming step of the program) when dealing with thousands of compounds with previously generated templates.

rewrite is a flag to make the program rewrite any pre-existing file. Whenever a pre-existing folder is used the program will search for and re-use any pre-existing file unless this option is used.

not_interactive is another flag option to avoid any possible question to the user. When this option is used the folder where the data will be generated must already exist.

log_file takes the complete path to the log file generated by the program. By default, the program will generate a file called sims_prep_log.txt inside the path where it has been launched.

debug is a flag to make the program print several messages at some of the program's steps. This option is of use only to the developers of the software.

PELE options

The arguments in this group are used to set up the fields in the PELE control file. All these fields will be encoded by the different variables present in the control file template. The arguments can represent values that change with the input, directly represent the value of a variable from the template file, change depending on where the PELE software is going to be run or are values susceptible to changes from one simulation to another.

pele_folders takes the complete path to the folder containing the `Data` and `Documents` folders from PELE.

pele_license takes the complete path to the PELE license folder to use for the simulation. It'll be used if the PELE control file template contains the variable license.

every or *fix_every_x_atoms* takes an integer. This integer is the number of residues without constraints and will be used only if the harmonic_constraints variable is present in the control file template. The default number of residues between constraints is 10; due to the implicit solvent scheme, PELE typically adds a small constraint every 10 alpha carbons, to prevent the collapse of the system while maximizing interchain contacts.

constraint or *constraint_strength* takes a float number as its only argument. This number is used as the constraint strength value to use when the harmonic_constraints variable is present in the PELE control file template. Its default value is 0.2 Kcal/$\mathring{A}^2$.

atoms2constraint takes a list of atom's names; these names must be specified within quotes, of 4 letters length, and include the necessary blank spaces. For each of the names provided with this option the program will generate a harmonic constraint every N (set by the option every)

residues on the control file as long as the harmonic_constraints variable is present in the PELE control file template. The default value consists of a list containing only the " CA " atom name, which is the most commonly used atom for constraints.

External Software paths

This group encompasses all the options used to specify the path for the external software required. All the default values for these parameters are set up in the `enviroment_parameters.py` file, and they should be updated by the user upon installation to match their enviroment. All the calls to the external programs are set inside the `external_software_calls.py` file and only should be modified if the user has knowledge about the external software used.

schrodinger_path takes the complete path to the folder where the Schrödinger software is installed (the academic version is enough for this program). This folder default value is set up with the variable schrodinger_path, since this is a proprietary software it may require a license that isn't provided with this platform.

plop_path takes the complete path to the `PlopRotTemp.py` script. This is a public software and can be freely downloaded from the repository PlopRotTemp_S_2017 in GitHub (`https://github.com/danielSoler93/PlopRotTemp_S_2017.git`)

mutations_program_path takes the complete path to the `mut_prep4pele.py` program. This external program was also developed during this thesis and it takes care of the structure preparation so it can be used with the PELE simulation software. It is available as open access software at GitHub at the address: `https://github.com/Jelisa/mut_prep4pele.git`

obc_param_generator takes the complete paths to the script to generate the OBC solvent parameters. This external script is called `solventOBCParamsGenerator.py` and has been developed by other PhD students from the group.

2.5 PELE Simulation

2.5.1 PELE

As mentioned in the introduction PELE is a simulation program to model the ligand-protein or ligand-DNA interactions. It's been used successfully to describe the mechanism of several enzymes and to find the right binding modes of drug candidates. The PELE program uses a MonteCarlo approach to explore the ligand's conformations performing random rotations and translations of the ligand and then adapting the sidechains of both the ligand and the protein. Finally, it performs a minimization and accepts or rejects the changes using a Metropolis criteria.

In order to model the ligand and proteins, the program uses a Template for: each aa in the complex, the ligand and one for each type of molecule not included on the standard aa. The

Template file contains the information about the aa that the program needs to model it. The information encloses the type of bond, the length, the angle and the dihedral of the bond for each atom in the aa; it also contains information about the radius of the atom and the different energies of the atom.

The program also uses a RotamerLibrary file containing the possible angles the sidechains can take for the protein and the ligand.

How to launch the simulation is quite easy, the user only has to call the appropriate version of the PELE program he wants to use and provide it with the configuration file. Although it's a simple step depending on where the software has been installed to launch automatically several instances of PELE will have to be done using different queue systems for the cluster or HPC system used. Thus, it's highly dependent on how the user has installed the program. This is the reason why this step hasn't been automatized or implemented in this platform.

2.6 Structure selection

2.6.1 Problem description

The PELE simulation generates from hundreds to thousands of structures, depending on the type of simulation: single core, mpi or adaptive; the length of the simulation and the number of processors used. The longer the simulation, and the more processors used, the more poses/structures will be generated.

This program reviews the output from any kind of PELE simulations and selects structures from the trajectory based on several criteria, such as minimum binding energy, distances, RMSD or energy clustering, etc. Then it creates one folder for each selected structure containing a pdb file with the structure. It also generates a file containing how many structures have been selected for each folder in the input, a file containing the PELE energy (binding or normal) of the selected structure and if the option is chosen it will also extract the initial PELE energy (binding or total energies).

2.6.2 Workflow

The Figure 2.7 shows the main program's general workflow. This program processes the output files generated by PELE.

The first step the program does is to parse the options provided to it; these options are used to specify the type of PELE simulation previously run, and whether the program should extract the structures or just the values.

If the program is asked to extract the poses' structures and not only the statistics it will generate a folder where each of the selected poses will be placed inside its own sub-folder.

Then, it iterates over all the paths provided as input until all of them have been processed. The

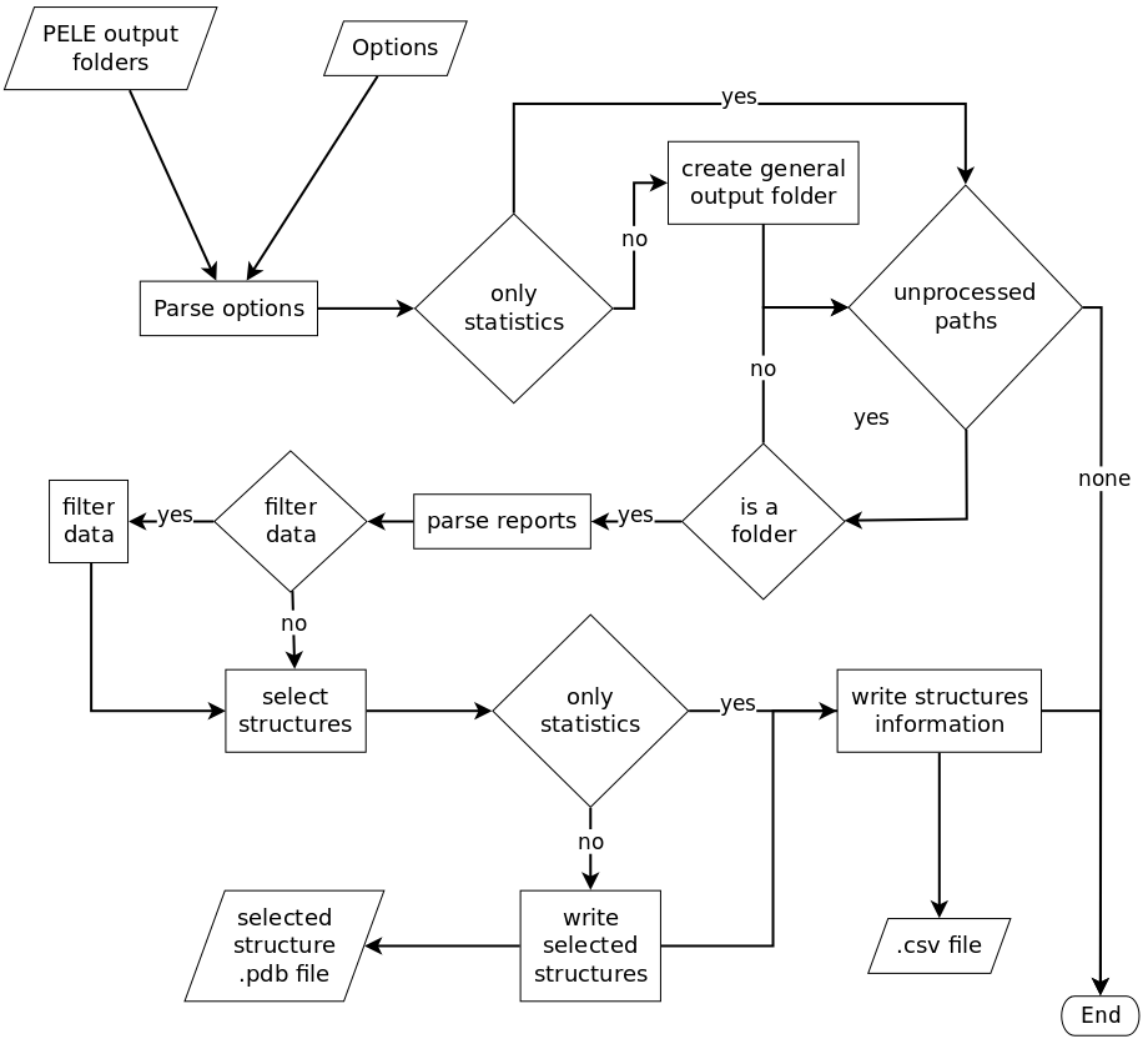


Figure 2.7: structure_selection.py flowchart.

first step in this iterative process is to check whether the path provided is actually a folder or not; if the path isn't a folder it will be skipped.

Next, the program parses the report files generated by PELE. All possible PELE runs generate similar report files, but how many there are, and how they are structured in folders, depends on the type of PELE simulation run. The single core PELE simulation only generates one report file and one trajectory file inside the output folder; while the mpi simulations generate one report and trajectory per CPU used for the simulation, and the adaptive methodology generates one folder per epoch (see section 1.3 for more information) and inside each of these folders it generates one report file and one trajectory per CPU used.

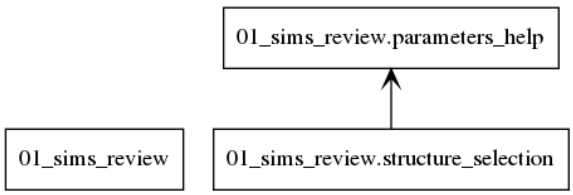


Figure 2.8: Modules relationship for the 01_sims_review package.

Afterwards, the program will filter the data if required to, and then it will select the best structures according to the selected methodology, using the report information.

Finally, it will generate the output files. First, it will create a subfolder inside the general output folder and write into it the selected poses, unless the option only_statistics has been selected. Secondly, it will generate one .csv file containing all the information present in the report files for each of the selected poses.

2.6.3 Implementation

The script has the following dependencies:

- The pandas library

Dependencies

Modules

This package is formed by two modules: the main module called `structure_selection` and the auxiliary module `parameters_help`. The `parameters_help` module contains the description of each of the options of the program while, the `structure_selection` module implements the workflow depicted in the Figure 2.7.

Input and Options

positional arguments:

minimum_energy, rmsd_clustering, energy_clustering The different criteria to select the structure(s).

optional arguments:

-h or *help* show the help message and exit

-input A list containing the path to the output folders from any kind of PELE simulations to analyse. I.E.: /path/to/pele/simulation/system_1/output_obc/ (default: None)

output_folder The directory where the program will create the folders containing the selected structures (one for each structure selected). It's mandatory (default:)

output_prefix The prefix to use for the output files containing the metrics from the report one for the initial model and another for the selected_model. The names are: (prefix_)initial_models_pele_metrics.csv and (prefix_)selected_models_pele_metrics.csv (default:)

ligand_chain The name of the chain where the ligand can be found. The ligand should be the only molecule in this chain. (default: Z)

only_statistics This option is a flag, thus, if present the program will only compute the number of structures selected and their energies according to the selection criteria. It won't extract the structures. (default: False)

total_energy_deviation The maximum deviation in KCal of the minimum total PELE energy possible. This criteria is always present and is combined with any selection criteria. With this we avoid picking unfeasible systems from the non-converged part of the simulation or the artefacts that the adaptive protocol may create. (default: 1std)

initial_energies When this option is present the script will write a file containing the PELE binding energy for the initial complex given to PELE called initial_PELE_binding_energies.csv. (default: False)

simulation_type single_core, adaptive, mpi This option specifies which kind of PELE simulation has generated the output folder given in the input: single_core refers to a PELE simulation run with only one processor (traditional PELE), mpi refers to the mpi version of PELE and adaptive to the adaptive script to launch PELE. (default: single_core)

log_file The name for the log file. (default: structures_extraction.log)

2.7 Re-score procedure

2.7.1 Problem description

PELE provides us with its own energy score, and we can use it to estimate the binding energy of compounds, but this energy is an all atom energy designed to score the poses in order to find the best one. As such, the larger the ligand, the larger the energy (in a somehow excessive additive manner); this energy also increases when the ligand presents charges able to interact with the protein. These two issues make the PELE energy unsuitable to compare the ability of two compounds to bind the protein.

Thus, in order to rank the compounds according to their Binding Energy (BE) we want to compute other SF designed to estimate the compounds ΔG value.

As stated in the introduction, there are several types of SF. This program is able to launch the computation of the following SF: vina, xscore, DSX, mmgbsa and glide, which we selected aiming for diversity of methods and to include the most used ones. It computes the selected scores for each of the structures provided by the input using the score in-place protocol.

It also prepares the files and folder structure needed to launch the RF-score manually. This score has to be launched one single time, unlike the other scores which need to be launched several times. In addition, the original software requires editing of the R script in order to use new files. Thus, we've decided to launch it manually from inside the Rstudio GUI. Due to the performance of this score we decided to stop using it, so it hasn't been fully integrated.

Finally, this program is capable of launching the program binana. This program isn't a SF, since it doesn't estimate the ligand affinity, instead, it describes the ligand-protein interaction with a series of important physical and electrostatic descriptors of the interaction, although the platform only extracts the physical ones. We can use this descriptors in order to generate a new SF, or to create a classification method.

2.7.2 Workflow

The Figure 2.9 depicts a graphical summary of the main module's workflow; which is explained in more detail in the following lines.

The program receives as inputs a list of files and several options, and the first step it does is to parse the options provided. From the options selected by the user it extracts the scores chosen to be computed.

If the user has chosen to compute the Glide score, the program will generate as many folders as needed, and one control file for each of them. The number of folders needed is computed by the formula N_F/n_{max} rounded up, where N_F is the number of files provided and n_{max} is the number provided by the option glide_max_structures_per_run. The reason to break the Glide calculation into several processes, one per folder created, is that we use the xglide script, which fails when too many structures are provided to score in place.

If the user has selected to compute the vina score, the program will read the templatized control file for vina and obtain all the information needed to create a system-specific control file later on.

If the user has chosen to prepare the RF-Score files, the program will try to read a file containing the experimental ΔG values of the compounds.

Next, for each of the files provided, it performs the following steps and checks: if the file isn't in pdb format it will mark the system as an error and skip it; otherwise, it will check if the Glide score has been chosen and if so it will create a symbolic link to the pdb file inside the previously prepared folder.

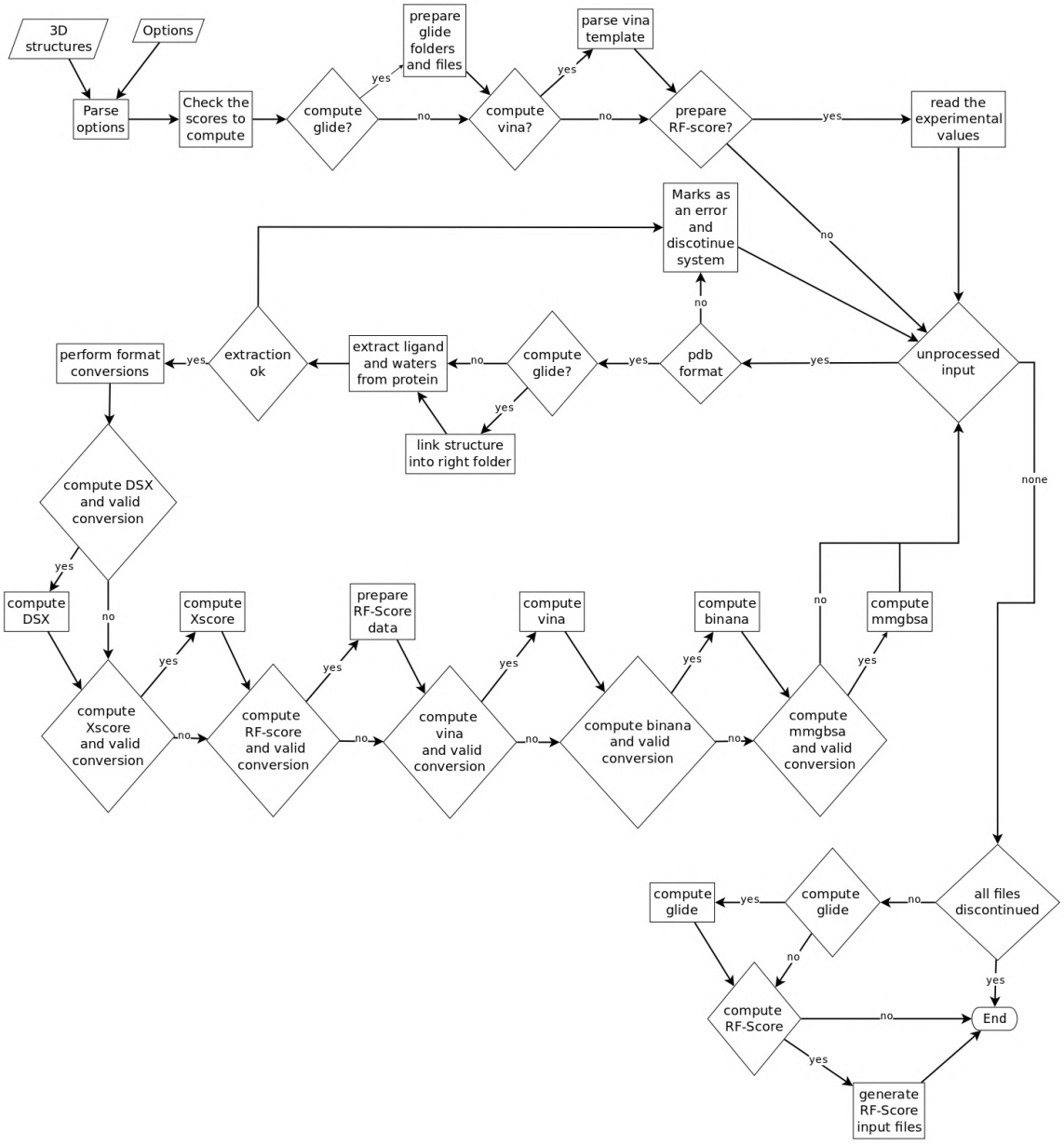


Figure 2.9: compute_scoring_functions.py flowchart.

Afterwards, it will extract the ligand from the file provided (1 file). It will also extract the protein with and without the water molecules (2 files) and the water molecules (1 file). Thus, in this step, it will generate from 2 files (if the complex doesn't contain water molecules) to 4 files (if it contains water molecules). In order to extract the ligand, it uses the information provided by the

option ligand_chain. If during this process anything fails, the program will mark the system as an error and skip it.

Next, the program performs all the format conversions needed to perform the chosen scores. Each of the scores needs the information about the protein-ligand complex in a specific manner, which can be checked on each score's own documentation.

Next, the program checks which scores have been chosen and launches each one of them. For each of the following scores: DSX, Xscore, Vina, binana and mmgbsa the program checks if they've been chosen, and if the conversions they require have been done correctly. If both conditions are met each score is launched one after the other. For the RF-score the program makes the same checks as for the other scores, but instead of launching the score, it will extract the information about the systems needed to generate the input file for the score.

If all the files have been skipped the program will end. Otherwise, it will check once again if Glide has been chosen, and if so, it will launch the score as many times as the number of folders created to this end. Afterwards, it will check for the RF-score and, if chosen, it will generate the file needed as input by this score.

2.7.3 Implementation

Dependencies

The script has the following dependencies:

- The DSX [67] software
- The X-Score [76] program
- The AutodockVina [74] program and its associated library MGLTOOLS.
- The binana script [15].
- The prime software from Schrödinger's suite [42].
- The Glide software from Schrödinger's suite [23] [33].

Modules

As we can see in Figure 2.10 this package is formed by 4 modules which interact among them; being the compute_scoring_functions module the main one, with three auxiliary modules.

The module external_software_paths contains the complete paths to the external dependencies this program has. This module should be modified upon installation of the software to match the user's paths to the folders. It should be noted that none of the scoring software packages is actually provided with this software due to licensing issues.

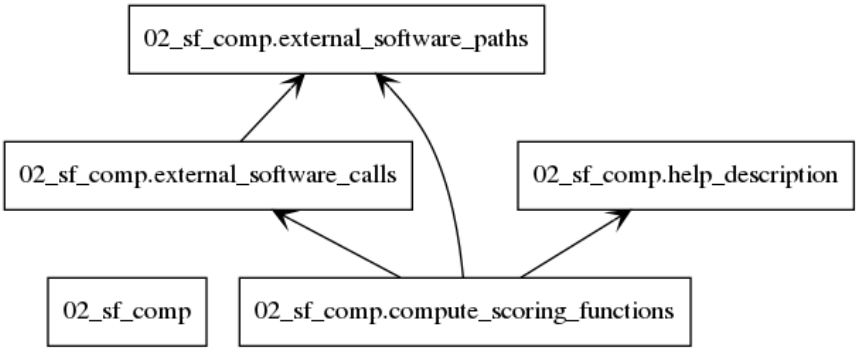


Figure 2.10: Modules relationship for the 02_sf_comp package.

The `external_software_calls` contains the strings used to launch the external SF software. These strings contain the general form of the commands to be used if we were calling them manually using the command line. They are completed using the paths provided by the `external_software_paths` and with the system specific information in the main module.

The `help_description` module provides the description of each of the program's options.

The main module `compute_scoring_functions` is the one that actually parses the options and creates the help, thus using the module `help_description`, it's also the one that launches the scores, thus using the other two modules.

Input and Options

h or *help* show the program's help message and exits.

input_files A list of pdb files containing ligand-protein complexes. (default: None)

folders_path The path where the Glide score computations will be prepared and launched. (default: ./scoring_functions_values)

scoring_functions A list containing the names of the SF to compute. The implemented ones are:[glide, vina, xscore, dsx, mmgbsa, binana, rf_score] (default: ['glide', 'vina', 'xscore', 'dsx', 'mmgbsa', 'binana', 'rf_score'])

schrodinger_host The name of the machine used to run the Glide computations, in most cases the default value will work correctly, but it all depends on the user's enviroment. (default: localhost)

schrodinger_cpus The number of cpus to be used in each xglide job. (default: 1)

glide_max_structures_per_run The maximum number of structures to run in one Glide job. If the user provides more input files than this number the program will launch as many xglide jobs as the number of input files divided by this number, rounded up. (default: 250)

vina_box_distance_to_ligand The number of angstroms used by Vina to define the docking box. (default: 20)

experimental_deltag The path to a .csv containing the δG values for each of the input files. (default:)

rf_score_output_file A string containing the name of the file where the input for RF-Score should be written. (default: -2)

debug A flag option that will print helpful messages to discover where the program is failing. Recommended only for developers. (default: False)

ligand_chain A string containing the name of the chain where the ligand is located on the input files. (default: Z)

log_file A string containing the complete path to the file where the log of the program should be written. (default: scoring_function_computations_log.txt)

rewrite When this option is present the program will rewrite any pre-existing file. Otherwise if a file already exists the program will use it, instead of generating it again. (default: False)

2.8 Scores extraction

2.8.1 Problem description

This script processes the output files generated by the `compute_scoring_functions.py` script. Each of the SF computed by the previous script generates a complex output file; which needs to be processed in order to extract the actual value of the SF. With this module the user can select to extract an individual score or any combination.

2.8.2 Workflow

The program's workflow is depicted in the Figure 2.11. For each of the provided file types the program will parse and process the corresponding files.

2.8.3 Implementation

Dependencies

The script has the following dependencies:

- Python 2.7

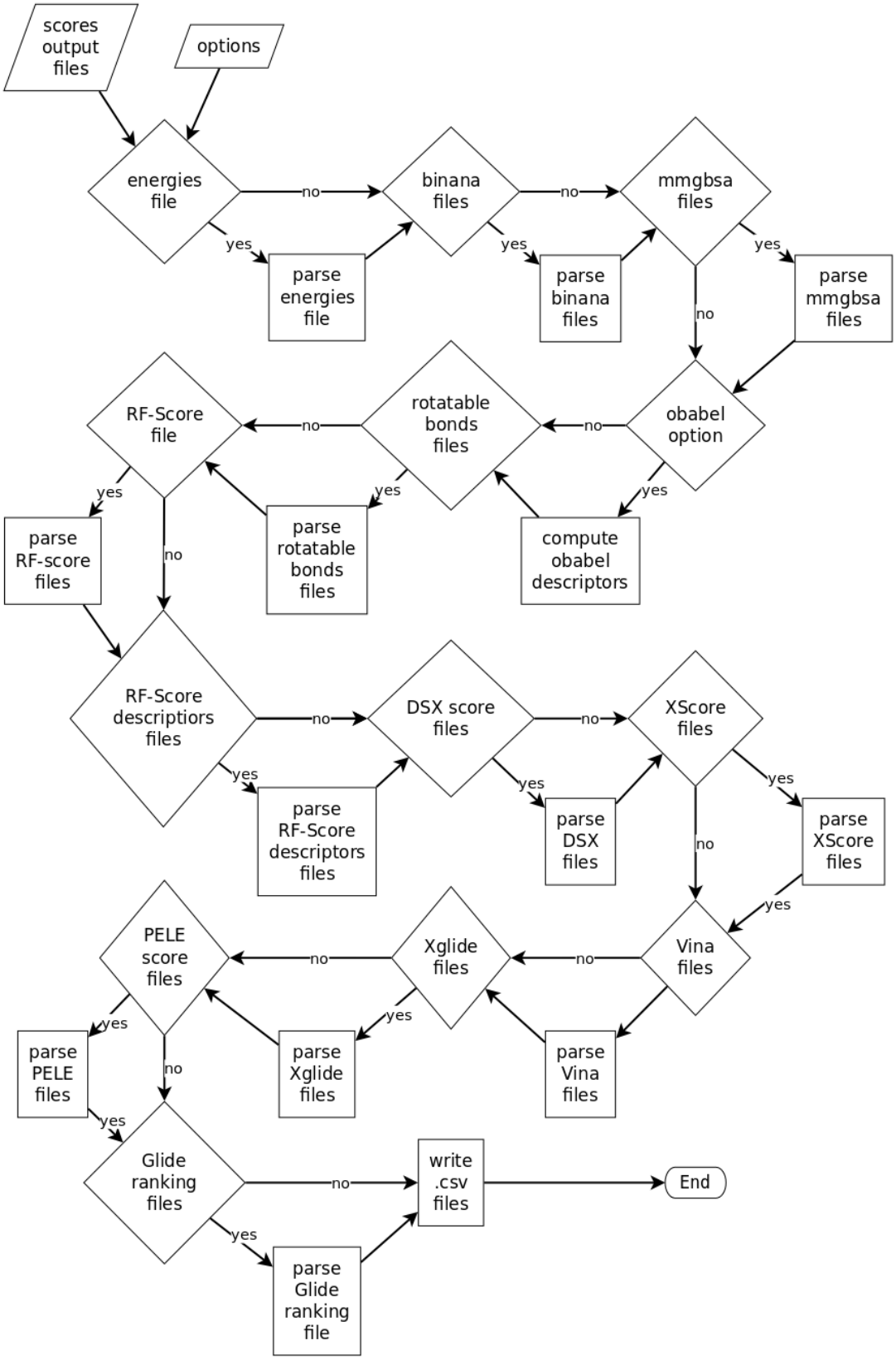


Figure 2.11: extract_sfs.py flowchart.

Modules

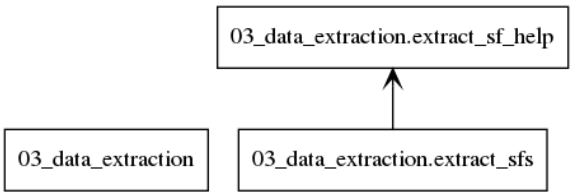


Figure 2.12: Modules relationship for the 03_data_extraction package.

This package presents only two modules: the main one called `extract_sfs` and an auxiliary module called `extract_sf_help`.

The main module implements the workflow depicted in Figure 2.9. It reads the options description from the module `extract_sf_help`.

Input and Options

h or **help** show this help message and exit

dsx_file This option specifies the files to use for the extraction of the dsx values. (default: False)

xscore_files This option specifies the files to use for the extraction of the xscore values. (default: False)

binana_file This option specifies the files to use for the extraction of the binana values. (default: False)

mmgbsa_files This option specifies the files to use for the extraction of the mmgbsa values. (default: False)

rf_as_score_file This option specifies the files to use for the extraction of the rf_score values. (default: False)

rf_descriptors_file A file containing the descriptors used by RF-Score. (default: False)

vina_files This option specifies the files to use for the extraction of the vina values. (default: False)

xglide_files This option specifies the log file from xglide to use for the extraction of Glide values. (default: False)

glide_ranking_csv_file This option expects two terms. One is a .csv file generated by exporting the spreadsheet from Maestro, which should contain the fields Title and docking score at least, and it should have the same order as the order used when extracting the compounds to launch the simulations. The other term is going to be used as the prefix for the simulations id, so

it should correspond with the prefix used as base name when exporting the structures from Maestro. (default: False)

pele_file The pele file containing the scores to use (default: False)

pele_mean_file The pele containing the mean PELE BE score to use for all systems (default: False)

mmgbsa_as_sf If this option is present the program will extract the total value of prime-mmgbsa as a scoring function, otherwise it will extract only the descriptors. (default: False)

rotable_bonds_file The .pdbqt files to analyse and extract the rotable bonds from. (default: False)

energies_file This option specifies the file containing the experimental energies for all the systems. (default: False)

obabel_desc This option computes the descriptors: logP, MW, TPSA, using pybel, from the ligand pdb files. (default: False)

convert or **convert2deltaG** This option establishes the scoring functions to convert to energy value from pkd. (default: ['xscore', 'nn_score'])

conversion_temperature or **temperature** The temperature value to use when converting from pkd to energy. (default: 300K)

conversion_r_value or **R** The r value to use when converting from pkd to energy. (default: $0.002\frac{Kcal}{K*mol}$)

output_general_name The complete path to where the output should be written plus the prefix for the files. (default: None)

ensemble A flag to select the id for the systems: if true, the pattern 'word_number_number' will be used, otherwise the pattern used will be 'word_number'. (default: False)

log_file A complete filename for the log file. (default: sf_extraction_log.txt)

common_systems_list When this option is selected the program will generate a file containing the list of the systems for which all the scoring functions and all descriptors were correctly computed. The name will be output_general_name+_common_systems.csv (default: False)

2.9 Conclusions

The PELE VS framework has been developed, allowing the use of PELE during VS campaigns in an easy and quick way.

This platform is capable of performing all the steps needed to automatically prepare and analyse PELE simulations in a quick and easy way.

In order to prepare the simulations of 2000 compounds from scratch, that is, generating all the Templates and Rotamer libraries (which is the most time-consuming step of the preparation), the platform uses around 1 to 2 hours depending on the compounds properties, while doing the same process manually can usually take 1 week of work. Thus, this platform considerably speeds up the process.

The modular structure of the platform allows the user to use different computers for each of the steps, and it allows for an easy addition of new steps and functions. In addition, all the code is open source, which allows users with new systems to adapt it to their needs.

The platform can easily be adapted to work with other simulation programs as AMBER [68] or CHARMM [9]. It can also be used just to perform a score procedure, without using any simulation program, in order to create a personalized consensus score or to choose the best score for the user's system.

Chapter 3

DUD-e Study

One of the initial objectives of this thesis was to obtain a general PELE simulation protocol capable of enriching the results of traditional VS campaigns. In order to obtain it, in this chapter we'll study how the introduction of PELE affects the VS process.

For this study we'll use part of the enhanced Directory of Useful Decoys (DUD-e dataset) [66], a standard dataset in the VS field to test whether we can use a single PELE simulation protocol to improve the ratio of true positives in the first 50 or 100 compounds of the docking results for all the different targets we study.

3.1 The DUD-e dataset

The DUD-e dataset is composed of 102 different targets which can be grouped into five different categories: GPCR, kinase, NHRs, proteases and diverse. Each of these categories represents a family of proteins, with the exception of the diverse category, which is a mix of proteins with different functions.

For each of the categories there is a variable number of proteins or targets; all of them are relevant proteins for the development of new drugs. For each target, there are an average of 224 ligands with known activity, and for each ligand there are 50 decoys with similar physico-chemical properties, but dissimilar 2-D topology based on the extended connectivity fingerprint. The physicochemical properties kept are: molecular weight, calculated LogP, hydrogen bond (h-bond) donors, h-bond acceptors, number of rotatable bonds and net molecular charge.

Due to the computational cost of our approach and the amount of resources needed to work with 102 different receptors, we decided to work only with a subset of the DUD-e dataset. Initially, we picked 3 receptors for each of the subsets provided by the DUD-e: for the GPCR coupled receptor family we picked the proteins Androgen receptor 1 (adrb1), Androgen receptor 2 (adrb2) and Dopamine D3 receptor (drd3); for the kinases family we selected the proteins Tyrosine-protein

kinase JAK2 (jak2), Cyclin-dependent kinase 2 (cdk2) and Serine/threonine-protein kinase WEE1 (wee1); for the proteases family we selected the Trypsin 1 (try1), Thrombin (thrb) and hiv protease; for the NHRs receptors family we initially selected the Peroxisome proliferator-activated receptor gamma (ppar), and mcr, later (due to our collaboration with AstraZeneca) we extended this family to include the Androgen Receptor 1 (andr1) Estrogen Receptor 1 (esr1), Glucocorticoid Receptor (gcr), Progesterone Receptor (prgr) Retinoid X Receptor Alpha (rxra) and Thyroid Hormone Receptor Beta-1 (thb) receptors. Most of the receptors structures we've used are the ones provided by the DUD-e itself, with two exceptions: the hiv protease and the Mineralocorticoid Receptor (mcr) proteins.

In the hiv protease case we started using the one provided by the DUD-e named hiv but some preliminary results (not shown) led to a more careful study of the protein. During this study, we discovered that this protein requires a water molecule in order to bind 99% of its ligands; but the structure provided by the DUD-e pertained to the complex between the hiv protease with one of the few ligands capable of displacing this otherwise crucial water. In order to better reproduce the correct binding pose of most ligands, we chose to use a .pdb file corresponding to the most common binding mode, which includes the interacting water molecule (which was kept for the simulations). This new receptor was called HIV protease containing one water molecule, which is responsible for the most common conformation of the binding pocket (hivw).

For the MCR receptor, we haven't used the structure provided by the DUD-e, but two different structures that have been used in our collaborations with AstraZeneca Sweden; which present significant differences in the pocket, which we'll be calling: MCR-in (PDB code 2OAX) and MCR-out (PDB code 4UDB). The MCR-in structure presents a smaller pocket with the MET-852 pointing towards the interior of the binding pocket, while the MCR-out presents a bigger pocket respectively with the MET-852 pointing outside the protein.

3.2 Initial Protocol

For each of the selected systems we performed a flexible-ligand rigid-protein docking using the docking program from Schrödinger Glide with the SP protocol [23, 33]. Out of the docked structures we've selected the best 1000 compounds to perform short PELE simulations with them, but since some of the known ligands had several protonated states or even several stereoisomers (called variant from now on) we ended up performing more than 1000 simulations per protein. Even though the DUD-e dataset has approximately around 2% of actives, the subset that we used had a variable percentage of actives ranging from 2.6% to 22.3%, the specific value for each protein can be checked in Table 3.1.

The huge variability of actives in our subset is a direct consequence of the variable performance

of the Glide protocol over different receptors. The assessment of the reasons for this variable performance, and its correction to obtain similar results over all families, is beyond the scope of this section. With the work described in this section we tried to develop a general method to improve the ratio of true positives for all the receptors.

The first test we performed was to run a single core short PELE simulation for the best 1000 compounds, using all their variants, over 14 selected receptors. This first simulation only had 200 steps, with steering of the ligand, a temperature value of 2000, constraints every 10 C_α (to avoid the collapse of the protein) and the VDGBNP solvent. This simulation protocol generates from 60 to 100 poses for each compound; from all these poses we selected the best ones according to their PELE Binding Energy (PELE BE).

Table 3.1: Selected compounds dataset characteristics

Family	Protein	# of Compounds	# of Actives	Actives %	# of Inactives
GPCR	Adrb1	998	93	9.3	905
	adrb2	1099	121	11.0	978
	Drd3	1000	47	4.7	953
KINASE	Cdk2	999	223	22.3	776
	jak2	1000	80	8.0	920
	wee1	1000	101	10.1	899
PROTEASE	hiv	999	222	22.2	777
	hivw	952	117	12.3	835
	try1	998	157	15.7	841
	thrb	999	187	18.7	812
NHRs	mcrin	998	26	2.6	972
	mcrout	999	38	3.8	961
	ppar	999	183	18.3	816
	andr1	1000	122	12.2	878
	esr1	997	191	19.2	806
	gcr	999	58	5.8	941
	prgr	1000	165	16.5	835
	rxra	999	106	10.6	893
	thb	1007	64	6.4	943
DIVERSE	aces	994	41	4.1	953
	hs90	999	42	4.2	957
	nram	999	50	5.0	949

For each of the poses coming from the Glide docking and for those selected from the simulation we computed the following SF: Glide [23, 33], Vina[74], DSX[67] and Xscore[76]. The Glide and Vina scores are usually associated with a docking protocol, but in our case we just computed the score using the structure derived from the Glide docking or the PELE simulation. The computation of the scores before and after performing the PELE simulation allows us to study how the rankings

of compounds change among scores, and with the IF effect.

3.3 Results

3.3.1 Enrichment Factor (EF)

To check how good a method is we use the Enrichment Factor (EF) computed by equation 3.1, where TP_s is the number of true positives in the S top compounds; S is the number of compounds at the top positions of the ranking that we select; TP_T is the number of true positives on the entire set and N is the total number of compounds in the set. This measure gives us an idea of how well a method behaves when compared with random selection.

This metric ranges from 0 to infinity, and it's highly dependent on the percentage of true positives in the sample. This lack of a closed range makes the comparison between the different receptors, which have a variable percentage of true positives, really hard. Thus, we'll use the $\%EF_{max}$ to compare the enrichment factors (EFs) between different systems, this metric normalizes the EFs value by dividing the EFs observed over the maximum EFs possible. The maximum EFs possible for a receptor is the case where all the compounds in the top S positions are positives, or it includes all the positive compounds in the first S top compounds.

$$EF_s = \frac{\frac{TP_S}{S}}{\frac{TP_T}{N}} \tag{3.1}$$

$$EF_{max} = \frac{1|\frac{TP_T}{S}}{\frac{TP_T}{N}} \tag{3.2}$$

$$\%EF_{max} = \frac{EF_S}{EF_{max}} \tag{3.3}$$

$$EF_{ratio} = \frac{EF_{IF}}{EF_{docking}} \tag{3.4}$$

The metric that we use to check how much does the PELE simulation affect the scoring is the EF_{ratio} defined in equation 3.4, where the EF_{IF} makes reference to the EFs observed at a given threshold (the top50 compounds) for the scores computed on the structures coming from the PELE simulation, and $EF_{docking}$ makes reference to the EFs at the same threshold, but for the scores computed on the structures coming from the initial docking process.

We'll use Glide score as the reference to check and improve the simulation, since it is the software most used by the industry. When we perform the same analysis for the other scores we can appreciate similar or better results for the scores DSX, Xscore and Vina, while we see that the force-field based scores likePELE BE get worse results .

As we can see in Figure 3.1a only 5 out of the 14 systems present an EF_{ratio} over 1, meaning they improve the enrichment from docking; 2 systems have an EF_{ratio} of 1 so they present the same enrichment as the docking methodology; 6 present an EF_{ratio} below 1, meaning the enrichment over the IF structures is lower than the enrichment of the docking, and thehivw system has no ratio meaning that both methodologies have no true positive on the top50 compounds .

In parallel, and in order to explore our sampling potential, we tried using the other implicit solvent model at our disposal: the OBC solvent. Thus, we tried the same simulation protocol but with the OBC solvent instead of the VDGBNP over the same set of proteins. When comparing the results of the simulations with this new solvent with the simulations with the VDGBNP solvent we observe a better EF_{ratio} for 9 of the systems, the same EF_{ratio} for 2 systems and a worse EF_{ratio} for another 2 systems, thus, we decided to use the OBC solvent for all future simulations.

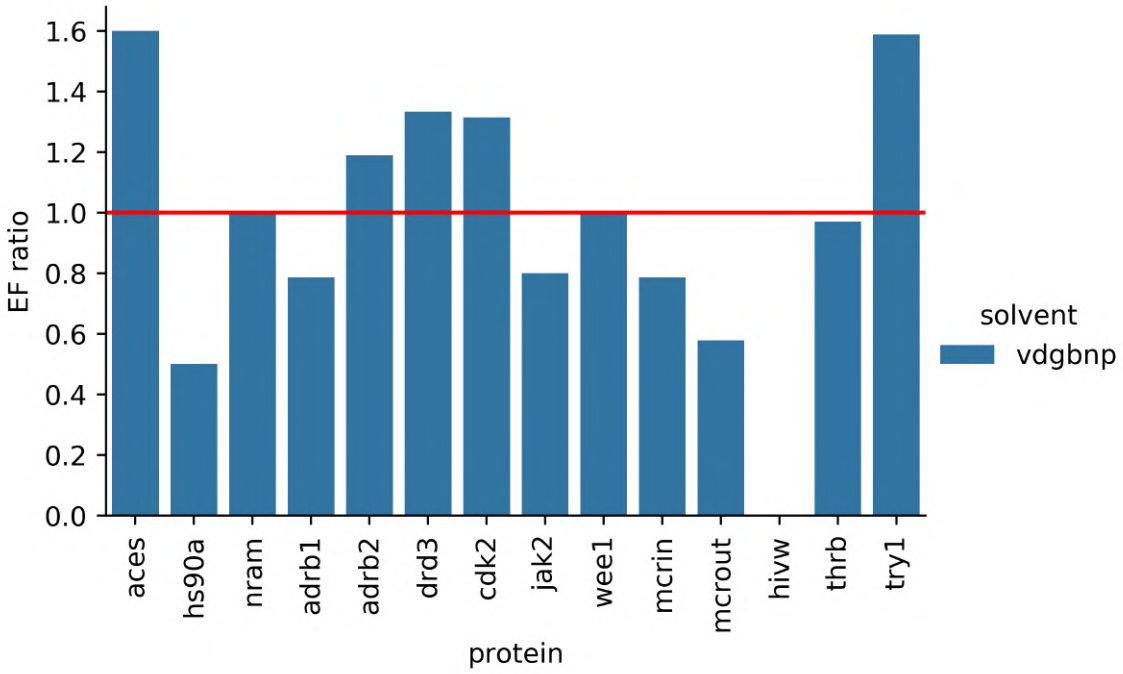
We decided to extend our study to include up to 21 different receptors by adding one more receptor to the diverse category, and 6 more receptors to the NHRs category. As stated, we prioritized human hormone receptors due to the interest in these systems from our collaborators, AstraZeneca. We performed the aforementioned simulation protocol with the OBC solvent for all the 21 receptors.

Table 3.2: $\%EF_{50}$ changes upon simulation with protocol 0

Family	Protein	Docking $\%EF_{max}$	I.F $\%EF_{max}$	EF_{ratio}	Threshold
DIVERSE	aces	0.12	0.15	**1.20**	50
	hs90	0.10	0.07	*0.75*	50
	nram	0.32	0.53	**1.67**	50
GPCR	adrb1	0.34	0.40	1.18	50
	adrb2	0.42	0.50	**1.19**	50
	drd3	0.07	0.15	**2.33**	50
KINASE	cdk2	0.70	0.80	**1.14**	50
	jak2	0.60	0.52	*0.87*	50
	wee1	1.00	1.00	1.00	50
NHRs	andr1	0.56	0.48	*0.86*	50
	esr1	0.98	0.46	*0.47*	50
	gcr	0.44	0.24	*0.55*	50
	mcrin	0.54	0.58	**1.07**	50
	mcrout	0.51	0.43	*0.84*	50
	ppar	0.24	0.44	**1.83**	50
	prgr	0.76	0.50	*0.66*	50
	rxra	0.86	0.48	*0.56*	50
	thb	0.42	0.30	*0.71*	50
PROTEASE	hivw	0.86	0.78	*0.91*	50
	thrb	0.72	0.74	**1.03**	50
	try1	0.36	0.72	**2.00**	50

Table 3.2 summarizes the results for the Glide score, considering only the top50 compounds. We now observe an improvement of the enrichment for receptors:aces, nram,adrb1, adrb2, drd3, cdk2, mcrin, ppar, thrb and try1; a loss of EFs for receptors: hs90, jak2, andr1, esr1, gcr, mcrout, prgr,

(a) VDGBNP solvent PELE simulation EF_{ratio} for each family



(b) VDGBNP vs OBC solvent PELE simulations EF_{ratio} for each family

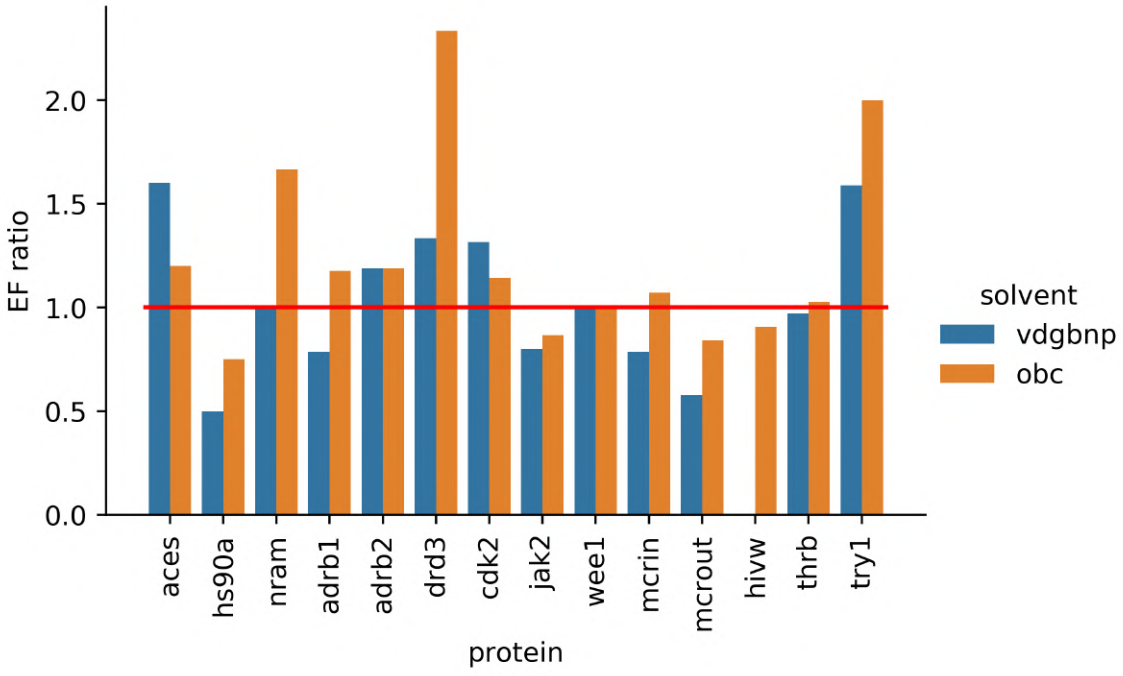


Figure 3.1: EF_{ratio} results for each family

Table 3.3: $\%EF_{100}$ changes upon simulation with protocol 0

Family	Protein	Docking $\%EF_{max}$	I.F $\%EF_{max}$	EF_{ratio}	Threshold
DIVERSE	aces	0.22	0.29	**1.33**	100.00
	hs90	0.19	0.10	*0.50*	100.00
	nram	0.47	0.77	**1.64**	100.00
GPCR	adrb1	0.32	0.33	**1.03**	100.00
	adrb2	0.34	0.43	**1.26**	100.00
	drd3	0.09	0.30	**3.50**	100.00
KINASE	cdk2	0.70	0.72	**1.03**	100.00
	jak2	0.45	0.44	*0.97*	100.00
	wee1	0.99	0.79	*0.80*	100.00
NHRs	andr1	0.47	0.43	*0.91*	100.00
	esr1	0.91	0.54	*0.59*	100.00
	gcr	0.50	0.28	*0.55*	100.00
	mcrin	0.58	0.73	**1.27**	100.00
	mcrout	0.62	0.59	*0.96*	100.00
	ppar	0.34	0.41	**1.21**	100.00
	prgr	0.59	0.39	*0.66*	100.00
	rxra	0.76	0.54	*0.71*	100.00
	thb	0.42	0.33	*0.78*	100.00
PROTEASE	hivw	0.50	0.65	**1.30**	100.00
	thrb	0.57	0.67	**1.18**	100.00
	try1	0.39	0.65	**1.67**	100.00

rxra, thb andhivw; the receptor wee1 presents a $\%EF_{max}$ of 100% (meaning all the compounds in the top50 compounds are actives) for the docking and the IF.

Table 3.3 summarizes the results for the Glide score, considering the top100 compounds. The number of systems with an improvement of the enrichment is now 11 and 10 systems present a loss of enrichment. The only receptors that change from category arehivw and wee1. Thehivw receptor now presents an enrichment of the top compounds, while the receptor wee1 presents a loss of enrichment.

Now, we have a protocol that improves the docking results for half of the systems (10 systems), but we still underperform for the other half of the systems (10 systems). We noticed that, out of these 10 failing systems, 7 belong to the NHR family. In order to discover the reason behind this split behaviour we studied several metrics that may explain the observed difference.

3.3.2 Accuracy

The metric the reader may be more familiar with is the accuracy. The accuracy values and the untreated EF values, can be found in the Supplementary Information (SI) tables A.1 to A.14. In our particular case, the accuracy metric can never become 1 for almost half of systems, because they present less active compounds than the sample size we're studying (50 or 100) thus, we've studied

the $\%EF_{max}$.

$$Accuracy = \frac{TP_S}{S} \tag{3.5}$$

The accuracy formula used for the computations is in equation 3.5 where TP_S is the number of true positives in the top S ranking and S is the number of compounds taken into account, or threshold, (50 or 100)

As shown in Table 3.1, the systems drd3, mcrin, mcrout,aces and hs90 contain less than 50 active compounds, and the systemsadrb1, jak2, gcr, thb and nram present less than 100 compounds. Thus, for the first group of systems, the accuracy can never become 1 when the sample is bigger than 50; while for the second group the accuracy and the $\%EF_{max}$ match at threshold 50 but no at threshold 100.

We can consider the $\%EF_{max}$ a normalized accuracy, both metrics present the same ratios and tendencies.

3.4 Dataset metric's relationships with the results

In this section we'll study the dataset properties in order to see if the difference in the enrichment is due to any particular characteristic of the dataset. It may be that for some families the actives are really potent (low ΔG values), maybe it's how exposed the cavity is or maybe the size of the pocket is what is causing the differences on the method performance.

For each of the metrics studied we've generated two boxplots images. The boxplot graph represents the data distribution, showing the range of the metric and each of the quartiles. The first image represents the metric distribution taking into account all the compounds of the protein subset; in the images the colour of the boxes represent whether the system presents an improvement of the EFs (in green colour), a loss of EFs (red colour) or no change (blue). The other image separates each protein subset into the actives compounds in a yellow box, and the inactive in a purple box, the borders of the boxes match the colours of the boxes in the first image.

The first metrics we checked are based only on the selected structure from the simulations. Thus, they provide information on the properties of the selected structures, and some of them give information about the morphology of the receptor. Nevertheless, they don't provide information about the changes introduced by the simulation.

3.4.1 Gibbs Free energy or ΔG

The Gibbs free energy (ΔG) is a measure of how much energy a chemical reaction needs (positive values) or produces (negative values). A negative ΔG means the reaction is spontaneous at room temperature. We're looking for compounds capable of binding the protein on their own with high

affinity (low concentrations of compounds needed), which means that, for us the lower the ΔG value is, the better. We can compute this value exactly from the K_i or K_d values of a compound, and we can approximate it using IC50 values instead of K_i or K_d.

Since we have K_i, K_d or IC_{50} values for all the active compounds in the DUD-e dataset, we can compute the ΔG for each of the actives using the following equation:

$$\Delta G = ln(K)RT \tag{3.6}$$

Where K is the K_i, K_d or IC_{50} of the compound in molar units, R is the ideal gas constant ($1.98 * 10^{-3} Kcal/mol$), and T is the room temperature in kelvin (298) degrees. For those compounds with no activity we assign a value of 0.

Figure 3.2a shows the distribution of ΔG values taking into account all the compounds of each receptor for the top50 compounds. We can appreciate that the average ΔG value between the systems that get improved by the methodology is -6.18 Kcal (horizontal green line) and the average for those that don't get improved is -6.43 Kcal (horizontal red line). The difference in the ΔG distributions between the systems with a gain of enrichment and those with a loss of enrichment is similar to the one present within the groups.

(a) ΔG distribution for each family for all the compounds

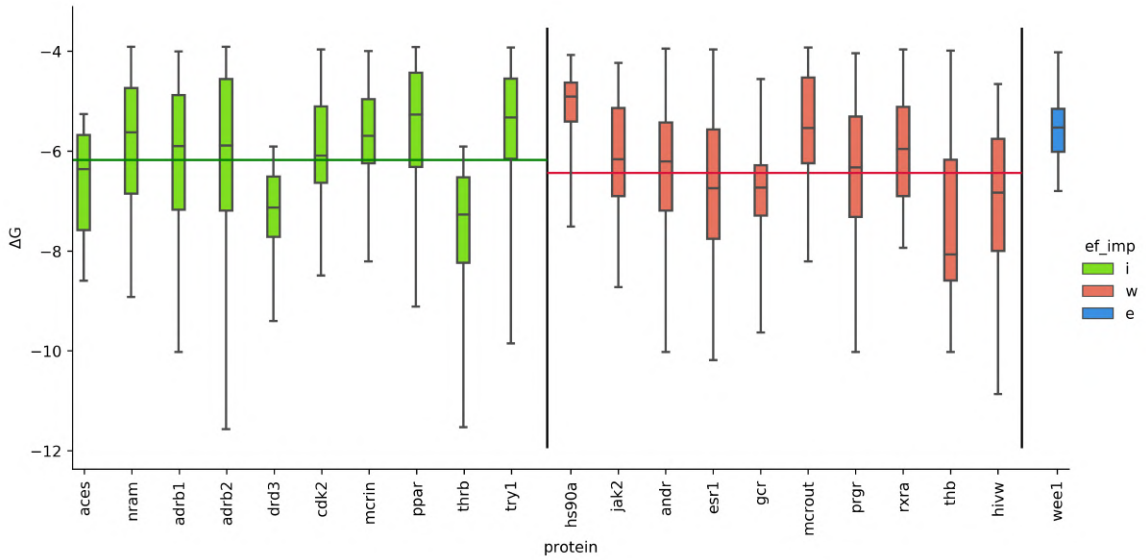


Figure 3.2: Boxplots with the ΔG values on the Y-axis and the different receptors studied on the X-axis.

3.4.2 PELE B.E

The first metric from the simulation we checked was the metric used to select the pose from the simulation: the PELE BE The PELE BE is a force-field based scoring function derived from the PELE total energy, used by the program to estimate the likelihood of a pose. The PELE BE is computed using the formula in equation 3.7, where $T.E_{complex}$ is the PELE total energy of the complex, $T.E_{protein}$ is the PELE total energy of the protein alone, and $T.E_{ligand}$ is the PELE total energy of the ligand.

$$PELEB.E. = T.E_{complex} - T.E_{protein} - T.E_{ligand} \tag{3.7}$$

In Figure 3.3a we can observe that this metric varies highly within the families, but also from one system to another. Both the systems that improve, and those that get worse in terms of enrichment, have similar average and distribution ofPELE BE So we cannot extract any correlation between the systems that get improved and their PELE BE distribution.

Another explanation could be that there's a significant difference in the distribution of thePELE BE between actives and inactives within the same receptor; and that this difference is the responsible of the differences observed on the EF_{ratio} .

In Figure 3.3b we can compare the distribution ofPELE BE between actives and inactives for each of the studied systems. We observe that, for the systems hs90 , gcr, and ppar, the medians are similar in value for the actives and inactives, and the distribution of thePELE BE is also similar. For the first two receptors the PELE simulations produce a loss in the EF; while for the ppar system it improves the EF. For the remaining receptors, we observed a lower median for the actives when compared to the inactives.

Given that both groups, the one with similar median and the one with a lower median for the actives, contain receptors with an improvement of the enrichment and receptors with a loss of enrichment, we cannot associate this metric with the improvement or loss of enrichment.

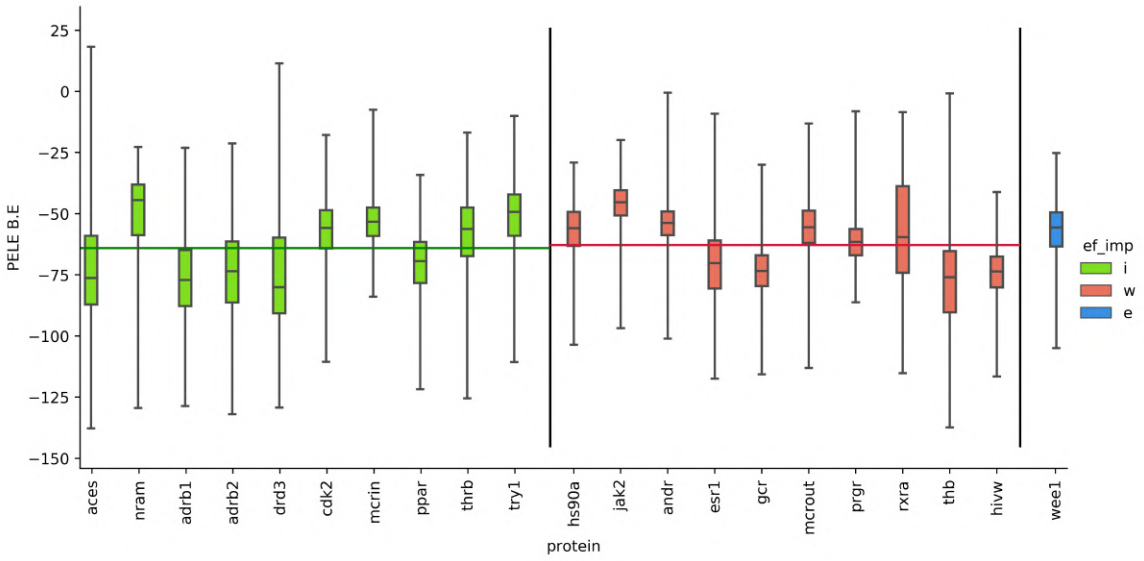
3.4.3 Solvent Accessible Surface Area (SASA)

The next metric we checked has to do with the solvent we're using. PELE uses implicit solvent models and we don't use any explicit water for the simulations, with the exception of the hiv protease. This receptor has an essential water molecule on the binding site, so we may be introducing a bigger error for those ligands that are more exposed to the solvent.

In order to check for the ligands exposure to the solvent, we map the distribution of the relative (compared to the full exposition in water) Solvent Accessible Surface Area (SASA) of the ligands inside the BP.

The mean SASA value through all the compounds is 0.09. In Figure 3.4a we can observe that for the systems andr1, esr1, gcr, mcr (both conformations: out and in) , prgr, rxra and thb the SASA

(a) PELE B.E distribution for each family for all the compounds



(b) PELE B.E distribution for each family for all the compounds by activity

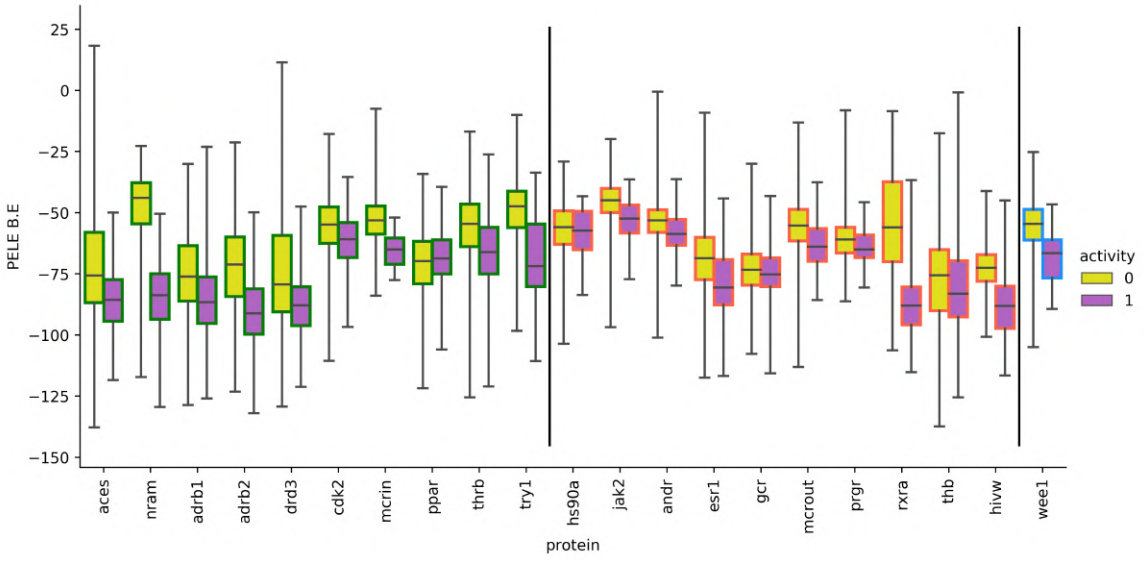


Figure 3.3: PELE B.E. distributions

is below 0.05 for at least 75% of the compounds. All of these systems belong to the NHRs family, whose BP is known for having almost no exposure to the solvent. From all these systems only the mcr-in has an improvement of the enrichment, and it's also the only system with a percentage of active compounds below the 3%.

In Figure 3.4b we can see that theaces, ppar and nram receptors present a significant difference in the distribution of SASA between the actives and inactives compounds. For theaces and ppar, the actives have an upper median than the inactive compounds, while the nrampresents a lower median of the active compounds, but the limits of the actives are included within the limits of the inactives compounds. The rest of receptors present similar medians for the actives and inactives.

3.4.4 General Pocket size

For each of the receptors we defined a general BP, which means we define a constant cavity throughout the simulation as the BP without taking into account the position of the ligand. In order to define this general BP we run the programs fpocket 3 [54] and SiteMap from Schrödinger[32, 34], which study the cavities in a protein, producing a list of ranked pockets. Next, we select the pockets that overlap with the ligand position present in the initial crystal, that is, the crystal used to extract the protein's structure for the initial docking step, and combine the pockets derived from both programs in order to have the biggest cavity possible as the BP.

One estimator of the pocket size can be how many residues are involved in the general BP. This measure doesn't allow us to estimate the volume since even receptors with the same number of residues in the general pocket can have different volume, depending on which are the residues present, but it can give us an idea of how wide it is.

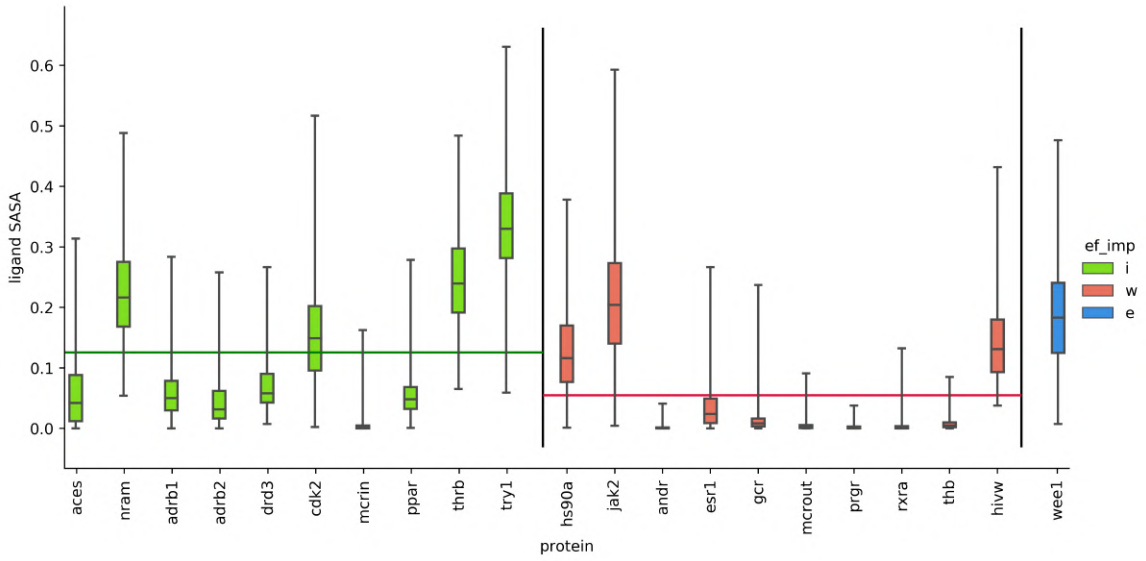
The average number of residues in the general pocket is 50.48 residues per pocket. Theaces receptor has exactly 50 residues within the BP. The receptorsadrb1, adrb2, drd3, ppar, wee1, thb, hs90, jak2, esr1, and gcr have more than 50 residues forming the general BP. While the receptors nram, mcrin, andr1, mcrout, prgr, rxra, thb and hivw have less than 50 residues in their general BP.

3.4.5 Molecular weight

The average MW of all compounds through all receptors is 415.72. Receptors nram, mcrin, andr1, mcrout and prgr have 75% of the compounds with molecular weights below 415.72. For receptors ppar,thrb, andhivw we find 75% of compounds with values above 415.72. The receptors in both groups encompass systems that present both an improvement and a loss of EF.

When comparing the distribution of the molecular weight between active and inactive compounds, we see that, for receptorsaces and rxra the active compounds have bigger molecular weights than the inactive. The rest of the receptors have similar distributions of molecular weight between the active compounds and the inactive compounds.

(a) Ligand SASA distribution for each family for all the compounds



(b) Ligand SASA distribution for each family for all the compounds by activity

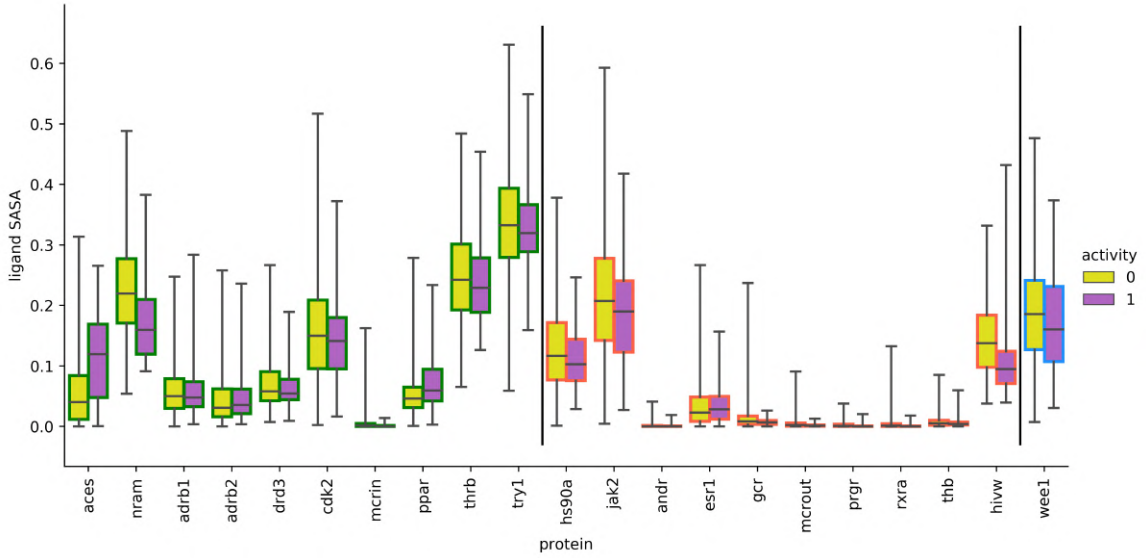
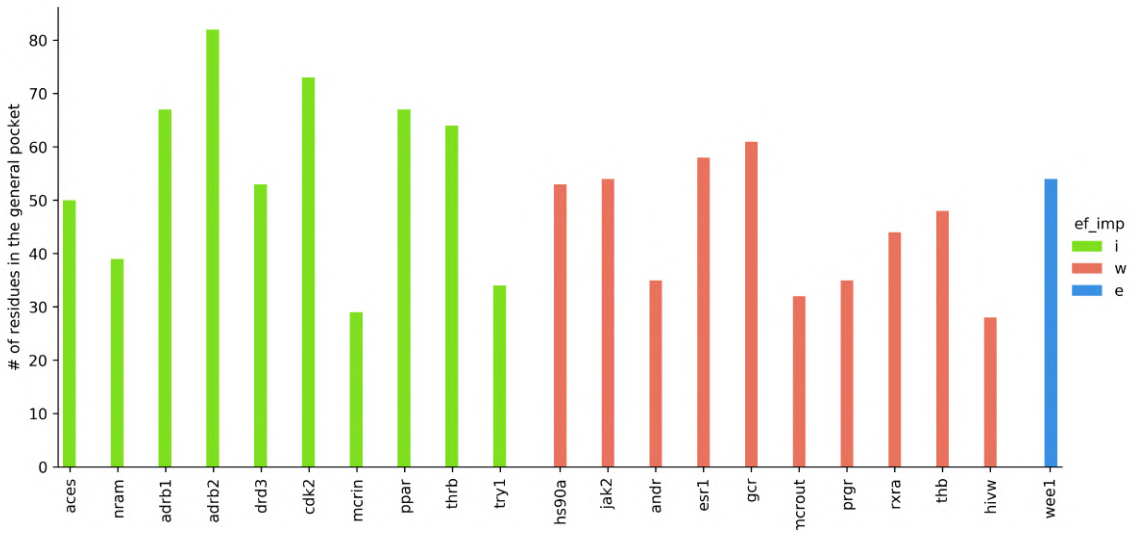


Figure 3.4: Ligand SASA distributions

Figure 3.5: Number of residues in the general pocket



3.5 Simulation influence on the results

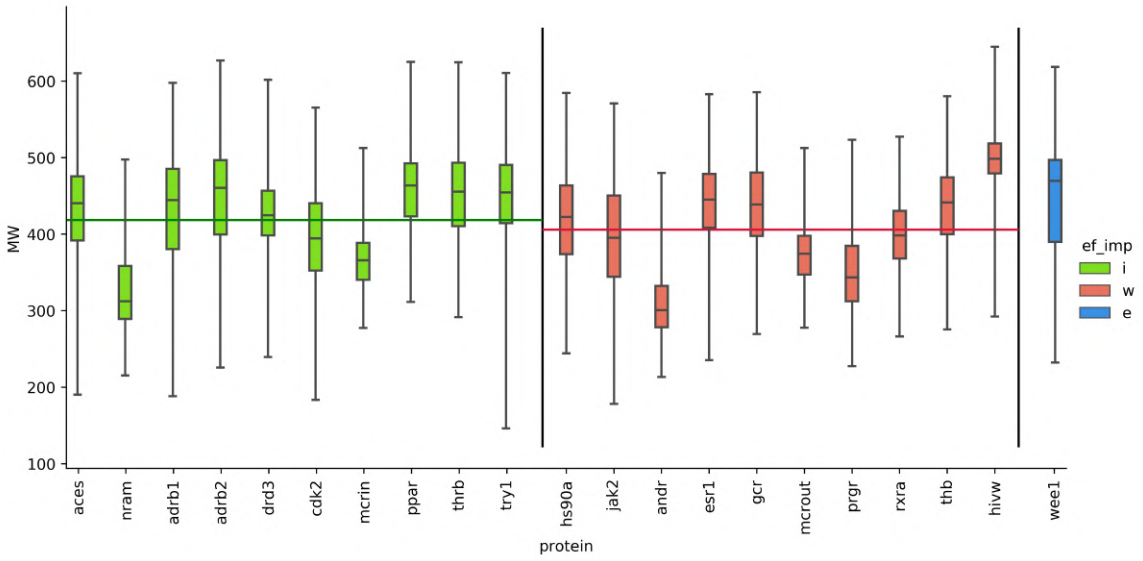
All the metrics we've checked so far are somewhat intrinsic to the system we're studying itself, even thePELE BE, since all of them are measured taking into account only the selected point of the simulation. In contrast, the next series of metrics we're going to study attempt to account for changes introduced by the simulation, since all of them are based on the Root Mean Square Deviation (RMSD) of heavy atoms of the complex (all atoms that aren't hydrogens).

The RMSD is computed using two structures and is the measure of the average distance between the heavy atoms in structures that occupy the same portion of the space. The RMSD formula is represented in equation 3.8, where N is the number of atoms used to compute the RMSD and δ is the distance between the two positions that a given atom occupies.

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2} \tag{3.8}$$

We want to estimate how much we change the complex structure from the initial docking to the new structure proposed (after IF). Thus, we'll compute the RMSD between the selected model of the protein-ligand complex and the initial one.

(a) MW distribution for each family for all the compounds



(b) MW distribution for each family for all the compounds by activity

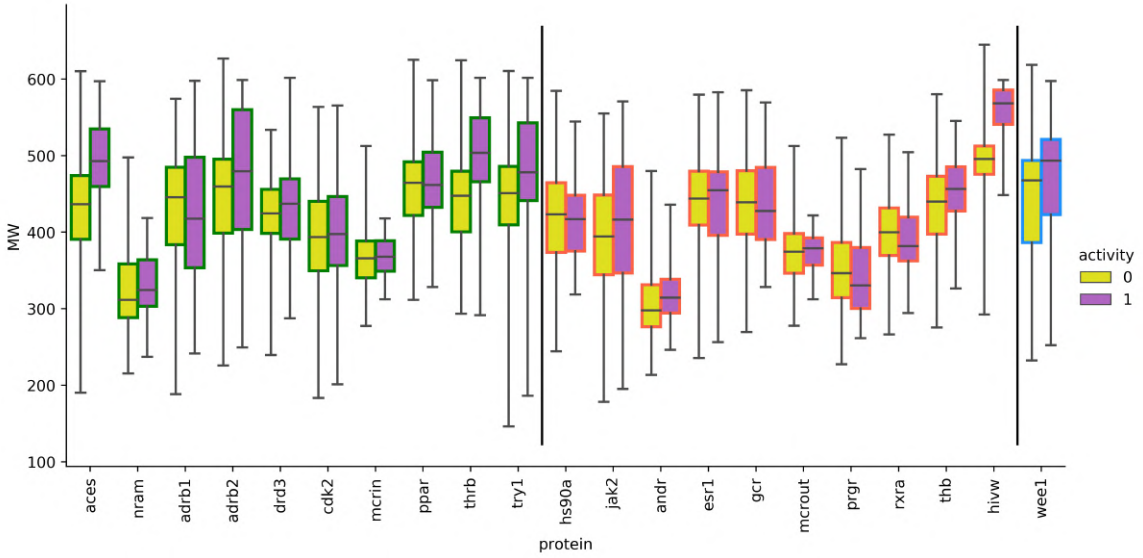


Figure 3.6: Molecular weight distributions

3.5.1 Ligand RMSD

First we checked how much we've modified the ligand pose. To do this, we compute the RMSD between the heavy atoms of the ligand, between the two poses (the initial and the selected). With this metric we estimate how much the ligand has been modified in order to fit into the BP.

When we compare the average RMSD for all the ligands of a given receptor (Figure 3.7a) we can appreciate that average RMSD of the systems with a loss of enrichment is 2.85 while the average RMSD of the systems with an improvement is 4.04, a significant difference that points to a "necessity" to correct the initial docking poses. However we observe a median value below 4 for the receptors: drd3, ppar, wee1, andr1, esr1, gcr, mcrout, prgr, rxra, and thb; From this list of receptors only drd3 and ppar have an improvement on the EF, while wee1 has the same EF. All the other systems show a loss of EF, but not all the systems with a loss of EFs are present in this list, thus, we cannot derive any explanation from these results.

In order to further explore this metric, we studied if there's an activity specific change on the RMSD of the ligand. With the exception of the ppar receptor all the other receptors present a median for the actives RMSD below the inactives median RMSD (Figure 3.7b). In receptors nram, mcrout, prgr and rxra, the median RMSD of the active ligands is below the RMSD value of the 3rd quartile of the inactives RMSD, which means that 75% of the inactives present in each of these receptors have a bigger RMSD than half of the actives in these systems; for the esr1 system the median of the actives' RMSD is slightly higher than the 3rd Quartile of the inactives, meaning it's a close case to the previous one. For the receptorsaces, jak2 and thb the medians are almost the same.

Figure 3.7 is a boxplot representation of the ligand heavy-atoms RMSD (Y-axis) for each of the receptors (X-axis). A) All ligands RMSD distribution coloured by the improvement (blue), no-change (green), or loss (yellow) of EF. B) Distribution of ligand RMSD coloured by activity, the inactives ligands in blue, and the actives in yellow.

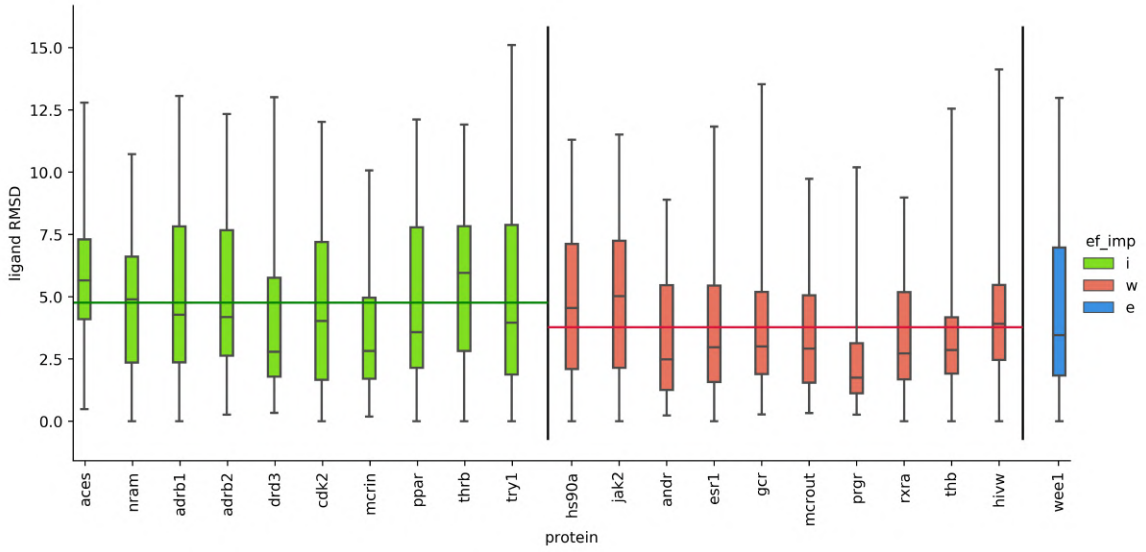
Given the low number of actives in some of the systems, we may think this is the cause for the observed differences. The percentage of actives for the ppar receptor is 18.3% and for the nram, mcrout, prgr, rxra and esr1 are 5.0%, 3.8%, 18.3%, 10.6% and 19.2%, respectively, while for theaces and jak2 receptors the percentages are 4.1% and 22.3%. As we can see all these groups have a heterogeneous percentage of actives, meaning this isn't the cause behind the behaviour we observe.

The only explanation left is that these changes are introduced by the simulation, which causes bigger changes into the conformation of most of the inactives compounds when compared to the changes introduced into the active compounds for all the systems but the ppar system.

3.5.2 General Pocket RMSD

Then we computed the RMSD of the residues present in this general pocket (the same as in subsection 3.4.4), for all of the compounds of each receptor.

(a) Ligand RMSD distribution for each family for all the compounds



(b) Ligand RMSD distribution for each family for all the compounds by activity

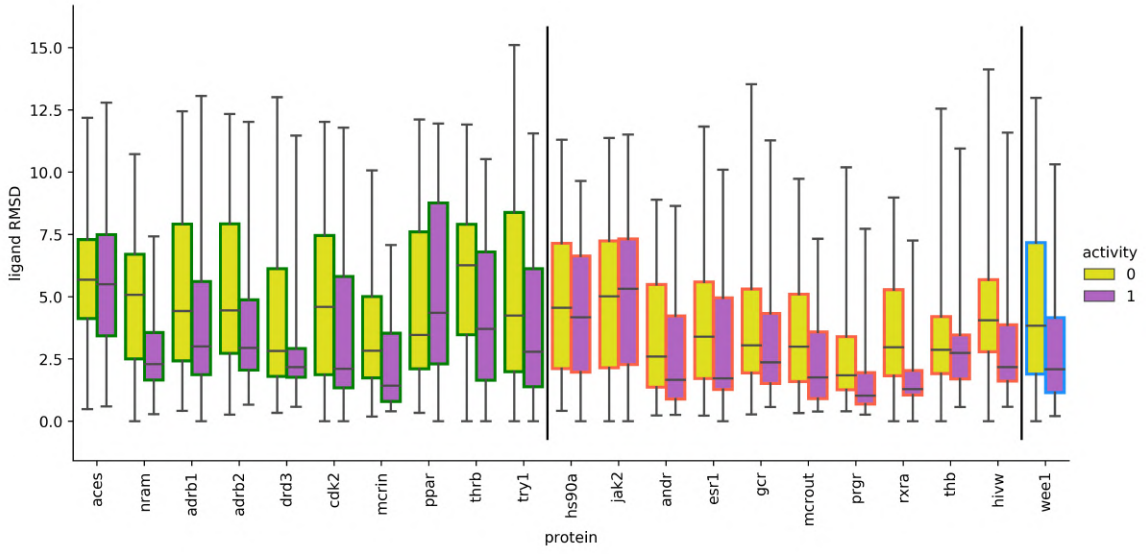
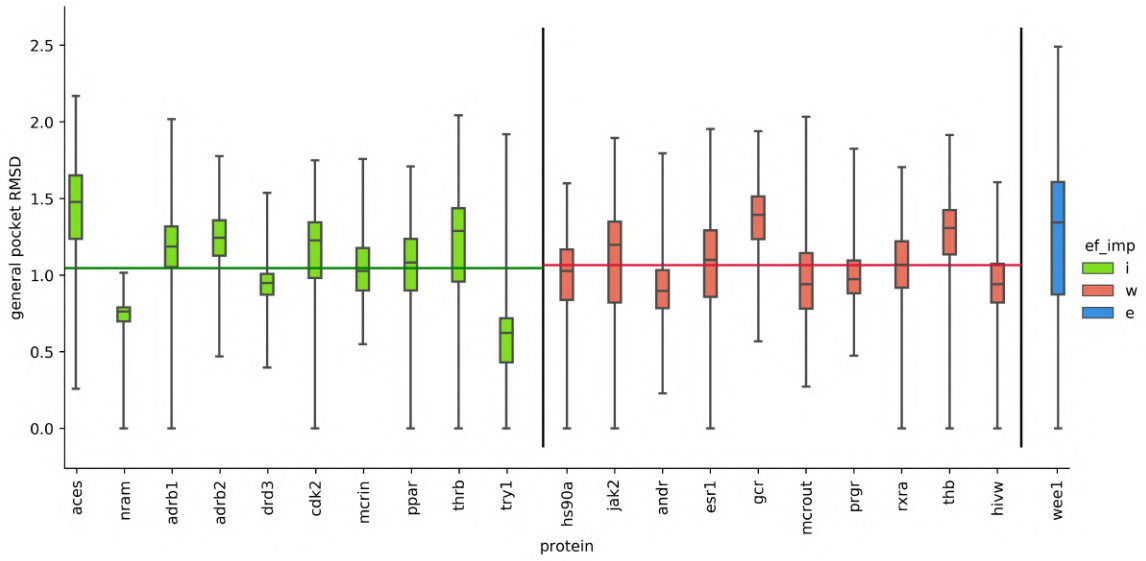


Figure 3.7: Ligand RMSD distributions

(a) General pocket RMSD distribution for each family for all the compounds



(b) General pocket RMSD distribution for each family for all the compounds by activity

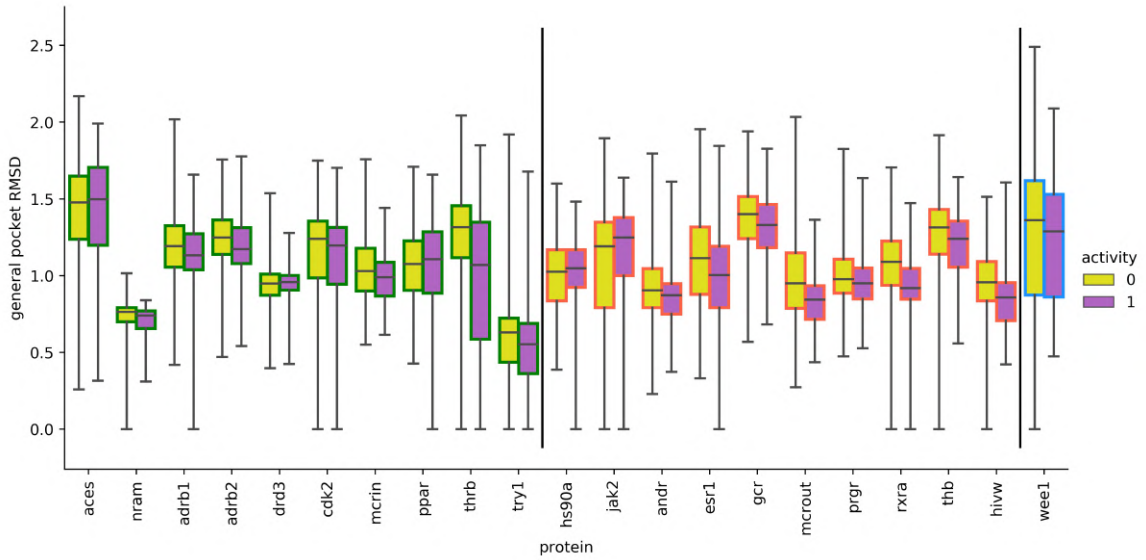


Figure 3.8: General pocket RMSD distributions

The average RMSD through all the receptors general pocket is 1.07 . From all receptors only nramhas an RMSD below 1.07 for all the compounds, for the receptors drd3, try1, andr1, mcrout andhivw have a median below 1.07A, receptors mcrin and hs90 are almost equal to 1.07, and for the other receptors the mean is over 1.07 but below 1.5. The RMSD distribution for the systems that get improved overlaps with those that don't improve. If we now study the differences between the general pocket RMSD distribution of actives and inactives for each receptor, it presents a similar median for all systems with the exception ofthrb and rxra, which present a median for the active compounds below the 3rd quartile of the inactive compounds. This result can be expected since the simulation protocol we're using is focused on optimizing the ligand pose while allowing for some rearrangement on the protein structure, but it isn't an extended exploratory protocol.

3.5.3 Ligand specific pocket

Next we studied the change of the ligand specific BP. To define this new BP we selected the residues with any heavy-atom within a distance of 4 of any heavy atom of the ligand from the initial structure in the simulation. We also perform the same selection of residues with the selected structure from the simulation, and we use the residues present in any of the selections as the BP. Finally, we compute the RMSD of the heavy-atoms of the selected residues.

The average RMSD for the ligand specific pocket through all the receptors is 1.28. The difference in the average RMSD between the receptors with an improvement in the enrichment and those with a loss is just 0.12. Receptorsnram, drd3, mcrin, andr1, mcrout and ppar have a RMSD below, or almost equal to, 1.28 for 75% of the compounds . Thehivw receptor has a RMSD above the 2.0 for all the compounds, while the rest of the receptors present a similar spread of the RMSD values for all compounds (Figure 3.9).

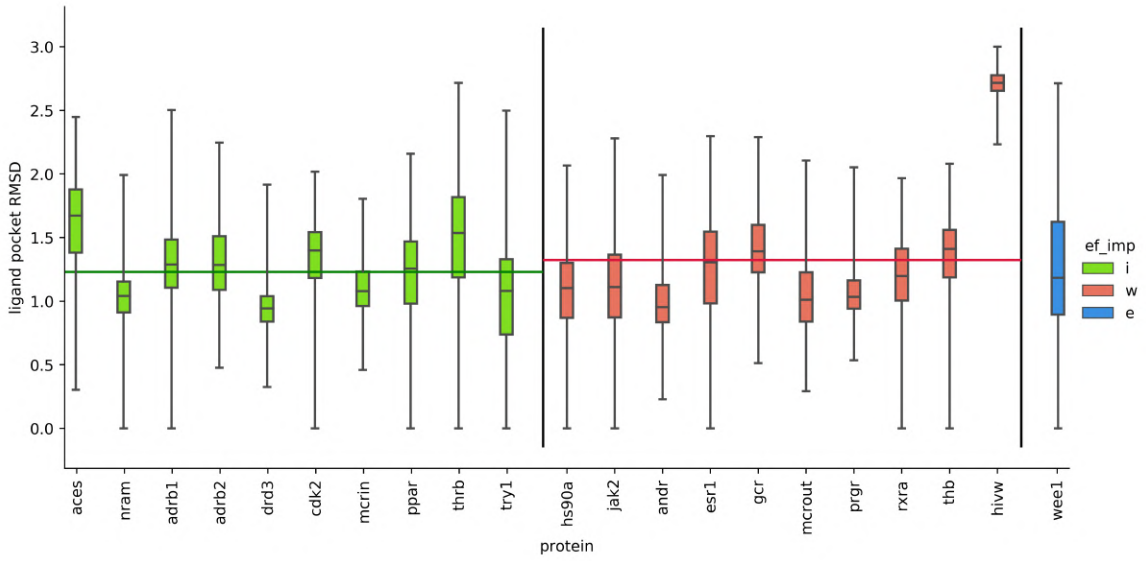
Next, we studied the distribution of ligand specific pocket RMSD distinguishing between active and inactive compounds for each of the receptors. We observe that receptorsaces, drd3, cdk2, ppar and hs90 present a similar median between actives and inactives. For receptors nram,adrb1, adrb2, mcrin, try1, andr1, esr1, gcr, mcrout, prgr, thb andhivw we observe a lower median for the active compounds. Only the jak2 receptor presents a higher median for the active compounds.

Another metric to study this ligand specific pocket is to check how many residues are present in both selections. That is, how many residues are within 4 Å of the two poses of the ligand: the docking pose and the I.F. pose. This metric gives us an idea of how much the environment of the ligand has changed during the simulation

3.5.4 Glide change

The difference in Glide score [23, 33] (ΔGlide) metric is the difference between the Glide score of the pose coming from the docking and the Glide score of the pose selected from the simulation (Equation 3.9). A value of 0 means there's been no-change in the score; a positive number means

(a) Ligand specific pocket RMSD distribution for each family for all the compounds



(b) Ligand specific pocket RMSD distribution for each family for all the compounds by activity

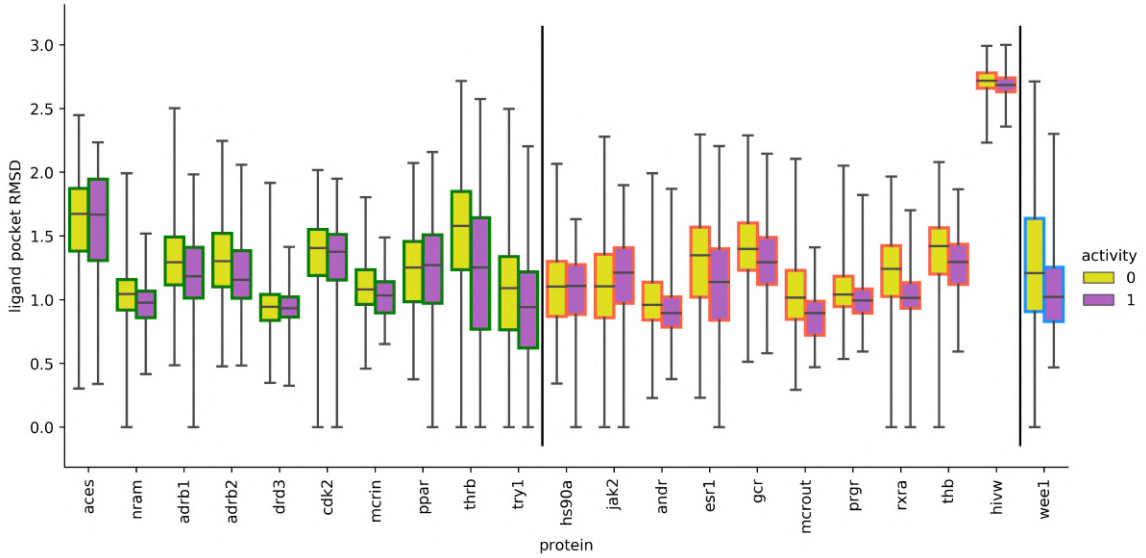
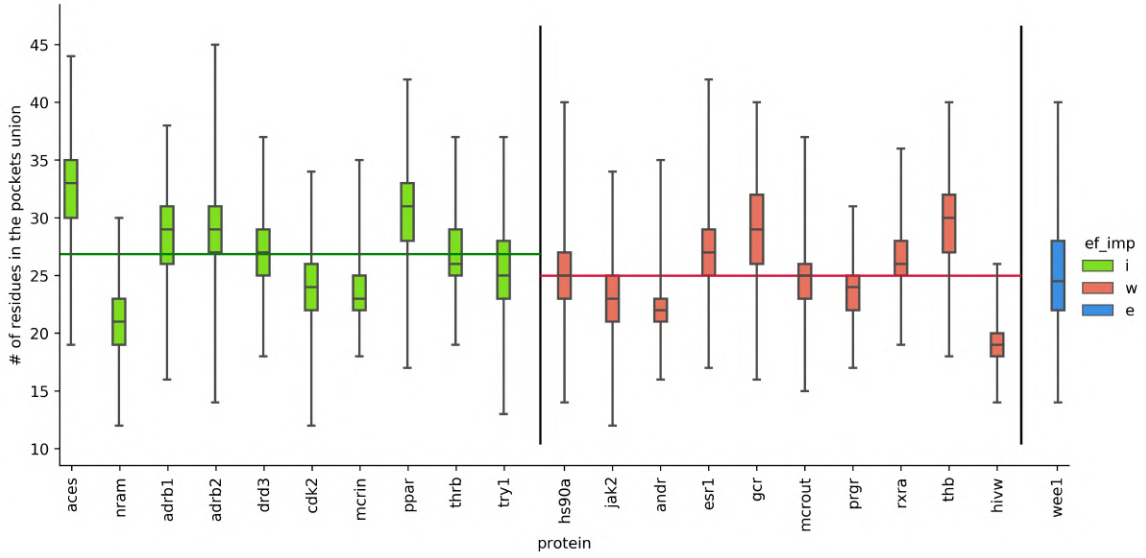


Figure 3.9: Ligand specific pocket RMSD distributions

(a) number of residues in the ligand specific pocket union distribution for each family for all the compounds



(b) number of residues in the ligand specific pocket union distribution for each family for all the compounds by activity

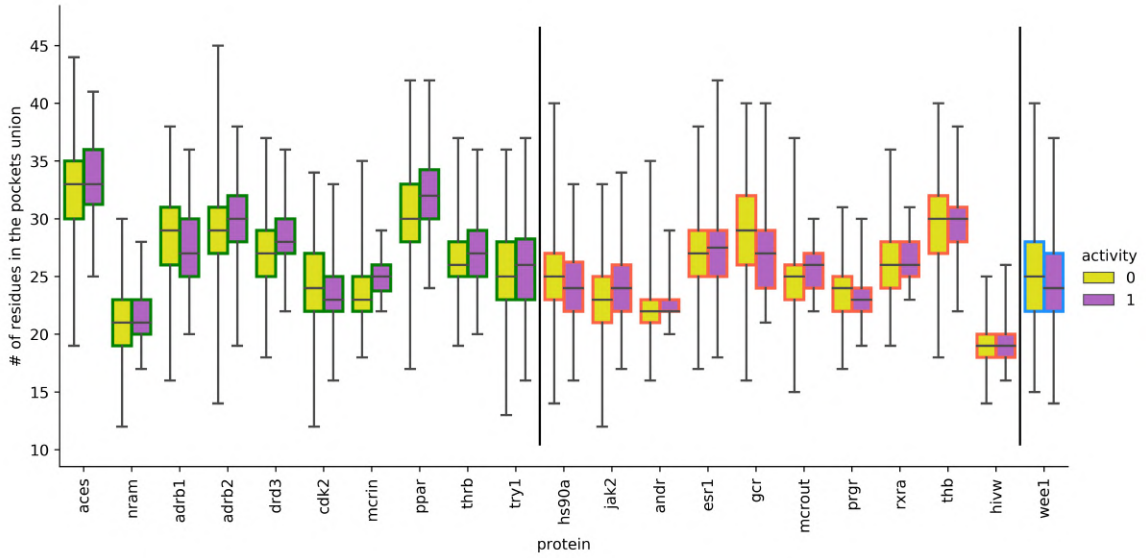
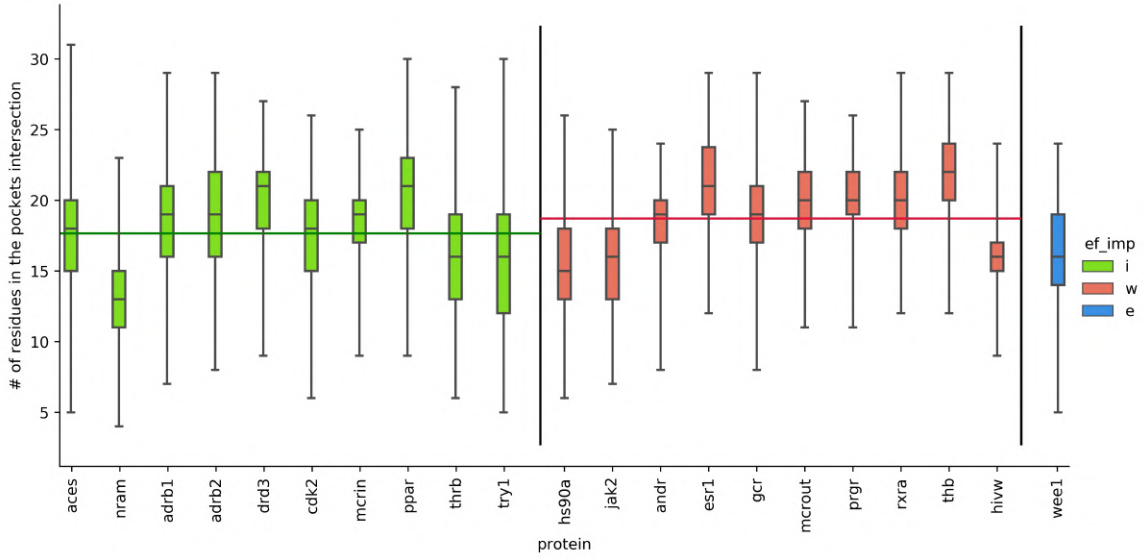


Figure 3.10: Ligand specific pocket number of residues in the union distributions

(a) number of residues in the ligand specific pocket intersection distribution for each family for all the compounds



(b) number of residues in the ligand specific pocket intersection distribution for each family for all the compounds by activity

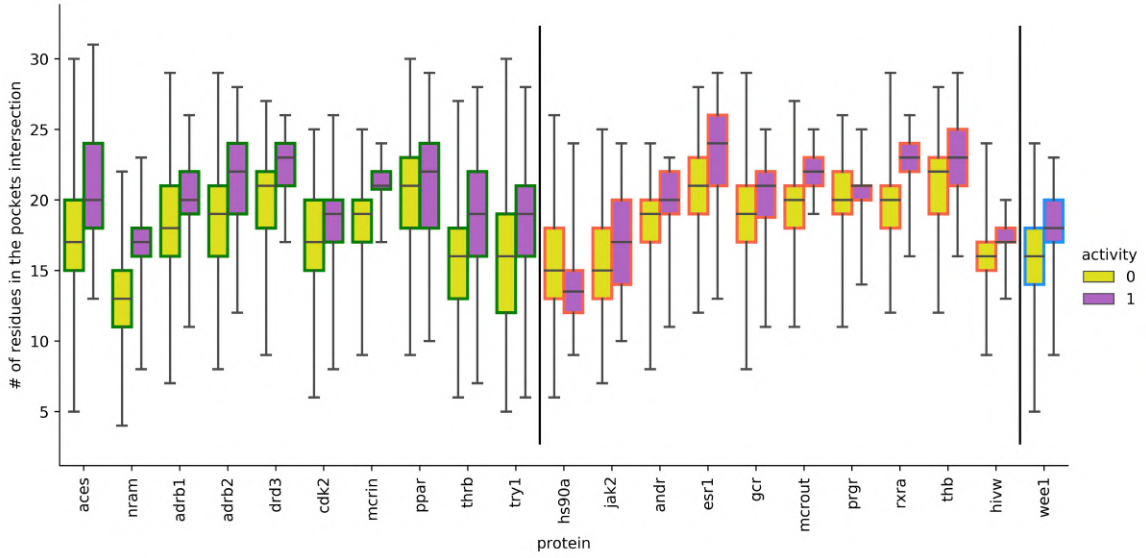
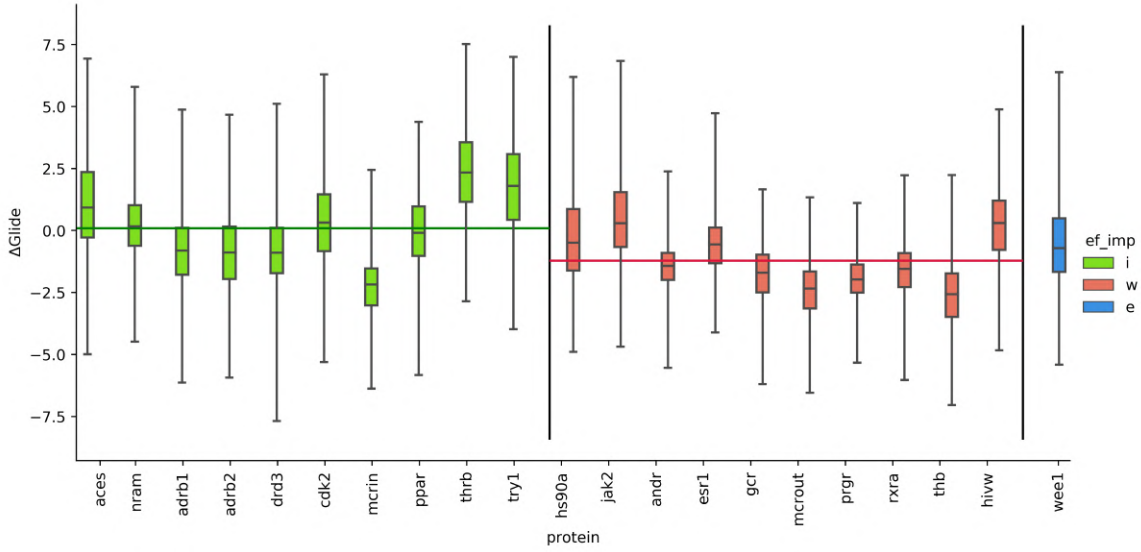


Figure 3.11: Ligand specific pocket number of residues in the intersection distributions

(a) ΔGlide distribution for each family for all the compounds



(b) ΔGlide distribution for each family for all the compounds by activity

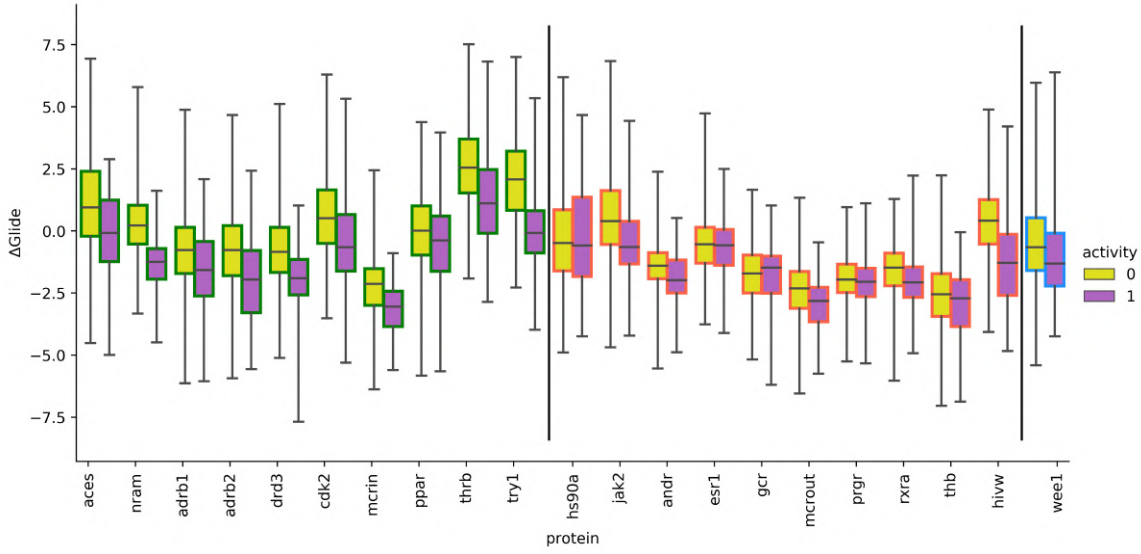


Figure 3.12: ΔGlide distributions

that the compound has a worse score, and a negative number means the selected structure has a better score.

$$\Delta Glide = Glide_{IF} - Glide_{docking} \tag{3.9}$$

As shown in figure 3.2 the average Glide difference through all the receptors is -0.59 . If we consider all the compounds of each receptor at the same time, only the receptors mcrin, andr1, gcr, mcrout, prgr rxra and thb have a ΔGlide below -0.59 for 75% of the compounds.

For the receptors with an improvement on the EFs, the median ΔGlide of the active compounds is smaller than, or equal to, the third quartile of the inactive compounds, with the exception ofadrb1 and ppar. for whom the median values are similar.

The fact that none of the systems present a loss of Glide score means that we're improving the poses in all the cases. But, for those systems with a loss of EFs, we're introducing noise, because we're adapting the protein to inactive compounds.

3.5.5 Conclusions

We cannot pinpoint a single metric that explains why the receptorsaces, nram,adrb1, adrb2, drd3, cdk2, mcrin, ppar and try1 present an improvement of the EFs while the receptors hs90, andr1, esr1, gcrr, mcrout, prgr rxra thb andhivw present a loss of the EFs on the top50 compounds. The fact that the receptor wee1 doesn't lose EFs values is the best possible scenario since the original EFs is already the maximum EFs that can be achieved.

When looking at the family relationships of the systems with a loss of EF we observe that most of them belong to the NHRs family, 7 out of 10, independently of the threshold. This elevated number suggests that the loss of EF is related to the characteristics of this family pocket. The NHRs has a completely enclosed pocket, as we can observe on the ligand's SASA distribution on Figure 3.4a. This means that the exploration of the pocket is very restricted and in consequence the protein may end up deformed during the simulation, in order to accommodate the ligand.

This made us think that our simulation protocol wasn't adequate to this kind of receptor, and we proceeded to try new simulation protocols.

3.6 Simulations fine tuning

The reason behind the loss of enrichment in half of the systems could be that the simulation protocol we're using wasn't the right one. In order to assess this option, we tried two more PELE simulations protocols. Due to the amount of computational resources needed to perform the same analysis we decided to use fewer systems.

The selected systems are: cdk2, jak2, ppar, mcrin and gcr. The first two systems belong to the

kinase family, while the other three belong to the NHRs family. The receptors jak2 and gcr present a loss of enrichment, while receptors cdk2, mcrin and ppar present an improvement. The selection of systems answers the need to make sure we not only improve the systems that go wrong, but that we also keep improving the systems we already improve.

Table 3.4: Table containing a summary of the PELE simulation parameters used for the different protocols

Protocol	Type	General parameters			Epochs	perturbation	
		# of processors	Temperature	# of pele Steps		spawning	steering
0	Single core	1	2000	200	n.a.	n.a.	2
1	Single core	1	1500	200	n.a.	n.a.	none
2	Adaptive	16	1500	8	15	inv prop	random

3.6.1 Protocols summary

The two protocols we tried are described in the table 3.4. Since we observe an elevated RMSD value of the pocket for both the actives and inactive compounds, we decided to try a PELE simulation reducing the "steering" of the simulation to reduce the protein adaptation, named protocol 1. The other protocol tried, protocol 2, uses the adaptive methodology and the reduced steering protocol in order to improve the sampling. As commented in the introduction, the adaptive methodology is used to improve the sampling of the ligand avoiding the meta-states and easing the transition between minima.

3.6.2 Results Protocol 1

Table 3.5: $\%EF$ changes upon simulation with protocol 1

Family	Protein	Docking $\%EF_{max}$	I.F $\%EF_{max}$	EF_{ratio}	Threshold
KINASE	cdk2	0.70	0.84	**1.20**	50
	jak2	0.60	0.42	*0.70*	50
NHRs	gcr	0.44	0.18	*0.41*	50
	mcrin	0.53	0.5	*0.93*	50
	ppar	0.24	0.56	**2.33**	50
KINASE	cdk2	0.70	0.72	**1.03**	100
	jak2	0.45	0.39	*0.86*	100
NHRs	gcr	0.50	0.24	*0.48*	100
	mcrin	0.58	0.65	**1.13**	100
	ppar	0.33	0.46	**1.39**	100

As we can observe in Table 3.5 the protocol 1 affects the different proteins in different ways. The proteins cdk2 and ppar improve their EF at the top50 compounds and the top100 compounds while the proteins jak2 and gcr suffer a loss of EF for both thresholds, and the mcrin protein loses EF at the top50 compounds, but it improves the EFs at the top100 compounds. We also observe that the accuracy presents the same trends as the EF regarding the methodology influence.

Given the differences in accuracy of the docking methodology depending on the family, as mentioned in the section 3.3.2, we could argue that it is easier to obtain an improvement for the NHRs family than for the kinase family. Nevertheless, we have observed that most of the systems with a loss of EFs are from the NHRs family.

The metric that can help us to better understand this contradictory behaviour is the ΔGlide. We observe that the average ΔGlide is the same for the systems with an improvement of the EF on the top50 structures and the ones with a loss. But, what we can observe, is that for the systems with a loss of EF, the average ΔGlide for the active and inactive compounds is either almost equal (the gcr protein) or the inactive compounds present a better improvement than the active compounds on the lowest quartile, while for all the systems with an improvement of the EF the averages of the active compounds is lower.

The differences between active and inactive compounds are similar to the ones observed for this metric when using protocol 0, and it indicates that the simulation not only improves the poses of the actives compounds, but it also overfits the protein in order to accommodate the inactive compounds.

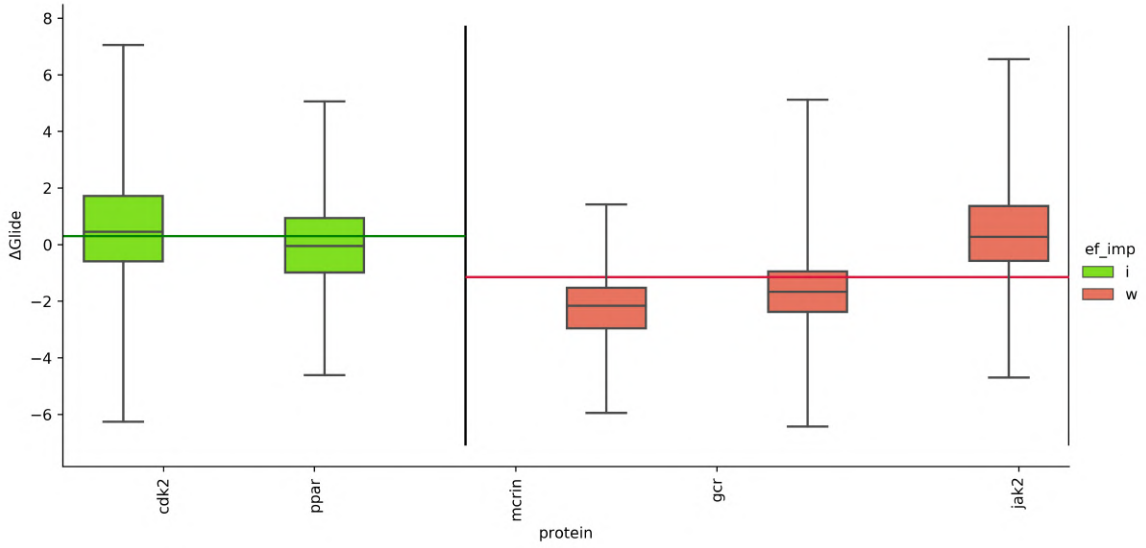
3.6.3 Results Protocol 2

Table 3.6: $\%EF$ changes upon simulation with protocol 2

Family	Protein	Docking $\%EF_{max}$	I.F $\%EF_{max}$	EF_{ratio}	Threshold
KINASE	cdk2	0.70	0.68	*0.97*	50
	jak2	0.60	0.60	1.00	50
NHRs	gcr	0.42	0.10	*0.24*	50
	mcrin	0.54	0.50	*0.93*	50
	ppar	0.28	0.44	**1.57**	50
KINASE	cdk2	0.69	0.60	*0.87*	100
	jak2	0.45	0.51	**1.14**	100
NHRs	gcr	0.50	0.22	*0.45*	100
	mcrin	0.58	0.65	**1.13**	100
	ppar	0.33	0.36	**1.09**	100

As we can observe in Table 3.6 protocol 2 affects the proteins in a different way than protocols 0 or 1. With this protocol, only three of the five proteins present the same trend for both the top50 and top100 compounds: the ppar is the only one that presents an improvement of the EF, while cdk2 and gcr present a loss. The jak2 protein presents the same EF as the initial docking in the top50 compounds, and an improvement on the top100, while the protein mcrin presents a loss of

(a) ΔGlide distribution for each family for all the compounds



(b) ΔGlide distribution for each family for all the compounds by activity

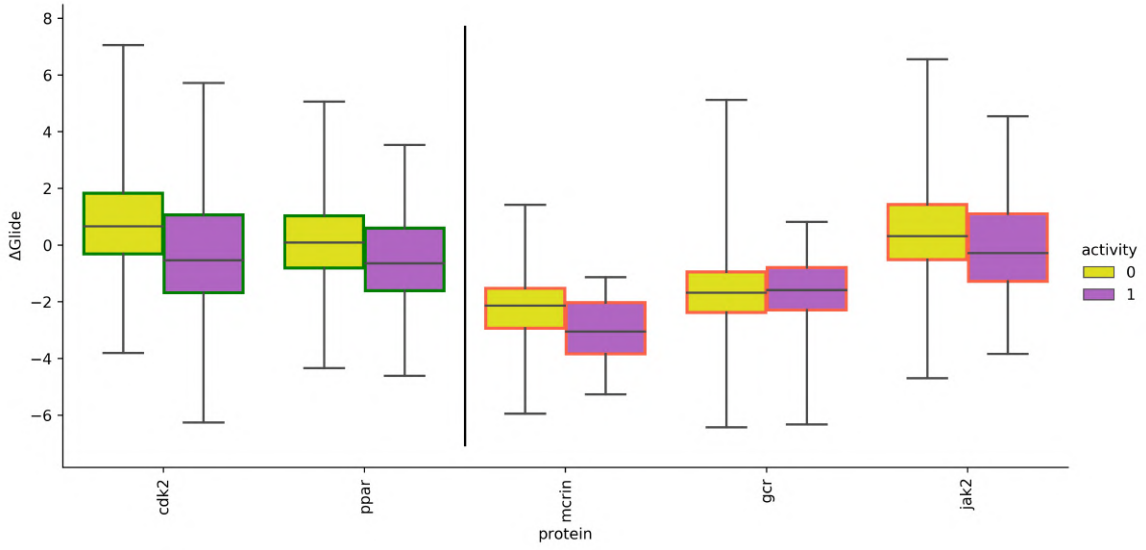


Figure 3.13: ΔGlide distributions for protocol 1

EF on the top50 and an improvement on the top100 compounds. The accuracy presents the same trends as the EF regarding the methodology influence, just like with the other protocols.

When we observe how the ΔGlide metric changes we can observe in Figure 3.14a that the average value of the systems with a loss of EF is lower than that of the systems with an improvement of, or the same, EF. We can also observe that the systems with a loss present a lower range of ΔGlide , or a similar average, while the ppar and jak2 proteins present a lower average and a lower minimum of ΔGlide for the active compounds. The cdk2 protein presents a behaviour similar to jak2 and ppar, despite its general loss of EF; when looking closely at the accuracy value, we can see that this behaviour is normal, since the loss is that of one single compound.

3.6.4 protocol comparison

Table 3.7: EF_{ratio} changes upon simulation with different protocols

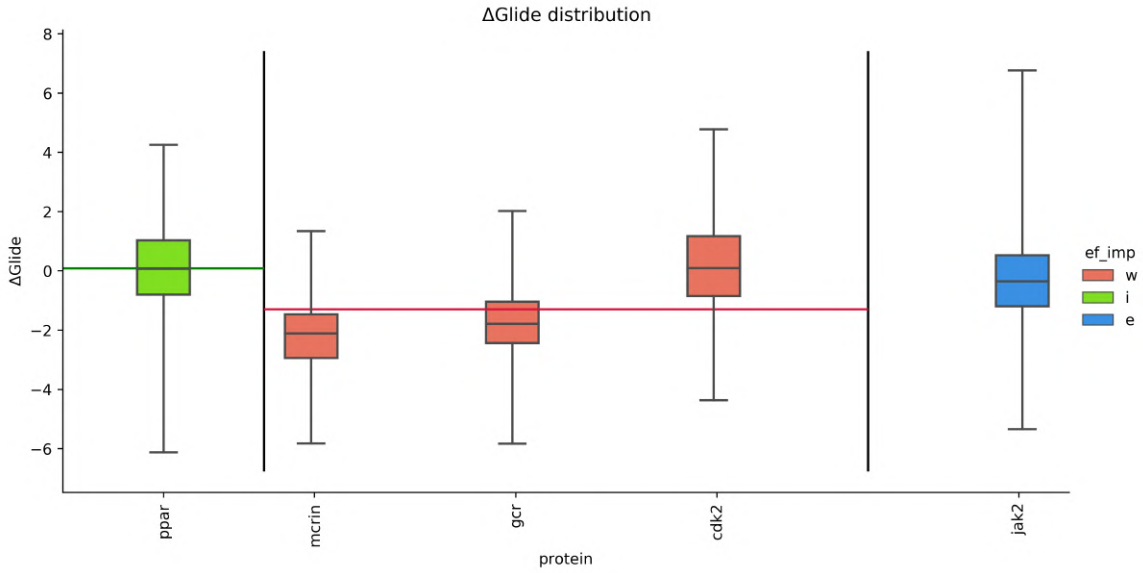
Family	Protein	Protocol 0	Protocol 1	Protocol 2	Threshold
KINASE	cdk2	**1.14**	**1.2**	*0.97*	50
	jak2	*0.87*	*0.7*	1.00	50
NHRs	gcr	*0.55*	*0.41*	*0.24*	50
	mcrin	**1.07**	*0.86*	*0.93*	50
	ppar	**1.83**	**2.33**	**1.57**	50
KINASE	cdk2	**1.03**	**1.03**	*0.87*	100
	jak2	*0.97*	*0.86*	**1.14**	100
NHRs	gcr	*0.55*	*0.48*	*0.45*	100
	mcrin	**1.27**	**1.13**	**1.13**	100
	ppar	**1.21**	**1.39**	**1.09**	100

An easy way to compare the influence of the protocol used for the simulation is to compare the EF_{ratio} of the different protocols for the same protein, since the EF_{ratio} tells us the how much have the EF and accuracy changed upon simulation. Table 3.7 summarizes the EF_{ratio} of the different protocols tried.

We see that the protocol 1 doesn't really change the EF tendency from protocol 0, with the exception of the mcrin protein all the other 4 proteins maintain the EF tendency on the top50 compounds and all 5 of them when looking at the top100 compounds. We observe a bigger loss of enrichment for the jak2 and gcr systems; while the cdk2 keeps the improvement previously observed, and the ppar system gets a slight improvement of the EF. The only system that presents a change of tendency is the mcrin at the top50 compounds, where it goes from an improvement of 7% to a loss of the 14% of active compounds.

The protocol 2 seems to introduce more changes on the EF tendency. For the top50 compounds this protocol is able to change the tendency for 3 of the 5 systems, and for the top100 compounds it inverts the tendency of the two kinases.

(a) ΔGlide distribution for each family for all the compounds



(b) ΔGlide distribution for each family for all the compounds by activity

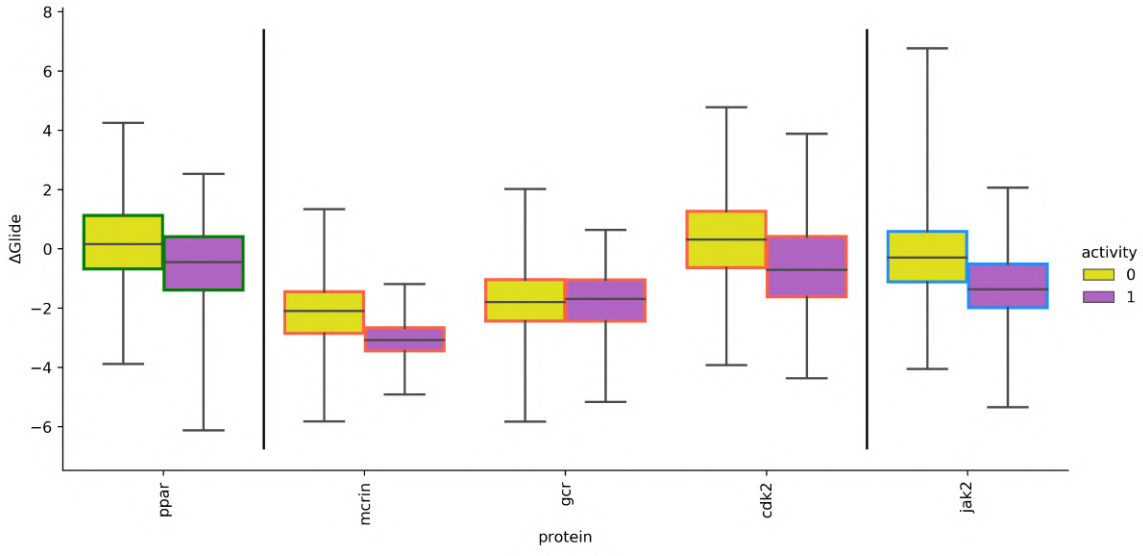


Figure 3.14: ΔGlide distributions for protocol 2

For the kinase family we can see that the EF change is almost reversed by protocol 2: the cdk2 loses EF for the top50 and top100 compounds while the jak2 maintains its EF on the top50 compounds, and improves it on the top100 compounds.

For the NHRs family, protocol 2 only changes the EF tendency from improvement to loss for the mcrin on the top50 compounds, while maintaining the improvement for the top100. It isn't able to change the tendency of the gcr nor the ppar proteins which retain their loss of EF and its improvement respectively.

3.7 Conclusions

Our first conclusion concerns the solvent to be used to perform the simulations. Even though the VDGBNP implicit solvent model should theoretically provide us with more accurate simulations than the OBC, which is a simpler model, we've consistently observed better results for the OBC solvent. The reason for this behaviour probably lays on the fact that VDGBNP has been parametrized to match the experimental values of a training set while the OBC hasn't; which makes it a more general solvent model.

In second place we have observed a split of the system regarding whether or not we can improve the results of the docking protocols using one simulation over all the systems. Upon performing the simulation half of the systems improve their rankings while the other half see it worsened or equal.

In third place, after reviewing the characteristics of the systems studied, we observe that the systems for which the simulation introduces error present: (i) really low Solvent Accessible Surface (SASA) values and (ii) most of them belong to the NHRs receptor, family that presents an almost completely occluded pocket.

In fourth place, changing the simulation protocol affects in different ways the improvements of EF observed, even to proteins belonging to the same family.

In fifth place, upon studying the changes in Glide score introduced by the PELE simulations, we observe that in all the cases this score is lowered. When this increase in the Glide score is separated by active and inactive compounds, we observe that in the systems with a loss of EF the inactive compounds improve their Glide score the same or even more than the active compounds. This trends in the Glide score means we're introducing noise with the PELE simulations. We're obtaining better poses for both the active and inactive compounds, making them harder to differentiate by their score.

Due to all these specific conclusions we can make the following general conclusions:

- The reason behind the improvement of EF for only half of the systems are related to the structural characteristics of the pockets.
- Finding a general protocol capable of improving the sampling for all the possible proteins is an extremely difficult task, due to the big differences on the proteins structure from one family to another.

- We see, for example, that increasing the RMSD to the initial Glide docking pose introduces significant enrichment. This, however, is quite difficult to obtain in very tight systems, with fully occluded active sites, like the NHRs. In these systems, we observe too many false positives as a result of enforcing the induced fit. While these cases should be filtered out when considering the total energy, all our attempts to do so did no work, as a result of inaccuracies in such a large number like the total energy.

- The many differentiating characteristics of the proteins obliges us to use a protein-specific protocol in order to improve the Glide docking results. With more detailed studies and a better understanding of which proteins have closely related pockets, we may be able to get to a pocket-type-specific protocol.

Chapter 4

Retos project: SilicoDerm

4.1 Introduction

This chapter presents the results from our collaborative retos project with Almirall, known as Silicoderm. The time scope comprehended is from the beginning of 2018 until October 2018. It constitutes one of the main goals of this PhD thesis: the application of our techniques to a real drug discovery campaign. The results shown in this chapter describe the application of the workflow developed, which has been previously explained, during this thesis to a real case of drug discovery.

The overall aim is to increase the number of true actives at the top position of the ranking in a VSs campaign. Most of the current docking technologies consider proteins as rigid entities, which means they generate one single pose for the protein-ligand complex (pink point in Figure 4.1). We account for the protein's flexibility and generate thousands of poses by performing a PELE simulation. Hopefully, the poses with a lower energy (circle in Figure 4.1) will be near-native poses.

Due to the confidentiality of this particular project we cannot reveal the real name of the protein, crystals or molecules studied. In this project we'll be focusing on the study of a kinase implied in dermatologic diseases as our target. We started the study with 6 public crystals named from A to F; later on, we got 1 in-house crystal, crystal G, from Almirall.

Performing all the work explained in this chapter has taken approximately 6 months of computations and implied the extensive use of the Marenostrum supercomputer and the purchase of three new computers in the group, of 20 cores each (to perform thousands of Glides' re-scoring calculations).

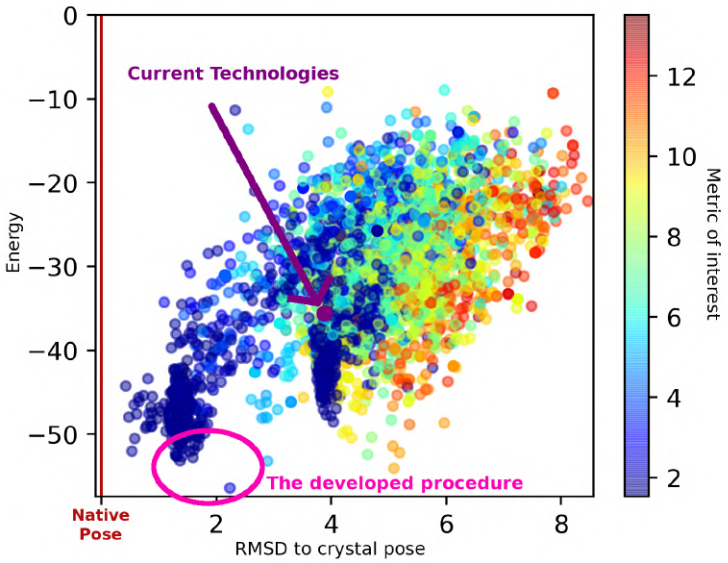


Figure 4.1: Visual summary of how our procedure aims at.

4.2 Simulations set up

4.2.1 Crystals preliminary study

Given our findings on the DUD-e dataset, we decided to study the simulation's behaviour on the 6 public crystals we had at the beginning of the project, developing a system specific simulation protocol. The first step was to prepare the crystal structures using the prepwizard program [69] from Schrödinger, to fill missing side chains, loops, and to optimize the hydrogen's positions; both with waters and ions present during the process, generating the WAT set, and without them, generating the NO_WAT set. The table 4.1 summarizes the basic characteristics of each crystal.

Table 4.1: Crystal's structures main characteristics.

Crystal	Receptor	Ligand	Waters in the crystal	Ions inside the BP	Public
crystal_A	receptor_A	ligand_A	yes	no	yes
crystal_B	receptor_B	ligand_B	no	no	yes
crystal_C	receptor_C	ligand_C	yes	yes	yes
crystal_D	receptor_D	ligand_D	yes	no	yes
crystal_E	receptor_E	ligand_E	yes	yes	yes
crystal_F	receptor_F	ligand_F	yes	no	yes
crystal_G	receptor_G	ligand_G	yes	no	no

While preparing the crystals we noticed that crystal_C presents a metal ion inside the BP. Since we know it isn't involved in the binding process, and that it is a crystallographic artefact, we cannot consider the ligand pose as the right one; thus, we're unable to ascertain whether or not we're

predicting the right pose with the simulation for this crystal. This led us to discard crystal_C from our data set.

We also noticed that the number of water molecules present in each crystal is variable. It varies so much that crystal_B doesn't present any water molecules, so it hasn't been included on the WAT set.

Next, we run PELE simulations for the two previously prepared sets of crystals: the WAT and the NO_WAT sets, using the adaptive protocol tested on the DUD-e dataset, protocol 0 in this chapter, which introduces no restraints, nor biases on the simulation. We expect different behaviours on the ALL_WAT and the NO_WAT sets for any simulation protocol we try. On one hand, we expect that any protocol applied to the crystals without waters or ions (NO_WAT set) will allow for an extensive exploration of the BP by the ligand; on the other hand, we expect that, when the protocols are applied to the WAT set, they will perform a more restricted exploration of the BP due to the presence of the less mobile waters.

We use the energy profiles to study the simulation's behaviour. The energy profiles depict the PELE BE on the y-axis, and the ligand RMSD on the x-axis, which is the heavy atoms RMSD between the simulation pose and the crystal pose. With this figure we can estimate how much conformational space does the ligand explore, and if it's exploring the desired region or not. In this type of profile, we would like to see a lot of points at the 1-2 Å region; which we consider the true positive's region, since we use the crystal pose as the right pose.

All the energy profiles shown in this chapter and its corresponding SI follow the colouring schema: blue means the WAT systems, cyan means the NO_WAT systems and teal means the STR_WAT systems (explained later on). The initial point of the simulation is coloured red in all cases; when it isn't present it means the initial structure has a positive PELE BE.

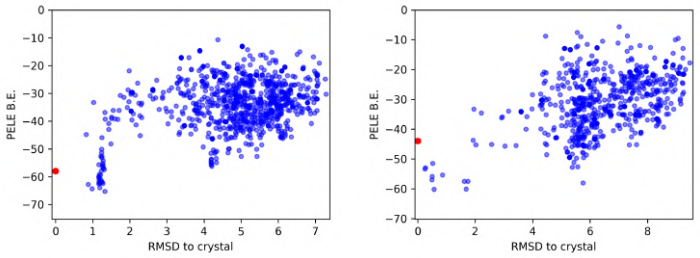
The origin of the ligand-protein complexes used in the simulations, the initial point of the simulations, and which protocol has been applied to each dataset is summarized on Table 4.2.

The energy profiles generated by protocol 0 (Figure 4.2) show significant differences between the WAT set and the NO_WAT set. It also shows a poor exploration of near native binding poses, as most poses present a RMSD over 3 Å in all simulations.

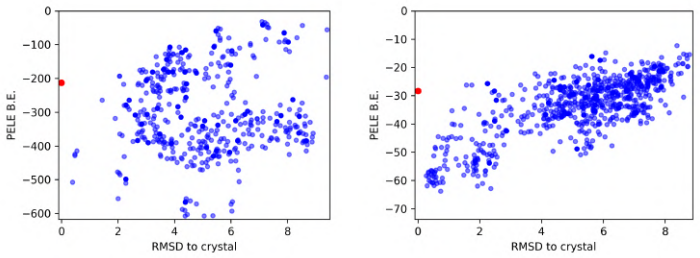
On the WAT simulations (panels from (a) to (d) in Figure 4.2) the poses with the best PELE BE do represent the binding pose for crystals with the exception of crystal_E, which doesn't present a well defined minimum, instead the poses with the lowest PELE BE are spread within the 4 Å to 6 Å region.

On the NO_WAT simulations, panels from (e) to (i) in Figure 4.2, we can observe that the poses with the best PELE BE do not represent the binding pose with the exception of crystal_A and crystal_F. Crystal_A presents one clear minimum on the 1 Å region, while crystal_F presents two minimums: a small one at the 1 Å region and another similarly small minimum at the 5.5 Å region. While for the crystals B and D the pose with the best PELE BE doesn't match that of the crystal,

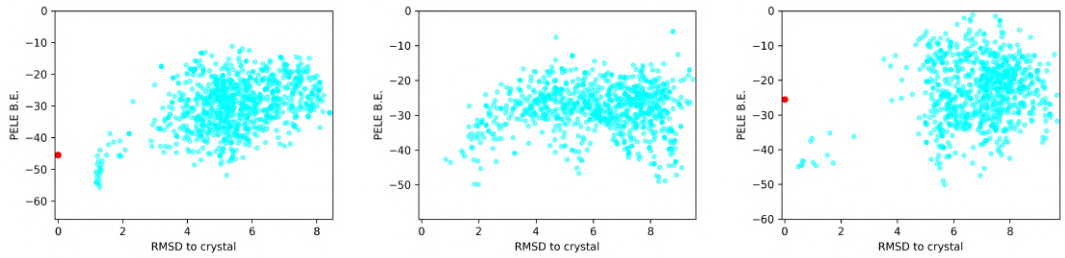
(a) ALL_WAT crystal_A protocol 0 (b) ALL_WAT crystal_D protocol 0



(c) ALL_WAT crystal_E protocol 0 (d) ALL_WAT crystal_F protocol 0



(e) NO_WAT crystal_A protocol 0 (f) NO_WAT crystal_B protocol 0 (g) NO_WAT crystal_D protocol 0



(h) NO_WAT crystal_E protocol 0 (i) NO_WAT crystal_F protocol 0

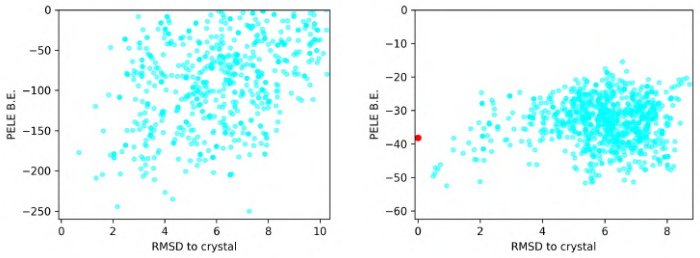


Figure 4.2: Energy profiles for the simulations using the protocol 0 over the crystals with all the water molecules but no ions (a to d) and the crystals with no waters nor ions (e to i)

Table 4.2: This table shows which is the origin of the structures in each set, which is the initial structure used for the simulations

set	structure generation	simulation initial pose	Protocols tried
ALL_WAT	crystal	crystal pose	0-9
NO_WAT	crystal	crystal pose	0-9,11,12
STR_WAT	crystal	crystal pose	8,9,12
crossdocking	crystal's crossdocking	docking pose	9-12

they present an RMSD over 4 Å to the crystal pose, the overall profile is good, unlike crystal_E.

All the energy profiles in 4.2 show a poor exploration of the region where we can find poses similar to the binding pose of the crystal. None of the WAT simulations has more than 50 points at the 1-2 Å region; the NO_WAT simulations present less than 50 points for crystals D and E, and less than 100 for crystals A and F, as shown in Figure 4.3. Thus, we needed to improve our simulation protocol in order to explore more this region.

We proceeded to try 7 different simulation protocols where we tried to improve the sampling by changing the simulation parameters without introducing any kind of bias just changing the duration, the number of processors, the number of epochs, and all their possible combinations. A summary of the characteristics of every protocol tried can be found at Table B.1.

The energy profiles, which can be checked on the supplementary data, show a bit more sampling in the 1-2 Å region but they still present significantly more sampling on other regions. We observe that the WAT models energy profiles barely change with independence of the simulation protocol, they're more or less populated but the general shape is kept; and they present better minimum than the no WAT simulations.

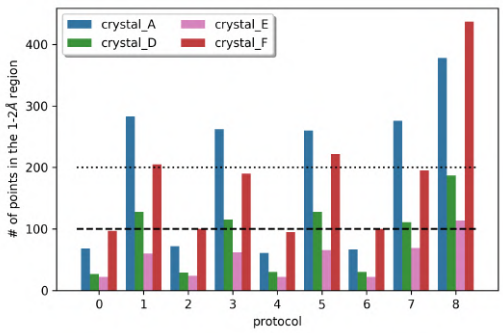
A summary of how well each protocol explores the 1-2 Å region is shown in Figure 4.3. Each of the plots show how many points are generated in the 1-2 Å region by each protocol over the different crystals sets.

When studying the reason for the better profiles for the ALL_WAT set we discovered that the presence of all the water molecules in the crystal significantly diminishes the sampling space of the ligands. From the crystals structures we know that some of these water molecules play an important role on the binding of ligands, so they should be kept in order to obtain the right binding pose. In order to take into account these key water molecules while removing the rest of waters we define the concept of structural waters.

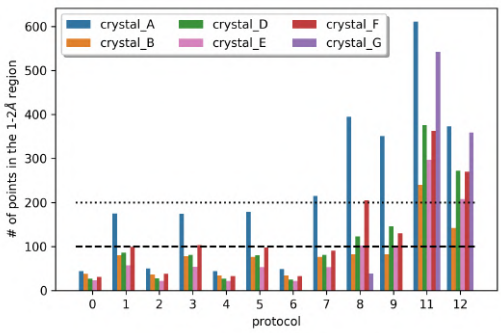
We define a structural water as any water that has its oxygen atom within 3.25 Å of 3 or more N or O atoms from the protein, or at least 1 N or O atom within 2 Å and another within 3.2-5 Å. We selected the waters automatically for each of the crystals using our own python script, and we prepared the resulting structures with prepwizard[69], to optimize the hydrogens' positions, creating a new set of models with only the structural waters present (STR_WAT).

In addition, upon further study of the profiles, we saw that the profiles of crystal_E present a

(a) ALL_WAT set



(b) NO_WAT set



(c) STR_WAT set

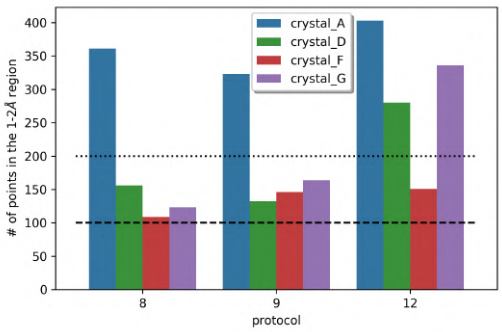


Figure 4.3: These bar-plots summarize how well each protocol explores the 1-2 Å region for each of crystals. The x-axis shows the protocol, the y-axis shows the number of points inside the 1-2 Å region and the colours represent the crystals.

spread population, with no clear minimum and almost no exploration of the 1-2 Å region, so we decided to study the characteristics of this compound further. The compound in this crystal is an ATP analog, which means that it is a long flexible compound with three phosphorus atoms; originally, it also presents a SO_4 molecule on the other side of the BP, which was removed upon the protein preparation. All these characteristics make this compound really hard to assess, because it will be highly difficult to dock, and any force-field used during the simulation will overestimate the energy due to the high charges; plus, none of these characteristics are desired in prospective drugs. Therefore crystal_E was eliminated from the simulation calibration.

Using Spawning: adding mechanistic information

So far we haven't used any information about the binding mechanism of these compounds, but we do know how they bind, and which interactions are important. This kind of information is used when performing the traditional methodologies such as docking or pharmacophores, and it'll be used later on when we apply these methods to the compounds pre-screening.

Thus, we decided to use the information about the key interaction that'll be used later on, which is the presence of an h-bond between the protein and the ligands, also present in all crystal structures. So, we now perform protocol 8 on the three sets; this protocol uses the epsilon parameter of adaptive PELE to steer the simulation toward sampling poses with a low distance between the atoms implied in the key h-bond, incorporating the aforementioned information.

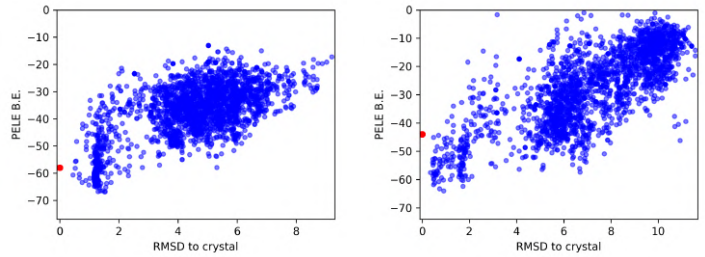
We can see the resulting energy profiles of protocol 8 on Figure 4.4. Thanks to the introduction of the bias we've obtained a protocol capable of sampling the 1-2 Å region quite well, but which still generates minimums in regions with higher RMSD values (false positives).

Up to this point we performed 85 different simulations and their analysis in order to obtain a protocol able to explore the 1-2 Å region properly, although it also samples other undesired minimums. This amount of computations was performed in 2.5 months thanks to the computational resources of the supercomputer Marenostrum IV, and to the PELE VS platform (developed during the first years of this thesis) which automatizes the whole procedure, allowing us to prepare, launch and analyse the simulations in a simple and quick way.

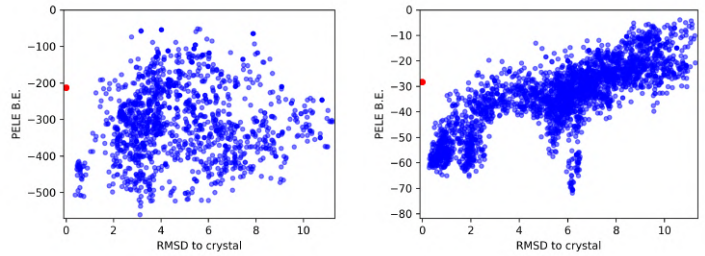
4.2.2 Cross-docking study

Once we start to work on the real VSs campaign we won't have crystals of the compounds to study, instead, we'll be working with structures derived from a docking process. Thus, we decided to apply the simulations to structures derived from a docking process and finish the refinement of the simulation protocol taking into account only this type of structure. The best way to generate this kind of structure while knowing the right binding pose, is to perform a cross-docking of all the crystals, that means, docking the compounds from each crystal against all the receptors we have.

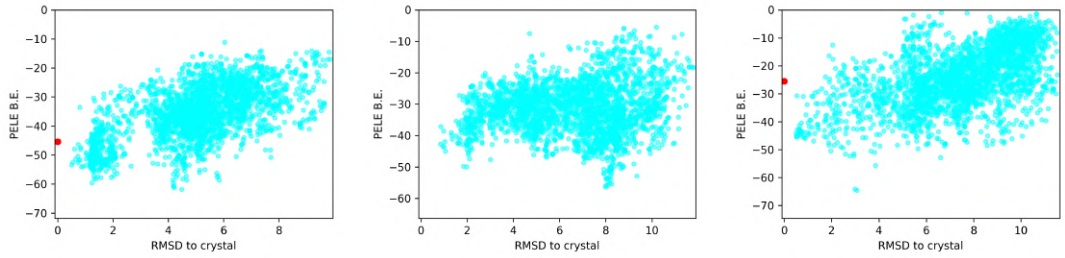
(a) ALL_WAT crystal_A protocol 8 (b) ALL_WAT crystal_D protocol 8



(c) ALL_WAT crystal_E protocol 8 (d) ALL_WAT crystal_F protocol 8



(e) NO_WAT crystal_A protocol 8 (f) NO_WAT crystal_B protocol 8 (g) NO_WAT crystal_D protocol 8



(h) NO_WAT crystal_E protocol 8 (i) NO_WAT crystal_F protocol 8 (j) NO_WAT crystal_G protocol 8

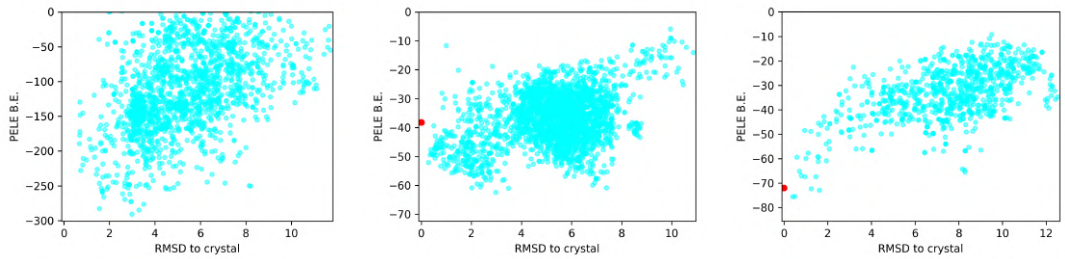
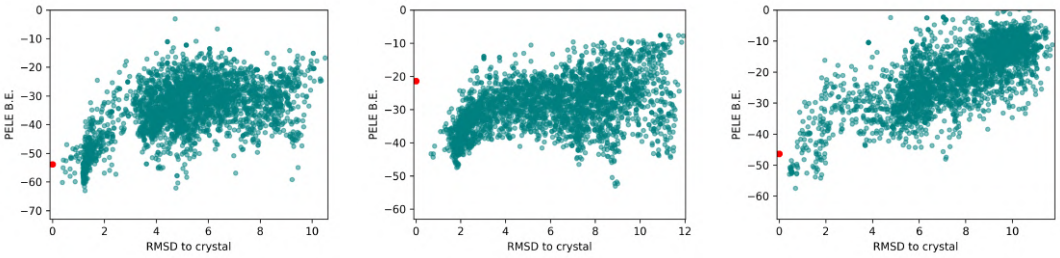


Figure 4.4: Energy profiles resulting from PELE simulations using protocol 8.

(k) STR_WAT crystal_A protocol 8 (l) STR_WAT crystal_B protocol 8 (m) STR_WAT crystal_D protocol 8



(n) STR_WAT crystal_F protocol 8 (o) STR_WAT crystal_G protocol 8

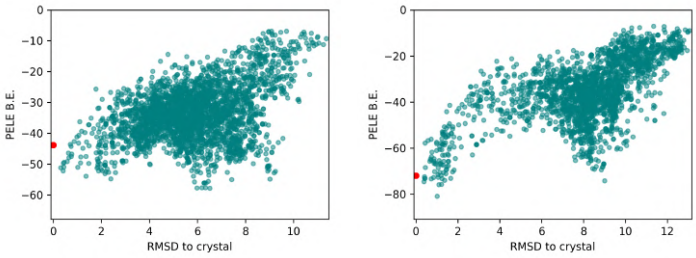


Figure 4.4: Continuation. Energy profiles resulting from PELE simulations using protocol 8.

In order to generate this new set of data, we opted for using only the proteins from each selected crystal as receptors. We took each crystal and eliminated the waters, ions and ligands leaving only one chain of protein, and prepared them again using prepwizard [69]. Then we prepared the grids and performed the docking process with the Glide [34, 23] software from Schrödinger suite-2018-1. For the docking we used the default parameters and allowed the program to place the ligand wherever it wanted on the BP, without taking into account the key h-bond.

We used all the receptors from all the crystals except receptor_C and receptor_E, and we performed the docking with the compounds from crystals: A, B, D, F and G. This combination of receptors and compounds generated 25 protein-ligand complexes.

When studying the structures derived from the docking we observe that the characteristics of each compound vary depending on which receptor is used for the docking. The main characteristics to consider are the RMSD of the heavy atoms between the docked and crystal poses, and the distance between the two atoms involved in the key hbond. With these two metrics we can estimate how good a pose is. We also use the PELE BE as a characteristic, because it allows us to see how energetically favourable a pose is in our force-field. All this information is summarized in Table 4.3.

We observe a better performance of the docking program when performing a self-docking experiment instead of a cross-docking. The self-docking means we dock the ligand into the protein obtained from the same crystal; while the cross-docking means we dock the ligands derived from the other crystals into the protein of one crystal.

Table 4.3: Crossdocking simulations starting structures' characteristics.

receptor	ligand	ligand RMSD	hbond distance	PELE BE
receptor_A	ligand_A	1.168360	0.79191	-50.3046
receptor_A	ligand_B	8.880020	0.44060	-30.6264
receptor_A	ligand_D	1.662200	0.21904	-24.7331
receptor_A	ligand_F	6.924620	0.29330	-49.9489
receptor_A	ligand_G	9.783690	0.20260	-32.2795
receptor_B	ligand_A	4.342870	0.78690	-33.1025
receptor_B	ligand_B	1.652870	0.97630	-41.4464
receptor_B	ligand_D	1.168550	0.96797	-24.8259
receptor_B	ligand_F	6.268810	0.65270	-35.7882
receptor_B	ligand_G	6.365760	0.93745	-22.6156
receptor_D	ligand_A	4.463980	0.22848	-38.3735
receptor_D	ligand_B	8.105680	0.75090	-37.6212
receptor_D	ligand_D	0.693984	0.16734	-39.3428
receptor_D	ligand_F	5.159510	0.18619	-36.8987
receptor_D	ligand_G	9.961790	0.53930	-32.6344
receptor_F	ligand_A	3.867120	0.83828	-34.3218
receptor_F	ligand_B	1.556140	0.94498	-43.5171
receptor_F	ligand_D	1.003210	0.17095	-13.7769
receptor_F	ligand_F	0.291288	0.01015	-41.6651
receptor_F	ligand_G	3.551170	0.59664	-44.4961
receptor_G	ligand_A	3.883020	0.95563	-35.6181
receptor_G	ligand_B	1.500830	0.02597	-45.6654
receptor_G	ligand_D	6.224360	0.65126	-27.4961
receptor_G	ligand_F	2.635000	0.80769	-35.7909
receptor_G	ligand_G	1.080900	0.05426	-55.2329

In order to avoid adding mechanistic information to the simulation, which is compound-specific, we decided to try first on this new dataset protocol 9. Protocol 9 is identical to protocol 8 with the only difference on the bias introduced. Protocol 9 s biased towards the structures with lower PELE BE instead of the ones with lower h-bond distance, as in protocol 8. With this, we hope to obtain an equally good exploration of the 1-2 Å region without having to add the mechanistic information.

The reason behind trying to avoid adding mechanistic information is that, while we know how the crystals in the compounds bind to the protein, we won't have this information for the VSs campaign. We know there will be, at least, one h-bond with a key hydrogen of the protein, but we can only guess which atom of the compound will be participating.

We also decided to try this same protocol on the crystal structures in order to compare protocol 9 with protocol 8. The reason we need to make this comparison is that, even for the self-docked structures, those where the ligand and the receptor are from the same crystal and which are the closest to the crystal structures, we observe that the starting pose presents an RMSD over 1 for all of them, with exception of ligand_D docked to receptor_D and ligand_F docked to receptor_F.

Given our findings about the presence of all waters we tried this protocol only on the NO_WAT and STR_WAT sets without crystal_E.

The profiles obtained from protocol 9 present a better exploration of the 1-2 Å region on the STR_WAT set than the NO_WAT set, with the exception of crystal_D, which presents a better exploration for the NO_WAT set. However, this protocol doesn't only improve the sampling on the desired region, but it also samples more the minimums in the 4-6 Å region, exaggerating what we know are incorrect minimums. Thus, by applying this protocol on the crystal's sets we enhance the sampling of low BE poses, whether they are true or false positives.

When studying the energy profiles of the crossdocking set we notice different outcomes for the same compound, depending on which receptor was used for the docking procedure, in a similar fashion to what happens when performing the initial docking. Only ligand_A (the ligand derived from crystal_A) presents a good exploration of the 1-2 Å region on almost all possible receptors, with the only exception of crystal_B as receptor, where it shows a poor exploration. All the other compounds present a good exploration of the 1-2 Å region when the initial docking procedure generates a pose with an RMSD value lower than 2 Å; otherwise, simulations get trapped in other local regions.

As we can see in Figure 4.5f, we observe a good exploration of the 1-2 Å region (more than 100 points) for compound A, and some of the other compounds just for one or two receptors. Thus, we decided to refine more the simulation protocol by: (i) changing the epsilon constraint from the PELE BE back to the distance between the two atoms forming the key hbond, (ii) reducing the perturbation radius, and (iii) by steering all the compounds towards the average center of mass from all the crystalline ligands poses).

With the first change we force that a 33% of the processors explore the poses with the lowest h-bond distance found along the simulation more; with the second and third changes we tried to minimize the exploration in other regions far away from the 1-2 Å region. Overall, with this protocol 10, we introduce system specific mechanistic experimental information that, we expect, should improve the exploration near-native poses.

When checking the new protocol 10, we observe that in most of the cases the profiles have changed; we noticed how the minimums outside the 1-2 Å region now present a smaller number of points and are not so deep as with protocol 9. We also observe a general improvement on the 1-2 Å region exploration. As we can see in Figure 4.5g, the number of compounds with more than 100 points in this region is still low, but now more compounds present points in this area.

In light of the changes observed, we decided to refine the protocol a bit more by (i) increasing the epsilon constraint, in order to increase to 75% the number of processors to explore the lower h-bond distance poses, and (ii) allowing the simulation to make bigger movements 20% of the time, as long as the ligand doesn't form the hbond. We've called this new simulation protocol: protocol 11.

We observe an improvement on the sampling at the 1-2 Å region for almost all of the 25 studied

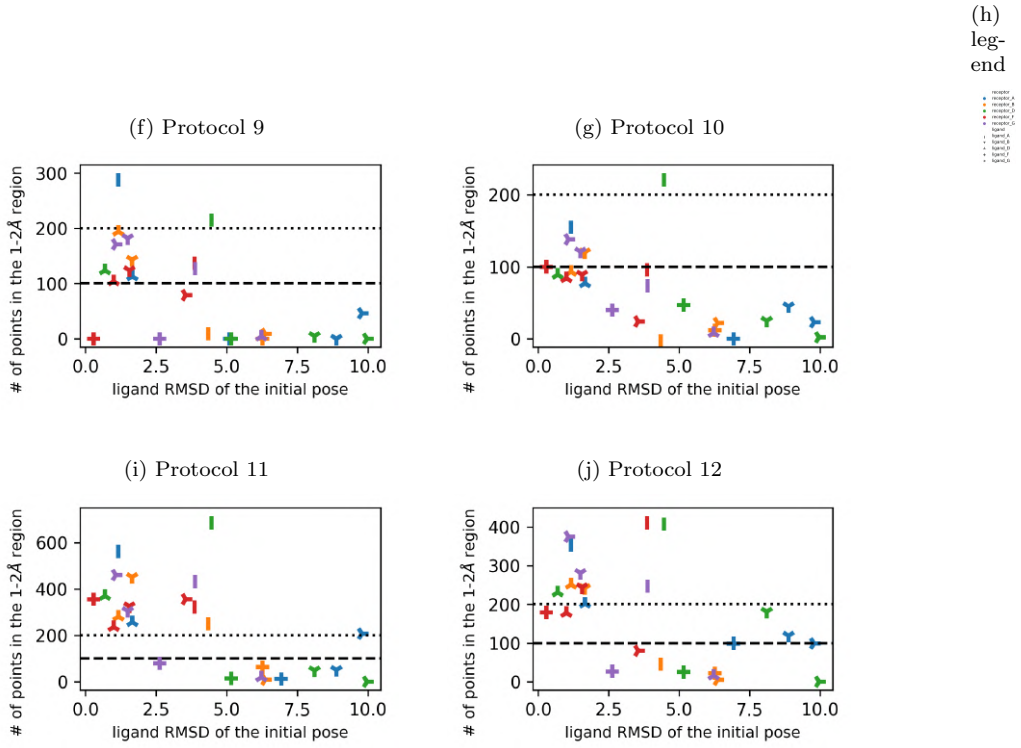


Figure 4.5: These plots show how many poses in the 1-2 Å RMSD region to the crystal pose are generated with the PELE simulation using the protocols 9 in figure 4.5f, 10 in figure 4.5g, 11 in figure 4.5i and 12 in figure 4.5j. The y-axis shows the number of points on the 1-2 Å region, the x axis shows the RMSD of the initial docking pose, the colour represents the receptor used during the docking process and the symbol the ligand. Pleas Note the change in scale in the different panels. The panels do not share the same scale.

structures; we also observe the apparition of other minima in other regions of the RMSD axis. Though the improvement of sampling is constant for all the structures the introduction of new minima aren't, it is highly dependant on the ligand-protein complex generated by the docking. I.E.: ligand_A presents a secondary minimum or not depending on which receptor is used for the docking, while ligand_G generates better profiles, with only one minimum in the desired region, no matter which receptor we use.

Even though the secondary minima isn't constant it is a general tendency. In order to avoid these other minima, we tried to apply the protocol 11 on the structures derived from the docking process using the protein receptor plus the structural waters. As mentioned before, the presence of water molecules is needed by some of the compounds in order to bind the protein, they also represent a steric constraint that can help to avoid the apparition of the second minima.

The profiles of the simulations with waters present less minima, or at least less exploration of the undesired minima, without the h-bond formed, and in the case of ligand_A docked to receptor_A the waters presence avoids the formation of a second minimum with the h-bond formed. Nevertheless, insertion of water molecules could significantly hinder the initial docking of general compounds, as well as the sampling of compounds capable of displacing water molecules; due to the water's reduced mobility.

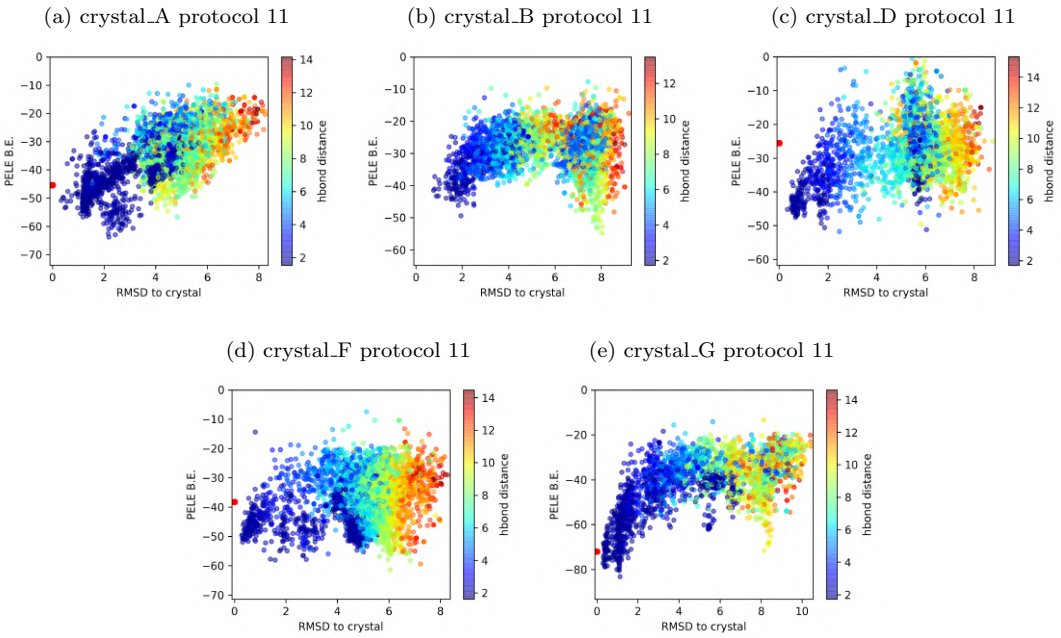


Figure 4.6: Energy profiles for the PELE simulations using protocol 11 on the NO_WAT set. In these images the colour represents the distance between the atoms involved in the key h-bond.

In order to make sure we have a good protocol; we performed the simulations with protocol 11

on the crystal structures without waters. The new protocols are biased towards structures with lower distances between the atoms involved in the key hbond, thus, we've decided to colour the plot using this metric in order to better study the simulation behaviour. In figure 4.6 we can observe the energy profiles derived from the simulation with protocol 11 on the NO_WAT set. These profiles present the secondary minima observed on the cross-docking simulations, in most of the cases the minima present short h-bond distances, which can mean we're introducing too much bias.

To reduce the secondary minima we decided to lower the epsilon parameter to 0.5, generating protocol 12, and applied it to the crossdocking set, where we observed that, even though the number of points in the 1-2 Å region has decreased considerably when compared to protocol 11, this protocol still provides a good exploration for most of the systems, with only 12 systems with an exploration below 100 points in the desired region, and 9 systems with more than 200 points. When looking at the profiles (Figures B.12 to B.14), with the exception of a couple systems, all the energy profiles have one clear minimum with the right h-bond distance and the secondary minima present a broken hbond, that is, the distance between the atoms is over 2.55 Å.

Given the good results we decided to try protocol 12 on the NO_WAT set. The results show clearer profiles. In most cases, the profiles only present one minima, located in the 1-2 Å region, and any remaining secondary minima are less populated and present a bad h-bond distance, that is, a distance over 2.5 Å, as we can see at Figure 4.7 left column.

We also tried protocol 12 on the STR_WAT to check whether the inclusion of water molecules could help eliminate the secondary minima or not. We observe that the inclusion of the structural water molecules helps to highly improve the energetic profile of crystal_D, but they don't significantly improve the profiles of the other crystals. Thus, with protocol 12, we've accomplished a protocol capable of good sampling with no need for water molecules.

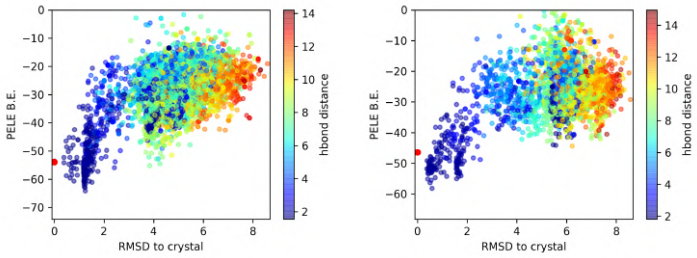
In summary, we have a simulation protocol capable of sampling sufficiently the 1-2 Å RMSD region, where the right binding pose and near-native poses are found. However, this protocol generates false positive minima in PELE BE, at regions with more than 3 Å RMSD away, which are not similar to the right binding pose. Importantly, we observe that most of the false minima do not present the key h-bond for binding, thus this could be a possible filter. Still, however, there are few false minima that present the key h-bond rendering this filter less useful.

4.3 Simulations re-scoring

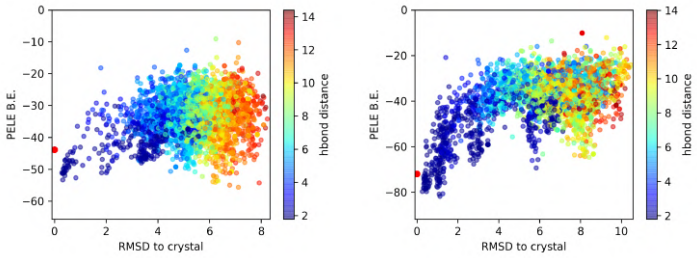
Given our incapability to avoid the formation of secondary minima when using the PELE BE to score the poses, and the good performance of Glide score on the kinase family: an average recall over 30%, we decided to re-score the whole simulation with the Glide scoring function.

In order to perform the re-score, we extracted each of the structures generated during the simulation and computed the Glide in place score. With this protocol, we use the poses from PELE, but

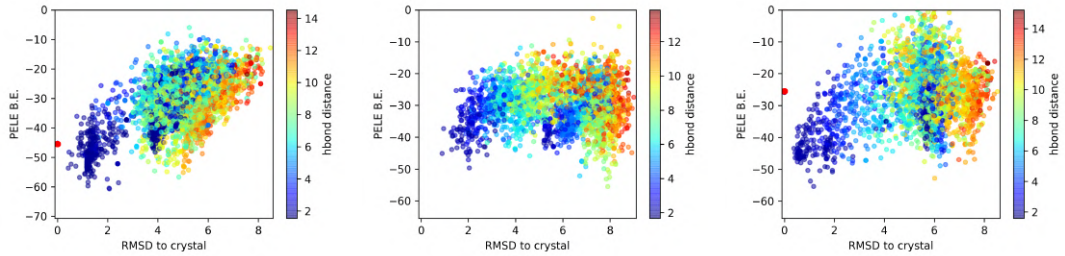
(a) STR_WAT crystal_A protocol 12 (b) STR_WAT crystal_D protocol 12



(c) STR_WAT crystal_F protocol 12 (d) STR_WAT crystal_G protocol 12



(e) NO_WAT crystal_A protocol 12 (f) NO_WAT crystal_B protocol 12 (g) NO_WAT crystal_D protocol 12



(h) NO_WAT crystal_F protocol 12 (i) NO_WAT crystal_G protocol 12

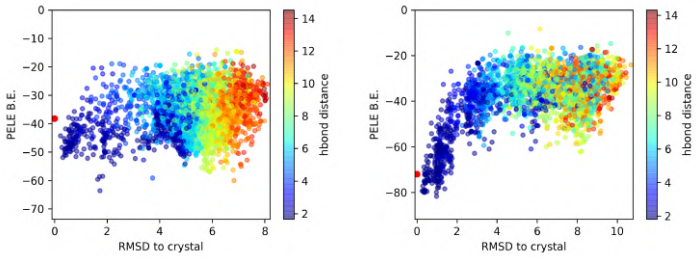


Figure 4.7: Energy profiles for the STR_WAT (panels from (a) to (d)) and NO_WAT (panels from (e) to (i) sets PELE simulations using protocol 12.

the binding energy estimation of Glide to generate the profiles.

The first set where we applied this new methodology was the crystals with structural waters and

protocol 9 simulation, which was the last simulation performed at that time. When comparing the energy profiles generated by PELE BE and Glide score, Figure 4.8, we observe that the minima present outside the 1-2 Å region have either disappeared or been reduced to secondary minima with higher scores than the ones in the desired region. Thus, the false positives have been removed from the selected poses.

Performing this re-score took us almost 2 weeks, while performing the simulations only took a couple of days, thus, we could try thrice the amount of simulation protocols than perform their re-score. The combination of the time needed to perform all the re-scores and the quick pace this project presents made us decide to use the re-score on the VSs data directly.

At this point we have a protocol that ensures a good sampling of near-native poses, while generating barely any secondary minima and the Glide re-score procedure that allows us to eliminate any possible secondary minima.

4.4 Application to VS data.

In this section we'll present the preliminary results of applying the procedure to a dataset forming part of a VS campaign. We'll work with 2.000 compounds, of which we know nothing about their activity or binding mode. This dataset, from now on the VS_set, is formed by compounds derived from a docking campaign previously performed by Almirall, who used different crystals as receptors and different water molecules depending on the receptor.

They provided us with the best 2.000 compounds based on their docking score plus their fitness to interact with the kinase, eliminating some false positives thanks to the group's know-how.

Even though the compounds derived from a docking procedure, we were provided only with the prepared 3D structure of the ligands, but no information on which was the receptor used to obtain them. Since PELE needs the structure of the protein-ligand complex we needed to perform a new docking with one of the receptors from the crystals. Due to the results of the cross-docking we chose the receptor_G as the receptor to use.

4.4.1 Compounds docking and study

We proceeded to perform the docking of all the 2.000 compounds to the receptor_G using the Glide SP from Schrödinger and used the Glide score to initially rank the compounds. We called this initial ranking the docking ranking. The docking protocol was set up to only write the best pose according to their score, which later on will be used to start the PELE simulation, and to generate poses with an h-bond with a specific hydrogen atom from the BP. With this set up. Glide will try to generate several poses for each compound, but it will only write the best one according to its internal score. For each compound, it will search for any atom capable of being an h-bond acceptor, and will try to place the compound with this atom within h-bond distance of the specified hydrogen from the

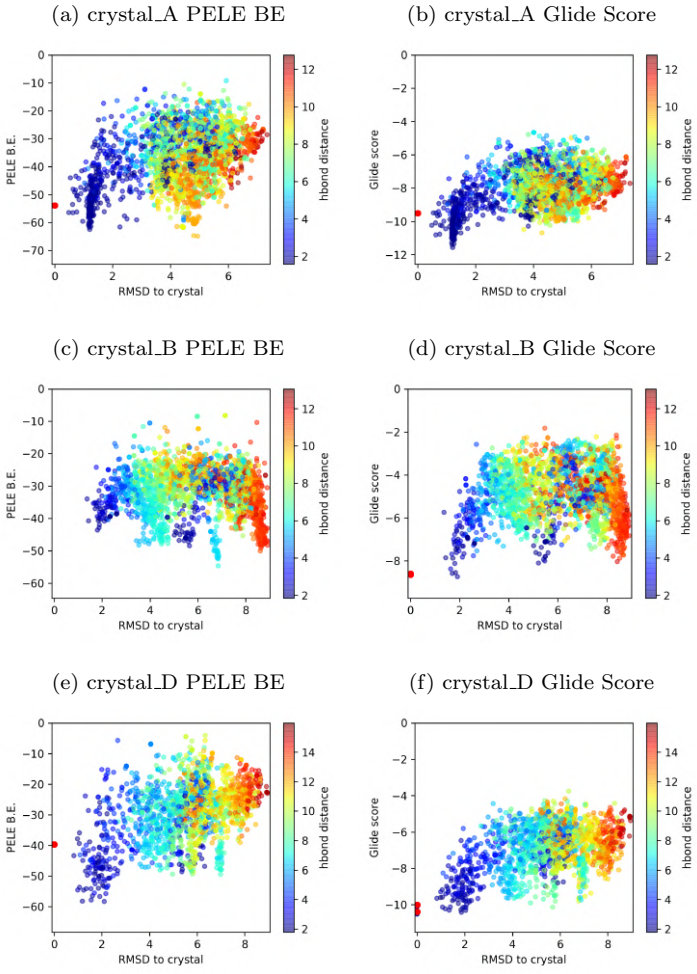


Figure 4.8: This figure shows the differences between the Glide score and the PELE BE. All the images show the structure derived of the PELE simulations using protocol 9, and x-axis shows the ligand RMSD to the crystal pose, and the colour indicates the h-bond distance. The y-axis shows the PELE BE for the images on the left column, and the Glide score for the images on the right column.

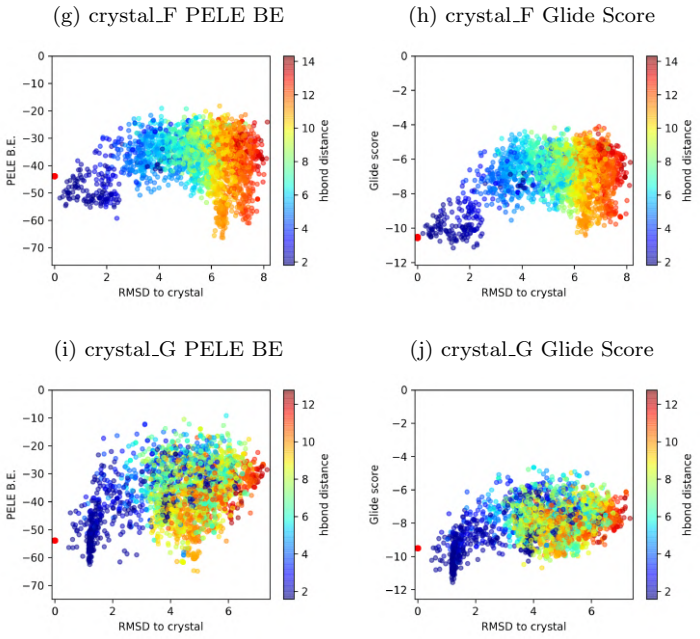


Figure 4.8: This figure shows the differences between the Glide score and the PELE BE. All the images show the structure derived of the PELE simulations using protocol 9, and x-axis shows the ligand RMSD to the crystal pose, and the colour indicates the h-bond distance. The y-axis shows the PELE BE for the images on the left column, and the Glide score for the images on the right column. Cont.

BP inside the protein, but without modifying the receptor_G at all. If by any chance the program isn't capable of placing the compound with an h-bond acceptor within the right distance of the key hydrogen, it will generate no pose for the given compound. We'll call this the constrained docking from now on.

When we performed the docking, Glide was able to generate a pose with the h-bond formed for most of the compounds, but there were 206 compounds for which Glide was unable to generate any pose mainly due to steric clashes with the receptor. It is important to note that the use of different docking programs (VINA, Glide or Rdock among many) will result in different poses and scores, and that even when using the same program Glide, the use of different precisions SP, XP or high-through put (which translates into different internal protocols) will generate different poses and/or scores, and even when using the same program and precision Glide SP, if the version is different (2018-1 or 2018-4), the pose may be the same but the score will change. It is due to this variability in score and pose generation, mixed with the difference in the receptor used, that even though all the compounds were derived from a docking campaign using Glide SP with the h-bond constraint, we were unable to dock part of the compounds. For the compounds we couldn't dock with the constrained docking, we performed exactly the same docking, but without the constraint (the non-constrained docking from now on), and we were able to obtain poses for these compounds.

Once we obtained the poses generated by glide, we started to prepare the PELE simulations with protocol 12. Since protocol 12 biases the simulation towards the poses with the lowest h-bond distance we need to specify the two atoms involved in the h-bond. The key hydrogen from the receptor_G remains constant for all the ligand-protein complexes, while the h-bond acceptor from the compound or ligand is specific and has to be extracted manually. In order to obtain this atom and check the fitness of the atom chosen by glide, we revised manually, one by one, the 2.000 compounds. While doing so we observed two interesting facts.

The first one was that the constrained docking sometimes picked atoms that couldn't be the right ones, due to how the known active compounds interact with the kinases; in this case we chose another h-bond acceptor from the compound as the interacting atom, and biased the simulation towards those poses with the atom closer to the key hydrogen. The second one was that, for the poses generated by the non-constrained docking, the program was able to place the compounds inside the cavity but far away from the region of interest, due to the big size of the BP. In both cases the problem lied in the incapability of Glide to generate poses with the right h-bond due to steric clashes with the receptor.

4.4.2 PELE simulations

Once we have the compound pose inside the protein complex and we have all the information needed to prepare the simulations, we automatically prepared all the simulations using the pose obtained by the constrained-docking as the starting pose. But we suspected that protocol 12 would not be able

to move the compounds with the initial pose obtained from the non-constrained docking towards the catalytic region, where the key h-bond is. To test this hypothesis, we performed a couple of simulations with different compounds and the result was positive. Protocol 12 starting from the non-constrained-docking didn't sample the catalytic region.

The reason why protocol 12 is unable to sample the desired region of the BP is that the compounds is placed far away from it, and protocol 12 has been designed as a local exploration simulation; which means it will move the compound inside a 4 Å box centered on the average center of mass of the crystals compounds. To solve this problem, and to sample the catalytic region of the receptor with the simulation, we decided to perform a short simulation before applying protocol 12; this simulation is set so it can bring the compounds towards the catalytic site. The protocol used for this simulation was called the equilibration step, because we will use the pose with the lowest energy from this simulation as the starting point of the PELE simulation with protocol 12. This new approach appears in table B.1 as the protocol 13.

All the preparatory steps to perform the PELE simulations were done using the VS_platform developed in this thesis, which, together with the computational resources of Marenostrum IV, allowed us to prepare, launch and analyse all the simulations (over 2.200) within one month and a half.

When studying the energy profiles of all the 2.000 simulations (data not shown), we observed that some profiles presented multiple minima with good h-bond distances. Since the re-score of the poses with glide, performing a score in place, had given us good results when applied to the energy profiles of protocol 9 on the NO_WAT set we decided to apply this methodology to the 2.000 compounds.

4.4.3 Glide Re-score

Since we're performing the score-in-place protocol, computing the Glide score implies creating one grid for each pose we re-score, because each of the poses generated by PELE not only presents a different pose of the ligand, it also presents a different receptor structure. If we were to re-score all the poses for all the compounds we would be re-scoring over 2.000.000 poses.

Given the computational cost involved in computing the Glide score in place for all the poses and compounds, we decided to reduce the number of structures to re-score for each compound. This reduction was done by clustering the poses generated by PELE by the RMSD among them. We expect that with this approach the number of poses to score is significantly reduced, without loosing any pose that may be of interest.

Since we don't know whether the compounds present activity or not, some compounds present more than one h-bond acceptor, and we don't know how they bind, we decided not to filter the poses by their h-bond distance.

In order to perform this re-score, the group bought 3 new machines of 20 cores each, that were

used intensively for 3 months. All the process was automatized, and 70% of the time was used in creating the receptors-grids.

4.4.4 Preliminary Results

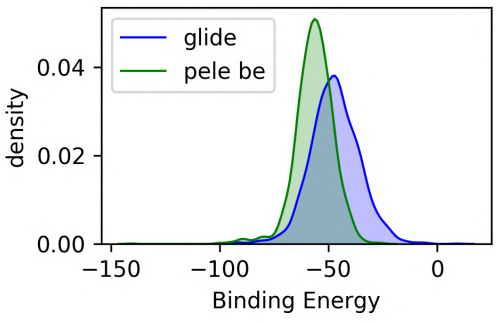
Since only a few hundred of compounds can be tested we now have to select the compounds with the highest probability to be active. This is the step where our methodology can influence the most. So far we have the docking ranking derived from our initial docking, but now we want to create a new ranking with more true binders.

With all the procedure we have done, we now have from 300 to 1000 poses for each compound. For each of these poses we know their PELE BE and their Glide score. In order to create the new ranking, we need to: (i) select one pose per compound and (ii) rank the compounds; both steps can be done according to either their Glide score or their PELE BE for the selected pose.

Pose selection

We have two different energy profiles for each of the 2000 compounds, the one generated by the PELE BE and the one generated by the Glide re-score. The decision now is whether to pick as the pose to study the one with the best PELE BE or the one with the best Glide score. From our previous studies about the re-score process we expect both methodologies will select the same pose in most cases. But we observed that sometimes the methodologies select different poses, like in Figures 4.8h and 4.8g where the pose with the minimum PELE BE and the pose with the minimum Glide score don't match. So, we decided to study how does the pose selection affect each of these metrics (Figure 4.9).

(a) PELE BE distribution



(b) Glide score distribution

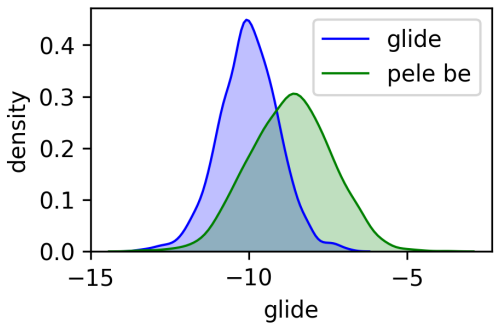


Figure 4.9: Scoring methods distributions. Each image shows the score distribution (PELE BE in (a) and Glide score in (b)) for each possible selection method. The bluish curves represent the score distribution generated by the poses selected according to Glide re-score and the greenish one represent the distribution generated when the pose is selected according to the PELE BE .

We've plotted the two scores distributions for each of the selected poses in 4.9, and even though the distributions overlap partially for the two scores (panels a and b), we see some differences in the distributions. In Figure 4.9a we can observe how the Glide score distribution changes depending on which score has been used to select the pose. We see that the poses selected by Glide have a higher PELE BE (blue curve) and they present a wider spread than the poses selected by the PELE BE (green curve). In Figure 4.9b we observe how the Glide score is affected by the pose selection. We see that the poses selected by the Glide score (bluish colour) present a narrower distribution, with lower Glide scores than the poses selected by the PELE BE (greenish colour).

The reader should be aware that, even though one may think the overlapping area is indicative of the number of compounds for which both selections render the same pose, it isn't. If we look carefully at the size of the overlapping area, we'll see that it changes depending on the score. When we check the number of compounds with the same pose for both selections, we discover that both selection methods (min glide and min PELE BE) select the same pose only for 10% of the compounds, while the overlapping areas on 4.9 seem to cover almost 50% of the compounds.

In Figure 4.10 we can observe the PELE BEs energy profile (panel (a)) and the Glide score profile (panel (b)) for the compound comp_1092, one of the many compounds that present two different best poses, depending on which score we use to do the selection of the pose.

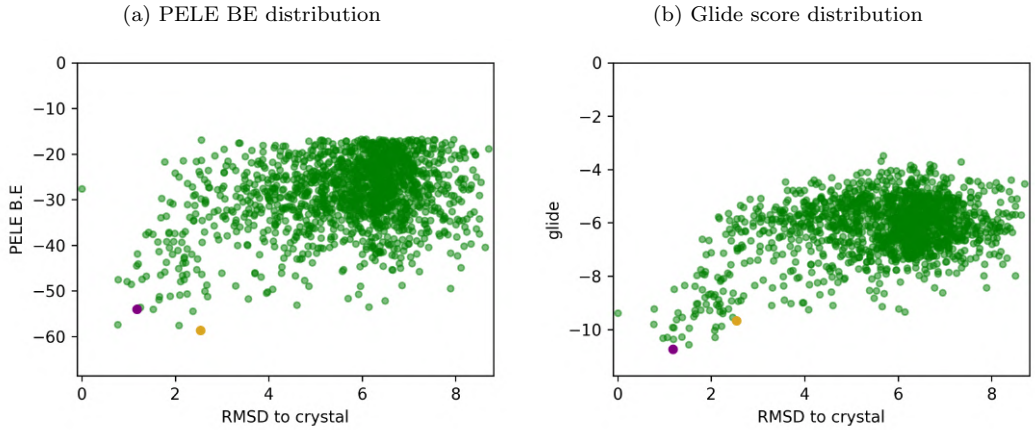


Figure 4.10: The image shows the PELE BE profile in (a) and the Glide profile in (b). These profiles differ from the previous ones due to the clustering applied to them. The purple dot indicates the pose with the minimum Glide score and the golden dot indicates the pose with the minimum PELE BE

Ranking changes

Given our two ranking scores (Glide and PELE BE), and the differences in their distribution caused by which one we use to select the pose to analyse (the one with the minimum score), for

each compound we obtain four new rankings. Figure 4.11 shows the five possible rankings (the four previously mentioned plus the initial one) we have: the *docking ranking*, which derives from our initial docking procedure; the *glide (glide selection)*, ranking derived from ranking by Glide score the poses selected according to their Glide score; the *glide (PELE BE selection)*, ranking derived from ranking by Glide score the poses selected by PELE BE ; the *PELE BE (glide selection)*, ranking derived from ranking the compounds according to their PELE BE using the pose selected by the Glide re-score process; and the *PELE BE (PELE BE selection)* ranking, which derives from ranking the compounds by their PELE BE computed on the best pose according to PELE BE.

Given the selection procedure the *glide (PELE BE selection)* and *PELE BE (PELE BE selection)* rankings score the poses selected by their PELE BE (golden points in Figure 4.10) according to their Glide score and PELE BE respectively generating two different rankings. On the other side, *glide (glide selection)* and *PELE BE (glide selection)* score the best poses selected by their Glide score (purple points in Figure 4.10) using the Glide and PELE BE respectively.

Figure 4.11 represents the changes of each ranking regarding the initial ranking. The x axis is divided in five parts, one for each possible ranking, while the position of the ranking is shown on the y-axis, and the position of the compounds on the initial ranking are represented by the colour scale. In this graphic, we can appreciate that each of the rankings are quite different among them and to the initial docking. It also shows that there's no correlation among the rankings since the colouration of the positions doesn't match from one ranking to another.

The influence of the pose selection and the score used to create the ranking is such that only 13 compounds share the same ranking according to any of these values. Only 6 compounds are not influenced by the pose selection and share the same position in the ranking generated by either the Glide or PELE BE scores, while only 7 compounds share the same ranking position independently of which score is used to generate the ranking.

We should note that, for 20% of the compounds the pose selected by PELE BE and Glide score is the same, but we're seeing that not even 10% of the compounds share the same ranking position, which indicates that not only the scores sample different conformational minima but they also differ greatly upon scoring the same poses.

Next, we'll study in more detail how does the top 2% of each ranking change, that means, we'll look at the top 40 compounds from each ranking, and their change regarding the initial docking. The compounds initial docking position for the best 40 compounds of each ranking are written in Table 4.4, the differences on the numbers give us an idea of how much the ranking has changed after the induced fit procedure. We're plotting the initial ranking position, as identifier, instead of the compounds name, due to confidentiality reasons, which means that if two rankings show the same number the same compound occupies that position in both rankings.

The most curious thing is that some of the compounds occupying the top positions in most of the rankings present a high initial docking position, over 1000, which means that if we only used

the traditional docking methodology we would not have studied those compounds. Upon careful examination, we've seen that some of those compounds are good candidates to become a drug, which indicates a certain degree of enrichment on the ranking.

Table 4.4: Ranking changes depending on the score used to select the pose and the score used to generate the ranking. The cell values indicate the ranking on the docking ranking.

glide (glide selection)	glide (PELE BE selection)	PELE BE (glide selection)	PELE BE (PELE BE selection)
1897	1897	1803	1965
1710	1710	155	1181
555	303	1659	458
303	909	303	320
8	1659	909	205
149	1188	320	1803
95	47	1000	155
1400	1295	205	1659
909	1181	458	555
458	860	1897	303
545	8	95	950
1504	245	971	909
1181	100	1181	106
1659	522	552	757
637	552	142	1897
935	870	1965	911
1188	156	391	1885
70	74	911	8
47	149	234	6
89	395	867	1873

As mentioned before (it can be observed in detail in table 4.4 and Figure 4.11) there's almost no consensus among the rankings, which makes the selection of compounds to experimentally test really hard. In light of all this it was decided that the methodology should be tested in a validation set. This new set will be composed of compounds for which we won't know their binding mode, that is, we don't have a crystal with the compound, but we'll have their activity value, we know they are active and present an effect on the protein.

4.5 Validation Set

In this section we'll study how our methodology affects the ranking of approximately 60 compounds with known activity. The main objective of this study is to set a selection criteria to later on apply onto the VS_set. Again, due to the confidentiality of this project we cannot disclose the names or formulas of the compounds used to this study.

4.5.1 Data preparation and PELE simulations

The first step in the process is to dock the compounds using the Glide program from Schrödinger, using the SP protocol set to generate one single pose for each compound, and form an h-bond with the key hydrogen from the protein. We performed the docking against the receptor_G, and, as happened with the VS_set, the program was unable to dock some compounds with the constraint on; for these compounds we performed an unconstrained docking.

The docking procedure provides us with a Glide score for each compound, and we use this score to generate a ranking of the compounds. We'll call this ranking the initial_glide, and whenever a value of this ranking is plotted or used it will be a reference to the Glide score of the compound, not its position in the ranking.

Then, the protocol 12 was applied to the compounds docked by the constrained docking and protocol 13 to those that underwent the unconstrained docking. The trajectories where then clustered according to the RMSD between the poses and the pose with the lowest PELE BE for each cluster was selected to be re-scored by glide.

From this new profiles two poses where selected for each compound: the best pose according to their PELE BE , min_PELE, and the best one according to their Glide scores, min_GLIDE. When all the combinations possible are taken into account, we end up with four different rankings, the same four types as in the VSs screening. According to which score is used to select the compound, we obtain two possible poses: the minGLIDE and the minPELE, where they are selected by their Glide score and their PELE BE respectively.

The most intuitive rankings are the ones where the selection of the compound and the ranking are made with the same score, where we would use the Glide score to rank the minGLIDE ranking and the PELE BE . When we look at these rankings and the initial_glide one way to see how much accuracy we gain with our methodology is to check their correlation with their experimental activity. A standard metric to measure the correlation between two variables is r^2; we'll use it here to measure the correlation between the experimental value pIC50 and the different scores: initial_glide, glide_minGLIDE and BE_minPELE.

The initial_glide has an r^2 value of 0.09 and, as we can see in Figure 4.12a, this means there's barely any correlation between the variables. When we look at the r^2 of the rankings we observe it has decreased even more, 0.07 for the glide_minGLIDE and 0.05 for the be_minPELE, which means that even though our simulations have been set up to obtain near-native poses to the crystals, the scores we are using are incapable of correctly sorting the compounds according to their real binding activity.

A first approach to solve this issue is to do some consensus between the scores we have, in this direction we have observed that when we try to correlate the PELE BE with the initial_glide score we can observe an improvement on the quality of the selected compounds. In Figure 4.13 we can see that, if we apply a combined criteria on the initial_glide and the PELE BE, we can avoid the

selection of compounds with low activity values.

4.6 Conclusions

The implementation of the technology into a real pharmacological development project has confirmed the conclusion that were obtained in the DUDE section: we need to use a specific simulation protocol tailored for the receptor. After a few calibration steps, we do obtain a protocol with PELE that is capable of sampling the bound species, correcting in many cases the wrong Glide rigid docking pose.

The scoring functions, however, still fail to have a strong correlation with the activity. While we do not know yet the results of the *in vitro* validation of our 2000 compound refinement (and, in fact, we might never know them due to confidentiality issues), the validation test set has raised significant doubts about each specific scoring function. Interestingly, however, applying a consensus ranking procedure seems to provide the best number of high active compounds. This has been the criteria that was finally used by the industrial partner to choose few compounds for lead optimization (as said, results are kept confidential).

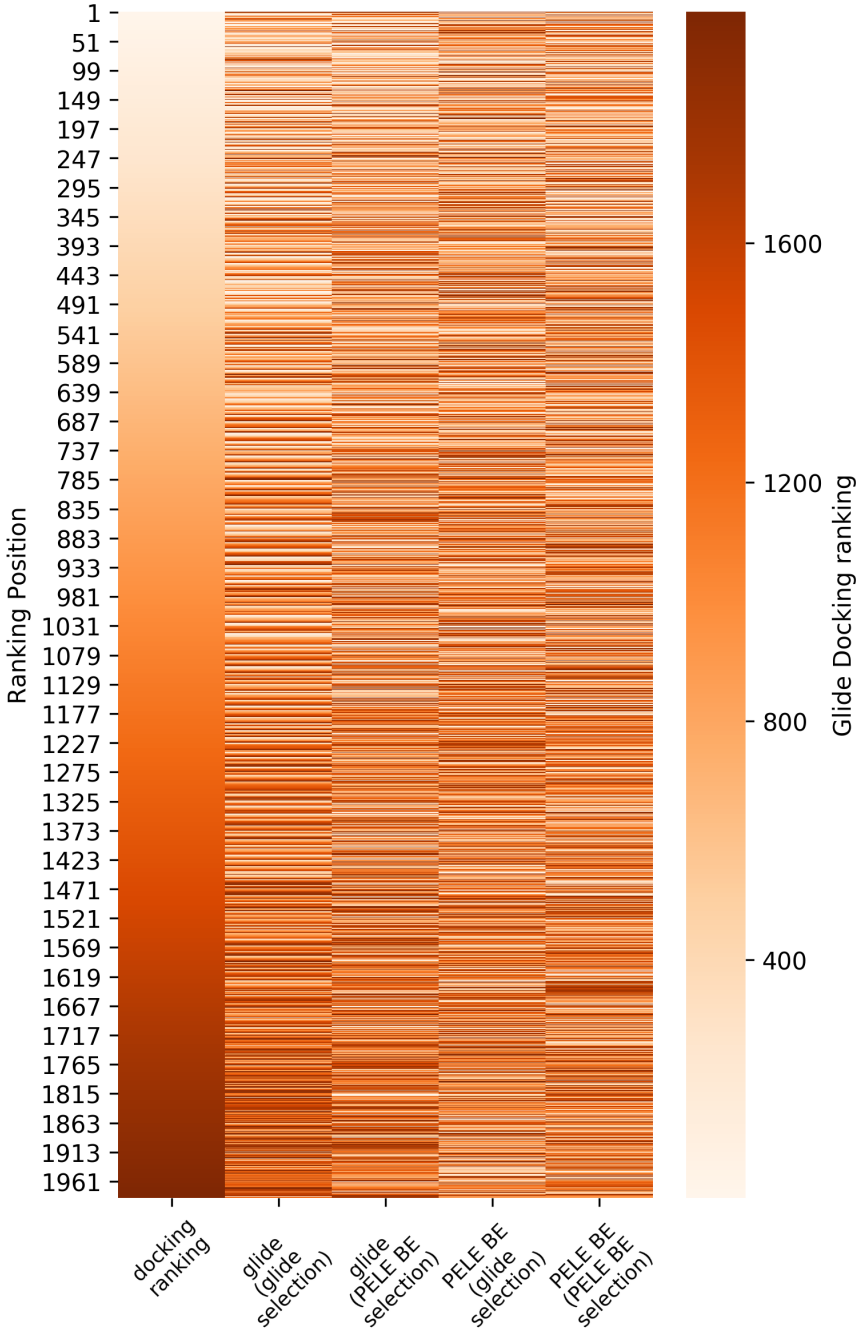
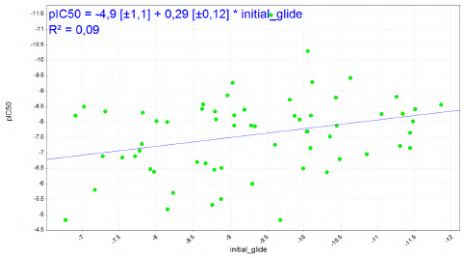
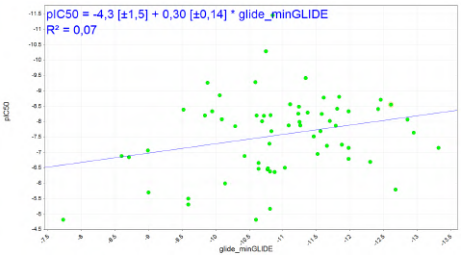


Figure 4.11: This image shows how much does each of the rankings generated after all the procedure change when compared to the initial docking. The x axis divides the space into the five ranking generated, the y-axis shows the position of the ranking and the colour represents the position of the compound in the initial ranking.

(a) pIC50 vs initial_glide



(b) pIC50 vs glide_minGLIDE



(c) pIC50 vs BE_minPELE

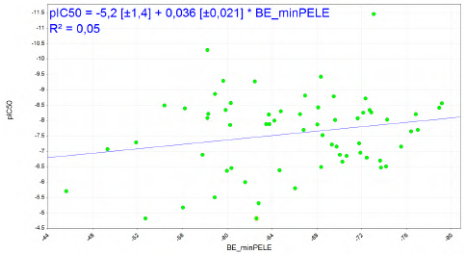


Figure 4.12: All three images show on the y-axis the pIC50 (the experimentally obtained activity value) and the x-axis shows one of the following scoring functions values: initial_glide (a), the Glide score from the best pose according to Glide (b) and the PELE BE from the best pose according to PELE BE (c).

(a) BE minPELE vs initial_glide

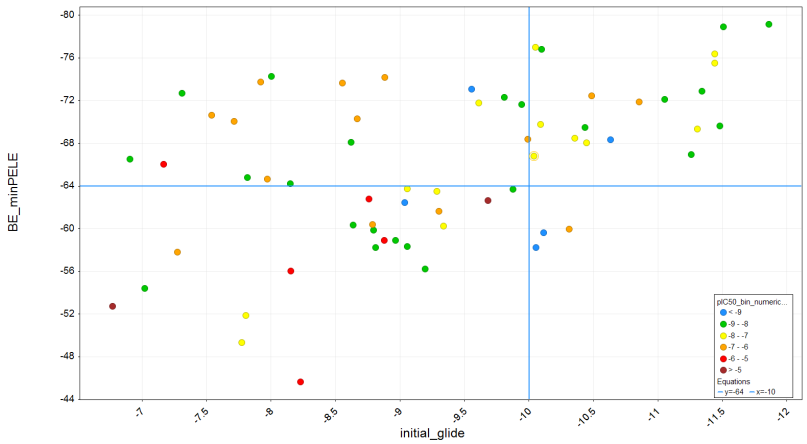


Figure 4.13: This figure shows in the y-axis the PELE BE of the poses selected by their PELE BE and in the x-axis it shows the Glide score from the docking pose of each compound. The colour represents the range of the compounds' activity.

Chapter 5

Conclusions

From the chapter 2 we have obtained the following conclusions:

- A platform that allows the use of PELE within VS campaigns have been developed.
- This platform reduces the time used to prepare, perform, analyse and re-score thousands of simulations from half a year to a couple of months.
- The platform design allows the user to use it not only to perform simulations but to use just re-score simulations.

From the chapter 3 the main conclusions are:

- The reason behind the improvement of EF for only half of the systems is related to the structural characteristics of the pockets.
- Finding a general protocol capable of improving the sampling for all the possible proteins is an extremely difficult task, due to the big differences on the proteins structure from one family to another.
- The many differentiating characteristics of the proteins obliges us to use a protein-specific protocol in order to improve the Glide docking results. With more detailed studies and a better understanding of which proteins have highly similar pockets we may be able to get to a pocket-type-specific protocol.

From the chapter 4 the main conclusions are:

- The implementation of the technology into a real pharmacological development project has confirmed the conclusions that were obtained in the 3: we need to use a specific simulation protocol tailored for the receptor. After a few calibration steps, we do obtain a protocol with PELE that is capable of sampling the bound species, correcting in many cases the wrong Glide rigid docking pose.

- The scoring functions, however, still fail to have a strong correlation with the activity. While we do not know yet the results of the *in vitro* validation of our 2000 compound refinement (and, in fact, we might never know them due to confidentiality issues), the validation test set has raised significant doubts about each specific scoring function. Interestingly, however, applying a consensus ranking procedure seems to provide the best number of high active compounds. This has been the criteria that was finally used by the industrial partner to choose few compounds for lead optimization (as said, results are kept confidential).

As an overall conclusion, we can summarise that the technology to perform an extensive sampling that reproduces the IF effect is ready, and with the proper automatization, by scripting platforms as the one developed in this thesis, it can be applied to VS campaigns as a final step, by refining a large number of compounds. However, the SFs do not seem to be mature enough to largely correlate with experimental biding affinities, thus, providing only qualitative filtering results.

List of Publications

J. Iglesias, S. Saen-oon, R. Soliva, and V. Guallar. Computational structure-based drug design: Predicting target flexibility. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 8(5), 2018.

Bibliography

[1] Giuseppina Andreotti, Israel Cabeza de Vaca, Angelita Poziello, Maria Chiara Monti, Victor Guallar, and Maria Vittoria Cubellis. Conformational response to ligand binding in phosphomannomutase2: insights into inborn glycosylation disorder. *The Journal of biological chemistry*, 289(50):34900–10, 12 2014.

[2] P. R. Andrews, D. J. Craik, and J. L. Martin. Functional Group Contributions to Drug-Receptor Interactions. *Journal of Medicinal Chemistry*, 27(12):1648–1657, 12 1984.

[3] Jurgen Bajorath. Pharmacophore. In *Encyclopedia of Cancer*, pages 2849–2852. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[4] Jay L. Banks, Hege S. Beard, Yixiang Cao, Art E. Cho, Wolfgang Damm, Ramy Farid, Anthony K. Felts, Thomas A. Halgren, Daniel T. Mainz, Jon R. Maple, Robert Murphy, Dean M. Philipp, Matthew P. Repasky, Linda Y. Zhang, Bruce J. Berne, Richard A. Friesner, Emilio Gallicchio, and Ronald M. Levy. Integrated Modeling Program, Applied Chemical Theory (IMPACT). *Journal of Computational Chemistry*, 26(16):1752–1780, 12 2005.

[5] Ewa Bielska, Xavier Lucas, Anna Czerwoniec, Joanna M. Kasprzak, Katarzyna H. Kaminska, and Janusz M. Bujnicki. Virtual screening strategies in drug design - methods and applications. *Biotechnologia*, 92(3):249–264, 2011.

[6] Kenneth W. Borrelli, Benjamin Cossins, and Victor Guallar. Exploring Hierarchical Refinement Techniques for Induced Fit Docking with Protein and Ligand Flexi. *Journal of computational chemistry*, 31(6):1224–35, 4 2010.

[7] Giovanni Bottegoni, Irina Kufareva, Maxim Totrov, and Ruben Abagyan. A new method for ligand docking to flexible receptors by dual alanine scanning and refinement (SCARE). *Journal of Computer-Aided Molecular Design*, 22(5):311–325, 5 2008.

[8] Rodolpho C Braga, Vinicius M Alves, Arthur C Silva, Marilia N Nascimento, Flavia C Silva, Luciano M Liao, and Carolina H Andrade. Virtual screening strategies in medicinal chemistry:

the state of the art and current challenges. *Current topics in medicinal chemistry*, 14(16):1899–912, 2014.

[9] B.R. Brooks, C.L Brooks, A. D Mackerell, L Nilsson, R. J Petrella, B Roux, Y Won, G Archontis, C Bartels, S Boresch, A Caflisch, L Caves, Q Cui, A. R Dinner, M Feig, S Fischer, J Gao, M Hodoscek, W Im, K Kuczera, T Lazaridis, J Ma, V Ovchinnikov, E Paci, R. W Pastor, C. B Post, J. Z Pu, M Schaefer, B Tidor, R. M Venable, H. L Woodcock, X Wu, W Yang, York D.M, and M Karplus. CHARMM: Molecular dynamics simulation package. *Journal of computational chemistry*, 30(10):1545–1614, 2009.

[10] Israel Cabeza De Vaca, Maria Ftima Lucas, and Victor Guallar. New Monte Carlo Based Technique to Study DNA-Ligand Interactions. *Journal of Chemical Theory and Computation*, 11(12):5598–5605, 2015.

[11] Claudio N. Cavasotto and Sharangdhar S. Phatak. Homology modeling in drug discovery: current trends and applications. *Drug Discovery Today*, 14(13-14):676–683, 7 2009.

[12] Nuno M.F.S.A. Cerqueira, Diana Gesto, Eduardo F. Oliveira, Diogo Santos-Martins, Natrcia F. Brás, Srgio F. Sousa, Pedro A. Fernandes, and Maria J. Ramos. Receptor-based virtual screening protocol for drug discovery. *Archives of Biochemistry and Biophysics*, 582:56–67, 9 2015.

[13] C. Da and D. Kireev. Structural protein-ligand interaction fingerprints (SPLIF) for structure-based virtual screening: Method and benchmark study. *Journal of Chemical Information and Modeling*, 54(9):2555–2561, 9 2014.

[14] J. A. Dimasi, L Feldman, A Seckler, and A Wilson. Trends in risks associated with new drug development: Success rates for investigational drugs. *Clinical Pharmacology and Therapeutics*, 87(3):272–277, 3 2010.

[15] Jacob D Durrant and J Andrew McCammon. BINANA: a novel algorithm for ligand-binding characterization. *Journal of molecular graphics & modelling*, 29(6):888–93, 4 2011.

[16] Karl Edman, Ali Hosseini, Magnus K. Bjursell, Anna Aagaard, Lisa Wissler, Anders Gunnarsson, Tim Kaminski, Christian Köhler, Stefan Bäckström, Tina J. Jensen, Anders Cavallin, Ulla Karlsson, Ewa Nilsson, Daniel Lecina, Ryoji Takahashi, Christoph Grebner, Stefan Geschwindner, Matti Lepistö, Anders C. Hogner, and Victor Guallar. Ligand Binding Mechanism in Steroid Receptors: From Conserved Plasticity to Differential Evolutionary Constraints. *Structure*, 23(12):2280–2290, 2015.

[17] M. Espona-Fiedler, V. Soto-Cerrato, A. Hosseini, J. M. Lizcano, V. Guallar, R. Quesada, T. Gao, and R. Pérez-Tomás. Identification of dual mTORC1 and mTORC2 inhibitors in melanoma cells: Prodigiosin vs. obatoclax. *Biochemical Pharmacology*, 83(4):489–496, 2 2012.

[18] Anna Maria Ferrari, Binqing Q Wei, Luca Costantino, and Brian K Shoichet. Soft docking and multiple receptor conformations in virtual screening. *Journal of medicinal chemistry*, 47(21):5076–84, 10 2004.

[19] Thomas Fichert, Mehran Yazdanian, and John R. Proudfoot. A structure-Permeability study of small drug-like molecules. *Bioorganic and Medicinal Chemistry Letters*, 13(4):719–722, 2003.

[20] Emil Fischer. Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der deutschen chemischen Gesellschaft*, 27(3):2985–2993, 10 1894.

[21] Xavier Fradera and Kerim Babaoglu. Overview of Methods and Strategies for Conducting Virtual Small Molecule Screening. *Current protocols in chemical biology*, 9(3):196–212, 1 2017.

[22] H Frauenfelder, F Parak, and R D Young. Conformational Substates in Proteins. *Annu. Rev. Biophys. Biophys. Chem.*, 17(1):451–479, 6 1988.

[23] 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, 3 2004.

[24] Emilio Gallicchio, Linda Yu Zhang, and Ronald M. Levy. The SGB/NP hydration free energy model based on the surface generalized born solvent reaction field and novel nonpolar hydration free energy estimators. *Journal of Computational Chemistry*, 23(5):517–529, 4 2002.

[25] Avijit Ghosh, Chaya Sendrovic Rapp, and Richard A. Friesner. Generalized Born Model Based on a Surface Integral Formulation. *The Journal of Physical Chemistry B*, 102(52):10983–10990, 2002.

[26] Valerio Guido Giacobelli, Emanuele Monza, M. Fatima Lucas, Cinzia Pezzella, Alessandra Piscitelli, Victor Guallar, and Giovanni Sannia. Repurposing designed mutants: a valuable strategy for computer-aided laccase engineering the case of POXA1b. *Catalysis Science & Technology*, 7(2):515–523, 1 2017.

[27] Joan F. Gilabert, Daniel Lecina, Jorge Estrada, and Victor Guallar. Monte Carlo Techniques for Drug Design: The Success Case of PELE. In *Biomolecular Simulations in StructureBased Drug Discovery*, chapter 5, pages 87–103. John Wiley & Sons, Ltd, 12 2018.

[28] Aleix Gimeno, Mara Jos Ojeda-Montes, Sarah Tomás-Hernández, Adri Cereto-Massagué, Ral Beltrán-Debón, Miquel Mulero, Gerard Pujadas, and Santiago Garcia-Vallvé. The light and dark sides of virtual screening: What is there to know? *International Journal of Molecular Sciences*, 20(6):1375, 3 2019.

[29] D Goldgaber, M. Lerman, O. McBride, U Saffiotti, and D. Gajdusek. Characterization and chromosomal localization of a cDNA encoding brain amyloid of Alzheimer's disease. *Science*, 235(4791):877–880, 2 1987.

[30] Christoph Grebner, Daniel Lecina, Victor Gil, Johan Ulander, Pia Hansson, Anita Dellsen, Christian Tyrchan, Karl Edman, Anders Hogner, and Victor Guallar. Exploring Binding Mechanisms in Nuclear Hormone Receptors by Monte Carlo and X-ray-derived Motions. *Biophysical Journal*, 112(6):1147–1156, 2017.

[31] Jason H Haga, Kohei Ichikawa, Susumu Date, Jason H. Haga, Kohei Ichikawa, and Susumu Date. Virtual Screening Techniques and Current Computational Infrastructures. *Current pharmaceutical design*, 22(23):3576–84, 2016.

[32] Thomas A Halgren. Identifying and characterizing binding sites and assessing druggability. *Journal of chemical information and modeling*, 49(2):377–89, 2 2009.

[33] Thomas A. Halgren, Robert B. Murphy, Richard A. Friesner, Hege S. Beard, Leah L. Frye, W. Thomas Pollard, and Jay Banks L. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *Journal of medicinal chemistry*, 47(7):1750–9, 3 2004.

[34] Tom Halgren. New method for fast and accurate binding-site identification and }analysis. *Chem. Biol. Drug Des.*, 69(2):146–148, 2 2007.

[35] Michael Hay, David W Thomas, John L Craighead, Celia Economides, and Jesse Rosenthal. Clinical development success rates for investigational drugs. *Nature Biotechnology*, 32(1):40–51, 1 2014.

[36] M C Hollstein, R A Metcalf, J A Welsh, R Montesano, and C C Harris. Frequent mutation of the p53 gene in human esophageal cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 87(24):9958–61, 12 1990.

[37] Ali Hosseini, Andreu Alibés, Marc Noguera-Julian, Victor Gil, Roger Paredes, Robert Soliva, Modesto Orozco, and Victor Guallar. Computational Prediction of HIV-1 Resistance to Protease Inhibitors. *Journal of Chemical Information and Modeling*, 56(5):915–923, 2016.

[38] Ali Hosseini, Moran Brouk, Maria Fatima Lucas, Fabian Glaser, Ayelet Fishman, and Victor Guallar. Atomic picture of ligand migration in toluene 4-monooxygenase. *Journal of Physical Chemistry B*, 119(3):671–678, 2015.

[39] Ali Hosseini, Margarita Espona-Fiedler, Vanessa Soto-Cerrato, Roberto Quesada, Ricardo Pérez-Tomás, and Victor Guallar. Molecular Interactions of Prodiginines with the BH3 Domain of Anti-Apoptotic Bcl-2 Family Members. *PLoS ONE*, 8(2):e57562, 2 2013.

[40] Sheng-You Huang and Xiaoqin Zou. Ensemble docking of multiple protein structures: Considering protein structural variations in molecular docking. *Proteins: Structure, Function, and Bioinformatics*, 66(2):399–421, 11 2006.

[41] Jelisa Iglesias, Suwipa Saen-oon, Robert Soliva, and Victor Guallar. Computational structure-based drug design: Predicting target flexibility. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 8(5), 2018.

[42] Matthew P. Jacobson, David L. Pincus, Chaya S. Rapp, Tyler J.F. F Day, Barry Honig, David E. Shaw, and Richard A. Friesner. A hierarchical approach to all-atom protein loop prediction. *Proteins*, 55(2):351–67, 5 2004.

[43] William P. Janzen. Screening technologies for small molecule discovery: The state of the art. *Chemistry and Biology*, 21(9):1162–1170, 9 2014.

[44] William L. Jorgensen and Julian Tirado-Rives. Molecular modeling of organic and biomolecular systems using BOSS and MCPRO. *Journal of Computational Chemistry*, 26(16):1689–1700, 12 2005.

[45] Kenneth W. Borrelli, Andreas Vitalis, Raul Alcantara, , Victor Guallar*, Kenneth W. Borrelli, Andreas Vitalis, Raul Alcantara, and Victor Guallar. PELE: Protein Energy Landscape Exploration. A Novel Monte Carlo Based Technique. *Journal of Chemical Theory and Computation*, 1(6):1304–1311, 11 2005.

[46] Douglas B Kitchen, Hlne Decornez, John R Furr, and Jrgen Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews. Drug discovery*, 3(11):935–49, 11 2004.

[47] Jana Kopečná, Israel Cabeza de Vaca, Nathan B P Adams, Paul A Davison, Amanda A Brindley, C Neil Hunter, Victor Guallar, and Roman Sobotka. Porphyrin Binding to Gun4 Protein, Facilitated by a Flexible Loop, Controls Metabolite Flow through the Chlorophyll Biosynthetic Pathway. *The Journal of biological chemistry*, 290(47):28477–88, 11 2015.

[48] Daniel E Koshland. The Key-Lock Theory and the Induced Fit Theory. *Angewandte Chemie International Edition*, 33(510):2375–2378, 1994.

[49] Martin Kotev, Daniel Lecina, Teresa Tarragó, Ernest Giralt, and Vctor Guallar. Unveiling prolyl oligopeptidase ligand migration by comprehensive computational techniques. *Biophysical Journal*, 108(1):116–125, 2015.

[50] Dennis M. Krüger and Andreas Evers. Comparison of Structure- and Ligand-Based Virtual Screening Protocols Considering Hit List Complementarity and Enrichment Factors. *ChemMedChem*, 5(1):148–158, 1 2010.

[51] Ashutosh Kumar and Kam Y.J. Zhang. Hierarchical virtual screening approaches in small molecule drug discovery. *Methods*, 71(C):26–37, 1 2015.

[52] Thierry. Langer and Remy D. Hoffmann. *Pharmacophores and pharmacophore searches*. Wiley-VCH, 2006.

[53] A. Lavecchia and C. Di Giovanni. Virtual screening strategies in drug discovery: A critical review. *Current Medicinal Chemistry*, 20(23):2839–2860, 2013.

[54] Vincent Le Guilloux, Peter Schmidtke, and Pierre Tuffery. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics*, 10:168, 6 2009.

[55] Daniel Lecina, Joan F. Gilabert, and Victor Guallar. Adaptive simulations, towards interactive protein-ligand modeling. *Scientific Reports*, 7(1):1–11, 12 2017.

[56] C a Lipinski, F Lombardo, B W Dominy, and P J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 46(1-3):3–26, 3 2001.

[57] Stephani Joy Y. Macalino, Vijayakumar Gosu, Sunhye Hong, and Sun Choi. Role of computer-aided drug design in modern drug discovery. *Archives of Pharmacal Research*, 38(9):1686–1701, 9 2015.

[58] Armin Madadkar-Sobhani and Victor Guallar. PELE web server: atomistic study of biomolecular systems at your fingertips. *Nucleic Acids Research*, 41(W1):W322–W328, 7 2013.

[59] Gerald Maggiora, Martin Vogt, Dagmar Stumpfe, and Jrgen Bajorath. Molecular similarity in medicinal chemistry. *Journal of Medicinal Chemistry*, 57(8):3186–3204, 2014.

[60] Laura Maintz and Natalija Novak. Histamine and histamine intolerance. *The American Journal of Clinical Nutrition*, 85(5):1185–1196, 5 2007.

[61] A Marchetti, F Buttitta, G Merlo, F Diella, S Pellegrini, S Pepe, P Macchiarini, A Chella, C A Angeletti, R Callahan, Maria Bistocchi, and Francesco Squartini. p53 alterations in non-small cell lung cancers correlate with metastatic involvement of hilar and mediastinal lymph nodes. *Cancer research*, 53(12):2846–51, 6 1993.

[62] Marcelo A Martí, Axel Bidon-Chanal, Alejandro Crespo, Syun-Ru Yeh, Victor Guallar, F Javier Luque, and Daro A Estrin. Mechanism of product release in NO detoxification from Mycobacterium tuberculosis truncated hemoglobin N. *Journal of the American Chemical Society*, 130(5):1688–93, 2008.

[63] Mark McGann. FRED pose prediction and virtual screening accuracy. *Journal of Chemical Information and Modeling*, 51(3):578–596, 2011.

[64] Mark McGann. FRED and HYBRID docking performance on standardized datasets. *Journal of Computer-Aided Molecular Design*, 26(8):897–906, 2012.

[65] Georgia B. McGaughey, Robert P. Sheridan, Christopher I. Bayly, J. Chris Culberson, Constantine Kreatsoulas, Stacey Lindsley, Vladimir Maiorov, Jean Francois Truchon, and Wendy D. Cornell. Comparison of topological, shape, and docking methods in virtual screening. *Journal of Chemical Information and Modeling*, 47(4):1504–1519, 2007.

[66] Michael M. Mysinger, Michael Carchia, John J. Irwin, and Brian K. Shoichet. Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594, 7 2012.

[67] Gerd Neudert and Gerhard Klebe. DSX: a knowledge-based scoring function for the assessment of protein-ligand complexes. *Journal of chemical information and modeling*, 51(10):2731–45, 10 2011.

[68] Romelia Salomon-Ferrer, David A. Case, and Ross C. Walker. An overview of the Amber biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 3(2):198–210, 2013.

[69] G Madhavi Sastry, Matvey Adzhigirey, Tyler Day, Ramakrishna Annabhimoju, and Woody Sherman. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *Journal of computer-aided molecular design*, 27(3):221–34, 3 2013.

[70] Michael Schaefer, Christian Bartels, and Martin Karplus. Solution conformations and thermodynamics of structured peptides: Molecular dynamics simulation with an implicit solvation model. *Journal of Molecular Biology*, 284(3):835–848, 1998.

[71] Yibing Shan, Eric T. Kim, Michael P. Eastwood, Ron O. Dror, Markus A. Seeliger, and David E. Shaw. How does a drug molecule find its target binding site? *Journal of the American Chemical Society*, 133(24):9181–9183, 2011.

[72] Woody Sherman, Tyler Day, Matthew P. Jacobson, Richard A. Friesner, and Ramy Farid. Novel procedure for modeling ligand/receptor induced fit effects. *Journal of Medicinal Chemistry*, 49(2):534–553, 1 2006.

[73] Maxim Totrov and Ruben Abagyan. Flexible ligand docking to multiple receptor conformations: a practical alternative. *Current Opinion in Structural Biology*, 18(2):178–184, 4 2008.

[74] Oleg Trott and Arthur J Olson. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–61, 1 2010.

[75] Daniel F Veber, Stephen R Johnson, Hung-Yuan Cheng, Brian R Smith, Keith W Ward, and Kenneth D Kopple. Molecular properties that influence the oral bioavailability of drug candidates. *Journal of medicinal chemistry*, 45(12):2615–23, 6 2002.

[76] Renxiao Wang, Luhua Lai, and Shaomeng Wang. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J Comput Aided Mol Des*, 16(1):11–26, 2002.

[77] Chi Heem Wong, Kien Wei Siah, and Andrew W. Lo. Estimation of clinical trial success rates and related parameters. *Biostatistics*, 20(2):273–286, 4 2019.

[78] C J Woods, J Michel, M Bodnarchuk, S Genheden, R Bradshaw, G A Ross, C Cave-Ayland, H Bruce-Macdonald, A I Cabedo Martinez, M Samways, and J Graham. ProtoMS 3.4, 2018.

Appendix A

DUD-e Supplementary Information

Table A.1: Diverse Family results for the top50 compounds of each receptor.

Protein	Score	EF			Accuracy		$\%EF_{max}$		Ratio	Threshold
		Docking	I.F.	maximum	Docking	I.F.	Docking	I.F.		
aces	dsx	91.79	81.59	418.16	0.18	0.16	0.22	0.20	0.89	50
	Glide	51.00	61.19	418.16	0.10	0.12	0.12	0.15	1.20	50
	mmgbsa	183.58	132.59	418.16	0.36	0.26	0.44	0.32	0.72	50
	pele	81.59	51.00	418.16	0.16	0.10	0.20	0.12	0.63	50
	vina	40.80	51.00	418.16	0.08	0.10	0.10	0.12	1.25	50
	Xscore-Average	81.59	91.79	418.16	0.16	0.18	0.20	0.22	1.13	50
	Xscore-HMScore	71.39	81.59	418.16	0.14	0.16	0.17	0.20	1.14	50
	Xscore-HPScore	112.19	101.99	418.16	0.22	0.20	0.27	0.24	0.91	50
	Xscore-HSScore	101.99	101.99	418.16	0.20	0.20	0.24	0.24	1.00	50
hs90a	dsx	0.00	9.96	418.16	0.00	0.02	0.00	0.02	∞	50
	Glide	39.82	29.87	418.16	0.08	0.06	0.10	0.07	0.75	50
	mmgbsa	189.17	29.87	418.16	0.38	0.06	0.45	0.07	0.16	50
	pele	159.30	39.82	418.16	0.32	0.08	0.38	0.10	0.25	50
	vina	0.00	0.00	418.16	0.00	0.00	0.00	0.00	n.a.	50
	Xscore-Average	0.00	0.00	418.16	0.00	0.00	0.00	0.00	n.a.	50
	Xscore-HMScore	9.96	9.96	418.16	0.02	0.02	0.02	0.02	1.00	50
	Xscore-HPScore	0.00	9.96	418.16	0.00	0.02	0.00	0.02	∞	50
	Xscore-HSScore	0.00	0.00	418.16	0.00	0.00	0.00	0.00	n.a.	50
nram	dsx	97.87	186.84	418.16	0.22	0.42	0.23	0.45	1.91	50
	Glide	133.46	222.43	418.16	0.30	0.50	0.32	0.53	1.67	50
	mmgbsa	240.22	204.63	418.16	0.54	0.46	0.57	0.49	0.85	50
	pele	284.70	258.01	418.16	0.64	0.58	0.68	0.62	0.91	50
	vina	80.07	62.28	418.16	0.18	0.14	0.19	0.15	0.78	50
	Xscore-Average	106.76	222.43	418.16	0.24	0.50	0.26	0.53	2.08	50
	Xscore-HMScore	71.18	195.73	418.16	0.16	0.44	0.17	0.47	2.75	50
	Xscore-HPScore	80.07	204.63	418.16	0.18	0.46	0.19	0.49	2.56	50
	Xscore-HSScore	133.46	240.22	418.16	0.30	0.54	0.32	0.57	1.80	50

Table A.2: GPCR Family results for the top50 compounds of each receptor.

Protein	Score	EF			Accuracy		$\%EF_{max}$		Ratio	Threshold
		Docking	I.F.	maximum	Docking	I.F.	Docking	I.F.		
adrb1	dsx	80.93	71.94	224.82	0.36	0.32	0.36	0.32	0.89	50
	Glide	76.44	89.93	224.82	0.34	0.40	0.34	0.40	1.18	50
	mmgbsa	22.48	22.48	224.82	0.10	0.10	0.10	0.10	1.00	50
	pele	157.37	67.45	224.82	0.70	0.30	0.70	0.30	0.43	50
	vina	13.49	17.99	224.82	0.06	0.08	0.06	0.08	1.33	50
	Xscore-Average	17.99	44.96	224.82	0.08	0.20	0.08	0.20	2.50	50
	Xscore-HMScore	8.99	53.96	224.82	0.04	0.24	0.04	0.24	6.00	50
	Xscore-HPScore	26.98	49.46	224.82	0.12	0.22	0.12	0.22	1.83	50
	Xscore-HSScore	17.99	31.47	224.82	0.08	0.14	0.08	0.14	1.75	50
adrb2	dsx	108.93	98.39	175.70	0.62	0.56	0.62	0.56	0.90	50
	Glide	73.79	87.85	175.70	0.42	0.50	0.42	0.50	1.19	50
	mmgbsa	24.60	24.60	175.70	0.14	0.14	0.14	0.14	1.00	50
	pele	91.36	105.42	175.70	0.52	0.60	0.52	0.60	1.15	50
	vina	14.06	17.57	175.70	0.08	0.10	0.08	0.10	1.25	50
	Xscore-Average	52.71	77.31	175.70	0.30	0.44	0.30	0.44	1.47	50
	Xscore-HMScore	38.65	63.25	175.70	0.22	0.36	0.22	0.36	1.64	50
	Xscore-HPScore	70.28	84.33	175.70	0.40	0.48	0.40	0.48	1.20	50
	Xscore-HSScore	49.20	49.20	175.70	0.28	0.28	0.28	0.28	1.00	50
drd3	dsx	45.45	109.09	418.16	0.10	0.24	0.11	0.26	2.40	50
	Glide	27.27	63.63	418.16	0.06	0.14	0.07	0.15	2.33	50
	mmgbsa	18.18	36.36	418.16	0.04	0.08	0.04	0.09	2.00	50
	pele	9.09	45.45	418.16	0.02	0.10	0.02	0.11	5.00	50
	vina	36.36	54.54	418.16	0.08	0.12	0.09	0.13	1.50	50
	Xscore-Average	18.18	45.45	418.16	0.04	0.10	0.04	0.11	2.50	50
	Xscore-HMScore	27.27	36.36	418.16	0.06	0.08	0.07	0.09	1.33	50
	Xscore-HPScore	18.18	54.54	418.16	0.04	0.12	0.04	0.13	3.00	50
	Xscore-HSScore	9.09	18.18	418.16	0.02	0.04	0.02	0.04	2.00	50

Table A.3: kinase Family results for the top50 compounds of each receptor.

Protein	Score	EF			Accuracy		$\%EF_{max}$		Ratio	Threshold
		Docking	I.F.	maximum	Docking	I.F.	Docking	I.F.		
cdk2	dsx	28.77	46.04	95.91	0.30	0.48	0.30	0.48	1.60	50
	Glide	67.14	76.73	95.91	0.70	0.80	0.70	0.80	1.14	50
	mmgbsa	47.95	38.36	95.91	0.50	0.40	0.50	0.40	0.80	50
	pele	38.36	32.61	95.91	0.40	0.34	0.40	0.34	0.85	50
	vina	53.71	55.63	95.91	0.56	0.58	0.56	0.58	1.04	50
	Xscore-Average	36.45	40.28	95.91	0.38	0.42	0.38	0.42	1.11	50
	Xscore-HMScore	40.28	47.95	95.91	0.42	0.50	0.42	0.50	1.19	50
	Xscore-HPScore	34.53	46.04	95.91	0.36	0.48	0.36	0.48	1.33	50
	Xscore-HSScore	24.94	24.94	95.91	0.26	0.26	0.26	0.26	1.00	50
jak2	dsx	67.95	88.86	261.35	0.26	0.34	0.26	0.34	1.31	50
	Glide	156.81	135.90	261.35	0.60	0.52	0.60	0.52	0.87	50
	mmgbsa	73.18	62.72	261.35	0.28	0.24	0.28	0.24	0.86	50
	pele	109.77	67.95	261.35	0.42	0.26	0.42	0.26	0.62	50
	vina	62.72	67.95	261.35	0.24	0.26	0.24	0.26	1.08	50
	Xscore-Average	62.72	94.09	261.35	0.24	0.36	0.24	0.36	1.50	50
	Xscore-HMScore	109.77	120.22	261.35	0.42	0.46	0.42	0.46	1.10	50
	Xscore-HPScore	78.41	94.09	261.35	0.30	0.36	0.30	0.36	1.20	50
	Xscore-HSScore	15.68	41.82	261.35	0.06	0.16	0.06	0.16	2.67	50
wee1	dsx	140.77	82.80	207.01	0.68	0.40	0.68	0.40	0.59	50
	Glide	207.01	207.01	207.01	1.00	1.00	1.00	1.00	1.00	50
	mmgbsa	207.01	136.63	207.01	1.00	0.66	1.00	0.66	0.66	50
	pele	194.59	103.50	207.01	0.94	0.50	0.94	0.50	0.53	50
	vina	202.87	149.05	207.01	0.98	0.72	0.98	0.72	0.73	50
	Xscore-Average	144.91	124.21	207.01	0.70	0.60	0.70	0.60	0.86	50
	Xscore-HMScore	149.05	144.91	207.01	0.72	0.70	0.72	0.70	0.97	50
	Xscore-HPScore	186.31	136.63	207.01	0.90	0.66	0.90	0.66	0.73	50
	Xscore-HSScore	33.12	37.26	207.01	0.16	0.18	0.16	0.18	1.13	50

Table A.4: NHRs Family results for the top50 compounds of each receptor. Part I

Protein	Score		EF		Accuracy		$\%EF_{max}$		Ratio	Threshold
	Doc-king	I.F.	maxi-mum	Doc-king	I.F.	Doc-king	I.F.			
andr	dsx	110.79	107.22	178.70	0.62	0.60	0.62	0.60	0.97	50
	Glide	100.07	85.78	178.70	0.56	0.48	0.56	0.48	0.86	50
	mmgbsa	75.05	75.05	178.70	0.42	0.42	0.42	0.42	1.00	50
	pele	57.18	53.61	178.70	0.32	0.30	0.32	0.30	0.94	50
	vina	103.65	117.94	178.70	0.58	0.66	0.58	0.66	1.14	50
	Xscore-Average	71.48	75.05	178.70	0.40	0.42	0.40	0.42	1.05	50
	Xscore-HMScore	53.61	57.18	178.70	0.30	0.32	0.30	0.32	1.07	50
	Xscore-HPScore	78.63	92.92	178.70	0.44	0.52	0.44	0.52	1.18	50
	Xscore-HSScore	75.05	78.63	178.70	0.42	0.44	0.42	0.44	1.05	50
esr1	dsx	44.49	64.50	111.21	0.40	0.58	0.40	0.58	1.45	50
	Glide	108.99	51.16	111.21	0.98	0.46	0.98	0.46	0.47	50
	mmgbsa	111.21	68.95	111.21	1.00	0.62	1.00	0.62	0.62	50
	pele	93.42	51.16	111.21	0.84	0.46	0.84	0.46	0.55	50
	vina	55.61	42.26	111.21	0.50	0.38	0.50	0.38	0.76	50
	Xscore-Average	57.83	73.40	111.21	0.52	0.66	0.52	0.66	1.27	50
	Xscore-HMScore	71.18	77.85	111.21	0.64	0.70	0.64	0.70	1.09	50
	Xscore-HPScore	60.05	75.62	111.21	0.54	0.68	0.54	0.68	1.26	50
	Xscore-HSScore	31.14	44.49	111.21	0.28	0.40	0.28	0.40	1.43	50
gcr	dsx	115.35	36.05	360.48	0.32	0.10	0.32	0.10	0.31	50
	Glide	158.61	86.52	360.48	0.44	0.24	0.44	0.24	0.55	50
	mmgbsa	136.98	43.26	360.48	0.38	0.12	0.38	0.12	0.32	50
	pele	158.61	28.84	360.48	0.44	0.08	0.44	0.08	0.18	50
	vina	122.56	43.26	360.48	0.34	0.12	0.34	0.12	0.35	50
	Xscore-Average	36.05	14.42	360.48	0.10	0.04	0.10	0.04	0.40	50
	Xscore-HMScore	36.05	36.05	360.48	0.10	0.10	0.10	0.10	1.00	50
	Xscore-HPScore	72.10	36.05	360.48	0.20	0.10	0.20	0.10	0.50	50
	Xscore-HSScore	7.21	0.00	360.48	0.02	0.00	0.02	0.00	0.00	50

Table A.5: NHRs Family results for the top50 compounds of each receptor, part II

Protein	Score	EF			Accuracy		$\%EF_{max}$		Ratio	Threshold
	Doc-king	I.F.	maxi-mum	Doc-king	I.F.	Doc-king	I.F.			
mcrin	dsx	225.16	241.25	418.16	0.28	0.30	0.54	0.58	1.07	50
	Glide	225.16	241.25	418.16	0.28	0.30	0.54	0.58	1.07	50
	mmgbsa	144.75	176.91	418.16	0.18	0.22	0.35	0.42	1.22	50
	pele	32.17	112.58	418.16	0.04	0.14	0.08	0.27	3.50	50
	vina	112.58	225.16	418.16	0.14	0.28	0.27	0.54	2.00	50
	Xscore-Average	160.83	209.08	418.16	0.20	0.26	0.38	0.50	1.30	50
	Xscore-HMScore	112.58	144.75	418.16	0.14	0.18	0.27	0.35	1.29	50
	Xscore-HPScore	225.16	273.41	418.16	0.28	0.34	0.54	0.65	1.21	50
	Xscore-HSScore	160.83	176.91	418.16	0.20	0.22	0.38	0.42	1.10	50
mcrout	dsx	124.32	113.02	418.16	0.22	0.20	0.30	0.27	0.91	50
	Glide	214.73	180.83	418.16	0.38	0.32	0.51	0.43	0.84	50
	mmgbsa	124.32	113.02	418.16	0.22	0.20	0.30	0.27	0.91	50
	pele	67.81	79.11	418.16	0.12	0.14	0.16	0.19	1.17	50
	vina	169.52	214.73	418.16	0.30	0.38	0.41	0.51	1.27	50
	Xscore-Average	67.81	101.71	418.16	0.12	0.18	0.16	0.24	1.50	50
	Xscore-HMScore	79.11	79.11	418.16	0.14	0.14	0.19	0.19	1.00	50
	Xscore-HPScore	101.71	101.71	418.16	0.18	0.18	0.24	0.24	1.00	50
	Xscore-HSScore	101.71	90.41	418.16	0.18	0.16	0.24	0.22	0.89	50
ppar	dsx	79.87	65.78	117.46	0.68	0.56	0.68	0.56	0.82	50
	Glide	28.19	51.68	117.46	0.24	0.44	0.24	0.44	1.83	50
	mmgbsa	42.29	37.59	117.46	0.36	0.32	0.36	0.32	0.89	50
	pele	23.49	18.79	117.46	0.20	0.16	0.20	0.16	0.80	50
	vina	23.49	25.84	117.46	0.20	0.22	0.20	0.22	1.10	50
	Xscore-Average	32.89	49.33	117.46	0.28	0.42	0.28	0.42	1.50	50
	Xscore-HMScore	35.24	37.59	117.46	0.30	0.32	0.30	0.32	1.07	50
	Xscore-HPScore	32.89	51.68	117.46	0.28	0.44	0.28	0.44	1.57	50
	Xscore-HSScore	42.29	49.33	117.46	0.36	0.42	0.36	0.42	1.17	50

Table A.6: NHRs Family results for the top50 compounds of each receptor. Part III

Protein	Score	EF			Accuracy		$\%EF_{max}$		Ratio	Threshold
	Doc-king	I.F.	maxi-mum	Doc-king	I.F.	Doc-king	I.F.			
prgr	dsx	23.38	23.38	129.86	0.18	0.18	0.18	0.18	1.00	50
	Glide	98.70	64.93	129.86	0.76	0.50	0.76	0.50	0.66	50
	mmgbsa	49.35	20.78	129.86	0.38	0.16	0.38	0.16	0.42	50
	pele	88.31	25.97	129.86	0.68	0.20	0.68	0.20	0.29	50
	vina	49.35	51.95	129.86	0.38	0.40	0.38	0.40	1.05	50
	Xscore-Average	20.78	20.78	129.86	0.16	0.16	0.16	0.16	1.00	50
	Xscore-HMScore	23.38	25.97	129.86	0.18	0.20	0.18	0.20	1.11	50
	Xscore-HPScore	12.99	18.18	129.86	0.10	0.14	0.10	0.14	1.40	50
	Xscore-HSScore	15.58	15.58	129.86	0.12	0.12	0.12	0.12	1.00	50
rxra	dsx	148.77	116.60	201.04	0.74	0.58	0.74	0.58	0.78	50
	Glide	172.89	96.50	201.04	0.86	0.48	0.86	0.48	0.56	50
	mmgbsa	193.00	164.85	201.04	0.96	0.82	0.96	0.82	0.85	50
	pele	112.58	124.64	201.04	0.56	0.62	0.56	0.62	1.11	50
	vina	168.87	60.31	201.04	0.84	0.30	0.84	0.30	0.36	50
	Xscore-Average	136.71	144.75	201.04	0.68	0.72	0.68	0.72	1.06	50
	Xscore-HMScore	120.62	132.69	201.04	0.60	0.66	0.60	0.66	1.10	50
	Xscore-HPScore	124.64	136.71	201.04	0.62	0.68	0.62	0.68	1.10	50
	Xscore-HSScore	156.81	156.81	201.04	0.78	0.78	0.78	0.78	1.00	50
thb	dsx	26.14	45.74	326.69	0.08	0.14	0.08	0.14	1.75	50
	Glide	137.21	98.01	326.69	0.42	0.30	0.42	0.30	0.71	50
	mmgbsa	98.01	71.87	326.69	0.30	0.22	0.30	0.22	0.73	50
	pele	58.80	71.87	326.69	0.18	0.22	0.18	0.22	1.22	50
	vina	91.47	58.80	326.69	0.28	0.18	0.28	0.18	0.64	50
	Xscore-Average	39.20	65.34	326.69	0.12	0.20	0.12	0.20	1.67	50
	Xscore-HMScore	65.34	71.87	326.69	0.20	0.22	0.20	0.22	1.10	50
	Xscore-HPScore	39.20	78.41	326.69	0.12	0.24	0.12	0.24	2.00	50
	Xscore-HSScore	45.74	39.20	326.69	0.14	0.12	0.14	0.12	0.86	50

Table A.7: Protease Family results for the top50 compounds of each receptor.

Protein	Score	EF			Accuracy		$\%EF_{max}$		Ratio	Threshold
		Docking	I.F.	maximum	Docking	I.F.	Docking	I.F.		
hivw	dsx	138.21	152.38	177.19	0.78	0.86	0.78	0.86	1.10	50
	Glide	148.84	138.21	177.19	0.84	0.78	0.84	0.78	0.93	50
	mmgbsa	145.29	163.01	177.19	0.82	0.92	0.82	0.92	1.12	50
	pele	63.79	148.84	177.19	0.36	0.84	0.36	0.84	2.33	50
	vina	24.81	31.89	177.19	0.14	0.18	0.14	0.18	1.29	50
	Xscore-Average	124.03	116.94	177.19	0.70	0.66	0.70	0.66	0.94	50
	Xscore-HMScore	127.57	127.57	177.19	0.72	0.72	0.72	0.72	1.00	50
	Xscore-HPScore	124.03	127.57	177.19	0.70	0.72	0.70	0.72	1.03	50
	Xscore-HSScore	120.49	113.40	177.19	0.68	0.64	0.68	0.64	0.94	50
thrb	dsx	73.12	70.84	114.25	0.64	0.62	0.64	0.62	0.97	50
	Glide	82.26	82.26	114.25	0.72	0.72	0.72	0.72	1.00	50
	mmgbsa	45.70	52.56	114.25	0.40	0.46	0.40	0.46	1.15	50
	pele	114.25	36.56	114.25	1.00	0.32	1.00	0.32	0.32	50
	vina	18.28	34.28	114.25	0.16	0.30	0.16	0.30	1.88	50
	Xscore-Average	45.70	52.56	114.25	0.40	0.46	0.40	0.46	1.15	50
	Xscore-HMScore	57.13	50.27	114.25	0.50	0.44	0.50	0.44	0.88	50
	Xscore-HPScore	52.56	54.84	114.25	0.46	0.48	0.46	0.48	1.04	50
	Xscore-HSScore	31.99	43.42	114.25	0.28	0.38	0.28	0.38	1.36	50
try1	dsx	30.26	57.77	137.55	0.22	0.42	0.22	0.42	1.91	50
	Glide	49.52	99.04	137.55	0.36	0.72	0.36	0.72	2.00	50
	mmgbsa	24.76	44.02	137.55	0.18	0.32	0.18	0.32	1.78	50
	pele	132.05	115.54	137.55	0.96	0.84	0.96	0.84	0.88	50
	vina	24.76	30.26	137.55	0.18	0.22	0.18	0.22	1.22	50
	Xscore-Average	22.01	57.77	137.55	0.16	0.42	0.16	0.42	2.63	50
	Xscore-HMScore	44.02	77.03	137.55	0.32	0.56	0.32	0.56	1.75	50
	Xscore-HPScore	22.01	49.52	137.55	0.16	0.36	0.16	0.36	2.25	50
	Xscore-HSScore	11.00	38.51	137.55	0.08	0.28	0.08	0.28	3.50	50

Table A.8: Diverse Family results for the top100 compounds of each receptor.

Protein	Score	EF			Accuracy		$\%EF_{max}$		Ratio	Threshold
		Docking	I.F.	maximum	Docking	I.F.	Docking	I.F.		
aces	dsx	76.49	71.39	209.08	0.15	0.14	0.37	0.34	0.93	100
	Glide	45.90	61.19	209.08	0.09	0.12	0.22	0.29	1.33	100
	mmgbsa	101.99	112.19	209.08	0.20	0.22	0.49	0.54	1.10	100
	pele	56.09	45.90	209.08	0.11	0.09	0.27	0.22	0.82	100
	vina	35.70	45.90	209.08	0.07	0.09	0.17	0.22	1.29	100
	Xscore-Average	51.00	61.19	209.08	0.10	0.12	0.24	0.29	1.20	100
	Xscore-HMScore	45.90	56.09	209.08	0.09	0.11	0.22	0.27	1.22	100
	Xscore-HPScore	61.19	81.59	209.08	0.12	0.16	0.29	0.39	1.33	100
	Xscore-HSScore	56.09	61.19	209.08	0.11	0.12	0.27	0.29	1.09	100
hs90a	dsx	0.00	14.93	209.08	0.00	0.03	0.00	0.07	∞	100
	Glide	39.82	19.91	209.08	0.08	0.04	0.19	0.10	0.50	100
	mmgbsa	109.52	34.85	209.08	0.22	0.07	0.52	0.17	0.32	100
	pele	94.58	24.89	209.08	0.19	0.05	0.45	0.12	0.26	100
	vina	4.98	0.00	209.08	0.01	0.00	0.02	0.00	0.00	100
	Xscore-Average	0.00	4.98	209.08	0.00	0.01	0.00	0.02	∞	100
	Xscore-HMScore	9.96	9.96	209.08	0.02	0.02	0.05	0.05	1.00	100
	Xscore-HPScore	0.00	4.98	209.08	0.00	0.01	0.00	0.02	∞	100
	Xscore-HSScore	0.00	0.00	209.08	0.00	0.00	0.00	0.00		100
nram	dsx	66.73	137.90	209.08	0.15	0.31	0.32	0.66	2.07	100
	Glide	97.87	160.15	209.08	0.22	0.36	0.47	0.77	1.64	100
	mmgbsa	169.04	169.04	209.08	0.38	0.38	0.81	0.81	1.00	100
	pele	164.59	177.94	209.08	0.37	0.40	0.79	0.85	1.08	100
	vina	75.62	62.28	209.08	0.17	0.14	0.36	0.30	0.82	100
	Xscore-Average	80.07	151.25	209.08	0.18	0.34	0.38	0.72	1.89	100
	Xscore-HMScore	71.18	137.90	209.08	0.16	0.31	0.34	0.66	1.94	100
	Xscore-HPScore	71.18	137.90	209.08	0.16	0.31	0.34	0.66	1.94	100
	Xscore-HSScore	106.76	155.70	209.08	0.24	0.35	0.51	0.74	1.46	100

Table A.9: GPCR Family results for the top100 compounds of each receptor.

Protein	Score	EF			Accuracy		$\%EF_{max}$		Ratio	Threshold
		Docking	I.F.	maximum	Docking	I.F.	Docking	I.F.		
adrb1	dsx	53.96	42.72	209.08	0.24	0.19	0.26	0.20	0.79	100
	Glide	67.45	69.69	209.08	0.30	0.31	0.32	0.33	1.03	100
	mmgbsa	24.73	29.23	209.08	0.11	0.13	0.12	0.14	1.18	100
	pele	103.42	56.20	209.08	0.46	0.25	0.49	0.27	0.54	100
	vina	15.74	15.74	209.08	0.07	0.07	0.08	0.08	1.00	100
	Xscore-Average	20.23	31.47	209.08	0.09	0.14	0.10	0.15	1.56	100
	Xscore-HMScore	22.48	35.97	209.08	0.10	0.16	0.11	0.17	1.60	100
	Xscore-HPScore	31.47	33.72	209.08	0.14	0.15	0.15	0.16	1.07	100
	Xscore-HSScore	17.99	31.47	209.08	0.08	0.14	0.09	0.15	1.75	100
adrb2	dsx	75.55	77.31	175.70	0.43	0.44	0.43	0.44	1.02	100
	Glide	59.74	75.55	175.70	0.34	0.43	0.34	0.43	1.26	100
	mmgbsa	26.35	38.65	175.70	0.15	0.22	0.15	0.22	1.47	100
	pele	68.52	77.31	175.70	0.39	0.44	0.39	0.44	1.13	100
	vina	12.30	22.84	175.70	0.07	0.13	0.07	0.13	1.86	100
	Xscore-Average	43.92	54.47	175.70	0.25	0.31	0.25	0.31	1.24	100
	Xscore-HMScore	42.17	50.95	175.70	0.24	0.29	0.24	0.29	1.21	100
	Xscore-HPScore	50.95	66.77	175.70	0.29	0.38	0.29	0.38	1.31	100
	Xscore-HSScore	38.65	47.44	175.70	0.22	0.27	0.22	0.27	1.23	100
drd3	dsx	45.45	81.81	209.08	0.10	0.18	0.22	0.39	1.80	100
	Glide	18.18	63.63	209.08	0.04	0.14	0.09	0.30	3.50	100
	mmgbsa	18.18	45.45	209.08	0.04	0.10	0.09	0.22	2.50	100
	pele	13.64	36.36	209.08	0.03	0.08	0.07	0.17	2.67	100
	vina	27.27	36.36	209.08	0.06	0.08	0.13	0.17	1.33	100
	Xscore-Average	31.82	50.00	209.08	0.07	0.11	0.15	0.24	1.57	100
	Xscore-HMScore	31.82	40.91	209.08	0.07	0.09	0.15	0.20	1.29	100
	Xscore-HPScore	31.82	63.63	209.08	0.07	0.14	0.15	0.30	2.00	100
	Xscore-HSScore	13.64	45.45	209.08	0.03	0.10	0.07	0.22	3.33	100

Table A.10: Kinase Family results for the top100 compounds of each receptor.

Protein	Score	EF			Accuracy		$\%EF_{max}$		Ratio	Threshold
		Docking	I.F.	maximum	Docking	I.F.	Docking	I.F.		
cdk2	dsx	27.81	35.49	95.91	0.29	0.37	0.29	0.37	1.28	100
	Glide	67.14	69.05	95.91	0.70	0.72	0.70	0.72	1.03	100
	mmgbsa	35.49	36.45	95.91	0.37	0.38	0.37	0.38	1.03	100
	pele	43.16	34.53	95.91	0.45	0.36	0.45	0.36	0.80	100
	vina	46.04	46.04	95.91	0.48	0.48	0.48	0.48	1.00	100
	Xscore-Average	26.85	38.36	95.91	0.28	0.40	0.28	0.40	1.43	100
	Xscore-HMScore	36.45	44.12	95.91	0.38	0.46	0.38	0.46	1.21	100
	Xscore-HPScore	30.69	41.24	95.91	0.32	0.43	0.32	0.43	1.34	100
	Xscore-HSScore	23.02	32.61	95.91	0.24	0.34	0.24	0.34	1.42	100
jak2	dsx	52.27	67.95	209.08	0.20	0.26	0.25	0.33	1.30	100
	Glide	94.09	91.47	209.08	0.36	0.35	0.45	0.44	0.97	100
	mmgbsa	49.66	57.50	209.08	0.19	0.22	0.24	0.28	1.16	100
	pele	99.31	70.56	209.08	0.38	0.27	0.48	0.34	0.71	100
	vina	57.50	60.11	209.08	0.22	0.23	0.28	0.29	1.05	100
	Xscore-Average	54.88	67.95	209.08	0.21	0.26	0.26	0.33	1.24	100
	Xscore-HMScore	70.56	62.72	209.08	0.27	0.24	0.34	0.30	0.89	100
	Xscore-HPScore	60.11	62.72	209.08	0.23	0.24	0.29	0.30	1.04	100
	Xscore-HSScore	18.29	41.82	209.08	0.07	0.16	0.09	0.20	2.29	100
wee1	dsx	120.07	86.94	207.01	0.58	0.42	0.58	0.42	0.72	100
	Glide	204.94	163.54	207.01	0.99	0.79	0.99	0.79	0.80	100
	mmgbsa	159.40	118.00	207.01	0.77	0.57	0.77	0.57	0.74	100
	pele	159.40	76.59	207.01	0.77	0.37	0.77	0.37	0.48	100
	vina	159.40	109.72	207.01	0.77	0.53	0.77	0.53	0.69	100
	Xscore-Average	124.21	109.72	207.01	0.60	0.53	0.60	0.53	0.88	100
	Xscore-HMScore	136.63	120.07	207.01	0.66	0.58	0.66	0.58	0.88	100
	Xscore-HPScore	142.84	120.07	207.01	0.69	0.58	0.69	0.58	0.84	100
	Xscore-HSScore	45.54	49.68	207.01	0.22	0.24	0.22	0.24	1.09	100

Table A.11: NHRs Family results for the top100 compounds of each receptor. Part I

Protein	Score	EF			Accuracy		$\%EF_{max}$		Ratio	Threshold
		Docking	I.F.	maximum	Docking	I.F.	Docking	I.F.		
andr	dsx	85.78	87.56	178.70	0.48	0.49	0.48	0.49	1.02	100
	Glide	83.99	76.84	178.70	0.47	0.43	0.47	0.43	0.91	100
	mmgbsa	53.61	55.40	178.70	0.30	0.31	0.30	0.31	1.03	100
	pele	39.31	57.18	178.70	0.22	0.32	0.22	0.32	1.45	100
	vina	75.05	80.42	178.70	0.42	0.45	0.42	0.45	1.07	100
	Xscore-Average	60.76	57.18	178.70	0.34	0.32	0.34	0.32	0.94	100
	Xscore-HMScore	53.61	53.61	178.70	0.30	0.30	0.30	0.30	1.00	100
	Xscore-HPScore	64.33	60.76	178.70	0.36	0.34	0.36	0.34	0.94	100
	Xscore-HSScore	64.33	60.76	178.70	0.36	0.34	0.36	0.34	0.94	100
esr1	dsx	48.93	57.83	111.21	0.44	0.52	0.44	0.52	1.18	100
	Glide	101.20	60.05	111.21	0.91	0.54	0.91	0.54	0.59	100
	mmgbsa	91.19	57.83	111.21	0.82	0.52	0.82	0.52	0.63	100
	pele	77.85	53.38	111.21	0.70	0.48	0.70	0.48	0.69	100
	vina	52.27	46.71	111.21	0.47	0.42	0.47	0.42	0.89	100
	Xscore-Average	52.27	62.28	111.21	0.47	0.56	0.47	0.56	1.19	100
	Xscore-HMScore	60.05	64.50	111.21	0.54	0.58	0.54	0.58	1.07	100
	Xscore-HPScore	53.38	67.84	111.21	0.48	0.61	0.48	0.61	1.27	100
	Xscore-HSScore	28.92	43.37	111.21	0.26	0.39	0.26	0.39	1.50	100
gcr	dsx	82.91	50.47	209.08	0.23	0.14	0.40	0.24	0.61	100
	Glide	104.54	57.68	209.08	0.29	0.16	0.50	0.28	0.55	100
	mmgbsa	75.70	43.26	209.08	0.21	0.12	0.36	0.21	0.57	100
	pele	86.52	32.44	209.08	0.24	0.09	0.41	0.16	0.38	100
	vina	86.52	43.26	209.08	0.24	0.12	0.41	0.21	0.50	100
	Xscore-Average	36.05	28.84	209.08	0.10	0.08	0.17	0.14	0.80	100
	Xscore-HMScore	21.63	21.63	209.08	0.06	0.06	0.10	0.10	1.00	100
	Xscore-HPScore	61.28	46.86	209.08	0.17	0.13	0.29	0.22	0.76	100
	Xscore-HSScore	14.42	14.42	209.08	0.04	0.04	0.07	0.07	1.00	100

Table A.12: NHRs Family results for the top100 compounds of each receptor. Part II

Protein	Score	EF			Accuracy		$\%EF_{max}$		Ratio	Threshold
		Docking	I.F.	maximum	Docking	I.F.	Docking	I.F.		
mcrin	dsx	152.79	160.83	209.08	0.19	0.20	0.73	0.77	1.05	100
	Glide	120.62	152.79	209.08	0.15	0.19	0.58	0.73	1.27	100
	mmgbsa	72.37	160.83	209.08	0.09	0.20	0.35	0.77	2.22	100
	pele	24.12	128.66	209.08	0.03	0.16	0.12	0.62	5.33	100
	vina	80.42	144.75	209.08	0.10	0.18	0.38	0.69	1.80	100
	Xscore-Average	144.75	136.71	209.08	0.18	0.17	0.69	0.65	0.94	100
	Xscore-HMScore	96.50	104.54	209.08	0.12	0.13	0.46	0.50	1.08	100
	Xscore-HPScore	144.75	152.79	209.08	0.18	0.19	0.69	0.73	1.06	100
	Xscore-HSScore	112.58	128.66	209.08	0.14	0.16	0.54	0.62	1.14	100
mcrout	dsx	101.71	90.41	209.08	0.18	0.16	0.49	0.43	0.89	100
	Glide	129.97	124.32	209.08	0.23	0.22	0.62	0.59	0.96	100
	mmgbsa	73.46	84.76	209.08	0.13	0.15	0.35	0.41	1.15	100
	pele	39.56	67.81	209.08	0.07	0.12	0.19	0.32	1.71	100
	vina	96.06	118.67	209.08	0.17	0.21	0.46	0.57	1.24	100
	Xscore-Average	73.46	90.41	209.08	0.13	0.16	0.35	0.43	1.23	100
	Xscore-HMScore	67.81	79.11	209.08	0.12	0.14	0.32	0.38	1.17	100
	Xscore-HPScore	56.51	90.41	209.08	0.10	0.16	0.27	0.43	1.60	100
	Xscore-HSScore	62.16	90.41	209.08	0.11	0.16	0.30	0.43	1.45	100
ppar	dsx	69.30	52.86	117.46	0.59	0.45	0.59	0.45	0.76	100
	Glide	39.94	48.16	117.46	0.34	0.41	0.34	0.41	1.21	100
	mmgbsa	38.76	35.24	117.46	0.33	0.30	0.33	0.30	0.91	100
	pele	34.06	18.79	117.46	0.29	0.16	0.29	0.16	0.55	100
	vina	30.54	25.84	117.46	0.26	0.22	0.26	0.22	0.85	100
	Xscore-Average	34.06	37.59	117.46	0.29	0.32	0.29	0.32	1.10	100
	Xscore-HMScore	29.37	29.37	117.46	0.25	0.25	0.25	0.25	1.00	100
	Xscore-HPScore	41.11	38.76	117.46	0.35	0.33	0.35	0.33	0.94	100
	Xscore-HSScore	41.11	38.76	117.46	0.35	0.33	0.35	0.33	0.94	100

Table A.13: NHRs Family results for the top100 compounds of each receptor. Part III

Protein	Score	EF			Accuracy		$\%EF_{max}$		Ratio	Threshold
		Docking	I.F.	maximum	Docking	I.F.	Docking	I.F.		
prgr	dsx	22.08	25.97	129.86	0.17	0.20	0.17	0.20	1.18	100
	Glide	76.62	50.65	129.86	0.59	0.39	0.59	0.39	0.66	100
	mmgbsa	36.36	20.78	129.86	0.28	0.16	0.28	0.16	0.57	100
	pele	68.83	28.57	129.86	0.53	0.22	0.53	0.22	0.42	100
	vina	41.56	48.05	129.86	0.32	0.37	0.32	0.37	1.16	100
	Xscore-Average	18.18	18.18	129.86	0.14	0.14	0.14	0.14	1.00	100
	Xscore-HMScore	23.38	23.38	129.86	0.18	0.18	0.18	0.18	1.00	100
	Xscore-HPScore	12.99	18.18	129.86	0.10	0.14	0.10	0.14	1.40	100
	Xscore-HSScore	15.58	14.28	129.86	0.12	0.11	0.12	0.11	0.92	100
rxra	dsx	124.64	104.54	201.04	0.62	0.52	0.62	0.52	0.84	100
	Glide	152.79	108.56	201.04	0.76	0.54	0.76	0.54	0.71	100
	mmgbsa	148.77	134.70	201.04	0.74	0.67	0.74	0.67	0.91	100
	pele	88.46	104.54	201.04	0.44	0.52	0.44	0.52	1.18	100
	vina	122.63	60.31	201.04	0.61	0.30	0.61	0.30	0.49	100
	Xscore-Average	112.58	122.63	201.04	0.56	0.61	0.56	0.61	1.09	100
	Xscore-HMScore	96.50	114.59	201.04	0.48	0.57	0.48	0.57	1.19	100
	Xscore-HPScore	104.54	110.57	201.04	0.52	0.55	0.52	0.55	1.06	100
	Xscore-HSScore	120.62	132.69	201.04	0.60	0.66	0.60	0.66	1.10	100
thb	dsx	35.94	39.20	209.08	0.11	0.12	0.17	0.19	1.09	100
	Glide	88.21	68.60	209.08	0.27	0.21	0.42	0.33	0.78	100
	mmgbsa	65.34	49.00	209.08	0.20	0.15	0.31	0.23	0.75	100
	pele	45.74	49.00	209.08	0.14	0.15	0.22	0.23	1.07	100
	vina	68.60	49.00	209.08	0.21	0.15	0.33	0.23	0.71	100
	Xscore-Average	58.80	62.07	209.08	0.18	0.19	0.28	0.30	1.06	100
	Xscore-HMScore	68.60	75.14	209.08	0.21	0.23	0.33	0.36	1.10	100
	Xscore-HPScore	58.80	65.34	209.08	0.18	0.20	0.28	0.31	1.11	100
	Xscore-HSScore	45.74	45.74	209.08	0.14	0.14	0.22	0.22	1.00	100

Table A.14: Protease Family results for the top100 compounds of each receptor.

Protein	Score	EF			Accuracy		$\%EF_{max}$		Ratio	Threshold
		Docking	I.F.	maximum	Docking	I.F.	Docking	I.F.		
hivw	dsx	106.31	116.94	177.19	0.60	0.66	0.60	0.66	1.10	100
	Glide	88.59	115.17	177.19	0.50	0.65	0.50	0.65	1.30	100
	mmgbsa	99.22	132.89	177.19	0.56	0.75	0.56	0.75	1.34	100
	pele	42.52	113.40	177.19	0.24	0.64	0.24	0.64	2.67	100
	vina	21.26	42.52	177.19	0.12	0.24	0.12	0.24	2.00	100
	Xscore-Average	118.71	108.08	177.19	0.67	0.61	0.67	0.61	0.91	100
	Xscore-HMScore	106.31	109.86	177.19	0.60	0.62	0.60	0.62	1.03	100
	Xscore-HPScore	111.63	104.54	177.19	0.63	0.59	0.63	0.59	0.94	100
	Xscore-HSScore	113.40	102.77	177.19	0.64	0.58	0.64	0.58	0.91	100
thrb	dsx	77.69	59.41	114.25	0.68	0.52	0.68	0.52	0.76	100
	Glide	65.12	76.55	114.25	0.57	0.67	0.57	0.67	1.18	100
	mmgbsa	38.85	52.56	114.25	0.34	0.46	0.34	0.46	1.35	100
	pele	102.83	42.27	114.25	0.90	0.37	0.90	0.37	0.41	100
	vina	20.57	35.42	114.25	0.18	0.31	0.18	0.31	1.72	100
	Xscore-Average	42.27	55.98	114.25	0.37	0.49	0.37	0.49	1.32	100
	Xscore-HMScore	51.41	52.56	114.25	0.45	0.46	0.45	0.46	1.02	100
	Xscore-HPScore	42.27	52.56	114.25	0.37	0.46	0.37	0.46	1.24	100
	Xscore-HSScore	31.99	44.56	114.25	0.28	0.39	0.28	0.39	1.39	100
try1	dsx	26.14	53.65	137.55	0.19	0.39	0.19	0.39	2.05	100
	Glide	53.65	89.41	137.55	0.39	0.65	0.39	0.65	1.67	100
	mmgbsa	28.89	48.14	137.55	0.21	0.35	0.21	0.35	1.67	100
	pele	119.67	110.04	137.55	0.87	0.80	0.87	0.80	0.92	100
	vina	24.76	28.89	137.55	0.18	0.21	0.18	0.21	1.17	100
	Xscore-Average	24.76	55.02	137.55	0.18	0.40	0.18	0.40	2.22	100
	Xscore-HMScore	39.89	72.90	137.55	0.29	0.53	0.29	0.53	1.83	100
	Xscore-HPScore	28.89	46.77	137.55	0.21	0.34	0.21	0.34	1.62	100
	Xscore-HSScore	16.51	39.89	137.55	0.12	0.29	0.12	0.29	2.42	100

Table A.15: $\%EF$ changes upon simulation with protocol 1

Family	Protein	EF			Accuracy		$\%EF_{max}$		Ratio	Threshold
		Docking	I.F.	maximum	Docking	I.F.	Docking	I.F.		
KINASE	cdk2	15.96	19.15	22.80	0.70	0.70	0.84	1.20	0.84	50
	jak2	36.93	25.85	61.55	0.60	0.42	0.60	0.42	0.70	50
NHR	gcr	37.35	15.28	84.90	0.44	0.18	0.44	0.18	0.41	50
	mcrin	53.03	45.45	98.48	0.28	0.24	0.54	0.46	0.86	50
	ppar	6.60	15.40	27.51	0.24	0.56	0.24	0.56	2.33	50
KINASE	cdk2	15.96	16.41	22.80	0.70	0.72	0.70	0.72	1.03	100
	jak2	22.16	19.08	49.24	0.36	0.31	0.45	0.39	0.86	100
NHR	gcr	24.62	11.89	49.24	0.29	0.14	0.50	0.24	0.48	100
	mcrin	28.41	32.20	49.24	0.15	0.17	0.58	0.65	1.13	100
	ppar	9.08	12.65	27.51	0.33	0.46	0.33	0.46	1.39	100

Table A.16: $\%EF$ changes upon simulation with protocol 2

Family	Protein	EF			Accuracy		$\%EF_{max}$		Ratio	Threshold
		Docking	I.F.	maximum	Docking	I.F.	Docking	I.F.		
KINASE	cdk2	15.40	14.96	22.00	0.70	0.68	0.70	0.68	0.97	50
	jak2	36.79	36.79	61.32	0.60	0.60	0.60	0.60	1.00	50
NHR	gcr	35.53	8.46	84.59	0.42	0.42	0.10	0.24	0.10	50
	mcrin	52.83	49.06	98.12	0.28	0.26	0.54	0.50	0.93	50
	ppar	7.85	12.34	28.03	0.28	0.44	0.28	0.44	1.57	50
KINASE	cdk2	15.18	13.20	22.00	0.69	0.60	0.69	0.60	0.87	100
	jak2	22.08	25.14	49.06	0.36	0.41	0.45	0.51	1.14	100
NHR	gcr	24.53	11.00	49.06	0.29	0.13	0.50	0.22	0.45	100
	mcrin	28.30	32.08	49.06	0.15	0.17	0.58	0.65	1.13	100
	ppar	9.25	10.09	28.03	0.33	0.36	0.33 e	0.36	1.09	100

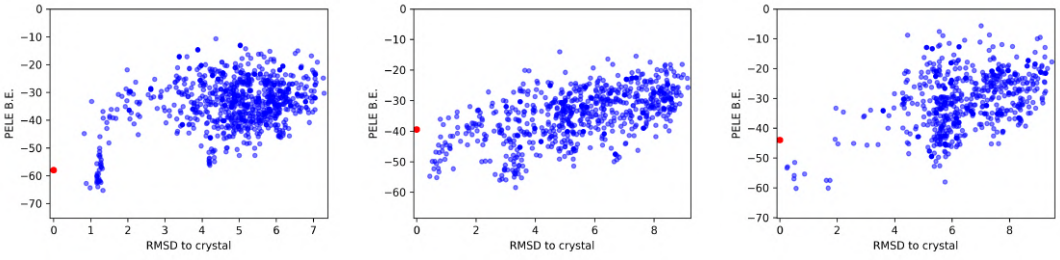
Appendix B

Retos Supplementary Information

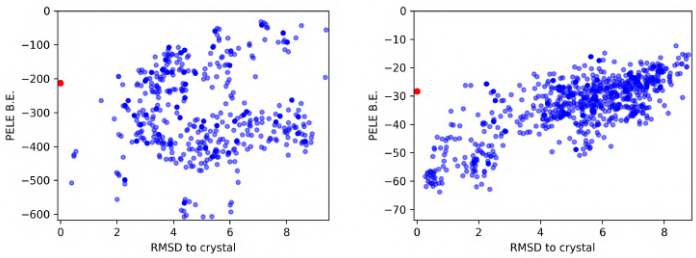
Table B.1: Table containing a summary of the PELE simulation parameters used for the different protocols

Pro-to-col	Equi-libra-ted	General parameters Type	# of processors	Tempe-rature	# of PELE Steps	Epochs	ra-dius	perturbation center	spawn-ing	# of trials	steering
0	No	adap-tive	16	1500	8	15	6	self COM	inv prop	10	random
1	No	adap-tive	32	1500	24	10	6	self COM	inv prop	10	random
2	No	adap-tive	16	1.000	8	15	6	self COM	inv prop	10	random
3	No	adap-tive	32	1.000	24	10	6	self COM	inv prop	10	random
4	No	adap-tive	16	1500	8	15	9	self COM	inv prop	10	random
5	No	adap-tive	32	1500	24	10	9	self COM	inv prop	10	random
6	No	adap-tive	16	1.000	8	15	9	self COM	inv prop	10	random
7	No	adap-tive	32	1.000	24	10	9	self COM	inv prop	10	random
8	No	adap-tive	32	1500	24	10	9	self COM	ep:0.33 hb	10	random
9	No	adap-tive	32	1500	24	10	4	self COM	ep:0.33 be	10	random
10	No	adap-tive	32	1500	24	10	4	average COM	ep:0.33 hb	10	random
11	No	adap-tive	32	1500	12	20	4	averg COM	ep:0.75 hb	10 / 25	random / 1
12	No	adap-tive	32	1500	12	20	4	average COM	ep:0.5 hb	10/25	random / 1
13	Yes	adap-tive	32	1500	12	20	4	average COM	ep:0.5 hb	10/25	random / 1

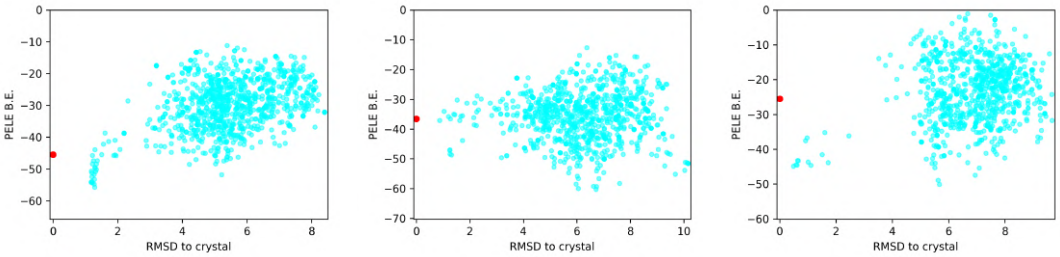
(a) crystal_A ALL_WAT protocol 0 (b) crystal_C ALL_WAT protocol 0 (c) crystal_D ALL_WAT protocol 0



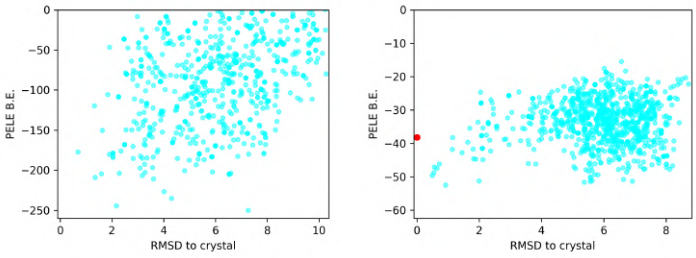
(d) crystal_E ALL_WAT protocol 0 (e) crystal_F ALL_WAT protocol 0



(f) crystal_A NO_WAT protocol 0 (g) crystal_C NO_WAT protocol 0 (h) crystal_D NO_WAT protocol 0



(i) crystal_E NO_WAT protocol 0 (j) crystal_F NO_WAT protocol 0



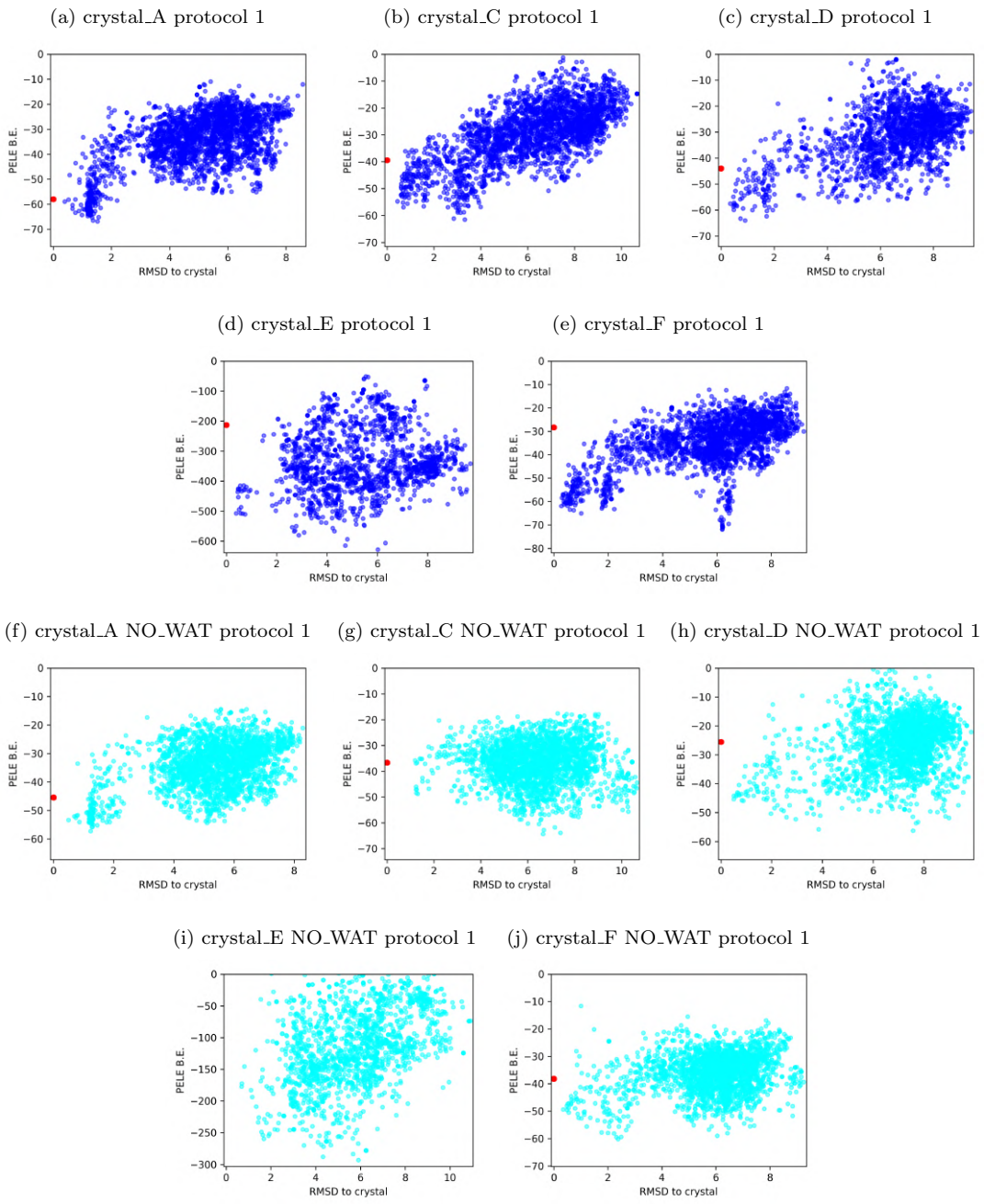


Figure B.1: Energy profiles for the simulations using the protocol 1 over the crystal structures with and without water

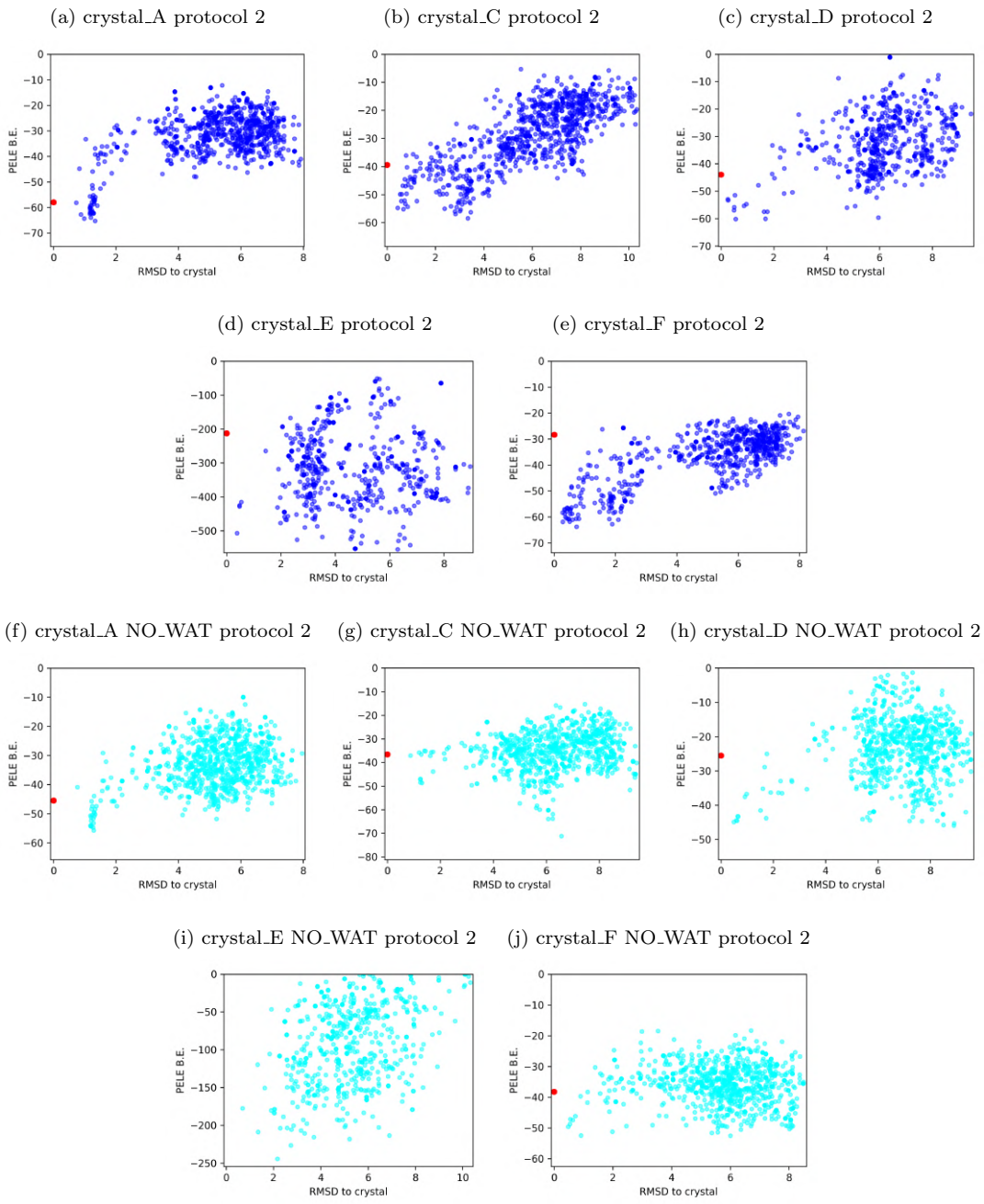


Figure B.2: Energy profiles for the simulations using the protocol 2 over the crystal structures with and without water

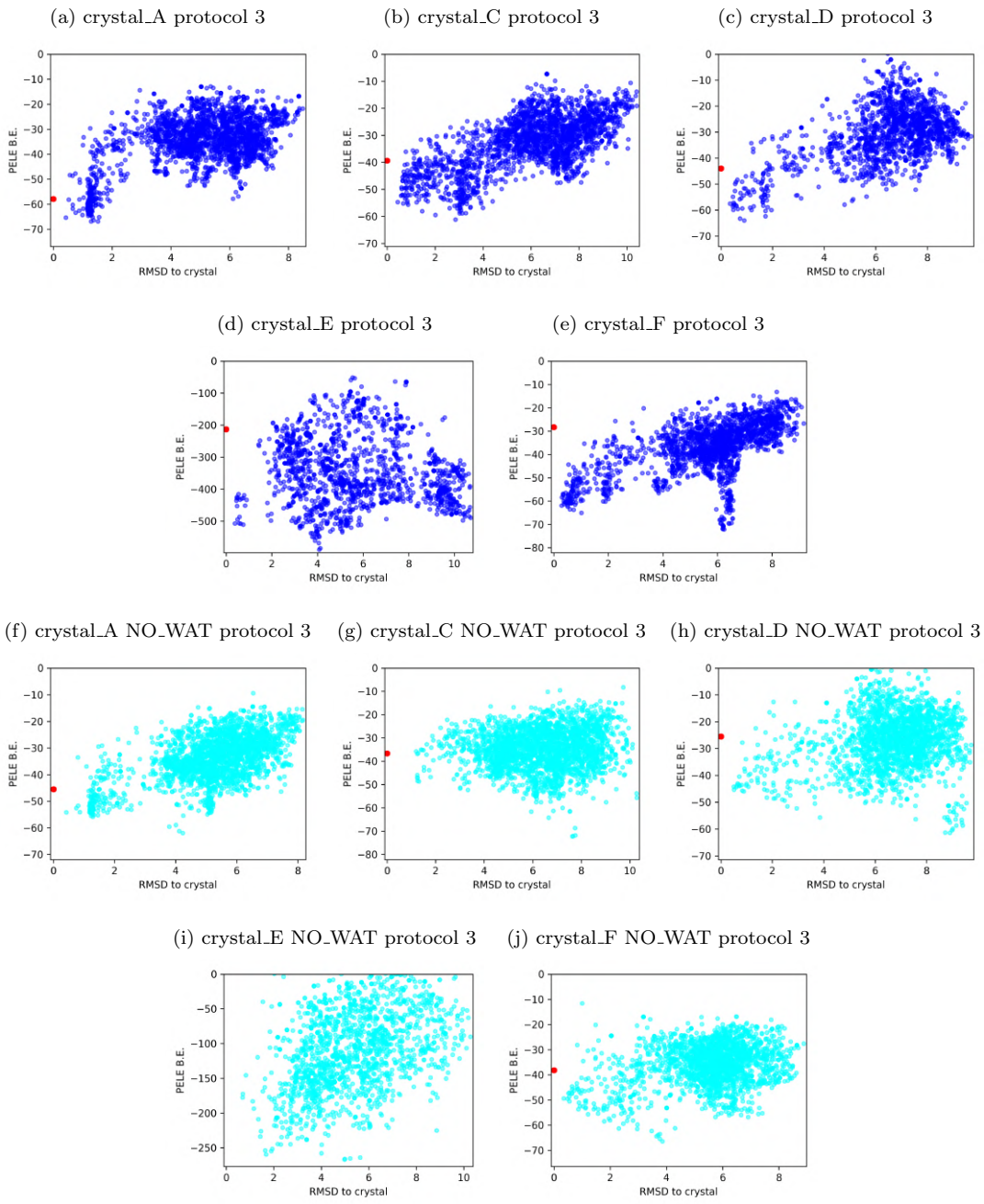


Figure B.3: Energy profiles for the simulations using the protocol 3 over the crystal structures with and without water

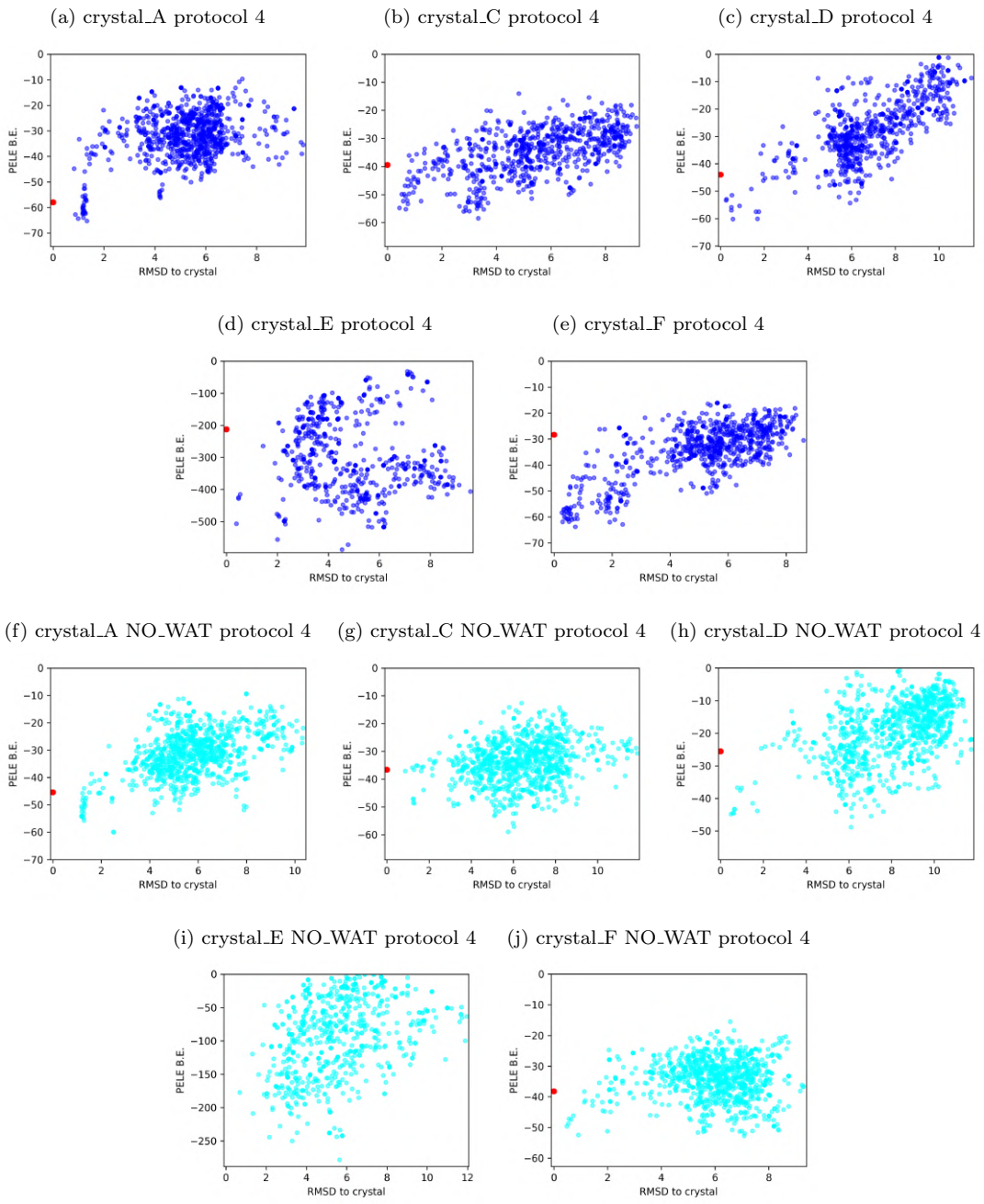


Figure B.4: Energy profiles for the simulations using the protocol 4 over the crystal structures with and without water

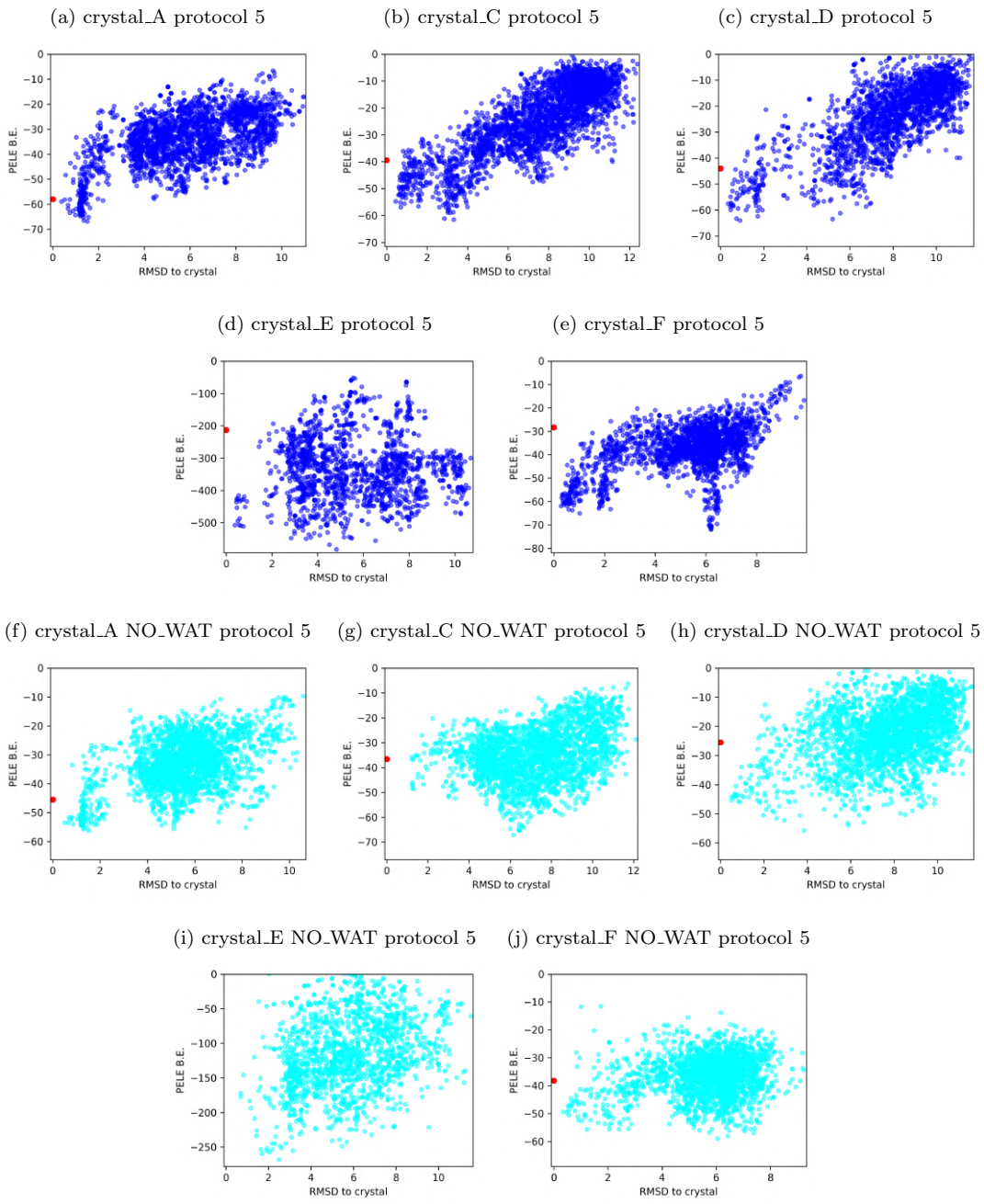


Figure B.5: Energy profiles for the simulations using the protocol 5 over the crystal structures with and without water

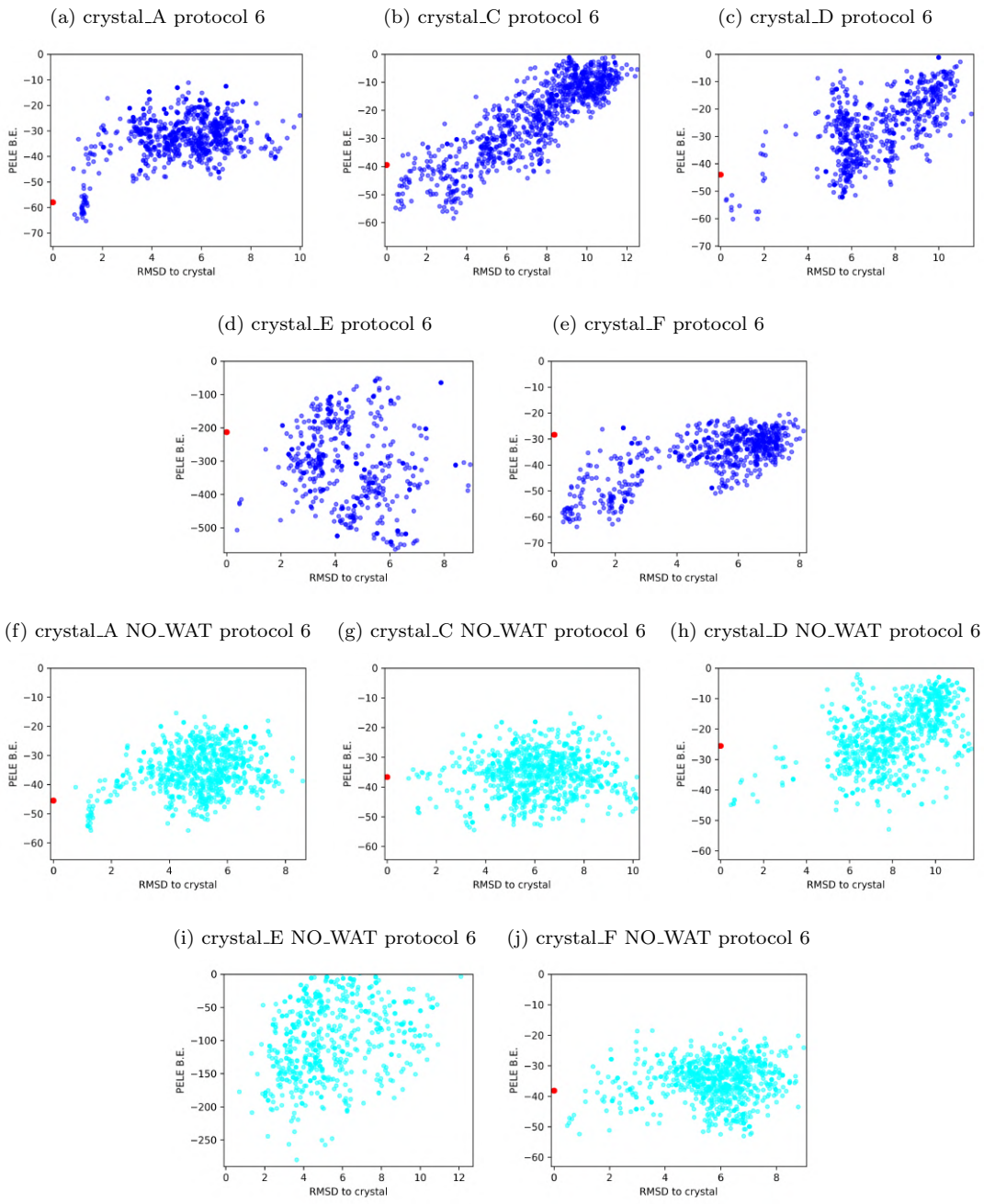


Figure B.6: Energy profiles for the simulations using the protocol 6 over the crystal structures with and without water

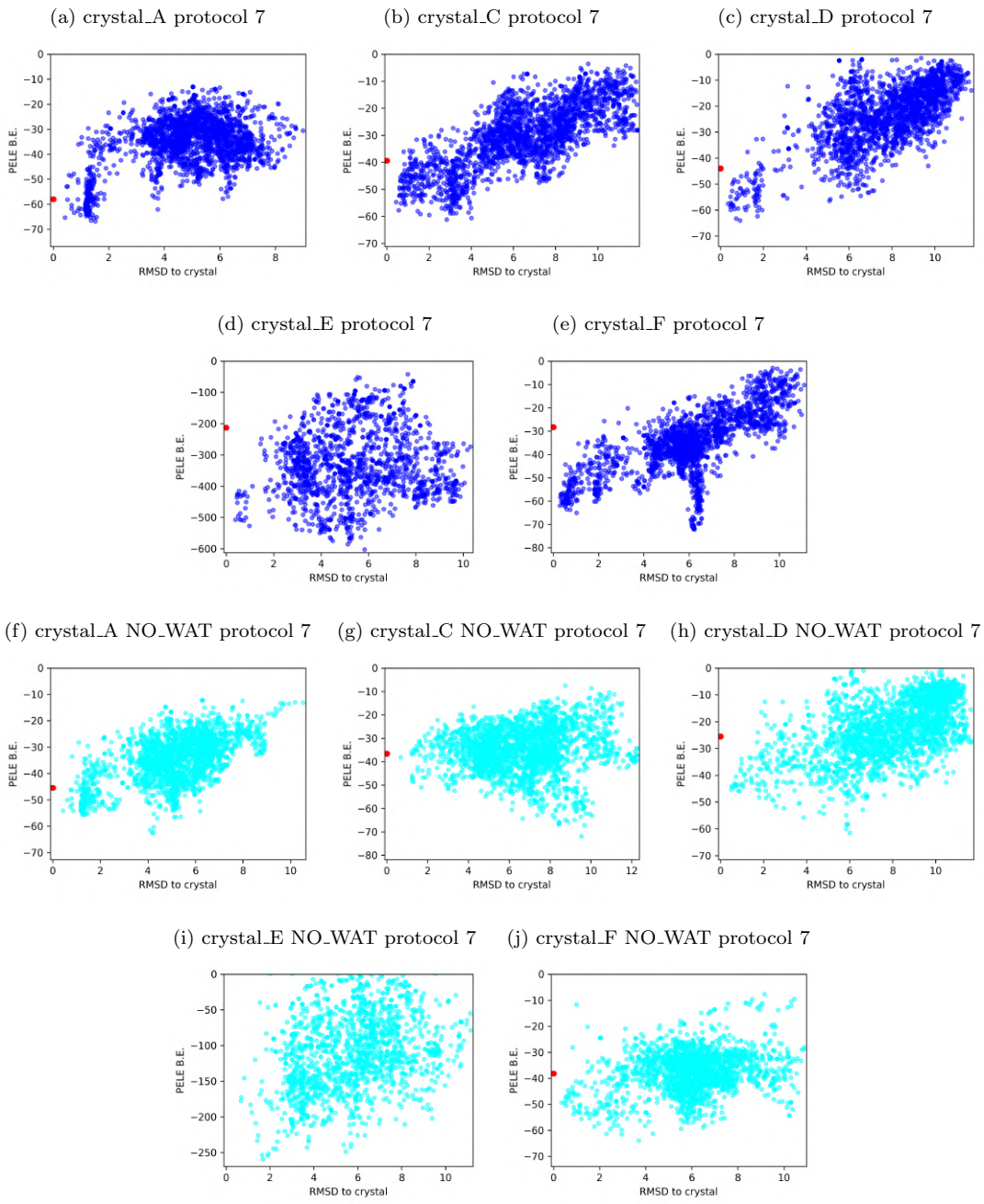


Figure B.7: Energy profiles for the simulations using the protocol 7 over the crystal structures with and without water

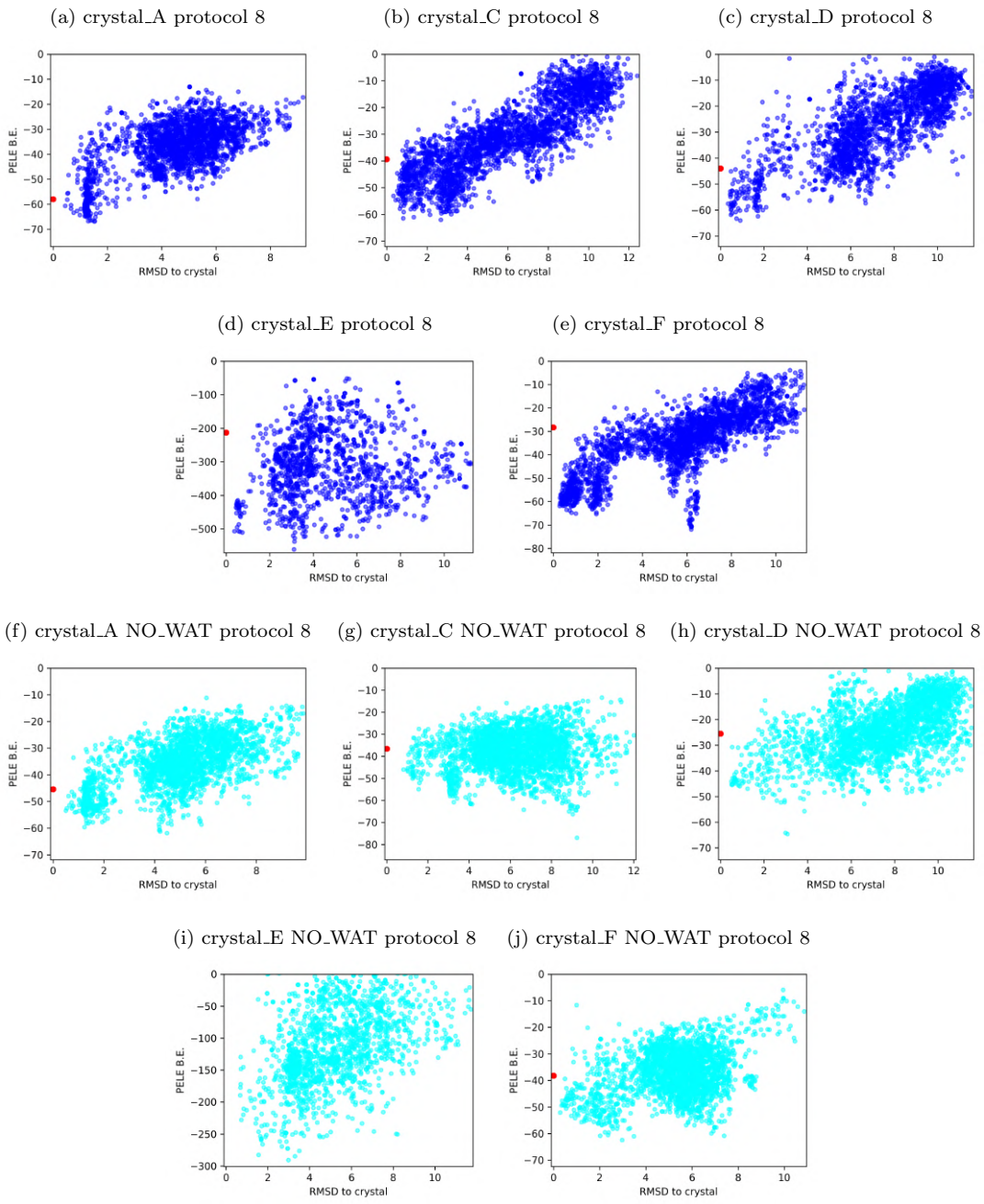
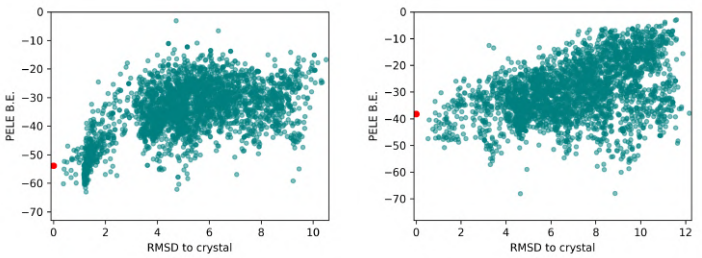


Figure B.8: Energy profiles for the simulations using the protocol 8 over the crystal structures with and without water

(k) crystal_A STR_WAT protocol 8 (l) crystal_C STR_WAT protocol 8



(m) crystal_D STR_WAT protocol 8 (n) crystal_F STR_WAT protocol 8

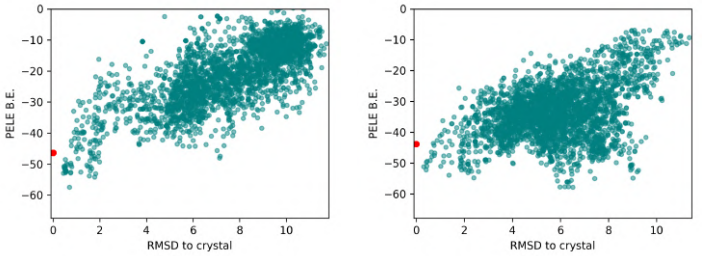
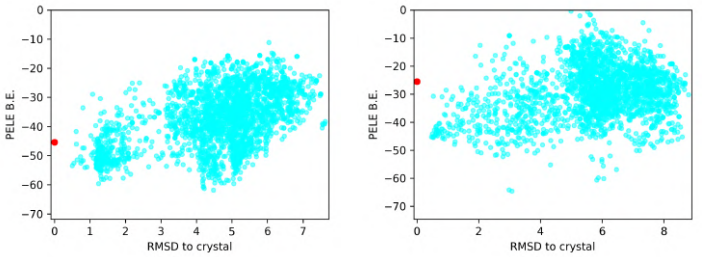


Figure B.8: Continuation. Energy profiles for the simulations using the protocol 8 over the crystal structures with and without water

(a) crystal_A NO_WAT protocol 9 (b) crystal_D NO_WAT protocol 9



(c) crystal_E NO_WAT protocol 9 (d) crystal_F NO_WAT protocol 9

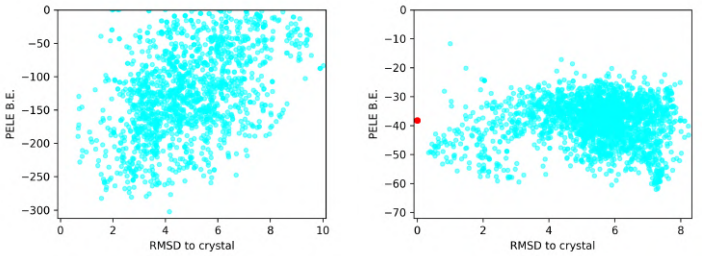
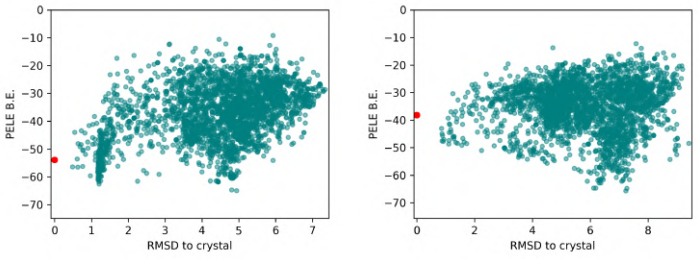


Figure B.9: Energy profiles for the simulations using the protocol 9 over the crystal structures with and without water

(e) crystal_A STR_WAT protocol 9 (f) crystal_C STR_WAT protocol 9



(g) crystal_D STR_WAT protocol 9 (h) crystal_F STR_WAT protocol 9

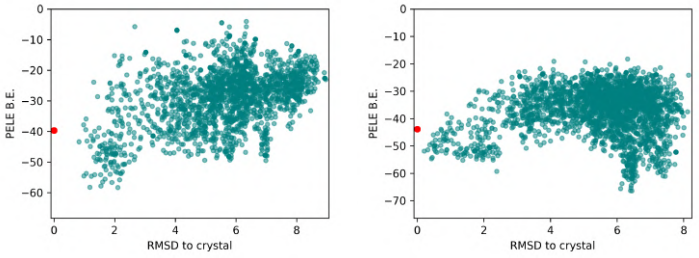
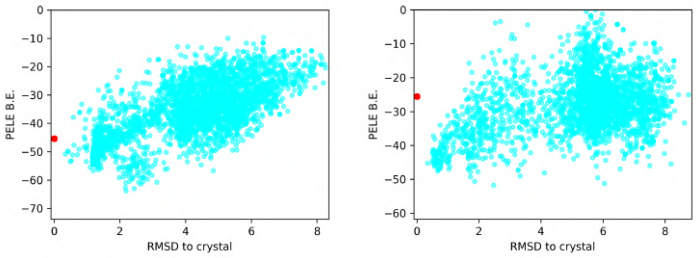


Figure B.9: Continuation. Energy profiles for the simulations using the protocol 9 over the crystal structures with and without water

(a) crystal_A NO_WAT protocol 11 (b) crystal_D NO_WAT protocol 11



(c) crystal_E NO_WAT protocol 11 (d) crystal_F NO_WAT protocol 11

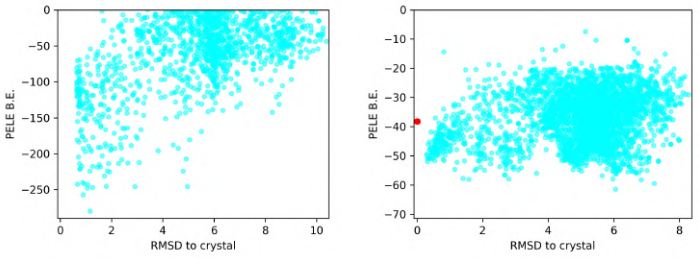
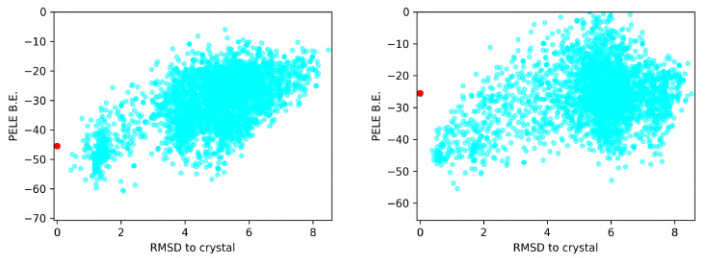


Figure B.10: Energy profiles for the simulations using the protocol 11 over the crystal structures with and without water

(a) crystal_A NO_WAT protocol 12 (b) crystal_D NO_WAT protocol 12



(c) crystal_E NO_WAT protocol 12 (d) crystal_F NO_WAT protocol 12

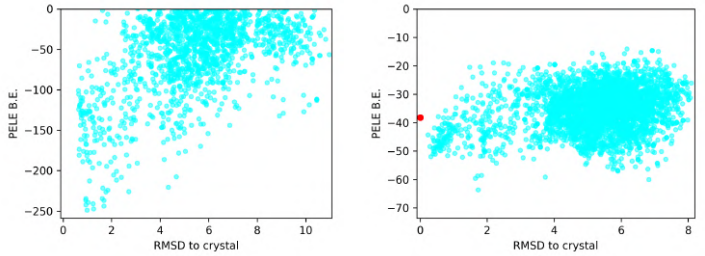
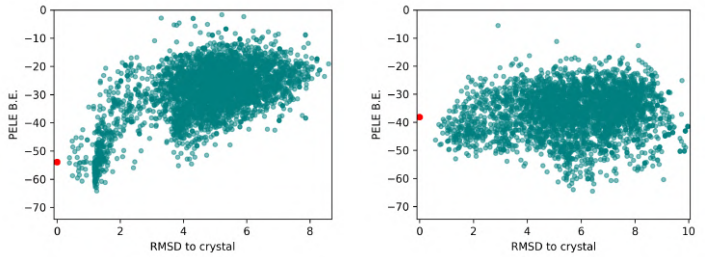


Figure B.11: Energy profiles for the simulations using the protocol 12 over the crystal structures with and without water

(e) crystal_A STR_WAT protocol 12 (f) crystal_C STR_WAT protocol 12



(g) crystal_D STR_WAT protocol 12 (h) crystal_F STR_WAT protocol 12

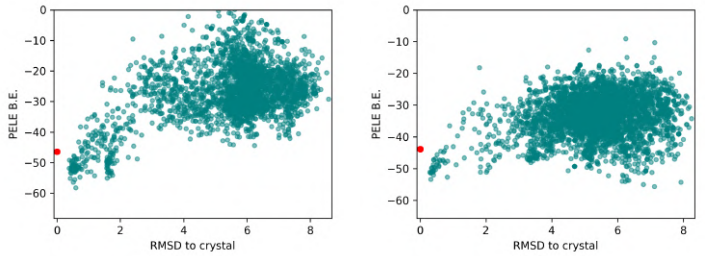
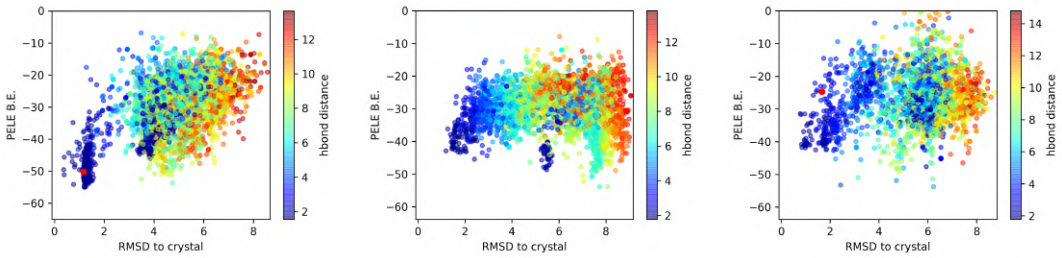


Figure B.11: Continuation. Energy profiles for the simulations using the protocol 12 over the crystal structures with and without water

(a) crossdocking ligand_A to receptor_A protocol 12

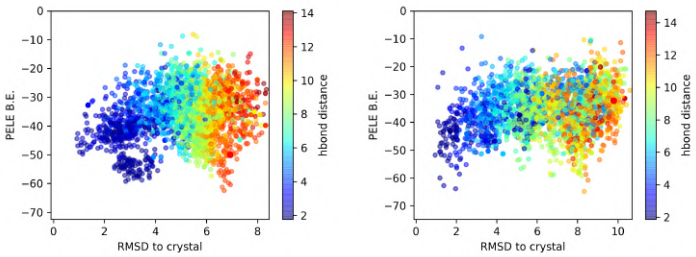
(b) crossdocking ligand_B to receptor_A protocol 12

(c) crossdocking ligand_D to receptor_A protocol 12



(d) crossdocking ligand_F to receptor_A protocol 12

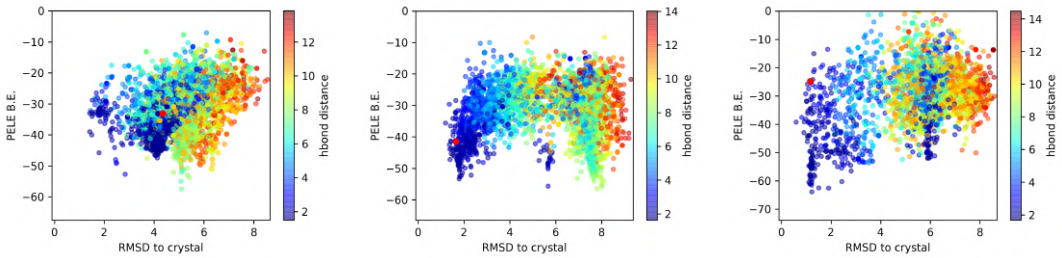
(e) crossdocking ligand_G to receptor_A protocol 12



(f) crossdocking ligand_A to receptor_B protocol 12

(g) crossdocking ligand_B to receptor_B protocol 12

(h) crossdocking ligand_D to receptor_B protocol 12



(i) crossdocking ligand_F to receptor_B protocol 12

(j) crossdocking ligand_G to receptor_B protocol 12

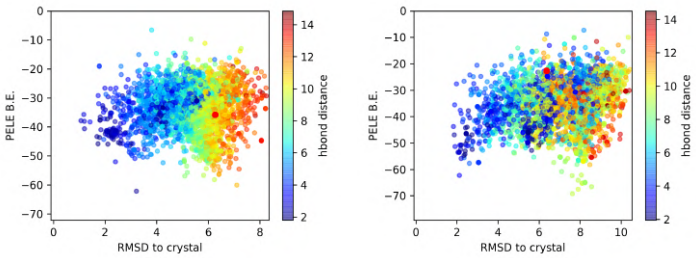
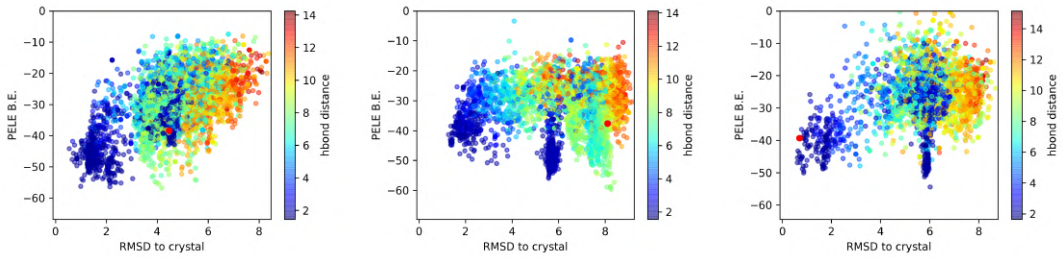


Figure B.12: Energy profiles for the simulations using the protocol 12 over the crossdocking structures with no water

(a) crossdocking ligand_A to receptor_D protocol 12

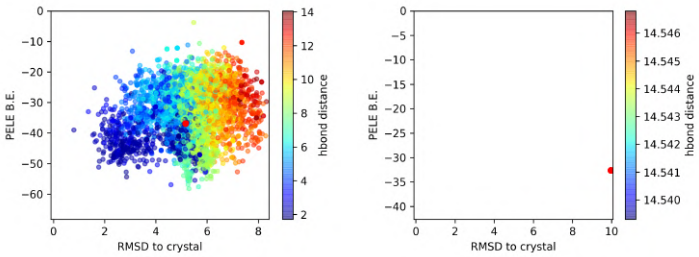
(b) crossdocking ligand_B to receptor_D protocol 12

(c) crossdocking ligand_D to receptor_D protocol 12



(d) crossdocking ligand_F to receptor_D protocol 12

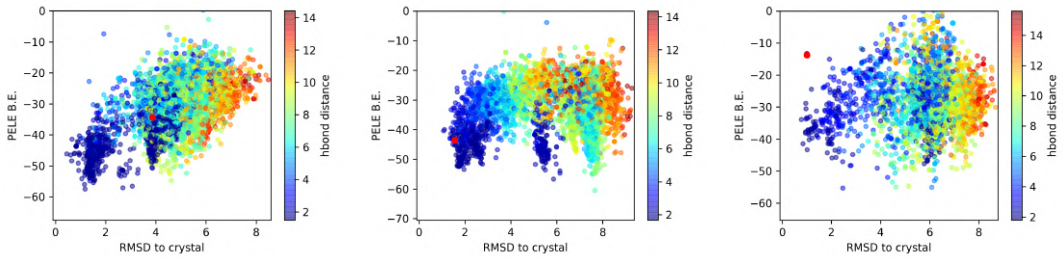
(e) crossdocking ligand_G to receptor_D protocol 12



(f) crossdocking ligand_A to receptor_F protocol 12

(g) crossdocking ligand_B to receptor_F protocol 12

(h) crossdocking ligand_D to receptor_F protocol 12



(i) crossdocking ligand_F to receptor_F protocol 12

(j) crossdocking ligand_G to receptor_F protocol 12

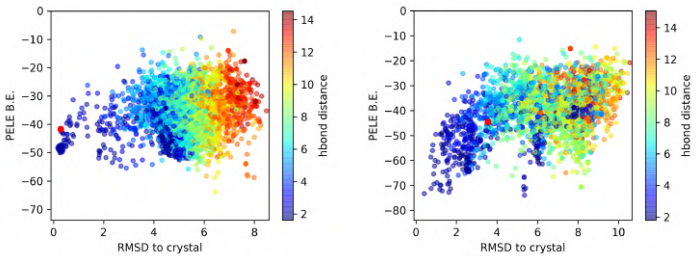
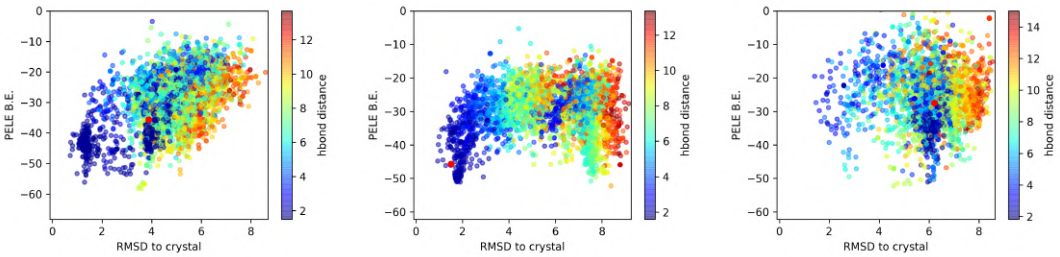


Figure B.13: Energy profiles for the simulations using the protocol 12 over the crossdocking structures with no water

(a) crossdocking ligand_A to receptor_G protocol 12

(b) crossdocking ligand_B to receptor_G protocol 12

(c) crossdocking ligand_D to receptor_G protocol 12



(d) crossdocking ligand_F to receptor_G protocol 12

(e) crossdocking ligand_G to receptor_G protocol 12

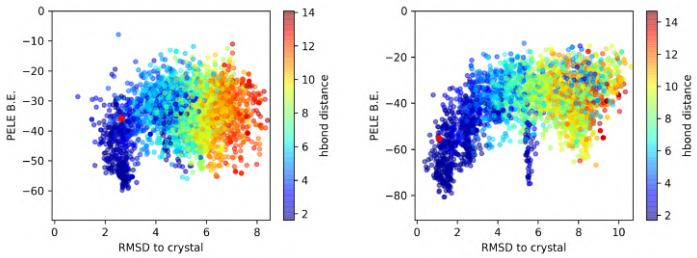


Figure B.14: Energy profiles for the simulations using the protocol 12 over the crossdocking structures with no water