Binding Free Energy Calculations and Molecular Dynamics Studies on Complexes of Viral Proteases with their Ligands

Inaugural-Dissertation
zur
Erlangung des Doktorgrades
Dr. rer. nat.

der Fakultät für
Biologie
an der

Universität Duisburg-Essen

vorgelegt von
Oliver Anselm Kuhn

aus Würzburg

January 2013

Die der vorliegenden Arbeit zugrunde liegenden Experimente wurden am Zentrum für Medizinische Biotechnologie (ZMB) in der Abteilung für Bioinformatik der Universität Duisburg-Essen durchgeführt.

1. Gutachter: Prof. Dr. Daniel Hoffmann

2. Gutachter: Prof. Dr. Holger Gohlke

Vorsitzender des Prüfungsausschusses: Prof. Dr. Markus Kaiser

Tag der mündlichen Prüfung: 15.7.2013

Für Simon.

Ein großer Teil dessen, was Menschen
Intelligenz nennen, ist am Ende Neugierde.

- Aaron Swartz -

Zusammenfassung

Ein Ziel der biomolekularen Modellierung ist die Berechnung der Affinität ΔG von Liganden an Proteine, insbesondere Enzyme. Das Spektrum der Methoden, die zu diesem Zweck entwickelt wurden, reicht von theoretisch genauen aber aufwändigen Verfahren zu einfachen, eher qualitativen Verfahren. Während letztere häufig empirische Scoring-Funktionen und eine einzelne Struktur als Eingabe verwenden, wird für kompliziertere Methoden der möglichst vollständige Konformationsraum eines Protein-Ligand-Komplexes benötigt. Dieser wird mit Sampling-Verfahren wie der Molekulardynamik (MD) durchmustert.

In dieser Promotionsarbeit sollten Verfahren zur Berechnung von ΔG, insbesondere Varianten der Molecular Mechanics Poisson-Boltzmann Surface Area (MMPBSA) Methode, getestet und nach Möglichkeit weiterentwickelt werden. Desweiteren sollte die Auswirkung bestimmter Resistenzmutationen auf Struktur und Dynamik von Proteinen mit unterschiedlichen Maßen aus MD Simulationen heraus erfasst werden.

Der erste Schritt der quantitativen Modellierung mit MD ist die Beschreibung der Moleküle durch die Parametrisierung eines Kraftfelds. Anhand des sulfatierten Tyrosins wurde eine solche molekulare Parametrisierung für ein Nicht-Standard-Molekül durchgeführt. Sodann wurden Varianten der tendenziell weniger aufwändigen MMPBSA-Methode getestet im Hinblick auf ihre Konvergenz und ihre Eignung zur Bestimmung genauer ΔG-Werte oder zumindest verschiedene Enzym-Ligand-Komplexe in eine richtige Rangfolge gemäß ihrer ΔG-Werte zu bringen. Die Varianten unterscheiden sich durch verschiedene Solvatisierungsmodelle und Methoden zur Berechnung der Entropie. Als molekulares Referenzsystem wurden Mutanten der HIV-Protease im Komplex mit Wirkstoffen verwendet, da es hierzu experimentelle Daten gibt, mit denen die berechneten Werte verglichen werden können. Am anderen Ende des methodischen Spektrums liegt die aufwändige Thermodynamische Integration (TI). Bei einer guten Kraftfeldparametrisierung sollte TI in der Lage sein, ΔG-Effekte in der Größenordnung weniger kJ/mol quantitativ zu bestimmen. Dies wurde anhand der Mutante L76V der HIV-Protease, die für einige Wirkstoffe zu einer Resensitivierung (erhöhte Affinität) führt, getestet. Schließlich sollten MD-Simulationen verwendet werden, um die molekularen Effekte von Mutationen der NS3/4A-Protease des humanen Hepatitis C Virus auf die Bindung von Liganden (Substrat, Inhibitoren) zu verstehen.

Abstract

A major aim of biomolecular modelling is the calculation of binding affinities ΔG of ligands to proteins, especially enzymes. The spectrum of methods that has been developed for this task ranges from theoretically exact but expensive to more simple and qualitative ones. While the latter are often empirical scoring functions using one single structure as an input, the more complex methods require the preferably complete conformational space of a protein-ligand complex which can be sampled using methods such as molecular dynamics (MD).

The intention of this thesis was to test and further develop methods for the calculation of ΔG, in particular variants of the molecular mechanics Poisson-Boltzmann surface area (MMPBSA) method. Furthermore, the effects of specific resistance mutations on the structure and dynamics of proteins should be determined using different metrics on MD simulation data.

The first step to quantitative modelling using MD is the description of the molecules by parameterizing a forcefield. Such a molecular parameterization was performed for the non-standard amino acid sulpho-tyrosine. Subsequently, variants of the less expensive MMPBSA method were tested with regard to their ability to converge and determine ΔG estimates or at least establish the correct ranking of ΔG values for a set of enzyme-ligand complexes. Different solvation models and procedures to calculate the entropy have been used. As a molecular reference system, mutants of the HIV protease complexed with inhibitors were used. For these systems, experimental data are available to which the calculated values can be compared. At the other end of the methodological spectrum is the more expensive thermodynamic integration (TI). With a proper forcefield parameterization, TI should be able to quantitatively determine ΔG effects in the order of a few kJ/mol. This was tested on the HIV protease mutation L76V which is known to lead to a resensitivation (increased affinity) for some drugs. Eventually, MD simulations were used to understand the molecular effects of mutations of the NS3/4A protease, an enzyme of the human hepatitis C virus, on the binding of ligands (substrate, inhibitors).

Contents

Zusammenfassung i

Abstract ii

1 Introduction 1
- 1.1 HIV and HCV Epidemiology . . . 1
- 1.2 HIV Structure and Life Cycle . . . 2
- 1.3 Antiviral Drugs and Resistance . . . 3
- 1.4 Research Motivation . . . 5

2 Biomolecular Modelling 6
- 2.1 Molecular Mechanics . . . 6
  - 2.1.1 From Quantum to Molecular Mechanics . . . 6
  - 2.1.2 Molecular Dynamics . . . 7
  - 2.1.3 Empirical Forcefields . . . 8
  - 2.1.4 Explicit Water Models . . . 9
- 2.2 Continuum Solvation . . . 12
  - 2.2.1 Poisson-Boltzmann . . . 13
  - 2.2.2 Generalized Born . . . 14
  - 2.2.3 Nonpolar Solvation . . . 15
- 2.3 Sampling . . . 16
  - 2.3.1 Thermodynamic Ensembles . . . 17
  - 2.3.2 Multiple Independent Simulations . . . 17
  - 2.3.3 Rotatable Dihedral Accelerated Molecular Dynamics . . . 18
  - 2.3.4 Replica Exchange Molecular Dynamics . . . 18
  - 2.3.5 Free Energy Guided Sampling . . . 19
  - 2.3.6 Performance Gains from Hardware . . . 19
- 2.4 Conformational Entropy . . . 20
  - 2.4.1 Normal Model Analysis . . . 21
  - 2.4.2 Alternatives . . . 22
- 2.5 Trajectory Analysis . . . 22
  - 2.5.1 Root Mean Square Deviation . . . 22
  - 2.5.2 Root Mean Square Fluctuation . . . 22
  - 2.5.3 Distance and RMSF Analysis using SAM . . . 23
  - 2.5.4 Concerted Motions from a Distance Covariance . . . 24

3 Free Energy of Ligand Binding 26
- 3.1 Binding Affinity and Equilibrium . . . 26
- 3.2 Measurement Methods . . . 26
  - 3.2.1 Inhibition Constant K_i . . . 27
  - 3.2.2 Inhibitory Concentration IC_{50} . . . 27
  - 3.2.3 Isothermal Titration Calorimetry ΔG_{ITC} . . . 28

3.3 Employed Free Energy Methods 29
3.3.1 Thermodynamic Integration 29
3.3.2 MMPBSA 30

4 Systems and Applications **35**
4.1 Derivation of Sulphotyrosine Forcefield Parameters 35
4.1.1 Introduction 35
4.1.2 Methods 35
4.1.3 Results and Discussion 36
4.1.4 Conclusion 38
4.2 MMPBSA on HIV Protease Complexes 39
4.2.1 Introduction 39
4.2.2 Methods 39
4.2.3 Results and Discussion 42
4.2.4 Conclusion 47
4.3 L76V Thermodynamic Integration Calculation 48
4.3.1 Introduction 48
4.3.2 Methods 48
4.3.3 Results and Discussion 50
4.3.4 Conclusion 51
4.4 Molecular Dynamics Study on HCV Protease 52
4.4.1 Introduction 52
4.4.2 Methods 60
4.4.3 Results and Discussion 62
4.4.4 Conclusion 70

Future Directions **72**

References **73**

Publications **85**

Acknowledgments **86**

Declarations **87**

List of Abbrevations

AA	Amino Acid
AIDS	Aquired Immune Deficiency Syndrome
APV	Amprenavir - HIV protease inhibitor
ART	Antiretroviral Therapy
ATV	Atazanavir - HIV protease inhibitor
B3LYP	Becke three-parameter Lee-Yang-Parr
BAR	Bennett Acceptance Ratio
BILN2061	Boehringer Ingelheim 2061- HCV protease inhibitor
CCR5	CC motive chemokine receptor 5
CD4	cluster of differentiation 4 - glycoprotein
CD	Cavity Dispersion
CPU	Central Processing Unit
CXCR4	CXC motive chemokine receptor 4
DiCC	Distance Correlation Coefficient
DNA	desoxyribonucleic acid
DRV	Darunavir - HIV protease inhibitor
ESP	Electrostatic Potential
FDA	Food and Drug Administration
FDR	False Discovery Rate
FEGS	Free Energy Guided Sampling
GAFF	Generalized Amber Forcefield
GB	Generalized Born
GCC	Generalized Correlation Coefficient
gp120	glycoprotein 120
gp41	glycoprotein 41
GPU	Graphics Processing Unit
HAART	Highly Active Antiretroviral Therapy
HCV	Hepatitis C Virus
HF	Hartree-Fock
HIV	Human Immune Deficiency Virus
HIV-PR	HIV protease
IC50	Inhibitory Concentration
ITC	Isothermal Titration Calorimetry
MD	Molecular Dynamics
MEP	Molecular Electrostatic Potential
MMGBSA	Molecular Mechanics Generalized Born Surface Area
MM	Molecular Mechanics
MMPBSA	Molecular Mechanics Poisson Boltzmann Surface Area
mRNA	messenger RNA
NMA	Normal Mode Analysis
NMR	Nuclear Magnetic Resonance
NS3-4A	non-structural protein 3-4A (HCV protease)

PARSE	Parameters for Solvation Energy
PBC	periodic boundary conditions
PB	Poisson-Boltzmann
PCA	Principial Component Analysis
PCM	Polarizable Continuum Model
PDB	Protein Data Bank
PME	particle mesh Ewald
QHA	Quasi-Harmonic Analysis
QM	Quantum Mechanics
REMD	Replica Exchange Molecular Dynamics
RESP	Restrained electrostatic potential
RF	Resistance Factor
RISM	Reference Interaction Site Model
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
RNA	ribonucleic acid
SAM	Significance Analysis of Microarrays
SASA	Solvent Accessible Surface Area
SCF	self-consistent field
SQV	Saquinavir - HIV protease inhibitor
TI	Thermodynamic Integration
TYS	sulpho-tyrosine
VCC	Vector Correlation Coefficient
WHO	World Health Organization

1 Introduction

1.1 HIV and HCV Epidemiology

According to the world health organization (WHO), in 2010, an estimated 34 million people worldwide were living with AIDS and 2.7 million were newly infected. Due to the availability of antiretroviral therapy (ART) the number of people dying from AIDS-related causes could be reduced from 2.2 million in 2005 down to 1.8 million in 2010 [1].
170 Mio people worldwide are infected with HCV, and the number is expected to increase dramatically in the next decade. In most cases (60–85%) the HCV infection progresses to chronic liver disease and eventually to liver cirrhosis and hepatocarcinoma [2]. The presently applied combination therapy with pegylated interferon-α together with Ribavirin is costly, prolonged and it is associated with severe side effects [3]. This therapy is able to eradicate the virus in approximately 80% for genotype 2, but only 50% cases of genotype 1 infected patients. Unfortunately, 70-80% in the United States and more than 60% in Europe and Asian are infected with genotype 1 making the current standard of care unsatisfactory for many of these patients. There is hence an urgent need for additional and in particular directly acting antiviral agents that target specific stages in the viral life cycle.

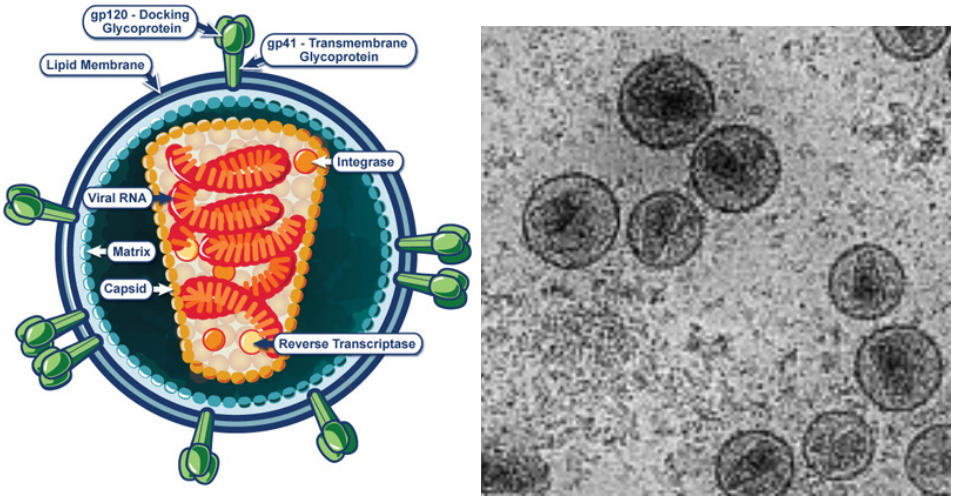


Figure 1.1: Structure of HIV. Left) Schematic representation of the HIV virion (source: National Institute of Allergy and Infectious Diseases (NIAID) [4]). Right) Cryo-electron microscopic image of mature HIV particles (source: Briggs et al. [5]).

1.2 HIV Structure and Life Cycle

The HIV virion particle is around 120 nm in diameter and has a roughly spherical shape. It has an outer coat, the viral envelope, that is composed of two lipid layers taken from the host cell membrane when a newly formed virus particle buds from the cell (figure 1.1). A various number of envelope proteins spike through its surface that consist of a cap made of three glycoproteins 120 (gp120) proteins and a stem of three glycoproteins 41 (gp41) which anchor the structure in the envelope. Inside that envelope, there is a cone-shaped core made up of roughly 2000 p24 proteins . Inside this core (or capsid), there are two copies of single-stranded RNA tightly bound to the nucleocapsid proteins p6 and p7 together with an arsenal of viral proteins needed for replication, most importantly the reverse transcriptase, integrase and protease.

A schematic representation of the HIV life cycle is depicted in figure 1.2 [6]. HIV specifically infects human T-cells, but other cells like macrophages or monocytes can also be used as hosts for viral replication [7]. The virion is directed towards the $CD4^+$ immune cells where its envelope proteins gp120 interact with the CD4 receptor. Additional interactions with one of the two chemokine receptors CCR5 or CXCR4 induce a conformational rearrangement in the HIV envelope leading to exposure of a hydrophobic domain on gp41 and the viral membrane in turn fuses with the cell membrane. The fusion process and viral entry induce uncoating of the viral core. Viral RNA and proteins are released into the cell, and the RNA is translated into DNA by the reverse transcriptase. The viral integrase inserts this DNA into the host cell genome. The cellular machinery initiates transcription into mRNA and produces the viral precursor proteins gag and gag-pol. In the following, these proteins diffuse to the cell membrane and the formation of new immature virus particles takes place. The precursor peptides are then cleaved at certain sites in a defined order by the HIV protease to yield mature virus particles.

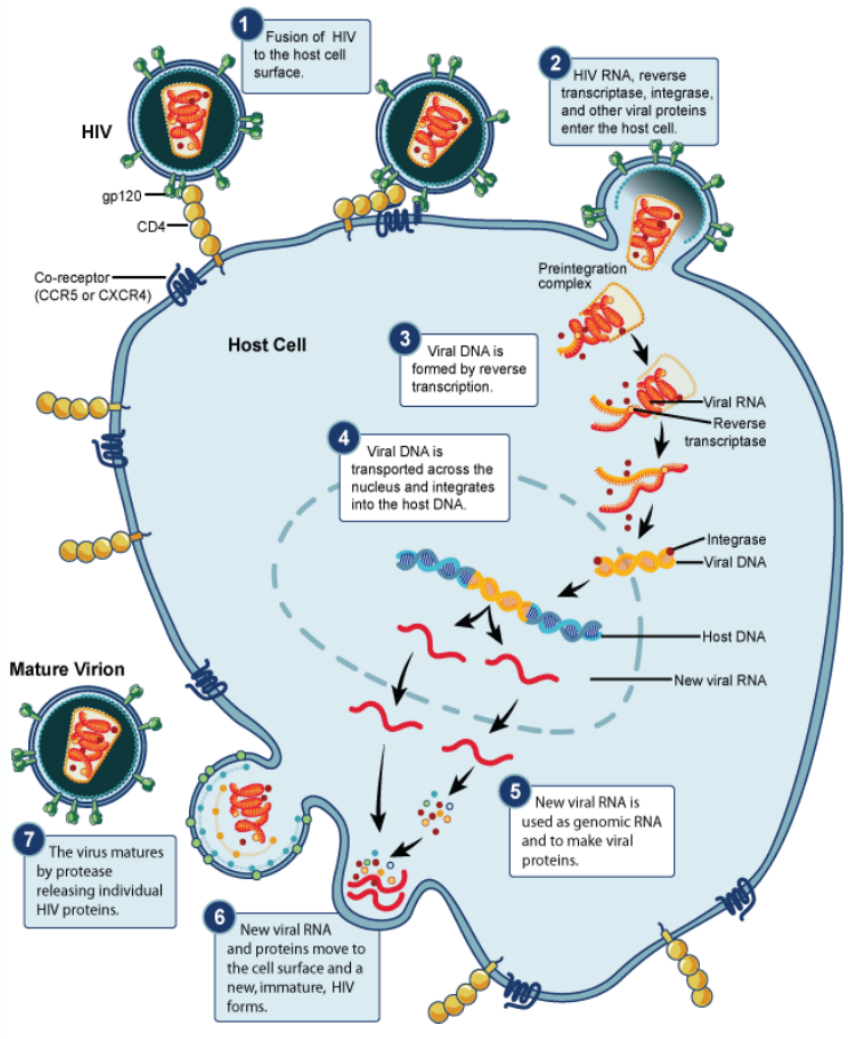


Figure 1.2: The HIV Lifecycle. Source: National Institute of Allergy and Infectious Diseases (NIAID) [4].

1.3 Antiviral Drugs and Resistance

Since the discovery of the human immunodeficiency virus (HIV) in 1983, several classes of drugs have been developed targeting viral entry, reverse transcription and in particular the maturation process by inhibiting the protease. The first protease inhibitor developed by Roche and approved by the food and drug administration (FDA) in 1995, saquinavir (SQV), mimics the intermediate state of the proteases natural substrate. Consequently, it inhibits protease activity by binding into the active site. Up to the present, nine protease inhibitors have been developed and this class of inhibitors con-

tinues to be the most effective class of HIV drugs. However, HIV recovers frequently its activity by developing resistance mutations. These mutations lead to a loss of drug binding affinity. It is therefore important to understand the structural mechanism of these resistance mutations to be able to construct new effective drugs. Since the introduction of the protease inhibitors, patients are treated with a combination of three drugs on average (highly active antiretroviral therapy, HAART) in order to avoid the development of resistant virus strains. On the annual Arevir Meeting in Bonn (www.genafor.org), clinicians and computer scientists meet to discuss most recent advances in the field of HIV diagnosis and therapy. One major topic there is the prediction of drug resistance that is used to guide clinicians with the compilation of their treatment regimes. These predictions are based on sequence data and different machine learning algorithms are adopted, e.g. decision trees [8], to predict the susceptibility to specific drugs from genotype. Whereas these systems serve very well for decision making, they have their limitations. In particular, these systems are knowledge-based and can therefore only predict mutations that have already been observed. Hence, resistance to newly developed drugs cannot be predicted without producing experimental data. These methods have recently been significantly improved by incorporating structural descriptors like the electrostatic potential in combination with hydrophobicity [9]. Because phenotypic resistance assays are time-consuming and costly, and genotypic rules-based interpretations may also fail to predict the effects of multiple mutations, a desirable goal of computational chemistry is the structure-based phenotype prediction [10], where the binding affinity of established drugs to the protease is calculated from a model based on available crystal structures.

The variety of resistance mechanisms is large. The major resistance mutations are typically located directly in the binding cavity with effects like contact losses, steric clashes, or alteration of hydrophobic clusters [11]. Other resistance mutations are more far off the active site and influence inhibitor binding by indirect geometric rearrangements or changes in flexibility [12]. Also more special mechanisms exist like the weakening of dimerization energy of the two HIV protease monomers [13].

The emergence of resistance is a trade-off of ligand binding loss and the balance of substrate processing efficiency. A paradigm that emerged over the last 10 years is the substrate envelope hypothesis [14]. Its message is that inhibitors that protrude from the substrate envelope, a representation of the consensus volume of the proteases natural substrates, are markedly more prone to the emergence of resistance mutations than inhibitors that stay within the envelope. The hypothesis has been proven to explain several major resistance mutations for both HIV and HCV proteases [15, 16] and it is already in use as a design paradigm for new protease inhibitors [17] that will be less prone to resistance.

1.4 Research Motivation

Experiments for the discovery of new drugs are very time-consuming and expensive tasks [18]. Especially, the chemical synthesis of new drug candidates can be very laborious. It is therefore desirable to have computational methods that can substitute for at least some of these experiments. Generally, the accurate and efficient calculation of the free energy of drug-target binding is a major goal of computational chemistry. Simple approximate computational methods, that are able to distinguish very weak binders from those that are possibly strong binders, are already in commercial use. It is however not possible to determine the binding affinity reliably without a full dynamical picture of a protein-drug complex. Hence, as computer power increases and forcefields become more accurate, molecular dynamics is a promising tool for drug design and can be expected to be used on a routine basis in future times [19, 20, 21]. For the development of a new drug, it would also be helpful to know the structural details of binding such as entropic or enthalpic energetics, van der Waals and electrostatic interactions. The molecular mechanics Poisson-Boltzmann surface area approach (MMPBSA) can in principle provide a residue-wise energetic decomposition giving clues for drug modifications that improve binding or make it less prone to resistance mutations. This has already quite often been done in the literature [22, 23, 24, 25].

The major concern of this thesis is to further improve and establish new strategies to understand resistance mutations on an energetic and mechanistic basis using molecular dynamics simulations.

2 Biomolecular Modelling

In the following chapter, I explain the theoretical concepts that are utilized within this thesis. The intention is to present a consistent picture in general. Therefore, some concepts are treated only in brief while other concepts of particular relevance are explained in higher computational detail.
Starting from quantum mechanics and its approximations, concepts of molecular mechanics and dynamics will be explained, followed by a section on advanced sampling strategies. Since needed for MMPBSA type free energy calculations, continuum solvation models and the calculation of conformational entropy based on normal modes are also explained in more detail.

2.1 Molecular Mechanics

To calculate medically relevant macroscopic properties one requires a proper physical description of the molecular system. The description has to be detailed enough to yield accurate quantitative results and must be computationally feasible.

2.1.1 From Quantum to Molecular Mechanics

In principle, the state of any molecular system is exactly described by a wave function ψ that satisfies the time-independent Schrödinger equation

$$H\psi = E\psi \tag{2.1}$$

where E is the energy and the Hamilton operator H is a structured operator describing the system in a formal fashion. It contains functions of the electronic and nucleic coordinates. This equation has analytical solutions only for simple systems such as the hydrogen atom and its solution has to be approximated for systems with more than a few atoms.
A commonly used simplification, the Born-Oppenheimer approximation, assumes that the motions of the electrons are directly coupled to the motions of the nuclei. This is reasonable because the mass of an electron is more than three orders of magnitude smaller than the mass of a nucleus. The kinetic energy of the electrons is therefore neglected.
A further simplification assumes that every single electron moves in the average field of all other electrons producing a self-consistent field (SCF). This is described by the Hartree-Fock equation. It can be used to calculate the molecular electrostatic potential (MEP). The MEP is used to calculate partial atomic point charges for empirical forcefields used in molecular dynamics simulations. Practically, the Hartree-Fock method is suited to approximate the Schrödinger equation for at most some hundreds of atoms.
Since it is not feasible to calculate the electronic structure for macromolecules with high accuracy by means of any type of quantum chemistry calculations,

most practical simulations use a set of simple classical functions to represent the energy, adjusting a large number of parameters to optimize agreement with experimental data and with quantum calculations on smaller molecules. The system can then be described with Newtonian mechanics.

2.1.2 Molecular Dynamics

The time evolution of a molecular system can be simulated by integrating Newtons equation of motion

$$\vec{F_i} = \frac{\delta V}{\delta \vec{r_i}} = m_i \frac{d^2 \vec{r_i}}{dt^2} \tag{2.2}$$

where m_i and $\vec{r_i}$ are the mass and position of particle i, respectively, F_i is the force acting on particle i and V is the potential energy function of the system. The potential energy is determined by the force field (subsection 2.1.3). The integration is carried out in a step-wise fashion where the time is discretized into time steps usually in the range of 1-2 femtoseconds. Several algorithms exist for this integration procedure [26]. In the Amber Molecular Dynamics Software, the leapfrog algorithm is used [27]. In this algorithm, velocities v at time $t + \frac{1}{2}\Delta(t)$ are calculated first and the positions r at time $t + \Delta(t)$ are calculated from these velocities.

$$\begin{aligned} v(t + \frac{1}{2}\Delta(t)) &= v(t - \frac{1}{2}\Delta(t)) + a(t)\Delta t \\ r(t + \Delta(t)) &= r(t) + v(t + \frac{1}{2}\Delta t)\Delta t \end{aligned} \tag{2.3}$$

In this way the velocities *leap* over the positions, then the positions *leap* over the velocities.

Periodic Boundary Conditions A molecular system is typically simulated in a box of some ten thousands of water molecules. The question arises how to handle the borders of this box. A solution to that problem is the periodic boundary condition (PBC) [28]. It uses an infinitely tileable neutrally charged box such that each box interacts on each side with the opposite side of its image. Any box shape that tiles space infinitely is possible. The truncated octahedron has the additional advantage that the number of water molecules is reduced compared to a cubic box. PBC is typically used in conjunction with Ewald summation methods such as the particle mesh Ewald method (PME) [29]. PME separates the pairwise particle-particle-interactions into a short range and a long range part. The short range part sums quickly in real space whereas the long range part is treated in its Fourier transform with the charge density discretized on a grid.

One single value, typically in the range of 8-12 Å, is set to define the cut-off between short and long range for both van der Waals and electrostatic interactions. Long-range van der Waals interactions are estimated by a continuum model. Care has to be taken that the size of the simulation box is large enough such that the simulated macromolecule does not interact with its own image in a neighboring box.

2.1.3 Empirical Forcefields

For a mechanical description of a molecular system, a potential energy function of the atomic coordinates is needed with a defined functional form and parameters. Because the forces on individual atoms are related to the gradient of this function, the potential function is commonly referred to as "forcefield" [30]. Different groups of forcefields exist. The most widely used are Gromos [31], Amber [32], Charmm [33] and OPLS [34]. Forcefields are parameterized on different characteristics, e.g. the Gromos forcefield is trimmed to accurately model solvation effects while the Amber forcefields are parameterized against ab initio data.
In this thesis, the two popular Amber forcefields ff99SB [35] and ff03 [36] have been used for proteins and the generalized Amber forcefield (GAFF) [37] for ligand molecules. The functional form of the Amber forcefield is

$$\begin{aligned} U(r) = & \sum_{\text{bonds}} k_b(l-l_0)^2 \\ & + \sum_{\text{angles}} k_a(\theta-\theta_0)^2 \\ & + \sum_{\text{torsions}} \frac{1}{2}V_n[1+\cos(n\omega-\gamma)] \\ & + \sum_{i}^{N}\sum_{j>i}^{N-1}\left\{\epsilon_{i,j}\left[\left(\frac{r_{0ij}}{r_{ij}}\right)^{12} - 2\left(\frac{r_{0ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}\right\} \end{aligned} \tag{2.4}$$

As can be seen, the potential contains relatively simple functions describing the different kinds of interatomic forces. The first three summations are over bonds (1-2 interactions), angles (1-3 interactions) and torsions (1-4 interactions) (figure 2.1). The first two are modelled by simple harmonic oscillators just like usual mechanical ball and spring models. The torsion term can also include so-called "improper torsions" where not all 4 interaction partners are connected via covalent bonds. Torsions are particularly important for correct protein dynamics and also hard to parameterize. The last summation models pairwise non-bonded interactions by a 6-12-Lennard-Jones potential

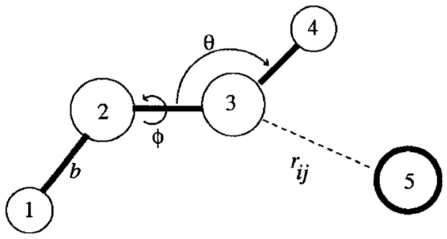


Figure 2.1: Schematic view of forcefield interactions. Scheme taken from Ponder et al. [30]. Heavy solid lines indicate covalent bonds, the light dashed line indicates nonbonded interactions.

depending on atomic radii and distances and an electrostatic Coulomb potential depending on partial atomic charges and distances. It iterates over all pairs of atoms that are separated by more than three covalent bonds and has usually special parameters for 1-4 interactions. The Lennard-Jones potential has a dispersion and a exchange repulsion component and is often called the "van der Waals" term.

The parameterization of a forcefield is a complex task and will not be discussed here at length. In the Amber 99 forcefield, van der Waals terms were adapted from fits to amide crystal data and liquid-state simulations. Force constants and equilibrium bond lengths and angles are taken from crystal structures and adapted to match normal mode frequencies for a number of peptide fragments. Generally, one goal of the parameterization is to obtain a balanced interaction between solute-water and water-water energies. Interestingly, fitting charges to the potentials at the HF/6-31G* level tends to overestimate bond-dipoles by amounts comparable to that in empirical water models such as SPC/E or TIP3P. Hence, fitting charges at the HF/6-31G* level tends to yield charges that are roughly consistent with these water models [30]. Charges for the Amber forcefields are derived by fitting atom-centered point charges to a quantum-chemically calculated electrostatic potential on the Connolly surface. This procedure has also been used for the calculation of sulpho-tyrosine partial atomic charges in subsection 4.1.
Both, the Amber ff99SB and ff03 forcefield, have recently been benchmarked and found to be particularly well performing in reproducing experimental residual dipolar couplings [38]. Other NMR observables and conformational populations of dipeptides available from vibrational spectroscopy have also been used for benchmarking [39]. Generally, the performance of fixed charge force fields is inherently limited, and polarizable forcefields such as Amoeba [40] are promising but not widely used yet.

2.1.4 Explicit Water Models

Hydration water at biomolecular surfaces plays a key role in protein-ligand interactions and enzymatic function [41]. At present, the most commonly used explicit solvent models treat the individual water molecules as rigid

bodies. The simplest models involve three interaction sites according to the positions of the oxygen atom and the two hydrogen atoms (figure 2.2).

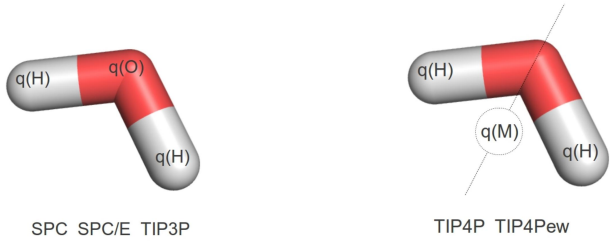


Figure 2.2: Geometries of 3-point (left) and 4-point (right) explicit water models. The interaction site in the 4-point model that lies on the bisector of the H-O-H angle is usually called the M-site.

They have a negative charge q on oxygen and positive charges $-q/2$ on the hydrogens. In the first transferable intermolecular potential (TIP) model, TIPS3 [42], the dimerization energy for two water molecules m and n is modelled as the sum of intermolecular charge-charge interactions and a Lennard-Jones term between the oxygens [43] as given in equation 2.5.

$$E_{mn} = \sum_{i}^{on\,m} \sum_{j}^{on\,n} \frac{q_i q_j}{r_{ij}} + \frac{A}{r_{OO}^{12}} - \frac{C}{r_{OO}^{6}} \tag{2.5}$$

The parameters (table 2.1) were optimized to give reasonable structural and energetic results for gas phase complexes of water and liquid water. Berendsen reparameterized the same model more thoroughly for liquid water yielding the single point charge model (SPC) [44]. The extended SPC model (SPCE) adds an average polarization correction to the potential energy function [45].

	q	r(OH)	r(OM)	α(HOH)	A	C
SPC [44]	-0.8200	1.0	-	109.47	2.6171e-3	2.6331e-6
SPCE [45]	-0.8476	1.0	-	109.47	2.6171e-3	2.6331e-6
Tip3P [43]	-0.8340	0.9572	-	104.52	2.4889e-3	2.4352e-6
Tip4P [42]	-1.0400	0.9572	0.15	104.52	2.5543e-3	2.5145e-6
Tip4P-Ew [46]	-1.0484	0.9572	0.125	104.52	2.7361e-3	2.7470e-6

Table 2.1: Geometry and potential function parameters for some important water models. The table lists the respective oxygen charge q, the lengths of O-H bonds r(OH), the displacement length of the M-site r(OM), the H-O-H angle α and Lennard-Jones constants A ($kJ\,\mathring{A}\,12\,mol^{-1}$) and C ($kJ\,\mathring{A}\,6\,mol^{-1}$).

Generally, all parameters (charges, OH-distances and Lennard-Jones constants) are optimized to reproduce several different types of experimental data, e.g. density, radial distribution functions, the enthalpy of vaporisation, heat capacity, diffusion coefficient and dielectric constant. However, it is not possible to satisfy all empirical restraints with one parameter set. For example, with the 3-point models, it is not possible to optimize the computed density without loosing the second peak in the O-O radial distribution function (not shown). Thus, 4-point models with an additional interaction site have been introduced, e.g. TIP4P [43]. In this model, the negative charge is moved off the oxygen towards the hydrogens at a point on the bisector of the HOH angle that is usually called the M-site (figure 2.2). The TIP4P-Ew is reparameterized to perform better in periodic boundary simulations [46]. There are also more complicated models, e.g. models incorporating 5 or 6 interaction sites, flexible water models including bond stretching and angle bending conformational changes of the single water molecules and polarizable water models. These are of course computationally more expensive than the simpler ones and are not yet widely used.
The performance of explicit water models has been assessed in several papers [47, 41]. According to van der Spoel et al. [47], SPCE seems to be the perfect model for bulk water simulations. But the authors add that they have made experience were SPCE gave dubious results in protein simulations. Thus, for the latter, they would prefer SPC. Vega et al. [48] show that from the simple water models (SPC, SPCE, TIP3P, TIP4P and TIP5P) only TIP4P provides a qualitatively correct phase diagram of water. A study of hydration thermodynamic properties [41] comes to the conclusion that for nearly all hydration properties considered, SPCE water performs best. Any of the models is parameterized against something specific and has therefore also its specific weakness. For example, SPC and TIP3P water models tend to underestimate the extent of water structuring close to hydrophobic groups, while the TIP4P-Ew model tends to yield overstructured hydration shells with too low enthalpies and entropies and too high heat capacities.
The most widely applied water models are the SPC/SPCE models and TIP3P in the Gromacs and Amber MD communities, respectively. Because the Amber forcefields are parameterized against TIP3P [49] and it is the most efficient among the TIP models, TIP3P is the default model in the Amber 11 software package. TIP3P water has been used throughout the present study.
Most recently, rigid body water potentials have been evaluated on their relevance for general purpose [50]. In this extensive test, the models have been inspected on their ability to reproduce a set of 17 phenomenological water properties. Although no model reproduces all properties, some models, in particular TIP4P/2005 [48] and SPCE perform better than the others. However, there is at present no evidence showing one water model to be clearly superior for simulations of proteins in water.

2.2 Continuum Solvation

Because usually approximately 80% of the molecules in a simulation box are water molecules for which all the interactions have to be calculated in each integration step, it is desirable to have an approximate model that is able to sufficiently represent the properties of water. Furthermore, the calculation of thermodynamic averages from explicit water models is limited by the need of integrating over the many solvent degrees of freedom. Because explicit solvent cannot be sampled sufficiently, the polar contribution to the solvation free energy is often calculated with continuum models. Continuum solvation models approximate the behavior of individual water molecules by representing the solvent as a dielectric continuum instead of individual (water and ion) molecules.

Solvation Energy A biomolecule's interaction with the solvent environment is a major determinant of its structure, dynamics and energetics. Consequently, solvation plays a crucial role in the energetics of ligand binding.

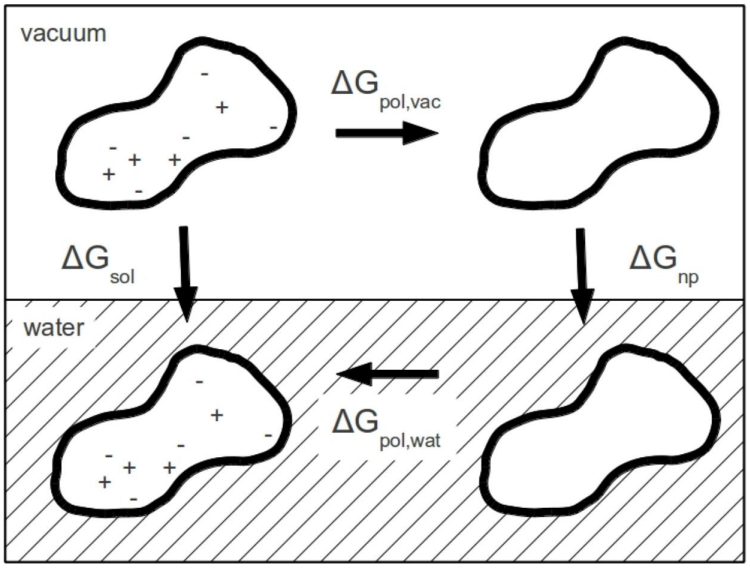


Figure 2.3: Thermodynamic cycle showing the breakdown of the solvation energy into electrostatic and nonpolar contributions. Figure is adapted from Sitkoff et al. [51].

Solvation free energy is the free energy difference of a molecule being in vacuum or in water. The process of bringing a biomolecule from vacuum into water can be envisioned as a three-step process (figure 2.3) [51]:

1. Discharging the solute in vacuo

2. Transferring the, now considered totally nonpolar, solute into water
3. Recharging the solute in water

Thus, the solvation energy can be evaluated as a sum of these single energies. The polar parts are usually treated together.

$$G_{sol} = G_{pol,vac} + G_{np} + G_{pol,wat} = G_{pol} + G_{np} \tag{2.6}$$

2.2.1 Poisson-Boltzmann

The Poisson Boltzmann model (PB) is based on the Poisson equation that allows to calculate the electrostatic potential directly from the molecular charge density in a homogeneous medium

$$\nabla \left[\epsilon(r)\nabla\phi(r)\right] = -4\pi\rho(r) \tag{2.7}$$

where $\phi(r)$ is the electrostatic potential, $\rho(r)$ is the charge density and $\epsilon(r)$ the permittivity. To account also for the impact of point charges (ions) in nonhomogeneous media, the Boltzmann part has to be introduced to yield the Poisson-Boltzmann equation

$$\nabla \left[\epsilon(r)\nabla\phi(r)\right] = -4\pi\rho^f(r) - 4\pi \sum_i c_i^\infty z_i q \lambda(r) e^{\frac{-z_i q\phi(r)}{kT}} \tag{2.8}$$

where $\rho^f(r)$ includes now only molecular charges, c_i is the concentration of ion i at an infinite distance from the molecule, z_i is its valency, q is the proton charge, k is the Boltzmann constant, T is the temperature and $\lambda(\vec{r})$ describes the accessibility to ions at point $\vec{r}$ [52]. $\epsilon(r)$ is discontinuous along the biomolecular surface and assumes solute dielectric values inside and bulk solvent values outside the surface (figure 2.4). The biomolecular surface is defined as the solvent excluded surface using Bondi radii [53].

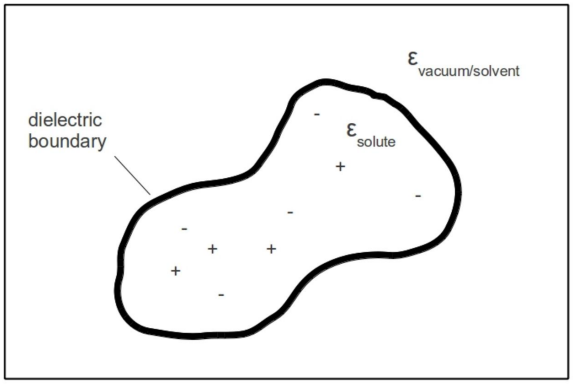


Figure 2.4: Definition of the dielectric boundary.

It is not straightforward to set the protein dielectric constant. An internal protein dielectric of 1 can be used. This means that there is no account for atomic polarizability. 2 would be a reasonable choice because the refraction index of most biomolecules is roughly 1.4 and the dielectric constant is the square of the refraction index. There is also experimental justification to use a dielectric of 4 [54].
When the Poisson equation began to become relevant for the electrostatics of biomolecules, the latter where first approximated as bodies of regular shapes with equal surface area, e.g. spheres. Numerical solutions for the Poisson-Boltzmann equation have made it possible to solve the Poisson-Boltzmann equation for arbitrary shapes. Different variants of the finite difference method are used for that. Generally, the molecular charges and dielectric are discretized on a grid [52]. To eventually evaluate the polar solvation energy of a molecule, the reaction field has to be calculated from electrostatic potentials both in solvent ($\epsilon = 80$) and in vacuum ($\epsilon = 1$) [52]. In the grid-based approach, this simplifies to a discrete formula.

$$G_{pol} = \frac{1}{2}\int_V \rho(r)\phi_{reac}(r)dV = \frac{1}{2}\sum_i q_i(\phi_{sol}(r_i) - \phi_{vac}(r_i)) \tag{2.9}$$

These calculations are computationally demanding both in CPU time and memory [54].

2.2.2 Generalized Born

The generalized Born model (GB) has become popular for MD applications because of its relative simplicity and computational efficiency [55]. The model is based on the formula of Max Born for the solvation energy of single ions [56]

$$\Delta G_{Born} = -\frac{q^2}{2a}(1 - \frac{1}{\epsilon_w}) \tag{2.10}$$

where q is the ion charge, a the radius and ϵ_w the solvent dielectric constant. Generalizing the Born model, a molecule is modelled consisting of charges $q_1 ... q_N$ embedded in spheres of radii $a_1 ... a_N$, and the polar solvation free energy can be given by a sum of individual Born terms and pairwise Coulombic terms [57]:

$$\Delta G_{pol} = \sum_i^N \frac{q_i^2}{2a_i}(\frac{1}{\epsilon_w} - 1) + \frac{1}{2}\sum_i^N \sum_{j\neq i}^N \frac{q_i q_j}{r_{ij}}(\frac{1}{\epsilon_w} - 1) \tag{2.11}$$

GB theory tries to resemble equation 2.11 by parameterizing an effective

function f_{GB}. The polar solvation energy then reads

$$\Delta G_{pol} = -\frac{1}{2}(\frac{1}{\epsilon_w} - 1)\sum_{ij}\frac{q_i q_j}{f_{GB}(r_{ij}, R_i, R_j)} \tag{2.12}$$

with f_{GB} being most often of the form

$$f_{GB}(r_{ij}) = [r_{ij}^2 + R_i R_j exp(-r_{ij}^2/4R_i R_j)]^{\frac{1}{2}} \tag{2.13}$$

where R are the so-called effective Born radii. These reflect the degree of burial of an atom inside the molecule. As they depend on the protein conformation, they must in principle be reevaluated at any time step in a MD simulation.
Different approaches to the calculation of the effective functions and radii exist. Several of them are implemented in the Amber software package, namely GB^{HCT} [58], the original Hawkins, Cramer, Truhlar approach with parameters by Tsui and Case, GB^{OBC} I and II [59], an improvement from Onufriev et al. with two different parameterizations intended to be a closer approximation to the true molecular volume and GBn [60], a model proposed by Mongan et al. introducing a more robust volume correction.

Although GB is an approximation to PB, it has been experienced that it performs slightly better on calculating solvation free energies on small drug-like molecules [61]. This is possible if the numerical inaccuracy in the solution of the PB model is larger than the approximation introduced by the assumptions of the GB model. For that reason, GB is not exclusively interesting for efficiency reasons and has therefore been included in the MMPBSA study in this thesis work.

2.2.3 Nonpolar Solvation

The nonpolar contribution to the solvation energy is composed of the cost of creating a cavity within the solvent to accomodate the solute together with nonpolar interactions (dispersion and exchange-repulsion) between the solute and solvent molecules.

Hermann et al. [62] found out, that the number of water molecules that can be packed around a solute molecule correlates well with the transfer energy of nonpolar molecules into water. A slightly more idealized measure is the cavity surface which contains the centers of the water molecules in the first layer around the solute. The logarithm of the transfer energy of alkanes into water correlates linearly with the alkanes surface areas [62]. Based on this finding, a model has been developed [51]

$$G_{np} = \gamma SASA + b \tag{2.14}$$

where γ and b are empirical parameters fitted to experimental solvation energies. For the calculation of the $SASA$, a new empirical atomic radii set, the parameters for solvation energy (PARSE) [51], has specifically been developed.
Other, more advanced approaches exist, namely the cavity-dispersion method (CD) [63] and the polarizable continuum model (PCM) [64, 65]. These have not been used within this study and are therefore not explained. The accuracies of nonpolar solvation methods have recently been investigated, revealing substantial qualitative differences between the different models [66].

Solvent-Accessible Surface Area (SASA) A molecule is defined as set of atoms with radii R_i and positions $r_i(x, y, z)$ (figure 2.5).

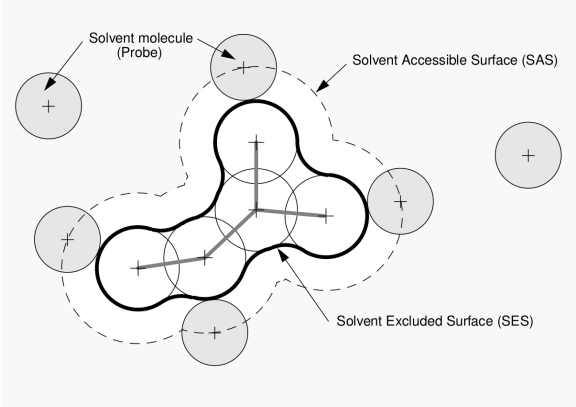


Figure 2.5: Definition of the solvent-accessible surface area.

The molecular surface is scanned by a probe sphere of a radius corresponding to a solvent molecule (1.4 Å for water). The surface defined by the center of the sphere is the solvent-accessible surface.

A general problem with implicit solvent models in the context of calculating protein-ligand binding affinities is that these models assume the ligand binding cavity to be filled with water in the free state. This may be simply incorrect for very small cavities or at least a very crude approximation because the hydration structure might be of particular importance [67].

2.3 Sampling

Physical macroscopic properties such as the free energy of binding cannot be accurately calculated from a single structure alone. Theoretically, a representative Boltzmann-averaged statistical ensemble is needed. However, conventional MD simulations in the canonical ensemble tend to get trapped

in local minima that are separated by high energy barriers which are rarely passed. Using conventional MD, the conformational space of a solvated protein system cannot sufficiently be explored in a practical amount of time. Thus, more advanced techniques have to be applied to overcome this problem. Many different approaches have been devised. One crucial criterion that differentiates these methods is if the Hamiltonian is changed during simulation or not. Most of these strategies such as REMD are mainly interesting to sample long time scale motions and are thus not particularly suited to enhance sampling for e.g. MMPBSA calculations. In the following, I will explain the thermodynamic ensembles together with the corresponding algorithms to obtain these ensembles and give a short overview on a subset of interesting MD-based sampling algorithms.

2.3.1 Thermodynamic Ensembles

NVT In the canonical ensemble (NVT), the number of particles, the volume and the temperature are constant. The temperature, that is defined by the ensemble average of kinetic energies, is kept constant by a thermostat that exchanges energy with the environment. Several approaches exist to perform the temperature control, mainly the Berendsen [68], Langevin [69] and Nosé-Hoover [70, 71] thermostats. The Berendsen thermostat, also called "coupling to an external heat bath", controls the temperature by rescaling the particle velocities by a constant factor. Nosé-Hoover is an integral type of thermostat. It introduces the heat bath in terms of additional degrees of freedom into the Hamiltonian of a system. The Langevin thermostat, that has been used throughout this thesis, regulates the temperature by adding random frictional forces from a Gaussian distribution to adjust the kinetic energy of the particles. The collision frequency controls the magnitude of these forces. Simulations performed with a higher collision frequency (5 ps^{-1} compared to 1 ps^{-1}) have small but noticeable higher relative RMSD (compared to the starting structure) than simulations performed with a lower collision frequency.

NPT The isothermic-isobaric ensemble (NPT) corresponds more to laboratory conditions, i.e. ambient (constant) temperature and pressure. The natural environment is able to compensate the volume change of a molecular system. Therefore, in addition to a thermostat, a barostat is needed. The pressure control is usually accomplished by extending and shrinking the simulation box volume by isotropic position scaling (Berendsen barostat).

2.3.2 Multiple Independent Simulations

One way to enhance sampling is to do multiple simulations using varying initial conditions. A number of quite arbitrary choices has to be made when

setting up MD simulations. The effects of some of these choices on the sampling diversity has been investigated by Genheden et al. [72]. From these considerations, three important possibilites to enhance sampling by using multiple simulations can be suggested:

1. **VIIT** - velocity induced independent trajectories
2. **CIIT** - conformation induced independent trajectories
3. **SIIT** - solvent induced independent trajectories

In VIIT, different initial random velocities are used, CIIT uses different initial conformations if these are available from given crystal structures, and SIIT uses different initial waterbox configurations. Combinations of these approaches have been used for the MMPBSA calculations and MD studies in this thesis. A justified criticism to this approach is that short simulations will not sample energy minima that are further away from the starting conformation on the potential energy landscape. These more distant energy minima are more probable to be reached by one long simulation. However, simulations with different initial conditions tend to propagate into different directions overcoming some energy barriers right in the beginning of the simulations. This is evident from the fact, that calculated MMPBSA energies from multiple simulations converge to distinct values in a broad range and they do not converge to one same value in an appropriate amount of simulation time. Consequently, the total amount of energy barriers crossed could be larger with the multiple short simulation setup. This is an interesting question that could be tested computationally.

2.3.3 Rotatable Dihedral Accelerated Molecular Dynamics

In accelerated molecular dynamics (aMD) [73], a bias potential is added to the original potential such that the potential energy wells are elevated whereas the transition state regions remain unaltered. It has been shown that a correct canonical distribution can be extracted by applying a reweighing scheme. The method has recently been shown to dramatically improve sampling providing routine access to millisecond events [74]. In the more recently presented RaMD approach [75], only the rotatable dihedrals that are most responsible for conformational changes of biomolecules are subjected to aMD yielding an accuracy comparable to aMD while greatly improving the efficiency.

2.3.4 Replica Exchange Molecular Dynamics

In the replica exchange method (REMD) [76], also called “parallel tempering”, a number of MD simulations (replicas) are started at different temperatures. Based on a decision made by a Metropolis criterion, the temperatures

between trajectories are exchanged at certain times. Exchanging conformations between high and low temperature replicas avoids the trapping in low energy conformations which is the typical problem of conventional MD. The pairs of trajectories that are exchanged need to have significantly overlapping energy distributions. REMD is particularly suited to investigate slow dynamic phenomena such as protein folding. It is on the other hand probably not suited for free energy calculations such as MMPBSA since we there want to sample locally around an equilibrium state. It is furthermore not straightforward to obtain a Boltzmann ensemble from REMD [77]. It has also been stated, that REMD has so far not shown to be efficient and has further theoretical problems, e.g. the entropic part of the temperature dependent transition rate is unaltered by increasing the temperature [78].

2.3.5 Free Energy Guided Sampling

Another promising approach is the free energy guided sampling algorithm (FEGS) developed by Zhou et al. [78]. FEGS explores conformational space using unbiased MD simulations, e.g. no collective variables or reaction coordinates are needed. In a first exploration stage, multiple short simulations are iteratively restarted from regions of the free energy surface that are visited rarely, thus increasing the probability for passing high energy barriers. In a second refinement stage, multiple independent runs are initiated from Boltzmann distributed conformations to yield an overall Boltzmann distributed ensemble. A mean first passage time cutoff can be used to control the kinetic range of sampling.

2.3.6 Performance Gains from Hardware

Developers of several MD software packages, in particular OpenMM and Amber, are trying to use the power of graphic processing units (GPU) to accelerate MD simulations. In April 2011, Amber 11 was released [79] with a GPU version of pmemd, the Amber MD routine. It uses CUDA and runs exclusively on NVIDIA cards. I have tested the implementation first on GTX295 and then on Tesla c1070 cards at the informatics department of the university at Münster and observed remarkable speedups (figure 2.6).

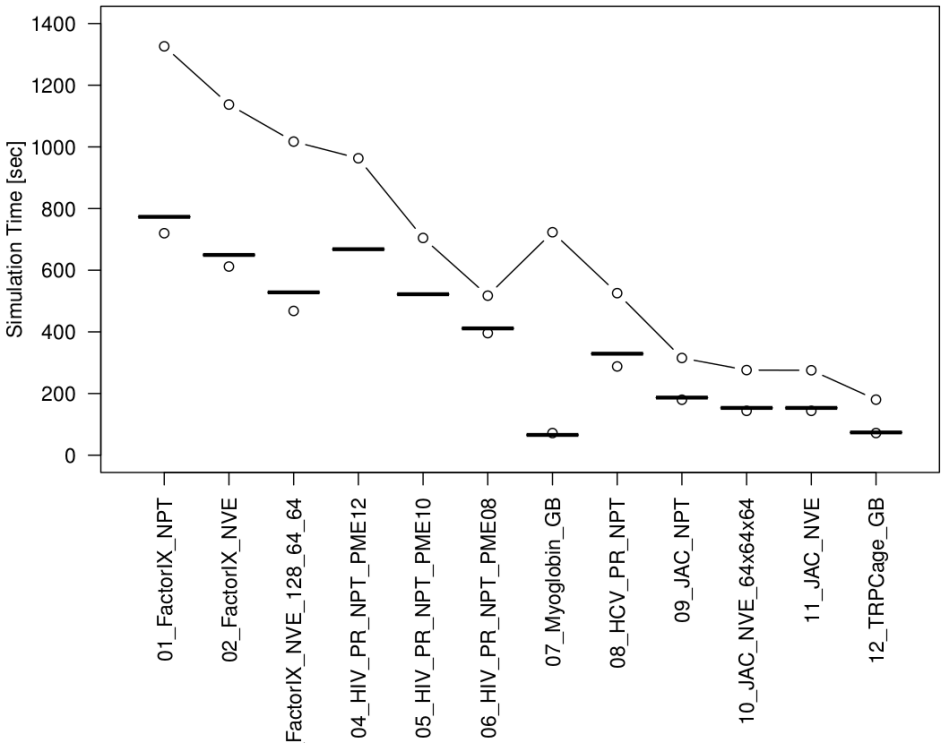


Figure 2.6: Performance gains using GPUs. Y-axis shows computer run time needed to simulate 20 ps of each system shown on the x-axis with an 8-CPU compute node (connected dots), a GTX295 GPU (bars) and a Tesla c1070 GPU (dots).

In August 2011, a new pmemd GPU version was released with a major code change (bugfix 17) with again substantial performance improvements and the 2012 version has support for some newer graphic cards [80]. At present, for typical production MD simulations, one compute node with two Tesla M2050 easily outperforms ten 8-CPU compute nodes (Xeon E5440 2.83GHz) while these components have roughly the same price of approximately 6500€. This is a crucial advance that should be noticed by anyone who is interested to buy hardware specifically for MD simulations (ambermd.org/gpus).

2.4 Conformational Entropy

To calculate the binding affinity by MMPBSA (equation 3.14), the conformational entropy has to be evaluated. The conformational entropy of a protein is the sum of translational, rotational and vibrational entropies.

$$S_{conf} = S_{trans} + S_{rot} + S_{vib} \tag{2.15}$$

While the first two terms are simply calculated with statistical mechanics

formulas alone [81], the latter needs vibrational frequencies ν as an input:

$$S_{vib} = R \sum_{i=1}^{3N-6} (\frac{h\nu_i}{k_B T} \frac{1}{e^{h\nu_i/k_B T} - 1} ln(1 - e^{h\nu_i/k_B T})) \tag{2.16}$$

where ν_i are the vibrational frequencies v_i, h and k are Planck and Boltzmann constants, respectively and T the absolute temperature. Within the MMPBSA framework, these frequencies are commonly calculated from normal modes.

2.4.1 Normal Model Analysis

The normal modes of a molecule are obtained as follows:
The Hessian matrix H contains the second partial derivatives of the potential energy function (equation 2.4). For a system of N atoms, it is a $3N \times 3N$ matrix.

$$H(r) = \left[\frac{\partial^2 V(r)}{\partial r_i \partial r_j}\right] = \begin{bmatrix} \frac{\partial^2 V(r)}{\partial x_1 \partial x_1} & \frac{\partial^2 V(r)}{\partial x_1 \partial y_1} & \cdots & \frac{\partial^2 V(r)}{\partial x_1 \partial z_n} \\ \frac{\partial^2 V(r)}{\partial y_1 \partial x_1} & \frac{\partial^2 V(r)}{\partial y_1 \partial y_1} & \cdots & \frac{\partial^2 V(r)}{\partial y_1 \partial z_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 V(r)}{\partial z_n \partial x_1} & \frac{\partial^2 V(r)}{\partial z_n \partial y_1} & \cdots & \frac{\partial^2 V(r)}{\partial z_n \partial z_n} \end{bmatrix} \tag{2.17}$$

The normal modes are eigenvectors of the Hessian matrix. The eigenvectors together with their corresponding eigenvalues are obtained by diagonalization of the Hessian. Each normal mode is then treated as an harmonic oscillator, thus, the frequencies of the harmonic normal modes (entering equation 2.16) can be expressed as

$$\nu_i = \frac{1}{2\pi} \sqrt{\frac{\lambda_i}{\mu}} \tag{2.18}$$

where λ_i are normal mode force constants and μ is the reduced molecular mass.

To put it more descriptive, NMA estimates the conformational entropy by measuring the widths of the potential energy wells of each MD snapshot. To do so, it assumes these wells to be harmonic as an approximation. To make the harmonic assumption valid, the molecule conformation must be minimized as close as possible to the true minimum of the nearest potential well [82, 83]. Otherwise, substantial error can occur. To this end, prior to the frequency calculation, a second-derivative based minimization approach

(Newton-Raphson) is usually applied. Both, geometric optimization and NMA are time-consuming and computer-memory demanding tasks.

2.4.2 Alternatives

The frequencies for the calculation of the vibrational entropy can in principle be taken from different sources, e.g. from principal components from the covariance matrix of the MD simulation. This approach is called quasi-harmonic analysis (QHA) [84]. Another approach is the Schlitter entropy that turned out to be an approximation which yields the result of the QHA in its upper limit [85, 86]. Another recent approximate and fast approach is to calculate the entropy directly from the $SASA$ [87].

2.5 Trajectory Analysis

2.5.1 Root Mean Square Deviation

A rough indicator for the stability of MD simulations is the root mean squared deviation (RMSD)

$$RMSD_j = \sqrt{\frac{1}{N}\sum_{i}^{N}(x_{j,i} - x_{R,i})^2} \tag{2.19}$$

where N is the number of atoms and x are atomic coordinates. The RMSD that is typically reported in this context is the minimized RMSD between each trajectory frame j and a reference structure R, most commonly the initial structure. The alignment algorithm by Kabsch and Sander [88] (together with an additional translation step) is used to perfectly align a set of chosen atoms by translation and rotation to minimize the RMSD. The selected group of atoms for the alignment may vary. Commonly used are $C\alpha$ atoms only, backbone atoms only or all atoms. Practically, the resulting RMSD differences among these variants are however marginal.

2.5.2 Root Mean Square Fluctuation

The root-mean square fluctuation (RMSF) is a measure for thermal motion. It is analogous to and often correlates well with experimental b-factors. It is based on the atom coordinates deviations from their time-average. Therefore, aligning all frames to the average structure is required prior to the RMSF calculation. Residue fluctuations can be based on the $C\alpha$ atom coordinates or the positions of the residues center of mass. The RMSF for each residue i is given by

$$RMSF_i = \sqrt{\frac{1}{T}\sum_{j}^{T}(x_i(t_j) - \bar{x}_i)^2} \tag{2.20}$$

where x are positional coordinates and T the number of simulation frames.

2.5.3 Distance and RMSF Analysis using SAM

To detect significant structural differences induced by point mutations, it is straightforward to analyze certain features such as e.g. average distances. Angles and dihedrals could also be taken into account but these can be expected to give redundant information. Additionally, changes in flexibility can be tracked by analyzing either the variances of distances or residue root mean square fluctuations. The most naive approach to detect significant differences of such features is an ordinary student's t-test on each individual feature. This is however problematic when testing a very large number of features. If one intends to test for example for differences of 100 features at a significance level of 0.05, then five false positives are expected even if there are no significant differences. This is a common problem of multiple testing. One typical means to handle this problem is the Bonferroni correction that corrects the significance level to adapt it to a large number of features. This method however still simply assumes the features to be independent from each other and it is more suited for a relatively small number of features (around 20) [89]. One of the most advanced statistical tools to handle the multiple testing problem is, besides the global rank test [90], the significance analysis of microarrays (SAM). The method has been developed to detect significant genes from microarray expression experiments where a large number of false positives cannot be tolerated. It was first applied to microarray data from investigations on the ionizing radiation response by Tusher et al. [91]. There it was compared to the Bonferroni correction (and one of its adaptations allowing for dependent tests by Westfall and Young [92]) and shown to give more useful results in this test case.

Significance Analysis of Microarrays (SAM) SAM has been implemented by Schwender in the R package siggenes [89, 93]. The procedure is described there and it goes as follows:

The input data is an $m \times n$ matrix comprising the expression values of m genes and n observations with a corresponding response (or class) vector of length n. B is the number of permutations to estimate the null distribution.

- For each gene i, $i = 1, ..., m$, the value d_i of a statistic appropriate for testing (moderated t-test or Wilcox) is computed.
- The null distribution $d^0_{(i)}$ is estimated by computing the m permuted d-statistics d_{ib} for each permutation b, $b = 1, ..., B$ of the n values of the response, and $d^0_{(i)}$ is set to $\sum_{b=1}^{B} d_{(i)b}/B$. The determination of a

correct theoretical null distribution is crucial for the determination of the FDR.

- For a set D of discrete positive tresholds Δ, an upper and a lower cutoff is determined. These cutoffs define a set of genes S_Δ called differentially expressed.
- As an error estimate, the false discovery rate (FDR) for each S_Δ is determined.

Figure 2.7 shows a typical SAM output.

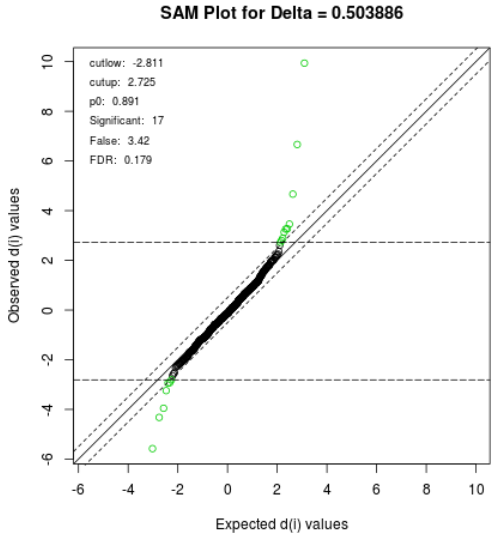


Figure 2.7: SAM example plot showing the expected versus observed d(i) values from the moderated t-statistic. Horizontal lines show the chosen cutoffs that define the set of genes called differentially expressed (green).

The described approach has been used for the identification of significant distance and RMSF differences induced by point mutations of the HCV protease. How the method was used on this data is more specifically described in subsection 4.4.

2.5.4 Concerted Motions from a Distance Covariance

In a recent work, the potential of a distance correlation coefficient (DiCC) to detect long-range correlated fluctuations has been investigated [94]. DiCC was compared to other variants of correlation coefficients, namely the displacement vector correlation coefficient (VCC) that is an extension to the

Pearson correlation coefficient for positional vectors and a generalized correlation coefficient (GCC). It has been shown that long-distance concerted motion could be observed with DiCC that could not be revealed with the other correlation coefficients. Likewise, the atom-by-atom correlation matrix exhibits a more pronounced profile than the matrices from VCC and GCC (figure 2.8).

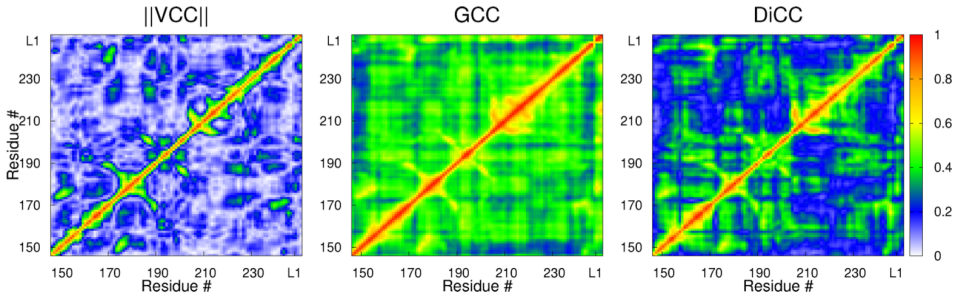


Figure 2.8: Comparison of Correlation Coefficients. Correlation matrices from VCC, GCC and DiCC between C_α atoms of Src SH2 domains. (Source: Roy et.al. [94])

The algorithm for the calculation of the DiCC is described in Roy et al. [94]. In this thesis, DiCC was used to investigate if a certain point mutation, in particular the HCV protease mutation I71T, induces changes in such long-range concerted motions. DiCC differences were investigated using the SAM method (subsection 4.4).

3 Free Energy of Ligand Binding

A drug is a small molecule that exerts its action by binding to a target protein. This protein is very often an enzyme. Most commonly, as in the case of the HIV protease, the drug molecule inhibits the enzyme competitively, which means that it blocks the binding of the physiological substrate to the active site. If the drug is a tight binder, many of the target proteins will be occupied at one instant of time at chemical equilibrium. Any small molecule that binds to a receptor protein is usually termed a ligand but the term is actually used more broadly for any molecule that binds another one.

3.1 Binding Affinity and Equilibrium

In an aqueous solution with fixed concentrations of protein and ligand at thermal equilibrium, there is constant complex formation and dissociation. In the non-covalent case, the complex formation of a ligand L and a protein P can then be formulated by the following chemical reaction:

$$P + L \underset{k_{dissociate}}{\overset{k_{associate}}{\rightleftharpoons}} PL \tag{3.1}$$

where $k_{associate}$ and $k_{dissociate}$ are the formation and dissociation rates, respectively. The reaction is governed by the binding constant K_B given by the ratio of concentrations of the free and bound species:

$$K_B = \frac{k_{associate}}{k_{dissociate}} = \frac{[PL]}{[P][L]} \tag{3.2}$$

The Gibbs free energy of binding is related to K_B by the equation

$$\Delta G = -RT \ln(K_B) \tag{3.3}$$

where R is the general gas constant and T the absolute temperature.

3.2 Measurement Methods

Several useful methods exist to determine the binding affinity of a ligand to a protein experimentally. I will explain here the three most relevant quantities: the inhibition constant K_i, the inhibitory concentration IC_{50} and the free energy of binding calculated from isothermal titration calorimetry, ΔG_{ITC}.

Enzyme inhibition assays measure the inhibition constant K_i or the inhibitory concentration IC_{50}. For enzyme inhibitors, the dissociation constant K_d, the reciprocal of the binding constant, is usually termed the inhibition constant. These assays are convenient on one hand because the enzyme acts as an amplifier and carries out many reactions that can be measured. On the other hand, a protein-specific procedure has to be established to detect

the release of products [95].

3.2.1 Inhibition Constant K_i

In the most simple setup, the K_i can be determined by measuring the fraction f of free protein (spectroscopically or based on its enzyme activity) as a function of the concentration of free ligand [95].

$$f = \frac{[P]_{eq}}{[PL]_{eq} + [P]_{eq}} = \frac{1}{1 + K_B[L]_{eq}} \tag{3.4}$$

3.2.2 Inhibitory Concentration IC_{50}

In the typical clinical setup of resistance testing, drug efficiency is measured phenomenologically. Cell cultures with increasing drug concentrations are infected with a resistant recombinant virus. For comparison, further cell cultures are infected with non-resistant (wildtype) virus in the same manner. The IC_{50} is the drug concentration that reduces the cell culture activity to 50% of that of the reference culture [96]. One option for the measurement of cell culture activity are so-called indicator viruses. In the case of HIV, the *n*ef gen, which is not needed for viability in immortalized T cell lines, is substituted by an indicator gen (e.g. β-galactosidase or luciferase). During replication, the viruses produce the corresponding indicator protein instead of the *n*ef protein. The amount of this protein can be determined by ELISA or luminometric testing and correlates directly to the number of viruses [96]. The IC_{50} depends strongly on the cell lines used. It is therefore reasonable to unite the IC_{50} values of wildtype and mutant in one parameter, the resistance factor (RF).

$$RF = \frac{IC_{50}^{WT}}{IC_{50}^{MT}} \tag{3.5}$$

The resistance factor (RF), that expresses how much more medicament is needed to decrease the mutant virus growth by 50% relative to the wildtype is directly related to the relative Gibbs free energy of binding.

$$\Delta\Delta G = \Delta G_{WT} - \Delta G_{MT} \approx -kT \ln(RF) \tag{3.6}$$

The measured IC_{50} is generally expected to be greater than K_i, but formally, the IC_{50} is related to the K_i via the Cheng-Prusoff equation [97]

$$K_i = \frac{IC_{50}}{1 + \frac{[L]}{K_m}} \tag{3.7}$$

where $[S]$ is a fixed substrate concentration and K_m is the substrate concentration with half-maximal enzyme activity. The equation implies that the IC_{50} becomes essentially equal to K_i for very low substrate concentrations. Corresponding ΔG values are obtained from enzyme assay data via the relations

$$\begin{aligned} \Delta G &= -RT \ln K_i \\ &\approx -RT \ln IC50 \end{aligned} \tag{3.8}$$

3.2.3 Isothermal Titration Calorimetry ΔG_{ITC}

The most exact method to determine ΔG is isothermal titration calorimetry (ITC). It is also the most expensive method because much more enzyme is needed for one measurement. However, the procedure is more generic and does not require any system specific adaptations.
A solution with a fixed enzyme concentration is kept at constant temperature by a thermostat. Ligand is then steadily injected and its binding to the enzyme leads to a temperature change of the solution. The amount of energy that is needed to keep the temperature constant can be accurately measured and calculated back to the binding free energy of one ligand and one enzyme molecule.

A thoroughly managed collection of binding affinities is administrated by Gilson and coworkers [98] available at www.bindingdb.org. Binding DB currently contains about 620000 binding data for 5500 proteins and over 270000 drug-like molecules.

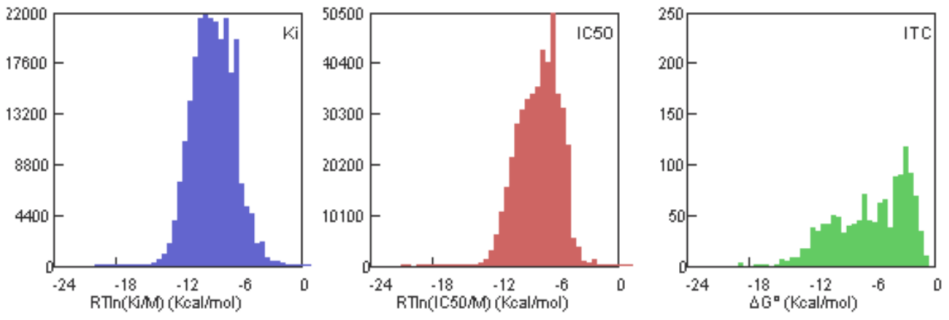


Figure 3.1: Distribution of available binding affinity measures from BindingDB.

The uncertainties of ΔG measures are not always given and they are usually around 0.4 kcal/mol [99].

3.3 Employed Free Energy Methods

Biomolecular modelling concepts can be applied to calculate the free energy of protein-ligand binding. Numerous concepts can be used to calculate free energies and in the following, I briefly explain the two I have used in this thesis.

3.3.1 Thermodynamic Integration

The thermodynamic integration method (TI) is a rigorous (theoretically exact) method to calculate relative binding affinities entirely based on statistical mechanics [100]. In this thesis, the relative free energy of binding between wildtype HIV protease bound to atazanavir and the mutant L76V has been calculated (subsection 4.3) and is used here as illustrative example. Instead of calculating the difference of absolute binding free energies of two protein-ligand complex variants, the relative binding affinity is calculated by estimating the energy differences of the alchemical transformations of one variant into the other in the bound and unbound state (figure 3.2).

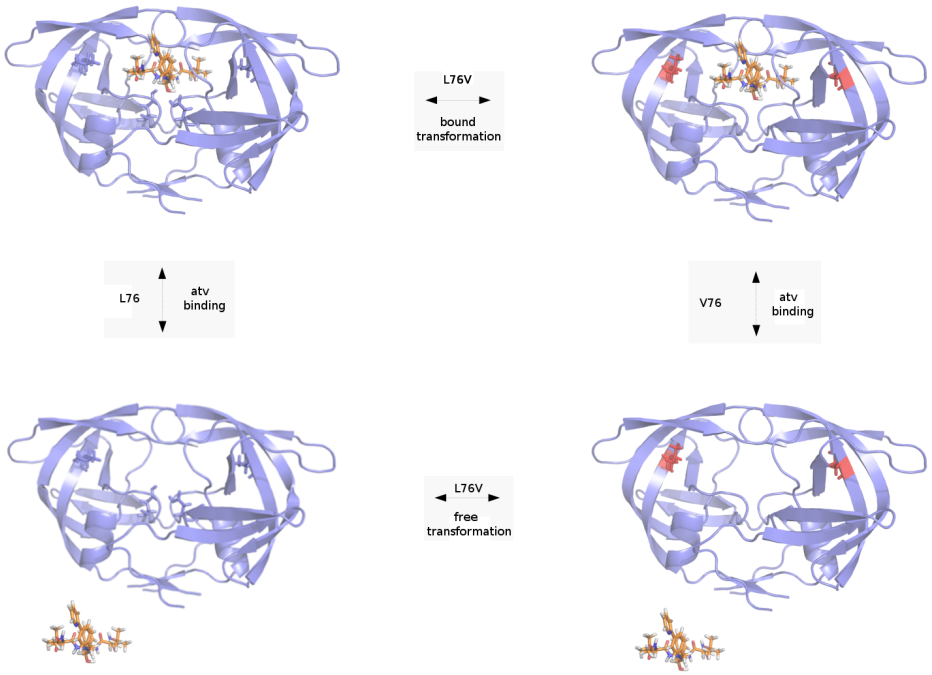


Figure 3.2: Thermodynamic cycle illustrating the TI formalism. The transformations of bound and unbound state can be substituted by alchemical transformations from wildtype to mutant. The point mutation is colored red.

The free energy difference between two λ-coupled states according to the TI formalism is

$$\Delta G_{TI} = \int_0^1 \langle \frac{\delta V(\lambda)}{\delta \lambda} \rangle_\lambda \, d\lambda \tag{3.9}$$

In this equation V is the λ-coupled potential function corresponding to V_{L76} for λ=0 and V_{76V} for λ=1. Since this equation can practically not be solved analytically for such a complex system, an integration scheme is used that numerically determines the value of the integral from simulations at discrete λ values. This scheme allows for efficient parallelization and also for additional λ values in regions where convergence has not been sufficiently achieved.
A simple way to couple the two end-point potential functions into the mixed potential V(λ) is

$$V(\lambda) = f(\lambda)V_1 + [1 - f(\lambda)]V_0 \tag{3.10}$$

with $f(\lambda) = \lambda$ yielding a linear coupling of the potential functions.
The two transformations, L76 to 76V, in the bound and the unbound state are furthermore split into three transformation steps. In the first step, charges on L76 are turned off, in the second step, van der Waals parameters are transformed and in a third step, 76V charges are turned on. The details of the implementation are given in subsection 4.3.

3.3.2 MMPBSA

A popular method to calculate the free energy of protein ligand binding is the molecular mechanics Poisson-Boltzmann surface area method (MMPBSA) pioneered by Massova and Kollman [101, 102, 103]. It combines molecular mechanics with continuum solvation models and an estimate for the free energy difference that results from a change in the configurational entropy. Molecular dynamics simulation is used to sample the configurational space of the protein ligand complex and the energy estimate is calculated as an average over an equally spaced subset of snapshots. In terms of its theoretical framework, MMPBSA is designed to calculate absolute binding free energies. However, since the calculated energies strongly depend on the solvation model that is used, only relative energies can be considered useful. In contrast to alchemical free energy calculations, MMPBSA is a so-called end-point method meaning that no intermediate states have to be taken into account reducing the states that have to be simulated to the bound and the unbound state. For that reason, MMPBSA is thought to be more efficient than alchemical calculations. Furthermore, the use of MMPBSA is not limited to the calculation of free energy differences of structurally similar systems. The basic idea can be illustrated by a thermodynamic cycle (figure 3.3).

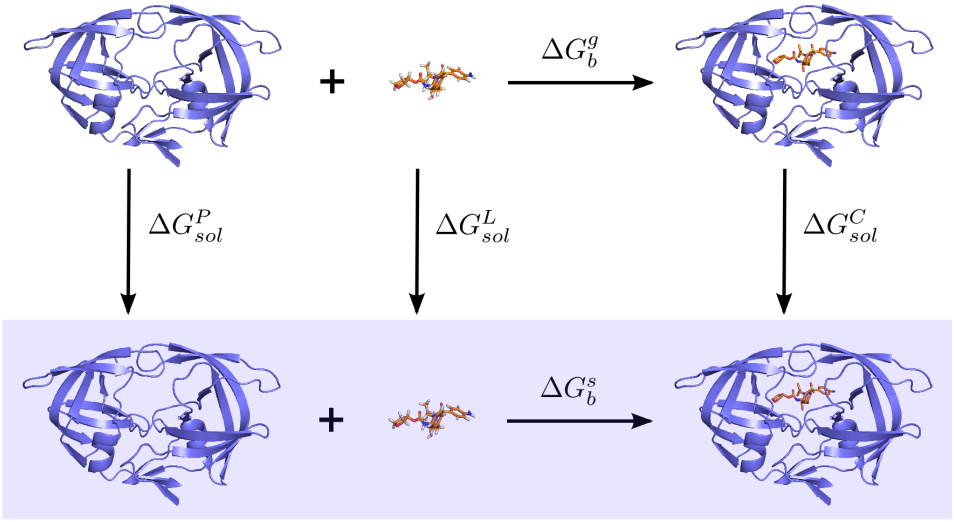


Figure 3.3: Thermodynamic cycle for the calculation of the binding free energy of a protein P and a ligand L forming a complex C in solution, ΔG_b^s. Molecules in the blue box are considered solvated. ΔG_b^g is the binding free energy in the gas phase, ΔG_{sol}^P, ΔG_{sol}^L and ΔG_{sol}^C the solvation energies of the protein, ligand and complex, respectively.

Given that the energy differences between states in a thermodynamic cycle by definition sum up to zero, the absolute Gibbs free energy of binding in solution ΔG_b^s can be calculated as a sum of the gas phase binding energy ΔG_b^g and the difference of solvation energies of the bound (C) and free (P,L) species:

$$\Delta G_b^s = \Delta G_b^g + \Delta G_{sol}^C - (\Delta G_{sol}^P + \Delta G_{sol}^L) \tag{3.11}$$

The gas phase binding energy is calculated as the sum of ensemble averages of molecular mechanics energies minus the configurational entropy S_{conf} multiplied by the absolute temperature T:

$$G_b^g = \langle E_{int} \rangle + \langle E_{vdw} \rangle + \langle E_{ele} \rangle - T\langle S_{conf} \rangle \tag{3.12}$$

where E_{int} are the internal energies (bond stretching, angle bending and dihedral torsion), and E_{vdw} and E_{ele} the van der Waals and electrostatics non-bonded energies, respectively. Angular brackets denote ensemble averages over snapshots from molecular dynamics trajectories. The solvation energy is subdivided in an electrostatic and a non-electrostatic part, G_{pol} and G_{np}.

$$G_{sol} = \langle G_{pol} \rangle + \langle G_{np} \rangle \tag{3.13}$$

In summary, we can write:

$$\Delta G_b^s = \langle \Delta E_{int} + \Delta E_{vdw} + \Delta E_{ele} - T\Delta S_{conf} + \Delta G_{pol} + \Delta G_{np} \rangle \quad (3.14)$$

The energy contributions are derived separately for each molecular dynamics snapshot. ΔE_{vdw} and ΔE_{ele} can be read directly from the already calculated potential energy function of the MD simulations that has been conducted to produce the snapshots. Theoretically, bound and unbound states would have to be simulated separately to account for conformational rearrangements that occur upon binding - the induced fit effect. This approach however practically brings about inaccuracies due to sampling issues which are typically much larger than the approximation of neglecting the induced fit effect. It is therefore common practice in MMPBSA calculations to only simulate the complex in solution and take the receptor and ligand coordinates from the complex. This is in turn slightly more efficient. This practice is referred to as the single-trajectory-approach. It leads to cancellation of ΔE_{int}. ΔG_{pol} is usually calculated with an implicit solvation model, either Poisson-Boltzmann (2.2.1) or Generalized Born (2.2.2). The performance of other approaches such as e.g. the reference interaction site model (RISM) have also been investigated [104]. ΔG_{np} is calculated as a linear correlate to the $SASA$ (2.2.3). Especially this term has been shown to be very approximate leading to severe inaccuracies [66, 67]. The theoretical foundation of this thermodynamic cycle has also been criticized [105, 106].
Over the last 10 years, the MMPBSA method has been extensively tested and used to elucidate details of protein-protein [22, 107] and protein-ligand interactions. The experiences made with the use of MMPBSA are diverse ranging from encouraging to unsatisfactory results. A pretty good correlation to experimental results has been obtained for a congeneric series of ligands to FKBP12 [108]. However, it is worth noting that FKBP12 is a relatively small and rigid receptor and the system might not be representative for a general benchmark. Furthermore, ranking of congeneric series is typically less difficult than ranking affinities of different protein variants to the same inhibitor, probably because the conformational space of the ligand is smaller than that of the receptor. Especially HIV protease inhibitor complexes have been extensively studied typically yielding good agreement with experimental values [109, 23, 24, 110]. Wittayanarakul et al. have stated that the identification of the correct protonation state of the HIV proteases catalytic aspartats for specific drugs is a prerequisite for the correct ranking of inhibitor affinities. While this is certainly true by itself, the outcome of this study is probably more governed by limited sampling [110]. An outstandingly accurate result has been obtained for HIV protease saquinavir complexes [111]. However, in a more recent paper, the authors state that this result was only possible by "choosing the right ... sampling trajectories" [112]. An appealing feature of the MMPBSA method is the energy

residue decomposition, i.e. the possibility to i) break down the binding energies into their separate van der Waals and electrostatic contributions and ii) partition these energy contributions on a per residue basis. This procedure gives information that is not amenable to non-computer experiments and it can give valuable hints and directions for the specific development of new drugs that are stronger binders and more robust against the emergence of resistance mutations [113, 24, 25, 114]. The experiences with MMPBSA are also often controversial. While Jenwith et al. experience that incorporating protein flexibility through the use of MD simulations improves correlation with experimental affinities [115], Rastelli et al. find, in an assessment for virtual screening, that using only one snapshot is often better than using MD [116]. Pearlman et al. experienced the three-trajectory approach to be superior over the single-trajectory-approach and MMPBSA to perform poorly compared to commonly used scoring functions [117]. Some special uses of MMPBSA include the investigation of the role of interface waters in binding [118], or the identification of conformational substates [119]. An extensive systematic study has been performed by Ryde and coworkers in a series of papers [120, 72, 121, 61, 66, 67, 122]. Most importantly, it has been convincingly shown that using an ensemble of many short simulations instead of one long simulation improves the convergence of binding free energy estimates and is necessary to get reproducible results [120]. Independent simulations can be obtained in several ways, e.g. using different initial atom velocities or different initial coordinates for the water molecules of the solvent box [72]. Considering the severe sampling issues, all MMPBSA studies using only one simulation per system have to be interpreted with caution. Only few studies exist using multiple trajectories to reach better convergence [120, 112, 123]. Besides the sampling issue the inaccuracy of the non-polar term [66, 67] and the convergence difficulties of the NMA entropy estimate are most crucial. In fact, it was often observed that neglecting the conformational entropy term gives results closer to experimental affinities (e.g. [112]). Given its intrinsic imprecision, MMPBSA has lost its greatest appeal, - that of being more efficient than rigorous methods.

From the literature study, it became apparent that MMPBSA can possibly be improved by the following means:

Multiple Independent Simulations. With improved simulation algorithms and especially the GPU code (2.3.6), more sampling is possible. Simulation times of above cited studies typically range from few hundred picoseconds to 10 ns. A set of multiple simulations is believed to improve the convergence of MMPBSA free energy estimates.

Truncation Approach with Fixed Buffer Region. To reduce computational complexity of the normal mode entropy calculation, one approach that has been introduced and often used, is, to truncate all residues that are farther away from the ligand than a given distance (usually 8-12 Å) [124, 125]. An issue with this approach is that it may lead to significant

change in the molecular geometry during minimization. Recently, Kongsted and Ryde [121] have proposed to use a fixed buffer region to prevent the geometry from being distorted during minimization.

In this thesis, these two MMPBSA improvements have been tested in a calculation of the binding free energies of HIV protease inhibitor complexes (subsection 4.2).

4 Systems and Applications

4.1 Derivation of Sulphotyrosine Forcefield Parameters

4.1.1 Introduction

The modified amino acid sulphotyrosine (O-sulphate-L-tyrosine, TYS) is contained in many important proteins. TYS residues are frequent extracellular modifications that play roles in many physiological processes including e.g. blood coagulation, cell attachment or viral entry into host cells [126]. Prominent examples of TYS-containing proteins are the HIV coreceptors CCR5 and CXCR4. At their N-termini there are several TYS residues that are essential for the binding of HIV gp120 protein and thus for viral entry [127]. Molecular dynamics simulations of CCR5 have already been performed using available phosphotyrosine parameters as an approximation [128]. Although this might work reasonably well, there is a need for more specific parameters that model the sulphotyrosine characteristics more accurately. To this end, TYS atomic partial charges and parameters suited for Cornell et al. forcefields [32] and its adaptions have newly been derived.

4.1.2 Methods

Restrained electrostatic potential (RESP) atomic charges for N-Acetyl-O-sulphate-L-tyrosine-N'-methylamide dipeptide (ACE-TYS-NME, capped amino acid), as well as for the central, (+)NH3-terminal and (-)OOC-terminal fragments of the O-sulphate-L-tyrosine amino acid were calculated using the R.E.D. IV server [129]. The procedure was executed analogous to the derivation of O-methyl-L-tyrosine charges described in Dupradeau et al. [130].
Two representative conformations close to those found in α-helix and β-sheet secondary structures were considered in the procedure [131]. Geometry optimization and molecular electrostatic potential (MEP) computation were carried out using the Gaussian 09 program (revision A.02) [132] in the gas phase and charge fitting was performed using the RESP program. The Hartree-Fock (HF) method with a 6-31G* basis set was used in geometry optimization, while HF/6-31G* and the Connolly surface algorithm were employed for MEP computation. The molecular orientation of each optimized geometry was controlled using the rigid-body reorientation algorithm implemented in the R.E.D. program. A geometric restraint was imposed on the dihedral angle C-CA-CB-CG during quantum-mechanical (QM) minimization to prevent the sulphate group from interacting with backbone atoms.
Forcefield parameters were adapted from previously derived phosphotyrosine parameters [133] available from the Amber parameter database [134]. To this end, dihedral parameters C-OS-S-O2 and CA-C-OS-S were fitted to a QM energy profile by a systematic search. The QM energy profile was built from 35 representative conformers where these two dihedral angles were varied

in steps of 15 degrees (figure 4.1). The conformers and corresponding QM energies were provided by Björn Thorwirth and Georg Jansen (University Duisburg-Essen).

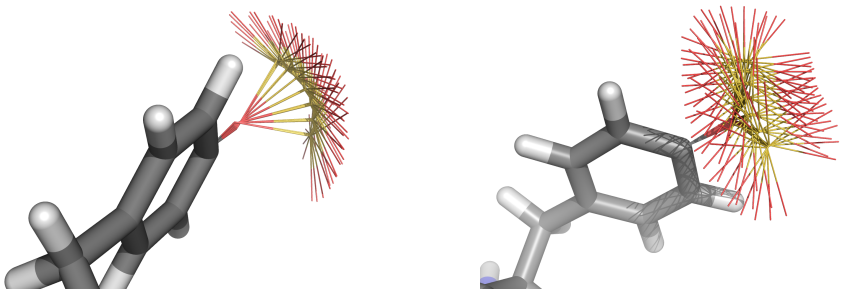


Figure 4.1: Representative Conformers.

4.1.3 Results and Discussion

Figure 4.2 shows the capped sulphotyrosine with the calculated RESP charges.

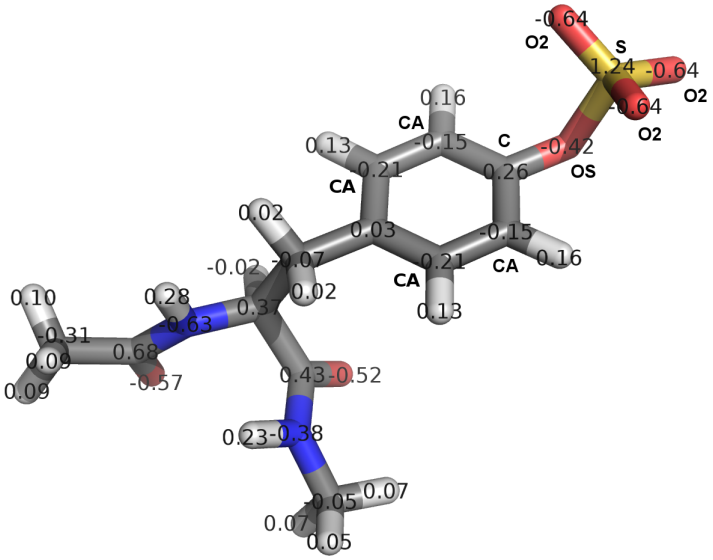


Figure 4.2: Sulphotyrosine Charges. Important atoms are labeled by Amber atom type (bold font).

Table 4.1 shows the new set of sulphotyrosine parameters in the format required in the Amber forcefield modification file (frcmod). K_r, r_{eq} are the bond and angle equilibrium lengths and K_Φ, Φ_{eq} the associated force constants, respectively. The representation of the torsional potential is $((V_n/2)/n_{bp}) \cdot (1 + \cos(n_p \cdot \Phi - \gamma))$ where V_n is the barrier height, n_{bp} is

the number of bond paths, a factor by which the barrier height is divided, γ the phase shift angle and n_p the periodicity of the torsional barrier. Improper torsions are represented in the same way as the torsions but not divided by a n_{bp} factor. A more detailed explanation for this representation is given in Weiner et al. [135]. The sources of parameters are given in brackets. Pp means that the parameter is adapted from an analogous parm99 phosphate parameter [133], P means taken from parm99 [35], G means taken from GAFF [37], Q means that these parameters are taken from a quantum energy-minimized structure and C means that these parameters were newly computed.

Table 4.1: Set of Sulphotyrosine Parameters.

Bonds	Atom types	K_r	r_{eq}		
	S-OS	525.0 (Pp)	1.66 (Q)		
	O2-S	525.0 (Pp)	1.43 (Q)		
Angles	Atom types	K_Φ	Φ_{eq}		
	C-OS-S	100.0 (Pp)	120.50 (Pp)		
	CA-C-OS	70.0 (G)	120.00 (G)		
	O2-S-O2	140.0 (Pp)	119.90 (Pp)		
	O2-S-OS	100.0 (Pp)	108.23 (Pp)		
Dihedrals	Atom types	$V_n/2$	n_{bp}	γ	n_p
	C-OS-S-O2	0.889 (C)	3	0.0	3
	CA-C-OS-S	0.333 (C)	2	180.0	2
Impropers	Atom types	$V_n/2$		γ	n_p
	CA-CA-C-OS	1.1 (P)		180.0	2

As can be seen in figure 4.3 on the right, the molecular mechanics energies (yellow) correlate very well with the energies from the quantum mechanics calculation at the Hartree-Fock level of theory (black dotted bold line) with r=0.97 and a maximum deviation of 2.3 kJ/mol. They also correlate well with quantum energies on the B3LYP level (black dotted timid line) indicating that the charges are reasonably consistent with the forcefield ff03. Furthermore, the calculated charges are consistent with previously derived charges for O-phosphate-L-tyrosine [133]. As an indication of improvement over the old parameters, the left figure shows the correlation of MM versus QM energies if phosphotyrosine torsion parameters are used.
As a first validation, we performed two 50 ns explicit water simulation of ACE-TYS-NME with the old and new parameter sets. We observed that the sulphate group is free to rotate (roughly twice per nanosecond). Using the phosphotyrosine parameters as an approximation, the energy barriers are much higher and the sulphate group rotated only twice in 50 ns.

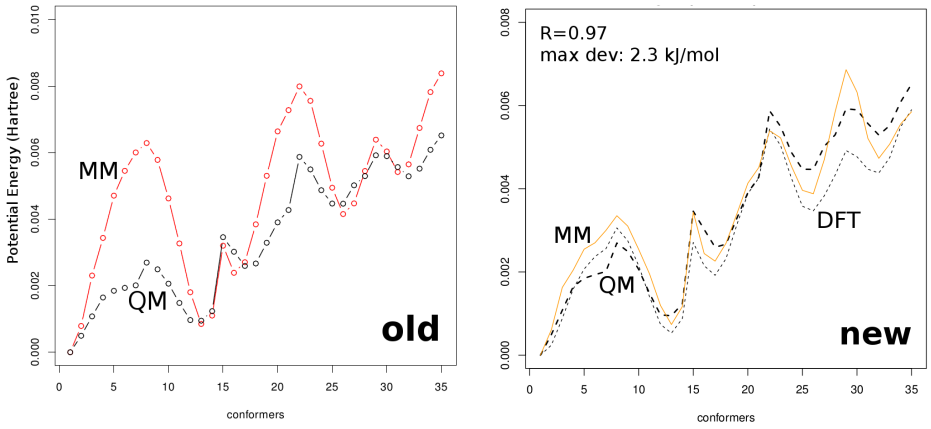


Figure 4.3: Fit of Quantum and Molecular Mechanical Energies. Left: MM (red) and Hartree-Fock QM (black) potential energies using the phosphotyrosine torsion parameters. Right) MM (yellow) versus Hartree-Fock (bold black) and B3LYP (timid black) QM using the new sulphotyrosine parameters.

4.1.4 Conclusion

Since the charges were derived with state-of-the-art tools consistent with the derivation procedure of Cornell et al. forcefields [32], the derived atomic partial charges are of high quality. The procedure for the derivation of missing forcefield parameters is certainly not perfect and especially not generic. However, the new parameters are at least a major improvement over the old onces and should be used when simulating TYS-containing protein systems. The parameters could certainly be improved by a more thorough optimization using for example a genetic algorithm. A ready-to-use protocol for the derivation of these parameters is not available at present but under development [136]. For further validation, TYS-containing protein systems should be simulated and analyzed for stability of certain crucial geometric features.

4.2 MMPBSA on HIV Protease Complexes

4.2.1 Introduction

With the methodological MMPBSA improvements outlined at the end of subsection 3.3.2 it has been tested whether it is now possible to rank binding affinities correctly. A set of four complexes, a wildtype (WT) and a double mutant (V82T/I84V, MT) HIV protease complexed with amprenavir (APV) and darunavir (DRV) where both experimental affinities from isothermal titration calorimetry (ITC) and crystal structures are available has been chosen as a test set. The range of the binding energies of these complexes is relatively narrow (3.5 kcal/mol) compared to other typical test systems such as congeneric series of Biotin analogues bound to Avidin [120]. This makes the calculation of a correct ranking very challenging, but it is on the other hand a real world problem. While the experimental precision of the free energies is limited and might be insufficient to distinguish the values, the correctness of the experimental ranking is verified by the resistance context.

4.2.2 Methods

Preparation of Complexes The HIV-PR simulations of WT-APV, MT-APV, WT-DRV, and MT-DRV were based on the corresponding crystal structures with PDB ids. 1HPV, 1T7J, 1T3R, and 1T7I [137, 138], respectively. Two different models (A,B) of each system were build by using available alternate side-chain conformations. Since there are no alternate side-chain conformations available in 1HPV, crystal structures with PDB ids. 1HPV and additionally 1T7J were used to build two different models for WT-APV. Crystal water molecules were removed, except one structurally important water molecule between the protease flaps and inhibitor. One of the catalytic Asp residues was protonated (Asp 25) on the basis of considerations made by Hou et al. [23], whereas all the other Asp and Glu residues were negatively charged and all Arg and Lys residues were positively charged. The two His residues were protonated on the NE2 atom. The protein was described with the Amber03 force field [36] and the ligands with GAFF [37]. The ligands were optimized at the HF/6-31G** level using Gaussian09 [132]. The electrostatic potential (ESP) was then calculated at the B3LYP/cc-pVTZ level with the polarizable continuum model [139] and a dielectric constant of $\epsilon = 4$ at points sampled with the Merz-Kollman scheme [140]. Point charges were fitted to the ESP using the RESP [141] procedure with the antechamber program. All complexes were immersed in a truncated octahedral box of TIP3P water molecules [43] extending 10 Å from the protein. Finally, five to seven chlorine ions were added to neutralize the system.

Molecular Dynamics Simulations The sander module of Amber 10 was used for minimization and equilibration and the GPU version of the Amber 11 pmemd module [79] was used for production. The temperature was kept constant at 300 K using a Langevin thermostat [69] with a collision frequency of 5 ps^{-1}, and the pressure was kept constant at 1 atm using an anisotropic scaling barostat [68] with a relaxation time of 1 ps. Particle mesh Ewald summation [29] was used to treat long range coulombic interactions. The cutoff for nonbonded interactions was set to 9 Å. The SHAKE algorithm was used to constrain bonds involving hydrogen atoms allowing for an integration time step of 2 fs. Water molecules and hydrogen atoms were relaxed first using 100 cycles steepest descent followed by 100 cycles conjugate gradient with all protein and ligand heavy atoms restrained with a force constant of 4 kcal $mol^{-1}Å^{-2}$. The solute was relaxed by 1000 cycles steepest descent followed by 100 cycles conjugate gradient with no restraints. Initial velocities were randomly assigned from a Boltzmann distribution. The system was then gradually heated from 0 to 300 K over 50 ps and equilibrated for another 100 ps in the NVT ensemble. 1.5 ns production simulations were carried out in the NPT ensemble. It was known from previous calculations that this simulation length is sufficient to converge ΔG estimates to within 0.5 kcal/mol in one individual simulation. For each system equilibration and production steps were repeated 28 times with both different random seeds (VIIT) and different initial water boxes (SIIT).

MMPBSA Implementation and Parameterization The MMPBSA.py module (Amber 11) was used for the calculation of all enthalpic terms and a modified version of the Amber 10 perl implementation (mm_pbsa.pl) together with additional perl scripts were used for the calculation of the entropy. The terms were calculated with all solvent molecules and ions stripped off the trajectory and without any periodic boundary conditions but with an infinite cutoff. The MM energies were estimated using the same forcefield as in the simulations. The polar solvation energy was calculated with several models, namely the Poisson-Boltzmann model with protein dielectric constants set to 1 or 2 using two different numerical solvers, the PBSA module from the Amber software (PB1, PB2) and APBS (APBS1, APBS2), and the Generalized Born model with a protein dielectric constant set to 1 (GB). The solvent dielectric constant was set to 80 in all cases. The non-polar solvation energy was calculated according to $\Delta G_{np} = \gamma SASA + b$ with $\gamma =$ 0.00542 kcal/mol·$Å^2$ and $b = 0.92$ kcal/mol [142]. The translational and rotational entropy was calculated with standard statistical mechanics formulas [81]. The vibrational entropies were estimated using the ideal-gas rigid-rotor harmonic-oscillator approximation [81]. The systems were minimized before the frequency calculations. Both the minimization and the frequency calculations were performed in a vacuum. The vibrational entropy calculations

were performed using the truncation approach suggested by Kongsted and Ryde [121] (figure 4.4). In this approach, all protein residues within a radius of 8 Å of the ligand (including the ligand itself) are included in the calculations together with a buffer that consists of all residues within 4 Å of the former residues, as well as all water molecules within 12 Å of the ligand. The buffer is meant to keep the geometry of the complex as close as possible to the original geometry during minimization.

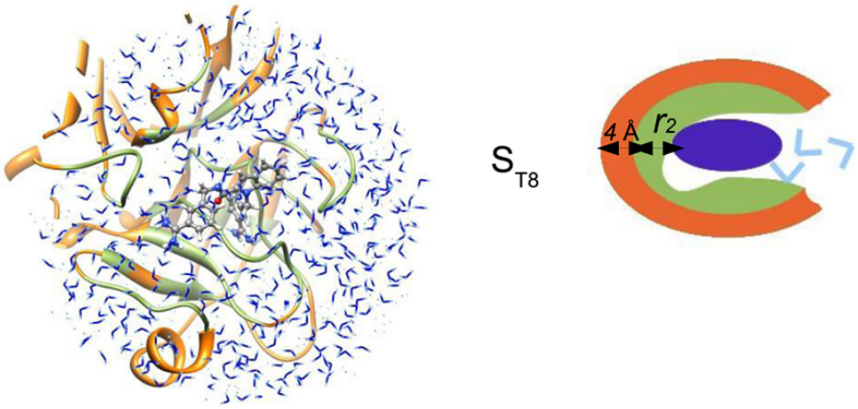


Figure 4.4: Entropy Calculation Truncation Approach Water molecules are shown in blue, protein residues that are included in the frequency calculations are shown in green, and protein residues that are in the buffer region are shown in orange. The ligand is shown in ball-and-stick representation. The protein is fXa. (Illustration adapted from [143] and produced by Samuel Genheden)

Analysis of the Energy Calculation From the 1.5 ns simulation length, snapshots in intervals of 10 ps were used for the energy calculation making a total of 300 snapshots in each individual simulation. For the final result shown in figure 4.5, 0.5 ns were considered additional equilibration and the last 1 ns was used to calculate the ensemble averages albeit the dependence of results on different chosen equilibration times has also been considered. For clarity, the data of each time series was subdivided into 12 blocks of 25 snapshots each (corresponding to a running mean with no overlap). I have chosen this variant because it formidably preserves the characteristics (especially the drift) of the data while smoothing out variance from thermal fluctuation. Alternatively, I have also used cumulative averages and reverse cumulative averages [144] which tend to hide drifts and give a too optimistic picture of convergence. The statistical error of the binding free energy estimates is given by the standard error of the mean with n being the number of simulations. This is deviant from the usual practice of giving the standard

error of the mean with n being the number of snapshots (e.g. [112]) and it is characteristically more conservative. The quality of the ranking was assessed with both the Pearson correlation and Kendalls rank correlation. According to the Gaussian law of error propagation for independent variables, the ΔG error bars were calculated as the sum of ΔH and $T\Delta S$ error bars.

4.2.3 Results and Discussion

Binding affinities for HIV-protease complexes using the MMPBSA method with ensembles of short simulations and a new entropy approach were calculated. The solvation models Poisson-Boltzmann using the Amber pbsa solver and APBS both with dielectric constants 2 and 4 and Generalized Born (dielectric 1) were tested.

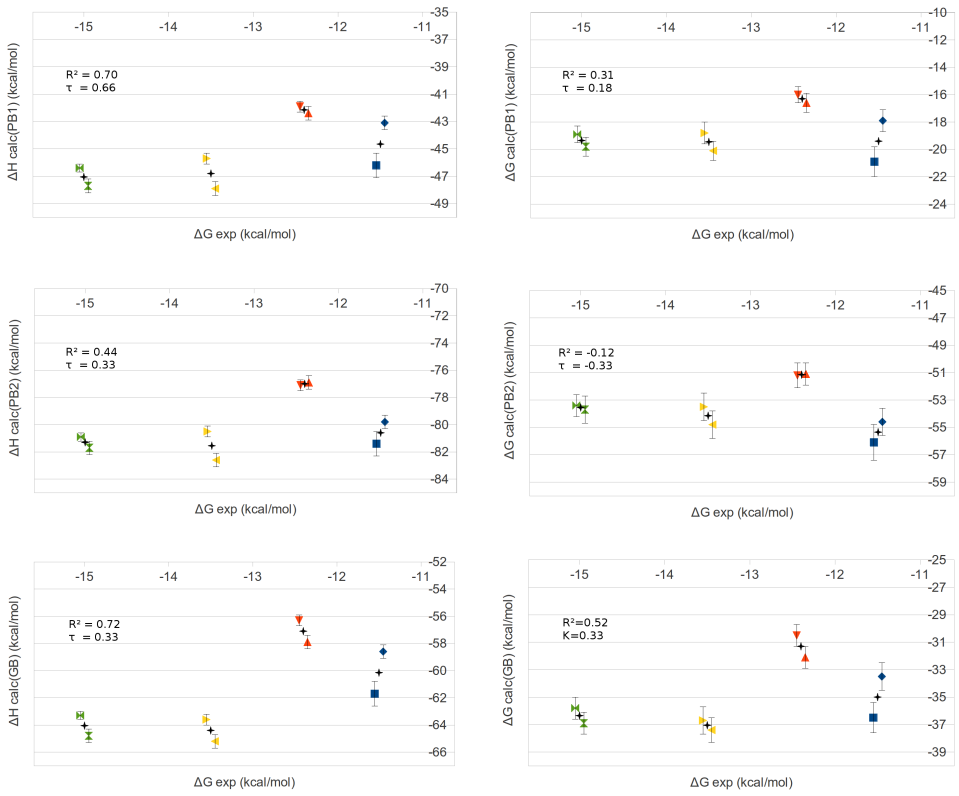


Figure 4.5: Ranking of Binding Affinities. MMPBSA results for the different solvation models PB1, PB2 and GB are shown from top to bottom (data for APBS not shown). Calculated ΔH and ΔG (including $-T\Delta S$) each vs. experimental affinities [145] are shown on the left and right panels of figures, respectively. For each protein-ligand system, the calculated affinities from models A and B are shown left and right, both with error bars (standard error of the mean), and the combined average is shown in the middle (black star). The scaling on the y-axis is equal for all plots.

Ranking of Binding Affinities Generally, the calculated binding affinities have converged to standard errors 0.6-1.1, 0.8-1.1 and 0.8-1.3 kcal/mol for PB1, PB2 and GB, respectively (figure 4.5). More simulations would be needed to reach the desired convergence to below 0.5 kcal/mol that would establish a statistically valid ranking. Interestingly, in many cases, the error bars of conformations A and B do not overlap indicating a strong dependence on the initial structures although the differences of these structures are relatively marginal.
It is important to note, that the affinity estimates for one same system vary in the range of roughly 10 kcal/mol for the 28 simulations indicating that the

values strongly depend on the chosen initial conditions. Calculations that I conducted previously showed that this circumstance does not change even if all simulations are prolonged to 10 ns.
The ranking of binding affinities is modest with Pearson correlation coefficients 0.31 (p=0.3), -0.12 (p=0.9) and 0.52 (p=0.5) and Kendalls rank correlation coefficients 0.18, -0.33 and 0.33 for PB1, PB2 and GB, respectively. However, none of the rankings is significant. As must be expected, the absolute binding affinities depend strongly on the solvation models while the absolute values using PB1 are closest to experiment (-18 to -22 kcal/mol). Moreover, also the established rankings depend on the solvation models as is reflected by the different Kendalls correlations.
Regarding the enthalpies, the APV complexes are clearly distinguished from the more stable DRV complexes which is probably a consequence of the favorable contacts of the additional DRV moiety. The large enthalpic experimental difference of 2.5 kcal/mol between the DRV MT and WT systems is reproduced by none of the solvation models and the values are almost equal instead. Also notably, the APV systems are more clearly distinguished from the DRV complexes by the GB model than by the PB model. However, the absolute enthalpic difference seems to be overestimated here. Surprisingly, the ranking of enthalpies of APV MT and WT systems is systematically wrong. The APBS solver (data not shown) generally gives larger fluctuations than the Amber pbsa solver and there were outliers in some cases. The cause of these outliers could not be determined.

Time Dependence of Calculated Affinities The dependence of enthalpic energies on different equilibration and production times were considered (figure 4.6). While the energies typically change in the first 250 ps (blocks 1 and 2) due to structural rearrangements, they are reasonably stable over the rest of simulation time. Averages from blocks 3-4 are highly correlated to the averages from blocks 5-12 (r=0.94-0.98). It can also be observed that APV complexes are by trend destabilized while the energies of the DRV complexes are more or less constant. The accuracy of the enthalpy values are lower than the aimed 0.5 kcal/mol in 7 of 8 complexes in the first 250 ps of the simulations but exceeds 0.5 kcal/mol in the further time course (data not shown), probably because the simulations diverge into different conformational subspaces. The GB results typically have larger variation than the PB results indicating that the GB model is more sensitive to conformational differences than the PB models. This effect can also be seen in the data of Genheden et al. [104].

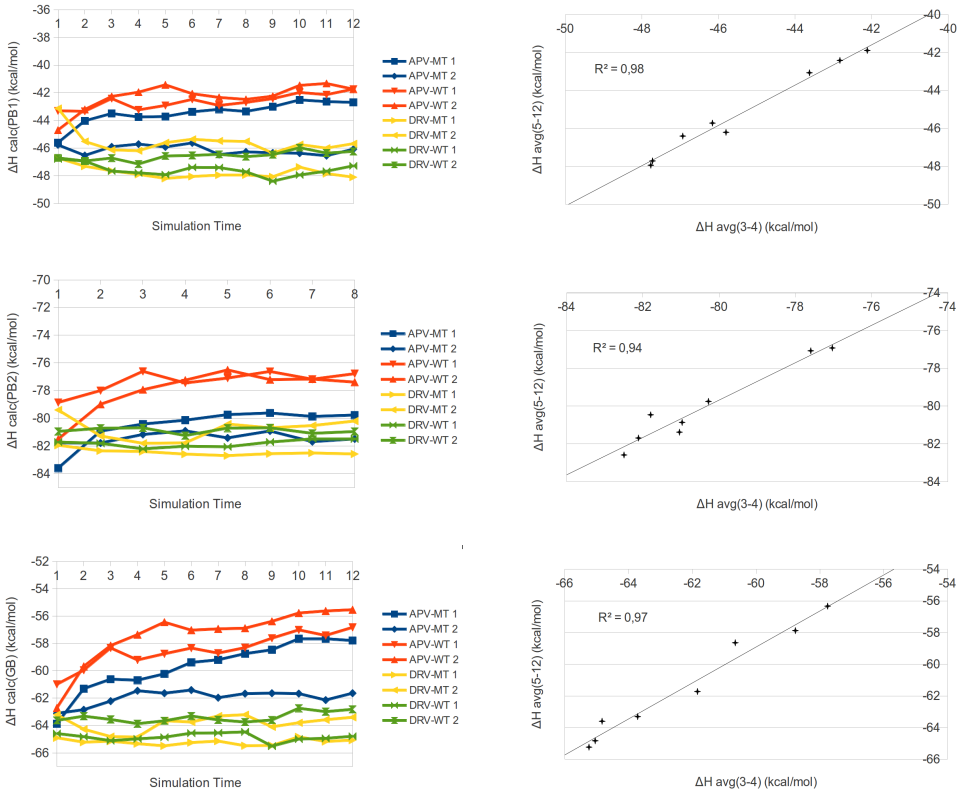


Figure 4.6: Time Dependence of Calculated Affinities. Results for the different solvation models PB1, PB2 and GB are shown from top to bottom. Left) Timeseries of ΔH blockwise averages. Right) Correlation of averages from blocks 3-4 and blocks 5-12.

Correlations between solvation models are shown in figure 4.7. The highest correlation (0.87) is observed for GB versus PB1, probably because the GB model is parameterized against an internal dielectric of 1.

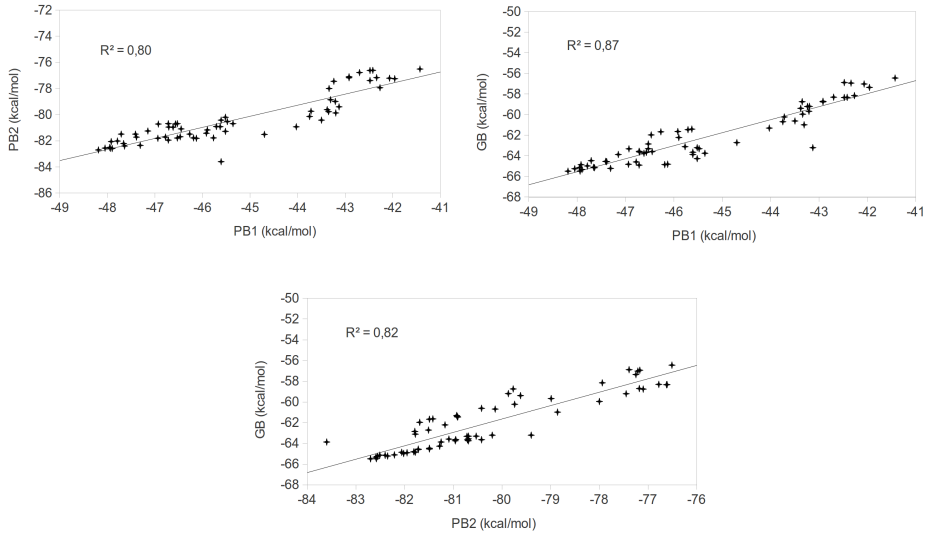


Figure 4.7: Correlation of calculated ΔH between the different solvation models. Blockwise averages from all systems were plotted.

Entropies from Truncation Approach The time dependence of calculated entropies is shown in figure 4.8. Compared to the enthalpies, the entropies are relatively stable over time. Interestingly, for the APV complexes, entropies are pretty much equal in the first half of the simulations while they tend to differ in the second half of the simulations. The standard error of the mean is very low (0.3 kcal/mol) compared to typical results from the conventional approach with no truncation and no water buffer.

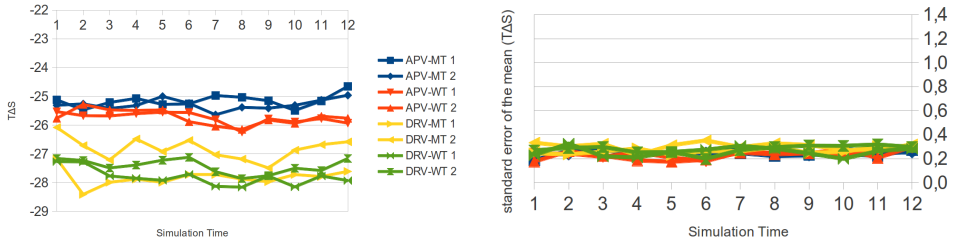


Figure 4.8: Entropy from Truncation Approach. On the left: Entropic component $T\Delta S$, on the right: standard error of the mean $T\Delta S$.

A more detailed analysis of the truncation entropy approach has additionally

been performed comprising analysis of the impact of buffer region, truncation radius and dielectric constant and was published in the Journal of Chemical Information and Modeling [143]. As a main result, this study shows, that while the truncation approach gives systematically lower entropies, the calculated relative affinities are not affected by the truncation and the estimates are typically much more precise, making the calculation more efficient.

4.2.4 Conclusion

A general important conclusion that has to be drawn from this study is that multiple trajectories are necessary to converge MMPBSA estimates with respect to the varying arbitrarily chosen initial conditions. I experienced that virtually all results from the literature that are based on only one simulation per system are not reproducible because the simulations depend on initial conditions, namely the conformation of the starting structure, the random velocity distribution of atoms and the configuration of the initial water box. Since the calculated affinities significantly depend on marginal structural differences of some rotamers of a small number of sidechains, it might be a better strategy to use one same starting structure and model the point mutations to calculate the relative affinities.
Furthermore, I conclude that using ensembles of simulations, while improving convergence, is still not a profound solution to fully converge MMPBSA estimates because the distributions of rotamers, especially those of buried side-chains, still depend on the initial rotamers. The time scale for rotation of buried side chains is 10^{-4} to 1 second. Hence, a complete Boltzmann-distributed ensembles seems to be far out of reach. Additionally, even more simulations would be needed to converge the estimates below 0.5 kcal/mol. As a consequence of its intrinsic imprecision, MMPBSA type calculations are not more efficient than rigorous free energy calculations which should be preferred if its possible. The truncation approach gives more precise entropy estimates than the conventional entropy approach and is a proper approach to increase the efficiency of NMA entropy calculations.
Further uncertainties exist in this study that hamper the evaluation of its outcome. Although the protonation state of the catalytic aspartates was chosen upon reasonable considerations, its validity is not guaranteed and this could possibly affect the calculated affinities. More importantly, the missing forcefield parameters for the inhibitors were taken from GAFF and not further validated. A consistent validated set of parameters for the approved HIV protease inhibitors would be of great value because the HIV protease serves as a prominent model system for numerous molecular dynamics investigations.

4.3 L76V Thermodynamic Integration Calculation

4.3.1 Introduction

An interesting mutation of the HIV protease is L76V. It has first been observed as a resistance mutation against the three inhibitors lopinavir, amprenavir and darunavir. Surprisingly, it turned out that L76V resensitivates the HIV protease against saquinavir and atazanavir (ATV). Hence, the L76V mutation is of direct therapeutic relevance and clinicians have strong interest in an explanation for this effect at the molecular level. Unfortunately, the relative free energy change could not be reproduced with MMPBSA calculations. In this work we investigated whether it is possible to capture the hypersensitization effect of the HIV-PR L76V mutation against ATV [146] using rigorous free energy calculation, and we chose thermodynamic integration (TI). The free energy difference for the wildtype has not been measured (at least not at the time of this study) and is only indirectly derived from z-score data out of geno2pheno [8] predictions. A z-score value has been determined by Alcaro et al. to -0.3 [147] corresponding to a free energy change of approximately 1 kcal/mol. In accordance to TI (explained in subsubsection 3.3.1), the relative free energy difference can be calculated from two transformations, one in the bound and one in the free state:

$$\Delta\Delta G_{bind,L76V} = \Delta G_{L76,binding} - \Delta G_{V76,binding} = \Delta G_{L76V,bound} - \Delta G_{L76V,free} \tag{4.1}$$

The L76-ATV (wildtype inhibitor complex) should be less stable than the V76-ATV (mutant inhibitor complex). Thus, if the calculation comes out right, a **positive** $\Delta\Delta G_{bind,L76V}$ has to be expected (corresponding to a negative z-score value), e.g.

$$\Delta\Delta G_{bind,L76V} = \Delta G_{L76,binding} - \Delta G_{V76,binding} = -12 - (-13) = 1 \tag{4.2}$$

4.3.2 Methods

Preparation of Input Files All calculations were conducted using Amber 10. HIV protease ATV complex models were built based on the crystal structure with PDB id. 2AQU [148]. One of the catalytic Asp residues was protonated (Asp 25), whereas all the other Asp and Glu residues were negatively charged and all Arg and Lys residues were positively charged. The two His residues were protonated on the NE2 atom. The protein was described with the Amber03 force field [36] and the ligands with GAFF. The ligands were optimized at the HF/6-31G** level using Gaussian09 [132]. The electrostatic potential (ESP) was then calculated at the B3LYP/cc-pVTZ level with the polarizable continuum model [139] and a dielectric constant of $\epsilon = 4$ at points sampled with the Merz-Kollman scheme [140]. Point charges

were fitted to the ESP using the RESP [141] procedure with the antechamber program. All complexes were immersed in a truncated octahedral box of TIP3P water molecules [43] extending 12 Å from the protein. Finally, eight chlorine ions were added to neutralize the system. Parameter and coordinate files were built as follows: In a first leap run, two pdb files were created with solvated and neutralized receptor and complex wildtype (L76) systems. Then, the additional methyl group atoms and the C_β atom were removed from these pdb files (in both monomers) and these LEU residues renamed to VAL yielding the mutant receptor and complex pdb files. The valine atoms added by leap were remodelled with Pymol [149] to get maximum overlap with the leucine residues. In a second leap run, topology and coordinate (restart) files were created for all four systems.

Free Energy Calculation The free energy was calculated according to the TI scheme described in subsubsection 3.3.1. Both transformations in the bound and unbound state were subdivided into three subtransformations for switching off partial charges on L76, transforming van der Waals parameters, and switching on the partial charges for V76. For each subtransformation, sampling was performed at λ = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9. Values for λ = 0 and 1 were extrapolated using the trapezoid rule. The total free energy is obtained by summing over all λ values. The following λ-dependent soft-core potential [150] was used for the van der Waals parameters of coupled atoms

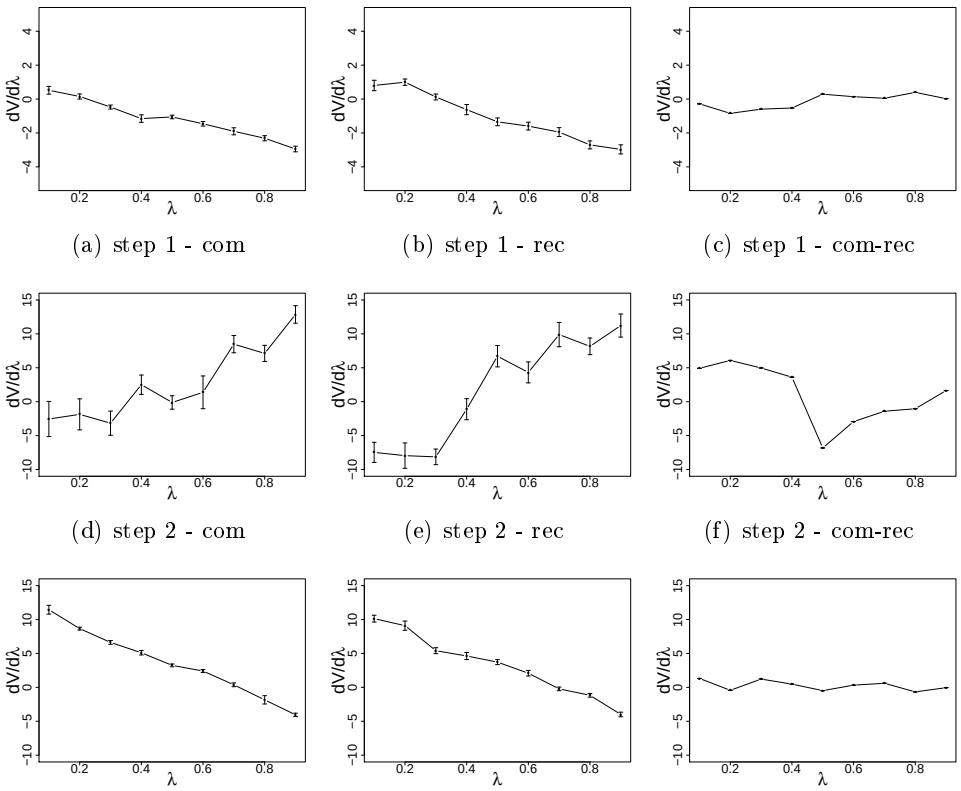
$$V_{softcore\_vdW} = 4\epsilon(1-\lambda)\left[\frac{1}{[\alpha\lambda + (r/\sigma)^6]^2} - \frac{1}{\alpha\lambda + (r/\sigma)^6}\right] \tag{4.3}$$

where ϵ and σ are the common Lennard-Jones parameters, r is the interatomic distance and α is an adjustable constant set to 0.5. No softcore potential for Coulombic interactions was used.

MD Simulations The multisander module was used for all simulations. The temperature was kept constant at 300 K using a Langevin thermostat [69] with a collision frequency of 5 ps^{-1}, and the pressure was kept constant at 1 atm using an anisotropic scaling barostat [68] with a relaxation time of 1 ps. Particle mesh Ewald summation [29] was used to treat long range Coulombic interactions. The cutoff for nonbonded interactions was set to 9 Å. The SHAKE algorithm was used to constrain bonds involving hydrogen atoms allowing for an integration time step of 2 fs. The whole system was minimized by 500 cycles steepest descent. Initial velocities were randomly assigned from a Boltzmann distribution. The density was then equilibrated for 25 ps in the NVT ensemble while heating the system gradually from 0 to 300 K. For each λ step of each subtransformation 6 ns production simulations were carried out in the NPT ensemble. The last 3 ns were used for the energy calculation.

4.3.3 Results and Discussion

Resulting $dV/d\lambda$ curves are shown in figure 4.9.



(a) step 1 - com (b) step 1 - rec (c) step 1 - com-rec

(d) step 2 - com (e) step 2 - rec (f) step 2 - com-rec

(g) step 3 - com (h) step 3 - rec (i) step 3 - com-rec

Figure 4.9: $dV/d\lambda$ curves assembled from 9 λ points each. The three transformations, turning off charges on L76, transforming vdW parameters and turning on charges on V76 are shown from top to bottom (named step1, step2 and step3). Bound and unbound transformation and the resulting difference are shown from left to right (named com, rec and com-rec).

It can be seen that the curves are generally relatively smooth indicating sufficient convergence. The electrostatics transformations are particularly smooth with very low standard deviations. This is reasonable because the partial charges of the changing atoms from the hydrophobic residues leucine to valine are very small. Accordingly, the free energy of electrostatic transformations are close to zero (-0.14 and 0.41 kcal/mol). However, there is a substantial free energy change for the van der Waals transformation. The kink in the $dV/d\lambda$ curve for this transformation indicates that there might be an abrupt conformational change. From visual inspection alone, however, this conformational change cannot be figured out and more advanced

analysis would be needed to understand the mechanistic effect of the L76V mutation.
Free energy changes calculated for the three transformations are shown in table 4.2. The total free energy change (1.53 kcal/mol) corresponds well to the expected experimental value (1 kcal/mol). Most importantly, the trend of the free energy change is correct identifying the L76V mutant as a stronger binder to ATV than the wildtype.

step	free energy change ΔG (kcal/mol)
1	-0.14
2	1.26
3	0.41
total	**1.53**

Table 4.2: Thermodynamic Integration Results.

4.3.4 Conclusion

Since the structural change introduced by the L76V mutation is rather small, this was a good example where TI could work. The TI calculation roughly reproduces the expected free energy change of 1 kcal/mol. Given that such a small free energy difference is approximately the margin of error for a TI calculation on biological systems, this is a pretty good result. Error estimation for TI is a tricky thing to do [150] and I left it with giving the standard errors for $dV/d\lambda$ values for the transformations. However, the smoothness of the $dV/d\lambda$ curves suggests that the result is not by chance.
A justified criticism to this study is the choice of the initial structure for the simulations of the free receptor state which has been taken from the HIV-ATZ complex with ATZ coordinates stripped off. It is well known that the apo HIV protease flaps adopt a semi-open conformation [151] and the chosen initial structure does possibly not sample the correct free equilibrium state. A mechanistic understanding of conformational rearrangement induced by the L76V mutation underlying the observed free energy change would be desirable. Although the transformation paths are unphysical, it could nevertheless well be valuable to do some form of distance analysis on this data.

4.4 Molecular Dynamics Study on HCV Protease

4.4.1 Introduction

The aim of the following study is to explain a set of interesting resistance data for the HCV protease, a major drug target for the therapy of Hepatitis C, from a mechanical point of view. The most crucial experience that I got from the MMPBSA calculations is that the calculated energies depend severely on the simulations initial conditions such as the randomly set starting velocities of the atoms, the configuration of the water box and the starting conformation. Since the MMPBSA energies depend on such initial conditions, it is also clear that geometrical features like distances, angles or root mean square fluctuations have similar convergence issues. From this experience, it is natural to not rely on the outcome of single simulations but rather use ensembles of multiple simulations instead. Hence, I have used a set of simulations for each protein system instead of only one simulation and evaluated the geometrical differences with a multiple testing approach. I start here by giving a topological overview of the protease structure and its specific characteristics. Subsequently, I discuss the available new resistance data and revisit an explanation of resistance mechanisms against BILN that is considered useful for this study. Then, I report on the simulation setup that I have performed and the analysis of geometrical features.

The HCV NS3-4A Protease Molecule The NS3 serine protease (HCV protease) is contained within the N-terminal 180 amino acid residues of the NS3 protein [152]. The C-terminal part (residues 180-630) of the NS3 protein has a helicase function. Crystallographic structures of the NS3 protease domain with [153] and without [154] the cofactor 4A have first been solved in 1996. The NS3 protease domain forms a double-barrel fold similar to that of serine proteases from the chymotrypsin/trypsin super-family with one N-terminal eight stranded (residues 1-93) and one C-terminal six stranded (residues 94-180) anti-parallel β-barrel (figure 4.10). The catalytic site is situated in the crevice between the two domains and is formed by the triad of residues H57, D81 on the N-terminal and S139 on the C-terminal domain. The NS4A cofactor peptide is one of the eight strands of the N-terminal β-barrel. The C-terminal domain contains one zinc binding domain at one end opposite to the catalytic triad. Two helices, α1 (residues 56-60) comprising the catalytic His57, and α2 (residues 131-137) are present with and without NS4A. Two additional helices, α0 (residues 13-21) and α3 (residues 172-180), are formed only in the NS3-4A complex structure.
From a comparison of C_α distances of the three catalytic site residues between NS3-4A and other serine proteases, it can be seen that while the distances are generally comparable, the His-Ser distance is relatively long compared to other small serine proteases [155]. This notion supports the hy-

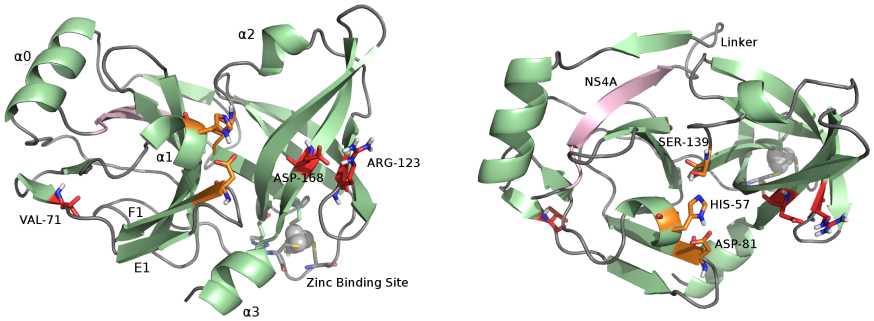


Figure 4.10: Structural Overview of NS3-4A. The figure is based on the initial energy-minimized structure that was used for the simulations. The catalytic serine is already remodelled. The figure on the right is rotated 90° around the horizontal axis with respect to the figure on the left. Secondary structure elements are colored green. The cofactor NS4A that is connected to the protease via a linker is shown in pink. Active site residues and resistance mutations are colored orange and red, respectively. Helices are labeled α_n and two important β-strands are labeled E1 and F1.

pothesis that the substrate induces proper active site geometry upon binding. "The commonly accepted mechanistic model of action of the serine proteinases implies a relay mechanism of hydrogen bonds involving, on one side, the carboxylate moiety of the Asp and the delta-HN of the His residue and, on the other side, the epsilon-N of the His and the gamma-HO of the Ser residues. This relay of H-bonds activates the gamma-O of the serine residue, which can produce the nucleophilic attack on the C atom of the scissile bond." [155]

Role of the NS4A Cofactor NS4A is a cofactor that has been shown to enhance protease activity in all cleavages via formation of a NS3-4A complex. „Interaction with the NS4A cofactor is required to perform the cleavages at NS3/NS4A, NS4A/NS4B, and NS4B/NS5A junctions but the protease in its uncomplexed state is still able to cleave at the NS5A/NS5B boundaries, although with much lower activity" [155]. A kinetic analysis conducted on different types of substrate-like inhibitors in the absence and presence of the NS4A cofactor has shown that the action of NS4A peptide is exerted only on the P'-side of the substrate [156]. From these findings, the authors conclude that NS4 modulates NS3 activity by alteration of the S' subsites. In the Barbato 1999 NMR structure set of the NS3 protease without cofactor [155], the N-terminal β-barrel seems to be less compact than the C-terminal one (figure 4.11). The mechanistic role of NS4A and substrate in the activation of the HCV NS3 protease has also been investigated by MD simulations

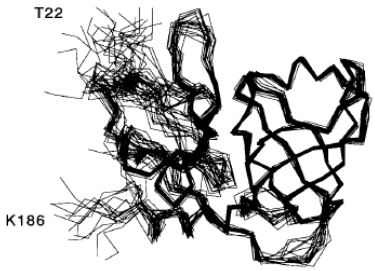


Figure 4.11: NMR Ensemble of HCV Protease without Cofactor NS4A. Strands A1-F1 are in the N-terminal and A2-F2 are in the C-terminal β-barrel [155].

[157]. Eventually, the NS4 cofactor has to be regarded as an integral part of the protein and has therefore no more particular role than the other protein parts.

Structure and Role of the Zinc Binding Domain In contrary to zinc proteases, the zinc binding in the HCV protease is generally believed to have only a structure-stabilizing role. It has been shown that the NS3 protein does not fold properly in the absence of zinc and is catalytically deficient [158]. The zinc-binding site in solution undergoes conformational exchange between an open and a closed conformation by switching the side-chain of His149 on the hundreds of milliseconds timescale [155]. There was some controversy on the His coordination site of the zinc. From the first crystal structures of NS3 and NS3-4A, the presence of a bridging water was postulated because the imidazole ring is too distant from the zinc ion [154]. In the NMR structure set of Barbato et al. [155] the imidazole ring directly coordinates the zinc ion. In the new substrate-bound crystal structures [16] the zinc ion is coordinated by the His via a bridging water.

Substrate Specificity NS3-4A is responsible for cleavage at four sites 4.12.

Substrate	P6	P5	P4	P3	P2	P1	P1′	P2′	P3′	P4′
3-4A	D	L	E	V	V	T	S	T	W	V
4A4B	D	E	M	E	E	C	S	Q	H	L
4B5A	E	C	T	T	P	C	S	G	S	W
5A5B	E	D	V	V	C	C	S	M	S	Y

Figure 4.12: Sequences of HCV Protease Substrate Cleavage Sites. Table from Romano et al. [16].

Unusually long substrates (P6-P4') are required for effective cleavage [159]. Importantly, there is always an acid (predominantly D/E) at P6, Cys/Thr at P1 and Ser/Ala at P1'.

Further Relevant Observations from the Literature The loop with the three residues 79-81 is solvent-exposed in the apo structure and has high temperature factors indicating a high degree of mobility [155, 16]. Hence, the hydrogen bond between the carboxylate group of Asp81 and the delta-NH of the His57 is not stable in the apo structure. It is generally believed that the Asp81 has to be shielded for the formation of a proper active site geometry. In fact, there are always large hydrophobic residues at the substrates P2 sites of small serine proteases. This residue together with the aliphatic methylene groups of R155 could contribute to shelter the Asp81. Helix $\alpha 3$ is crucial for the correct packing of the strand F1 that positions the catalytic Asp81 [160]. In many of the conformations from the Barbato NMR structure set, there is an alternative conformation in which the β-sheet is twisted and Asp79 coordinates the catalytic His57 instead of Asp81. This could possibly be an alternative mode of His coordination. Given the fact that Asp79 is only conserved in HCV genotype 1 this could be a reason why the genotype 1 is more robust than the other genotypes. It is also worth noting that the residue position 71 is polymorphic in the wildtype. Several hydrophobic residues, in particular I, V and L have been observed [152]. In direct proximity, I72 and T72 instead of V72 have also been observed.

Substrate-bound Structure Recently, structures of complexes of HCV protease with the N-terminal product of the four natural peptide substrates were solved [16] at resolutions 1.6-1.9 Å. For this structure determination, the NS3-4A single-chain construct was used in which the essential NS4A cofactor is covalently linked to the N-terminus by a flexible linker. The catalytic serine has been substituted with an alanine to prevent catalysis [16]. Because even with this substitution, the P7-P5' substrate was cleaved, presumably mediated by a water molecule, only the complex with the N-terminal cleavage product P7-P1 could be determined [16].
The availability of this structure gives us the opportunity to not only investigate the mutation-induced geometrical differences on the apo structures but also on the substrate-bound state. This is particularly relevant because the I71T mutation might affect the active site, and it has been pointed out that the N-terminal substrate possibly induces proper active site geometry.

For this study, I have used one of these structures, the one with the 4B5A substrate, as template (figure 4.13) because it has a particularly direct relationship to the mutation R123K. „In all structures but product complex 4B5A, K165 forms salt-bridges with the P6 acids, while residues R123, D168, R155, and the catalytic D81 form an ionic network along one surface of the bound products. In complex 4B5A, R123 interacts directly with the P6 acid, while D168 reorients and no longer contacts R155.“ [16].

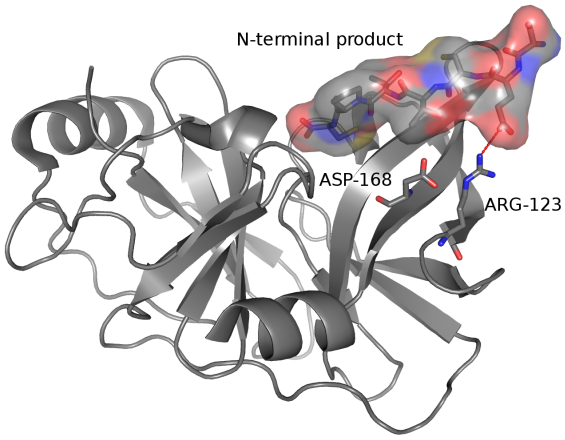


Figure 4.13: HCV protease complexed with the N-terminal product of the 4B-5A substrate. The product is depicted with its van der Waals volume. Major resistance mutations are shown as sticks. The direct polar contact between the substrate and Arg123 is shown as a red dash (PDB id 3M5N).

Resistance Profile against BILN In in vitro studies conducted by Jörg Timm at the University Clinical Center Essen, the following resistance profile has been found (figure 4.14) (unpublished data):

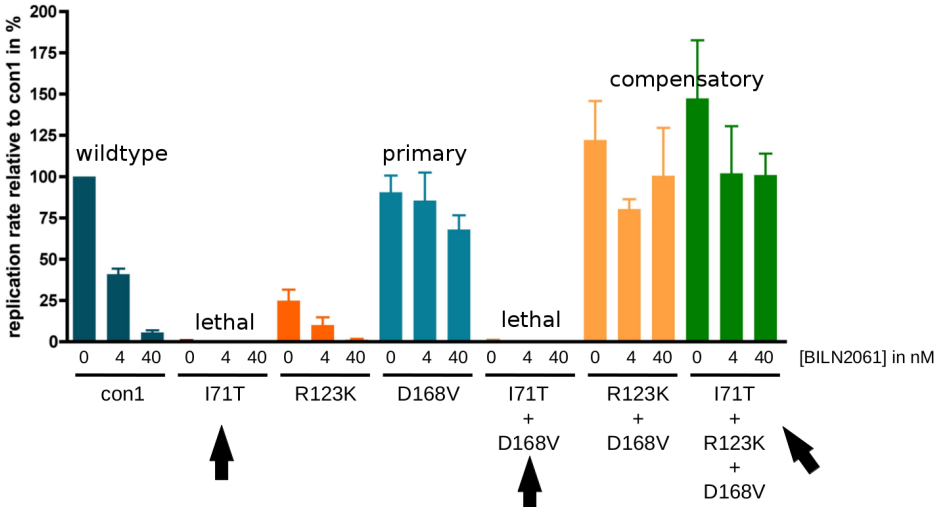


Figure 4.14: Resistance to the HCV protease inhibitor BILN2061. The profile shows the replication rate for each protease variant in dependence to different drug concentrations. Rightmost is the data of the wildtype (con1) as reference. Other bar plots show the resistance profile of different combinations of mutations that were investigated in in vitro studies.

It is experimental evidence that the mutation I71T promotes viral fitness when occurring along with two resistance mutations R123K and D168V in

HCV protease while it is lethal to the virus as a single mutation.
Given the primary resistance mutation D168V, under ongoing drug pressure, two further mutations, R123K and I71T, emerge leading to successively increased fitness. Surprisingly, a R123K single mutant has significantly decreased fitness and a I71T single mutant is not viable at all. Also the D186V/I71T double mutant is not viable. Only the triple mutant is viable when the I71T mutation is present. Because the two further mutations are not resistance mutations by themselves, it seems that they compensate the fitness-decreasing effect of the primary resistance mutation in some way. An astonishing fact is, that the I71T mutation is situated roughly at the opposite of the molecule from the resistance mutations. Characteristically, long-range effects are more difficult to explain from any type of analysis. On the other hand, much could be learned if the full compensatory effect can be explained.

Resistance Mechanism against BILN A circumstance that hampers the development of inhibitors for the HCV protease is its shallow binding site. The first proof-of-concept inhibitor against NS3-4A was the macrocyclic BILN2061 (BILN) [161]. It was found to induce, withing 48 h, 2-3 log10 viral load reduction in HCV infected patients. It was, however, stopped in a Phase II clinical trial (in 2005) due to drug-induced cardiac toxicity. More recently, other PIs have emerged such as VX-950 and SCH-6. However, BILN-2061 is still a reference compound for HCV-related PIs [162].
Courcambeck et al [162] have described the molecular mechanisms of four resistance mutations R155Q, A156T, D168A and D168V that had been found in vitro previously [163, 164, 165] (figure 4.15). The authors distinguish a direct mechanism for R155Q and A156T because these residues physically contact the inhibitor, and an indirect mechanism for D168A and D168V because Asp168 interacts with residues connected to the drug (Arg155 and Arg123). Arg123 and Val158 define a solvent exposed binding pocket for BILNs hydrophobic cyclopentyl moiety. Arg123 participates in the orientation of this moiety towards Val158. Arg155 exhibits both hydrophilic (alkyl side chain) and polar (guanidinium side chain extremity) characteristics and contracts with BILNs P2 moiety these two types of interactions together with additional π-π interactions resulting in a strong important contact. R155Q modifies the hydrophobic and electrostatic environments of BILNs S2 binding pocket. Gln155 is uncharged, whereas Arg155 is positively charged on its guanidinium moiety. The electrostatic interaction between Arg155 and BILN are therefore disrupted in the R155Q mutant. In the case of the A156T mutation, the molecular mechanism of viral resistance to BILN relies on a steric conflict. D168A prevents the formation of two salt-bridges between Asp168 and Arg123 as well as between Asp168 and Arg155. While Asp168 does not significantly interact with BILN, it stabilizes the conformations of

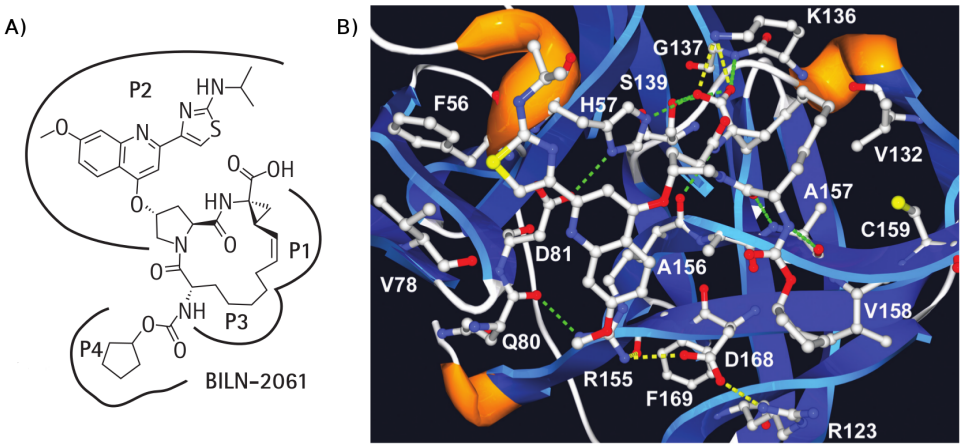


Figure 4.15: Resistance against BILN. A) Two dimensional chemical structure of BILN with schematic representation of its P1 to P4 binding moieties. **B)** HCV protease complexed with inhibitor BILN. BILN, active site residues and residues relevant to inhibitor binding are shown as ball-and-stick. Hydrogen bonds are highlighted as green and salt-bridges and electrostatic interactions as yellow dashes. The figure is adapted from Courcambeck et al. 2006 [162].

both Arg155 and Arg123. In the D168A mutant, Arg123 is exposed to the solvent and no more orients the BILN cyclopentyl moiety towards the Val158. D168V introduces some additional unfavorable van der Waals interactions (steric repulsion) with Arg155 and with R123 compared to D168A. Hence, R123 is even more solvent-exposed and no more orients Val158 to make the van der Waals contact to BILN.

Putative effect of the R123K mutation Next, we consider the mutation R123K in the context of substrate binding (figure 4.16). In the D168V single mutant, Arg123 is no more pre-positioned by D168 and is free to move making a 123-P6 contact harder to be established, possibly leading to a loss of fitness. With the shorter lysine side-chain in the D168V/R123K double mutant, formation of such a contact becomes more probable again. Additional van der Waals packing of Val168 with P4-Val may strengthen the contact and improve the positioning of the substrate in a slightly different mode. On the contrary, R123K is unfavorable as a single mutation because the lysine side-chain is too short giving rise to a steric conflict of D168 with P4-Val and preventing contact formation. This might be the reason why the R123K single mutant has low fitness with and without drug pressure.

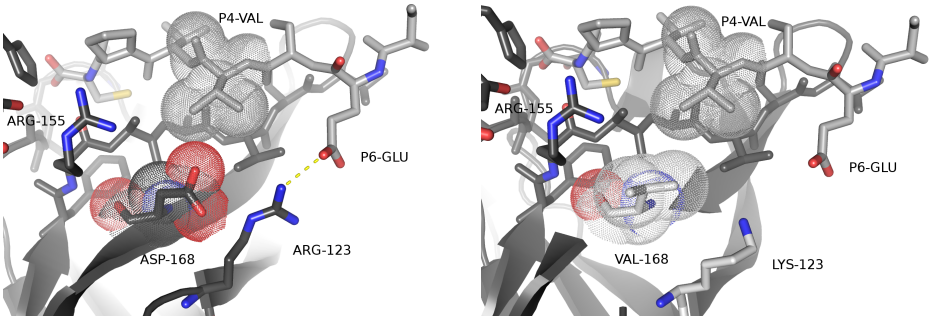


Figure 4.16: Protease-Substrate Contact 123-P6. The left-hand side figure shows the wildtype in which the R123-P6 contact is established (yellow dotted line). In the right-hand side figure, D168V and R123K mutations are indicated by simple side-chain substitutions. Van der Waals volumes of residue 168 and P4-Val are indicated as dotted spheres.

The compensatory mutation I71T was first mentioned by Lu et al [165]. "How mutations at these non-active-site positions could impact BILN 2061 binding, if there is any effect, is unclear. The location of residues 71, 72, and 88 near the NS4A activation protein suggests that they may play a role in influencing enzyme activation, but further experimentation would be required to evaluate this possibility." The mechanism was however so far not further investigated. It has been stated that T72I and P88L possibly have a comparable effect as I71T, presumably increasing the E1/F1 β-sheet flexibility influencing the His57-coordinating Asp81.

Using a multiple testing approach on multiple trajectories To perform a robust analysis of protein dynamics properties, especially when investigating distal effects of mutations [166], multiple simulation approaches (or ensemble approaches) were shown to be useful in diverse MD simulation studies including conformational sampling in general [167, 168], rigorous [169, 170, 171] and end-state [120, 112, 143] free energy calculations and mechanistic trajectory analysis [166, 157]. In the majority of these studies, the production phases of the several simulations are concatenated to a macro-trajectory. Given that the single simulations get trapped in energy minima and are thus not independent from each other, this approach might be disadvantageous for t-testing where the independence of values is a prerequisite. This may lead to a large number of false positives when analyzing geometric features because the distributions get unrealistically sharp. Instead, it might be an improvement to regard only the values from simulations with varying initial conditions to be truly independent from each other.

4.4.2 Methods

Setup of MD Simulations A set of MD simulations was conducted. All simulations were based on the HCV protease structure complexed with its 4B5A N-terminal product (PDB id 3M5N). This structure is one of the highest resolution structures of HCV protease reported to date [16]. It contains four protease molecules and three of them have the N-terminal product bound (chains F, G and H). Both apo and holo type were considered. In a first set of simulations, the inactivating Ala was not remodelled to serine because I wanted the initial structure to stay close to the crystal structure. However, in the course of analysis, the impression emerged that the active site serine is crucial to the compensatory effect. Hence, in a second run, it was planned to model the apo type as an active protease and the holo type as intermediary state where the products C-terminus is covalently bound to the catalytic serine as illustrated in Barbato et al. [160]. The active apo state simulations were accomplished and used for analysis. However, for the holo state, I would have needed the Amber software version 12 which makes it possible to use restraints in the GPU version of pmemd (pmemd.cuda). For both apo and holo types, eight receptor variants, the wildtype, single mutants D168V, R123K and double mutant D168V/R123K, with and without the I71T mutation, were considered. For each of these 16 systems, 12 independent simulations, F, G and H complexes, each with four different velocity random seeds, were performed to provide a means to figure out statistically relevant differences.

Preparation of Complexes and MD Simulation From all template structures, the poly-lysine tail was removed. All crystal waters were also removed except the bridging water between the His149 and the zinc ion. The zinc coordination site was modelled using the cationic dummy atom method [172], i.e. the 2+ charge of the zinc ion is not commonly modelled as a point charge on the atom center but distributed on four tetrahedrally liganded dummy cations with no mass. This approach mimics the zinc's $4s4p^3$ orbitals that accommodate lone-pair electrons and it has been shown to preserve best the geometric arrangement of protein regions with bounded zinc ions. For the apo type, the product peptide was removed and for the active state simulations Ala139 was resubstituted by a serine. The catalytic His57 is delta-N monoprotonated. All the mutations (S1, S2, DT w/o I71T) were built by side-chain substitution conducted using the Pymol mutagenesis feature [149].
All protein residues (except the ones constituting the zinc coordination site) were parameterized using the Amber ff99SBildn forcefield [32, 35, 173]. The complexes were solvated in octahedral boxes of TIP3P water molecules [43] with a minimum distance of 10 Å between the solute and the box boundary. Amber11 [79] was used for all simulations. Water, hydrogens and all built

atoms were minimized by 100 steps steepest descent. Water and hydrogens were then equilibrated over 20 ps NVT. Side-chains were equilibrated over 30 ps NVT. The whole system was equilibrated 50 ps NVT and 50 ps NPT. Then, 5 ns production simulations were performed at 300 K, 1 bar and 2 fs integration time step in the NPT ensemble using a Langevin thermostat [69] with a collision frequency of 5 ps^{-1} and a Berendsen barostat [68]. Electrostatic interactions were calculated using the PME [29] summation scheme. The cutoff for non-bonded interactions was set to 8 Å. Snapshots were recorded any 1 ps. The total number of simulations is 192 and the total production simulation time is 960 ns. All simulations were performed on a compute node with two Tesla M2050 GPUs using pmemd.cuda.

Significant Differences of Distances and RMSF Several data sets of distances and residue RMSF were investigated using the significance analysis of microarrays (SAM) method (subsubsection 2.5.3) for multiple testing. The distance data sets were defined considering the chemico-physical context. To find out if the overall geometry has changed C_α distances were analyzed. Three data sets of C_α distances with ranges $<$ 4 Å (short), 10-12 Å (middle) and 24-25 Å (long) were defined. Hydrogen bonds and salt bridges were analyzed by defining sets of H-O and N-O distances, respectively, both in the range $<$ 4 Å. The integrity of the hydrophobic core was investigated by defining a set of C-C atoms of exclusively hydrophobic residues in the distance range $<$ 4.6 Å. In principle, any atom-atom distance could be used as feature, but this would be much more compute-intensive than using only subsets of distances. Instead, the range of distances for the C_α atoms has been chosen to have a reasonable coverage of distances representing the geometry. The sets are illustrated in figure 4.17.
The data from all distance timeseries from each 5 ns simulation was subdivided into three blocks, the first nanosecond, nanoseconds 2-3 and nanoseconds 4-5. This is in order to assess if the set of significant differences obtained by the significance analysis changes over time and in particular if the effects introduced by the modelling of the point mutations needs further time to equilibrate. From these blocks, the averages and variances were calculated from each single simulation. To figure out differences induced by point mutations, the data has been rearranged into SAM input data sets that each oppose 48 simulations with and 48 simulations without the mutation under consideration, namely with and without the I71T, R123K and D168V mutations. Further compilations were also considered, namely sets opposing the three single mutants versus wildtype, lethal versus viable variants, double mutant versus wildtype, triple mutant versus wildtype and triple mutant versus wildtype. There is however in these setups a considerably lower amount of data available, only 12 versus 12 simulations rendering statistical analysis less powerful.

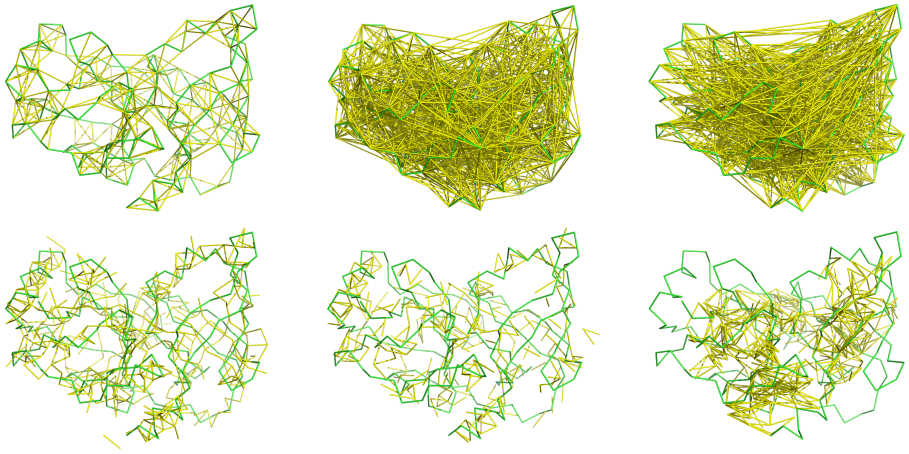


Figure 4.17: Distance sets for the SAM analysis. Figures on top from left to right show the short, middle and long range C_α distance sets. Figures on bottom show from left to right the H-O, N-O and hydrophobic C-C distances.

Significant Differences of Concerted Motions Correlated motions from a distance covariance (DiCC) were measured as described in subsection 2.5.4. The analysis was based on all C_α-C_α atom pairs. The sets of DiCC values were used as feature sets for a SAM analysis analogous to the distance analysis described above.

4.4.3 Results and Discussion

Stability of the Simulations The general stability of simulations has been assessed by the C_α RMSD (figure 4.18, black curves). RMSD values are in the range of 1-1.5 Å with a moderate increase and divergence. This means that the simulations are stable and all simulations naturally tend to move away from the geometry of the starting structure exploring conformational space.

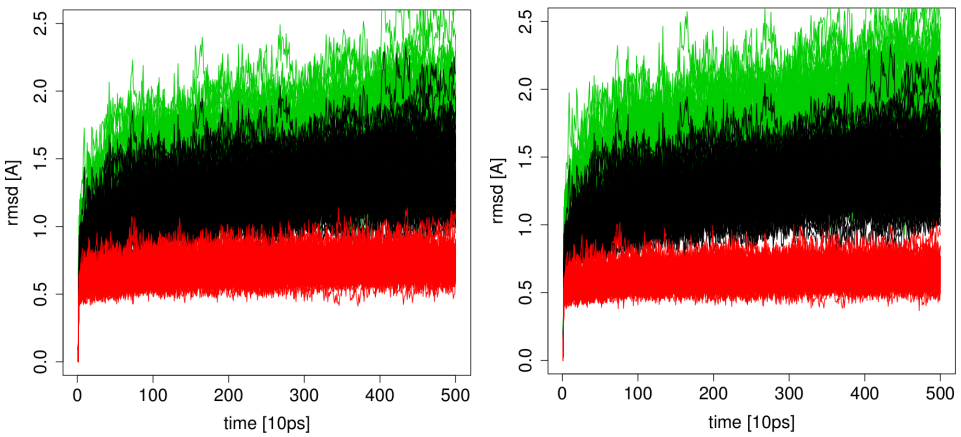


Figure 4.18: Root Mean Square Deviation. RMSD of MD snapshots with respect to the starting structure based on a common C_α alignment (left figure) and based on an alignment to residues with RMSF < 0.5 Å (right figure). RMSD from all residues, from residues with RMSF < 0.5 Å and residues with RMSF > 0.5 Å are colored black, red and green, respectively.

The calculation of DiCC values was based on an alignment to residues with low RMSF in order to increase the positional variances of exterior residues. The cutoff for the definition of residues with low RMSF was chosen as the maximum of the RMSF density function (figure 4.19, right). These residues resemble roughly the stable core of the protein (figure 4.20). While the RMSD difference of the low-RMSF alignment to a common C_α alignment is marginal ($2 * 10^{-3}$ Å), there is at least a modest increase of RMSD in the exterior protein residues (figure 4.18, green curve) that might lead to the detection of an increased number of significant DiCC differences.

Root Mean Square Fluctuation Figure 4.19 shows on the left a RMSF comparison of apo and holo simulations. The residue fluctuations in the region 120-152 are increased for the apo simulations. This means that these residues are more free to move because the substrate is not bound. Further comparisons were accomplished in the search for RMSF differences that are induced by the mutations under investigation. There are however no significant differences to be found.

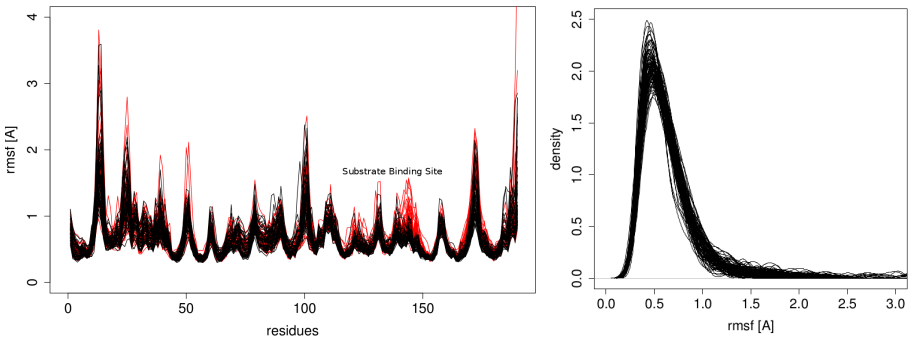


Figure 4.19: Root Mean Square Fluctuations. Left) RMSF from 48 apo and 48 holo simulations are colored black and red, respectively. Right) RMSF density function from all apo simulations.

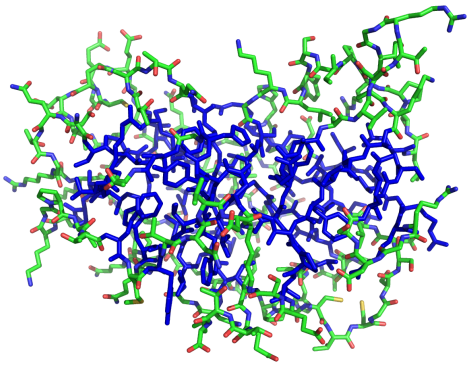


Figure 4.20: Low RMSF Residues. Residues with low and high RMSF defined by a cutoff of 0.5 Å are colored blue and green, respectively.

SAM Trajectory Analysis - Distances and RMSF Distance averages, distance variances and RMSF were analyzed using the SAM method based on the data from the apo simulations with a correct active site geometry. The result of the significance analysis is illustrated in figure 4.21.
Generally, it can be stated that virtually all differences that are called significant by SAM are either in spatial proximity to the investigated mutations or, if the differences are over a long range, they have a traceable reference to it indicating a low level of false positives. This is not trivial because typical simpler feature analysis suffer from false positives caused by large motions of regions with high thermal fluctuation, e.g. loop regions. Such type of false positives is however successfully suppressed by the applied procedure. Furthermore, the sets of significant differences based on data from the three different time intervals are pretty much comparable. Taking into account, that the apo simulations were based on the holo structure with the substrate removed, relaxation effects in the first time interval can be expected. Especially the pattern of significant differences for the first time interval of the D168V mutation is considered to be some sort of relaxation effect. For

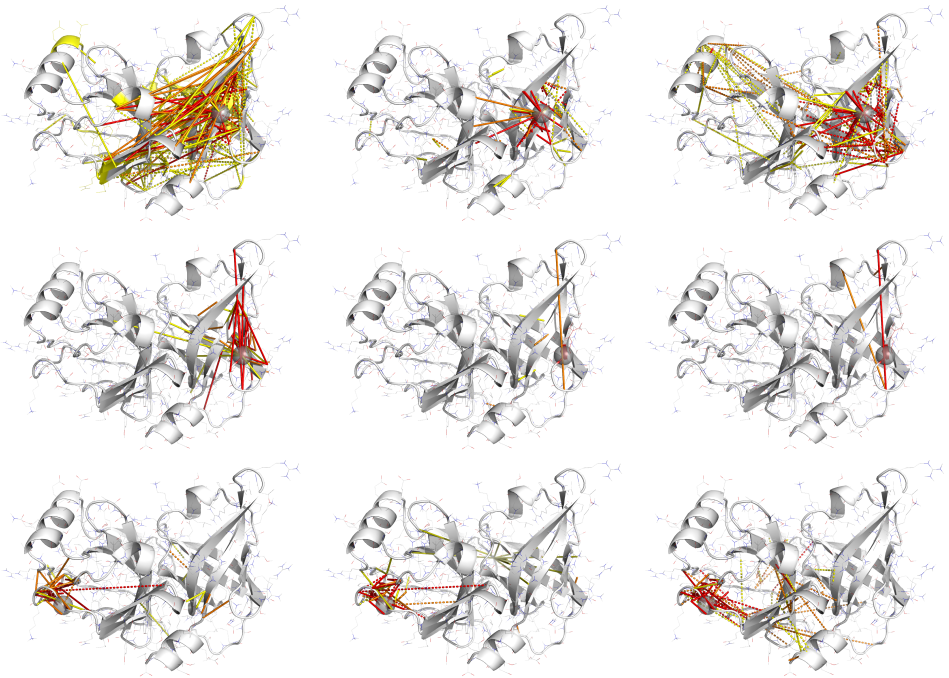


Figure 4.21: Significant Distance and RMSF Differences. Significant differences of average distances (solid lines), distance variances (dashed lines) and RMSF (residue coloring) are shown at FDR tresholds 0.01 (red), 0.05 (orange) and 0.1 (yellow). The location of each point mutation under investigation is shown as a half-transparent grey sphere. The arrangement of figures shows from left to right the results for the first nanosecond, nanoseconds 2-3 and nanoseconds 4-5 and top to bottom the mutations D168V, R123K and I71T.

that reason, we focus on the analysis from the last time interval for all three mutations.

Sign and extend of significant distance differences are illustrated in figure 4.22.
D168V: In the D168V mutant, the distances from the 168-C_α atom to the opposite β-sheet and further atoms of the active site are decreased indicating a slight size reduction of the crevice between the two barrels. Furthermore, some long ranged distance variances are increased.
R123K: Two distances are decreased in the R123K mutant. I refrain from an interpretation of these because the role of the R123K mutation is more convincingly explained in the context of substrate binding (paragraph 4.4.1).
I71T: The I71T mutation seems to loosen its local environment reflected by some increased local distances. It furthermore slightly decreases the fluctu-

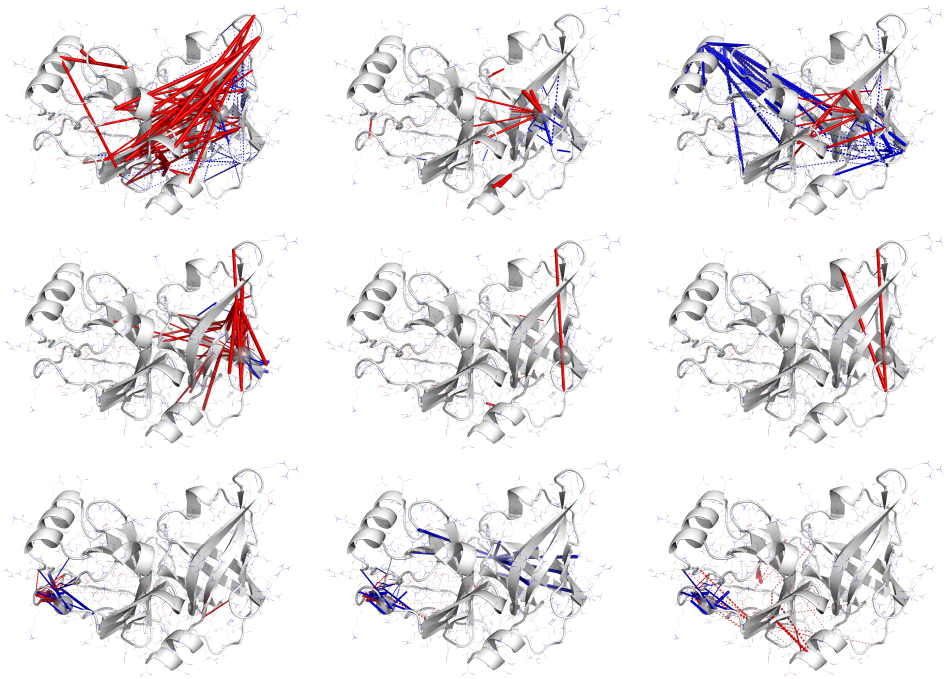


Figure 4.22: Sign and Extend of Significant Distance Differences. Sign and extend of significant differences of average distances (solid lines) and distance variances (dashed lines). Longer distances and higher variances are colored blue, shorter distances and lower variances are colored red. The width of dashes represents the extend to which the features have changed. The location of each point mutation under investigation is shown as a half-transparent grey sphere. The arrangement of figures shows from left to right the results for the first nanosecond, nanoseconds 2-3 and nanoseconds 4-5 and top to bottom the mutations D168V, R123K and I71T.

ation towards the active site reflected by some decreased distance variances. Considering the time evolution, it is probable that this effect would be more pronounced if the simulations would be prolonged. The local effect of I71T is further analyzed in paragraph 4.4.3.

Distance Analysis using randomForest The same dataset that was used for SAM can also be used as input for the machine-learning method randomForest [174, 175]. Briefly explained, randomForest classifies the features from a majority vote of an ensemble of decision tree outcomes. The features can then be ranked by a measure indicating the importance for the classification result, e.g. the mean decrease Gini. Figures 4.23 and 4.24 show the results using randomForest for the D168V and I71T mutations,

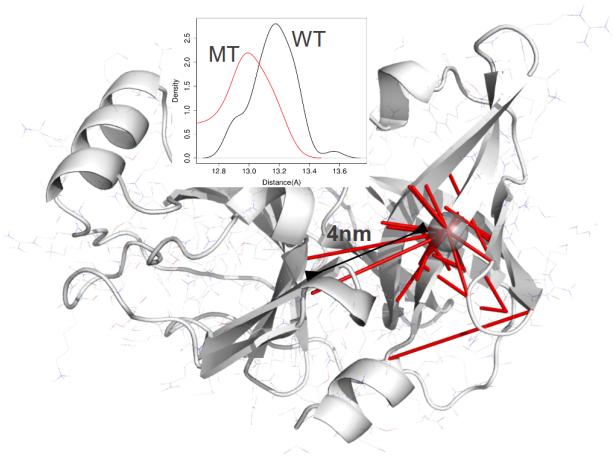


Figure 4.23: Significant distance differences for the D168V mutation using randomForest. Red lines indicate distance differences at a Mean Decrease Gini treshold of 0.28. The black arrow indicates the induced distance shift.

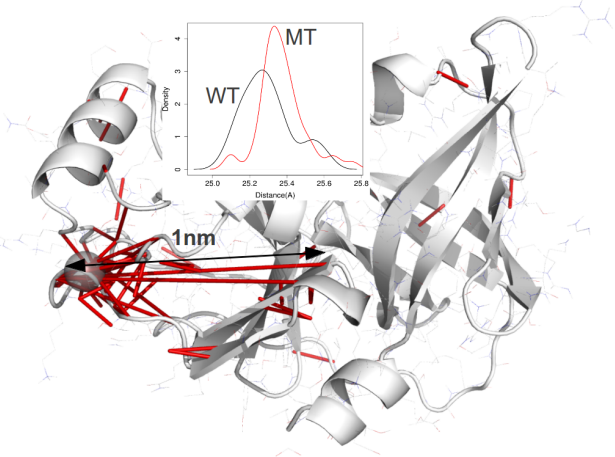


Figure 4.24: Significant distance differences for the I71T mutation using randomForest. Red lines indicate distance differences at a Mean Decrease Gini treshold of 0.28. The black arrow indicates the induced distance shift.

respectively. Only the average distances were used for this analysis.
The randomForest result for the I71T mutation shows two new significant distance differences that are directed towards the hairpin that coordinates the catalytic Histidine in the active site and which were not found by SAM. These distances are particularly interesting because they suggest a connection to the effects introduced by the D168V mutation. The average distances from the D168V mutation site to the hairpin is decreased by roughly 4 nm (figure 4.23) while the average distance from the I71T mutation site to the hairpin is increased by roughly 1 nm (figure 4.24). The full truth concerning the compensatory effect is probably quite complicated, because the I71T/D168V double mutant is also lethal to the virus and the role of the R123K mutation with regard to the I71T mutation remains unclear. The two mutations D168V and R123K probably have an effect that slightly modulates substrate binding and this together with the distance shift effects described here could possibly yield a functioning protease.

Significant differences between apo and holo structures The SAM analysis has also been performed for a comparison of apo versus holo simulations (figure 4.25). This analysis was performed only with short and no long ranged distances and no distance variances. It is clear that substrate binding induces relatively large structural changes. Hence, this comparison may serve as a proof of concept. SAM clearly detects structural changes that are induced by the binding of the substrate. Importantly, this test case shows that significant RMSF differences are also reliably detected if present.

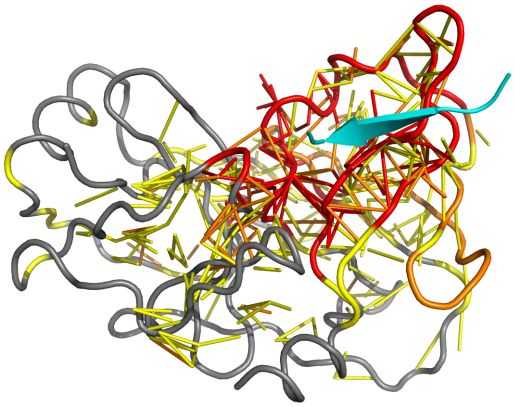


Figure 4.25: Significant Distance and RMSF differences - Apo versus Holo. Significant differences of average distances (solid lines) and RMSF (residue coloring) are shown at FDR tresholds 0.01 (red), 0.05 (orange) and 0.1 (yellow). The substrate is colored cyan.

Significant DiCC differences Significant differences of concerted motions were analyzed by a SAM analysis on DiCC [94] values. The result for I71T is illustrated in figure 4.26. It can be seen that there are only few differences at FDR rates up to 0.1. With increasing FDR rate up to 0.7, there are many more significant differences that have reasonable reference to I71T. The result suggests that the concertedness of motions is increased between these residues. It is however decreased for some residues in the location of the resistance mutations. Only the difference colored red is significant at a FDR rate 0.7 and the other two (colored orange and yellow) were manually added to check if these are also in the same location. This is true only for the orange one. The yellow one might be a false positive. Generally, it is surprising that these DiCC differences are detected while there are no observable differences in the residue RMSFs.

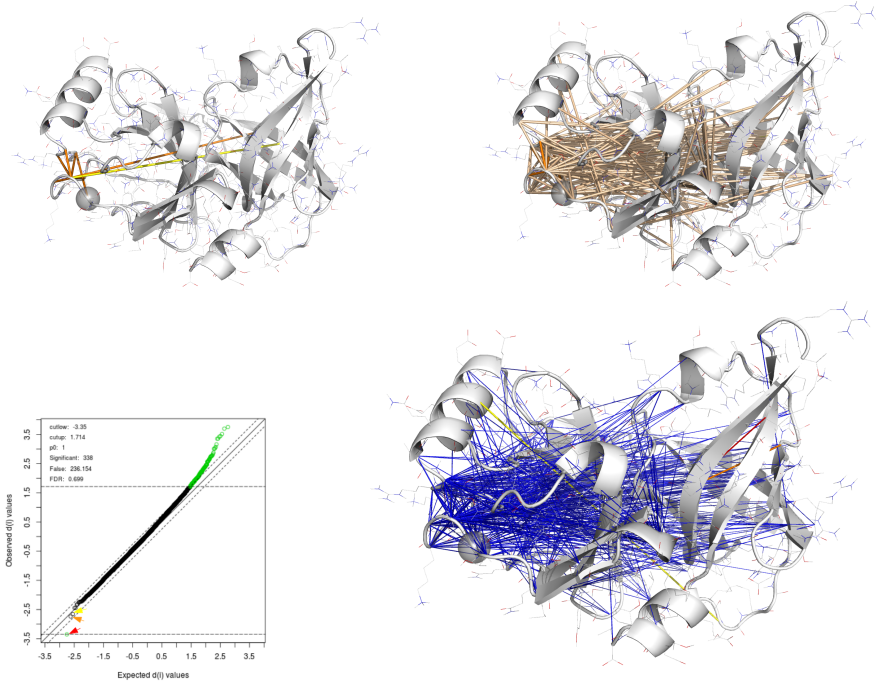


Figure 4.26: Significant DiCC differences. The top left figure shows significant DiCC differences at FDR rates 0.05 and 0.1 in orange and yellow, respectively. The top right figure shows additional significant DiCC differences at a FDR rate 0.7 colored wheat. The bottom right figure shows decreased DiCC differences in blue and increased DiCC differences in red, orange and yellow. The bottom left figure shows corresponding SAM plot.

Local effect of the I71T mutation Threonine 71 atoms OG1 and HG1 introduce both a H-bond acceptor and donor into the system (figure 4.27). From visual inspection of the trajectories it is obvious that these two atoms vividly change their interaction partners and the bond between Gly69-O and Ser66-HG tends to break. To grasp this quantitatively the bond breakage and formation of this bond was measured. This bond is present in both systems with and without the I71T mutation and can therefore be compared. A distance cutoff was set to 3 Å and the number of formations counted. It turns out that bond breakage and formation is 6-fold increased for the 71T systems.

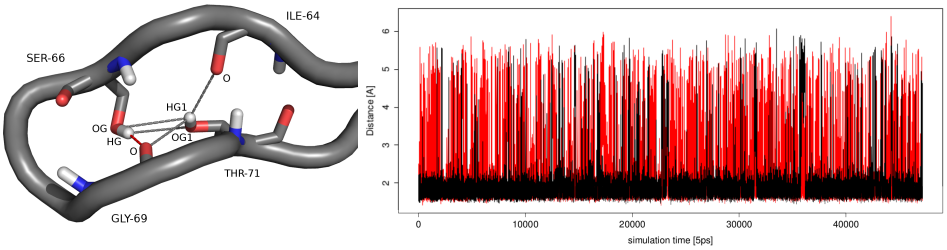


Figure 4.27: Bond Breakage and Formation induced by the I71T Mutation. Right) Hydrogen bond acceptors and donors. Left) Distance time series of the bond between Gly69-O and Ser66-HG for all systems with I71 (black) and 71T (red).

4.4.4 Conclusion

Using a multiple testing approach, SAM or randomForest, on ensembles of trajectories revealed at least some indications of local and long range effects for the I71T mutation. However, no satisfactory explanation for its lethal or fitness-increasing effects can be derived from this analysis. The synergistic effect of all three mutations might be too subtle to be equilibrated by simulations of 5 ns length. Therefore, longer simulations would be needed. Roughly four times more simulations would be needed to perform comparisons between pairs of the eight mutational species to yield comparable informations as when comparing simulation sets with and without one of the mutations. An attempt to bootstrap the data resulted in a large number of false positives.
Generally, my impression is that a few microseconds of data would be sufficient to yield a more detailed explanation. Sets of 50 times 10 ns for each species would probably be sufficient. For the eight species, simulation time would some up to 4 microseconds for the apo state and 8 microseconds if an analysis of the holo state would be included. This is roughly ten times more simulation time that has been used for the present analysis.
This study has at least shown that a multiple testing approach on simulation ensembles improves the detection of mutation induced geometric features compared to analysis of data from only one simulation for each system. The reason for this is that snapshots taken from one simulation are highly correlated even after 10 nanoseconds [120]. Several short simulations differentially perturbed by different initial conditions yield a more diverse set of snapshots. Therefore, false positives in regions with high thermal motions are suppressed.
It is certainly disadvantageous that for computational reasons the choice of distance subsets was restricted to these relatively narrow distance intervals. Increasing the ranges of these intervals or ideally analyzing all distances would give a clearer picture of induced geometric differences. Regarding

the problem of equilibrating effects of modelled point mutations when using multiple simulations, it would probably be a good idea to run a smaller number of longer master trajectories that initially equilibrate these effects and start the multiple simulations from the last snapshots of these master trajectories.

Future Directions

MMPBSA has often been regarded as a suitable trade-off between accuracy and efficiency for binding affinity calculations because only end states have to be simulated. However, a correct ranking of binding affinities is typically not obtained using conventional MD mainly because a Boltzmann-averaged ensemble is not achieved. The usefulness of more advanced sampling algorithms such as accelerated molecular dynamics (aMD) or free energy guided sampling (FEGS) for MMPBSA calculations has not been explored yet and this will have to be assessed in the future. Given the fact that an ensemble of simulations is needed to yield sufficient sampling to get free energy estimates converged with respect to varying initial conditions, the computational effort that has to be undertaken approaches or even exceeds that of rigorous energy calculations. Rigorous free energy calculations on the contrary have the potential to determine relative free energy differences induced by point mutation quite precisely. As a consequence, rigorous methods have to be preferred if the introduced differences are small, - this is expected for numerous relevant mutations altering the binding energy of a protein receptor to inhibitors. In the last years, the Bennett Acceptance Ratio method (BAR) [176] has found its way to free energy calculations of biomolecules [177]. It has been proven to be the most efficient scheme to estimate the free energy from lambda-perturbed molecular dynamics simulations and is expected to replace TI on the long run. Compared to TI, fewer lambda steps are needed for convergence and furthermore, BAR has an intrinsic error estimate.

Given that a high amount of simulation data is needed for both free energy calculations and to accurately derive geometric effects of point mutations, a possible perspective would be to do both calculations on the same data. A comprehensive strategy for the investigations of resistance mutations on both an energetic and geometric basis could be as follows: First, a BAR calculation could be performed to calculate the free energy difference between wildtype and mutant inhibitor complexes. For further analysis, the correct determination of the free energy difference should be a precondition. It is assumed, that if the free energy difference comes out right, the geometric rearrangement introduced by the mutation is correctly simulated over the alchemical transformation path. Subsequently, geometric features, e.g. a distances, RMSF or DiCC values could be analyzed over the transformation path. To do so, distance distributions from all lambda steps could be evaluated with corresponding averages and standard deviations. The distance data could then be analyzed comparable to a multiple testing approach.

References

[1] UNAIDS, 2011, *World AIDS Day Report 2011: How to Get to Zero: Faster, Smarter, Better* 1

[2] Shepard, C. W., Finelli, L., Alter, M. J., 2005, Global epidemiology of hepatitis C virus infection, *Lancet Infect Dis*, 5(9):558–567 1

[3] Pawlotsky, J. M., Chevaliez, S., McHutchison, J. G., 2007, The hepatitis C virus life cycle as a target for new antiviral therapies, *Gastroenterology*, 132(5):1979–1998 1

[4] *HIV Lifecycle*, National Institute of Allergy and Infectious Diseases HIV/AIDS, http://www.niaid.nih.gov/topics/HIVAIDS 1, 3

[5] Briggs, J. A., Wilk, T., Welker, R., Krausslich, H. G., Fuller, S. D., 2003, Structural organization of authentic, mature HIV-1 virions and cores, *EMBO J.*, 22(7):1707–1715 1

[6] Arts, E. J., Hazuda, D. J., 2012, HIV-1 Antiretroviral Drug Therapy, *Cold Spring Harb Perspect Med*, 2(4):a007161 2

[7] Choudhary, S. K., Margolis, D. M., 2011, Curing HIV: Pharmacologic approaches to target HIV-1 latency, *Annu. Rev. Pharmacol. Toxicol.*, 51:397–418 2

[8] Beerenwinkel, N., Daumer, M., Oette, M., Korn, K., Hoffmann, D., Kaiser, R., Lengauer, T., Selbig, J., Walter, H., 2003, Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes, *Nucleic Acids Res.*, 31(13):3850–3855 4, 48

[9] Dybowski, J. N., Heider, D., Hoffmann, D., 2010, Prediction of co-receptor usage of HIV-1 from genotype, *PLoS Comput. Biol.*, 6(4):e1000743 4

[10] Shenderovich, M. D., Kagan, R. M., Heseltine, P. N., Ramnarayan, K., 2003, Structure-based phenotyping predicts HIV-1 protease inhibitor resistance, *Protein Sci.*, 12(8):1706–1718 4

[11] Shen, C. H., Wang, Y. F., Kovalevsky, A. Y., Harrison, R. W., Weber, I. T., 2010, Amprenavir complexes with HIV-1 protease and its drug-resistant mutants altering hydrophobic clusters, *FEBS J.*, 277(18):3699–3714 4

[12] Liu, F., Kovalevsky, A. Y., Tie, Y., Ghosh, A. K., Harrison, R. W., Weber, I. T., 2008, Effect of flap mutations on structure of HIV-1 protease and inhibition by saquinavir and darunavir, *J. Mol. Biol.*, 381(1):102–115 4

[13] Louis, J. M., Zhang, Y., Sayer, J. M., Wang, Y. F., Harrison, R. W., Weber, I. T., 2011, The L76V drug resistance mutation decreases the dimer stability and rate of autoprocessing of HIV-1 protease by reducing internal hydrophobic contacts, *Biochemistry*, 50(21):4786–4795 4

[14] Prabu-Jeyabalan, M., Nalivaika, E., Schiffer, C. A., 2002, Substrate shape determines specificity of recognition for HIV-1 protease: analysis of crystal structures of six substrate complexes, *Structure*, 10(3):369–381 4

[15] Chellappan, S., Kairys, V., Fernandes, M. X., Schiffer, C., Gilson, M. K., 2007, Evaluation of the substrate envelope hypothesis for inhibitors of hiv-1 protease, *Proteins: Structure, Function, and Bioinformatics*, 68(2):561–567 4

[16] Romano, K. P., Ali, A., Royer, W. E., Schiffer, C. A., 2010, Drug resistance against HCV NS3/4A inhibitors is defined by the balance of substrate recognition versus inhibitor binding, *Proc. Natl. Acad. Sci. U.S.A.*, 107(49):20986–20991 4, 54, 55, 60

[17] Altman, M. D., Ali, A., Reddy, G. S., Nalam, M. N., Anjum, S. G., Cao, H., Chellappan, S., Kairys, V., Fernandes, M. X., Gilson, M. K., Schiffer, C. A., Rana, T. M., Tidor, B., 2008, HIV-1 protease inhibitors from inverse design in the substrate envelope exhibit subnanomolar binding to drug-resistant variants, *J. Am. Chem. Soc.*, 130(19):6099–6113 4

[18] Munos, B., 2009, Lessons from 60 years of pharmaceutical innovation, *Nat Rev Drug Discov*, 8(12):959–968 5

[19] Durrant, J. D., McCammon, J. A., 2011, Molecular dynamics simulations and drug discovery, *BMC Biol.*, 9:71 5

[20] Borhani, D. W., Shaw, D. E., 2012, The future of molecular dynamics simulations in drug discovery, *J. Comput. Aided Mol. Des.*, 26(1):15–26 5

[21] Harvey, M. J., De Fabritiis, G., 2012, High-throughput molecular dynamics: the powerful new tool for drug discovery, *Drug Discov. Today*, 17(19-20):1059–1062 5

[22] Gohlke, H., Kiel, C., Case, D. A., 2003, 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 5, 32

[23] Hou, T., Yu, R., 2007, Molecular dynamics and free energy studies on the wild-type and double mutant HIV-1 protease complexed with amprenavir and two amprenavir-related inhibitors: mechanism for binding and drug resistance, *J. Med. Chem.*, 50(6):1177–1188 5, 32, 39

[24] Cai, Y., Schiffer, C. A., 2010, Decomposing the energetic impact of drug resistant mutations in HIV-1 protease on binding DRV, *J Chem Theory Comput*, 6(4):1358–1368 5, 32, 33

[25] Cai, Y., Schiffer, C., 2012, Decomposing the energetic impact of drug-resistant mutations: the example of HIV-1 protease-DRV binding, *Methods Mol. Biol.*, 819:551–560 5, 33

[26] Allen, M. P., Tildesley, D. J., 1987, *Computer Simulation of Liquids*, Clarendon, Oxford 7

[27] Swope, W. C., Andersen, H. C., Berens, P. H., Wilson, K. R., 1982, A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters, *J. Chem. Phys.*, 76:637–649 7

[28] Adcock, S. A., McCammon, J. A., 2006, Molecular dynamics: Survey of methods for simulating the activity of proteins, *Chem. Rev.*, 106:1589–1615 7

[29] Darden, T., York, D., Pedersen, L., 1993, Particle mesh Ewald - an N.Log(N) method for Ewald sums in large systems, *J. Chem. Phys.*, 98:10089–10092 7, 40, 49, 61

[30] Ponder, J. W., Case, D. A., 2003, Force fields for protein simulations, *Adv. Proteins Chem.*, 66:27–85 8, 9

[31] Oostenbrink, C., Villa, A., Mark, A. E., Van Gunsteren, W. F., 2004, A biomolecular force field based on the free enthalpy of hydration and solvation: The gromos force-field parameter sets 53A5 and 53A6, *J. Comput. Chem.*, 25:1656–1676 8

[32] Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., Kollman, P. A., 1995, A second generation force field for the simulation of proteins, nucleic acids, and organic molecules, *J. Am. Chem. Soc.*, 117:5179–5197 8, 35, 38, 60

[33] MacKerell, A. D., Bashford, D., Bellott, 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., Wiórkiewicz-Kuczera, J., Yin, D., Karplus, M., 1998, All-atom empirical potential for molecular modeling and dynamics studies of proteins, *J. Phys. Chem. B*, 102:3586–3616 8

[34] Jorgensen, W. L., Tirado-Rives, J., 1988, The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin, *J. Am. Chem. Soc.*, 110:1657–1666 8

[35] Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., Simmerling, C., 2006, Comparison of multiple amber force fields and development of improved protein backbone parameters, *Proteins: Structure, Function, and Bioinformatics*, 65(3):712–725 8, 37, 60

[36] Duan, Y., Wu, C., Chowdhury, S., Lee, M. C., Xiong, G., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J., Kollman, P., 2003, A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations, *Journal of Computational Chemistry*, 24(16):1999–2012 8, 39, 48

[37] Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., Case, D. A., 2004, Development and testing of a general Amber force field, *J. Comput. Chem.*, 25:1157–1174 8, 37, 39

[38] Lange, O. F., van der Spoel, D., de Groot, B. L., 2010, Scrutinizing molecular mechanics force fields on the submicrosecond timescale with NMR data, *Biophys. J.*, 99(2):647–655 9

[39] Beauchamp, K. A., Lin, Y.-S., Das, R., Pande, V. S., 2012, Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements, *J. Chem. Theory Comput.*, 8(4):1409–1414 9

[40] Ponder, J. W., Wu, C., Ren, P., Pande, V. S., Chodera, J. D., Schnieders, M. J., Haque, I., Mobley, D. L., Lambrecht, D. S., DiStasio, R. A., Head-Gordon, M., Clark, G. N. I., Johnson, M. E., Head-Gordon, T., 2010, Current Status of the AMOEBA Polarizable Force Field, *The Journal of Physical Chemistry B*, 114(8):2549–2564 9

[41] Hess, B., van der Vegt, N. F., 2006, Hydration thermodynamic properties of amino acid analogues: a systematic comparison of biomolecular force fields and water models, *J Phys Chem B*, 110(35):17616–17626 9, 11

[42] Jorgensen, W. L., 1981, Quantum and statistical mechanical studies of liquids. 10. Transferable intermolecular potential functions for water, alcohols, and ethers. Application to liquid water, *Journal of the American Chemical Society*, 103(2):335–340 10

[43] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., Klein, M. L., 1983, Comparison of simple potential functions for simulating liquid water, *J. Chem. Phys.*, 79:926–935 10, 11, 39, 49, 60

[44] Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., Hermans, J., 1981, *Intermolecular forces*, Reidel, Dordrecht, 2nd edition 10

[45] Berendsen, H. J. C., Grigera, J. R., Straatsma, T. P., 1987, The missing term in effective pair potentials, *The Journal of Physical Chemistry*, 91(24):6269–6271 10

[46] Horn, H., Swope, W., Pitera, J., Madura, J., Dick, T., Hura, G., Head-Gordon, T., 2004, Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew, *J. Chem. Phys.*, 120:9665–9678 10, 11

[47] van der Spoel, D., van Maaren, P., C., B. H. J., 1998, A systematic study of water models for molecular simulation: Derivation of water models optimized for use with a reaction field, *J Chem Phys*, 108:10220–10231 11

[48] Vega, C., Abascal, J. L. F., Sanz, E., MacDowell, L. G., McBride, C., 2005, Can simple models describe the phase diagram of water?, *Journal of Physics: Condensed Matter*, 17(45):S3283 11

[49] Hu, Z., Jiang, J., 2010, Assessment of biomolecular force fields for molecular dynamics simulations in a protein crystal, *Journal of Computational Chemistry*, 31(2):371–380 11

[50] Vega, C., Abascal, J. L. F., 2011, Simulating water with rigid non-polarizable models: a general perspective, *Phys. Chem. Chem. Phys.*, 13:19663–19688 11

[51] Sitkoff, D., Sharp, K. A., Honig, B., 1994, Accurate calculation of free energies using macroscopic solvent models, *J. Phys. Chem.*, 98:1978–1988 12, 15, 16

[52] Fogolari, F., Brigo, A., Molinari, H., 2002, The Poisson–Boltzmann equation for bimolecular electrostatics: A tool for structural biology, *J. Mol. Recog.*, 15:377–392 13, 14

[53] Bondi, A., 1964, Van der Waals Volumes and Radii, *The Journal of Physical Chemistry*, 68(3):441–451 13

[54] Lu, B. Z., Zhou, Y. C., Holst, M. J., McCammon, J. A., 2008, Recent Progress in Numerical Methods for the Poisson-Boltzmann Equation in Biophysical Applications, *Communications in Computational Physics*, 3(5):973–1009 14

[55] Tsui, V., Case, D. A., 2000, Theory and applications of the generalized born solvation model in macromolecular simulations, *Biopolymers*, 56(4):275–291 14

[56] Born, M., 1920, Volumen und Hydratationswärme der Ionen, *Zeitschrift für Physik*, 1:45–49 14

[57] Bashford, D., Case, D. A., 2000, Generalized Born models of macromolecular solvation effects, *Ann. Rev. Phys. Chem.*, 51:129–152 14

[58] Hawkins, G. D., Cramer, C. J., Truhlar, D. G., 1996, Parametrized Models of Aqueous Free Energies of Solvation Based on Pairwise Descreening of Solute Atomic Charges from a Dielectric Medium, *J. Phys. Chem.*, 100(51):19824–19839 15

[59] Onufriev, A., Bashford, D., Case, D. A., 2004, Exploring protein native states and large-scale conformational changes with a modified generalized born model, *Proteins: Structure, Function, and Bioinformatics*, 55(2):383–394 15

[60] Mongan, J., Simmerling, C., McCammon, J. A., Case, D. A., Onufriev, A., 2007, Generalized Born Model with a Simple, Robust Molecular Volume Correction, *J. Chem. Theory Comput.*, 3(1):156–169 15

[61] Kongsted, J., Söderhjelm, P., Ryde, U., 2009, How accurate are continuum solvation models for drug-like molecules?, *Journal of Computer-Aided Molecular Design*, 23:395–409 15, 33

[62] Hermann, R. B., 1972, Theory of hydrophobic bonding. II. Correlation of hydrocarbon solubility in water with solvent cavity surface area, *J. Phys. Chem.*, 76:2754–2759 15

[63] Tan, C., Tan, Y.-H., Luo, R., 2007, Implicit nonpolar solvent models, *J. Phys. Chem. B*, 111:12263–12274 16

[64] Barone, V., Cossi, M., Tomasi, J., 1997, A new definition of cavities for the computation of solvation free energies by the polarizable continuum model, *J. Chem. Phys.*, 107:3210–3221 16

[65] Floris, F., Tomasi, J., 1989, Evaluation of the dispersion contribution to the solvation energy. a simple computational model in the continuum approximation, *J. Comput. Chem.*, 10:616–627 16

[66] Genheden, S., Kongsted, J., Soderhjelm, P., Ryde, U., 2010, Nonpolar Solvation Free Energies of Protein-Ligand Complexes, *J. Chem. Theory Comput.*, 6(11):3558–3568 16, 32, 33

[67] Genheden, S., Mikulskis, P., Hu, L., Kongsted, J., Soderhjelm, P., Ryde, U., 2011, Accurate predictions of nonpolar solvation free energies require explicit consideration of binding-site hydration, *J. Am. Chem. Soc.*, 133(33):13081–13092 16, 32, 33

[68] Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., Dinola, A., Haak, J. R., 1984, Molecular-dynamics with coupling to an external bath, *J. Chem. Phys.*, 81:3684–3690 17, 40, 49, 61

[69] Wu, X., Brooks, B. R., 2003, Self-guided Langevin dynamics simulation method, *Chem. Phys. Lett.*, 381:512–518 17, 40, 49, 61

[70] Nose, S., 1984, A unified formulation of the constant temperature molecular dynamics methods, *The Journal of Chemical Physics*, 81(1):511–519 17

[71] Hoover, W. G., 1985, Canonical dynamics: Equilibrium phase-space distributions, *Phys. Rev. A*, 31:1695–1697 17

[72] Genheden, S., Ryde, U., 2011, A comparison of different initialization protocols to obtain statistically independent molecular dynamics simulations, *J Comput Chem*, 32(2):187–195 18, 33

[73] Hamelberg, D., Mongan, J., McCammon, J. A., 2004, Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules, *J Chem Phys*, 120(24):11919–11929 18

[74] Pierce, L. C., Salomon-Ferrer, R., Augusto F de Oliveira, C., McCammon, J. A., Walker, R. C., 2012, Routine Access to Millisecond Time Scale Events with Accelerated Molecular Dynamics, *J Chem Theory Comput*, 8(9):2997–3002 18

[75] Doshi, U., Hamelberg, D., 2012, Improved Statistical Sampling and Accuracy with Accelerated Molecular Dynamics on Rotatable Torsions, *Journal of Chemical Theory and Computation*, 8(11):4004–4012 18

[76] Sugita, Y., Okamoto, Y., 1999, Replica-exchange molecular dynamics method for protein folding, *Chemical Physics Letters*, 314(1–2):141–151 18

[77] Cooke, B., Schmidler, S. C., 2008, Preserving the Boltzmann ensemble in replica-exchange molecular dynamics, *J Chem Phys*, 129(16):164112 19

[78] Zhou, T., Caflisch, A., 2012, Free Energy Guided Sampling, *Journal of Chemical Theory and Computation*, 8(6):2134–2140 19

[79] Case, D. A., Darden, T. A., Cheatham, T. E., Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Crowley, M., Walker, R. C., Zhang, W., Merz, K. M., Wang, B., Hayik, S., Roitberg, A., Seabra, G., Kolossváry, I., Wong, K. F., Paesani, F., Vanicek, J., Wu, X., Brozell, S. R., Steinbrecher, T., Gohlke, H., Yang, L., Tan, C., Mongan, J., Hornak, V., Cui, G., Mathews, D. H., Seetin, M. G., Sagui, C., Babin, V., Kollman, P. A., 2010, *Amber 11*, University of California, San Francisco 19, 40, 60

[80] Götz, A. W., Williamson, M. J., Xu, D., Poole, D., Le Grand, S., Walker, R. C., 2012, Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born, *J. Chem. Theory Comput.*, 8(5):1542–1555 20

[81] Jensen, F., 1999, *Introduction to Computational Chemistry*, Wiley, Chichester 21, 40

[82] Bahar, I., Rader, A. J., 2005, Coarse-grained normal mode analysis in structural biology, *Curr. Opin. Struct. Biol.*, 15(5):586–592 21

[83] Yang, L., Song, G., Jernigan, R. L., 2007, How well can we understand large-scale protein motions using normal modes of elastic network models?, *Biophys. J.*, 93(3):920–929 21

[84] Chang, C.-E., Chen, W., Gilson, M. K., 2005, Evaluating the Accuracy of the Quasi-harmonic Approximation, *Journal of Chemical Theory and Computation*, 1(5):1017–1028 22

[85] Schlitter, J., 1993, Estimation of absolute and relative entropies of macromolecules using the covariance matrix, *Chemical Physics Letters*, 215(6):617–621 22

[86] Andricioaei, I., Karplus, M., 2001, On the calculation of entropy from covariance matrices of the atomic fluctuations, *The Journal of Chemical Physics*, 115(14):6289–6292 22

[87] Wang, J., Hou, T., 2012, Develop and Test a Solvent Accessible Surface Area-Based Model in Conformational Entropy Calculations, *Journal of Chemical Information and Modeling*, 52(5):1199–1212 22

[88] Kabsch, W., 1976, A Solution for the Best Rotation to Relate Two Sets of Vectors, *Acta Cryst.*, A32:922–923 22

[89] Schwender, H., 2007, *Statistical Analysis of Genotype and Gene Expression Data*, Ph.D. thesis, Department of Statistics of the University of Dortmund 23

[90] Zhou, Y., Cras-Meneur, C., Ohsugi, M., Stormo, G. D., Permutt, M. A., 2007, A global approach to identify differentially expressed genes in cDNA (two-color) microarray experiments, *Bioinformatics*, 23(16):2073–2079 23

[91] Tusher, V. G., Tibshirani, R., Chu, G., 2001, Significance analysis of microarrays applied to the ionizing radiation response, *Proc. Natl. Acad. Sci. U.S.A.*, 98(9):5116–5121 23

[92] Westfall, P., Young, S., 1993, *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustments.*, Wiley, New York 23

[93] Schwender, H., 2011, *siggenes: Multiple testing using SAM and Efron's empirical Bayes approaches*, r package version 1.28.0 23

[94] Roy, A., Post, C. B., 2012, Detection of Long-Range Concerted Motions in Protein by a Distance Covariance, *Journal of Chemical Theory and Computation*, 8(9):3009–3014 24, 25, 68

[95] Gilson, M., 2010, An Introduction to Protein-Ligand Binding for BindingDB Users 27

[96] Oette, M., 2003, *Resistenz in der HIV-Therapie. Diagnostik und Management*, UniMed, Bremen 27

[97] Cheng, Y.-C., William, H. P., 1973, Relationship between the inhibition constant (ki) and the concentration of inhibitor which causes 50 per cent inhibition (i50) of an enzymatic reaction, *Biochemical Pharmacology*, 22(23):3099–3108 27

[98] Liu, T., Lin, Y., Wen, X., Jorissen, R. N., Gilson, M. K., 2007, BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities, *Nucleic Acids Res.*, 35(Database issue):198–201 28

[99] Brown, S. P., Muchmore, S. W., Hajduk, P. J., 2009, Healthy skepticism: assessing realistic model performance, *Drug Discov. Today*, 14(7-8):420–427 28

[100] Steinbrecher, T., Labahn, A., 2010, Towards accurate free energy calculations in ligand protein-binding studies, *Curr. Med. Chem.*, 17(8):767–785 29

[101] Massova, I., Kollman, P. A., 1999, Computational alanine scanning to probe protein-protein interactions – a novel approach to evaluate binding free energies, *J. Am. Chem. Soc.*, 121(36):8133–8143 30

[102] Kollman, P. A., 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, D. A., Cheatham, T. E., 2000, Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models, *Acc. Chem. Res.*, 33:889–897 30

[103] Massova, I., Kollman, P., 2000, Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding, *Pers. Drug Discov. Des.*, 18:113–135 30

[104] Genheden, S., Luchko, T., Gusarov, S., Kovalenko, A., Ryde, U., 2010, An MM/3D-RISM approach for ligand binding affinities, *J Phys Chem B*, 114(25):8505–8516 32, 44

[105] Luo, H., Sharp, K., 2002, On the calculation of absolute macromolecular binding free energies, *Proc. Nat. Ac. Sci. U.S.A*, 99:10399–10404 32

[106] Swanson, J. M., Henchman, R. H., McCammon, J. A., 2004, Revisiting Free Energy Calculations: A Theoretical Connection to MM/PBSA and Direct Calculation of the Association Free Energy, *Biophys. J.*, 86:67–74 32

[107] Gohlke, H., Case, D. A., 2004, Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf, *J Comput Chem*, 25(2):238–250 32

[108] Xu, Y., Wang, R., 2006, A computational analysis of the binding affinities of FKBP12 inhibitors using the MM-PB/SA method, *Proteins*, 64(4):1058–1068 32

[109] Aruksakunwong, O., Wolschann, P., Hannongbua, S., Sompornpisut, P., 2006, Molecular dynamic and free energy studies of primary resistance mutations in HIV-1 protease-ritonavir complexes, *J Chem Inf Model*, 46(5):2085–2092 32

[110] Wittayanarakul, K., Hannongbua, S., Feig, M., 2008, Accurate prediction of protonation state as a prerequisite for reliable MM-PB(GB)SA binding free energy calculations of HIV-1 protease inhibitors, *J Comput Chem*, 29(5):673–685 32

[111] Stoica, I., Sadiq, S. K., Coveney, P. V., 2008, Rapid and accurate prediction of binding free energies for saquinavir-bound HIV-1 proteases, *J. Am. Chem. Soc.*, 130(8):2639–2648 32

[112] Sadiq, S. K., Wright, D. W., Kenway, O. A., Coveney, P. V., 2010, Accurate ensemble molecular dynamics binding free energy ranking of multidrug-resistant HIV-1 proteases, *J. Chem. Inf. Model.*, 50:890–905 32, 33, 42, 59

[113] Yam, W. K., Wahab, H. A., 2009, Molecular insights into 14-membered macrolides using the MM-PBSA method, *J Chem Inf Model*, 49(6):1558–1567 33

[114] Yang, Y., Qin, J., Liu, H., Yao, X., 2011, Molecular dynamics simulation, free energy calculation and structure-based 3D-QSAR studies of B-RAF kinase inhibitors, *J Chem Inf Model*, 51(3):680–692 33

[115] Jenwitheesuk, E., Samudrala, R., 2003, Improved prediction of HIV-1 protease-inhibitor binding energies by molecular dynamics simulations, *BMC Struct. Biol.*, 3:2 33

[116] Rastelli, G., Del Rio, A., Degliesposti, G., Sgobba, M., 2010, Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA, *J Comput Chem*, 31(4):797–810 33

[117] Pearlman, D. A., 2005, Evaluating the molecular mechanics poisson-boltzmann surface area free energy method using a congeneric series of ligands to p38 MAP kinase, *J. Med. Chem.*, 48(24):7796–7807 33

[118] Wong, S., Amaro, R. E., McCammon, J. A., 2009, MM-PBSA Captures Key Role of Intercalating Water Molecules at a Protein-Protein Interface, *J Chem Theory Comput*, 5(2):422–429 33

[119] Zhou, W., Motakis, E., Fuentes, G., Verma, C. S., 2012, Macrostate Identification from Biomolecular Simulations through Time Series Analysis, *J Chem Inf Model*, 52(9):2319–2324 33

[120] Genheden, S., Ryde, U., 2010, How to obtain statistically converged MM/GBSA results, *J Comput Chem*, 31(4):837–846 33, 39, 59, 70

[121] Kongsted, J., Ryde, U., 2009, An improved method to predict the entropy term with the MM/PBSA approach, *Journal of Computer-Aided Molecular Design*, 23:63–71 33, 34, 41

[122] Genheden, S., Nilsson, I., Ryde, U., 2011, Binding affinities of factor Xa inhibitors estimated by thermodynamic integration and MM/GBSA, *J Chem Inf Model*, 51(4):947–958 33

[123] Wan, S., Coveney, P. V., 2011, Rapid and accurate ranking of binding affinities of epidermal growth factor receptor sequences with selected lung cancer drugs, *J R Soc Interface*, 8(61):1114–1127 33

[124] Wang, J., Morin, P., Wang, W., Kollman, P. A., 2001, Use of MM-PBSA in reproducing the binding free energies to HIV-1 RT of TIBO derivatives and predicting the binding mode to HIV-1 RT of efavirenz by docking and MM-PBSA, *J. Am. Chem. Soc.*, 123(22):5221–5230 33

[125] Hou, T., Wang, J., Li, Y., Wang, W., 2011, Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations, *J. Chem. Inf. Model.*, 51(1):69–82 33

[126] Seibert, C., Sakmar, T. P., 2008, Toward a framework for sulfoproteomics: Synthesis and characterization of sulfotyrosine-containing peptides, *Biopolymers*, 90(3):459–477 35

[127] Seibert, C., Cadene, M., Sanfiz, A., Chait, B. T., Sakmar, T. P., 2002, Tyrosine sulfation of CCR5 N-terminal peptide by tyrosylprotein sulfotransferases 1 and 2 follows a discrete pattern and temporal sequence, *Proc. Natl. Acad. Sci. U.S.A.*, 99(17):11031–11036 35

[128] Da, L. T., Wu, Y. D., 2011, Theoretical studies on the interactions and interferences of HIV-1 glycoprotein gp120 and its coreceptor CCR5, *J Chem Inf Model*, 51(2):359–369 35

[129] Dupradeau, F. Y., *RESP ESP charge Derive Server*, http://q4md-forcefieldtools.org/RED/ 35

[130] Dupradeau, F. Y., Pigache, A., Zaffran, T., Savineau, C., Lelong, R., Grivel, N., Lelong, D., Rosanski, W., Cieplak, P., 2010, The R.E.D. tools: advances in RESP and ESP charge derivation and force field library building, *Phys Chem Chem Phys*, 12(28):7821–7839 35

[131] Cieplak, P., Cornell, W. D., Bayly, C., Kollman, P. A., 1995, Application of the multimolecule and multiconformational RESP methodology to biopolymers: Charge derivation for DNA, RNA, and proteins, *Journal of Computational Chemistry*, 16(11):1357–1377 35

[132] Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Scalmani, G., Barone, V., Mennucci, B., Petersson, G. A., Nakatsuji, H., Caricato, M., Li, X., Hratchian, H. P., Izmaylov, A. F., Bloino, J., Zheng, G., Sonnenberg, J. L., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Vreven, T., Montgomery, J. A., Jr., Peralta, J. E., Ogliaro, F., Bearpark, M., Heyd, J. J., Brothers, E., Kudin, K. N., Staroverov, V. N., Kobayashi, R., Normand, J., Raghavachari, K., Rendell, A., Burant, J. C., Iyengar, S. S., Tomasi, J., Cossi, M., Rega, N., Millam, J. M., Klene, M., Knox, J. E., 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., Martin, R. L., Morokuma, K., Zakrzewski, V. G., Voth, G. A., Salvador, P., Dannenberg, J. J., Dapprich, S., Daniels, A. D., Farkas, O., Foresman, J. B., Ortiz, J. V., Cioslowski, J., J., F. D., 2009, Gaussian 09 Revision A.1, gaussian Inc. Wallingford CT 2009 35, 39, 48

[133] Homeyer, N., Horn, A. H., Lanig, H., Sticht, H., 2006, AMBER force-field parameters for phosphorylated amino acids in different protonation states: phosphoserine, phosphothreonine, phosphotyrosine, and phosphohistidine, *J Mol Model*, 12(3):281–289 35, 37

[134] Bryce, R., *AMBER Parameter Database*, http://www.pharmacy.manchester.ac.uk/bryce/amber/ 35

[135] Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S., Weiner, P., 1984, A new force field for molecular mechanical simulation of nucleic acids and proteins, *Journal of the American Chemical Society*, 106(3):765–784 37

[136] Dupradeau, F. Y., *FFParmDev*, http://q4md-forcefieldtools.org/FFParmDev/ 38

[137] Kim, E. E., Baker, C. T., Dwyer, M. D., Murcko, M. A., Rao, B. G., Tung, R. D., Navia, M. A., 1995, Crystal structure of HIV-1 protease in complex with VX-478, a potent and orally bioavailable inhibitor of the enzyme, *Journal of the American Chemical Society*, 117(3):1181–1182 39

[138] Surleraux, D. L. N. G., Tahri, A., Verschueren, W. G., Pille, G. M. E., de Kock, H. A., Jonckers, T. H. M., Peeters, A., De Meyer, S., Azijn, H., Pauwels, R., de Bethune, M.-P., King, N. M., Prabu-Jeyabalan, M., Schiffer, C. A., Wigerinck, P. B. T. P., 2005, Discovery and Selection of TMC114, a Next Generation HIV-1 Protease Inhibitor, *Journal of Medicinal Chemistry*, 48(6):1813–1822 39

[139] Miertuš, S., Scrocco, E., Tomasi, J., 1981, Electrostatic interaction of a solute with a continuum. A direct utilizaion of Ab Initio molecular potentials for the prevision of solvent effects, *Chem. Phys.*, 55:117–129 39, 48

[140] Besler, B. H., Merz, K. M., Kollman, P. A., 1990, Atomic charges derived from semiempirical methods, *Journal of Computational Chemistry*, 11(4):431–439 39, 48

[141] Bayly, C. I., Cieplak, P., Cornell, W., Kollman, P. A., 1993, A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model, *The Journal of Physical Chemistry*, 97(40):10269–10280 39, 49

[142] Kuhn, B., Kollman, P. A., 2000, Binding of a diverse set of ligands to avidin and streptavidin: an accurate quantitative prediction of their relative affinities by a combination of molecular mechanics and continuum solvent models, *J. Med. Chem.*, 43(20):3786–3791 40

[143] Genheden, S., Kuhn, O., Mikulskis, P., Hoffmann, D., Ryde, U., 2012, The Normal-Mode Entropy in the MM/GBSA Method: Effect of System Truncation, Buffer Region, and Dielectric Constant, *Journal of Chemical Information and Modeling*, 52(8):2079–2088 41, 47, 59

[144] Yang, W., Bitetti-Putzer, R., Karplus, M., 2004, Free energy simulations: Use of reverse cumulative averaging to determine the equilibrated region and the time required for convergence, *J. Chem. Phys.*, 120:2618–2628 41

[145] King, N. M., Prabu-Jeyabalan, M., Nalivaika, E. A., Wigerinck, P., de Bethune, M. P., Schiffer, C. A., 2004, Structural and thermodynamic basis for the binding of TMC114, a next-generation human immunodeficiency virus type 1 protease inhibitor, *J. Virol.*, 78(21):12012–12021 43

[146] Wiesmann, F., Vachta, J., Ehret, R., Walter, H., Kaiser, R., Sturmer, M., Tappe, A., Daumer, M., Berg, T., Naeth, G., Braun, P., Knechten, H., 2011, The L76V mutation in HIV-1 protease is potentially associated with hypersusceptibility to protease inhibitors Atazanavir and Saquinavir: is there a clinical advantage?, *AIDS Res Ther*, 8:7 48

[147] Alcaro, S., Artese, A., Ceccherini-Silberstein, F., Ortuso, F., Perno, C. F., Sing, T., Svicher, V., 2009, Molecular dynamics and free energy studies on the wild-type and mutated HIV-1 protease complexed with four approved drugs: mechanism of binding and drug resistance, *J Chem Inf Model*, 49(7):1751–1761 48

[148] Clemente, J. C., Coman, R. M., Thiaville, M. M., Janka, L. K., Jeung, J. A., Nukoolkarn, S., Govindasamy, L., Agbandje-McKenna, M., McKenna, R., Leelamanit, W., Goodenow, M. M., Dunn, B. M., 2006, Analysis of HIV-1 CRF 01 A/E protease inhibitor resistance: structural determinants for maintaining sensitivity and developing resistance to atazanavir, *Biochemistry*, 45(17):5468–5477 48

[149] Schrödinger, LLC, 2010, The PyMOL molecular graphics system, version 1.3r1 49, 60

[150] Steinbrecher, T., Mobley, D. L., Case, D. A., 2007, Nonlinear scaling schemes for Lennard-Jones interactions in free energy calculations, *J. Chem. Phys.*, 127:214108 49, 51

[151] Heaslet, H., Rosenfeld, R., Giffin, M., Lin, Y. C., Tam, K., Torbett, B. E., Elder, J. H., McRee, D. E., Stout, C. D., 2007, Conformational flexibility in the flap domains of ligand-free HIV protease, *Acta Crystallogr. D Biol. Crystallogr.*, 63(Pt 8):866–875 51

[152] Lin, C., 2006, HCV NS3-4A Serine Protease., *Hepatitis C Viruses: Genomes and Molecular Biology*. 52, 55

[153] Kim, J. L., Morgenstern, K. A., Lin, C., Fox, T., Dwyer, M. D., Landro, J. A., Chambers, S. P., Markland, W., Lepre, C. A., O'Malley, E. T., Harbeson, S. L., Rice, C. M., Murcko, M. A., Caron, P. R., Thomson, J. A., 1996, Crystal structure of the hepatitis C virus NS3 protease domain complexed with a synthetic NS4A cofactor peptide, *Cell*, 87(2):343–355 52

[154] Love, R. A., Parge, H. E., Wickersham, J. A., Hostomsky, Z., Habuka, N., Moomaw, E. W., Adachi, T., Hostomska, Z., 1996, The crystal structure of hepatitis C virus NS3 proteinase reveals a trypsin-like fold and a structural zinc binding site, *Cell*, 87(2):331–342 52, 54

[155] Barbato, G., Cicero, D. O., Nardi, M. C., Steinkuhler, C., Cortese, R., De Francesco, R., Bazzo, R., 1999, The solution structure of the N-terminal proteinase domain of the hepatitis C virus (HCV) NS3 protein provides new insights into its activation and catalytic mechanism, *J. Mol. Biol.*, 289(2):371–384 52, 53, 54, 55

[156] Landro, J. A., Raybuck, S. A., Luong, Y. P., O'Malley, E. T., Harbeson, S. L., Morgenstern, K. A., Rao, G., Livingston, D. J., 1997, Mechanistic role of an NS4A peptide cofactor with the truncated NS3 protease of hepatitis C virus: elucidation of the NS4A stimulatory effect via kinetic analysis and inhibitor mapping, *Biochemistry*, 36(31):9340–9348 53

[157] Zhu, H., Briggs, J. M., 2011, Mechanistic role of NS4A and substrate in the activation of HCV NS3 protease, *Proteins*, 79(8):2428–2443 54, 59

[158] De Francesco, R., Urbani, A., Nardi, M. C., Tomei, L., Steinkuhler, C., Tramontano, A., 1996, A zinc binding site in viral serine proteinases, *Biochemistry*, 35(41):13282–13287 54

[159] Steinkuhler, C., Urbani, A., Tomei, L., Biasiol, G., Sardana, M., Bianchi, E., Pessi, A., De Francesco, R., 1996, Activity of purified hepatitis C virus protease NS3 on peptide substrates, *J. Virol.*, 70(10):6694–6700 54

[160] Barbato, G., Cicero, D. O., Cordier, F., Narjes, F., Gerlach, B., Sambucini, S., Grzesiek, S., Matassa, V. G., De Francesco, R., Bazzo, R., 2000, Inhibitor binding induces active site stabilization of the HCV NS3 protein serine protease domain, *EMBO J.*, 19(6):1195–1206 55, 60

[161] Tsantrizos, Y. S., Bolger, G., Bonneau, P., Cameron, D. R., Goudreau, N., Kukolj, G., LaPlante, S. R., Llinas-Brunet, M., Nar, H., Lamarre, D., 2003, Macrocyclic inhibitors of the NS3 protease as potential therapeutic agents of hepatitis C virus infection, *Angew. Chem. Int. Ed. Engl.*, 42(12):1356–1360 57

[162] Courcambeck, J., Bouzidi, M., Perbost, R., Jouirou, B., Amrani, N., Cacoub, P., Pepe, G., Sabatier, J. M., Halfon, P., 2006, Resistance of hepatitis C virus to NS3-4A protease inhibitors: mechanisms of drug resistance induced by R155Q, A156T, D168A and D168V mutations, *Antivir. Ther. (Lond.)*, 11(7):847–855 57, 58

[163] Lin, C., Lin, K., Luong, Y. P., Rao, B. G., Wei, Y. Y., Brennan, D. L., Fulghum, J. R., Hsiao, H. M., Ma, S., Maxwell, J. P., Cottrell, K. M., Perni, R. B., Gates, C. A., Kwong, A. D., 2004, In vitro resistance studies of hepatitis C virus serine protease inhibitors, VX-950 and BILN 2061: structural analysis indicates different resistance mechanisms, *J. Biol. Chem.*, 279(17):17508–17514 57

[164] Lin, C., Gates, C. A., Rao, B. G., Brennan, D. L., Fulghum, J. R., Luong, Y. P., Frantz, J. D., Lin, K., Ma, S., Wei, Y. Y., Perni, R. B., Kwong, A. D., 2005, In vitro studies of cross-resistance mutations against two hepatitis C virus serine protease inhibitors, VX-950 and BILN 2061, *J. Biol. Chem.*, 280(44):36784–36791 57

[165] Lu, L., Pilot-Matias, T. J., Stewart, K. D., Randolph, J. T., Pithawalla, R., He, W., Huang, P. P., Klein, L. L., Mo, H., Molla, A., 2004, Mutations conferring resistance to a potent hepatitis C virus serine protease inhibitor in vitro, *Antimicrob. Agents Chemother.*, 48(6):2260–2266 57, 59

[166] Papaleo, E., Pasi, M., Tiberti, M., De Gioia, L., 2011, Molecular Dynamics of Mesophilic-Like Mutants of a Cold-Adapted Enzyme: Insights into Distal Effects Induced by the Mutations, *PLoS ONE*, 6(9):e24214 59

[167] Smith, P. E., Pettitt, B. M., Karplus, M., 1993, Stochastic dynamics simulations of the alanine dipeptide using a solvent-modified potential energy surface, *The Journal of Physical Chemistry*, 97(26):6907–6913 59

[168] Caves, L. S. D., Evanseck, J. D., Karplus, M., 1998, Locally accessible conformations of proteins: Multiple molecular dynamics simulations of crambin, *Protein Science*, 7(3):649–666, doi:10.1002/pro.5560070314 59

[169] Fujitani, H., Tanida, Y., Ito, M., Jayachandran, G., Snow, C., Shirts, M., Sorin, E., Pande, V., 2005, Direct calculation of the binding free energies of FKBP ligands, *J. Chem. Phys.*, 123:084108 59

[170] Zagrovic, B., van Gunsteren, W. F., 2007, Computational Analysis of the Mechanism and Thermodynamics of Inhibition of Phosphodiesterase 5a by Synthetic Ligands, *Journal of Chemical Theory and Computation*, 3(1):301–311 59

[171] Lawrenz, M., Baron, R., McCammon, J. A., 2009, Independent-trajectories thermodynamic-integration free-energy changes for biomolecular systems: Determinants of H5N1 avian influenza virus neuraminidase inhibition by peramivir, *J. Chem. Theory Comput.*, 5:1106–1116 59

[172] Pang, Y.-P., 2001, Successful molecular dynamics simulation of two zinc complexes bridged by a hydroxide in phosphotriesterase using the cationic dummy atom method, *Proteins: Structure, Function, and Bioinformatics*, 45(3):183–189 60

[173] Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O., Shaw, D. E., 2010, Improved side-chain torsion potentials for the Amber ff99SB protein force field, *Proteins: Structure, Function, and Bioinformatics*, 78(8):1950–1958 60

[174] Breiman, L., 2001, Random Forests, *Machine Learning*, 45:5–32 66

[175] Liaw, A., Wiener, M., 2002, Classification and Regression by randomForest, *R News*, 2(3):18–22 66

[176] Bennett, C. H., 1976, Efficient estimation of free energy differences from Monte Carlo data, *J. Comput. Phys.*, 22:245–268 72

[177] Shirts, M. R., Bair, E., Hooker, G., Pande, V. S., 2003, Equilibrium Free Energies from Nonequilibrium Measurements Using Maximum-Likelihood Methods, *Phys. Rev. Lett.*, 91:140601 72

Publications

Peer-reviewed articles

Jonas Winkler, Giuliano Armano, Jan Nikolaj Dybowski, **Oliver Kuhn**, Filippo Ledda, Dominik Heider. Computational Design of a DNA- and Fc-Binding Fusion Protein. *Advances in Bioinformatics*, 2011, 457578.

Samuel Genheden, **Oliver Kuhn**, Paulius Mikulskis, Daniel Hoffmann and Ulf Ryde. The Normal-Mode Entropy in the MM/GBSA Method: Effect of System Truncation, Buffer Region, and Dielectric Constant, *Journal of Chemical Information and Modeling*, 2012, 52(8):2079-2088.

Posters

Oliver Kuhn, Daniel Hoffmann. MMPBSA using ensembles of independent trajectories. *Joint Meeting of the Swedish and German Biophysical Societies.* Hünfeld, 2011.

Acknowledgments

Eventually, I like to thank all the people that were more or less directly involved in this work.
I thank all present and former collegues from the **Bioinformatics Department** of the University Duisburg-Essen, especially **Manuel Prinz** for numerous helpful discussion and technical assistance and **Karsten Sewczyk**, **Jonas Winkler**, **Dominik Heider** for fruitful collaboration.
Thanks to **Samuel Genheden** for the cooperation project on the NMA entropy estimate, **Jörg Timm** for the HCV protease resistance data, **Amitava Roy** for kindly providing his DiCC Fortran subroutine. All **Amber community members** answering the Amber mailing list, in particular **Jason Swails**, **Carlos Simmerling**, **Ross Walker** and **Thomas Steinbrecher**.
I like to thank **Daniel Hoffmann** for advising me and being always available for questions of any kind.
Last and not least, I thank all my **family** for any support.

Declarations

Erklärung:
Hiermit erkläre ich, gem. § 6 Abs. (2) f) der Promotionsordnung der Fakultäten für Biologie und Geografie, Chemie und Mathematik zur Erlangung des Dr. rer. nat., dass ich das Arbeitsgebiet, dem das Thema "Free Energy Calculations and Molecular Dynamics Studies on Complexes of Viral Proteases with their Ligands" zuzuordnen ist, in Forschung und Lehre vertrete und den Antrag von Oliver Kuhn befürworte und die Betreuung auch im Falle eines Weggangs, wenn nicht wichtige Gründe dem entgegenstehen, weiterführen werde.

Essen, den _____ _____

Unterschrift eines Mitgliedes der Universität Duisburg-Essen

Erklärung:
Hiermit erkläre ich, gem. § 7 Abs. (2) c) + e) der Promotionsordnung der Fakultäten für Biologie und Geografie, Chemie und Mathematik zur Erlangung des Dr. rer. nat., dass ich die vorliegende Dissertation selbständig verfasst und mich keiner anderen als der angegebenen Hilfsmittel bedient habe.

Essen, den _____ _____

Unterschrift des Doktoranden

Erklärung:
Hiermit erkläre ich, gem. § 7 Abs. (2) d) + f) der Promotionsordnung der Fakultäten für Biologie und Geografie, Chemie und Mathematik zur Erlangung des Dr. rer. nat., dass ich keine anderen Promotionen bzw. Promotionsversuche in der Vergangenheit durchgeführt habe und dass diese Arbeit von keiner anderen Fakultät/Fachbereich abgelehnt worden ist.

Essen, den _____ _____

Unterschrift des Doktoranden