



Computational Strategies for Improvement of Protein-Ligand Docking: Optimization Algorithm, Scoring Function, and Protein Flexibility

Uehara, Shota

(Degree)

博士 (計算科学)

(Date of Degree)

2017-03-25

(Date of Publication)

2019-03-25

(Resource Type)

doctoral thesis

(Report Number)

甲第6939号

(URL)

<https://hdl.handle.net/20.500.14094/D1006939>

※ 当コンテンツは神戸大学の学術成果です。無断複製・不正使用等を禁じます。著作権法で認められている範囲内で、適切にご利用ください。



PhD Thesis

Computational Strategies for Improvement of
Protein-Ligand Docking: Optimization Algorithm,
Scoring Function, and Protein Flexibility

タンパク質-リガンドドッキングの精度向上のための計算科学的アプローチ：
最適化アルゴリズム，スコア関数，及びタンパク質の柔軟性

Graduate School of System Informatics,
Department of Computational Science, Kobe University

Shota UEHARA

January 2017

Contents

Chapter 1	General Introduction	1
Chapter 2	Protein-Ligand Docking Using Fitness Learning–based Artificial Bee Colony with Proximity Stimuli.....	9
2.1	Introduction	9
2.2	Material and Methods	11
2.2.1	Protein-ligand Docking	11
2.2.2	Classical Artificial Bee Colony Algorithm.....	14
2.2.3	Fitness Learning-based Artificial Bee Colony with Proximity Stimuli (FIABCps) Algorithm.....	18
2.2.4	Simulation Set-up	23
2.2.4.1	Comparative Algorithms and Parameter Setting.....	23
2.2.4.2	Astex Diverse Dataset.....	24
2.2.4.3	Evaluation of Docking Accuracy.....	25
2.3	Results and Discussion.....	26
2.3.1	Docking Accuracy of FIABCps	26
2.3.2	Structural Analysis of the Binding Pocket of Neprilysin	27
2.4	Conclusions	32
Chapter 3	AutoDock-GIST: Incorporating Thermodynamics of Active-Site Water into Scoring Function for Accurate Protein-Ligand Docking	34

Contents

3.1 Introduction	34
3.2 Material and Methods	37
3.2.1 Grid Inhomogeneous Solvation Theory (GIST)	37
3.2.2 AutoDock4	40
3.2.3 Development of GIST-based Desolvation Function	43
3.2.4. Datasets and Preparation	46
3.2.4.1 Structure Preparation and MD Simulation for FXa	46
3.2.4.2 GIST Calculation and Docking Set-up	47
3.2.4.3 Dataset Preparation and Docking Metrics	48
3.3 Results and Discussion	50
3.3.1 Parameter Fitting for GIST-based Desolvation Function	50
3.3.2 Accuracy of Binding Affinity Prediction for FXa ligands	54
3.3.3 Docking Success Rate for FXa ligands	57
3.3.4 Virtual Screening Performance of AutoDock-GIST	60
3.4 Conclusions	62
Chapter 4 Cosolvent-based Molecular Dynamics for Ensemble Docking: Practical Method for Generating Flexible Protein Conformations	65
4.1 Introduction	65
4.2 Material and Methods	71
4.2.1 Target Protein and Structure Preparation	71
4.2.2 Choice of Cosolvents and System Set-up for MD	72
4.2.3 Molecular Dynamics Protocols	74
4.2.4 Selection of Conformational Ensemble from MD Trajectory	75

4.2.5 Dataset for Virtual Screening Experiment	77
4.2.6 Docking Protocols	78
4.2.7 Ensemble Docking and Scoring.....	78
4.2.8 Enrichment Measurements.....	79
4.3 Results and Discussion.....	80
4.3.1 Virtual Screening Performances	80
4.3.2 Binding Pocket Conformation and Probe Concentration	86
4.3.3 Protein Motion of Cosolvent-based Molecular Dynamics	88
4.3.4 Ensemble Selection from MD trajectory by RSPI	89
4.4 Conclusions	90
Chapter 5 General Conclusions	92
Appendix A	94
Appendix B	96
Appendix C	104
Acknowledgements	113
Bibliography.....	114

Chapter 1

General Introduction

Interactions between a macromolecule (protein) and a small molecule (ligand) play a central role in many biological processes such as enzymatic reactions, signal transduction, gene transcription, and physiological regulations. Since a lot of key biological functions are regulated by protein-ligand interactions, receptor proteins often become prime targets for pharmaceutical research. According to the advent of the human genome project and functional genomics, the number of new therapeutic targets has been continuously increased. In response to these newly discovered targets and great efforts of pharmaceutical industry, many innovative drugs have been successfully developed. However, drug discovery is a highly time-consuming and expensive process. The average cost for the recent discovery of a new drug is estimated approximately \$1.8 billion and still rising rapidly, and it takes over 12 years to launch a new drug in the market [1]. Thus, predictive techniques aimed at reducing the cost and accelerating the process of drug discovery are highly valued for further development. Computational approaches for drug design, called *in silico* drug design, are among the most mature and predictive, and yet areas for improvement remain. Coupled with a rapidly rising number of structures for target proteins and an increasing evolution of computational power, structure-based drug design (SBDD) has become prominent in the successful drug discovery [2]. There are currently more than 100,000 entries of X-ray or nuclear magnetic resonance (NMR)

structures of proteins or nucleic acids available in the Protein Data Bank (PDB) [3], including many biological targets for the drug discovery [4]. Since understanding the principles by which ligands recognize and interact with protein is of great importance in pharmaceutical research, the three-dimensional structure of a protein target is much informative and beneficial for the rational drug design [5]. Furthermore, using such structural information, the recent computational methods enable to predict which compound is truly bind to a protein target [6].

Among the various SBDD methods, the principal one is protein-ligand docking. Protein-ligand docking is a widely used computational tool in drug discovery efforts that tries to accurately predict the three-dimensional structure of a binding ligand to a target protein and to correctly estimate its strength of binding [7]. Computational docking also offers a relatively fast and economic alternative to standard experimental techniques (*in vitro* experiments). For example, in the early phase of the drug design, large compound libraries are screened via experimental methods such as high-throughput screening (HTS) for the discovery of “hit” compounds to a specific target. Instead of such an experimental approach, virtual screening

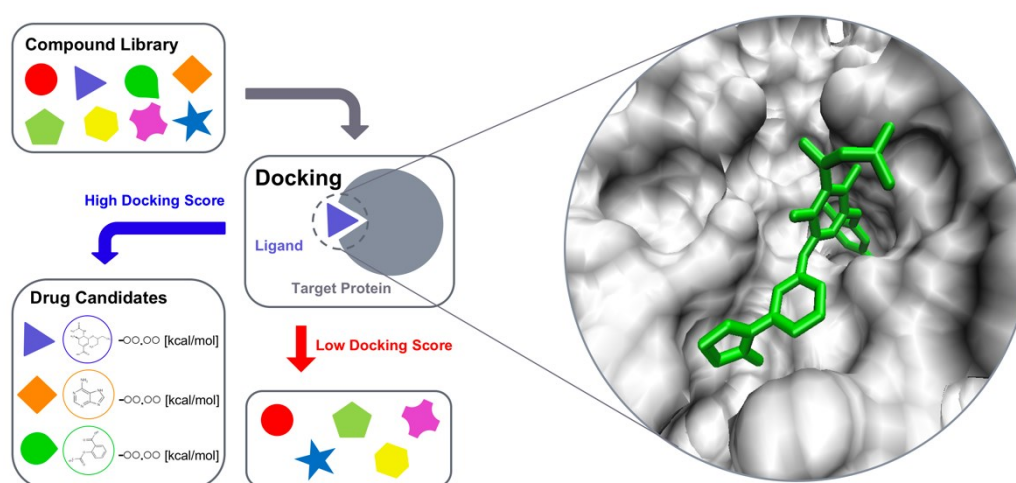


Figure 1.1 Diagram of structure-based virtual screening (SBVS) method by protein-ligand docking.

(VS) method is capable of selecting promising compounds from a huge chemical database by using the protein-ligand docking [8] (Figure 1.1). To date, the protein-ligand docking methods have been widely applied to many drug discovery efforts [9]. Despite many successes of protein-ligand docking, several aspects have remained important challenges, with significant margin for improvement [10,11]. In this thesis, three distinct strategies are presented for the further development and improvement of protein-ligand docking, focusing on optimization algorithm, scoring function, and protein flexibility.

In chapter 2, we attempt to improve docking accuracy by applying a novel optimization algorithm. Protein-ligand docking is an optimization problem which aims to identify the binding pose of a ligand with the lowest energy in the active site of a target protein (Figure 1.2). Hence, two essential components of the successful docking method are an accurate scoring function and an efficient optimization algorithm. Recent docking programs consider the

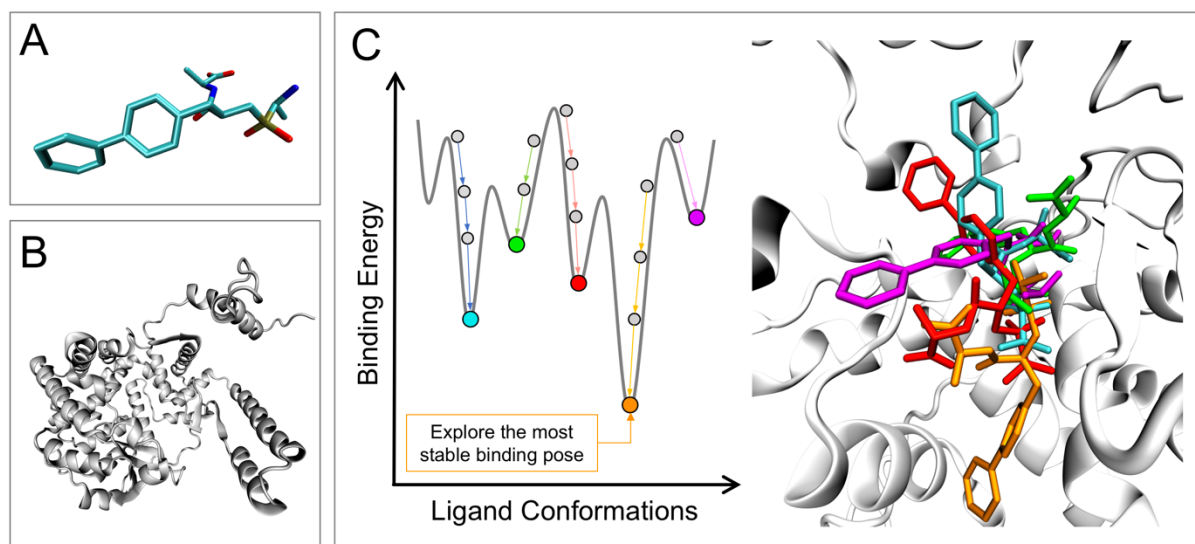


Figure 1.2 Diagram of protein-ligand docking as an optimization problem. (A) Input structure of a ligand. (B) Three-dimensional structure of a target protein. (C) Docking programs explore stable poses of the ligand bound to a specific site of the target protein using various optimization algorithms and rank their binding energies by scoring functions.

flexibility of a ligand and use a rigid structure of target protein so that the available conformations of the binding ligand are sufficiently explored with a rational computational cost [12]. However, since energy landscapes of the scoring functions are usually complicated and exhibit rugged funnel shape, it is still difficult to identify the correct binding pose of a ligand, in particular for the highly flexible ligand with many optimization parameters. Even though a large number of docking programs have been developed, it has been reported that major docking programs can identify the correct docking pose with an accuracy of only about 60% for the diverse protein-ligand complexes [13]. Traditionally, some variants of genetic algorithm (GA) have often been used to solve the docking problem. However, optimization algorithms based on GA do not have enough search ability in dealing with highly complicated or multi-modal problems like docking [14]. Hence, we focused on the novel optimization algorithms of swarm intelligence (SI). In last decades, wide varieties of SI-based optimization algorithms have been proposed, such as artificial bee colony algorithm (ABC), particle swarm optimization (PSO), and ant colony optimization (ACO), and it has been reported that they show superior performance to GA especially for the highly multi-dimensional optimization problems. In this study, we apply a variant of ABC to the protein-ligand docking, called fitness learning-based artificial bee colony with proximity stimuli (F/ABCps) [15]. F/ABCps is a powerful optimization algorithm based on the intelligent behavior of honey bee swarm, which has higher global search ability than other algorithms. The docking performance of F/ABCps was evaluated using 85 protein-ligand complexes and compared with four state-of-the-arts docking algorithms. Simulation results revealed that F/ABCps significantly improved the success rate of docking compared to four state-of-the-art algorithms. The present results also showed superior docking performance of F/ABCps, in particular for dealing with the highly flexible

ligands and the protein targets which have the wide and shallow binding pocket.

In chapter 3, we introduce the thermodynamics of active-site water into the scoring function of the docking program. In the docking calculation, the binding strength between protein and ligand is estimated by a scoring function which is typically described as the summation of various pairwise interatomic potentials, such as hydrogen bonding, electrostatic, van der Waals, etc. The interaction between protein and ligand is a principal source of the molecular binding, but it is not all factor of the molecular binding. For example, an indispensable participant is a water molecule. It has been widely recognized that water plays a significant role in the binding process between protein and ligand [16]. However, the thermodynamics of water molecules are often underestimated, or even ignored in protein-ligand docking. In the physiological environment, the active sites of protein are filled with water molecules, and thermodynamics of these water molecules are diverse and quite different from those of bulk water. When a small molecule binds to a protein, it causes the displacement of water molecules from the active site to the bulk region, and the thermodynamics of this displacement process significantly contributes to the free energy change of protein-ligand binding (Figure 1.3). In recent years, it has become possible to calculate the free energy of

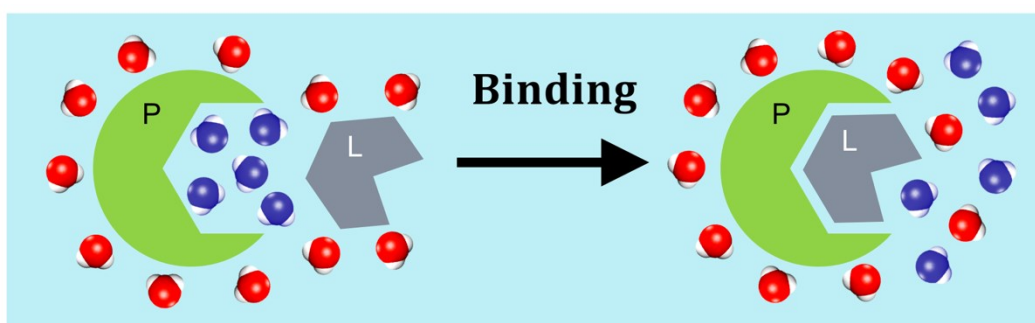


Figure 1.3 Displacement of active-site water molecules upon ligand binding. Each colored object represents the following: target protein (green), ligand (gray), displaced active-site water (blue), and bulk water (red).

active-site water molecules by some computational methods, such as grid inhomogeneous solvation theory (GIST) [17]. Here, we show a case study of the combination of GIST and a docking program and discuss the effectiveness of the displacing gain of unfavorable water in protein-ligand docking. We combined the GIST-based desolvation function with the scoring function of AutoDock4, which is called AutoDock-GIST. The proposed scoring function was assessed employing 51 ligands of coagulation factor Xa (FXa), and results showed that both scoring accuracy and docking success rate were improved, thus finding that the displacing gain of unfavorable water is effective for a successful docking campaign.

In chapter 4, we tackle the problem of protein flexibility in the protein-ligand docking. Protein flexibility is a major hurdle in current the protein-ligand docking methods that need to be more efficiently accounted for. Traditionally, the most docking methods only consider ligand flexibility and use a rigid structure of target protein for efficient calculations of numerous drug candidates. However, since proteins are intrinsically flexible and frequently undergo conformational changes on ligand binding, the static view of protein structure in classical docking is far from reality. In the last decades, the importance of protein flexibility upon ligand binding has been widely recognized [18]. The ideal approach is incorporating full protein flexibility to the docking. However, such a method requires the highly expensive cost of computation and thus impractical for large-scale docking studies like a VS experiment. Hence, a simplified model has been presented to incorporate limited protein motions while keeping computational time practical, called ensemble docking [19]. Ensemble docking makes use of multiple and discrete structures of a target protein. In standard ensemble docking procedure, each compound is sequentially docked to a set of protein conformers (i.e., ensemble) to find the best-fit protein structure for a particular ligand (Figure 1.4). Consequently, the flexibility of

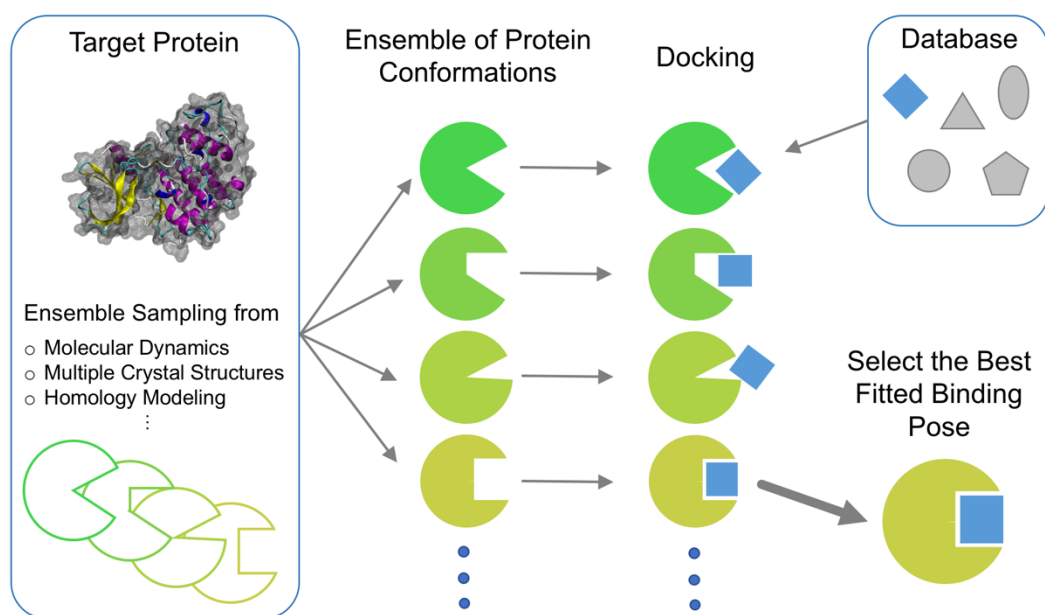


Figure 1.4 Diagram of ensemble docking procedure. By using the multiple conformations of a target protein, ensemble docking implicitly introduces flexibility of the target protein into protein-ligand docking.

target protein is implicitly introduced into the docking method. Although the ensemble docking is capable of accounting for any scale of protein motion, in practical, the coverage of protein flexibility completely depends on the quality of the structural ensemble. Thus, the critical issue of ensemble docking is how to select and/or generate a high-quality ensemble structure of the target protein. In this context, MD simulation is useful to produce distinct protein conformations without abundant experimental structures. In this study, we present a novel strategy that makes use of cosolvent-based molecular dynamics (CMD) simulation for the ensemble docking. CMD is a simple computational method which uses water and organic probe molecules for the solvent when performing the MD simulation of a protein target [20]. By mixing small organic molecules into a solvent, CMD can stimulate dynamic protein motions and induces partial conformational changes of binding pocket residues appropriate for the binding of diverse

ligands. In other words, CMD is capable of generating diverse conformations of a target protein and is expected to enhance the druggability of protein conformations [21]. The simulation results revealed that the present method is capable of generating diverse protein conformations and identifying many active ligands than the previous methods. Furthermore, the results also showed that present method is widely applicable for the diverse protein targets.

Chapter 2

Protein-Ligand Docking Using Fitness Learning-based Artificial Bee Colony with Proximity Stimuli

2.1 Introduction

Protein-ligand docking plays an essential role in structure-based drug design (SBDD), which aims to identify the binding structure of a ligand with the high affinity to a target protein using computer simulation. In lead identification, virtual screening based on docking simulation enables us to perform more efficient drug screening than experimental high-throughput screening (HTS) in terms of cost and efficiency [22,23]. Also in lead optimization, successful docking leads to a rational molecular design based on the three-dimensional structure of a target protein and binding ligands [24]. Incorporating SBDD, a number of drugs have been successfully developed [25-31].

Protein-ligand docking is regarded as an optimization problem, which identifies the binding pose of a ligand with the lowest energy (i.e., the most stable binding conformation) in an active site of a target protein. Thus, successful docking program requires two essential components, an accurate scoring function, and an efficient optimization algorithm. In past decades, many scoring functions have been developed for an accurate estimation of binding affinity [32-35]. However, since energy landscapes of the scoring functions are usually complicated and exhibit rugged funnel shape [36], successful docking calculation requires an

efficient optimization algorithm which correctly finds the lowest energy conformation of ligand. Inefficient optimization algorithms often give solutions trapped in some local optimum points of a scoring function, which results in an incorrect binding pose of a ligand and a wrong estimation of the binding affinity. In particular, highly flexible ligands with many rotational bonds are known to be more difficult for the docking simulation, due to their large number of optimization parameters [37].

In recent years, various optimization algorithms have been developed for the protein-ligand docking. Genetic algorithm (GA) based approaches are the most general, which are implemented, *e.g.*, in GOLD [38] and AutoDock [39]. On the other hand, swarm intelligence (SI) [40] based algorithms are highly attracted in the field of optimizations recently. SI based approaches simulate the collective behavior of simple agents or boids interacting locally with one another and with their environment. Such behaviors are often found in nature, especially biological systems so that SI based approaches are also called nature-inspired algorithms or metaheuristics. The famous nature-inspired algorithms include particle swarm optimization (PSO) [41], ant colony optimization (ACO) [42], firefly algorithm (FA) [43], cuckoo search (CS) [44] and artificial bee colony (ABC) optimization [45]. In previous studies, some variants of PSO have been developed and introduced into protein-ligand docking programs, such as SODOCK [37] and PSO@AutoDock [46]. It was reported that the PSO based approaches improve the docking accuracy better than GA. Both GA and PSO quickly find the global optimum point for a simple problem, because of their high convergence ability. However, these algorithms potentially have the risk of premature convergence to some local optimum point, in particular for the multi-modal, non-convex or highly multi-dimensional problems [14,47]. In this meaning, a more efficient optimization algorithm is strongly required for protein-ligand

docking.

In this study, we attempted to apply a novel nature-inspired optimization algorithm, called fitness learning-based artificial bee colony with proximity stimuli (F/ABCps) [15], to the protein-ligand docking. Artificial bee colony (ABC) algorithm is a simple and powerful optimization algorithm for the multi-dimensional and multi-modal functions, inspired from intelligent behaviors of the honey bee swarm. ABC has been widely applied to various optimization problems, such as neural network [48], spanning tree [49], digital filter [50], clustering [51], constrained optimization problem [52]. It has been also reported that the ABC based algorithms give better results for some optimization problems than the conventional algorithms [53,54]. F/ABCps is a variant of the ABC algorithm, extending its applicability to more complicated optimization problems like the protein-ligand docking.

The docking performance of F/ABCps was examined in comparison with four state-of-the-arts algorithms: ABC, SODOCK, PSO and LGA. Lamarckian genetic algorithm (LGA) [55] is a variant of GA, which is implemented in AutoDock as a default algorithm. The present results revealed that F/ABCps improved the success rate of the docking compared to the other algorithms, in particular for highly flexible ligands with many optimization parameters. In addition, we analyzed the relationship between the structure of the binding pocket and the energy landscape of the scoring function. This analysis clearly showed that F/ABCps is a suitable algorithm for dealing with proteins which have a wide and shallow binding pocket.

2.2 Material and Methods

2.2.1 Protein-ligand Docking

Protein-ligand docking searches the most stable conformation of binding ligand in the

active site of a target protein. The calculation of protein-ligand docking is considered as an optimization problem, which minimizes the interaction energy of protein and ligand in the consideration of ligand flexibility. The interaction energy is then estimated using scoring function described as the summation of pairwise atomic potentials between protein and ligand [56]. The present study used AutoDock4 scoring function [32] which is a force field based semiempirical scoring function. The AutoDock4 scoring function consists of five energy terms: hydrogen bonding, electrostatic, van der Waals, conformational entropy, and desolvation. The detailed description of the AutoDock4 scoring function is described in Chapter 3.

Originally, protein-ligand docking should deal full flexibility of target protein, since the conformational change of protein significantly affects ligand binding [57]. However, proteins are large molecules and have numerous degree of freedom compared with that of small ligand, and a much higher computational cost is needed to search their available conformations. Then, most of the recent docking programs only consider the flexibility of ligand for rapid calculation, which enables docking calculation time from seconds to minutes order per ligand. The flexibility and rigid motion of ligand are described with translation, orientation and rotatable bonds (Figure 2.1). The following represents for each optimization parameters.

- **Translation:** the three-dimensional coordinates of the mass center atom of ligand which are described with t_x, t_y, t_z . These three parameters are constrained within a user defined cubic region, which covers the binding pocket of a target protein.
- **Orientation:** rigid rotation of ligand with the quaternion, r_x, r_y, r_z, r_w . Here, three parameters of $r_x, r_y, r_z \in [0,1]$ are unit vectors determining the direction of ligand; the parameter of $r_w \in [-\pi, \pi]$ represents rotation around the unit vector r_x, r_y, r_z . The

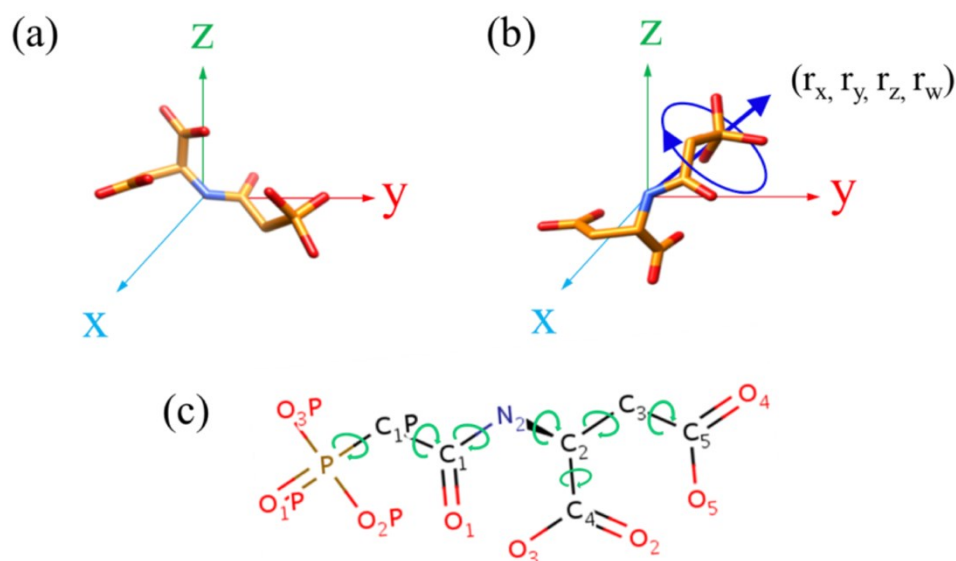


Figure 2.1 Optimization parameters in flexible docking: (a) translation, (b) orientation with quaternion, (c) rotatable bonds which represent flexibility of ligand.

orientational representation of quaternion is useful to avoid gimbal lock problem occurred in Euler angle [58].

- **Rotatable bonds:** conformational changes of ligand which is defined as any single non-ring bond, bounded to nonterminal heavy atom. Usually, amide C-N bonds are not considered because of their high rotational energy barrier [59]. Although a ligand has other degrees of freedoms, such as bond-stretching and angle-bending, these motions are sufficiently small compared with the bond rotations. Hence, protein-ligand docking programs realize efficient conformational sampling by allowing the degree of freedom of rotational bonds.

Accordingly, the total number of optimization parameter is $D = 7 + T$ in protein-ligand docking, where T is the number of rotatable bonds of ligand. The number of rotatable bonds

is quite different by the type of ligand. For instance, a small ligand sometimes has no rotatable bond ($T = 0$), whereas a large ligand has over 20 rotatable bonds. Since many rotational bonds correspond to many optimization parameters, a highly flexible ligand is a difficult target for the protein-ligand docking.

2.2.2 Classical Artificial Bee Colony Algorithm

The artificial bee colony (ABC) is a swarm based meta-heuristic algorithm proposed by Karaboga et al. [14] for numerical optimization problems. It was inspired by the intelligent foraging behavior of honey bees. ABC is composed of three kinds of honey bees: employed bees, onlooker bees and scout bees (Figure 2.2). First, an employed bee is assigned to a particular food source. She carries nectar to the hive and shares information on the nectar amount of the food source with onlooker bees waiting on the hive. Second, an onlooker bee

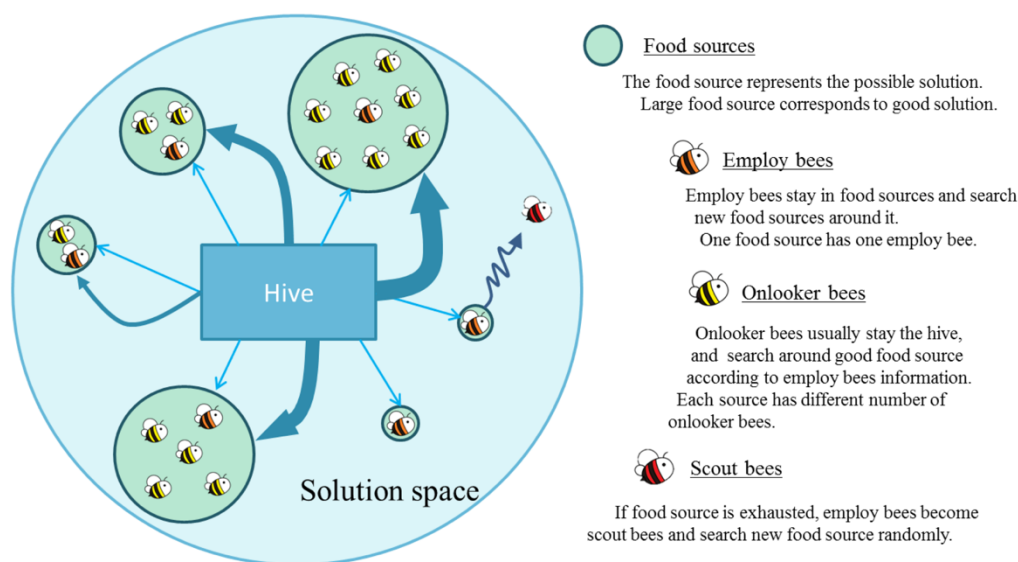


Figure 2.2 Diagram of ABC algorithm. The position of food source represents a possible solution to optimization problem, and the nectar amount of a food source corresponds to the quality of the associated solution. Employ bees, onlooker bees and scout bees are the foragers which search optimal solutions.

chooses a rich food source, based on the nectar information. If one food source has much nectar amounts, a large number of onlooker bees are assigned to the source. Finally, a scout bee carries out a random search for discovering new food sources.

In ABC, a colony of artificial honey bees (agents) searches for rich food sources (good solutions to a given problem). The position of a food source represents a solution vector of the optimization problem, and a quality of the food source (nectar amount) is represented by a fitness value calculated with the scoring function. The number of food sources SN is equal to the number of employed bees or onlooker bees. The three kinds of bees search for a global optimum point in D -dimensional real parameter space, where D corresponds to the number of optimization parameters (*e.g.*, translation, orientation and conformation of the ligand for the flexible protein-ligand docking). A D -dimensional solution vector on a food source is described as

$$\boldsymbol{\theta}_i^C = [\theta_{i,1}^C, \theta_{i,2}^C, \theta_{i,3}^C, \dots, \theta_{i,D}^C], \quad (2.1)$$

where $i = 1, 2, \dots, SN$ is an index of food sources and $C = 1, 2, \dots, MCN$ (maximum count number) is a current cycle number. In the beginning of optimization ($C = 0$), each parameter of food sources is initialized with uniformly distributed random numbers restricted to certain ranges. A fitness value for a food source is then calculated as

$$fitness_i = \begin{cases} 1/(1 + f_i) & \text{if } f_i \geq 0 \\ 1 + \text{abs}(f_i) & \text{if } f_i < 0 \end{cases}, \quad (2.2)$$

where f_i is an actual value of scoring function F to be optimized ($f_i = F(\boldsymbol{\theta}_i^C)$). Since we consider a minimization condition here, a food sources with a lower score of the scoring function have a higher fitness value (Figure 2.3). After the initialization, ABC performs the optimization

process through cycles of three exploration steps by employed bees, onlooker bees and scout bees until the termination criteria are satisfied.

In the employed bee phase, the employed bees seek a new food source around the assigned food sources, where a new food source is explored in the direction to another food source by perturbing a single optimization parameter as

$$v_{i,j}^C = \theta_{i,j}^C + \phi(\theta_{k,j}^C - \theta_{i,j}^C), \quad (2.3)$$

Here, $k \in \{1,2,\dots,SN\}$ is an index of randomly selected food source except for i . Similarly, $j \in \{1,2,\dots,D\}$ is an index randomly selected from the D -dimensional parameters. ϕ is a random number in the range of $[-1,1]$. If a new food source v_i^C has a higher fitness value than the current food source θ_i^C , an employed bee updates θ_i^C to v_i^C . After all the employed bees finish exploiting, they go back to the hive and share the information on the food sources (nectar amounts) with the onlooker bees waiting on the hive.

In the onlooker bee phase, the onlooker bees perform a probabilistic selection of food sources for exploiting. A probability of a food source to be selected is calculated with the fitness

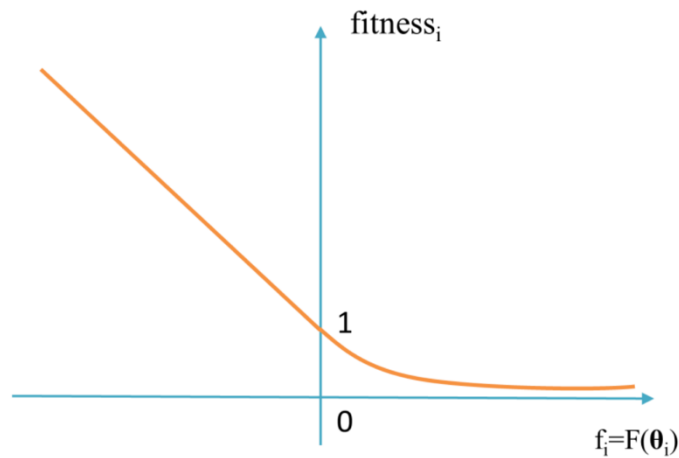


Figure 2.3 Plot of fitness value $fitness_i$ according to the scoring function $f_i = F(\theta_i^C)$ in a minimization condition.

values, given by

$$p_i = \frac{fitness_i}{\sum_{l=1}^{SN} fitness_l} \cdot \quad (2.4)$$

Based on this probability, the onlooker bees perform the roulette wheel selection for the decision of the food source, so that a higher fitness food source is intensively explored by a large number of the onlooker bees. The onlooker bee searches for a new food source around the selected food source using Equation (2.3), and updates the current food source with the greedy selection in the same way as the employed bee.

In the scout bee phase, a food source which cannot be improved anymore is replaced by a new food source created with random numbers. To find these exhausted food sources, a trial counter t_i is used at each i th food source. If the employed or onlooker bee is unable to

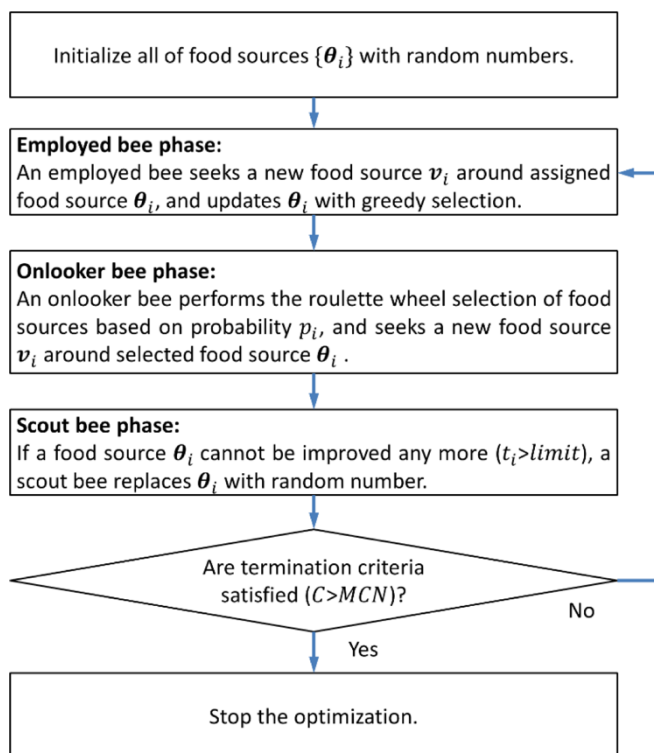


Figure 2.4 Flowchart of the artificial bee colony (ABC) algorithm.

improve the previous fitness value of the i th food source, t_i is increased by unity. The trial counter t_i is reset to zero when the i th food source is successfully improved. When t_i reaches the maximum trial number, *limit*, the i th food source is replaced with random numbers and t_i is reset to zero. In this way, the scout bees play an important role in keeping the diversity of population. The algorithm of ABC is summarized as flowchart in Figure 2.4.

2.2.3 Fitness Learning-based Artificial Bee Colony with Proximity Stimuli (FIABCps) Algorithm

Fitness learning-based ABC with proximity stimuli (FIABCps) is a variant of ABC proposed by Das et al. [15]. They introduced three vital modifications to the classical ABC, to achieve the superior performance for real-world optimization problems.

First, an improved positional modification scheme is introduced. This scheme is developed on the basis of the fitness learning mechanism and the directive component towards adjacent food sites. In the classical ABC, the positional modification given by Equation (2.3) is performed with a randomly selected food source θ_k^C . Alternatively, elite food sources and neighbor food sources are used in FIABCps for the positional modification, which gives a superior balance of bee's exploration between global search and local search.

Second, a multi-dimensional perturbation scheme is introduced to the positional modification. As mentioned above, the single parameter perturbation is used in the classical ABC, which sometimes leads to the slow convergence for highly multi-dimensional problems. [60] On the other hands, all optimization parameters are updated in PSO and GA, which result in the premature convergence for the complicated problems (*i.e.*, trapped solutions in some local optimum points of scoring function). In FIABCps, a subset of the D -dimensional

parameters is randomly selected for the positional modification, based on the Rechenberg's 1/5th mutation rule [61]. It helps in an efficient convergence of solutions, properly avoiding the premature convergence.

Third, proximity-based stimuli are employed for the food site selection by the onlooker bees. In the classical ABC, the onlooker bees perform the roulette wheel selection of food sources using the probability of Equation (2.4), which contributes to an intensive exploitation around a high fitness food sources. However, this selection scheme sometimes causes the overcrowding of the onlooker bees at the best-so-far food source, which results in the premature convergence. To circumvent this problem, F/ABCps introduces a weighted probability based on the proximity-based stimuli. Since the weighted probability reflects the locality of the food sources, neighbor food sources around the high fitness food sources get more chances to be selected by the onlooker bees.

F/ABCps proceeds in the same way as the classical ABC through the employed bee phase, the onlooker bee phase and the scout bee phase. The position of a food source represents a solution vector to the optimization problem, and the quality of a food source (nectar amount) corresponds to the fitness value calculated with the scoring function. The number of food sources SN is equal to the number of the employed bees or the onlooker bees. The three kinds of bees search for a global optimum point in D -dimensional real parameter space, where D corresponds to the number of optimization parameters. Each D -dimensional solution vector at the i th food sources, θ_i^C , is described as Equation (2.1). In the beginning of optimization ($C=0$), each parameter of food sources is initialized with uniformly distributed random numbers which are restricted to certain ranges. After the initialization, the following procedures are repeated in each cycle until the termination criteria are satisfied.

First, the fitness value of a food sources, $fitness_i$, is calculated as Equation (2.2). Since we consider a minimization condition, a food source with the lower score of function has a higher fitness value. Fitness learning mechanism is described with mixing with the elite components

$$\theta_{i,j}^{FC} = \begin{cases} \theta_{r1,j}^C & \text{if } fitness_{r1} \geq fitness_{r2} \\ \theta_{r2,j}^C & \text{if } fitness_{r1} < fitness_{r2} \end{cases}, \quad (2.5)$$

where $j=1,2,\dots,D$ is D -dimensional parameter index of a food source. Indices of $r1$ and $r2$ represent two different elite food sources randomly selected from the top q % of the population. The value of q is varied from the top 20 % (0.2) members initially, to the 10 % (0.1) at the end of the cycle. This variation of q that occurs nonlinearly is given by

$$q = 0.2 - 0.1 \left(\frac{e^{m \cdot C/MCN} - 1}{e^m - 1} \right), \quad (2.6)$$

with an uniform random number m , lying in the range $[0,1]$. In addition, F/ABCps uses a selective parameter scheme for multi-dimensional perturbation, based on the Rechenberg's 1/5th mutation rule [45] The perturbation parameters for the positional modification are selected by

$$J^* = \{j_1, j_2, j_3, \dots, j_n\}; \quad j_v \in \{1, 2, \dots, D\}; \quad 1 \leq n \leq \left\lceil \frac{1}{5} D \right\rceil \quad (2.7)$$

Here, n is a random integer corresponding to the number of components of J^* . It is noted that J^* is a subset of D -dimensional parameters which are composed of randomly selected n indices.

In the employed bee phase, the employed bee seeks a new food source v_i^C around the assigned food source θ_i^C by using the perturbation parameters J^* . The positional modification scheme in F/ABCps is performed with a combination of the directive component towards

adjacent food sites and the fitness learning mechanism of θ_i^{FC} (Equation (2.5))

$$v_{i,J^*}^C = \theta_{i,J^*}^C + \phi_G(\theta_{k_i,J^*}^{NC} - \theta_{i,J^*}^C) + \phi_C(\theta_{i,J^*}^{FC} - \theta_{i,J^*}^C), \quad (2.8)$$

where $\theta_{k_i}^{NC}$ represents one of the k th-nearest food sources from θ_i^C according to the Euclidean distance. The parameter k is a random integer, lying in the range $[0, \sqrt{SN/2}]$. Two control parameters, ϕ_G and ϕ_C , are different random numbers, generated as

$$\begin{aligned} \phi_G &= N(\mu, \sigma^2); & \mu &= 0, \sigma^2 = 1, \\ \phi_C &= Q(r; x_0, \gamma); & x_0 &= 0, \gamma = 0.5, r \in (0,1), \end{aligned} \quad (2.9)$$

where $N(\mu, \sigma^2)$ denotes the Gaussian distributed number with mean μ and variance σ^2 ; $Q(r; x_0, \gamma)$ denotes the quantile function of Cauchy distribution with location x_0 , scale γ and restrict range r . The Gaussian distribution has a short tail property, and is suitable for the fine local search. On the other hand, the Cauchy distribution has a far wider tail than the Gaussian distribution, and is useful when the global optimum is far away from the current search point. If a new food source v_i^C has the higher fitness value than the current food source θ_i^C , the employed bee updates θ_i^C to v_i^C .

In the onlooker bee phase, the onlooker bee performs a probabilistic selection of the food source for exploitation. In the classical ABC, a probability of a food source to be selected, p_i , is calculated with the fitness values, given by Equation (2.4). The selection scheme using Equation (2.4) sometimes causes the overcrowding of the onlooker bees at the best-so-far food source, which results in the premature convergence. To circumvent this problem, F/ABCps introduces a weighted probability based on the proximity-based stimuli

$$p_i^w = \frac{1}{2m_i} \sum_{l=1}^{m_i} (p(N_l^i) + p(F_l)), \quad (2.10)$$

where $p(\cdot)$ represents the probability Equation (2.4) of a selected food source taken as an argument ($p(i) = p_i$). N_l^i is an index representing the l th-nearest food sources calculated with the Euclidean distance from the l th food source. Similarly, F_l is an index which refers to the l th-best food source calculated with the fitness value. The parameter m_i is a random integer, lying in the range $[0, SN/\sqrt{D}]$. If the weighted probability p_i^w is larger than p_i , the i th food source is selected by an onlooker bee for exploitation. The onlooker bee searches for a new food source v_i^c around the selected food source θ_i^c using Equation (2.10), and updates θ_i^c to v_i^c with the greedy selection in the same way as the employed bee. This selection is repeated until all the onlooker bees are assigned to any of the food sources.

In the scout bee phase, the food source that cannot be improved anymore is replaced randomly by a scout bee, which is the same procedure as the classical ABC. To find these exhausted food sources, a trial counter t_i is used at each i th food source. If an employed or onlooker bee is unable to improve the previous fitness value of the i th food source, t_i is increased by unity. The trial counter t_i is reset to zero when the i th food source is successfully improved. When t_i reaches the maximum trial number *limit*, the i th food source is replaced with random numbers and t_i is reset to zero.

The performance of F/ABCps was examined for two real-world optimization problems including numerous local peaks, non-linearity, interdependence and bound constraints [15]. As a result, F/ABCps provided the best solutions among nine state-of-the-arts optimization algorithms. Pseudocode of F/ABCps is described in Appendix A (Table A1).

2.2.4 Simulation Set-up

2.2.4.1 Comparative Algorithms and Parameter Setting

The docking performance of *F/ABCps* was evaluated by comparison with four state-of-the-art algorithms: ABC, SODOCK, PSO and LGA. Lamarckian genetic algorithm (LGA) is a variant of GA and default algorithms of AutoDock4, which enhanced global search efficiency by combining local search algorithms with classical GA [55]. The same strategy is applied in SODOCK. SODOCK is a hybrid algorithm of PSO and local search algorithm [37]. Other two algorithms represent classical ABC and PSO. LGA and PSO are available in the docking program of AutoDock4 [62]. AutoDock4 is one of most famous docking program, which is open source and freely available for academic users. We introduced ABC, SODOCK and *F/ABCps* into AutoDock4 for comparison of docking accuracy. We assessed five algorithms under the identical conditions: (I) Examinations were performed in the framework of AutoDock4; (II) A flexibility of a ligand was described with translation, orientation, and conformation, and a protein was treated as a rigid object; (III) 85 complexes in Astex diverse set [63] was used for the evaluation of docking performances; (IV) A binding pocket was set with a cubic box ($22.5 \times 22.5 \times 22.5 \text{ \AA}^3$) centered at the crystal ligand; (V) AutoDock4 scoring function [32] was used; (VI) The maximum number of energy evaluations was set to 2,500,000. The parameters for *F/ABCps* were determined empirically, so that the population number *SN* and the maximum trial number *limit* were set to 500 and 200, respectively. The parameters for other algorithms were basically default values. Setting parameters for the five algorithms are summarized in Appendix A (Table A2).

2.2.4.2 Astex Diverse Dataset

In this work, we used astex diverse set for the data set of protein-ligand docking. Astex diverse set was developed for the evaluation of docking programs, which consists of high-quality X-ray crystallographic structures of 85 protein-ligand complex [63]. All of the ligands in astex diverse set have drug-like structures; 23 ligands are approved drug, and 6 ligands are under clinical trial. These 85 ligands cover wide chemical structure, and 85 target proteins also cover diverse protein families. Note that hydrogens were properly added for all proteins and ligands, and astex diverse set is available in Chembridge Crystallographic Data Center (<http://www.ccdc.cam.ac.uk>).

Since our study focused on optimization, the number of rotatable bonds of ligand is an important index for the evaluation of docking accuracy. We divided astex diverse set into three

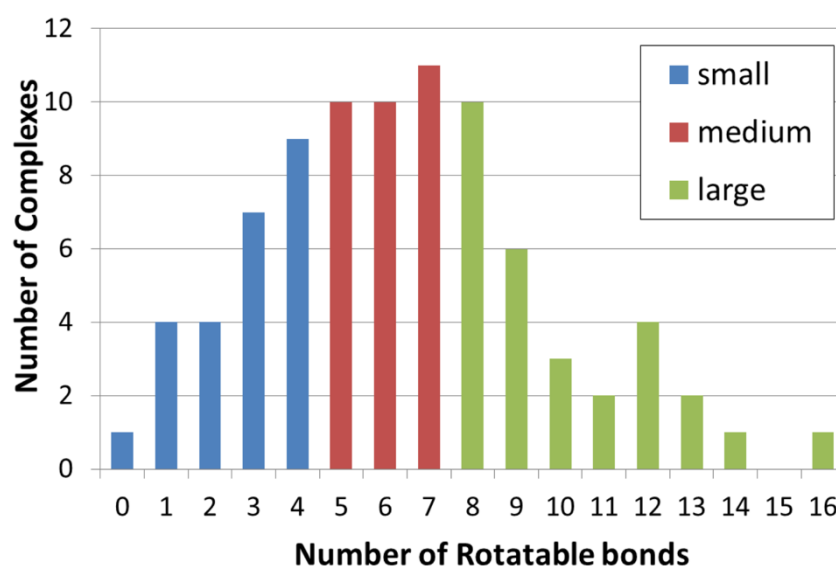


Figure 2.5 The histogram of complexes for astex diverse set in term of the number of rotatable bonds of ligand. These ligands are clustered in three groups: blue colored bars represent “small” group which is composed 25 complexes, red colored bars represent “medium” group which is composed 31 complexes, green colored bars represent “large” group which is composed 29 complexes.

groups, small, medium and large, on the basis of the number of the rotatable bonds of ligand (Figure 2.5); Small group consists of 25 ligands with the number of rotatable bonds 0~4; Medium group consists of 31 ligands with the number of rotatable bonds 5~7; Large group consists of 31 ligands with the number of rotatable bonds over 8.

2.2.4.3 Evaluation of Docking Accuracy

We evaluated docking performance based on the success rate of binding pose prediction. In other word, it is called re-docking experiments, which performs docking calculation to known protein-ligand complex, and evaluates how well to reproduce crystalized pose of bound ligand (Figure 2.6). The structural similarity is measured by root mean square deviation (RMSD) described as:

$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^n r_i^2}{n}} \quad (2.11)$$

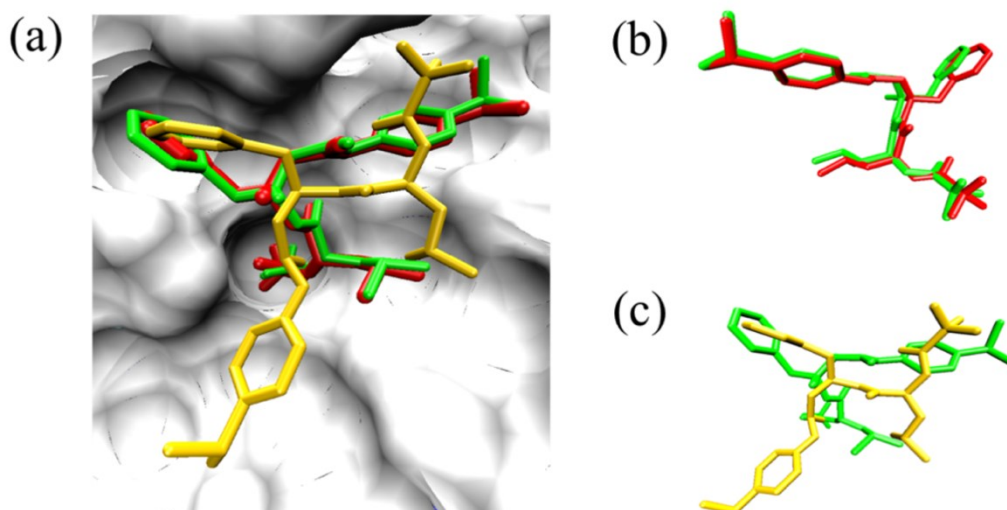


Figure 2.6 Re-docking experiment for evaluation of docking program: (a) Results of re-docking calculation. (b) Successful docking with small RMSD. (c) Unsuccessful docking with large RMSD. Green colored pose represents crystallized structure, and red and yellow colored poses correspond to two different binding poses which are obtained by docking.

where, n is the number of ligand atom; r_i is the distance between corresponding atom i . In this work, we set RMSD criteria 2 Å. In other word, if RMSD between docking pose and native pose of ligand is less than 2 Å, the binding pose prediction is success.

2.3 Results and Discussion

2.3.1 Docking Accuracy of FIABCps

Table 2.1 shows the results of the docking calculations obtained with FIABCps, ABC, SODOCK, PSO and LGA for 85 complexes of Astex diverse set. The docking performances were examined in terms of the success rate of the pose prediction and the searching ability of the lowest energy. In addition, 85 complexes in Astex diverse set were divided into three groups according to the number of rotational bonds of ligands, which were used for examining the dependence of the docking accuracy on the number of optimization parameters. The success rate of the docking was evaluated with root mean square deviation (RMSD) of the predicted ligand pose from the crystal structure. The simulation results showed that FIABCps provided the best performance of all the five algorithms with the success rate of 87.1 %. In general, the docking for highly flexible ligands is more difficult than that for less flexible ligands, due to their large number of optimization parameters [37]. Even for such highly flexible ligands ($N_r = 8\sim 16$), FIABCps can successfully find the correct binding poses with 89.7 %, whereas the other methods lowered their success rates. This result indicated that FIABCps might be extended to more complicated systems, such as the partially flexible protein docking including side-chain flexibility of proteins [64-66] or the docking under explicit water molecules [67-69].

Table 2.1 Docking results by comparison of five algorithms for 85 complexes of Astex diverse set.

N _r ^a	N _c ^b	Success rate [%] ^c					No. of wins ^d				
		F/ABCps	ABC	SODOCK	PSO	LGA	F/ABCps	ABC	SODOCK	PSO	LGA
0~4	25	84.0	84.0	84.0	80.0	84.0	19	4	1	0	1
5~7	31	87.1	80.6	83.9	64.5	77.4	23	3	1	2	2
8~16	29	89.7	79.3	79.3	44.8	55.2	17	4	5	3	0
Total	85	87.1	81.2	82.4	62.4	71.8	59	11	7	5	3

^a N_r represents the number of rotational bonds for ligands. ^b N_c represents the number of complexes. ^c Rate of successful docking that RMSD from the crystal structure is less than 2 Å. ^dThe number of wins in finding the lowest energy in the scoring function among the five algorithms.

The present results also showed that F/ABCps gave the best results (*i.e.*, the lowest energy) for the 59 complexes. Assuming that the scoring function can describe the correct binding energy, the lowest energy in the scoring function corresponds to the actual binding affinity between a ligand and a protein. Thus, F/ABCps is found to give more accurate estimations of the binding affinity, compared with the other four algorithms. The classical ABC gave the success rate of 81.2 %, which was better than PSO and LGA. Thus, the basic strategy of ABC is superior to that of GA and PSO for the protein-ligand docking. From these results, F/ABCps is found to be a more suitable algorithm for solving the protein-ligand docking than the conventional algorithms.

2.3.2 Structural Analysis of the Binding Pocket of Neprilysin

Next, we analyzed the performance of F/ABCps with respect to the binding pocket structure and the energy landscape of the scoring function. The performance of F/ABCps was compared with LGA which is a major algorithm implemented in AutoDock. For this analysis, we used the crystal structure of neprilysin (PDB-ID: 1R1H) [70] and its potent inhibitor N -[3

-[(1-aminoethyl) (hydroxyl) phosphoryl] -2 -(1, 1' -biphenyl -4 -ylmethyl) propanoyl] alanine (BIR), because LGA could hardly find the correct binding pose of this ligand. We performed 1000 times of docking calculations with LGA and sampled 1000 different docking poses of the ligand. As a result, two specific clusters named cluster-1 and cluster-2 were found on the basis of the structural similarity of their binding poses (Figure 2.7A). The population of cluster-1 and cluster-2 totally accounts for 58 % of all the sampled poses. The main difference between the two clusters was in the direction of two aromatic rings of the ligand (Figure 2.7B).

The 1000 sampled poses were also calculated with F/ABCps, which resulted in the same two clusters as the LGA ones. Figure 2.8 shows distributions of cluster-1 and cluster-2 with respect to the RMSD from the crystal structure of the ligand. The distributions obtained with

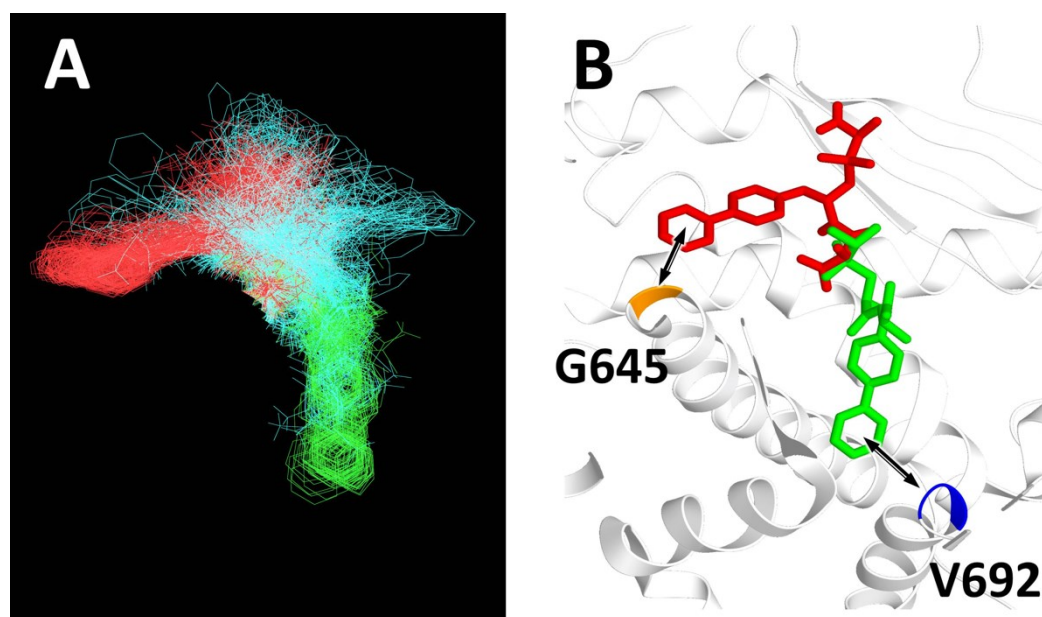


Figure 2.7 (A) Superposition of 1000 sampled poses of BIR. Poses of cluster-1, cluster-2, and the others are shown in green, red and cyan, respectively. (B) Definitions of cluster-1 and cluster-2. White ribbon represents the backbone of neprilysin. Green colored pose is a representative structure of cluster-1 where the distance between the center of the aromatic ring of the ligand and VAL692 (blue) is less than 6.5 Å. Red colored pose is a representative structure of cluster-2 where the distance between the center of the aromatic ring of the ligand and GLY645 (orange) is less than 6.5 Å.

F/ABCps are completely different from those with LGA. In LGA, we found 10 % population for cluster-1 and 48 % population for cluster-2, whereas 40 % population for cluster-1 and 9 % population for cluster-2 were observed in F/ABCps. In common, the docking pose with the RMSD less than 2 Å is regarded as the successful reproduction of the crystal structure of the ligand. Therefore, the poses of cluster-1 correspond to the crystal structure (see Figure 2.9). The lowest binding energies for cluster-1 and cluster-2 were -15.53 kcal/mol and -11.97 kcal/mol, respectively. These results showed that F/ABCps successfully found the correct binding poses at the global minimum (poses of cluster-1). In contrast, LGA gave the binding poses trapped in the local minimum of the scoring function (poses of cluster-2).

Regarding the molecular structures, the neprilysin has two specific docking regions in its binding pocket: the wide and shallow region on which the poses of cluster-2 are located, and the narrow and deep region on which the poses of cluster-1 and the crystal ligand are located

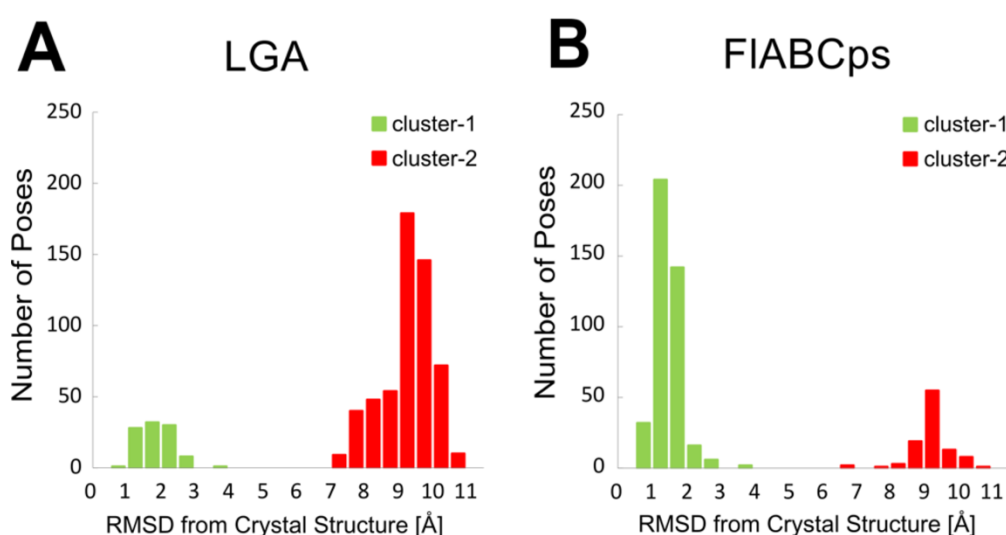


Figure 2.8 Distribution of cluster-1 and cluster-2 with respect to the RMSD from the crystal structure of BIR: **(A)** 1000 docking poses with LGA; **(B)** 1000 docking poses with F/ABCps. Green and red colored bars refer to cluster-1 and cluster-2, respectively.

(Figure 2.9). In other words, the narrow and deep region corresponds to the global minimum, and the wide and shallow region corresponds to one of the local minima of the scoring function.

Next, we analyzed the energy landscape of the scoring function around the two clusters. Here, we used the RMSD from the crystal structure for simplicity. Figure 2.10 shows the energy distributions of cluster-1 and cluster-2 with respect to the RMSD from the crystal structure. The energy distributions plotted on the RMSD space can approximate the multi-dimensional energy landscape of the scoring function. In addition, the RMSD standard deviations of two clusters can be regarded as the widths of the energy wells in the multi-dimensional spaces. Supposing that these distributions refer to the normal distribution, the energy landscape can be approximated by a Gaussian function. If the centers of these energy wells are set to the

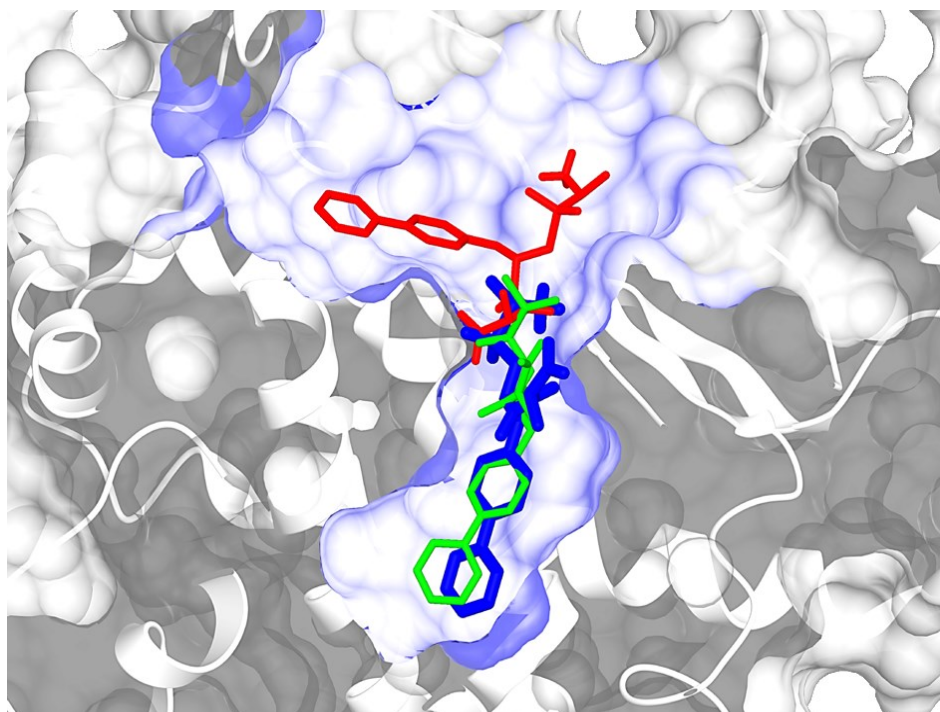


Figure 2.9 Molecular structures of the binding pocket of neprilysin and BIR. Blue colored pose is the crystal structure, green colored pose is a representative structure of cluster-1, and red colored pose is a representative structure of cluster-2.

individual lowest energy structures, these bell curves reflect the shapes of the multi-dimensional energy landscape around cluster-1 and cluster-2. The standard deviations of cluster-1 and cluster-2 from the individual lowest energy structures were 1.8 Å and 2.9 Å, respectively. Therefore, the poses of cluster-1 were located on the narrow and steep energy well of the global minimum, whereas the poses of cluster-2 were trapped in the wide and gradual energy well of the local minimum. These results can be interpreted as follows. The neprilysin has the wide and shallow region in its binding pocket on which the poses of cluster-2 are located. Around this region, the scoring function gives the wide and gradual energy well of the local minimum. The conventional algorithms, including GA and PSO, usually show the high convergence ability for simple problems. However, they often give solutions trapped in some local minima, when

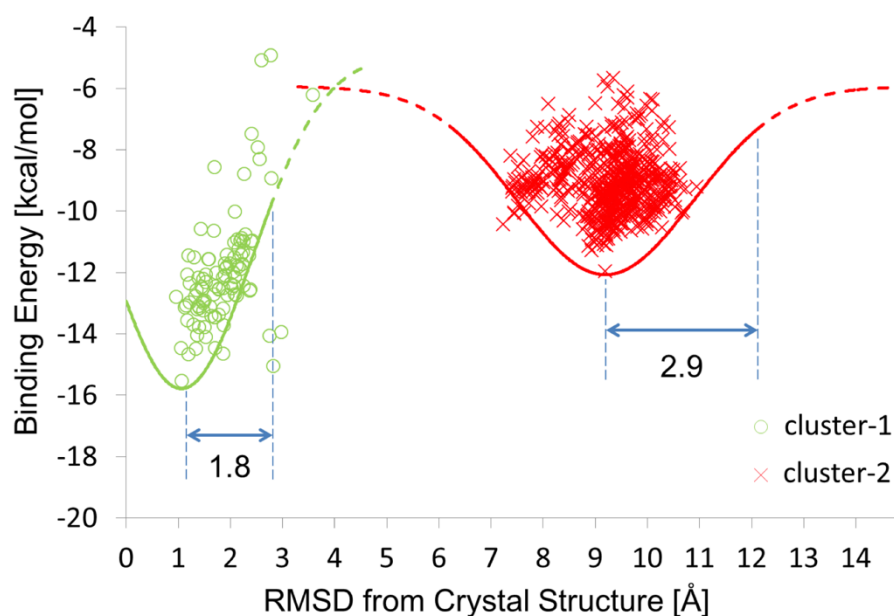


Figure 2.10 Scatter plots of the binding energies of cluster-1 and cluster-2 with respect to the RMSD from the crystal structure. Poses of cluster-1 are shown by green circles and those for cluster-2 are shown by red crosses. Solid lines in the individual clusters show Gaussian distributions of the RMSD from the pose with the lowest energy; standard deviation of cluster-1 is 1.8 Å and that of cluster-2 is 2.9 Å.

dealing with multi-modal and multi-dimensional problems [35,36]. Also in our simulation results, most of the LGA calculations gave the binding poses trapped in the local minimum (cluster-2). In contrast, *F/ABCps* successfully found the correct binding poses existing in the global minimum (cluster-1), properly avoiding such a local minimum. Some kinds of proteins which have a wide and shallow binding pocket were supposed to provide a challenging task for *in silico* docking. This is because such kinds of proteins usually contain a large number of local minima on their energy landscape of the scoring function. *F/ABCps* would be a suitable algorithm for such proteins with these features as kinases.

The scoring (objective) functions for protein-ligand docking are generally constructed by summation of interatomic potentials between all pairs of protein and ligand atoms [56]. Eventually, these functions with numerous terms describe non-convex and multi-modal solution space, even if the pairwise interatomic potentials are simple convex functions. These kinds of objective functions are often used for optimization problems of molecular sciences in which any interatomic potentials are calculated. Nature-inspired metaheuristic optimization algorithms are then developed to solve such kinds of problems with non-convex or multi-modal functions that are not amenable to the approach via differentiations as in the steepest descent method. *F/ABCps* is one of the most robust optimization algorithms for the problems containing a number of local minima and/or highly multi-dimensional solution space.

2.4 Conclusions

In this work, we introduced a novel optimization algorithm *F/ABCps* for the protein-ligand docking. The performance of *F/ABCps* was assessed in comparison with the four state-of-the-art docking algorithms. Simulation results revealed that *F/ABCps* gave significantly accurate

docking poses of the ligands, compared with the other four algorithms. The results also showed that *F/ABCps* provided the best performance for the highly flexible ligands with many optimization parameters. In addition, we analyzed the simulation regarding the energy landscape of the scoring function and the shape of the binding pocket of the receptor protein. Some kinds of proteins were supposed to be a challenging task for the docking because they usually possess a large number of wide and gradual energy wells corresponding to the local minima in the scoring function. For these proteins, the conventional optimization algorithms can hardly find the correct binding pose of ligand. In contrast, *F/ABCps* successfully find the correct binding poses, properly avoiding such local minima. Consequently, *F/ABCps* would become a useful algorithm for more complicated optimization problems concerning *in silico* drug discovery.

Chapter 3

AutoDock-GIST: Incorporating Thermodynamics of Active-Site Water into Scoring Function for Accurate Protein-Ligand Docking

3.1 Introduction

Water is an indispensable participant in the binding process of a protein and a small molecule [71–77]. In an *in vivo* environment, the active sites of a proteins are filled with water molecules, and thermodynamics of these water molecules are diverse and quite different from those of bulk water [78-80]. When a small molecule binds to a protein, it causes the displacement of water molecules from the active site to the bulk region, and the thermodynamics of this displacement process is a principal source of binding free energy of ligands [81-83]. For instance, a water molecule enclosed by hydrophobic residues of protein that cannot make appropriate hydrogen bonds is enthalpically unfavorable, and the displacement of such water earns an enthalpic contribution in binding free energy. On the other hand, a water molecule forming tight hydrogen bonds to hydrophilic residues of a protein is enthalpically favorable, and the displacement of such water may incur the penalty of protein-ligand binding. Thus, the role of active-site water molecules is widely appreciated in the study of molecular recognitions [84-88].

Computational approaches for analyzing active-site water properties have become

essential to our better understanding of protein-ligand binding [89-92]. Many computational methods have been developed to predict the location of binding-site water and/or its binding properties [93-98]. In recent years, molecular dynamics (MD) based methods have led to important advances in the study of active-site water and its thermodynamic role in ligand binding. The early key contributions include WaterMap [99], STOW [100], WATsite [101], and other approaches [102,103]. These methods usually determine high-density water locations as a spherical site, termed the “hydration site”, by analyzing the MD trajectory of protein and explicit water molecules, and calculate various thermodynamic quantities. For example, WaterMap locates hydration sites using a clustering algorithm, and calculates enthalpic and entropic contributions of individual hydration sites based on inhomogeneous solvation theory (IST) [104,105]. The hydration site analysis (HSA) helps researchers intuitively understand crucial water upon ligand binding, although it cannot represent the complex shape of high-density water regions by a collection of spheres [17]. Moreover, there is another MD-based approach called grid inhomogeneous solvation theory (GIST) [17,106]. Instead of locating hydration sites, GIST discretizes the continuous distribution of water density and thermodynamic properties onto three-dimensional grids. Accordingly, compared to HSA based methods, GIST can capture the complex shape of water distribution, covering high- and low-occupancy water regions.

Protein-ligand docking simulation is a powerful tool for the rational and efficient design of small molecules in structure-based drug design (SBDD) [107,108]. The atom-atom pairwise potentials, used in most of the scoring functions of docking programs, give a relevant approximation of interaction energy between proteins and ligands. However, the accurate estimation of thermodynamics of water molecules is still challenging due to the highly

expensive cost of computation for virtual screening [109]. In recent years, the precise modeling and scoring of water molecules has become a critical issue of protein-ligand docking [110-112]. For example, some early works introduced hydration water molecules which remain in the binding site and form hydrogen bonds to proteins and ligands into docking program, and improved docking performances [113,114]. However, thermodynamics of displaced water molecules are still underestimated or even ignored in protein-ligand docking. Many scoring functions of docking software, including AutoDock, use an implicit solvent model in the form of a continuous desolvation function [32,115,116] which cannot describe inhomogeneous active-site water molecules. The thermodynamics of displaced water molecules is a fundamental component of protein-ligand binding that contributes not only to the binding affinity but also to the binding conformation of ligands, since the ligand replaces unfavorable water molecules more easily than tightly bound water molecules [117]. Thus, the appropriate description of active-site water molecules should be essential for the improvement of docking performance.

Here, we incorporate thermodynamics of active-site water molecules into AutoDock4 [62] by combining a new desolvation function based on grid inhomogeneous solvation theory (GIST), which is called AutoDock-GIST. The GIST-based desolvation function was designed to formulate the driving force for unfavorable water molecules displaced by the binding ligand. Similar desolvation functions were proposed in previous studies of WaterMap and GIST [99,106]. Notably, they estimated the affinity difference between the closely related congeneric pair of ligands, where the difference in binding affinity results from dominant contributions of solvation rather than protein-ligand interaction [99,106]. Following these two key studies, the present work attempted to estimate binding affinities of diverse ligands and to improve docking

success rates by combining the scoring function of AutoDock and GIST-based desolvation function. Since AutoDock uses a gridded energy map for fast calculation of scoring function [118], the grid water properties of GIST are tractable to be incorporated into AutoDock. Furthermore, after calculating active-site water properties from single MD trajectory of the apoprotein and explicit water, the GIST-based desolvation function can be used for virtual screening campaign via docking with almost the same computational cost as in AutoDock4.

To validate the capability of our proposed scoring function, we study the complex system of coagulation factor Xa (FXa) and its small molecule inhibitors of 51 ligands which have experimentally measured binding affinities and X-ray crystal structures, including 28 ligands used in a previous work by WaterMap [99]. Using this dataset, we discuss the performance of AutoDock-GIST concerning the binding affinity estimation and the binding pose prediction. Furthermore, we evaluate the virtual screening performance, employing 793 active and 20,418 decoy compounds of FXa from the directory of useful decoys-enhanced (DUD-E) [119]. The results have revealed that scoring accuracy, docking success rate, and screening performance are significantly improved. Note that our work is a case study for a single target protein of FXa, but the finding generally supports the applicability of GIST for successful docking campaign.

3.2 Material and Methods

3.2.1 *Grid Inhomogeneous Solvation Theory (GIST)*

Grid inhomogeneous solvation theory (GIST) is a powerful and tractable computational method to calculate the hydration structure and thermodynamics of water around macromolecules, proposed by Nguyen et al. [17]. The thermodynamic properties of water molecules can be calculated based on inhomogeneous solvation theory (IST) [104,105], using

the snapshots of trajectories obtained from MD simulation of explicit water and protein. Most other computational methods, except GIST, use hydration site analysis (HSA) to identify the high-density and localized water region, called the hydration site. Although HSA-based approaches provide valuable insights into the role of specific water sites, they still have a significant limitation that they do not provide information on larger high-density water regions and other regions where the water density is low, rather than high, relative to bulk value [121]. To overcome these limitations, GIST discretizes IST onto a three-dimensional grid that fills the active site of protein, covering all occupancy regions of water (Figure 3.1). Thus, GIST provides more informative pictures of hydration water as the distribution of density and its thermodynamic properties.

GIST calculates various thermodynamic quantities of water molecules on the three-dimensional rectangular grid of cubic voxel k in the region of interest. The complete description of the GIST method is compiled in the original paper [17]. In the present work, we studied the following five properties of water molecules in voxel k , computed by GIST:

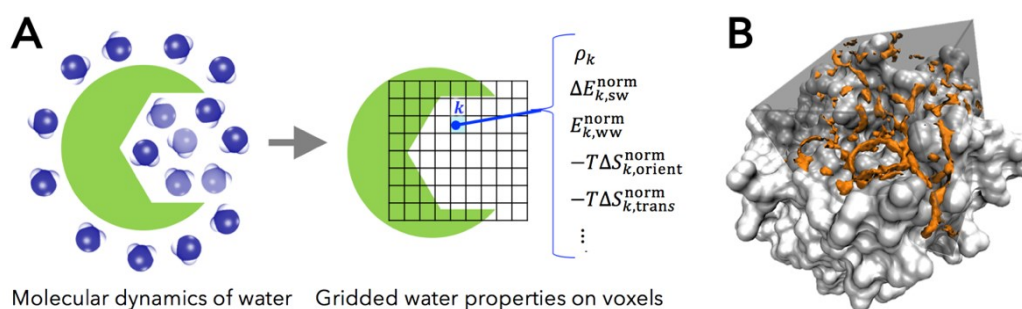


Figure 3.1 (A) Diagram of grid inhomogeneous solvation theory (GIST) calculation. The grid water properties of GIST are calculated using molecular dynamics (MD) trajectory of protein and explicit water; (B) The two-fold denser water regions (orange) than bulk in the active site of coagulation factor Xa (FXa) (gray) calculated by GIST. Figure prepared by using Visual Molecular Dynamics (VMD) [120].

- ρ_k , the number density of oxygen atom of water molecule found in a voxel k , in units of the density in bulk region (*i.e.*, the number density of bulk water $\rho_{\text{bulk}} = 1$).
- $\Delta E_{k,sw}^{\text{norm}}$, the mean energy of solute-water interaction per water molecule in a voxel k (kcal/mol/water). This quantity is referenced to bulk water, in the trivial sense that the energetic contribution of solute-water interaction is zero in bulk region.
- $E_{k,ww}^{\text{norm}}$, one-half the mean energy of water-water interaction per water molecule in a voxel k with all other water molecules (kcal/mol/water). The factor 1/2 prevents double counting of two water-water interaction and preserves the total energy of neat water being written as the sum of the single water energy [106].
- $-T\Delta S_{k,\text{orient}}^{\text{norm}}$, first-order orientational entropy per water molecule in a voxel k (kcal/mole/water), referenced to bulk water (*i.e.*, the orientational entropy of bulk water is set to be zero).
- $-T\Delta S_{k,\text{trans}}^{\text{norm}}$, first-order translational entropy per water molecules in a voxel k (kcal/mol/water), referenced to bulk water (*i.e.*, the translational entropy of bulk water is set to be zero).

Based on these quantities, thermodynamic properties of water molecules are described by following equations. Here, we regard the interaction energy as enthalpic contribution in this paper. The total enthalpy of a water molecule in a voxel k , relative to bulk, is defined as

$$\Delta H_k^{\text{norm}} = \Delta E_{k,sw}^{\text{norm}} + 2(E_{k,ww}^{\text{norm}} - E_{\text{bulk},ww}^{\text{norm}}) , \quad (3.1)$$

where $E_{\text{bulk},ww}^{\text{norm}}$ represents the mean energy of water-water interaction in bulk region. The value of ΔH_k^{norm} represents the mean interaction of a water molecule with the protein and all

other water molecules, referenced to that of bulk, $2E_{\text{bulk,ww}}^{\text{norm}}$. Similarly, the total entropy of a water molecule in voxel k , relative to bulk, is defined as

$$-T\Delta S_k^{\text{norm}} = -T\Delta S_{k,\text{orient}}^{\text{norm}} - T\Delta S_{k,\text{trans}}^{\text{norm}} , \quad (3.2)$$

where T is the absolute temperature (that is included in the entropy terms of GIST by default). Accordingly, the free energy of a water molecule in voxel k , relative to bulk, is the sum of total enthalpy and entropy written as

$$\Delta G_k^{\text{norm}} = \Delta H_k^{\text{norm}} - T\Delta S_k^{\text{norm}} . \quad (3.3)$$

Then, the unfavorable water molecule has a positive free energy ($\Delta G_k^{\text{norm}} > 0$); in contrast, the favorable water molecule has a negative free energy ($\Delta G_k^{\text{norm}} < 0$). As mentioned above, these thermodynamic quantities represent the differences from those of bulk water, which means that the displacement of high free-energy water is considered to be a driving force of protein-ligand binding.

3.2.2 *AutoDock4*

Our present method incorporates the GIST result into AutoDock4. AutoDock is one of the most widely used docking programs which is capable of quickly and accurately predicting bound conformation and binding energies [62]. In addition, AutoDock is widely used as a platform for the development of novel docking methodologies [114,122,123]. Two essential components of a docking program are an efficient search algorithm to find the conformation of the binding ligand and an accurate scoring function to estimate the binding free energy. AutoDock4 employs Lamarckian Genetic Algorithm (LGA) [39] for search algorithm and AutoDock4.2 force field [32] for the scoring function. The scoring function of AutoDock4 is a

semiempirical free-energy force field written as:

$$\begin{aligned}
 \Delta G_{\text{bind}}^{\text{AutoDock}} &= \Delta H_{\text{vdW}} + \Delta H_{\text{hbond}} + \Delta H_{\text{elec}} + \Delta S_{\text{conf}} + \Delta G_{\text{desolv}} \\
 &= W_{\text{vdW}} \sum_i^{\text{lig}} \sum_j^{\text{prot}} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right) \\
 &\quad + W_{\text{hbond}} \sum_i^{\text{lig}} \sum_j^{\text{prot}} E(\theta) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) \\
 &\quad + W_{\text{elec}} \sum_i^{\text{lig}} \sum_j^{\text{prot}} \left(\frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} \right) \\
 &\quad + W_{\text{conf}} N_{\text{tor}} \\
 &\quad + W_{\text{desolv}} \sum_i^{\text{lig}} \sum_j^{\text{prot}} (S_i V_j + S_j V_i) e^{(-r_{ij}^2/2\sigma^2)}.
 \end{aligned} \tag{3.4}$$

Here, the scoring function consists of five potential energy terms, including van der Waals ΔH_{vdW} , hydrogen bonding ΔH_{hbond} , electrostatic ΔH_{elec} , the conformational entropy of ligand ΔS_{conf} , and desolvation ΔG_{desolv} . The intermolecular potentials are calculated by summation over all pairs of ligand atom i and protein atom j as the function of their distance. The van der Waals term is a typical Lennard-Jones 12-6 dispersion/repulsion potential. The parameters A and B are taken from Amber force field [124]. The hydrogen bonding term is a Lennard-Jones 12-10 dispersion/repulsion potential with the directionality of hydrogen bond $E(\theta)$ depending on the angle θ and the parameters C and D [125]. The electrostatic term is a screened Coulomb potential with the distance-dependent dielectric function $\varepsilon(r_{ij})$ [126]. The conformational entropy term represents the loss of torsional entropy upon binding, depending on the number of rotatable bonds of ligand N_{tor} . The last term is a desolvation potential based on the volume V of atoms that surround a given atom and shelter it from the

solvent, weighted by the charge-based solvation parameter S and the exponential term with distance-weighting factor σ [127]. The coefficients W are weight factors fitted using the training set of the crystal structure of protein-ligand complexes and the experimentally measured binding affinities. Since the scoring function of AutoDock4 has these weight factors, it is called a semiempirical scoring function.

Using this scoring function and the optimization algorithm, AutoDock4 searches the most stable (*i.e.*, the lowest energy) binding conformation of the ligand in the user-defined cubic docking site (Figure 3.2). To enable searching for a large conformational space available to a ligand in protein, AutoDock4 introduced a grid-based energy calculation method. In this approach, the binding site of a target protein is embedded in the grid map. Before the docking simulation, a probe atom is sequentially set on each grid center, and the interaction energy between a probe atom and the target protein is calculated and stored in the grid map. This grid

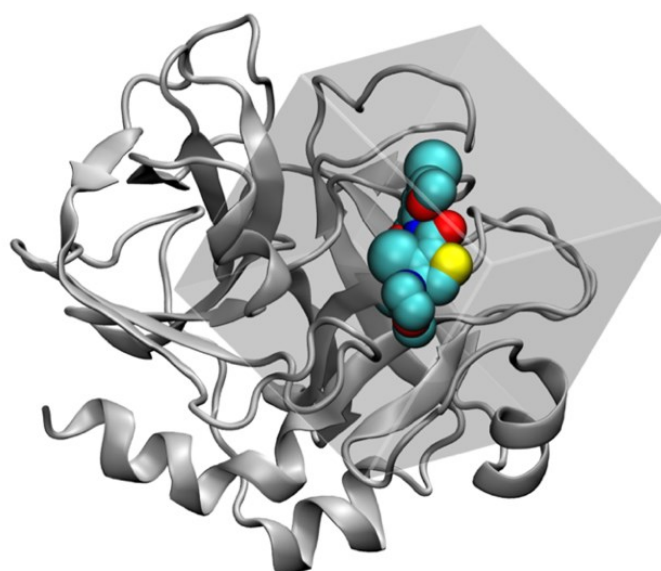


Figure 3.2 Co-crystal structure of FXa (cartoon) with ligand, Protein Data Bank (PDB) ligand-id (HET) XLD, in van der Waals representation (cyan). The cubic region represents the docking site of AutoDock4 (gray). Figure prepared by using VMD [120].

map is used as a lookup table during the docking simulation for rapid energy evaluation of ligand conformations. This cubic docking region and grid-based potential calculation approach are quite suitable to be combined with the description of water properties by GIST.

3.2.3 Development of GIST-based Desolvation Function

Although the free energy change of displacing water can be calculated by GIST results directly, many previous studies reported that there was no direct correlation between the free energy of water molecules in the binding site and the affinity of bound ligands and that the use of the simplified scoring function performed well [81,99,106]. Hence, we developed GIST-based desolvation function according to a simple physical principle: If a heavy atom of ligand displaced a high-occupancy and unfavorable water molecule, the ligand earned a favorable contribution in binding free energy. The unfavorable water in this context corresponds to the high free-energy water for which the enthalpy-entropy compensation breaks down and either enthalpy or entropy is significantly unfavorable. Based on this physical principle, we design and propose a desolvation function suitable for grid-based energy calculation of AutoDock4. Once running MD simulation of apoprotein and explicit water and calculating thermodynamics of water with GIST, the grid water properties are readily converted to the map of unfavorable water according to two criteria as follows: (I) The free energy of a water molecule in a voxel k , ΔG_k^{norm} , is higher than a cutoff value ΔG_{co} ; (II) The number density of a water molecule in a voxel k , ρ_k , is greater than a cutoff value ρ_{co} . Using this map of unfavorable water, the displacing gain of an unfavorable water molecule is calculated as:

$$\Delta G_{\text{watdisp}} = \sum_i^{\text{lig}} \delta_i \Delta G_{\text{aff}} , \quad (3.5)$$

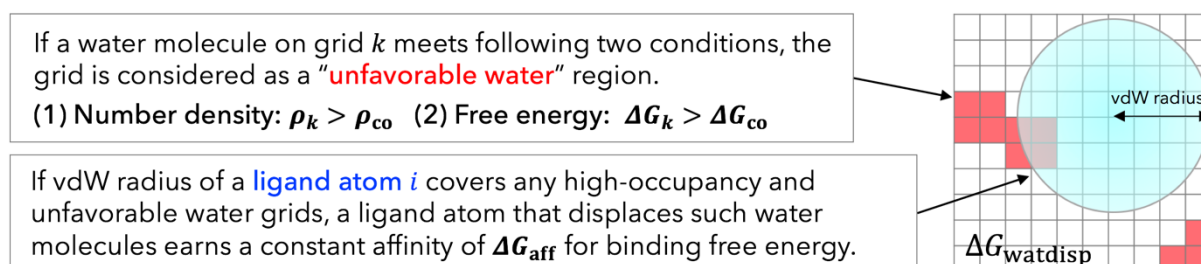
$$\delta_i = \begin{cases} 1 & \text{if vdW radius of ligand atom } i \text{ covers unfavorable water grid } k \\ 0 & \text{otherwise} \end{cases} . \quad (3.6)$$

Here, ΔG_{aff} is a fitting parameter which specifies the free energy gain by displacement of the unfavorable water molecule; δ_i is a binary displacement indicator which equals 1 if the vdW radius of a ligand atom i covers any unfavorable water grid k and 0 otherwise. Figure 3.3 shows the diagram of this method. Note that our proposed method has three parameters, ρ_{co} , ΔG_{co} , and ΔG_{aff} , which have to be fitted according to the binding thermodynamics of ligands.

The GIST-based solvation function $\Delta G_{\text{watdisp}}$ was incorporated into the scoring function of AutoDock4, which is called AutoDock-GIST. This incorporation is achieved by a simple summation of the AutoDock4.2 force field and the GIST-based solvation function, expressed as:

$$\Delta G_{\text{bind}}^{\text{AutoDock-GIST}} = \Delta G_{\text{bind}}^{\text{AutoDock}} + \Delta G_{\text{watdisp}} \quad (3.7)$$

Note that we retained the original desolvation term ΔG_{desolv} of AutoDock4 (Equation (3.4)) in the proposed scoring function. Since the desolvation term ΔG_{desolv} is based on the



continuous solvation model and represents a penalty of binding free energy [32], we assumed that the displacing gain of unfavorable water $\Delta G_{\text{watdisp}}$ does not conflict with ΔG_{desolv} . The AutoDock-GIST approach takes advantage of on-the-fly evaluation to search binding conformations in the docking process, as compared with other rescoring-after-docking models [128-130]. Since AutoDock uses the optimization algorithm to search the binding poses of ligand, the scoring function significantly affects the conformations of sampled binding poses. The pre-configured scoring function of AutoDock-GIST is capable of docking the ligand while taking into account the displacement of unfavorable water molecules. Furthermore, once calculating the GIST-based desolvation function, the AutoDock-GIST calculation can be implemented in high-throughput docking with almost the same computational cost as in AutoDock4.

The three fitting parameters of proposed scoring function, ρ_{co} , ΔG_{co} , and ΔG_{aff} , were adjusted and validated using 51 ligands of FXa consisting of 28 training set ligands and 23 test set ligands. In this work, we sought two sets of optimal parameters for protein-ligand docking: (I) Affinity parameter set, which maximized the correlation between calculated score $\Delta G_{\text{bind}}^{\text{AutoDock-GIST}}$ and experimentally measured binding affinity ΔG_{exp} ; (II) Pose parameter set, which maximized the success rate of binding pose prediction yielding root mean square deviation (RMSD) between docking pose and native pose of ligands less than 2 Å. To find these parameters, we scanned the value of ρ_{co} from 1.0 to 6.0 by increments of 0.1, the value of ΔG_{co} from 0.0 to 4.0 kcal/mol by increments of 0.1 kcal/mol, and the value of ΔG_{aff} from 0.0 to -2.0 kcal/mol by decrements of 0.01 kcal/mol, respectively. This scan yields $61 \times 41 \times 201 = 50,271$ combinations of the three parameters. For each combination, the training set ligands are calculated with AutoDock-GIST, and evaluated by each of the two

conditions above. The optimal parameters found in this procedure were then validated using the test set ligands.

3.2.4. Datasets and Preparation

3.2.4.1 Structure Preparation and MD Simulation for FXa

In this work, we studied the coagulation factor Xa (FXa) to assess the performance of AutoDock-GIST. To analyze thermodynamics of active-site water molecules of FXa by GIST, we performed MD simulation of apoprotein and explicit surrounding water. The crystal structure of FXa was obtained from Protein Data Bank (PDB) [3] entry 1FJS [131], as studied previously [99,106]. First, we removed all crystallographic water molecules and bound ligands, keeping ions, from the system, and added hydrogens using the program Reduce [132]. We also removed the chain L of the crystal structure. We then used Tleap program from AmberTools [133] to prepare the system. We assigned protein parameters from AMBER99SB force field [134] and solvated the system in a TIP3P [135] water box with the periodic boundary condition, keeping the minimum distance of 10 Å away from any atom of the protein. Four disulfide bonds were set up and two crystal ions, Ca^{2+} and Cl^- , were restrained at their original positions.

After preparing the system, we minimized the energy of the system and ran MD simulation. All following procedures were carried out with the Amber 14 software using pmemd.cuda [136]. First, we minimized the system energy in two steps: (I) Only the water while restraining all protein atoms; (II) The water and the protein hydrogen atoms while restraining the protein heavy atoms. Both minimization steps used 1500 cycles of the steepest descent algorithm followed by the conjugate gradient method for the maximum of 20,000 cycles, where the atoms

were harmonically restrained with force constant of 100 kcal/mol/Å. Next, the system was heated for 200 ps from 0 K to 50 K in the NVT ensemble with the first simulation and the temperature was incremented by 50 K for 200 ps in the NPT ensemble until 300 K was reached. The system was then equilibrated for 10 ns at 300 K in the NPT ensemble. At the final volume, the system was equilibrated again for 5 ns at 300 K in the NVT ensemble. The final production MD run of 100 ns was performed in the NVT ensemble, and snapshots of this simulation were saved every 1 ps, for a total 10,000 frames of snapshots stored. Notably, during all MD simulations, all protein atoms were harmonically restrained with a force constant of 100 kcal/mol/Å. A time step of 2 fs was employed with SHAKE algorithm [137]. The temperature was regulated by Langevin thermostat; the nonbonded interactions were truncated at 9 Å and the particle mesh Ewald method was implemented to account for the long-range electrostatic interaction [138]. After all, for the GIST calculation, the trajectory of production MD was aligned across all frames referenced to the initial position of the protein, using the cpptraj program [139].

3.2.4.2 GIST Calculation and Docking Set-up

Before the GIST calculation, we prepared the FXa structure for docking simulation, following the standard AutoDock protocol. First, the protein structure of 1FJS was aligned to the initial coordinate of MD trajectory, to superpose the GIST region and docking region of AutoDock. The bound ligand, water, and ions were removed from the system and polar hydrogens were added to the protein using AutoDockTools [62]. The docking site was set to $22.5 \times 22.5 \times 22.5 \text{ \AA}^3$ cubic region centered at bound ligand of 1FJS, which was the range to cover the active site of FXa. In this docking site, grid-based potential maps were calculated by

AutoGrid (included in AutoDock suit). We then used a default grid size of 0.375 Å (approximately a quarter of the vdW radius of carbon atom) to calculate the grid-based potential maps of AutoDock, which resulted in the number of grid points of map $60 \times 60 \times 60$.

The GIST calculation was performed using `cpptraj` program included in AmberTools [139,140]. The cubic region of GIST analysis was set to the active site of FXa, corresponding to the docking region set above. The grid centroid position was the center of docking site. The grid size was $60 \times 60 \times 60$. The voxel side length (grid spacing) was 0.375 Å, the same as default grid size of AutoDock4. The thermodynamic properties of active-site water molecules were then calculated by GIST using MD snapshots, and the free energies of water molecules were calculated based on Equations (3.1)–(3.3), and subsequently, the GIST-based desolvation function was adjusted using the training set described below.

3.2.4.3 Dataset Preparation and Docking Metrics

For the evaluation of proposed method, we used diverse 51 ligands of FXa for which both experimentally measured binding affinities and X-ray crystal structures are known. The 51 ligands were grouped into the training set and the test set to optimize and validate fitting parameters of GIST-based desolvation function. First, for the training set, we used 28 ligands of FXa which were used in a previous computational study [99] (see Table B1 in Appendix B). Next, for the test set, we selected an additional 23 ligands of FXa from PDBbind 2007 refined set [141] (see Table B2 in Appendix B). Note that we then eliminated some FXa ligands from original PDBbind dataset, which have the adverse correlation between molecular weight and binding affinity (*e.g.*, a ligand with low molecular weight but high binding affinity or a ligand with high molecular weight but low binding affinity), since with such ligands it is quite difficult

to estimate the correct binding affinity by scoring functions of docking programs [142,143]. As a result, the correlations between molecular weight and binding affinity of the training set and the test set are 0.48 and 0.33 in R^2 values, respectively. All ligands in the training set and the test set were carefully aligned on the initial structure of simulated protein (1FJS) and their energies were minimized by AMBER12:EHT force field using Molecular Operating Environment (MOE) [144]. In addition, we used a compound dataset of FXa obtained from the directory of useful decoys-enhanced (DUD-E) [119] to validate the virtual screening performance of AutoDock-GIST. The virtual screening dataset includes 793 active and 20418 decoy compounds of FXa. All of the ligands used in this study were prepared for docking simulation, by using AutoDockTools [62].

The capability of AutoDock-GIST was assessed in terms of binding affinity prediction, docking success rate, and virtual screening performance. First, the accuracy of binding affinity prediction was measured by the correlation between the calculated score of native pose ligand and the experimentally measured binding affinity, for the R^2 value of Pearson correlation coefficient. Next, the docking calculation was performed 10 times for each ligand, and the lowest energy conformation was selected. The docking success rate was then calculated based on RMSD between the predicted binding pose and crystal pose of ligand. In this work, an RMSD of less than 2 Å was regarded as a success of binding pose prediction. At last, the performance of virtual screening was evaluated by area under the curve (AUC) of receiver operating characteristic (ROC) [145] and enrichment factor (EF) [146]. The ROC curve plots the true positive rate against the false positive rate of virtual screening results, and the context of AUC represents the area under the ROC curve. The range of the AUC is 0 to 1: the value 1 represents ideal virtual screening result, and the value 0.5 represents random selection. The

enrichment factor is a characteristic of a rank-ordered list of a given first $x\%$ subset, calculated as:

$$EF(x\%) = \frac{\text{hits}_x/N_x}{\text{hits}_t/N_t}, \quad (3.8)$$

where hits_x is the number of actives found in the first $x\%$ subset, N_x is the total number of compounds at first $x\%$ subset; hits_t and N_t are the total number of actives and the total number of compounds in the entire docked dataset, respectively. Therefore, $EF(x\%)$ estimates how many times a docking program can pick out actives relative to random, in the first $x\%$ subset of a rank-ordered docking result.

3.3 Results and Discussion

3.3.1 Parameter Fitting for GIST-based Desolvation Function

In this section, we discuss adjusted parameters of GIST-based desolvation function and the unfavorable water distributions in the active site of FXa. As mentioned above, we constructed the two sets of parameters: (I) The affinity parameter set which maximized the correlation between docking score and experimentally measured binding affinity; and (II) The pose parameter set which maximized the success rate of binding pose prediction in docking. The values of three parameters were systematically searched from the parameter space using 28 training set ligands of FXa. As a result, we found optimal values of ρ_{co} , ΔG_{co} , and ΔG_{aff} for each parameter set (Table 3.1), so that docking and scoring performances were significantly improved, as will be discussed in the following sections.

AutoDock-GIST: Incorporating Thermodynamics of Active-Site Water into Scoring Function for Accurate Protein-Ligand Docking

Table 3.1 Adjusted parameters for GIST-based desolvation function of AutoDock-GIST.^a

Parameter set	ρ_{co}	ΔG_{co} [kcal/mol/water]	ΔG_{aff} [kcal/mol]
Affinity parameter set	4.8	1.0	-0.50
Pose parameter set	4.3	1.9	-0.25

^a ρ_{co} is a density cutoff parameter for unfavorable water molecules in active site; ΔG_{co} is a free-energy cutoff parameter for active-site water; ΔG_{aff} is a free-energy gain of unfavorable water molecule displaced by a ligand heavy atom. Parameter fitting methods are described in Materials and Methods section.

In both parameter sets, density cutoff parameters ρ_{co} have high values beyond 4, in other words, the unfavorable water region of GIST-based desolvation function has over four-fold higher density than that for bulk water. The value of ρ_{co} in the affinity parameter set is slightly greater than that in the pose parameter set. On the other hand, the value of free-energy cutoff parameter ΔG_{co} in the affinity parameter set is approximately a half of that in the pose parameter set, that is, the affinity parameter set picks up less unfavorable water molecules than the pose parameter set. The displacing gain of unfavorable water, ΔG_{aff} , is two-fold higher in the affinity parameter set than that in the pose parameter set. In summary, the affinity parameter set gives high free-energy gain to the displacement of unfavorable water molecules, while the pose parameter set gives low free-energy gain to displacement of highly unfavorable water molecules.

The active site of FXa and the distribution of unfavorable water for each parameter set are shown in Figure 3.4. The active site of FXa includes two important subpockets for bound inhibitors, S1 and S4 [147] (Figure 3.4A). The S1 pocket is a deeply concave region and determines the major component of selectivity and binding by residues Asp189, Ser195, and Tyr228. The S4 pocket, called hydrophobic box, is formed from three aromatic residues Tyr99,

Phe174, and Trp215. FXa inhibitors are generally bound in an L-shaped conformation, where one group of the ligand occupies the anionic S1 pocket, and another group of the ligand occupies the aromatic S4 pocket; a fairly rigid linker group connects these two interaction sites [148]. The unfavorable water region of GIST-based desolvation function was determined by two cutoff parameters, density cutoff parameter ρ_{co} and free-energy cutoff parameter ΔG_{co} . In both parameter sets, the unfavorable water molecules were found in both S1 and S4 pockets; in other words, GIST analysis indicated that high-occupancy and high free-energy water molecules exist in S1 and S4 pockets. This result coincides with an early computational study

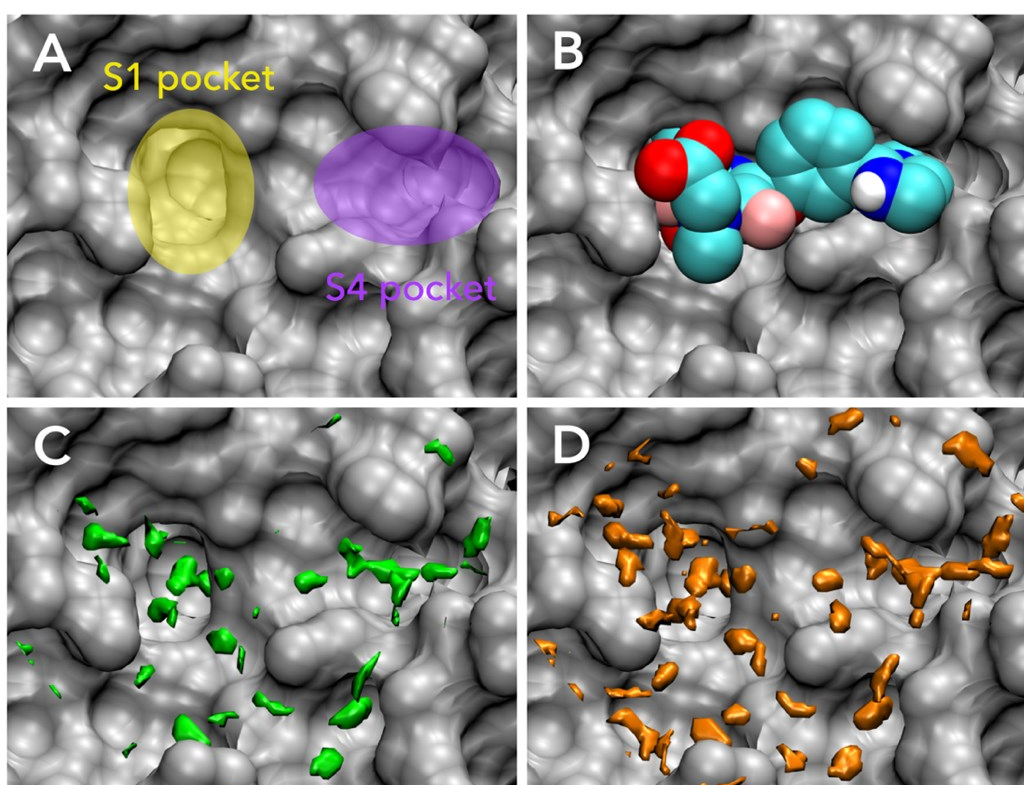


Figure 3.4 Binding ligand and distributions of unfavorable water for GIST-based desolvation function in the active site of FXa (PDB-id: 1FJS, gray): (A) Binding hot spots of FXa, S1 pocket (yellow), and S4 pocket (purple); (B) The bound ligand of 1FJS (residue-id: Z34), in van der Waals representation (cyan); (C) The unfavorable water distribution for pose parameter set (green); (D) The unfavorable water distribution for affinity parameter set (orange). Figure prepared by using VMD [120].

of FXa by WaterMap [99]. However, the unfavorable water regions of the pose parameter set and affinity parameter set showed somewhat different configurations. For the pose parameter set, the high value of ΔG_{co} caused the tight distribution of unfavorable water on the binding hot spots of FXa (Figure 3.4C). In contrast, for the affinity parameter set, the low value of ΔG_{co} caused the broad water distribution covering the active-site surface of FXa (Figure 3.4D).

For further discussion, we analyzed the free energy components of unfavorable waters in the active site of FXa. We have discussed the unfavorable active-site water in a term of high free energy so far. However, there are two types of unfavorable water regions which comprise enthalpically unstable water or entropically unstable water in an active-site of protein. It is widely known that enthalpy and entropy compensate each other in biomolecular systems [87,149-151]. For instance, a water molecule placed on a hydrophobic surface is enthalpically unfavorable, since it cannot make appropriate hydrogen bonds. However, at the same time, such water molecules are entropically favorable, because the missing hydrogen bond relaxes its orientation and earns orientational entropy. In contrast, a tightly bound water molecule is enthalpically favorable but entropically unfavorable due to its fixed orientation. The high free-energy water then causes the breakdown of enthalpy-entropy compensation and either enthalpy or entropy is significantly unfavorable. For each parameter set of GIST-based desolvation function, we decomposed unfavorable water region into an enthalpically unfavorable water and an entropically unfavorable water regions (Figure 3.5). Here, the enthalpically dominant water represents $\Delta H_k^{norm} > -T\Delta S_k^{norm}$, whereas the entropically dominant water represents

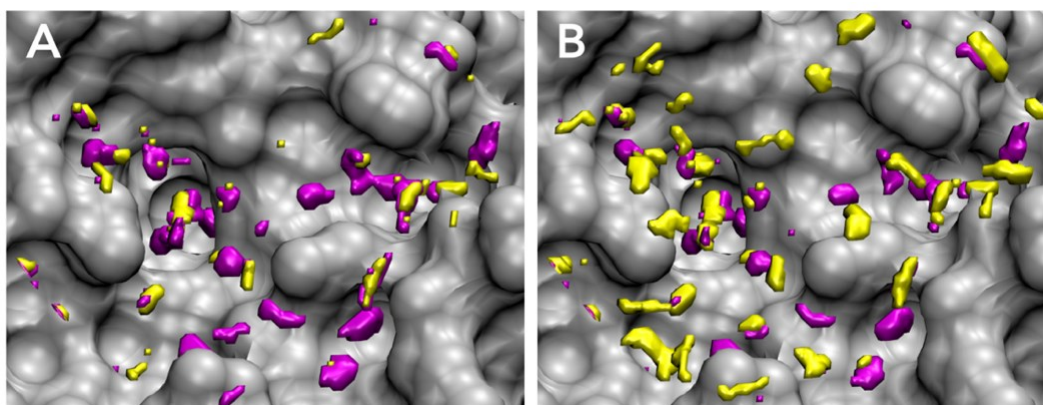


Figure 3.5 Enthalpy-entropy decomposition of unfavorable water distributions for GIST-based desolvation function in the active site of FXa (gray): (A) pose parameter set; (B) affinity parameter set. More enthalpically unfavorable water regions are shown in purple ($\Delta H_k^{\text{norm}} > -T\Delta S_k^{\text{norm}}$), whereas more entropically unfavorable water regions are shown in yellow ($\Delta H_k^{\text{norm}} < -T\Delta S_k^{\text{norm}}$). Figure prepared by using VMD [120].

$\Delta H_k^{\text{norm}} < -T\Delta S_k^{\text{norm}}$. The results showed that the unfavorable water for pose parameter set was more enthalpically unfavorable (Figure 3.5A), whereas that of affinity parameter set was more entropically unfavorable (Figure 3.5B). The main difference in two parameter sets was the value of free-energy cutoff ΔG_{co} : The value of ΔG_{co} in the affinity parameter set is approximately a half of that in the pose parameter set. Hence, the results also indicate that the enthalpically unfavorable water is highly unfavorable in its free energy more than the entropically unfavorable water. In other words, the entropically unfavorable water is not so unfavorable in its free energy than the enthalpically unfavorable water.

3.3.2 Accuracy of Binding Affinity Prediction for FXa ligands

After the fitting parameters of GIST-based desolvation function were adjusted by 28 training sets ligands, the scoring accuracy of AutoDock-GIST was assessed for 23 test set ligands. Figure 3.6 shows the results of binding affinity predictions for FXa ligands. The R^2

values between calculated score of AutoDock4 and experimentally measured binding affinity were 0.38 for the training set ligands and 0.49 for the test set ligands, respectively. In contrast, the affinity parameter set of AutoDock-GIST found the optimal parameters achieving the R^2 value 0.60 for the training set ligands, and also improved the R^2 value to 0.58 for the test set of ligands. Hence, this result has proved that the displacing gain of unfavorable water is an essential factor to improve the scoring function of docking. Some typical improvements are highlighted in Figure 3.6. For instance, AutoDock4 scoring function underestimated the binding free energy for the ligand of 1FJS (blue), since its interaction energy with the protein was not so high. On the other hand, the ligand of 1FJS successfully displaced some unfavorable water molecules and earned favorable free energy gain whose value of $\Delta G_{\text{watdisp}}$ was -17.5 kcal/mol. A similar improvement was observed in the ligand of 2Y5F (red), which had poor interaction with the protein but displaced a great deal of unfavorable water. The value of $\Delta G_{\text{watdisp}}$ was -16.0 kcal/mol for 2Y5F ligand. In contrast, the binding free energy of ligand of 2J34 (green) was overestimated by AutoDock4 scoring function, since it had favorable vdW interactions with protein atoms. However, the ligand of 2J34 earned little displacing gain of unfavorable water molecules so that the value of $\Delta G_{\text{watdisp}}$ was -14.0 kcal/mol. As a result, these differences in the values of $\Delta G_{\text{watdisp}}$ significantly improved the scoring accuracy of the affinity parameter set.

The same calculation was performed with the pose parameter set of AutoDock-GIST. Even though the pose parameter set was not adjusted in consideration of the accuracy of binding affinity prediction, interestingly, the R^2 values were slightly improved, which are 0.41 and 0.50 for the training set and the test set, respectively. This result also supported the fact that the GIST-based desolvation function correctly described an essence of binding thermodynamics of

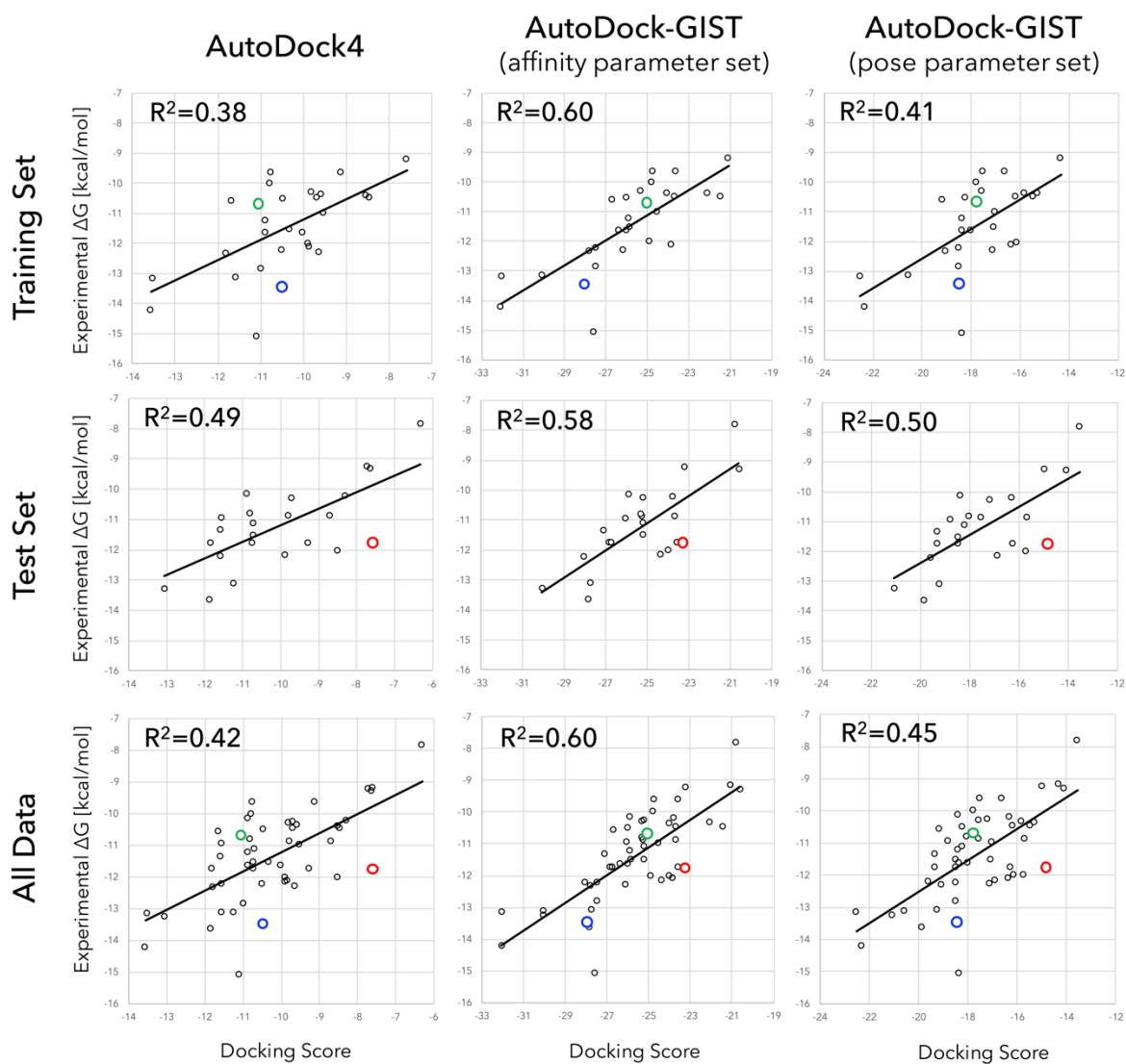


Figure 3.6 Scatter plots and regression lines of experimentally measured binding affinities versus docking scores of AutoDock4 (**left**), AutoDock-GIST with affinity parameter set (**middle**), and AutoDock-GIST with pose parameter set (**right**) for FXa ligands of training set (**upper**), test set (**middle**) and all data (**lower**). R^2 values represent the squares of Pearson correlation coefficients. Color plots show specific examples of improvements: blue, green, and red circles represent the ligands of 1FJS, 2J34, and 2Y5F, respectively.

ligand. On the other hand, since the pose parameter set had the higher value of free-energy cutoff ΔG_{co} and the lower value of displacing gain ΔG_{aff} than those of the affinity parameter set, the calculated result showed that these parameters did not significantly affect scoring accuracy. In other words, the result suggested that low free-energy cutoff value of unfavorable water and high displacing gain were effective for quantitative scoring of binding free energy. Notably, in this study, the absolute values of AutoDock-GIST scores were greater than those of AutoDock4 since we did not scale them in comparison to the experimental values.

3.3.3 Docking Success Rate for FXa ligands

We expected that the displacement of unfavorable water molecules might contribute to the favorable conformation of binding ligand and inclusion of displacing gain should improve the docking performance. Based on this assumption, the pose parameter set of the GIST-based desolvation function was adjusted to be optimal for binding pose prediction using 28 training set ligands, and evaluated by 23 test set ligands. Table 3.2 shows the results of docking calculation of pose prediction success rates for FXa ligands. The docking success rates of AutoDock4 were 75.0% and 82.6% for the training set and the test set, respectively. As expected, the pose parameter set of AutoDock-GIST found the suitable parameters for binding pose prediction which resulted in a docking success rate 89.3% for the training set ligands, and also improved the docking success rate up to 95.7% for the test set ligands. On the other hand, for the affinity parameter set of AutoDock-GIST, the docking success rates were almost unchanged or were a little bit improved, which were 71.4% for the training set and 90.4% for the test set.

Table 3.2 Accuracies of binding pose predictions: docking success rates^a for FXa ligands.

Data set	AutoDock4	AutoDock-GIST (pose parameter set)	AutoDock-GIST (affinity parameter set)
Training set	75.0%	89.3%	71.4%
Test set	82.6%	95.7%	90.4%
All data	78.4%	92.1%	80.4%

^a RMSD between the native structure and the docking pose of ligand being less than 2Å was regarded as a success of binding pose prediction.

For further discussion, we carefully analyzed docking results and found three typical cases that displacement of unfavorable water molecules affected conformations of docking ligands (Figure 3.7). First, for the ligand of 1NFX (residue-id: RDR), AutoDock4 successfully found the native-like pose of the bound ligand with RMSD of 1.24, and the pose parameter of AutoDock-GIST also reproduced the native-like pose of the bound ligand with RMSD of 1.42 (Figure 3.7A). However, the affinity parameter set of AutoDock-GIST failed to dock the ligand with RMSD of 6.18. In the other two cases for the ligands of 1MQ6 (residue-id: XLD) and 1NFU (residue-id: RRP), only the pose parameter set of AutoDock-GIST successfully reproduced the native-like poses, whereas AutoDock4 and the affinity set of AutoDock-GIST docked the ligands at far from native pose (Figure 3.7B,C). As mentioned above, the unfavorable water regions of the pose parameter set were mostly placed on the important binding pockets of FXa, S1, and S4 (see Figure 3.4). The docking results clearly showed that the displacing gain of such unfavorable water molecules was an essential factor in determining the binding conformations of FXa ligands. In fact, the displacement of some unfavorable water in the pose parameter set indicated with blue arrows in Figure 3.7 seem to contribute to successful docking simulations. On the other hand, the affinity parameter set of AutoDock-

GIST did not improve the docking success rate significantly, and found some unusual docking poses that were different from those of AutoDock4. In the cases of the affinity parameter set, we supposed possibilities that broad distribution of the unfavorable water and high displacing gain might yield unnecessary local minima in the free energy landscape of scoring function and merely caused docking failures.

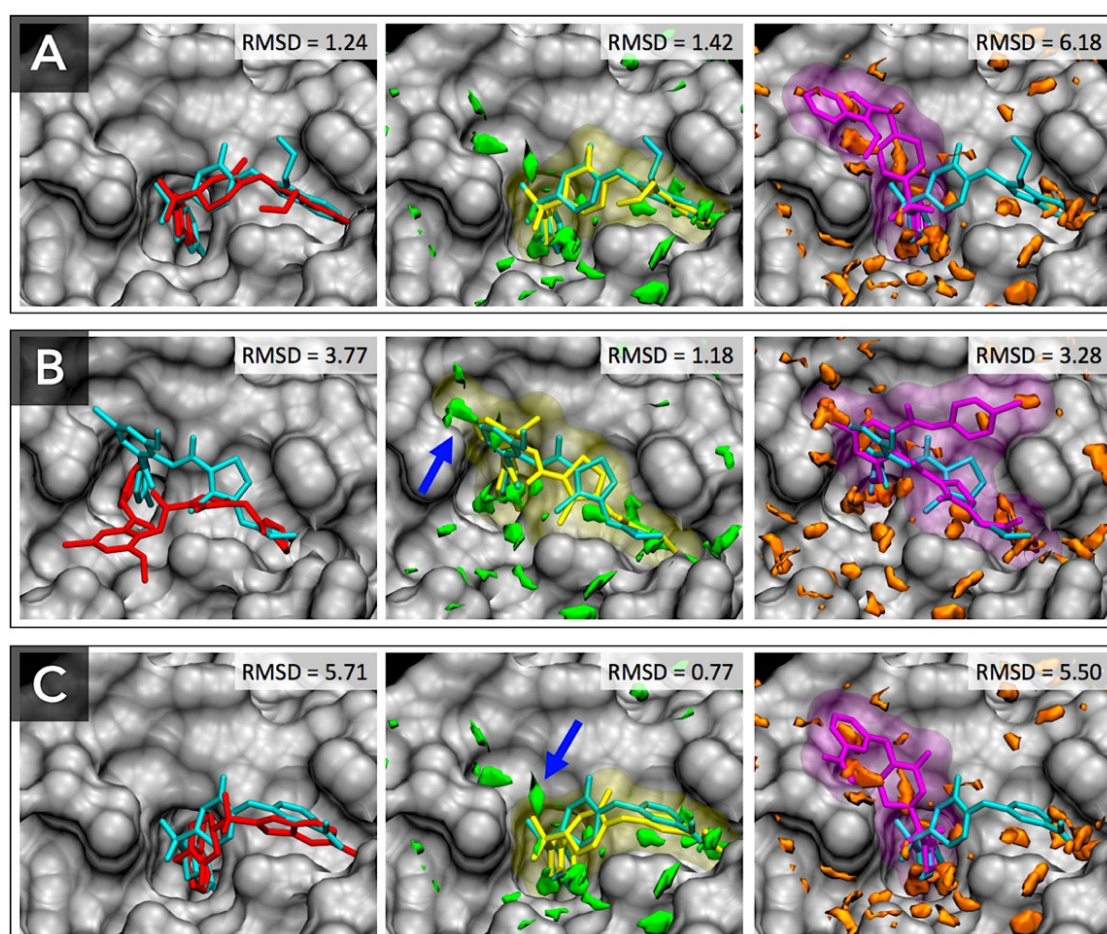


Figure 3.7 Docking results for FXa ligands of PDB entries: (A) 1NFX, (B) 1MQ6 and (C) 1NFU. Native crystallographic structures of bound ligands are shown as cyan sticks. Docking results of AutoDock4 are shown as red sticks (**left**), those of AutoDock-GIST with the pose parameter set are shown as yellow sticks and transparent surface (**middle**), and those of AutoDock-GIST with the affinity parameter set are shown as purple sticks and transparent surface (**right**). The unfavorable water distributions for the pose parameter set and the affinity parameter set are shown as green and orange regions, respectively. The blue arrows point to unfavorable water molecules which contribute to the successful docking. Figure prepared by using VMD [120].

3.3.4 Virtual Screening Performance of AutoDock-GIST

Another key measure of the docking performance is the enrichment of ligands among the top ranking docked compounds. We evaluated virtual screening performance of AutoDock-GIST through the docking calculation for 793 active and 20,418 decoy compounds of FXa from the directory of useful decoys-enhanced (DUD-E) [119]. Figure 3.8 shows the ROC plot and AUC of docking results. Though AutoDock4 showed good screening performance with AUC of 80.4%, both parameter sets of AutoDock-GIST improved the value of AUC compared with AutoDock4, which were 85.6% for the affinity parameter set and 86.4% for the pose parameter set. Interestingly, even though the pose parameter set of AutoDock-GIST was not adjusted in consideration of quantitative binding affinity of FXa ligands, it showed a slightly better performance than that of the affinity parameter set.

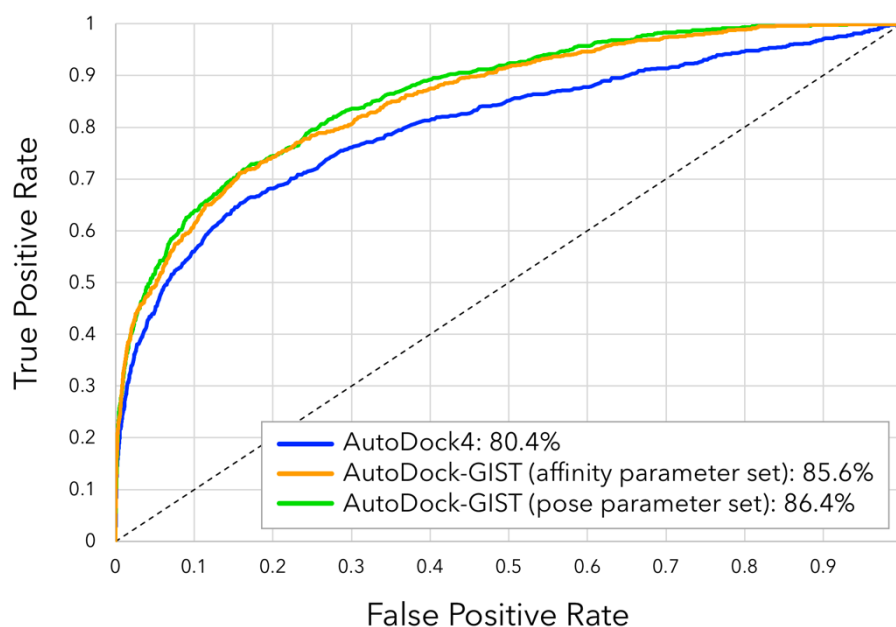


Figure 3.8 Receiver operating characteristic (ROC) plots of docking results for FXa ligands: AutoDock4 (blue), AutoDock-GIST with the affinity parameter set (orange), and AutoDock-GIST with the pose parameter set (green). The values represent the percentages of the AUC.

We also assessed the early enrichment of docking results by enrichment factor EF (Table 3.3). For all subset sizes, AutoDock-GIST resulted in superior performance to AutoDock4. The values 26.75 of EF(0.1%) in both parameter sets of AutoDock-GIST represent that 21 compounds of the top 0.1% subset were all active compounds, which are calculated by $EF(0.1\%) = (21/21)/(793/(20,418 + 793))$ with Equation (3.8). The affinity parameter set of AutoDock-GIST showed the best value of EF(0.5), 25.23. For the larger subset, the pose parameter set of AutoDock-GIST performed the best. As mentioned above, the affinity parameter set of AutoDock-GIST tended to cause the docking failure frequently compared with that of the pose parameter set, and the incorrect binding pose then resulted in the wrong estimation of the binding affinity [13]. In other words, improvement of docking success rate with the pose parameter set contributed more positively to the virtual screening campaign than the affinity parameter set. Eventually, the virtual screening results indicated that our method was feasible to deal with diverse ligands of FXa and inclusion of displacing gain of unfavorable water molecules had a significant advantage in the docking campaign.

Table 3.3 Enrichment factors for DUD-E ligands of FXa.^a

Metrics	AutoDock4	AutoDock-GIST (affinity parameter set)	AutoDock-GIST (pose parameter set)
EF(0.1%)	25.47	26.75	26.75
EF(0.5%)	23.22	25.23	24.73
EF(1%)	21.45	22.84	23.34
EF(2%)	16.65	17.92	19.11
EF(5%)	10.30	10.57	12.26
EF(10%)	6.44	6.61	7.59

^a The percentage in parenthesis represents the ratio of subset of rank-ordered list in docking result.

3.4 Conclusions

Although the thermodynamics of active-site water molecules are widely appreciated in the studies of molecular recognition, it is still challenging to estimate its contributions in protein-ligand docking quantitatively. Here, we showed a case study of the combination of GIST and AutoDock4, called AutoDock-GIST, and discussed the effectiveness of displacing gain of unfavorable water in protein-ligand docking. Following early key studies of GIST [106] and WaterMap [99], the present GIST-based desolvation function was designed on the basis of a simple physical principle: if a heavy atom of ligand displaced a high-occupancy and unfavorable water molecule, the ligand earned a favorable contribution in binding free energy. We studied diverse ligands of FXa by the proposed docking method and concluded that displacing gain of unfavorable water molecules was an essential factor for protein-ligand docking. The computational results showed that inclusion of water thermodynamics could improve not only quantitative scoring of binding affinity but also a conformational prediction of binding ligand. The result also indicated that the proposed method had a significant advantage in the virtual screening of the large compound set of FXa via docking.

Another interesting finding was that the high free-energy water molecules in the active site of FXa were mostly enthalpically unfavorable, rather than entropically unfavorable. This result is consistent with many previous studies that enthalpically unfavorable water molecules are more important for molecular recognition when they are displaced by a binding ligand [81,106]. In addition, our result revealed that the entropically unfavorable water molecules are also effective for quantitative binding affinity calculation when we consider the free energy of water molecules. However, our enthalpy dominant water model for the pose parameter set did not significantly improve the accuracy of binding affinity calculation. It implies that the

displacement gain of enthalpically unfavorable water has a similar property to the scoring function of AutoDock4. It is possible that, since empirical or semiempirical scoring functions fit their interaction potentials to experimentally measured binding affinity ignoring the displacement of water molecules, they implicitly include a part of water replacement energies [33,152]. In fact, the weight factor of vdW potential in AutoDock4 scoring function is 1.37 times higher than that of hydrogen bonding potential. It might be modeling the difference of the displacement energy of active-site water molecules between a hydrophobically enclosed one (enthalpically unfavorable) and a hydrogen bonded one (enthalpically favorable). Hence, this presumption indicates that we should re-adjust potential parameters of scoring function with explicit water replacement terms for a more rigorous description of displacing water molecules.

Though the displacement of unfavorable water molecules is a principal driving force of the protein-ligand binding, it is worth mentioning that it is only a part of water thermodynamics upon ligand binding. For instance, some research groups reported that the displacement of tightly bound water molecules incurs a penalty in binding free energy [153-155]. It is also important to consider the contribution of hydrated water molecules, which remain and form a bridge of hydrogen bonds to proteins and binding ligands [72,156-158]. In both cases, GIST is capable of capturing such water molecules. However, an accurate modeling of water molecules becomes even more complicated in consideration of various thermodynamics of active-site water. It also needs a large dataset of protein-ligand complexes for further development of the scoring function because different protein binding site affect water differently so that a different result might be obtained for a different protein target. Some scoring functions attempted to cope with these kinds of challenging work, such as the WScore developed by Schrödinger, Inc., New

York City, NY, USA [159]. Notably, while our work is a case study for a single target protein of FXa and further studies would be needed to show that this is a general result, our result supports the applicability of GIST for successful docking campaigns. We also hope that the present results would activate more quantitative studies of molecular docking for drug design.

Chapter 4

Cosolvent-based Molecular Dynamics for Ensemble Docking: Practical Method for Generating Flexible Protein Conformations

4.1 Introduction

Protein-ligand docking is one of the most promising computational tools in the large-scale discovery of compound hits for target macromolecules, which potentially reduces the costs and improves the efficiency of modern high-throughput screening (HTS) for drug design [160]. The docking calculation is applied to rank database compounds for a specific target, and the use of high-quality compound libraries and appropriately constructed docking model can lead hit rate several folds above random [161-163]. Docking-based virtual screening (VS) methods have successfully contributed to the discovery of novel inhibitors of various targets, including HIV-1 integrase [164], human estrogen receptor alpha [165], cytochrome P450 aromatase [166], and many others [167-172]. Despite these successes, docking still has some limitations in its applicability for diverse target proteins. The weakness often comes from the deficient representation of protein flexibility. Traditionally, most of the docking methods only consider ligand flexibility and use a single and rigid structure of target protein for fast calculation. However, since proteins are intrinsically flexible and frequently undergo conformational changes on ligand binding, the static view of protein structure in classical docking is far from

reality. For instance, protein kinases are widely known as a difficult target for docking due to its flexible binding pocket [173]. Moreover, some cross-docking studies have shown that docking ligand to the non-native structure of target protein leads to failure of docking in mode/affinity prediction [174-177]. These results also imply that the use of single protein structure might lead the poor enrichment of VS experiments. To overcome this limitation, recent approaches for improving docking methods have focused on the efficient incorporation of protein flexibility [178-183].

In the last decades, the importance of protein flexibility upon ligand binding has been widely recognized [18,184,185]. The first proposal was the induced fit model proposed by Koshland [186] in 1958, which suggested that ligand binding induces a conformational change of protein. A more recent explanation was given by the conformational selection model, in which a ligand binds to a particular conformation of the unbound protein and stabilizes the potential energy of such a conformation by forming protein-ligand complex [187-189]. It is well understood that the binding process of protein and ligand is not so simple, and the induced fit model and the conformational selection model are not contradictory. For instance, a ligand may bind to a particular conformation of protein fluctuating between several metastable conformations and may induce a small conformational change of protein to stabilize it. A small induced fit effect has been successfully introduced into docking methods by allowing the rearrangement of several protein side chains when the docking calculation is performed [190,191]. However, for some targets, major backbone movements are observed, in which case the full receptor flexibility in docking calculation might be required [192-194]. The ideal approach to incorporating protein flexibility would be to explore the full degrees of freedom of the protein-ligand system, using the computational simulations such as Monte Carlo (MC) or

molecular dynamics (MD). Unfortunately, such a method is computationally expensive and impractical for large-scale docking experiments like a virtual screening [195]. Thus, a simplified model has been presented to incorporate limited protein motions while keeping computational time practical; that is an ensemble docking [19].

Recently, numerous studies have focused on the ensemble docking approach [196-200]. In contrast to explicit modeling of protein flexibility, the ensemble docking makes use of multiple and discrete structures of a target protein. In the standard ensemble docking procedure, each compound is sequentially docked to a set of protein conformers (*i.e.*, ensemble) to find the best-fit protein structure for a particular ligand. Consequently, the flexibility of target protein is implicitly introduced into the docking method. The ensemble docking has been successfully applied to diverse protein targets, for example, nuclear receptors [201], protein kinases [202-204], proteases [205-207], and oxidoreductases [208,209]. A certain advantage is that the ensemble docking is capable of accounting for any scale of protein motions, including large-scale backbone movement, loop activation of protein kinases, and side-chain rearrangement around the bound ligand. Nevertheless, in practical, the coverage of protein flexibility completely depends on the quality of the structural ensemble. In other words, the ensemble docking method never finds a new protein conformation not included in the prepared ensemble. Thus, the critical issue of ensemble docking is how to select and/or generate a high-quality ensemble structures of the target protein.

In early studies of ensemble docking, multiple protein conformations have often been collected from the experimental structures determined by X-ray crystallography or/and nuclear magnetic resonance (NMR), and most of these studies have concluded that the use of multiple protein structures is beneficial in VS experiments [201-205,210-213]. Similarly, some studies

using homology modeling for ensemble construction have also shown a better performance of ensemble docking than the use of single protein structure [214-216]. Since experimentally determined protein structures contain diverse ligand-bound conformations, the methods above are capable of including those distinct protein conformations in the ensemble appropriate for the binding of diverse database ligands. However, it is worth noting that the success of such approaches requires rich experimental structures of the target protein.

On the other hand, a more challenging approach can be provided by the molecular dynamics (MD) simulation to generate multiple protein conformations [217-220]. The use of MD simulation has two certain advantages. First, the MD simulation only needs one experimental structure of the target protein. Hence, it is widely applicable to diverse targets, even though the target structure is few, of low resolution, or even a computationally modeled one. Second, the MD simulation might find a completely new conformation of the target protein superior to the experimental one for the VS study. In fact, some early studies reported that the best snapshot of MD simulation was more predictive than the X-ray or NMR structures [221,222]. However, at the same time, MD snapshots include many poor structures, and it is still difficult to select the promising structure for the VS experiments. Therefore, a rational selection method of protein conformations from the MD trajectory is needed for the successful ensemble docking. Another question is whether the MD simulation with pure water molecules is the best approach to generating druggable protein conformations. As mentioned above, there are two fundamental models of protein flexibility upon ligand binding: induced fit model and conformational selection model. For instance, the use of an apo-protein for the MD simulation might cause conformational changes of the binding pocket and represent the conformational selection model, whereas it could not take into account the induced-fit effects of a specific

ligand binding. In contrast, the use of holoprotein successfully accounts for a specific induced fit model, whereas it might restrict dynamic motions of the target protein. Hence, our interest is focused on those more sophisticated simulation methods to generate the druggable conformations of a target protein. In this study, we present novel ensemble docking procedures by combining the inexpensive conformational selection method and the cosolvent-based molecular dynamics (CMD) simulation.

The selection of multiple protein conformations is an essential process for the success of ensemble docking. Since the MD simulation generates numerous protein conformations, they have to be narrowed down to an appropriate size of ensemble. The use of a large conformational ensemble is tremendously time-consuming and increases the false positive rate of the VS experiment. Applying a clustering algorithm is a general approach to picking up distinct protein conformations from the MD trajectory. However, it has been reported that the clustered protein conformations include not only good structures but also poor structures for VS study [223]. One promising technique for relevantly selecting protein structures involves docking a library containing known actives and a large number of decoys to the target protein [224,225]. The larger the number of actives found among the top hits and the higher the rank of the actives, the better is the receptor structure for virtual screening. However, this method could be computationally expensive when this analysis is repeated to the numerous snapshots of MD simulation. A more practical method has been proposed by Huang and Wong, called screening performance index (SPI) [226]. They showed SPI is capable of selecting a good experimental structure for the VS experiments through the docking of a set of known actives only. We introduce a little modification to SPI to be stable for the selection of MD snapshots and apply it to our study.

In addition to the conformational selection method, we propose the use of cosolvent-based molecular dynamics (CMD) simulation for the generation of druggable protein conformations. Cosolvent-based MD, so-called mixed-solvent MD or ligand-mapping MD, is a simple but a highly attracting computational method which uses water and organic probe molecules for the solvent when performing the MD simulation of a protein target. The first cosolvent-based simulation method was reported by Barril and co-workers in 2009, called MDmix [227]. Following this study, some similar methods have been proposed to date, such as SILCS [228], MixMD [229], pMD [230], and many others [231-234]. Various probe molecules, resembling certain chemical moieties found in drug-like ligands, have been used for mapping the protein surface, finding the binding hotspots, and identifying the pharmacophore feature of hotspots where high-affinity ligands are attainable. At the same time, the CMD simulation often brings about the probe-induced conformational changes of a target protein. In fact, it has been reported that the cosolvent simulations are capable of finding the allosteric sites or new binding hotspots which the standard MD could not identify [235-237]. More recently, Yang and co-workers have applied the CMD to the complex system of Bcl-xL and showed that the use of conformations obtained from the CMD improved the docking performance for the known ligands of Bcl-xL [21]. Based on these impressive studies, we attempt to incorporate the CMD simulation into the ensemble docking, expecting the CMD generates more druggable conformations of a target protein than the experimental structures and enhances the enrichment of the VS study.

In the present study, CMD was performed for six diverse protein targets using three different probe molecules to evaluate the applicability of the CMD-based ensemble docking. We then used apo forms of protein structures for the input of MD simulations so that the difference of conformational changes could be clearly observed. The present method was

validated by the VS performance using diverse active and decoy compounds taken from DEKOIS 2.0 library [238] and compared with the single structure docking to the X-ray structures of the apo and holo form proteins. The present method was also assessed in comparison with the standard MD simulation (i.e., only using water and ions for solvents) based ensemble docking. The results have revealed that the present method is capable of identifying more diverse active ligands than the previous methods, and is widely applicable for the diverse protein targets.

4.2 Material and Methods

4.2.1 Target Protein and Structure Preparation

Six diverse target proteins taken from DEKOIS 2.0 were used for the ensemble docking studies, which include progesterone receptor (PR), cyclin-dependent kinase 2 (CDK2), NAD-dependent protein deacetylase sirtuin-2 (SIRT2), human immunodeficiency virus-1 protease (HIV1PR), thymidine phosphorylase (TP), and epidermal growth factor receptor (EGFR). These six proteins have been shown to be difficult targets for the virtual screening in the original DEKOIS 2.0 study [238]. We selected each one of apo (or apo-like) and holo form (inhibitor bound) structures for the six target proteins from the protein data bank (PDB). The apoproteins were used for the MD simulations and the ensemble docking studies, whereas the holoproteins were only used in the single structure docking for the comparison of the ensemble docking performances. The six target proteins and the selected twelve structures are summarized in Table 4.1. Note that since an apo form structure of PR was not available in PDB, we selected progesterone (hormone) bound structures for PR (PDB-ID: 1A28). For the descriptive purposes, we call this structure “apo” in this paper.

Table 4.1 X-ray structures for the six target proteins used in this study.

Target protein	Protein structure (PDB ID)	
	Apo	Holo
Progesterone receptor (PR)	1A28 ^a	2W8Y ^b
Cyclin-dependent kinase 2 (CDK2)	1HCL	1CKP ^b
NAD-dependent protein deacetylase sirtuin-2 (SIRT2)	1J8F ^b	5D7P
Human immunodeficiency virus-1 protease (HIV1PR)	2PC0	3NU3 ^b
Thymidine phosphorylase (TP)	2WK5	1UOU ^b
Epidermal growth factor receptor (EGFR)	5EDP	1M17 ^b

^a Progesterone (hormone) is bound in the binding pocket. ^b Structures used in the original DEKOIS 2.0 study.

In the present study, all protein structures were prepared by using molecular operating environment (MOE) [114]. First, all the twelve structures were prepared for the docking calculation. All water molecules, ions, and, bound ligands were removed from the systems (including the progesterone), and hydrogens were added by using Protonate3D [239] in MOE. Next, the six apo form structures were prepared for the MD simulation as follows: (I) compensating the missing atoms and residues; (II) modeling the missing loops; (III) fixing engineered mutations. Following the structure preparations, the systems of six apoproteins were set up for the MD simulation.

4.2.2 Choice of Cosolvents and System Set-up for MD

In addition to the standard MD, we tested three different CMDs with isopropanol, benzene, and purine. We selected these probe molecules based on the size and their chemical features. The isopropanol is the most widely used probe molecule for the cosolvent simulations, which is miscible in the water but capable of interacting with the hydrophobic surfaces of proteins. In

contrast, the benzene is an insoluble molecule in water, whereas its aromatic interaction is essential for the molecular recognition between the protein and the drug-like ligand. The purine is also an aromatic probe but a little bit larger than benzene, and capable of forming hydrogen bonds. While the purine has rarely been used for the cosolvent simulation, it is a representative moiety often found in the biomolecules or the approved drugs. In this study, the probe-water concentrations were set to $\sim 0.25\text{M}$ for all simulations, since some previous studies concluded that a low concentration resulted in a clear occupancy of probe molecules in the binding hotspots [240-242].

Consequently, we tested the total 24 systems (the four different solvents for the six protein systems) for the MD simulation. All the systems were prepared with the identical procedure. First, the systems were solvated using Packmol [243]. The target proteins were randomly shelled with the probe molecules and solvated in the cubic box with water molecules, and the minimal number of Na or Cl ions were then added to electrically neutralize the systems. Next, the force field parameters were assigned to the solvated systems using the Tleap program from AmberTools 16 [244]. We used AMBER14SB force field [245] for the proteins, TIP3P water model [135] for the solvent water molecules, and the generalized amber force field (GAFF) [246] for the probe molecules. The partial atomic charges of the probe molecules were then calculated by the restrained electrostatic potential (RESP) method, using quantum-mechanically derived electrostatic potentials at the Hartree-Fock level with the 6-31G* basis set. At last, the periodic boundary condition was set to the cubic box. As an example, Figure 4.1 shows the prepared system and the final system of cosolvent simulation for PR and benzene probes. Throughout the simulation, most of probe molecules spread into water solvent, whereas some probes concentrate on a particular surface of the protein target.

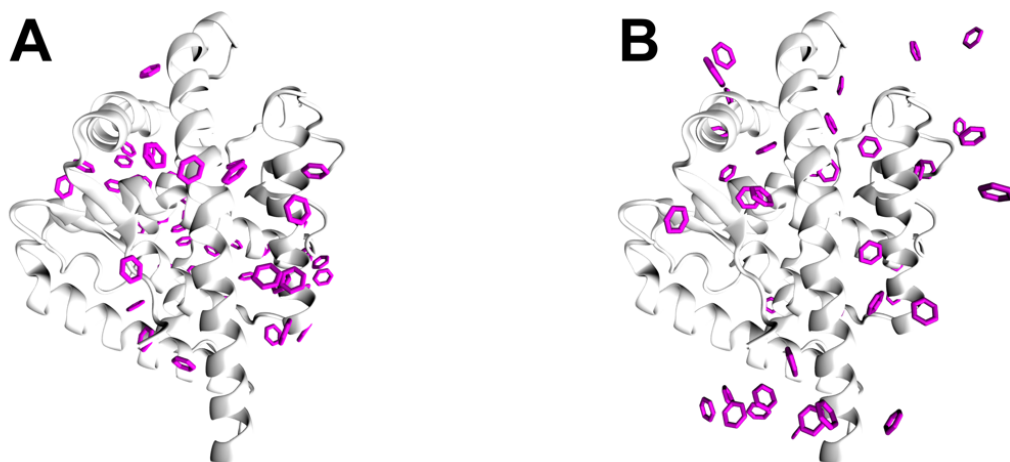


Figure 4.1 Snapshots of cosolvent simulation for PR (white ribbon) and benzene probes (magenta sticks). (A) Initial structure of MD simulation. (B) Final structure (after 50 ns product run) of MD simulation. Water and ions are not displayed.

4.2.3 Molecular Dynamics Protocols

To generate multiple protein conformations, the MD simulations of the prepared systems were performed with the identical protocols. Following procedures were carried out with the Amber 14 software using pmemd.cuda [136]. First, the system energies were minimized in two steps: (I) Only solvents and protein hydrogens while restraining protein heavy atoms; (II) The whole system without any restraint. Both minimization steps used 1500 cycles of the steepest descent algorithm followed by the conjugate gradient method for the maximum of 20000 cycles, and the restraint was harmonic with force constant of 100 kcal/mol/Å. Next, the system was heated in three steps: (I) for 50 ps from 0 K to 50 K in the NPT ensemble while tightly restraining protein heavy atoms with force constant of 100 kcal/mol/Å; (II) for 150 ps from 50 K to 150 K in the NPT ensemble while weakly restraining protein heavy atoms with force

constant of 10 kcal/mol/Å; (III) for 200 ps from 150 K to 300 K in the NPT ensemble. The first two steps were performed to relax the solvents, in particular for probe molecules. The system was then equilibrated for 2 ns at 300 K in the NPT ensemble. At the final volume, the system was equilibrated again for 2 ns at 300 K in the NVT ensemble. The final production MD run of 50 ns was performed in the NVT ensemble, and snapshots of this simulation were saved every 1 ps, for a total 50,000 frames of snapshots stored. During the MD simulations, a time step of 2 fs was employed with the SHAKE algorithm [137], a temperature was regulated by Langevin thermostat, the nonbonded interactions were truncated at 9 Å, and the particle mesh Ewald method was employed to account for the long-range electrostatic interaction [138]. As a result, we performed totally 1.2 μs production MD run (50ns for the 24 systems). On the post-processing, all water, probes, and ions were stripped from the trajectory, and each frame of protein snapshot was aligned on the initial structure to remove translational and orientational movement of the entire protein using the Cpptraj program from AmberTools 16.

4.2.4 Selection of Conformational Ensemble from MD Trajectory

We here present an ensemble selection procedure by combining the rough clustering and the inexpensive structure ranking method. First, we performed the structure clustering to the MD trajectory. The purpose of this procedure was the reduction of MD snapshots with maintaining the structural diversity of a target protein. Since we focused on the dynamics of the binding pocket, we only used atoms around the binding pocket for the clustering. The positions of binding pockets were selected according to the bound ligands of holoproteins used in this study. The binding pocket atoms were then selected using the fpocket program [247], which selected the atoms contacting to the alpha spheres. The alpha sphere was used in the binding

pocket detection by the fpocket algorithm, which is defined as a sphere that contacts four atoms on its boundary and contains no internal protein atoms. The selected pocket atoms for each target are summarized in Appendix C (Table C1). Using these pocket atoms, we applied k-means clustering algorithm to the snapshots of MD simulation using the Cpptraj from AmberTools 16 and selected centroids from each cluster for the candidates of the conformational ensemble. We then used the RMSD-based pairwise distance and set the cluster number k to 500. Throughout this process, 50,000 frames of MD snapshots were narrowed down to the 500 specific protein conformations.

Before the ensemble selection, we introduced a structure selection measurement. An efficient conformational selection method, screening performance index (SPI), was proposed by Huang and Wong [226] in 2015. They showed that SPI is capable of selecting a good structure for VS experiments through the docking of a few known active compounds. The rationale of SPI is simple: If many actives can dock to a protein structure with docking energies more favorable than the overall average docking energies to all protein structures, such structure might be more likely to pick out many actives in virtual screening. On the basis of this rationale, SPI is formulated as:

$$\text{SPI}_j = \frac{\sum_{i=1}^n x_i}{l}, \quad x_i = \begin{cases} 1 & \text{if } E_i \leq \bar{E} \\ 0 & \text{otherwise} \end{cases}. \quad (4.1)$$

Here, $i \in [1, 2, \dots, n]$ and $j \in [1, 2, \dots, m]$ represent the indices of an active and a protein structure, respectively. Terms l , n , and m are the total number of actives, the actives successfully docked to a specific protein structure, and the total number of protein structures, respectively. \bar{E} is the overall averaged docking energies across all actives and all protein

structures. The range of SPI is [0,1], and a good structure shows the high value. In the previous study, SPI was applied to the several X-ray structures of eight target proteins, and its effectivity was validated [226]. However, we found that SPI was not predictive for the large number of protein structures generated by the MD simulation, in particular for the high-rank region near the SPI value of 1. In other words, it means that many protein structures easily reach the SPI value of 1. Hence, we introduced a slight modification into SPI, called ranking-based SPI (RSPI). The idea is simple: The more actives dock to a protein structure with lower docking energies relative to the other structures, the more favorably such structure picks out many actives in virtual screening. RSPI is described as:

$$\text{RSPI}_j = 1 - \frac{\sum_{i=1}^l r_{ji}}{l \times m}, \quad (4.2)$$

where i , j , l , and m are the same as SPI; r_{ji} is the rank of the i th active docked to j th protein structure in the descending ordered list of docking energies of i th active against all protein structures m . RSPI is highly correlated to SPI, but more distinguishable for high ranked structures.

The clustered 500 protein structures were then ranked for the ensemble construction using RSPI. We used 40 known actives from DEKOIS 2.0 for each one of the six target proteins, and the docking was performed using AutoDock Vina [248]. The compounds and docking protocols used in this work are described in the following sections. Eventually, we selected top ten protein structures from each MD trajectory based on RSPI for the use of ensemble docking studies.

4.2.5 Dataset for Virtual Screening Experiment

DEKOIS 2.0 [238] was used in this work, which is a useful benchmarking dataset for the

evaluation of VS performance through the docking. In DEKOIS 2.0, the benchmarking set was constructed using an original protocol for the selection of both actives and decoys. For each target, the DEKOIS 2.0 includes 40 known actives and 1200 decoys. The active ligands were retrieved from the BindingDB [249] using several filters. The decoy sets were selected from the 15 million ZINC [250] compounds to be of similar physiochemical properties but structurally dissimilar to the actives. It is worth noting that decoy sets of DEKOIS 2.0 are not true inactive compounds and may adversely affect the evaluation of the VS performance [62].

4.2.6 Docking Protocols

All docking calculations were carried out using AutoDock Vina [248]. The totally $6 \times 1240 = 7440$ dataset compounds were prepared for the docking study using the `prepare_ligand4.py` program of AutoDockTools [251]. The preparation procedure of protein targets was mentioned above. To define the search volume, all the protein structures were aligned to the holo form of the same target protein, and then a cubic box of $22.5 \text{ \AA} \times 22.5 \text{ \AA} \times 22.5 \text{ \AA}$ was placed around the center of the co-crystallized ligands. Default settings were used for all docking calculations, and the highest score (*i.e.*, the lowest energy) was selected from each docking run and used for the compound ranking.

4.2.7 Ensemble Docking and Scoring

In ensemble docking protocol, each compound is sequentially docked to a set of protein conformers (ensemble), resulting in multiple docking scores obtained depending on the number of protein structures. Hence, a method to determine the single scoring value of a given compound is needed for ensemble docking. Several different methods for combining multiple

docking scores into a single docking score have been suggested. Reported protocols include selecting the best score across all structures [199,208], creating composite grids of all ensemble members [19,252], and using different weighted averages which include arithmetic [253] and Boltzmann weighted averages [254] as well as averages using weights determined by knowledge-based methods [196]. In this work, we used two simple approaches for the ensemble scoring: (I) Minimum scoring method, which adopts the best scoring function value (*i.e.*, minimum energy) across all ensemble members; (II) Average scoring method, which calculates the arithmetic mean of docking scores across all ensemble members. These two ensemble scoring methods are compared in Results and Discussion section.

4.2.8 Enrichment Measurements

Early enrichment is an essential measurement of the VS performance. In the structure-based virtual screening, a large number of compounds in a database are sequentially docked to a target protein and ranked by its docking score. Usually, only top few percent compounds are selected from the rank ordered list of the large compound database for more rigorous evaluation of *in vitro* or *in vivo* experiments. Hence, a metric to measure how many true actives are included on the top ranked list is suitable for the evaluation of VS methods. In this study, we used the Boltzmann-enhanced discrimination receiver operating characteristic (BEDROC) [255] for the statistical measurement of screening efficiency. BEDROC is regarded as one of the most useful metrics for gauging the performance of screening models, in particular for the measurement of early recognition problem. The metric is given by

$$\text{BEDROC} = \frac{\sum_{i=1}^n e^{-\alpha r_i/N}}{\frac{n}{N} \left(\frac{1 - e^{-\alpha}}{e^{\alpha/N} - 1} \right)} \frac{R_a e^{\alpha R_a} (e^{\alpha} - 1)}{(e^{\alpha} - e^{\alpha R_a})(e^{\alpha R_a} - 1)} + \frac{1}{1 - e^{\alpha(1-R_a)}} , \quad (4.3)$$

where R_a is the ratio of the total number of actives n to the total number of database compounds N , and r_i is the relative rank of the i th active in the rank ordered list. BEDROC gives the probability that an active is ranked ahead of a compound randomly selected from a hypothetical exponential probability distribution function with parameter α . It is bound by the interval $[0,1]$, with 1 reflecting the best possible screening performance. In this work, we select the most widely used value, $\alpha = 20.0$, which corresponds to the top 8% of the relative rank accounting for 80% of the BEDROC score.

4.3 Results and Discussion

4.3.1 Virtual Screening Performances

Using the snapshots of individual MD runs, conformational ensembles were selected based on RSPI score. In this work, we constructed an ensemble with top ten structures of RSPI score for each simulation. The performances of docking methods were assessed by the VS experiments using 40 actives and 1200 decoys of the DEKOIS 2.0 library for each protein target. The ensemble dockings were then performed for the 24 ensemble systems (the six protein targets for the three different CMDs and the standard MD) and compared with single structure docking to the X-ray structure of apo and holo form proteins. The BEDROC values of VS experiments are provided in Table 4.2. The results revealed that, compared to virtual screening using the single X-ray structure, the use of the CMD-based ensembles resulted in significant improvement of early enrichments. The results also showed the superior performance by using

the cosolvent simulation to the standard water simulation.

Table 4.2 BEDROC values of virtual screening experiments for the six target proteins. ^a

Protein structure	PR	CDK2	SIRT2	HIV1PR	TP	EGFR	Average
X-ray structure (single docking)							
Apoprotein	0.104	0.018	0.125	0.223	0.116	0.180	0.128
Holoprotein	0.153	0.088	0.131	0.229	0.083	0.153	0.139
Minimum scoring (ensemble docking)							
Pure water MD	0.083	0.028	0.138	0.195	0.068	0.068	0.097
Isopropanol probe MD	0.137	0.027	0.193	0.134	0.155	0.165	0.136
Benzene probe MD	0.255	0.030	0.197	0.233	0.200	0.233	0.191
Purine probe MD	0.247	<u>0.055</u>	0.208	0.285	0.186	0.187	0.195
Average scoring (ensemble docking)							
Pure water MD	0.127	0.020	0.195	0.240	0.052	0.052	0.113
Isopropanol probe MD	0.097	0.028	0.195	0.326	0.157	0.194	0.156
Benzene probe MD	0.186	0.015	0.164	0.281	0.216	0.223	0.180
Purine probe MD	0.231	<u>0.051</u>	0.229	0.329	0.013	0.191	0.170
Best snapshot ^b (single docking)							
Pure water MD	0.179	0.073	0.173	0.352	0.198	0.234	0.202
Isopropanol probe MD	0.156	<u>0.094</u>	0.219	0.393	0.189	0.229	0.213
Benzene probe MD	0.217	0.051	0.209	0.392	0.269	0.263	0.234
Purine probe MD	0.276	0.092	0.202	0.389	0.094	0.209	0.210

^a **Bold** represent superior BEDROC values to X-ray structures. Underlines represent the best BEDROC values among the four different MDs in the same scoring protocols. ^b Best structure represents the best result of single structure docking among the ten ensemble structures obtained by different probe simulation.

First, we discuss the result of the minimum scoring method. The log-scaled receiver operating characteristic (ROC) plots for the minimum scoring results are shown in Figure 4.2. The significant improvements were found especially in the benzene probe CMD and purine probe CMD. These two probe-based ensembles successfully improved BEDROC values of five of all the six protein targets, PR, SIRT2, HIV1PR, TP, and EGFR. For example, the ensemble

of benzene probe CMD improved BEDROC value (0.255) for PR, which is approximately 2.5-fold higher than that of single apo form X-ray structure (BEDROC = 0.104). Similarly, the ensemble of purine probe CMD resulted in BEDROC value 0.247 for PR. Next to the PR, the early enrichments of SIRT2 and TP were clearly improved by utilizing the cosolvent simulation for the ensemble docking. On the other hand, the ensemble docking of isopropanol probe based CMD only improved the VS performance only for two targets, SIRT2 and TP, and interestingly, its BEDROC values were worse than those of benzene and purine probes across all the six targets.

Another considerable finding was that the pure water MD-based ensemble docking lowered the BEDROC value than the single docking of X-ray structure, only except for SIRT2. Introducing protein flexibility by ensemble docking may enable accurate prediction of native binding poses and quantification of ligand binding affinities [178,181]. However, it also results in an increased number of false positives in the VS study [208]. In other words, the latter can result in a poor enrichment performance compared to using a single static structure for the VS study. The ensemble docking results of pure water MD indicated such increases of false positive rates. In fact, the best single docking result of ensemble structures was always better than the ensemble docking result in the pure water MD systems (see Table 4.2). Similarly, in other ensembles produced by CMDs, the single docking result of the best snapshot in the ensemble structures tended to be better than the ensemble docking results. Nevertheless, by combining the cosolvent simulation, our results suggested that the CMD-based ensemble docking is capable of resulting in comparable performance to the best single docking (see Figure C3-C8 in Appendix C). Furthermore, in the case of benzene probe simulation for PR, the ensemble docking led to the improvement of screening performance compared to every single docking

Cosolvent-based Molecular Dynamics for Ensemble Docking: Practical Method for Generating Flexible Protein Conformations

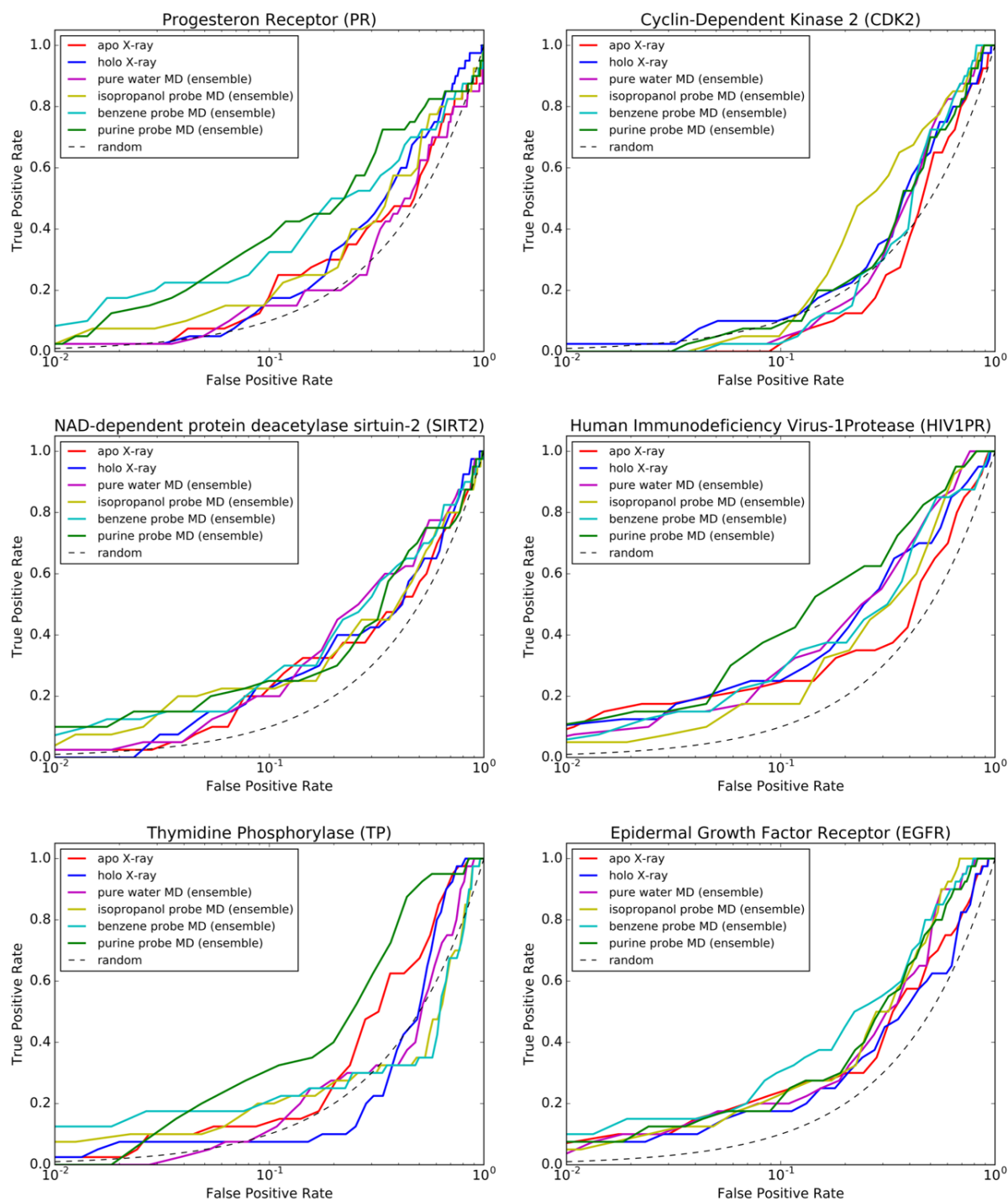


Figure 4.2 Log-scaled receiver operating characteristic (ROC) plots with the minimum scoring method of ensemble docking for the six targets. Each line represents the following: apo X-ray (red), holo X-ray (blue), pure water MD (magenta), isopropanol probe MD (yellow), benzene probe MD (cyan), purine probe MD (green), and theoretical result of random selection (black dots).

(Figure C3 in Appendix C). The similar results were found in the case of purine probe simulation for TP (Figure C7 in Appendix C). Surprisingly, in the case of purine probe simulation for TP, the minimum score of ensemble docking improved the BEDROC value to 0.186, which is approximately twice higher than the best single docking (BEDROC = 0.094). These two results were typical cases successfully introducing protein flexibility into protein-ligand docking.

Regarding the average scoring method, the results were not significantly different from the minimum scoring method. The log-scaled ROC plots for the average scoring results are shown in Figure 4.3. For the pure water MD and isopropanol probe CMD, the BEDROC values of the average scoring results were slightly better than those of the minimum scoring results. In contrast, for the benzene probe CMD and purine probe CMD, the average scoring method resulted in lower BEDROC values than those of the minimum scoring. Significant improvements were found in the case of HIVPR, that the average scoring method showed clear improvements of the BEDROC value compared to the minimum scoring method across all the simulation methods. In particular, the average score of isopropanol probe based ensemble achieved the BEDROC value of 0.326 for HIV1PR, which is over 2-fold higher than the that of the minimum score (BEDROC = 0.134). On the contrary, in the purine probe simulation, the BEDROC value of average scoring for TP was severely got worse (0.031). This result suggested that the multiple conformations of TP generated by the purine probe simulation might be substantially different from each other and highly selective for the typical ligands. Owing to such reasons, an active might be successfully docked to a particular conformation of the ensemble but could not fit other conformations, resulting in the worse average score and the good minimum score of the ensemble docking (Figure C7 in Appendix C). It is speculated that

Cosolvent-based Molecular Dynamics for Ensemble Docking: Practical Method for Generating Flexible Protein Conformations

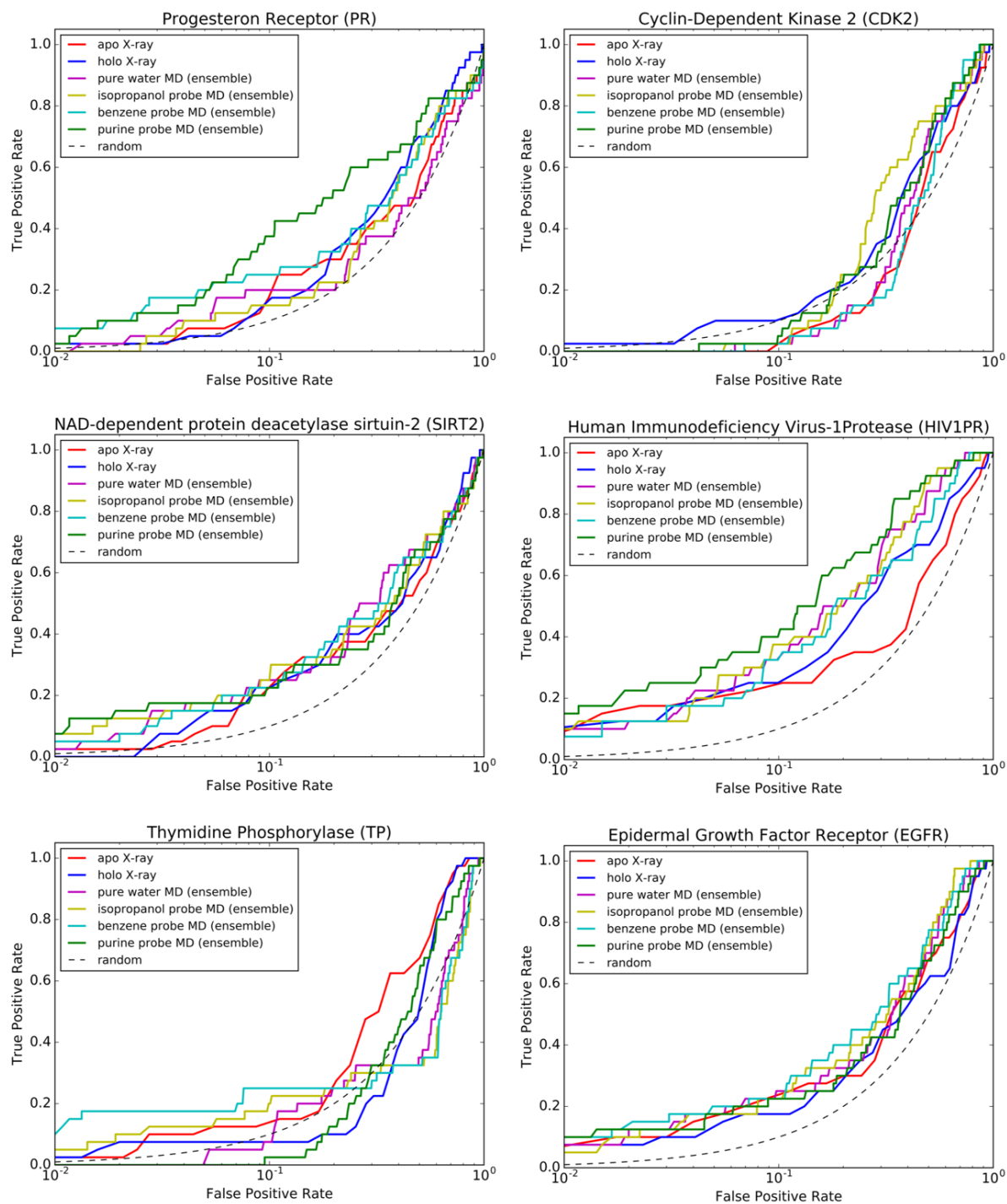


Figure 4.3 Log-scaled receiver operating characteristic (ROC) plots with the average scoring method of ensemble docking for the six targets. Each line represents the following: apo X-ray (red), holo X-ray (blue), pure water MD (magenta), isopropanol probe MD (yellow), benzene probe MD (cyan), purine probe MD (green), and theoretical result of random selection (black dots).

the contrastive situation, good average score and worse minimum score, was found in the ensemble docking to HIV1PR, in particular for the isopropanol probe CMD (Figure C6 in Appendix C).

To sum up, comparing the two ensemble scoring methods, the minimum scoring method was slightly better than the average scoring method for the benzene probe CMD and purine probe CMD. In contrast, for the pure water MD and the isopropanol probe CMD, the average scoring method was superior to the minimum scoring method. The difference of two scoring method did not seem to depend on targets or probe molecules. It still needs further studies to find the consistent rule for selecting ensemble scoring methods.

4.3.2 Binding Pocket Conformation and Probe Concentration

For further analysis of the effect of the probe molecules, the principal component analysis (PCA) was performed to all the ensemble structures. PCA is a valuable tool for comparing conformations obtained through the MD simulations to the experimental structures. In this study, PCA was carried out on the Cartesian coordinates of binding pocket atoms, using the ptraj module in AmberTools 16. The resulting projections along the first two principal components (PC1 and PC2) are plotted in Figure 4.4. As expected, PCA results clearly showed distinct distributions of binding pocket conformations in response to the different probe molecules except for HIV1PR. This result also suggested that the probe molecules induced conformational changes of binding pockets.

Particular probe bindings have been found in several snapshots of MD trajectories. For instance, Figure 4.5 shows the benzene and purine probes concentrating to the binding pocket of the PR system. Figure 4.5A shows a snapshot of the benzene probe simulation at around 46

Cosolvent-based Molecular Dynamics for Ensemble Docking: Practical Method for Generating Flexible Protein Conformations

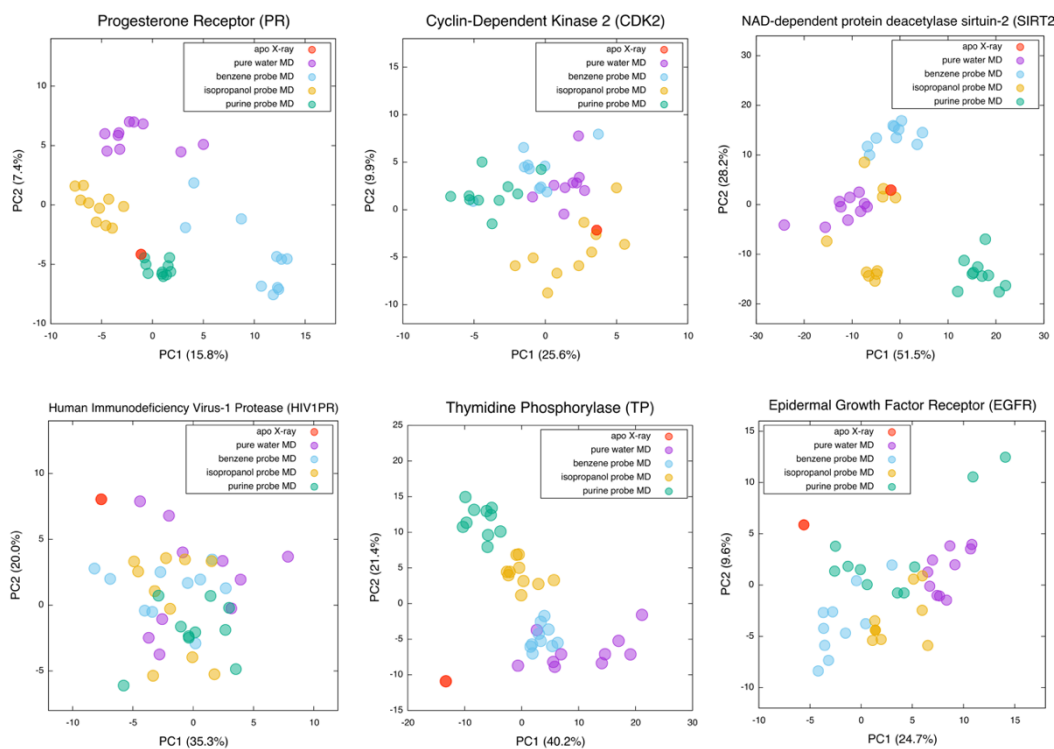


Figure 4.4 Principal component analysis (PCA) of binding pocket atoms for top 10 structures of RSPI obtained from each MD run. Each color circle represents the following: apo form X-ray structure (red), pure water MD (magenta), isopropanol probe MD (yellow), benzene probe MD (cyan), and purine probe MD (green).

ns, which has the best BEDROC value of 0.217 of the VS experiments in the benzene systems. Through the MD simulation, three benzene probes were bound to the binding pocket of PR. Interestingly, the bound structures of these benzene probes overlapped well the ring moieties of the co-crystallized ligand of PR (PDB ID: 2W8Y). The result suggested that the benzene probes reproduced a part of essential interactions between protein and ligand, and such probe-induced conformational change was beneficial for the docking of diverse ligand. The similar probe concentrations were found in the purine probe system of PR. The purine probes were bound to the binding pocket during the equilibrium phase of the MD simulation and stayed in the binding pocket during 50 ns of the production run. Figure 4.5B shows a snapshot of this simulation at

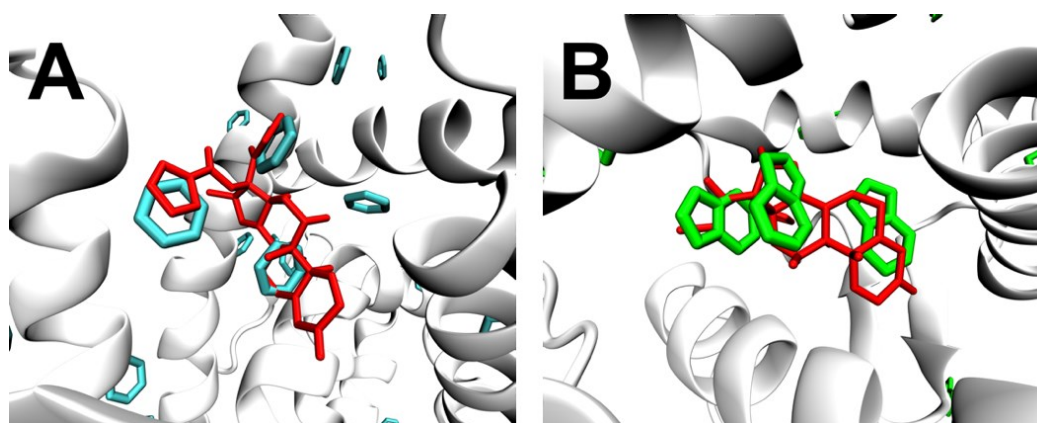


Figure 4.5 Concentrations of probe molecules in the binding pocket of progesterone receptor (white ribbon) and superposed ligand of PDB entry 2W8Y (red sticks). (A) Snapshot of cosolvent molecular dynamics simulation with benzene probe (cyan sticks). (B) Snapshot of cosolvent molecular dynamics simulation with purine probe (green sticks).

around 25 ns, which has the best BEDROC value of 0.276 of the VS study in the purine system. The number of purine probe in the binding pocket was the same as that of the benzene system, but the bound structures were different. The purine probes successfully overlapped the two-membered ring moieties of the same ligand. Such kinds of probe concentrations were similarly found in other systems, and it might affect the binding pocket conformations of protein targets.

4.3.3 Protein Motion of Cosolvent-based Molecular Dynamics

Although the CMD simulation might cause the protein unfolding or aggregation of probe molecules [232], all the MD runs performed in this work have been finished without such difficulties. We have checked protein dynamics of each MD run in terms of the root mean square deviation (RMSD) plots of backbone C α atoms and binding pocket atoms with the lapse of time (see Figure C1, C2 in Appendix C). Contrary to our expectations, there has not been any consistent change of protein dynamics according to the difference of probe molecules. The motions of backbone and binding pocket atoms have also not been correlated for all cases.

However, depending on the combination of protein and probes, the results have clearly shown that the different probe molecules induce distinct protein motions. For example, purine probe significantly stabilized the binding pocket of PR (average RMSD was 1.02 Å), whereas it destabilized the binding pocket movement of SIRT2 compared with other probe molecules. Interestingly, in the cases of HIV1PR and TP, the probe molecules suppressed the backbone movement of proteins in comparison to the standard MD without probes, even though we expected that the hydrophobic probes enhance the protein dynamics. We then found that some probe molecules concentrate into specific sites of the protein surface. Accordingly, these results have suggested that such probe binding might stabilize the whole dynamics of protein target. In fact, the similar mechanism has been reported that the cosolvent of water and glycerol enhances the protein stability [256].

4.3.4 Ensemble Selection from MD trajectory by RSPI

Although the CMD-based ensemble docking has showed the superior VS performances to the use of single X-ray structure, it should be noted that the structure selection method presented here still has a room for improvement, and further discussion is needed. The RSPI method successfully selects many useful structures from the MD snapshots, whereas it also includes some poor structures which enhance the false positive rate of the VS experiments. In fact, there were no correlations between the BEDROC values and the RSPI scores in the selected ten structures across all cases (Table 4.3). It might be concluded from the results that our assumption for the RSPI score is not correct. It also indicated the fact that; if many actives dock to a protein structure with lower docking energies relative to the other structures, such structure not always distinguish actives and decoys in virtual screening. However, at the same time, it

means that there remains a possibility for further improvement of the CMD-based ensemble docking. For instance, Xu and Lill [224] reported that the use of a small subset of actives and decoys is capable of selecting appropriate protein structures for the VS study. They also suggested that a very small number (3-5) of protein structures can perform good ensemble docking with the feasible training process.

Table 4.3. Correlations (R^2) between the RSPI scores and BEDROC values among top ten protein structures of RSPI score obtained from each MD run. ^a

Target	Pure water	Benzene probe	Isopropanol probe	Purine probe
PR	0.025 (1)	0.001 (4)	0.036 (1)	0.205 (8)
CDK2	0.003 (0)	0.029 (0)	0.444 (1)	0.018 (1)
SIRT2	0.362 (6)	0.000 (7)	0.229 (10)	0.188 (8)
HIV1PR	0.017 (5)	0.051 (2)	0.162 (3)	0.066 (8)
EGFR	0.052 (3)	0.001 (5)	0.025 (2)	0.005 (6)
TP	0.162 (2)	0.001 (9)	0.141 (6)	0.001 (0)

^a The values in the parentheses represent the number of structures which show better BEDROC values than that of the X-ray structures, among the ten ensemble structures used in the VS experiments.

4.4 Conclusions

In this study, we have presented a novel ensemble docking strategy by combining the inexpensive conformational selection method and the cosolvent-based MD simulation. The present method has been evaluated using the six diverse protein targets with the three different probe molecules. The ensemble docking results revealed that multiple protein conformations produced by the CMD simulations are surely suitable to be used in the VS studies. Moreover, the PCA of binding pocket atoms has shown that different probe molecules induce different binding pocket movements, and such a difference significantly affects the VS performance. The

results could lead to a conclusion that the use of the CMD simulation is more beneficial than standard MD with pure water. It was also more predictive than the single structure docking with X-ray structures of both apo and holo form proteins. In almost all cases, using an ensemble of proteins performs better than the average of a single protein structure result. Furthermore, in some cases, such as PR and SIRT2, the use of ensemble conformations outperformed the best or almost the same as the best BEDROC values among all the single structure dockings. This result also indicated that the present method appropriately introduces the essential protein conformational changes into the protein-ligand docking. However, the choice of probe molecule is still a delicate issue. In this work, we tested three probe molecules of isopropanol, benzene, and purine for the CMD simulations. The ensemble docking results showed that the use of different probe molecules significantly affects the VS performance. Although the use of the benzene and purine probes seems good in this study, more sophisticated selection of probe molecules may further improve the VS performance. We suggest the utilization of a small moiety of a high-affinity ligand or a small fragment hit for the advance. On the other hand, the simple ensemble selection method has been far from satisfactory. We presented the RSPI method for the structure selection of the ensemble docking. The present method successfully selects many useful structures from the MD snapshots, whereas it also includes some poor structures which enhance the false positive rate of the VS experiments. For further improvement of the MD-based ensemble docking method, an advanced structure selection method is strongly needed. Although there remain several challenges to brush up the cosolvent simulation to a practical tool for the structure-based VS study, we believe that the present study would contribute to the future drug design.

Chapter 5

General Conclusions

The great goals of protein-ligand docking are accurately predicting the bonding pose of ligand, correctly estimating binding strength of drug candidate, and ranking the huge database compounds for promising “hit” discovery. Although the protein-ligand docking is a powerful tool for the rational drug design, it is still far away from ideal performance. The challenges remain especially with efficient search algorithm, desolvation energy of scoring function, and protein flexibility upon ligand binding. In this thesis, the three strategies have been presented for the further improvement of protein-ligand docking. In chapter 2, a swarm-based optimization algorithm, fitness learning-based artificial bee colony with proximity stimuli (F/ABCps) have been introduced to the docking program, finding that F/ABCps significantly improve the accuracy of pose prediction in particular for the highly flexible ligands with many optimization parameters. In chapter 3, the thermodynamics of active-site water molecules have been incorporated into the scoring function of the docking program. The computational experiments have revealed that the incorporation of water thermodynamics is substantially beneficial for the pose prediction and the affinity estimation, further the success of virtual screening. In chapter 4, novel ensemble docking method have been presented by combining the cosolvent-based molecular dynamics simulation, aiming to the practical incorporation of the protein flexibility into protein-ligand docking. The simulation results have been revealed that

the present method is capable of generating diverse protein conformations and identifying many active ligands than the previous methods. The methods presented here are not definitive and still on the way to the ultimate goal of protein-ligand docking. However, I believe the finding and results of this thesis would encourage and support further development of protein-ligand docking methodologies.

Appendix A

Table A1 Pseudocode of F/ABCps algorithm.

Initialization:

Set the parameters SN , $limit$ and MCN .
 Generate initial food sources $\{\theta^0\}$ with random numbers.
 Evaluate the fitness value of all food sources according to the Equation (2.2).
 Set the value of trial counter t_i to 0 for all food sources.

While $C < MCN$

Increment the current cycle number C by 1.
 Update the value of q by Equation (2.6).

//Calculate fitness learning sources.

Sort the food sources $\{\theta^C\}$ order in the fitness values in descending.
For $i=1$ to SN
 Initialize the fitness learning source θ_i^F .
 For $j=1$ to D
 Chose different indicators $r1$ and $r2$ randomly in the top range $[1, \text{ceil}(q/100)SN]$.
 Set the value of $\theta_{i,j}^F$ by using eqn Equation (2.5).
 End for
End for

//Employed bee phase

For $i=1$ to SN
 Simulate the perturbation parameters J^* by Equation (2.7).
 Obtain the k -nearest neighbor θ_k^{NC} with respect to food source θ_i^C .
 Search a new food source v_i^C by using Equation (2.8).
 Evaluate the fitness value of v_i^C according to Equation (2.2).
 Perform the greedy selection between θ_i^C and v_i^C according to their fitness values.
 Update the trial counter t_i .
End for

//Onlooker bee phase

Obtain p according to Equation (2.4).
For $i=1$ to SN
 Obtain p_i^w by Equation (2.10) using N^i and F .
End for
 Set $i=1$ and $l=0$. *//Then l corresponds to the index of onlooker bees.*
While $l < SN$
 If $p_i < p_i^w$
 Increment the value of l by 1.
 Simulate the perturbation parameters J^* by Equation (2.7).
 Obtain the k -nearest neighbor θ_k^{NC} with respect to food source θ_i^C .
 Search a new food source v_i^C by using Equation (2.8).
 Evaluate the fitness value of v_i^C according to Equation (2.2).
 Perform the greedy selection between θ_i^C and v_i^C according to their fitness values.
 Update the trial counter t_i .
 End if
 Increment the value of i by 1.
 If $i > SN$ *//Food site selection is repeated till all onlooker bees have been allocated.*
 $i=1$.
 End if
End while

//Scout bee phase

For $i=1$ to SN
 If $t_i \geq limit$
 Reinitialize the food source θ_i^C with random number.
 Evaluate the fitness value of θ_i^C according to Equation (2.2).
 Reset the trial counter t_i to 0.
 End if
End for

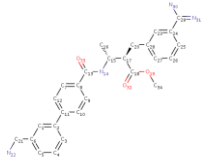
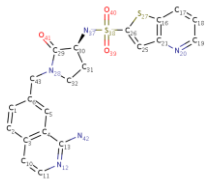
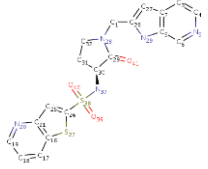
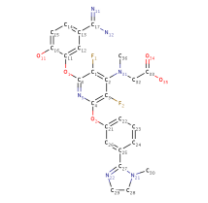
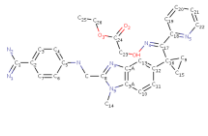
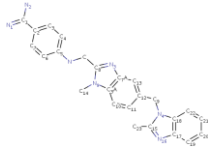
End while

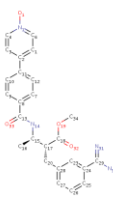
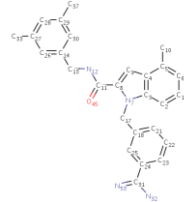
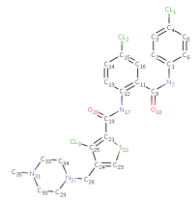
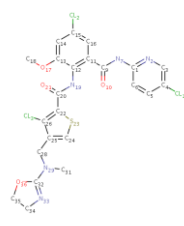
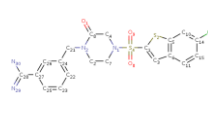
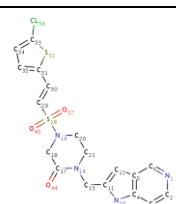
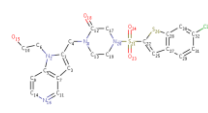
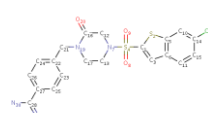
Table A2 Setting parameters of the five algorithms for docking experiments.

F/ABCps	
Number of food sources, SN	500
Maximum trial number, $limit$	200
<hr/>	
ABC	
Number of food sources, SN	500
Maximum trial number, $limit$	200
<hr/>	
SODOCK	
Number of particles, Np	500
Number of immediate neighbors, K	4
Inertia weight, w	0.9~0.4 (liner decreasing)
Cognitive weight, $c1$	2.0
Social weight, $c2$	2.0
Maximal velocity, $Vmax$	2.0 Å (for translation) 1.0, 180 deg (for orientation) 50 deg (for conformation)
Maxmal steps of local search	50
<hr/>	
PSO	
Number of particles, Np	150
Inertia weight, w	0.9~0.4 (liner decreasing)
Cognitive weight, $c1$	2.0
Social weight, $c2$	2.0
Maximal velocity, $Vmax$	2.0 Å (for translation) 1.0, 180 deg (for orientation) 50 deg (for conformation)
<hr/>	
LGA	
Population size (ga_pop_size)	150
Survive elite (ga_elitism)	1
Mutation rate (ga_mutation_rate)	0.02
Crossover rate (ga_crossover_rate)	0.8
Window size (ga_window_size)	10
ga_cauchy_alpha	0.0
ga_cauchy_beta	1.0
Maximum iteration of local search (sw_max_its)	300
Maximum number of success (sw_max_succ)	4
Maximum number of fail (sw_max_fail)	4
sw_rho	1.0
sw_lb_rho	0.01
ls_search_freq	0.06

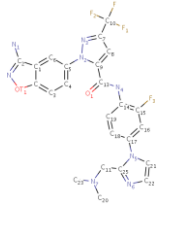
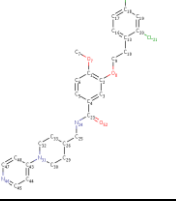
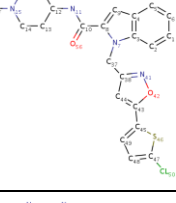
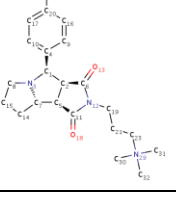
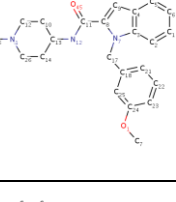
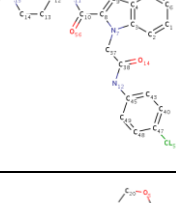
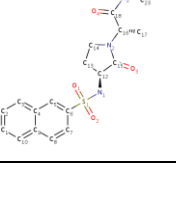
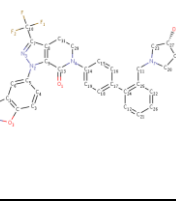
Appendix B

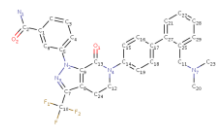
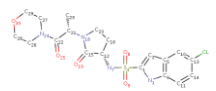
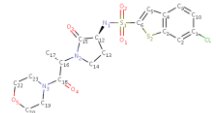
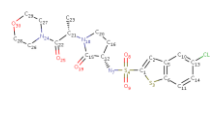
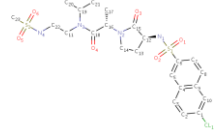
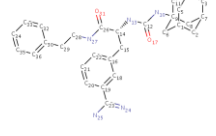
Table B1 28 ligands of coagulation factor Xa used in this work for the training set.

ID	PDB-ID/ res-ID	Chemical Structure	Compound Name	MW (g/mol)	ΔG_{exp} (kcal/mol) [reference]
1	1EZQ/RPR		3-[(3'-AMINOMETHYL-BIPHENYL-4-CARBONYL)-AMINO]-2-(3-CARBAMIMIDOYL-BENZYL)-BUTYRIC ACID METHYL ESTER	458.5	-12.21 [257]
2	1F0R/815		THIENO[3,2-B]PYRIDINE-2-SULFONIC ACID [1-(1-AMINO-ISOQUINOLIN-7-YLMETHYL)-2-OXO-PYRROLDIN-3-YL]-AMIDE	453.5	-10.34 [258]
3	1F0S/PR2		THIENO[3,2-B]PYRIDINE-2-SULFONIC ACID [2- OXO-1-(1H-PYRROLO[2,3-C]PYRIDIN-2-YLMETHYL)- PYRROLIDIN-3-YL]-AMIDE	427.5	-10.45 [258]
4	1FJS/Z34		N-[2-[5-[AMINO(IMINO)METHYL]-2-HYDROXYPHENOXY]-3,5-DIFLUORO-6-[3-(4,5-DIHYDRO-1-METHYL-1H-IMIDAZOL-2-YL)PHENOXY]PYRIDIN-4-YL]-N-METHYLGLYCINE	526.5	-13.44 [259]
5	1G2L/T87		[(1-{2[(4-CARBAMIMIDOYL-PHENYLAMINO)-METHYL]-1-METHYL-1H-BENZOIMIDAZOL-5-YL}-CYCLOPROPYL)-PYRIDIN-2-YL-METHYLENEAMINOXY]-ACETIC ACID ETHYL ESTER	525.6	-10.28 [260]
6	1G2M/R11		4-{[1-METHYL-5-(2-METHYL-BENZOIMIDAZOL-1-YLMETHYL)-1H-BENZOIMIDAZOL-2-YLMETHYL]-AMINO}-BENZAMIDINE	423.5	-10.49 [260]

7	1KSN/FXV		METHYL-3-(4'-N- OXOPYRIDYLPHENOYL)-3- METHYL- 2-(M- AMIDINO BENZYL)-PROPIONATE	447.5	-12.82 [261]
8	1LQD/CMI		1-(3-CARBAMIMIDOYL-BENZYL)- 4-METHYL-1H-INDOLE- 2- CARBOXYLIC ACID 3,5- DIMETHYL-BENZYLAMIDE	424.5	-10.97 [262]
9	1MQ5/XLC		3-CHLORO-N-[4-CHLORO-2-[[4- CHLOROPHENYL)AMINO]CARBO NYL]PHENYL]- 4-[(4-METHYL-1- PIPERAZINYL)METHYL]-2- THIOPHENECARBOXAMIDE	537.9	-12.31 [263]
10	1MQ6/XLD		3-CHLORO-N-[4-CHLORO-2-[[5- CHLORO-2- PYRIDINYL)AMINO]CARBONYL]- 6-METHOXYPHENYL]-4-[[4,5- DIHYDRO-2- OXAZOLYL)METHYLAMINO]MET HYL]- 2- THIOPHENECARBOXAMIDE	568.9	-15.06 [263]
11	1NFU/RRP		3-({4-[(6-CHLORO-1-BENZOTHIEN- 2-YL)SULFONYL]- 2- OXOPIPERAZIN-1- YL}METHYL)BENZENECARBOXI MIDAMIDE	463.0	-10.45 [257]
12	1NFW/RRR		4- {(E)-2-(5-CHLOROTHIEN-2- YL)VINYL]SULFONYL}- 1-(1H- PYRROLO[3,2-C]PYRIDIN-2- YLMETHYL)PIPERAZIN- 2-ONE	436.9	-12.09 [257]
13	1NFX/RDR		4-[(6-CHLORO-1-BENZOTHIEN-2- YL)SULFONYL]- 1-{[1-(2- HYDROXYETHYL)-1H- PYRROLO[3,2-C]PYRIDIN- 2- YL]METHYL}PIPERAZIN-2-ONE	505.0	-11.51 [257]
14	1NFY/RTR		3-({4-[(6-CHLORO-1-BENZOTHIEN- 2-YL)SULFONYL]- 2- OXOPIPERAZIN-1- YL}METHYL)BENZENECARBOXI MIDAMIDE	463.0	-12.00 [257]

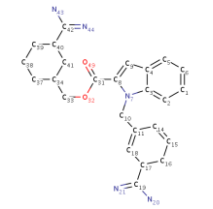
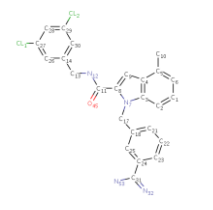
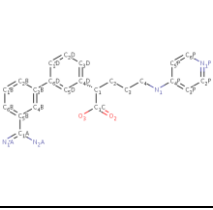
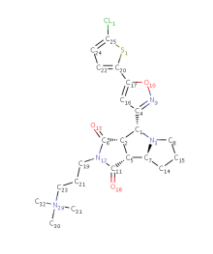
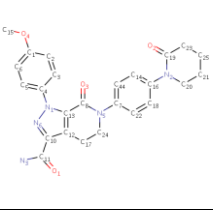
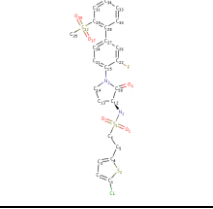
Appendix B

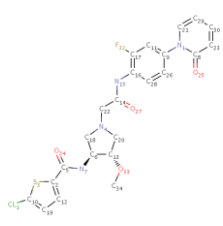
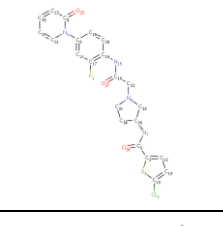
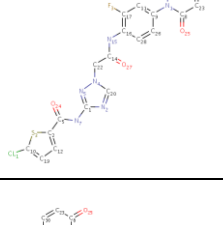
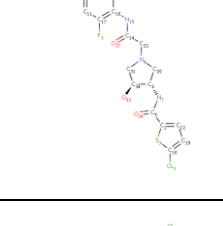
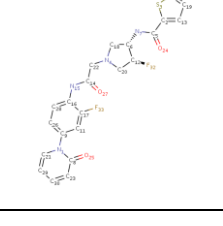
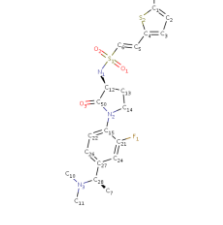
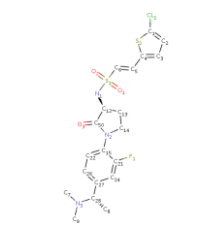
15	1Z6E/IK8		1-(3-AMINO-1,2-BENZISOXAZOL-5-YL)-N-(4-{2-[(DIMETHYLAMINO)METHYL]-1H-IMIDAZOL-1-YL}-2-FLUOROPHENYL)-3-(TRIFLUOROMETHYL)-1H-PYRAZOLE-5-CARBOXAMIDE	528.5	-13.12 [264]
16	2BMG/I1H		3-[2-(2,4-DICHLOROPHENYL)ETHOXY]-4-METHOXY- N-[(1-PYRIDIN-4-YL)PYRIDIN-4-YL)METHYL]BENZAMIDE	514.4	-10.56 [265]
17	2BOH/IIA		1-{[5-(5-CHLORO-2-THIENYL)ISOXAZOL-3-YL]METHYL}- N-(1-ISOPROPYLPYRIDIN-4-YL)-1H-INDOLE-2- CARBOXAMIDE	483.0	-11.62 [266]
18	2BOK/784		[AMINO (4-{(3AS,4R,8AS,8BR)-1,3-DIOXO-2- [3-(TRIMETHYLAMMONIO)PROPYL]DECAHYDROPYRROLO[3,4- A] PYRROLIZIN-4-YL} PHENYL) METHYLENE]AMMONIUM	398.5	-9.173 [267]
19	2BQ7/IID		N-(1-ISOPROPYLPYRIDIN-4-YL)-1-(3-METHOXYBENZYL)- 1H-INDOLE-2-CARBOXAMIDE	405.5	-9.61 [266]
20	2BQW/IEE		1-{2-[(4-CHLOROPHENYL)AMINO]-2-OXOETHYL}- N-(1-ISOPROPYLPYRIDIN-4-YL)-1H-INDOLE-2- CARBOXAMIDE	453.0	-11.62 [266]
21	2CJI/GSK		6-CHLORO-N-{(3S)-1-[(1S)-1-METHYL-2-(4-MORPHOLINYL)- 2-OXO ETHYL]-2-OXO-3-PYRROLIDINYL}-2-NAPHTHALENESULFONAMIDE	466.0	-11.21 [268]
22	2FZZ/5QC		1-(3-AMINO-1,2-BENZISOXAZOL-5-YL)-6-(2'-{[(3R)- 3-HYDROXYPYRROLIDIN-1-YL]METHYL}BIPHENYL- 4-YL)-3-(TRIFLUOROMETHYL)-1,4,5,6-TETRAHYDRO- 7H-PYRAZOLO[3,4-C]PYRIDIN-7-ONE	588.6	-14.21 [269]

23	2G00/4QC		3-[6-{2'-[(DIMETHYLAMINO)METHYL]BIPHENYL- 4-YL}-7-OXO-3-(TRIFLUOROMETHYL)-4,5,6,7-TETRAHYDRO- 1H-PYRAZOLO[3,4-C]PYRIDIN-1-YL]BENZAMIDE	533.5	-13.16 [269]
24	2J2U/GSQ		5-CHLORO-N-[(3S)-1-[(1S)-1-METHYL-2-MORPHOLIN- 4-YL-2-5-CHLORO-N-[(3S)-1-[(1S)-1-METHYL- 2-MORPHOLIN-4-YL-2-SULFONAMIDE	454.9	-9.61 [270]
25	2J34/GS6		6-CHLORO-N-[(3S)-1-[(1S)-1-METHYL-2-MORPHOLIN- 4-YL-2-OXOETHYL]-2-OXOPYRROLIDIN-3-YL]-1- BENZOTHIOPHENE-2-SULFONAMIDE	472.0	-10.67 [270]
26	2J38/GS5		5-CHLORO-N-[(3S)-1-[(1S)-1-METHYL-2-MORPHOLIN- 4-YL-2-OXOETHYL]-2-OXOPYRROLIDIN-3-YL]-1- BENZOTHIOPHENE-2-SULFONAMIDE	472.0	-9.99 [270]
27	2J4I/GSJ		1-PYRROLIDINEACETAMIDE, 3-[[[(6-CHLORO-2-NAPHTHALENYL)SULFONYL]AMINO]-ALPHA-METHYL-N-(1-METHYLETHYL)-N-[2-[(METHYLSULFONYL)AMINO]ETHYL]-2-OXO-, (ALPHA,3S)-	559.1	-12.27 [268]
28	3LIW/RUP		(R)-2-(3-ADAMANTAN-1-YL-UREIDO)-3-(3-CARBAMIMIDOYL-PHENYL)-N-PHENETHYL-PROPIONAMIDE	487.6	-10.37 [271]

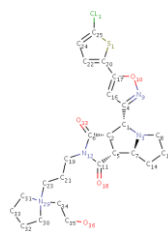
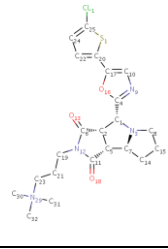
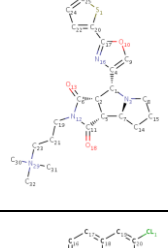
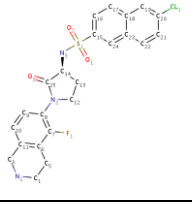
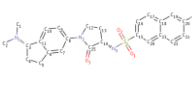
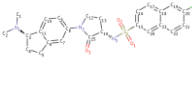
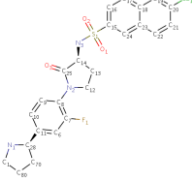
Appendix B

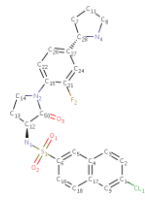
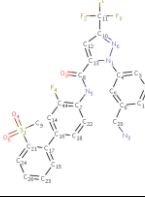
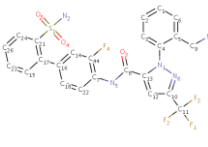
Table B2 23 ligands of coagulation factor Xa used in this work for the test set.

ID	PDB-ID/res-ID	Chemical Structure	Compound Name	MW (g/mol)	ΔG_{exp} (kcal/mol) [reference]
1	1LPK/CBB		1-(3-CARBAMIMIDOYL-BENZYL)-1H-INDOLE-2-CARBOXYLIC ACID 3-CARBAMIMIDOYL-BENZYLESTER	433.6	-10.20 [262]
2	1LPZ/CMB		1-(3-CARBAMIMIDOYLBENZYL)-N-(3,5-DICHLOROBENZYL)-4-METHYL-1H-INDOLE-2-CARBOXAMIDE	465.4	-10.26 [262]
3	1XKA/4PP		(2S)-(3'-AMIDINO-3-BIPHENYL)-5-(4-PYRIDYLAMINO)PENTANOIC ACID	388.5	-9.29 [272]
4	2JKH/BI7		3-[(3AS,4R,8AS,8BR)-4-[5-(5-CHLORO-2-THIENYL)ISOXAZOL-3-YL]-1,3-DIOXOCTAHYDROPYRROLO[3,4-A]PYRROLIZIN-2(3H)-YL]-N,N,N-TRIMETHYLPROPAN-1-AMINIUM	464.0	-10.86 [273]
5	2P16/GG2		1-(4-METHOXYPHENYL)-7-OXO-6-[4-(2-OXOPIPERIDIN-1-YL)PHENYL]-4,5,6,7-TETRAHYDRO-1H-PYRAZOLO[3,4-C]PYRIDINE-3-CARBOXAMIDE	459.5	-13.63 [274]
6	2VH6/GSV		2-(5-CHLOROTHIOPHEN-2-YL)-N-[(3S)-1-[3-FLUORO-2-(METHYLSULFONYL)BIPHENYL-4-YL]-2-OXOPYRROLIDIN-3-YL]ETHANESULFONAMIDE	557.1	-13.09 [275]

7	2VVC/LZF		5-CHLORO-N-[(3S,4S)-1-(2-([2-FLUORO-4-(2- OXOPYRIDIN-1(2H)-YL)PHENYL]AMINO)-2- OXOETHYL)- 4- METHOXYPIRROLIDIN-3- YL]THIOPHENE-2-CARBOXAMIDE	505.0	-11.51 [276]
8	2VVU/H22		5-CHLORO-N-[(3R)-1-(2-([2-FLUORO-4-(2-OXOPYRIDIN- 1(2H)- YL)PHENYL]AMINO)-2- OXOETHYL)PYRROLIDIN- 3- YL]THIOPHENE-2-CARBOXAMIDE	474.9	-10.93 [276]
9	2VVV/H21		5-CHLORO-N-[1-(2-([2-FLUORO-4-(2- OXOPYRIDIN- 1(2H)- YL)PHENYL]AMINO)-2- OXOETHYL)-1H-1,2,4- TRIAZOL-3- YL]THIOPHENE-2-CARBOXAMIDE	472.9	-11.10 [276]
10	2VWN/H25		5-CHLORO-THIOPHENE-2- CARBOXYLIC ACID ((3S,4S)- 1-([2- FLUORO-4-(2- OXO-2H-PYRIDIN-1- YL)-PHENYLCARBAMOYL]- METHYL}-4-HYDROXY- PYRROLIDIN-3-YL)-AMIDE	490.9	-10.80 [276]
11	2VWO/LZ G		5-CHLORO-THIOPHENE-2- CARBOXYLIC ACID ((3S,4S)- 4- FLUORO- 1-([2-FLUORO-4-(2-OXO- 2H-PYRIDIN- 1-YL)- PHENYLCARBAMOYL]-METHYL}- PYRROLIDIN- 3-YL)-AMIDE	492.9	-10.14 [276]
12	2WYG/461		(E)-2-(5-CHLOROTHIOPHEN-2-YL)-N- [(3S)-1-{4- [(1R)-1- (DIMETHYLAMINO)ETHYL]-2- FLUOROPHENYL}- 2- OXOPYRROLIDIN-3- YL]ETHENESULFONAMIDE	472.0	-11.74 [277]
13	2WYJ/898		(E)-2-(5-CHLOROTHIOPHEN-2-YL)-N- [(3S)-1-{4- [(1S)-1- (DIMETHYLAMINO)ETHYL]-2- FLUOROPHENYL}- 2- OXOPYRROLIDIN-3- YL]ETHENESULFONAMIDE	472.0	-12.15 [277]

Appendix B

14	2Y5F/XWG		(3AS,4R,5S,8AS,8BR)-4-[5-(5-CHLOROTHIOPHEN-2-YL)-1,2-OXAZOL-3-YL]-2-[3-[1-(2-HYDROXYETHYL)PYRROLIDIN-1-IUM-1-YL]PROPYL]-4,6,7,8,8A,8B-HEXAHYDRO-3AH-PYRROLO[3,4-A]PYRROLIZINE-1,3-DIONE	520.1	-11.74 [278]
15	2Y5G/FJD		3-[(3AS,4R,5S,8AS,8BR)-4-[5-(5-CHLOROTHIOPHEN-2-YL)-1,3-OXAZOL-2-YL]-1,3-DIOXO-4,6,7,8,8A,8B-HEXAHYDRO-3AH-PYRROLO[3,4-A]PYRROLIZIN-2-YL]PROPYL-TRIMETHYL-AZANIUM	464.0	-9.23 [278]
16	2Y5H/Y5H		3-[(3AS,4R,5S,8AS,8BR)-4-[2-(5-CHLOROTHIOPHEN-2-YL)-1,3-OXAZOL-4-YL]-1,3-DIOXO-4,6,7,8,8A,8B-HEXAHYDRO-3AH-PYRROLO[3,4-A]PYRROLIZIN-2-YL]PROPYL-TRIMETHYL-AZANIUM	464.0	-7.82 [278]
17	2Y7X/MZA		6-CHLORO-N-[(3S)-1-(5-FLUORO-1,2,3,4-TETRAHYDROISOQUINOLIN-6-YL)-2-OXO-PYRROLIDIN-3-YL]NAPHTHALENE-2-SULFONAMIDE	474.0	-12.00 [279]
18	2Y7Z/C0Z		6-CHLORO-N-[(3S)-1-[(1S)-1-DIMETHYLAMINO-2,3-DIHYDRO-1H-INDEN-5-YL]-2-OXO-PYRROLIDIN-3-YL]NAPHTHALENE-2-SULFONAMIDE	484.0	-11.74 [280]
19	2Y80/439		6-CHLORO-N-[(3S)-1-[(1S)-1-(DIMETHYLAMINO)-2,3-DIHYDRO-1H-INDEN-5-YL]-2-OXO-3-PYRROLIDINYL]-2-NAPHTHALENESULFONAMIDE	484.0	-10.86 [280]
20	2Y81/931		6-CHLORO-N-[(3S)-2-OXO-1-{4-[(2R)-2-PYRROLIDINYL]PHENYL}-3-PYRROLIDINYL]-2-NAPHTHALENESULFONAMIDE	488.0	-11.74 [280]

21	2Y82/930		6-CHLORO-N-((3S)-2-OXO-1-{4-[(2S)-2-PYRROLIDINYL]PHENYL}-3-PYRROLIDINYL)-2-NAPHTHALENESULFONAMIDE	488.0	-11.34 [280]
22	3M36/M35		1-[3-(AMINOMETHYL)PHENYL]-N-[3-FLUORO-2'-(METHYLSULFONYL)BIPHENYL-4-YL]-3-(TRIFLUOROMETHYL)-1H-PYRAZOLE-5-CARBOXAMIDE	532.5	-13.26 [264]
23	3M37/M37		1-[2-(AMINOMETHYL)PHENYL]-N-(3-FLUORO-2'-(SULFAMOYL)BIPHENYL-4-YL)-3-(TRIFLUOROMETHYL)-1H-PYRAZOLE-5-CARBOXAMIDE	533.5	-12.21 [281]

Appendix C

Table C1 Selected binding pocket atoms for the six target proteins.

Targets	Selected pocket atoms (residue-id@atom-name)
PR (62 atoms)	39@OD1:38@O:38@C:38@CB:210@CD1:210@CE1:114@CE1:35@CD2:117@CD2:114@CD1:76@SD:207@CD2:207@CD1:121@SD:207@CB:207@O:76@CE:207@CA:211@CB:211@SG:211@CA:225@CE2:229@CE:38@CD1:114@CZ:98@CZ:121@CE:98@CE2:79@CB:98@CD2:38@CD2:45@NE2:86@NH2:41@CD2:98@O:45@OE1:41@CG:42@N:45@CD:83@CB:79@O:83@CD2:79@C:80@CA:80@CB:80@CG2:80@N:76@O:79@SD:75@CZ3:42@CA:35@O:35@CB:39@ND2:225@CZ:214@CB:214@CG2:214@OG1:35@CD1:223@CG1:211@N:210@CB
CDK2 (103 atoms)	80@CD1:145@N:145@OD1:144@CB:64@CG1:64@CB:31@CB:134@CD1:33@CE:33@NZ:80@CB:81@O:80@CG:80@CD2:127@OD2:127@CB:15@OH:154@CB:149@CB:154@CG2:14@CG2:14@OG1:15@CE2:14@CB:132@ND2:145@OD2:145@CB:129@NZ:125@CE1:125@ND1:125@O:145@O:149@N:148@CB:16@O:33@CB:17@O:34@O:16@N:18@CG2:35@CD1:33@CD:15@N:13@CA:15@CB:15@CD2:14@N:148@CD2:78@CD1:80@CE2:134@CD2:131@O:132@CA:132@CB:131@CB:132@N:131@C:83@N:83@O:82@CD1:18@CG1:10@CG2:10@CD1:82@CA:82@CE1:86@OD2:11@CA:10@O:18@CB:11@C:13@N:11@N:10@C:12@N:12@O:131@CG:12@C:162@OE2:129@CD:163@O:158@CG2:162@CB:162@O:13@C:129@CE:164@CG2:164@CA:165@N:162@CD:131@NE2:162@OE1:131@CD:164@CG1:165@CG2:84@O:85@CA:85@N:84@C:85@C:89@NZ:86@N:86@CB:85@O
SIRT2 (120 atoms)	42@N:41@O:41@C:40@O:36@OG1:36@CG2:42@CB:32@O:35@C:35@CB:40@CB:40@CG2:36@N:43@CE1:41@CD:84@O:41@CG:50@CD1:85@CA:85@O:90@CE2:85@CD2:43@CZ:40@CD1:90@CD2:117@CB:86@O:87@CA:117@O:50@CD2:116@O:116@CG1:85@CB:116@C:117@CA:117@N:115@OD1:115@CB:117@OD2:115@CG:43@CE2:116@N:114@O:43@CD2:32@CB:115@CA:32@CA:85@CD1:36@CA:42@CA:44@NH2:42@CG:42@OD1:237@O:211@CD1:237@N:236@CB:236@CA:212@OE1:237@CB:236@CG:235@O:231@CD2:210@CA:211@CG:211@N:209@O:256@CE2:231@CD1:233@ND2:208@O:33@N:233@OD1:33@CA:234@CB:235@CD:234@CD:235@OE2:235@OE1:235@CG:235@CB:237@CA:236@C:236@O:210@CB:210@N:44@NH1:44@CZ:235@N:209@CA:208@C:182@CZ:182@N:180@O:182@CG:182@CE1:182@CE2:182@CD2:134@CE1:134@CG:134@ND1:66@CE2:66@CD2:134@CD2:179@CG2:181@CD1:66@CB:66@CD1:66@CG:66@CE1:66@CZ:134@O:134@CB:116@CD1:134@NE2:66@O:182@CD1:116@CG2:116@CG1:114@O
HIV1PR (23 atoms)	84@CG2:28@CB:84@CD1:30@O:32@CG1:30@CB:30@OD2:30@N:47@CD1:29@N:27@O:48@CA:29@OD2:48@O:48@N:25@OD2:28@CA:50@CD1:29@CB:47@CB:32@CB:76@CD1:47@CG2
TP (102 atoms)	221@CG:88@CG2:225@CD1:184@CG2:221@CD2:204@CG1:221@O:204@CG2:205@O:218@CB:212@CB:212@N:88@CA:88@C:88@O:118@CD2:211@CG2:88@CB:211@CB:184@CD1:88@OG1:184@CG1:178@CG1:172@NH2:211@CG1:188@CG1:187@OG:86@NE2:169@CE2:169@OH:87@O:188@CD1:184@CA:187@CB:184@O:211@N:212@CD1:210@CB:178@CB:179@N:179@OD1:176@CA:179@OD2:119@CA:179@CG:178@N:176@C:176@O:178@CG2:173@O:172@O:120@ND1:176@CB:179@CB:120@CE1:118@O:119@N:121@OG1:118@CB:121@CG2:120@N:211@O:211@CA:90@CA:93@CB:122@O:90@O:127@NZ:89@O:123@CA:169@OH:116@O:115@O:124@N:124@OG1:114@OG:124@CG2:87@N:87@CB:96@OG:93@O:93@OD1:85@NZ:87@OG:121@O:88@O:121@OG1:118@CB:118@CD2:86@CE1:86@NE2:87@O:115@N:118@N:87@C:112@CE:86@ND1:86@CA:93@N:115@C:115@CA:85@CE
EGFR (65 atoms)	159@CG2:146@O:160@OD1:149@CD2:147@OD1:97@CD1:23@CD1:98@O:48@CB:98@N:98@CB:149@CD1:101@O:31@CG1:101@CA:31@CG2:24@CA:31@CB:23@O:105@OD1:102@N:28@CZ:102@CB:102@SG:28@CE2:146@CG:146@CB:23@CB:98@CG:96@O:101@C:146@CD:95@OG1:71@SD:82@CD1:95@CG2:93@CB:93@CD1:67@OE1:50@NZ:71@CE:159@OG1:50@CE:50@CB:48@O:93@O:48@C:50@N:49@C:95@CB:80@SG:96@CB:98@SD:96@OE1:80@CB:81@O:96@N:81@N:80@CA:157@CD:68@N:67@C:67@O:68@CA:93@CD2

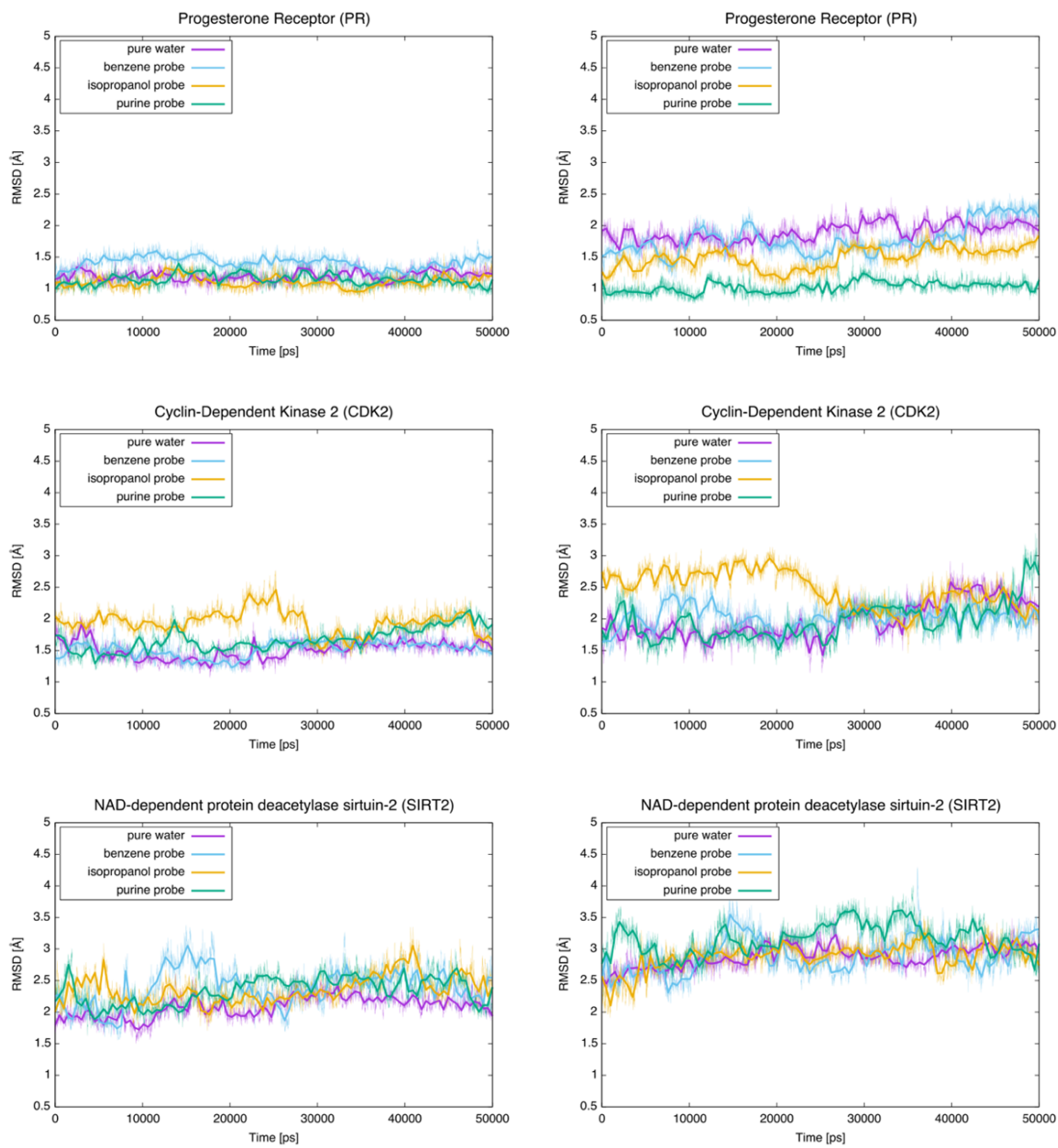


Figure C1 Root mean square deviation (RMSD) plots of molecular dynamics simulation for PR (**top**), CDK2 (**middle**), and SIRT2 (**bottom**); Bilateral represents RMSD of backbone Ca atoms (**left**) and RMSD of binding pocket atoms (**right**).

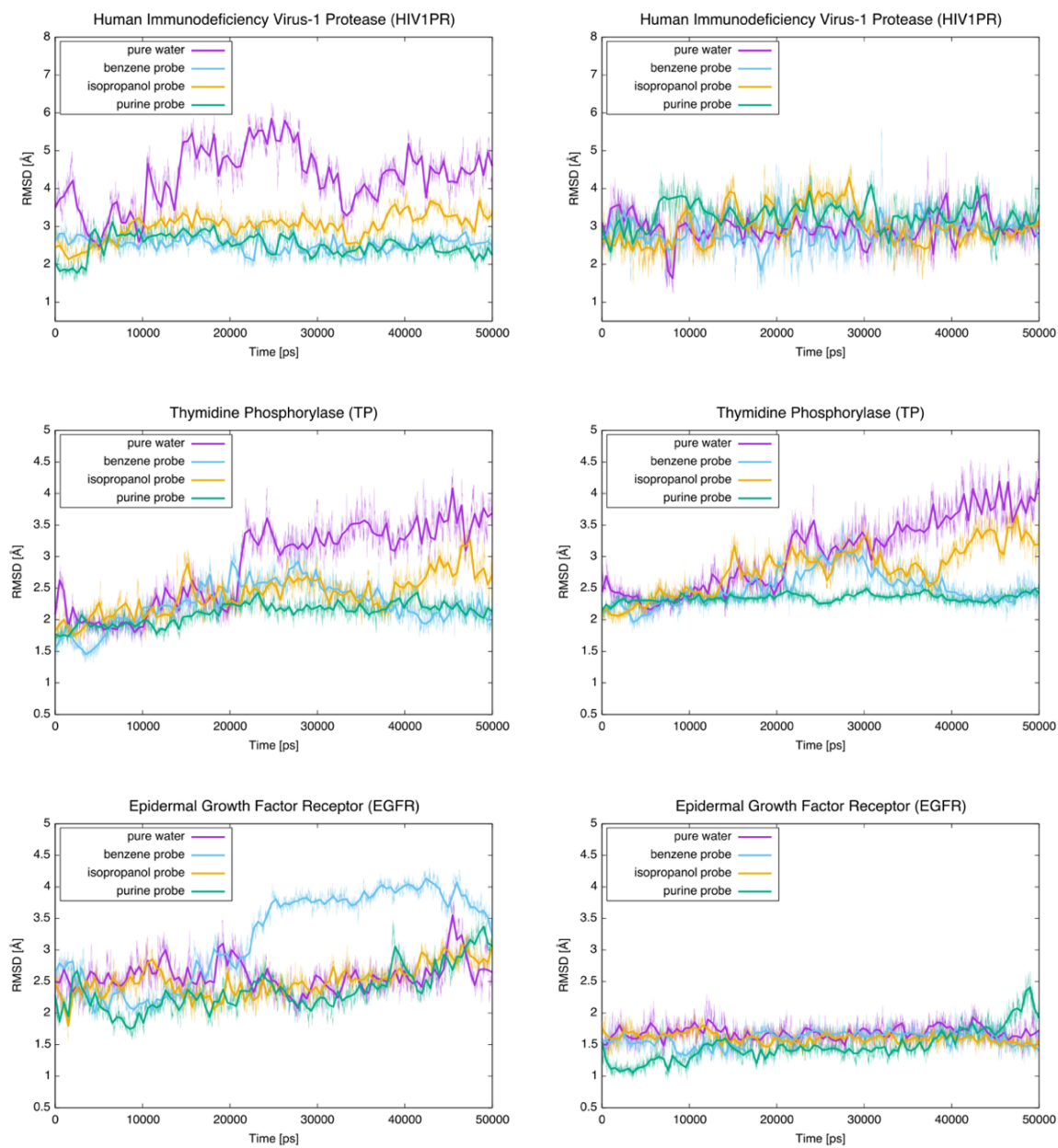


Figure C2 Root mean square deviation (RMSD) plots of molecular dynamics simulation for HIV1PR (**top**), TP (**middle**), and EGFR (**bottom**); Bilateral represents RMSD of backbone Ca atoms (**left**) and RMSD of binding pocket atoms (**right**).

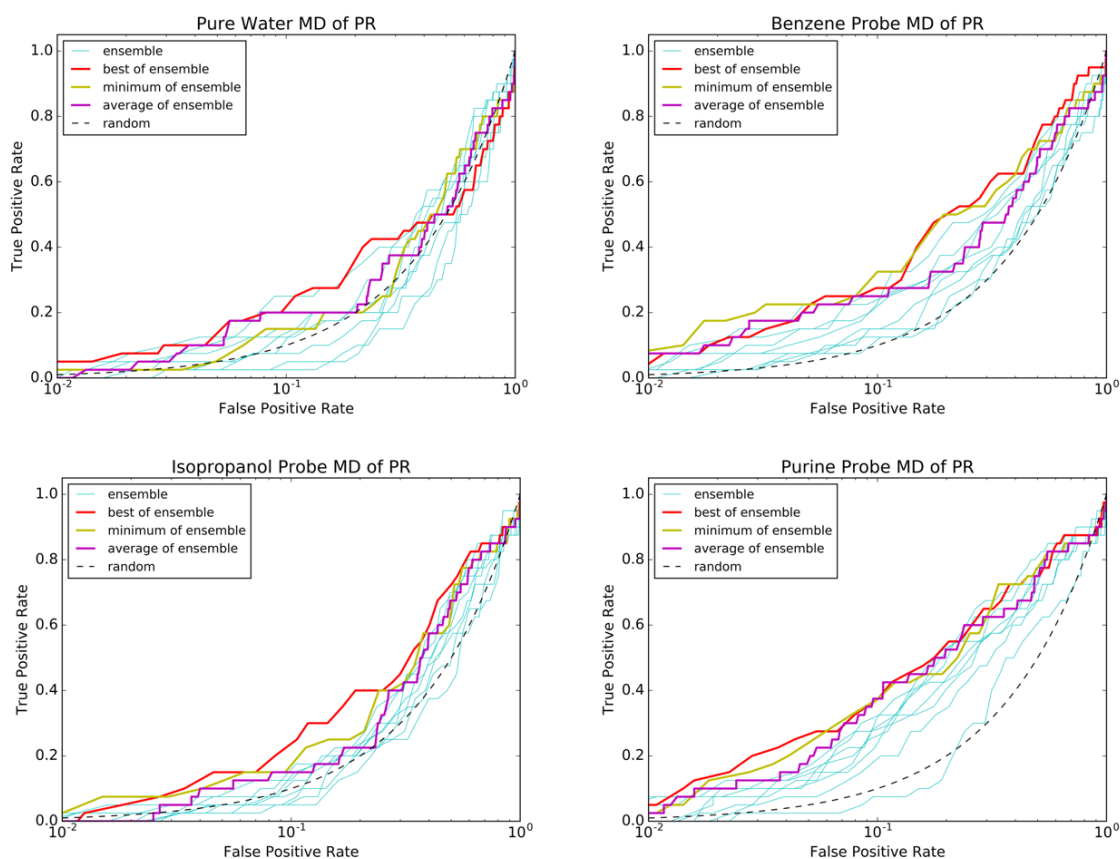


Figure C3 Log-scaled receiver operating characteristic (ROC) plots for all MD-generated structures of progesterone receptor (PR): standard MD (**upper left**), benzene probe MD (**upper right**), isopropanol probe MD (**bottom left**), purine probe MD (**bottom right**). Each line represents the following: the best snapshot (red), minimum score of ensemble docking (yellow), average score of ensemble docking (magenta), single docking score of ensemble structures (cyan), and theoretical result of random selection (black dots).

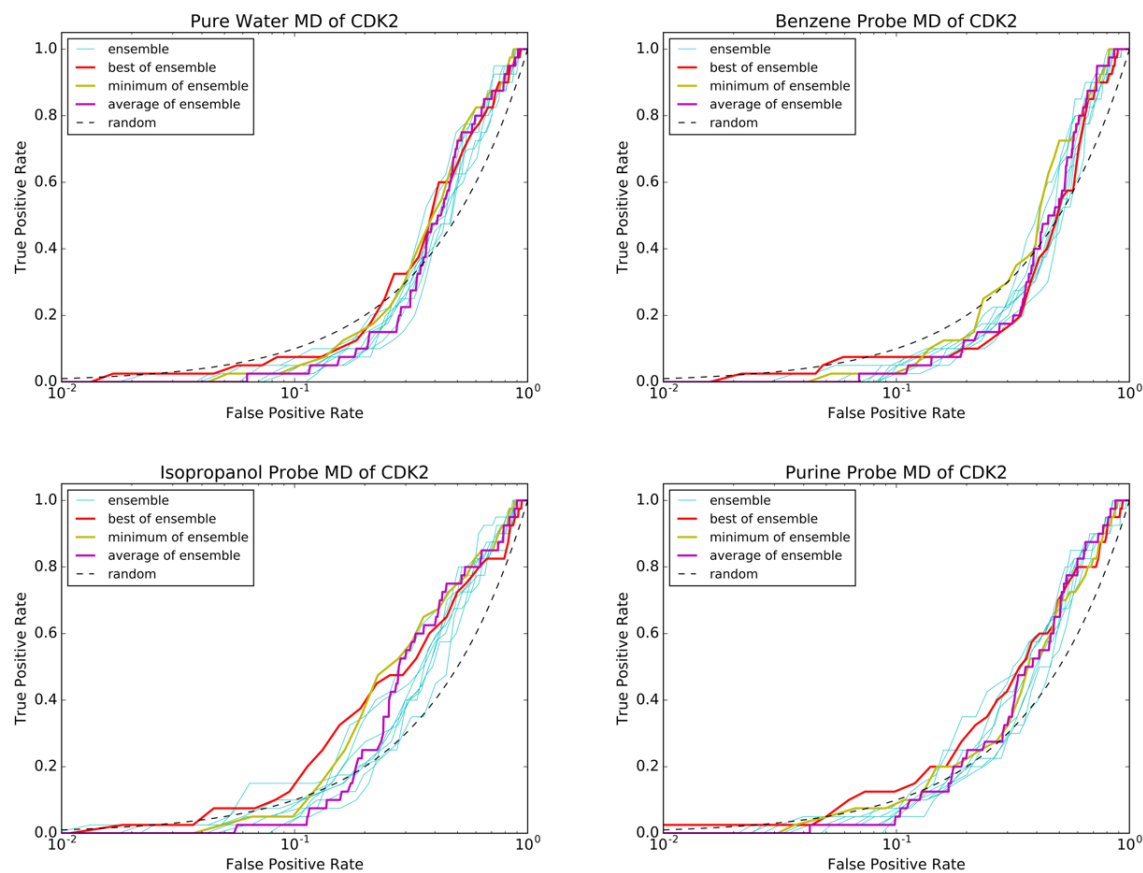


Figure C4 Log-scaled receiver operating characteristic (ROC) plots for all MD-generated structures of cyclin-dependent kinase 2 (CDK2): standard MD (**upper left**), benzene probe MD (**upper right**), isopropanol probe MD (**bottom left**), purine probe MD (**bottom right**). Each line represents the following: the best snapshot (red), minimum score of ensemble docking (yellow), average score of ensemble docking (magenta), single docking score of ensemble structures (cyan), and theoretical result of random selection (black dots).

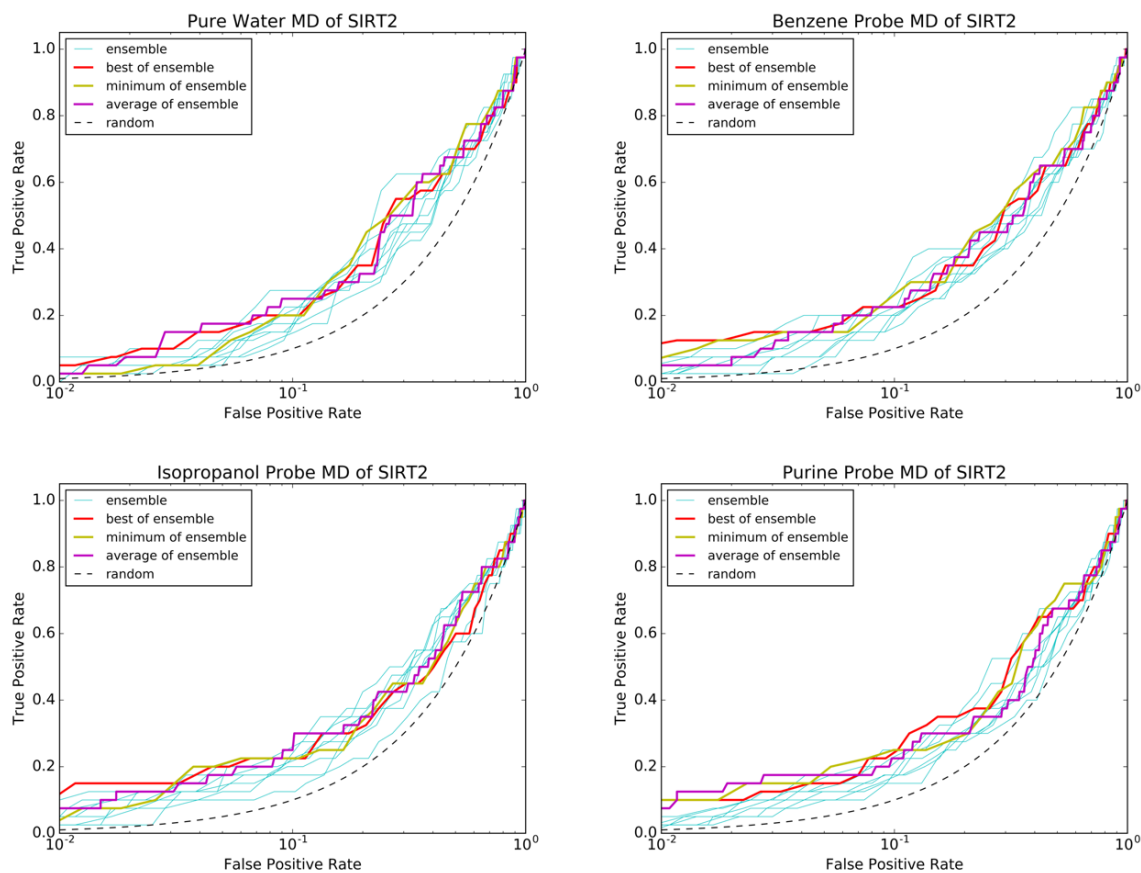


Figure C5 Log-scaled receiver operating characteristic (ROC) plots for all MD-generated structures of NAD-dependent protein deacetylase sirtuin-2 (SIRT2): standard MD (**upper left**), benzene probe MD (**upper right**), isopropanol probe MD (**bottom left**), purine probe MD (**bottom right**). Each line represents the following: the best snapshot (red), minimum score of ensemble docking (yellow), average score of ensemble docking (magenta), single docking score of ensemble structures (cyan), and theoretical result of random selection (black dots).

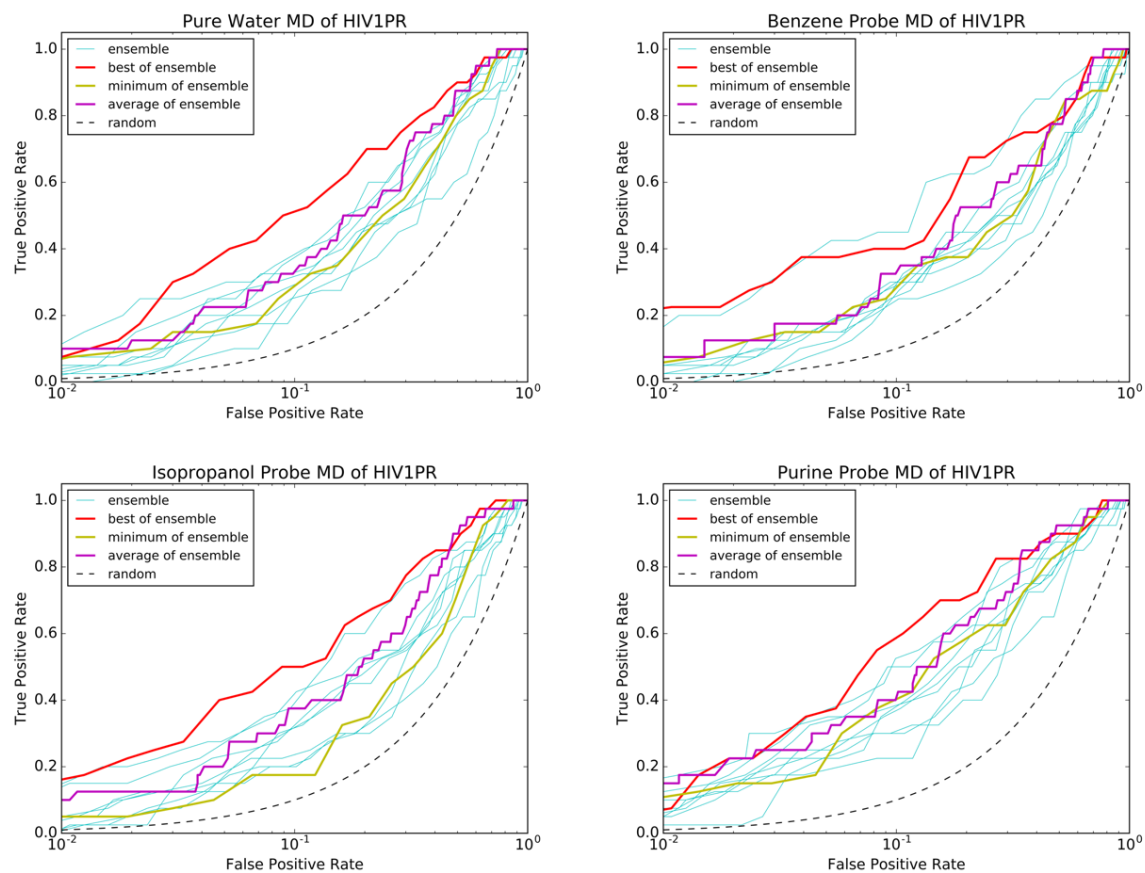


Figure C6 Log-scaled receiver operating characteristic (ROC) plots for all MD-generated structures of human immunodeficiency virus-1 protease (HIV1PR): standard MD (**upper left**), benzene probe MD (**upper right**), isopropanol probe MD (**bottom left**), purine probe MD (**bottom right**). Each line represents the following: the best snapshot (red), minimum score of ensemble docking (yellow), average score of ensemble docking (magenta), single docking score of ensemble structures (cyan), and theoretical result of random selection (black dots).

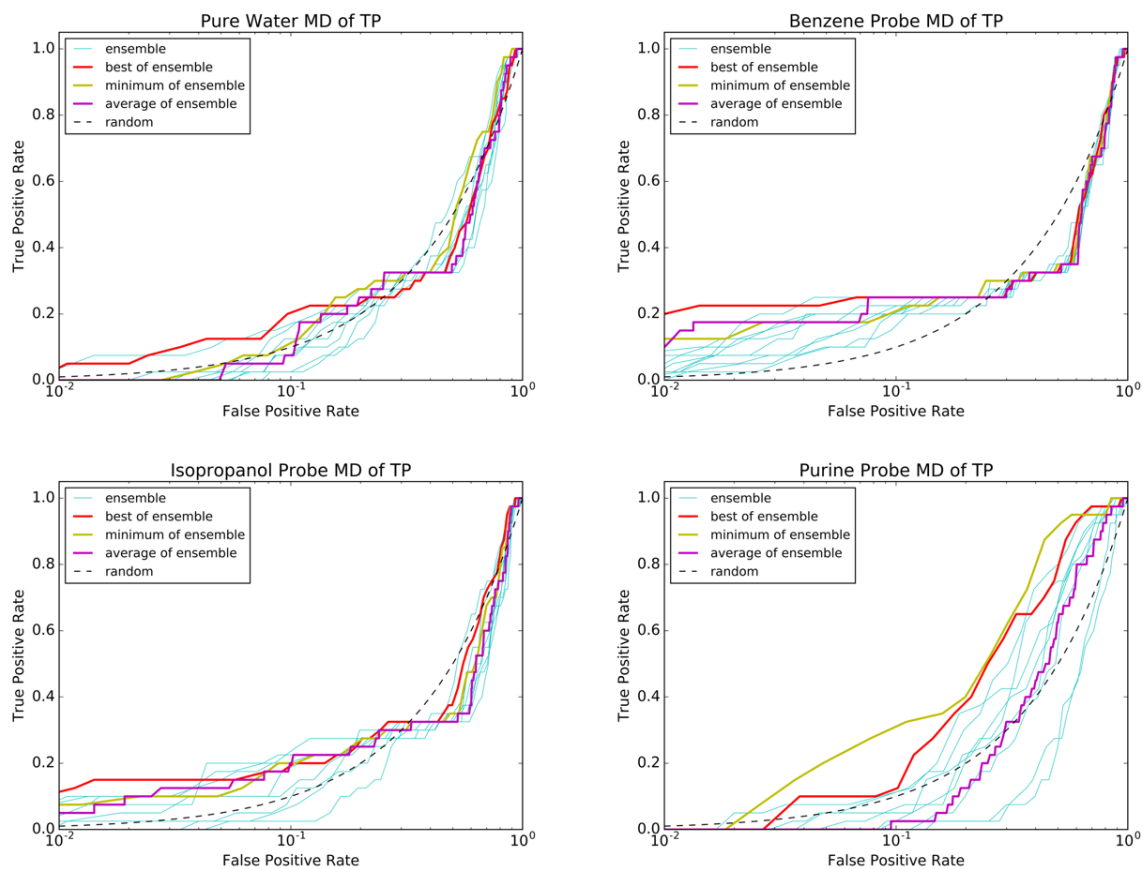


Figure C7 Log-scaled receiver operating characteristic (ROC) plots for all MD-generated structures of thymidine phosphorylase (TP): standard MD (**upper left**), benzene probe MD (**upper right**), isopropanol probe MD (**bottom left**), purine probe MD (**bottom right**). Each line represents the following: the best snapshot (red), minimum score of ensemble docking (yellow), average score of ensemble docking (magenta), single docking score of ensemble structures (cyan), and theoretical result of random selection (black dots).

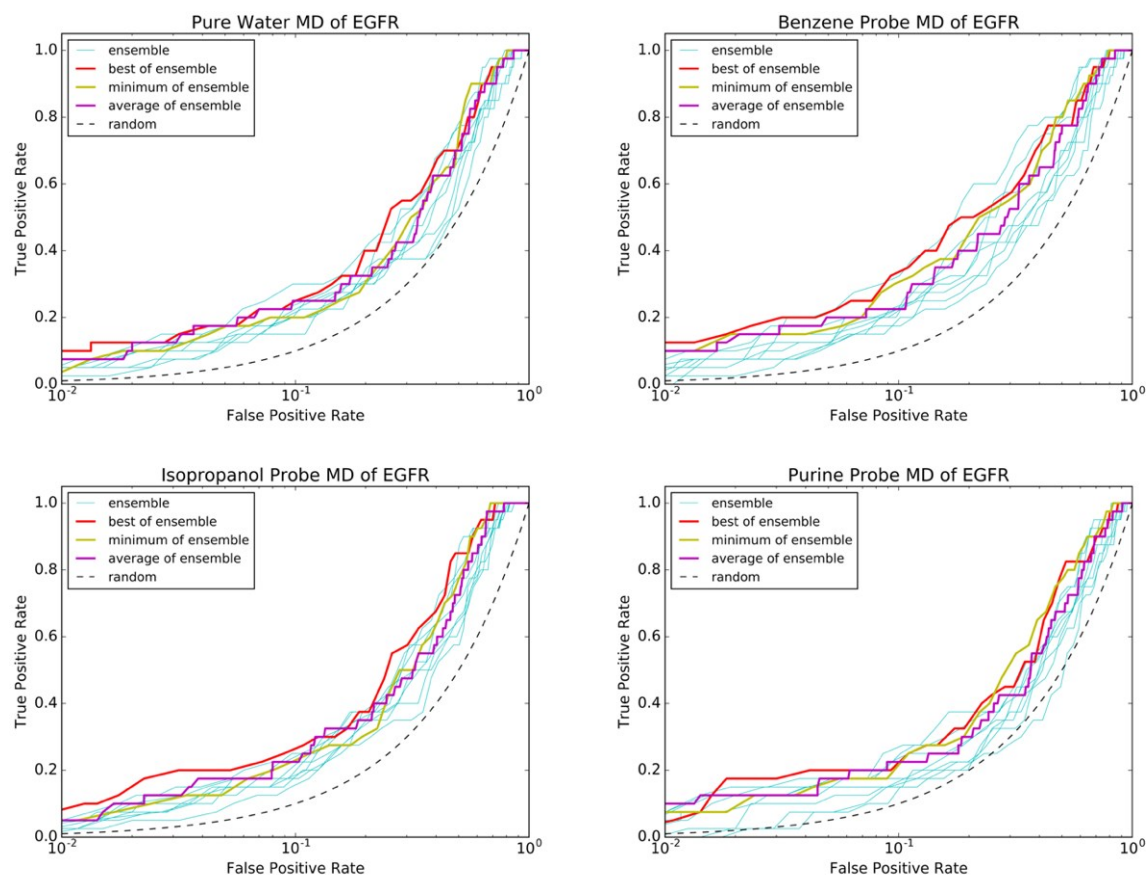


Figure C8 Log-scaled receiver operating characteristic (ROC) plots for all MD-generated structures of epidermal growth factor receptor (EGFR): standard MD (**upper left**), benzene probe MD (**upper right**), isopropanol probe MD (**bottom left**), purine probe MD (**bottom right**). Each line represents the following: the best snapshot (red), minimum score of ensemble docking (yellow), average score of ensemble docking (magenta), single docking score of ensemble structures (cyan), and theoretical result of random selection (black dots).

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor, Prof. Shigenori Tanaka for his invaluable guidance, continuous support, and insightful advice throughout my PhD study at Kobe University. I also like to thank Prof. Kuniyoshi Ebina for his in-depth advice and suggestive comments during my graduate course. My deepest appreciation goes to Dr. Kazuhiro Fujimoto for his enthusiastic guidance and fruitful advice throughout my undergraduate course and master course. Without his guidance and persistent help, my PhD work would not have been possible. I also would like to express my gratitude to Prof. Kohei Shimamura for his helpful advice and being approachable for my question at any time. Furthermore, many thanks to the members of Tanaka research group for providing a helpful, friendly, and interactive working atmosphere. I am particularly grateful to Ritsuko Ushio who support and encouraged me throughout my work at Tanaka laboratory. I would like to special thanks to Rena Suzuki for her encouragement and support during the last years. Finally, I would like to show my greatest appreciation to my family, especially for my parents, Masaaki Uehara and Keiko Uehara for their kind support throughout my times at Kobe University.

Bibliography

- [1] Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* **2010**, *9*, 203–214.
- [2] Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: A molecular modeling perspective. *Med. Res. Rev.* **1996**, *16*, 3–50.
- [3] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- [4] Rose, P. W.; Prlić, A.; Bi, C.; Bluhm, W. F.; Christie, C. H.; Dutta, S.; Green, R. K.; Goodsell, D. S.; Westbrook, J. D.; Woo, J.; Young, J.; Zardecki, C.; Berman, H. M.; Bourne, P. E.; Burley, S. K. The RCSB Protein Data Bank: Views of structural biology for basic and applied research and education. *Nucleic Acids Res.* **2015**, *43*, D345–D356.
- [5] Kuntz, I. D. Structure-Based Strategies for Drug Design and Discovery. *Science* **1992**, *257*, 1078–1082.
- [6] Lyne, P. D. Structure-based virtual screening: An overview. *Drug Discov. Today* **2002**, *7*, 1047–1055.
- [7] Waszkowycz, B.; Clark, D. E.; Gancia, E. Outstanding challenges in protein-ligand docking and structure-based virtual screening. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 229–259.
- [8] Shoichet, B. K. Virtual Screening of Chemical Libraries. *Nature* **2004**, *432*, 862–865.
- [9] Villoutreix, B.; Eudes, R.; Miteva, M. Structure-Based Virtual Ligand Screening: Recent Success Stories. *Comb. Chem. High Throughput Screen.* **2009**, *12*, 1000–1016.
- [10] Yuriev, E.; Agostino, M.; Ramsland, P. A. Challenges and advances in computational docking: 2009 in review. *J. Mol. Recognit.* **2011**, *24*, 149–164.
- [11] Sousa, S. F.; Ribeiro, A. J. M.; Coimbra, J. T. S.; Neves, R. P. P.; Martins, S. A.; Moorthy, N. S. H. N.; Fernandes, P. A.; Ramos, M. J. Protein-Ligand Docking in the New Millennium – A Retrospective of 10 Years in the Field. *Curr. Med. Chem.* **2013**, *20*, 2296–2314.
- [12] Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. Protein-ligand docking: Current status and future challenges. *Proteins Struct. Funct. Bioinforma.* **2006**, *65*, 15–26.
- [13] Plewczynski, D.; Łażniewski, M.; Augustyniak, R.; Ginalski, K. Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *J. Comput. Chem.* **2011**, *32*,

742–755.

- [14] Andre, J.; Siarry, P.; Dognon, T. An improvement of the standard genetic algorithm fighting premature convergence in continuous optimization. *Adv. Eng. Softw.* **2001**, *32*, 49–60.
- [15] Das, S.; Biswas, S.; Kundu, S. Synergizing fitness learning with proximity-based food source selection in artificial bee colony algorithm for numerical optimization. *Appl. Soft Comput.* **2013**, *13*, 4676–4694.
- [16] Ladbury, J. E. Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chem. Biol.* **1996**, *3*, 973–980.
- [17] Nguyen, C. N.; Young, T. K.; Gilson, M. K. Grid inhomogeneous solvation theory: Hydration structure and thermodynamics of the miniature receptor cucurbit[7]uril. *J. Chem. Phys.* **2012**, *137*, 149901.
- [18] Teague, S. J. Implications of protein flexibility for drug discovery. *Nat. Rev. Drug Discov.* **2003**, *2*, 527–541.
- [19] Knegtel, R. M.; Kuntz, I. D.; Oshiro, C. M. Molecular docking to ensembles of protein structures. *J. Mol. Biol.* **1997**, *266*, 424–440.
- [20] Ghanakota, P.; Carlson, H. A. Driving Structure-Based Drug Discovery through Cosolvent Molecular Dynamics. *J. Med. Chem.* **2016**, acs.jmedchem.6b00399.
- [21] Kalenkiewicz, A.; Grant, B. J.; Yang, C.-Y. Enrichment of druggable conformations from apo protein structures using cosolvent-accelerated molecular dynamics. *Biology* **2015**, *4*, 344–366.
- [22] Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* **2004**, *3*, 935–949.
- [23] Schneider, G.; Böhm, H. J. Virtual screening and fast automated docking methods. *Drug Discov. Today* **2002**, *7*, 64–70.
- [24] Joseph-McCarthy, D.; Baber, J. C.; Feyfant, E.; Thompson, D. C.; Humblet, C. Lead optimization via high-throughput molecular docking. *Curr. Opin. Drug Discov. Devel.* **2007**, *10*, 264–724.
- [25] Wlodawer, A.; Vondrasek, J. Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. *Annu. Rev. Biophys. Biomol. Struct.* **1998**, *27*, 249–284.
- [26] Kryger, G.; Silman, I.; Sussman, J. L. Structure of acetylcholinesterase complexed with E2020 (Aricept®): implications for the design of new anti-Alzheimer drugs. *Structure* **1999**, *7*, 297–307.
- [27] Babu, Y. S.; Chand, P.; Bantia, S.; Kotian, P.; Dehghani, A.; El-Kattan, Y.; Lin, T. H.; Hutchison, T. L.; Elliott, A. J.; Parker, C. D.; Ananth, S. L.; Horn, L. L.; Laver, G. W.; Montgomery, J. A. BCX-1812 (RWJ-270201): discovery of a novel, highly potent, orally active, and selective influenza neuraminidase inhibitor through structure-based drug design. *J. Med. Chem.* **2000**, *43*, 3482–3486.

Bibliography

- [28] Pierce, A. C.; Jacobs, M.; Stuver-Moody, C. Docking study yields four novel inhibitors of the protooncogene Pim-1 kinase. *J. Med. Chem.* **2008**, *51*, 1972–1975.
- [29] Costantino, G.; Macchiarulo, A.; Camaioni, E.; Pellicciari, R. Modeling of Poly(ADP-ribose)polymerase (PARP) Inhibitors. Docking of Ligands and Quantitative Structure – Activity Relationship Analysis. *J. Med. Chem.* **2001**, *44*, 3786–3794.
- [30] Vangrevelinghe, E.; Zimmermann, K.; Schoepfer, J.; Portmann, R.; Fabbro, D.; Furet, P. Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking. *J. Med. Chem.* **2003**, *46*, 2656–2662.
- [31] Doman, T. N.; McGovern, S. L.; Witherbee, B. J.; Kasten, T. P.; Kurumbail, R.; Stallings, W. C.; Connolly, D. T.; Shoichet, B. K. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J. Med. Chem.* **2002**, *45*, 2213–2221.
- [32] Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S. A semiempirical free energy force field with charge-based desolvation. *J. Comput. Chem.* **2007**, *28*, 1145–52.
- [33] Wang, R.; Lai, L.; Wang, S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput. Aided. Mol. Des.* **2002**, *16*, 11–26.
- [34] Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337–356.
- [35] Friesner, R. a; Banks, J. L.; Murphy, R. B.; Halgren, T. a; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739–49.
- [36] Miller, D. W.; Dill, K. A. Ligand binding to proteins: the binding landscape model. *Protein Sci.* **1997**, *6*, 2166–2179.
- [37] Chen, H.-M.; Liu, B.-F.; Huang, H.-L.; Hwang, S.-F.; Ho, S.-Y. SODOCK: Swarm optimization for highly flexible protein–ligand docking. *J. Comput. Chem.* **2007**, *28*, 612–623.
- [38] Jones, G.; Willett, P.; Glen, R. C.; Leach, a R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727–748.
- [39] Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639–1662.
- [40] Garnier, S.; Gautrais, J.; Theraulaz, G. The biological principles of swarm intelligence. *Swarm Intell.* **2007**, *1*, 3–31.
- [41] *Encyclopedia of Machine Learning*; Claude Sammut; Webb, G. I., Eds.; Springer US: NY, **2010**; Vol. 33.

-
- [42] Dorigo, M.; Blum, C. Ant colony optimization theory: A survey. *Theor. Comput. Sci.* **2005**, *344*, 243–278.
- [43] Yang, X.-S. Firefly Algorithms for Multimodal Optimization. In Proceedings of the 5th International Conference on Stochastic Algorithms: Foundations and Applications; **2009**; pp. 169–178.
- [44] Yang, X.-S.; Deb, S. Cuckoo Search via Levy Flights. *World Congr. Nat. Biol. Inspired Comput.* **2009**, 210–214.
- [45] Karaboga, D.; Basturk, B. A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. *J. Glob. Optim.* **2007**, *39*, 459–471.
- [46] Namasivayam, V.; Günther, R. PSO@autodock: A fast flexible molecular docking program based on Swarm intelligence. *Chem. Biol. Drug Des.* **2007**, *70*, 475–484.
- [47] Alatas, B.; Akin, E.; Ozer, a. B. Chaos embedded particle swarm optimization algorithms. *Chaos, Solitons & Fractals* **2009**, *40*, 1715–1734.
- [48] Karaboga, D.; Akay, B.; Ozturk, C. Artificial Bee Colony (ABC) Optimization Algorithm for Training Feed-Forward Neural Networks. In *Modeling Decisions for Artificial Intelligence*; Springer Berlin Heidelberg: Berlin, Heidelberg; **2007**, pp. 318–329.
- [49] Singh, A. An artificial bee colony algorithm for the leaf-constrained minimum spanning tree problem. *Appl. Soft Comput.* **2009**, *9*, 625–631.
- [50] Karaboga, N. A new design method based on artificial bee colony algorithm for digital IIR filters. *J. Franklin Inst.* **2009**, *346*, 328–348.
- [51] Karaboga, D.; Ozturk, C. A novel clustering approach: Artificial Bee Colony (ABC) algorithm. *Appl. Soft Comput.* **2011**, *11*, 652–657.
- [52] Karaboga, D.; Akay, B. A modified Artificial Bee Colony (ABC) algorithm for constrained optimization problems. *Appl. Soft Comput. J.* **2011**, *11*, 3021–3031.
- [53] Akay, B. A study on particle swarm optimization and artificial bee colony algorithms for multilevel thresholding. *Appl. Soft Comput.* **2013**, *13*, 3066–3091.
- [54] Ma, R.; Xu, X.; Zhao, L.; Cao, R.; Fang, Q. Mutual Artificial Bee Colony Algorithm for Molecular Docking. *Int. J. Biomath.* **2013**, *6*, 1350038.
- [55] Fuhrmann, J.; Rurainski, A.; Lenhof, H.-P.; Neumann, D. A new Lamarckian genetic algorithm for flexible ligand-receptor docking. *J. Comput. Chem.* **2010**, *31*, 1911–1918.
- [56] Huang, S.-Y.; Grinter, S. Z.; Zou, X. Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions. *Phys. Chem. Chem. Phys.* **2010**, *12*, 12899–12908.
- [57] Weikl, T. R.; Von Deuster, C. Selected-fit versus induced-fit protein binding: Kinetic differences and

Bibliography

- mutational analysis. *Proteins Struct. Funct. Bioinforma.* **2009**, *75*, 104–110.
- [58] Shoemake, K. Animating rotation with quaternion curves. *ACM SIGGRAPH Comput. Graph.* **1985**, *19*, 245–254.
- [59] Veber, D. F.; Johnson, S. R.; Cheng, H.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45*, 2615–2623.
- [60] Akay, B.; Karaboga, D. A modified Artificial Bee Colony algorithm for real-parameter optimization. *Inf. Sci. (Ny).* **2012**, *192*, 120–142.
- [61] Thomas, B.; and David, B. F.; Zbigniew, M.; Handbook of Evolutionary Computation, Oxford University Press, Oxford, **1997**, A2.3:6-A2.3:7.
- [62] Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **2009**, *30*, 2785–2791.
- [63] Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726–741.
- [64] Zavodszky, M. I.; Kuhn, L. A. Side-chain flexibility in protein-ligand binding: the minimal rotation hypothesis. *Protein Sci.* **2005**, *14*, 1104–1014.
- [65] Najmanovich, R.; Kuttner, J.; Sobolev, V.; Edelman, M. Side-chain flexibility in proteins upon ligand binding. *Proteins Struct. Funct. Genet.* **2000**, *39*, 261–268.
- [66] Shin, W. H.; Seok, C. GalaxyDock: Protein-ligand docking with flexible protein side-chains. *J. Chem. Inf. Model.* **2012**, *52*, 3225–3232.
- [67] Verdonk, M. L.; Chessari, G.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R. Modeling water molecules in protein-ligand docking using GOLD. *J. Med. Chem.* **2005**, *48*, 6504–15.
- [68] Roberts, B. C.; Mancera, R. L. Ligand-protein docking with water molecules. *J. Chem. Inf. Model.* **2008**, *48*, 397–408.
- [69] Lemmon, G.; Meiler, J. Towards Ligand Docking Including Explicit Interface Water Molecules. *PLoS One* **2013**, *8*, 1–12.
- [70] Oefner, C.; Roques, B. P.; Fournie-Zaluski, M.-C.; Dale, G. E. Structural analysis of neprilysin with various specific and potent inhibitors. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2004**, *60*, 392–396.
- [71] Poornima, C.S.; Dean, P.M. Hydration in drug design. 1. Multiple hydrogen-bonding features of water molecules in mediating protein-ligand interactions. *J. Comput. Aided Mol. Des.* **1995**, *9*, 500–512.

-
- [72] Hummer, G. Molecular binding: Under water's influence. *Nat. Chem.* **2010**, *2*, 906–907.
- [73] Baron, R.; Setny, P.; McCammon, J.A. Water in Cavity–Ligand Recognition. *J. Am. Chem. Soc.* **2010**, *132*, 12091–12097.
- [74] Chung, E.; Henriques, D.; Renzoni, D.; Zvelebil, M.; Bradshaw, J.M.; Waksman, G.; Robinson, C.V.; Ladbury, J.E. Mass spectrometric and thermodynamic studies reveal the role of water molecules in complexes formed between SH2 domains and tyrosyl phosphopeptides. *Structure* **1998**, *6*, 1141–1151.
- [75] McPhalen, C.A.; James, M.N. Structural comparison of two serine proteinase-protein inhibitor complexes: Eglin-c-subtilisin Carlsberg and CI-2-subtilisin Novo. *Biochemistry* **1988**, *27*, 6582–6598.
- [76] Quioco, F.A.; Wilson, D.K.; Vyas, N.K. Substrate specificity and affinity of a protein modulated by bound water molecules. *Nature* **1989**, *340*, 404–407.
- [77] Baron, R.; Setny, P.; Paesani, F. Water Structure, Dynamics, and Spectral Signatures: Changes upon model cavity–ligand recognition. *J. Phys. Chem. B* **2012**, *116*, 13774–13780.
- [78] Barillari, C.; Taylor, J.; Viner, R.; Essex, J.W. Classification of water molecules in protein binding sites. *J. Am. Chem. Soc.* **2007**, *129*, 2577–2587.
- [79] Michel, J.; Tirado-Rives, J.; Jorgensen, W.L. Prediction of the water content in protein binding sites. *J. Phys. Chem. B* **2009**, *113*, 13337–13346.
- [80] Raman, E.P.; Mackerell, A.D. Spatial analysis and quantification of the thermodynamic driving forces in protein-ligand binding: Binding site variability. *J. Am. Chem. Soc.* **2015**, *137*, 2608–2621.
- [81] Haider, K.; Huggins, D.J. Combining solvent thermodynamic profiles with functionality maps of the Hsp90 binding site to predict the displacement of water molecules. *J. Chem. Inf. Model.* **2013**, *53*, 2571–2586.
- [82] Li, Z.; Lazaridis, T. Thermodynamics of buried water clusters at a protein-ligand binding interface. *J. Phys. Chem. B* **2006**, *110*, 1464–1475.
- [83] Chen, J.M.; Xu, S.L.; Wawrzak, Z.; Basarab, G.S.; Jordan, D.B. Structure-based design of potent inhibitors of scytalone dehydratase: Displacement of a water molecule from the active site. *Biochemistry* **1998**, *37*, 17735–17744.
- [84] Robinson, D.; Bertrand, T.; Carry, J.-C.; Halley, F.; Karlsson, A.; Mathieu, M.; Minoux, H.; Perrin, M.-A.; Robert, B.; Schio, L.; et al. Differential Water Thermodynamics Determine PI3K-Beta/Delta Selectivity for Solvent-Exposed Ligand Modifications. *J. Chem. Inf. Model.* **2016**, *56*, 886–894.
- [85] Ladbury, J.E.; Klebe, G.; Freire, E. Adding calorimetric data to decision making in lead discovery: A hot tip. *Nat. Rev. Drug Discov.* **2010**, *9*, 23–27.
- [86] Barandun, L.J.; Ehrmann, F.R.; Zimmerli, D.; Immekus, F.; Giroud, M.; Grunenfelder, C.; Schweizer,

Bibliography

- W.B.; Bernet, B.; Betz, M.; Heine, A.; et al. Replacement of water molecules in a phosphate binding site by furanoside-appended lin-benzoguanine ligands of tRNA-guanine transglycosylase (TGT). *Chem. A Eur. J.* **2015**, *21*, 126–135.
- [87] Biela, A.; Nasief, N.N.; Betz, M.; Heine, A.; Hangauer, D.; Klebe, G. Dissecting the hydrophobic effect on the molecular level: The role of water, enthalpy, and entropy in ligand binding to thermolysin. *Angew. Chem. Int. Ed.* **2013**, *52*, 1822–1828.
- [88] Betz, M.; Wulsdorf, T.; Krimmer, S.G.; Klebe, G. Impact of surface water layers on protein-ligand binding: how well are experimental data reproduced by molecular dynamics simulations in a thermolysin test case? *J. Chem. Inf. Model.* **2016**, *56*, 223–233.
- [89] Pearlstein, R.A.; Hu, Q.Y.; Zhou, J.; Yowe, D.; Levell, J.; Dale, B.; Kaushik, V.K.; Daniels, D.; Hanrahan, S.; Sherman, W.; et al. New hypotheses about the structure-function of proprotein convertase subtilisin/kexin type 9: Analysis of the epidermal growth factor-like repeat a docking site using watermap. *Proteins Struct. Funct. Bioinform.* **2010**, *78*, 2571–2586.
- [90] Pearlstein, R.A.; Sherman, W.; Abel, R. Contributions of water transfer energy to protein-ligand association and dissociation barriers: Watermap analysis of a series of p38 α MAP kinase inhibitors. *Proteins Struct. Funct. Bioinform.* **2013**, *81*, 1509–1526.
- [91] Bortolato, A.; Tehan, B.G.; Bodnarchuk, M.S.; Essex, J.W.; Mason, J.S. Water network perturbation in ligand binding: Adenosine A2A antagonists as a case study. *J. Chem. Inf. Model.* **2013**, *53*, 1700–1713.
- [92] Barillari, C.; Duncan, A.L.; Westwood, I.M.; Blagg, J.; van Montfort, R.L.M. Analysis of water patterns in protein kinase binding sites. *Proteins Struct. Funct. Bioinforma.* **2011**, *79*, 2109–2121.
- [93] Imai, T.; Hiraoka, R.; Kovalenko, A.; Hirata, F. Locating missing water molecules in protein cavities by the three-dimensional reference interaction site model theory of molecular solvation. *Proteins Struct. Funct. Bioinform.* **2006**, *66*, 804–813.
- [94] Bayden, A.S.; Moustakas, D.T.; Joseph-McCarthy, D.; Lamb, M.L. Evaluating Free Energies of Binding and Conservation of Crystallographic Waters Using SZMAP. *J. Chem. Inf. Model.* **2015**, *55*, 1552–1565.
- [95] Goodford, P.J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.
- [96] Goodfellow, J.M.; Vovelle, F. Biomolecular energy calculations using transputer technology. *Eur. Biophys. J.* **1989**, *17*, 167–172.
- [97] Pitt, W.R.; Murray-Rust, J.; Goodfellow, J.M. AQUARIUS2: Knowledge-based modeling of solvent sites around proteins. *J. Comput. Chem.* **1993**, *14*, 1007–1018.
- [98] Verdonk, M.L.; Cole, J.C.; Taylor, R. SuperStar: A knowledge-based approach for identifying interaction sites in proteins. *J. Mol. Biol.* **1999**, *289*, 1093–1108.

-
- [99] Abel, R.; Young, T.; Farid, R.; Berne, B.J.; Friesner, R.A. Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. *J. Am. Chem. Soc.* **2008**, *130*, 2817–2831.
- [100] Li, Z.; Lazaridis, T. Computing the Thermodynamic Contributions of Interfacial Water. *Methods Mol. Biol.* **2012**, *819*, 393–404.
- [101] Hu, B.; Lill, M.A. WATsite: Hydration site prediction program with PyMOL interface. *J. Comput. Chem.* **2014**, *35*, 1255–1260.
- [102] Czapiewski, D.; Zielkiewicz, J. Structural properties of hydration shell around various conformations of simple polypeptides. *J. Phys. Chem. B* **2010**, *114*, 4536–4550.
- [103] Henchman, R.H.; McCammon, J.A. Extracting hydration sites around proteins from explicit water simulations. *J. Comput. Chem.* **2002**, *23*, 861–869.
- [104] Lazaridis, T. Inhomogeneous fluid approach to solvation thermodynamics. 1. Theory. *J. Phys. Chem. B* **1998**, *102*, 3531–3541.
- [105] Lazaridis, T. Inhomogeneous fluid approach to solvation thermodynamics. 2. Applications to Simple Fluids. *J. Phys. Chem. B* **1998**, *102*, 3542–3550.
- [106] Nguyen, C.N.; Cruz, A.; Gilson, M.K.; Kurtzman, T. Thermodynamics of water in an enzyme active site: Grid-based hydration analysis of coagulation factor xa. *J. Chem. Theory Comput.* **2014**, *10*, 2769–2780.
- [107] Meng, X.-Y.; Zhang, H.-X.; Mezei, M.; Cui, M. Molecular docking: A powerful approach for structure-based drug discovery. *Curr. Comput. Aided-Drug Des.* **2011**, *7*, 146–157.
- [108] Yuriev, E.; Agostino, M.; Ramsland, P.A. Challenges and advances in computational docking: 2009 in review. *J. Mol. Recognit.* **2011**, *24*, 149–164.
- [109] Okimoto, N.; Futatsugi, N.; Fuji, H.; Suenaga, A.; Morimoto, G.; Yanai, R.; Ohno, Y.; Narumi, T.; Taiji, M. High-performance drug discovery: Computational screening by combining docking and molecular dynamics simulations. *PLoS Comput. Biol.* **2009**, *5*, e1000528.
- [110] Graaf, D.C.; Pospisil, P.; Pos, W.; Folkers, G.; Vermeulen, N.P.E. Binding mode prediction of cytochrome P450 and thymidine kinase protein-ligand complexes by consideration of water and rescoring in automated docking. *J. Med. Chem.* **2005**, *48*, 2308–2318.
- [111] Huang, S.; Zou, X. Inclusion of solvation and entropy in the knowledge-based scoring function for protein—Ligand interactions. *J. Chem. Inf. Model.* **2010**, *50*, 262–273.
- [112] Kumar, A.; Zhang, K.Y.J. Investigation on the effect of key water molecules on docking performance in CSARdock exercise. *J. Chem. Inf. Model.* **2013**, *53*, 1880–1892.
- [113] Huang, N.; Shoichet, B.K. Exploiting ordered waters in molecular docking. *J. Med. Chem.* **2008**, *51*, 4862–4865.

Bibliography

- [114] Forli, S.; Olson, A.J. A force field with discrete displaceable waters and desolvation entropy for hydrated ligand docking. *J. Med. Chem.* **2012**, *55*, 623–638.
- [115] Mysinger, M.M.; Shoichet, B.K. Rapid context-dependent ligand desolvation in molecular docking. *J. Chem. Inf. Model.* **2010**, *50*, 1561–1573.
- [116] Shoichet, B.K.; Leach, A.R.; Kuntz, I.D. Ligand solvation in molecular docking. *Proteins Struct. Funct. Genet.* **1999**, *34*, 4–16.
- [117] Alvarez-Garcia, D.; Barril, X. Molecular simulations with solvent competition quantify water displaceability and provide accurate interaction maps of protein binding sites. *J. Med. Chem.* **2014**, *57*, 8530–8539.
- [118] Goodsell, D.S.; Olson, A.J. Automated docking of substrates to proteins by simulated annealing. *Proteins Struct. Funct. Genet.* **1990**, *8*, 195–202.
- [119] Mysinger, M.M.; Carchia, M.; Irwin, J.J.; Shoichet, B.K. Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *J. Med. Chem.* **2012**, *55*, 6582–6594.
- [120] Humphrey, W.; Dalke, A.; Schulten, K. VDM: Visual molecular dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38.
- [121] Young, T.; Hua, L.; Huang, X.; Abel, R.; Friesner, R.; Berne, B.J. Dewetting transitions in protein cavities. *Proteins Struct. Funct. Bioinform.* **2010**, *78*, 1856–1869.
- [122] Uehara, S.; Fujimoto, K.J.; Tanaka, S. Protein-ligand docking using fitness learning-based artificial bee colony with proximity stimuli. *Phys. Chem. Chem. Phys.* **2015**, *17*, 16412–16417.
- [123] Santos-Martins, D.; Forli, S.; Ramos, M.J.; Olson, A.J. AutoDock4Zn: An improved AutoDock force field for small-molecule docking to zinc metalloproteins. *J. Chem. Inf. Model.* **2014**, *54*, 2371–2379.
- [124] Weiner, S.J.; Kollman, P.A.; Case, D.A.; Singh, U.C.; Ghio, C.; Alagona, G.; Profeta, S.; Weinerl, P. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **1984**, *106*, 765–784.
- [125] Huey, R.; Goodsell, D.; Morris, G.; Olson, A. Grid-based hydrogen bond potentials with improved directionality. *Lett. Drug Des. Discov.* **2004**, *1*, 178–183.
- [126] Mehler, E.L.; Solmajer, T. Electrostatic effects in proteins: Comparison of dielectric and charge models. *Protein Eng. Des. Sel.* **1991**, *4*, 903–910.
- [127] Stouten, P.F.W.; Frömmel, C.; Nakamura, H.; Sander, C. An Effective Solvation Term Based on Atomic Occupancies for Use in Protein Simulations. *Mol. Simul.* **1993**, *10*, 97–120.
- [128] Sgobba, M.; Caporuscio, F.; Anighoro, A.; Portioli, C.; Rastelli, G. Application of a post-docking procedure based on MM-PBSA and MM-GBSA on single and multiple protein conformations. *Eur. J. Med. Chem.* **2012**, *58*, 431–440.

-
- [129] Thompson, D.C.; Humblet, C.; Joseph-McCarthy, D. Investigation of MM-PBSA rescoring of docking poses. *J. Chem. Inf. Model.* **2008**, *48*, 1081–1091.
- [130] Sun, H.; Zhao, L.; Peng, S.; Huang, N. Incorporating replacement free energy of binding-site waters in molecular docking. *Proteins Struct. Funct. Bioinform.* **2014**, *82*, 1765–1776.
- [131] Adler, M.; Davey, D.D.; Phillips, G.B.; Kim, S.H.; Jancarik, J.; Rumennik, G.; Light, D.R.; Whitlow, M. Preparation, characterization, and the crystal structure of the inhibitor ZK-807834 (CI-1031) complexed with factor Xa. *Biochemistry* **2000**, *39*, 12534–12542.
- [132] Word, J.M.; Lovell, S.C.; Richardson, J.S.; Richardson, D.C. Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* **1999**, *285*, 1735–1747.
- [133] Case, D.A.; Babin, V.; Berryman, J.T.; Betz, R.M.; Cai, Q.; Cerutti, D.S.; Cheatham, T.E., III; Darden, T.A.; Duke, R.E.; Gohlke, H.; et al. AMBER 14; University of California: San Francisco, CA, USA, 2014.
- [134] Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins Struct. Funct. Bioinform.* **2006**, *65*, 712–725.
- [135] Jorgensen, W.L.; Chandrasekhar, J.; Madura, J.D.; Impey, R.W.; Klein, M.L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.
- [136] Salomon-Ferrer, R.; Götz, A.W.; Poole, D.; Le Grand, S.; Walker, R.C. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh ewald. *J. Chem. Theory Comput.* **2013**, *9*, 3878–3888.
- [137] Ryckaert, J.P.; Ciccotti, G.; Berendsen, H.J.C. Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.
- [138] Essmann, U.; Perera, L.; Berkowitz, M.L.; Darden, T.; Lee, H.; Pedersen, L.G. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577–8593.
- [139] Roe, D.R.; Cheatham, T.E., III; PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* **2013**, *9*, 3084–3095.
- [140] Ramsey, S.; Nguyen, C.; Salomon-Ferrer, R.; Walker, R.C.; Gilson, M.K.; Kurtzman, T. Solvation thermodynamic mapping of molecular surfaces in AmberTools: GIST. *J. Comput. Chem.* **2016**, *37*, 2029–2037.
- [141] Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R. PDB-wide collection of binding data: Current status of the PDBbind database. *Bioinformatics* **2014**, *31*, 405–412.
- [142] Pan, Y.; Huang, N.; Cho, S.; MacKerell, A.D. Consideration of molecular weight during compound

Bibliography

- selection in virtual target-based database screening. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 267–272.
- [143] Kim, R.; Skolnick, J. Assessment of programs for ligand binding affinity prediction. *J. Comput. Chem.* **2008**, *29*, 1316–1331.
- [144] *Molecular Operating Environment (MOE), 2013.08*; Chemical Computing Group Inc.: Montreal, QC, Canada, 2016.
- [145] Triballeau, N.; Acher, F.; Brabet, I.; Pin, J.P.; Bertrand, H.O. Virtual screening workflow development guided by the “receiver operating characteristic” curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* **2005**, *48*, 2534–2547.
- [146] Bender, A.; Glen, R.C. A discussion of measures of enrichment in virtual screening: Comparing the information content of descriptors with increasing levels of sophistication. *J. Chem. Inf. Model.* **2005**, *45*, 1369–1375.
- [147] Perzborn, E.; Roehrig, S.; Straub, A.; Kubitz, D.; Misselwitz, F. The discovery and development of rivaroxaban, an oral, direct factor Xa inhibitor. *Nat. Rev. Drug Discov.* **2011**, *10*, 61–75.
- [148] Rai, R.; Sprengeler, P.; Elrod, K.; Young, W. Perspectives on Factor Xa Inhibition. *Curr. Med. Chem.* **2001**, *8*, 101–119.
- [149] Lumry, R.; Rajender, S. Enthalpy-entropy compensation phenomena in water solutions of proteins and small molecules: A ubiquitous property of water. *Biopolymers* **1970**, *9*, 1125–1227.
- [150] Meloun, M.; Ferenčíková, Z. Enthalpy–entropy compensation for some drugs dissociation in aqueous solutions. *Fluid Phase Equilib.* **2012**, *328*, 31–41.
- [151] Ahmad, M.; Helms, V.; Lengauer, T.; Kalinina, O.V. Enthalpy-entropy compensation upon molecular conformational changes. *J. Chem. Theory Comput.* **2015**, *11*, 1410–1418.
- [152] Friesner, R.A.; Murphy, R.B.; Repasky, M.P.; Frye, L.L.; Greenwood, J.R.; Halgren, T.A.; Sanschagrin, P.C.; Mainz, D.T. Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.* **2006**, *49*, 6177–6196.
- [153] Kadirvelraj, R.; Foley, B.L.; Dyekjær, J.D.; Woods, R.J. Involvement of water in carbohydrate–protein binding: concanavalin a revisited. *J. Am. Chem. Soc.* **2008**, *130*, 16933–16942.
- [154] Li, Z.; Lazaridis, T. Thermodynamic contributions of the ordered water molecule in HIV-1 protease. *J. Am. Chem. Soc.* **2003**, *125*, 6636–6637.
- [155] Michel, J.; Tirado-Rives, J.; Jorgensen, W.L. Energetics of displacing water molecules from protein binding sites: Consequences for ligand optimization. *J. Am. Chem. Soc.* **2009**, *131*, 15403–15411.
- [156] Baldwin, E.T.; Bhat, T.N.; Gulnik, S.; Liu, B.; Topol, I.A.; Kiso, Y.; Mimoto, T.; Mitsuya, H.; Erickson, J.W. Structure of HIV-1 protease with KNI-272, a tight-binding transition-state analog containing allophenylnorstatine. *Structure* **1995**, *3*, 581–590.

-
- [157] Blum, A.P.; Lester, H.A.; Dougherty, D.A. Nicotinic pharmacophore: The pyridine N of nicotine and carbonyl of acetylcholine hydrogen bond across a subunit interface to a backbone NH. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 13206–13211.
- [158] Adachi, M.; Ohhara, T.; Kurihara, K.; Tamada, T.; Honjo, E.; Okazaki, N.; Arai, S.; Shoyama, Y.; Kimura, K.; Matsumura, H.; et al. Structure of HIV-1 protease in complex with potent inhibitor KNI-272 determined by high-resolution X-ray and neutron crystallography. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 4641–4646.
- [159] Murphy, R.B.; Repasky, M.P.; Greenwood, J.R.; Tubert-Brohman, I.; Jerome, S.; Annabhimoju, R.; Boyles, N.A.; Schmitz, C.D.; Abel, R.; Farid, R.; et al. WScore: A flexible and accurate treatment of explicit water molecules in ligand–receptor docking. *J. Med. Chem.* **2016**, *59*, 4364–4384.
- [160] Lavecchia, A.; Di Giovanni, C. Virtual screening strategies in drug discovery: A critical review. *Curr. Med. Chem.* **2013**, *20*, 2839–2860.
- [161] Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. Comparison of several molecular docking programs: Pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* **2009**, *49*, 1455–1474.
- [162] Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins Struct. Funct. Genet.* **2004**, *57*, 225–242.
- [163] Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.
- [164] Hu, G.; Li, X.; Zhang, X.; Li, Y.; Ma, L.; Yang, L. M.; Liu, G.; Li, W.; Huang, J.; Shen, X.; Hu, L.; Zheng, Y. T.; Tang, Y. Discovery of inhibitors to block interactions of HIV-1 integrase with human LEDGF/p75 via structure-based virtual screening and bioassays. *J. Med. Chem.* **2012**, *55*, 10108–10117.
- [165] Nose, T.; Tokunaga, T.; Shimohigashi, Y. Exploration of endocrine-disrupting chemicals on estrogen receptor α by the agonist/antagonist differential-docking screening (AADS) method: 4-(1-Adamantyl)phenol as a potent endocrine disruptor candidate. *Toxicol. Lett.* **2009**, *191*, 33–39.
- [166] Caporuscio, F.; Rastelli, G.; Imbriano, C.; Del Rio, A. Structure-based design of potent aromatase inhibitors by high-throughput docking. *J. Med. Chem.* **2011**, *54*, 4006–4017.
- [167] Schapira, M.; Abagyan, R.; Totrov, M. Nuclear hormone receptor targeted virtual screening. *J. Med. Chem.* **2003**, *46*, 3045–3059.
- [168] Xing, L.; McDonald, J. J.; Kolodziej, S. A.; Kurumbail, R. G.; Williams, J. M.; Warren, C. J.; O’Neal, J. M.; Skepner, J. E.; Roberds, S. L. Discovery of Potent Inhibitors of Soluble Epoxide Hydrolase by Combinatorial Library Design and Structure-Based Virtual Screening. *J. Med. Chem.* **2011**, *54*, 1211–1222.

Bibliography

- [169] Vidler, L. R.; Filippakopoulos, P.; Fedorov, O.; Picaud, S.; Martin, S.; Tomsett, M.; Woodward, H.; Brown, N.; Knapp, S.; Hoelder, S. Discovery of novel small-molecule inhibitors of BRD4 using structure-based virtual screening. *J. Med. Chem.* **2013**, *56*, 8073–8088.
- [170] Rodrigues, T.; Moreira, R.; Gut, J.; Rosenthal, P. J.; O'Neill, P. M.; Biagini, G. A.; Lopes, F.; Dos Santos, D. J. V. A.; Guedes, R. C. Identification of new antimalarial leads by use of virtual screening against cytochrome bc 1. *Bioorganic Med. Chem.* **2011**, *19*, 6302–6308.
- [171] Feng, J.; Jin, K.; Zhu, H.; Zhang, X.; Zhang, L.; Liu, J.; Xu, W. A novel aminopeptidase N inhibitor developed by virtual screening approach. *Bioorganic Med. Chem. Lett.* **2012**, *22*, 5863–5869.
- [172] Chen, Y.; Wen, D.; Huang, Z.; Huang, M.; Luo, Y.; Liu, B.; Lu, H.; Wu, Y.; Peng, Y.; Zhang, J. 2-(4-Chlorophenyl)-2-oxoethyl 4-benzamidobenzoate derivatives, a novel class of SENP1 inhibitors: Virtual screening, synthesis and biological evaluation. *Bioorganic Med. Chem. Lett.* **2012**, *22*, 6867–6870.
- [173] Cavasotto, C. N.; Abagyan, R. A. Protein Flexibility in Ligand Docking and Virtual Screening to Protein Kinases. *J. Mol. Biol.* **2004**, *337*, 209–225.
- [174] Verdonk, M. L.; Mortenson, P. N.; Hall, R. J.; Hartshorn, M. J.; Murray, C. W. Protein-ligand docking against non-native protein conformers. *J. Chem. Inf. Model.* **2008**, *48*, 2214–2225.
- [175] Murray, C. W.; Baxter, C. A.; Frenkel, A. D. The sensitivity of the results of molecular docking to induced fit effects: application to thrombin, thermolysin and neuraminidase. *J. Comput. Aided. Mol. Des.* **1999**, *13*, 547–562.
- [176] Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* **2006**, *49*, 534–553.
- [177] Erickson, J. A.; Jalaie, M.; Robertson, D. H.; Lewis, R. A.; Vieth, M. Lessons in Molecular Recognition: The Effects of Ligand and Protein Flexibility on Molecular Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 45–55.
- [178] Totrov, M.; Abagyan, R. Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr. Opin. Struct. Biol.* **2008**, *18*, 178–184.
- [179] B-Rao, C.; Subramanian, J.; Sharma, S. D. Managing protein flexibility in docking and its applications. *Drug Discov. Today* **2009**, *14*, 394–400.
- [180] A. Sotriffer, C. Accounting for Induced-Fit Effects in Docking: What is Possible and What is Not? *Curr. Top. Med. Chem.* **2011**, *11*, 179–191.
- [181] Lill, M. A. Efficient incorporation of protein flexibility and dynamics into molecular docking simulations. *Biochemistry* **2011**, *50*, 6157–6169.
- [182] Bottegoni, G.; Kufareva, I.; Totrov, M.; Abagyan, R. A new method for ligand docking to flexible receptors by dual alanine scanning and refinement (SCARE). *J. Comput. Aided. Mol. Des.* **2008**, *22*,

-
- 311–325.
- [183] Broughton, H. B. A method for including protein flexibility in protein-ligand docking: Improving tools for database mining and virtual screening. *J. Mol. Graph. Model.* **2000**, *18*, 247–257.
- [184] Carlson, H. A.; McCammon, J. A. Accommodating protein flexibility in computational drug design. *Mol. Pharmacol.* **2000**, *57*, 213–218.
- [185] Carlson, H. A. Protein flexibility and drug design: How to hit a moving target. *Curr. Opin. Chem. Biol.* **2002**, *6*, 447–452.
- [186] Koshland, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci. U. S. A.* **1958**, *44*, 98–104.
- [187] Frauenfelder, H.; Sligar, S.; Wolynes, P. The energy landscapes and motions of proteins. *Science.* **1991**, *254*, 1598–1603.
- [188] Tsai, C. J.; Tsai, C. J.; Kumar, S.; Kumar, S.; Ma, B.; Ma, B.; Nussinov, R.; Nussinov, R. Folding funnels, binding funnels, and protein function. *Protein Sci* **1999**, *8*, 1181–1190.
- [189] Ma, B.; Kumar, S.; Tsai, C. J.; Nussinov, R. Folding funnels and binding mechanisms. *Protein Eng.* **1999**, *12*, 713–720.
- [190] Leach, A. R. Ligand docking to proteins with discrete side-chain flexibility. *J. Mol. Biol.* **1994**, *235*, 345–356.
- [191] Kokh, D. B.; Wenzel, W. Flexible side chain models improve enrichment rates in in silico screening. *J. Med. Chem.* **2008**, *51*, 5919–5931.
- [192] Sandak, B.; Wolfson, H. J.; Nussinov, R. Flexible docking allowing induced-fit in proteins: Insights from an open to closed isomers. *Proteins Struct. Funct. Genet.* **1998**, *32*, 159–174.
- [193] Koska, J.; Spassov, V. Z.; Maynard, A. J.; Yan, L.; Austin, N.; Flook, P. K.; Venkatachalam, C. M. Fully Automated Molecular Mechanics Based Induced Fit Protein-Ligand Docking Method. *J. Chem. Inf. Model* **2008**, *48*, 1965–1973.
- [194] De Paris, R.; Frantz, F. A.; Norberto De Souza, O.; Ruiz, D. D. A. WFRoDoW: A cloud-based web environment to handle molecular docking simulations of a fully flexible receptor model. *Biomed Res. Int.* **2013**, *2013*, 1-12.
- [195] Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins Struct. Funct. Genet.* **2002**, *47*, 409–443.
- [196] Yoon, S.; Welsh, W. J. Identification of a Minimal Subset of Receptor Conformations for Improved Multiple Conformation Docking and Two-Step Scoring. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 88–96.
- [197] Carlson, H. Protein Flexibility is an Important Component of Structure-Based Drug Discovery. *Curr. Pharm. Des.* **2002**, *8*, 1571–1578.

Bibliography

- [198] Teodoro, M.; Kavraki, L. Conformational Flexibility Models for the Receptor in Structure Based Drug Design. *Curr. Pharm. Des.* **2003**, *9*, 1635–1648.
- [199] Korb, O.; Olsson, T. S. G.; Bowden, S. J.; Hall, R. J.; Verdonk, M. L.; Liebeschuetz, J. W.; Cole, J. C. Potential and limitations of ensemble docking. *J. Chem. Inf. Model.* **2012**, *52*, 1262–1274.
- [200] Evangelista, W.; Weir, R. L.; Ellingson, S. R.; Harris, J. B.; Kapoor, K.; Smith, J. C.; Baudry, J. Ensemble-based docking: From hit discovery to metabolism and toxicity predictions. *Bioorg. Med. Chem.* **2016**, *24*, 4928–4935.
- [201] Park, S. J.; Kufareva, I.; Abagyan, R. Improved docking, screening and selectivity prediction for small molecule nuclear receptor modulators using conformational ensembles. *J. Comput. Aided. Mol. Des.* **2010**, *24*, 459–471.
- [202] Tian, S.; Sun, H.; Pan, P.; Li, D.; Zhen, X.; Li, Y.; Hou, T. Assessing an Ensemble Docking-Based Virtual Screening Strategy for Kinase Targets by Considering Protein Flexibility. *J. Chem. Inf. Model.* **2014**, *54*, 2664–2679.
- [203] Rao, S.; Sanschagrin, P. C.; Greenwood, J. R.; Repasky, M. P.; Sherman, W.; Farid, R. Improving database enrichment through ensemble docking. *J. Comput. Aided. Mol. Des.* **2008**, *22*, 621–627.
- [204] Yadav, I. S.; Nandekar, P. P.; Shrivastava, S.; Sangamwar, A.; Chaudhury, A.; Agarwal, S. M. Ensemble docking and molecular dynamics identify knoevenagel curcumin derivatives with potent anti-EGFR activity. *Gene* **2014**, *539*, 82–90.
- [205] Huang, S.; Zou, X. Efficient molecular docking of NMR structures: Application to HIV-1 protease. *Protein Sci* **2007**, *16*, 43–51.
- [206] Limongelli, V.; Marinelli, L.; Cosconati, S.; Braun, H. A.; Schmidt, B.; Novellino, E. Ensemble-docking approach on BACE-1: Pharmacophore perception and guidelines for drug design. *ChemMedChem* **2007**, *2*, 667–678.
- [207] Hritz, J.; De Ruiter, A.; Oostenbrink, C. Impact of plasticity and flexibility on docking results for cytochrome P450 2D6: A combined approach of molecular dynamics and ligand docking. *J. Med. Chem.* **2008**, *51*, 7469–7477.
- [208] Bolstad, E. S. D.; Anderson, A. C. In pursuit of virtual lead optimization: The role of the receptor structure and ensembles in accurate docking. *Proteins Struct. Funct. Genet.* **2008**, *73*, 566–580.
- [209] Huang, S.-Y.; Zou, X. Ensemble docking of multiple protein structures: Considering protein structural variations in molecular docking. *Proteins Struct. Funct. Bioinforma.* **2006**, *66*, 399–421.
- [210] Craig, I. R.; Essex, J. W.; Spiegel, K. Ensemble docking into multiple crystallographically derived protein structures: An evaluation based on the statistical analysis of enrichments. *J. Chem. Inf. Model.* **2010**, *50*, 511–524.
- [211] Barril, X.; Morley, S. D. Unveiling the full potential of flexible receptor docking using multiple

-
- crystallographic structures. *J. Med. Chem.* **2005**, *48*, 4432–4443.
- [212] Ferrari, A. M.; Wei, B. Q.; Costantino, L.; Shoichet, B. K. Soft docking and multiple receptor conformations in virtual screening. *J. Med. Chem.* **2004**, *47*, 5076–5084.
- [213] Abagyan, R.; Rueda, M.; Bottegoni, G. Recipes for the selection of experimental protein conformations for virtual screening. *J. Chem. Inf. Model.* **2010**, *50*, 186–193.
- [214] Fan, H.; Irwin, J. J.; Webb, B. M.; Klebe, G.; Shoichet, B. K.; Sali, A. Molecular docking screens using comparative models of proteins. *J. Chem. Inf. Model.* **2009**, *49*, 2512–2527.
- [215] Novoa, E. M.; De Pouplana, L. R.; Barril, X.; Orozco, M. Ensemble docking from homology models. *J. Chem. Theory Comput.* **2010**, *6*, 2547–2557.
- [216] Bolstad, E. S. D.; Anderson, A. C. In pursuit of virtual lead optimization: Pruning ensembles of receptor structures for increased efficiency and accuracy during docking. *Proteins Struct. Funct. Bioinforma.* **2009**, *75*, 62–74.
- [217] Park, I. H.; Li, C. Dynamic ligand-induced-fit simulation via enhanced conformational samplings and ensemble dockings: A survivin example. *J. Phys. Chem. B* **2010**, *114*, 5144–5153.
- [218] Cheng, L. S.; Amaro, R. E.; Xu, D.; Li, W. W.; Arzberger, P. W.; McCammon, J. A. Ensemble-based virtual screening reveals potential novel antiviral compounds for avian influenza neuraminidase. *J. Med. Chem.* **2008**, *51*, 3878–3894.
- [219] Wong, C. F.; Kua, J.; Zhang, Y.; Straatsma, T. P.; McCammon, J. A. Molecular docking of balanol to dynamics snapshots of protein kinase A. *Proteins Struct. Funct. Bioinforma.* **2005**, *61*, 850–858.
- [220] Wassman, C. D.; Baronio, R.; Demir, Ö.; Wallentine, B. D.; Chen, C.-K.; Hall, L. V.; Salehi, F.; Lin, D.-W.; Chung, B. P.; Hatfield, G. W.; Richard Chamberlin, A.; Luecke, H.; Lathrop, R. H.; Kaiser, P.; Amaro, R. E. Computational identification of a transiently open L1/S3 pocket for reactivation of mutant p53. *Nat. Commun.* **2013**, *4*, 1407.
- [221] Nichols, S. E.; Baron, R.; Ivetac, A.; McCammon, J. A. Predictive power of molecular dynamics receptor structures in virtual screening. *J. Chem. Inf. Model.* **2011**, *51*, 1439–1446.
- [222] Gao, C.; Uzelac, I.; Gottfries, J.; Eriksson, L. A. Exploration of multiple Sortase A protein conformations in virtual screening. *Sci. Rep.* **2016**, *6*, 20413.
- [223] Ellingson, S. R.; Miao, Y.; Baudry, J.; Smith, J. C. Multi-conformer ensemble docking to difficult protein targets. *J. Phys. Chem. B* **2015**, *119*, 1026–34.
- [224] Xu, M.; Lill, M. A. Utilizing experimental data for reducing ensemble size in flexible-protein docking. *J. Chem. Inf. Model.* **2012**, *52*, 187–198.
- [225] Li, Y.; Kim, D. J.; Ma, W.; Lubet, R. A.; Bode, A. M.; Dong, Z. Discovery of Novel Checkpoint Kinase 1 Inhibitors by Virtual Screening Based on Multiple Crystal Structures. *J. Chem. Inf. Model.* **2011**, *51*, 2904–2914.

Bibliography

- [226] Huang, Z.; Wong, C. F. Inexpensive Method for Selecting Receptor Structures for Virtual Screening. *J. Chem. Inf. Model.* **2016**, *56*, 21–34.
- [227] Seco, J.; Luque, F. J.; Barril, X. Binding site detection and druggability index from first principles. *J. Med. Chem.* **2009**, *52*, 2363–2371.
- [228] Guvench, O.; MacKerell, A. D. Computational fragment-based binding site identification by ligand competitive saturation. *PLoS Comput. Biol.* **2009**, *5*, e1000435.
- [229] Lexa, K. W.; Carlson, H. A. Full protein flexibility is essential for proper hot-spot mapping. *J. Am. Chem. Soc.* **2011**, *133*, 200–202.
- [230] Prakash, P.; Hancock, J. F.; Gorfe, A. A. Binding hotspots on K-ras: Consensus ligand binding sites and other reactive regions from probe-based molecular dynamics analysis. *Proteins Struct. Funct. Bioinforma.* **2015**, *83*, 898–909.
- [231] Yang, C.-Y.; Wang, S. Computational analysis of protein hotspots. *ACS Med. Chem. Lett.* **2010**, *1*, 125–129.
- [232] Bakan, A.; Nevins, N.; Lakdawala, A. S.; Bahar, I. Druggability Assessment of Allosteric Proteins by Dynamics Simulations in the Presence of Probe Molecules. *J. Chem. Theory Comput.* **2012**, *8*, 2435–2447.
- [233] Huang, D.; Caflisch, A. Small Molecule Binding to Proteins: Affinity and Binding/Unbinding Dynamics from Atomistic Simulations. *ChemMedChem* **2011**, *6*, 1578–1580.
- [234] Basse, N.; Kaar, J. L.; Settanni, G.; Joerger, A. C.; Rutherford, T. J.; Fersht, A. R. Toward the Rational Design of p53-Stabilizing Drugs: Probing the Surface of the Oncogenic Y220C Mutant. *Chem. Biol.* **2010**, *17*, 46–56.
- [235] Tan, Y. S.; Spring, D. R.; Abell, C.; Verma, C. S. The Application of Ligand-Mapping Molecular Dynamics Simulations to the Rational Design of Peptidic Modulators of Protein-Protein Interactions. *J. Chem. Theory Comput.* **2015**, *11*, 3199–3210.
- [236] Ghanakota, P.; Carlson, H. A. Moving Beyond Active-Site Detection: MixMD Applied to Allosteric Systems. *J. Phys. Chem. B* **2016**, *120*, 8685–8695.
- [237] Yang, C.-Y. Identification of Potential Small Molecule Allosteric Modulator Sites on IL-1R1 Ectodomain Using Accelerated Conformational Sampling Method. *PLoS One* **2015**, *10*, e0118671.
- [238] Bauer, M. R.; Ibrahim, T. M.; Vogel, S. M.; Boeckler, F. M. Evaluation and optimization of virtual screening workflows with DEKOIS 2.0 - A public library of challenging docking benchmark sets. *J. Chem. Inf. Model.* **2013**, *53*, 1447–1462.
- [239] Labute, P. Protonate3D: Assignment of ionization states and hydrogen coordinates to macromolecular structures. *Proteins Struct. Funct. Bioinforma.* **2009**, *75*, 187–205.
- [240] Ung, P. M. U.; Ghanakota, P.; Graham, S. E.; Lexa, K. W.; Carlson, H. A. Identifying binding hot

-
- spots on protein surfaces by mixed-solvent molecular dynamics: HIV-1 protease as a test case. *Biopolymers* **2016**, *105*, 21–34.
- [241] Tan, Y. S.; Śledź, P.; Lang, S.; Stubbs, C. J.; Spring, D. R.; Abell, C.; Best, R. B. Using ligand-mapping simulations to design a ligand selectively targeting a cryptic surface pocket of polo-like kinase 1. *Angew. Chemie - Int. Ed.* **2012**, *51*, 10078–10081.
- [242] Raman, E. P.; Yu, W.; Lakkaraju, S. K.; Mackerell, A. D. Inclusion of multiple fragment types in the site identification by ligand competitive saturation (SILCS) approach. *J. Chem. Inf. Model.* **2013**, *53*, 3384–3398.
- [243] Martínez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M. PACKMOL: A package for building initial configurations for molecular dynamics simulations. *J. Comput. Chem.* **2009**, *30*, 2157–2164.
- [244] Case D. A.; Betz, R. M.; Cerutti, D. S.; Cheatham, T. E., III; Darden, T. A.; Duke, R. E.; Giese, T. J.; Gohlke, H.; et. al. *AMBER 2016*, University of California, San Francisco, 2016.
- [245] Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.
- [246] Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
- [247] Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics* **2009**, *10*, 168.
- [248] Trott, O.; Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2009**, *31*, 455–461.
- [249] Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, 198–201.
- [250] Irwin, J. J.; Shoichet, B. K. ZINC – A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model* **2005**, *45*, 177–182.
- [251] Österberg, F.; Morris, G. M.; Sanner, M. F.; Olson, A. J.; Goodsell, D. S. Automated docking to multiple target structures: Incorporation of protein mobility and structural water heterogeneity in autodock. *Proteins Struct. Funct. Genet.* **2002**, *46*, 34–40.
- [252] Bottegoni, G.; Kufareva, I.; Totrov, M.; Abagyan, R. Four-dimensional docking: A fast and accurate account of discrete receptor flexibility in ligand docking. *J. Med. Chem.* **2009**, *52*, 397–406.
- [253] Swift, R. V.; McCammon, J. A. Substrate induced population shifts and stochastic gating in the PBCV-1 mRNA capping enzyme. *J. Am. Chem. Soc.* **2009**, *131*, 5126–5133.
- [254] Paulsen, J. L.; Anderson, A. C. Scoring ensembles of docked protein:ligand interactions for virtual lead optimization. *J. Chem. Inf. Model.* **2009**, *49*, 2813–2819.

Bibliography

- [255] Truchon, J. F.; Bayly, C. I. Evaluating virtual screening methods: Good and bad metrics for the “early recognition” problem. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.
- [256] Vagenende, V.; Yap, M. G. S.; Trout, B. L. Mechanisms of Protein Stabilization and Prevention of Protein Aggregation by Glycerol. *Biochemistry* **2009**, *48*, 11084–11096.
- [257] Maignan, S.; Guilloteau, J.-P.; Choi-Sledeski, Y.M.; Becker, M.R.; Ewing, W.R.; Pauls, H.W.; Spada, A.P.; Mikol, V. Molecular structures of human factor Xa complexed with ketopiperazine inhibitors: Preference for a neutral group in the S1 pocket. *J. Med. Chem.* **2003**, *46*, 685–690.
- [258] Maignan, S.; Guilloteau, J.-P.; Pouzieux, S.; Choi-Sledeski, Y.M.; Becker, M.R.; Klein, S.I.; Ewing, W.R.; Pauls, H.W.; Spada, A.P.; Mikol, V. Crystal structures of human factor Xa complexed with potent inhibitors. *J. Med. Chem.* **2000**, *43*, 3226–3232.
- [259] Adler, M.; Davey, D.D.; Phillips, G.B.; Kim, S.H.; Jancarik, J.; Rumennik, G.; Light, D.R.; Whitlow, M. Preparation, characterization, and the crystal structure of the inhibitor ZK-807834 (CI-1031) complexed with factor Xa. *Biochemistry* **2000**, *39*, 12534–12542.
- [260] Nar, H.; Bauer, M.; Schmid, A.; Stassen, J.-M.; Wienen, W.; Priepke, H.W.; Kauffmann, I.K.; Ries, U.J.; Huel, N.H. Structural basis for inhibition promiscuity of dual specific thrombin and factor Xa blood coagulation inhibitors. *Structure* **2001**, *9*, 29–37.
- [261] Guertin, K.R.; Gardner, C.J.; Klein, S.I.; Zulli, A.L.; Czekaj, M.; Gong, Y.; Spada, A.P.; Cheney, D.L.; Maignan, S.; Guilloteau, J.-P.; et al. Optimization of the β -Aminoester class of factor Xa inhibitors. Part 2: Identification of FXV673 as a potent and selective inhibitor with excellent In vivo anticoagulant activity. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 1671–1674.
- [262] Matter, H.; Defossa, E.; Heinelt, U.; Blohm, P.-M.; Schneider, D.; Müller, A.; Herok, S.; Schreuder, H.; Liesum, A.; Brachvogel, V.; et al. Design and quantitative structure-activity relationship of 3-amidinobenzyl-1 H -indole-2-carboxamides as potent, nonchiral, and selective inhibitors of blood coagulation factor Xa. *J. Med. Chem.* **2002**, *45*, 2749–2769.
- [263] Ye, B.; Arnaiz, D.O.; Chou, Y.-L.; Griedel, B.D.; Karanjawala, R.; Lee, W.; Morrissey, M.M.; Sacchi, K.L.; Sakata, S.T.; Shaw, K.J.; et al. Thiophene-anthranilamides as highly potent and orally available factor Xa inhibitors 1. *J. Med. Chem.* **2007**, *50*, 2967–2980.
- [264] Quan, M.L.; Lam, P.Y.S.; Han, Q.; Pinto, D.J.P.; He, M.Y.; Li, R.; Ellis, C.D.; Clark, C.G.; Teleha, C.A.; Sun, J.-H.; et al. Discovery of 1-(3'-Aminobenzisoxazol-5'-yl)-3-trifluoromethyl-N-[2-fluoro-4-[(2'-dimethylaminomethyl)imidazol-1-yl]phenyl]-1H-pyrazole-5-carboxamide Hydrochloride (Razaxaban), a highly potent, selective, and orally bioavailable factor Xa inhibitor. *J. Med. Chem.* **2005**, *48*, 1729–1744.
- [265] Matter, H.; Will, D.W.; Nazaré, M.; Schreuder, H.; Laux, V.; Wehner, V. Structural Requirements for factor Xa inhibition by 3-Oxybenzamides with neutral P1 substituents: Combining X-ray crystallography, 3D-QSAR, and tailored scoring functions. *J. Med. Chem.* **2005**, *48*, 3290–3312.
- [266] Nazaré, M.; Essrich, M.; Will, D.W.; Matter, H.; Ritter, K.; Urmann, M.; Bauer, A.; Schreuder, H.;

-
- Dudda, A.; Czech, J.; et al. Factor Xa inhibitors based on a 2-carboxyindole scaffold: SAR of neutral P1 substituents. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 4191–4195.
- [267] Schärer, K.; Morgenthaler, M.; Paulini, R.; Obst-Sander, U.; Banner, D.W.; Schlatter, D.; Benz, J.; Stihle, M.; Diederich, F. Quantification of Cation- π interactions in protein-ligand complexes: Crystal-structure analysis of factor Xa bound to a quaternary ammonium ion ligand. *Angew. Chem. Int. Ed.* **2005**, *44*, 4400–4404.
- [268] Young, R.J.; Campbell, M.; Borthwick, A.D.; Brown, D.; Burns-Kurtis, C.L.; Chan, C.; Convery, M.A.; Crowe, M.C.; Dayal, S.; Diallo, H.; et al. Structure- and property-based design of factor Xa inhibitors: Pyrrolidin-2-ones with acyclic alanyl amides as P4 motifs. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 5953–5957.
- [269] Pinto, D.J.P.; Galembo, R.A.; Quan, M.L.; Orwat, M.J.; Clark, C.; Li, R.; Wells, B.; Woerner, F.; Alexander, R.S.; Rossi, K.A.; et al. Discovery of potent, efficacious, and orally bioavailable inhibitors of blood coagulation factor Xa with neutral P1 moieties. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 5584–5589.
- [270] Senger, S.; Convery, M.A.; Chan, C.; Watson, N.S. Arylsulfonamides: A study of the relationship between activity and conformational preferences for a series of factor Xa inhibitors. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 5731–5735.
- [271] Mueller, M.M.; Sperl, S.; Stürzebecher, J.; Bode, W.; Moroder, L. (R)-3-Amidinophenylalanine-Derived inhibitors of Factor Xa with a novel active-site binding mode. *Biol. Chem.* **2002**, *383*, 1185–1191.
- [272] Kamata, K.; Kawamoto, H.; Honma, T.; Iwama, T.; Kim, S.H. Structural basis for chemical inhibition of human blood coagulation factor Xa. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 6630–6635.
- [273] Salonen, L.M.; Bucher, C.; Banner, D.W.; Haap, W.; Mary, J.-L.; Benz, J.; Kuster, O.; Seiler, P.; Schweizer, W.B.; Diederich, F. Cation- π interactions at the active site of factor Xa: Dramatic enhancement upon stepwise n-alkylation of ammonium ions. *Angew. Chem. Int. Ed.* **2009**, *48*, 811–814.
- [274] Pinto, D.J.P.; Orwat, M.J.; Koch, S.; Rossi, K.A.; Alexander, R.S.; Smallwood, A.; Wong, P.C.; Rendina, A.R.; Luetgen, J.M.; Knabb, R.M.; et al. Discovery of 1-(4-Methoxyphenyl)-7-oxo-6-(4-(2-oxopiperidin-1-yl)phenyl)-4,5,6,7-tetrahydro-1H-pyrazolo[3,4-c]pyridine-3-carboxamide (Apixaban, BMS-562247), a highly potent, selective, efficacious, and orally bioavailable inhibitor of blood coagulation. *J. Med. Chem.* **2007**, *50*, 5339–5356.
- [275] Young, R.J.; Borthwick, A.D.; Brown, D.; Burns-Kurtis, C.L.; Campbell, M.; Chan, C.; Charbaut, M.; Chung, C.; Convery, M.A.; Kelly, H.A.; et al. Structure and property based design of factor Xa inhibitors: Pyrrolidin-2-ones with biaryl P4 motifs. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 23–27.
- [276] Zbinden, K.G.; Anselm, L.; Banner, D.W.; Benz, J.; Blasco, F.; Décoret, G.; Himber, J.; Kuhn, B.; Panday, N.; Ricklin, F. Design of novel aminopyrrolidine factor Xa inhibitors from a screening hit.

Bibliography

- Eur. J. Med. Chem.* **2009**, *44*, 2787–2795.
- [277] Kleanthous, S.; Borthwick, A.D.; Brown, D.; Burns-Kurtis, C.L.; Campbell, M.; Chaudry, L.; Chan, C.; Clarte, M.-O.; Convery, M.A.; Harling, J.D.; et al. Structure and property based design of factor Xa inhibitors: pyrrolidin-2-ones with monoaryl P4 motifs. *Bioorg. Med. Chem. Lett.* **2010**, *20*, 618–622.
- [278] Salonen, L.M.; Holland, M.C.; Kaib, P.S.J.; Haap, W.; Benz, J.; Mary, J.-L.; Kuster, O.; Schweizer, W.B.; Banner, D.W.; Diederich, F. Molecular recognition at the active site of factor Xa: Cation- π interactions, stacking on planar peptide surfaces, and replacement of structural water. *Chem. A Eur. J.* **2012**, *18*, 213–222.
- [279] Watson, N.S.; Adams, C.; Belton, D.; Brown, D.; Burns-Kurtis, C.L.; Chaudry, L.; Chan, C.; Convery, M.A.; Davies, D.E.; Exall, A.M.; et al. The discovery of potent and long-acting oral factor Xa inhibitors with tetrahydroisoquinoline and benzazepine P4 motifs. *Bioorg. Med. Chem. Lett.* **2011**, *21*, 1588–1592.
- [280] Young, R.J.; Adams, C.; Blows, M.; Brown, D.; Burns-Kurtis, C.L.; Chan, C.; Chaudry, L.; Convery, M.A.; Davies, D.E.; Exall, A.M.; et al. Structure and property based design of factor Xa inhibitors: Pyrrolidin-2-ones with aminoindane and phenylpyrrolidine P4 motifs. *Bioorg. Med. Chem. Lett.* **2011**, *21*, 1582–1587.
- [281] Pruitt, J.R.; Pinto, D.J.P.; Galembo, R.A.; Alexander, R.S.; Rossi, K.A.; Wells, B.L.; Drummond, S.; Bostrom, L.L.; Burdick, D.; Bruckner, R.; et al. Discovery of 1-(2-Aminomethylphenyl)-3-trifluoromethyl-*N*-[3-fluoro-2'-(aminosulfonyl)[1,1'-biphenyl]-4-yl]-1H-pyrazole-5-carboxamide (DPC602), a potent, selective, and orally bioavailable factor Xa inhibitor 1. *J. Med. Chem.* **2003**, *46*, 5298–5315.

Doctor Thesis, Kobe University

“Computational Strategies for Improvement of Protein-Ligand Docking: Optimization Algorithm, Scoring Function, and Protein Flexibility”, 134 pages Submitted on January, 19th, 2017

The date of publication is printed in cover of repository version published in Kobe University Repository Kernel.

© Shota Uehara

All Right Reserved, 2017