

Computational Analysis of Protein-Ligand Binding: From single continuous trajectories to multiple parallel simulations

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

Holmfridur B. Thorsteinsdottir

aus Reykjavik, Island

Basel, 2010

Originaldokument gespeichert auf dem Dokumentenserver der Universität Basel
edoc.unibas.ch



Dieses Werk ist unter dem Vertrag „Creative Commons Namensnennung-Keine kommerzielle
Nutzung-Keine Bearbeitung 2.5 Schweiz“ lizenziert. Die vollständige Lizenz kann unter
creativecommons.org/licences/by-nc-nd/2.5/ch
eingesehen werden.

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von:

Prof. Dr. Torsten Schwede
Prof. Dr. Markus Meuwly

Basel 25.November 2008

Prof. Dr. E. Parlow (Dekan)



Namensnennung-Keine kommerzielle Nutzung-Keine Bearbeitung 2.5 Schweiz

Sie dürfen:



das Werk vervielfältigen, verbreiten und öffentlich zugänglich machen

Zu den folgenden Bedingungen:



Namensnennung. Sie müssen den Namen des Autors/Rechteinhabers in der von ihm festgelegten Weise nennen (wodurch aber nicht der Eindruck entstehen darf, Sie oder die Nutzung des Werkes durch Sie würden entlohnt).



Keine kommerzielle Nutzung. Dieses Werk darf nicht für kommerzielle Zwecke verwendet werden.



Keine Bearbeitung. Dieses Werk darf nicht bearbeitet oder in anderer Weise verändert werden.

- Im Falle einer Verbreitung müssen Sie anderen die Lizenzbedingungen, unter welche dieses Werk fällt, mitteilen. Am Einfachsten ist es, einen Link auf diese Seite einzubinden.
- Jede der vorgenannten Bedingungen kann aufgehoben werden, sofern Sie die Einwilligung des Rechteinhabers dazu erhalten.
- Diese Lizenz lässt die Urheberpersönlichkeitsrechte unberührt.

Die gesetzlichen Schranken des Urheberrechts bleiben hiervon unberührt.

Die Commons Deed ist eine Zusammenfassung des Lizenzvertrags in allgemeinverständlicher Sprache: <http://creativecommons.org/licenses/by-nc-nd/2.5/ch/legalcode.de>

Haftungsausschluss:

Die Commons Deed ist kein Lizenzvertrag. Sie ist lediglich ein Referenztext, der den zugrundeliegenden Lizenzvertrag übersichtlich und in allgemeinverständlicher Sprache wiedergibt. Die Deed selbst entfaltet keine juristische Wirkung und erscheint im eigentlichen Lizenzvertrag nicht. Creative Commons ist keine Rechtsanwalts-gesellschaft und leistet keine Rechtsberatung. Die Weitergabe und Verlinkung des Commons Deeds führt zu keinem Mandatsverhältnis.

1 Abstract

The interaction of proteins with other proteins or small molecules is essential for biological functions. Understanding the molecular basis of protein-ligand binding is of a vast interest for drug discovery, and computational methods to estimate protein-ligand binding are starting to play an increasingly important role. In order to apply atomistic computational methods to the drug discovery process it is necessary to have accurate three-dimensional structures of the target protein and a fast and reliable method to estimate the binding affinity between the target protein and potential inhibitors. Unfortunately, three-dimensional structures are not available for all proteins of interest, but often their coordinates can be predicted by computational methods such as homology modeling.

In this thesis we study the effect of inaccuracies of homology models to ligand binding using HIV-1 protease as a model system. Homology models of decreasing accuracy are built and additional errors are introduced by misplacing side chains during rotamer modeling. We establish a MM-GBSA approach to estimate protein-ligand binding free energies, and apply this method to the different homology models.

Although MM-GBSA methods are significantly faster than traditional MM-PBSA methods, still the required computational effort is significant as it is based on the calculation of a continuous molecular dynamics trajectory. In this study, we establish a novel approach based on multiple independent short simulations, which is suitable for execution of a distributed grid of computers and thereby dramatically reduces the computation time needed. This workflow is validated using the HIV-1 protease model system, and then applied to the estrogen receptor. Novel methods to assess the sampling of the different trajectory approaches and potential application to docking problems are presented and discussed.

Contents

1	Abstract	1
2	Overview	5
3	Introduction	9
3.1	Protein and Ligand Coordinates	9
3.1.1	Homology Modeling	10
3.1.2	Docking	12
3.2	Molecular Recognition and Binding Affinities	15
3.2.1	Experimental approaches	17
3.2.2	Computational approaches	18
3.3	Protein Dynamics	23
3.3.1	Multiple Molecular Dynamics Simulations	25
3.4	Protein Systems Studied	26
3.4.1	HIV-1 Protease	27
3.4.2	Estrogen Receptor beta	28
4	Methods	31
4.1	Parameterization	32
4.2	Input Structures	32
4.3	Molecular Dynamics	34
4.3.1	Standard long MD simulations	35
4.3.2	Multiple MD simulations	35
4.4	Analysis	36
4.4.1	Binding Energy Evaluation	36
4.4.2	Correlation to Experimental Values	38
4.4.3	Principle Component Analysis	38
4.4.4	Clustering	38
4.4.5	Analysis of Clusters	43

5	Results	45
5.1	How Inaccuracies in Protein Structure Models Affect Estimates of Protein-Ligand Interactions	46
5.2	MM-GBSA on a PC GRID: Setup, Validation and Applications	64
6	Discussion	103
7	Summary & Outlook	107
8	Appendix	125
9	Acknowledgments	129

2 Overview

Proteins are organic macromolecules that are essential to all organisms. They are comprised of amino acids linearly combined into polypeptides that fold into specific three-dimensional structures. Proteins in cells perform a variety of tasks, ranging from structural and mechanical functions such as in the cytoskeleton and in muscles to controlling biological processes through interacting with other proteins or small molecules that bind to them. Among biological processes that proteins control through molecular recognition are cell signaling, signal transduction, immunological responses and enzyme catalysis. The activity of proteins can often be modified by the use of drugs that target specific interaction sites on the protein. This makes the detailed understanding of the three-dimensional structure, the dynamics of proteins and how they interact with their natural ligands and potential drugs essential to aid the discovery and development of novel specific inhibitors for disease related proteins in any drug discovery effort.

Ultimately it is the three-dimensional structure of proteins that determines their function. In order to be able to accurately describe the interactions between proteins and ligands a detailed three-dimensional structure of the protein is crucial. In spite of considerable advances in the last decades, the mechanism in which proteins fold is still not fully understood, and therefore the determination of a protein structure directly from its amino acid sequence is still not possible with computational methods. The three-dimensional structures of proteins must therefore be determined experimentally with methods such as x-ray crystallography, nuclear magnetic resonance or cryo-electro microscopy. These methods, however, are slow in comparison to the rapid accumulation of protein sequences that are being determined, which results in a huge gap between known protein sequences and the actual three-dimensional structures of the corresponding proteins. In those cases where no experimentally determined three-dimensional structures are available, homology modeling methods provide insights and are becoming increasingly important. Such methods are based on the observation that the three-dimensional structures of proteins within a protein

family are better conserved than their amino acid sequence¹. This enables a three-dimensional model of a protein sequence to be built based on its similarity to related proteins with a known structure.

Drug discovery and development is a multi-billion dollar business and although the numbers of actively pursued drug targets vary in the literature, the consensus number is at least 324 drug targets that are currently being pursued². In order to identify potential drug leads, drug discovery makes use of methods such as high-throughput screening, where large libraries of compounds are tested for their ability to interact with the target protein. To reduce the number of compounds that have to be synthesized and tested, computational methods such as virtual screening are becoming increasingly important to the drug discovery process. In virtual screening molecules from large libraries of available compounds are docked computationally into the binding site of the protein target. To identify compounds that are potential drug leads, the binding energy between the protein and compounds is calculated and those that have the most favorable interaction are selected for further analysis. It is therefore very important to have a fast and reliable way of calculating these interactions accurately. The drawback of such virtual screening methods is that the algorithms currently used have to make a number of approximations in order to be able to screen large numbers of compounds in a reasonable time, which results in less accurate description of the binding energies.

One way to overcome these limitations is to use more accurate force field based methods to determine the binding energies of the compounds and thereby more reliably reject those poses that do not have a favorable binding energy. Such force field methods have their limitations as well, in addition of requiring reliable force field parameters for the calculations of interaction energies these methods tend to be very computationally expensive which limits their applicability to the drug discovery process.

In this thesis we aim to address some of the limitations of protein-ligand affinity calculations: first by analyzing the usefulness of homology models for binding affinity calculations or in other words how sensitive are these calculations to erroneous structures; then by modifying the molecular dynamics protocol to sample protein-ligand conformations more efficiently and finally by addressing the question whether the lowest scoring docking pose is necessarily the best one and if the poses can be improved by running a molecular dynamics simulation on them.

This thesis is organized as follows: First, a general introduction to the acquire-

ment of protein and ligand coordinates, methods to estimate the binding energy and ways to enhance conformational sampling is given, followed by a brief introduction to the protein systems used in this work. Second, the methods applied in this work are introduced and described in detail. Finally, the main results are presented, first the question of how inaccuracies in protein structure models affect the estimate of protein-ligand interactions is addressed, and finally the development of multiple short trajectory approach to speed up computationally intensive calculations is reported.

3 Introduction

3.1 Protein and Ligand Coordinates

In recent years molecular biology has moved from studying only one gene at a time, to the study of whole genomes and biological systems³. Microarray and proteomic methods help to identify dysregulated genes and proteins that are abnormally expressed and with the assistance of bioinformatics tools an increasing amount of lead compounds that could potentially be novel drug candidates are being identified^{4,5}. This has made molecular biology extremely important for drug discovery, not only for the development of high-throughput assays and designing clinical trials, but also for the identification of novel drug targets.

With this increase in identification of potential drug targets, computational methods of discovering likely drugs are becoming increasingly important. It is vital for most virtual screening methods that an accurate and reliable three-dimensional structure of the target protein is available. Unfortunately, the mechanism in which proteins fold into their functional three-dimensional structures from the amino acid chain is still poorly understood⁶. In spite of advances in *ab initio* prediction of protein structures directly from the amino acid sequence⁷ and through efforts of distributed computing project such as Folding@home, the computational simulation of the folding of proteins is still limited to small peptides and a very time-consuming process⁸. The three-dimensional structures of proteins must therefore be determined experimentally. Even though techniques such as x-ray crystallography, NMR spectroscopy and cryo-electron microscopy are becoming faster, there is still a large gap between the number of available structures and the protein sequences. In these cases where no experimental structure exists, homology modeling proves to be an important alternative.

3.1.1 Homology Modeling

In the SwissProt databases there are currently ~ 400.000 non-redundant annotated sequences⁹ and if the computer annotated TrEMBL database is included there are over 6.5 million protein sequences available in these databases (<http://www.expasy.ch/sprot>). In contrast, there are only around 53.000 proteins structures that have been solved experimentally¹⁰. There is obviously a huge gap between the available experimentally determined structures and the sequences that have been determined.

Homology modeling (HM) is based on the observation that structures within a given protein fold family are better conserved than their sequences¹. This observation enables known experimental structures to be used as a **template** protein to create a three-dimensional model for a similar protein for which the structure is unknown, or the **target** protein¹¹. Figure 3.1 shows the four steps that are common to all methods in building homology models.

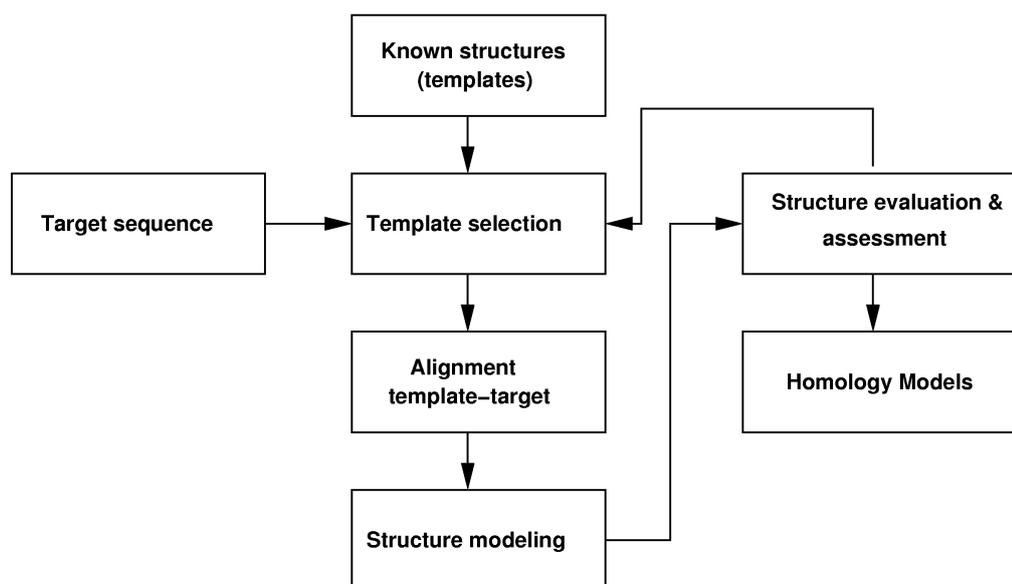


Figure 3.1: The main steps involved in building a homology model.

In the first step the template protein is selected from the protein data bank (PDB)¹⁰, using the target protein sequence as a query. Sequence similarity is a well established

measure to select suitable template structures¹². If the sequence similarity is high (>30%) this is relatively straight forward with pairwise sequence comparison tools such as BLAST¹³ and FASTA¹⁴ most commonly used for that task. However, if the sequence identity between the target and template sequences is below 30% alternative strategies based on multiple sequence alignment such as profile analysis^{15,16} have to be used.

The second step involves aligning the target and template sequences. The alignment from the template selection is not necessarily the optimal one, because the algorithms used to detect remote homologs are not necessarily the best ones for the best possible alignment. Once a suitable template has been selected it should therefore be aligned to the target using specialized programs such as t-coffee¹⁷. For closely related proteins with sequence similarity over 50% this alignment is almost always correct. However, if the sequence similarity falls below 35%, the so called "twilight zone" is entered and the accuracy of the alignment drops significantly where regions of low sequence similarity increase¹⁸. In general it can be assumed that as the sequence similarity decreases the alignment contains increasing number of gaps and alignment errors. Great care must therefore be taken to obtain an accurate alignment because no modeling procedure can recover from incorrect alignment; consequently manual inspection of the alignment is often required and strongly encouraged.

Based on the template structure(s) and the alignment between the target and template(s), a model for the target sequence is built. To assist with the homology modeling, a number of automated methods are available. In general, they can be divided in methods based on rigid fragment assembly such as SWISS-MODEL^{19,20} or Composer²¹ and spatial restraint methods as used in Modeller²². Modeling based on rigid fragment assembly makes use of the possibility to identify structurally conserved core sections within a protein family that can be used to build the model. The loops and side-chains are however variable and have to be modeled by other means. Modeling by satisfaction of spatial restraints involves generating numerous constraints or restraints on the structure of the target sequence based on bond lengths, angles, dihedrals and non-bonded contacts that are obtained from a molecular mechanics force fields. The models are built by minimizing the penalty of these restraints. Side-chains are normally modeled using rotamer libraries²³ which is often backbone dependent²⁴. Loop modeling is either based on *ab initio* methods where a large number of possible conformations are generated and scored^{25,26}, or database methods which try to find available protein structure where there is a significant similarity in the stem

of the loop^{27,28}.

Often observed errors in homology models can be roughly divided into five classes²⁹. Errors in side-chain prediction, alignment errors, errors in the regions without a template, errors due to misalignments and finally the use of incorrect templates. There are many available programs and webservers developed to assist in the evaluation of homology models. PROCHECK³⁰ and WHATIF³¹ check for correct stereochemistry and structural packing quality of the models and VERIFY3D³² analyzes the compatibility of an atomic model with its own amino acid sequence. While programs such as ANOLEA³³ checks for fitness of the sequence to the structure and evaluate the non-local environment of each of the heavy atom in the molecule. It is often necessary to go back to the template selection step or the template-target alignment to improve the quality of the resulting model. Building a number of different models and then selecting the best one is also advisable.

3.1.2 Docking

Docking methods are used to identify possible drugs against a given protein target. In general there are two aims of docking studies³⁴. The first is to identify novel ligands by virtual screening (docking), this also includes finding possible poses within the protein binding site, and the second is to predict the binding affinities of the suggested binding modes (scoring). The binding affinity calculations will be addressed in the next chapter, but here we will look at how different ligand poses are generated with virtual screening methods.

All docking programs aim to find the best fit of a ligand to the binding site of the protein, and they can roughly be divided into three categories depending on which approach they use for this purpose³⁵. These are systematic methods, random or stochastic methods and simulation methods. Table 3.1 shows an example of docking programs that utilize each of these methods³⁶.

Systematic search methods

If the systematic search approaches would explore all the possible degrees of freedom for a given molecule the number of possible combinations would very quickly grow so large that it would result in combinatorial explosion⁴⁶. To avoid this systematic search methods normally use stepwise or incremental searches, for example dividing the ligands into rigid (core) fragments that are docked first into the active site and

Table 3.1: Common docking programs that utilize systematic, random or simulation methods in their search for optimal ligand pose, adapted from³⁶

Systematic	Random	Simulation
DOCK ³⁷	AutoDock ³⁸	DOCK
FlexX ³⁹	MOE-Dock ⁴⁰	Glide
Glide ⁴¹	GOLD ⁴²	MOE-Dock
Hammerhead ⁴³	PRO_LEADS ⁴⁴	AutoDock
FLOG ⁴⁵		Hammerhead

flexible parts (side-chains) which are added in incremental fashion^{39,47} as used in the program DOCK³⁷. The docking program Glide was used in this thesis and will be described in more detail here.

Glide is a novel docking program that approximates a complete systematic search approach, the algorithm is shown in figure 3.2⁴¹.

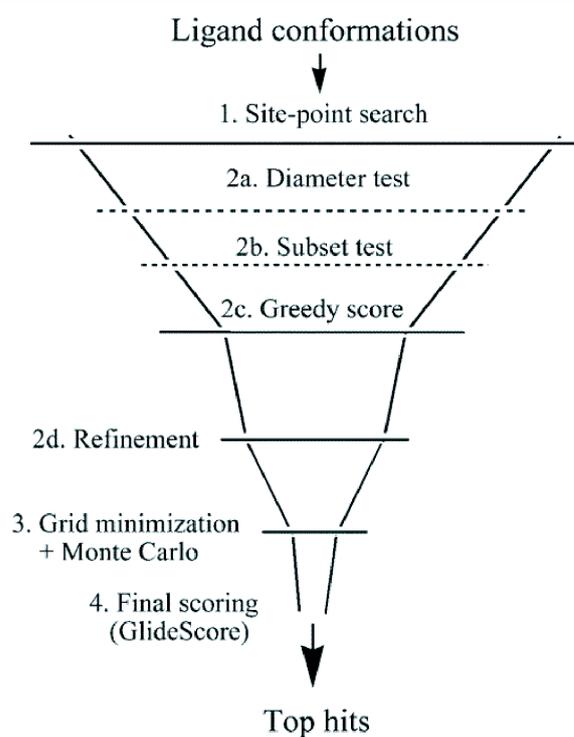


Figure 3.2: The Glide docking approach. From⁴¹

Each ligand is separated into a core region, which is represented by a number of conformations, and a number of rotamer groups, which are attached to the core by

a rotatable bond. The core and in addition all of the possible rotamer group conformations is then docked as a single object, which can contain thousands of molecules. Then an initial screen of possible ligand positions over the active site of the protein is performed. The positions of ligands are then scored with respect to orientations of the ligand, steric clashes and hydrogen bonds with so-called "greedy-scoring". The top scoring poses are refined and re-scored and the best scoring of those are energy minimized using a molecular mechanics scoring function and the best scoring ones finally scored using a specialized scoring function as shown in figure 3.2⁴¹.

Random search methods

In random search, or stochastic methods the ligand is generally considered as a whole and a ligand pose is evaluated by a pre-defined probability function³⁶. Examples of random approach algorithms are Monte Carlo, genetic algorithms and Tabu search. The Monte Carlo algorithm in its simplest form generates a random configuration of a ligand that is scored, then a new conformation is generated which is scored as well. The Metropolis algorithm is then used to determine whether the new conformation is accepted or not. This procedure is repeated until enough conformations are obtained³⁸. Genetic algorithms model conformations of ligands as "chromosomes" that are stochastically varied and evaluated by a fitness function. The chromosomes that correspond to the best intermediate solutions are subjected to "crossover and mutation" operations to produce the next generation⁴². Tabu search on the other hand takes the conformational space already explored into consideration^{44,48}. In order to determine whether a given configuration is accepted or not the root mean square deviation between that conformation and all others is calculated⁴⁴.

Simulation methods

In the simulation methods the ligand explores the conformational space while the protein is kept rigid. The conformations that have the lowest energies are then picked as potential poses. These methods have the limitation of often being unable to cross high-energy barriers and therefore being bound to a local minimum on the energy surface³⁵. Number of different ways to circumvent this problem have been suggested, for example to simulate different parts of the protein-ligand system at different temperatures⁴⁹ or to start molecular dynamics calculations from different ligand positions as implemented in the approach by Pak *et al*⁵⁰.

Receptor flexibility

It is not entirely accurate to view the protein target as completely rigid, which is the approximation of many docking algorithms. Proteins are frequently observed to adapt to ligand binding and the importance of that should not be neglected. Various approaches have been developed to treat protein flexibility, but its treatment is not as advanced as ligand flexibility. These methods include molecular dynamics and Monte Carlo calculations^{51,52}, use of rotamer libraries^{53,54} and protein ensemble grids⁵⁵. Including the receptor flexibility is computationally intensive so those methods are mostly limited to side-chain flexibility of the binding pocket.

3.2 Molecular Recognition and Binding Affinities

The mutual molecular recognition of proteins with their ligands is the beginning of most of biological processes. Therefore one of the main objectives of structure-based drug design is to be able to reliably and accurately predict the binding affinity of compounds that bind to the target protein. The "lock-and-key" analogy to substrates binding to enzymes like a key fits into a lock was first introduced by Emil Fischer in 1894⁵⁶. Although this is still a valid analogy since a certain shape complementarity has to be present between the substrate and the protein, it is also clear that there are many other factors that play a part in the ligand binding process. In this work we will only focus on non-covalent binding, which can be viewed as the association of Protein (P) with Ligand (L) to form a Protein-Ligand complex (PL). Figure 3.3 shows a schematic picture of non-covalent ligand binding to a protein, and some of the factors that contribute to protein-ligand binding.

It has been suggested that the electrostatic interactions mainly determines molecular recognition and noncovalent binding⁵⁷, but this is by no means a general rule. There are equal evidence to the importance of shape complementarity⁵⁸. Molecular recognition can be therefore attributed to contributions of electrostatic and van der Waals interactions, solvation/desolvation and flexibility of ligand and protein. These major contributors to molecular recognition are described briefly below

Noncovalent Interactions

Noncovalent interactions are a number of relatively weak chemical interactions that stabilize the conformations and the interactions between molecules.

Hydrogen bonds result from electrostatic attraction between an electronegative

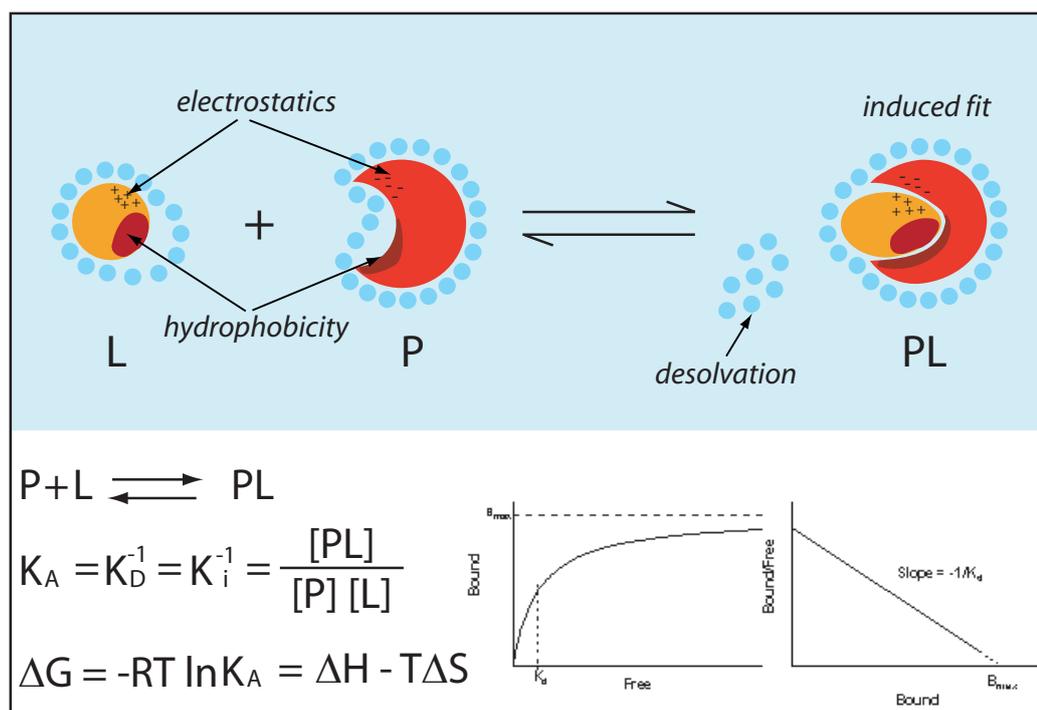


Figure 3.3: A schematic overview of the main factors of protein-ligand binding and how to derive binding energies from them. K_A is the association constant, and K_D , K_i the dissociation and inhibition constants. ΔG is the Gibbs free energy. The plots show the kinetic curves used to extract the values for K_D and B_{\max} .

atom and hydrogen that is connected to an electronegative atom, which is usually oxygen or nitrogen and less frequently fluorine, or from a π - π -interactions, or stacking. This also implies that it is very important for theoretical calculations to have the protonation states of arginine, lysine, aspartic and glutamic acids, as well as histidine correctly determined for an accurate description of electrostatic interactions. Distances of hydrogen bonds are normally 2.5-3.2 Å and angles of 130°-180° are typically found⁵⁹. The strength of a hydrogen bond depends on its directionality and its surroundings. The hydrogen bonds in the interior of proteins are stronger than the ones in the solvent-exposed regions⁶⁰. In addition ionic bonds are very important to ligand-protein binding, but their strength is considerably reduced in water due to shielding.

Water plays an important role in the hydrophobic forces, by forcing hydrophobic groups together it abolishes disruptive effects on the hydrogen bonded network in water, which is also known as the hydrophobic effect.

The van der Waals interactions, or London dispersion forces are used to model the

attractive and repulsive forces between molecules. If two atoms are too close to each other they will repel each other, which makes it possible to define a fixed radius for the "size" of each atom (van der Waals radius). The contact distance between two atoms is then the sum of their van der Waals radii. Van der Waals interactions can be very important when two surfaces of molecules fit well together.

Solvation and Desolvation

Water plays an important role in the formation of protein-ligand complexes. Before a protein-ligand complex is formed, the individual partners that are not a part of hydrophobic surface are involved in hydrogen bonds with the surrounding water. Once the complex is formed, these hydrogen bonds are replaced with hydrogen bonds between the ligand and the protein. The contribution of hydrophobic interactions to protein-ligand binding is normally regarded to be proportional to the size of the hydrophobic surface buried during complex formation^{61,62}. Hydrophobic interactions are also regarded to be the main driving force of conformational change of the receptor upon ligand binding⁵⁷.

Entropy

The change in the degrees of freedom of the ligand and protein upon binding results in a change of the entropy⁶³. It can be viewed as the ligand and protein both losing three degrees of translational and rotational freedom, while six new vibrational degrees of freedom are created for the complex⁶⁴

3.2.1 Experimental approaches

The ability of theoretical methods to predict the binding affinity of protein-ligand interactions can be evaluated on how well they are able to reproduce experimental data.

The general equation for binding association is shown in figure 3.3. If the free concentration of the ligand [L] reaches the value of K_D , the receptor binding sites are half saturated with ligand. The value of K_D is also half of the B_{max} or the maximal specific binding. In the same way the IC_{50} concentration is the free ligand concentration at 50% receptor saturation.

The K_D and B_{max} can be determined with a saturation assay, where different concentrations of labeled ligands are added to the assay. The data is normally analyzed by Scatchard analysis where the amount of bound ligand divided by the amount of free ligand is plotted against the amount of bound ligand. The slope is then the K_D

and the x-axis intercept is the B_{\max} . More recently this method is being replaced by a non-linear regression analysis.

The IC_{50} concentration is determined with a competition assay to a labeled ligand. This has the advantage that each of the ligands that is being tested does not have to be explicitly labeled but rather tested against an already labeled molecule. The K_i constant is then determined from the Cheng-Prusoff equation⁶⁵

Numerous binding assays have been designed to experimentally determine binding affinities, these can be divided into methods that require separation of the components or do not require separation of the components for analysis.

Separation Assays

In separation assays the target protein is incubated with a labeled ligand (radioactive or fluorescence) until equilibrium is reached. The amount of bound ligand is estimated after separating the unbound ligand from the bound normally by centrifugation or filtration. These assays can either be saturation assays or competition assays, an example of a separation assay is the radioligand-binding assay which has been used very successfully for the high throughput drug discovery process⁶⁶

Direct Assays

In direct assays the protein-ligand interaction is measured in real-time. This is convenient because it does not require separation of the components prior to measurement. One method commonly used is fluorescence polarization which measures the change in polarization of light emitted from a fluorescent labeled ligand from when it is rotating freely to when it is bound to a protein target. The light emitted from the freely rotating ligand is depolarized due to the rapid motion of the ligand, whereas the rotational speed decreases when it is bound to a high-molecular weight protein target and the emitted light stays polarized. This method has also been successfully used in high throughput drug discovery and has the advantage over radioligand assays that it is non-radioactive and considerably cheaper⁶⁷.

3.2.2 Computational approaches

There is a need for fast, accurate and reliable methods to calculate the binding affinity of ligands to proteins⁶⁸. These methods should ideally help the drug discovery process by enabling the pre-screening of potential drugs and as a consequence reduce the number of compounds that need to be screened by in vitro/vivo experimental methods. The available computational methods differ considerably in their accuracy

and complexity. Affinity calculations in the docking context have to be able to deal with a very large number of compounds in a relatively short time. These methods therefore tend to only deal with a single structure of the docked pose and focus on a fast calculation, making numerous approximations to obtain a reasonable ranking of compounds, rather than accurately calculate absolute binding affinities. On the other hand simulation methods use techniques such as Monte Carlo or Molecular Dynamics to obtain conformational sampling for increased accuracy, which in turn makes them more computationally intensive and therefore not easily applicable to high-throughput virtual screening.

The sections below describe briefly a number of methods that are commonly used to calculate binding affinities between protein and ligands. They are divided in whether they are methods that aim to calculate binding affinities from one structure such as a docked pose from a virtual screening or using molecular dynamics or Monte Carlo sampling to calculate the binding affinities from an ensemble of structures.

Single Structure from Docked Poses

The scoring functions that are used in virtual screening have to be able to deal with large number of compounds in short time. It is important to have a scoring function that is able to correctly rank the poses generated and thereby dismiss the poses that are incorrect. The most commonly used scoring functions used in virtual screening can be divided into knowledge-based, empirical and force-field based methods. Each one of those methods has its approximations and limitation so often the results can be improved by combining various scoring functions⁶⁹.

Knowledge-based scoring

In knowledge-based scoring the focus is more towards reproducing the actual structures instead of the energies. They are based on atomic interaction-pair potentials which are derived from observations of contacts from known ligand-protein complexes available in databases such as the PDB¹⁰ and Cambridge Structural Database (CSD)⁷⁰. Scoring functions that use this method are for example potential of mean force (PMF)⁷¹ and DrugScore⁷². These methods tend to be simple to compute efficiently, but a drawback is that the number of available protein-ligand complexes needed to derive the parameters is limited³⁶.

Empirical scoring

The empirical scoring scheme uses combination of several parameters that are then

fitted to reproduce experimental energies⁷³. The parameters are selected to best cover all the important contributions of the total binding free energy, such as hydrogen bonds, hydrophobic and ionic interactions, and an entropic term to account for the loss of conformational freedom. The coefficient of the parameters used to fit the data are obtained using regression analysis from experimental binding free energies and x-ray structure information as atomic interaction-pair potentials³⁶. Examples of scoring functions that use this method are LUDI⁷³ and ChemScore⁷⁴. An advantage of this scoring function is that it has a simple functional form but it is dependent on data sets for fitting resulting in different weighting factors³⁶.

Force-field based scoring

Force field based scoring functions avoid specific parameterization by using well established molecular mechanics force fields to estimate the binding energy of non-bonded interactions (vdw, hydrogen bond and electrostatics)³⁶. An example of force field based scoring functions are G-Score⁷⁵ and AutoDock⁷⁶. These scoring functions are generally limited by the exclusion of solvent and entropic terms, although recent implementation include an entropy term and solvation energy using PBSA or GBSA models⁷⁷.

Molecular Mechanics Methods

Estimating binding free energies accurately is a very time-consuming process. The most accurate results are obtained with methods such as Free Energy Perturbation (FEP) / Thermodynamics Integration (TI), and similar results can be obtained at a lower computational cost with methods such as MM-PBSA/MM-GBSA or Linear Interaction Energy (LIE). These methods are however still considered too computationally intensive to be of much use in virtual screening approaches.

Free Energy Pathway methods

FEP and TI methods are often referred to as "computational alchemy", in the sense that they evaluate the difference between the binding energy of two similar ligands by using pathways to compute the change in free energy when ligand A is changed to ligand B within the binding site and in solution⁷⁸. These methods generally give a very good estimate of the binding energy, with errors below 1kcal/mol^{79,80}. By slowly "growing" the ligand into the binding site it is also possible to calculate the absolute binding free energy, but this is a very time consuming process⁸¹.

Figure 3.4 shows the thermodynamic cycle that is used for chemical alchemy calculations.

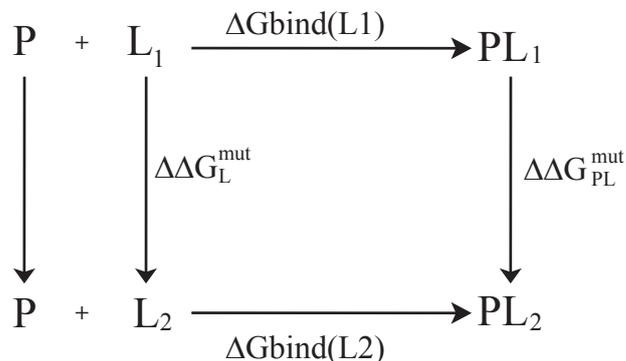


Figure 3.4: The thermodynamic cycle used for TI calculations, adapted from⁸². P is the protein and L is the ligand, PL is the protein-ligand complex.

By using this thermodynamics cycle the difference binding free energy between ligand1 and ligand2 can be calculated as:

$$\begin{aligned}
 \Delta\Delta G_{\text{bind}} &= \Delta G_{\text{bind}}(\text{L}_1) - \Delta G_{\text{bind}}(\text{L}_2) \\
 &= \Delta\Delta G_{\text{PL}}^{\text{mut}} - \Delta\Delta G_{\text{L}}^{\text{mut}}
 \end{aligned}
 \tag{3.1}$$

MM-PBSA and MM-GBSA

MM-PBSA and MM-GBSA are so called end-point methods in the sense that they only evaluate the initial and final states of the system instead of the path between the states. They can use molecular dynamics or Monte Carlo simulations to obtain snapshots of the protein-ligand complex which are used to calculate the average binding free energy. If the configurational entropy is included it is estimated by minimizing a small number of snapshots and from them calculate the entropy with a rigid-rotor/harmonic-oscillator approximation⁶⁸, but often this term is neglected if only relative binding affinities are required because it is very time consuming.

Figure 3.5 shows the thermodynamics cycle used in MM-PBSA/GBSA calculations.

The binding free energy is decomposed into contributions from the gas phase and solvation free energy that arises from the gas phase to water transition:

$$\Delta G_{\text{bind}} = \Delta G_{\text{PL}} - (\Delta G_{\text{P}} + \Delta G_{\text{L}})
 \tag{3.2}$$

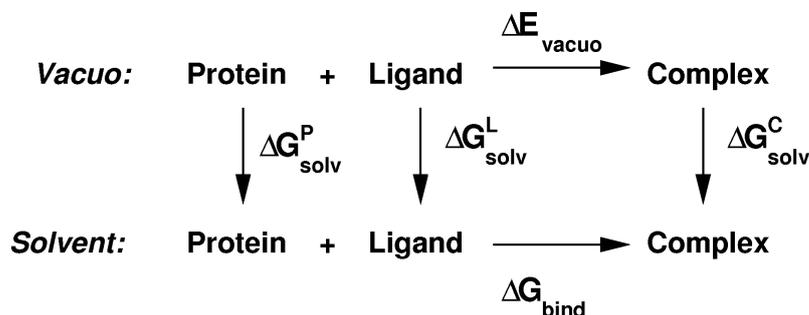


Figure 3.5: The thermodynamic cycle used for MM-PBSA/GBSA calculations

where

$$G = G_{\text{gas}} + G_{\text{solv}} = (H_{\text{gas}} - TS) + G_{\text{solv}} \quad (3.3)$$

The solvation contribution G_{solv} can be decomposed into contributions from electrostatics and nonpolar interactions.

$$G_{\text{solv}} = G_{\text{elec}} + G_{\text{nonpolar}} \quad (3.4)$$

The nonpolar contribution is assumed to be proportional to the solvent accessible surface area (SASA). The electrostatics contribution is solved using an implicit solvent model either PBSA⁸³ or GBSA⁸⁴. The enthalpic contributions are calculated using a force field approximation:

$$H_{\text{gas}} \approx E_{\text{gas}} = E_{\text{bond}} + E_{\text{angle}} + E_{\text{torsion}} + E_{\text{elec}} \quad (3.5)$$

The energies are normally averaged over conformations sampled from molecular dynamics simulations, often with explicit water. These methods have been applied to a variety of ligand-protein complexes and proven to give good estimates of the binding energy⁸⁵.

Linear Interaction Energy

The Linear Interaction Energy (LIE) methods⁸⁶ are end-point method like MM-PBSA/GBSA

and also use averaged conformations from molecular dynamics simulations. The binding free energy is estimated as:

$$\Delta G_{\text{bind}} \approx \alpha(\langle E_{\text{elec}} \rangle_{\text{bound}} - \langle E_{\text{elec}} \rangle_{\text{free}}) + \beta(\langle E_{\text{vdw}} \rangle_{\text{bound}} - \langle E_{\text{vdw}} \rangle_{\text{free}}) \quad (3.6)$$

where the brackets denote averages from molecular dynamics trajectory. The factors α and β account for changes in the internal energy of the solvent and protein and are determined empirically⁸⁶. This method has been shown to give accurate results⁸⁶ and newer implementations include a solvation energy term to increase accuracy⁸⁷. A drawback of this method is that there is no universal value for the factors α and β , they have to be determined independently for each case and require experimental data.

3.3 Protein Dynamics

Molecular dynamics simulations are commonly used to get information on the time evolution of conformations of biological macromolecules such as proteins. They provide information on motions of individual atoms as a function of time and thereby enable us to better understand the properties of molecules and the interactions between them. The first molecular dynamics simulation was performed in 1957 using a so-called hard sphere model⁸⁸. In 1964 Rahman improved this rough model by applying a smooth, continuous potential which better mimics the atomic interactions⁸⁹. With increasing computer power the possibility of molecular dynamics simulations has also steadily increased and since 1976 when the first MD simulation of a protein was performed⁹⁰ the use of MD simulations has steadily increased and today molecular dynamics simulations are widely and commonly used in physical and biological sciences.

The interaction within a protein or between a protein and a ligand is normally described by a force field, such as the CHARMM force field⁹¹. It contains several discrete terms, which describe the intermolecular and intramolecular forces in the system:

$$\begin{aligned}
U(\vec{R}) = & \sum_{bonds} K_b(b - b_0)^2 + \sum_{Urey-Bradley} K_{UB}(S - S_0)^2 + \sum_{angle} K_\theta(\theta - \theta_0)^2 + \\
& \sum_{dihedrals} K_\chi(1 + \cos(n\chi - \delta)) + \sum_{impropers} K_\phi(\phi - \phi_0)^2 + \\
& \sum_{nonbond} \epsilon_{ij} \left[\left(\frac{R_{ij}^{min}}{r_{ij}} \right)^{12} - \left(\frac{R_{ij}^{min}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{\epsilon r_{ij}}
\end{aligned} \tag{3.7}$$

where for the bonded interactions K_b , K_{UB} , K_θ , K_χ , and K_ϕ are the constants for bond length, Urey-Bradley (1-3 bond length), bond angle, dihedral angle, and improper dihedral angle force, respectively. In the same way d , S , θ , χ and ϕ refer to the bond length, Urey-Bradley (1-3 bond length), bond angle, dihedral angle and improper dihedral angle values, the zero refers to the equilibrium values. For the non-bonded interactions, ϵ_{ij} refers to the Lennard-Jones well depth, R_{ij}^{min} to the distance at which the interparticle potential is zero, q_i is the partial atomic charge of atom i , ϵ is the effective dielectric constant, and r_{ij} is the distance between atoms i and j .

The CHARMM force field contains a set of parameters for the standard 20 amino acids, nucleic acids and a number of organic molecules⁹¹. These parameters make it possible to simulate proteins and any system that is described within these parameters⁹¹. However, if the intention is to simulate molecules that are not described within this set as is often the case in drug discovery processes, the relevant parameters need to be obtained⁹². Manually determining parameters is very time-consuming task and not well adapted for a large-scale simulation of protein-ligand complexes. Fortunately, there exist a number of programs to assist with parameterization such as Wit!P (A.Widmer, Novartis) and Antechamber⁹³.

The energy landscape of proteins has often been described as containing one global minimum and a number of local minima. These minima are separated by energy barriers which the protein must cross in order to adapt to a new conformation. In this sense the shape and roughness of the landscape determine the dynamics behaviors of the protein. It has been suggested that the potential energy surface that describes the native state of proteins contains multiple minima that correspond to very similar conformations⁹⁴. This observation was confirmed by molecular dynamics simulation⁹⁵ and later experimentally⁹⁶. Molecular dynamics studies must therefore be able to explore states of interest. In contrast, to protein folding/unfolding where exhaustive sampling of the entire energy landscape is necessary, in the work presented here

it is sufficient to sample the protein-ligand binding ensembles to obtain a realistic sampling. There are numerous molecular dynamics methods that have been developed to efficiently explore the energy surface of interest. Among these are methods that modify the energy landscape such as metadynamics⁹⁷ or stochastic tunneling⁹⁸, methods that increase the temperature to attempt to overcome barriers, such as simulated annealing⁹⁹ or multiple simulations. In addition, increased computer power which often uses parallel environment makes it possible to perform molecular dynamics simulations that extend into the microsecond timescale¹⁰⁰.

3.3.1 Multiple Molecular Dynamics Simulations

An alternative way to enhance the conformational sampling is to perform multiple short molecular dynamics simulations, instead of a single long one^{101,102}. In other words, instead of running one simulation using many CPUs the simulation is split up into multiple, independent simulations that are calculated on separate computers. These independent simulations can then later be combined and analyzed as a complete set. The length of the short molecular dynamics simulations limits the processes that can be observed. Therefore it is important to have knowledge of the system to choose suitable time scales. Table 3.2 lists time scales and movements commonly observed in proteins.

Table 3.2: Timescales and motions commonly observed in proteins

Time scale	Amplitude	Description
femto to pico	0.001 - 0.1 Å	bond stretching, angle bending constraint dihedral motion
pico to nano	0.1 - 10 Å	unhindered surface side chain motion loop motion, collective motion
nano to micro	1 - 100 Å	folding in small peptides helix coil transition
micro to second	10 - 100 Å	protein folding

There have been numerous attempts of running multiple short trajectories and comparing the results to those obtained from a single long one. Auffinger *et al* looked at the hydration of tRNA(Asp) anticodon by running six 500 ps multiple molecular dynamics trajectories. They found that even if the trajectories were similar in dynamical characteristics they displayed different local hydration patterns which reflects

the conformational landscape that is explored¹⁰³. Caves *et al* looked at the differences in sampling between running an individual trajectory of 5ns or ten independent trajectories of 120 ps each, which only differ in the initial velocities. They found that the overall sampling was improved by using the multiple independent trajectory approach and suggest that this should be used to obtain better sampling¹⁰⁴. Loccisano *et al* looked at the $A_1 \rightarrow A_{1,3}$ transition in MbCO. They found that by running ten 400 ps simulations they were able to observe this transition frequently, while using two 1.2 ns simulations they were only able to observe it once. The initial structures came from five x-ray structures with random initial velocities¹⁰⁵. Other attempts have shown improved sampling by running multiple molecular dynamics simulations, time scales ranging from 100 ps to few nanoseconds and the number of short simulations from being less than 10 to around 1000. These simulations vary from studying the conformational space of small peptides and proteins^{106,107} to studying ion channels and lipid bilayer^{108,109}.

Until now not much has been done to investigate multiple short trajectory approaches in the context of improving conformational sampling for protein-ligand interactions. Brown *et al* showed that they could obtain comparable correlation to experimental binding free energies of biotin to avidin by running six 13ps short molecular dynamics simulations on a Grid.¹¹⁰ They extended this approach to show that they could obtain reasonable correlation to experimental data for 18 ligands of urokinase¹¹¹. In both cases there is only one short trajectory for each of the minimized structures and the analysis only consists on monitoring the RMSD and correlation to experimental values.

3.4 Protein Systems Studied

In order to be able to further our understanding of molecular recognition and to develop algorithms that improve scoring functions, it is necessary to have well established protein systems available for study. In this work we use two protein systems, the HIV-1 protease and the Estrogen Receptor β , both of which have numerous structures available in the protein data bank and a large number of known ligands that bind to them.

main-chain NH of the opposing loop. Thr26 also donates a hydrogen bond to the carbonyl O atom of residue 24 on the opposite loop¹¹⁴. The HIV-1 protease contains two flaps which cover the active site and participate in the binding of their natural substrates and inhibitors. A common feature to most complexes of HIV-1 protease is a water molecule that is separated from the bulk solvent and bridges the inhibitor and Ile50 and Ile150 NH groups of the flaps¹¹⁴.

Most of the HIV-1 protease inhibitors that are being designed are inhibitors that compete with the natural substrate for the same active site¹¹⁵. An alternative approach has been suggested with ligands that bind at the subunit interface and thereby destabilize the dimeric structure, but so far these efforts have not been very successful¹¹⁶.

3.4.2 Estrogen Receptor beta

The Estrogen Receptor belongs to the nuclear receptor family of transcription factors (NR). It functions as a ligand-regulated transcription factor by binding to cis-regulatory DNA elements in the promoter¹¹⁷.

Estrogen receptors control many physiological processes that can be influenced by agonist or antagonist ligands. Estrogen agonists are for example used in the treatment of postmenopausal osteoporosis¹¹⁸, atherosclerosis¹¹⁹ and Alzheimer's disease¹²⁰. However, the activation of ER can also increase the risk of breast and uterine cancer^{121,122}. Drug discovery efforts concerning the ER α have already resulted in successful drugs that are currently used in therapy^{123,124}.

A second ER subtype or ER β was recently discovered and isolated in 1996¹²⁵. Its discovery caused some excitement in the drug discovery field due to the successes in developing drugs against ER α . The tissue distribution differs considerably¹²⁶ and so does their biological function^{127,128}. Figure 3.7 shows the structure of ER β (A) and the sequence similarity of the two ERs (B).

The two ER receptors share 95% sequence identity in their DNA binding domain, but only $\approx 57\%$ identity in their ligand binding domain (LBD)¹²⁹. The binding pocket itself is however very similar, only two amino acid difference, L384 and M421 in ER α are M336 and L373 in ER β respectively.

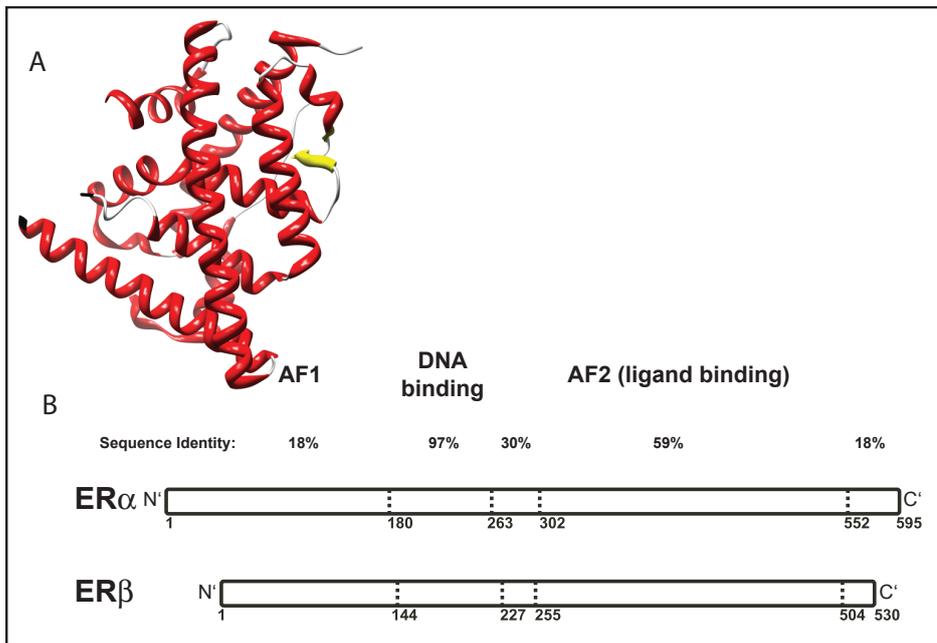


Figure 3.7: The Estrogen Receptor β (A) and its sequence identity to ER α (B)

4 Methods

This chapter describes the theoretical methods used in this work. Figure 4.1 shows the methodological organization of the work.

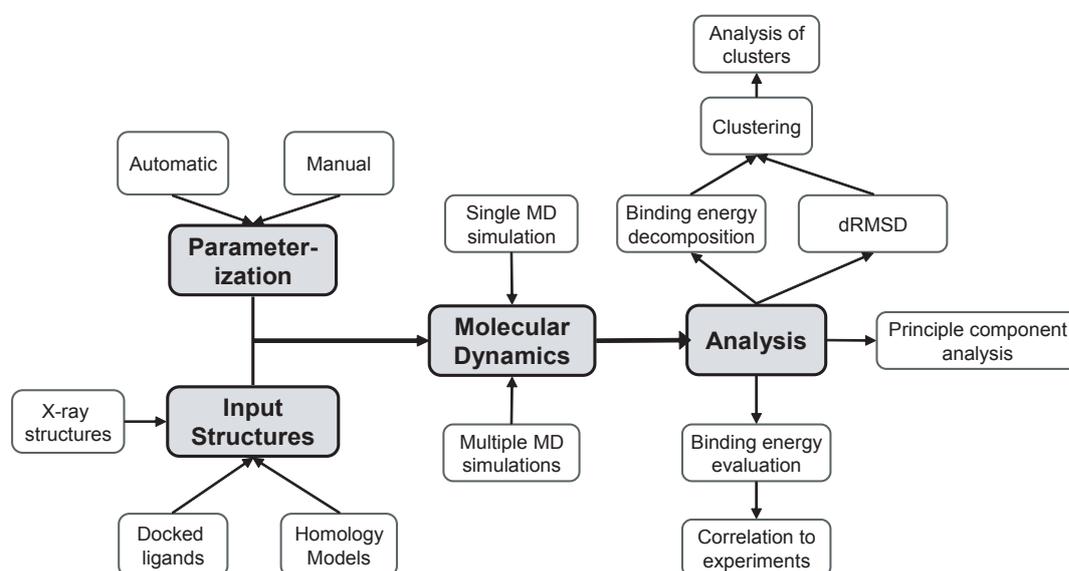


Figure 4.1: A flowchart describing the workflow used in this thesis. The gray squares refer to the common aspects of all the projects, and the white squares refer to the parts that vary between different approaches.

The general workflow for the molecular dynamics simulations is always the same. The input structure (protein and ligands) has to be parameterized before a molecular dynamics simulation is performed. After the molecular dynamics simulation has completed an analysis is performed. The input structures can come from structures that have been solved experimentally (x-ray) or by homology modeling and the lig-

ands are either the natural ligands or poses obtained from docking. These structures have to be parameterized and are done so either by manual or automatic means. The molecular dynamics simulations are either run as a single long simulation or multiple short ones. Finally, the type of analysis that is performed varies according to the problem that is being addressed.

4.1 Parameterization

The ligands used in this work are not a part of the standard CHARMM parameter set⁹¹. As a consequence their specific force field parameters have to be determined, either manually or by an automated procedure.

Manual parameterization

The manual parameterization was done by analogy to known parameters. The atoms of the ligand were compared to similar atoms of known types already existing in the force field and values for partial charges, similar bond, angle and dihedral values were used.

Automatic parameterization

The software package Antechamber was used for the automatic parameterization of partial charges and force field parameters⁹³. The workflow is as following; a pdb file of the ligand to be parameterized first has to be renumbered so that the atom numbers are corresponding to AMBER. Then the Mulliken partial charges are calculated from basic laws of quantum mechanics using the electronic structure program Gaussian¹³⁰. Antechamber is then used to produce the necessary residue topology file and parameter file. The atom type naming is different in AMBER than in CHARMM so all of the atom types of each ligand were renamed and the corresponding files checked for errors and missing parameters.

4.2 Input Structures

All of the work in this thesis is based on structures where three-dimensional coordinates are available. Either the protein structures have been solved experimentally by x-ray crystallography or they are modeled using homology modeling. For the ligands, either they come from known x-ray structures or they come from docking algorithms.

X-ray structures

The experimentally determined structures used for the HIV-1 protease and its ligands are the same as used in a previous study¹³¹ with pdb access codes 1HVL, 1HVK, 1HVI, 1HVJ, 1DIF, 1OHR, 1AJX, 1AJV, 1HTF, 1HPX, 1HSG, 2BPV, 1HBV, 1HOS, 1HPS and 1HPV. For the Estrogen Receptor the structures were selected from the pdbind database based on their experimentally determined binding affinities¹³², pdb access codes 2BJ4, 1B1V, 1X7E, 2FAI, 1GWR, 1GWQ, 1XQC, 1UOM, 1XPC, 1XP9, 1SJ0, 1XP1 and 1XP6. All the coordinates of the protein structures were then obtained from the Protein Data Bank (PDB)¹⁰

Homology Models

The homology models for the HIV-1 protease were based on templates from the Simian Immunodeficiency Virus (1SIP, 51% sequence identity)¹³³, HIV-2 protease (1IDA, 48% sequence identity)¹³⁴, Rous sarcoma virus (1BAI, 40% sequence identity)¹³⁵, Equine infectious anemia virus protease (1FMB, 32% sequence identity)¹³⁶ and Feline immunodeficiency virus (3FIV, 32% sequence identity)¹³⁷. The homology models were built using the comparative modeling server SWISS-MODEL¹⁹. The project submission mode²⁰ was used and all models were validated using Procheck³⁰ to assess the stereochemical quality of the models and Anolea mean force potential assessment³³ to evaluate the environment of heavy atoms.

Docked ligands

First the ligands are prepared for docking using the LigPrep module of Maestro in the Schrodinger suite of software afterwards the bond order of the ligands have to be manually corrected. The protein structure 1HVI was used for the HIV-1 protease as the reference structure in which all the ligands were docked, and the structure 1XPC for the Estrogen receptor. The protein preparation workflow was used to correctly set up the protein by assigning bond order, add hydrogens, generate ionization states and minimize the protein structure. A docking grid was set up with the receptor grid generation tool and the docking was performed with Glide v4.5⁴¹. For the ligands of HIV-1 protease two docking runs were performed, one for the ligands that contain a structural water in the binding pocket and a separate run for the two ligands (AHA001 and AHA006) that do not contain the structural water. The aspartic acid 25 of chain 1, or Asp25 was protonated on OD1.

4.3 Molecular Dynamics

For both of the standard long and multiple short molecular dynamics simulations, the first steps of solvating, equilibration and heating are the same as described below.

The molecular dynamics (MD) simulations were performed using CHARMM¹³⁸ version 30b1 and the all-atom force field CHARMM22⁹¹. The starting coordinates for the MD simulations were the coordinates of the ligands in their experimentally determined structures, or the structures obtained from docking experiments, respectively. All protein-ligand complexes were superimposed, and the coordinates of the ligands were transferred into the chosen reference structures. Partial atomic charges and the all-hydrogens parameters for van der Waals and bonded energy terms for the ligands of HIV-1 protease have been previously determined¹³¹ and applied to validate the MM-GBSA approach¹³⁹. The ligands for the estrogen receptor were automatically parameterized using Antechamber⁹³.

Each of the protein-ligand complexes was solvated with a water sphere of radius 24 Å, centered on a ligand atom which is in the center of the protein-ligand complex. The stochastic boundary method was used¹⁴⁰, with a reaction region of 20 Å and a buffer region of 4 Å. Electrostatic interactions were shifted and calculated with a 12 Å cutoff and the van der Waals interactions were switched at long range.

The water molecules were equilibrated at 300 K for 80 ps with the protein and ligand fixed. Then the complex was minimized for 200 steps of steepest descent minimization and then heated and equilibrated from 0 K to the target temperature in steps of 40ps for each 100K increase. The target temperatures for the validation were 300K and 500K. Harmonic restraints of 5 kcal/mol initially applied to the ligand and protein atoms within the reaction region were slowly removed during the heating. At the target temperature all restraints were removed from the reaction region and equilibration continued for 40ps at constant temperature. Heavy atoms in the buffer region were coupled to a heat bath, using the Langevin equation of motion and a 250 ps⁻¹ friction constant¹⁴¹. A friction constant of 62 ps⁻¹ was applied to the water oxygens.

After the heating and equilibrium steps have completed different molecular dynamics approaches are applied.

4.3.1 Standard long MD simulations

For the standard long molecular dynamics simulation, the simulations were carried out at the target temperature using Langevin dynamics during 500 ps with a time step of 1 fs.

4.3.2 Multiple MD simulations

The general methodology for the parallelization of the multiple short molecular dynamics simulations is shown in figure 4.2.

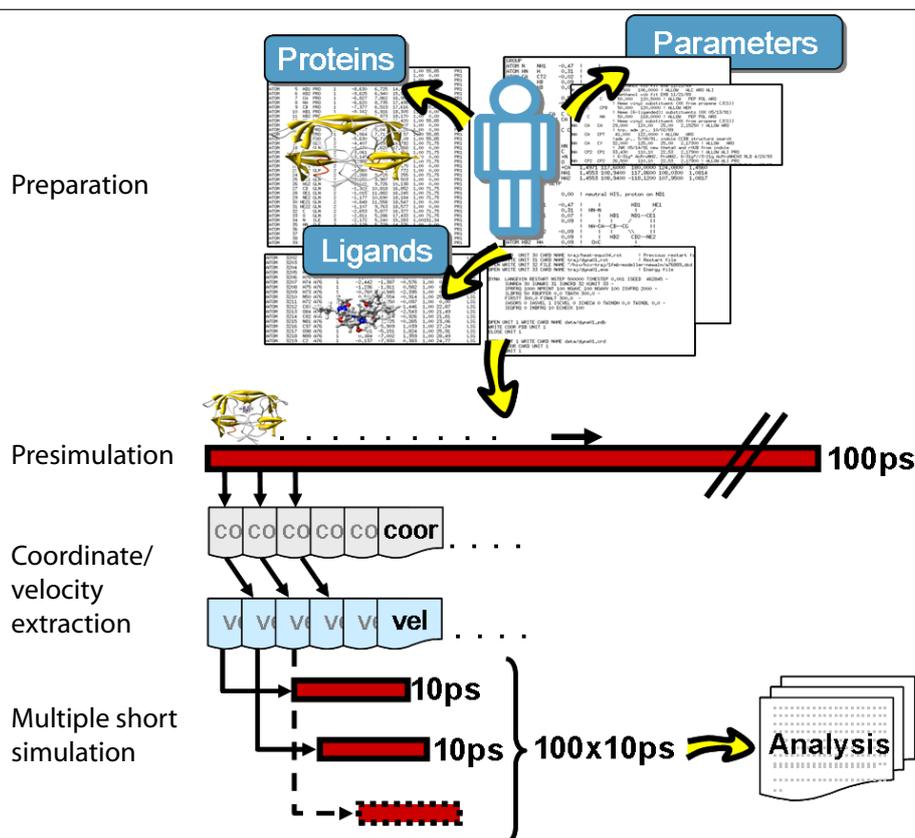


Figure 4.2: An overview of the workflow for the multiple short molecular dynamics simulations approach. First as with standard long molecular dynamics, the protein and ligand coordinates, along with the necessary force field parameters are used to run a so-called pre-MD simulation. This pre-simulations is normally run 100ps and every 1ps coordinates and velocities are extracted. These coordinates and velocities are then recombined and used as starting points for for multiple short molecular dynamics simulations of 10ps each, which can be run on a PC GRID

As for a standard long molecular dynamics simulation, the necessary force field

parameters along with the coordinates of the protein and a number of ligands are used as an input to run a molecular dynamics pre-simulation. This pre-simulation is performed for 100ps with coordinates and velocities are extracted every 1ps. These coordinates and velocities are then recombined and used as starting points for independent short MD simulations of 10ps each, totaling in 100 short MD simulations. As a first attempt the coordinates and velocities are simply combined as coordinate i with velocity $i+1$. These short simulations are then suitable for execution on a PC-GRID, and their total combined simulation time of 1ns is presumed to produce results comparable to a single continuous MD simulation of 500ps^{102,104}.

4.4 Analysis

4.4.1 Binding Energy Evaluation

Frames from both the long and the short MD simulations were extracted every 1 ps, resulting in 500 frames for the long simulation and 1000 for the multiple short ones. The binding free energy ΔG_{bind} was calculated using the MM-GBSA approach⁸⁴. ΔG_{bind} can be decomposed into a sum of the gas phase contribution, ΔE_{vacuo} , the desolvation energy upon binding, ΔG_{desolv} , and an entropy contribution, $T\Delta S$:

$$\Delta G_{\text{bind}} = \langle \Delta E_{\text{vacuo}} \rangle + \langle \Delta G_{\text{desolv}} \rangle - \langle T\Delta S \rangle \quad (4.1)$$

The brackets indicate ensemble averages of the quantities calculated from the extracted frames from each of the trajectories. In this approach, the gas phase contribution is equal to the van der Waals (ΔE_{vdw}) and electrostatic (ΔE_{elec}) interaction energies between the ligand and the protein, and the difference of the internal energy, ΔE_{intra} between the complex and the separated ligand/protein system (deformation energy). However, since we only use the trajectory of the complex, $\Delta E_{\text{intra}} = 0$ and it does not enter the subsequent calculations.

The desolvation energy, ΔG_{desolv} , is the difference between the solvation energy of the complex, $\Delta G_{\text{solv}}^{\text{C}}$, and of the ligand and protein, $\Delta G_{\text{solv}}^{\text{L}}$ and $\Delta G_{\text{solv}}^{\text{P}}$, respectively. The solvation for all three systems can be further divided into electrostatic ($\Delta G_{\text{solv,elec}}$) and nonpolar ($\Delta G_{\text{solv,np}}$) contributions. The electrostatic contribution to the solvation free energy ($\Delta G_{\text{solv,elec}}$) was estimated from the analytical Generalized Born (GB) GB-MV2 model implemented in CHARMM^{142,143}. Calculations using this model are much faster compared to numerically solving the Poisson equation (as required for

PBSA calculations^{83,144}) and were found to reproduce solvation energies with 1% accuracy compared to the Poisson model. Recent results show that the deviation in the desolvation energy between the GB-MV2 model and the PB model is constant, which means that the use of GB-MV2 does not alter the ranking of the ligands¹⁴⁵. Also, using the GB-MV2 model allows for easier decomposition of the electrostatic energy than is possible with PBSA. The nonpolar contribution is assumed to be proportional to the solvent accessible surface area (SASA). This approximation is based on the observation that the solvation of saturated nonpolar hydrocarbons is linearly related to the SASA¹⁴⁶. A value of 0.0072 kcal/mol Å² was used for the surface tension^{84,147,148}. The solvent accessible surface areas were calculated analytically with CHARMM, with a solvent probe radius of 1.4 Å.

The entropic contribution, $T\Delta S$, corresponds to the contributions of translational, ΔS_{trans} , rotational, ΔS_{rot} and vibrational, ΔS_{vib} entropies:

$$-T\Delta S = -T\Delta S_{\text{vib}} - T\Delta S_{\text{trans}} - T\Delta S_{\text{rot}} \quad (4.2)$$

where each of these terms is calculated from standard equations of statistical mechanics^{149,150}. The entropic contribution is required for calculating absolute binding free energies. $T\Delta S_{\text{trans}}$ and $T\Delta S_{\text{rot}}$ are functions of the mass and moments of inertia, while $T\Delta S_{\text{vib}}$ is calculated from a normal mode analysis¹⁵⁰. Since the normal mode calculations are computationally very demanding, $-T\Delta S$ was only averaged over 20 frames of the trajectory. The VIBRAN module of CHARMM was used to determine the normal modes and normal mode frequencies. The normal modes were calculated *in vacuo* for a fully minimized structure of the molecule (complex and protein:ligand) with a distance dependent dielectric, $\epsilon = 4r$. The minimization was performed using the Adopted Basis Newton-Raphson algorithm, until the root mean square of the energy gradient reached 10^{-7} kcal/molÅ. This gradient value is expected to give only real frequencies¹⁵⁰. The minimized structures were also used to calculate the moments of inertia. Since $T\Delta S_{\text{vib}}$ is computationally very demanding to calculate and we are only interested in relative energies, the entropic term was calculated for all the ligands docked in the reference protein 1HVI to verify that a ranking of the ligands can be obtained without including the entropic term. Thus, the working equation to

estimate binding free energies is

$$\langle \Delta G_{\text{bind}} \rangle = \langle \Delta E_{\text{elec}} \rangle + \langle \Delta E_{\text{vdw}} \rangle + \langle \Delta G_{\text{desolv,elec}} \rangle + \langle \Delta G_{\text{desolv,np}} \rangle . \quad (4.3)$$

4.4.2 Correlation to Experimental Values

The correlation of the calculated binding free energies was based on its linear regression to the experimental binding free energy values.

The correlation coefficient is determined by first calculating the standard deviations of the x and y datasets:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2} \quad (4.4)$$

The standard deviation is calculated the same way for y. Then the covariance between the two data sets is determined by:

$$\sigma_{x,y} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 (y_i - \bar{y})^2 \quad (4.5)$$

Finally the correlation coefficient is defined as:

$$r = \frac{\sigma_{x,y}}{\sigma_x \sigma_y} \quad (4.6)$$

4.4.3 Principle Component Analysis

The program Gromacs¹⁵¹ was used to examine the positional differences of the coordinates from different molecular dynamics approaches and simulations at different temperatures. The *g_covar* module was used to calculate and diagonalize the covariance matrix and corresponding eigenvectors are determined. The eigenvectors were analyzed with the *g_anaeig* module.

4.4.4 Clustering

We are interested in finding out if the different simulations approaches (standard long and multiple short) are equivalent. Clustering is a commonly used method of

grouping together data that are similar in some way, which is determined by a given criteria or cluster metric. In this work we have used two different cluster metrics. They both deal with residues that are in the binding pocket, within 5 Å of the ligand and consider each frame of the molecular dynamics simulation separately. The first metric is based on a per residue binding free energy decomposition and the second on a simple root mean square deviation (dRMSD) to a reference structure,

Binding Energy Decomposition

To estimate the contribution of each of the residues of the protein to the total binding free energy, an energy decomposition of the binding free energy for each residue was performed.

For the electrostatic interaction energy between the residues of the protein and the ligand, one half of the energy is attributed to the ligand and the other half to the protein residue. The contribution of atom i to the total electrostatic interaction energy is given by:

$$E_{\text{elec}}^i = \frac{1}{2} \sum_j \frac{q_i q_j}{r_{ij}} \quad (4.7)$$

where j loops over all the atoms of the component to which i does not belong, i.e. over the protein if i belongs to the ligand, or over the ligand if i belongs to the protein. r_{ij} is the distance between atoms i and j and q_i and q_j are their atomic charges. For the pairwise van der Waals interactions between the protein and ligand, one half is attributed to each atom involved in the interaction.

The solvent accessible surface area of atom i in the complex, $SASA^{iC}$, and in the protein, $SASA^{iP}$, and ligand, $SASA^{iL}$, is calculated by CHARMM. The contribution of atom i to the nonpolar desolvation term is thus $\Delta G_{\text{np,solv}}^i = \sigma(SASA^{iC} - (SASA^{iP} + SASA^{iL}))$. The electrostatic solvation term is calculated using the GB-MV2 approach, which uses the Still *et al.*¹⁴⁷ expression for the electrostatic solvation term:

$$\Delta G_{\text{elec,solv}} = k \sum_{i,j} \frac{q_i q_j}{\sqrt{r_{ij}^2 + \alpha_i \alpha_j \exp(-r_{ij}^2 / K_s \alpha_i \alpha_j)}} \quad (4.8)$$

where $k = -166.0(\epsilon_{\text{solute}}^{-1} - \epsilon_{\text{solvent}}^{-1})$. ϵ_{solute} and $\epsilon_{\text{solvent}}$ are the dielectric constant, i.e. 1 and 80, respectively and α_i and α_j are the Born radii of atoms i and j , calculated with the GB-MV2 approach. i and j loop over all atoms of the system. The constant K_s is equal to 8 in the GB-MV2 approach. The contribution of atom i to $\Delta G_{\text{elec,solv}}^X$ can

therefore be written as:

$$\Delta G_{\text{elec,solv}}^{i,X} = k \frac{q_i^2}{\alpha_i} + \frac{1}{2} k \sum_{j \neq i} \frac{q_i q_j}{\sqrt{r_{ij}^2 + \alpha_i \alpha_j \exp(-r_{ij}^2 / K_s \alpha_i \alpha_j)}} \quad (4.9)$$

where X stands for the complex, the protein or the ligand. The contribution of atom i to the electrostatic desolvation energy is $\Delta G_{\text{elec,desolv}}^i = \Delta G_{\text{elec,solv}}^{i,C} - (\Delta G_{\text{elec,solv}}^{i,P} + \Delta G_{\text{elec,solv}}^{i,L})$. Adding all these atomic contributions over the atoms of a given residue or side chain or backbone fragment yields its contribution to the total binding free energy.

This binding free energy is calculated for those residues that fall within 5 Å of the ligand for all frames of the molecular dynamics simulations that are being compared.

Distance RMSD

For calculating RMSD as distance metric, only atoms of the side-chains of the residues that fall within 5 Å of the ligand are considered. First for the reference structure (x-ray structure), all the pairwise distances between each each residue atom to each ligand atom are calculated by:

$$D_{\text{distance,ref}}^{\text{res}}(\Gamma^{\text{ref}}) = \sqrt{\frac{1}{N_{\text{pair}}} \sum_{i,j} (r_i^{\text{res}} - r_j^{\text{lig}})^2}, \quad (4.10)$$

where Γ^{ref} correspond to the reference structure. N_{pair} the total number of distances compared, that is all the interacting pairs and $r_i^{\text{res}} - r_j^{\text{lig}}$ is the distance between atoms i of the protein residue and atoms j of the ligand.

The $D_{\text{distance,ref}}^{\text{res}}$ is calculated for all of the residues of the reference structure and then for all residues of all frames for the molecular dynamics simulations ($D_{\text{distance,frame}}^{\text{res}}$). The difference in distance for each the residue for each the frames during the molecular dynamics simulation can then be defined as:

$$\Delta D_{(\text{ref-frame})}^{\text{res}} = |D_{\text{distance,ref}}^{\text{res}} - D_{\text{distance,frame}}^{\text{res}}| \quad (4.11)$$

In the same way as with the binding free energy criteria, the $\Delta D_{(\text{ref-frame})}^{\text{res}}$ values are given an integer score depending on which distance interval they belong to.

Scoring

Once the different clustering metrics have been defined and calculated, the corresponding results from the molecular dynamics simulations can be compared and

clustered. The analysis workflow is depicted in figure 4.3.

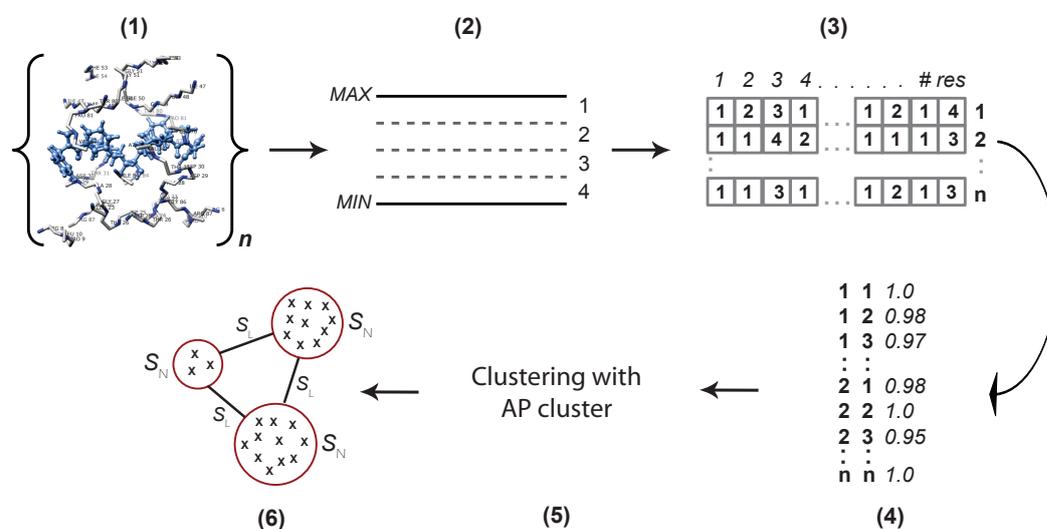


Figure 4.3: An overview of the fingerprint definition scheme. The first step identifies all residues that fall within a certain radius of the ligand (1). In the second step the maximum and minimum values for the criteria of choice are defined and the values are split into even intervals (2). In the third step a fingerprint is created for each of the frame under investigation (4). The fingerprints are then scored all-to-all (4) and clustered as described in the text (5-7)

The first part of figure 4.3 (1), or the definition of a clustering metric and calculation of values has already been discussed. The following sections describe the fingerprint definition figure 4.3 (2 and 3), scoring of the results figure 4.3 (4) and the affinity propagation algorithm to cluster the data 4.3 (5-6).

Fingerprint Definition

As described above, for each of the residues of the protein that falls within 5 Å of the ligand a binding free energy and distance metric is calculated for all frames considered. In order to simplify the fingerprint definition the resulting continuous values from the rmsd or binding free energy metric are assigned an integer score depending on which interval they fall into. For the binding free energy metric the intervals are divided for each 1 kcal/mol energy contributions up to ± 8 kcal/mol. In this way residues that have an energy contribution of 0.5 kcal/mol in one frame get the score 1

(the interval from 0-1 kcal/mol), whereas if in the next frame the energy contribution is 1.6 kcal/mol it would get the score 2 (the interval from 1-2 kcal/mol), until ± 8 kcal/mol all contributions over +8 kcal/mol get the same score and in the same way the contributions that are less than -8 kcal/mol.

The fingerprint for the distance metric is defined in a similar way, by dividing the dRMSD value of each residue to the ligand for each frame of the trajectory to the same distance in the reference structure into intervals of 0.5 Å up to the distance difference of ± 3 Å

Once these scores have been defined the fingerprints can be produced as shown in figure 4.3 (3). Each of the columns refers to a residue and each of the lines corresponds to a single frame of a molecular dynamics simulation. For each of the metrics a fingerprint represents a profile of the interactions of residues to the ligand for the molecular dynamics simulation frame under consideration.

The fingerprints between frames of a molecular dynamics simulations can now be compared pairwise and given a score depending on how distant they are from each other (Hamming score) as described below.

Hamming scoring

The Hamming scoring function is defined as followed:

$$D_{AB} = \sum_{i=0}^m | A_i - B_i | \quad (4.12)$$

where D_{AB} is the distance score between fingerprint A and B and $| A_i - B_i |$ is the absolute difference between the values of column i of fingerprint A and B.

The calculation by this scoring functions results in a list of pairwise comparisons as depicted in figure 4.3 (4).

Clustering

Once the fingerprints have been defined and have been scored according to the Hamming or Tanimoto score, they can be clustered based on their similarity. In this work we use the affinity propagation clustering.

Affinity Propagation

The affinity propagation algorithm¹⁵² identifies exemplars among data points and forms clusters of data points around these exemplars. It operates by simultaneously

considering all data point as potential exemplars and exchanging messages between data points until a good set of exemplars and clusters emerges.

4.4.5 Analysis of Clusters

Once a reliable set of clusters has been obtained, it is interesting to analyze the different clusters to get an idea of which of the residues differ between the clusters and which stay the same throughout and between molecular dynamics simulation. A novel method to analyze multiple sequence alignment was adapted for that task¹⁵³.

Proteinkeys

Proteinkeys¹⁵³ is a method to identify specificity residues in sets of proteins related by evolution, where the specificity residues are defined as those conserved within a subfamily (cluster) but differ between subfamilies, that is encode functional diversity. This method can be adapted to analyze the clusters obtained by Affinity propagation.

The adapted proteinkeys method is as follows; The total number of permutations in a column i of cluster k is given by a combinatorial formula:

$$Z_{i,k} = \frac{N_k!}{\prod_{\alpha=1,..,n} N_{\alpha,i,k}!} \quad (4.13)$$

Where N_k is the number of members in each cluster k , $N_{\alpha,i,k}$ is the number of residues with a value of type α in column i of cluster k .

Then combinatorial entropy is used as an additive measure for comparing different distributions of fingerprint values between clusters:

$$S_i = \sum_i \ln Z_{i,k} \quad (4.14)$$

Specificity residues are those that have similar properties within a cluster but differ between clusters. At one extreme the column specific S_i is zero if values of one type populate this column. So $S_i = 0$ for properties that are the same or perfect specificity residues in column i .

At the other extreme, uniformly distributed residues, S_i has a maximal value given by the background entropy \tilde{S}_i :

$$\tilde{S}_i = \sum_k \ln \tilde{Z}_{i,k} = \sum_k \frac{N_k!}{\prod_{\alpha=1,..,n} \tilde{N}_{\alpha,i,k}!} \quad (4.15)$$

Where $\tilde{N}_{\alpha,i,k}$ is the expected number of the residue of a type α in the column i of cluster k , provided that all the residues in the column are uniformly mixed, or:

$$\tilde{N}_{\alpha,i,k} = \frac{N_k N_{\alpha,i}}{N} \quad (4.16)$$

where $N_{\alpha,i,k}$ is the number of the residues of a type α in column i and N is the total number of cluster members. Since $\tilde{N}_{\alpha,i,k}$ as shown in equation 5.17 can be non integer numbers, $\tilde{N}_{\alpha,i,k}!$ is calculated using the relation $X! = \Gamma(X + 1)$.

As the numerical measure of order over disorder, the entropy difference can be calculated: $\Delta S_i = S_i - \tilde{S}_i$

5 Results

5.1 How Inaccuracies in Protein Structure Models Affect Estimates of Protein-Ligand Interactions

Thorsteinsdottir HB, Schwede T, Zoete V, Meuwly M.
Proteins. (2006) 65:407-23

The work presented here addresses one of the bottlenecks of structure-based drug design, which is the limited availability of experimentally determined protein structures. In the cases where no protein structures are available, an alternative can be to build a homology model, but they need to be sufficiently accurate to be of use for drug discovery. Validation of these homology models is therefore a crucial aspect in drug development. One important question we aim to address is how errors and inaccuracies of the homology models affect the subsequent molecular modeling of protein-ligand interaction. We study this by utilizing a well characterized protein system with ligands whose binding free energy has been determined experimentally. By doing this we can study the effects of sequence variations and introduce systematic errors in the protein model. This enables us to simulate the typical errors that occur during homology modeling, e.g. sub-optimal template selection or side chain placement, to quantify the effect on ligand binding affinity and ranking of ligands

How Inaccuracies in Protein Structure Models Affect Estimates of Protein–Ligand Interactions: Computational Analysis of HIV-I Protease Inhibitor Binding

Holmfridur B. Thorsteinsdottir,^{1,2} Torsten Schwede,^{1,2} Vincent Zoete,³ and Markus Meuwly^{4*}

¹Biozentrum, University of Basel, Klingelbergstrasse 50, 4056 Basel, Switzerland

²Swiss Institute of Bioinformatics, Biozentrum, University of Basel, Switzerland

³Swiss Institute of Bioinformatics, BEP-UNIL, CH-1015 Lausanne, Switzerland

⁴Department of Chemistry, University of Basel, Klingelbergstrasse 80, 4056 Basel, Switzerland

ABSTRACT The influence of possible inaccuracies that can arise during homology modeling of protein structures used for ligand binding studies were investigated with the molecular mechanics generalized Born surface area (MM-GBSA) method. For this, a family of well-characterized HIV-I protease–inhibitor complexes was used. Validation of MM-GBSA led to a correlation coefficient ranging from 0.72 to 0.93 between calculated and experimental binding free energies ΔG . All calculated ΔG values were based on molecular dynamics simulations with explicit solvent. Errors introduced into the protein structure through misplacement of side-chains during rotamer modeling led to a correlation coefficient between ΔG_{calc} and ΔG_{exp} of 0.75 compared with 0.90 for the correctly placed side chains. This is in contrast to homology models for members of the retroviral protease family with template structures ranging in sequence identity between 32% and 51%. For these protein models, the correlation coefficients vary between 0.84 and 0.87, which is considerably closer to the original protein (0.90). It is concluded that HIV-I low sequence identity with the template structure still allows creating sufficiently reliable homology models to be used for ligand-binding studies, although placement of the rotamers is a critical step during the modeling. *Proteins* 2006;65:407–423. © 2006 Wiley-Liss, Inc.

Key words: molecular dynamics; homology modeling; HIV; ligand binding

INTRODUCTION

One of the bottlenecks of structure-based drug design is the limited availability of experimentally determined protein structures. In virtual screening, a widely used approach in drug discovery, a library of small molecules is docked into the active sites of proteins. Accurate and reliable atomic representations of the receptor proteins are required to correctly model possible ligand–protein interactions. In spite of advances in X-ray crystallography and NMR protein determination techniques, there is still a considerable gap between the number of

sequences and structures available. Although the UniProt nonredundant protein sequence database¹ contains just below 2 million sequences, the Protein Data Bank (PDB)² contains about 12,500 different structures (both numbers based on 90% sequence identity). Thus, for most protein sequences, there are no experimental structures available. In such cases, homology models can provide a valuable alternative. Homology modeling (HM) uses known experimental structures as a template to create a three-dimensional model for a similar protein for which the structure is unknown.³ This is possible because structures within a given protein fold family are better conserved than their sequences.⁴ Validation of homology models is a crucial aspect in drug development. One important question is to assess to what extent errors and inaccuracies of the homology models affect the subsequent molecular modeling of protein–ligand interactions. To address this question, we have selected the HIV-I protease system.

HIV-I protease is an essential component in the life cycle of the HIV virus. The function of the protein is to cleave viral polyproteins so that the viral particles can mature. Because of this important role, HIV-I protease is one of the main targets for AIDS therapy, and currently there are four inhibitors of the HIV-I protease in clinical use and more that are undergoing clinical trials.⁵ HIV-I protease is a member of the aspartyl protease family.⁶ The protein is active as a homodimer of two identical 99 amino acid chains. More than 150 experimental structures are publicly available in the Protein Data Bank (<http://www.rcsb.org/pdb/>) or the HIV PR database (<http://mcl1.ncifcrf.gov/hivdb/>).

In a recent study, the binding free energies of 16 HIV-I protease inhibitor complexes were investigated computationally using molecular dynamics simulations and

Grant sponsor: Swiss National Science Foundation.

*Correspondence to: Markus Meuwly, Department of Chemistry, University of Basel, Klingelbergstrasse 80, 4056 Basel, Switzerland. E-mail: m.meuwly@unibas.ch

Received 6 October 2005; Revised 15 March 2006; Accepted 9 May 2006

Published online 28 August 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21096

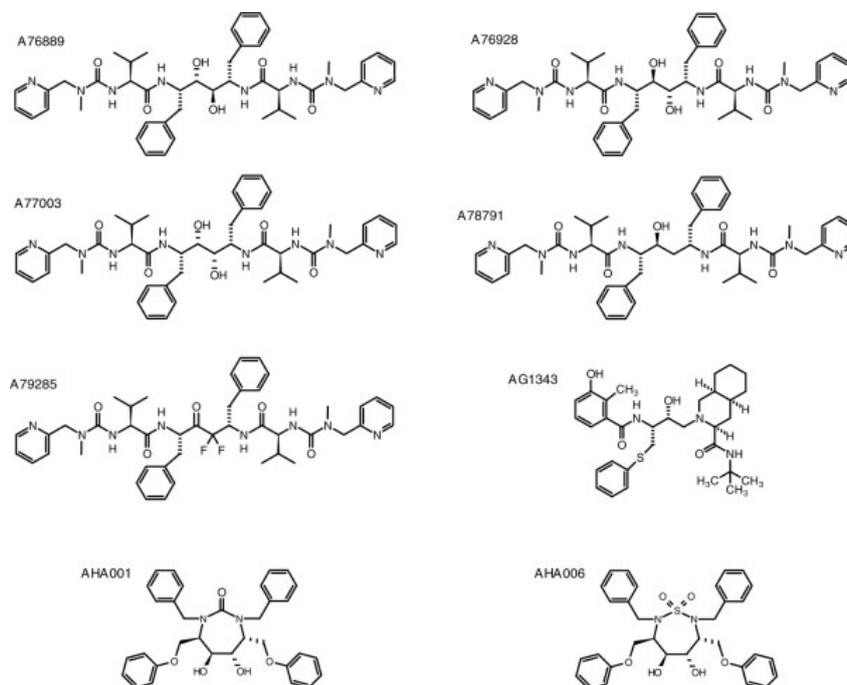


Fig. 1. Structures of ligands 1–8 used for the present study.

techniques based on a linear interaction energy (LIE)-like method. A favorable correlation of 0.91 between calculated and experimental values was observed.⁷ In the present study, a protocol based on the molecular mechanics-generalized Born surface area (MM-GBSA) approach⁸ to calculate binding free energies of the same 16 HIV-I protease inhibitor complexes is first investigated. Such an efficient approach to estimate ligand binding free energies allows studying the effects of variations in the protein structure on the affinity calculations. Here, MM-GBSA is first validated and then applied to explore the effect of structural variations in view of typical inaccuracies that can occur during protein structure homology modeling.

The ligands used in our study have different chemical properties; there are diol inhibitors and analogues of those ligands with thiophenyl ether and phenol–amide substituents, penicillin-derived, amides and amino sulfonamides. The experimental binding free energies of ligands differ by ≈ 5 kcal/mol, and the experiments were carried out by a number of different groups. Where given, errors in experimental binding free energies are estimated between 10 and 20%.^{9,10} Here we attribute a conservative error of 10% for all the experimentally determined binding free energies.

The present work is structured as follows. First, the computational and theoretical methods used are presented. Next, the MM-GBSA approach to estimate free energies and to rank a number of ligands is validated

with respect to experimental data. The results are analyzed in detail and in a next step, MM-GBSA is applied to erroneous protein structures. They come either from suboptimally placed side chains or from structures based on homology models with different degrees of sequence similarities. Finally, the results are discussed in view of experiments and other modeling studies.

THEORETICAL METHODS

For the present study, 16 well-characterized inhibitors (Figs. 1 and 2) of the HIV-I protease were selected. Crystal structures for each of the 16 ligands complexed to HIV-I protease and experimental values for the binding free energy are available. Table I provides PDB accession code, resolution, and the experimentally determined binding free energies for each of the 16 ligand–protein complexes. In all complexes, the protease corresponds to the wild type sequence.

As shown by Zoete et al.,²³ the conformation for these 16 ligand–protein complexes is well conserved, with the most rigid part around the binding site of the ligand (Table I reports the root mean square deviations (RMSD) of the backbone atoms for each of the ligand–protein complexes to the reference structure of 1HVI).

Molecular Dynamics Simulations

The molecular dynamics (MD) simulations were performed using CHARMM²⁴ version 30b1 and the “all-atom”

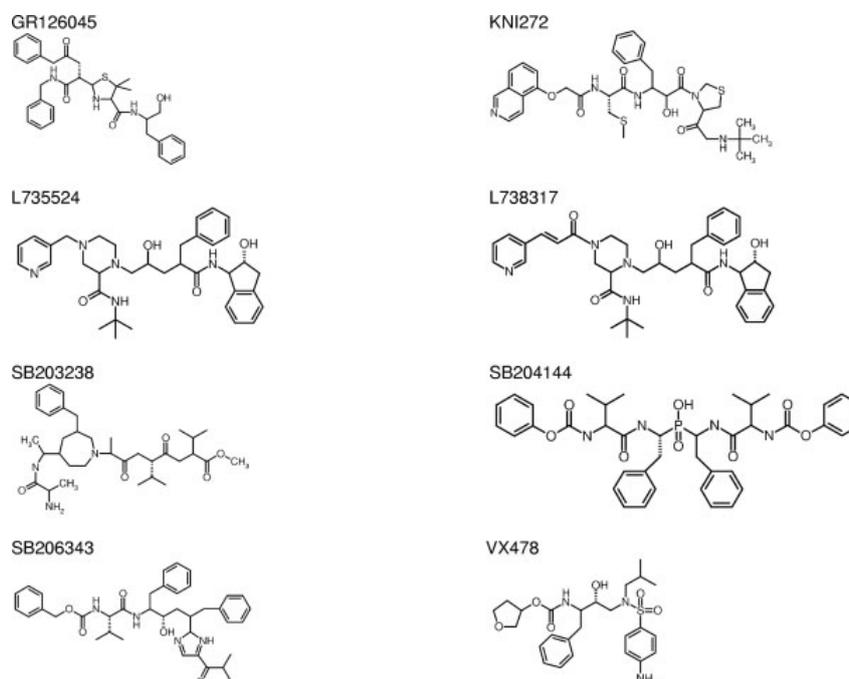


Fig. 2. Structures of ligands 9–16 used for the present study.

force field CHARMM22.²⁵ The starting coordinates for the MD simulations were the coordinates of the 16 ligands in their experimentally determined structures, the ones of 1HVI, and 1OHR, or the structures obtained from homology modeling, respectively. All protein–ligand complexes were superimposed, and the coordinates of the ligands were transferred into the chosen reference structures. Partial atomic charges and the all-hydrogens parameters for van der Waals and bonded energy terms for the 16 ligands have been previously determined.⁷ Depending on the inhibitor, either residue Asp25 or Asp25' of the HIV-I protease can be protonated in a complex.^{16,26–28} In our work, to facilitate comparisons, protonation of Asp25 for all protein–ligand complexes is assumed.

Each of the protein–ligand complexes was solvated with a water sphere of radius 24 Å, centered at the oxygen O48 atom of ligand A77003, which is in the center of the protein–ligand complex. Depending on the ligand present, a total of 2880–2985 water molecules were added, resulting in a system of 6104–6208 atoms. The stochastic boundary method was used²⁹ with a reaction region of 20 Å and a buffer region of 4 Å. Electrostatic interactions were shifted and calculated with a 12 Å cut-off and the van der Waals interactions were switched at long range.

The water molecules were equilibrated at 300 K for 80 ps with the protein and ligand fixed. Then the complex was minimized for 200 steps of steepest descent minimization and then heated and equilibrated from 0 to 300 K during 150 ps. Harmonic restraints of 5 kcal/mol ini-

TABLE I. PDB Codes, Ligand Names, and the Experimentally Determined Binding Free Energy of all the Ligand–Protein Complexes

PDB	Ligand	Resolution (Å)	RMSD (Å) ^a	$\Delta G_{\text{bind}}^{\text{exp}}$	Ref.
1HVL	A76889	1.80	0.15	–14.16	[11]
1HVK	A76928	1.80	0.18	–15.60	[11]
1HVI	A77003	1.80	0.00	–15.54	[11]
1HVJ	A78791	2.00	0.14	–16.22	[11]
1DIF	A79285	1.70	0.17	–15.17	[12]
1OHR	AG1343	2.10	0.21	–12.38	[13]
1AJX	AHA001	2.00	0.68	–11.26	[14]
1AJV	AHA006	2.00	0.75	–10.98	[15]
1HTF	GR126045	2.20	0.55	–9.82	[15]
1HPX	KNI272	2.00	0.60	–16.02	[16]
1HSG	L735524	2.00	0.62	–13.21	[17]
2BPV	L738317	1.90	0.61	–10.51	[18]
1HBV	SB203238	2.30	0.60	–9.06	[19]
1HOS	SB204144	2.30	0.57	–12.17	[20]
1HPS	SB206343	2.30	0.64	–13.12	[21]
1HPV	VX478	1.90	0.64	–13.12	[22]

Energies are in kcal/mol.

^aRMSD of backbone atoms to the reference structure 1HVI1.

tially applied to the ligand and protein atoms within the reaction region were slowly removed during the heating. At 300 K, all restraints were removed from the reaction region. Heavy atoms in the buffer region were coupled to a heat bath, using the Langevin equation of motion and a 250 ps^{–1} friction constant.³⁰ A friction constant of

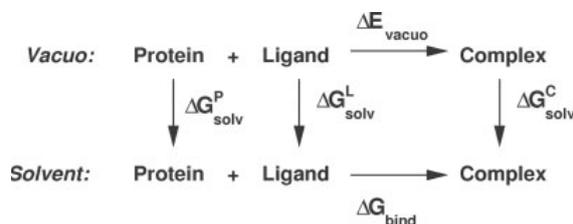


Fig. 3. The thermodynamic cycle used to estimate the binding free energy of the ligand–protein complexes in vacuo, ΔE_{vacuo} , and in water, ΔG_{bind} , $\Delta G_{\text{solv}}^{\text{P}}$, $\Delta G_{\text{solv}}^{\text{L}}$, and $\Delta G_{\text{solv}}^{\text{C}}$, are the solvation energies of the protein, ligand, and complex, respectively.

62 ps⁻¹ was applied to the water oxygens. MD simulations used for the analysis were carried out at 300 K, using Langevin dynamics during 500 ps with a time step of 1 fs. For the validation simulations (ligands in their respective protein) all trajectories were extended to 2 ns.

Binding Free Energy Estimation

Frames from the MD simulation were extracted every 1 ps, which gives a total of 500 frames per trajectory (2000 for the validation runs). The binding free energy ΔG_{bind} was calculated using the MM-GBSA approach,⁸ according to the thermodynamic cycle shown in Figure 3. ΔG_{bind} can be decomposed into a sum of the gas phase contribution, ΔE_{vacuo} , the desolvation energy upon binding, ΔG_{desolv} , and an entropy contribution, $T\Delta S$:

$$\Delta G_{\text{bind}} = \langle \Delta E_{\text{vacuo}} \rangle + \langle \Delta G_{\text{desolv}} \rangle - \langle T\Delta S \rangle \quad (1)$$

The brackets indicate ensemble averages of the quantities calculated from the 500 frames along the trajectory. In this approach, the gas phase contribution is equal to the van der Waals (ΔE_{vdw}) and electrostatic (ΔE_{elec}) interaction energies between the ligand and the protein, and the difference of the internal energy, ΔE_{intra} , between the complex and the separated ligand/protein system (deformation energy). However, since we only use the trajectory of the complex, $\Delta E_{\text{intra}} = 0$, and it does not enter the subsequent calculations.

The desolvation energy, ΔG_{desolv} , is the difference between the solvation energy of the complex, $\Delta G_{\text{solv}}^{\text{C}}$, and of the ligand and protein, $\Delta G_{\text{solv}}^{\text{L}}$ and $\Delta G_{\text{solv}}^{\text{P}}$, respectively. The solvation for all three systems can be further divided into electrostatic ($\Delta G_{\text{solv,elec}}$) and nonpolar ($\Delta G_{\text{solv,np}}$) contributions. The electrostatic contribution to the solvation free energy ($\Delta G_{\text{solv,elec}}$) was estimated from the analytical Generalized Born (GB) GB-MV2 model implemented in CHARMM.^{31,32} Calculations using this model are much faster compared with those solved numerically using the Poisson equation (as required for PBSA calculations^{33,34}) and were found to reproduce solvation energies with 1% accuracy compared with the Poisson model. Recent results show that the deviation in the desolvation energy between the GB-MV2 model and the PB model is constant, which means that the use of GB-MV2 does not alter the ranking of the ligands.³⁵ Also, using the GB-MV2 model allows for

easier decomposition of the electrostatic energy than is possible with PBSA. The nonpolar contribution is assumed to be proportional to the solvent accessible surface area (SASA). This approximation is based on the observation that the solvation of saturated nonpolar hydrocarbons is linearly related to the SASA.³⁶ A value of 0.0072 kcal/mol Å² was used for the surface tension.^{8,37,38} The solvent accessible surface areas were calculated analytically with CHARMM, with a solvent probe radius of 1.4 Å.

The entropic contribution, $T\Delta S$, corresponds to the contributions of translational, ΔS_{trans} , rotational, ΔS_{rot} , and vibrational, ΔS_{vib} , entropies:

$$-T\Delta S = -T\Delta S_{\text{vib}} - T\Delta S_{\text{trans}} - T\Delta S_{\text{rot}} \quad (2)$$

where each of these terms is calculated from standard equations of statistical mechanics.^{39,40} The entropic contribution is required for calculating absolute binding free energies. $T\Delta S_{\text{trans}}$ and $T\Delta S_{\text{rot}}$ are functions of the mass and moments of inertia, while $T\Delta S_{\text{vib}}$ is calculated from a normal mode analysis.⁴⁰ Since the normal mode calculations are computationally very demanding, $-T\Delta S$ was only averaged over 20 frames of the trajectory. The VIBRAN module of CHARMM was used to determine the normal modes and normal mode frequencies. The normal modes were calculated in vacuo for a fully minimized structure of the molecule (complex and protein:ligand) with a distance-dependent dielectric, $\epsilon = 4r$. The minimization was performed using the Adopted Basis Newton-Raphson algorithm, until the root mean square of the energy gradient reached 10⁻⁷ kcal/mol Å. This gradient value is expected to give only real frequencies.⁴⁰ The minimized structures were also used to calculate the moments of inertia. Since $T\Delta S_{\text{vib}}$ is computationally very demanding to calculate and as we are only interested in relative energies, the entropic term was calculated for all the ligands docked in the reference protein 1HVI to verify that a ranking of the ligands can be obtained without including the entropic term. Thus, the working equation to estimate binding free energies is

$$\langle \Delta G_{\text{bind}} \rangle = \langle \Delta E_{\text{elec}} \rangle + \langle \Delta E_{\text{vdw}} \rangle + \langle \Delta G_{\text{desolv,elec}} \rangle + \langle \Delta G_{\text{desolv,np}} \rangle \quad (3)$$

Binding Free Energy Decomposition

To estimate which residues of the protein make the most important contributions to the ligand binding, an energy decomposition of the binding free energy was performed. For the electrostatic interaction energy between the residues of the protein and the ligand, one half of the energy is attributed to the ligand and the other half to the protein residue. The contribution of atom i to the total electrostatic interaction energy is given by

$$E_{\text{elec}}^i = \frac{1}{2} \sum_j \frac{q_i q_j}{r_{ij}} \quad (4)$$

where j loops over all the atoms of the component to which i does not belong, that is, over the protein if i

belongs to the ligand, or over the ligand if i belongs to the protein. r_{ij} is the distance between atoms i and j and q_i and q_j are their atomic charges. For the pairwise van der Waals interactions between the protein and ligand, one half is attributed to each atom involved in the interaction.

The solvent accessible surface area of atom i in the complex, $SASA^{i,C}$, and in the protein, $SASA^{i,P}$, and ligand, $SASA^{i,L}$, is calculated by CHARMM. The contribution of atom i to the nonpolar desolvation term is thus $\Delta G_{np,solv}^i = \sigma(SASA^{i,C} - (SASA^{i,P} + SASA^{i,L}))$. The electrostatic solvation term is calculated using the GB-MV2 approach, which uses the expression of Still et al.³⁸ for the electrostatic solvation term:

$$\Delta G_{elec,solv} = k \sum_{i,j} \frac{q_i q_j}{\sqrt{r_{ij}^2 + \alpha_i \alpha_j \exp(-r_{ij}^2 / K_s \alpha_i \alpha_j)}} \quad (5)$$

where $k = -166.0(\epsilon_{solute}^{-1} - \epsilon_{solvent}^{-1})$. ϵ_{solute} and $\epsilon_{solvent}$ are the dielectric constant, that is, 1 and 80, respectively, and α_i and α_j are the Born radii of atoms i and j , calculated with the GB-MV2 approach. i and j loop over all atoms of the system. The constant K_s is equal to 8 in the GB-MV2 approach. The contribution of atom i to $\Delta G_{elec,solv}^X$ can therefore be written as

$$\Delta G_{elec,solv}^{i,X} = k \frac{q_i^2}{\alpha_i} + \frac{1}{2} k \sum_{j \neq i} \frac{q_i q_j}{\sqrt{r_{ij}^2 + \alpha_i \alpha_j \exp(-r_{ij}^2 / K_s \alpha_i \alpha_j)}} \quad (6)$$

where X stands for the complex, the protein, or the ligand. The contribution of atom i to the electrostatic desolvation energy is $\Delta G_{elec,desolv}^i = \Delta G_{elec,solv}^{i,C} - (\Delta G_{elec,solv}^{i,P} + \Delta G_{elec,solv}^{i,L})$. Adding all these atomic contributions over the atoms of a given residue or side chain or backbone fragment yields its contribution to the total binding free energy.

Homology Modeling

Homology modeling (HM) aims at predicting the three-dimensional structure of a given protein based on information derived from experimentally determined structures of homologous proteins. HM methods consist of four steps: template selection, alignment, model building, and model evaluation. For a given target sequence, templates are selected from a database of known structures. In general, sequence similarity between the target and template is a good indicator for template selection. Next, the template sequence(s) and the target sequence are aligned. This is a crucial step in HM, since an incorrect alignment leads to errors that cannot be corrected later on. Typically, for closely related proteins with sequence identity above 50%, the generated alignments are correct, in particular for the better conserved active sites of a protein. For lower sequence identities, it is advisable to manually inspect the alignment to ensure that insertions and deletions have not been placed in structurally unfavorable regions.⁴¹ Based on the template structure(s) and the alignment between the target and

template(s), a model for the target sequence can be built. To assist with the homology modeling, automated methods are available. In general, they can be divided in methods based on rigid fragment assembly such as SWISS-MODEL^{42,43} or Composer,⁴⁴ and spatial restraint methods as used in Modeller.⁴⁵

Errors in HM can occur at various stages. Typically, errors in models increase with decreasing sequence similarity between the target and template. A low sequence identity between the target and template sequences increases the risk of alignment errors and incorrect side chain placements. Template selection is an important factor in homology modeling. Minor or major structural variations between different templates can influence the subsequent modeling. For example, the template structure can be in the apo or complex forms, open or closed conformation, and have different bound ligands.

In the present work, homology models for HIV-I protease were based on template structures from (a) Simian Immunodeficiency Virus protease (1SIP, 2.3 Å resolution) sharing 51% sequence identity,⁴⁶ (b) HIV-2 protease (1IDA, 1.7 Å resolution) sharing 48% sequence identity,⁴⁷ (c) Rous Sarcoma Virus Protease (1BAI, 2.4 Å resolution) sharing 43% sequence identity,⁴⁸ (d) equine infectious anemia virus protease (1FMB, 1.8 Å resolution) sharing 32% sequence identity,⁴⁹ and (e) Feline Immunodeficiency Virus (3FIV, 1.85 Å resolution) sharing 32% sequence identity with HIV-I protease.⁵⁰

The homodimeric assembly of the template 1FMB was generated by applying crystallographic symmetry. In this first approach, the best possible models were generated using a pairwise structural alignment between the target and the template structure of the reference protein 1HVI as starting alignments. The sequence alignments used in the modeling are provided in Figure 4. Homology models were built using the comparative modeling server SWISS-MODEL⁴² using the project submission mode.⁵¹ All models were validated using Procheck⁵² to assess the stereochemical quality of the models and Anolea mean force potential assessment⁵³ to evaluate the environment of heavy atoms. For the model based on 1SIP, 1FMB, and 1IDA, no residues with bad stereochemistry were found in the proximity of the binding pocket. The model based on 1BAI had one residue (Met145) within 5 Å of the ligand in the “generously allowed” region of the Ramachandran plot. The model based on 3FIV contained two residues (Val82 and Val181) within 5 Å of the ligand in the “disallowed” region.

Accurate prediction of side-chain conformation is expected to play an important role in HM for drug-design purposes. Most comparative modeling methods, such as SWISS-MODEL or Modeller, extract information about side chain placement from template structures. Rotamer libraries are used to assist the modeling of side chain conformations when only insufficient template information is available. To assess the influence of correct rotamer placement for calculating the relative binding free energy of ligands, all side-chain conformations were

TABLE II. Experimental and Calculated Binding Free Energies of the 16 Ligands

Ligand	Own protein	1HVI	Sub-opt rotamer	1SIP (HM)	1FMB (HM)	1BAI (HM)	3FIV (HM)	1IDA (HM)	Exp
A76889	-70.4 ± 7.9	-71.8 ± 10.1	-83.8 ± 7.8	-65.4 ± 8.4	-66.2 ± 7.8	-66.8 ± 8.2	-63.3 ± 8.1	-60.4 ± 9.0	-14.2
A76928	-77.3 ± 8.1	-79.2 ± 8.0	-82.6 ± 7.0	-74.0 ± 7.7	-70.4 ± 8.3	-71.9 ± 12.6	-75.3 ± 8.6	-79.6 ± 8.0	-15.6
A77003	-76.2 ± 7.3	-76.2 ± 7.3	-84.3 ± 6.0	-65.9 ± 8.2	-75.4 ± 7.4	-76.6 ± 8.9	-76.9 ± 8.4	-76.2 ± 7.9	-15.5
A78791	-65.4 ± 7.8	-67.6 ± 7.5	-55.0 ± 8.9	-56.3 ± 8.4	-58.8 ± 8.4	-71.2 ± 7.1	-60.7 ± 8.6	-66.0 ± 7.8	-16.2
A79285	78.0 ± 6.4	-77.0 ± 7.9	-64.5 ± 6.5	-85.3 ± 6.9	-80.1 ± 6.5	-80.2 ± 6.3	-66.0 ± 7.3	-65.0 ± 9.6	-15.2
AG1343	-62.7 ± 6.0	-58.0 ± 6.4	-60.4 ± 5.5	-54.6 ± 8.0	-53.9 ± 6.5	-54.4 ± 6.2	-49.9 ± 7.4	-57.9 ± 6.9	-12.4
AHA001	-54.4 ± 7.0	-53.6 ± 6.5	-59.0 ± 6.4	-46.7 ± 10.0	-55.1 ± 6.5	-51.7 ± 6.6	-52.4 ± 6.9	-54.9 ± 6.3	-11.3
AHA006	-62.1 ± 5.6	-55.1 ± 6.3	-58.9 ± 5.5	-60.5 ± 6.0	-58.9 ± 6.7	-62.8 ± 5.5	-51.5 ± 6.6	-55.7 ± 6.8	-11.0
GR126045	-37.6 ± 7.6	-43.8 ± 6.2	-42.0 ± 10.2	-33.6 ± 7.7	-50.9 ± 7.2	-51.4 ± 5.5	-42.7 ± 7.1	-47.4 ± 6.1	-9.8
KNI272	-59.6 ± 9.6	-62.0 ± 8.6	-60.0 ± 6.4	-56.3 ± 8.2	-60.1 ± 6.8	-61.2 ± 6.6	-62.4 ± 11.2	-63.0 ± 6.9	-16.0
L735524	-77.0 ± 7.2	-50.2 ± 7.3	-67.3 ± 6.1	-64.2 ± 6.5	-60.0 ± 6.5	-64.6 ± 6.5	-50.2 ± 7.5	-71.1 ± 8.4	-13.2
L738317	-58.8 ± 6.5	-48.9 ± 7.6	-69.1 ± 5.4	-55.7 ± 7.5	-56.3 ± 6.6	-57.7 ± 6.6	-42.6 ± 9.7	-57.9 ± 9.8	-10.5
SB203238	-53.8 ± 8.5	-40.5 ± 8.7	-48.5 ± 7.2	-38.9 ± 10.2	-43.5 ± 6.4	-34.0 ± 8.0	-30.0 ± 8.1	-50.6 ± 7.5	-9.1
SB204144	-29.7 ± 11.9	-58.2 ± 10.2	-50.8 ± 9.2	-58.5 ± 8.0	-59.5 ± 9.6	-57.4 ± 8.2	-33.6 ± 9.5	-48.8 ± 11.8	-12.2
SB206343	-59.7 ± 10.5	-56.9 ± 8.0	-57.8 ± 6.1	-63.4 ± 6.9	-69.6 ± 6.6	-70.5 ± 7.4	-55.1 ± 8.8	-62.6 ± 7.1	-13.1
VX478	-55.7 ± 5.7	-50.3 ± 6.2	-54.3 ± 6.2	-49.4 ± 6.8	-53.6 ± 6.1	-52.0 ± 6.9	-52.2 ± 8.2	-53.9 ± 6.7	-13.1

Energies are in kcal/mol.

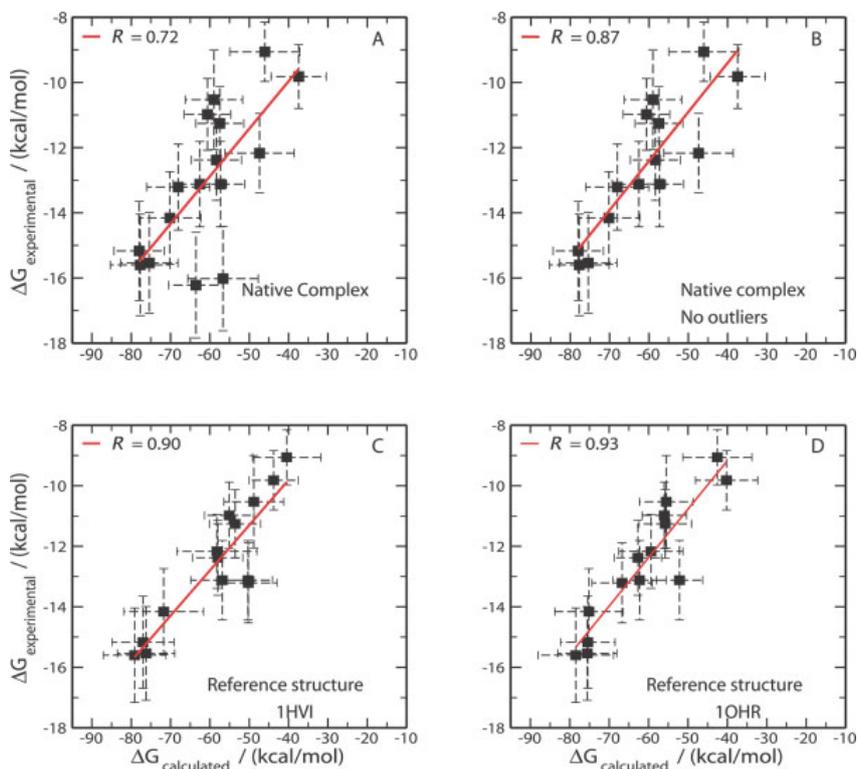


Fig. 5. Correlation between the experimentally determined binding free energy and the calculated one for the ligands in their respective structures, with (A) and without ligands A78791 and KNI272 (B), and in reference structures 1HVI (C) and 1OHR (D). Energies are in kcal/mol. Error bars are 10% for the experimental values (see text).

comparison is less favorable. The correlation without these two improves to $R = 0.87 \pm 0.02$. Figure 5(A,B) shows the correlation (with and without the ligands A78791 and KNI272) of the calculated binding free

energy ΔG_{calc} for each ligand. The uncertainty in the correlation coefficients R was calculated from a leave-one-out procedure. The two outliers are discussed in more detail in the Appendix.

TABLE IV. Calculated Binding Free Energies of the Ligands L738317, SB203238, and SB204144 for Both the Reference Structure 1HVI and the Respective Structure

Ligand	VDW	Elec.	Elec. desolv.	Nonpolar desolv.	Total
Reference protein (1HVI)					
L738317	-64.05	-50.60	74.03	-8.23	-48.85
SB203238	-59.22	-40.47	67.44	-8.23	-40.48
SB204144	-67.58	-60.06	78.10	-8.66	-58.21
Respective structures					
L738317	-67.45	-63.93	80.79	-8.17	-58.78
SB203238	-67.83	-44.52	66.82	-8.24	-53.76
SB204144	-62.78	-43.40	87.77	-8.26	-29.68

Energies are in kcal/mol.

move during the MD simulations. However, energy terms in MM-GBSA are evaluated between *all* atoms of the system. If the ligands are docked into *one* reference structure (e.g., 1HVI), the structure of the most distant part is invariant. This is not the case if ligands are docked into their proper protein structures; there, each ligand has an individual, slightly different long-range environment which influences the interaction energies and thus the ranking. This effect is illustrated in Table IV, where the binding free energy decomposition for the ligands L738317, SB203238, and SB204144 is shown. When docked into the reference protein, the ranking of the three ligands is in agreement with experiment, while docking them into their respective structures changes the ranking.

Previous simulations of three HIV-I protease inhibitor complexes used 1 ns of MD simulations with calculated and experimental binding free energies of (-30.1, -45.0, -45.0) kcal/mol versus (-14.2, -13.4, -14.6) kcal/mol for (SQV, IDV, QF34), respectively, which yields $R = -0.2$.⁵⁶ It is also of interest to compare the binding free energy previously determined using a LIE-like method and the binding free energies calculated here. The correlation between the two computational approaches is 0.89, including both A78791 and KNI272, which is reassuring. It is worthwhile noting that the same two ligands were also amongst the outliers when the more accurate LIE-like method for estimating ΔG_{calc} was used.⁷

It was also tested whether including entropic effects reduces the observed deviations. The entropic contribution to ΔG_{calc} was calculated for the ligands A78791 and KNI272 for each of the ligand-protein complexes in the reference protein 1HVI for 20 frames along the molecular dynamics trajectory. However, including the entropic term in the binding free energy does not improve the correlation, and the ligands A78791 and KNI272 are still outliers.

Application to Assess Comparative Models

Based on the good correlation between experimentally determined and calculated binding free energies for 14 of the 16 ligands in their respective complex structures and docked into the reference proteins 1HVI and 1OHR, MM-GBSA as applied here is a computationally feasible and sufficiently accurate approach to rank ligands based

on ΔG values. The molecular dynamics and binding free energy calculations for a single protein-ligand complex takes 48 h on a 2.6-GHz AMD Opteron computer using a single CPU. This makes the procedure applicable to assess the effects of typical errors observed in HM to computing binding free energies of ligand-protein complexes. For comparison, a previous study of peptide analogue binding to HIV-I protease reports a timing of 30 h for the determination of one binding constant on a cluster of Linux computers with 1.3-GHz Intel Pentium III processors using one CPU for the MD simulations and two CPUs for solving the Poisson equation.⁵⁷ These simulations used a weighted histogram analysis to estimate the free energy of binding and were carried out for 5500 time steps in implicit solvent with fixed protein atoms further than 5 Å away from any ligand atom. The correlation coefficient from this study is $R = 0.89$ for the five peptide analogues considered.

The correlation between calculated and observed binding free energies was found to be satisfactory for 14 out of the 16 ligands. Possible reasons for the observation that A78791 and KNI272 do not correlate with experimental observations are discussed in more detail in the Appendix. In the following, the performance of various homology models and perturbed backbone structures will be assessed. Since the two outliers could obscure the major point of this investigation, which is to assess the influence of possible inaccuracies in the protein structure on estimates for ligand affinities, the results are discussed for the validated set of the 14 ligands that correlate well with experiment. However, all calculations were carried out for the 16 ligands and correlation coefficients are given for both, the set of all 16 ligands and the set without the two outliers.

Suboptimal rotamer placement

A protein structure model with suboptimally placed side-chain conformations was generated by remodeling all side chains based on the backbone of the reference structure 1HVI using SQWRL.⁵⁴ Residues within 5 Å of the ligand superimpose to 1.26 Å (backbone and side-chains) compared with the structure of 1HVI. The ligands were docked in this suboptimal rotamer model and the

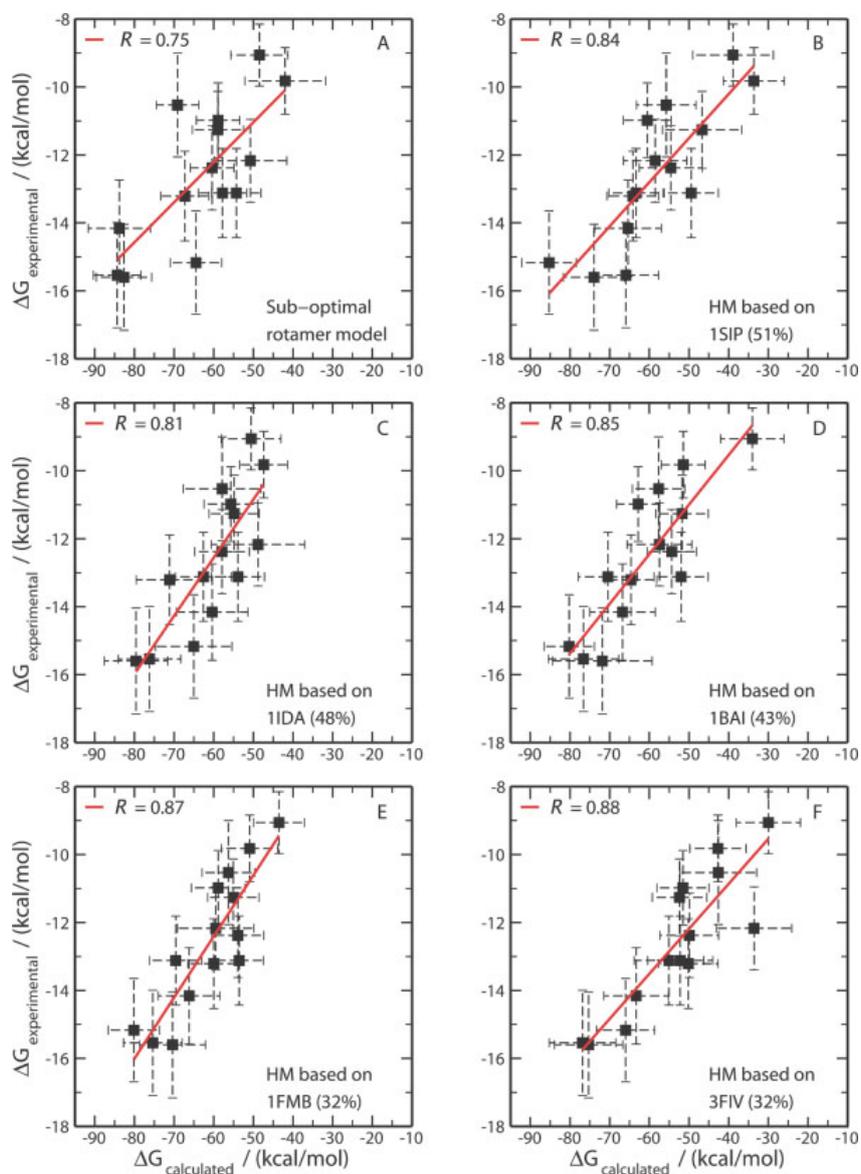


Fig. 7. Correlation between observed and calculated ligand-binding energies for the suboptimal protein structures; (A) the rotamer structure, (B) homology model based on 1SIP (51% sequence identity), (C) 1IDA (48% sequence identity), (D) 1BAI (43% sequence identity), (E) 1FMB (32% sequence identity), (F) 3FIV (32% sequence identity). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

relative ΔG calculated as described before. Figure 7(A) shows the correlation between the experimentally determined binding free energies and the calculated ones. A correlation of 0.75 ± 0.04 (0.54 ± 0.05 for all 16 ligands) is obtained, which compares with a correlation of 0.90 (0.85) for the reference protein. Thus, the suboptimal rotamer model destroys the correlation for some of the ligands, in particular for ligands A79285, SB203238, and L738317. Table V analyzes ΔG for these three ligands, in

comparison with the binding free energies in the reference structure of 1HVI. For A79285, ΔG decreases by ≈ 12.5 kcal/mol with electrostatics and desolvation energies contributing most to the difference, while for L738317 a stabilization of about 20 kcal/mol is found, again mostly from the electrostatics and desolvation energy terms. However, for SB203238 the favorable contribution of 8 kcal/mol originates mostly from the van der Waals interactions.

TABLE V. Calculated Binding Free Energies of the Ligands A79285, L738317, and SB203238 for Both the Reference Structure 1HVI and the Suboptimal Rotamer Model

Ligand	VDW	Elec.	Elec. desolv.	Nonpolar desolv.	Total
Reference protein (1HVI)					
A79285	-80.01	-81.85	93.97	-9.15	-77.02
L738317	-64.05	-50.60	74.03	-8.23	-48.85
SB203238	-59.22	-40.47	67.44	-8.23	-40.48
Suboptimal rotamer model					
A79285	-73.67	-58.61	76.81	-9.01	-64.49
L738317	-67.55	-62.81	69.16	-7.93	-69.13
SB203238	-66.03	-39.72	65.43	-8.15	-48.47

Energies are in kcal/mol.

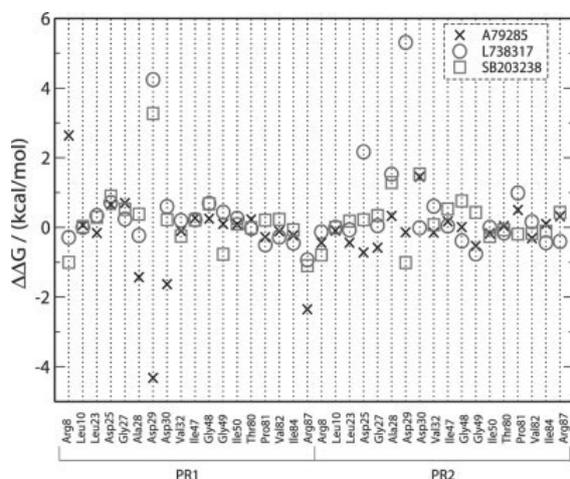


Fig. 8. The difference in total energy contribution ($\Delta\Delta G$) between the reference structure 1HVI and the suboptimal rotamer model for residues within 5 Å of ligands A79285, L738317, and SB203238. The x-axis represents the residues of each of the monomer that fall within 5 Å of the ligands and y-axis is the difference in total energy contribution in kcal/mol.

It is instructive to consider the binding free energy decomposition for residues within 5 Å of the active site. This gives an impression of the origin for the free energy differences between the calculations in the reference structure and the rotameric model. Figure 8 shows the difference $\Delta\Delta G$ in total energy contributions between the three ligands in the reference structure of 1HVI and the suboptimal rotamer placement. Considerable differences (larger than 2 kcal/mol) are found for residues Asp29, Arg87, Asp25', and Asp29'. For Asp25' and Asp29', the same rotamer conformation as in the reference structure is generated with SQWRL. The RMSD difference between the side-chains generated with SQWRL and those of the reference protein is only 0.45 Å and 0.36 Å respectively. It is only for the ligand L738317 that a considerable difference is observed between the rotamer model and the reference protein, in both cases resulting in better electrostatic interactions between the residues and the ligand. However, for Asp29 and Arg87, a different rotamer conformation is generated. The RMSD difference between the side-chains generated with SQWRL and those of the reference protein is

2.3 Å and 2.0 Å, which results in considerable difference in energy contribution of these two residues for all three ligands A79285, SB203238, and L738317.

Models Based on Suboptimal Templates

In general, the accuracy of homology models is observed to decrease with decreasing sequence similarity between target and template.⁵⁸ In the following, the influence of inaccuracies of HM based on templates sharing different degrees of sequence conservation is assessed. For this, protein structures from the retroviral protease subfamily, which contains eight members (including HIV-1), were selected as templates to build HMs of HIV-I protease.

The structures selected were (a) Simian Immunodeficiency Virus protease (1SIP, 2.3 Å resolution) sharing 51% sequence identity,⁴⁶ (b) HIV-2 protease (1IDA, 1.7 Å resolution) sharing 48% sequence identity,⁴⁷ (c) Rous Sarcoma Virus Protease (1BAI, 2.4 Å resolution) sharing 43% sequence identity,⁴⁸ (d) equine infectious anemia virus protease (1FMB, 1.8 Å resolution) sharing 32% sequence identity,⁴⁹ and (e) Feline Immunodeficiency Virus (3FIV, 1.85 Å resolution) sharing 32% sequence identity with HIV-I protease.⁵⁰

Models were built using SWISS-MODEL as described earlier. The overall RMSD of the backbone atoms between the models and the reference protein 1HVI is 1.02 Å for the model based on 1SIP (51% ID), 0.83 Å for the model based on 1IDA (48% ID), 1.18 Å for the model based on 1BAI (43% ID), 1.14 Å for the model based on 1FMB (32% ID), and 1.28 Å for the model based on 3FIV (32% ID). If only residues within 5 Å of the ligand are considered, the corresponding RMSD values are 1.53, 0.69, 1.63, 1.30, and 0.94 Å, respectively. No suitable dimeric X-ray structures that could serve as a template are available for the Myeloblastosis-associated viral protease and the Mason-Pfizer monkey virus.

Figure 7(B–F) show the correlation between the experimentally determined binding free energies and the calculated values for the five homology models. For the model based on Simian Immunodeficiency Virus protease (1SIP) the correlation is 0.84 ± 0.02 (0.68 ± 0.04), for the model based on HIV-2 protease (1IDA) the correlation is 0.81 ± 0.02 (0.76 ± 0.02), for the model based on Rous Sarcoma Virus Protease (1BAI) it is 0.85 ± 0.02

(0.79 ± 0.03), for the model based on equine infectious anemia virus protease (1FMB) the correlation is 0.87 ± 0.02 (0.81 ± 0.04), and using Feline Immunodeficiency Virus (3FIV) as a template a correlation of 0.88 ± 0.02 (0.85 ± 0.02) is obtained. It is interesting to note that even with a HM based on a template sharing only 32% sequence identity, a satisfactory correlation is obtained and that there is virtually no difference between the correlation obtained with a model based on a template with 51% sequence identity and one based on a template with 32% sequence identity. This observation might be specific for the aspartic proteases, and further protein families will have to be studied to draw more general conclusions.

CONCLUSIONS

In this work, a method for estimating the relative binding free energy between ligand and protein of the HIV-I protease system was validated and applied to estimate ΔG between systematically perturbed protein structures and a set of ligands. The method is based on the MM-GBSA approach with docking of different ligands into the same protein structure. Docking into the same protein X-ray structure gave correlation coefficients between 0.78 and 0.85 including all 16 ligands and 0.90 to 0.93 omitting the two outliers (A78791 and KNI272) already found in a previous study.⁷ To put the present results into perspective, it is instructive to compare these correlation coefficients with correlations from previous work. In a recent study of five peptidic inhibitors, the correlation coefficient was $R = 0.89$.⁵⁷ A study which investigated the influence of water molecules to the ranking of ligands found correlation coefficients between 0.30 and 0.61 for calculations without and with water, respectively.⁵⁹ Finally, a study of seven different ligands gave a correlation coefficient of 0.89.⁶⁰ Thus, the present results are well within previous attempts to describe ligand-binding in HIV-I.

Previously, MM-GBSA using the GB-MV2 model has been validated compared with results from solving the Poisson equation in estimating the total solvation energy changes in insulin dimerization.³⁵ Since MM-GBSA is a relatively rapid method (scoring of one structure takes 20 s), it can be applied to assess the influence of variations in protein structures on the computation of ligand affinities. The analysis of protein structures with different cocrystallized ligands and homology models of different accuracy shows that MM-GBSA is a useful tool to assess the influence of modeling inaccuracies in computational drug design. Previously, it has been postulated that HM down to 50% sequence identity between target and template can be used in drug design.^{61,62} However, in the case of the HIV-I protease system studied in this work, it was shown that even models based on much lower sequence identity can provide quite reliable results (correlation of 0.85), while other factors (such as the computational methods used for modeling side chains) can have a much more detrimental effect. The

rapid and accurate method to calculate relative binding free energies between ligands presented in this work will allow analyzing the effects of inaccuracies of homology models by systematically introducing errors which are typically observed during homology modeling. These results will provide the basis to specifically improve the accuracy of homology modeling methods targeted for applications in structure-based drug discovery.

In the present work, MM-GBSA was validated for application to estimating ligand binding energies in protein structures and structures derived from homology models, the target system HIV-I was used because it represents a particularly well-tested and experimentally well-characterized system. The method has been applied to increasingly difficult docking problems. First, the method was validated for 16 ligands in their own protein structures and in two reference structures (1HVI, 1OHR). Fourteen out of the 16 ligands were found to correlate favorably compared with the experimental data. As a next level of complexity, ligand-binding energies were estimated for comparative models for HIV-I protease built based on alignments to other retroviral protease structures (SIV, HIV-2, RSV, EIAV, FIV) belonging to the same subclass. Finally, the most challenging application was to calculate ligand affinities for suboptimal side-chain conformations. It was shown that the modeling of the side chains is crucial for ligand-binding studies based on homology models for HIV-I protease.

Because this investigation showed that MM-GBSA is a useful level to rank ligands, it should be possible to apply this methodology to cases where the amount of experimental data is smaller and templates for the homology models structurally more diverse. Work on less favorable cases is under way.

REFERENCES

1. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LL. Uniprot: the universal protein knowledgebase. *Nucleic Acids Res* 2004;32:115–119.
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. *Nucleic Acids Res* 2000;28:235–242.
3. Blundell TL, Sibanda B, Sternberg MJ, Thornton JM. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* 1987;326:347–352.
4. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. *EMBO J* 1986;5:823–826.
5. Wlodawer A, Vondrasek J. Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. *Ann Rev Biophys Biomol Struct* 1998;27:249–284.
6. Pearl LH, Taylor WR. A structural model for the retroviral proteases. *Nature* 1987;329:351–354.
7. Zoete V, Michielin O, Karplus M. Protein-ligand binding free energy estimation using molecular mechanics and continuum electrostatics. Application to HIV-1 protease inhibitors. *J Comput Aided Mol Des* 2003;17:861–880.
8. Gohlke H, Kiel C, Case DA. Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes. *J Mol Biol* 2003;330:891–913.
9. Tame JRH. Scoring functions: a view from the bench. *J Comput Aided Mol Des* 1999;13:99–108.

10. Ajay, Murcko MA. Computational methods to predict binding free energy in ligand-receptor complexes. *J Med Chem* 1995;38:4953-4967.
11. Hosur MV, Bhat TN, Kempf DJ, Baldwin ET, Liu B, Gulnik S, Wideburg NE, Norbeck DW, Appelt K, Erickson JW. Influence of stereochemistry on activity and binding modes for C2 symmetry-based diol inhibitors of HIV-1 protease. *J Am Chem Soc* 1994;116:847-855.
12. Silva AM, Cachau RE, Sham HL, Erickson JW. Inhibition and catalytic mechanism of HIV-1 aspartic protease. *J Mol Biol* 1996;255:321-346.
13. Kaldor SW, Kalish VJ, Davies JF, II, Shetty BV, Fritz JE, Appelt K, Burgess JA, Campanale KM, Chirgadze NY, Clawson DK, Dressman BA, Hatch SD, Khalil DA, Kosa MB, Lubbehusen PP, Muesing MA, Patrick AK, Reich SH, Su KS, Tatlock JH. Viracept (nelfinavir mesylate, AG1343): a potent, orally bioavailable inhibitor of HIV-1 protease. *J Med Chem* 1997;40:3979-3985.
14. Backbro K, Lowgren S, Osterlund K, Atepo J, Unge T, Hulsten J, Bonham NM, Schaal W, Karlen A, Hallberg A. Unexpected binding mode of a cyclic sulfamide HIV-1 protease inhibitor. *J Med Chem* 1997;40:898-902.
15. Jhota H, Singh OM, Weir MP, Cooke R, Murray-Rust P, Wonacott A. X-ray crystallographic studies of a series of penicillin-derived asymmetric inhibitors of HIV-1 protease. *Biochemistry* 1994;33:8417-8427.
16. Baldwin E, Bhat TN, Gulnik S, Liu B, Topol IA, Kiso Y, Mimoto T, Mitsuya H, Erickson J. Structure of HIV-1 protease with KNI-272, a tight-binding transition-state analog containing allophenylnorstatine. *Structure* 1995;3:581-590.
17. Chen Z, Li Y, Chen E, Hall DL, Darke PL, Culberson C, Shafer JA, Kuo LC. Crystal structure at 1.9 Å resolution of human immunodeficiency virus (HIV) II protease complexed with L-735,524, an orally bioavailable inhibitor of the HIV proteases. *J Biol Chem* 1994;269:26344-26348.
18. Munshi S, Chen Z, Li Y, Olsen DB, Fraley ME, Hungate RW, Kuo LC. Rapid X-ray diffraction analysis of HIV-1 protease-inhibitor complexes: inhibitor exchange in single crystals of the bound enzyme. *Acta Crystallogr Sect D* 1998;54:1053-1060.
19. Hoog SS, Zhao B, Winborne E, Fisher S, Green DW, DesJarlais RL, Newlander KA, Callahan JF, Moore ML, Huffman WF, Abdel-Meguid SS. A check on rational drug design: crystal structure of a complex of human immunodeficiency virus type 1 protease with a novel γ -turn mimetic inhibitor. *J Med Chem* 1995;38:3246-3252.
20. Abdel-Meguid SS, Zhao B, Murthy KH, Winborne E, Choi JK, DesJarlais RL, Minnich MD, Culp JS, Debouck C, Tomaszek TA, Jr, Meek TD, Dreyer GB. Inhibition of human immunodeficiency virus-1 protease by a C2-symmetric phosphinate. Synthesis and crystallographic analysis. *Biochemistry* 1993;32:7972-7980.
21. Thompson SK, Murthy KH, Zhao B, Winborne E, Green DW, Fisher SM, DesJarlais RL, Tomaszek TA, Jr, Meek TD, Gleason JG, Abdel-Meguid SS. Rational design, synthesis, and crystallographic analysis of a hydroxyethylene-based HIV-1 protease inhibitor containing a heterocyclic P1'-P2' amide bond isostere. *J Med Chem* 1994;37:3100-3107.
22. Kim EE, Baker CT, Dwyer MD, Murcko MA, Rao BG, Tung RD, Navia MA. Crystal structure of HIV-1 protease in complex with VX-478, a potent and orally bioavailable inhibitor of the enzyme. *J Am Chem Soc* 1995;117:1181,1182.
23. Zoete V, Michielin O, Karplus M. Relation between sequence and structure of HIV-1 protease inhibitor complexes: a model system for the analysis of protein flexibility. *J Mol Biol* 2002;315:21-52.
24. Brooks BR, Brucoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization and dynamics calculations. *J Comput Chem* 1983;4:187-217.
25. MacKerell AD, Jr, Bashford D, Bellott M, Dunbrack RL, Jr, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau TK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, III, Roux B, Schlenker M, Smith JC, Stote R, Straub J, Watanabe M, Workiewicz-Kuczera J, Yin D, Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* 1998;102:3586-3616.
26. Harte WE, Beveridge DL. Mechanism for the destabilization of the dimer interface in a mutant HIV-1 protease - a molecular-dynamics study. *J Am Chem Soc* 1993;115:1231-1234.
27. Miller M, Geller M, Gribskov M, Kent SB. Analysis of the structure of chemically synthesized HIV-1 protease complexed with a hexapeptide inhibitor. Part 1: crystallographic refinement of 2 Å data. *Proteins* 1997;27:184-194.
28. Wang YX, Freedberg DI, Yamazaki T, Wingfield PT, Stahl SJ, Kaufman JD, Kiso Y, Torchia DA. Solution NMR evidence that the HIV-1 protease catalytic aspartyl groups have different ionization states in the complex formed with the asymmetric drug KNI-272. *Biochemistry* 1996;35:9945-9950.
29. Brooks CL, III, Brunger A, Karplus M. Active site dynamics in protein molecules: a stochastic boundary molecular-dynamics approach. *Biopolymers* 1985;24:843-865.
30. Brooks CL, III, Karplus M. Solvent effects on protein motion and protein effects on solvent motion. *J Mol Biol* 1989;208:159-181.
31. Lee MS, Salsbury FR, Jr, Brooks CL, III. Novel generalized Born methods. *J Chem Phys* 2002;116:10606-10614.
32. Lee MS, Feig M, Salsbury FR, Jr, Brooks CL, III. New analytical approximation to the standard molecular volume definition and its application to generalized born calculations. *J Comput Chem* 2003;24:1348-1356.
33. Srinivasan J, Cheatham TE, III, Cieplak P, Kollman PA, Case DA. Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices. *J Am Chem Soc* 1998;120:9401-9409.
34. Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L, Lee M, Lee T, Duan Y, Wang W, Donini O, Cieplak P, Srinivasan J, Case DA, Cheatham TE, III. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc Chem Res* 2000;33:889-897.
35. Zoete V, Meuwly M, Karplus M. Study of the insulin dimerization: binding free energy calculations and per-residue free energy decomposition. *Proteins* 2005;61:79-93.
36. Hermann RB. Theory of hydrophobic bonding. II. correlation of hydrocarbon solubility in water with solvent cavity surface area. *J Phys Chem* 1972;76:2754-2759.
37. Hasel W, Hendrikson TF, Still WC. A rapid approximation to the solvent accessible surface areas of atoms. *Tetrahedron Comput Methodol* 1988;1:103-116.
38. Still WC, Tempczyk A, Hawley RC, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J Am Chem Soc* 1990;112:6127-6129.
39. McQuarrie DA. *Statistical mechanics*. New York: Harper and Row; 1976.
40. Tidor B, Karplus M. The contribution of vibrational entropy to molecular association. *J Mol Biol* 1994;238:405-414.
41. Rost B. Twilight zone of protein sequence alignments. *Protein Eng* 1999;12:85-94.
42. Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res* 2003;31:3381-3385.
43. Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* 2006;22:195-201.
44. Johnson M, Srinivasan N, Sowdhamini R, Blundell T. Knowledge-based protein modeling. *Crit Rev Biochem Mol Biol* 1994;29:1-68.
45. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 1993;234:779-815.
46. Wilderspin A, Sugrue RJ. Alternative native flap conformation revealed by 2.3 Å resolution structure of SIV proteinase. *J Mol Biol* 1993;239:97-103.
47. Tong L, Pav S, Mui S, Lamarre D, Yoakim C, Beaulieu P, Anderson PC. Crystal structures of HIV-2 protease in complex with inhibitors containing the hydroxyethylamine dipeptide isostere. *Structure* 1995;3:33-40.
48. Wu J, Adomat JM, Ridky TW, Louis JM, Leis J, Harrison RW, Weber IT. Structural basis for specificity of retroviral proteases. *Biochemistry* 1998;37:4518-4526.
49. Gustchina A, Kervinen J, Powell DJ, Zdanov A, Kay J, Wlodawer A. Structure of equine infectious anemia virus proteinase complexed with an inhibitor. *Protein Sci* 1996;5:1453-1465.
50. Laco GS, Schalk-Hihi C, Lubkowski J, Morriss G, Zdanov A, Olson A, Elder JH, Wlodawer A, Gustchina A. Crystal struc-

- tures of the inactive D30N mutant of feline immunodeficiency virus protease complexed with a substrate and an inhibitor. *Biochemistry* 1997;36:10696–10707.
51. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-Pdb Viewer: an environment for comparative protein modelling. *Electrophoresis* 1997;18:2714–2723.
 52. Laskowski RA, MacArthur MW, Moss D, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr* 1998;26:283–291.
 53. Melo F, Feytmans E. Assessing protein structures with a non-local atomic interaction energy. *J Mol Biol* 1998;277:1141–1152.
 54. Canutescu AA, Shelenkov AA, Dunbrack RL, Jr. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci* 2003;12:2001–2014.
 55. Bursulaya BD, Totrov M, Abagyan R, Brooks CL, III. Comparative study of several algorithms for flexible ligand docking. *J Comput Aided Mol Des* 2003;17:755–763.
 56. Lepsik M, Kriz Z, Havlas Z. Efficiency of a second-generation HIV-1 protease inhibitor studied by molecular dynamics and absolute binding free energy calculations. *Proteins* 2004;57:279–293.
 57. Bartels C, Widmer A, Ehrhardt C. Absolute free energies of binding of peptide analogs to the HIV-1 protease from molecular dynamics simulations. *J Comput Chem* 2005;26:1294–1305.
 58. Cozzetto D, Tramontano A. Relationship between multiple sequence alignments and quality of protein comparative models. *Proteins* 2005;58:151–157.
 59. Fornabai M, Spyraakis F, Mozzarelli A, Cozzini P, Abraham DJ, Kellogg GE. Simple, intuitive calculations of free energy of binding for protein-ligand complexes. 3. The free energy contribution of structural water molecules in HIV-1 protease complexes. *J Med Chem* 2004;47:4507–4516.
 60. Lee CY, Yang PK, Tzou WS, Hwang MJ. Estimates of relative binding free energies for HIV protease inhibitors using different levels of approximations. *Protein Eng* 1998;6:429–437.
 61. Hillisch A, Pineda LF, Hilgenfeld R. Utility of homology models in the drug discovery process. *Drug Discov Today* 2004;9:659–669.
 62. Kopp J, Schwede T. Automated protein structure homology modeling: a progress report. *Pharmacogenomics* 2004;4:405–416.
 63. Verkhivker G, Appelt K, Freer ST, Villafranca JE. Empirical free energy calculations of ligand-protein crystallographic complexes. I. Knowledge-based ligand-protein interaction potentials apply to the prediction of human immunodeficiency virus 1 protease binding affinity. *Protein Eng* 1995;8:677–691.
 64. Bardi JS, Luque I, Freire E. Structure-based thermodynamic analysis of HIV-1 protease inhibitors. *Biochemistry* 1997;36:6588–6596.
 65. Tawa GJ, Topol IA, Burt SK, Erickson JW. Calculation of relative binding free energies of peptidic inhibitors to HIV-1 protease and its I84V mutant. *J Am Chem Soc* 1998;120:8856–8863.
 66. Kageyama S, Mimoto T, Murakawa Y, Nomizu M, Ford H Jr., Shirasaka T, Gulnik S, Erickson J, Takada K, Hayashi H, et al. In vitro anti-human immunodeficiency virus (HIV) activities of transition state mimetic HIV protease inhibitors containing allophenylnorstatine. *Antimicrob Agents Chemother* 1993;37:810–817.

APPENDIX

Analysis of the Ligand A78791

For A78791 a more detailed analysis can be carried out, since binding free energies for the chemically similar ligands A76889, A76928, A77003, and A78791 are available, which were determined by the same group using identical protocols.¹¹ All four ligands are C2 symmetry-based diol inhibitors, and A78791 is a deshydroxy analogue of A77003. The main structural differences concern the central portion of the ligands and primarily affect the interactions with the active site pocket formed by Asp25/25' and Gly27/27'¹¹; (see Fig. 1). Experimentally, A78791 was found to be the best binder (−16.2 kcal/mol,

Table II). Previous computational ligand binding studies of all or some of the four ligands were based on empirical free energy calculations,⁶³ various approximations to the binding free energy calculated from static structures and MD simulations in implicit solvent,⁶⁰ thermodynamic computations of binding free energies,⁶⁴ scoring of a series of protein–ligand complexes,⁵⁹ and a combination of molecular mechanics, dielectric continuum solvation, and surface area based methods.⁶⁵ The first two studies^{60,63} agree in that A78791 is the best inhibitor (although this is not true for all approximate methods in Ref. 60). Although the first study finds the other three ligands to bind almost equally well (which is not in agreement with experiment), the second study finds some differences between them. The structure-based thermodynamic analysis finds A77003 to bind better than A78791 and A76928.⁶⁴ Unfortunately, Kellogg et al. do not report individual binding free energies, while Erickson and coworkers report data only on A77003.⁶⁵

In the present calculations A76889, A76928, and A77003 correlate well with the experimental values, which are $\Delta G = -14.2$, -15.6 , and -15.5 kcal/mol. The binding mode of all four ligands is determined by how well the ligand can optimize hydrogen bonding with active site carboxylate groups and the van der Waals contacts with the neighboring backbone atoms. In the X-ray structures, the conformations of A77003 and A78791 are virtually identical. This suggests that any difference in the binding free energy is not due to differences in protein–inhibitor interactions.¹¹ According to Hosur et al., the presence of the second hydroxyl group on the ligand A77003 leads to larger desolvation energies and a greater entropy loss upon binding relative to the A78791 analogue. These effects, together with the limited hydrogen-bonding compensation provided by the interaction between Asp25' and the second hydroxyl group, are most likely responsible for the decreased potency of A77003.¹¹

Some of these proposals can be addressed by MD simulations. During the MD simulations in the reference protein 1HVI, the four protein–ligand structures (A76889, A76928, A77003, and A78791) superimpose on average to within 0.15 Å. To determine the possible reasons why A78791 binds least strongly, ΔG is decomposed into the electrostatic, van der Waals, nonpolar, and desolvation energy terms between the ligand and each residue of the protein. Table AI shows each of these energy terms for the residues Asp25/25' and Gly27/27' that previously have been determined to be the most important for binding.¹¹ Overall and in agreement with the above proposition, Asp25' has a positive (destabilizing) contribution to the binding free energy of A77003. However, this contribution is even more destabilizing for A78791. It was suggested that for A77003 the second hydroxyl group (which is not present for A78791) leads to a larger desolvation energy, which is not compensated for by the Asp25' interaction to the hydroxyl group. Although ΔG_{desolv} is larger for A77003 than for A78791 by 10 kcal/mol (Asp25') and 5 kcal/mol (for the ligand), respectively, in agreement with the proposition, the electrostatic

5.1 How Inaccuracies in Protein Structure Models Affect Estimates of Protein-Ligand Interactions

TABLE AI. Per Residue Decomposition of the Interaction Energy Between the Ligands A76889, A76928, A77003, and A78791 and Residues Asp25/25' and Gly27/27'

	VDW	Elec.	Elec. desolv.	Nonpolar desolv.	Total
A76889					
Asp25	-0.37	0.11	-0.28	-0.04	-0.58
Gly27	-0.87	0.46	-1.50	-0.01	-1.92
Asp25'	2.60	-17.95	15.45	-0.06	0.05
Gly27'	-0.90	0.66	-0.07	-0.06	-0.37
Ligand	-35.38	-40.09	47.24	-6.73	-34.96
A76928					
Asp25	-0.45	-0.94	0.66	-0.06	-0.79
Gly27	-1.02	0.49	-0.78	-0.02	-1.33
Asp25'	2.17	-18.07	17.85	-0.08	1.88
Gly27'	-1.29	0.33	-0.12	-0.09	-1.18
Ligand	-38.93	-39.03	44.56	-6.89	-40.30
A77003					
Asp25	-0.48	-0.47	0.57	-0.06	-0.44
Gly27	-1.42	-0.65	0.86	-0.14	-1.35
Asp25'	1.74	-20.27	22.53	-0.10	3.90
Gly27'	-0.70	0.13	-1.25	-0.02	-1.85
Ligand	-37.14	-43.68	45.96	-6.72	-41.58
A78791					
Asp25	-0.83	0.70	-1.18	-0.04	-1.35
Gly27	-0.80	0.23	-0.76	-0.04	-1.38
Asp25'	0.48	-7.40	12.34	-0.05	5.38
Gly27'	-1.13	0.12	0.00	-0.07	-1.08
Ligand	-39.18	-28.83	40.43	-6.70	-34.27

Energies are in kcal/mol.

TABLE AII. Per Residue Energy Decomposition for the Residues Asp25/25' and Gly27/27' for the Four Different Protonation States of Asp25/25' Indicated by the /res/OD Code Which Gives the Residue and Which One of the Oxygen Atoms Is Protonated

	VDW	Elec.	Elec. desolv.	Nonpolar desolv.	Total
A78791/25/OD1					
Asp25	-0.52	-1.13	0.74	-0.05	-0.95
Gly27	-0.77	0.70	-0.57	-0.03	-0.67
Asp25'	0.50	-7.00	12.21	-0.06	4.84
Gly27'	-1.46	0.15	0.30	-0.11	-1.11
Ligand	-38.99	-29.95	39.67	-6.63	-35.91
A78791/25/OD2					
Asp25	-0.77	0.95	-1.60	-0.06	-1.48
Gly27	-0.71	0.49	-1.07	-0.02	-1.31
Asp25'	0.64	-7.76	14.55	-0.08	7.35
Gly27'	-1.24	0.24	-0.39	-0.08	-1.47
Ligand	-37.69	-27.89	40.40	-6.68	-31.87
A78791/25'/OD1					
Asp25	0.87	-9.49	13.92	-0.05	5.25
Gly27	-0.69	1.02	-0.68	-0.02	-0.37
Asp25'	-0.96	0.65	-0.25	-0.06	-0.62
Gly27'	-1.21	0.09	-0.20	-0.09	-1.41
Ligand	-39.04	-29.48	40.65	-6.68	-34.55
A78791/25'/OD2					
Asp25	0.05	-8.40	13.20	-0.05	4.81
Gly27	-0.82	1.12	-0.22	-0.03	0.05
Asp25'	-1.02	2.46	-1.19	-0.06	0.20
Gly27'	-0.84	-0.57	-0.92	-0.02	-2.35
Ligand	-40.13	-28.51	41.36	-6.76	-34.05

Energies are in kcal/mol.

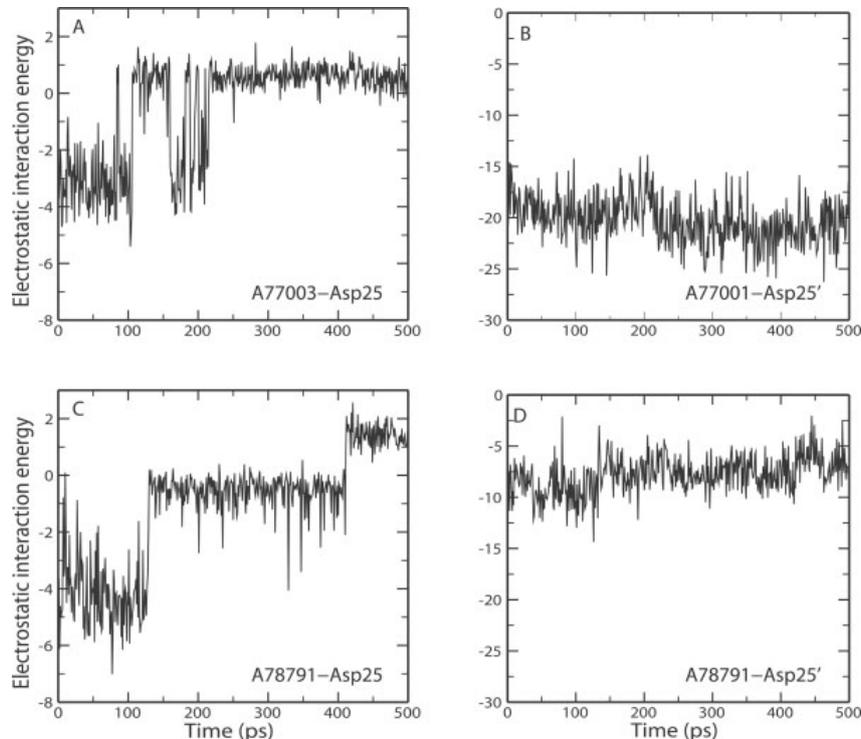


Fig. A1. Electrostatic interaction energy between the residues Asp25 and Asp25' of the reference protein 1HVI and the ligands A78791 (A,B) and A77003 (C,D), calculated for 500 frames (500 ps) along the trajectory. Energies are in kcal/mol.

422

H.B. THORSTEINSDOTTIR ET AL.

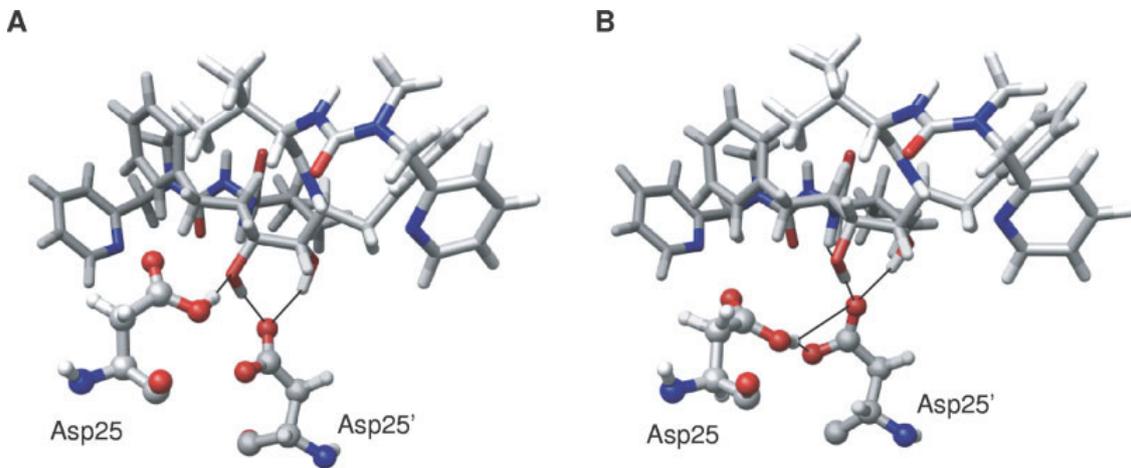


Fig. A2. Hydrogen bonding network for the two configurations found for A77033 and Asp25/25' during the MD simulation.

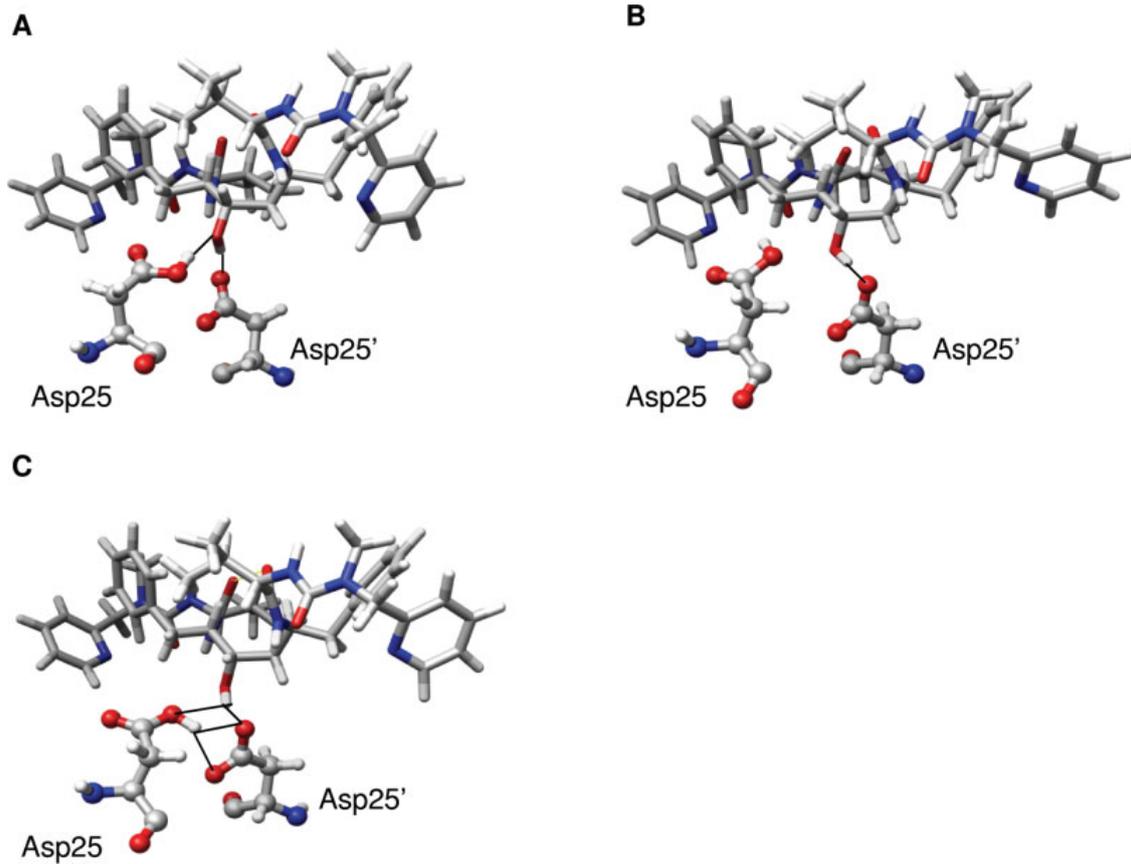


Fig. A3. Hydrogen bonding network for the three configurations found for A78791 and Asp25/25' during the MD simulation.

interaction between A78791 and Asp25' does *not* compensate for this. One possibility is that the protonation state of Asp25/Asp25' depends on the ligand. To exclude this possibility, ΔG for A78791 has been calculated from 500 ps MD simulations for each of the four possible protonation states. Table AII shows the results for each of the energy terms for the four possible protonation states for residues Asp25/25' and Gly27/27'. The initially chosen protonation state of OD1 of Asp25 gives the best overall binding free energies, supporting the choice of this protonation state. Finally, it is possible to follow the electrostatic interaction, E_{elec} , between the two ligands and residues Asp25/Asp25' and Gly27/Gly27'.

Figure A1(A) shows E_{elec} between Asp25 of the protein and the ligand A77003 for 500 frames along the trajectory. E_{elec} increases from around -3.5 kcal/mol to ≈ 0.5 kcal/mol for most of the trajectory. This is related to a structural change which removes the hydrogen bond between Asp25 and the ligand [Fig. A2(B)]. On the other hand, Figure A1(B) shows that E_{elec} between Asp25' and A77003 is slowly decreasing (with an average around -20 kcal/mol). Towards the end of the trajectory, there are three hydrogen bonds between Asp25' and the ligand, one more than at the beginning. Figure A2 shows snapshots of the structures for the two Asp residues and the ligand taken from frames of the trajectory that represent each of the energy states visited in Figure A1(A,B).

A similar analysis for A78791 shows that during the MD simulations, the hydrogen bond between the ligand and Asp25 is lost [see Fig. A1(C)], while Asp25' has only one hydrogen bond with the ligand throughout the simulation [Fig. A1(D)]. Figure A3 shows snapshots of the active site taken from frames of the trajectory that represent each of the three energy states visited in Figure A1(C,D). In addition, hydrogen bonding between Asp25 and Asp25' is observed. This explains, in part, why A78791 is a less

favorable binder than A77003. Hydrogen bonding between ligand and protein is in competition with H-bond formation between protein residues. For A78791, Asp25' is not able to form as many and strong hydrogen bonds to the ligand as A77003, and the desolvation energy does not compensate for that. Systematic errors in the simulations are unlikely, since the same two ligands (A78791 and KNI272 for which no detailed analysis was made) are outliers based on independent simulations and on different evaluation schemes for the binding free energies (LIE-like method and MM-GBSA).

In their MD simulations, Hwang et al. kept the protein frozen which will not lead to the possible complications observed here. The only other simulation where MD simulations were analyzed found A78791 to favor A77003 by 0.23 kcal/mol compared with 0.68 kcal/mol from experiment.⁷ Finally, it is worthwhile to note that inhibition constants were measured at pH 4.7, which is different from the computations. In summary, conformational changes during the MD simulations could contribute to the finding that A78791 is not the best ligand as observed in the experiment. The same applies to ligand KNI272 for which a similar, but less detailed analysis showed that after 750 ps of MD simulation (see Fig. 6), the H-bonding pattern between the ligand-hydroxyl group and the aspartic acid residues surrounding the active site changed. This leads to a decreased electrostatic interaction between the ligand and the protein, which lowers the calculated ΔG and gives a lower affinity than that found in the experiment. The results on KNI272 were determined in an independent study, using the same protocol as for A76889, A76928, A77003, and A78791.⁶⁶ In conclusion, the more detailed simulations for the two outliers suggest that the computational approach chosen is likely not responsible for the disagreement between the experiment and simulation.

5.2 MM-GBSA on a PC GRID: Setup, Validation and Applications

Thorsteinsdottir HB, Podvinec M, Meuwly M, Schwede T.
manuscript in preparation.

One of the main obstacles in the routine application of molecular dynamics simulations to virtual screening or structure-based drug design is the long calculation time that is required and therefore computational biology is often limited by the available computational resources. The work presented here aims to reduce the time that is needed for a single molecular dynamics simulation by developing a method to run multiple short molecular dynamic simulations on a PC Grid. This methodology is first validated with the protein system used previously (HIV-1 protease) and then applied to a ligand docking study. There we want to answer the question if the docking algorithms really score the lowest energy poses accurately and if molecular dynamics approaches can help to improve both the pose from the docking program and to obtain more accurate binding free energies for better ranking of the poses. This is then applied to the estrogen rdeceptor β .

MM-GBSA on a PC GRID: Setup, Validation and Applications

Holmfridur B. Thorsteinsdottir^[a,b], Michael Podvynec^[a,b], Markus Meuwly*^[c],
Torsten Schwede^[a,b]

November 25, 2008

^[a] Biozentrum, University of Basel, Klingelbergstrasse 50, 4056 Basel Switzerland

^[b] Swiss Institute of Bioinformatics, Biozentrum, University of Basel Switzerland

^[c] * To whom correspondence should be addressed;
Department of Chemistry, University of Basel, Klingelbergstrasse 80, 4056 Basel
Switzerland
E-mail: m.meuwly@unibas.ch

Keywords:
Molecular Dynamics, PC GRID, Docking, Ligand Binding

1 Introduction

Proteins play an essential role in the lifecycle of cells. Of particular interest to drug discovery are the many diseases that are a result of a faulty recognition of a ligand or a specific pathway. These proteins could be inhibited by designing an inhibitor that blocks an essential part of that pathway.

Virtual screening is commonly used to identify novel drugs that can be used to fight diseases, often very successfully [1]. However, in order to give valuable information these virtual screening algorithms have to dock potential ligands in a correct pose and accurately score a very large number of compounds. This means that for sake of efficiency that these scoring functions only address a single pose and do not take any sampling into account. In addition the scoring functions contain many approximations and therefore are limited in their accuracy [2]. This can result in non-optimal pose selection and unrealistic energies.

An alternative would be to use more accurate scoring functions to evaluate the binding energies and thereby eliminate those poses that are obviously energetically incorrect. The most precise methods are so-called free energy methods, such as free energy perturbation or thermodynamic integration [3], but they are far too time consuming to be used for a high-throughput analysis of different ligands. The MM-GBSA approach is commonly used to simulate protein-ligand interactions [5] and has previously been applied successfully to estimate the relative binding free energies of 16 ligands of the HIV-1 protease and to investigate the influence of possible inaccuracies that can arise during homology modeling [4]. Although the MM-GBSA method offers quite fast computation, the total computation time needed for a single ligand binding study is still a bottleneck. In particular, the excessive calculation time for sampling conformations prevents any large scale investigation, such as virtual screening. Traditionally, MD simulations are run as a single continuous trajectory of relatively long timescales (nanosecond scale), requiring excessive CPU time. In the approach presented here, we want to reduce the time needed for computation by splitting up the MD simulation into smaller units which then can be calculated simultaneously and independent from each other on a distributed GRID of desktop PCs.

The idea of running multiple molecular simulations for increased sampling is not new. Caves *et al* looked at the differences in sampling between running an individual trajectory of 5 ns or ten independent trajectories of 120 ps each, which only differ in the initial velocities. They found that the overall sampling was improved by using the multiple independent trajectory approach and suggest that it should be used to obtain better sampling [6]. Using the same approach Loccisano *et al* looked at the $A_1 \rightarrow A_{1,3}$ transition in MbCO. They found that by running ten 400 ps simulations they were able to observe this transition frequently, while using two 1.2 ns simulations they were only able to observe it once. The

initial structures came from five x-ray structures with random initial velocities [7]. Many other attempts have shown the improved sampling by running multiple molecular simulations, especially to observe the conformational space of small peptides and proteins [8,9]. All of these multiple molecular dynamics simulations still have rather long timescales of at least 100 ps and often even on a nanosecond scale and so far not much has been done to speed up conformational sampling for protein-ligand binding studies.

In this work we first validate the multiple short trajectory approach by applying it to the previously validated HIV-1 protease [4] at two different temperatures to see if increasing the temperature results in increased sampling. Then we apply the methodology to docked poses of the 16 ligands firstly to see if the overall pose can be improved by running a molecular dynamics simulation on it, and secondly to see if the scoring can be improved by employing a more advanced scoring function. This methodology is then applied to a different protein system, i.e. the estrogen receptor in order to see if the method is transferable to other systems.

2 Theoretical Methods

The theoretical methods used here are all described in chapter 5.

3 Results

Accurate calculation of protein-ligand binding energies is an essential part of the drug discovery process. Unfortunately, the docking algorithms and subsequent scoring functions need to make numerous approximations in order to screen a large number of compounds on a reasonable time scale. Molecular dynamics simulations could potentially improve the ligand poses and provide sufficient sampling to obtain more accurate binding energies using scoring functions such as MM-GBSA. These calculations are however still too computationally demanding to be of much use for large scale calculations. An additional limitation to molecular dynamics approaches is the need for a reliable set of parameters for the ligands under investigation, since manually parameterizing the ligands is a difficult and time consuming process.

In this work a new approach to rapidly perform molecular dynamics simulations on a PC Grid is proposed. The aim is that this short simulation approach (multiple short trajectory approach) will give comparable results to running a single continuous trajectory (standard long trajectory approach) but in a shorter time. This method is validated using the HIV-1 protease and a set of 16 ligands that previously has been validated for the MM-GBSA approach. Additionally, we val-

idate the automated parameterizing of Antechamber [12] for molecular dynamics simulations. This method is then applied to docked poses of these 16 ligands and finally to the estrogen receptor.

3.1 HIV-1 Protease

3.1.1 Validation of Short Molecular Dynamics Simulation approach

The standard approach of running molecular dynamics simulations for conformational sampling of protein-ligand interactions is to run a single continuous trajectory of a relatively long time scale. In our approach we aim to shorten the computation time needed by running independent multiple short molecular dynamics simulations which are then suitable to be calculated on a distributed GRID of PC computers. By this we hope both to shorten the overall computation time needed and secondly to obtain comparable or even improved correlation to the experimental binding free energies, compared to using the standard long molecular dynamics approach.

The multiple short trajectory approach was validated by correlation to experimental energies. Molecular simulations of both the standard long and multiple short approaches were first run at 300 K and in addition, both approaches were run at 500 K to test if increased conformational sampling could be obtained with increased temperature. For the short molecular dynamics simulation approach, the increased temperature was only used for the 100 ps of the simulation used to generate starting conformations, the actual short molecular dynamics simulation of 10 ps each were performed at 300 K. Frames were extracted every 1 ps and the binding energy over each trajectory was calculated.

Figure 1 shows the correlation coefficient between the calculated binding free energy to the experimental binding free energy calculated every 100 ps of the total simulation time of the short trajectory approach at 300 K (black line) and 500 K (red line). Figure 1 A shows this for the complete set of ligands while 1 B excludes the known outliers (A79785 and KNI272). The solid lines show the correlation coefficient calculated for every 100 ps up to 1000 ps for the multiple short dynamic approach, and the broken lines are the reference correlation coefficient calculated for a standard long molecular dynamics simulation. The results show that the multiple short trajectory approach is able to improve the correlation coefficient to experimental binding free energies at both temperatures. Another interesting observation is that for the multiple short molecular dynamics simulation at 300 K the correlation coefficient decreases from 0.86 to 0.84 until 600ps of simulated time, but for the simulation at 500 K it increases from 0.76 to 0.83 until ca 500 ps total simulation time.

Figure 2 shows the correlation graphs between the experimentally determined

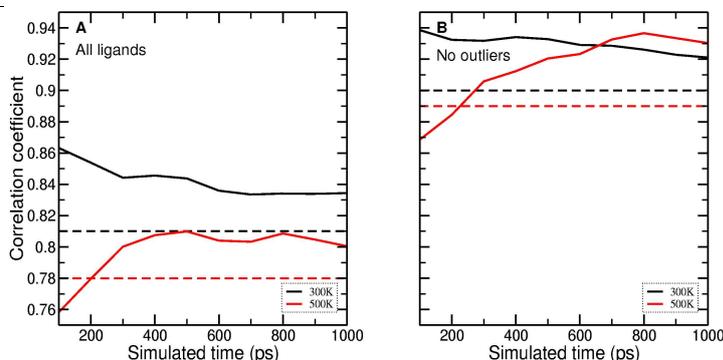


Figure 1: Correlation coefficient of calculated relative binding free energy (excluding the outliers) to the experimental binding free energies as a function of simulated time for A , simulations including all ligands and B for simulations excluding the two outliers (A79785 and KNI272) The solid lines show the correlation coefficient calculated for every 100ps up to 1000ps for the multiple short dynamic approach, and the broken lines are the reference correlation coefficient calculated for a standard long molecular dynamics simulation

binding free energy and the calculated binding free energy at 300 K for 500 ps of the standard long molecular dynamics simulation (A), 50x10ps of the multiple short molecular dynamics simulation (B) and at 500 K for 500 ps of the standard long molecular dynamics simulation (C), 50x10 ps of the multiple short molecular dynamics simulation (D) including all the ligands. At both temperatures the multiple short molecular dynamics simulation approach is able to improve the correlation coefficient between calculated binding energies and experimentally determined energies. Figure 3 shows the same excluding the two outliers and again the same improvement of the correlation coefficient is observed.

The tables for the energy contributions for the molecular dynamics simulations of each of the ligand for the standard long trajectory approach and the multiple short are in the supplementary material for both the simulations at 300 K and 500 K. A general observation is that the total binding free energies are quite similar between the standard long molecular dynamics approach and the multiple short dynamics approach at 300 K. However, there are some energy differences between the standard long molecular dynamics approach and the multiple short dynamics approach at 500K and between simulations at 300 K and 500 K. For example, the ligand A76889 has a total energy of -77.61 ± 6.75 for the standard long simulation at 300 K, and a very similar value of -76.02 ± 6.92 for the multi-

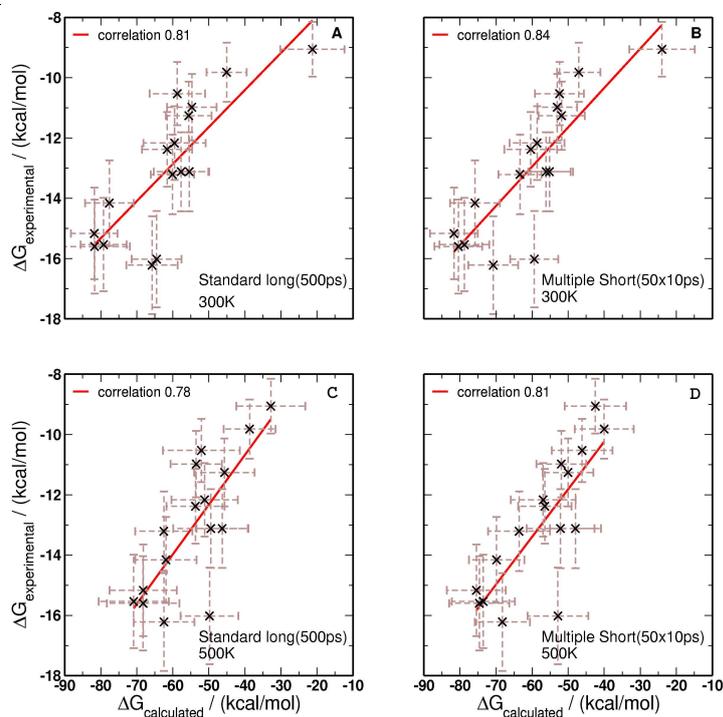


Figure 2: Correlation between the experimentally determined binding free energy and the calculated one for all the ligands for standard long molecular dynamics simulations and multiple short molecular dynamics simulations at 300 K (A, B) and 500 K (C,D)

ple short simulation at the same temperature. For the standard long simulation at 500K it gets the considerably higher energy of -61.90 ± 8.49 and -69.98 ± 7.70 for the multiple short trajectory approach at the same temperature.

These results show that although the binding energy values differ between different target temperatures, the multiple short molecular dynamics simulations approach is able to obtain good correlation to experimental values. In some cases this correlation is improved from what is obtained with the standard long trajectory approach.

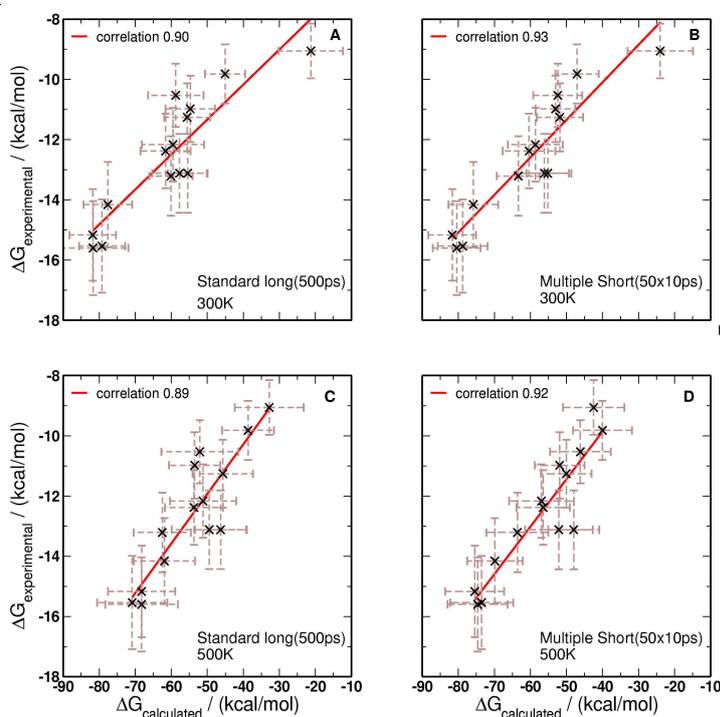


Figure 3: Correlation between the experimentally determined binding free energy and the calculated one excluding the outliers (A79785 and KNI272) for standard long molecular dynamics simulations and multiple short molecular dynamics simulations at 300 K (A, B) and 500 K (C, D)

3.1.2 Validation of automatic parameterization

To validate the approach of using Antechamber to automatically parameterize ligands for molecular dynamics with CHARMM, first the previously validated ligands of the HIV-1 protease were re-parameterized with Antechamber. Figure 4 shows the experimental binding free energy vs. the calculated relative binding free energy for the Antechamber parameterized ligands (A all 16 ligands, B without the two outliers A79785 and KNI272). The correlation coefficient for the complete dataset is 0.73 which is considerably lower than using previous manually parameterized parameters (0.83). But if the two known outliers are excluded the correlation coefficient increases to 0.85, compared to 0.9 with the original parameters. Considering that this is a much faster method with a simpler force fields,

these results are quite adequate. These results are very encouraging and promising for the application of this automated parameterization for future work.

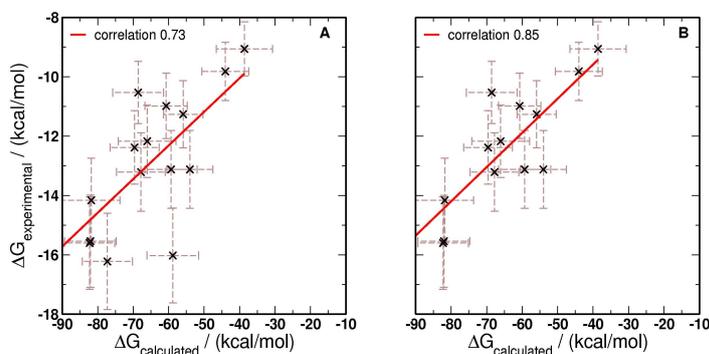


Figure 4: Correlation between the experimentally determined binding free energy and the calculated one for multiple short molecular dynamics simulations for all ligands (A) and excluding the outliers (A79785 and KNI272) (B).

3.1.3 Application to Docking

A Glide docking of the 16 ligands in the HIV-1 protease was performed using Glide's SP docking procedure. Since the HIV-1 protease ligands are quite large, the number of rotatable bonds and atoms was increased from the standard values to ensure that the ligands would be docked without any problems. Two docking runs were set up, one for ligands that require a structural water to be present in the protein, and one for the two ligands that require the absence of the structural water (aha001 and aha006). The protonation state of the catalytic Asp residues was set to be on OD1 of Asp25. The respective docking times were 162 minutes for the 14 ligands with water, and 5.4 minutes for the 2 ligands without the structural water.

Glide assigns a Emodel score for each of the docked poses and since we know the position of the ligand in the crystal structure we can calculate the rmsd of the docked pose to the reference structure. For each ligand a number of poses were selected for further study. The pose with maximum and minimum rmsd to the reference structure and the maximum and minimum Emodel score are selected and in addition at equal intervals based on the rmsd values.

The poses that were selected were then submitted to a MM-GBSA calculation on the PC-GRID. The docking results are all in the Supplementary material. It is observed, that in general after molecular dynamics and evaluation of MM-GBSA

the binding free energies are better. The trend towards obtaining good binding energies for the low rmsd poses and worse energies for the high rmsd poses is much more pronounced for the MM-GBSA approach than with only using Glide, especially for the poses with rmsd values lower than 4 Å. This is illustrated in figure 5. There are some exceptions to this such as the ligand AHA006 where the results are more or less random both in the case of MM-GBSA and Glide. The ligand AHA001 is also an interesting case since it has two distinct populations, up to a rmsd of 4 Å, then there is a population at a rmsd of 6.5 Å that obtains very good energies as seen in figure 6. This will be analyzed in more detail in the next section.

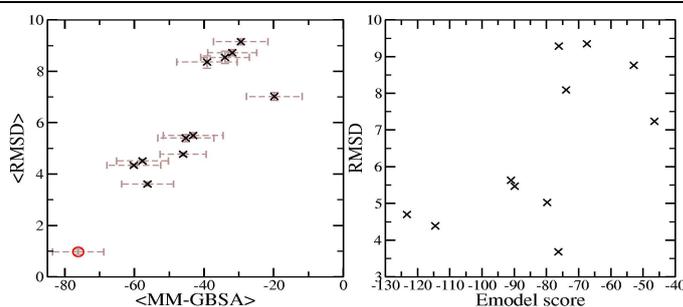


Figure 5: Docking results for the ligand A76889. The left panel shows the average rmsd vs the average binding energy for MM-GBSA calculations. The right panel shows the rmsd vs. the Emodel score for poses from Glide docking.

3.1.4 Analysis of results

For the analysis of the different trajectory approaches we first look at the difference between the energy sampling of the standard long and multiple short trajectories of ligand A76889 at 300K and then the same for these at 500K, and finally if the energy sampling between the multiple short trajectory approaches at 300K and 500K are different. Then we perform a principle component analysis to see if any conformational sampling differences can be identified in order to find important structural features. Finally, we have a look at two of the docked ligands to see how they compare to reference calculations of the crystal pose of the ligand.

Analysis based on energy contribution

In order to investigate the contribution of residues to the binding free energy between the different approaches, cluster analysis based on per residue binding en-

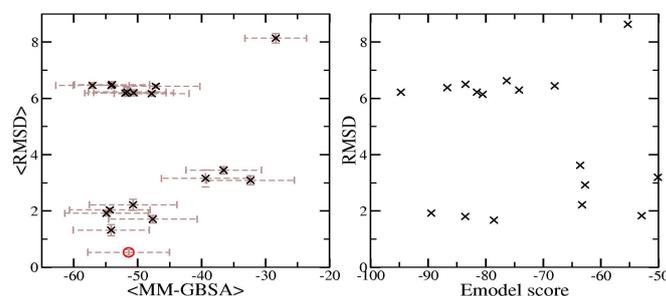


Figure 6: Docking results for the ligand AHA001. The left panel shows the average rmsd vs the average binding energy for MM-GBSA calculations. The right panel shows the rmsd vs. the Emodel score for poses from Glide docking

ergy contributions. Figure 7 shows the results from the analysis. The left panel shows the average energy contribution of the residues within 5 Å of the ligand and the right panel shows the clusters obtained based on the energy metric. The size of the cluster corresponds to the number of frames it contains, and the color depends on the contributions from each of the approaches. For the first comparison of the standard long trajectory approach and multiple short approach at 300 K the average energy contribution of the residues is very similar with only residues 124 and 129 that show some differences (figure 7 (A)). This is reflected in the clustering, where the clusters are quite mixed in the contributions from each approach. For the comparison of the standard long trajectory approach and multiple short approach at 500 K (figure 7 (B)) the average energy contribution of the residues is in general a bit lower for the multiple short trajectory approach (red circles) and again there are only a couple of residues that show a big spread in the energies, namely 124 and 128. The largest clusters are again quite mixed but a lot of the smaller clusters are specific for either of the approaches. Finally for the comparison of the multiple short trajectory approaches at 300 K and 500 K (figure 7 (C)) from the average energy contribution it can be seen that some of the residues such as 28, 124, 128 and 129 differ considerably in their energy values. The clustering in this case is also distinguishing a bit more between the two approaches, but since the overall energy differences aren't that big they are quite mixed as well.

Analysis based on PCA

Figure 8 shows the results from a principle component analysis which was used to highlight major difference in the motions of the dynamics approaches. For the comparison of the standard long and multiple short trajectory approach for 300 K figure 8(A) the plot shows that the multiple short trajectory approach only sam-

ples a fraction of the standard long trajectory approach. The structural analysis shows however that for the first principle component the conformational differences are very small. For the comparison of the standard long and multiple short trajectory approach for 500 K figure 8(B) the plot shows a larger area is sampled, and the multiple short trajectory approach again only samples a part of what the standard long trajectory approach does. The structural analysis shows that there are differences; Phe152 is very flexible in the standard long approach, and so are the residues in the flap (Gly51/150), and the residues in the binding pocket (Ala28/127 and Gly27/126) show some variation as well. Finally, we look at the differences between the multiple short trajectory approach at 300 K and at 500 K (figure 8(C)). Here, based on the plot of the first two principle components the approaches seem to be sampling two separated areas. The structural analysis confirms this finding, The side-chains of Met145 and Asp25 are oriented differently, and Gly126 shows a considerable backbone difference between the two approaches.

Analysis of two docked ligands

First, we analyse a two of docked poses for the ligand A76889. Figure 9 shows the results for the analysis. In figure 9(A) the plot of the average rmsd to the reference ligand vs the average binding energy is shown and to the right the structures of a few of the docked poses. Two of those were selected for further analysis, pose 19 which has the lowest rmsd and pose 3 which has the highest rmsd. Although pose 19 is overall structurally more similar, there are some large differences between poses, in particular pose 3 is very different from the reference. These poses were clustered in comparison to the MM-GBSA simulation of the reference structure. Figure 9(B) shows that for both of the cases there is a complete separation between the clusters. This is to be expected since the energy difference is quite large even for the pose with the lowest rmsd value (around 20 kcal/mol difference). Figure 9(C) shows the average energy contribution of each residue within 5 Å of the ligand to the binding free energy. Here it can clearly be seen that only a few residues are responsible for this large differences in the binding energy. The main contributors to this difference are Asp124 and 128 and to a lesser extent residues 25, 28, 30 and 127.

Results for the ligand aha001 are shown in figure 10. In figure 10(A) the plot of the average rmsd to the reference ligand vs the average binding energy is shown and to the right the structures the two poses that were selected for further analysis. Pose 59 shows considerable overall similarity, but pose 45 shows large structural difference, mainly in the central part of the ligand. These poses were clustered in comparison to the MM-GBSA simulation of the reference structure. In contrast to the results for ligand a86889, the clustering for ligand aha001 shown in figure 10 (B) has no clear separation of the clusters, but still a tendency for either of the approaches is observed. Figure 10(C) shows the average energy contribution of

each residue within 5 Å of the ligand to the binding free energy. Although pose 45 has a large rmsd against the crystal structure the calculated binding energy is very favorable. This is due to improved energy contributions of residues 29 and 124, while residues 25, 30, 50 128 and 129 stay mostly unchanged.

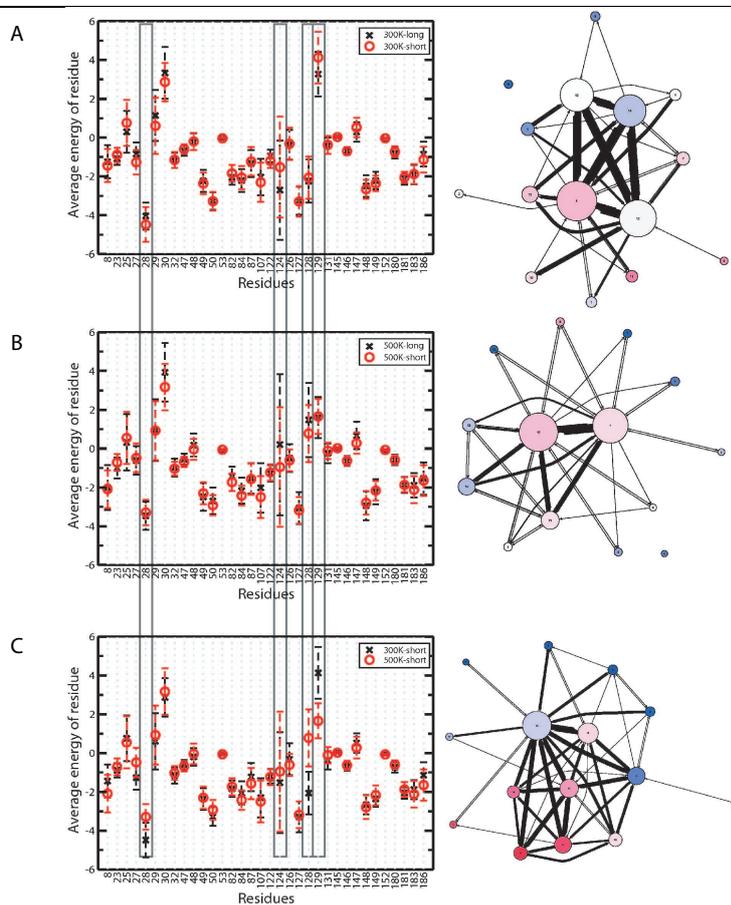


Figure 7: Results of the clustering analysis based on the per residue binding energy contribution for the standard long and multiple short trajectory approaches at 300 K and 500 K. The left panel shows the average energy contribution of the residues within 5\AA of the ligand and the right panel shows the clusters obtained by clustering based on the energy metric. The sizes of the clusters correspond to the number of frames it contains and the color to the ratio between the different approaches (red is fully populated by the long trajectory (A and B) or the short trajectory approach at 300K (C)).

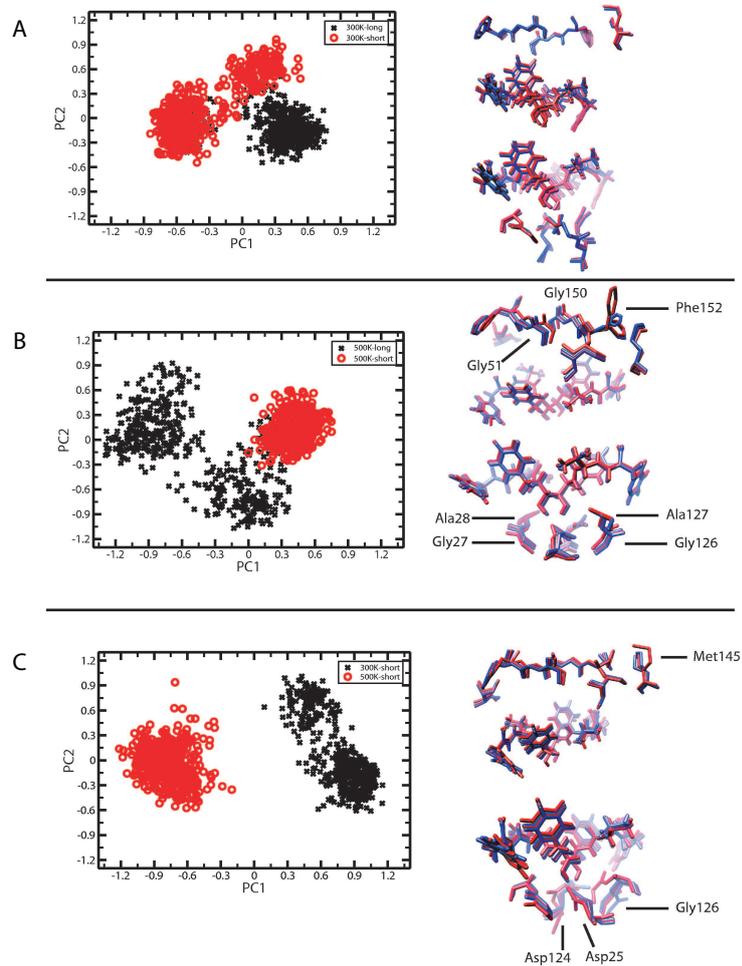


Figure 8: Principle component analysis of the standard long and multiple short trajectory approaches at 300K and 500K. The left panel shows the plots for the first two principle components and the right panel shows the extreme atomic displacement of the first principle component (red for the long trajectory approach and blue for the short trajectory approach, last panel blue for 300K and red for 500K)

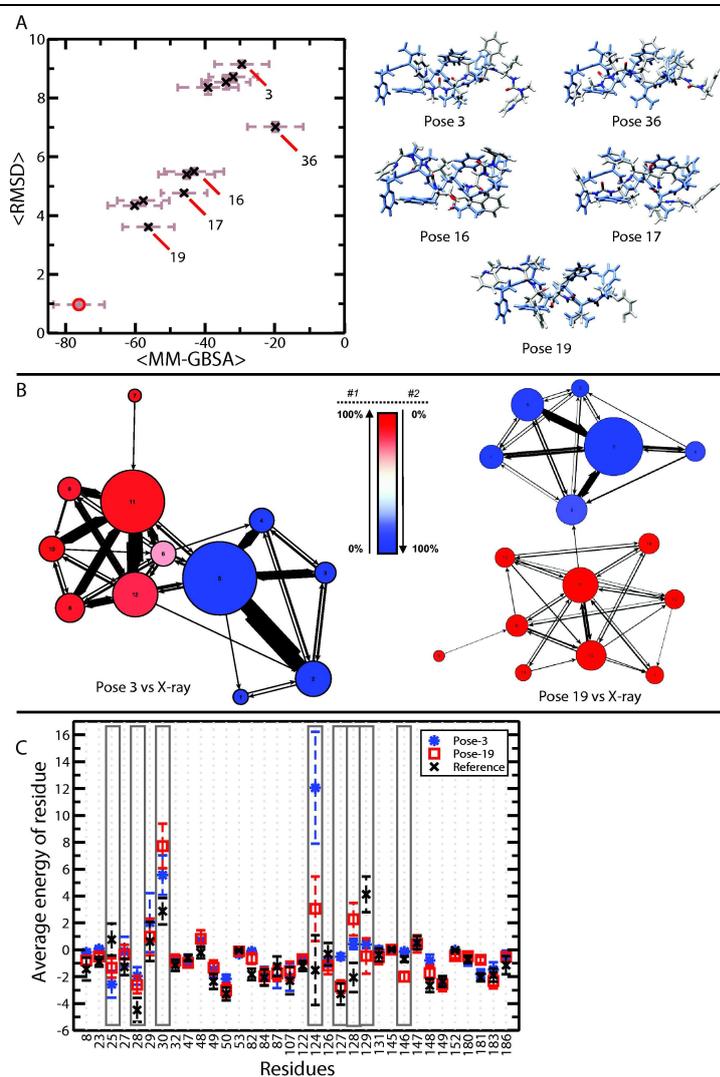


Figure 9: Docking results for the ligand A76889. (A) The left panel shows the average rmsd vs the average binding energy for MM-GBSA calculations and the right panel shows the rmsd vs the Emodel score for the poses from Glide docking. (B) Shows the results from clustering based on per residue energy contributions as compared to multiple short dynamics simulation the crystal pose of the ligand. (C) Shows the average energy of residues between the poses compared and the crystal structure pose.

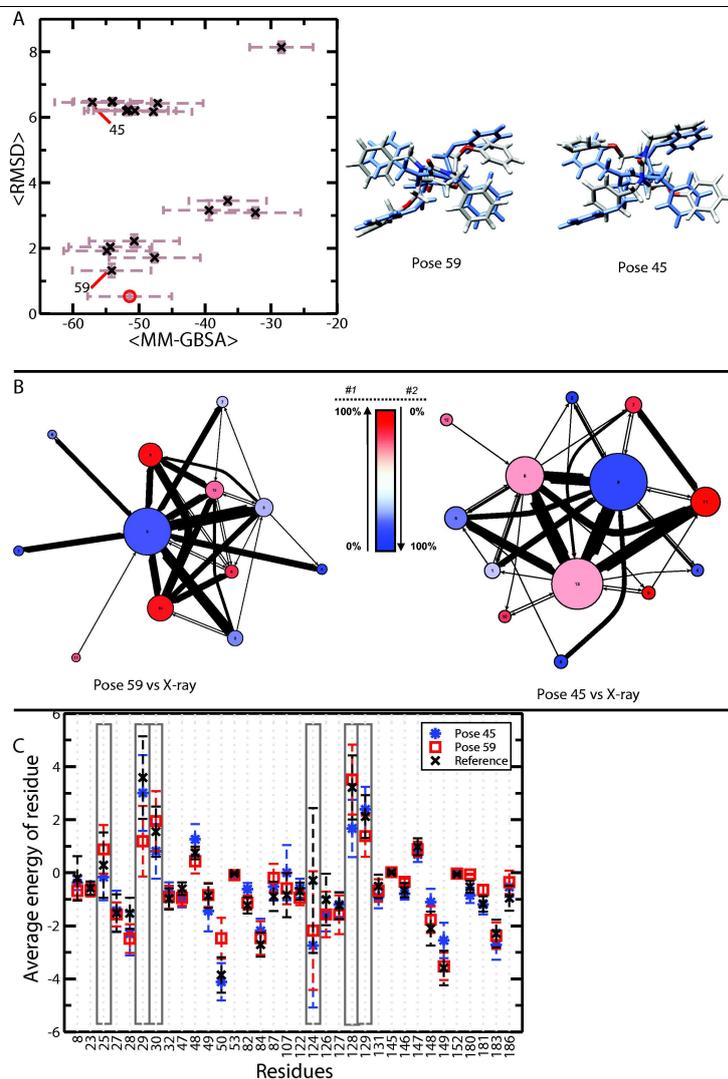


Figure 10: Docking results for the ligand AHA001. (A) The left panel shows the average rmsd vs the average binding energy for MM-GBSA calculations and the right panel shows the rmsd vs the Emodel score for the poses from Glide docking. (B) Shows the results from clustering based on per residue energy contributions as compared to multiple short dynamics simulation the crystal pose of the ligand. (C) Shows the average energy of residues between the poses compared and the crystal structure pose.

3.2 Estrogen Receptor

Motivated by the good correlation obtained for the HIV-1 ligands, the Antechamber parameterization approach was applied to 13 ligands of the estrogen receptor. Figure 11 shows the chemical structure of these 13 ligands. The protein with PDB entry 1XPC was selected as the reference structure in which all the other ligands were docked. The pdbbind database (<http://www.pdbbind.org>) is a collection of experimentally measured binding affinity data (K_d , K_i , and IC_{50}) for the protein-ligand complexes available in the Protein Data Bank (PDB) [11]. It contains binding data for the estrogen receptor as shown in table 1.

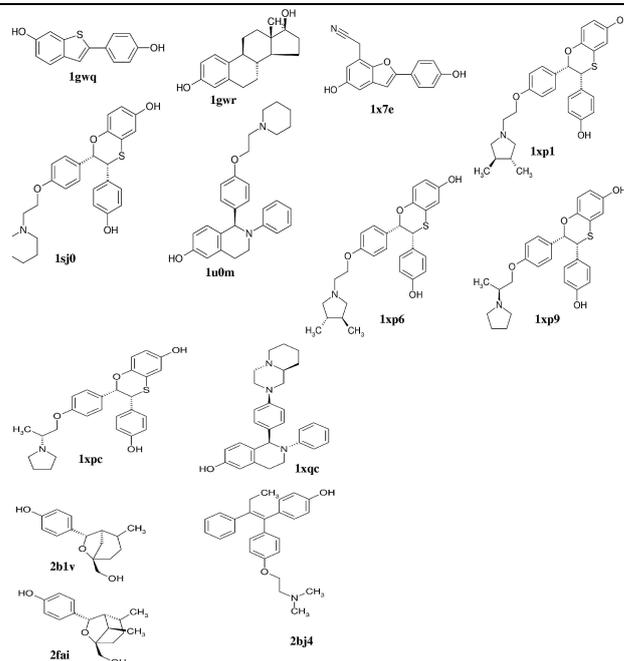


Figure 11: Schematic overview of the estrogen receptor ligands used in this study

The protonation states of the histidine residues was taken from the paper of Oostenbrink et al [30] and are shown in table 2

As previously done, the molecular dynamics simulations were performed at two different temperatures to investigate whether increased temperature results in better conformational sampling.

Table 7 and 8 in the supplementary material show the calculated energy terms

Table 1: Binding data for the estrogen receptor from the pdbname database, the affinity as given in the pdbname database, the affinity in kcal/mol, resolution of the crystal structure and the root mean square deviation to the reference structure selected (1XPC)

pdb entry	Affinity (pKd)	Affinity (kcal/mol)	Resolution (Å)	RMSD to ref
2bj4	4.47	-6.10	2	0.49
2b1v	5.74	-7.83	1.8	0.50
1x7e	5.97	-8.14	2.8	0.57
2fai	6.24	-8.51	2.1	0.42
1gwr	6.60	-9.00	2.4	0.55
1gwq	7.12	-9.71	2.45	0.54
1xqc	7.20	-9.82	2.05	0.69
luom	7.72	-10.53	2.28	0.51
1xpc	8.77	-11.96	1.6	0.00
1xp9	8.89	-12.12	1.8	0.27
1sj0	9.10	-12.41	1.9	0.29
1xpl	9.30	-12.68	1.8	0.19
1xp6	9.40	-12.82	1.7	0.17

for the multiple short simulations of the the 13 ligands of Estrogen receptor at 300 K and 500 K.

Figure 12 shows the correlation between the experimentally determined binding free energy and the calculated relative binding free energy, where the black stars are binding free energies calculated from simulations at 300 K and the red circles are the binding free energies calculated at 500 K. Figure 12 (A) shows the correlation coefficients for the standard long molecular dynamics at 300 K and 500 K. The correlation coefficient is 0.71 at 300K which improves to 0.75 for the simulation at 500 K. Figure 12 (B) shows the same improvement for high temperature simulations for the multiple short simulations. The correlation coefficient is a bit lower in this approach for the simulation at 300 K, or 0.68, but increases to 0.85 for the simulation at 500 K. There is one ligand (BJ4) that is an outlier in the simulation but for the multiple short simulations it results in much lower binding energies which gives a considerably better correlation coefficient to the experimental values (especially at 500 K). Figures 12 (C) and 12 (D) show the same graphs, but excluding this outlier. For the standard long simulation this increases slightly the correlation coefficient to 0.86 at 300 K and 0.89 at 500 K (see fig 12 (C)). For the multiple short simulations this brings the correlation coefficient up

Table 2: Protonation states of histidine residues in Estrogen Receptor. HSD means that the delta oxygen is protonated, and HSE the epsilon oxygen.

	356	373	377	398	474	476	488	501	513	516	524	547
Prot. state	HSE	HSD	HSD	HSD	HSE	HSD	HSE	HSD	HSD	HSE	HSE	HSE

to 0.85 for simulations at 300 K and to 0.90 for the simulations from 500 K (fig 12 (D)).

These results show that for the estrogen receptor a very good correlation can be obtained to experimental values using the multiple short trajectory approach. This correlation is as good as with the standard long approach. Indeed, the correlation coefficient of 0.85 is obtained for the multiple short trajectory approach at 300K and can even be increased to 0.9 at 500K.

3.2.1 Application to Docking

A Glide docking was set up for these 12 ligands of the estrogen receptor using the SP docking procedure. As before a number of poses for each ligand were selected for further study. The pose with maximum and minimum rmsd to the reference structure and the maximum and minimum Emodel score are selected and in addition at equal intervals based on the rmsd values.

The poses that were selected were then submitted to a MM-GBSA calculation on the PC-GRID. The docking results are all in the Supplementary material.

Compared to the HIV-1 protease there doesn't seem to be any clear trend towards better energies with lower rmsd for the estrogen receptor. The results are rather random regardless of the method used to evaluate the energies. Figure 13 illustrates an example of this.

In this figure the poses that have a very low rmsd ($\sim 2 \text{ \AA}$) have a very large spread in their energy values, which cannot be improved with the MM-GBSA approach.

4 Conclusions

Conclusions that can be drawn from this work are first of all that the multiple short molecular dynamics approach is successful in reproducing results using the standard long trajectory approach of a single 500 ps molecular dynamics simulation. Using the standard long approach a calculation of one ligand in the HIV-protease takes up to 4 days to complete, in contrast for a single ligand in the HIV-1 protease

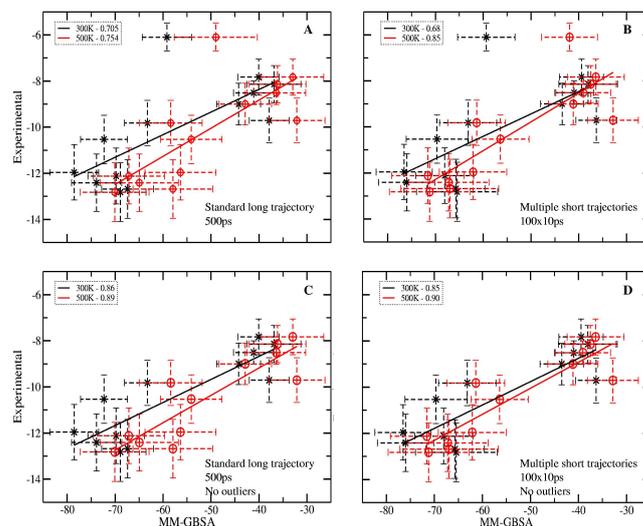


Figure 12: Correlation between the experimentally determined binding free energy and the calculated one for multiple short molecular dynamics simulations.

the calculation using the multiple short trajectories takes only 48 hours. The most time is spent on the pre-MD simulation, equilibrating the protein-ligand complex and producing the starting coordinate and velocities for subsequent calculations (24 hours). Thus, the multiple short trajectory approach allows a significant speed up of the calculation and therefore is applicable to the computationally intensive task of rescoring top docked poses.

This work demonstrates that the sampling of the energies of the HIV-1 protease is equal for both approaches. It is therefore hard to distinguish some preferred energy values or ranges that are more specific for either of them. In general the residue Asp124 shows the largest variety, and in addition residues 28, 128 and 129 are responsible for most of the variation in the energy sampling between these

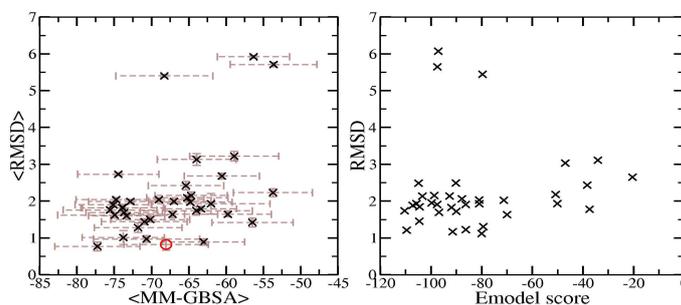


Figure 13: Docking results for the ligand XP9. The left panel shows the average rmsd vs the average binding energy for MM-GBSA calculations and the right panel shows the rmsd vs the Emodel score for the poses from Glide docking.

approaches. From the principle component analysis it is observed that the actual conformational difference between the standard long and multiple short trajectory approach for 300 K is very small. It is slightly larger for the same sampling at 500 K and in comparing the multiple short trajectory approaches for 300K and 500 K some structural changes can be observed. These are however not large enough to have a large impact on the average energy of the protein-ligand binding.

For the docking of ligands in the HIV-1 protease the MM-GBSA approach is able to distinguish nicely between good and bad poses from Glide docking as seen from the general trend of high rmsd poses obtaining very bad energies from the MM-GBSA calculations. There are exceptions from this, for example the ligand aha001 where the highest rmsd pose also has one of the best binding free energies. This is due to its ability to optimize exceptionally good energy contributions to residues 124 and 128 in spite of a worse overall rmsd value for the pose.

In general it can be seen from this work that clustering is a nice tool to identify residues contribution most to the binding free energy. There are only a handful of residues that are deterministic of the binding free energy for a given ligand in the HIV-1 protease. Once these are identified they could potentially be used to reduce the sampling space needed for flexible docking and by that obtain a more reliable pose from docking programs.

On the other hand, for the Estrogen Receptor we are not able to obtain such a clear distinction between high rmsd poses and those with low rmsd.

5 Acknowledgments

References

- [1] Jorgensen WL. The many roles of computation in drug discovery. *Science* 2004;303:1813–1188.
- [2] Sousa SF, Fernandes PA, Ramos MJ. Protein-ligand docking: current status and future challenges. *Proteins* 2006;65:15–26.
- [3] Tembe BL, McCammon JA. Ligand-receptor interactions. *Comput Chem* 1984;8:281–2836.
- [4] Thorsteinsdottir HB, Zoete V, Schwede, T and Meuwly M. How inaccuracies in protein structure models affect estimates of protein-ligand interactions: computational analysis of HIV-I protease inhibitor binding. *Proteins* 2006;65:407–23.
- [5] Gohlke H, Kiel C, Case DA. Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes. *J Mol Biol* 2003;330:891–913.
- [6] Caves LSD, Evanseck JD and Karplus M. Locally accessible conformations of proteins: Multiple molecular simulations of crambin. *Protein Science* 1997;7:649–666.
- [7] Loccisano AE, Acevedo O, DeChancie J, Schulze BG, Evanseck JD. Enhanced sampling by multiple molecular dynamics trajectories: carbomony myoglobin $10\ \mu\text{s}$ $A_1 - > A_{1,3}$ transition from ten 400 picosecond simulations. *J Mol Graph Model* 2004;22:369–76.
- [8] Schulze BG, Evanseck JD. Cooperative role of Arg45 and His64 in the spectroscopic A(3) state of carbonmonoxy myoglobin: Molecular dynamics simulations, multivariate analysis, and quantum mechanical computations. *J Am Chem Soc* 1999;121:6444–6454.
- [9] Monticelli L, Sorin EJ, Tieleman DP, Pande VS, Colombo G. Molecular simulation of multistate peptide dynamics: A comparison between microsecond timescale sampling and multiple shorter trajectories. *J Comput Chem* 2008;29:1740–1752.
- [10] Zoete V, Michielin O, Karplus M. Protein-ligand binding free energy estimation using molecular mechanics and continuum electrostatics. application to HIV-1 protease inhibitors. *J Comput Aided Mol Des* 2003;17:861–80.

- [11] Wang R, Fang X, Lu Y, Wang S. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J Med Chem* 2004;47:2977–2980.
- [12] Wang J, Wang W, Kollman PA, Case DA. Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graphics Modell* 2006;25:247–260.
- [13] Wang, J., Wolf, R. M.; Caldwell, J. W.; Kollman, PA.; Case, DA. Development and testing of a general AMBER force field. *J Comp Chem* 2004;25:1157–74.
- [14] MacKerell Jr. AD, Bashford D, Bellott M, Dunbrack Jr. RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau TK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher III WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* 1998;102:3586–3616.
- [15] Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus, M. CHARMM: A program for macromolecular energy, minimization and dynamics calculations. *J Comput Chem* 1983;4:187–217.
- [16] Lee MS, Salsbury Jr FR, Brooks III, CL. Novel generalized Born methods. *J Chem Phys* 2002;116:10606–14.
- [17] Lee MS, Feig M, Salsbury FR, Brooks CL. New analytic approximation to the standard molecular volume definition and its application to generalized born calculations. *J Comput Chem* 2003;24:1821–1821.
- [18] Srinivasan J, Cheatham III TE, Cieplak P, Kollman PA, Case DA. Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices. *J Am Chem Soc* 1998;120:9401–9.
- [19] Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L, Lee M, Lee T, Duan Y, Wang W, Donini O, Cieplak P, Srinivasan J, Case DA, Cheatham III TE. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc Chem Res* 2000;33:889–97.
- [20] Zoete V, Meuwly M, Karplus M. Study of the insulin dimerization: binding free energy calculations and per-residue free energy decomposition. *Proteins* 2005;61:79-93.

- [21] Hermann RB. Theory of hydrophobic bonding. II. correlation of hydrocarbon solubility in water with solvent cavity surface area. *J Phys Chem* 1972;76:2754–9.
- [22] Hasel W, Hendrikson TF, Still WC. A rapid approximation to the solvent accessible surface areas of atoms. *Tetrahedron Comput Methodol* 1988;1:103–116.
- [23] Still WC, Tempczyk A, Hawley RC, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J Am Chem Soc* 1990;112:6127–6129.
- [24] McQuarrie DA. *Statistical Mechanics*. Harper & Row, New York, , 1976.
- [25] Tidor B, Karplus M. The contribution of vibrational entropy to molecular association. *J Mol Biol* 1994;238:405–14.
- [26] Lindahl E, Hess B, van der Spoel D. GROMACS 3.0: a package for molecular simulation and trajectory analysis *J Mol Model* 2001;7:306–317.
- [27] Frey BJ, Dueck D. Clustering by Passing Messages Between Data Points. *Science* 2007;315:972–976.
- [28] Reva B, Antipin Y, Sander C Determinants of protein function revealed by combinatorial entropy optimization *Genome Biology* 2007;8:R.232-1–15.
- [29] Elofsson A, Nilsson L.. How consistent are molecular dynamics simulations? Comparing structure and dynamics in reduced and oxidized *Escherichia coli* thioredoxin. *J Mol Biol* 1993;233:766–80.
- [30] Oostenbrink BC, Pitera JW, van Lipzig MM , Meerman JH, van Gunsteren WF Simulations of the estrogen receptor ligand-binding domain: affinity of natural ligands and xenoestrogens. *J Med Chem* 2000;43:4594–4605.
- [31] Brooks III CL, Brunger A, Karplus M. Active site dynamics in protein molecules: a stochastic boundary molecular-dynamics approach. *Biopolymers* 1985;24:843–65.
- [32] Brooks III CL, Karplus M. Solvent effects on protein motion and protein effects on solvent motion. *J Mol Biol* 1989;208:159–181.
- [33] Kruger P, Luke, M, Szameit A SIMLYS – a software package for trajectory analysis of molecular dynamics simulations *Comput Phys Commun* 1991;62:371–80.

- [34] Diamond A Real-space Refinement of the Structure of Hen Egg-white Lysozyme J Mol Biol 1974;82:371–91.

6 Supplementary Information

6.1 HIV-1 Protease

6.1.1 Binding Energy Estimation

Tables 3 - 6 show the energy contributions for the standard long and multiple short trajectory approaches at 300K and 500K.

Table 3: Energy contributions for the standard long trajectory approach of HIV-1 protease at 300K

Ligand	VDW	Elec.	Elec. desolv.	Nonpolar desolv.	Total
a76889	-73.17 ± 4.13	-78.25 ± 6.51	82.85 ± 6.07	-9.04 ± 0.07	-77.61 ± 6.75
a76928	-79.22 ± 4.36	-72.43 ± 12.49	79.07 ± 6.18	-9.11 ± 0.11	-81.70 ± 9.78
a77003	-75.23 ± 3.99	-88.03 ± 7.09	93.06 ± 5.62	-9.02 ± 0.08	-79.22 ± 6.38
a78791	-77.17 ± 3.60	-60.38 ± 10.06	80.78 ± 7.27	-8.98 ± 0.08	-65.75 ± 7.10
a79285	-82.93 ± 3.93	-87.28 ± 9.91	97.50 ± 6.79	-9.07 ± 0.08	-81.78 ± 6.46
ag1343	-64.71 ± 3.47	-53.18 ± 5.88	63.97 ± 5.28	-7.67 ± 0.10	-61.59 ± 6.93
aha001	-56.09 ± 3.43	-54.11 ± 5.80	61.80 ± 5.89	-7.18 ± 0.10	-55.58 ± 6.32
aha006	-61.65 ± 3.22	-52.89 ± 4.29	67.30 ± 5.56	-7.48 ± 0.08	-54.73 ± 6.83
gr126045	-48.08 ± 3.66	-58.84 ± 5.56	68.52 ± 5.04	-6.67 ± 0.08	-45.07 ± 5.59
kni272	-68.06 ± 3.79	-68.95 ± 5.14	80.49 ± 5.62	-7.96 ± 0.09	-64.48 ± 6.92
l735524	-66.02 ± 3.47	-58.26 ± 5.60	72.43 ± 5.72	-8.24 ± 0.12	-60.09 ± 5.99
l738317	-62.84 ± 3.42	-67.74 ± 6.81	80.12 ± 6.26	-8.29 ± 0.11	-58.75 ± 7.68
sb203238	-60.96 ± 2.76	-36.69 ± 7.04	84.49 ± 8.76	-8.12 ± 0.11	-21.27 ± 8.93
sb204144	-67.81 ± 4.34	-64.68 ± 9.72	81.67 ± 6.58	-8.72 ± 0.15	-59.53 ± 8.60
sb206343	-62.32 ± 3.64	-69.04 ± 5.84	81.91 ± 6.64	-8.25 ± 0.11	-57.70 ± 7.55
vx478	-54.88 ± 3.45	-57.34 ± 5.84	63.78 ± 5.65	-7.00 ± 0.07	-55.44 ± 5.51

6.1.2 Application to docking

Figures 14 - 17 show the docking results for the 16 ligands of the HIV-1 protease.

Figures 14 - 17 show the results of the docking. The left panel shows the average rmsd to the reference structure vs the average binding energy value from the MM-GBSA calculations and the right panel shows the rmsd to the reference structure vs the Emoldel score for the poses docked with Glide. In figure 14 the ligands are generally docked with not too high accuracy, it is only ligand A76928

Table 4: Energy contributions for the multiple short trajectory approach of HIV-1 protease at 300K

Ligand	VDW	Elec.	Elec. desolv.	Nonpolar desolv.	Total
a76889	-74.36 ± 4.10	-76.44 ± 6.58	83.79 ± 6.72	-9.01 ± 0.08	-76.02 ± 6.92
a76928	-78.43 ± 4.44	-81.44 ± 6.68	88.42 ± 6.07	-9.14 ± 0.08	-80.59 ± 6.64
a77003	-75.30 ± 4.31	-89.01 ± 6.51	94.38 ± 5.56	-9.01 ± 0.08	-78.94 ± 6.93
a78791	-75.84 ± 4.28	-71.52 ± 5.77	85.45 ± 5.99	-9.01 ± 0.09	-70.92 ± 6.95
a79285	-83.05 ± 4.14	-86.00 ± 9.10	96.33 ± 6.78	-9.14 ± 0.09	-81.86 ± 6.60
ag1343	-65.42 ± 3.45	-51.22 ± 7.47	63.94 ± 5.26	-7.79 ± 0.09	-60.49 ± 7.32
aha001	-54.99 ± 3.18	-52.46 ± 5.90	62.78 ± 6.02	-7.25 ± 0.09	-51.92 ± 6.40
aha006	-60.33 ± 3.41	-51.53 ± 3.77	66.20 ± 5.16	-7.49 ± 0.08	-53.15 ± 5.55
gr126045	-49.76 ± 3.38	-61.23 ± 5.70	70.58 ± 5.29	-6.75 ± 0.11	-47.17 ± 6.01
kni272	-66.02 ± 3.49	-68.05 ± 5.91	82.56 ± 6.78	-7.98 ± 0.10	-59.50 ± 6.66
l735524	-65.48 ± 3.44	-54.86 ± 5.39	64.94 ± 5.02	-8.10 ± 0.11	-63.50 ± 5.99
l738317	-63.35 ± 3.42	-59.18 ± 6.08	78.36 ± 6.51	-8.35 ± 0.11	-52.53 ± 6.80
sb203238	-58.09 ± 3.66	-51.16 ± 7.63	93.41 ± 9.89	-8.20 ± 0.15	-24.03 ± 9.07
sb204144	-69.45 ± 4.02	-63.54 ± 7.60	83.12 ± 6.80	-8.86 ± 0.14	-58.74 ± 7.62
sb206343	-62.06 ± 3.75	-67.95 ± 6.00	82.10 ± 7.24	-8.32 ± 0.09	-56.24 ± 7.47
vx478	-55.05 ± 3.43	-56.02 ± 5.28	62.73 ± 5.82	-6.97 ± 0.08	-55.30 ± 5.89

that has one pose below 3Å. In all of these cases the higher the rmsd gets, the worse the energies become. For figure 15 the same is observed for A79285 and AG1343. Ligand AHA001 shows the same trend up to 4 Å but there is a small cluster of poses with high rmsd values that get very good binding free energies. The poses of ligand AHA006 are behaving very erratically and impossible to extract some general trend from those values. For the ligands in figure 16 the ligand L738317 shows a trend of worse energies with higher rmsd values and to some extent GR126045 shows the same, but the ligands KNI272 and L735524 behave more randomly. Although in the case of ligand KNI272 the rmsd values are all very high. Lastly for figure 17 the ligands SB203238 and SB204144 have only high rmsd poses.

Table 5: Energy contributions for the standard long trajectory approach of HIV-1 protease at 500K

Ligand	VDW	Elec.	Elec. desolv.	Nonpolar desolv.	Total
a76889	-69.78 ± 5.01	-65.65 ± 8.53	82.52 ± 7.44	-8.99 ± 0.11	-61.90 ± 8.49
a76928	-72.53 ± 4.98	-71.19 ± 11.94	84.53 ± 7.88	-9.06 ± 0.12	-68.24 ± 10.05
a77003	-70.54 ± 5.21	-82.35 ± 8.80	91.06 ± 8.28	-9.01 ± 0.13	-70.84 ± 9.77
a78791	-74.80 ± 4.93	-56.38 ± 7.31	77.68 ± 7.19	-8.96 ± 0.13	-62.45 ± 8.43
a79285	-76.22 ± 5.52	-83.26 ± 12.83	100.40 ± 10.93	-9.15 ± 0.12	-68.22 ± 9.32
ag1343	-61.88 ± 4.54	-47.81 ± 9.00	63.62 ± 6.96	-7.66 ± 0.13	-53.73 ± 8.05
aha001	-52.86 ± 4.42	-43.95 ± 8.30	58.34 ± 7.34	-7.20 ± 0.12	-45.67 ± 8.40
aha006	-60.98 ± 4.62	-48.66 ± 5.72	63.64 ± 6.89	-7.50 ± 0.10	-53.51 ± 7.11
gr126045	-46.93 ± 4.48	-46.93 ± 8.52	61.83 ± 7.38	-6.67 ± 0.14	-38.70 ± 7.23
kni272	-64.14 ± 4.52	-58.63 ± 8.13	80.77 ± 6.91	-7.83 ± 0.14	-49.82 ± 8.04
l735524	-62.32 ± 4.96	-56.32 ± 9.69	64.40 ± 7.15	-8.25 ± 0.16	-62.49 ± 7.94
l738317	-59.49 ± 4.69	-57.58 ± 11.12	73.22 ± 7.73	-8.22 ± 0.18	-52.08 ± 10.66
sb203238	-58.51 ± 4.13	-36.38 ± 8.02	70.10 ± 8.44	-8.03 ± 0.15	-32.82 ± 9.57
sb204144	-66.37 ± 5.32	-53.46 ± 9.27	77.50 ± 6.93	-8.85 ± 0.19	-51.18 ± 9.18
sb206343	-59.09 ± 5.17	-61.68 ± 8.11	79.52 ± 9.27	-8.21 ± 0.17	-49.46 ± 10.42
vx478	-53.98 ± 3.98	-44.84 ± 6.38	59.52 ± 6.35	-7.00 ± 0.11	-46.31 ± 7.08

Table 6: Energy contributions for the multiple short trajectory approach of HIV-1 protease at 500K

Ligand	VDW	Elec.	Elec. desolv.	Nonpolar desolv.	Total
a76889	-72.75 ± 4.29	-69.68 ± 6.35	81.47 ± 6.51	-9.02 ± 0.10	-69.98 ± 7.70
a76928	-75.64 ± 4.54	-69.10 ± 8.68	79.06 ± 7.00	-9.12 ± 0.11	-74.80 ± 8.35
a77003	-72.24 ± 4.79	-83.83 ± 7.39	91.36 ± 8.00	-9.00 ± 0.12	-73.71 ± 8.75
a78791	-77.53 ± 4.43	-58.52 ± 5.93	76.69 ± 6.43	-9.03 ± 0.11	-68.40 ± 7.67
a79285	-80.47 ± 4.99	-82.29 ± 9.85	96.36 ± 8.14	-9.22 ± 0.12	-75.62 ± 8.16
ag1343	-61.36 ± 4.10	-54.59 ± 5.98	67.15 ± 6.61	-7.81 ± 0.13	-56.60 ± 7.40
aha001	-54.40 ± 3.73	-47.35 ± 5.78	58.95 ± 5.83	-7.33 ± 0.14	-50.14 ± 7.01
aha006	-61.11 ± 3.57	-50.11 ± 4.89	66.62 ± 5.63	-7.43 ± 0.10	-52.03 ± 6.91
gr126045	-45.39 ± 3.73	-52.48 ± 9.16	64.60 ± 6.77	-6.82 ± 0.14	-40.09 ± 8.18
kni272	-65.42 ± 4.03	-66.62 ± 7.36	87.19 ± 7.35	-8.09 ± 0.13	-52.95 ± 8.45
l735524	-66.35 ± 4.65	-54.56 ± 9.65	65.63 ± 6.83	-8.44 ± 0.12	-63.73 ± 8.61
sb203238	-57.19 ± 3.44	-44.07 ± 8.28	66.78 ± 7.92	-8.05 ± 0.16	-42.53 ± 8.49
sb204144	-66.69 ± 4.68	-59.09 ± 8.04	77.62 ± 7.01	-8.85 ± 0.15	-57.02 ± 9.00
sb206343	-60.96 ± 4.29	-60.94 ± 7.91	78.06 ± 7.95	-8.36 ± 0.18	-52.20 ± 9.42
vx478	-55.39 ± 3.50	-46.59 ± 5.64	60.99 ± 6.65	-7.10 ± 0.10	-48.09 ± 7.07

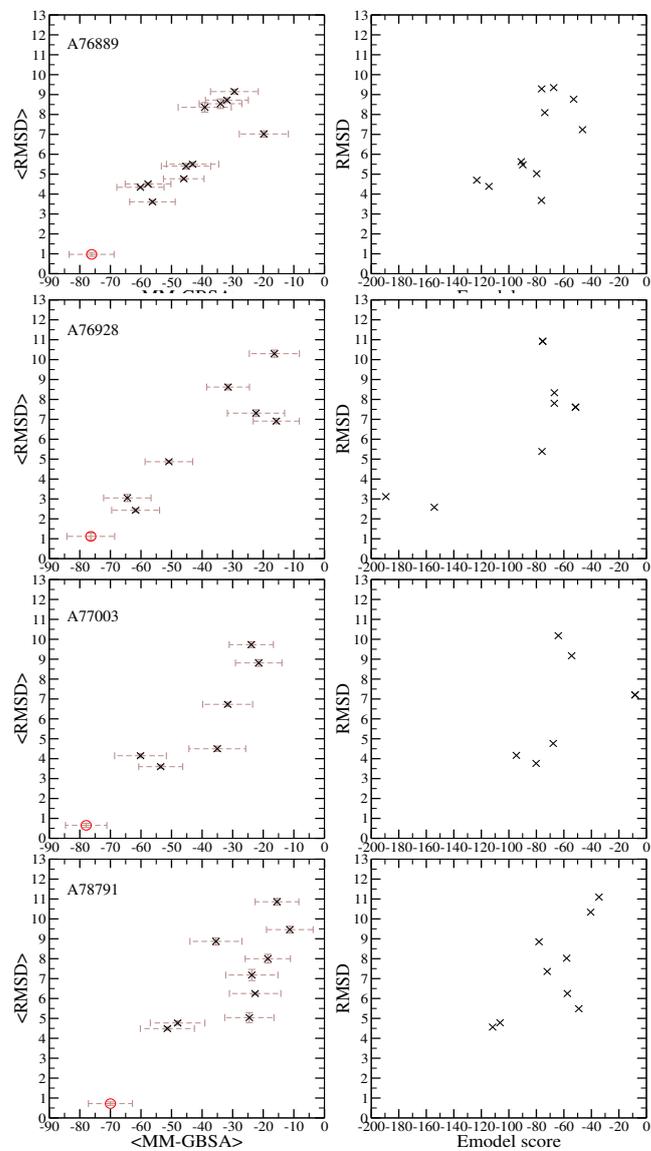


Figure 14: Docking results for the ligands: a76889, a76928, a77003 and a78791. The left panel shows the average rmsd vs the average binding energy for MM-GBSA calculations and the right panel shows the rmsd vs the Emodel score for the poses from Glide docking.

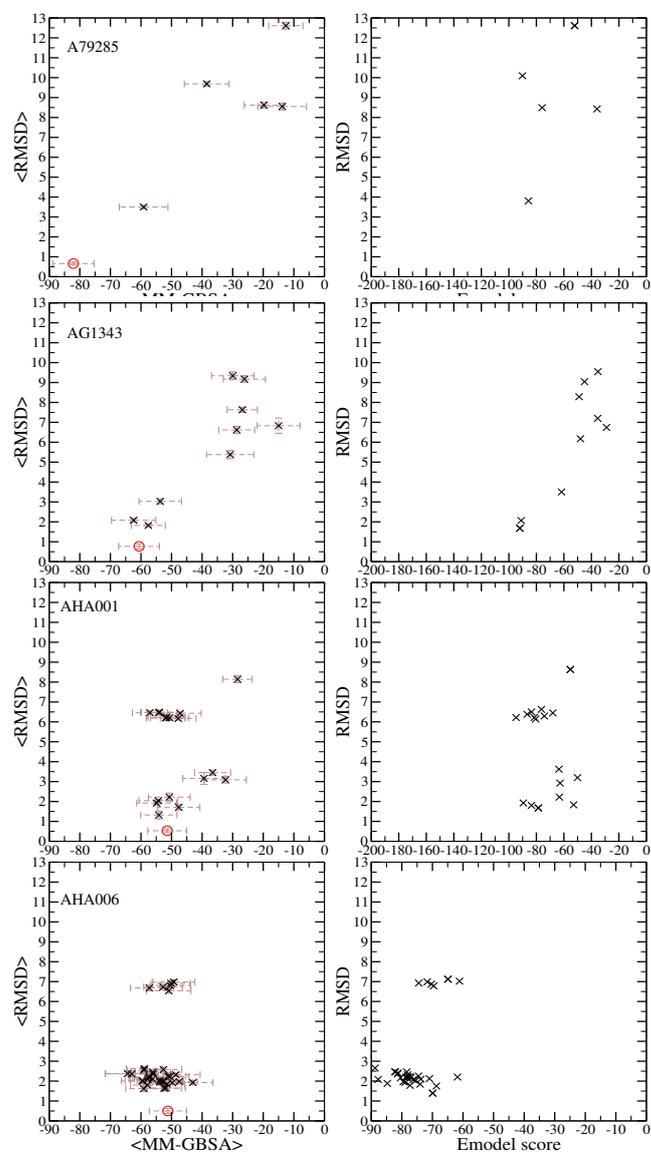


Figure 15: Docking results for the ligands: a79285, ag1343, aha001 and aha006. The left panel shows the average rmsd vs the average binding energy for MM-GBSA calculations and the right panel shows the rmsd vs the Emodel score for the poses from Glide docking.

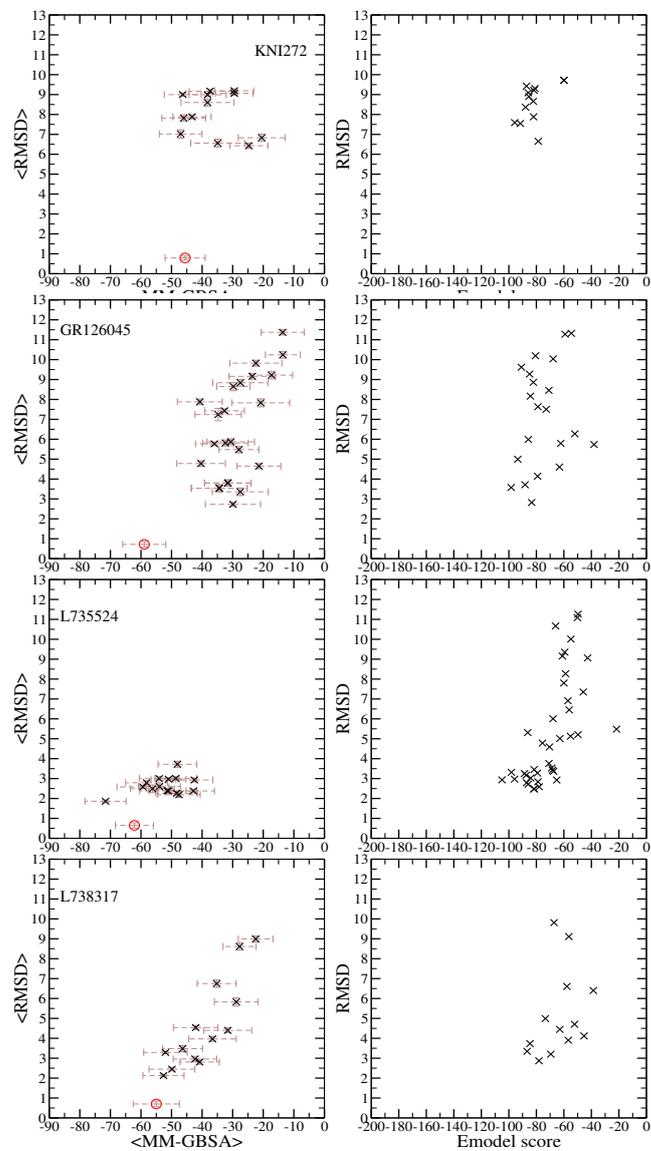


Figure 16: Docking results for the ligands: kni272, gr126034, 1735524 and 1738317. The left panel shows the average rmsd vs the average binding energy for MM-GBSA calculations and the right panel shows the rmsd vs the Emodel score for the poses from Glide docking.

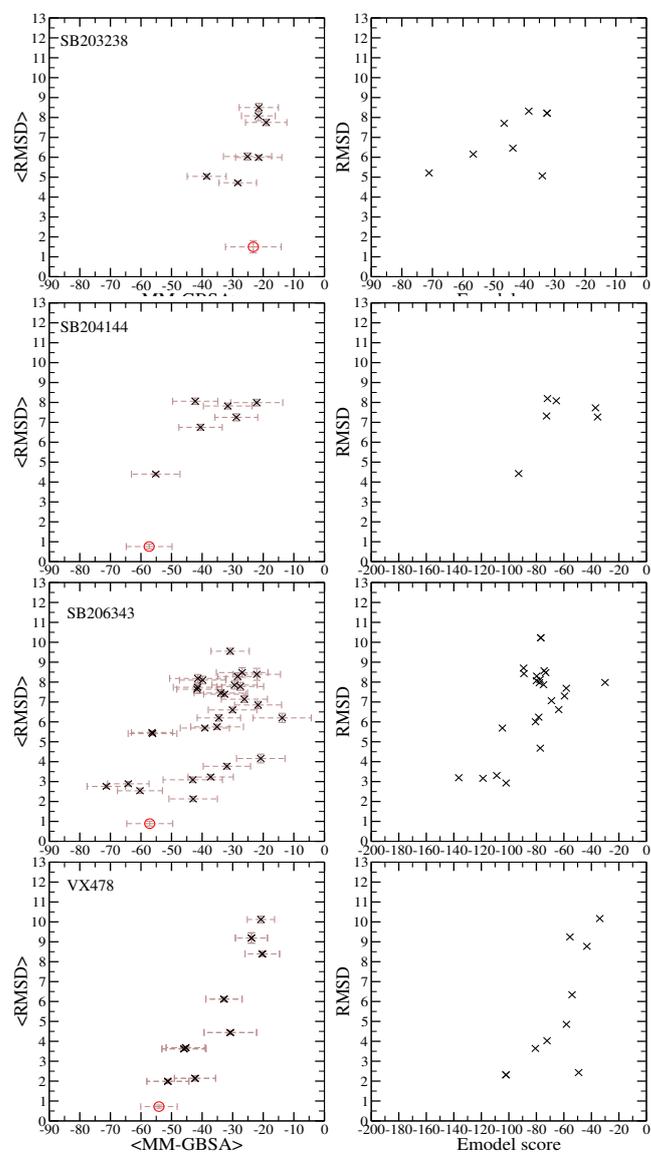


Figure 17: Docking results for the ligands: sb203238, sb204144, sb206343 and vx478. The left panel shows the average rmsd vs the average binding energy for MM-GBSA calculations and the right panel shows the rmsd vs the Emodel score for the poses from Glide docking.

6.2 Estrogen Receptor

6.2.1 Binding Energy Estimation

Table 7 and table 8 show the energy contributions for the standard long and multiple short trajectory approaches at 300K and 500K.

Table 7: Energy contributions for the Estrogen receptor at 300K

Ligand	VDW	Elec.	Elec. desolv.	Nonpolar desolv.	Total
gwq	-29.89 ± 3.00	-29.31 ± 3.71	27.07 ± 4.09	-4.17 ± 0.15	-36.31 ± 4.27
gwr	-37.34 ± 3.19	-27.26 ± 4.57	25.79 ± 4.54	-4.67 ± 0.16	-43.47 ± 4.50
sj0	-60.89 ± 4.02	-201.23 ± 9.56	192.63 ± 8.78	-6.51 ± 0.22	-76.00 ± 5.87
uom	-60.82 ± 3.70	-185.09 ± 9.39	182.84 ± 8.49	-6.59 ± 0.23	-69.66 ± 6.48
x7e	-32.86 ± 3.04	-38.38 ± 4.69	37.79 ± 5.39	-4.59 ± 0.16	-38.04 ± 6.12
xp1	-62.61 ± 4.12	-181.92 ± 10.51	185.55 ± 8.51	-6.72 ± 0.23	-65.70 ± 6.07
xp6	-61.98 ± 5.39	-183.52 ± 17.27	186.83 ± 14.04	-6.79 ± 0.44	-65.45 ± 8.51
xp9	-58.81 ± 3.97	-188.01 ± 9.33	185.45 ± 8.58	-6.68 ± 0.23	-68.05 ± 6.07
xpc	-61.09 ± 4.08	-207.11 ± 10.14	198.26 ± 8.89	-6.60 ± 0.22	-76.54 ± 5.85
xqc	-63.27 ± 3.52	-153.21 ± 8.54	159.90 ± 9.98	-6.58 ± 0.22	-63.16 ± 5.98
b1v	-35.44 ± 2.94	-24.34 ± 4.52	24.83 ± 4.45	-4.49 ± 0.15	-39.44 ± 4.63
bj4	-51.73 ± 3.49	-180.26 ± 9.28	178.87 ± 8.60	-6.23 ± 0.21	-59.35 ± 6.02
fai	-34.23 ± 3.14	-27.92 ± 4.52	25.82 ± 3.73	-4.66 ± 0.16	-40.98 ± 4.03

Table 8: Energy contributions for the Estrogen receptor at 500K

Ligand	VDW	Elec.	Elec. desolv.	Nonpolar desolv.	Total
gwq	-28.74 ± 3.12	-27.50 ± 5.06	27.70 ± 4.79	-4.29 ± 0.15	-32.83 ± 5.29
gwr	-36.52 ± 3.41	-25.25 ± 4.68	25.36 ± 4.70	-4.72 ± 0.16	-41.14 ± 5.21
sj0	-60.00 ± 4.45	-193.25 ± 19.01	192.68 ± 14.04	-6.63 ± 0.23	-67.21 ± 8.43
uom	-61.07 ± 4.17	-155.79 ± 10.15	167.10 ± 9.59	-6.64 ± 0.25	-56.40 ± 5.98
x7e	-33.99 ± 3.34	-35.88 ± 4.94	36.96 ± 4.67	-4.58 ± 0.17	-37.50 ± 5.33
xp1	-61.43 ± 4.30	-170.76 ± 22.04	172.11 ± 15.93	-6.88 ± 0.25	-66.97 ± 10.23
xp6	-60.63 ± 4.38	-183.47 ± 12.59	179.90 ± 12.66	-7.01 ± 0.24	-71.21 ± 8.44
xp9	-57.90 ± 4.26	-184.15 ± 13.88	177.16 ± 11.20	-6.69 ± 0.24	-71.57 ± 7.60
xpc	-58.94 ± 4.39	-166.80 ± 12.97	170.41 ± 12.55	-6.78 ± 0.24	-62.11 ± 7.06
xqc	-62.10 ± 4.01	-154.48 ± 11.27	161.89 ± 13.14	-6.68 ± 0.24	-61.37 ± 6.03
b1v	-34.64 ± 3.27	-23.81 ± 6.38	26.56 ± 4.99	-4.55 ± 0.16	-36.44 ± 5.94
bj4	-50.62 ± 3.58	-133.67 ± 13.24	148.78 ± 14.64	-6.41 ± 0.22	-41.92 ± 5.85
fai	-37.41 ± 3.62	-22.38 ± 5.07	25.32 ± 4.99	-4.60 ± 0.17	-39.07 ± 5.75

6.2.2 Application to Docking

Figures 18 - 20 show the docking results of the Estrogen Receptor.

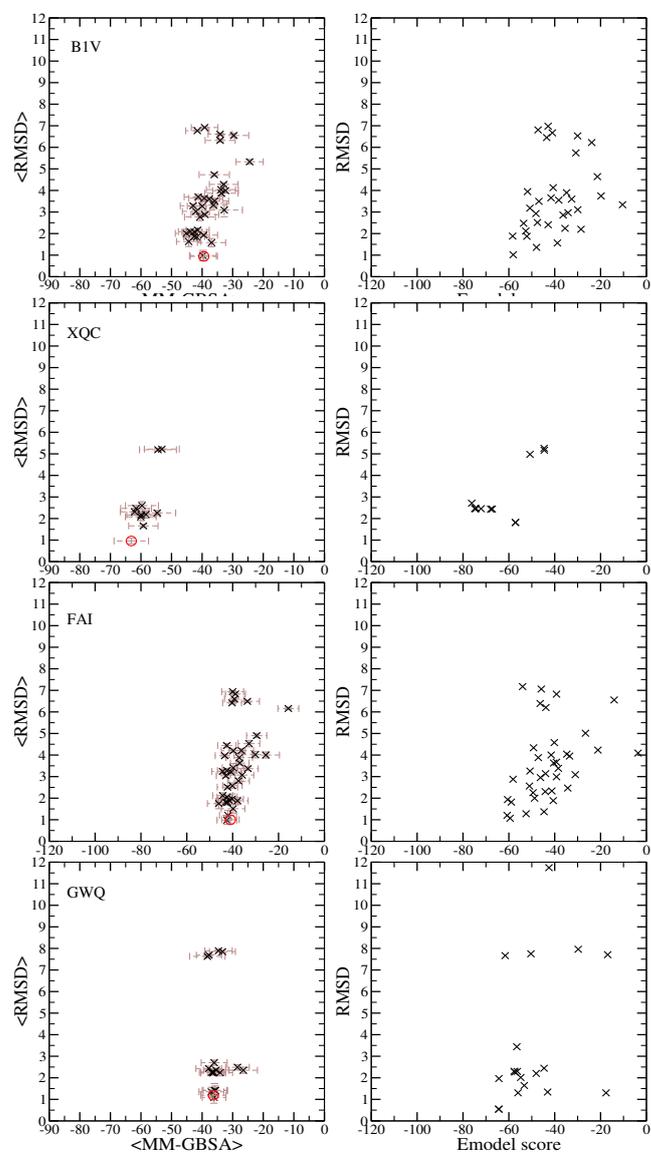


Figure 18: Docking results for the ligands: B1V, XQC, FAI and GWQ. The left panel shows the average rmsd vs the average binding energy for MM-GBSA calculations and the right panel shows the rmsd vs the Emodel score for the poses from Glide docking.

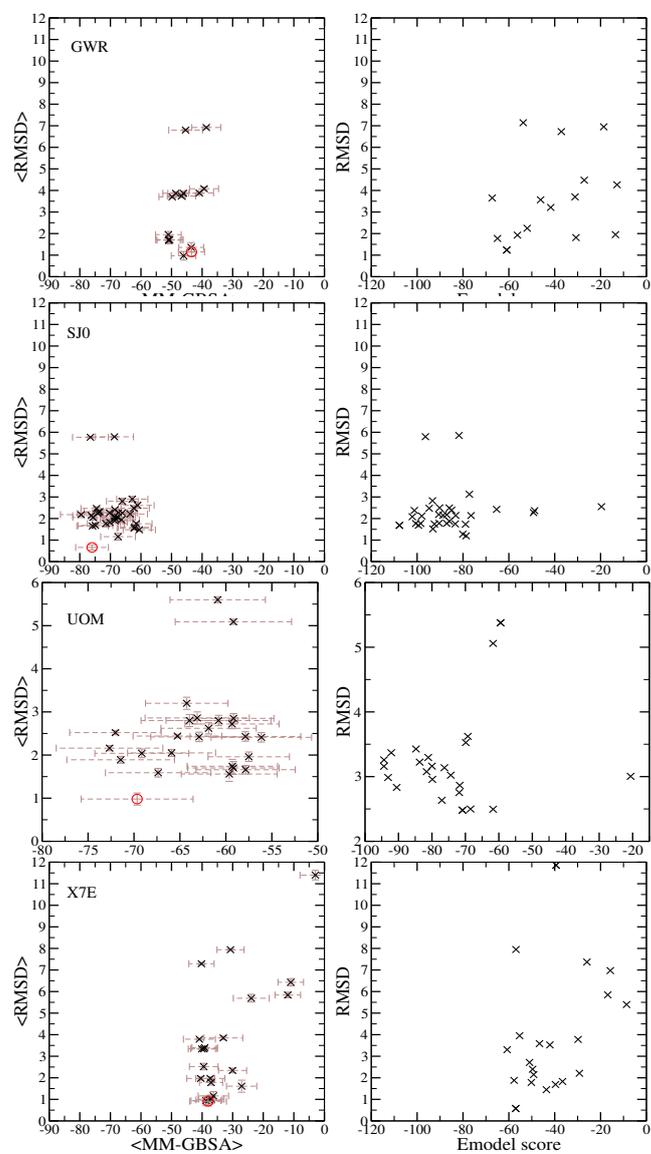


Figure 19: Docking results for the ligands: GWR, SJ0, UOM and X7E. The left panel shows the average rmsd vs the average binding energy for MM-GBSA calculations and the right panel shows the rmsd vs the Emodel score for the poses from Glide docking.

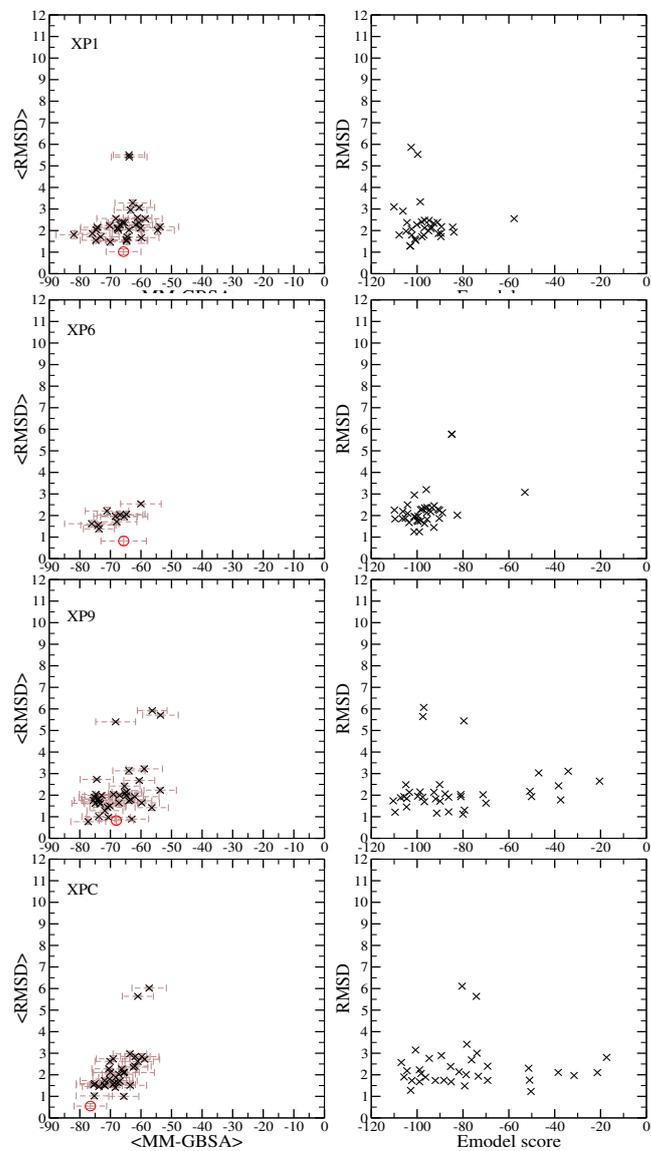


Figure 20: Docking results for the ligands: XP1, XP6, XP9 and XPC. The left panel shows the average rmsd vs the average binding energy for MM-GBSA calculations and the right panel shows the rmsd vs the Emodel score for the poses from Glide docking.

6 Discussion

Accurate three-dimensional structures are essential for estimating binding free energy between protein and their ligands using atomistic simulations such as molecular mechanics. In the cases where no experimentally determined structures are available, computational methods for modeling their three-dimensional coordinates can be applied. There are various methods to do this, homology modeling currently being the most accurate.

In the first part of this thesis we address the question of how accurate homology models need to be in order to give useful insights into binding energy calculations. A prerequisite to any such investigation is to have a reliable protein system in which errors can be introduced and investigated. After the protein system has been verified, homology models of decreasing back-bone accuracy are built. Finally, the effect of incorrectly placed side-chains is investigated. The HIV-1 protease was validated as a model system for estimating ligand binding energies. Homology models were built for this protein based on its sequence similarity to other members of the retroviral protease family. We found that for the HIV-1 protease we are able to obtain a very good correlation to experimental binding energies, even for sequence similarities as low as 32 %. Kairys *et al* explored the use of homology models in docking. They find that using homology models for the docking can lead to as good enrichment of known ligand comparable with using the actual crystal structure¹⁵⁴. In addition they find that the sequence similarity to the template used to build the homology model is not predictive of the accuracy of the results¹⁵⁴. This is comparable with our own findings where binding energy calculations using homology models built on templates with only 32 % sequence identity obtain a very good correlation to experimental data. Furthermore, we found that for a correct estimation of binding free energies it is crucial to have the side-chains placed correctly. The worst correlation to the experimental values was found when the side-chains were modeled using only a rotamer library without taking template conformation into account. This is in accord with McGovern *et al* who found that often small structural errors such as incorrect side-chain rotamer

were the reason for a poor performance of some homology models¹⁵⁵. This is also a general observation for homology models as highlighted in recent CASP experiment¹⁵⁶. It is therefore not surprising that in our work the worst correlation to the experimental binding energies arises from models where the only difference to the crystal structure was the placement of side-chains. The MM-GBSA approach validated here was shown to be very promising to rank ligands and is therefore suitable for application to protein systems where the amount of experimental data is smaller and the quality of the structures is more diverse. However, further investigations to narrow down specific side-chains would be of interest.

One significant limitation for the routine application of atomistic simulations in the drug discovery process is their excessive computational cost. It is common to dock large numbers of compounds into the binding sites of proteins and to score the best poses obtained. These docking and scoring functions often contain a number of approximations which can decrease the accuracy of the outcome. It would be very valuable to have a fast and accurate scoring function to calculating the interactions between proteins and ligands accurately. One way to achieve more accurate scoring is to use conformational sampling and force field based scoring functions, but their long computational times make them less suitable for dealing with the large number of compounds that is required for drug discovery. For the second part of the thesis we modify the previously validated MM-GBSA approach in order to reduce the computational time needed to estimate binding energies. A novel method is developed that allows multiple short molecular dynamics simulations to be run on a grid of distributed computers (PC Grid) instead of the traditional way of running a single long molecular dynamics simulation. We find that using this multiple short molecular dynamics approach we are able to obtain a very good correlation to experimental binding free energies for the HIV-1 model system. This methodology was found to shorten the computational time needed from around 4 days with the traditional approach to around 2 days for the multiple short approach. We found that the multiple short molecular dynamics simulation approach at 300K samples very similar conformations and energies as the traditional long approach. By raising the temperature to 500K the multiple short and traditional long approaches sample slightly different conformations but not enough to have an impact on the energies.

A major part of analysis is to find methods that distinguish between characteristics of a system, for example conformational variability. There are various ways that have been developed to investigate differences in sampling and structures. Monticelli *et*

al pre-defined certain conformational state and then compared how they were populated between different simulation approaches¹⁰⁷. An interesting method to analyze the structural differences and similarities in alternative structural models based on C α or side chain centroid variability¹⁵⁷. However, we found this method is ill adapted to the large scale analysis needed here. A common method to estimate the conformational space sampled is principle component analysis, which reduces dimensionality to identify major contributions to the atomic displacements¹⁰⁴.

Clustering is a popular tool to divide a large dataset into subsets based on a given criteria. These subsets can then be analyzed further to find unique features that characterize specificities of the subset. The advantage of using clustering tools is that only one property is needed as criteria, but in addition information can be inferred from other contributions. In this work a novel analysis was developed that is specific to the protein-ligand binding residues. This analysis is based on defining a cluster metric, which in our case is either the per-residue binding energy contribution or per-residue distance RMSD. The values for each of the metrics for all residues in the binding pocket then give a fingerprint of the interaction in a particular frame of the simulation. These fingerprints for all frames of simulations that are being compared can then be clustered to identify subsets that share common qualities. From these subsets we can then identify the residues that are specific for each subset by using a modified version of the proteinkey analysis¹⁵³.

A prerequisite for using the MM-GBSA approach on other systems is obtaining parameters for the ligands that are not described in the CHARMM force field⁹¹. We successfully automated the Antechamber program to generate parameters suitable for calculations using CHARMM and validated the results. In order to test if the MM-GBSA method is applicable to other protein systems, we applied the same methodology to experimental ligand binding data to the estrogen receptor β . Using this protein system we found the same, which is that by using the multiple short molecular dynamics approach we are able to obtain very good correlation to experimental binding energies. This is comparable to previous findings where it has been suggested that a good sampling of the conformational space of protein can be achieved by running multiple short trajectories instead of a single long trajectory^{101,102}.

These findings of reduced computational time and transferability to the estrogen receptor encouraged us to apply our relatively fast and accurate scoring function to the question of scoring conformations of ligands derived from docking simulations. The objective is twofold; to see if the docked pose can be improved by molecular sim-

ulation and to see if a better ranking of the poses can be obtained with a more accurate scoring function. For the docking of ligands in the HIV-1 protease the MM-GBSA approach is able to improve the ranking of the poses and distinguish between poses that are similar to the actual crystal structure pose and those that are not. There are however exceptions from this, cases where low rmsd poses and those with very high rmsd all obtain very favorable binding energies. This is explained by the ability of the MM method to optimize energetically favorable interactions to the protein in spite of a large overall rmsd difference. In a study by Nervall et al, a series of HIV-1 reverse transcriptase inhibitors was cross-docked into a non-native crystal structure. They were able to distinguish between correct binding mode and an incorrect one which displayed a flipped heterocyclic group using molecular dynamics simulation and the LIE method¹⁵⁸. Graves et al found that rescoring docked poses with MMGBSA better distinguished known ligands from known decoys as compared to the scoring functions¹⁵⁹. They also found that rescoring introduced new false positives in spite of rescuing docking false positives.

However, for the docking of ligands to the estrogen receptor we didn't see improvement in the ranking. It has been suggested that alchemical relative free energy methods are outperforming approximate methods such as MM-GBSA in ligand binding studies to the estrogen receptor^{80,160}. Since the computation time of such methods is considerable it is worthwhile to identify limitations of MM-GBSA in order to develop a reliable universal method.

The MM-GBSA method in combination to the clustering analysis has the potential to identify a small number of protein residues that are necessary to obtain favorable binding free energies to a docked compound. Once these residues are identified they can be used in flexible docking to increase accuracy but reduce the time needed if the whole binding site was allowed to stay flexible. However, there are some limitations highlighted with its inability to improve the ranking of the estrogen receptor ligands. These limitations have to be investigated further to identify their source. Further studies with different protein systems are required to elaborate, if and for which systems MM-GBSA based methods can provide significant improvements over simple scoring functions for ligand binding.

7 Summary & Outlook

This thesis has addressed some important limitations of the applications of computational modeling methods in the drug discovery process. For the homology modeling aspects, we and others have shown that even models with low sequence similarity can be used to obtain very good correlation to experimental data. However, we only looked at the HIV-1 protease which belongs to a very structurally conserved family. It would be of interest to look into other less favorable cases. Of even more interest would be to investigate in more detail what exactly is the limitation of the side-chain accuracy, to narrow it down to specific interactions.

We have successfully reduced the computational time required for MM-GBSA calculations by using a multiple short trajectory approach for the conformational sampling. This method has also been shown to give rise to good correlation to experimental data. In its applications to docked poses it is able to distinguish between good and bad poses for the HIV-1 protease but is unable to improve the poses significantly. It would be of great interest to look into improving the poses by alternative means of sampling. The MM-GBSA inability of improving the ranking of the ligands for estrogen receptor needs detailed analysis to answer the question whether it is a limitation of the scoring function or the docking program. In order to do that, more systems have to be included and analyzed to find the weaknesses of the approach. Ultimately it is the goal to have a method that not only will improve the pose obtain from docking but in addition correctly predict the binding energy. In addition this method would need to be transferable between different protein systems and take solvation, entropy and allosteric effects into account. Finally it would be valuable to have an analysis pipeline suitable for non experts where top ranking hits from docking runs could be analyzed in more detail.

Bibliography

1. Chothia, C. and Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *Embo J* **5**(4), 823–6 (1986).
2. Overington, J. P., Al-Lazikani, B., and Hopkins, A. L. How many drug targets are there? *Nature reviews* **5**(12), 993–6 (2006).
3. Friedman, A. and Perrimon, N. Genome-wide high-throughput screens in functional genomics. *Curr Opin Genet Dev* **14**(5), 470–6 (2004).
4. Allison, D. B., Cui, X., Page, G. P., and Sabripour, M. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet* **7**(1), 55–65 (2006).
5. Searls, D. B. Using bioinformatics in gene and drug discovery. *Drug Discov Today* **5**(4), 135–143 (2000).
6. Daggett, V. and Fersht, A. The present view of the mechanism of protein folding. *Nat Rev Mol Cell Biol* **4**(6), 497–502 (2003).
7. Venclovas, C., Zemla, A., Fidelis, K., and Moult, J. Assessment of progress over the casp experiments. *Proteins* **53 Suppl 6**, 585–95 (2003).
8. Ensign, D. L., Kasson, P. M., and Pande, V. S. Heterogeneity even at the speed limit of folding: large-scale molecular dynamics study of a fast-folding variant of the villin headpiece. *J Mol Biol* **374**(3), 806–16 (2007).
9. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic Acids Res* **31**(1), 365–70 (2003).
10. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The protein data bank. *Nucleic Acids Res* **28**(1), 235–42 (2000).

11. Blundell, T. L., Sibanda, B. L., Sternberg, M. J., and Thornton, J. M. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* **326**(6111), 347–52 (1987).
12. Tramontano, A. and Morea, V. Assessment of homology-based predictions in casp5. *Proteins* **53 Suppl 6**, 352–68 (2003).
13. Pearson, W. R. and Lipman, D. J. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* **85**(8), 2444–8 (1988).
14. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res* **25**(17), 3389–402 (1997).
15. Barton, G. J. Protein multiple sequence alignment and flexible pattern matching. *Methods Enzymol* **183**, 403–28 (1990).
16. Suyama, M., Matsuo, Y., and Nishikawa, K. Comparison of protein structures using 3d profile alignment. *J Mol Evol* **44 Suppl 1**, S163–73 (1997).
17. Notredame, C., Higgins, D. G., and Heringa, J. T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**(1), 205–17 (2000).
18. Rost, B. Twilight zone of protein sequence alignments. *Protein Eng* **12**(2), 85–94 (1999).
19. Schwede, T., Kopp, J., Guex, N., and Peitsch, M. C. Swiss-model: An automated protein homology-modeling server. *Nucleic Acids Res* **31**(13), 3381–5 (2003).
20. Arnold, K., Bordoli, L., Kopp, J., and Schwede, T. The swiss-model workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22**(2), 195–201 (2006).
21. Johnson, M. S., Srinivasan, N., Sowdhamini, R., and Blundell, T. L. Knowledge-based protein modeling. *Crit Rev Biochem Mol Biol* **29**(1), 1–68 (1994).
22. Sali, A. and Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**(3), 779–815 (1993).
23. Ponder, J. W. and Richards, F. M. Tertiary templates for proteins. use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol* **193**(4), 775–91 (1987).

24. Dunbrack, R. L., J. and Karplus, M. Backbone-dependent rotamer library for proteins. application to side-chain prediction. *J Mol Biol* **230**(2), 543–74 (1993).
25. Fiser, A., Do, R. K., and Sali, A. Modeling of loops in protein structures. *Protein Sci* **9**(9), 1753–73 (2000).
26. Rohl, C. A., Strauss, C. E., Chivian, D., and Baker, D. Modeling structurally variable regions in homologous proteins with rosetta. *Proteins* **55**(3), 656–77 (2004).
27. Wojcik, J., Mornon, J. P., and Chomilier, J. New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. *J Mol Biol* **289**(5), 1469–90 (1999).
28. Chothia, C. and Lesk, A. M. Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol* **196**(4), 901–17 (1987).
29. Sanchez, R. and Sali, A. Evaluation of comparative protein structure modeling by modeller-3. *Proteins Suppl* **1**, 50–8 (1997).
30. Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. Procheck: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26**, 283–291 (1993).
31. Vriend, G. What if: a molecular modeling and drug design program. *J Mol Graph* **8**(1), 52–6, 29 (1990).
32. Eisenberg, D., Luthy, R., and Bowie, J. U. Verify3d: assessment of protein models with three-dimensional profiles. *Methods Enzymol* **277**, 396–404 (1997).
33. Melo, F. and Feytmans, E. Assessing protein structures with a non-local atomic interaction energy. *J Mol Biol* **277**(5), 1141–52 (1998).
34. Leach, A. R., Shoichet, B. K., and Peishoff, C. E. Prediction of protein-ligand interactions. docking and scoring: successes and gaps. *J Med Chem* **49**(20), 5851–5 (2006).
35. Brooijmans, N. and Kuntz, I. D. Molecular recognition and docking algorithms. *Annu Rev Biophys Biomol Struct* **32**, 335–73 (2003).
36. Kitchen, D. B., Decornez, H., Furr, J. R., and Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews* **3**(11), 935–49 (2004).

37. Ewing, T. J., Makino, S., Skillman, A. G., and Kuntz, I. D. Dock 4.0: search strategies for automated molecular docking of flexible molecule databases. *J Comput Aided Mol Des* **15**(5), 411–28 (2001).
38. Goodsell, D. S., Lauble, H., Stout, C. D., and Olson, A. J. Automated docking in crystallography: analysis of the substrates of aconitase. *Proteins* **17**(1), 1–10 (1993).
39. Rarey, M., Kramer, B., Lengauer, T., and Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* **261**(3), 470–89 (1996).
40. MOE. Chemical computing group, (2003).
41. 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., and Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *J Med Chem* **47**(7), 1739–49 (2004).
42. Gold. Version 1.2, (2003).
43. Welch, W., Ruppert, J., and Jain, A. N. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chem Biol* **3**(6), 449–62 (1996).
44. Westhead, D. R., Clark, D. E., and Murray, C. W. A comparison of heuristic search algorithms for molecular docking. *J Comput Aided Mol Des* **11**(3), 209–28 (1997).
45. Kearsley, S. K., Underwood, D. J., Sheridan, R. P., and Miller, M. D. Flexibases: a way to enhance the use of molecular docking methods. *J Comput Aided Mol Des* **8**(5), 565–82 (1994).
46. Leach, A. R. *Molecular Modelling: Principles and Applications*. Addison Wesley Longman Limited, Harlow, (1996).
47. DesJarlais, R. L., Sheridan, R. P., Dixon, J. S., Kuntz, I. D., and Venkataraghavan, R. Docking flexible ligands to macromolecular receptors by molecular shape. *J Med Chem* **29**(11), 2149–53 (1986).
48. Baxter, C. A., Murray, C. W., Clark, D. E., Westhead, D. R., and Eldridge, M. D. Flexible docking using tabu search and an empirical estimate of binding affinity. *Proteins* **33**(3), 367–82 (1998).

49. Di Nola, A., Roccatano, D., and Berendsen, H. J. Molecular dynamics simulation of the docking of substrates to proteins. *Proteins* **19**(3), 174–82 (1994).
50. Pak, Y. S. and Wang, S. M. Application of a molecular dynamics simulation method with a generalized effective potential to the flexible molecular docking problems. *Journal of Physical Chemistry B* **104**(2), 354–359 (2000).
51. Hart, T. N. and Read, R. J. A multiple-start monte carlo docking method. *Proteins* **13**(3), 206–22 (1992).
52. Oshiro, C. M., Kuntz, I. D., and Dixon, J. S. Flexible ligand docking using a genetic algorithm. *J Comput Aided Mol Des* **9**(2), 113–30 (1995).
53. Leach, A. R. Ligand docking to proteins with discrete side-chain flexibility. *J Mol Biol* **235**(1), 345–56 (1994).
54. De Maeyer, M., Desmet, J., and Lasters, I. The dead-end elimination theorem: mathematical aspects, implementation, optimizations, evaluation, and performance. *Methods Mol Biol* **143**, 265–304 (2000).
55. Knegtel, R. M., Kuntz, I. D., and Oshiro, C. M. Molecular docking to ensembles of protein structures. *J Mol Biol* **266**(2), 424–40 (1997).
56. Fischer, E. *Ber. Dtsch. Chem. Ges* **27**, 2985–2993 (1894).
57. Gohlke, H. and Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angewandte Chemie-International Edition* **41**(15), 2645–2676 (2002).
58. Homans, S. W. Water, water everywhere—except where it matters? *Drug discovery today* **12**(13-14), 534–9 (2007).
59. Jeffrey, G. and Saenger, W. *Hydrogen bonding in biological structures*. Springer, Berlin, (1991).
60. Stahl, M. and Bohm, H. J. Development of filter functions for protein-ligand docking. *Journal of Molecular Graphics & Modelling* **16**(3), 121–132 (1998).
61. Chothia, C. Hydrophobic bonding and accessible surface area in proteins. *Nature* **248**(446), 338–9 (1974).

62. Wang, J. M., Morin, P., Wang, W., and Kollman, P. A. Use of mm-pbsa in reproducing the binding free energies to hiv-1 rt of tibo derivatives and predicting the binding mode to hiv-1 rt of efavirenz by docking and mm-pbsa. *Journal of the American Chemical Society* **123**(22), 5221–5230 (2001).
63. Brady, G. P. and Sharp, K. A. Entropy in protein folding and in protein-protein interactions. *Curr Opin Struct Biol* **7**(2), 215–21 (1997).
64. Page, M. I. and Jencks, W. P. Entropic contributions to rate accelerations in enzymic and intramolecular reactions and the chelate effect. *Proc Natl Acad Sci U S A* **68**(8), 1678–83 (1971).
65. Cheng, Y. and Prusoff, W. H. Relationship between the inhibition constant (k_i) and the concentration of inhibitor which causes 50 per cent inhibition (i_{50}) of an enzymatic reaction. *Biochemical pharmacology* **22**(23), 3099–108 (1973).
66. Strachan, R. T., Ferrara, G., and Roth, B. L. Screening the receptorome: an efficient approach for drug discovery and target validation. *Drug discovery today* **11**(15-16), 708–16 (2006).
67. Allen, M., Hall, D., Collins, B., and Moore, K. A homogeneous high throughput nonradioactive method for measurement of functional activity of gs-coupled receptors in membranes. *Journal of biomolecular screening* **7**(1), 35–44 (2002).
68. Gilson, M. K. and Zhou, H. X. Calculation of protein-ligand binding affinities. *Annual Review of Biophysics and Biomolecular Structure* **36**, 21–42 (2007).
69. Charifson, P. S., Corkery, J. J., Murcko, M. A., and Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J Med Chem* **42**(25), 5100–9 (1999).
70. Allen, F. H. The cambridge structural database: a quarter of a million crystal structures and rising. *Acta crystallographica* **58**(Pt 3 Pt 1), 380–8 (2002).
71. Muegge, I. A knowledge-based scoring function for protein-ligand interactions: Probing the reference state. *Perspectives in Drug Discovery and Design* **20**(1), 99–114 (2000).
72. Gohlke, H., Hendlich, M., and Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *Journal of Molecular Biology* **295**(2), 337–356 (2000).

73. Bohm, H. J. Ludi: rule-based automatic design of new substituents for enzyme inhibitor leads. *J Comput Aided Mol Des* **6**(6), 593–606 (1992).
74. Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V., and Mee, R. P. Empirical scoring functions .1. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *Journal of Computer-Aided Molecular Design* **11**(5), 425–445 (1997).
75. Kramer, B., Rarey, M., and Lengauer, T. Evaluation of the flexx incremental construction algorithm for protein-ligand docking. *Proteins-Structure Function and Genetics* **37**(2), 228–241 (1999).
76. Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., and Olson, A. J. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry* **19**(14), 1639–1662 (1998).
77. Shoichet, B. K., Leach, A. R., and Kuntz, I. D. Ligand solvation in molecular docking. *Proteins* **34**(1), 4–16 (1999).
78. Tembe, B. and McCammon, J. A. Ligand-receptor interactions. *Comput. Chem.* **8**, 281–283 (1984).
79. Rao, B. G., Kim, E. E., and Murcko, M. A. Calculation of solvation and binding free energy differences between vx-478 and its analogs by free energy perturbation and amsol methods. *J Comput Aided Mol Des* **10**(1), 23–30 (1996).
80. Oostenbrink, C. and van Gunsteren, W. F. Free energies of binding of polychlorinated biphenyls to the estrogen receptor from a single simulation. *Proteins* **54**(2), 237–46 (2004).
81. Helms, V. and Wade, R. C. Computational alchemy to calculate absolute protein-ligand binding free energy. *Journal of the American Chemical Society* **120**(12), 2710–2713 (1998).
82. Aqvist, J., Luzhkov, V. B., and Brandsdal, B. O. Ligand binding affinities from md simulations. *Accounts of Chemical Research* **35**(6), 358–365 (2002).
83. Srinivasan, J., Cheatham, T. E., Cieplak, P., Kollman, P. A., and Case, D. A. Continuum solvent studies of the stability of dna, rna, and phosphoramidate - dna helices. *Journal of the American Chemical Society* **120**(37), 9401–9409 (1998).

84. Gohlke, H., Kiel, C., and Case, D. A. Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the ras-raf and ras-ralgds complexes. *J Mol Biol* **330**(4), 891–913 (2003).
85. Wang, J., Wang, W., Huo, S., Lee, M., and Kollman, P. *J. Phys. Chem. B* **105**, 5055–5067 (2001).
86. Aqvist, J., Medina, C., and Samuelsson, J. E. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng* **7**(3), 385–91 (1994).
87. Zhou, R. H., Friesner, R. A., Ghosh, A., Rizzo, R. C., Jorgensen, W. L., and Levy, R. M. New linear interaction method for binding affinity calculations using a continuum solvent model. *Journal of Physical Chemistry B* **105**(42), 10388–10397 (2001).
88. Alder, B. J. and Wainwright, T. E. Phase transition for a hard sphere system. *Journal of Chemical Physics* **27**(5), 1208–1209 (1957).
89. Rahman, A. Correlations in motion of atoms in liquid argon. *Physical Review a-General Physics* **136**(2A), A405–& (1964).
90. McCammon, J. A., Gelin, B. R., and Karplus, M. Dynamics of folded proteins. *Nature* **267**(5612), 585–90 (1977).
91. MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiorcikiewicz-Kuczera, J., Yin, D., and Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry B* **102**(18), 3586–3616 (1998).
92. Karplus, M. and McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat Struct Biol* **9**(9), 646–52 (2002).
93. Wang, J., Wang, W., Kollman, P. A., and Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graph Model* **25**(2), 247–60 (2006).

94. Cooper, A. Thermodynamic fluctuations in protein molecules. *Proceedings of the National Academy of Sciences of the United States of America* **73**(8), 2740–2741 (1976).
95. Elber, R. and Karplus, M. Multiple conformational states of proteins - a molecular-dynamics analysis of myoglobin. *Science* **235**(4786), 318–321 (1987).
96. Frauenfelder, H., Sligar, S. G., and Wolynes, P. G. The energy landscapes and motions of proteins. *Science* **254**(5038), 1598–603 (1991).
97. Laio, A. and Parrinello, M. Escaping free-energy minima. *Proceedings of the National Academy of Sciences of the United States of America* **99**(20), 12562–12566 (2002).
98. Wenzel, W. and Hamacher, K. Stochastic tunneling approach for global minimization of complex potential energy landscapes. *Physical Review Letters* **82**(15), 3003–3007 (1999).
99. Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. Optimization by simulated annealing. *Science* **220**(4598), 671–680 (1983).
100. Daggett, V. Long timescale simulations. *Current Opinion in Structural Biology* **10**(2), 160–164 (2000).
101. Straub, J. E., Rashkin, A. B., and Thirumalai, D. Dynamics in rugged energy landscapes with applications to the s-peptide and ribonuclease-a. *Journal of the American Chemical Society* **116**(5), 2049–2063 (1994).
102. Elofsson, A. and Nilsson, L. How consistent are molecular dynamics simulations? comparing structure and dynamics in reduced and oxidized escherichia coli thioredoxin. *Journal of molecular biology* **233**(4), 766–80 (1993).
103. Auffinger, P. and Westhof, E. Rna hydration: Three nanoseconds of multiple molecular dynamics simulations of the solvated trna(asp) anticodon hairpin. *Journal of Molecular Biology* **269**(3), 326–341 (1997).
104. Caves, L. S., Evanseck, J. D., and Karplus, M. Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. *Protein Sci* **7**(3), 649–66 (1998).

105. Loccisano, A. E., Acevedo, O., DeChancie, J., Schulze, B. G., and Evanseck, J. D. Enhanced sampling by multiple molecular dynamics trajectories: carbonmonoxy myoglobin 10 microsecond $\alpha_0 \rightarrow \alpha(1-3)$ transition from ten 400 picosecond simulations. *J Mol Graph Model* **22**(5), 369–76 (2004).
106. Schulze, B. G. and Evanseck, J. D. Cooperative role of arg45 and his64 in the spectroscopic $\alpha(3)$ state of carbonmonoxy myoglobin: Molecular dynamics simulations, multivariate analysis, and quantum mechanical computations. *Journal of the American Chemical Society* **121**(27), 6444–6454 (1999).
107. Monticelli, L., Sorin, E. J., Tieleman, D. P., Pande, V. S., and Colombo, G. Molecular simulation of multistate peptide dynamics: A comparison between microsecond timescale sampling and multiple shorter trajectories. *Journal of Computational Chemistry* **29**(11), 1740–1752 (2008).
108. Law, R. J., Forrest, L. R., Ranatunga, K. M., La Rocca, P., Tieleman, D. P., and Sansom, M. S. P. Structure and dynamics of the pore-lining helix of the nicotinic receptor: Molecular dynamics simulations in water, lipid bilayers, and transbilayer bundles. *Proteins-Structure Function and Genetics* **39**(1), 47–55 (2000).
109. Schneider, M. J. and Feller, S. E. Molecular dynamics simulations of a phospholipid-detergent mixture. *Journal of Physical Chemistry B* **105**(7), 1331–1337 (2001).
110. Brown, S. P. and Muchmore, S. W. High-throughput calculation of protein-ligand binding affinities: Modification and adaptation of the mm-pbsa protocol to enterprise grid computing. *Journal of Chemical Information and Modeling* **46**(3), 999–1005 (2006).
111. Brown, S. P. and Muchmore, S. W. Rapid estimation of relative protein-ligand binding affinities using a high-throughput version of mm-pbsa. *Journal of Chemical Information and Modeling* **47**(4), 1493–1503 (2007).
112. Kohl, N. E., Emini, E. A., Schleif, W. A., Davis, L. J., Heimbach, J. C., Dixon, R. A., Scolnick, E. M., and Sigal, I. S. Active human immunodeficiency virus protease is required for viral infectivity. *Proceedings of the National Academy of Sciences of the United States of America* **85**(13), 4686–90 (1988).

113. Abdel-Meguid, S. S., Metcalf, B. W., Carr, T. J., Demarsh, P., DesJarlais, R. L., Fisher, S., Green, D. W., Ivanoff, L., Lambert, D. M., Murthy, K. H., and et al. An orally bioavailable hiv-1 protease inhibitor containing an imidazole-derived peptide bond replacement: crystallographic and pharmacokinetic analysis. *Biochemistry* **33**(39), 11671–7 (1994).
114. Wlodawer, A. and Vondrasek, J. Inhibitors of hiv-1 protease: a major success of structure-assisted drug design. *Annu Rev Biophys Biomol Struct* **27**, 249–84 (1998).
115. Todd, M. J., Semo, N., and Freire, E. The structural stability of the hiv-1 protease. *J Mol Biol* **283**(2), 475–88 (1998).
116. Zutshi, R., Franciskovich, J., Shultz, M., Schweitzer, B., Bishop, P., Wilson, M., and Chmielewski, J. Targeting the dimerization interface of hiv-1 protease: Inhibition with cross-linked interfacial peptides. *Journal of the American Chemical Society* **119**(21), 4841–4845 (1997).
117. Pettersson, K. and Gustafsson, J. A. Role of estrogen receptor beta in estrogen action. *Annu Rev Physiol* **63**, 165–92 (2001).
118. Cosman, F. and Lindsay, R. Selective estrogen receptor modulators: clinical spectrum. *Endocr Rev* **20**(3), 418–34 (1999).
119. Barrett-Connor, E., Cox, D. A., and Anderson, P. W. The potential of serms for reducing the risk of coronary heart disease. *Trends Endocrinol Metab* **10**(8), 320–325 (1999).
120. Yaffe, K., Sawaya, G., Lieberburg, I., and Grady, D. Estrogen therapy in postmenopausal women: effects on cognitive function and dementia. *Jama* **279**(9), 688–95 (1998).
121. Davidson, N. E. and Lippman, M. E. The role of estrogens in growth regulation of breast cancer. *Crit Rev Oncog* **1**(1), 89–111 (1989).
122. Zumoff, B. Does postmenopausal estrogen administration increase the risk of breast cancer? contributions of animal, biochemical, and clinical investigative studies to a resolution of the controversy. *Proc Soc Exp Biol Med* **217**(1), 30–7 (1998).

123. Clemons, M., Danson, S., and Howell, A. Tamoxifen ("nolvadex"): a review. *Cancer Treat Rev* **28**(4), 165–80 (2002).
124. Fabian, C. J. and Kimler, B. F. Selective estrogen-receptor modulators for primary prevention of breast cancer. *J Clin Oncol* **23**(8), 1644–55 (2005).
125. Kuiper, G. G., Enmark, E., Peltö-Huikko, M., Nilsson, S., and Gustafsson, J. A. Cloning of a novel receptor expressed in rat prostate and ovary. *Proc Natl Acad Sci U S A* **93**(12), 5925–30 (1996).
126. Couse, J. F., Lindzey, J., Grandien, K., Gustafsson, J. A., and Korach, K. S. Tissue distribution and quantitative analysis of estrogen receptor- α (er α) and estrogen receptor- β (er β) messenger ribonucleic acid in the wild-type and er α -knockout mouse. *Endocrinology* **138**(11), 4613–21 (1997).
127. Curtis Hewitt, S., Couse, J. F., and Korach, K. S. Estrogen receptor transcription and transactivation: Estrogen receptor knockout mice: what their phenotypes reveal about mechanisms of estrogen action. *Breast Cancer Res* **2**(5), 345–52 (2000).
128. Dupont, S., Krust, A., Gansmüller, A., Dierich, A., Chambon, P., and Mark, M. Effect of single and compound knockouts of estrogen receptors α (er α) and β (er β) on mouse reproductive phenotypes. *Development* **127**(19), 4277–91 (2000).
129. Pike, A. C., Brzozowski, A. M., Hubbard, R. E., Bonn, T., Thorsell, A. G., Engström, O., Ljunggren, J., Gustafsson, J. A., and Carlquist, M. Structure of the ligand-binding domain of oestrogen receptor β in the presence of a partial agonist and a full antagonist. *Embo J* **18**(17), 4608–18 (1999).
130. Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Montgomery, Jr., J. A., Vreven, T., Kudin, K. N., Burant, J. C., Millam, J. M., Iyengar, S. S., Tomasi, J., Barone, V., Mennucci, B., Cossi, M., Scalmani, G., Rega, N., Petersson, G. A., Nakatsuji, H., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Klene, M., Li, X., Knox, J. E., Hratchian, H. P., Cross, J. B., Bakken, V., Adamo, C., Jaramillo, J., Gomperts, R., Stratmann, R. E., Yazyev, O., Austin, A. J., Cammi, R., Pomelli, C., Ochterski, J. W., Ayala, P. Y., Morokuma, K., Voth, G. A., Salvador, P., Dannenberg, J. J., Zakrzewski, V. G., Dapprich, S., Daniels, A. D., Strain, M. C.,

- Farkas, O., Malick, D. K., Rabuck, A. D., Raghavachari, K., Foresman, J. B., Ortiz, J. V., Cui, Q., Baboul, A. G., Clifford, S., Cioslowski, J., Stefanov, B. B., Liu, G., Liashenko, A., Piskorz, P., Komaromi, I., Martin, R. L., Fox, D. J., Keith, T., Al-Laham, M. A., Peng, C. Y., Nanayakkara, A., Challacombe, M., Gill, P. M. W., Johnson, B., Chen, W., Wong, M. W., Gonzalez, C., and Pople, J. A. Gaussian 03, Revision C.02.
131. Zoete, V., Michielin, O., and Karplus, M. Protein-ligand binding free energy estimation using molecular mechanics and continuum electrostatics. application to hiv-1 protease inhibitors. *J Comput Aided Mol Des* **17**(12), 861–80 (2003).
 132. Wang, R., Fang, X., Lu, Y., and Wang, S. The pdbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J Med Chem* **47**(12), 2977–80 (2004).
 133. Wilderspin, A. F. and Sugrue, R. J. Crystallization and preliminary x-ray investigation of recombinant simian immunodeficiency virus proteinase. *J Mol Biol* **231**(4), 1139–42 (1993).
 134. Tong, L., Pav, S., Mui, S., Lamarre, D., Yoakim, C., Beaulieu, P., and Anderson, P. C. Crystal structures of hiv-2 protease in complex with inhibitors containing the hydroxyethylamine dipeptide isostere. *Structure* **3**(1), 33–40 (1995).
 135. Wu, J., Adomat, J. M., Ridky, T. W., Louis, J. M., Leis, J., Harrison, R. W., and Weber, I. T. Structural basis for specificity of retroviral proteases. *Biochemistry* **37**(13), 4518–26 (1998).
 136. Gustchina, A., Kervinen, J., Powell, D. J., Zdanov, A., Kay, J., and Wlodawer, A. Structure of equine infectious anemia virus proteinase complexed with an inhibitor. *Protein Sci* **5**(8), 1453–65 (1996).
 137. Laco, G. S., Schalk-Hihi, C., Lubkowski, J., Morris, G., Zdanov, A., Olson, A., Elder, J. H., Wlodawer, A., and Gustchina, A. Crystal structures of the inactive d30n mutant of feline immunodeficiency virus protease complexed with a substrate and an inhibitor. *Biochemistry* **36**(35), 10696–708 (1997).
 138. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. Charmm - a program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry* **4**(2), 187–217 (1983).

139. Thorsteinsdottir, H. B., Schwede, T., Zoete, V., and Meuwly, M. How inaccuracies in protein structure models affect estimates of protein-ligand interactions: computational analysis of hiv-i protease inhibitor binding. *Proteins* **65**(2), 407–23 (2006).
140. Brooks, C. L., Brunger, A., and Karplus, M. Active-site dynamics in protein molecules - a stochastic boundary molecular-dynamics approach. *Biopolymers* **24**(5), 843–865 (1985).
141. Brooks, C. L. and Karplus, M. Solvent effects on protein motion and protein effects on solvent motion - dynamics of the active-site region of lysozyme. *Journal of Molecular Biology* **208**(1), 159–181 (1989).
142. Lee, M. S., Salsbury, F. R., and Brooks, C. L. Novel generalized born methods. *Journal of Chemical Physics* **116**(24), 10606–10614 (2002).
143. Lee, M. S., Feig, M., Salsbury, F. R., and Brooks, C. L. New analytic approximation to the standard molecular volume definition and its application to generalized born calculations (vol 24, pg 1348, 2003). *Journal of Computational Chemistry* **24**(14), 1821–1821 (2003).
144. Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S. H., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D. A., and Cheatham, T. E. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Accounts of Chemical Research* **33**(12), 889–897 (2000).
145. Zoete, V., Meuwly, M., and Karplus, M. Study of the insulin dimerization: Binding free energy calculations and per-residue free energy decomposition. *Proteins-Structure Function and Bioinformatics* **61**(1), 79–93 (2005).
146. Hermann, R. B. Theory of hydrophobic bonding .2. correlation of hydrocarbon solubility in water with solvent cavity surface-area. *Journal of Physical Chemistry* **76**(19), 2754– (1972). N4716 Times Cited:650 Cited References Count:29.
147. Still, W. C., Tempczyk, A., Hawley, R. C., and Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *Journal of the American Chemical Society* **112**(16), 6127–6129 (1990).

148. Hasel, W., Hendrikson, T., and Still, W. Hasel w, hendrikson tf, still wc. a rapid approximation to the solvent accessible surface areas of atoms. *Tetrahedron Comput Methodol* **1**, 103–116 (1998).
149. McQuarrie, D. *Statistical mechanics*. Harper and Row, New York, (1976).
150. Tidor, B. and Karplus, M. The contribution of vibrational entropy to molecular association - the dimerization of insulin. *Journal of Molecular Biology* **238**(3), 405–414 (1994).
151. Lindahl, E., Hess, B., and van der Spoel, D. Gromacs 3.0: a package for molecular simulation and trajectory analysis. *Journal of Molecular Modeling* **7**(8), 306–317 (2001).
152. Frey, B. J. and Dueck, D. Clustering by passing messages between data points. *Science* **315**(5814), 972–6 (2007).
153. Reva, B., Antipin, Y., and Sander, C. Determinants of protein function revealed by combinatorial entropy optimization. *Genome biology* **8**(11), R232 (2007).
154. Kairys, V., Fernandes, M. X., and Gilson, M. K. Screening drug-like compounds by docking to homology models: a systematic study. *Journal of chemical information and modeling* **46**(1), 365–79 (2006).
155. McGovern, S. L. and Shoichet, B. K. Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. *J Med Chem* **46**(14), 2895–907 (2003).
156. DeWeese-Scott, C. and Moulton, J. Molecular modeling of protein function regions. *Proteins* **55**(4), 942–61 (2004).
157. Domingues, F. S., Rahnenfuhrer, J., and Lengauer, T. Conformational analysis of alternative protein structures. *Bioinformatics* **23**(23), 3131–8 (2007).
158. Nervall, M., Hanspers, P., Carlsson, J., Boukharta, L., and Aqvist, J. Predicting binding modes from free energy calculations. *J Med Chem* **51**(9), 2657–67 (2008).
159. Graves, A. P., Shivakumar, D. M., Boyce, S. E., Jacobson, M. P., Case, D. A., and Shoichet, B. K. Rescoring docking hit lists for model cavity sites: predictions and experimental testing. *J Mol Biol* **377**(3), 914–34 (2008).

160. Oostenbrink, C. and van Gunsteren, W. F. Free energies of ligand binding for structurally diverse compounds. *Proceedings of the National Academy of Sciences of the United States of America* **102**(19), 6750–4 (2005).

8 Appendix

Curriculum Vitae

Hólmfríður Berglind Þorsteinsdóttir

Nationality: Icelandic
Date of Birth: 21.06 1974
E-mail: holmfridur.thorsteinsdottir@unibas.ch

Education

- From 2003- **PhD student in Bioinformatics**, Computational and Systems Biology, Biozentrum, University of Basel, Switzerland.
Thesis title: Computational Analysis of Protein-Ligand Binding: From single continuous trajectories to multiple parallel simulations
- 2000-2001 **MSc in Bioinformatics**, Institute of Biomedical and Life Sciences, University of Glasgow, Scotland
Masters project: Sequence motifs and patterns, work performed at DeCODE genetics, Reykjavik, Iceland
- 1993-1997 **BSc in Biochemistry**, University of Iceland
Final year project: Apoptosis in various cell types, work performed at the Icelandic Cancer Society

Work Experience

- 2001-2003 **DeCODE genetics, Iceland**
Microarray database development
- 1999-2000 **DeCODE genetics, Iceland**
Affymetrix experimental development
- 1997-1999 **Lund University, Sweden**
Antigenic libraries, construction and selection
- 1996-1997 **Institute for Experimental Pathology, Iceland**
Antigenic variants in the maedi-visna virus
- 1996 **Faculty of Odontology, University of Iceland**
Quantification of acid, fluoride, calcium and phosphate in beverages.
- 1995 **Lyckeby Stärkelseförädling, Sweden**
An IAESTE summer student in Kristianstad Sweden.

Publications

Thorsteinsdottir HB, Podvinec M, Arnold K, Meuwly M and Schwede T. Application of GRID computing to molecular modeling of protein-ligand interactions using MM-GBSA. *Manuscript in preparation*.

Thorsteinsdottir HB, Schwede T, Zoete V and Meuwly M. How inaccuracies in protein structure models affect estimates of protein-ligand interactions: Computational analysis of HIV-I protease inhibitor binding. *Proteins* 2006; 65:407-23

Andresdottir V, Skraban R, Matthiasdottir S, Lutley R, Agnarsdottir G, and **Thorsteinsdottir H**. Selection of antigenic variants in maedi-visna virus infections. *Journal of General Virology* 2002; 83:2543-2551.

Agnarsdottir G, **Thorsteinsdottir HB**, Oskarsson T, Mattiasdottir S, Haflidadottir BS, Andresson OS and Andresdottir V. The Long Terminal Repeat is a Determinant of Cell Tropism of Maedi-Visna Virus. *Journal of General Virology* 2000;81:1901– 1905.

Ohlin M, Jirholt P, **Thorsteinsdottir H**, and Borrebaeck CA. Understanding Human Immunoglobulin Repertoires in vivo and Evolving Specificities in vitro. In: *The Antibodies* 1999, Volume 6, eds M. Zanetti and J.D. Carpa. Harwood Academic Publishers.

Ohlin M, Jirholt P, **Thorsteinsdottir HB**, Söderlind E and Borrebaeck CAK. Targeting CDR in synthetic antibody design. In: *Antibody Engineering*. New technology, Application and Commercialisation 1998. IBC UK Conferences, Inc.

Gudmundsson K, Holbrook WP, **Thorsteinsdottir H**. Acid, fluoride, calcium and phosphate in beverages. *Journal of Dental Research* 1998;77: 1340-1340

Ohlin M, Jirholt P, **Thorsteinsdottir H**, Lantto J, Lindroth Y, and Borrebaeck CAK. CDR-shuffling: targeting hypervariable loops for library construction and selection. In: *Proceedings of the 10th International Congress of Immunology* 1998, eds G.P Talwar, I. Nath, N.K. Ganguly, and K.V.S. Rao. Monduzzy Editore S.p.A., Bologna, pp. 1525-1529

Seminars & Posters

Swiss Chemical Society meeting, Zurich, Switzerland, September 2008. **Oral presentation**

Swiss Institute of Bioinformatics meeting, Grindelwald, Switzerland. September 2007.

Poster presentation.

ISBC meeting, Vienna, Austria. July 2007. **Poster presentation.**

GRID meeting, Basel, Switzerland. April 2006. **Oral presentation.**

ECCB meeting, Madrid, Spain. September 2005. **Poster presentation**

Swiss Chemical Society meeting, Lausanne, Switzerland. October 2005. **Poster presentation**

Swiss Institute of Bioinformatics meeting, Leysin, Switzerland, October 2004. **Oral presentation**

Swiss Chemical Society meeting, Zurich, Switzerland, October 2004. **Oral presentation**

9 Acknowledgments

I would like to thank my two supervisors, prof. Torsten Schwede and prof. Markus Meuwly for giving me the opportunity work as a PhD student in their groups. They provided me with a very interesting project and were always ready to guide and support me.

I also would like to say thanks to my colleagues from both groups for making a pleasant and fruitful working environment. In particular I would like to thank Poedi and Tobias for many scientific (and non-scientific) discussions, Jürgen for reading the thesis and Franziska for her valuable insights.

The Icelanders in Basel, in particular "FIBLin" (Hafdis, Gunni, Ragnhildur, Siggi and Mummi) and of course Elin my "foster mother" for being a family away from my family.

My friends in Basel, Pernilla, Maria, Rikke and especially Melle for all the help and support. And all my friends around the world, thanks for putting up with me for so long, I couldn't have done this without you.

Finally and especially I would like to thank my family for their endless support and love. Especially my brother Þröstur and my father who continues to teach me that you can overcome all obstacles with determination and optimism.