

Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN  
CHIMICA

Ciclo XXVI

**Settore Concorsuale di afferenza:** 03/D1

**Settore Scientifico disciplinare:** CHIM/08

Computational Methods in Biophysics and Medicinal  
Chemistry: Applications and Challenges

**Presentata da:** Giovanni Paolo Di Martino

**Coordinatore Dottorato**

Prof. Aldo Roda

**Relatore**

Prof. Maurizio Recanatini

**Correlatore**

Dr. Matteo Masetti

Prof. Andrea Cavalli

**Esame Finale anno 2014**



*To my father*



# Table of Contents

Preface.....	1
<b>Chapter 1. Introduction.....</b>	<b>3</b>
1.1 Computational Methods in Medicinal Chemistry.....	4
1.1.1 Application to hERG Potassium Channel	10
1.2 Computational Methods in Biophysics.....	12
1.2.1 Application to the Peptidyl-Prolyl <i>cis-trans</i> Isomerase Pin1	15
<b>Chapter 2. Theoretical Background.....</b>	<b>17</b>
2.1 System Representation.....	18
2.1.1 Molecular Mechanics	18
2.1.1.1 The Amber Forcefield	20
2.1.2 Quantum Mechanics	22
2.2 Sampling of Microstates.....	25
2.2.1 Statistical Mechanics	25
2.2.2 Monte Carlo	27
2.2.3 Molecular Dynamics	28
2.2.4 Free Energy Calculation	31
2.2.4.1 Enhanced Sampling	32
2.2.4.1.1 Umbrella Sampling	33

2.3 Approximate Methods.....	35
2.3.1 Molecular Docking	36
2.3.2 MM-GB(PB)SA	37
<b>Chapter 3. An Automated Docking Protocol for hERG Channel Blockers.</b>	<b>40</b>
3.1 Introduction.....	41
3.1.1 Aim of the Project and Protocol Presentation	43
3.2 Computational Methods.....	48
3.2.1 Channel Models Generation	48
3.2.2 Shape-based Cluster Analysis	49
3.2.3 Docking	51
3.2.4 Post-processing of the Docking Outcome	51
3.2.5 Building and Evaluation of hERG-Blockers Models	53
3.2.6 MM-PBSA Refinement	55
3.3 Results.....	58
3.3.1 Development of the Protocol	58
3.3.2 Protocol Validation	70
3.4 Discussion.....	72
3.4.1 Pore Shape of the hERG Channel Models	72
3.4.2 Binding Modes of the Sertindole Analogues	74
3.4.3 Performance of the Structure-Based Models	75
3.5 Conclusions.....	77
3.6 Appendix.....	79
3.6.1 Presentation of the Automated Protocol CoRK <sup>+</sup>	79
3.6.2 Supporting Materials	81

---

<b>Chapter 4. Insights on the Pin1 Peptidyl-Prolyl <i>Cis-Trans</i> Isomerization..</b>	<b>89</b>
4.1 Introduction.....	90
4.1.1 Prolyl <i>Cis-Trans</i> Isomerization	95
4.1.2 PPIase Catalyzed Prolyl <i>Cis-Trans</i> Isomerization	101
4.1.2.1 Cyclophilins	101
4.1.2.2 FKBP	105
4.1.2.3 Parvulins	107
4.1.3 Aim of the Work and Project Presentation	111
4.2 Computational Details.....	112
4.2.1 Substrate and Models Setup	112
4.2.2 Testing the C and NC Reaction Mechanisms	113
4.2.3 Investigation of the Bulk and NC <i>Cis-Trans</i> Isomerization	114
4.2.4 Prolyl Nitrogen Pyramidal Conformations	115
4.2.5 Free Energy Difference between Bulk and NC <i>Cis</i> States	117
4.2.6 Energy Barrier to Disrupt the Intramolecular Hydrogen Bond	117
4.3 Results.....	118
4.3.1 Testing the C and NC Reaction Mechanisms	118
4.3.2 Bulk <i>Cis-Trans</i> Isomerization	121
4.3.3 NC <i>Cis-Trans</i> Isomerization	127
4.3.3.1 Entropic Effect	129
4.3.3.2 <i>Cis</i> Ground State Destabilization	132
4.3.3.3 Intramolecular Hydrogen Bond Investigation and Effects on the Barrier	134
4.3.3.4 Intermolecular Interactions during the Isomerization	137
4.4 Discussion.....	140
4.4.1 The Entropy Trap	140
4.4.2 <i>Cis</i> Ground State Destabilization	142

4.4.3 The <i>Hydrogen Bond Shuttle-Assisted Mechanism</i>	143
4.5 Conclusions.....	145
<b>Chapter 5. Conclusions.....</b>	<b>147</b>
Acknowledgements.....	150
References.....	151




# Preface

*Chemists used to create models of molecules using plastic balls and sticks. Today, the modelling is carried out in computers. [...] Today the computer is just as important a tool for chemists as the test tube [1].*

From the second half of the last century, computational chemistry has reached a fast growth facilitated by the extraordinary progress of computational technologies and software developments. Advances in computational resources, and the easy way to get access to powerful workstations, have produced a remarkable increasing in the number of scientists routinely using *in silico* techniques. A clear indicator of this phenomena is represented by the quantity of published papers and references to computational methods in scientific journals. Nowadays, each research group has at least one or more collaborators doing computational experiments.

The main goal of my dissertation consists in providing a detailed description of several computational methods, ranging from molecular mechanics (MM) to quantum mechanics (QM) approaches, showing *when* and *how* they could be applied for getting useful information in the field of medicinal chemistry and biophysics. One has to keep in mind that, in switching from experimental to a theoretical approach, different approximations have necessarily to be assumed. The treatment of very large chemical systems could become computationally demanding, and simplifications in the model representation, as well as approximations in the algorithms, are strongly required. In the light of that, choosing which aspects or properties of a particular real system could be ignored or have to be taken into account is not trivial, and represents an open computational challenge.

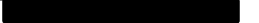
My Ph.D. thesis consists in five sections. In Chapter 1, I will briefly present an overview of current computational methods and their applications in medicinal chemistry and biophysics, with a particular focus on introducing the topics studied during my Ph.D.: the development of an automated docking protocol for the voltage-gated hERG potassium channel blockers, and the investigation of the catalytic mechanism of the human peptidyl-prolyl *cis-trans* isomerase Pin1. In Chapter 2, I will provide the theoretical background of the methodologies used in my projects. Finally, in Chapter 3 and Chapter 4, I will present details and results obtained working on my two projects introduced before. The latter two sections could be therefore considered totally independent from each other. Conclusions, considerations, and challenges will be summarized in Chapter 5.



---

**Chapter 1.**  
**Introduction**

---



## 1.1 Computational Methods in Medicinal Chemistry

Computational methods have found a wide application in medicinal chemistry, providing valuable information in academic, as well as in pharmaceutical companies, drug discovery and development research. The usage of computational techniques for medicinal chemical purposes, is generally referred as computer-aided drug design (CADD).

CADD plays a key role in discovering biologically active compounds, in structural optimizations, in predicting molecular properties, like logP, solubility, and ADME/T parameters [2,3]. The reduction of costs and time for doing *in silico* experiments, is one of the main advantages of CADD. Advances in computational methods, and the advent of high performance parallel computing, have determined an acceleration of all the steps in the drug discovery process, starting from the hits identification to the lead compounds optimization [4]. The success of CADD and cheminformatics in the pharmaceutical industry, is also due to the extraordinary amount of data to be stored and analysed, and by the necessity to design and manage databases and chemical libraries. For instance, CADD has led to the discovery of several compounds with therapeutic activity against different pathological conditions. The ACE inhibitor *captopril* (Bristol Myers-Squibb), used for the treatment of hypertension, or the carbonic anhydrase inhibitors *dorzolamide* (Merck), used against the ocular disease glaucoma, represent just two historical cases in which computational techniques have successfully played a crucial role in the identification of molecules that have gained approval by US Food and Drug Administration (FDA) for therapeutic use [4,5]. By the way, CADD should not be considered an alternative to the conventional high-throughput screening (HTS) assays. Given their similar tasks and goals, CADD and HTS methodologies should

be used in combination in drug discovery, improving the hit rates, maximizing the output, and reducing the waste of resources in synthesis and screening [6-9]. Ideally, CADD has to be introduced in a pre-filtering stage, for the identification of compounds for HTS assays, before their synthesis or purchase [10]. Because they result to be cheaper and faster than experimental assays, CADD techniques are suitable for pre-screening large virtual libraries of compounds, providing valuable information for subsequent *in vitro* tests [11]. CADD offers several approaches and strategies which could be successfully applied in drug discovery, for target prediction, lead optimization, hit identification, molecular properties and affinity prediction, as well as for understanding the dynamical evolution of a biological system. These computational techniques are conventionally divided into *structure-based* (SB) and *ligand-based* (LB) approaches. SB drug design exploits structural information of the drug target which represents a pre-requisite for its application. The 3D structure of a target is generally determined by means of experimental X-ray crystallography or NMR. Whenever no crystal structure is available, computational methods like homology modelling (or comparative modelling), could be used for predicting the protein 3D structure based on the sequence alignment with templates [12-14]. The basic idea of this approach, is that similar sequence corresponds to similar a structure, and the closer is the sequence identity, the higher is the quality of the model [15]. In case of low sequence identity, building a structure by means of homology modelling, represents a very challenging task. On the other hand, LB techniques can provide crucial insights in lead discovery and optimization, exploiting the information coming from established ligands of the biological target of interest. The most popular LB approaches, are the quantitative structure-activity relationship (QSAR) method and pharmacophore modelling, allowing to build predictive models that are suitable in drug discovery [16,17].

Here, I will briefly present challenging tasks in drug discovery and development, and the widely used computational approaches in CADD.

### *Target identification*

Drug target prediction represents a challenging step in the drug discovery pipeline [18]. The number of targets for all approved therapeutic drug has been long debated.

Recently, in 2006, Overington et al., finally proposed a consensus number of 324 targets for all the classes of US FDA-approved therapeutic drugs [19]. The presence of drugs which display a therapeutic activity by modulating multiple targets, has significantly contributed to worsen this scenario. Computational methods have acquired a crucial role in identifying targets for small molecules during the last years. In this context, *reverse docking* has gained significant applications [20,21]. This method consists in an inverse-docking protocol: the term *molecular docking* is used for identify a promising computational technique which allows to dock a ligand into a protein binding site, predicting the binding mode. This approach could be exploit for identifying novel compounds with unknown therapeutic activity against a target of interest. Conversely, reverse docking attempts to dock a single small molecule into a pool of protein structures, leading to the identification of potential protein targets, or the prediction of possible side effects of a drug candidate [20]. Because of the considerable amount of protein structures in the Protein Data Bank (PDB), and the computational costs required for job submissions, protein cavity databases have been developed in order to make reverse docking search faster. In particular, these cavity models, derived from an overlapping of spheres which fill up that cavities [20,22]. The computational approach utilized for generating a spatial arrangement of essential features necessary for ensuring the optimal protein-ligand interactions, is known as *pharmacophore modelling* [23]. An *in silico* pharmacophore model consists in an ensemble of chemical features like hydrogen bond acceptors or donors, hydrophobic regions, positively or negatively charged functional groups. Screening a small molecule against a set of pharmacophore models, which represent the binding sites of several protein structures, leads to a drastically reduction of computational costs and resources required by reverse docking [23,24]. A pharmacophore could be derived by means of a LB, or a SB approach. In the first case, by superimposing compounds, with known therapeutic activity against a protein, and extracting the common chemical features, while, in the second case, by probing possible interaction points between ligands and target.

#### *Protein-ligand binding mode prediction*

Due to its ability in predicting the interactions between protein-ligand complexes, molecular docking has acquired a crucial role for hit discovery and lead optimization. [25,26]. Binding mode prediction is described by a two-steps process. Starting from a set of compounds, molecular docking allows a configurational exploration into a protein binding site. This step is a very challenging task, as even small compounds could be characterized by many conformational degrees of freedom. Therefore, an accurate sampling of all the degrees of freedoms represents a computationally expensive process. Molecular docking should be a strategy which provides a reliable conformational search, leading to a correct identification of binding poses, but, in the same time, fast enough to allow the evaluation of thousands of compounds in a docking run, for *virtual screening* application [25]. This step is then followed by the usage of a scoring function in order to evaluate the binding affinity, and provide a correct ranking of the poses. The development of a scoring function able to accurately describe protein-ligand interaction, represents an important task in computationally driven drug discovery. In a recent review regarding the evaluation methods for protein-ligand interaction, Huang et al. have reported three important applications of energy scoring function in molecular docking [27]. The first application is related to the ability of a scoring function to predict protein-ligand interactions, ideally ranking as best poses the experimentally determined binding modes [28]. The second, consists in the prediction of the absolute binding affinity, a challenging task, which could enhance the accuracy in lead optimization [29]. The third and last application of a scoring function, is to support the identification of potential hits against a target of interest. An ideal scoring function has to successfully identify compounds with known, experimentally proved, activity, during a large database screening. In this context, it has been stated that the evaluation of the solvation contribution in drug binding, plays a critical role in the accuracy of the results [30]. Three classes of scoring functions are applied in docking: force field based, empirical, and knowledge based functions, whose details are reported in Chapter 2. Another important aspect which should be taken into account in molecular docking, consists in the treatment of ligand and protein flexibility [25]. Ligand flexibility could be explored by means of systematic search methods, like incremental construction algorithms [31]; stochastic methods, like Monte Carlo [32-34] or genetic algorithm [35]; by usage of molecular dynamics [36]. Protein flexibility is a more challenging

issue. Monte Carlo approach and molecular dynamics are commonly applied for sampling the local flexibility of the binding site in the protein structure. Another useful approach consists in modelling side-chain conformations of the binding site residues by means of libraries of pre-defined rotamers [37]. Protein flexibility could be also successfully solved before carrying out a docking simulation, using an ensemble of protein conformations of the target of interest [38].

### *Hit identification*

Computational strategies which allow an automatically evaluation of large libraries of ligands, aimed at the identification of compounds with relevant biological activity against a target, are generally referred as virtual screening (VS) related methods [39,40]. Contrary to the previous task regarding the binding mode prediction, in which high computational costs are secondary to the obtainment of high accuracy in the results, in VS the optimization of the speed of the calculations is of primary importance. This issue could be overcome by means of more simple scoring functions in which some features, necessary for a better affinity prediction, are omitted [41]. VS methods are conventionally classified as ligand-based (LBVS) or structure-based virtual screening (SBVS), based on the direct knowledge of the 3D target structure [11,26]. When a target protein is provided, SBVS can be employed, allowing explicit molecular docking of each database compound into the target binding site. At the end, after a rigorous post-docking filtering criteria, a subset of potential active compounds are selected for experimental tests. The success of a SBVS strongly depends on the methods selected for defining the screening protocol, on the way they are combined, and on the chemical databases used [42]. SBVS is also widely applied in the context of fragment-based drug discovery (FBDD). The direct role of FBDD, is to identify low molecular weight scaffolds (molecular mass less than 300 Da) which are able to bind only weakly the binding site of interest. These fragments are then chosen and subjected to several optimization steps, in order to model high affinity compounds [43,44]. On the other hand, once a target structure is not provided, a pharmacophore model, derived extracting common features from known active ligands, could be used for the identification of novel hits. This method, known as pharmacophore-based VS, represents a LBVS approach that



has been found a wide use in drug discovery. Even though several successful applications of the method are reported in literature, one of the most important problem of this technique, is represented by its strong dependence on the quality of the pharmacophore model used for the screening, which could dramatically compromise the hit identification process [23,45].

### *Lead optimization*

In drug development, lead optimization covers computational strategies exploited for enhancing the binding affinity of the identified hit compounds. One could envisage using molecular docking for this purpose. Ideally, once a docking protocol is able to provide reliable binding modes and highly accurate ranked poses, the same procedure could be used for predicting the binding affinities of modified ligands (structurally related compounds), and providing, in this way, useful information for synthesis and experimental testing [27]. Lead optimization is therefore straightforwardly influenced by the scoring function used in the simulations. Unfortunately, inclusion of key contributions like entropic and solvation effects in docking scoring functions, is still a challenge. Higher accuracy could be achieved by means of more computationally expensive method like free-energy perturbation (FEP), or higher level of theory (quantum mechanical) approaches [41].

### *Molecular properties prediction*

Quantitative structure-activity relationship (QSAR) methods [46] offer the possibility to predict physicochemical, pharmacokinetic and toxicological properties of molecules, gaining a crucial role in drug design. These methods attempt to find an empirical relationship between molecular structures and biological properties, aiming at achieving valuable information for the development of novel drug candidates. In particular, QSAR models represent a useful tool for the prediction of the activities of untested compounds [47,48]. These methods are generally carried out following two different approaches. 2D QSAR attempts to find a simple correlation between a set of independent variables, referred as chemical descriptors, and a dependent variable. The latter, represents the value for which the model provides prediction,

and is routinely expressed as biological activity,  $pK_a$ , or  $\log P$  (estimation of drug hydrophobicity) [49]. Several methods could be exploited for the selection of 2D QSAR descriptors. In this context, widely used approaches are regression-analysis, multivariate analysis algorithms, heuristic, and genetic algorithms [11]. By the way, this kind of approach has a limited utility in designing new molecules. This is mainly due to the lack of information regarding the 3D structure of the compounds. More computationally complex than 2D QSAR, is the 3D QSAR methodology. In this case, a QSAR model is built by means of molecular descriptors which are directly derived by the compound conformations. For building a 3D QSAR model, a training set of molecules, with a wide range of activity, has to be provided. The next step, consists in the generation of the molecular conformations for each member of the training set. For this purpose, molecular mechanics and experimental data could be exploited. After alignment in space, a dimensionality reduction step leads to acquire the 3D distribution of electrostatic and steric fields [11,50]. A validation of the generated model, is commonly carried out with a test set of experimentally known active compounds. The latter represents an important step for proving the robustness of the 3D QSAR model. QSAR methods could assist lead optimization phase, explaining the activities of known compounds and providing valuable information regarding structure-activity relationships. Combined approach 3D QSAR/VS are also reported in literature [51-53].

### **1.1.1 Application to hERG Potassium Channel**

In this section I will present the application of computational methods, and related challenges, for achieving new insights on the blockade of the hERG potassium channel, described in details in Chapter 3.

The voltage-gated hERG potassium channel, also known as Kv11.1, is expressed in several organs and tissues. In the heart, it plays a key role in modulating the rapid component of the potassium current  $I_{Kr}$ , which is responsible for

the repolarization phase of the cardiac potential action [54]. During the last years, a remarkable number of studies on hERG has been reported in literature.

This channel represents an important target in drug discovery and safety: alterations of its functionality have been associated to the long QT syndrome-type 2 (LQTS2), a potentially lethal pro-arrhythmic condition [55]. This alteration could be caused by inherited mutations, or induced by an accidental block by drugs [55]. Although the dysfunction caused by a block by drugs represents an extremely rare event, it is of primary importance in terms of drug safety and drug discovery, acquiring a detailed knowledge of the molecular features at the basis of the channel block. In this context, assessing the blockade activity at the early stages of the drug discovery process, is necessary for limiting waste of time and for reducing costs and resources in the development of compounds which potentially carrying a hERG toxicity [56]. The application of computational methods for overcoming this issue is a very challenging task, since no crystallographic structure of hERG are currently available. In this scenario, ligand-based approaches have represented the most appropriate choice for performing prediction. 2D [57-61] or 3D [62-66] QSAR models have been widely applied in academia and pharmaceutical companies. As introduced on the previous paragraph, several are the limitations of ligand-based approaches. In fact, 2D descriptors lead to a difficult chemical interpretation, whereas 3D models remarkably depend on the conformations chosen for building the QSAR model and for the alignment. During the last years, a considerable effort has also been spent for developing structure-based models for hERG blockade prediction. The main difficulty encountered with these kind of approaches is the lack of a crystal structure, and the low percentage of sequence identity between hERG and available templates, making homology modelling particularly challenging. By means of structure-based methods, both open and closed states of the channel have been modelled. Österberg and Boukharta have docked a series of derivatives of the antipsychotic sertindole, a potent blocker of the channel, in an open state hERG model [67,68]. A more general set of blockers has also been exploited within a closed conformation of the channel by Farid and Coi [69,70]. Because of the lack of structural information, these structure-based models are usually validated by their ability to reproduce the observed trend in binding free energy over the considered set of blockers [67,68,71]. Despite these remarkable efforts, the lack of a consistent

binding mode for the most potent blockers entails an incomplete comprehension of the phenomenon under study.

## **1.2 Computational Methods in Biophysics**

Computational techniques, such as molecular dynamics simulations (MD), have become essential theoretical tools in biophysics, providing atomic details of the structures and behaviours of biological systems, and allowing for computing systems dynamics, as well as thermodynamic properties.

X-ray crystallography, electron microscopy and other techniques, are powerful tools for revealing the three-dimensional arrangements of the atoms of a molecule or biological system. The limitation of these techniques, is represented by the fact that they are able to determine static structures, but only limited information about protein dynamics, which are often crucial for biological function [72]. Proteins, for example, are highly dynamic, and the resulting conformational changes are directly linked to their functions, to their ability to recognize and bind a substrate, or to catalyse a reaction mechanism [73-75]. MD simulations allow to overcome this issue, modelling atomic-level motions computationally. In MD simulations, positions and velocities of the atoms of a system evolve in time according to the laws of classical mechanics. All the forces acting on atoms are evaluated by means of forcefields, which consist of a combination of parameters fitted with experimental or quantum mechanical data, and first-principles physics [72]. By the way, this approach requires high computational costs, and moreover, the simulations are in general limited in time and model size [76]: in fact MD are typically performed on biological systems containing thousands or millions of atoms, with accessible timescale ranging from nanoseconds to few microseconds. These timescales are clearly shorter than those required by biological events, such as protein folding, protein-drug binding, protein conformational changes, membrane transport, limiting in this way the applicability of

this technique. These events generally take place on the timescales of microseconds to milliseconds. Modelling few microseconds of dynamics could require several months of simulations, exploiting modern supercomputers and high-end hardware. Advances achieved in computer technologies and in the development of new parallelization algorithms have recently increased the timescale accessibility, allowing the first milliseconds scale of simulations which lead to investigate key biochemical processes [72]. In this context, the new machine Anton, developed at D. E. Shaw Research, is able to perform about 20  $\mu\text{s}/\text{day}$  all atoms MD simulations, extremely increasing the simulations rates achieved by common software and hardware [77]. Another way to speed up the simulations, is to use software designed to run on graphics processing units (GPUs) on graphics cards. The bio-molecular dynamics software ACEMD developed by Gianni De Fabritiis's group, has been optimized to reach the microsecond time scale even on cost-effective workstation hardware using the power of GPUs [78]. Anyway, another important aspect that has to be considered is that MD simulations are completely under the control of the user. Indeed, one can easily decide to modify a potential or to remove a particular contribution in the energy function for the determination and examination of a specific property of the analysed system [79]. Over the years, improved force fields have been developed, and the longer accessible timescales have allowed a rigorous validations against experimental data [80-82]. Comparison between MD simulations and experimental data is indeed crucial for testing the accuracy of the results, and improving force field parameters.

Although the recent advances in classical simulation methodologies, alternative approaches could be used to accelerate the sampling and reducing the calculation time required for simulating long-timescales dynamics of biomolecular systems. The so called *enhanced sampling* techniques, are widely used for sampling states separated by large barriers. These techniques are particularly suitable for the calculation of the free energy associated to rare events. In fact, by means of a classical *unbiased* MD, the configuration space around a minimum is well sampled, whereas higher energy regions, are sampled rarely. In order to obtain a free energy profile, a probability density around high energy regions is necessary, and unbiased MD is unfeasible [83]. Commonly used enhanced sampling techniques could be distinct in two main groups: approaches based on *collective variable (CV) biasing*,

and approaches that rely on *tempering* [84], briefly introduced in the following sections. The applicability of a particular method, strongly depends on the system which has to be analysed [85,86].

#### *Approaches based on CV biasing*

These approaches include methods like thermodynamic integration [87,88], umbrella sampling [89,90], and metadynamics [91,92], which are based on the idea of accelerating the sampling by introduction of a bias potential on a limited number of degrees of freedom. These degrees of freedom, or CVs, represent the reaction coordinates ( $\xi$ ) which describe the transition between states that are difficult to sample, and have to be properly chosen. In thermodynamic integration, or *Blue Moon* method [87,88], the simulation over a barrier is achieved by constraining  $\xi$  at different values among a number of predefined windows, and sampling the orthogonal degrees of freedom of the system. In this way, the force on the frozen reaction coordinate could be estimated, and the resulting mean force represents the negative of the derivative of the free energy with respect to  $\xi$ . By integration of the mean force, it is possible to evaluate the potential of mean force, PMF, which corresponds to the free energy profile as function of the reaction coordinate. In umbrella sampling [89,90], a bias potential is applied to the system for ensuring the sampling along the reaction coordinate. Contrary to the Blue Moon approach, here the reaction coordinate is not frozen, but only harmonically restrained to target values by application of the bias in a series of windows. On the other hand, metadynamics works by adding a history-dependent potential, which is built as a sum of Gaussians, whose role consists in discouraging the system to explore regions already visited in the CV space, accelerating in this way rare events. The free energy surface can be easily reconstructed as the opposite of the sum of all Gaussians [91,92].

#### *Approaches based on tempering*

These approaches are generally preferred when the choice of appropriate CVs for describing a transition is not straightforward. This is the case of protein

conformational analysis or protein folding studies. Methods belonging to this class, exploit the increasing of temperature of the system to overcome barriers during MD simulations [84]. The strong dependence of temperature with the rate of a barrier crossing is stated by the Arrhenius equation. In one of these approaches, the temperature of a system is dramatically increased during sampling to easily explore high energy regions. This step is followed by a cooling phase, in which the temperature is therefore decreased to pull down the simulation toward a local energy minimum. This approach, better known as *simulated annealing*, is successfully used to address many optimization problems [93]. A critical parameter of this technique, is represented by the speed of the cooling phase: the probability to reach a global minimum is directly linked to the decreasing of the speed. An alternative approach is the *parallel tempering* [94], in which, instead of changing the temperature of a single system, several replicas of the original system of interest, are simulated at different temperatures. High temperature systems are able to explore large volumes of phase space, whereas low temperature replicas could sample low energy regions and local minima. During the simulation, the replicas are allowed to swap configurations if an acceptance probability criteria is satisfied [95]. One of the advantages of parallel tempering is that, this technique could be efficiently carried out on CPU clusters, running simultaneously all the replicas in parallel.

### **1.2.1 Application to the Peptidyl-Prolyl *cis-trans* Isomerase Pin1**


Here I briefly present the peptidyl-prolyl *cis-trans* isomerase Pin1, an enzyme for which the application of enhanced sampling techniques has been shown to successfully shed light on the catalytic mechanism. Details are reported in Chapter 4.

The peptidyl-prolyl isomerase (PPIase) Pin1 is a member of the parvulin sub-family, which specifically recognizes phospho(p)-Ser/Thr-Pro proteins and catalyzes the *cis-trans* isomerization of the proline amide bond [96-98]. This *cis-trans* interconversion is a rather slow spontaneous process, whose rate is dramatically

accelerated by the presence of Pin1 by several order of magnitude in the timescale of seconds [99,100].

Several studies have established that alterations of Pin1 functionality are coupled with pathological conditions outbreaks. Pin1 is over-expressed in human cancers, including breast, lung and prostate, playing a critical role in oncogenesis, and seems to be also implicated in neurological disorders such as Alzheimer's disease, asthma and inflammation [101-103]. Despite a great interest has arisen on this enzyme, the catalytic mechanism has long been debated. In this context, two models of catalysis have been hypothesized: a covalent [97], and a non-covalent [104] model of reaction, each of them supported by mutagenesis data. Recently, theoretical studies have provided new valuable insights for unravelling the reaction mechanism [105,106], but unanswered questions still remain and further in-depth investigations are required to clarify the catalytic process.






---

## **Chapter 2.**

# **Theoretical Background**

---



## 2.1 System Representation

In this section, I will give a brief overview of the basic principles of (classical) molecular mechanics and quantum mechanics, showing the differences of the two approaches in the treatments of molecular systems, and their applicability.

### 2.1.1 Molecular Mechanics

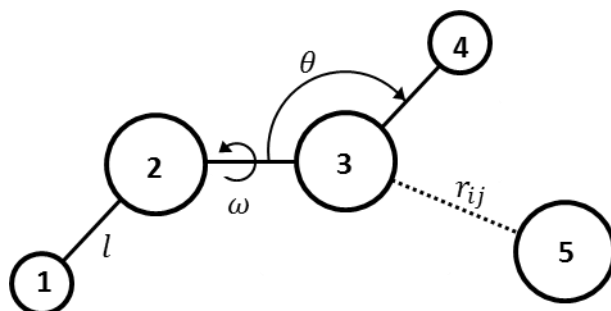
Molecular mechanics (MM) could be expressed as a mathematical model which considers a molecule as a simple collection of balls (atoms) held together by springs (bonds) [107]. Within such a representation, the energy of a molecular system is going to change as a result of the geometry deformation: stretching or bending of the springs, for example, can push away the system from “standard values” (the equilibrium). The MM energy expression is a simple algebraic equation, characterized by constants obtained either from experimental data or *ab initio* calculations. An important assumption of MM is the transferability of the parameters. The energy contribution related to a particular molecular motion, for example the stretching between two carbon atoms defining a single bond, will be the same from one molecule to the next [108]. This leads to a very simple energy calculation and allows the application to very large molecular systems. On the contrary, *ab initio* methods can be applied just for modelling limited size systems, because of the high computational resources required. Obviously, the performance of MM approach, strongly depends on several factors: 1) the functional form of the energy equation, 2) the constants used in this equation, 3) the way in which they are parameterized, and 4) the ability of the user in applying this technique for solving biological tasks [109].

The mathematical expression of the energy function and the parameters in it, take the name of *forcefield*. Commonly used forcefields, could differ on the number of terms in the energy expression, and on the constant parameters used. Since no electrons are explicitly taken into account, the electronic processes, like bond breaking, cannot be simulated. A forcefield is generally composed by bonded and non-bonded terms. Bonded terms allow the evaluation of energy penalties associated with deviations of bonds, angles, and torsions, from equilibrium values, whereas the non-bonded ones, describe interactions between different molecules, or pairs of atoms not directly connected and separated by at least three bonds. These terms are illustrated in Figure 1.1. A general functional form of the forcefield, which describes the potential energy ( $\mathcal{V}$ ) of a molecule, is shown in the following equation [108]:

$$\begin{aligned} \mathcal{V} = & \sum_{bonds} \frac{k_i}{2} (l_i - l_{i,0})^2 + \sum_{angles} \frac{k_i}{2} (\theta_i - \theta_{i,0})^2 + \sum_{torsions} \frac{V_n}{2} (1 + \cos(n\omega - \gamma)) \\ & + \sum_{i=1}^N \sum_{j=i+1}^N \left( 4\varepsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right) \end{aligned} \quad (2.1)$$

The first term in (2.1), represents the *bond stretching* term, a harmonic potential term which describes the increasing in energy due to the deviation of the bond length, between two atoms, from the equilibrium value  $l_{i,0}$ . The second term defines the *angle bending*, modelled again by means of a harmonic potential. The third term is the *torsional potential*, which defines the rotation around a bond, and usually described by a cosine expansion. The latter term, represents the *non-bonded* interactions, which are modelled using the Lennard-Jones 12/6 potential for van der Waals interactions, and a Coulomb potential, for electrostatic interactions. The charges,  $q_i$  and  $q_j$ , in Coulomb's law, are the partial atomic charges, usually centered on atom nuclei, designed to reproduce the electrostatic properties of the molecule. Furthermore, the Lennard-Jones 12/10 potential is also widely used for hydrogen bond interactions. Forcefields could use the same functional form, as the one shown before, but different constants (see  $k_i$ ,  $V_n$ ,  $\sigma_{ij}$ , in Equation (2.1),

parameterized in order to reproduce experimental data), or presenting a totally different functional form.



**Figure 1.1** Representation of the bonded (solid line) and non-bonded terms (dashed) in a classical forcefield.

A fundamental aspect that has to be considered, is that forcefields are empirical and there is no a universal functional form. In particular, some forcefields could perform better than others, and their choice strongly depend on the system that has to be analysed [108,109]. A forcefield which is parameterized against a specific class of molecules, like nucleotides, cannot provide a reasonable description of proteins, and vice versa. Moreover, some simplified forcefields, the so called *united atoms*, do not include explicit representation of nonpolar hydrogen atoms. These approximated models are generally used for speeding up highly demanding conformational sampling, such as protein folding, or protein-protein binding [110]. Several forcefields have been developed [109]. Amber (Assisted Model Building with Energy Refinement) forcefield is one of the most applied for proteins and nucleic acids.

### 2.1.1.1 The Amber Forcefield

In the last decades, a great effort was spent in the development of forcefields resulting from a compromise between accuracy and the limited computational resources available at that time.

In this context, in 1984, Weiner et al. developed a united atom forcefield for simulation of nucleic acids and proteins, incorporated in the amber software package [111]. In the proposed model, bond lengths, and angles parameters, were taken from crystal structures and adapted to match the normal mode frequencies for a number of peptide fragments. Torsion parameters, were adapted to match torsional barriers evaluated by means of experimental measurements or quantum mechanical calculations. A Hartree-Fock STO-3G level of theory, was used to derive the charges, while van der Waals parameters were adapted from Hagler et al. parameters [112]. The impossibility of a suitable description of the quadrupolar charge distribution of benzene, and therefore of  $\pi$ - $\pi$  and  $\pi$ -cation interactions, led to move to an all-atom approach. In 1986, an all-atom extension of the CH united atom forcefield was published, including parameters derived from gas phase simulations [113]. Advances in computational resources made possible the development of the *ff94*, a new MM forcefield for simulations of proteins, nucleic acids, and organic molecules [114]. In the *ff94*, a new set of charges, determined by means of a Hartree-Fock 6-31G\* level of theory and RESP (restrained electrostatic potential) fitting, are able to accurately reproduce interaction energies, conformational energies, and free energy of solvation of small molecules. New van der Waals parameters obtained from liquid simulations, were also derived. These improved parameters, made the Lennard-Jones 12/10 potential no longer necessary for the hydrogen bonds description. The bonded parameters were also modified to reproduce vibrational frequency data. Improvements in the description of longer-range effects, were then provided in the *ff96* [115] and *ff99* [116] forcefields. An important aspect of these forcefields, was an inadequate parameterization of backbone dihedral terms. The *ff94* dihedral parameters, for example, were derived using few glycine and alanine dipeptide conformers in the fitting procedure. New improved phi/psi dihedral terms were included in the *ff99SB* forcefield [117], by fitting the energies of multiple conformations of glycine and alanine tetrapeptides, exploiting high level quantum mechanical simulations. In 2010, optimized side-chains torsions parameters for isoleucine, leucine, aspartate, and asparagine, were provided in an extension version of the previous forcefield, the *ff99SBildn* [118]. Recently, the *ff12SB* offers side-chain corrections for lysine, arginine, glutamate, glutamine, methionine, serine, threonine, valine, tryptophan, cysteine, phenylalanine,

tyrosine, and histidine, improving in this way, the reproduction of experimental geometries [119]. Although the high accuracy achieved in molecular modelling of biological macromolecules, the limitations of these forcefields consist in a poor treatment of the electrostatic. Traditional forcefields use fixed point charges centered on atoms, therefore they may not accurately describe varying in the electrostatic properties of the environment [120]. Improvements in the fixed point charge models were provided by Duan et al. [121] in the development of the *ff03* forcefield. In particular, the authors applied a continuum solvent model to calculate the electrostatic potentials in organic solvent for the derivation of partial charges. Moreover, a great effort was also made in the development of forcefields taking into account polarization effects, as in the case of the Amber *ff02* [122], in which an additional polarization term is included in the energy function.

## 2.1.2 Quantum Mechanics

The description of the electronic behaviour of atoms and molecules, and their reactivity, is one of the application of quantum mechanics (QM). The main goal of QM methods, is to determine the electronic structure, the probability distribution of electrons in chemical systems. However QM equations can be exactly solved just for two interacting particles, and several approximations have been introduced for many electron problems [108,123]. The electronic structure is determined by solving the Schrödinger equation, associated with the electronic molecular Hamiltonian. The time-independent formulation of the Schrödinger equation is given by:

$$\hat{H}\Psi = E\Psi \quad (2.2)$$

where  $\hat{H}$  is the Hamiltonian operator, which acts upon the wave function  $\Psi$ , and returns  $E$ , the energy of the system. In mathematical notation, Equation (2.2) represents an eigen equation, where  $\Psi$  is the eigenfunction, and  $E$  the eigenvalue. Importantly, the product of the wave function  $\Psi$  with its complex conjugate (i.e.,  $|\Psi^* \Psi|$ ) has units of probability density. Thus, the probability of finding a particle

within some region of multi-dimensional space is equal to the integral of  $|\Psi|^2$  over that region of space. The Hamiltonian operator  $\hat{H}$ , can be expressed as:

$$\hat{H} = - \sum_i^{\text{particles}} \frac{\nabla_i^2}{2m_i} + \sum_{i<j}^{\text{particles}} \sum \frac{q_i q_j}{r_{ij}} \quad (2.3)$$

where,

$$\nabla_i^2 = \frac{\delta^2}{\delta x_i^2} + \frac{\delta^2}{\delta y_i^2} + \frac{\delta^2}{\delta z_i^2} \quad (2.4)$$

$\nabla_i^2$  is the Laplacian operator acting on particles  $i$  (nuclei and electrons) with mass  $m_i$  and charge  $q_i$ . Distances between particles are defined by  $r_{ij}$ . The first term in (2.3) represents the kinetic energy, while the second a potential term, the Coulombic attraction or repulsion between particles. By the way, this formulation could be analytically solved just for hydrogen atom, and an application to molecules, requires approximations, simplification of the electronic description.

The Born-Oppenheimer (BO) approximation [124] assumes that the motions of nuclei and electrons can be decoupled, because, under typical physical condition, nuclei move much slowly than electrons (neutrons are about 1800 times more massive than electrons). Therefore, only the motions of electrons are considered, whereas nuclei are considered fixed. By means of the BO approximation, some of the terms coupling electrons and nuclei are omitted in the Hamiltonian:

$$\hat{H}_{\text{BO}} = - \sum_i^{\text{electrons}} \frac{\nabla_i^2}{2} - \sum_i^{\text{nuclei}} \sum_j^{\text{electrons}} \frac{Z_i}{r_{ij}} + \sum_{i<j}^{\text{electrons}} \sum \frac{1}{r_{ij}} \quad (2.5)$$

The Hamiltonian includes just the kinetic energy of electrons, the coupling term between nuclei and electrons (attraction), the repulsion between electrons, while the term which describes the repulsion between nuclei is added at the end of the calculation. Two QM approaches are generally used: *ab initio* and *semi-empirical*.

Both of them rely on the BO approximations, but different types of approximations are used.

*Ab initio* approach, computationally more demanding, implies a non-empirical solution of the time-independent Schrödinger equation: it doesn't make use of any experimentally derived parameter in solving the Schrödinger equation for electrons. During this expensive process, the molecular geometry is considered as a fixed parameter. Once the optimal electronic wavefunction is determined, one can evaluate the gradient on the nuclei, which represents the derivative of the total energy with respect to nuclear positions. The positions of the nuclei can therefore be updated, until the process reaches the convergence. The most common *ab initio* method, is the Hartree-Fock (HF), in which the Coulombic repulsion between electrons is taken into account in an averaged way (mean field approximation). This is a variational calculation: the resulting approximate energies, which are expressed in terms of the system's wavefunction, are equal or greater than the exact energy, tending to a limiting value, known as HF limit. One can start a calculation with the HF method, and then correct the missing electronic calculation using post-HF methods, like Møller-Plesset perturbation theory or Coupled Cluster.

Density Functional Theory (DFT), is usually classified as *ab initio*, even though functionals derived from empirical data, are used [125]. DFT is based on the idea that the energy of a molecule, and all the observables, can be determined by the electron density  $\rho(\mathbf{r})$ , instead of a wavefunction. DFT has many advantages compared to other *ab initio* methods. It is less computationally expensive. In a system of  $N$  electrons and  $M$  nuclei,  $\psi$  results to be a function of  $4N + 3M$  degrees of freedom:  $3N$  spatial coordinates and  $N$  spin coordinates of electrons, and  $3M$  spatial coordinates of nuclei. However, ignoring the spin and considering the BO approximation,  $\psi$  results to be function of  $3N$  coordinates. On the contrary, the density  $\rho(\mathbf{r})$  is just a function of 3 coordinates, the spatial coordinates  $x$ ,  $y$  and  $z$ , reducing the computational costs, and providing similar accuracy than the other approaches.

Other methods called *semi-empirical* replace costly integrals in HF calculations, with empirical data. To correct these approximations, additional empirical terms are introduced in the Hamiltonian. These methods, result to be faster than *ab initio*, but the obtained results could be not accurate if these approaches are



applied to a set of molecules structurally different from the one used to derive the parameters [123].

## 2.2 Sampling of Microstates

After a brief introduction on the fundamentals of *statistical mechanics*, I will provide, in this section, a description of the molecular simulation methods, *Monte Carlo* and *Molecular Dynamics*. These are computational approaches which allow to evaluate macroscopic properties of the system of interest, from microscopic information (as the distribution of microscopic states).

### 2.2.1 Statistical Mechanics

The ultimate goal of statistical mechanics, is to model and predict the thermodynamic properties of materials, from the structures of the atoms and molecules of which they are composed. In general, such properties depend upon the position  $\mathbf{x}$ , and momenta  $\mathbf{p}$  of the particles composing the system. Hence, the instantaneous value of a property  $f$ , is given by  $f(\mathbf{x}(t), \mathbf{p}(t))$ . This value fluctuates over time, as a result of the interactions between the particles. Therefore, the experimental measure of such property, is an average of over time  $f_{\text{ave}}$ , known as *time average*. If the measurement is made over a time approaching infinity,  $f_{\text{ave}}$  is given by [108]:

$$f_{\text{ave}} = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{t=0}^T f(\mathbf{x}(t), \mathbf{p}(t)) dt \quad (2.6)$$

The calculation of  $f_{\text{ave}}$  requires to simulate the dynamic behaviour of the system. However, because of the strong dependence of time averaged property to the initial

configuration of a system (considering that a simulation could be performed in a finite time), time average calculations are virtually impossible to carry out. A way to overcome this task, is to replace the time average of a single system by an *ensemble average* of a large collection of systems. Therefore, at a given instance of time, a collection of large number of systems (having different microstates) are considered: for a sufficiently long simulation, the observed property of a single system over a period of time, is the same as the average over all microstates (ergodic hypothesis). The average of a property  $f$  over all replications of the ensemble generated by the simulation,  $\langle f \rangle$ , could be expressed as [108]:

$$\langle f \rangle = \iint d\mathbf{p} d\mathbf{x} f(\mathbf{p}, \mathbf{x}) \rho(\mathbf{p}, \mathbf{x}) \quad (2.7)$$

Therefore,  $\langle f \rangle$  could be determined by integrating over all possible configurations of a system. In (2.7),  $\rho(\mathbf{p}, \mathbf{x})$  represents the probability of finding a configuration with position  $\mathbf{x}$ , and momenta  $\mathbf{p}$ , which is referred to as probability density of the ensemble. Under conditions of constant number of particles  $N$ , volume  $V$ , and temperature  $T$  (*canonical ensemble*),  $\rho(\mathbf{p}, \mathbf{x})$  is given by [108]:

$$\rho_{NVT}(\mathbf{p}, \mathbf{x}) = \frac{e^{-\frac{E(\mathbf{p}, \mathbf{x})}{k_B T}}}{Q} \quad (2.8)$$

where,  $E(\mathbf{p}, \mathbf{x})$  is the total energy (the sum of the kinetic energy of the system,  $K(\mathbf{p})$ , and the potential energy  $\mathcal{V}(\mathbf{x})$ ),  $Q$  the partition function,  $k_B$  the Boltzmann's constant ( $k_B = 1.38 \times 10^{-23}$  J/K =  $1.38 \times 10^{-16}$  erg/K), and  $T$  the absolute temperature of the system. The partition function is commonly written in terms of the Hamiltonian  $H$  [108]:

$$Q_{NVT} = \frac{1}{N! h^{3N}} \iint d\mathbf{p} d\mathbf{x} e^{-\frac{H(\mathbf{p}, \mathbf{x})}{k_B T}} \quad (2.9)$$

In (2.9),  $N!$  is used to account for the indistinguishability of the particles, while the factor  $h^{3N}$  represents the elementary volume of a microstate.

Computing the equilibrium properties of classical many-body systems, is the main purpose of molecular simulations. *Monte Carlo* (MC) and *Molecular Dynamics* (MD) represent the two most common simulation techniques used in molecular modelling. These simulation techniques consider small replications of the macroscopic system, allowing to predict structural and thermodynamic properties with a feasible amount of computation.

## 2.2.2 Monte Carlo

MC methods generate configurations of the system in a stochastic way, and new configurations are chosen according to a statistical distribution. In a MC simulation, therefore each configuration depends only upon its predecessor. In a classical MC approach, the following scheme is proposed [108,123]:

1. Randomly choose an initial set of atomic positions  $\mathbf{x}_1$ .
2. Calculation of the potential energy  $\mathcal{V}_1$  for the generated configuration.
3. Making a random trial in the microstate. This step corresponds to a perturbation phase leading to a new configuration  $\mathbf{x}_2$ .
4. Compute the potential energy  $\mathcal{V}_2$  of the new arrangement of the atoms.
5. Keep  $\mathbf{x}_2$  if the acceptance criteria is satisfied. In the Metropolis scheme [126], the acceptance of a new configuration is based on a Boltzmann-weighted probability. The probability for accepting  $\mathbf{x}_2$  could be expressed as:

$$p = \min \left[ 1, \frac{e^{-\mathcal{V}_2/k_B T}}{e^{-\mathcal{V}_1/k_B T}} \right] \quad (2.10)$$

If the energy  $\mathcal{V}_2$  related to the new configuration, is lower than  $\mathcal{V}_1$ ,  $\mathbf{x}_2$  is always accepted. On the other hand, if  $\mathcal{V}_2$  results to be higher than the energy of the previous point  $\mathcal{V}_1$ ,  $\mathbf{x}_2$  is kept if  $p \geq \text{rand}(0, 1)$ , a random number between 0 and 1. In the case in which the new configuration is rejected, a perturbation

phase is attempted again, starting from  $\mathbf{x}_1$ . This strong dependence of a new point with the preceding one, is called *Markov chain*.

6. Continue the process and collecting data for computing the desired thermodynamic property of the analysed system.

The moves which characterize the perturbation phase (step 3), play a crucial role in the efficiency of a MC calculation. If the move size is too small, the sampling proceeds very slowly and the calculation requires high computational resources. On the contrary, if it tends to become too large, the number of rejected moves will grow, compromising the efficiency. Another important aspect is that, in contrast to MD, MC simulations cannot be used for calculating time-dependent quantities, such as transport coefficient or viscosity. By the way, MC represents a useful method for significantly explore different areas of the phase space (which consists of all possible values of position and momentum variables), allowing to sample energy states separated by high barriers. The ability of crossing barriers is not a feature of (*unbiased*) MD simulations, which, however, are suitable for exploring the local phase space.

### 2.2.3 Molecular Dynamics

MD simulations represent a powerful tool for following the temporal evolution of a system in the phase space, providing atomic details of the motions of a many-body system, and allowing the evaluation of thermodynamic properties. I will consider, in this section, MD simulations based on classical mechanics, then making usage of empirical force fields (Section 2.1.1.1) for describing the potential energy of the system: an approach referred as *classical* MD. In MD, the evolution of the system is described by Newton's equation of motion, which states that:

$$\mathbf{F}_i = m_i \frac{d^2 \mathbf{x}_i}{dt^2} = - \frac{\delta \mathcal{V}}{\delta \mathbf{x}_i} \quad (2.11)$$

where  $F_i$  is the force acting on atom  $i$  with position  $\mathbf{x}_i$ , mass  $m_i$ , and  $\mathcal{V}$  represents the potential energy of the system of interest. Computing the classical trajectory exactly is a challenging task, because it requires solving a  $3N$  coupled  $2^{nd}$  order differential equations, with  $N$  representing the total number of atoms of the system. It is too costly for large  $N$ . Several approximated methods, based on time discretization, are required to overcome this difficulty. The *finite difference methods*, allow to integrate the equations of motion in different stages, separated in time by a fixed time  $\delta t$ . This is known as *timestep*, which should be properly chosen to avoid discretization errors. A  $\delta t$  too large cause atoms moving too far along a trajectory, leading to an inaccurate simulation of motions. On the other side, a  $\delta t$  too small, increases the number of iterations required for acquiring the trajectory. The timestep of a MD simulation, is dictated by the highest frequency motions present in the system, like bond vibrations, which, moreover, are usually of less interest than the lower frequency modes. A common way to increase the timestep without altering the accuracy of the simulation, is the application of constraints to “fix” some internal coordinates of the system. A widely used procedure is the *SHAKE algorithm* [127], which allows the introduction of constraints on some degrees of freedom of the system. The main advantage achieved with SHAKE, is represented by the possibility of constraining intramolecular bond lengths, like C-H, reducing the complexity of the simulation. Bond vibrations are high frequency motions, and therefore, freezing bond lengths allows the usage of a larger timestep, speeding up the calculation. In classical simulations of molecules, typical timesteps are in the order of femtoseconds.

Several algorithms for integrating the equation of motions were developed. They approximate position and dynamic properties, like velocities and accelerations, as a Taylor expansion. Among them, the *Verlet integration scheme* [128], exploits positions and accelerations from the previous step at time  $t$ , to calculate the new positions at time  $t + \delta t$ :

$$\mathbf{x}(t + \delta t) = \mathbf{x}(t) + \delta t \mathbf{v}(t) + \frac{1}{2} \delta t^2 \mathbf{a}(t) + \dots \quad (2.12)$$

$$\mathbf{x}(t - \delta t) = \mathbf{x}(t) - \delta t \mathbf{v}(t) + \frac{1}{2} \delta t^2 \mathbf{a}(t) - \dots \quad (2.13)$$

From (2.12) and (2.13) we obtain the relation:

$$\mathbf{x}(t + \delta t) = 2\mathbf{x}(t) - \mathbf{x}(t - \delta t) + \delta t^2 \mathbf{a}(t) \quad (2.14)$$

Basically, for each particle of the system, the subsequent position  $\mathbf{x}(t + \delta t)$  is determined by the current position  $\mathbf{x}(t)$ , by the previous one  $\mathbf{x}(t - \delta t)$ , and by the acceleration  $\mathbf{a}(t)$ , the latter computing from the forces on the particle: no explicit velocities are taken into account. This procedure could be advantageous just when one is interested in the evaluation of a property which is independent of momentum. Several variations of the Verlet algorithm have been developed. A commonly used variation, is the *velocity Verlet* scheme [129], which uses the Taylor expansion truncated beyond the quadratic term for the coordinates:

$$\mathbf{x}(t + \delta t) = \mathbf{x}(t) + \delta t \mathbf{v}(t) + \frac{1}{2} \delta t^2 \mathbf{a}(t) \quad (2.15)$$

$$\mathbf{v}(t + \delta t) = \mathbf{v}(t) + \frac{1}{2} \delta t [\mathbf{a}(t) + \mathbf{a}(t + \delta t)] \quad (2.16)$$

In the *velocity* method, positions, velocities and accelerations at time  $t + \delta t$  are obtained from the same quantities at time  $t$ .

If the total number of atoms  $N$ , the volume  $V$ , and the total energy of the system  $E$ , are kept constant during the simulation, MD are said to be performed in the *microcanonical ensemble* (NVE). NVE is useful for exploring the constant energy surface of the conformational space, however most of the biological events occur at constant pressure and temperature. In particular, the *canonical ensemble* (NVT), is obtained by coupling the system to a heat-bath, while the *isothermal-isobaric ensemble* (NPT) by introducing a barostat.

## 2.2.4 Free Energy Calculation

The Helmholtz free energy  $A$ , can be expressed in terms of the partition function  $Q$ , as follows [130]:

$$A = -\beta^{-1} \ln Q_{NVT} \quad (2.17)$$

where  $\beta = 1/k_B T$ . (2.17) represents a connection between thermodynamics and statistical mechanics in the canonical ensemble NVT. The evaluation of the partition function  $Q_{NVT}$  is a very difficult task. However, one is interested in the estimation of the free energy differences  $\Delta A$ , between a reference (0) and a target (1) system state, which can be expressed by the partition functions  $Q_0$  and  $Q_1$ :

$$\Delta A = -\beta^{-1} \ln \frac{Q_1}{Q_0} \quad (2.18)$$

If the masses of the particles in the two systems are the same, (2.18) can be equivalently written considering the configurational integrals ratio  $Z_1/Z_0$ :

$$\Delta A = -\beta^{-1} \ln \frac{Z_1}{Z_0} \quad (2.19)$$

One way to determine  $\Delta A$ , consists in the estimation of the appropriate probability densities of the two states. Assuming that the system 0 can be transformed to system 1 through the modification of a parameter  $\xi$  (e.g. a generalized coordinate, as a torsion, or a distance), the probability density function for system 0 can be written as:

$$P_0 = P(\xi_0) = \frac{\int \exp(-\beta H) \delta(\xi - \xi_0) d\mathbf{x} d\mathbf{p}_x}{\mathcal{N}} = \frac{Q_0}{\mathcal{N}} \quad (2.20)$$

where  $\mathcal{N}$  is a normalization constant. In (2.20) the Hamiltonian  $H$ ,  $\beta$ , or  $\mathbf{x}$ ,  $\mathbf{p}_x$ , could be functions of the parameter  $\xi$ . In particular, (2.20) represents the connection between probability density and partition function. The probability density function for system 1 could be written by replacing the subscript 0 to 1. By a combination of (2.18) and (2.20) it is possible to obtain:

$$\Delta A = -\beta^{-1} \ln \frac{P_1}{P_0} \quad (2.21)$$

The probability distribution function  $P(\xi)$ , for the range comprised between  $\xi_0$  and  $\xi_1$ , could be obtained by computer simulations as a histogram. This allows the evaluation of the ratio  $P_1/P_0$ , and hence the estimation of  $\Delta A$ .

### 2.2.4.1 Enhanced Sampling

Traditional MD simulations are suitable to perform an exhaustive sampling of the local phase space. However, these approaches are limited by the short timescales accessible with computational methods and resources currently available. Many biological systems are characterized by the presence of states separated by high energy barriers. Sampling these states by crossing large barriers, represents a very challenging task for standard simulation methods. In classical computer simulations, the volume in phase space covered during the sampling is not sufficient to provide a reliable estimation of the statistical averages of the property of interest. Therefore, a direct application of these methods might not allow a correct estimation of free energies. Advanced strategies are needed to guarantee a suitable exploration of the phase space regions, including rare events, that are important for free energy calculations. Several methods were developed, based on tempering, on the modification of the energy expression for reducing the barriers, or on the restriction of the sampling space by constraining degrees of freedom, with the exception of the reaction coordinates describing the rare event. All these approaches, referred to as



enhanced sampling methods, have been shown to accelerate the sampling of configuration space [83,84,131].

Among them, the *umbrella sampling* method [89,90], based on the addition of a biased potential to the potential energy function of a system in order to sample high energy states, has been widely used to calculate the free energies in chemical processes. Umbrella sampling allows the calculation of the free energy along a reaction coordinate, also known as potential of mean force (PMF).

### 2.2.4.1.1 Umbrella Sampling

In umbrella sampling a biased potential  $\omega_i$  is applied to a given reaction coordinate  $\xi$ , known as collective variable CV, opportunely chosen to describe the transition from states that are difficult to sample. The reaction coordinate is therefore harmonically restrained to a target value during the simulation. Usually, to ensure a rigorous sampling along the whole reaction coordinate  $\xi$ , umbrella sampling simulations are carried out in a series of different windows  $i$ , which are run in independent simulations. The bias potential  $\omega_i$ , takes the form of a classical harmonic potential:

$$\omega_i(\xi) = \frac{k_i}{2} (\xi - \xi_i^{ref})^2 \quad (2.22)$$

where  $\xi_i^{ref}$  is the  $\xi$  reference value of window  $i$ , kept by the bias  $\omega_i(\xi)$ , while  $k_i$  represents the force constant, thus the strength of the bias. An optimal choice of  $k_i$ , is necessary for achieving a correct overlap between the distributions of the different windows. In particular, too large  $k_i$ , will cause narrow distributions, while a too small constant will not sufficiently bias the simulation over the barriers. An optimal overlap between neighbouring windows is necessary to guarantee that a continuous energy function can later be derived from these simulations.

With the introduction of the bias, the effective energy of the system  $E^b(\mathbf{x})$  is given by:

$$E^b(\mathbf{x}) = E^u(\mathbf{x}) + \omega_i(\xi) \quad (2.23)$$

where the additional term  $\omega_i(\xi)$  is added to the original unbiased potential  $E^u(\mathbf{x})$ . Using the biased potential, the simulation of the system provides a biased distribution  $P_i^b(\xi)$  for the window  $i$ , along the reaction coordinate. In the hypothesis that the biased system is ergodic:

$$P_i^b(\xi) = \frac{\int \exp\{-\beta[E(\mathbf{x}) + \omega_i(\xi'(\mathbf{x}))]\} \delta(\xi'(\mathbf{x}) - \xi) d\mathbf{x}}{\int \exp[-\beta E(\mathbf{x})] d\mathbf{x}} \quad (2.24)$$

where  $\omega_i(\xi'(\mathbf{x}))$  is the bias potential, and  $\beta = 1/k_B T$ . The unbiased distribution  $P_i^u(\xi)$ , which is necessary to derive the unbiased free energy  $A_i(\xi)$ , is obtained as:

$$P_i^u(\xi) = \frac{\int \exp[-\beta E(\mathbf{x})] \delta(\xi'(\mathbf{x}) - \xi) d\mathbf{x}}{\int \exp[-\beta E(\mathbf{x})] d\mathbf{x}} \quad (2.25)$$

The relation between (2.24) and (2.25) can be written as:

$$P_i^u(\xi) = P_i^b(\xi) \exp[\beta \omega_i(\xi)] \langle \exp[-\beta \omega_i(\xi)] \rangle \quad (2.26)$$

The unbiased free energy could be easily derived as:

$$A_i(\xi) = -\left(\frac{1}{\beta}\right) \ln P_i^b(\xi) - \omega_i(\xi) + F_i \quad (2.27)$$

here,  $F_i$  is a normalizing constant, numerically estimated by means of the *Weighted Histogram Analysis Method* (WHAM) [132]. WHAM evaluates the global unbiased distribution  $P^u(\xi)$ , as a weighted sum of the unbiased probability  $P_i^u$ , over all the windows  $i$ :

$$P^u(\xi) = \sum_i^{\text{windows}} p_i(\xi) P_i^u(\xi) \quad (2.28)$$

In (2.28),  $p_i$  represent the weights, chosen so that the statistical error of  $P^u(\xi)$  is minimized, under the condition  $\sum_i p_i = 1$ :

$$p_i = \frac{a_i}{\sum_j a_j}, \text{ with } a_i = N_i \exp[-\beta \omega_i(\xi) + \beta F_i] \quad (2.29)$$

with  $N_i$  equals to the number of steps for window  $i$ . Finally, the equation related to the computation of  $F_i$  is given as follows:

$$F_i = -\left(\frac{1}{\beta}\right) \int P^u(\xi) \exp[-\beta \omega_i(\xi)] d\xi \quad (2.30)$$

Therefore, because  $P^u(\xi)$  is essential for deriving  $F_i$ , and  $F_i$  is also involved in the evaluation of  $P^u(\xi)$ , (2.28) and (2.30) have to be iterated until convergence. Once it has been achieved,  $F_i$  could be used for the calculation of the free energy by means of (2.27).

## 2.3 Approximate Methods

Simulation methods described before, are suitable approaches for sampling the conformational space, and hence, for obtaining statistical distributions of the system of interest. However, considering the computational resources required for carrying out such simulations, faster techniques, representing a crude approximation of the real dynamics of the system, have been developed. Among them, here, I will discuss the theoretical basis of *molecular docking* methods, based in particular on genetic algorithm, and *MM-PBSA* approaches. These “approximate methods” represent suitable tools in drug discovery, mainly used for modelling protein-ligand binding affinities.

### 2.3.1 Molecular Docking

One of the most important aspect in drug design, is the ability to predict the interactions, and therefore, the affinity of binding between small molecules and a biological target [29]. Computational chemistry tools widely applied for this purpose, are referred to as *molecular docking* techniques.

A molecular docking protocol is characterized by two steps, a search strategy, and a scoring phase. In the first step, a search algorithm is applied to sample the configurational space of the candidate pose. During this phase, all possible binding modes between ligand and receptor are sampled, and subsequently evaluated by a scoring function. Usually, only the conformational space of the ligand is considered, assuming the receptor as rigid body. The high number of degrees of freedom during the searching phase, in fact, increases the computational costs, but also affects the success of the optimization algorithm [133]. Although several docking algorithms take into account protein flexibility, it has been shown the difficulties encountered in obtaining an exhaustive sampling of the protein conformational states [133]. However, when a protein target is derived by homology modelling, because of the absence of a 3D crystal structure, protein flexibility has to be considered in a docking protocol in order to obtain reliable results. One strategy, is to dock ligands in an ensemble of pre-generated conformations of the target, treating them as rigid bodies [134,135].

Among the plethora of search algorithms currently available, *genetic algorithms* (GA) represent the most popular in docking procedure. GA require the random generation of a first population of *individuals*, which are points in the search space. Individuals are simple binary strings, where all the information regarding the parameters that have to be optimized (usually, the degrees of freedom of the ligand) are encoded as genes in chromosome [35,136]. The evolution of the population is obtained via two genetic operators, *crossover*, and *mutation*. The crossover operator, exchanges set of genes from one parent individual to another, while the mutation operator, randomly changes the values of genes. By means of the genetic

operators, a new generation of individuals could be derived from the preceding. An optimisation step is then followed by a scoring phase. During the latter phase, a fitness, based on the application of a scoring function, is assigned to each individual. Several optimisation-scoring cycles are generally required for achieving a docking solution.

Autodock [137], a popular open source docking software, uses a *Lamarckian genetic algorithm* (LGA), an hybrid algorithm which combines the standard GA as global optimiser, with energy minimisation for a local search. For the fitting evaluation, Autodock exploits an empirical scoring function composed by five terms. In particular it is characterized by: a Lennard-Jones 12/6 potential accounting for dispersion/repulsion interactions; a Lennard-Jones 12/10 for hydrogen bond; a Coulomb potential for electrostatic interactions; a term for estimate the entropic contribution in binding: this term is proportional to the number of rotatable bonds in ligand, accounting for the loss of degree of freedom upon binding; a desolvation term, which is function of the solvent accessible surfaces of both ligand, and protein. Protein and ligand parameters are taken from the Amber forcefield.

### 2.3.2 MM-GB(PB)SA

Computational methods combining molecular mechanics and implicit solvation models, such as *Molecular Mechanics/Poisson–Boltzmann Surface Area* (MM-PBSA) [138] and *Molecular Mechanics/Generalized Born Surface Area* (MM-GBSA) [139], are widely used for free energy of binding calculation. These methods are also referred to as *endpoint* methods, because for computing the binding free energy only the starting and final states (bound and unbound) are required. They are less computational expensive than more rigorous methods exploited for the same purpose, like free energy perturbation (FEP), hence widely applied for protein – ligand complex systems. Considering a ligand (L), a protein (P), and related complex (PL), the binding free energy is given by:

$$\Delta G_{\text{bind}} = G^{\text{PL}} - G^{\text{P}} - G^{\text{L}} \quad (2.31)$$

where the free energy contribution of each molecular system X (PL, P, or L), is obtained as:

$$G^{\text{X}} = E_{\text{MM}}^{\text{X}} + G_{\text{solv}}^{\text{X}} - TS^{\text{X}} \quad (2.32)$$

In (2.32),  $E_{\text{MM}}^{\text{X}}$ ,  $G_{\text{solv}}^{\text{X}}$ , and  $S$ , are respectively the molecular mechanics energy, the solvation free energy, and entropy contributions of the molecular system X. Typically, these contributions are evaluated along a set of “snapshots”, conformations coming from MD simulations performed for the complex PL, the free protein P, and the free ligand L [140,141]. However, the application of MM-GB(PB)SA using a single energy-minimized structure, has been shown to be an adequate and sometimes better approach than the standard free energy averaging over molecular dynamics snapshots, allow a fast and accurate prediction of binding free energy, with consequent save of computing time [142,143].

The first term of (2.32) is composed by:

$$E_{\text{MM}}^{\text{X}} = E_{\text{bonded}}^{\text{X}} + E_{\text{elec}}^{\text{X}} + E_{\text{vdW}}^{\text{X}} \quad (2.33)$$

These correspond to the bonded and non-bonded terms of classical forcefields (see section 2.1.1), obtained by energy minimizations in gas-phase or implicit solvent model. The second term of (2.32) is in turn decomposed in two contributions:

$$G_{\text{solv}}^{\text{X}} = G_{\text{PB(GB)}}^{\text{X}} + G_{\text{SASA}}^{\text{X}} \quad (2.34)$$

which correspond to the polar  $G_{\text{PB(GB)}}^{\text{X}}$ , and non-polar  $G_{\text{SASA}}^{\text{X}}$  contributions to the solvation free energy  $G_{\text{solv}}^{\text{X}}$ . The polar contribution could be calculated using either the generalized Born (GB) or Poisson-Boltzmann (PB) continuum-electrostatic models, while the non-polar term, is proportional to the solvent-accessible surface area (SASA). In PB or GB models the solute X (complex, protein, or ligand) is

treated as a low-dielectric body, which is embedded in a high dielectric medium (water dielectric,  $\epsilon = 78.54$ ). The solvation free energy is then expressed as :

$$G_{PB(GB)}^X = \frac{1}{2} \sum_{i,j} q_i q_j g_{i,j}^{PB(GB)} \quad (2.35)$$

where  $q_i$  and  $q_j$  represent the atomic charges, while  $g_{i,j}^{PB(GB)}$ , is a term calculated exploiting either the PB model, and hence the numerical solution of the Poisson-Boltzmann equation, or the GB model, by solution of the following equation:

$$g_{i,j}^{GB} = \left( \frac{1}{\epsilon} - 1 \right) \left[ r_{i,j}^n + B_{i,j} \exp\left(-\frac{r_{i,j}^n}{AB_{i,j}}\right) \right]^{-1/n} \quad (2.36)$$

Here,  $r_{i,j}$  is the distance between atoms  $i$  and  $j$ ,  $\epsilon$  is the dielectric of the solvent,  $B_{i,j}$  is a parameter depending on atom positions, and finally  $n$  and  $A$ , which are constants set to 2 and 4 respectively. In PB model, the dielectric of the solute is taken into account in  $g_{i,j}^{PB}$ , affecting the calculation of the solvation free energy in (2.35). Finally, the computation of the entropy term  $S$ , the last term of (2.32), is computational expensive, because requires an extensive minimization of the conformations for the complex, protein, and ligand, followed by application of normal mode analysis. For a fast evaluation of the solvation free energy in protein-ligand binding, as, for example, in the case of postprocessing of docking outcome, the calculation of the entropic contribution  $S$  is generally omitted in MM-GB(PB)SA routines [143].

Both MM-PBSA and MM-GBSA, are widely exploited in drug design for binding free energy prediction. Recently, the impact of different forcefields and partial charge models on the performance of these methods was investigated, providing useful guidance for their applications [144].

---

**Chapter 3.**  
**An Automated Docking Protocol for hERG  
Channel Blockers**

---



## 3.1 Introduction

As introduced in section 1.1.1, the hERG potassium channel is of great interest in drug development and drug safety. The drug-induced hERG blockade can lead to QT prolongation, increasing the incidence of potentially fatal arrhythmias referred to as *Torsades de Pointes* (TdP) [145,146]. Although it has been estimated that the frequency of occurrence of these cardiotoxic side effects is less than 1/100'000 [147], assessing the QT liability of drug candidates in a pre-marketing stage represents a major safety issue, which has attracted attention from the drug regulatory authorities. Several restrictions have been therefore placed on the use of many drugs. In particular, during the last decades several torsadogenic potential drugs have been identified and withdrawn from the market [148]. A considerable effort has been spent in the development of *in silico* protocols in order to provide a rapid and cheap prediction of the potential hERG toxicity of drug candidates. Unlike ligand-based approaches, structure-based models can provide a richer picture of the chemical requirements for hERG blockade, which is necessary for understanding the risk factors, and hence, the incidence of these pathological conditions. In addition, such knowledge could be useful for the identification of potential channel blockers in chemical collection, and for reducing hERG toxicity during lead optimization. However, the lack of a hERG crystal structure, and the low percentage of sequence identity with available templates, make the development of reliable structure-based protocol, an extremely challenging task.

To date, many structure-based models have been proposed, exploiting both open and closed homology modelling-based states of the channel. In 2006, Farid and co-workers [69], docked a series of five sertindole analogues in the hERG cavity, the binding site, achieving a good relationship between the predicted binding affinities and experimental data. In particular, their results suggested a crown shaped conformation, perpendicular to the channel axis, for hERG blockers. Similar binding

modes were also identified by Choe et al. [149] for clozapine, another hERG channel blocker. In 2005, Österberg et al. [67], by means of a combined molecular docking-MD-linear interaction energy approach, identified extended parallel solutions for the docked outcome. An extended binding mode, but opposite to the previous one, was also suggested two years later by Stansfeld and co-workers [150]. In particular, starting from a “hybrid” state of the channel, generated by a rotation of the S6 transmembrane helix in a closed state homology model, they obtained binding modes for twenty known blockers consistent with mutagenesis data [151,152] and previous ligand-based *in silico* studies [62,153]. Furthermore, a “multiple” state approach was adopted by Rajamani et al. [71], in order to capture the flexibility of the channel to correctly evaluate the binding affinity of the ligands. Recently in 2011, Boukharta et al. [68], used a docking protocol similar to the one adopted by Österberg [67], in order to rationalize the structure-activity relationships of a series of nine sertindole derivatives. They obtained a good reproduction of the observed binding free energies. In particular, their protocol led to achieve a  $r^2 = 0.60$ , representing an excellent result in the context of hERG structure-based approaches. Because of the absence of a 3D crystal structure, these structure-based models are usually validated by their ability to predict hERG binding affinities that are consistent with experimental values, over the considered set of blockers [67,68,71]. However, despite the remarkable efforts in developing *in silico* models, further investigations are required for achieving a consistent binding mode for the most potent blockers, and hence a detailed comprehension of the molecular features at the basis of hERG toxicity. The major difficulty in the application of structure-based approaches lies in the features of hERG binding site. In fact, the hERG cavity, which is involved in the interaction with blockers, represents an uncommon binding site [154]. The intrinsic symmetry of the channel, the conformations of key aromatic residues, and the importance of unspecific interactions in ligand binding are issues that strongly challenge docking simulations. Moreover, the impact of these aspects on the quality of the derived structure-based models has not been explicitly addressed yet. For instance, it is customary to perform docking on channel models generated by satisfying the C4 point group symmetry [69,150], whereas when using the relaxed complex scheme (docking studies carried out on an ensemble of protein

configurations sampled by means of molecular dynamics simulations) [155], the symmetry of the channel is naturally broken [67,68,156].

### 3.1.1 Aim of the Project and Protocol Presentation

Here, I will present a docking protocol aimed to address some of the issues reported above. Starting from a well-grounded homology model of the open state channel [157], several putative hERG-blocker models were built, employing different protein conformations. The quality of each model was evaluated using as a figure of merit the squared correlation coefficient ( $r^2$ ) between experimental activities and docking scores obtained over a selected set of compounds. The proposed protocol consists of the following steps (see Figure 3.1):

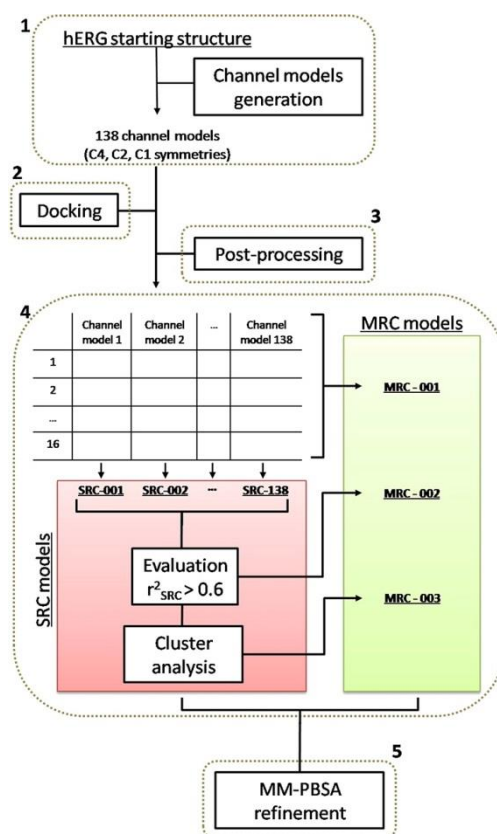
(1) Extensive conformational sampling of the binding site's amino acids was performed imposing the C4, C2, or C1 point group symmetries, thus leading to as many families of channels (C4, C2, and C1, respectively). A total number of 138 models passed a filter criterion meant to assess the stereochemical quality of the protein and to prune possible conformational redundancy.

(2) A series of congeneric sertindole derivatives ( $n = 16$ , series 1 in Table 3.1), for which experimental blocking activity is available and measured in controlled and consistent conditions [65], was docked into the channel models as obtained in step 1.

(3) The poses obtained in step 2 were post-processed to resolve redundancy due to protein symmetry (in case of C4 and C2 families of models) and rescored by taking into account their configurational entropy [158,159].

(4) The docking outcome was finally evaluated via both a single-receptor and a multiple-receptor conformation approach (hereafter referred to as SRC and MRC, respectively) [160]. Specifically, each model relying upon an SRC description consisted of a single channel model associated to a unique binding solution for every ligand belonging to the set, thus leading to a total number of 138 models (SRC-001 to SRC-138). An MRC model was instead built to account for protein flexibility and induced fit phenomena, and it was obtained using information provided by the whole

set of channels (MRC-001). Then, by exploiting the information achieved through a careful analysis of the SRC models, two additional MRC models were obtained (MRC-002 and MRC-003).



**Figure 3.1** Flowchart describing the entire procedure: (1) channel models generation; (2) automated docking of ligands against the 138 hERG channel models; (3) automated post-processing; (4) construction of the hERG-blocker models; (5) refinement of the fittest models.

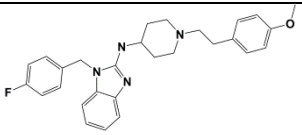
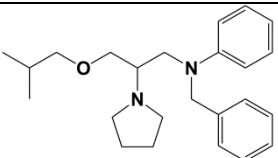
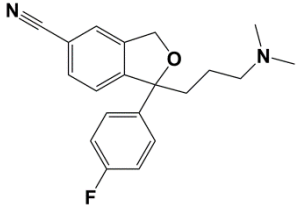
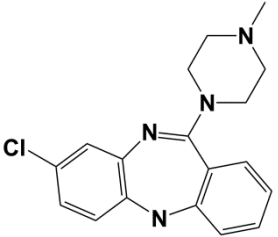
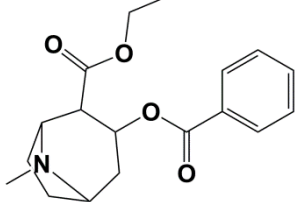
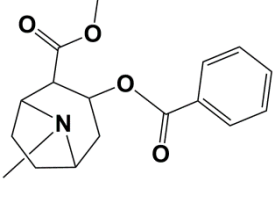
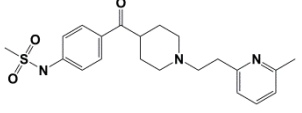
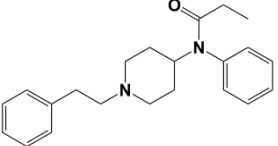
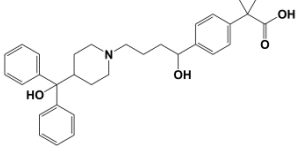
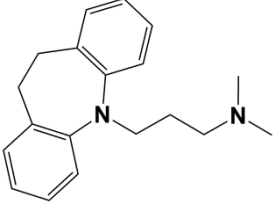
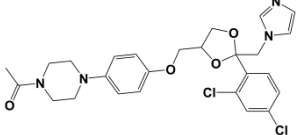
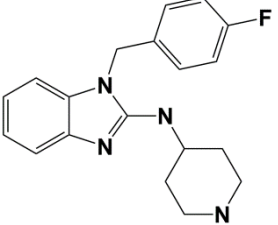
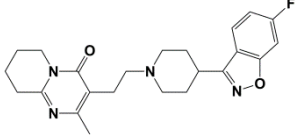
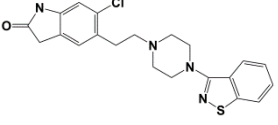
(5) The fittest hERG-blocker models (either SRC or MRC) were further refined to better describe solvation effects upon binding via a single-configurational MM-PBSA rescoring scheme performed using a multiple dielectric description [138,161]. Steps 2-4, and their input/output flows, were automatized and generalized to be used with any kind of series of blockers (see section 3.6.1 Presentation of the automated protocol CoRK<sup>+</sup>, for more details).

The main result of this work is the obtainment of a strategy to achieve a small set of putative hERG-blocker models to quantitatively relate docking scores with

blockers activity in a fully automated way. It is also shown that the use of a limited amount of knowledge-based information to derive a minimal subset of channel models is required to significantly improve the quality of the MRCs compared to a more simplistic SRC description. As a corollary of the effectiveness of the protocol, the results demonstrate that the symmetry of the channel conformations has a non-negligible impact on the performance of the structure-based models derived. Moreover, it is also shown that among the many possible channel models, there is a high percentage of nonvaluable conformations to derive structure–activity relationships. In retrospect, the application of the protocol allowed to highlight seven channel conformations as a subset of relevant and structurally diverse candidates to perform MRC-based docking studies efficiently and without relevant loss of information for the sertindole series of analogues. The protocol was then validated using a series of structurally unrelated blockers (series 2 in Table 3.2. The validation procedure is then described in section 3.3.6).

Compd	Structure	IC <sub>50</sub>	Compd	Structure	IC <sub>50</sub>
Sertindole		3	14		204
1		88	16		26,000
2		10	17		1480
3		7	18		4550
4		579	19		1947
6		137	20		15,700
7		131	21		2200
13		23.5	22		3500

**Table 3.1** Set of sertindole derivatives and their experimental  $IC_{50}$  (nM) used in the development of the docking protocol (series 1). The ligand numbering was adopted according to the work of Pearlstein [65].

Compd	Structure	$IC_{50}$	Compd	Structure	$IC_{50}$
Astemizole		1	Bepidil		501
Citalopram		3981	Clozapine		199
Cocaethylene		1259	Cocaine		7943
E-4031		8	Fentanyl		1995
Fexofenadine		19,953	Imipramine		3162
Ketoconazole		1585	Norastemizole		25
Risperidone		158	Ziprasinone		158

**Table 3.2** Set of structurally unrelated blockers, and their experimental  $IC_{50}$  (nM), used in the validation of the docking protocol (series 2).

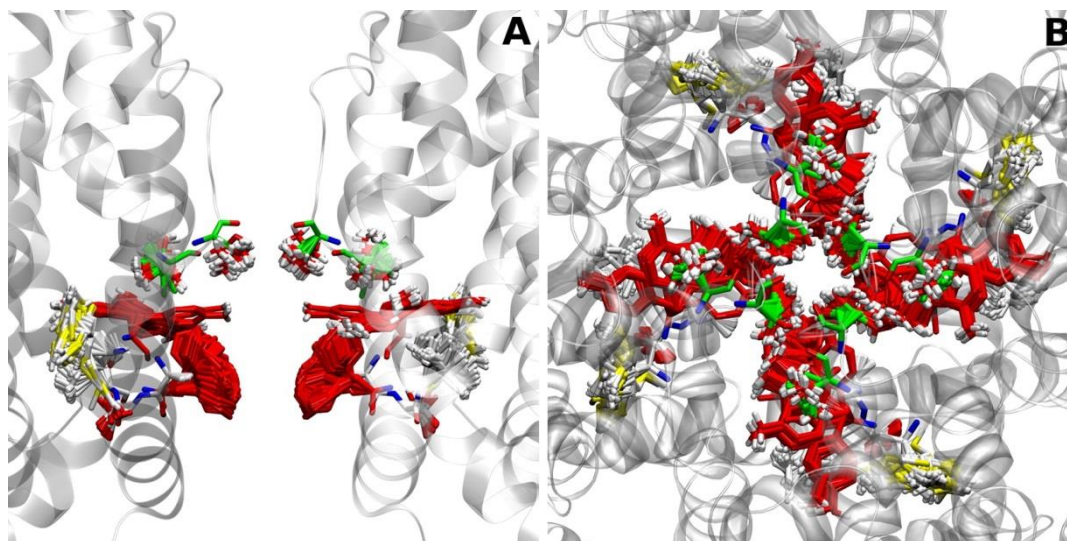
## 3.2 Computational Methods

In this section I will provide a detailed description regarding the strategies used for the development of the automated protocol for the study of hERG channel blockers. The protocol, briefly presented in section 3.1.1, was applied on the two set of compounds listed in Table 3.1 (series 1) and 3.2 (series 2).

### 3.2.1 Channel Models Generation

Since no crystal structures of the hERG channel are currently available, a rigorously validated homology model of the channel was used [157]. In order to characterize the channel cavity flexibility, an exhaustive sampling of the amino acids side chains (residues 623–624 of the selectivity filter and 651–656 of the S6 helix) was carried out by means of Modeller 9v8 [162]. The refinement routine consisted of multiple optimization and molecular dynamics cycles. Specifically, an initial conjugate gradient optimization was followed by a heating-cooling phase of molecular dynamics with simulated annealing. An additional step of conjugate gradient completed the cycle. A set of 296 channel models (Figure 3.2 A and B) was therefore generated by satisfying different point group symmetries ( $C_4$ ,  $C_2$ , and  $C_1$ ). The stereochemical quality of the models was assessed by comparing parameters such as bond lengths, bond angles, torsion angles, and chirality with those derived by high resolution protein structures [163,164] with the Procheck software [165]. A total number of 215 channel models turned out to satisfy the stereochemical quality filter.





**Figure 3.2** (A) Side view (only two out of four subunits are shown for clarity) and (B) top view of the 296 channel models generated. The conformational search was performed on the side chains of amino acids of the inner cavity: Thr623, Ser624 shown in green; Tyr652, Phe656 shown in red; and Met651, Ala653, Ser654, Ile655 shown in yellow.

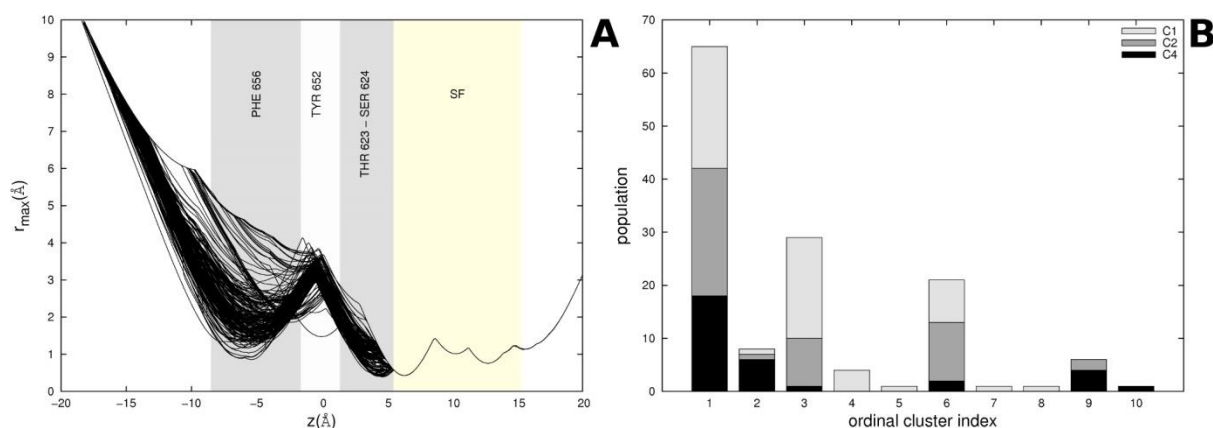
### 3.2.2 Shape-based Cluster Analysis

Rather than using a well-established but problematic clustering analysis in the RMSD space, a shape-based cluster analysis was employed. The accessibility profile of the channel pore for ligand binding (here simply referred to as shape profile) was described by measuring the maximum radius ( $r_{max}$ ) along the channel axis ( $z$ ) using the HOLE program (see Figure 3.3 A) [166]. The local dissimilarity between each channel model  $i$  and  $j$  was defined by taking the modulo difference of the maximum radius of the internal cavity sampled at a given position of the pore axis. Then, by summing these local dissimilarities along  $z$ , the pairwise distance used for the clustering was obtained:

$$d_{i,j} = \sum_z |r_{i,max}(z) - r_{j,max}(z)| \quad (3.1)$$

The cluster analysis was performed with the R software environment [167] using the average linkage method [168]. The metric introduced is a description of the accessible space available for ligand binding, rather than a direct and detailed measure of the shape of the internal cavity. Moreover, the metric adopted, simply based on the maximum radius along the pore axis, can be employed in virtue of a substantially cylindrical binding site and because of the 4-fold symmetry displayed by the tertiary structure of potassium channels, which was not altered during the protein conformational analysis procedure. In other words, the assumption is that different channel conformations, showing similar internal accessible cavities, would also display similar chemical environments, which is indeed reasonable in light of the above-reported peculiar features of the binding site.

Two clustering thresholds were employed in this work, depending on the purpose of the analysis performed. A height cutoff of 15 Å turned out to be effective in pruning the redundancy in the channel models, which were therefore considerably reduced from a number of 215 to 138. Such a cutoff was chosen in a heuristic way, by carefully evaluating the trade-off between the ability of the method to detect redundancy (either due to conformational duplication or symmetry multiplicity) and the inevitable loss of information. Besides such a fine-grained clustering, a coarser clustering analysis was actually employed to classify the channel models for the subsequent analysis of the SRCs. Rather than explicitly using a height cutoff, in this case the cluster dendrogram was cut so as to obtain a convenient number of clusters able to describe the elementary shape of the pore cavity. A number of 10 clusters was arbitrarily chosen as a compromise between manageability of the outcome and accuracy of the description.



**Figure 3.3** (A) Shape profile of the pore of the 138 channel models. The important areas of the pore are shown as grayish rectangles (the Phe656, Tyr652, and Thr623–Ser624 rings). For clarity, the region spanned by the selectivity filter is also shown (in beige). (B) Population of the clusters represented in terms of the different symmetry families of the channel models.

### 3.2.3 Docking

An automated docking protocol was applied to a set of 16 congeneric sertindole derivatives, listed in Table 3.1 (series 1), for which experimental hERG blocking activity is available [65]. These compounds were selected in order to uniformly cover a quite broad range of activity (about 4 pIC<sub>50</sub> units) and to span a significantly large chemical space (although necessarily limited by the congenericity relationship).

The docking simulations were performed using the Autodock 4.2 software [137]. The ligands were treated using Gasteiger partial charges [169], whereas Kollman charges [111] were used for the protein. An all-atoms representation was adopted. The *autogrid* box was built so as to include the cavity of the channel models which were considered as rigid bodies during simulations. The search was carried out with the Lamarckian genetic algorithm (more details in section 2.3.1), which allowed an exhaustive sampling of the conformational space, and the docking runs were set to 250 with the initial population of 150 individuals and a mutation rate of 0.02. Maximum number of generations and energy evaluations were set to 27 000 and  $2.5 \times 10^6$ , respectively. The same parameters were used for series 2 (Table 3.2).

### 3.2.4 Post-processing of the Docking Outcome

#### *Channel Symmetry*

To cope with the docking poses redundancy arising as a consequence of the C2 and C4 channel symmetries, the docking outcome was subject to rotation along the channel axis ( $z$ ) by 90, 180, and 270°. Then, among the different solutions, only the orientation showing the lowest RMSD with respect to a reference structure was kept.

### *Colony Energy*

The pose rescoring was performed using the CE (Colony Energy) method proposed by Xiang et al. [158] in the context of protein loop modelling and later extended to protein-ligand docking problems [159]. The central assumption of the method is that each sampled conformation represents a colony of states on the potential energy surface and that similar configurations (binding poses) belong to the same basin. The size of the colonies, depending on the density of poses which are close in the configurational space (in the limit of an exhaustive and uniformly distributed sampling), represents a statistical assessment of the configurational entropy for a given docking solution. The CE score assigned to  $i$ th configuration is expressed as [158,159]:

$$CE_i \equiv -k_B T \ln \left[ \sum_{j=1}^N \exp \left( -\frac{\text{RMSD}_{i,j}^3}{t^3} \right) \exp \left( -\frac{1}{k_B T E_j} \right) \right] \quad (3.2)$$

where  $k_B$  is the Boltzmann constant, and  $T$  is the temperature (300 K). As it can be noticed from (3.2), the CE score has the form of a free energy and the argument of the logarithm plays the role of a partition function where the energy (score) of the  $j$ th docking solution is weighted by a function that takes into account the similarity of the given docking solution to the colony leading configuration ( $i$ th pose). In the weighting function, the  $t$  parameter was set to 1.8 Å [170]. To properly employ the CE approach both the symmetry around the channel axis and the internal symmetry of the ligands must be taken into account. To this aim, all the symmetric atom pairs were swapped in coordinates, and the lowest RMSD was used in the CE calculation. The use of 250 runs in the docking procedure was motivated by the need to obtain a large ensemble

of binding modes for each ligand, so as to meet the statistical requirements of an exhaustive and uniform sampling. For each ligand, the AD score corresponding to the lowest (top-ranked) CE pose was used to describe the binding.

### 3.2.5 Building and Evaluation of hERG-Blockers Models

Two kinds of correlative structure-based models were built using the information provided by the docking procedure: the more naïve SRCs and the more physically sound MRCs.

#### *Single Receptor Conformation (SRC) Model*

For the generation of the SRC models, only the results obtained by the docking performed against each individual receptor conformation were considered, so that all the 138 channel models were treated independently. For each channel, the AD score obtained for the top-ranking binding modes highlighted by the CE method was stored and employed in evaluating the performance of the derived structure-based model. As figure of merit, the squared correlation coefficient  $r^2$  between the ligands' docking scores and experimental hERG blocking activities expressed as  $\text{pIC}_{50}$  was used. To simplify the analysis of the hERG-blocker models, a classification of the fitness depending on the observed  $r^2$  was adopted. Accordingly, structure-based models showing an  $r^2$  lower than 0.4 were classified as “bad” performing models, whereas those displaying a squared correlation coefficient greater than 0.6 were considered as “good” performing models.

#### *Multiple Receptor Conformation (MRC) Model*

Three MRC models were built by differently exploiting the information provided by the docking procedure and the analysis of the channel models. The MRC-001 model was derived combining all the information obtained by the docking procedure against the 138 channel conformations. The MRC-002 was obtained upon

a selection of the channel models leading to the fittest SRCs ( $r^2 > 0.6$ : 24 channel conformations). Finally, the MRC-003 was built by combining the latter selection with the results of the cluster analysis, leading to a total number of 7 channel structures. Unlike SRCs, in the MRCs evaluation, three statistical data treatments across the ensemble members were used [171]: the best scores, the arithmetic mean, and the Boltzmann-weighted average (see Table 3.3 for the sertindole series). In the latter, the Boltzmann distribution function was applied to the ligands' scores within the ensemble of conformations:

$$p = \frac{N_i}{\sum N_i} = \frac{\exp(-\frac{E_i}{k_B T})}{\sum \exp(-\frac{E_i}{k_B T})} \quad (3.3)$$

Besides the performance of the hERG-blocker models, their predictive ability was also examined through two different statistical approaches: the widely used leave-one-out  $q^2$  and PI [172]. The  $q^2$  was estimated by a correlation between ligands' scores and related predicted activities, derived from the leave-one-out technique: each molecule was in turn removed from the initial set of compounds, and then, its activity was predicted using the remainder of the data. While the  $q^2$  proved the robustness of a model, the PI was employed to quantify the ability of a model to rank the compounds according to their binding affinities. The PI was calculated with the equation:

$$PI = \frac{\sum_{j>i} \sum_i \omega_{ij} C_{ij}}{\sum_{j>i} \sum_i \omega_{ij}} \quad (3.4)$$

with

$$\omega_{ij} = |E(j) - E(i)| \quad (3.5)$$

and

$$C_{ij} = \begin{cases} 1 & \text{if } [E(j) - E(i)]/[P(j) - P(i)] < 0 \\ -1 & \text{if } [E(j) - E(i)]/[P(j) - P(i)] > 0 \\ 0 & \text{if } [P(j) - P(i)] = 0 \end{cases} \quad (3.6)$$

where  $E(i)$  is the experimental  $\text{pIC}_{50}$  and  $P(i)$ , the score related to the reference  $i$ th compound. The PI assumes values ranging from  $-1$  to  $+1$ : a value of  $+1$  indicates perfect predictions; a value of  $-1$  indicates wrong predictions; a  $\text{PI} = 0$ , suggests completely random predictions.

The procedure consisting in docking, post-processing of the docking outcome, and building and evaluation of the correlative models was entirely automated through the Unix BASH-shell scripting language with multiple Tcl and Awk calls, see 3.6.1 Presentation of the automated protocol CoRK<sup>+</sup>.

### 3.2.6 MM-PBSA Refinement

In order to better treat solvation effects, the fittest hERG-blocker models were refined using a single-configurational MM-PBSA [138] rescoring scheme. Accordingly, the  $\Delta G_{\text{bind}}$  for each ligand was computed as:

$$\Delta G_{\text{bind}} = G_{\text{complex}} - G_{\text{protein}} - G_{\text{ligand}} \quad (3.7)$$

where each term was evaluated as the sum of the force field energy ( $G_{MM}$ ) and the polar ( $G_{PB}$ ) and nonpolar ( $G_{SA}$ ) contributions to solvation free energy. For sake of simplicity, no entropic contributions were taken into account [173,174]. The force field energy was calculated after 1000 steps of minimizations (500 steps of steepest descent followed by 500 steps of conjugate gradient) of the complexes by means of the *sander* module of the Amber11 package [175]. The generalized Born model of Hawkins, Cramer, and Truhlar was used as solvation model [176,177]. The AMBER ff99SB-ildn force field [118] was used for the protein, and the General Amber Force Field (GAFF) [178] together with RESP charges [179,180] was adopted to treat ligands. RESP charges were calculated with the G09 package [181] at the B3LYP/6-31G\*\*/B3LYP/6-31G\* level of theory. The polar contributions to solvation free energy were estimated by solving the linearized Poisson–Boltzmann (PB) equation for the minimized structures using APBS [182]. In doing so, a multiple dielectric description of the systems was adopted: water, solute, and membrane were treated as different

environments using  $\epsilon$  values of 80, 4, and 2, respectively. A 3-level of focusing approach which starts by solving the PB equation on a coarse grid of large size, then on a medium grid, and finally on a fine grid, was used for this purpose. Two types of calculations were performed: the first, in a multielectric environments using an ionic strength of 0.1 M of both +1/-1 ions with a radius of 2.0 Å in the aqueous environment, and the second in vacuum (no membrane, homogeneous dielectric  $\epsilon = 1$  for solute and solvent, and null ionic strength). The solvation energy was estimated by subtracting the latter term (the Coulombic contribution, in vacuum) to the first calculation. Finally, the nonpolar components of the solvation free energy were computed with APBS according to the following equation [182]:

$$G_{SA} = \gamma SASA + b \quad (3.8)$$

where  $SASA$  is the solvent-accessible surface area estimated using a probe with radius of 1.4 Å,  $\gamma$  is the solvent surface tension parameter ( $0.00542 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ ), and  $b$  is the free energy of nonpolar solvation for a point solute ( $0.92 \text{ kcal mol}^{-1}$ ) [142,173].



Compd	SRC-001	SRC-002	SRC-008	MRC-001			MRC-002			MRC-003		
				best	avg	bz	best	avg	bz	best	avg	bz
sertindole	-5.07	-5.33	-5.41	-5.41	-3.56	-4.60	-5.41	-4.66	-4.93	-5.41	-4.93	-5.11
<b>1</b>	-4.42	-5.23	-4.85	-5.23	-3.47	-4.48	-5.23	-4.43	-4.76	-5.23	-4.68	-4.89
<b>2</b>	-4.79	-5.23	-5.31	-5.34	-3.58	-4.67	-5.31	-4.67	-4.95	-5.31	-4.91	-5.05
<b>3</b>	-4.81	-4.92	-4.84	-5.05	-3.30	-4.39	-5.01	-4.46	-4.64	-4.92	-4.62	-4.72
<b>4</b>	-4.02	-4.36	-4.36	-5.93	-3.91	-4.80	-5.55	-4.38	-4.75	-4.97	-4.38	-4.54
<b>6</b>	-4.82	-5.43	-5.44	-5.73	-3.86	-4.88	-5.59	-4.71	-5.09	-5.44	-5.03	-5.22
<b>7</b>	-4.29	-4.35	-4.50	-5.21	-3.19	-4.15	-5.21	-4.03	-4.47	-4.80	-4.24	-4.43
<b>13</b>	-4.49	-4.86	-4.06	-5.38	-3.83	-4.60	-5.26	-4.47	-4.78	-4.94	-4.60	-4.73
<b>14</b>	-4.50	-4.76	-4.98	-5.63	-3.78	-4.79	-5.10	-4.40	-4.63	-4.98	-4.60	-4.68
<b>16</b>	-2.71	-2.83	-2.69	-3.71	-2.61	-2.99	-3.34	-2.91	-3.00	-3.28	-2.91	-2.98
<b>17</b>	-2.51	-2.67	-2.65	-3.29	-2.14	-2.68	-3.10	-2.62	-2.78	-2.76	-2.64	-2.65
<b>18</b>	-3.31	-2.97	-2.89	-3.63	-2.72	-3.01	-3.47	-2.99	-3.14	-3.42	-3.07	-3.16
<b>19</b>	-2.99	-3.41	-2.85	-3.83	-2.74	-3.24	-3.68	-3.18	-3.28	-3.43	-3.15	-3.23
<b>20</b>	-2.28	-2.62	-2.24	-3.66	-2.08	-2.81	-3.35	-2.58	-2.88	-2.81	-2.55	-2.62
<b>21</b>	-3.35	-2.84	-2.39	-3.68	-2.48	-2.99	-3.57	-2.95	-3.16	-3.35	-2.95	-3.12
<b>22</b>	-2.99	-2.91	-2.70	-3.58	-2.63	-2.96	-3.24	-2.92	-3.00	-3.22	-2.97	-3.02
$r^2$	<b>0.83</b>	<b>0.79</b>	<b>0.74</b>	<b>0.57</b>	<b>0.53</b>	<b>0.65</b>	<b>0.67</b>	<b>0.76</b>	<b>0.73</b>	<b>0.73</b>	<b>0.77</b>	<b>0.76</b>

**Table 3.3** Fittest correlative structure-based models for series 1. The docking score ( $\text{kcal mol}^{-1}$ ) of the 16 compounds used to develop the protocol for the three fittest SRC (SRC-001, SRC-002, and SRC-008) and MRC models, are reported. For the MRC models, the results obtained using the three data treatment methods employed are also shown (best scores, arithmetic mean, and Boltzmann-weighted average). The performance of each model measured in terms of  $r^2$  is reported at the bottom of the table.

## 3.3 Results

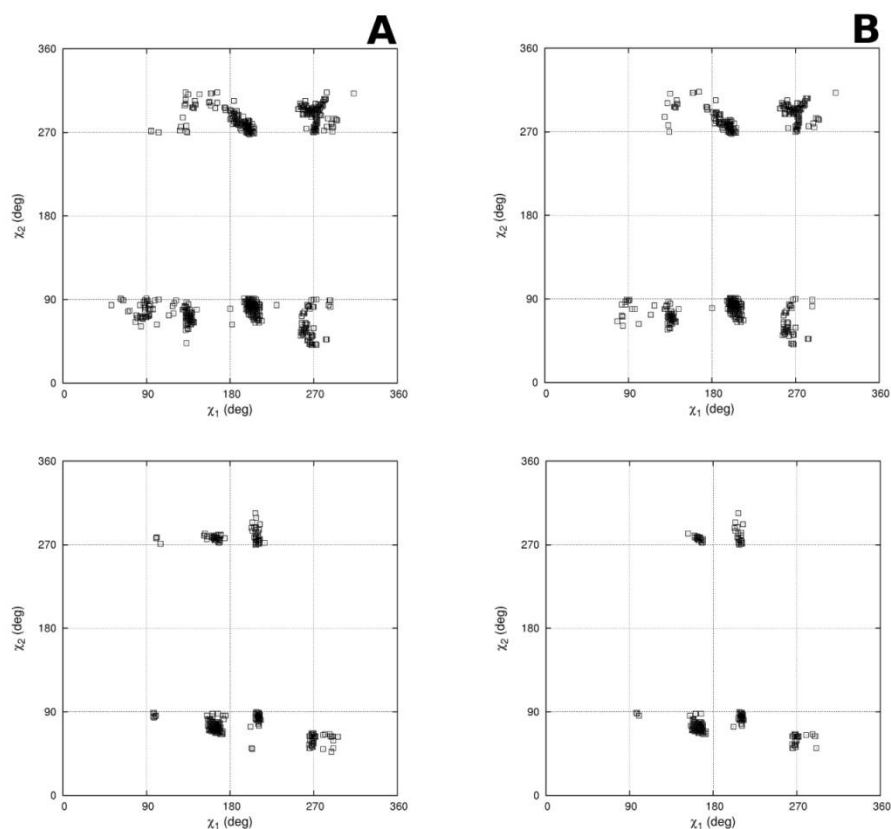
In this section, I will report the main results achieved in the development and application of the docking protocol previously described. The protocol was developed on the series of congeneric sertindole derivatives, series 1, and validated on a series of structurally unrelated blockers, series 2.

### 3.3.1 Development of the Protocol

#### 1. Channel Models

To account for flexibility, the local conformational space of the cavity of the open state channel was extensively explored leading to a set of channel models (Figure 3.2). Under the assumption that the starting configuration represented an optimal geometric assembly of the protein in agreement with experimental data [157], only the side chains of key amino acids (see section 3.2.1 Channel Models Generation) were relaxed. In particular, the attention was focused on the conformations explored by Tyr652 and Phe656, since their importance in ligand binding has largely been discussed in the literature [152]. From the analysis of the  $\chi_1/\chi_2$  plot for these residues before and after applying the stereochemical quality filter (shown in Figure 3.4 A and B, respectively) it has been highlighted that, taking into account all the models, these amino acids sampled the most of their conformational space. Notably, the  $\chi_1/\chi_2$  plots are in good agreement with those recently obtained by Knape and co-workers and derived from 50 ns of molecular dynamics simulation [183]. In Figure 3.3 A, the shape of the cavity measured as

maximum pore radius plotted against channel axis ( $z$ ) is reported for all the channel models. It clearly emerged that most of the variability in shape was in proximity of the phenyl ring of Phe656. From a docking standpoint, the four Phe656 residues could constitute a potential pore restriction that might critically affect the generation of reliable hERG-blocker models. Indeed, in this region of the pore, in some cases we observed a maximum radius as small as about 1 Å, underlying protein conformations virtually unproductive to accommodate the majority of the larger blockers. For sake of clarity, we refer to these conformations as “narrow channel models” to be distinguished by the “wide channel models”, bearing in this region a maximum radius greater than 3 Å. Conversely, Tyr652 protruded much less toward the cavity, and the sampled conformational space was significantly more limited (see Figure 3.4 B).



**Figure 3.4**  $\chi_1/\chi_2$  plot for Phe656 (top) and Tyr652 (bottom) before (A) and after (B) applying the stereochemical quality filter.

In order to classify the channel conformations and to prune their possible redundancy, a cluster analysis was performed. Because of the channel overall

symmetry and the internal symmetry of amino acids' side chains such as phenylalanine and tyrosine, the RMSD metric appeared to be particularly ill suited to address the problem. Therefore for measuring the dissimilarity among channel models, was used a metric describing the local difference in shape calculated over the pore axis (see section 3.2.2 Shaped-Based Cluster Analysis, for more details). The proposed metric turned out to be quite effective for our purposes and provided an intuitive picture of the classification of channel models into clusters. In Figure 3.3 B, the population of the 10 clusters along with their internal population in terms of symmetry families is reported. As it can be seen, the symmetry of the channel models was transversally parted on the clusters. In other words, there is no a direct relationship between the symmetry family and a particular shape of the cavity, implying that a similar shape could be obtained by imposing different symmetries during the generation of the channel models.

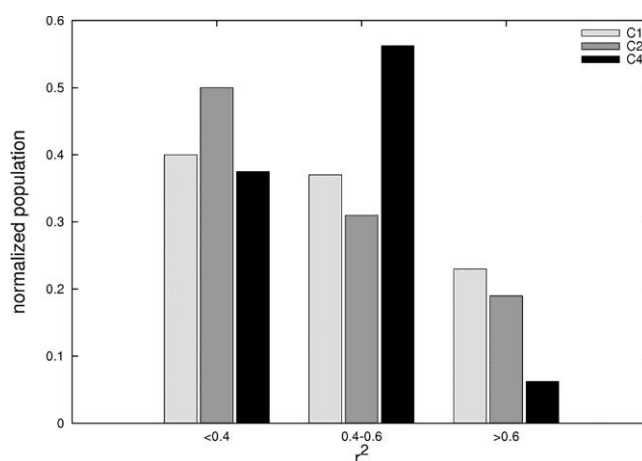
## 2. SRC Models

The performance of the hERG-blocker models was evaluated by calculating the correlation (in terms of  $r^2$ ) between the docking scores and the experimental blocking activities. The  $r^2$  values for all the SRC models along with the information concerning the symmetry family of the corresponding channel model are reported in section 3.6.2 Supporting Material, Table S1. As explained in section 3.2.5 of Computational Methods,  $r^2$  values were calculated employing the Autodock [137] score (AD score) after applying a pose rescoring, which took into account the ligands' configurational entropy. This approach was based on the assumption that a better description of binding would be obtained by reweighting the score associated to a given pose by its relative probability to be achieved during extensive docking simulations (i.e., the so-called Colony Energy, CE) [159]. Within the CE formalism, the weight assigned to a specific binding mode statistically incorporates (in an approximate way) the contribution of the ligand configurational entropy in a rescoring scheme-like fashion [159].

In Figure S1 (section 3.6.2 Supporting Material), the performance of the SRC models is reported for each symmetry family (C1, C2, and C4 in panel A, B, and C,

respectively). In the same figure, the  $r^2$  values calculated before (i.e., AD solutions) and after applying the CE rescoring scheme, were also compared.

The rescoring procedure either slightly or significantly decreased the fitness of the SRC model. However, for a limited number of channel conformations, an interesting performance increasing was recorded. By classifying the fitness of the hERG-blocker models as bad ( $r^2 < 0.4$ ), intermediate ( $0.4 < r^2 < 0.6$ ), and good ( $r^2 > 0.6$ ), the performances of the SRCs were reported as histograms for each symmetry family. Figure 3.5 shows that C1 channel models performed on average better than those belonging to other families. However, the best SRC models of the entire ensemble belonged to the C2 family (SRC-001 and SRC-002,  $r^2$  of 0.83 and 0.79, respectively, see Figure S3 B in section 3.6.2 Supporting Material), even though, at the same time, this family owned the largest amount of poorly performing models. Generally, the C4 family of models performed much worse.

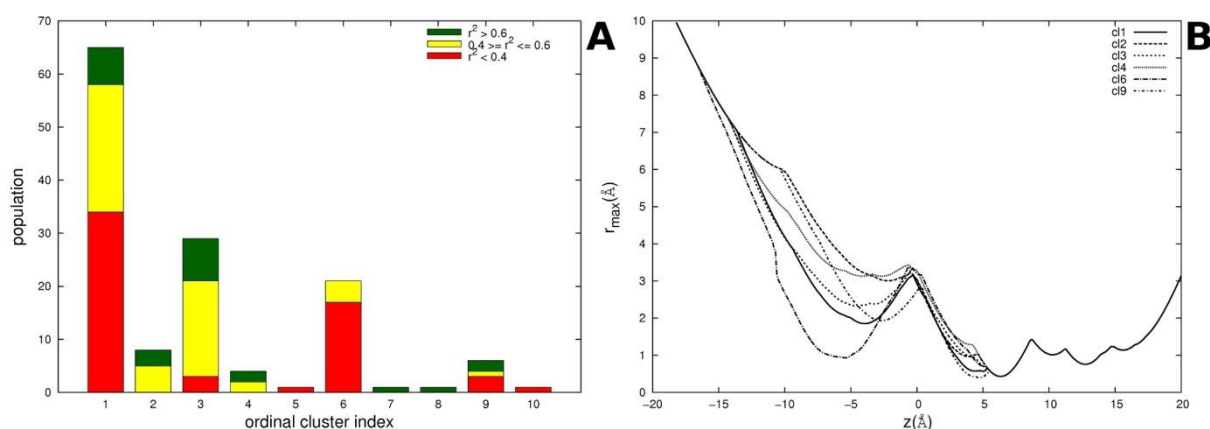


**Figure 3.5** Performance evaluation of the SRC models for series 1 based on the squared correlation coefficient and reported in terms of the symmetry family displayed by the corresponding channel model.

Since no clear-cut connections between symmetry and cavity shape, or symmetry and performance of the derived SRC models were obtained, the attention was focused on investigating whether any relationships between shape and fitness could be derived. In Figure 3.6 A, the clusters of the channel models along with the quality of the corresponding SRCs are presented. Apart from the trivial singleton clusters, a certain correspondence could be inferred. From a hERG-blocker model

standpoint, clusters 2, 3, and 4 clearly underlined effective channel model shapes, whereas cluster 6 definitely did not. The shape profiles of the channel conformations associated to the fittest SRC models are shown in Figure 3.6 B for the nonsingleton clusters. The channel model representative of cluster 6 exhibited an extremely narrow cavity conformation, thus explaining the poor performances of SRC models within this cluster. Indeed, the restriction provided by the Phe656 ring hampered the access of ligands to the upper portion of the binding site, thus preventing a proper binding and in turn an efficient channel block.

Such results could be an indirect confirmation of the health of the protocol. In fact, because the AD docking box was generously extended toward the intracellular side of the protein, ligands were allowed to bind even in the presence of the restriction. Therefore, the fact that the narrowest channels systematically provided much worse SRC models might be interpreted as a signal of a correct quantitative response of the protocol. Notably, while narrow channels always provided poor SRC models, the opposite did not hold true. In fact, it was possible to identify moderately wide channel models providing a poor correlation, even though in general they rather returned average or good SRC models.



**Figure 3.6** (A) Population of the clusters represented in terms of the associated SRCs' fitness calculated for series 1. (B) Shape profiles of the best SRC models for each cluster. Singleton clusters are not shown.

### 3. MRC Models

Even though analysing the performance of SRC models is an informative exercise, it is clear and straightforward that a most relevant picture of binding should be achieved through a correlative hERG-blocker model based on an MRC description. The first MRC model (MRC-001), obtained using the information coming from all the docking simulations in a standard ensemble docking approach, turned out to be rather poor. In fact, either using the best scores and the average scores data treatment, an unsatisfactory correlation was obtained ( $r^2$  of 0.57 and 0.53, respectively; see Table 3.3). Only when using the Boltzmann weighted averages an acceptable correlation was reached ( $r^2 = 0.65$ ). Although disappointing, these results were not completely unexpected, as it is well-known that using a large number of conformations might have a detrimental effect on the performances of the ensemble docking approach [184]. The second MRC model generated (MRC-002) was built employing only the information coming from the docking simulations performed on a worthwhile subset of channel models yielding the best correlations between calculated and experimental data in the SRC description ( $r^2 > 0.60$ ). The rationale of this procedure was based on the assumption that protein conformations responsible for the fittest SRC models, on average, would carry more relevant information to describe the binding of blockers than the others. In other words, the identification of a minimal subset of channel models would statistically improve the performance of the MRC description by reducing the noise possibly related to scoring function inaccuracies or limitations in sampling. In this respect, this approach is similar in spirit (although slightly less rigorous but much simpler) to the method proposed by Yoon and Welsh [185]. Accordingly, the MRC-002 statistics significantly improved (see Table 3.3). Notably, only 24 channel models out of 138 were used in this description. In pursuing the definition of the minimal subset of channel conformations, MRC-003 was built using only the channel models belonging to the previous subset leading at the same time to the fittest SRCs along different clusters. Compared to MRC-002, MRC-003 displayed better statistics. For example, for the best scores data treatment, the  $r^2$  increased from 0.67 to 0.73 (see Table 3.3). This result demonstrates that a good MRC model for the hERG channel can be successfully obtained with an extremely limited subset of channel conformations. In this respect, the analysis of SRC models turned out to be instrumental in defining a strategy to select the MRC channel conformations.

Examining the methods used for the data treatment, averaging docking scores across the channel conformations turned out to provide better statistical performances rather than using exclusively the best energy scores. However, no substantial differences were noticed whether a simple arithmetic mean or a Boltzmann-weighted mean were used.

In the best case scenario, MRCs would be the only hERG-blocker models to be considered, since their derivation is solely driven by the physics of the scoring function, while SRCs are obviously conditional models.

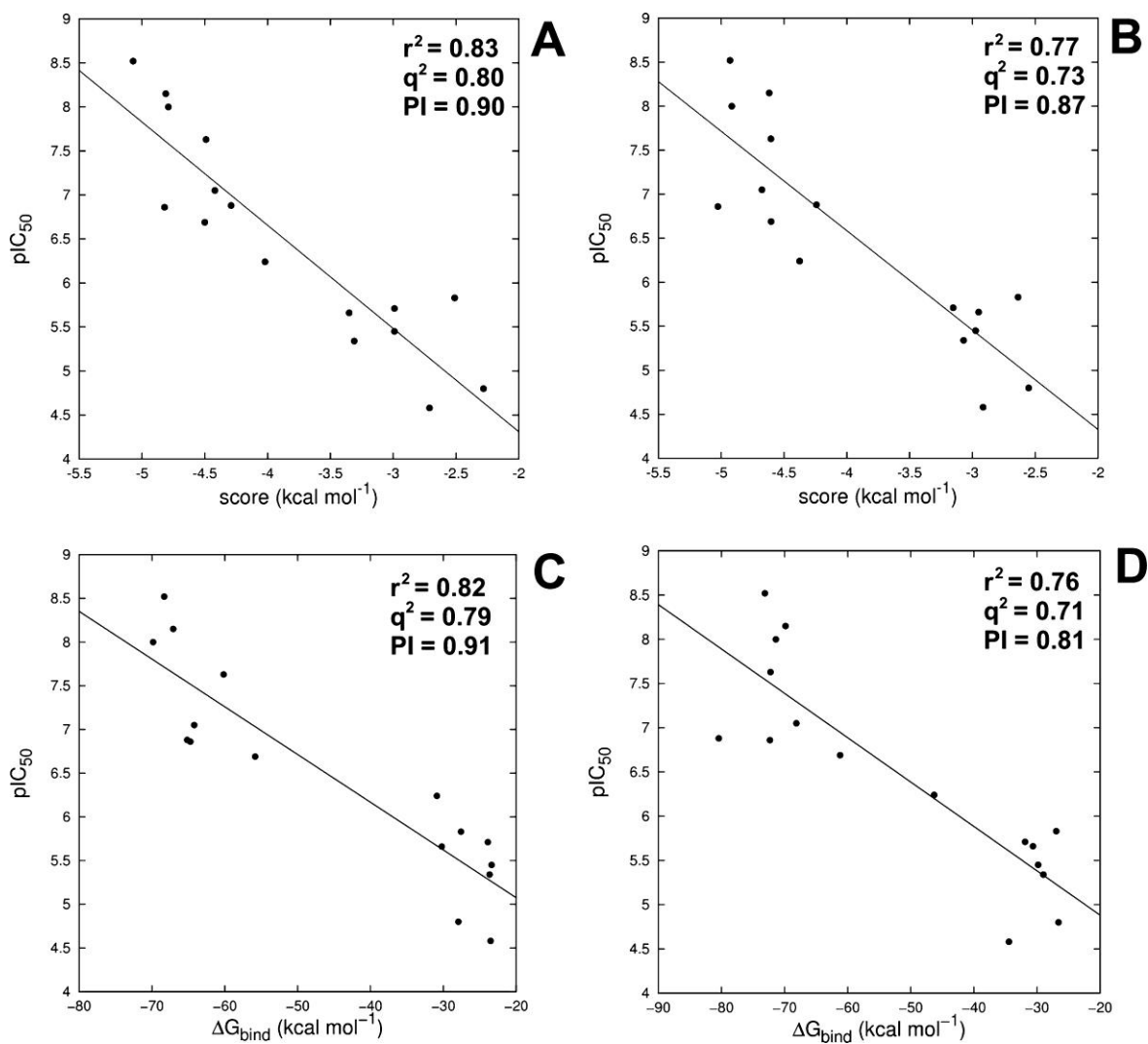
#### 4. *Structure-Activity Relationships*

Here, I will present the binding modes adopted by the sertindole analogues in the fittest structure-based models: SRC-001 ( $r^2 = 0.83$ , see Figure 3.7 A) and MRC-003 ( $r^2 = 0.77$ , adopting the arithmetic mean data treatment method, Figure 3.7 B). Despite the slightly better performance of SRC-001, a more consistent binding mode was achieved using the MRC approach.

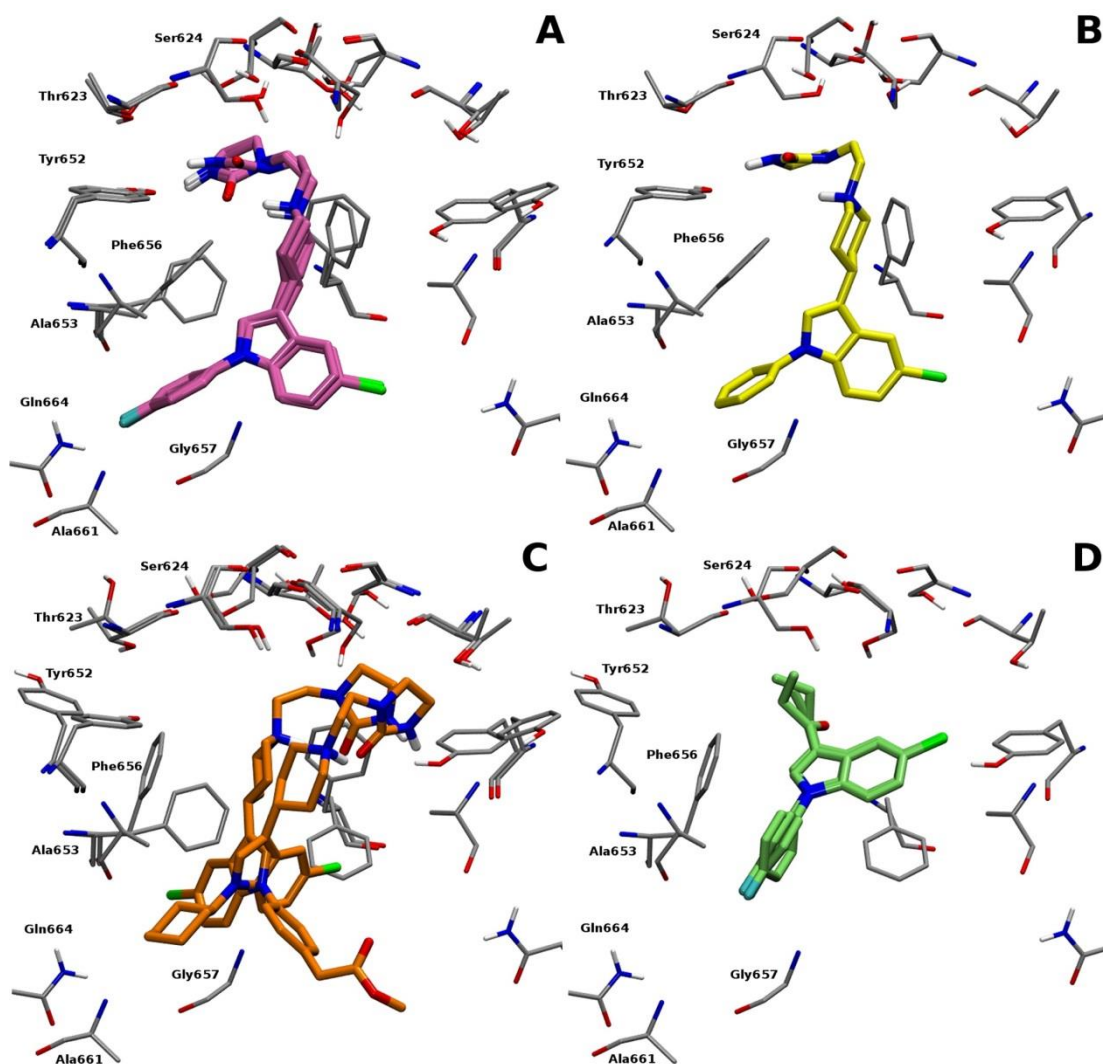
In MRC-003, all the compounds displayed a binding mode similar to that proposed by Österberg and Åqvist [67], by fitting into the cavity parallel to the channel axis, adopting an extended conformation, and pointing the imidazolidinone moiety toward the selectivity filter. The basic nitrogen was found to occupy a central position in the area surrounded by Phe656–Tyr652 and, in agreement with other studies [67,68], no cation- $\pi$  interactions with Tyr652 were observed. The most potent blockers of the set of compounds, sertindole ( $IC_{50} = 3$  nM), **3** ( $IC_{50} = 7$  nM), and **2** ( $IC_{50} = 10$  nM) showed similar binding modes though different channel models were chosen (channel model C1\_43 for sertindole and **3** and C2\_48 for **2**; Figure 3.8 A). In particular, the imidazolidinone group was placed at the bottom of the filter. The amidic nitrogen of this group was involved in an H-bond interaction with the hydroxyl group of Tyr652, whereas the aliphatic cyclic moiety showed hydrophobic interactions with the adjacent Tyr652–Phe656 and with the backbone of Thr623–Ser624. In addition, the indole moiety formed hydrophobic or  $\pi$ -stacking interactions with at least three neighbouring Phe656, while the fluorobenzyl group took electrostatic contacts with the side chain of Gln664 and hydrophobic interactions with



Phe656 and Ala653. In the resulting binding modes, Gln664 together with Tyr652 seemed to play a key role in determining the blocking affinity. This was evident by the poses adopted by compounds **13** ( $IC_{50} = 23.5$  nM, Figure S2 A in section 3.6.2 Supporting Material) and **1** ( $IC_{50} = 88$  nM, Figure S2 B), where the absence of the H-bond either with Gln664 (in the case of **1**, lacking the fluorine) or with Tyr652 (in the case of **13**) reflected the reduction of affinity. The presence of a methyl phenylacetate instead of the fluorobenzyl group, together with the absence of polar contact with Tyr652 in **7** ( $IC_{50} = 131$  nM) or the total substitution of the halo-aromatic ring with a cyclohexyl moiety in **6** ( $IC_{50} = 137$  nM), were found to be further detrimental for the blocking activity (Figure 3.8 C). Although the H-bond with Gln664 was preserved, **4** ( $IC_{50} = 579$  nM) did not interact with Tyr652 and fewer hydrophobic contacts between the indole group and Phe656 were observed (Figure S2 B). Concerning the less potent binders, which were also the smaller molecules of the set (**14–22**), the indole moiety replaced the piperazine group in the binding site, occupying the accessible volume delimited by the four Phe656. The fluorobenzyl portion of the ligands was mainly located in the lower side of the cavity in between Ala653 and two copies of adjacent Phe656 (Figure 3.8 D). An exception is represented by **18** ( $IC_{50} = 4550$  nM), which displayed the fluorobenzyl toward the selectivity filter and was involved in nonpolar contacts with two neighboring Thr623, and in a  $\pi$ - $\pi$  T-shaped interaction with one copy of Phe656 (Figure S2 C). Compound **14** ( $IC_{50} = 204$  nM) was the most potent blocker among the smaller compounds, and the only one showing a positively charged nitrogen sequestered by the polar area consisting in Thr623 and Ser624 at the bottom of the filter. The alkyl portions were mainly located in the hydrophobic pocket formed by two adjacent subunits consisting in the residues Thr623, Ser624, Tyr652 and Phe656, while the presence of the hydroxyl functional groups, in **20** ( $IC_{50} = 15,700$  nM) and **21** ( $IC_{50} = 2200$  nM), resulted in additional polar contacts with Ser624 (Figure S2 D).

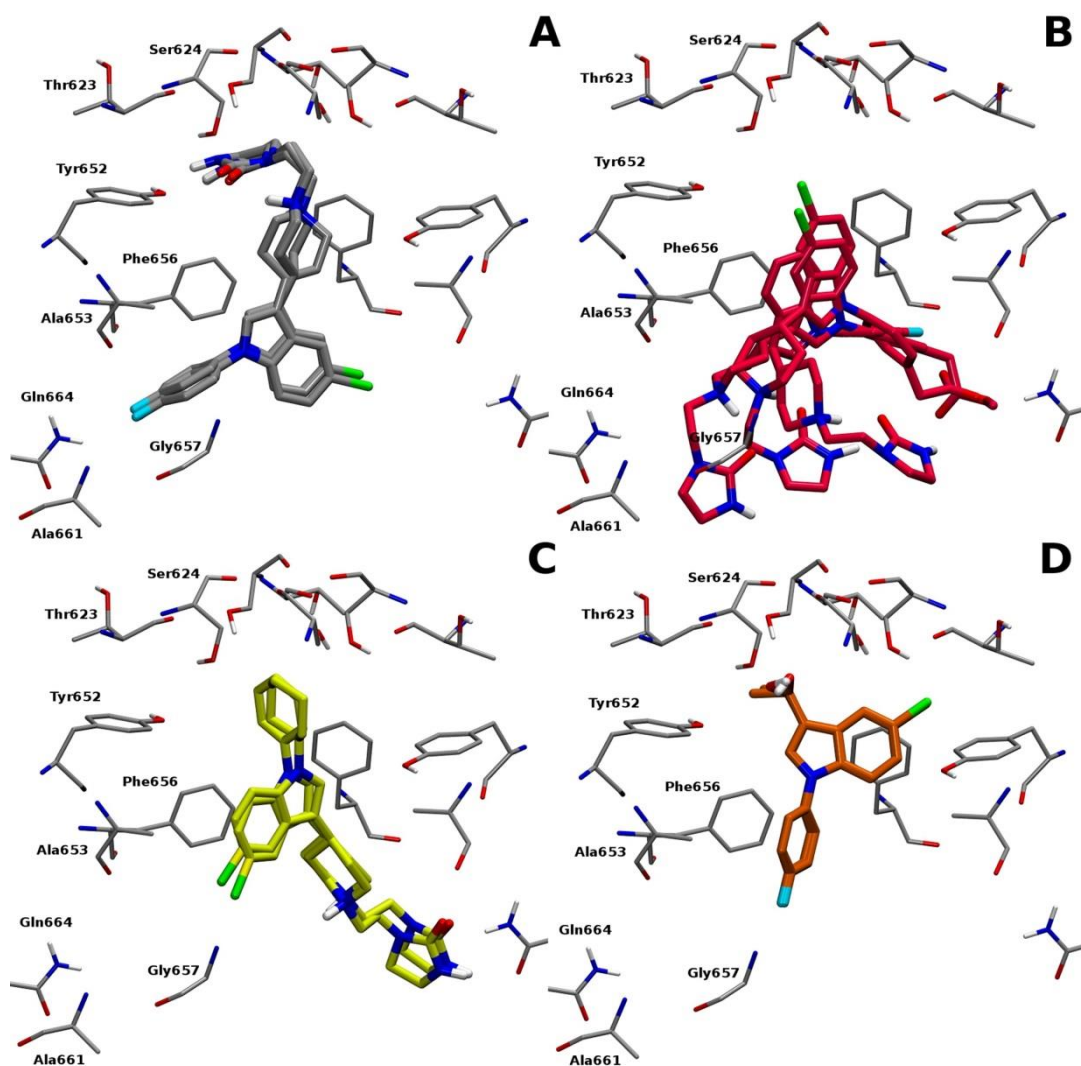


**Figure 3.7** Correlation plots between the experimental  $pIC_{50}$  and the calculated binding energies for series 1 before (upper panels) and after (lower panels) applying the MM-PBSA refinement. Panels A and C refer to SRC-001, whereas panels B and D show the results for MRC-003.



**Figure 3.8** Most relevant binding modes displayed by MRC-003 in series 1: (A) sertindole, 2, and 3 (mauve); (B) compound 1 (yellow); (C) compounds 6 and 7 (orange); and (D) compounds 16, 17, 19, and 22 (lime). The multiple conformations adopted by the protein are shown in gray.

As previously mentioned, the binding picture derived from the SRC-001 model showed less consistency when compared to MRC-003. In particular, three different binding modes could be observed among diverse group of molecules. In analogy with the MRC-003 representation, the highest affinity compounds, sertindole, **3**, and **2**, retained an Österberg-like binding mode [67] with the imidazolidinone group pointed toward the filter, and the fluorobenzyl moiety connected to Gln664 through an H-bond interaction (Figure 3.9 A). Together with a progressive reduction of affinity, different binding modes emerged. Similarly to the binding mode observed by Boukharta and co-workers [68], compounds **13**, **7**, and **4**, placed the indole moiety in the central area of the cavity surrounded by Phe656, which was also involved (in the case of **7** and **4**) in  $\pi$ -stacking interactions with the chlorine oriented toward the selectivity filter (Figure 3.9 B). Furthermore, the fluorobenzyl, the phenylacetate, or the methyl phenylacetate groups in compounds **13**, **4**, and **7**, respectively, were accommodated in the area beneath the Phe656, taking favourable polar interactions with Gln664. However, unlike the binding mode described by Boukharta, the imidazolidinone moiety of these compounds was not placed close to the bottom of the filter, but rather it contacted the C $\alpha$  of Gly657 and the side chain of Ala661 via multiple hydrophobic interactions. A different scenario was provided by compounds **1** and **6**, which showed a binding mode overall in agreement with the one commented by Stansfeld and co-workers [150]. As shown in Figure 3.9 C, these compounds adopted an extended conformation, parallel to the channel axis, and pointing the phenyl group (**1**) or the cyclohexyl ring (**6**) toward the bottom of the filter, while the imidazolidinone interacted via polar contacts with the Gln664 side chain. The indole moiety of these compounds was sequestered in the central core of the cavity by  $\pi$ - $\pi$  interactions involved with the four Phe656. In contrast with the MRC-003 description, the binding modes adopted by smaller compounds showed less consistency. For most of the molecules (compounds **16**, **17**, **19**, **20**, and **22**), the alkyl portions were located in the hydrophobic pocket defined by two copies of adjacent Phe656 and Ala653, in the opposite side with respect to the one occupied by the fluorophenyl group (Figure S3 A in section 3.6.2 Supporting Material). The additional presence of a hydroxyl functional group (**18** and **21**, shown in Figure 3.9 D) or a protonated nitrogen (**14**, Figure S3 B) was responsible to drive the compounds toward the polar area at the bottom of the filter.



**Figure 3.9** Most relevant binding modes displayed by SRC-001 in series 1: (A) sertindole, 2, and 3 (silver); (B) compounds 13, 7, and 4 (magenta); (C) compounds 1 and 6 (yellow); and (D) compounds 18 and 21 (orange).

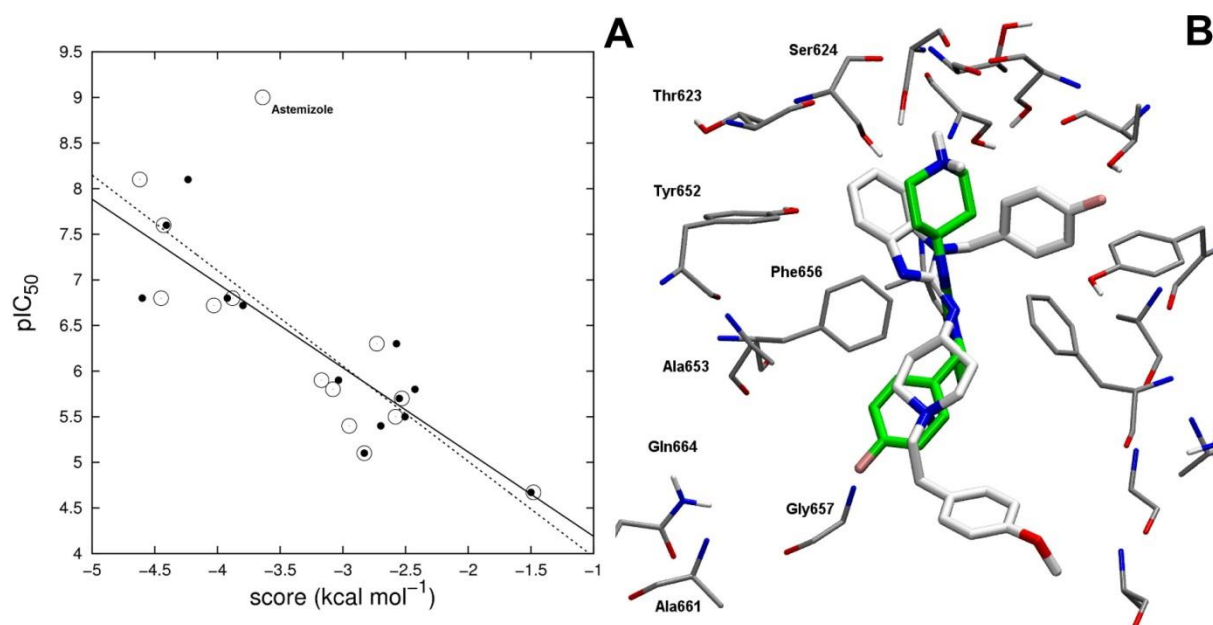
### 5. Assessment of the Solvation Free Energy Contribution

To further confirm the reliability of the fittest hERG-blocker models, an additional force-field based rescoring scheme was employed. In Table S2 (section 3.6.2), the estimated MM-PBSA free energies of binding ( $\text{kcal mol}^{-1}$ ) for the docked compounds in the fittest structure-based models, SRC-001 and MRC-003, was reported. For each model, the molecular mechanics ( $\Delta G_{MM}$ ) and both the polar ( $\Delta G_{PB}$ ) and nonpolar ( $\Delta G_{SA}$ ) contributions to the solvation free energy are also reported (see section 3.2.6 for details). As it can be seen, the favorable formation of the complexes was driven by the molecular-mechanics energy  $\Delta G_{MM}$  and by the nonpolar component of the solvation energy  $\Delta G_{SA}$ , which assumed negative values, while  $\Delta G_{PB}$  component showed positive values. Interestingly, while  $\Delta G_{SA}$  contributions appeared always small and similar among the compounds (ranging from  $-2.79$  to  $-4.62 \text{ kcal mol}^{-1}$  for SRC-001 and from  $-2.71$  to  $-5.00 \text{ kcal mol}^{-1}$  for MRC-003), although somehow correlated with molecular size,  $\Delta G_{MM}$  and  $\Delta G_{PB}$  values were clearly different through the set and were directly dependent on the net charge of the compounds: higher contributions were observed for the positive charged sertindole and compounds **1**, **2**, **3**, **6**, **7**, **13**, and **14**. In Figure 3.7 C and D, the regression plots of the  $\Delta G_{\text{bind}}$  versus  $\text{pIC}_{50}$  for SRC-001 and MRC-003 are respectively reported. In general, the results showed that MM-PBSA did not significantly influence the scenario depicted by the docking scoring function. In particular, the  $r^2$  changed from 0.83 ( $s = 0.51$ ) to 0.82 ( $s = 0.52$ ) in SRC-001, while it passed from 0.77 ( $s = 0.59$ ) to 0.76 ( $s = 0.60$ ) in MRC-003 before and after refinement. Concerning the predictive statistical parameters, while the  $q^2$  slightly decreased, the predictive index (PI) [172] retained satisfactory values (larger than 0.80).

### 3.3.2 Protocol Validation

In order to show the general applicability of the proposed docking strategy, the reported protocol was applied to a series of structurally unrelated blockers (series 2).

The compounds were those of the test set employed by Cavalli et al. for the validation of a 3D-QSAR model [62]. To achieve a uniform distribution of experimental activities, other classical hERG blockers were also taken into account, leading to a validation set of 14 structurally unrelated compounds (see Table 3.2). Series 2 covered a range of more than four  $pIC_{50}$  units of activity (from 4.67 to 9.00). The correlation between experimental activity and predicted docking score for the fittest SRC model turned out to be fairly good ( $r^2 = 0.60$ ; see Figure 3.10 A). Figure 3.10 A, clearly showed that the major outlier was astemizole, and when excluded from the set, very good correlations were obtained. In particular, the fittest SRC displayed an  $r^2$  of 0.83, while the corresponding MRC-003 showed an  $r^2$  of 0.77 (for the average score data treatment, see Figure 3.10 A and Table S3 in section 3.6.2 Supporting Materials). In Figure 3.10 B, the binding mode of some selected compounds belonging to series 2 is reported.



**Figure 3.10** (A) Correlation plots between the experimental  $pIC_{50}$  and the calculated binding energies for series 2 before ( $n = 14$ , empty circles, dotted line) and after ( $n = 13$ , filled circles, solid line) the removal of the outlier astemizole. (B) Representative binding modes for selected blockers: astemizole (white) and norastemizole (green).

## 3.4 Discussion

In this chapter, a docking protocol aimed to predict the putative binding mode for a series of hERG blockers, was presented. The protocol was first developed using a series of sertindole analogues that has become a standard reference for structure-based models of hERG block [67,68]. Then, the same strategy was applied to a more challenging set of structurally unrelated molecules. The protocol consisted in the automatic generation of an ensemble of binding modes for the considered set of compounds and in their evaluation according to the  $r^2$  calculated between docking scores and experimental blocking activities. Equivalently, the proposed procedure generated and compared an ensemble of possible correlative models to quantify the hERG blocking activity. These models were obtained using both SRC and MRC descriptions. In the latter, different conformations of the pore amino acid side chains were taken into account. Specifically, the statistical parameters of the MRC models were improved by using the information obtained by an in depth analysis of the relevant SRCs.

### 3.4.1 Pore Shape of the hERG Channel Models

As previously stated, the features and the statistical performance of the MRC models were strongly dependent upon both the results achieved by the SRCs and the analysis strategy undertaken. In particular, the employed description of the shape of the cavity turned out to be a simple but successful approach to address the problem of an uncommon binding site such as that of the hERG channel. In addition, the results suggested the existence of a relationship between the shape of the pore and the fitness of the SRC model derived from series 1. While a posteriori it was not



surprising that narrow channel models always led to poor SRCs, it was interesting to note that the fitness of the hERG-blocker models did not monotonically improve with the pore width at the constriction region. Despite the importance of the punctual arrangement of amino acids in determining a certain binding mode, and the fact that a given shape could be obtained with different side chain conformations, it is tempting to advance an explanation of the general performance of the SRC models in terms of the shape of the binding site. In Figure 3.6 B, it has been shown that the best SRCs for series 1 were generated basing on wide channel models, even though the widest one (C4\_25, belonging to cluster 2; see Table S1 in section 3.6.2) displayed a slightly worse performance than the others ( $r^2 = 0.52$ ). It is possible to speculate that an optimal SRC model would be obtained with a moderate amount of restriction in the channel pore. A possible explanation could be that a right compromise between wide and narrow channel volume and shape had to be obtained to allow the present series of compounds to properly dock into hERG. Indeed, the Phe656 ring had to achieve an optimal ensemble of conformations to interact productively with the most potent ligands. Concerning this point, it is worth commenting on the overwhelming population of cluster 1 and the wide range of  $r^2$  displayed by the related SRC models. Cluster 1 entailed a quite broad range of conformations of the channel, whose shape could be crudely described as encompassing both narrow and slightly open pore conformations, showing a maximum radius in the restriction region lower and greater than 2 Å, respectively. Notably, the maximum radius never exceeded the value of 2.2 Å in this cluster. For most of the channel models belonging to cluster 1, the Phe656 restriction prevented the achievement of effective SRC models. However, because of the rearrangement of these amino acids, whenever the pore radius exceeded a critical threshold (of about 2 Å), highly productive SRC models became within reach. These considerations hold for both the series of compounds, and as such, they can be considered of general significance. Clearly, taking into account Phe656 conformations in binding is a necessary but not sufficient condition to explain at best the experimental blocking activities, as the conformations of several other residues comes into play at a molecular level.

As already pointed out in the Results section, no straightforward relationships between the shape of the pore and the symmetry family were identified. In this

respect, an interesting exception was provided by cluster 2. As previously discussed, cluster 2 represented the ensemble of widest channel models, and it was mostly constituted by C4 models. Interestingly, the channel conformation leading to the fittest SRC within this cluster belonged to the C1 point group symmetry for both the series of compounds (C1\_20, see Table S1), demonstrating that the shape, rather than the symmetry used in the channel model generation, is relevant to build effective structure-based models based on an SRC description. Indeed channel model C1\_20 virtually displayed C4 symmetry, whereas no symmetry restraints were imposed in its derivation. In this context, the fact that C4 channels apparently performed slightly worse than those belonging to the C2 and C1 class (see Figure 3.5 for series 1) should not be overestimated. Indeed, the worse performance displayed by C4 channel models is only apparent and is due to the fact that, because of the high number of restraints that must be satisfied in their derivation, there is a greater chance to incur either in narrower or wider pore shapes compared to other symmetries, which were generally associated to less fit SRCs.

### **3.4.2 Binding Modes of the Sertindole Analogues**

From the perspective of the development of the protocol, the MRC-003 model represents the main achievement of this work, as it provided at the same time a consistent binding mode for structurally similar blockers and satisfying structure-activity relationships. Even though the strategy employed to reach this description of binding was not free from a certain degree of knowledge-based subjectivity [185], it must be stressed that the resulting model was obtained in a completely unbiased way. The fact that the performance of some SRC models turned out to be comparable (and in few cases better) to MRC-003 could be interpreted as an evidence of the existence of alternate or multiple binding modes for hERG channel blockers. Although such a perspective is undoubtedly attractive, it is reasonable to think that it is too speculative to be embraced without skepticisms. Indeed, further studies relying on free energy calculations should be carried out to properly address such a hypothesis.

In line with the survey carried out by Zachariae and co-workers [186], it is interesting to compare the binding modes obtained in the fittest structure-based model with those previously presented in the literature and attempt to relate them with the available pharmacophores for the hERG blocking activity. Updating the scenario depicted by previous work of Recanatini et al. [154], several binding modalities for potent hERG blockers have been proposed so far. Therefore, considering sertindole as a reference compound, it is possible to mention:

(1) the perpendicular (with respect to the channel axis) curled solution (Farid-like binding mode) [69], where sertindole adopts a crown shaped conformation delimited by the underneath Phe656 ring;

(2) the parallel extended solution where the fluorophenyl group interacts with Ser624 (Stansfeld-like binding mode) [150];

(3) the parallel extended solution opposite to the previous one, where the imidazolidine group interacts with Ser624 (Österberg-like binding mode) [67];

(4) the perpendicular extended solution where the imidazolidine group interacts with Ser624 (Boukharta-like binding mode) [68].

Accordingly, the binding mode displayed by MRC-003 agrees with the Österberg-like binding solution, which can be at some extent consistent with the pharmacophore model proposed by Cavalli et al. [62,187]. Furthermore, along the many possibilities found in the SRC models, several Stansfeld-like and Boukharta-like binding modes could be identified, whereas no Farid-like solutions were obtained.

### 3.4.3 Performance of the Structure-Based Models

The automated docking protocol was purposely developed using a series of congeneric derivatives, in an attempt of capturing the relevant features of the channel-blocker interactions that might lead to meaningful structure-activity relationship models. A series of analogue compounds should reduce the scoring function uncertainty related to an approximate estimation of entropic and/or solvation contributions. Because of the large amount of unspecific interactions between hERG

channel and blockers, a pose re-ranking based on a statistical (approximated) evaluation of the ligands' configurational entropy was here utilized. The good correlations achieved by the fittest SRCs and MRC-003 ( $r^2$  more than 0.7), together with the consistency of the binding modes found in the latter model, it is a proof of the feasibility of the approach. The fact that SRC-001 and MRC-003 retained good performances after the MM-PBSA rescoring was a further confirmation about the performance of our protocol. One could also envision a protocol where each binding mode is rescored with the MM-PBSA method. However, it should be underlined that such strategy can be highly CPU demanding.

The protocol was then validated using a series of structurally unrelated blockers, covering several different classes of drugs. Indeed, obtaining a good correlation with a general set of blockers is a much more challenging than using a series of analogous compounds. As expected, the docking protocol returned less satisfying results, even though an acceptable SRC model could be obtained ( $r^2 = 0.60$ ). Figure 3.10 A clearly showed that astemizole was a major outlier, and its removal led to a significantly improved correlation with an  $r^2$  more than 0.8 (channel C1\_53, belonging to cluster 3). The relatively high energies attributed to astemizole by the AD score, which in turn led to poor correlations, was likely due to the fact that the docking algorithm was unable to properly pose a relatively large drug into a crowded binding site. In spite of this, the encouraging correlations pointed to an acceptable general applicability of the present protocol to series of compounds in order to achieve predictive qualitative and quantitative insights about their hERG blocking ability.

### 3.5 Conclusions

Achieving a reliable description of the molecular interactions responsible for the hERG channel blockade by drugs represents one of the hottest topics in drug discovery. In recent years, a number of docking studies attempting to unravel the most likely binding modes for several series of hERG blockers have appeared in the literature. In most of these studies, the reliability of the docking solutions has been assessed by the ability to explain the experimental activity either directly, by using docking scoring functions, or indirectly, by relying on more sophisticated rescoring schemes. Eventually, these procedures implicitly provide different structure-based models that might be useful in prospect to predict the blockade potency of newly designed compounds, thus accelerating the identification of LQTS liabilities. Notwithstanding the importance of these docking studies, it is nonetheless surprising to notice the low agreement between binding modes proposed by different authors, especially considering the fact that the overall features of the starting channel models appear to be quite similar. The latter observation, together with the consideration that no systematic docking studies on multiple conformations of the hERG channel has been undertaken yet, has encouraged the development of a specific protocol aimed to overcome several modeling difficulties associated to the problem under investigation [154]. On the one hand, the main achievement of the present work is the unbiased obtainment of a consistent binding modality for a congeneric series of sertindole derivatives which supports one of the most accredited pictures of binding (Österberg-like binding mode). On the other hand, the protocol itself proved to be successful and has allowed to elucidate some methodological aspects of docking so far not explicitly addressed. Concerning this point, it has been found that the conformational symmetry imposed to the amino acids of the binding site can have an impact on the reliability of the structure-based models derived. However, since similar shapes of the channel cavity can be

achieved with different symmetries, the shape of the pore turned out to be a simpler and better descriptor to analyse the relevant structural features of the channel. Since the results show that channel models belonging to the C1 class on average provided the fittest SRCs, it is reasonable to suggest that there is no real need to impose symmetry restraints in the channel models generation. This finding could be considered as a general validity result for symmetrical binding sites, not only in the context of hERG docking. In the attempt to overcome the tendency to achieve geometrically sparse binding modes for structurally similar compounds, a rescoring method suited to reweight the docking score by the probability to achieve a given binding mode, was used and implemented in the protocol. In the present case, the approach turned out to successfully increase the consistency of the docking outcomes and the possibility to rationalize structure-activity relationships for the SRC description. Finally, the results achieved by means of this protocol, have shown the possibility to obtain a minimal subset of channels spanning the relevant conformations of the cavity that are suitable for predicting hERG liability on a series of analogues compounds. These channel conformations would be in turn instrumental to reduce the computational cost in view of the lead optimization phase of drug discovery.

## 3.6 Appendix

### 3.6.1 Presentation of the Automated Protocol CoRK<sup>+</sup>

Here, I will present some features regarding the execution of the automated protocol, named as CoRK<sup>+</sup> (Correlative Re-ranking hERG K<sup>+</sup> channel), for studying hERG blockers.

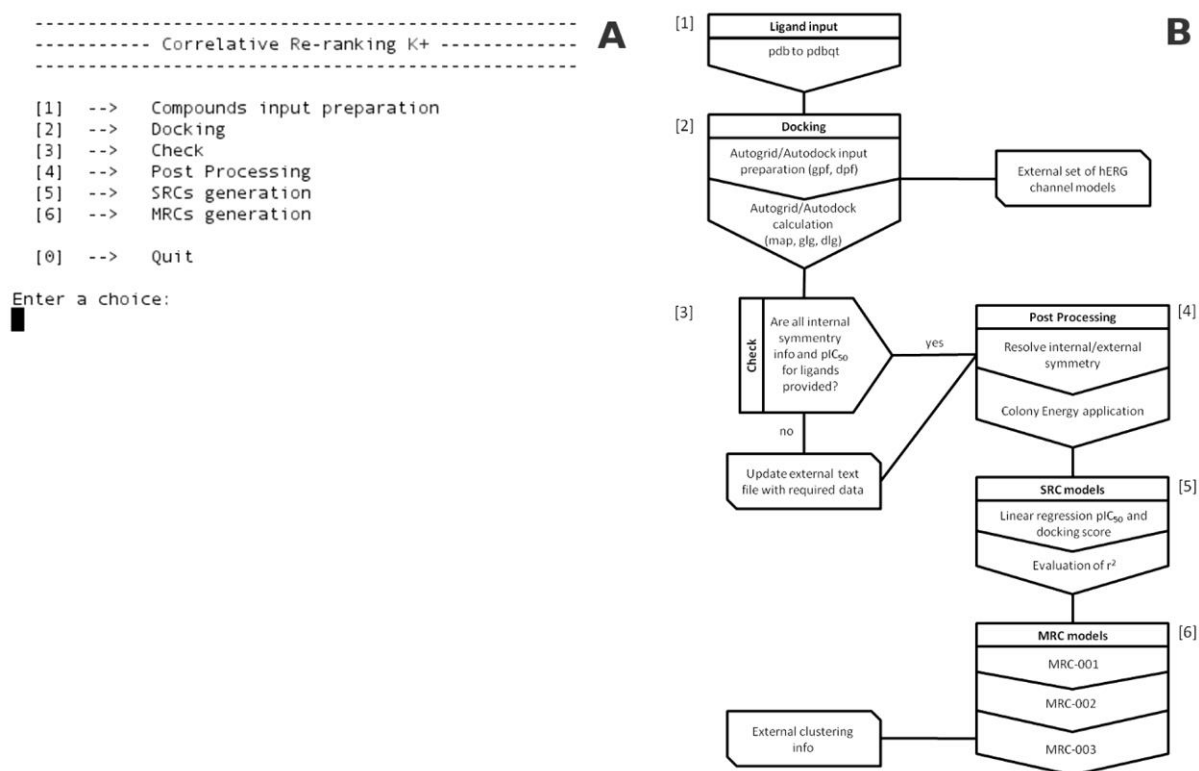
The protocol must be executed in the same folder containing the set of compounds (in the PDB file type). The core of the protocol is a BASH user-friendly interface which manages and calls other BASH, Awk, and Tcl scripts.

A complete description of options 1-6 is reported in the followings (Figure 3.11):

- (1) PDB to PDBQT file type conversion required for Autodock [137].
- (2) Generation of the input files for autogrid/autodock and subsequent docking calculations by using the external stored set of hERG channels.
- (3) The checking phase verifies that all data regarding experimental activities and internal symmetry for each compound are provided in an external file.
- (4) Post-processing of docking outcome. The internal/external symmetry of the docking outcome is resolved by means of Tcl scripts exploiting the text-only displaying device of the VMD environment [188]. The Colony Energy method [158,159] is also applied in this step.
- (5) Evaluation of the SRC models performance: the resulting output file contains the linear regression parameters required for the following step.

(6) Generation of the MRC models by means of a data treatment of docking scores according to the SRCs performance and clustering information, opportunely stored in an external input file.

The protocol is interactive (see Figure 3.11 A) and implemented in such a way that steps 1-6 can be either executed sequentially or modularly in order to optimally take advantage of distributed computational resources. In the latter case, typically, steps 1-2 are executed in separate machines, whereas steps 3-6 are performed after output recollection. In both cases, the total number of autoproductors to be used in step 2 (parallel docking) can be specified via an input flag.



**Figure 3.11** The interactive dialog box (A) Input/output flows (B) of the automated procedure.



### 3.6.2 Supporting Materials

#### Contents:

**Table S1.** Complete list of the SRC models (series 1).

**Table S2.** MM-PBSA free energy contributions for the fittest structure-based models (series 1).

**Table S3.** Fittest correlative models (series 2).

**Figure S1.** Performance of the SRC models before and after applying the CE method (series 1).

**Figure S2.** Binding modes for the MRC-003 model (series 1).

**Figure S3.** Binding modes for the SRC-001 model (series1).

SRC-model	Symmetry	Channel model	Cluster index	$r^2$	SRC-model	Symmetry	Channel model	Cluster index	$r^2$
1	C2	17	3	0.83	70	C4	3	1	0.58
2	C2	48	9	0.79	71	C2	31	6	0.47
3	C2	32	3	0.63	72	C2	26	6	0.37
4	C1	28	1	0.62	73	C2	40	6	0.34
5	C1	4	1	0.72	74	C2	43	6	0.36
6	C1	11	3	0.70	75	C4	5	1	0.54
7	C1	45	3	0.69	76	C2	11	1	0.24
8	C1	43	4	0.74	77	C4	31	6	0.51
9	C4	24	2	0.68	78	C1	7	1	0.26
10	C1	35	3	0.64	79	C1	58	4	0.54
11	C1	44	3	0.47	80	C2	33	3	0.48
12	C1	14	3	0.65	81	C4	7	1	0.43
13	C1	48	7	0.70	82	C1	32	3	0.33
14	C1	54	3	0.59	83	C2	2	1	0.31
15	C1	51	8	0.62	84	C2	47	1	0.41
16	C1	8	1	0.50	85	C4	12	1	0.42
17	C1	53	3	0.56	86	C2	20	1	0.26
18	C4	14	2	0.56	87	C4	18	1	0.49
19	C1	21	3	0.54	88	C2	46	6	0.19
20	C1	17	1	0.61	89	C1	42	6	0.35
21	C4	23	2	0.61	90	C4	21	1	0.43
22	C1	24	3	0.50	91	C4	27	1	0.31
23	C2	10	1	0.56	92	C2	37	1	0.23
24	C1	56	3	0.72	93	C4	16	9	0.31
25	C2	41	1	0.63	94	C2	34	1	0.35
26	C1	1	3	0.40	95	C2	12	1	0.14
27	C4	19	2	0.59	96	C2	4	6	0.13
28	C2	29	1	0.53	97	C4	29	9	0.26
29	C1	26	1	0.52	98	C2	18	1	0.27
30	C2	21	9	0.61	99	C4	30	1	0.33
31	C1	55	3	0.52	100	C1	6	1	0.29

32	C4	9	1	0.53	101	C4	32	1	0.41
33	C4	10	3	0.59	102	C2	45	6	0.30
34	C2	44	1	0.74	103	C2	42	1	0.13
35	C2	28	3	0.52	104	C1	3	1	0.25
36	C2	15	3	0.72	105	C2	13	1	0.21
37	C1	20	2	0.69	106	C1	39	1	0.13
38	C2	9	1	0.56	107	C2	6	1	0.23
39	C1	19	3	0.51	108	C4	6	10	0.27
40	C4	25	2	0.52	109	C1	34	1	0.30
41	C2	27	3	0.57	110	C2	38	6	0.19
42	C2	8	3	0.58	111	C4	13	9	0.30
43	C4	1	1	0.44	112	C2	24	1	0.17
44	C2	23	1	0.58	113	C4	4	1	0.22
45	C2	22	2	0.49	114	C1	49	6	0.24
46	C1	41	6	0.57	115	C4	15	1	0.29
47	C1	36	1	0.41	116	C2	35	6	0.37
48	C1	9	3	0.44	117	C1	52	1	0.42
49	C4	11	1	0.40	118	C4	22	1	0.19
50	C1	23	4	0.45	119	C2	7	1	0.21
51	C2	14	1	0.53	120	C2	16	6	0.22
52	C2	3	3	0.46	121	C1	16	6	0.11
53	C1	22	3	0.28	122	C1	31	1	0.14
54	C1	12	3	0.50	123	C4	26	1	0.19
55	C1	2	3	0.45	124	C1	18	1	0.20
56	C2	30	3	0.46	125	C1	13	6	0.08
57	C1	15	1	0.48	126	C2	39	1	0.16
58	C4	17	1	0.41	127	C2	36	6	0.17
59	C4	8	9	0.50	128	C4	28	1	0.12
60	C4	20	2	0.50	129	C1	27	1	0.20
61	C2	5	1	0.66	130	C1	50	1	0.17
62	C1	29	3	0.50	131	C1	40	6	0.13
63	C1	30	4	0.72	132	C1	47	6	0.13
64	C1	38	1	0.52	133	C1	5	1	0.02

65	C1	46	6	0.44	134	C1	37	1	0.01
66	C2	1	1	0.60	135	C1	10	5	0.04
67	C1	33	1	0.38	136	C2	25	1	0.14
68	C4	2	6	0.31	137	C1	57	1	0.05
69	C2	19	1	0.47	138	C1	25	1	0.00

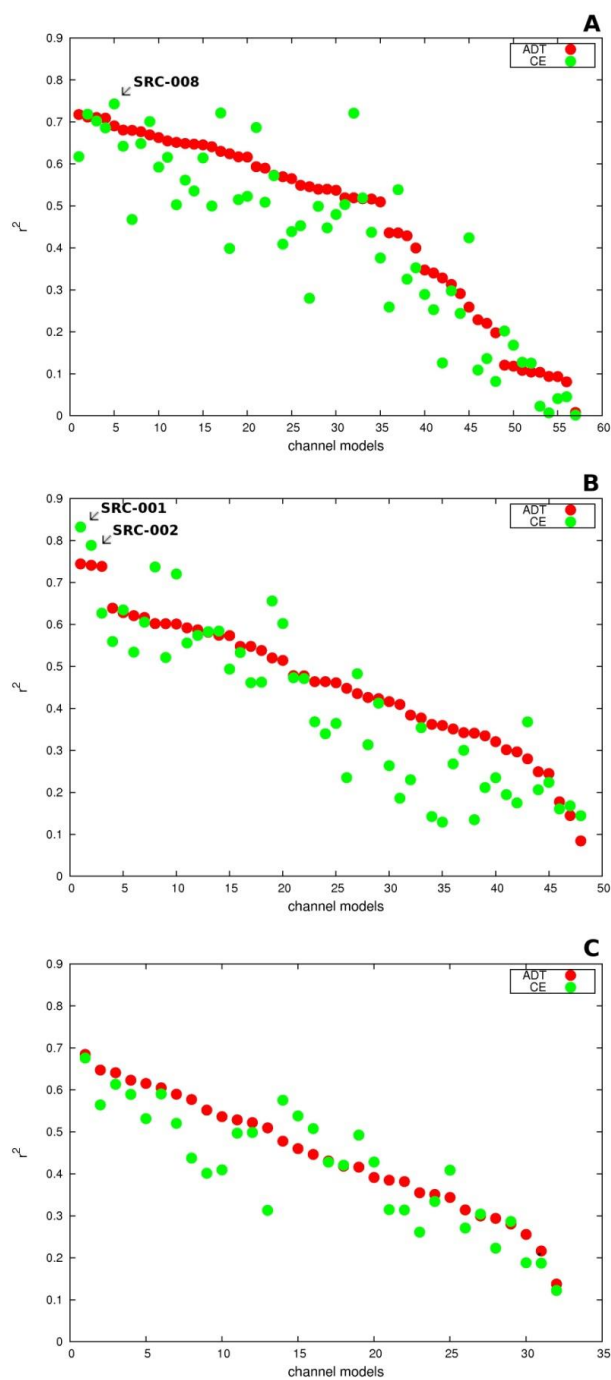
**Table S1.** Complete list of the SRC models. The table summarizes the  $r^2$  calculated for all the 138 SRC models together with information regarding the symmetry class of the corresponding channel models and cluster index. The SRC models are arbitrarily numbered according to their performance before applying the re-ranking procedure.

Compd	SRC-001				MRC-003			
	C2_17				$\Delta G_{MM}$	$\Delta G_{PB}$	$\Delta G_{SA}$	$\Delta G_{bind}$
	$\Delta G_{MM}$	$\Delta G_{PB}$	$\Delta G_{SA}$	$\Delta G_{bind}$				
sertindole	-237.92	173.93	-4.33	-68.31	-245.84	177.18	-4.47	-73.12
<b>1</b>	-195.60	135.76	-4.36	-64.20	-232.75	168.67	-4.04	-68.13
<b>2</b>	-234.11	168.47	-4.21	-69.85	-241.05	174.10	-4.45	-71.40
<b>3</b>	-235.76	172.82	-4.15	-67.09	-230.18	164.52	-4.19	-69.85
<b>4</b>	-43.88	17.02	-4.01	-30.88	-100.47	58.42	-4.24	-46.29
<b>6</b>	-199.57	139.20	-4.36	-64.73	-232.99	164.93	-4.26	-72.32
<b>7</b>	-172.38	111.85	-4.62	-65.16	-233.78	158.34	-5.00	-80.43
<b>13</b>	-167.60	111.10	-3.67	-60.17	-219.14	150.85	-3.94	-72.23
<b>14</b>	-234.43	182.00	-3.40	-55.83	-251.21	193.23	-3.24	-61.21
<b>16</b>	-25.92	5.41	-3.00	-23.51	-37.97	6.36	-2.84	-34.45
<b>17</b>	-30.78	6.46	-3.26	-27.58	-27.66	3.69	-3.00	-26.97
<b>18</b>	-31.88	11.02	-2.80	-23.66	-49.44	23.24	-2.80	-29.00
<b>19</b>	-25.35	4.56	-3.10	-23.89	-37.74	8.72	-2.88	-31.90
<b>20</b>	-32.31	7.67	-3.29	-27.94	-43.48	20.31	-3.44	-26.61
<b>21</b>	-42.24	14.96	-2.93	-30.21	-42.21	14.49	-2.92	-30.64
<b>22</b>	-26.39	5.81	-2.79	-23.38	-34.47	7.35	-2.71	-29.83

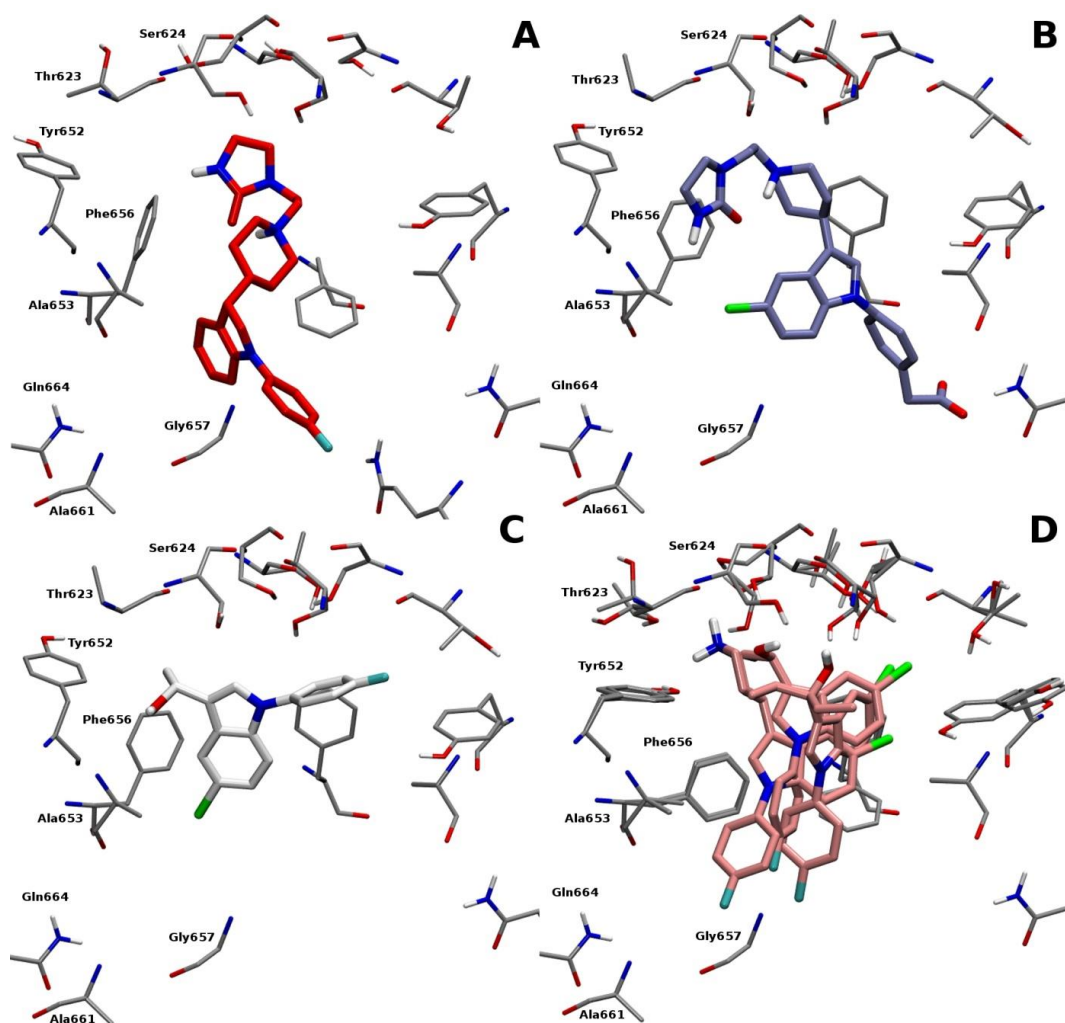
**Table S2.** MM-PBSA free energy contributions for the fittest structure-based models. The table summarizes the various energetic contributions ( $\text{kcal mol}^{-1}$ ) to the free energy of binding. For simplicity, no entropic contributions were considered.

Compd	Fittest SRC	MRC-001	MRC-002	MRC-003
<b>Bepridil</b>	-2.73	-1.81	-2.41	-2.57
<b>Citalopram</b>	-2.95	-2.23	-2.58	-2.70
<b>Clozapine</b>	-4.03	-3.21	-3.69	-3.80
<b>Cocaethylene</b>	-3.17	-2.88	-2.76	-3.03
<b>Cocaine</b>	-2.83	-2.98	-2.81	-2.83
<b>E-4031</b>	-4.62	-3.20	-4.13	-4.23
<b>Fentanyl</b>	-2.53	-2.02	-2.59	-2.55
<b>Fexofenadine</b>	-1.48	-1.04	-1.27	-1.50
<b>Imipramine</b>	-2.58	-2.24	-2.45	-2.50
<b>Ketoconazole</b>	-3.08	-2.11	-2.40	-2.42
<b>Norastemizole</b>	-4.43	-3.60	-4.37	-4.41
<b>Risperidone</b>	-4.45	-3.99	-4.58	-4.60
<b>Ziprasidone</b>	-3.88	-3.42	-3.78	-3.92
<b><math>r^2</math></b>	<b>0.83</b>	<b>0.50</b>	<b>0.75</b>	<b>0.77</b>

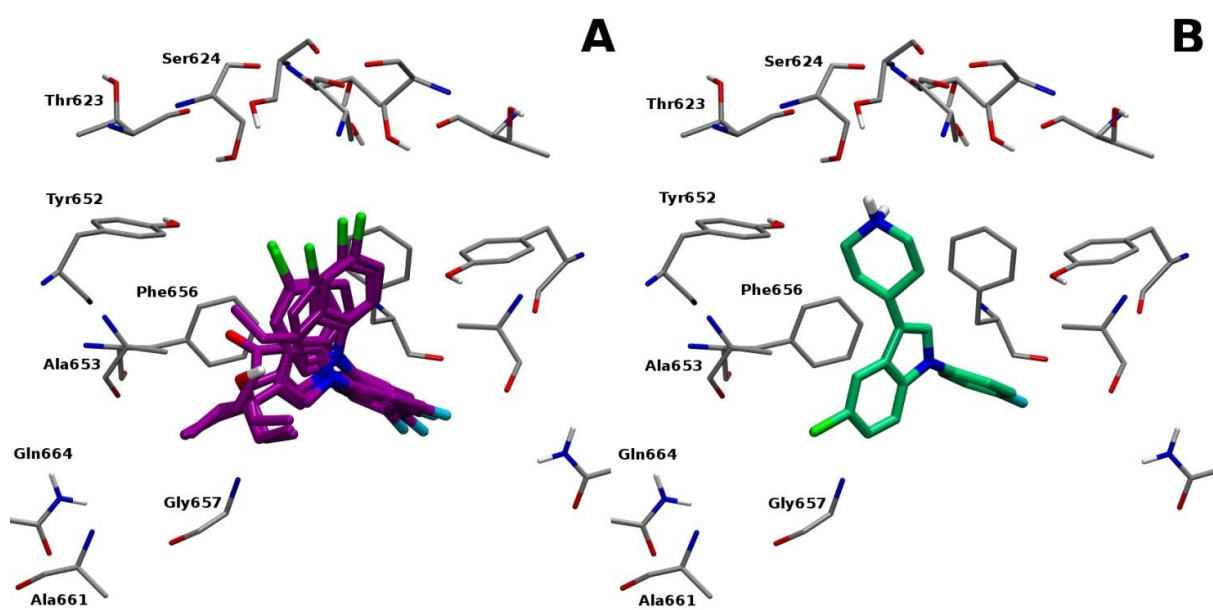
**Table S3.** Fittest correlative structure-based models for series 2. The table summarizes the docking score ( $\text{kcal mol}^{-1}$ ) referred to the set of structurally unrelated compounds for the fittest SRC and MRC models. For the data treatment of the MRC models, the arithmetic mean was employed. The performance of each model measured in terms of  $r^2$  is reported at the bottom of the table.



**Figure S1.** Performance ( $r^2$ ) of the SRC models before (red points) and after (green points) applying the CE re-ranking method. The channel models belong to symmetry classes C1, C2 and C4 are reported in panels (A), (B) and (C) respectively.



**Figure S2.** Binding modes displayed by MRC-003 for **13** (A, magenta), **4**, (B, light blue), **18** (C, white), and **14**, **20**, **21** (D, pink).



**Figure S3.** Binding modes displayed by SRC-001 for 16, 17, 19, 20, 22 (A, violet), and for 14 (B, green).



---

**Chapter 4.****Insights on the Pin1 Peptidyl-Prolyl *Cis-Trans* Isomerization**

---

## 4.1 Introduction

The modulation of protein activity plays a key role for the proper functioning of cellular processes. Over the years, several modulation mechanisms have been established, including autoinhibition, post-translational modifications, allosteric regulations, prolyl *cis-trans* isomerization. Here I will briefly present these protein regulating strategies, focusing on the prolyl *cis-trans* isomerization.

It has been stated that intramolecular interactions between separable elements within a single polypeptide could provide a regulatory strategy for protein function. Within this modulating strategy, known as autoinhibition control, the protein activity is negatively regulated by intramolecular interactions between two different regions of the same protein [189,190]. This strategy is a widespread mechanism, adopted by proteins involved in different biological processes, like signaling, transcription, and transport phenomena. Among these proteins, receptor tyrosine kinases (RTKs) exploit autoinhibition for enabling their inactive configurations and preventing them from high-affinity binding of ligands [190]. On the other hand, another strategy for protein regulation, is represented by allosteric modulation, as in the case of G protein-coupled receptors (GPCRs). It has been demonstrated that the allosteric modulators could promote a certain signalling pathway, stabilizing a certain receptor conformation [191]. Moreover, protein modulation could be also achieved by means of post-translational modifications (PTMs), which reversibly or irreversibly alter the structure, stability and function of proteins, through biochemical reactions, including glycosylation, phosphorylation, acetylation and methylation [192]. In particular, PTMs induce covalently modifications of protein side chains, by addition or modification of chemical groups. The list of protein modifications reported in literature consists of more than 200 entries [193,194]. PTMs have also an important role for the maturation and folding of newly synthesized proteins: this is the case of

reaction mechanisms as glycosylation, lipidation and disulfide bridge formation [195]. Recently it has also been established the role of PTMs in regulating pluripotent stem cells [192].

Contrary to the covalent modifications of PTMs, proline isomerization represents a conformational change process able to control protein function without an alteration of the covalent structure of the protein. Proline residues can exist in two distinct forms, *cis* and *trans*, and their conformational switch is controlled by the prolyl *cis-trans* isomerization process [96]: this phenomena could produce dramatic effects on protein structure and function. Prolyl *cis-trans* isomerization is a rather slow process in normal condition, and can be catalysed by specific enzymes, known as peptidyl-prolyl *cis-trans* isomerases (PPIases). This family of proteins comprises three structurally unrelated subfamilies: cyclophilins (CyPs), FK506-binding proteins (FKBPs), parvulins, and Ser/Thr phosphatase 2A (PP2A) activator PTPA [96,103]. In particular a remarkable number of studies have been focused on CyPs and FKBPs, as they represent the cellular targets of the immunosuppressant drugs cyclosporin A and FK506, respectively [196,197]. Contrary to the other members of the PPIase family, the parvulin Pin1 specifically recognises and binds phosphorylated Ser/Thr-Pro (pSer/Thr-Pro) motifs in the protein substrate. After binding, Pin1 catalyses the *cis-trans* isomerization of proline amide bonds (Figure 4.1), regulating in this way the conformation of its substrate [198,199]. The isomerization is associated with the switching of the  $\omega$  dihedral bond from  $0^\circ$  (*cis*) to  $180^\circ$  (*trans*), passing to a transition state (TS,  $\omega = 90^\circ$ ), and vice versa. Several studies have established the role of Pin1 in diverse cellular processes. In particular, Pin1 has referred to as a “molecular timer” which profoundly affects several biological processes, including cell signalling, gene expression, ion channel gating, neuronal differentiation [96,101,200]. Pin1 has been also implicated in the outbreaks of pathological conditions, including cancer, Alzheimer’s disease, asthma, infection [103,201]. Pin1 is a 18 kDa protein, composed by a single chain of 163 residues, comprising two functional domains separated by a flexible linker (Figure 4.2): a C-terminal PPIase domain, responsible for the catalytic activity, and an N-terminal WW domain, the recognition module involved in protein-protein interactions. It has been determined that the absence of the WW domain, does not affect the catalytic activation and binding of the catalytic

domain [202,203]. However, among the parvulins, the presence of the WW domain is unique to Pin1 [104].

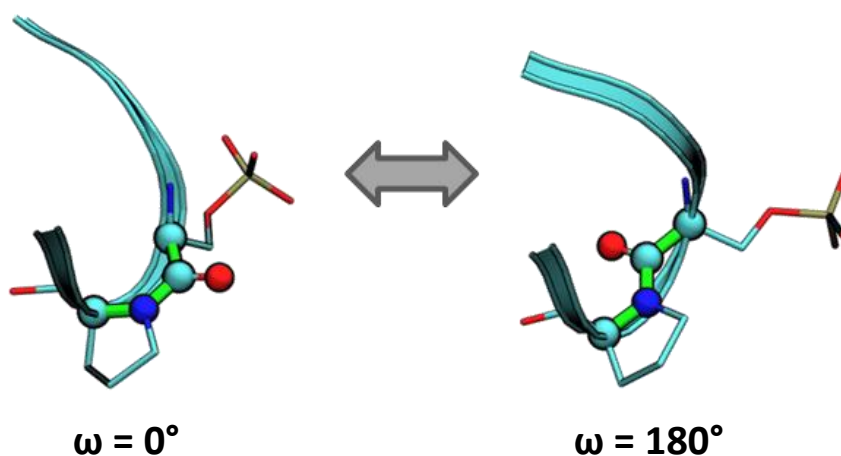
Regarding the structural features of the catalytic domain, it is possible to identify three essential regions (Figure 4.3).

(1) A basic patch, composed by Lys63, Arg68, and Arg69 (net charge 3+), which interacts with the phosphate moiety of the substrate. Interestingly, FKBP and Cyp5 show hydrophobic residues at the positions of the basic cluster [97]. Moreover, mutagenesis studies have suggested that Arg68 is not critical for substrate recognition [104].

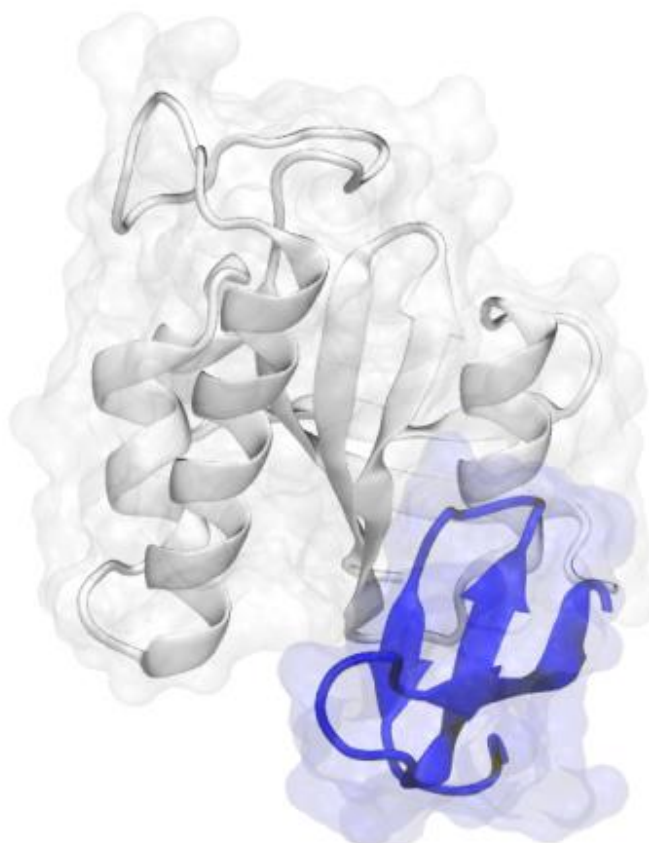
(2) A hydrophobic pocket, composed by Phe134, Met130, and Leu122, representing the binding site for the cyclic side chain of the proline residue of the substrate.

(3) The proper catalytic pocket, consisting in Cys113, His59, His157, and Ser154. In order to investigate the role played by each residue in catalysis, several mutagenesis studies have been performed [97,104,204]. In particular, the direct substitution of Cys113 with Ala, and Ser, revealed a dramatic loss of catalytic activity [97], suggesting a nucleophilic role for Cys113, and leading to hypothesize a covalent mechanism for Pin1 *cis-trans* isomerization. On the other hand, the mutations of Cys113 and H59 with Asp, and Leu, respectively, are found to be functional, allowing to propose a non-covalent mechanism for the catalysed isomerization [104,204]. Details about Pin1 reaction mechanism, and related theoretical studies, are reported in section 4.1.2.3 Parvulins.

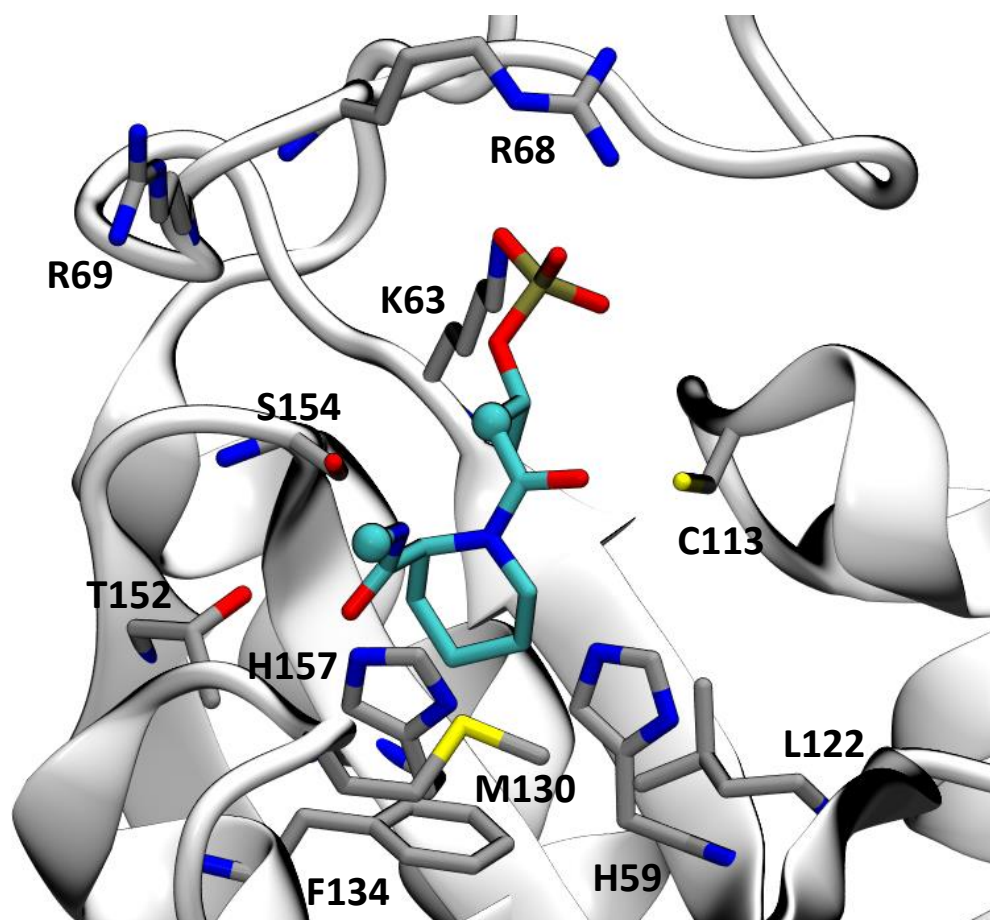
In the next section I will discuss the prolyl *cis-trans* isomerization and non-enzymatic contributions which could accelerate the process, before switching to the PPIase catalysed reaction mechanism.



**Fig 4.1** Pin1-catalyzed *cis* ( $\omega = 0^\circ$ ) to *trans* ( $\omega = 180^\circ$ ) isomerization of pSer/Thr-Pro substrates. The  $\omega$  dihedral bond is reported in green. The picture highlights the changing in conformation, due to the interconversion, in a pSer-Pro peptide.



**Fig 4.2** X-Ray crystal structure of peptidyl-prolyl *cis-trans* isomerase Pin1 (Protein Data Bank (PDB): 2Q5A). The C-terminal catalytic domain is reported in white, while the N-terminal WW, in blue.



**Fig 4.3** Active site of human Pin1 bound to non-natural peptide inhibitor (PDB: 2Q5A). The substrate (in cyan) is partially shown to allow a clear view of the catalytic site: CPK atoms represent the cut-points.

### 4.1.1 Prolyl *Cis-Trans* Isomerization

As introduced before, prolyl *cis-trans* isomerization is a slow process under normal conditions (non-enzymatic isomerization). The high energy barriers associated to the interconversion between *cis-trans* isomers, are mainly due to the partial double bond character of the prolyl amide bond [205]. X-Ray and NMR experiments have revealed that the peptide bond is principally in *trans* state in folded protein [206]. However, it has been shown that 5% of prolyl amide bonds in proteins are in the *cis* states [207]. The *cis-trans* isomerization of the prolyl bond plays a critical role in protein folding, and, as discussed in section 4.1, it is crucial for several biological processes, and for the function of enzymes [208,209].

Over the years, several experimental and theoretical studies have been performed with the aim to elucidate this mechanism of interconversion, as well as the chemical properties of prolyl bond. In particular, the N-acetylproline methylamide (Ace-Pro-NMe) was commonly used in these studies, as it represents the simplest model for a proline containing peptide fragment (see Figure 4.4 A). NMR investigations, performed on a population of *cis* and *trans* isomers, have led to determine a free energy barrier ( $\Delta G^\ddagger$ ) associated to the isomerization in Ace-Pro-Nme, of 20.4 kcal/mol, and a free energy difference between the two isomers ( $\Delta G_{cis-trans}$ ) of 0.57 kcal/mol, with a predominant population of *trans* isomer [210]. The two conformations are therefore close in energy. On the contrary, in non-proline peptide bonds, the *trans* configuration has a higher frequency of occurrence, and the free energy difference between the isomers is approximately 0.5-4 kcal/mol, mainly due to higher steric clashes involving substituents on the C $_{\alpha}$  in *cis* configuration [211,212]. In addition, in 1981, Grathwohl and Wuthrich [213], performed H-NMR studies to measure the rate of *cis-to-trans* interconversion ( $k_{cis \rightarrow trans}$ ) of proline amide bond in linear and cyclic oligopeptides. They found a rate  $k_{cis \rightarrow trans} = 0.0025 \text{ s}^{-1}$  at 25°C, for the zwitterionic form of H-Ala-Pro-OH. Similar kinetics was reported, in 1996, by Eberhardt et al. [214]. Using the N-acetylproline methyl ester (Ace-Pro-OMe) as model for NMR studies, the authors found a rate for the *cis-to-trans*

isomerization in water at 37 °C, equal to  $0.0024 \pm 0.017 \text{ s}^{-1}$ , while a value of  $0.0010 \pm 0.008 \text{ s}^{-1}$ , was determined for the *trans-to-cis* conversion.

Theoretical methods were also widely applied in studying the *cis-trans* isomerization of prolyl amide bond. From a computational point of view, the main challenging task is represented by a proper choice of the degrees of freedom describing the *cis-trans* isomerization process. In 1994, Fischer and co-workers [205] presented a detailed study regarding the isomerization of Ace-Pro-Nme in vacuo, using empirical energy functions and *ab initio* calculations. They showed that an essential element characterizing the transition state is the presence of a dipole moment on the proline nitrogen, which is due to a pyramidalization effect resulting from the shift of the nitrogen lone pair from a  $p_z$  to an  $sp^3$  orbital (nitrogen out-of-plane). Moreover, they found favourable hydrogen bond interactions between this dipole and the C-terminal amide of proline dipeptide, during the isomerization (distance  $d_1$  in Figure 4.4 B). The effect of such interactions is to provide an autocatalytic contribution to the amide bond rotation, lowering the activation free energy of the process. In particular Fischer and co-workers, evaluated a reduction of barrier, due to this effect, of about 1.4 kcal/mol. However, the autocatalysis phenomena was already known by scientific community, as it was first observed during the refolding process of denatured dihydrofolate reductase [215]. Because of the  $\omega$  bond results to be coupled with the out-of-plane of the nitrogen, the improper dihedral  $\zeta$  ( $C_\alpha-C_\delta-O_1-C_1$ , see Figure 4.4 A) was found to better describe the isomerization, in association with the dihedral  $\Psi_{\text{PRO}}$  ( $N_5-C_4-C_\alpha-N_3$ ), which, otherwise, allowed to describe the formation of the autocatalytic intramolecular hydrogen bond. The authors also proposed four theoretically possible transition states for the *cis-trans* prolyl bond isomerization: *syn/exo*, *syn/endo*, *anti/exo*, *anti/endo*. Depending on whether  $C_1$  is on the same or opposite side of  $C_4$  with respect to the averaged plane of the proline, it is possible to obtain the two conformations *syn* and *anti*, respectively. Moreover, the out-of-plane of the prolyl nitrogen  $N_3$  is associated to a change of the orientation of two carbonyl carbons preceding ( $C_2$ ) and following ( $C_4$ ) the Pro residue. The same or opposite orientation of  $C_2$  and  $C_4$ , leads to define the *endo* or *exo* conformations, respectively (Figure 4.4 B). To better understand the isomerization process and the location of the different conformations assumed by a proline containing substrate model on the CVs space, I will present, in Figure 4.5, the



free energy profile (PMF) of Ace-Pro-NMe as a function of the dihedrals  $\zeta$  and  $\Psi_{\text{PRO}}$ , derived by means of the umbrella sampling technique [89,90] (see section 4.2 for computational details). The PMF presented four ground states: two *cis* states (CIS1 at  $\zeta = 0^\circ / \Psi_{\text{PRO}} = -30^\circ$ , CIS2 at  $\zeta = 0^\circ / \Psi_{\text{PRO}} = 150^\circ$ ), and two *trans* states (TRANS1 at  $\zeta = 180^\circ / \Psi_{\text{PRO}} = -30^\circ$ , TRANS2 at  $\zeta = 180^\circ / \Psi_{\text{PRO}} = 150^\circ$ ). CIS1 and TRANS1, have shown the intramolecular hydrogen bond  $\text{H}_1 - \text{N}_3$  compatible with the autocatalytic process. In addition, four saddle points were identified. The transition states TS1 and TS3, both of them with a *syn/exo* configuration, were located at  $\zeta = 90^\circ$ , representing the *syn* area of the PMF. On the other side, TS2 and TS4, located at  $\zeta = -90^\circ$ , have shown an *anti/exo* configuration (see the different conformations of ACE-Pro-NMe in Figure 4.5). The preference of the transition states for *syn/exo* and *anti/exo* configurations, was in line with previous *ab initio* and DFT studies on the proline dipeptide [216]. Among the four saddle points, TS1 and TS2, represented the stabilized transition states. In particular, TS1 was stabilized by the autocatalytic interaction  $\text{H}_1 - \text{N}_3$ , while TS2 by the intramolecular hydrogen bond between  $\text{H}_1 - \text{O}_1$ , which, on the other hand, was responsible to weaken the interaction  $\text{H}_1 - \text{N}_3$ . The computed free energy barriers were almost iso-energetic and approximately equal to 18 kcal/mol. Therefore the isomerization could proceed from *cis* to *trans* passing either through the *syn* ridge at  $\zeta = 90^\circ$  (referred to as counter-clockwise direction), or through the *anti* ridge, centered at  $\zeta = -90^\circ$  (clockwise direction).

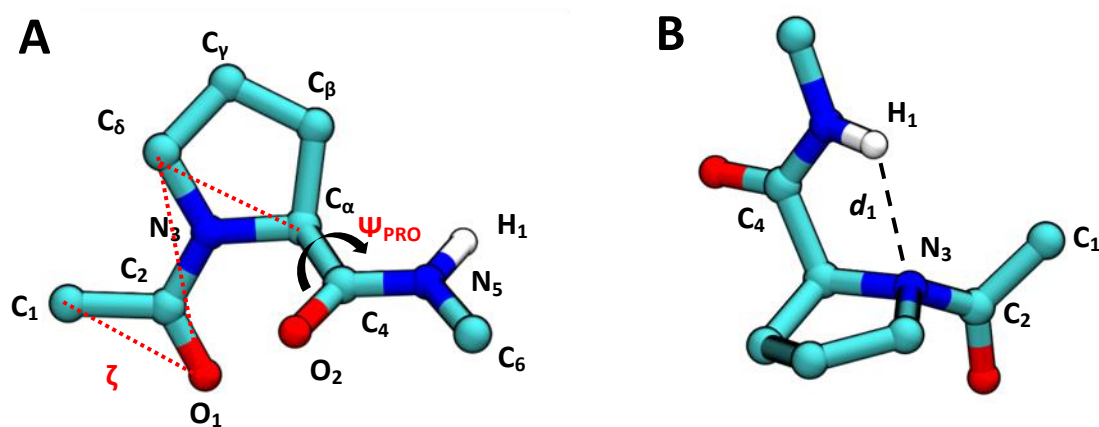
Advances in computing technology and resources, allowed to perform extensive *ab initio* calculations, including also solvent effects, with the aim to achieve new insights on the isomerization process. In this regard, Kang published several works, presenting the results obtained on Ace-Pro-NMe and analogous models, using *ab initio* HF and density functional level of theory, with the conductor-like polarizable continuum model (CPCM) of self-consistent reaction field methods [217-219]. He found that the activation barriers to the *cis-trans* isomerization, strongly depends on the polarity of the solvent: in particular, it has been shown that the rotational barriers increase with the increasing of polarity of the solvent, due to the reduction of the intramolecular hydrogen bond formation between prolyl nitrogen and the C-terminal amide group, suggested as an important factor to stabilize the transition state. In 2008, a hybrid QM/MM molecular dynamics simulation study on

Ace-Pro-NMe, in explicit water, was presented by Yonezawa and co-workers [220]. The free energy profile obtained with the umbrella sampling method, showed that the *trans* configuration is more stable of about 4 kcal/mol than the *cis*, and an activation barrier of 20 kcal/mol. Furthermore they analysed the pyramidalization of the prolyl nitrogen during the amide bond rotation. They found that the proline peptide can assume two typical conformations during the isomerization, stabilized by intramolecular hydrogen bonds and hydration effects: a “positive” pyramidal conformation, in which the prolyl nitrogen moves upward to the plane of the ring, stabilized by the intramolecular hydrogen bond with the C-terminal amide group (distance  $d_1$  in Figure 4.4 B), and a “negative” one, which is the inverse conformation with the distance  $d_1$  compromised. They found that positive pyramidalization typically occurs at  $\omega \cong -90^\circ$  and  $90^\circ$ , while the negative at  $\omega \cong -60^\circ$  and  $60^\circ$  [220]. In 2009, Melis and co-workers [221] evaluated the free energy landscapes of a series of proline analogues. In particular, they performed metadynamics simulations using the dihedrals  $\zeta$  and  $\Psi_{\text{PRO}}$  as reaction coordinates, within a classical description with ESP atomic partial charges [222] calculated at a DFT level. In the case of the proline dipeptide Ace-Pro-NMe, the free energy difference between *cis* and *trans* isomers in water, was found equal to  $1.0 \pm 0.3$  kcal/mol, lower than the value observed in vacuum ( $2.4 \pm 0.2$  kcal/mol). In addition, they reported that the two isomers are separated by a barrier of 15.6 kcal/mol in aqueous solution, lower than the barrier calculated in vacuum (13.8 kcal/mol). A higher activation energy in water is the result of a competition between the formation of the intramolecular hydrogen bond, which promotes the proline nitrogen out-of-plane and the catalysis, and the intermolecular interactions with water molecules [221].

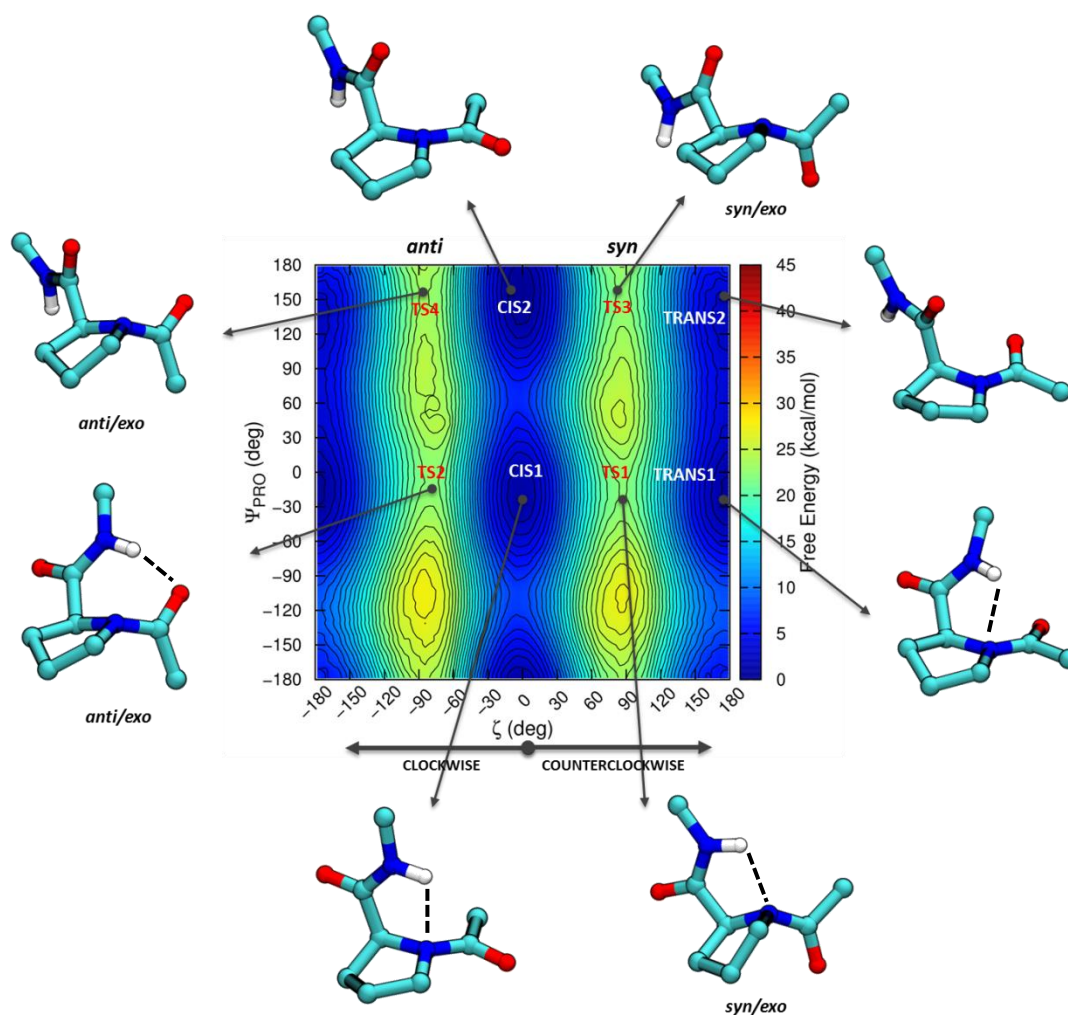
The high barrier to rotation in peptide amide bonds is due to a resonance stabilization, resulting from a delocalization of the lone electron pair of the nitrogen over the peptide bond. An interesting article, published in 2004 by Fanghänel and Fischer [223], reported a series of non-enzimatic effects which could accelerate the prolyl *cis-trans* isomerization, most of them discussed before. In particular, they presented several environmental factors which contribute to enhance the rate of the peptide bond rotation. First of all, the isomerization rate is found to be faster in low polar solvents. A similar effect was observed for the isomerization of N,N-dimethylacetamide (DMA), by shifting the solvent from water to cyclohexane [224]. In

this context, it has been shown, that reducing the possibility of hydrogen bond interactions with the peptide bond oxygen, could significantly decrease the activation free

energy. The isomerization rate is also affected by the nature of the neighbouring aminoacids of the prolyl amide bond. In this case, electron donating substituents close to the peptide carbonyl have the effect to enhance the rate of isomerization, while a deceleration is observed if they are located near the prolyl nitrogen. Moreover, Fanghänel and Fischer also discussed the role of the autocatalysis phenomena, favoured by intra- or intermolecular protonation of the prolyl nitrogen, in the energy barrier of *cis-trans* interconversion: an event fully characterized by the previously reported theoretical studies [205,220,221].



**Fig 4.4** (A) N-acetylproline methylamide (Ace-Pro-NMe), the simplest model for a proline containing peptide, used for *cis-trans* isomerization studies. The dihedrals  $\zeta$  and  $\Psi_{\text{PRO}}$ , are shown in red. (B) The intramolecular bond  $d_1$  which stabilizes and enhances the proline nitrogen pyramidalization, reducing the barrier to rotation. The carbon atoms  $C_1$ ,  $C_2$ , and  $C_4$ , defining the *syn/anti* – *exo/endo* configurations, are also shown.



**Fig 4.5** Free energy profile of Ace-Pro-NMe as a function of  $\zeta / \Psi_{\text{PRO}}$ . The conformations assumed by the peptide model in the *cis* and *trans* basins, as well as in the transition states, are also reported.

## 4.1.2 PPlase Catalyzed Prolyl *Cis-Trans* Isomerization

As discussed before, prolyl *cis-trans* isomerization is a slow process. However, several spontaneously occurring non-enzymatic factors can contribute to increase the rate of the isomerization, hence to promote the catalysis. In the next sections, I will briefly introduce the role played by enzymes belonging to the PPlase family, in prolyl *cis-trans* isomerization. In particular, I will consider the most studied PPlases: Cyps, FKBP, and parvulins. Regarding the latter subfamily, the attention will be focused on the peptidyl-prolyl *cis-trans* isomerase Pin1, which represents the topic of my project.

### 4.1.2.1 Cyclophilins

A great interest has arisen on Cyps subfamily. This is mainly due to the fact that they represent the cellular target for the immunosuppressive drug cyclosporin A (CsA) [225]. However the action of CsA does not directly involve the inhibition of PPlase activity [225]. In particular, human CypA (also known as Cyp18), is also considered as a potential drug target in the treatment of HIV infection, because of its specific interaction with a Gly-Pro motif on the HIV capsid [226]. However the exact mechanism of CypA is not fully clear, and more investigations are required. In the last years, a considerable number of Cyps from different organisms were purified and characterized. Kinetics investigations for almost all Cyps, highlighted a second order rate constant ( $k_{cat}/K_M$ ) between  $10^5$  and  $10^7$   $M^{-1}s^{-1}$ , for the *cis-trans* isomerization, estimated using the oligopeptide substrate Succinyl(Succ)-Ala-Ala-Pro-Phe-p(NA)Nitroanilide [227]. In particular, in the case of human CypA, kinetics analysis performed by using dynamic NMR spectroscopy, have reported a  $K_M^{cis} = 80$   $\mu M$  and  $k_{cat}^{cis} = 620$   $s^{-1}$  for *cis* to *trans* isomerization, and  $K_M^{trans} = 220$   $\mu M$  and  $k_{cat}^{trans}$

= 680 s<sup>-1</sup>, for *trans* to *cis* isomerization [228]. These parameters highlight a higher affinity of CypA for the *cis* isomer than for the *trans*.

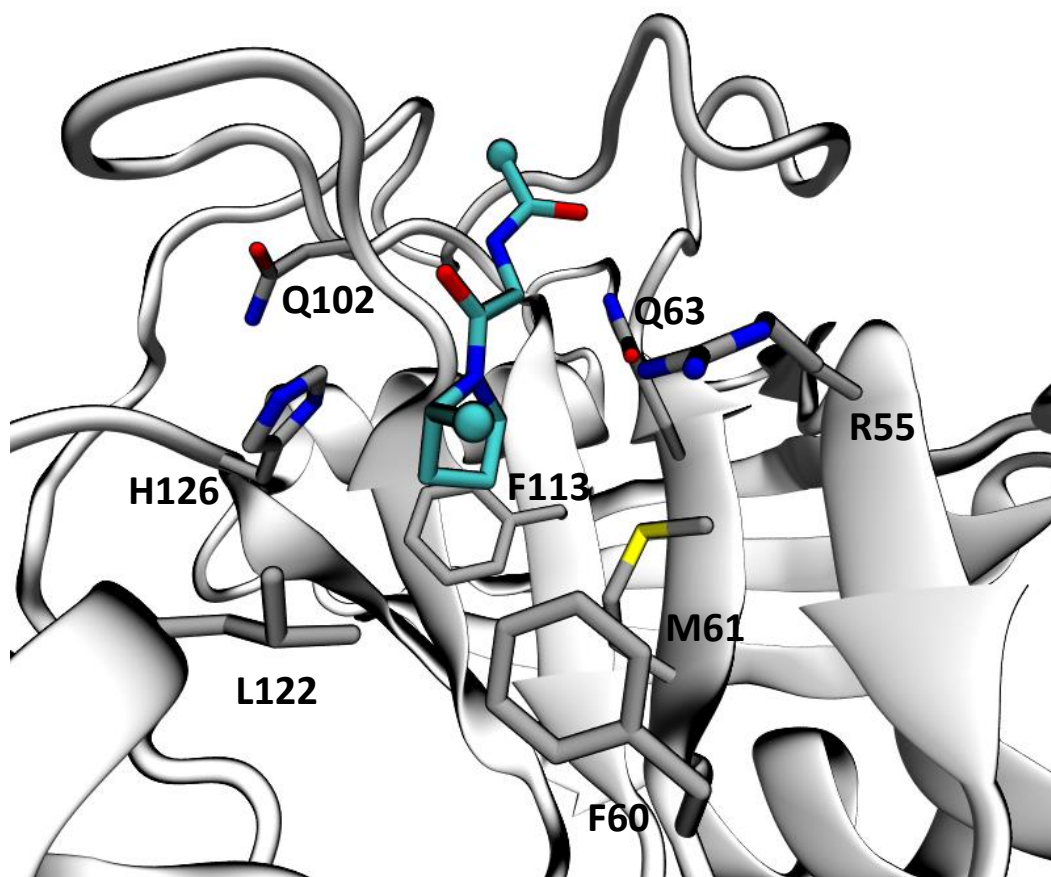
From a structural point of view, the cavity that accommodates the proline pyrrolidine ring consists in four hydrophobic residues: F60, M61, F113, and L122 (Figure 4.6). It has been shown that replacing the proline with a four- or six-membered ring resulted in a dramatic reduction of the efficiency of CypA, highlighting that a correct orientation of the prolyl ring in the TS plays a key role in catalysis [227,229]. Furthermore, mutagenesis studies have shown the importance of R55, F60, H126, in the reaction mechanism [230]. The reduction of activity evaluated for the H126Q mutant, has pointed out that hydrophobic contributions of H126, more than its ability to participate in hydrogen bonds, are essential for substrate binding. Over the years, several reaction mechanisms for CypA *cis-trans* isomerization were proposed on the basis of structural and mutagenesis evidences. In 1989, a nucleophilic catalysis, involving a covalent bond formation between a deprotonated cysteine of the active site and the prolyl bond carbonyl carbon, was proposed [231]. In the resulting tetrahedral intermediate, the double bond character of the prolyl amide bond is compromised, and the effect is a dramatic reduction of the barrier to rotation. This mechanism was supported by experimental studies involving the usage of sulfhydryl group modifying agents [231], and by kinetic experiments [232]. However, mutagenesis data showed that all four cysteine residues of CypA are not essential for the PPIase activity [233]. Furthermore, the closest cysteine is 8 Å far from the substrate, making a nucleophilic attack to the prolyl carbonyl carbon unlikely [234]. In 1990, a “catalysis by distortion” was proposed. This mechanism is based on the idea that the binding energy provided by the formation of the complex CypA/substrate, is exploited by the enzyme to induce a distortion in the substrate (strain), allowing the formation of a twisted prolyl bond [235]. This was supported by the observation of a low enthalpy of activation and a notably large negative entropy of activation. Furthermore the authors pointed out on the important role of the hydrophobic binding pocket to stabilize the twisted conformation. In 1993, Ke and co-workers, proposed a “solvent-assisted” mechanism of *cis-trans* isomerization [236]. This new mechanism was based on the observation of a water molecule hydrogen bonded to the Q63 and the carbonyl oxygen of the prolyl bond, in the CypA/Ala-Pro crystal structure (PDB: 1CYH). The

authors suggested that this hydrogen bond plays a key role to stabilize a twisted transition state, and hence to promote the catalysis. Moreover, the authors observed an hydrogen bond between R55 and the prolyl nitrogen, which reduces the delocalization of the electron cloud along the pseudo-double peptide bond, and therefore contributes to the lowering of the rotational barrier (prolyl nitrogen pyramidalization). This is supported by the reduction of CypA PPLase activity caused by the mutation of R55 with the hydrophobic alanine [230].

The CypA catalyzed *cis-trans* isomerization has been extensively studied theoretically, in order to shed light on this complex scenario. Recently, using the accelerated molecular dynamics method, Hamelberg and McCammon [237] showed that the catalytic mechanism of CypA is mainly due to the stabilization and preferential binding of the transition state, achieved by means of favourable hydrogen bond interactions between the carbonyl oxygen of proline and R55. In addition, the observed proximity of the guanidinium moiety of this residue to the prolyl nitrogen, during the simulation, was proposed as a further enzymatic contribution to the catalytic process, which strengthens the crucial role played by R55. The computed free energy barrier in enzyme was about 10.2 kcal/mol, with a reduction of 6.3 kcal/mol compared to the barrier evaluated for the free substrate (Ace-His-Ala-Gly-Pro-Ile-Ala-NMe) in water (16.5 kcal/mol). However, as commented by the authors, their classical method was not able to capture electronic or desolvation effect on the barrier height [237]. In 2009, Leone and co-workers [238], proposed a novel hypothesis of reaction mechanism for CypA, based on the results achieved by metadynamics simulations. The isomerization in both water and enzyme, was explored using the dihedrals  $\Psi_{\text{PRO}}/\zeta$  of the substrate, as reaction coordinates. They found that in the enzyme, the more stabilized configuration is the *cis* located at values  $\zeta \cong 0^\circ$  and  $\Psi_{\text{PRO}} \cong 180^\circ$ , referred by the authors to as *cis*<sub>180</sub>, which corresponds to the most unfavourable state in water (where the *trans*<sub>0</sub> is identified to be the global minimum). Moreover, they found that the only possible interconversion accelerated by the enzyme, is the isomerization from *trans*<sub>180</sub> to *cis*<sub>180</sub> with an activation barrier of about 12 kcal/mol, 4 kcal/mol lower than the barrier evaluated for the free substrate in aqueous solution. The global minimum in water (*trans*<sub>0</sub>) is therefore sequestered by CypA and interconverted into the *trans*<sub>180</sub>, which is more populated in enzyme. Then, CypA catalyzes the isomerization from *trans*<sub>180</sub>

to *cis*<sub>180</sub>, which could be immediately released or after the interconversion into *cis*<sub>0</sub> [238]. In 2012, Ladani and Hamelberg [239] explored the conformational space of a CypA substrate analogue (Ace-Ala-Ala-Pro-Phe-NMe), in free solution and in the active site of the enzyme, using accelerated molecular dynamics [240]. In particular, they investigated the role assumed by entropy and intramolecular polarizability of the substrate in the catalytic mechanism of CypA. They showed that the conformational space of the substrate in the enzyme, is more restricted than in aqueous solution. Comparing the backbone  $\Phi/\Psi$  conformational phase space of Ala and Pro in the free substrate and in the complex, they found that whereas in water the energetic barrier between  $\alpha$ - and  $\beta$ -regions is low enough to allow an easy interconversion, in the enzyme the  $\beta$ -region results to be predominantly populated by the substrate. Moreover, their results suggested that the relative change in conformational entropy at the transition state, contributes favourably to the free energy of binding. They also found that the intramolecular polarization of the substrate during the reaction mechanism (described as a redistribution of the partial charges at the transition state), contributes only about -1.0 kcal/mol to the stabilization of the transition state. They finally demonstrated that the usage of a classical fixed charge forcefield provides a reliable description of this type of biological systems in which the substrate binding pocket is mainly hydrophobic [239].



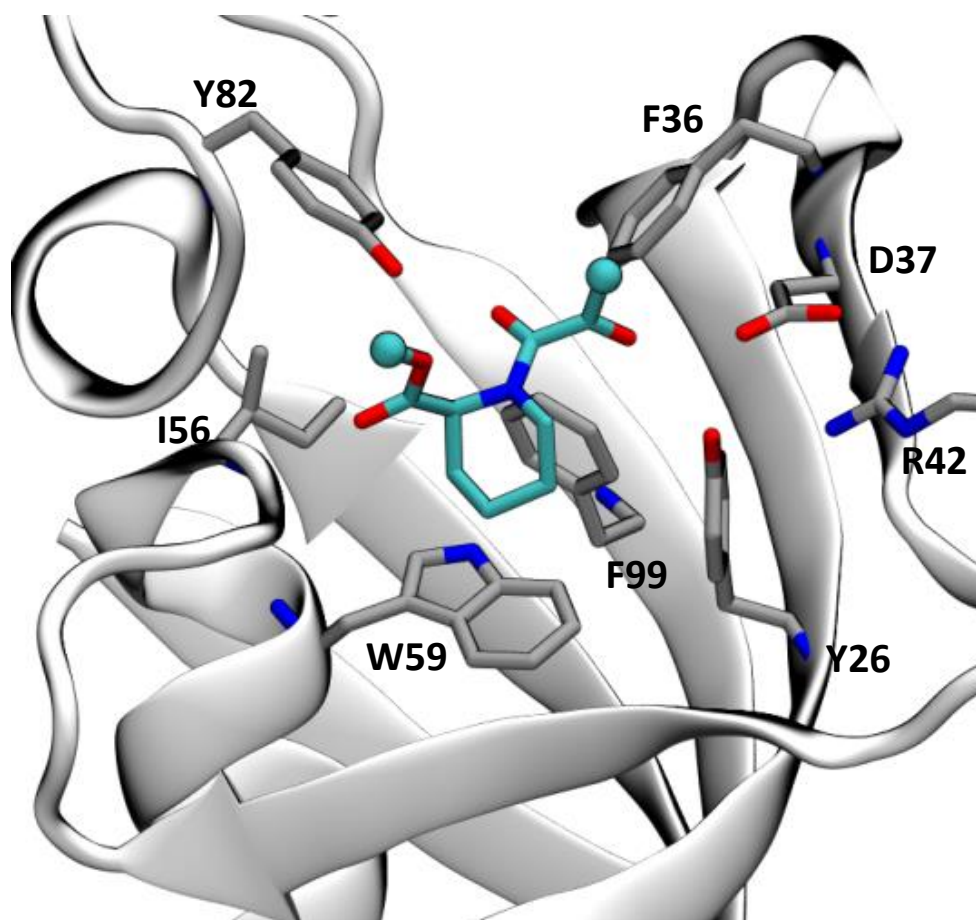


**Fig 4.6** Active site of human CypA complexed with His-Ala-Gly-Pro-Ile-Ala (PDB: 1AWR). The substrate (in cyan) is partially shown to allow a clear view of the catalytic site.

#### 4.1.2.2 FKBP

FKBPs represent another known class of enzymes that catalyze the *cis-trans* isomerization of prolyl peptide bonds. Almost all FKBP investigated so far show a second order rate constant ( $k_{cat}/K_M$ ) between  $10^4$  and  $10^7$   $M^{-1}s^{-1}$  [223]. It has been shown that the observed rate enhancement is achieved by lowering the energy barrier to rotation of about 6.6 kcal/mol [241,242], highlighting a close analogy with Cyps. FKBP are believed to employ a “catalysis by distortion” reaction mechanism, proposed by Harrison and Stein for Cyp-catalyzed *cis-trans* isomerization [223,235].

According to this mechanism, the binding into the enzyme active site induces a strain in the substrate. In addition to this destabilization effect, computational studies led to identify a large number of nonbonded interactions which act together to stabilize preferentially the twisted transition state [243,244]. In particular, Trp59 and Asp37 have been suggested to be directly involved in the substrate binding and catalysis [244]. However, FKBP's catalysis has also several features that are different from Cyp's. In the FKBP's-catalyzed reaction, the activation barrier is characterized by a large enthalpic and a small entropic contribution [229]. Furthermore, Cyp-catalyzed reactions, do not depend on the nature of the residue preceding proline. It has been shown, that hydrophobic residue preceding proline, like leucine or phenylalanine, could increase the FKBP's rate constant of about 100 to 1000 fold, compared to substrate presenting a charged residue in the same position [242,245]. From a structural point of view, the FKBP's active site is made of mainly hydrophobic residues, that are highly conserved throughout the family (Figure 4.7). In particular, Tyr26 and Phe99 form a hydrophobic pocket which accommodates the substrate proline residue [223]. Molecular dynamics and free energy perturbation methods highlighted that the substrate adopt a VIa  $\beta$ -turn conformation within the active site, allowing the formation of an intramolecular hydrogen bond interaction with the prolyl nitrogen [244]. As discussed in section 4.1.2.1, this interaction is thought to lower the energy barrier to rotation (autocatalysis mechanism). Furthermore, these studies suggested the crucial role played by Asp37 to stabilize the transition state, by means of charge-dipole interaction with the carbonyl of the prolyl amide bond. This finding is in total agreement with the reduction of activity achieved by mutating Asp37 with a leucine [246]. However, the main difficulty encountered by computational approaches, is related to the absence of a resolved crystal structure of FKBP-substrate complex, and therefore the substrate binding has been modelled by analogy with inhibitors [247]. Recently, Alag and co-workers [248] provided a crystallographic structure of the human malarial parasite *Plasmodium vivax* FK506 binding protein 35 (PvFKBP35) in complex with the tetrapeptide substrate succinyl-Ala-Leu-Pro-Phe-*p*-nitroanilide (sALPFp) determined at 1.65 Å resolutions. This provides new insightful information regarding the enzyme-substrate binding, which could be exploited for further investigations into the FKBP's *cis-trans* isomerization mechanism.



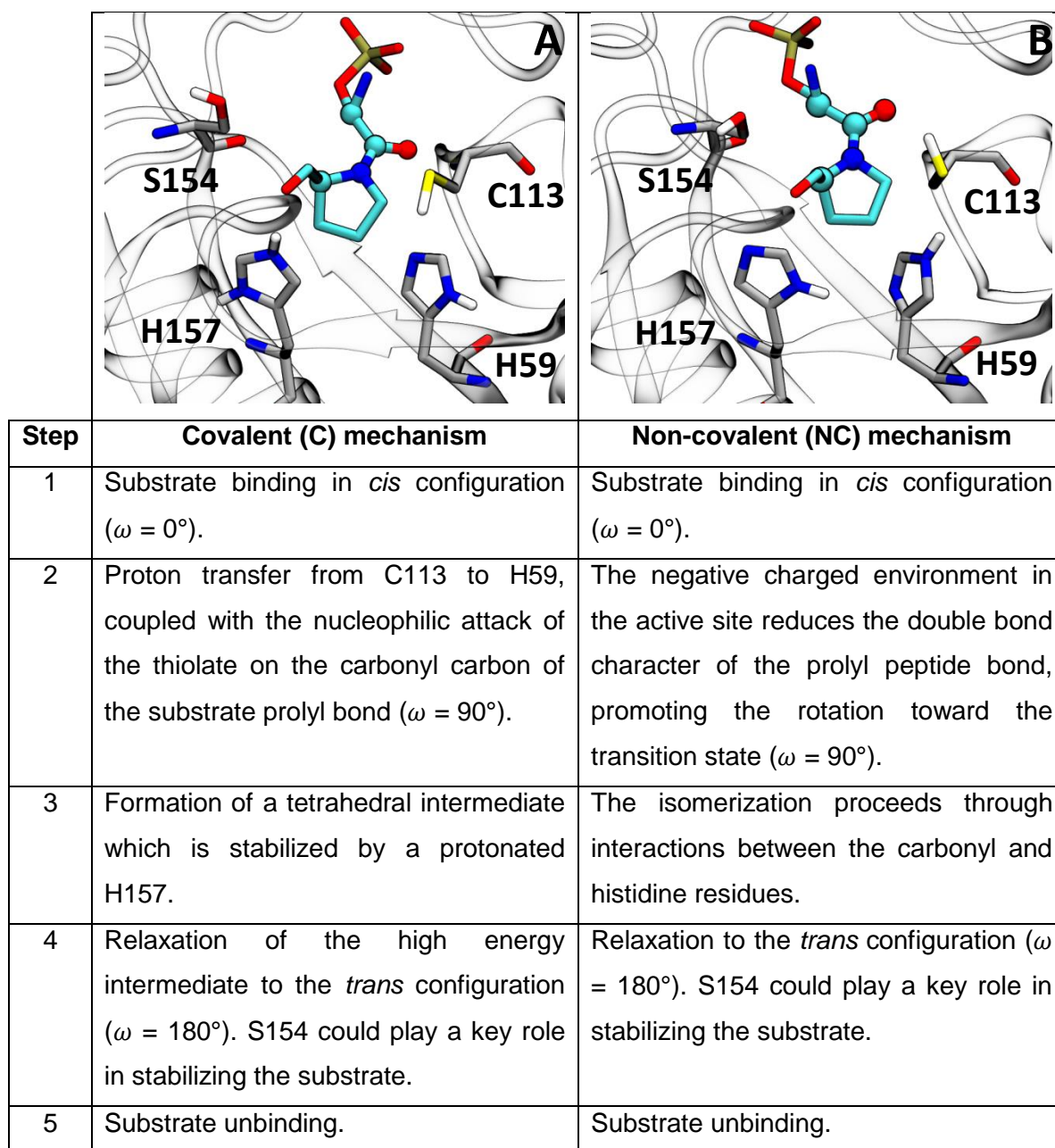
**Fig 4.7** Active site of the human FKBP-12 complexed with the immunosuppressant drug rapamycin (PDB: 1FKB). The substrate (in cyan) is partially shown to allow a clear view of the catalytic site.

### 4.1.2.3 Parvulins

Parvulins represent a third class of peptidyl-prolyl *cis-trans* isomerases, conserved from bacteria to man [249]. The first member of this family was identified in 1994, in *Escherichia Coli* Par10 [250], which is a very low molecular mass enzyme (10.1 kDa) compared to other PPlases: the name parvulin was derived from Latin word *parvulus*, meaning “very small”. In fact, *E. Coli* Par10 consists of the minimal number of amino acid residues (92 residues) facilitating the *cis-trans* isomerization of prolyl amide bonds. Regarding substrate specificity, *E. Coli* Par10 has shown a great analogy with FKBP PPlase family, revealing a preference for bulky, hydrophobic side chains preceding the proline residue [249]. Human Pin1 (also known as Par18) was

identified in 1996 [251]. The differences in substrate specificity for Pin1 and Par10 suggested the existence of two parvulin subfamilies [223,252]. Pin1-type enzymes show a striking preference for protein substrates containing phosphorylated side chains of serine or threonine residues preceding the proline position (pSer/Thr-Pro) [97,98,199,253]. In particular, enzymes that are responsible for Ser/Thr-Pro motifs phosphorylation, an event which plays a key role during cell division and signal transduction, belong to the Pro-directed protein kinases family, including cyclin-dependent protein kinases (CDKs), mitogen-activated protein kinases (MAPKs), Jun N-terminal protein kinases (JNKs) and glycogen synthase kinase-3 (GSK-3) [201]. It has been shown that Pin1 preferentially isomerizes proline residues preceded by pSer/Thr with up to 1300-fold selectivity, compared with non-phosphorylated substrates [98]. Moreover, it has been determined that Pin1 accelerates the *cis-trans* isomerization of pSer/Thr-Pro bonds with a catalytic efficiency ( $k_{cat}/K_M$ ) of  $10^7 \text{ M}^{-1}\text{s}^{-1}$ , measured using a Ala-Ala-pSer-Pro-Arg-pNa substrate [98,223]. Furthermore, NMR measurements on pThr-Pro peptide have demonstrated, besides an acceleration of the isomerization by over 1000-fold over the typical uncatalyzed rate (to the timescale of seconds), that the catalyzed *cis* to *trans* rate is 10-fold faster than the inverse, *trans* to *cis* [254]. As discussed in section 4.1, human Pin1 has been subjected of intense biochemical and clinical research as it seems also to be involved in the outbreaks of different pathological conditions, like cancer and Alzheimer's disease [102,103,198,201]. Despite the great interest on this enzyme, the reaction mechanism has been long debated. In particular, two models for the isomerization process have been hypothesized (Scheme 4.1). Based on the decrease in  $k_{cat}/K_M$  as result of mutating the Cys113 to Ala and Ser, in 1997 Ranganathan and co-workers suggested a covalent (C) mechanism of reaction for Pin1 isomerization [97]. According to this mechanism (Scheme 4.1 A), the deprotonation of Cys113 by the nitrogen of His59 leads to the formation of a thiolate side chain, which acts as the nucleophile of the reaction. This step is followed by a nucleophilic attack on the carbonyl carbon of the amide bond of the substrate, involving the formation of a tetrahedral intermediate. Interestingly, the authors proposed the presence of a charged His157 within Pin1 active site, which stabilizes the tetrahedral intermediate. The typical Pin1 active site setup for the suggested C mechanism (as shown in Scheme 4.1 A), is hereafter referred to as "model A". This

C mechanism for the catalysed isomerization was called into question by Behrsin and co-workers [104]. In particular, they showed that Pin1 catalytic activity is retained after the mutation of the Cys113 in Asp, relieving Cys113 of its role as nucleophile. A non-covalent (NC) mechanism was therefore suggested (Scheme 4.1 B). According to this hypothetical model, the transition from *cis* to *trans* state is promoted by a negatively charged environment in the active site, which disfavours the double bond character of the proline amide bond of the substrate, speeding up the isomerization. The structural features of Pin1 active site which are compatible with the NC mechanism, are referred to belong to the “model B” setup (see representation in Scheme 4.1 B). Recent theoretical studies on the catalyzed and uncatalyzed reaction mechanisms, have been reported by Vöhringer-Martinez et al. [105] and Velazquez et al. [106]. By means of a QM/MM – Mean Reaction Force protocol [255], Vöhringer-Martinez and co-workers, were able to identify structural and electronic contributions to the free energy barrier during isomerization. Nevertheless, as reported by the authors, the method provided similar activation barriers of the *cis* isomer for the catalyzed and uncatalyzed reactions, in contrast with NMR measurements: it has been experimentally determined that Pin1 reduces the activation barrier of about 7 kcal/mol [99]. On the other hand, by application of accelerated molecular dynamics [240], Velazquez and Hamelberg showed a lower barrier for the enzymatic isomerization, due to the fact that Pin1 preferentially binds the transition state configuration of the substrate. In this context, they obtained a free energy barrier for the catalyzed reaction of about 13 kcal/mol, compared to approximately 20 kcal/mol for the uncatalyzed isomerization. Moreover, the authors highlighted the crucial roles of the phosphate binding pocket aminoacids, Lys63 and Arg69, in stabilizing the transition state.



**Scheme 4.1** Schematic presentation of the two mechanism proposed for the Pin1 *cis-trans* isomerization: (A) C mechanism suggested by Ranganathan and co-workers [97]; and (B) the NC proposed by Behrsin et al. [104]. The tautomeric states of the active site residues, models A and B, are also shown. This scheme considers the *cis* to *trans* isomerization path.

### 4.1.3 Aim of the Work and Project Presentation

Is the Pin1 catalytic mechanism closely related with the catalytic pathways used by the other PPlase enzymes? Do PPlase family of enzymes share a common reaction mechanism? Despite the remarkable efforts spent in understanding the Pin1-catalyzed *cis-trans* isomerization of the prolyl amide bond, by experimental and theoretical studies, these questions are still not fully answered, and more investigations are therefore required. The aim of this project consists in comparing, and hence testing, the two models proposed for the isomerization process catalyzed by Pin1 (Scheme 4.1). For this purpose, extensive unbiased molecular dynamics simulations were carried out on two Pin1/substrate complexes, in which the tautomeric states of the active site residues were opportunely modelled according to models A and B clearly shown in Scheme 4.1. An accurate analysis of resulting trajectories, in terms of distances between substrate and binding site aminoacids, suggested that model B is the only model suitable for describing the isomerization. These findings straightforwardly allowed to achieve a reliable starting structure of Pin1/substrate complex to study the catalyzed isomerization, which therefore follows a NC reaction mechanism. To this aim, the umbrella sampling method [89,90] was exploited to investigate both the catalyzed (NC) and uncatalyzed (bulk) reactions.

## 4.2 Computational Details

In this section I will report the details regarding the setup of the two models, A and B, and the computational strategies used for studying the Pin1 *cis-trans* isomerization.

### 4.2.1 Substrate and Models Setup

The 1.5 Å resolution X-ray crystal structure of Pin1 bound to a non-natural peptide inhibitor (PDB: 2Q5A) [256] was used as starting structure in this study. The side chains of the inhibitor were modified in order to match with the substrate model sequence Ace-Ala-Ala-pSer-Pro-Phe-NMe, preserving the initial coordinates and position inside the catalytic site. The choice of this substrate was dictated by the high binding specificity and the isomerase activity exhibited by Pin1 against this pSer-Pro containing peptide [98]. In the crystal structure, the prolyl amide bond of the peptide was in the *cis* state with  $\omega = 0^\circ$ . The substrate was then built using TLEAP package from AmberTools13 [119], and Homeyer et al. parameters [257] were assigned to the phosphorylated serine pSer. The substrate was solvated in a 36 Å x 41 Å x 40 Å TIP3P [258] water model box, and neutralized with two Na<sup>+</sup> ions. This led to the preparation of the substrate-water model, the bulk system, for studying the uncatalyzed reaction mechanism. For the catalytic models setup (models A and B in Scheme 4.1), the WW domain was removed from the crystal structure, and the tautomeric states of His157 and His59 opportunely modified. In particular, model A (Scheme 4.1 A) was prepared according to the tautomeric states of the active site residues before step 2, representing the step of the proton transfer between Cys113 and His59 [97]. Specifically, Cys113 was considered in its neutral state, His157



charged (HIP), and His59 mono-protonated with the nitrogen in  $\epsilon$ -position (HID), as in Scheme 4.1 A. On the other hand, for model B, representing the setup for the NC mechanism, both the histidine residues were modelled as mono-protonated, with the nitrogen in  $\epsilon$ -position (HID) for His157, and in  $\delta$ -position (HIE) for His59 (Scheme 4.1 B). In particular, this setup was chosen according to the hydrogen-bonding network showed by the 0.8 Å high resolution X-ray crystal structure of Par14, a non-phosphospecific PPIase member of the parvulin family [259]. From a structural point of view, the catalytic sites of the two isomerases present a different residue in position 113. In fact, Pin1 Cys113 is replaced by an aspartate in Par14. Because the mutation Cys113 in aspartate [104] has been shown not to compromise Pin1 isomerase activity, the same tautomeric states for the histidine motif, as observed in the crystal structure of Par14, were used in model B. The resulted Pin1/substrate complexes were subsequently solvated in a 74 Å x 74 Å x 74 Å TIP3P [258] water model cubic box. For all the systems, the Amber ff99SB-ildn forcefield [118], and reoptimized dihedral parameters for the peptide  $\omega$ -bond angle [260] were used.

## 4.2.2 Testing the C and NC Reaction Mechanisms

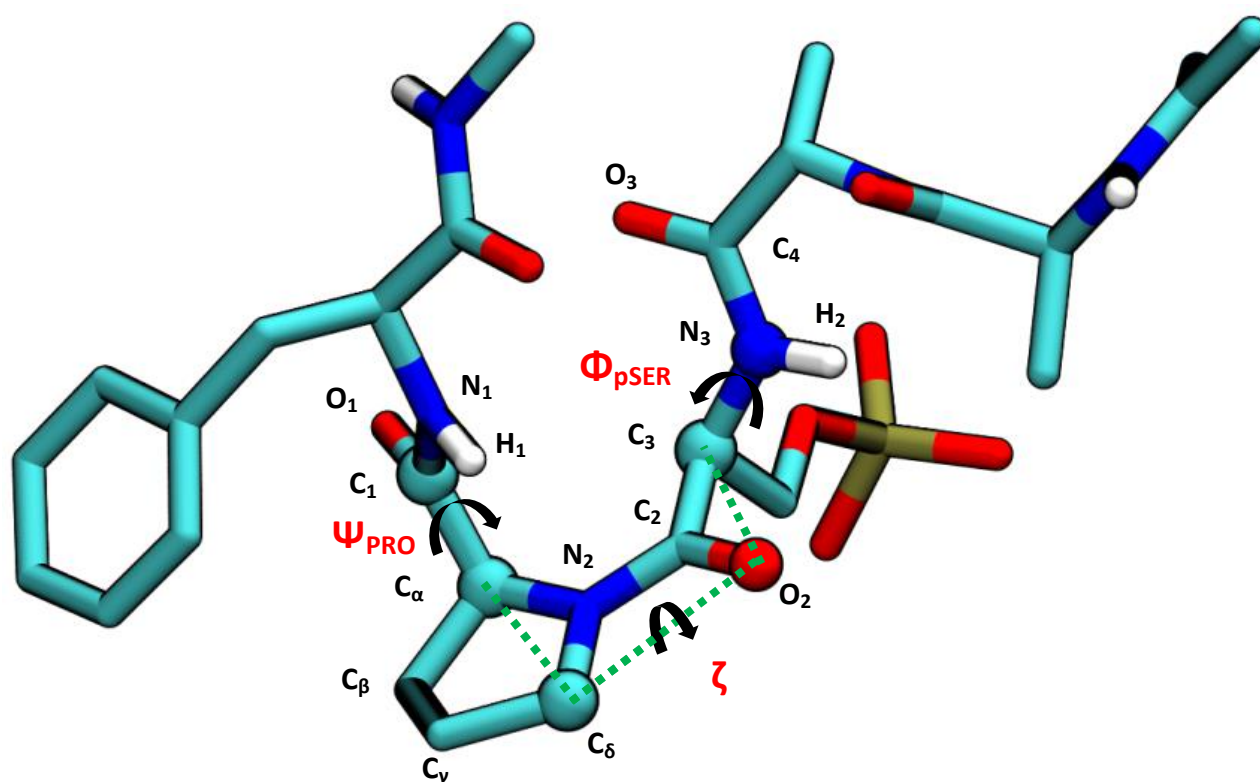
For testing the C and NC reaction mechanisms, classical unbiased MD simulations were performed on the two complexes (model A and B, bound to the peptide substrate), using Amber12 package [119] (Table 4.1). The systems were first minimized and subsequently equilibrated for 1 ns in the isothermal-isobaric NPT ensemble. In particular, the pressure was kept to 1 bar, and the Langevin thermostat [261] with a collision frequency of 2.0 ps<sup>-1</sup> was used to maintain the temperature at 300 K. The SHAKE algorithm [127] kept constrained bonds involving hydrogen atoms. The Particle Mesh Ewald scheme [262] was used to treat long-range electrostatics, while short-range interactions were calculated with a nonbonded cutoff of 10.0 Å. The Langevin equation of motion was integrated every 2 fs using the Verlet integrator. For each system, the equilibration was followed by a production phase of 100 ns within the canonical ensemble NVT (Table 4.1).

Mechanism	System	Unbiased MD
C	Model A + substrate	100 ns, NVT, 300 K
NC	Model B + substrate	100 ns, NVT, 300 K

**Table 4.1** System setup for the hypothesized reaction mechanisms and related unbiased MD simulation details.

### 4.2.3 Investigation of the Bulk and NC *Cis-Trans* Isomerization

The bulk and NC reaction mechanisms for the prolyl bond isomerization was investigated using umbrella sampling [89,90]. The dihedral angle  $\Psi_{\text{PRO}}$  ( $\text{N}_1\text{-C}_1\text{-C}_\alpha\text{-N}_2$ ) and the improper angle  $\zeta$  ( $\text{C}_\alpha\text{-C}_\delta\text{-O}_2\text{-C}_3$ ) of the substrate, were chosen as collective variables (Figure 4.8). These angles have been shown to properly describe the *cis-trans* isomerization [205]. Despite the *cis-trans* isomerization involves the rotation around the prolyl amide bond, the  $\omega$  dihedral is not properly efficient in describing the reaction mechanism, as it results to be coupled to the out-of-plane of the proline nitrogen (pyramidalization). A properly use of  $\omega$  requires the specification of a second reaction coordinate  $\eta$  ( $\text{C}_3\text{-C}_\alpha\text{-N}_2\text{-C}_\delta$ ), an improper dihedral angle which describes the pyramidalization. Furthermore, the selection of the torsion  $\Psi_{\text{PRO}}$  as collective variable is justified to the fact that this degree of freedom controls the interaction between carbonyl group and nitrogen lone pair of the proline, leading to the nitrogen pyramidalization which reduces the free energy barrier to rotation (autocatalysis [205]). The umbrella sampling was performed, for the bulk and NC isomerizations, with Amber 12 [119], harmonically restraining the reaction coordinates by steps of  $15^\circ$  for a total number of 576 windows, and a simulation time of 230 ns. A force constant of  $0.043 \text{ kcal/mol/deg}^2$  was used to allow a satisfactorily sampling of all the configurational space defined by the selected collective variables. The free energy profile along the reaction coordinates, or potential of mean force (PMF), was extracted by means of the Weighted Histogram Analysis Method (WHAM) [132].



**Figure 4.8** The substrate model Ace-Ala-Ala-pSer-Pro-Phe-NMe, and the dihedral angles essential for describing the prolyl isomerization, in red. The improper dihedral angle  $\zeta$  is shown in green dotted line.

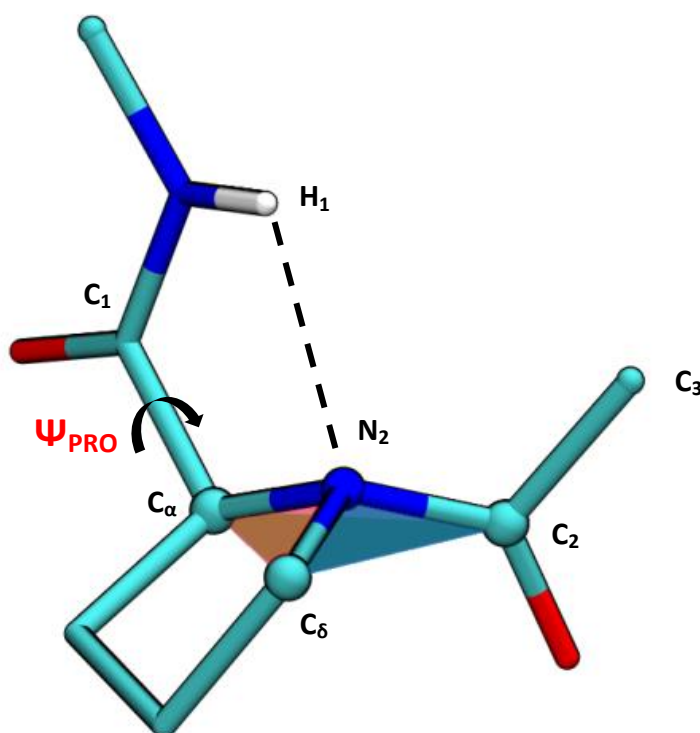
#### 4.2.4 Prolyl Nitrogen Pyramidal Conformations

To analyse the degree of pyramidalization of the prolyl nitrogen during the *cis-trans* isomerization process, the method proposed by Yonezawa et al. [220] was exploited. The authors defined the degree of pyramidalization, referred to as pyramidalization, as:

$$\text{pyramidalization} = \det[\mathbf{a1}, \mathbf{a2}, \mathbf{a3}] \quad (4.1)$$

where  $\mathbf{a1}$ ,  $\mathbf{a2}$ , and  $\mathbf{a3}$ , are unit vectors, whose directions are from  $N_2$  to  $C_\alpha$ ,  $C_\delta$ , and  $C_2$ , respectively (Figure 4.9). In other words, the pyramidalization is defined as the volume of the tetrahedron having a triangular base delimited by the atoms  $C_\alpha$ ,  $C_\delta$ ,

and  $C_2$ , and the prolyl  $N_2$  as the apex. The sign of the determinant allowed to gain information regarding the directionality of the nitrogen out-of-plane, and hence to discriminate between positive and negative pyramidal conformations. As shown in Figure 4.9, for  $\Psi_{\text{PRO}} \cong 0^\circ$ , the positive pyramidalization results to be stabilized by the intramolecular hydrogen bond between  $H_1 - N_2$ . The implementation of equation (4.1) was carried out by means of Tcl scripting language, allowing a fully automated evaluation of the pyramidalization among the MD trajectories.



**Figure 4.9** Typical positive pyramidal conformation of the prolyl nitrogen. The atoms  $C_\alpha$ ,  $C_\delta$ , and  $C_2$ , define the base of the tetrahedron, while the prolyl  $N_2$  the apex. The directionality of the pyramidalization is given by the sign of the determinant in (4.1), which results to be positive in this case.

### 4.2.5 Free Energy Difference between Bulk and NC *Cis* States

For the evaluation of the free energy difference between the conformations assumed by the substrate, in Bulk and NC systems, in the starting *cis* configuration, an umbrella sampling in the space of the RMSD [263] was performed by means of Plumed 1.3 patch [264] for Amber 11 [265]. Using the NC conformation as reference structure, the RMSDs of the peptide in bulk and within Pin1 active site, were harmonically restrained by steps of 0.05 Å, starting from 0.00 to 5.00 Å (100 windows), using a force constant of 150 kcal/mol/Å<sup>2</sup>. In particular, for the RMSD calculation, the structures were first aligned on the prolyl heavy atoms (non-hydrogen atoms), while the measure was performed on all the peptide heavy atoms. The total simulation time consisted of more than 20 ns for each simulation, the bulk and the NC. Flat-bottom dihedral restraints were also applied on the equivalent units of the substrate, for a more rigorous estimation of the RMSD [263]. The PMFs in the space of the RMSD were then extracted by means of WHAM [132].

### 4.2.6 Energy Barrier to Disrupt the Intramolecular Hydrogen Bond

Contrary to the NC isomerization, an intramolecular hydrogen bond between H<sub>1</sub>-O<sub>3</sub> (see Figure 4.8 for atom numbering) was found during the reaction in bulk solvent. To estimate the energy necessary to disrupt this interaction, an umbrella sampling on the space of the dihedral  $\Phi_{\text{pSER}}$ , found to be responsible of such interaction, was performed. Starting from the conformation of the peptide in the bulk *cis* state, the umbrella sampling was carried out with Amber 12 [119], harmonically restraining  $\Phi_{\text{pSER}}$  by steps of 10°, until the value the dihedral assumed in Pin1 active site was reached (from 80° to -150°). The same parameters reported in section 4.2.3 were used, and the mono-dimensional PMF was extracted by means of WHAM [132].

## 4.3 Results

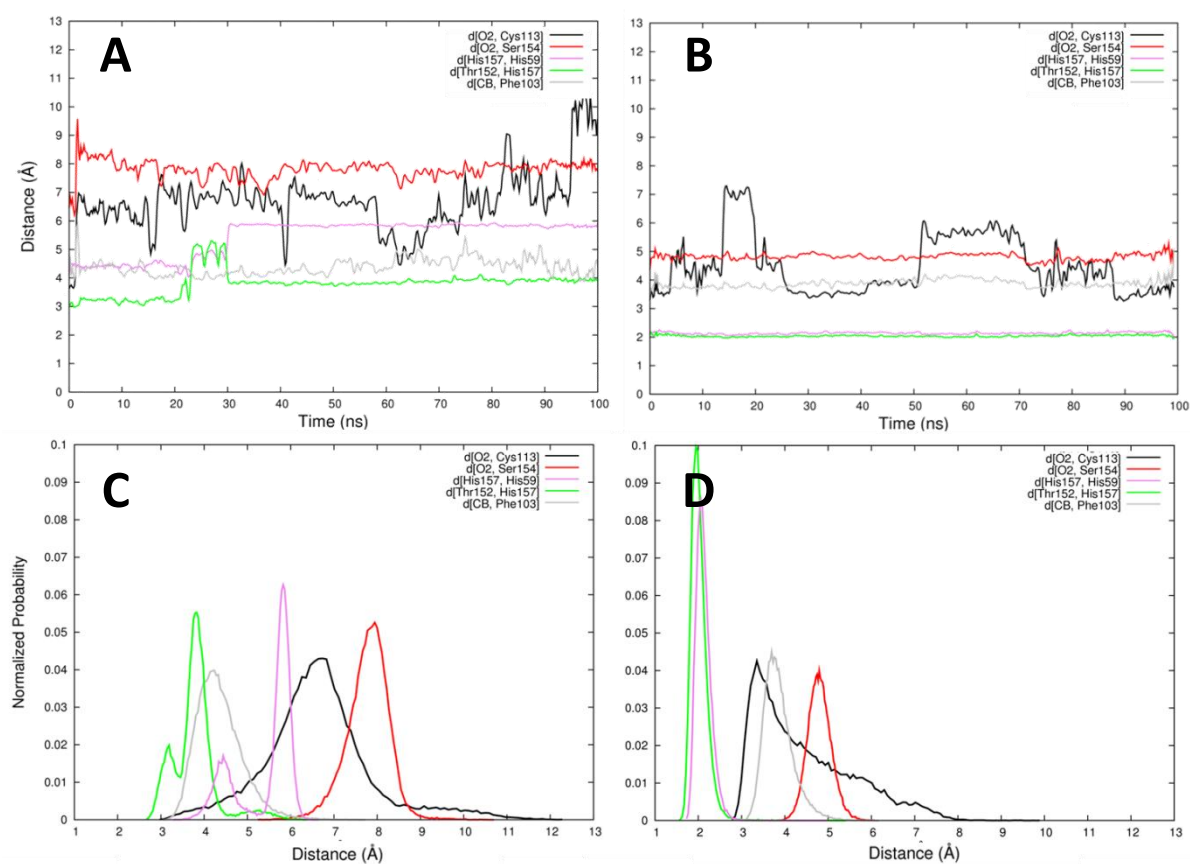
To understand Pin1 catalysis, an investigation of the two hypothesized (C and NC) reaction mechanisms is required. Once a preferential reaction mechanism is identified, it is possible to proceed in comparing the *cis-trans* isomerization results obtained in bulk water and in Pin1 active site.

### 4.3.1 Testing the C and NC Reaction Mechanisms

To test the proposed C and NC mechanisms of *cis-trans* isomerization [97,104], unbiased MD simulations were performed for both model A, and B, complexed with the peptide substrate (see Scheme 4.1 and Table 4.1). Each system was simulated for 100 ns in the canonical ensemble at the temperature of 300 K. The starting configuration of the substrate was *cis*, with the prolyl amide bond  $\omega = 0^\circ$ , and the tautomeric states of the aminoacids of the binding sites were modelled according to the evidences outlined in section 4.2.1 of Computational Details. The intermolecular distances between substrate and key catalytic residues, as well as the distances crucial for the active site folding, are reported in Figure 4.10 A and B, while their distributions are shown in Figure 4.10 C and D. Regarding the NC model B complex (Figure 4.10 B and D), the hydrogen bond network between Pin1 residues, as well as the position of the substrate within the binding site, were satisfactorily preserved during the simulation. In Figure 4.10 B the distances between the carbonyl oxygen O<sub>2</sub> of the substrate and the sulphur atom of Cys113 (black line), the hydroxylic oxygen of Ser154 (red line), as well as the distance between the prolyl ring C<sub>β</sub> and the center of mass of Phe103 (gray line) as function of the simulation time, highlighted that the peptide maintained its relative position within the active site

---

along the simulation. In particular, the distance O<sub>2</sub> - Cys113 (black line in Figure 4.10 B) underwent to reversible fluctuations between 3.5 – 6 Å, revealing that Cys113 could populate two interchangeable conformational states. However this behaviour didn't affect the substrate binding, as it could be observed from the analysis of the other distances in Figure 4.10 B, and from the tall and narrow distributions in Figure 4.10 D. On the other hand, a different scenario was observed for the covalent model A complex (Figure 4.10 A and C). Switching from model B to model A, only the distances between C<sub>β</sub> and Phe103 (gray line) centered at 4 Å, was retained. Despite this evidence proved that no unbinding occurred during the simulation, the wider distributions, and the increasing of the distances between active site residues (Figure 4.10 A and C), revealed an alteration of the active site shape, and therefore of the suitable folding for catalysis. In particular, the mean distance between the carbonyl group of the substrate and Cys113 was found to be centered at 6.8 Å, resulting incompatible with the suggested nucleophilic attack (step 2 in Scheme 4.1). These results strongly ruled out the possibility of a C reaction mechanism for peptidyl-prolyl *cis-trans* isomerization.



**Figure 4.10** Intermolecular distances between the peptide, in *cis* configuration, and Cys113 (black line), Ser154 (red line), Phe103 (gray line), for the C (A) and NC (B) model after 100 ns of unbiased MD (NVT, 300K). The distances between the aminoacids of the catalytic site, critical for binding site folding (Thr152 – His157, His157 – His59), are also shown. At the bottom, the probability distributions of such distances are reported for the C (C) and NC (D) mechanism.



### 4.3.2 Bulk *Cis-Trans* Isomerization

For the uncatalyzed (bulk) reaction mechanism, the PMF as function of the dihedral angles  $\Psi_{\text{PRO}}$  and  $\zeta$ , is reported in Figure 4.11. The free energy profile showed four minima: CIS1 ( $\zeta \cong 0^\circ / \Psi_{\text{PRO}} \cong -30^\circ$ ), CIS2 ( $\zeta \cong 0^\circ / \Psi_{\text{PRO}} \cong 150^\circ$ ), TRANS1 ( $\zeta \cong 180^\circ / \Psi_{\text{PRO}} \cong -30^\circ$ ), TRANS2 ( $\zeta \cong 180^\circ / \Psi_{\text{PRO}} \cong 150^\circ$ ). TRANS1 was found to be the global minimum, and approximately 3 kcal/mol more stable than TRANS2. The estimated free energies of the two *cis* configurations, CIS1 and CIS2, are about 5 and 3 kcal/mol higher than TRANS1, respectively. These results showed that CIS2 and TRANS2 are energetically equivalent. Similar results were also obtained by previous theoretical studies. In this regard, Velazquez and Hamelberg [74,106] obtained similar free energy plots by means of accelerated molecular dynamics. In particular, they showed that the phosphorylation of a Ser-Pro containing peptide affected the population of CIS1 [74]. A comparable free energy map was also shown by Melis et al. [221], in the context of a *cis-trans* investigation for a proline dipeptide. From the PMF, four saddle points could be identified: TS1 ( $\zeta \cong 90^\circ / \Psi_{\text{PRO}} \cong -15^\circ$ ), TS2 ( $\zeta \cong -90^\circ / \Psi_{\text{PRO}} \cong 0^\circ$ ), TS3 ( $\zeta \cong 90^\circ / \Psi_{\text{PRO}} \cong 150^\circ$ ), and TS4 ( $\zeta \cong -90^\circ / \Psi_{\text{PRO}} \cong 150^\circ$ ). Similarly to Ace-Pro-NMe (Figure 4.5), using the notation proposed by Fischer et al. [205], all the transition states have shown an *exo* configuration. In particular, TS1 and TS3 could be classified as *syn/exo* transition states, while TS2 and TS4, as *anti/exo*. These transition states were almost isoenergetic, and the computed free energy barriers were approximately 22 kcal/mol, for the isomerization in bulk water. Therefore, the reaction mechanism could proceed following the counter-clockwise direction, as well as through the clockwise pathway, in a symmetric-like isomerization model.

The conformations adopted by the peptide in the four minima, and transition states, are also reported in Figure 4.11. All the configurations belonging to the  $\Psi_{\text{PRO}}$  space ranging from  $-60^\circ$  to  $60^\circ$  (CIS1, TRANS1, TS1, and TS2, see Figure 4.11) were found to be compatible with the autocatalytic process [205,215,220], showing the amino group hydrogen of the residue following the proline  $H_1$  in proximity of the

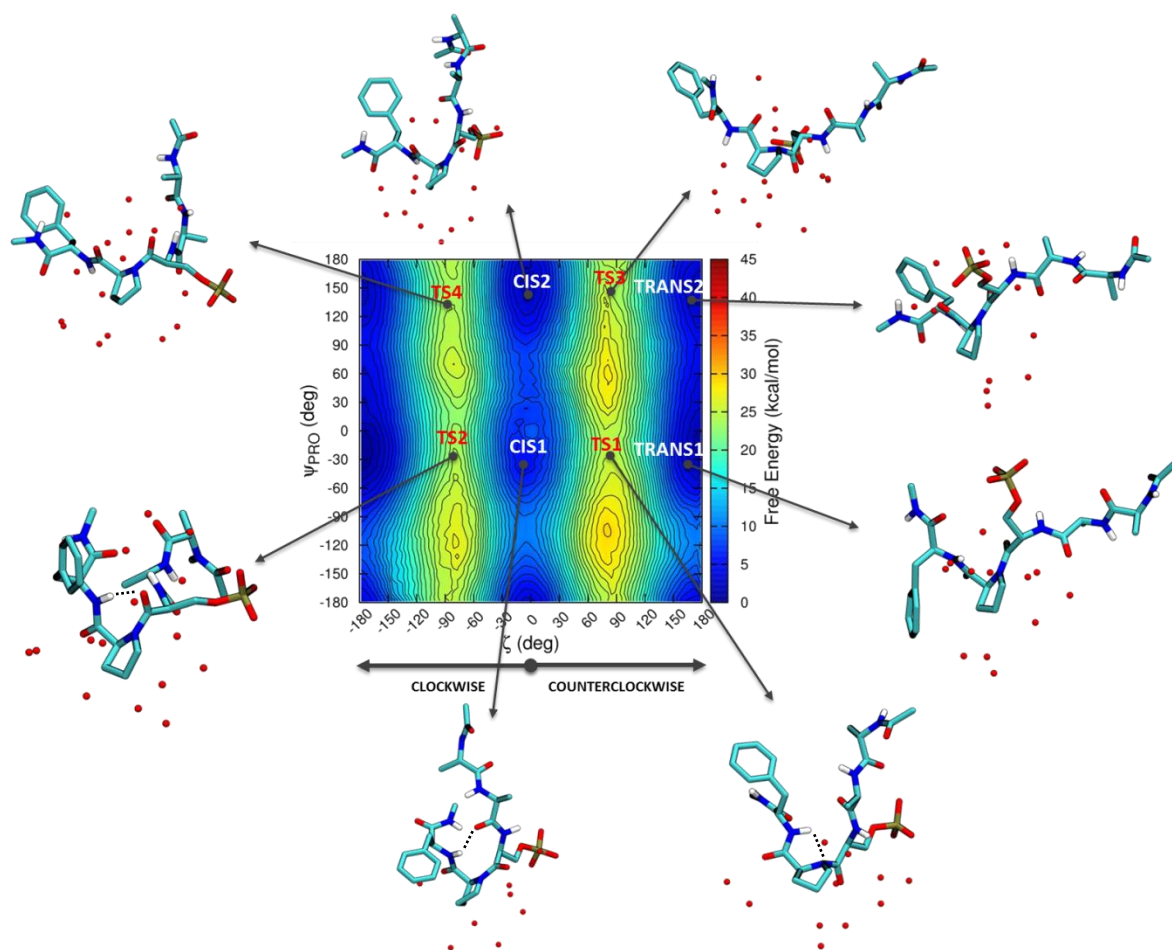
prolyl nitrogen  $N_2$ . However, this interaction, which is responsible to promote the nitrogen pyramidalization, could be weakened by the presence of an intramolecular hydrogen bond between the carbonyl oxygen of pSer,  $O_2$ , and the hydrogen  $H_1$ . In this regard, in Figure 4.12 A, the distance  $O_2 - H_1$  along the space of the collective variables, is reported. This plot highlighted that  $O_2 - H_1$  bond occurred with an high frequency along the path connecting the TRANS1 configuration with TS2, in clockwise direction (see also TS2 conformation in Figure 4.11). Hence, during the isomerization  $TRANS1 \rightarrow TS2 \rightarrow CIS1$ , this intramolecular bond had to be disrupted. Moreover along the path  $CIS1 \rightarrow TS1$ , another hydrogen bond interaction between the amino group hydrogen  $H_1$  and the carbonyl oxygen  $O_3$  of the residue following pSer, was found to be involved in the isomerization. As shown in Figure 4.12 B, starting from CIS1, the intramolecular bond  $O_3 - H_1$  was preserved during the counter-clockwise rotation of the proline amide bond, until the TS1 was reached. Therefore, to complete the isomerization toward the TRANS1 basin, this interaction had to be broken.

In *trans* states, the peptide adopted more extended conformations than in *cis*, with an increasing of the shell of solvation (see Figure 4.11 for a qualitatively evaluation). A similar increasing in hydration was also observed passing toward conformations belonging to the space characterized by higher  $\Psi_{PRO}$  values ( $60^\circ < \Psi_{PRO} < 180^\circ$ ). This evidence suggested that the water molecules could play an important roles in the isomerization in this space of the CVs, where no autocatalytic bond formation was observed.

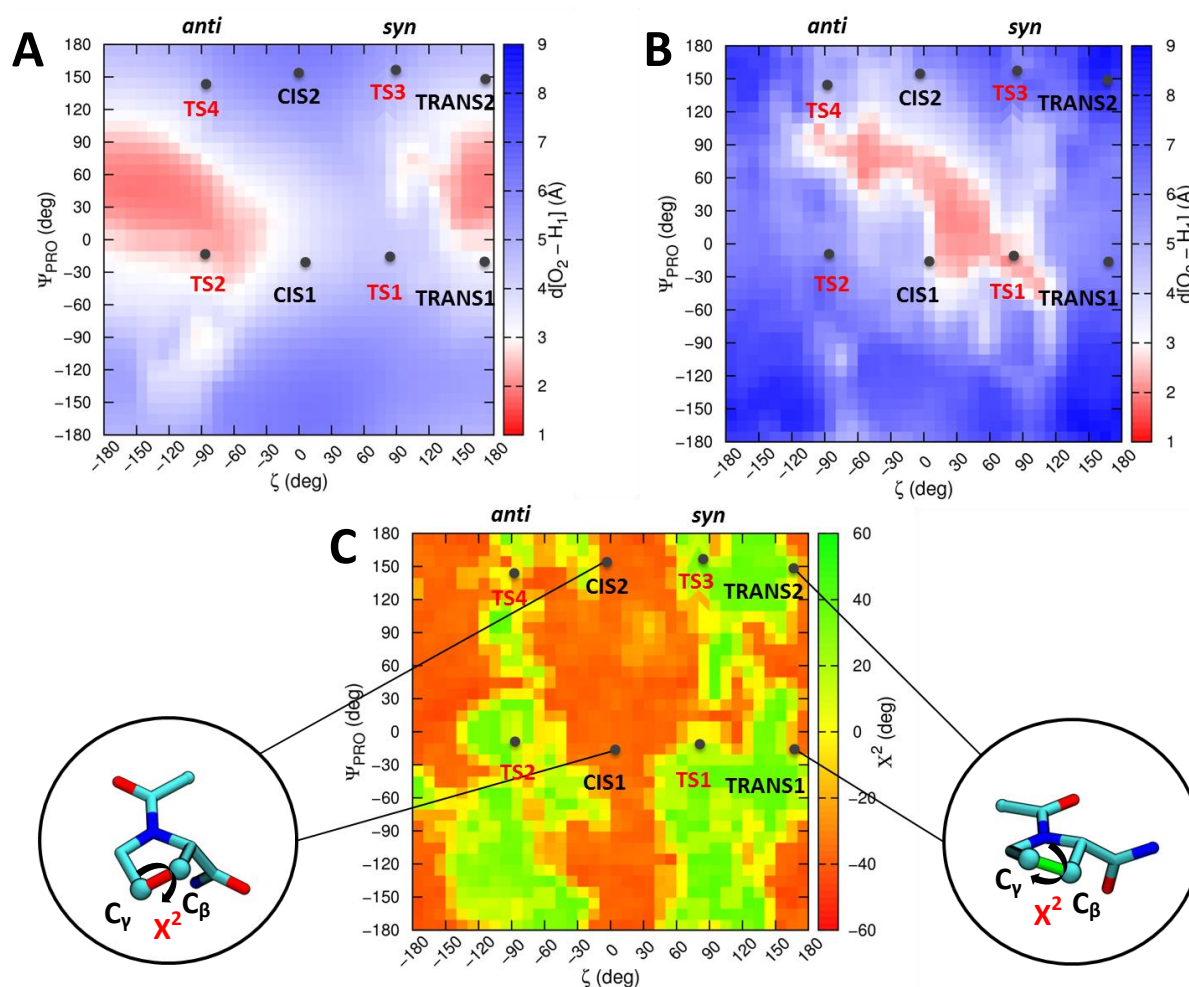
The conformations adopted by the proline residue during the isomerization were also investigated (Figure 4.12 C). It has been established that the proline ring could adopt two distinct up- and down-puckered conformations, based on the position of  $C_\gamma$  above or below the plane of the ring defined by  $C_\alpha$ ,  $C_\beta$ ,  $C_\delta$ , and  $N_2$  [266-268]. As proposed by Ho and co-workers [268], the dihedral  $X^2$  (Figure 4.12 C) was used as figure of merit in order to determine the puckers:  $X^2 > 10^\circ$  for up-puckers,  $X^2 < -10^\circ$  for down-puckers, while  $-10^\circ < X^2 < 10^\circ$  for the definition of planar conformations. In Figure 4.10 C,  $X^2$  was mapped in the space of the reaction coordinates  $\Psi_{PRO} / \zeta$ , showing that *cis* conformations mainly populated the down-pucker state, while *trans* ones were found to be compatible with the up- and down-

pucker geometries without preferences. This observed trend was found in line with previous investigations reported in literature [268,269].

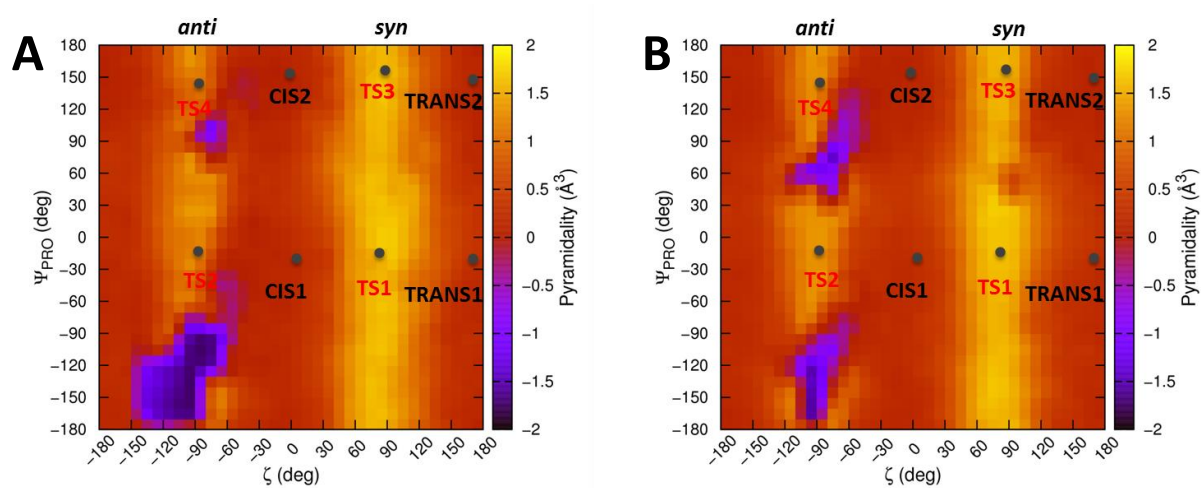
In Figure 4.13 A, the degree of the prolyl nitrogen pyramidalization, for the peptide substrate, was mapped on the space of the reaction coordinates (see section 4.2.4 for computational details). The absence of nitrogen pyramidalization in the space areas centered at  $\zeta$  values of  $0^\circ$  and  $180^\circ$ , is in agreement with the planarity of the amide bond in *cis* and *trans* ground states. On the other side, the highest degree of pyramidalization was evaluated on the *syn* and *anti* ridges (at  $\zeta = 90^\circ$ , and  $\zeta = -90^\circ$ , respectively), corresponding to the twisted amide bond conformations. A positive pyramidalization was evaluated for all the four transition states. Interestingly, the *syn/exo* transition state TS1 have shown a higher pyramidalization than the *anti/exo* TS2. This finding was in line with the stabilization effect of the pyramidal conformation due to the autocatalytic hydrogen bond  $H_1 - N_2$ , in the case of TS1. On the contrary, in TS2, this stabilization effect was altered by the hydrogen bond  $O_2 - H_1$  (Figure 4.12 A), leading to a lower degree of the nitrogen pyramidalization (Figure 4.12 A). The negative pyramidalization resulted to be not directly involved in the isomerization process, as it was confined on the *anti* ridge corresponding to the free energy maxima. In this regard, a slightly decrease of the negative area was observed at  $\zeta \cong -90^\circ / \Psi_{\text{PRO}} \cong 90^\circ$ , whereas a sharp increase was evaluated at  $\zeta \cong -90^\circ / \Psi_{\text{PRO}} \cong -150^\circ$ , compared with the pyramidalization map of Ace-Pro-NMe (Figure 4.12 B). This phenomena was strictly connected to the presence of residues preceding and following the proline. In particular, the intramolecular hydrogen bond  $H_1 - O_3$ , was found to stabilize a positive pyramidal conformation for the prolyl nitrogen at  $\Psi_{\text{PRO}} \cong 90^\circ$ , and the inverse pyramidalization at  $\Psi_{\text{PRO}} \cong -150^\circ$ .



**Figure 4.11** PMF for the uncatalyzed *cis-trans* isomerization of the prolyl amide bond using  $\Psi_{\text{PRO}}$  and  $\zeta$  as reaction coordinates. The conformations assumed by the peptide in the different states of the dihedrals space, are also reported.



**Figure 4.12** (A) The distance  $d[\text{O}_2 - \text{H}_1]$  (Å) as a function of the reaction coordinates  $\Psi_{\text{PRO}} / \zeta$ . This interaction has to be broken during the isomerization  $\text{TRANS1} \rightarrow \text{TS2} \rightarrow \text{CIS1}$ . (B) The distance  $d[\text{O}_3 - \text{H}_1]$  (Å) as a function of  $\Psi_{\text{PRO}} / \zeta$ . This interaction has to be disrupted during the isomerization  $\text{CIS1} \rightarrow \text{TS1} \rightarrow \text{TRANS1}$ . (C) The dihedral  $X^2$  as a function of  $\Psi_{\text{PRO}} / \zeta$ . The typical down (red), and up (green), conformations for proline ring are also shown.



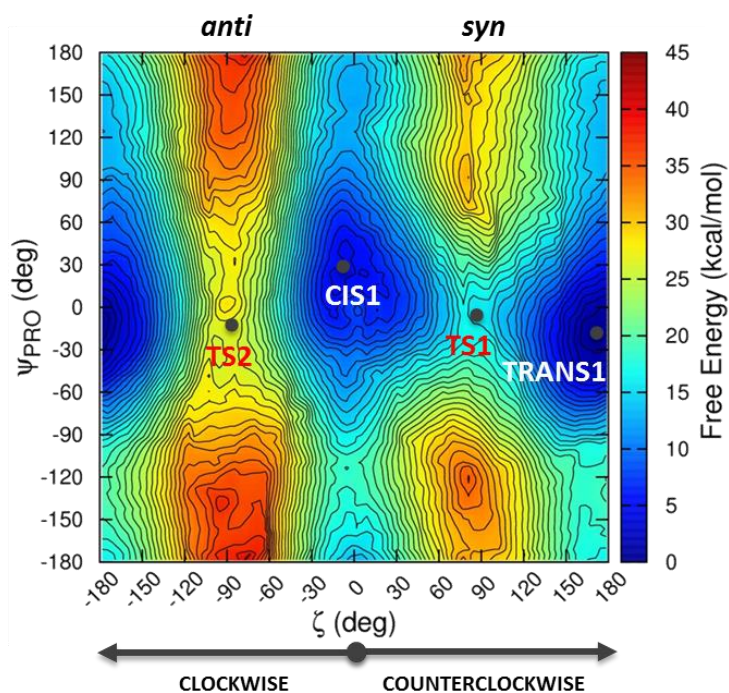
**Figure 4.13** Map of the prolyl nitrogen pyramidalization as a function of  $\Psi_{\text{PRO}} / \zeta$ , for the peptide substrate (A), and for Ace-Pro-NMe model (B).

### 4.3.3 NC *Cis-Trans* Isomerization

The derived free energy profile for the enzymatic isomerization of the prolyl amide bond within the NC reaction mechanism, is reported in Figure 4.14. From a comparison with the PMF of the reaction in bulk water (Figure 4.11), it was possible to observe a dramatic alteration of the ground state basins. In fact, Figure 4.14 showed just two minima: the CIS1 configuration, placed at value  $\zeta \cong -15^\circ / \Psi_{\text{PRO}} \cong 30^\circ$ , and, as for the reaction in water, TRANS1, was located in the space of the collective variables centered at  $\zeta \cong 180^\circ / \Psi_{\text{PRO}} \cong -30^\circ$ . Moreover, similarly to the non-enzymatic isomerization, TRANS1 represented the global minimum, and hence the predominant isomer. No ground states basins were found at higher values of  $\Psi_{\text{PRO}}$  dihedral space. Moreover, only two transition states were observed, the TS1 at  $\zeta \cong 90^\circ / \Psi_{\text{PRO}} \cong 0^\circ$  (*syn* ridge), and the TS2, at  $\zeta \cong -90^\circ / \Psi_{\text{PRO}} \cong -15^\circ$  (*anti* ridge). Contrary to the reaction mechanism in bulk, the results highlighted a *syn/exo* configuration for TS1, whereas an *anti/endo* was found for TS2. In this regard, the degree of pyramidalization was mapped on the CVs space in Figure 4.15. As shown, an enhancement of positive pyramidalization was reported on the *syn* ridge, compared to the non-enzymatic isomerization (4.13 A), resulting from the stabilization effect of the autocatalytic hydrogen bond between  $H_1 - N_2$ . Moreover, contrary to Figure 4.3 A, a negative pyramidalization was found to occur on TS2 (*anti* ridge). This is due to the stabilization effect of a different hydrogen bond, between the amino group  $H_1$  and the prolyl bond carbonyl oxygen  $O_2$ , favoured by the conformation assumed by the peptide within the restricted Pin1 active site space.

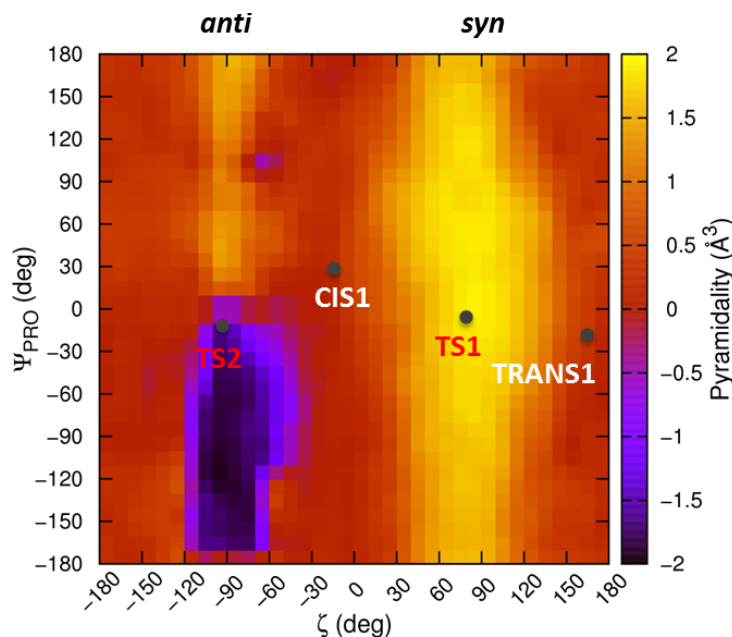
The estimated energy barrier from CIS1 to TRANS1 was found to be approximately 15 kcal/mol, highlighting a reduction of the activation free energy to rotation of about 7 kcal/mol, compared to the non-enzymatic isomerization. This finding was in good agreement with previous experimental [99] and theoretical studies [106]. In opposition to the scenario depicted in Figure 4.11, the free energy profile for Pin1-catalyzed reaction suggested that the isomerization had to proceed from CIS1 (reference state) to TRANS1 following the counter-clockwise direction,

through the more stabilized saddle point TS1. On the other side, the CIS1  $\rightarrow$  TS2  $\rightarrow$  TRANS1 path (in clockwise direction) was strongly prohibited by the high energy associated to the transition state TS2 (see Figure 4.14). In particular, this latter state, was found to be more than 10 kcal/mol higher in energy than TS1, and approximately 5 kcal/mol higher than the corresponding TS2 in bulk solution.



**Figure 4.14** PMF for Pin1-catalyzed *cis-trans* isomerization of the prolyl amide bond using  $\Psi_{\text{PRO}}$  and  $\zeta$  as reaction coordinates.



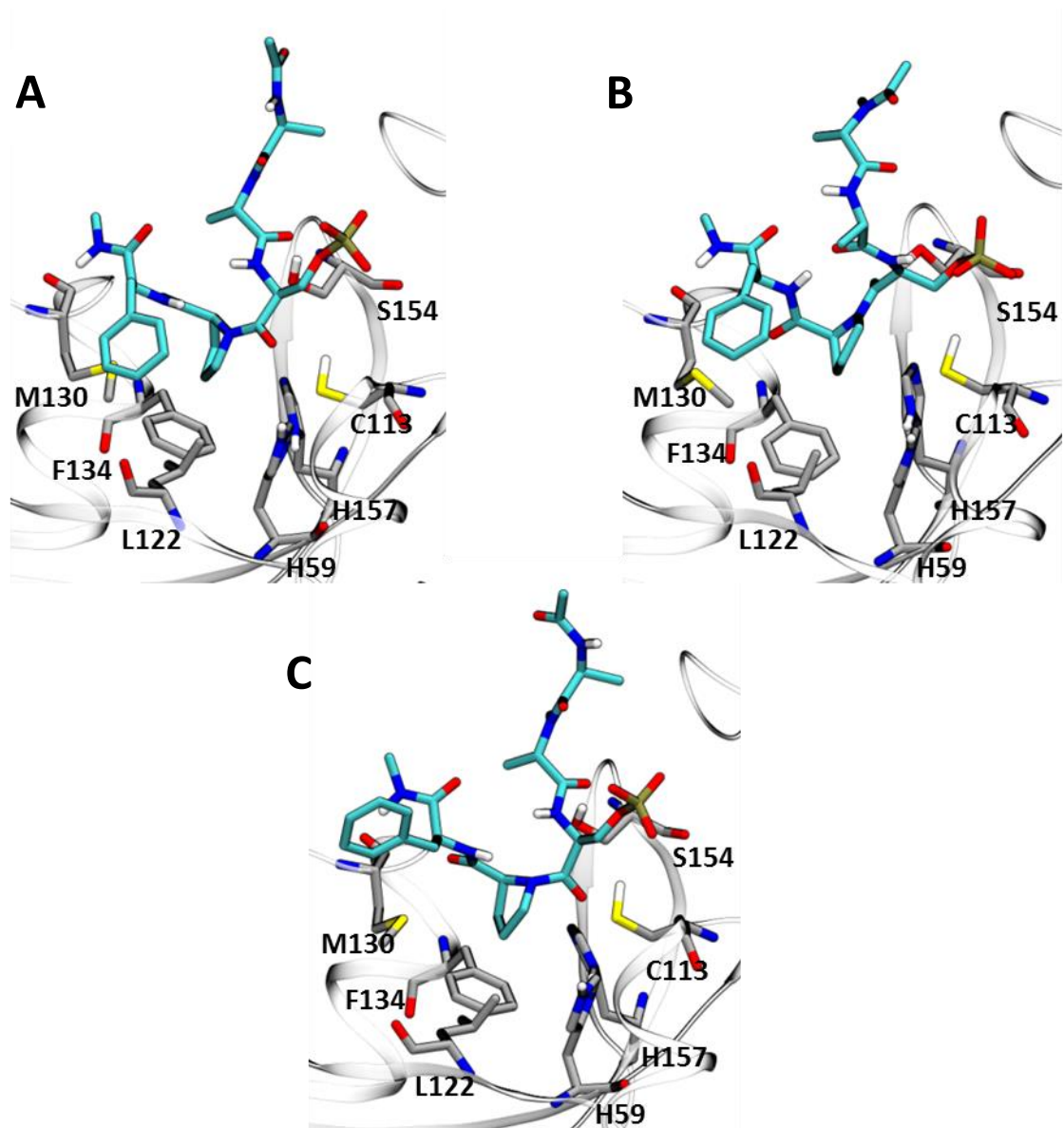


**Figure 4.15** Map of the prolyl nitrogen pyramidalization as a function of  $\Psi_{\text{PRO}} / \zeta$ , for the peptide substrate within Pin1 active site.

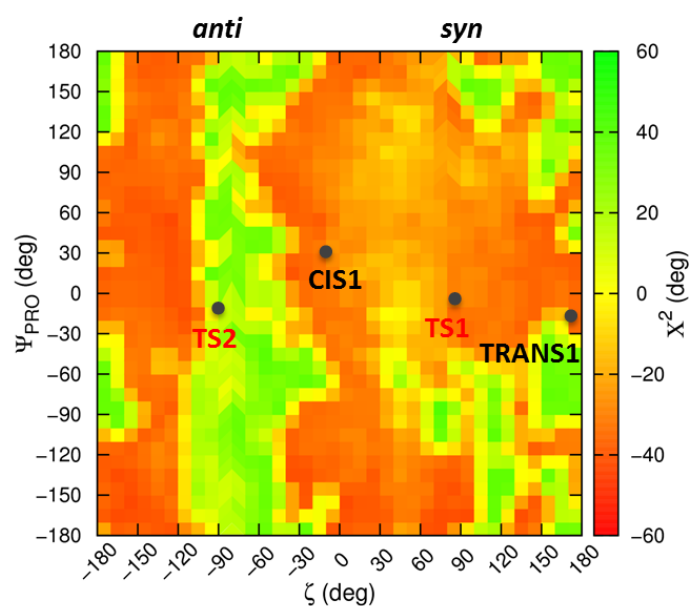
### 4.3.3.1 Entropic Effect

Upon binding, the protein environment significantly reduced the torsional degrees of freedom of the peptide substrate. As shown in Figure 4.14, Pin1 active site allowed the substrate to sample a restricted area of the configurational space, corresponding to the  $\Psi_{\text{PRO}}$  region ranging from  $-60^\circ$  to  $60^\circ$ . This evidence was found to dramatically influence the prolyl nitrogen configuration on the *syn/anti* ridges, compared to the non-enzymatic reaction (Figure 4.15). On the other side, the sampling of the remaining space was strongly avoided by steric clashes occurring between the C-terminal portion of the substrate and Pin1 active site residue side chains. From the trajectories analysis, it was determined that these bad contacts directly involved the substrate phenylalanine side chain, whose orientation was found to be critically influenced by the values assumed by  $\Psi_{\text{PRO}}$  dihedral. In Figure 4.16, the conformations adopted by the peptide in the configurational states CIS1, TS1, and TRANS1 are reported. Because of the high energy associated to TS2, the conformations relative to this state were not considered, and the attention was

focused on the more populated states during Pin1 isomerization. Interestingly, in contrast with the non-enzymatic reaction, no extended conformations could be observed for the *trans* state (Figure 4.16 B), as a consequence of the restricted space within the binding site. In a certain way, Pin1 environment seemed to generate a sort of “entropic trap” by capturing the substrate, reducing its degrees of freedom, and driving the isomerization along the path connecting CIS1 to TRANS1 passing through the TS1, in both the counter-clockwise or clockwise directions. In Figure 4.17, the dihedral  $X^2$  was mapped on the space of the reaction coordinates  $\zeta / \Psi_{\text{PRO}}$ . By making a comparison with the same plot derived from the bulk isomerization (Figure 4.12 C), it was possible to note an impressive alteration of the proline ring states, which were found to be no longer spread on the map in a symmetric-like way. Interestingly, Figure 4.17 showed a substantial preference for the down-pucker proline ring conformation along all the path connecting CIS1 to TRANS1, contrary to the non-enzymatic reaction where an higher frequency of occurrence for the up-pucker was observed (in particular for the TRANS1 state). This finding had be considered as a direct consequence of the entropy loss within PIN1 binding site, leading *cis* and *trans* prolyl isomers to preferentially populate the down-pucker state.



**Figure 4.16** Conformations assumed by the substrate in (A) CIS1, (B) TRANS1, and (C) in the saddle point TS1.

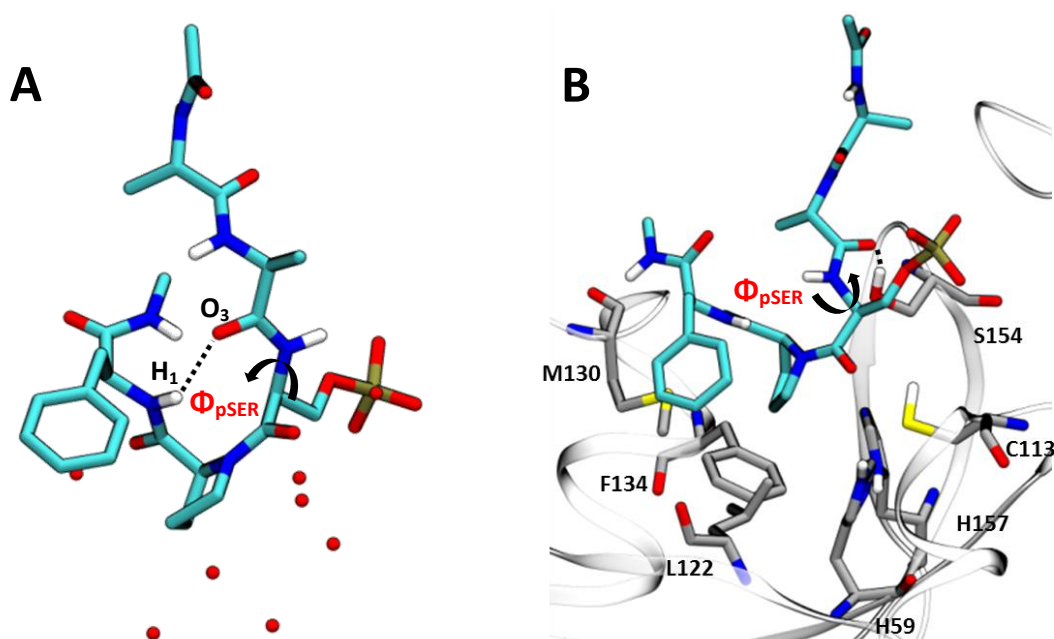


**Figure 4.17** Dihedral  $X^2$  mapped on the  $\Psi_{\text{PRO}} / \zeta$  configurational space. The typical down-, and up-pucker conformations for proline ring are labelled in red and green, respectively.

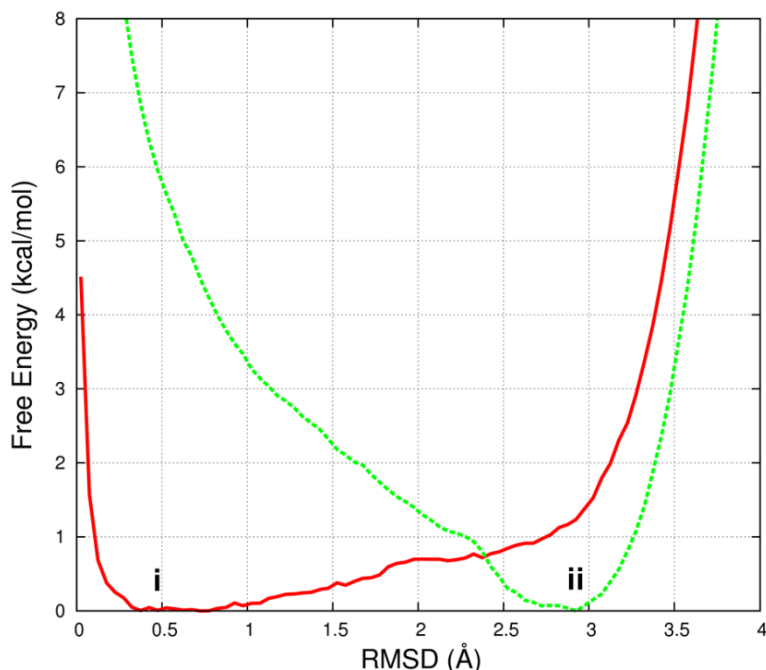
### 4.3.3.2 *Cis* Ground State Destabilization

Another important aspect that emerged from the PMF analysis (Figure 4.14), was the shifting of the CIS1 configuration in the collective variables space, compared to the corresponding *cis* state for the isomerization in bulk water (Figure 4.11). In fact, regarding the NC isomerization, CIS1 was centered at  $\Psi_{\text{PRO}} \cong 30^\circ$ , with a displacement of about  $60^\circ$ , relative to the position assumed in water. The corresponding conformations adopted by the peptide in the two *cis* configurational states, are reported in Figure 4.18. In Pin1 active site, the peptide assumed a slightly more extended conformation (Figure 4.18 B) than in water (Figure 4.18 A), due to the absence of the intramolecular hydrogen bond between the amino group hydrogen  $H_1$  of the residue preceding the proline, and the carbonyl oxygen  $O_3$ , of the residue following pSer. Therefore, Pin1 stabilized a *cis* conformation which has shown a low frequency of occurrence in water, where the intramolecular interaction  $H_1 - O_3$  was disrupted, and replaced by an hydrogen bond with Ser154 (Figure 4.18). In other words, the *cis* ground state destabilization has to be considered as

the result of the active site environmental effect. Following the approach described in section 4.2.5 of Computational Details, the difference in free energy between the conformations assumed by the peptide in water and within the Pin1 binding pocket, was quantified. The PMFs calculated in the space of the RMSD have been reported in Figure 4.19. The red curve represented the PMF for the peptide substrate in the Pin1 catalytic site. The global minimum *i* was found at RMSD value of 0.5 Å. The dotted green curve, was the free energy derived for the peptide in bulk solvent. In this case, the global minimum, labelled as *ii*, was located at higher value of the reaction coordinate, approximately at 3.0 Å. In particular, at RMSD = 0.5 Å, corresponding to the global minimum *i* in the enzyme environment, the peptide in solution resulted to be placed at higher energy value, around 6 kcal/mol. This quantity allowed to estimate the free energy required to convert the conformation assumed by the substrate in CIS1 state in bulk solution (Figure 4.18 A), to the one in the binding site (4.18 B). In other words, this quantity represented the strain energy, and therefore the energy cost paid by Pin1 for stabilizing a less probable conformation in water for the CIS1 state. This conformation, as reported in the next section, has been shown to be catalytically competent, favouring the barrier crossing.



**Figure 4.18** Conformations adopted by the substrate in CIS1 state for the isomerization in bulk water (A), and for the NC Pin1-catalyzed reaction (B).

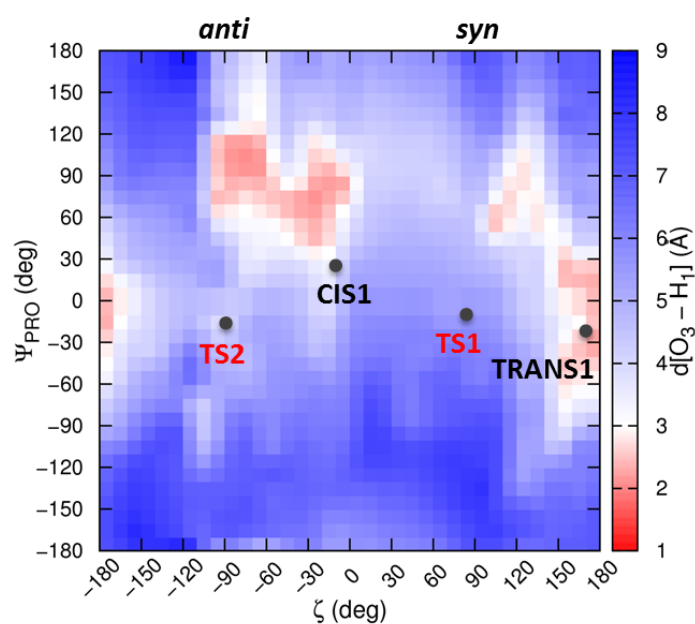


**Figure 4.19** PMF calculated in the space of the RMSD. The red line represents the PMF for the peptide in Pin1 binding site, while the green dotted line, the PMF for the peptide in bulk solution. The respectively minima have been labelled as *i* and *ii*.

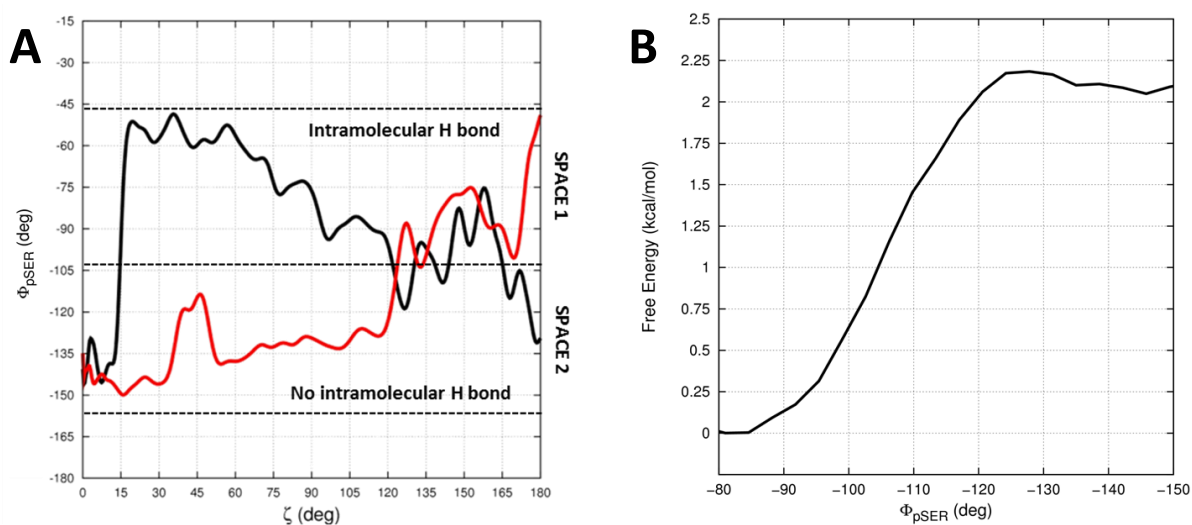
### 4.3.3.3 Intramolecular Hydrogen Bond Investigation and Effects on the Barrier

The frequency of occurrence of the intramolecular hydrogen bond, which has been shown to characterize the CIS1 state in bulk solvent, was evaluated mapping the distance between  $H_1 - O_3$  in the space of the collective variable  $\zeta / \Psi_{\text{PRO}}$  (Figure 4.20). The same plot for the uncatalyzed reaction mechanism (Figure 4.12 B), revealed that this hydrogen bond was preserved along the path connecting CIS1 to TS1, on the *syn* ridge. Therefore, this interaction had to be broken to reach the TRANS1 basin, completing in this way the isomerization process. A different scenario emerged for Pin1-catalyzed reaction mechanism (Figure 4.20). As a matter

of fact, no intramolecular hydrogen bond was found during the twisting of the proline amide bond. The dihedral  $\Phi_{\text{pSER}}$ , which allowed the rotation along the  $\text{C}_3 - \text{N}_3$  bond (the alpha carbon and amide nitrogen of pSer), was found to be responsible for the formation of such interaction. In particular, the value assumed by  $\Phi_{\text{pSER}}$  in the *cis* configuration in water was about  $-80^\circ$ , against the  $-150^\circ$  measured for the same configuration state in enzyme. In Figure 4.21 A, the dihedral  $\Phi_{\text{pSER}}$  has been plotted as function of  $\zeta$  for both the reactions, in bulk water (black line) and Pin1-catalyzed (red line). This analysis allowed to identify two areas in the  $\Phi_{\text{pSER}}$  space to discriminate between formation or absence of the intramolecular bond  $\text{H}_1 - \text{O}_3$ . In particular, such interaction occurred at  $-105^\circ < \Phi_{\text{pSER}} < -45^\circ$  (dihedral space 1, in Figure 4.21 A), gradually disappearing for  $\Phi_{\text{pSER}} < -105^\circ$  (dihedral space 2). Interestingly, for the peptide in Pin1 active site, such dihedral started at value  $-150^\circ$  in CIS1 state, gradually reaching the value  $-130^\circ$  in TS1 ( $\zeta \cong 90^\circ$ ). After the transition state, for  $\zeta > 120^\circ$ ,  $\Phi_{\text{pSER}}$  sharply increased at suitable values for the hydrogen bond formation. These results were in line with the evidences coming from Figure 4.20. On the other side, an opposite behaviour was found for the peptide in water (black line in Figure 4.21 A). In order to estimate the energy necessary to disrupt this hydrogen bond, an umbrella sampling on the space of the dihedral  $\Phi_{\text{pSER}}$ , was performed (see section 4.2.6 of Computational Details). The derived mono-dimensional PMF has been shown in Figure 4.21 B, revealing an energy barrier of approximately 2.2 kcal/mol. Therefore, due to the absence of the intramolecular interaction  $\text{H}_1 - \text{O}_3$  during the Pin1-catalyzed *cis-trans* isomerization, this quantity contributed to reduce the activation barrier to rotation.



**Figure 4.20** The distance  $d[\text{O}_3 - \text{H}_1]$  (Å) as a function of the reaction coordinates  $\Psi_{\text{PRO}} / \zeta$  during the NC *cis-trans* isomerization.



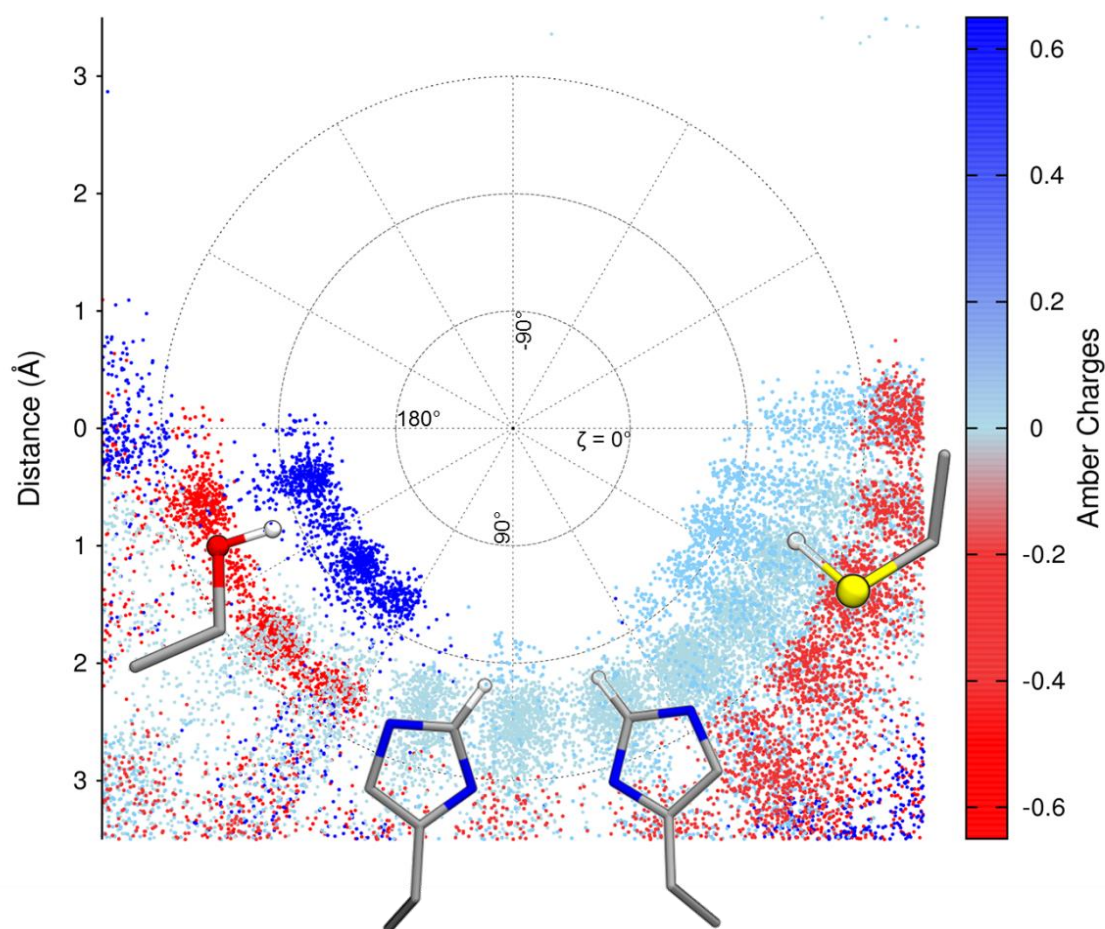
**Figure 4.21** (A) The dihedral  $\Phi_{\text{pSER}}$  as a function of  $\zeta$ , for the isomerization in bulk solvent (black line), and in Pin1 active site (red line). Space 1 represents the  $\Phi_{\text{pSER}}$  space which allows the formation of the intramolecular hydrogen bond  $\text{H}_1 - \text{O}_3$  in the peptide. On the contrary, space 2, indicates no intramolecular bond formation. (B) Free energy profile along the dihedral  $\Phi_{\text{pSER}}$ , describing the formation ( $\Phi_{\text{pSER}} = -80^\circ$ ), and breaking ( $\Phi_{\text{pSER}} = -150^\circ$ ) of  $\text{H}_1 - \text{O}_3$  interaction, evaluated during the isomerization in bulk solvent.



#### 4.3.3.4 Intermolecular Interactions during the Isomerization

The contacts between the peptidyl-prolyl amide bond oxygen  $O_2$  and the active site side chains, are reported as a function of the dihedral  $\zeta$  in the polar plot in Figure 4.22. In the polar grid, the dihedral  $\zeta$  was marked in  $30^\circ$  increments starting from the CIS1 configuration at  $\zeta = 0^\circ$ , and returning to the same state after a full rotation in the counter-clockwise direction, passing sequentially through the TS1 ( $\zeta = 90^\circ$ )  $\rightarrow$  TRANS1 ( $\zeta = 180^\circ$ )  $\rightarrow$  TS2 ( $\zeta = -90^\circ$ ) states. The distances of the amide bond carbonyl with the four residues, Cys113, His59, His157, and Ser154, have been considered in this analysis. In particular, as shown in Table 4.2, the contacts between  $O_2$  and the hydrogen and sulfur atoms of Cys113 sulfhydryl group, the epsilon carbon-linked hydrogen of His59 and His157 imidazole groups, and both the atoms of Ser154 hydroxyl side chain, have been monitored, and labelled on the basis of their fixed parm94 Amber partial charges [114]. Figure 4.22 highlighted a competition, at  $\zeta = 0^\circ$ , of attractive and repulsive interactions between the carbonyl oxygen and Cys113 sulfhydryl side chain, with contact distances ranging from 2 to 3 Å. Interestingly, at  $\zeta \cong 15^\circ$ , corresponding to the initial alteration of the amide bond planarity, an increasing of more favourable electrostatic interactions with Cys113, have been shown. As a matter of fact, the intermolecular hydrogen bond  $O_2 - HS$  occurred with a distance  $< 2$  Å. Moreover, these contacts have been retained until the value  $\zeta \cong 60^\circ$  was reached. At  $60^\circ < \zeta < 120^\circ$ , corresponding to the TS1 configuration, the carbonyl oxygen  $O_2$  has shown to interact with the less positive charged epsilon C-H (H5, see Table 4.2), with distances ranging from 2 to 3 Å. However, a slightly reduction of the these interaction distances were reported at  $90^\circ \leq \zeta < 120^\circ$ , (distances  $< 2$  Å). In particular, in this space of the reaction coordinate, it was observed an initial formation of the hydrogen bond with the hydroxyl group of Ser154 ( $O_2 - HO$ ). These interactions were found to be predominant at  $\zeta > 120^\circ$ , compared with the repulsive  $O_2 - OH$  contacts, revealing distances spread on a narrow 1.5 – 2 Å range. These results suggested a non-negligible role of Pin1 active site residue on *cis-trans* isomerization of prolyl amide bond. On the other hand, in

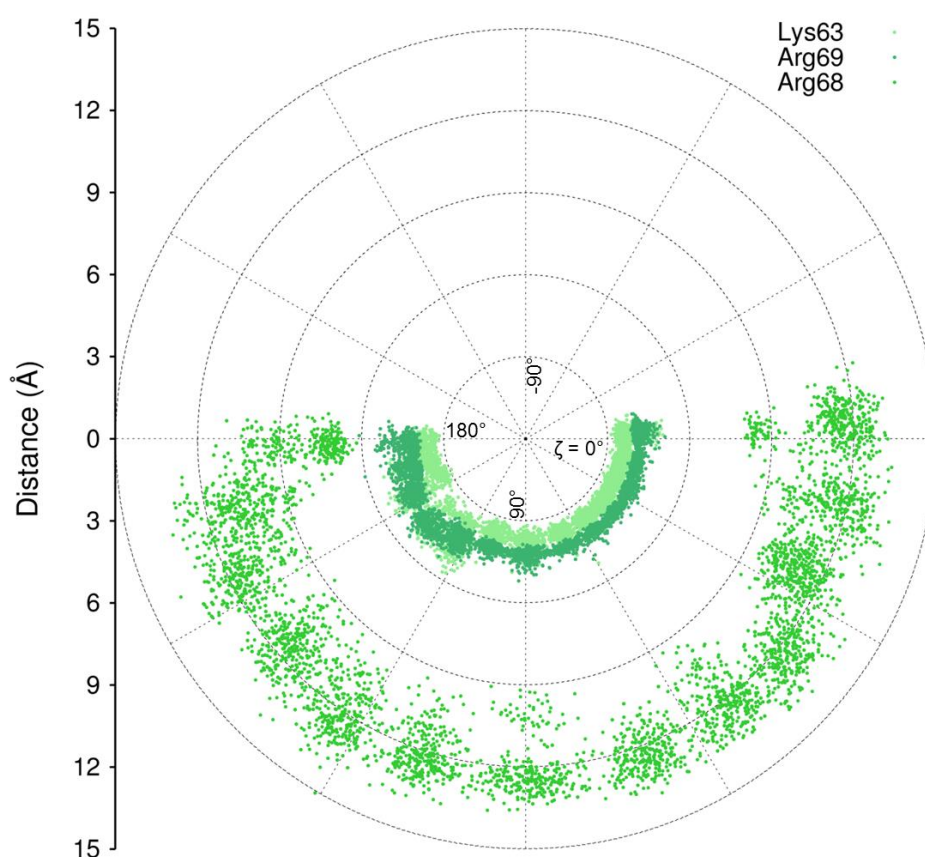
Figure 4.23, a similar polar plot for the interactions between the phosphorus atom P of the substrate phosphate moiety and the basic pocket aminoacids (Lys63, Arg68, and Arg69) was reported. In this case, because of the absence of a competition between attractive/repulsive interactions, the fixed point charges were not used as labels. Unlike Arg68, which has shown to preferentially bind the substrate in the TRANS1 state (distance ranging approximately from 6 to 13 Å), Arg69 and Lys63 were found to form tighter interactions with the phosphate during all the isomerization process. Both the distances P – Arg69 guanidinium carbon atom, and P – Lys63 nitrogen, have been reported in the range between 3 – 5 Å. These interactions resulted not to be influenced by the different configurations of the substrate.



**Figure 4.22** Polar plot of the contacts between the amide bond carbonyl oxygen O<sub>2</sub> and the side chain atoms of Pin1 active site residues, as a function of the reaction coordinate  $\zeta$ . The distances (Å) are labelled on the basis of the fixed point charges on the atoms (parm94) [114].

Residue	Side chain parm94 atoms type / Residue [114]	Charges (e) [114]
Cys113	SH / CYS	-0.312
	HS / CYS	0.193
His59 – His157	H5 / HIE	0.143
Ser154	OH / SER	-0.655
	HO / SER	0.427

**Table 4.2** Side chain atom types and relative charges for Cys113, His59, His157, Ser154, based on the Amber parm94 [114].



**Figure 4.23** Polar plot of the contacts between the phosphate moiety of the substrate and the side chains of Lys63, Arg68, and Arg69. For Lys63, the distance between the phosphorus atom P and the terminal nitrogen, was monitored during the isomerization process. For Arg68, and Arg69, the distances were evaluated with respect to the carbon atom of the guanidinium groups.

## 4.4 Discussion

The main result of this study was the identification of a combination of factors which have been shown to contribute to the Pin1-catalyzed *cis-trans* isomerization of the peptidyl-prolyl amide bond. In particular, the cooperation of these enzymatic effects, made a significant contribution both in enhancing the rate of the catalysis, and in lowering the activation barrier. In the free energy landscape along the reaction coordinates  $\zeta / \Psi_{\text{PRO}}$  for Pin1 isomerization (Figure 4.14), there were clearly two energy minima corresponding to the CIS1 and TRANS1 configurations. Along the pathway connecting CIS1 to TRANS1, passing through the TS1 (*syn/exo* transition state), the energy barrier was found equal to 15 kcal/mol, approximately 7 kcal/mol lower than the reaction in bulk solvent. On the contrary, the inverse path through the *anti/endo* TS2 was strongly disfavoured by the high energy associated to the transition state. These findings were in line with recent experimental [99] and theoretical [106] studies. In particular, Greenwood and co-workers [99] by means of NMR lineshape analysis, determined activation barriers for the isomerization in presence or absence of Pin1, of 13.2 and 20.2 kcal/mol, respectively.

### 4.4.1 The Entropy Trap

As clearly shown from a comparison of the PMFs in Figure 4.11 and 4.14, the configurational space available for the substrate in bulk solvent, has been drastically restricted in Pin1 active site. Pin1 allowed the substrate just to sample the reaction coordinate space, enclosed approximately by  $-30^\circ \leq \zeta \leq 180^\circ$  and  $-60^\circ \leq \Psi_{\text{PRO}} \leq 60^\circ$ , straightforwardly driving the isomerization toward the CIS1 - TS1 - TRANS1 pathway, in both the clockwise and counter-clockwise directions.

This entropy loss, arising upon substrate binding, is the result of the enzyme environment, and in particular, of the steric clashes mainly occurring between the substrate C-terminal portion and the active site residues. In a certain way, Pin1 active site seemed to generate a sort of entropy trap, reducing motions and torsional degrees of freedom. The idea that the entropic contributions could determine an acceleration of the enzymatic reactions, has been well-established [270-273]. In particular, regarding the Pin1-catalyzed reaction, this entropic factor has shown several important implications. First of all, it was found to reduce the low-energy conformational changes of the substrate. As a matter of fact, Figure 4.17 highlighted a significant preference for the proline ring to adopt the down-pucker conformation along all the isomerization process. This finding was in contrast with the evidences coming from the reaction in bulk solvent, where an higher frequency of occurrence of the up-puckers was observed (Figure 4.12 C). Therefore, Pin1 induced a low-frequency puckering motions, preferentially stabilizing the down-pucker state: notice that the free energy barrier to ring flip from the down-to-up conformations, was estimated to be 2.5 and 3.2 kcal/mol for *trans* and *cis* isomers, respectively [217]. The second important effect of the substrate entropy loss, was found in the stabilization of pro-catalytic conformations of the peptide. In particular, in all the configurational states, no extended conformations were found. Moreover, an optimal orientation of the amide hydrogen of the residue preceding the proline, H<sub>1</sub>, toward prolyl nitrogen N<sub>2</sub>, was observed. It has been established the role of this intramolecular interaction to speed up the reaction, by favouring the prolyl nitrogen pyramidalization (autocatalysis) [205,215,220]. Interestingly, in this context, an increasing of the positive degree of the nitrogen pyramidalization has been clearly shown in Figure 4.15, compared to the non-enzymatic *cis-trans* isomerization (Figure 4.13 A).

The entropic contribution seemed to play a key role in the isomerization process of other PPIase enzymes. Indeed, the effect of the restriction of the substrate conformational space within the active site, was also characterized by Ladani and co-workers for the reaction mechanism of CypA [239]. In particular, they investigated the relative change in conformational entropy between transition state and *cis/trans* states, which was found to contribute favorable to the free energy of stabilizing the transition state.

## 4.4.2 *Cis* Ground State Destabilization

Another important effect, strictly related to the restriction of the reaction coordinate space induced by Pin1 active site, was the shifting of the CIS1 basin from  $\zeta \cong 0^\circ / \Psi_{\text{PRO}} \cong -30^\circ$  to  $\zeta \cong -15^\circ / \Psi_{\text{PRO}} \cong 30^\circ$ , as shown in Figure 4.14. The analysis of the corresponding conformation assumed by the substrate in this state, pointed out to the stabilization of a less probable conformation occurring in bulk solvent, which has been estimated to be approximately 6 kcal/mol higher in energy. In other words, these results suggested a destabilization of the *cis* ground state upon binding. A similar enzymatic effect was also observed for the peptidyl-prolyl *cis-trans* reaction mechanisms of human CypA [229,235] and FKBP [242,243], by means of experimental or theoretical procedures. However, Pin1 destabilization effect seemed not to be implicated in favouring a distorted prolyl amide bond, as observed for the other enzymes. The results obtained from umbrella sampling simulations, highlighted the absence of the intramolecular hydrogen bond  $\text{H}_1 - \text{O}_3$  in CIS1 (Figure 4.18), which has been shown an high frequency of occurrence during the uncatalyzed isomerization. In particular, within Pin1 active site, the dihedral  $\Phi_{\text{P}_{\text{SER}}}$  of the *cis* substrate was forced to go from  $-80^\circ$  to  $-150^\circ$ , leading to the formation of the hydrogen bond between  $\text{O}_3$  and the hydroxyl group of Ser154. Moreover, the absence of the intramolecular  $\text{H}_1 - \text{O}_3$  was also observed in TS1 state (Figure 4.20). This finding brought to light an important mechanistic aspect of Pin1-catalyzed reaction, with a significant impact on the energy barrier to rotation. In fact, no intramolecular hydrogen bonds have to be disrupted during the isomerization, leading to an estimated reduction of the activation barrier of about 2.2 kcal/mol. In other words, upon binding Pin1 forced the substrate to adopt a conformation which has shown a low frequency of occurrence in water, to disfavour the intramolecular hydrogen bond  $\text{H}_1 - \text{O}_3$ , and promoting in this way the peptidyl-prolyl *cis-trans* isomerization.

### 4.4.3 The *Hydrogen Bond Shuttle-Assisted* Mechanism

In order to investigate the role assumed by the enzyme environment during the isomerization process, the distances between the prolyl amide bond oxygen O<sub>2</sub> and the active site aminoacids, have been reported as a function of the reaction coordinate  $\zeta$  in the polar plot in Figure 4.22. This analysis revealed the formation of multiple hydrogen bond interactions between the carbonyl oxygen and the partial positive charges of the active site side chain atoms, during the *cis-trans* isomerization. However, these interactions have been shown to take place approximately at  $\zeta \geq 30^\circ$ , and therefore when the peptide amide bond was already in a partially twisted conformation. Which factors are responsible to promote an initial deviation of the prolyl bond planarity? As shown in Figure 4.22, when the substrate was in CIS1 configuration, the carbonyl oxygen was in close contact with both the hydrogen (HS), and the sulphur atom (SH) of the sulfhydryl side chain of Cys113. The former has a stabilizing effect on the ground state, in contrast to the latter. In particular, local fluctuations of Cys113 side chain could shift this equilibrium of electrostatic interactions toward the repulsive components. This sudden and temporary break of the equilibrium toward the O<sub>2</sub> – SH contacts, could induce a starting rotation of the prolyl amide bond, leading to a partially twisted conformation. At this point (starting from  $\zeta \geq 30^\circ$ ), the carbonyl oxygen was involved in a series of subsequent hydrogen bond interactions with Cys113, His59, His 157, and Ser154, which stabilized the distorted amide bond, drove the isomerization through the TS1, and assisting the barrier crossing until the TRANS1 was reached. This *hydrogen bond shuttle-assisted* mechanism, revealed a remarkable similarity with the solvent-assisted catalysis proposed by Ke and co-workers for CypA [236]. In particular, the solvent-assisted mechanism arose from the observation, in CypA/Ala-Pro complex, of a conserved structural water molecule which was placed in a good orientation to stabilize the substrate transition state through hydrogen bond to the carbonyl oxygen.

The role played by the basic aminoacids of the substrate phosphate binding pocket in the isomerization, was also investigated. In Figure 4.23 the distances between phosphate moiety and Lys63, Arg68, and Arg69, were reported in a polar plot, using  $\zeta$  as reaction coordinate. Recently, Velazquez and co-workers [106], pointed out to the crucial roles of such aminoacids to stabilize the transition state configuration of the substrate, and therefore to promote the catalysis. In particular, their results highlighted that Arg69 and Lys63, were involved in tighter interactions with the phosphate when the substrate was in the transition state. From Figure 4.23, it has been shown that although a slightly preference for the TRANS1 state, Arg68 resulted not to get involved in short-ranged interactions (distances ranging from 6 to 13 Å). This finding was in line with mutagenesis data, suggesting a negligible role of Arg68 for the PPlase activity [104,202]. On the other hand, tighter interactions were evaluated between the phosphate group and Arg69/Lys63 (distances of about 3 – 5 Å). However these short-ranged contacts were retained during all the isomerization process, occurring without a preference for a particular configurational state of the substrate. These results supported the well-established roles of Arg69 and Lys63 in anchoring the substrate during the *cis-trans* isomerization, excluding their potential involvement in the transition state stabilization. In this regard, Figure 4.22 suggested that the stabilization of the twisted amide bond was mainly due to the active site *hydrogen bond shuttle*.




## 4.5 Conclusions

In this chapter, the investigation of the catalytic mechanism of the peptidyl-prolyl *cis-trans* isomerase Pin1 was presented. First of all, unbiased MD simulations were carried out in order to test the proposed C [97] and NC [104] mechanisms for the isomerization. Large fluctuations registered for the substrate and for binding site residues during dynamics, excluded the feasibility of the nucleophilic catalysis. Umbrella sampling was therefore carried out for both the isomerization in bulk solvent, and Pin1-catalyzed within the NC reaction mechanism. The free energy profiles, showed activation barriers consistent with experimental NMR measurements [99], and with previous theoretical studies [106]. The activation barrier for the isomerization within Pin1 active site was found to be around 7 kcal/mol lower than the uncatalyzed mechanism. Several enzymatic effects, directly linked to the acceleration of the prolyl bond isomerization, were identified and characterized. These contributions have been also proposed for CypA and FKBP reactions, suggesting a common driving force behind the catalytic power of the PPIase family members. The conformational entropy loss of the substrate upon binding and the *cis* ground state destabilization, were found to promote the isomerization. This was mainly due to the reduction of low-energy conformational changes, and the stabilization of a pro-catalytic conformation for the substrate. In particular, this conformation was characterized by 1) the autocatalytic interaction, 2) an high degree of positive pyramidalization of the proline nitrogen, and 3) the absence of intramolecular hydrogen bond  $H_1 - O_3$ , which has shown to reduce the barrier to rotation of approximately 2.2 kcal/mol. Moreover, the observation of multiple hydrogen bond interactions between the carbonyl oxygen of the twisted proline amide bond and the side chains of the active site residues, suggested an *hydrogen-bond shuttle-assisted* model of catalysis. These interactions have been found to

stabilize the distorted amide bond, and to assist the carbonyl rotation until the ground state was reached.

Therefore, considering the *cis* configuration as reference state, the following reaction mechanism for the *cis* to *trans* isomerization was proposed:

- (i) Binding of the substrate in *cis* configuration;
- (ii) reduction of the conformational entropy of the substrate, and stabilization of a pro-catalytic conformation for the *cis* state;
- (iii) *hydrogen bond shuttle-assisted* isomerization;
- (iv) unbinding of the substrate in *trans* configuration.




---

## **Chapter 5.**

# **Conclusions**

---



In this thesis I described the theory and application of several computational methods in solving medicinal chemistry and biophysical tasks. I pointed out to the valuable information which could be achieved by means of computer simulations and to the possibility to predict the outcome of traditional experiments. Nowadays, computer represents an invaluable tool for chemists. Presenting the major fields of application of computational methods as well as their theoretical backgrounds, represented the main topic of Chapter 1, and Chapter 2.

The development of an automated docking protocol for hERG potassium channel blockers has been presented in Chapter 3. hERG is a target of great interest in drug development and drug safety. The drug-induced hERG blockade has been associated to potentially lethal proarrhythmic conditions. Providing a fast and cheap strategy to assess the blockade activity at the early stages of the drug discovery process, is a challenging task, and the aim of this project.

In particular, a strategy that explicitly takes into account the conformations of the channel, their possible intrinsic symmetry, and the role played by the configurational entropy of ligands, was designed. The protocol was developed on a series of congeneric sertindole derivatives, and it has been shown to satisfactorily explain the structure-activity relationships for this set of blockers, and to provide qualitative and quantitative insights about their blocking ability. The protocol was then successfully applied to a series of structurally unrelated blockers.

In Chapter 4, I have presented the investigation of the catalytic mechanism of peptidyl-prolyl *cis-trans* isomerase Pin1. Protein phosphorylation has been shown to be involved in a variety of cellular signalling pathways. In this context, human Pin1, has a key role in the regulation of pSer/Thr-Pro proteins, acting as a molecular timer of the cell cycle. After recognition of the phosphorylated motifs, Pin1 catalyzes the rapid *cis-trans* isomerization of proline amide bonds of substrates, playing a critical role in maintaining the equilibrium between the two isoforms. Although the great interest arisen on this enzyme, mainly due to the well-known impact of Pin1 functionality on the onset of several pathological disorders, its catalytic mechanism has long been debated.

The application of umbrella sampling techniques allowed to shed lights on the catalytic process, highlighting several enzymatic effects which have been shown to accelerate the reaction, and providing new mechanistic insights on the isomerization. The combination of entropic effects and ground state destabilization, has been found to play a crucial role in Pin1-catalyzed reaction. These effects were also observed in the reaction mechanisms of other enzymes with isomerase activity, suggesting a very close catalytic pathway. Moreover, during the isomerization, the formation of multiple hydrogen bonds between the carbonyl oxygen of the twisted prolyl bond and the active site residues, have been observed. These interactions have been shown to stabilize the transition state and to drive the peptide bond rotation. This finding suggested a new *hydrogen bond shuttle-assisted* model of catalysis.

## Acknowledgements

I am very grateful to Prof. Maurizio Recanatini for giving me the possibility to join his research group, for trusting me all the way, for his encouragement, guidance, and all the fruitful discussions. I would like to thank Prof. Andrea Cavalli for the constructive suggestions, and insightful discussions during all my Ph.D. period.

Special thanks to Dr. Matteo Masetti for supervising all my research projects, for his constant guidance, critical revisions, and inspiring discussions. Thanks Matteo, for your invaluable support through this long research period, and for your contagious enthusiasm for research and facing new challenges.

I would like to thank my colleagues, Dr. Federico Falchi, Rosa Buonfiglio, Mariarosaria Ferraro, Elisa Uliassi, Mattia Bernetti, Dario Gioia, for the pleasant working atmosphere. In particular, I would like to thank Federico and Rosa, for all the nice discussions, and unforgettable moments spent together.

Most importantly, I would like to express my gratitude to my family for their love, support, and encouragement.

## References

1. The Nobel Prize in Chemistry 2013 - Press Release. Nobelprize.org. Nobel Media AB 2013. Web. 27 Feb 2014.
2. Cumming JG, Davis AM, Muresan S, Haeberlein M, Chen H (2013) Chemical predictive modelling to improve compound quality. *Nat Rev Drug Discov* 12: 948-962.
3. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* 46: 3-26.
4. Talele TT, Khedkar SA, Rigby AC (2010) Successful applications of computer aided drug discovery: moving drugs from concept to the clinic. *Curr Top Med Chem* 10: 127-141.
5. Clark DE (2006) What has computer-aided molecular design ever done for drug discovery? *Expert Opin Drug Discov* 1: 103-110.
6. Doman TN, McGovern SL, Witherbee BJ, Kasten TP, Kurumbail R, et al. (2002) Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J Med Chem* 45: 2213-2221.
7. Ferreira RS, Simeonov A, Jadhav A, Eidam O, Mott BT, et al. (2010) Complementarity between a docking and a high-throughput screen in discovering new cruzain inhibitors. *J Med Chem* 53: 4891-4905.
8. Polgar T, Baki A, Szendrei GI, Keseru GM (2005) Comparative virtual and experimental high-throughput screening for glycogen synthase kinase-3beta inhibitors. *J Med Chem* 48: 7946-7959.
9. Tidten-Luksch N, Grimaldi R, Torrie LS, Frearson JA, Hunter WN, et al. (2012) IspE Inhibitors Identified by a Combination of *In Silico* and *In Vitro* High-Throughput Screening. *PLoS ONE* 7: e35792.
10. Bajorath J (2002) Integration of virtual and high-throughput screening. *Nat Rev Drug Discov* 1: 882-894.
11. Liao C, Sitzmann M, Pugliese A, Nicklaus MC (2011) Software and resources for computational medicinal chemistry. *Future Med Chem* 3: 1057-1085.
12. Kairys V, Gilson MK, Fernandes MX (2006) Using protein homology models for structure-based studies: approaches to model refinement. *ScientificWorldJournal* 6: 1542-1554.
13. Cavasotto CN, Phatak SS (2009) Homology modeling in drug discovery: current trends and applications. *Drug Discov Today* 14: 676-683.
14. Hillisch A, Pineda LF, Hilgenfeld R (2004) Utility of homology models in the drug discovery process. *Drug Discov Today* 9: 659-669.
15. Wilson CA, Kreychman J, Gerstein M (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *Journal of Molecular Biology* 297: 233-249.
16. Acharya C, Coop A, Polli JE, Mackerell AD, Jr. (2011) Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach. *Curr Comput Aided Drug Des* 7: 10-22.

17. Zhang W, Xia S, Ye J, Tang Y, Li Z, et al. (2013) Structural features of GABAA receptor antagonists: pharmacophore modeling and 3D-QSAR studies. *Medicinal Chemistry Research* 22: 5961-5972.
18. Hajduk PJ, Huth JR, Tse C (2005) Predicting protein druggability. *Drug Discovery Today* 10: 1675-1682.
19. Overington JP, Al-Lazikani B, Hopkins AL (2006) How many drug targets are there? *Nat Rev Drug Discov* 5: 993-996.
20. Chen YZ, Zhi DG (2001) Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule. *Proteins: Structure, Function, and Bioinformatics* 43: 217-226.
21. Li H, Gao Z, Kang L, Zhang H, Yang K, et al. (2006) TarFisDock: a web server for identifying drug targets with docking approach. *Nucleic Acids Res* 34: W219-224.
22. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE (1982) A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology* 161: 269-288.
23. Yang SY (2010) Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discov Today* 15: 444-450.
24. Zheng R, Chen TS, Lu T (2011) A comparative reverse docking strategy to identify potential antineoplastic targets of tea functional components and binding mode. *Int J Mol Sci* 12: 5200-5212.
25. Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov* 3: 935-949.
26. Zheng M, Liu X, Xu Y, Li H, Luo C, et al. (2013) Computational methods for drug design and discovery: focus on China. *Trends Pharmacol Sci* 34: 549-559.
27. Huang S-Y, Grinter SZ, Zou X (2010) Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions. *Physical Chemistry Chemical Physics* 12: 12899-12908.
28. Rajamani R, Good AC (2007) Ranking poses in structure-based lead discovery and optimization: current trends in scoring function development. *Curr Opin Drug Discov Devel* 10: 308-315.
29. Shoichet BK, McGovern SL, Wei B, Irwin JJ (2002) Lead discovery using molecular docking. *Current Opinion in Chemical Biology* 6: 439-446.
30. Seifert MH, Kraus J, Kramer B (2007) Virtual high-throughput screening of molecular databases. *Curr Opin Drug Discov Devel* 10: 298-307.
31. Rarey M, Kramer B, Lengauer T, Klebe G (1996) A Fast Flexible Docking Method using an Incremental Construction Algorithm. *Journal of Molecular Biology* 261: 470-489.
32. Gupta A, Gandhimathi A, Sharma P, Jayaram B (2007) ParDOCK: an all atom energy based Monte Carlo docking protocol for protein-ligand complexes. *Protein Pept Lett* 14: 632-646.
33. Hart TN, Read RJ (1992) A multiple-start Monte Carlo docking method. *Proteins* 13: 206-222.
34. Liu M, Wang S (1999) MCDOCK: a Monte Carlo simulation approach to the molecular docking problem. *J Comput Aided Mol Des* 13: 435-451.
35. Ördög R, Grolmusz V (2008) Evaluating Genetic Algorithms in Protein-Ligand Docking. In: Mändouli I, Sunderraman R, Zelikovsky A, editors. *Bioinformatics Research and Applications*: Springer Berlin Heidelberg. pp. 402-413.
36. Alonso H, Bliznyuk AA, Greedy JE (2006) Combining docking and molecular dynamic simulations in drug design. *Medicinal Research Reviews* 26: 531-568.
37. Desmet J, Maeyer MD, Hazes B, Lasters I (1992) The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* 356: 539-542.
38. Huang S-Y, Zou X (2007) Ensemble docking of multiple protein structures: Considering protein structural variations in molecular docking. *Proteins: Structure, Function, and Bioinformatics* 66: 399-421.
39. Shoichet BK (2004) Virtual screening of chemical libraries. *Nature* 432: 862-865.



40. Walters WP, Stahl MT, Murcko MA (1998) Virtual screening—an overview. *Drug Discovery Today* 3: 160-178.
41. Stroud RM, Finer-Moore J (2008) *Computational and Structural Approaches to Drug Discovery: Ligand-protein Interactions*: RSC Publishing.
42. Lyne PD (2002) Structure-based virtual screening: an overview. *Drug Discovery Today* 7: 1047-1055.
43. Congreve M, Chessari G, Tisi D, Woodhead AJ (2008) Recent developments in fragment-based drug discovery. *J Med Chem* 51: 3661-3680.
44. Pellecchia M (2009) Fragment-based drug discovery takes a virtual turn. *Nat Chem Biol* 5: 274-275.
45. Waltenberger B, Wiechmann K, Bauer J, Markt P, Noha SM, et al. (2011) Pharmacophore Modeling and Virtual Screening for Novel Acidic Inhibitors of Microsomal Prostaglandin E2 Synthase-1 (mPGES-1). *Journal of Medicinal Chemistry* 54: 3163-3174.
46. Hansch C (1969) Quantitative approach to biochemical structure-activity relationships. *Accounts of Chemical Research* 2: 232-239.
47. Perkins R, Fang H, Tong W, Welsh WJ (2003) Quantitative structure-activity relationship methods: perspectives on drug discovery and toxicology. *Environ Toxicol Chem* 22: 1666-1679.
48. Winkler DA (2002) The role of quantitative structure - activity relationships (QSAR) in biomolecular discovery. *Briefings in Bioinformatics* 3: 73-86.
49. Puzyn T, Leszczynski J, Cronin MTD (2010) *Recent Advances in QSAR Studies: Methods and Applications*: Springer.
50. Dudek AZ, Arodz T, Galvez J (2006) Computational methods in developing quantitative structure-activity relationships (QSAR): a review. *Comb Chem High Throughput Screen* 9: 213-228.
51. Dube D, Periwai V, Kumar M, Sharma S, Singh T, et al. (2012) 3D-QSAR based pharmacophore modeling and virtual screening for identification of novel pteridine reductase inhibitors. *Journal of Molecular Modeling* 18: 1701-1711.
52. Bhatt HG, Patel PK (2012) Pharmacophore modeling, virtual screening and 3D-QSAR studies of 5-tetrahydroquinolinylidene aminoguanidine derivatives as sodium hydrogen exchanger inhibitors. *Bioorganic & Medicinal Chemistry Letters* 22: 3758-3765.
53. Brogi S, Papazafiri P, Roussis V, Tafi A (2013) 3D-QSAR using pharmacophore-based alignment and virtual screening for discovery of novel MCF-7 cell line inhibitors. *European Journal of Medicinal Chemistry* 67: 344-351.
54. Tseng GN (2001) I(Kr): the hERG channel. *J Mol Cell Cardiol* 33: 835-849.
55. Sanguinetti MC, Tristani-Firouzi M (2006) hERG potassium channels and cardiac arrhythmia. *Nature* 440: 463-469.
56. Recanatini M, Poluzzi E, Masetti M, Cavalli A, De Ponti F (2005) QT prolongation through hERG K(+) channel blockade: current knowledge and strategies for the early prediction during drug development. *Med Res Rev* 25: 133-166.
57. Yoshida K, Niwa T (2006) Quantitative structure-activity relationship studies on inhibition of HERG potassium channels. *J Chem Inf Model* 46: 1371-1378.
58. Su BH, Shen MY, Esposito EX, Hopfinger AJ, Tseng YJ (2010) In silico binary classification QSAR models based on 4D-fingerprints and MOE descriptors for prediction of hERG blockage. *J Chem Inf Model* 50: 1304-1318.
59. Keseru GM (2003) Prediction of hERG potassium channel affinity by traditional and hologram qSAR methods. *Bioorg Med Chem Lett* 13: 2773-2775.
60. Coi A, Massarelli I, Murgia L, Saraceno M, Calderone V, et al. (2006) Prediction of hERG potassium channel affinity by the CODESSA approach. *Bioorg Med Chem* 14: 3153-3159.
61. Du-Cuny L, Chen L, Zhang S (2011) A critical assessment of combined ligand- and structure-based approaches to HERG channel blocker modeling. *J Chem Inf Model* 51: 2948-2960.

62. Cavalli A, Poluzzi E, De Ponti F, Recanatini M (2002) Toward a Pharmacophore for Drugs Inducing the Long QT Syndrome: Insights from a CoMFA Study of HERG K<sup>+</sup> Channel Blockers. *Journal of Medicinal Chemistry* 45: 3844-3853.
63. Ekins S, Crumb WJ, Sarazan RD, Wikel JH, Wrighton SA (2002) Three-Dimensional Quantitative Structure-Activity Relationship for Inhibition of Human Ether-a-Go-Go-Related Gene Potassium Channel. *Journal of Pharmacology and Experimental Therapeutics* 301: 427-434.
64. Aronov AM (2006) Common pharmacophores for uncharged human ether-a-go-go-related gene (hERG) blockers. *J Med Chem* 49: 6917-6921.
65. Pearlstein RA, Vaz RJ, Kang J, Chen XL, Preobrazhenskaya M, et al. (2003) Characterization of HERG potassium channel inhibition using CoMSiA 3D QSAR and homology modeling approaches. *Bioorg Med Chem Lett* 13: 1829-1835.
66. Durdagi S, Duff HJ, Noskov SY (2011) Combined receptor and ligand-based approach to the universal pharmacophore model development for studies of drug blockade to the hERG1 pore domain. *J Chem Inf Model* 51: 463-474.
67. Österberg F, Åqvist J (2005) Exploring blocker binding to a homology model of the open hERG K<sup>+</sup> channel using docking and molecular dynamics methods. *FEBS Letters* 579: 2939-2944.
68. Boukharta L, Keränen H, Stary-Weinzinger A, Wallin Gr, de Groot BL, et al. (2011) Computer Simulations of Structure-Activity Relationships for hERG Channel Blockers. *Biochemistry* 50: 6146-6156.
69. Farid R, Day T, Friesner RA, Pearlstein RA (2006) New insights about HERG blockade obtained from protein modeling, potential energy mapping, and docking studies. *Bioorganic & Medicinal Chemistry* 14: 3160-3173.
70. Coi A, Massarelli I, Testai L, Calderone V, Bianucci AM (2008) Identification of "toxicophoric" features for predicting drug-induced QT interval prolongation. *European Journal of Medicinal Chemistry* 43: 2479-2488.
71. Rajamani R, Tounge BA, Li J, Reynolds CH (2005) A two-state homology model of the hERG K<sup>+</sup> channel: application to ligand binding. *Bioorg Med Chem Lett* 15: 1737-1741.
72. Dror RO, Dirks RM, Grossman JP, Xu H, Shaw DE (2012) Biomolecular simulation: a computational microscope for molecular biology. *Annu Rev Biophys* 41: 429-452.
73. Liu H, Dastidar SG, Lei H, Zhang W, Lee MC, et al. (2008) Conformational changes in protein function. *Methods Mol Biol* 443: 258-275.
74. Velazquez HA, Hamelberg D (2011) Conformational Selection in the Recognition of Phosphorylated Substrates by the Catalytic Domain of Human Pin1. *Biochemistry* 50: 9605-9615.
75. Goh CS, Milburn D, Gerstein M (2004) Conformational changes associated with protein-protein interactions. *Curr Opin Struct Biol* 14: 104-109.
76. Durrant JD, McCammon JA (2011) Molecular dynamics simulations and drug discovery. *BMC Biol* 9: 71.
77. Shaw DE, Dror RO, Salmon JK, Grossman JP, Mackenzie KM, et al. (2009) Millisecond-scale molecular dynamics simulations on Anton. *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*. Portland, Oregon: ACM. pp. 1-11.
78. Harvey MJ, Giupponi G, Fabritiis GD (2009) ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. *Journal of Chemical Theory and Computation* 5: 1632-1639.
79. Karplus M, McCammon JA (2002) Molecular dynamics simulations of biomolecules. *Nat Struct Biol* 9: 646-652.
80. Guvench O, MacKerell AD (2008) Comparison of Protein Force Fields for Molecular Dynamics Simulations. pp. 63-88.
81. Lindorff-Larsen K, Maragakis P, Piana S, Eastwood MP, Dror RO, et al. (2012) Systematic Validation of Protein Force Fields against Experimental Data. *PLoS ONE* 7: e32131.

82. Beauchamp KA, Lin YS, Das R, Pande VS (2012) Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. *J Chem Theory Comput* 8: 1409-1414.
83. Kästner J (2011) Umbrella sampling. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1: 932-942.
84. Abrams C, Bussi G (2013) Enhanced Sampling in Molecular Dynamics Using Metadynamics, Replica-Exchange, and Temperature-Acceleration. *Entropy* 16: 163-199.
85. Shirts MR, Pande VS (2005) Comparison of efficiency and bias of free energies computed by exponential averaging, the Bennett acceptance ratio, and thermodynamic integration. *The Journal of Chemical Physics* 122: -.
86. Trzesniak D, Kunz A-PE, van Gunsteren WF (2007) A Comparison of Methods to Compute the Potential of Mean Force. *ChemPhysChem* 8: 162-169.
87. Carter EA, Ciccotti G, Hynes JT, Kapral R (1989) Constrained reaction coordinate dynamics for the simulation of rare events. *Chemical Physics Letters* 156: 472-477.
88. Ciccotti G, Ferrario M (2004) Blue Moon Approach to Rare Events. *Molecular Simulation* 30: 787-793.
89. Torrie GM, Valleau JP (1977) Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics* 23: 187-199.
90. Torrie GM, Valleau JP (1974) Monte Carlo free energy estimates using non-Boltzmann sampling: Application to the sub-critical Lennard-Jones fluid. *Chemical Physics Letters* 28: 578-581.
91. Laio A, Parrinello M (2002) Escaping free-energy minima. *Proceedings of the National Academy of Sciences* 99: 12562-12566.
92. Barducci A, Bonomi M, Parrinello M (2011) Metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1: 826-843.
93. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by Simulated Annealing. *Science* 220: 671-680.
94. Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters* 314: 141-151.
95. Earl DJ, Deem MW (2005) Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics* 7: 3910-3916.
96. Lu KP, Finn G, Lee TH, Nicholson LK (2007) Prolyl cis-trans isomerization as a molecular timer. *Nat Chem Biol* 3: 619-629.
97. Ranganathan R, Lu KP, Hunter T, Noel JP (1997) Structural and functional analysis of the mitotic rotamase Pin1 suggests substrate recognition is phosphorylation dependent. *Cell* 89: 875-886.
98. Yaffe MB, Schutkowski M, Shen M, Zhou XZ, Stukenberg PT, et al. (1997) Sequence-Specific and Phosphorylation-Dependent Proline Isomerization: A Potential Mitotic Regulatory Mechanism. *Science* 278: 1957-1960.
99. Greenwood AI, Rogals MJ, De S, Lu KP, Kovrigin EL, et al. (2011) Complete determination of the Pin1 catalytic domain thermodynamic cycle by NMR lineshape analysis. *J Biomol NMR* 51: 21-34.
100. Tchaicheyan O (2004) Is peptide bond cis/trans isomerization a key stage in the chemo-mechanical cycle of motor proteins? *FASEB J* 18: 783-789.
101. Lu KP, Zhou XZ (2007) The prolyl isomerase PIN1: a pivotal new twist in phosphorylation signalling and disease. *Nat Rev Mol Cell Biol* 8: 904-916.
102. Ryo A, Liou YC, Lu KP, Wulf G (2003) Prolyl isomerase Pin1: a catalyst for oncogenesis and a potential therapeutic target in cancer. *J Cell Sci* 116: 773-783.
103. Yeh ES, Means AR (2007) PIN1, the cell cycle and cancer. *Nat Rev Cancer* 7: 381-388.

104. Behrsin CD, Bailey ML, Bateman KS, Hamilton KS, Wahl LM, et al. (2007) Functionally important residues in the peptidyl-prolyl isomerase Pin1 revealed by unigenic evolution. *J Mol Biol* 365: 1143-1162.
105. Vöhringer-Martinez E, Duarte F, Toro-Labbé A (2012) How Does Pin1 Catalyze the Cis–Trans Prolyl Peptide Bond Isomerization? A QM/MM and Mean Reaction Force Study. *The Journal of Physical Chemistry B* 116: 12972-12979.
106. Velazquez HA, Hamelberg D (2013) Conformation-Directed Catalysis and Coupled Enzyme–Substrate Dynamics in Pin1 Phosphorylation-Dependent Cis–Trans Isomerase. *The Journal of Physical Chemistry B* 117: 11509-11517.
107. Lewars E (2010) *Computational Chemistry: Introduction to the Theory and Applications of Molecular and Quantum Mechanics*: Springer.
108. Leach AR (2001) *Molecular Modelling: Principles and Applications*: Prentice Hall.
109. Young D (2004) *Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems*: Wiley.
110. Yang L, Tan C-h, Hsieh M-J, Wang J, Duan Y, et al. (2006) New-Generation Amber United-Atom Force Field. *The Journal of Physical Chemistry B* 110: 13166-13176.
111. Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, et al. (1984) A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society* 106: 765-784.
112. Hagler AT, Huler E, Lifson S (1974) Energy functions for peptides and proteins. I. Derivation of a consistent force field including the hydrogen bond from amide crystals. *Journal of the American Chemical Society* 96: 5319-5327.
113. Weiner SJ, Kollman PA, Nguyen DT, Case DA (1986) An all atom force field for simulations of proteins and nucleic acids. *Journal of Computational Chemistry* 7: 230-252.
114. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, et al. (1995) A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* 117: 5179-5197.
115. Kollman P, Dixon, R., Cornell, W., Fox, T., Chipot, C., and Pohorille, A. (1997) In "Computer Simulations of Biomolecular Systems," Vol. 3 (W. F. van Gunsteren, P. K. Weiner, and A. J. Wilkinson, Eds.), pp. 83–96. Kluwer Academic Publishers, Dordrecht, The Netherlands.
116. Wang J, Cieplak P, Kollman PA (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry* 21: 1049-1074.
117. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, et al. (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics* 65: 712-725.
118. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, et al. (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function, and Bioinformatics* 78: 1950-1958.
119. D. A. Case TAD, T. E. Cheatham, III, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, R. C. Walker, W. Zhang, K. M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A. W. Goetz, I. Kolossváry, K. F. Wong, F. Paesani, J. Vanicek, R. M. Wolf, J. Liu, X. Wu, S. R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D. R. Roe, D. H. Mathews, M. G. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, and P. A. Kollman (2012) *AMBER 12* (University of California, San Francisco, CA).
120. Wang Z-X, Zhang W, Wu C, Lei H, Cieplak P, et al. (2006) Strike a balance: Optimization of backbone torsion parameters of AMBER polarizable force field for simulations of proteins and peptides. *Journal of Computational Chemistry* 27: 781-790.

121. Duan Y, Wu C, Chowdhury S, Lee MC, Xiong G, et al. (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of Computational Chemistry* 24: 1999-2012.
122. Cieplak P, Caldwell J, Kollman P (2001) Molecular mechanical models for organic and biological systems going beyond the atom centered two body additive approximation: aqueous solution free energies of methanol and N-methyl acetamide, nucleic acid base, and amide hydrogen bonding and chloroform/water partition coefficients of the nucleic acid bases. *Journal of Computational Chemistry* 22: 1048-1057.
123. Cramer CJ (2005) *Essentials of Computational Chemistry: Theories and Models*: Wiley.
124. Born M, Oppenheimer R (1927) Zur Quantentheorie der Molekeln. *Annalen der Physik* 389: 457-484.
125. Gross EKV, Dreizler RM, Division NATOSA (1995) *Density Functional Theory*: Springer.
126. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* 21: 1087-1092.
127. Ryckaert J-P, Ciccotti G, Berendsen HJC (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* 23: 327-341.
128. Verlet L (1967) Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review* 159: 98-103.
129. Swope WC, Andersen HC, Berens PH, Wilson KR (1982) A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of Chemical Physics* 76: 637-649.
130. Chipot C, Pohorille A (2007) *Free Energy Calculations: Theory and Applications in Chemistry and Biology*: Physica-Verlag.
131. Gao YQ (2008) An integrate-over-temperature approach for enhanced sampling. *The Journal of Chemical Physics* 128: -.
132. Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, Kollman PA (1992) THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry* 13: 1011-1021.
133. Taylor RD, Jewsbury PJ, Essex JW (2002) A review of protein-small molecule docking methods. *J Comput Aided Mol Des* 16: 151-166.
134. Andrusier N, Mashich E, Nussinov R, Wolfson HJ (2008) Principles of flexible protein-protein docking. *Proteins* 73: 271-289.
135. Huang SY, Zou X (2007) Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking. *Proteins* 66: 399-421.
136. Clark DE, Mannhold R, Kubinyi H, Timmerman H (2008) *Evolutionary Algorithms in Molecular Design*: Wiley.
137. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, et al. (2009) AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry* 30: 2785-2791.
138. Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, et al. (2000) Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Accounts of Chemical Research* 33: 889-897.
139. Gohlke H, Kiel C, Case DA (2003) Insights into Protein-Protein Binding by Binding Free Energy Calculation and Free Energy Decomposition for the Ras-Raf and Ras-RalGDS Complexes. *Journal of Molecular Biology* 330: 891-913.
140. Gilson MK, Zhou H-X (2007) Calculation of Protein-Ligand Binding Affinities\*. *Annual Review of Biophysics and Biomolecular Structure* 36: 21-42.

141. Hou T, Wang J, Li Y, Wang W (2010) Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 1. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations. *Journal of Chemical Information and Modeling* 51: 69-82.
142. Kuhn B, Gerber P, Schulz-Gasch T, Stahl M (2005) Validation and Use of the MM-PBSA Approach for Drug Discovery. *Journal of Medicinal Chemistry* 48: 4040-4048.
143. Rastelli G, Rio AD, Degliesposti G, Sgobba M (2010) Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA. *Journal of Computational Chemistry* 31: 797-810.
144. Xu L, Sun H, Li Y, Wang J, Hou T (2013) Assessing the Performance of MM/PBSA and MM/GBSA Methods. 3. The Impact of Force Fields and Ligand Charge Models. *The Journal of Physical Chemistry B* 117: 8408-8421.
145. Yap YG, Camm AJ (2003) Drug induced QT prolongation and torsades de pointes. *Heart* 89: 1363-1372.
146. De Ponti F, Poluzzi E, Montanaro N, Ferguson J (2000) QTc and psychotropic drugs. *The Lancet* 356: 75-76.
147. Raschi E, Vasina V, Poluzzi E, De Ponti F (2008) The hERG K<sup>+</sup> channel: target and antitarget strategies in drug development. *Pharmacological Research* 57: 181-195.
148. Shah R (2005) Drugs, QT Interval Prolongation and ICH E14. *Drug Safety* 28: 115-125.
149. Choe H, Nah KH, Lee SN, Lee HS, Lee HS, et al. (2006) A novel hypothesis for the binding mode of HERG channel blockers. *Biochemical and Biophysical Research Communications* 344: 72-78.
150. Stansfeld PJ, Gedeck P, Gosling M, Cox B, Mitcheson JS, et al. (2007) Drug block of the hERG potassium channel: Insight from modeling. *Proteins: Structure, Function, and Bioinformatics* 68: 568-580.
151. Mitcheson JS, Chen J, Lin M, Culberson C, Sanguinetti MC (2000) A structural basis for drug-induced long QT syndrome. *Proceedings of the National Academy of Sciences* 97: 12329-12333.
152. Fernandez D, Ghanta A, Kauffman GW, Sanguinetti MC (2004) Physicochemical Features of the hERG Channel Drug Binding Site. *Journal of Biological Chemistry* 279: 10120-10127.
153. Pearlstein RA, Vaz RJ, Kang J, Chen X-L, Preobrazhenskaya M, et al. (2003) Characterization of HERG potassium channel inhibition using CoMSiA 3D QSAR and homology modeling approaches. *Bioorganic & Medicinal Chemistry Letters* 13: 1829-1835.
154. Recanatini M, Cavalli A, Masetti M (2008) Modeling HERG and its interactions with drugs: recent advances in light of current potassium channel simulations. *ChemMedChem* 3: 523-535.
155. Lin J-H, Perryman AL, Schames JR, McCammon JA (2002) Computational Drug Design Accommodating Receptor Flexibility: The Relaxed Complex Scheme. *Journal of the American Chemical Society* 124: 5632-5633.
156. Masetti M, Cavalli A, Recanatini M (2008) Modeling the hERG potassium channel in a phospholipid bilayer: Molecular dynamics and drug docking studies. *Journal of Computational Chemistry* 29: 795-808.
157. Tseng GN, Sonawane KD, Korolkova YV, Zhang M, Liu J, et al. (2007) Probing the outer mouth structure of the HERG channel with peptide toxin footprinting and molecular modeling. *Biophys J* 92: 3524-3540.
158. Xiang Z, Soto CS, Honig B (2002) Evaluating conformational free energies: The colony energy and its application to the problem of loop prediction. *Proceedings of the National Academy of Sciences* 99: 7432-7437.
159. Lee J, Seok C (2008) A statistical rescoring scheme for protein-ligand docking: Consideration of entropic effect. *Proteins: Structure, Function, and Bioinformatics* 70: 1074-1083.
160. Totrov M, Abagyan R (2008) Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr Opin Struct Biol* 18: 178-184.

161. Wang J, Morin P, Wang W, Kollman PA (2001) Use of MM-PBSA in Reproducing the Binding Free Energies to HIV-1 RT of TIBO Derivatives and Predicting the Binding Mode to HIV-1 RT of Efavirenz by Docking and MM-PBSA. *Journal of the American Chemical Society* 123: 5221-5230.
162. Šali A, Potterton L, Yuan F, van Vlijmen H, Karplus M (1995) Evaluation of comparative protein modeling by MODELLER. *Proteins: Structure, Function, and Bioinformatics* 23: 318-326.
163. Morris AL, MacArthur MW, Hutchinson EG, Thornton JM (1992) Stereochemical quality of protein structure coordinates. *Proteins: Structure, Function, and Bioinformatics* 12: 345-364.
164. Engh RA, Huber R (1991) Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallographica Section A* 47: 392-400.
165. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography* 26: 283-291.
166. Smart OS, Neduvellil JG, Wang X, Wallace BA, Sansom MSP (1996) HOLE: A program for the analysis of the pore dimensions of ion channel structural models. *Journal of Molecular Graphics* 14: 354-360.
167. Ihaka R, Gentleman R (1996) R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics* 5: 299-314.
168. Everitt BS, Landau S, Leese M (2001) *Cluster Analysis*: Wiley.
169. Gasteiger J, Marsili M (1980) Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron* 36: 3219-3228.
170. Fogolari F, Tosatto SCE (2005) Application of MM/PBSA colony free energy to loop decoy discrimination: Toward correlation between energy and root mean square deviation. *Protein Science* 14: 889-901.
171. Paulsen JL, Anderson AC (2009) Scoring Ensembles of Docked Protein:Ligand Interactions for Virtual Lead Optimization. *Journal of Chemical Information and Modeling* 49: 2813-2819.
172. Pearlman DA, Charifson PS (2001) Are Free Energy Calculations Useful in Practice? A Comparison with Rapid Scoring Functions for the p38 MAP Kinase Protein System†. *Journal of Medicinal Chemistry* 44: 3417-3423.
173. Ferrari AM, Degliesposti G, Sgobba M, Rastelli G (2007) Validation of an automated procedure for the prediction of relative free energies of binding on a set of aldose reductase inhibitors. *Bioorganic & Medicinal Chemistry* 15: 7865-7877.
174. Weis A, Katebzadeh K, Söderhjelm P, Nilsson I, Ryde U (2006) Ligand Affinities Predicted with the MM/PBSA Method: Dependence on the Simulation Method and the Force Field. *Journal of Medicinal Chemistry* 49: 6596-6606.
175. Case DA, Darden TA, Cheatham ITE, Simmerling CL, Wang J, et al. (2010) AMBER 11.
176. Hawkins GD, Cramer CJ, Truhlar DG (1995) Pairwise solute descreening of solute charges from a dielectric medium. *Chemical Physics Letters* 246: 122-129.
177. Hawkins GD, Cramer CJ, Truhlar DG (1996) Parametrized Models of Aqueous Free Energies of Solvation Based on Pairwise Descreening of Solute Atomic Charges from a Dielectric Medium. *The Journal of Physical Chemistry* 100: 19824-19839.
178. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) Development and testing of a general amber force field. *Journal of Computational Chemistry* 25: 1157-1174.
179. Bayly CI, Cieplak P, Cornell W, Kollman PA (1993) A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry* 97: 10269-10280.
180. Cornell WD, Cieplak P, Bayly CI, Kollmann PA (1993) Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation. *Journal of the American Chemical Society* 115: 9620-9631.
181. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, et al. (2009) Gaussian 09.

182. Sitkoff D, Sharp KA, Honig B (1994) *J Phys Chem* 98: 1978.
183. Knape K, Linder T, Wolschann P, Beyer A, Stary-Weinzinger A (2011) *In silico* Analysis of Conformational Changes Induced by Mutation of Aromatic Binding Residues: Consequences for Drug Binding in the hERG K<sup>+</sup> Channel. *PLoS ONE* 6: e28778.
184. Craig IR, Essex JW, Spiegel K (2010) Ensemble Docking into Multiple Crystallographically Derived Protein Structures: An Evaluation Based on the Statistical Analysis of Enrichments. *Journal of Chemical Information and Modeling* 50: 511-524.
185. Yoon S, Welsh WJ (2003) Identification of a Minimal Subset of Receptor Conformations for Improved Multiple Conformation Docking and Two-Step Scoring. *Journal of Chemical Information and Computer Sciences* 44: 88-96.
186. Zachariae U, Giordanetto F, Leach AG (2009) Side Chain Flexibilities in the Human Ether-a-go-go Related Gene Potassium Channel (hERG) Together with Matched-Pair Binding Studies Suggest a New Binding Mode for Channel Blockers. *Journal of Medicinal Chemistry* 52: 4266-4276.
187. Cavalli A, Buonfiglio R, Ianni C, Masetti M, Ceccarini L, et al. (2012) Computational Design and Discovery of "Minimally Structured" hERG Blockers. *Journal of Medicinal Chemistry* 55: 4010-4014.
188. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. *J Mol Graph* 14: 33-38, 27-38.
189. Pufall MA, Graves BJ (2002) AUTOINHIBITORY DOMAINS: Modular Effectors of Cellular Regulation. *Annual Review of Cell and Developmental Biology* 18: 421-462.
190. Schlessinger J (2003) Autoinhibition Control. *Science* 300: 750-752.
191. Gao Z-G, Jacobson KA (2013) Allosteric modulation and functional selectivity of G protein-coupled receptors. *Drug Discovery Today: Technologies* 10: e237-e243.
192. Wang Y-C, Peterson SE, Loring JF (2014) Protein post-translational modifications and regulation of pluripotency in human stem cells. *Cell Res* 24: 143-160.
193. Kho Y, Kim SC, Jiang C, Barma D, Kwon SW, et al. (2004) A tagging-via-substrate technology for detection and proteomics of farnesylated proteins. *Proceedings of the National Academy of Sciences of the United States of America* 101: 12479-12484.
194. Mann M, Jensen ON (2003) Proteomic analysis of post-translational modifications. *Nat Biotech* 21: 255-261.
195. Deribe YL, Pawson T, Dikic I (2010) Post-translational modifications in signal integration. *Nat Struct Mol Biol* 17: 666-672.
196. Harding MW, Galat A, Uehling DE, Schreiber SL (1989) A receptor for the immuno-suppressant FK506 is a cis-trans peptidyl-prolyl isomerase. *Nature* 341: 758-760.
197. Takahashi N, Hayano T, Suzuki M (1989) Peptidyl-prolyl cis-trans isomerase is the cyclosporin A-binding protein cyclophilin. *Nature* 337: 473-475.
198. Lu KP, Suizu F, Zhou XZ, Finn G, Lam P, et al. (2006) Targeting carcinogenesis: A role for the prolyl isomerase Pin1? *Molecular Carcinogenesis* 45: 397-402.
199. Zhou XZ, Lu PJ, Wulf G, Lu KP (1999) Phosphorylation-dependent prolyl isomerization: a novel signaling regulatory mechanism. *Cellular and Molecular Life Sciences CMLS* 56: 788-806.
200. Nicholson LK, Lu KP (2007) Prolyl cis-trans Isomerization as a Molecular Timer in Crk Signaling. *Molecular Cell* 25: 483-485.
201. Lu KP (2004) Pinning down cell signaling, cancer and Alzheimer's disease. *Trends in biochemical sciences* 29: 200-209.
202. Daum S, Fanghänel J, Wildemann D, Schiene-Fischer C (2006) Thermodynamics of Phosphopeptide Binding to the Human Peptidyl Prolyl cis/trans Isomerase Pin1. *Biochemistry* 45: 12125-12135.



203. Namanja AT, Peng T, Zintsmaster JS, Elson AC, Shakour MG, et al. (2007) Substrate Recognition Reduces Side-Chain Flexibility for Conserved Hydrophobic Residues in Human Pin1. *Structure* 15: 313-327.
204. Bailey ML, Shilton BH, Brandl CJ, Litchfield DW (2008) The Dual Histidine Motif in the Active Site of Pin1 Has a Structural Rather than Catalytic Role†. *Biochemistry* 47: 11481-11489.
205. Fischer S, Dunbrack RL, Karplus M (1994) Cis-Trans Imide Isomerization of the Proline Dipeptide. *Journal of the American Chemical Society* 116: 11931-11937.
206. Ramachandran GN, Sasisekharan V (1968) Conformation of Polypeptides and Proteins. In: C.B. Anfinsen MLAJTE, Frederic MR, editors. *Advances in Protein Chemistry*: Academic Press. pp. 283-437.
207. Weiss MS, Jabs A, Hilgenfeld R (1998) Peptide bonds revisited. *Nat Struct Mol Biol* 5: 676-676.
208. Wedemeyer WJ, Welker E, Scheraga HA (2002) Proline Cis-Trans Isomerization and Protein Folding†. *Biochemistry* 41: 14637-14644.
209. Takahashi K, Uchida C, Shin RW, Shimazaki K, Uchida T (2008) Prolyl isomerase, Pin1: new findings of post-translational modifications and physiological substrates in cancer, asthma and Alzheimer's disease. *Cellular and Molecular Life Sciences* 65: 359-375.
210. Beausoleil E, Lubell WD (1996) Steric Effects on the Amide Isomer Equilibrium of Prolyl Peptides. *Synthesis and Conformational Analysis of N-Acetyl-5-tert-butylproline N'-Methylamides*. *Journal of the American Chemical Society* 118: 12902-12908.
211. Pal D, Chakrabarti P (1999) Cis peptide bonds in proteins: residues involved, their conformations, interactions and locations. *Journal of Molecular Biology* 294: 271-288.
212. Tchaicheeyan O (2004) Is peptide bond cis/trans isomerization a key stage in the chemo-mechanical cycle of motor proteins? *The FASEB Journal* 18: 783-789.
213. Grathwohl C, Wüthrich K (1981) Nmr studies of the rates of proline cis-trans isomerization in oligopeptides. *Biopolymers* 20: 2623-2633.
214. Eberhardt ES, Panasik N, Raines RT (1996) Inductive Effects on the Energetics of Prolyl Peptide Bond Isomerization: Implications for Collagen Folding and Stability. *Journal of the American Chemical Society* 118: 12261-12266.
215. Texter FL, Spencer DB, Rosenstein R, Matthews CR (1992) Intramolecular catalysis of a proline isomerization reaction in the folding of dihydrofolate reductase. *Biochemistry* 31: 5687-5691.
216. Kang YK (2004) Ab initio and DFT conformational study of proline dipeptide. *Journal of Molecular Structure: THEOCHEM* 675: 37-45.
217. Kang YK, Young Choi H (2004) Cis-trans isomerization and puckering of proline residue. *Biophysical Chemistry* 111: 135-142.
218. Kee Kang Y, Sook Park H (2005) Ab initio conformational study of N-acetyl-l-proline-N',N'-dimethylamide: a model for polyproline. *Biophysical Chemistry* 113: 93-101.
219. Kang YK (2006) Conformational Preferences of Non-Prolyl and Prolyl Residues. *The Journal of Physical Chemistry B* 110: 21338-21348.
220. Yonezawa Y, Nakata K, Sakakura K, Takada T, Nakamura H (2009) Intra- and Intermolecular Interaction Inducing Pyramidalization on Both Sides of a Proline Dipeptide during Isomerization: An Ab Initio QM/MM Molecular Dynamics Simulation Study in Explicit Water. *Journal of the American Chemical Society* 131: 4535-4540.
221. Melis C, Bussi G, Lummi SCR, Molteni C (2009) Trans-cis Switching Mechanisms in Proline Analogues and Their Relevance for the Gating of the 5-HT3 Receptor. *The Journal of Physical Chemistry B* 113: 12148-12153.
222. Singh UC, Kollman PA (1984) An approach to computing electrostatic charges for molecules. *Journal of Computational Chemistry* 5: 129-145.
223. Fanghanel J, Fischer G (2004) Insights into the catalytic mechanism of peptidyl prolyl cis/trans isomerases. *Front Biosci* 9: 3453-3478.

224. Drakenberg T, Dahlqvist KI, Forsen S (1972) Barrier to internal rotation in amides. IV. N,N-Dimethylamides. Substituent and solvent effects. *The Journal of Physical Chemistry* 76: 2178-2183.
225. Siekierka JJ, Staruch MJ, Hung SH, Sigal NH (1989) FK-506, a potent novel immunosuppressive agent, binds to a cytosolic protein which is distinct from the cyclosporin A-binding protein, cyclophilin. *The Journal of Immunology* 143: 1580-1583.
226. Thali M, Bukovsky A, Kondo E, Rosenwirth B, Walsh CT, et al. (1994) Functional association of cyclophilin A with HIV-1 virions. *Nature* 372: 363-365.
227. Fanghänel JF, G. (2004) Insights Into The Catalytic Mechanism of Peptidyl Prolyl Cis/Trans Isomerases. *Frontiers in Bioscience* 9: 3453-3578.
228. Janowski B, Wöllner S, Schutkowski M, Fischer G (1997) A Protease-Free Assay for Peptidyl Prolylcis/transomerases Using Standard Peptide Substrates. *Analytical Biochemistry* 252: 299-307.
229. Stein RL (1993) Mechanism of Enzymatic and Nonenzymatic Prolyl Cis-Trans Isomerization. In: C.B. Anfinsen JTEFMRDSE, George L, editors. *Advances in Protein Chemistry: Academic Press*. pp. 1-24.
230. Zydowsky LD, Etzkorn FA, Chang HY, Ferguson SB, Stolz LA, et al. (1992) Active site mutants of human cyclophilin A separate peptidyl-prolyl isomerase activity from cyclosporin A binding and calcineurin inhibition. *Protein Science* 1: 1092-1099.
231. Fischer G, Wittmann-Liebold B, Lang K, Kiefhaber T, Schmid FX (1989) Cyclophilin and peptidyl-prolyl cis-trans isomerase are probably identical proteins. *Nature* 337: 476-478.
232. Fischer G, Berger E, Bang H (1989) Kinetic  $\beta$ -deuterium isotope effects suggest a covalent mechanism for the protein folding enzyme peptidylprolyl cis/trans-isomerase. *FEBS Letters* 250: 267-270.
233. Liu J, Albers MW, Chen CM, Schreiber SL, Walsh CT (1990) Cloning, expression, and purification of human cyclophilin in *Escherichia coli* and assessment of the catalytic role of cysteines by site-directed mutagenesis. *Proceedings of the National Academy of Sciences* 87: 2304-2308.
234. Kallen J, Walkinshaw MD (1992) The X-ray structure of a tetrapeptide bound to the active site of human cyclophilin A. *FEBS Letters* 300: 286-290.
235. Harrison RK, Stein RL (1990) Mechanistic studies of peptidyl prolyl cis-trans isomerase: evidence for catalysis by distortion. *Biochemistry* 29: 1684-1689.
236. Ke H, Mayrose D, Cao W (1993) Crystal structure of cyclophilin A complexed with substrate Ala-Pro suggests a solvent-assisted mechanism of cis-trans isomerization. *Proceedings of the National Academy of Sciences* 90: 3324-3328.
237. Hamelberg D, McCammon JA (2008) Mechanistic Insight into the Role of Transition-State Stabilization in Cyclophilin A. *Journal of the American Chemical Society* 131: 147-152.
238. Leone V, Lattanzi G, Molteni C, Carloni P (2009) Mechanism of Action of Cyclophilin A Explored by Metadynamics Simulations. *PLoS Comput Biol* 5: e1000309.
239. Ladani ST, Hamelberg D (2012) Entropic and Surprisingly Small Intramolecular Polarization Effects in the Mechanism of Cyclophilin A. *The Journal of Physical Chemistry B* 116: 10771-10778.
240. Hamelberg D, Mongan J, McCammon JA (2004) Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *The Journal of Chemical Physics* 120: 11919-11929.
241. Kofron JL, Kuzmic P, Kishore V, Colon-Bonilla E, Rich DH (1991) Determination of kinetic constants for peptidyl prolyl cis-trans isomerases by an improved spectrophotometric assay. *Biochemistry* 30: 6127-6134.
242. Park ST, Aldape RA, Futer O, DeCenzo MT, Livingston DJ (1992) PPIase catalysis by human FK506-binding protein proceeds through a conformational twist mechanism. *J Biol Chem* 267: 3316-3324.

243. Fischer S, Michnick S, Karplus M (1993) A mechanism for rotamase catalysis by the FK506 binding protein (FKBP). *Biochemistry* 32: 13830-13837.
244. Orozco M, Tirado-Rives J, Jorgensen WL (1993) Mechanism for the rotamase activity of FK506 binding protein from molecular dynamics simulations. *Biochemistry* 32: 12864-12874.
245. Albers MW, Walsh CT, Schreiber SL (1990) Substrate specificity for the human rotamase FKBP: a view of FK506 and rapamycin as leucine-(twisted amide)-proline mimics. *The Journal of Organic Chemistry* 55: 4984-4986.
246. Tradler T, Stoller G, Rücknagel KP, Schierhorn A, Rahfeld J-U, et al. (1997) Comparative mutational analysis of peptidyl prolyl cis/trans isomerases: active sites of *Escherichia coli* trigger factor and human FKBP12. *FEBS Letters* 407: 184-190.
247. Dugave C (2006) *cis-trans Isomerization in Biochemistry*: Wiley.
248. Alag R, Balakrishna AM, Rajan S, Qureshi IA, Shin J, et al. (2013) Structural Insights into Substrate Binding by PvFKBP35, a Peptidylprolyl cis-trans Isomerase from the Human Malarial Parasite *Plasmodium vivax*. *Eukaryotic Cell* 12: 627-634.
249. Mueller JW, Bayer P (2008) Small Family with Key Contacts: Par14 and Par17 Parvulin Proteins, Relatives of Pin1, Now Emerge in Biomedical Research. *Perspectives in Medicinal Chemistry* 2: 11-20.
250. Rahfeld J-U, Rücknagel KP, Schelbert B, Ludwig B, Hacker J, et al. (1994) Confirmation of the existence of a third family among peptidyl-prolyl cis/trans isomerases Amino acid sequence and recombinant production of parvulin. *FEBS Letters* 352: 180-184.
251. Ping Lu K, Hanes SD, Hunter T (1996) A human peptidyl-prolyl isomerase essential for regulation of mitosis. *Nature* 380: 544-547.
252. Metzner M, Stoller G, Rücknagel KP, Lu KP, Fischer G, et al. (2001) Functional Replacement of the Essential ESS1 in Yeast by the Plant Parvulin DPar13. *Journal of Biological Chemistry* 276: 13524-13529.
253. Schutkowski M, Bernhardt A, Zhou XZ, Shen M, Reimer U, et al. (1998) Role of Phosphorylation in Determining the Backbone Dynamics of the Serine/Threonine-Proline Motif and Pin1 Substrate Recognition†. *Biochemistry* 37: 5566-5575.
254. Pastorino L, Sun A, Lu P-J, Zhou XZ, Balastik M, et al. (2006) The prolyl isomerase Pin1 regulates amyloid precursor protein processing and amyloid- $\beta$  production. *Nature* 440: 528-534.
255. Vöhringer-Martinez E, Toro-Labbé A (2011) The mean reaction force: A method to study the influence of the environment on reaction mechanisms. *The Journal of Chemical Physics* 135: -.
256. Zhang Y, Daum S, Wildemann D, Zhou XZ, Verdecia MA, et al. (2007) Structural Basis for High-Affinity Peptide Inhibition of Human Pin1. *ACS Chemical Biology* 2: 320-328.
257. Homeyer N, Horn AH, Lanig H, Sticht H (2006) AMBER force-field parameters for phosphorylated amino acids in different protonation states: phosphoserine, phosphothreonine, phosphotyrosine, and phosphohistidine. *J Mol Model* 12: 281-289.
258. Jorgensen WL (1982) Revised TIPS for simulations of liquid water and aqueous solutions. *The Journal of Chemical Physics* 77: 4156-4163.
259. Mueller JW, Link NM, Matena A, Hoppstock L, Ruppel A, et al. (2011) Crystallographic proof for an extended hydrogen-bonding network in small prolyl isomerases. *J Am Chem Soc* 133: 20096-20099.
260. Doshi U, Hamelberg D (2009) Reoptimization of the AMBER force field parameters for peptide bond (Omega) torsions using accelerated molecular dynamics. *J Phys Chem B* 113: 16590-16595.
261. Feller SE, Zhang Y, Pastor RW, Brooks BR (1995) Constant pressure molecular dynamics simulation: The Langevin piston method. *The Journal of Chemical Physics* 103: 4613-4621.
262. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, et al. (1995) A smooth particle mesh Ewald method. *The Journal of Chemical Physics* 103: 8577-8593.

263. Wang J, Deng Y, Roux B (2006) Absolute Binding Free Energy Calculations Using Molecular Dynamics Simulations with Restraining Potentials. *Biophysical journal* 91: 2798-2814.
264. Bonomi M, Branduardi D, Bussi G, Camilloni C, Provasi D, et al. (2009) PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Computer Physics Communications* 180: 1961-1972.
265. Case D, Darden TA, Cheatham TE, Simmerling C, Wang J, et al. Amber 11.
266. Momany FA, McGuire RF, Burgess AW, Scheraga HA (1975) Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids. *The Journal of Physical Chemistry* 79: 2361-2381.
267. Kang YK (2004) Ring Flip of Proline Residue via the Transition State with an Envelope Conformation. *The Journal of Physical Chemistry B* 108: 5463-5465.
268. Ho BK, Coutsias EA, Seok C, Dill KA (2005) The flexibility in the proline ring couples to the protein backbone. *Protein Science* 14: 1011-1018.
269. Vitagliano L, Berisio R, Mastrangelo A, Mazzarella L, Zagari A (2001) Preferred proline puckerings in cis and trans peptide groups: Implications for collagen stability. *Protein Science* 10: 2627-2632.
270. Page MI, Jencks WP (1971) Entropic Contributions to Rate Accelerations in Enzymic and Intramolecular Reactions and the Chelate Effect. *Proceedings of the National Academy of Sciences* 68: 1678-1683.
271. Hammes GG (2008) How Do Enzymes Really Work? *Journal of Biological Chemistry* 283: 22337-22346.
272. Bruice TC (1976) Some Pertinent Aspects of Mechanism as Determined with Small Molecules. *Annual Review of Biochemistry* 45: 331-374.
273. Siegel JB, Zanghellini A, Lovick HM, Kiss G, Lambert AR, et al. (2010) Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. *Science* 329: 309-313.