Protein-protein binding affinities calculated using the LIE method

KJE-3900

Tor Arne Heim Andberg

Master's Thesis in Chemistry
Faculty of Science and Technology
University of Tromsø
November 2011

Protein-protein binding affinities calculated using the LIE method

KJE-3900

Tor Arne Heim Andberg

Master's Thesis in Chemistry
Faculty of Science and Technology
University of Tromsø
November 2011

Protein-protein binding affinities calculated using the LIE method

Keywords: protein-protein interaction; serine proteinase; docking; molecular dynamics; linear interaction energy; molecular modelling.

Abstract: Protein-protein interactions are important for many biological functions, such as enzyme regulation and immune response. Knowledge about these interactions is important both to understand the biological processes but also to aid in drug design. This project has studied the interaction of the third domain of the turkey ovomucoid inhibitor (OMTKY3) in complex with Streptomyces griseus proteinase B (SGPB) and porcine pancreatic elastase (PPE). Point mutations were carried out on the primary binding residue (P1) and the models were subjected to molecular dynamics simulations. Absolute binding energies were calculated using the linear interaction energy (LIE) method and 70 % of the P1 variants yielded calculated binding free energies with error less than 2.0 kcal/mol in respect to the experimental binding energies.

Acknowledgements

I want start with thanking my family for their help and unconditional support. During the latter part of the project I have worked part time as a teacher and I want to thank the school, Nordkapp maritime fagskole og videregående skole, for their understanding and willingness to work out solutions (finding substitutes) whenever I went to Tromsø.

I also want to thank everyone who has in some way helped me, encouraged me or simply crossed paths with me during the course of the project. This includes Valentina Vollan for helping me with administrative issues, Erik Axel Vollan whose help with my computers has been invaluable, Ronny Helland for organising practical matters and I also want to thank Magne Olufsen who introduced me to the molecular dynamics software and who is always available for a chat. A few others I want to mention are Annfrid Sivertsen and Geir Isaksen.

But most of all I want to thank my supervisors Arne Oskar Smalås and Bjørn Olav Brandsdal for bearing with me and encouraging me when things looked bleak. I especially want to thank Bjørn Olav Brandsdal for his help, guidance and insight, and without whom this project would never have been completed.

This thesis is dedicated to my uncle Egil Henrik Heim who passed away unexpectedly in March 2005.

Tor Arne Heim Andberg

Abbreviations

BO	Born-Oppenheimer
BPTI	Bovine Pancreatic Trypsin Inhibitor
FEP	Free Energy Perturbation
HLE	Human Leukocyte Elastase
IUPAC	International Union of Pure and Applied Chemistry
LIE	Liner Interaction Energy
LRF	Local Reaction Field
MC	Monte Carlo
MD	Molecular Dynamics
MM	Molecular Mechanics
MM-PBSA	Molecular Mechanics/Poisson-Boltzmann/Surface Area
OMTKY3	Third Domain of the Turkey Ovomucoid Inhibitor
PDB	Protein Data Bank
PPE	Porcine Pancreatic Elastase
QM	Quantum Mechanics
RMSD	Root Mean Square Deviation
SGPB	Streptomyces Griseus Proteinase B
TI	Thermodynamic Integration
vdW	van der Waals

Contents

1 Introduction

Proteins are polypeptides, which are essential in living organisms. They occur as structural proteins like keratin found in hair and nails or globular proteins such as haemoglobin, which transports oxygen in the bloodstream. Another group of globular proteins are enzymes and they catalyse chemical processes by reducing the activation energy.

Interactions involving globular proteins are critical in many biological processes. Computer based approaches are well suited to probe protein-ligand and protein-protein interactions, and a large number of techniques have been developed to study these interactions. Two problems are typically associated with protein interactions: Structural prediction of a protein-ligand or protein-protein complex and accurate calculation of the binding energies for the complex. The former is referred to as docking, and the latter as scoring. Computer based methods for docking and binding energy calculations are useful tools to supplement traditional experiments in fields such as drug design. To allow accurate prediction of binding energy a model needs to describe intra- and intermolecular forces accurately, which is perhaps only possible through the use of quantum mechanical models. Secondly, sufficient sampling of configurations is necessary to generate a truly representative ensemble of structures. Accurate quantum mechanical methods are notoriously computer intensive and protein sized systems involving thousands of atoms present a very difficult challenge for today's methods and computers.

Current transistor based computers have come a long way since the advent of the microprocessor in the early 1970's. But even today's powerful computers are based on the same fundamental design and unless new technology is developed the exponential growth in computer power, as stipulated by Moore's law, will slow down as the transistors approach their theoretical minimum size. To best take advantage of current and future computers, it is important to find a healthy compromise between computational cost and accuracy. Many methods exist today and improving them and devising new methods is one of the goals of computational chemistry.

This thesis focus on protein-protein interactions and computer based models for studying the binding affinities. An overview is given of the actual protein systems to be studied and important interaction types are discussed. Special care is given to the

modelling of the interactions for protein-protein complexes and methodologies for calculating binding free energies for protein complexes.

1.1 Proteolytic enzymes

Proteolytic enzymes catalyse the hydrolysis of peptide bonds and are important in many physiological processes in most living organisms. There are two classes of proteolytic enzymes, exopeptidases and endopeptidases. Exopeptidases hydrolyse peptide bonds at either the N- or C-terminal side of polypeptide chains, while endopeptidases hydrolyse peptide bonds within the polypeptide chain. Proteolytic enzymes are classified according to the principal catalytic residue (serine, threonine, cysteine, aspartic and metallo peptidases) and are further divided into limited and unlimited proteolysis. Limited proteolysis is when a proteinase cleaves a limited number of peptide bonds of a target protein to produce an active form of the given protein. This allows for control of proteolytic activity. Unlimited proteolysis is complete degradation from protein to amino acids, as is the case for digestive enzymes.

1.1.1 Serine proteinases

Serine proteinases are one of the most studied classes of proteolytic enzymes. This is because serine proteinases have important biological functions both in digestion and proteolytic regulation. 40 different families of serine proteinases have been identified based on amino acid composition [1, 2]. The chymotrypsin and subtilisin families are the two most studied serine proteinase families. Both families utilize the same arrangement of catalytic residues, His, Ser and Asp, but, although the catalytic region is virtually identical the proteins themselves are structurally very different. Due to the intense scrutiny of serine proteinases there is an abundance of available x-ray structures and experimental binding data, for example the work of Laskowski et al. which is of particular importance for this thesis [3, 4].

In 1967 Schechter and Berger proposed a model for the interactions between an enzyme and a substrate or inhibitor [5]. The model describes the interaction in relation to the primary or catalytic binding site being flanked by secondary binding sites as illustrated in Figure 1.1. The primary binding site is termed S1 and the

secondary sites are labelled S2-SN towards the N-terminal and S1'-SN' towards the C-terminal sides, respectively. The binding sites on the substrate or inhibitor follow a similar scheme with P1 being the name for the primary binding residue and P2-PN and P1'-PN' for the secondary residues.

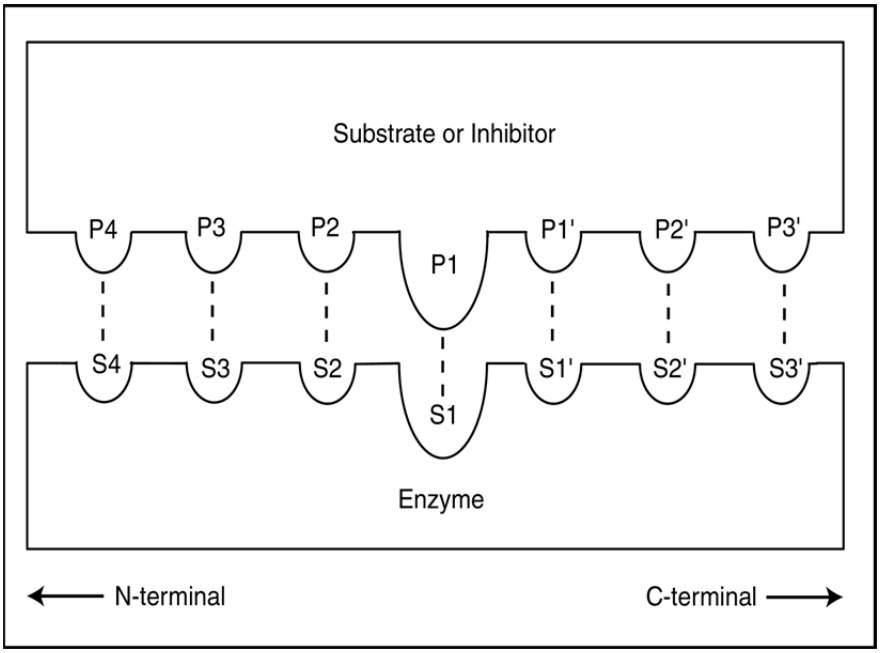

Figure 1.1: Binding subsites complexes formed by serine proteinases and their substrates/inhibitors, using nomenclature of Schechter and Berger [5].

Chymotrypsin-like serine proteinases

The fold of chymotrypsin-like serine proteinases are similar, consisting of two six stranded β-barrels with the catalytic triad situated between the two domains (Figure 1.2). Although the amino acid sequence can differ substantially the fold is highly conserved.

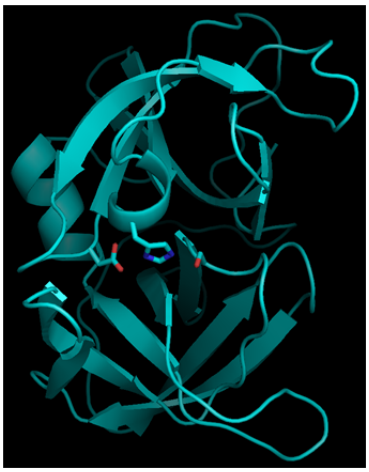

Figure 1.2: SGPB a chymotrypsin-like serine proteinase with catalytic residue drawn as sticks (1SGQ [3]). Figure generated using PyMol [6].

Catalytic mechanism in serine proteinases

The catalytic triad is composed of His57, Asp102 and Ser195. During catalysis Ser195 acts as a nucleophile and reacts with the carbonyl oxygen of the P1 residue and leads to cleaving of the peptide bond, also referred to as the scissile bond. The lone pair on the His57 nitrogen can accept the hydrogen from Ser195's hydroxyl group. Asp102 form a hydrogen bond through the carboxyl group with His57 and increase the electronegativity of the lone pair. The catalytic action relies on stabilisation of the transition state and the mechanism is described by Branden and Tooze in "Introduction to Protein Structure" from 1999 [7]. An important feature in the catalysis is the "oxyanion hole" which stabilizes the negatively charged carbonyl oxygen in the tetrahedral transition state. The oxyanion hole consists of backbone nitrogen atoms that donate their hydrogen atoms to stabilize the carbonyl oxygen as the reaction takes place. The reaction mechanism is outlined in Figure 1.3.

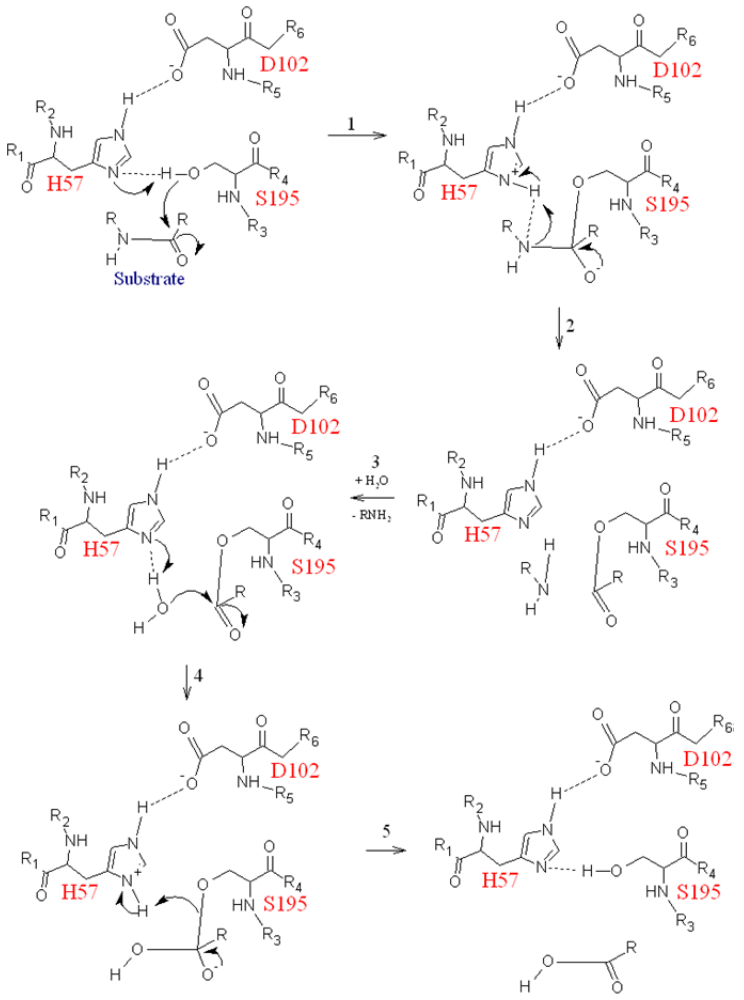

Figure 1.3: Reaction mechanism in serine proteinases (figure adapted from [8]).

1.1.2 Substrate specificity

The chymotrypsin super family is comprised of very similar enzymes but with very different specificities. This is typically illustrated by exploring chymotrypsin, trypsin and elastase. The enzymes themselves are very similar in structure, the catalytic residues are placed almost identical in space and the mechanism is the same. Chymotrypsin has high affinity for large aromatic side chains, trypsin favours positively charged side chains and elastase binds best to small uncharged side chains. These differences are caused by the mutation of a small number of amino acids at and around the primary binding site. Of particular importance are residues 189, 216 and [8]226. Residue 189 is situated at the bottom of the binding site and 216 and 226 sits on either side. In chymotrypsin and trypsin residues 216 and 226 are glycines and allow large and wide residues access to the binding pocket. Residue 189 is Ser in chymotrypsin and Asp in trypsin. Serine is small and do not hinder aromatic residues from binding. Whereas in trypsin residue 189 is mutated to Asp and it interacts beneficially with Arg and Lys residues, explaining trypsin's preference for positively charged side chains. Elastase cleaves peptide bonds adjacent to small uncharged residues, and this is because the glycines found in chymotrypsin and trypsin at position 216 and 226 are replaced with Val216 and Thr226. The binding pocket in elastase is narrower and do not allow large residues to fit at all [4].

1.1.3 Protein inhibitors

Proteolytic enzymes are essential in living organism but their proteolytic activity needs to be controlled as unregulated proteolysis would degrade the all proteins into their constituent amino acids. Proteinases are expressed as inactive precursors and activated when needed by cleaving off one or more peptide bonds at the N-terminal side of the catalytic site. The activated enzyme is then controlled by protein inhibitors that reduce or stop the proteolytic activity upon complex formation.

Proteinase inhibitors are classified according to function, into either active site inhibitors or α_2-macroglobulins (α_2Ms) [9, 10]. α_2Ms inhibit a large range of endopeptidases of varying origins and catalytic mechanisms. The target proteinase binds to the α_2M and cleaves a peptide bond in the "bait region", causing the α_2M to change conformation and enfold the proteinase.

Active site inhibitors are either canonical or non-canonical, as outlined by Bode and Huber [11]. Canonical inhibitors have the same Ramachandran angles in the combining loop in both bound and unbound states. Canonical inhibitors targeting chymotrypsin-like and subtilisin-like serine proteinases can be divided into at least 18 different groups based on sequence homology, structural similarity and binding mechanism [9, 12]. When the canonical inhibitors bind to serine proteinases they follow "the standard mechanism" [9, 13]. The enzyme-inhibitor interaction can be written as an equilibrium reaction:

$$E + I \rightleftharpoons C \rightleftharpoons E + I^*$$

where E is the enzyme, I is the intact inhibitor, I^* is the modified inhibitor with hydrolysed peptide bond and C is the enzyme-inhibitor complex.

Canonical inhibitors of this group are small to medium sized proteins, ranging from 14 to about 200 residues [14], and bind through an exposed combining loop that surrounds the P1 residue. The binding interaction can be divided into primary and secondary interactions in relation to the primary and secondary binding sites, as shown in Figure 1.1. At the primary binding site the P1 side chain interacts with the specificity pocket or S1 pocket. The nature of this interaction is highly dependent on the side chain and specificity pockets composition. Secondary interactions include hydrogen bonds between the enzyme and inhibitor main chains as well as electrostatic and van der Waals interactions. Hydrophobic interaction also plays a part in the formation of the enzyme-inhibitor complex, although not specific to the binding loop it relates to the hydrophobic surface area which is buried upon formation of the complex.

1.2 Intra- and intermolecular forces

Shape complementarity

Shape complementarity is sometimes referred to as "key in lock" or "hand in glove" analogies, because matching surfaces are necessary to form stable complexes. This is important for protein-ligand and protein-protein complexes. It is not the shape complementarity itself that causes favourable interactions, but rather the contacting surface allows the target and the ligand to interact via electrostatic, van der Waals (vdW) and also hydrophobic interactions. When protein-protein complexes form the water molecules solvating the proteins must be expelled to allow for favourable

interactions to occur. The local environment at the contact interface of protein-protein complexes is also expected to change; in particular the dielectric constant will be reduced. As a consequence the electrostatic interactions will increase in strength due to less screening of the charges.

Many experimental structures of protein-protein complexes have been determined during the past decades, which has led to activities aimed at predicting the three dimensional structure of such complexes. A number of methods have been developed to find the best binding mode between a protein and a target, either a ligand or another protein [15]. To simplify the docking proteins are often treated as rigid bodies, and such methods are quick and computationally efficient, especially for protein-ligand systems but also for some protein-protein complexes. The drawback with rigid bodies is that proteins sometimes undergo conformational change upon binding and rigid body models are unable to predict these effects.

In order to form a stable complex the surfaces must match and the close contacts must also include attractive interactions, such as electrostatic attraction between permanent charges or multipoles, attractive vdW forces between non polar atoms and hydrogen bonds. The interactions that stabilize protein-protein complexes can be grouped into several different categories, main chain-main chain, main chain-side chain and side chain-side chain. Main chain-main chain interactions typically involve hydrogen bonds between the N-H group and the carbonyl oxygen.

Electrostatic interactions

Electrostatic interactions are important not only to provide specificity but also for the overall stability of protein-protein complexes. These interactions arise from the attraction and repulsion of electrically charged particles. Electrostatic interactions can be further divided into different types, including hydrogen bonds, salt bridges, multipole and vdW interactions that are typically found in protein systems. Hydrogen bonds and vdW interactions are discussed below.

Electrostatic interactions between pairs of point charges are described mathematically by Coulomb's law and form the basis for understanding interactions between ions. Salt bridges are formed when two ionisable amino acids of opposite charge are close in space, typically around 3 Å, and are of significant importance for protein stability as well as protein-protein complex formation. It is generally found that locally solvent exposed salt bridges contribute little to stability, and may even

destabilize proteins, while completely buried salt bridges contribute with several kcal/mol to the free energy of folding. Global salt bridges occur between pairs of ionisable amino acids that are separated in sequence, and thus connect different parts of the protein together. These are also expected to contribute more to the stability. Differences in electronegativity between elements result in uneven charge distributions in molecules. This gives rise to dipoles, quadrupoles, octopoles, etc. collectively called multipoles, which interact with permanent charges or other multipoles. For simplicity the phenomenon is commonly referred to as dipole interactions although multipole interactions are technically a better name. Dipole interactions are classified as either dipole-dipole or dipole-induced dipole, where the first is the electrostatic interactions between the partial charges of the atoms in the dipole. In the latter case, the electrostatic influence of one dipole induces a dipole in otherwise neutral atoms by polarizing the covalent bond between them.

van der Waals interactions

van der Waals forces are of electrostatic origin and composed of attractive and repulsive forces unrelated to point charges, dipole-dipole or dipole-induced dipole interactions. The attractive component is called dispersive forces and is a result of instantaneous dipole interactions caused by fluctuations of the electron clouds. It was first explained through quantum mechanics (QM) by Fritz London in 1930, and is sometimes called the London force [16].

van der Waals repulsion can be explained by Pauli's exclusion principle, which states that no two electrons in a given system can have the same quantum numbers. This prevents electrons with the same spin from occupying the same area of space and if two atoms are too close the electron density between the corresponding nuclei is reduced due to overlap. When the electron density is reduced there is less shielding between the positively charged nuclei and this leads to electrostatic repulsion [17].

vdW forces are important not only for the interaction between different proteins but also between proteins and small ligands. These interactions occur between any pair of atoms, and will increase in strength as the number of atoms increase. While they decay rapidly as the distance between the interacting particles increases, the large number of interacting particles in protein-protein complexes

results in a significant contribution from vdW forces to the overall stability of the complexes.

Hydrogen bonds

Hydrogen bonds are perhaps the most important interaction in biological systems, and are responsible for the peculiar properties of water. Hydrogen bonds play a pivotal role in stabilizing secondary structure elements of proteins, they are central in DNA base-paring and at interfaces of protein-protein complexes. A hydrogen bond is typically described as an attractive force between a hydrogen atom from a molecule where the hydrogen is covalently bound to N, O or F and a lone pair in an adjacent N, O or F atom of another molecule or part of the same molecule. It is usually depicted as: $X-H\cdots Y$ where the dots denote the bond and X and Y are N, O or F. The International union of Pure and Applied Chemistry (IUPAC) recently updated their definition of hydrogen bonds [18] to give a more general and universal view of the phenomenon.

Hydrogen bonds are mainly electrostatic, but also include contributions that are covalent in nature as well as dispersion effects [19]. The $X-H\cdots Y$ unit tends to be linear and the bond is oriented towards a lone pair in Y. Typical bond lengths measured from H to Y are about 2 Å. The binding energy depends on which elements the hydrogen is bound to, $O-H\cdots O$ has a binding energy of about 5 kcal/mol. Hydrogen bonds in protein ligand systems are more favourable than hydrogen bonds between the protein and solvent [20] and hydrogen bonds are also important to determine the specificity [20, 21].

Hydrophobic interactions

Hydrophobic interactions are the seemingly attractive interactions that arise between hydrophobic substances when they are dissolved in water. It is worth noting that hydrophobic interactions are not an attractive force between the particles but rather an effect caused by the solvent. Upon dissolving a hydrophobic unit, either a non-polar molecule or a non-polar part of a molecule, will be surrounded by water molecules that form a network of hydrogen bonds between themselves. This ordered structure reduces the entropy, making the above case energetically unfavourable.

When multiple hydrophobic units come together the combined surface area is smaller than for the units individually. A smaller surface area means fewer water molecules are locked and this causes an increase in entropy. Thus maximising hydrophobic contacts minimises the disruption of water and is energetically favourable. Proteins are known to have hydrophobic cores and hydrophobic interactions are important for the correct fold. Hydrophobic interactions are also important for protein-ligand and protein-protein binding. Protein-protein binding interfaces are found to be more hydrophobic than the solvated surface area [22] and facilitate the complex formation.

1.3 Modelling of intra- and intermolecular forces

Many methodologies are presently available to model biological systems of interest, and the interactions occurring in such systems. The most accurate way to model a molecular system is by quantum mechanics. All QM methods are based on the Schrödinger equation and consider the electrons in the system explicitly. This combined with the general difficulty of solving the equation results in high computational cost. Ideally one would like to apply quantum mechanical models on every system, but many systems are unfortunately too large to be considered by QM approaches. This is particularly true for complex biological systems, which often consist of several thousands of atoms. Solvents effects may also be difficult to reveal through QM calculations, as a large amount of water molecules are needed for proper solvation of biological molecules.

1.3.1 Molecular Mechanics

Molecular mechanics (MM), also known as force field methods, ignore the electrons and write the energy as a function of nuclear coordinates only. MM is based on several assumptions, among them the Born-Oppenheimer (BO) approximation is of particular importance. The BO approximation states that the electrons immediately adapt to any changes in the nuclear positions, and since a proton is roughly 1800 times as heavy as an electron this is regarded as a good approximation. MM is usually based on simple models, typically a sum of components from stretching of bonds, bending angles and rotation of single bonds among others. Even using simple

functions to model the contributions, e.g. Hooke's law, the accuracy can be good and comparable to QM at a much lower computational cost for some properties.
A typical MM force field can look as follows:

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

$V(r^N)$ is the potential energy of a system as a function of the positions (r) of N particles (usually atoms). The first term is contributions from bond stretching, the second is angle bending and the third is bond rotation or torsions. The last term models non-bonded interactions between atom pairs, specifically van der Waals and electrostatic interactions. Contributions from bond stretching, angle bending, torsions, electrostatic and van der Waals interactions make up a typical force field and is discussed in greater detail below.

Bond stretching

An accurate description of the potential energy curve for bonds is given by the Morse potential:

$$v(l) = D_e[1 - \exp(-a[l - l_0])]^2 \quad (2)$$

$$a = \omega\sqrt{\mu / 2D_e} \quad (3)$$

where D_e is the depth of the potential energy minimum, μ is the reduced mass and ω is the frequency of the bond vibration. l_0 is the reference length of the bond, and is the length of the bond when all other terms in the force field are zero. In contrast the equilibrium bond length results in the lowest potential energy when all terms are considered. Although accurate the Morse potential is computationally inefficient and it requires three parameters per bond.

Because bond lengths rarely change significantly from the reference length the simpler Hooke's law (or harmonic potentials) is often used:

$$v(l) = \frac{k}{2}(l - l_0)^2 \quad (4)$$

where k is the force constant, l is the bond length and l_0 is the reference bond length. The functional form of Hooke's law is a good approximation to the curve of the

Morse potential at the bottom of the potential energy well. It is however less accurate away from the reference length.

Sometimes cubic, quartic and higher order terms are included in the Hooke's law potential to increase the accuracy. These higher order functions will be more accurate around the reference length, but have certain drawbacks as well. As the order increases additional parameters are needed for each bond type and as the function gets more complicated the computational cost increases. Cubic functions can cause another problem if the bond is stretched beyond the maximum of the cubic function. In this case the potential energy will become increasingly negative as the bond length increases and break the model.

Angle bending

As with bond stretching a Hooke's law formula can be used to describe the potential:

$$v(l) = \frac{k}{2}(\theta - \theta_0)^2 \tag{5}$$

where k is the force constant, θ is the bond angle and θ_0 is the reference bond angle.

The energy required to bend angles is less than the energy necessary to change the bond length. As such the parameters for angle bending is comparatively smaller, and typically 1/10 of the force constants used to model bond stretching. Higher order terms can be included to increase the accuracy of the force field, but again, at the cost of increased computing time

Torsional terms

Torsional terms are usually included to construct rotational barriers. The energy required to rotate a bond is significantly smaller than deviations in bond length and angle, often referred to as hard degrees of freedom. It is not strictly necessary to include explicit torsional terms since non-bonded interactions often are sufficient to provide the required energy barrier. Most force fields dealing with "organic" molecules usually have a torsional component for each quartet A-B-C-D in the system. Torsional potentials can be written as follows:

$$v(\omega) = \sum_{n=0}^{N} \frac{V_n}{2}[1 + \cos(n\omega - \gamma)] \tag{6}$$

where V_n is often referred to as the barrier height, ω is the torsion angle and n is the multiplicity which defines the number of minimum points through all 360°. γ is the phase factor and defines when the torsion angle passes through the energy minima.

Improper torsions and out-of-plane bending

Improper torsion and out-of-plane bending terms are necessary to allow force fields to reproduce planar geometries. In some cases involving sp^2 hybridized carbon atoms such as the cyclobutanone molecule, the carbon and oxygen atoms lie in a plane due to the π bond. A force field with only bond stretching and angle bending terms predict the oxygen atom to be placed at an angle rather than in the plane. One way to deal with this effect is through the use of an "improper" torsional angle, which is a torsion angle between four atoms not bound in sequence. To keep the improper torsional angle at either 0° or 180° a torsional component can be used:

$$v(\omega) = k(1 - \cos 2\omega) \tag{7}$$

where k is the force constant and ω is improper torsion angle.

A way to calculate the out-of-plane bending effect is to calculate the angle or distance the atom in question is away from the plane. Two typical harmonic potential formulas are:

$$v(\theta) = \frac{k}{2}\theta^2 \tag{8}$$

$$v(h) = \frac{k}{2}h^2 \tag{9}$$

where k is the force constant, θ is the angle between the atom and the plane and h is the distance between the angle and the plane.

Cross terms

Cross terms are contributions relating to coupling between various terms in the force field. Common cross terms include stretch-stretch, stretch-bend, stretch-torsion, bend-torsion and bend-bend couplings. MM is based on a sum of different energy contributions such as bond stretching, angle bending among others. These contributions are calculated separately regardless of any other terms, and additivity is assumed. From QM we know that changes in one part of a molecule will affect the molecule as a whole and the idea to break down the potential energy to a set of independent contributions is an approximation. However, not all properties are

affected in the same manner. Certain structural properties for small molecules can be accurately reproduced with only a few cross terms, while other properties such as vibrational frequencies are even affected by small changes to the bond length and angles require more cross terms, yet not all force fields include any cross terms. AMBER [23] and OPLS [24] are two commonly used force fields to model biological molecules and neither incorporates cross terms. Indeed a force field requires cross terms between all contributions if it is to be completely accurate. The drawback is extensive parameterization and computational cost. In general interactions that are far away from each other in the molecule can be treated as zero, while cross terms are included for close interactions.

The purpose of the force field is also important, a force field designed to calculate vibrational spectra would include many cross terms as well as higher order expansions for bond stretching, etc. This would result in a complex but accurate force field suitable for vibrational spectra, but would be too computational costly to use in a large scale Monte Carlo (MC) or Molecular Dynamics (MD) simulation where a simpler and faster force field would allow for longer simulations and better sampling.

Non-bonded interactions

Non-bonded interactions are typically divided into van der Waals and electrostatic contributions. The charge distribution in molecules determines the electrostatics and can be represented in many ways, most commonly through the use of partial atomic point charges. Obtaining the partial atomic charges can be accomplished in a number of ways [17]. Electrostatic contributions are often calculated using Coulomb's law, as a sum of interactions between pairs of point charges between two molecules or between different parts of the same molecule.

$$V = \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \tag{10}$$

q_i and q_j are point charges, ε_0 is the vacuum permittivity and r_{ij} is the distance between the point charges. N_A is the number of point charges in the first molecule and N_B is the number of point charges in the second molecule.

van der Waals interactions are in a sense electrostatic in nature. There are both attractive and repulsive van der Waals forces and both can be explained by electrostatic interactions. The attractive forces are explained with quantum mechanics

as fluctuations in the electron cloud around the atom, which in turn induce dipoles in nearby atoms. As a result induced dipole-induced dipole interactions occur. The effect increase with the number of electrons, so heavier elements will contribute significantly more. Quantum mechanical models predict this force potential is related with the distance as $1/r^6$. At short range repulsive exchange forces prevent the atoms from coming too close. There are many ways to model this effect but using a relationship of $1/r^{12}$ is common. This leads to the Lennard-Jones 12-6 function:

$$v(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{11}$$

This potential is a function of the distance r, and ε is the depth of the well, while σ is the collision diameter. This gives the force field a fast and roughly accurate description of the van der Waals interactions. Due to the nature of orbitals, van der Waals forces are direction specific. This is not included in the Lennard-Jones potential, and to model this effect additional terms are required.

Transferability

Ideally a force field should include unique parameters for every possible combination of atoms from every possible molecule. For example the C-H bonds in methane would be expected to be subtly different from the C-H bonds in let say ethane. However this level of parameterization would result in a very large and cumbersome force field, if it was possible to make one at all. Such a force field would not be able to predict properties of new molecules and its usefulness would be limited. Instead properties are assumed to be transferable from one system to another. Meaning a C-H bond for example would be treated identically in both methane and ethane or simply whenever a hydrogen atom is bound to a sp^3 hybridized carbon atom.

Transferability allows a force field to use a small set of parameters, typically obtained through QM calculations of a series of test molecules, to predict properties in other similar molecules. In fact transferability is the premise on which all MM force fields rely.

Polypeptides can consist of hundreds and thousands of atoms and it is virtually impossible for current QM methods to analyse entire proteins to elucidate the MM parameters. Force fields for polypeptides are usually constructed through the study of dipeptides and tripeptides. Atoms inside di- or tripeptide will experience similar

surroundings as if they were inside a larger polypeptide and the low number of atoms makes it easier to use QM calculations. The parameters are then assumed to be applicable to atoms from any polypeptide.

1.3.2 Molecular Dynamics

Molecular dynamics is a method used to generate successive configurations of a system by integrating Newton's laws of motion.

$$\frac{d^2 x_i}{dt^2} = \frac{F_{x_i}}{m_i} \tag{12}$$

which describes the movement of a particle of mass m_i, along the direction x_i where F_x is the force along the given direction.

An alternative method is Monte Carlo simulations. Whereas MD integrates Newton's laws of motion MC employ random sampling algorithms to produce similar results.

Several different MD models have been made. The first of these is the hard-sphere model, and in this model particles move at constant speed in straight lines until they collide with another particle. The collisions are perfectly elastic and after a collision the new velocities are calculated through conventional mechanics applying conservation of linear momentum. This model has provided useful insights into the microscopic understanding of fluids, but suffers from many deficiencies.

Molecular dynamics with continuous potentials

In reality particles are under the influence of continuous potentials and the force on each particle changes as it moves in relation to its surroundings. The motions are coupled and the movement of one particle will affect all the others. Models using continuous potential will in all but the simplest cases give rise to a many-body problem that cannot be solved analytically.

Finite difference methods avoid the many-body problem by breaking down the integration into small steps of time, δt. At a given time t the total force acting on all particles are calculated. When the force is known the acceleration can be calculated, which in turn is used to calculate the speed and position of each particle. When the position and velocity has been calculated for $t + \delta t$ the procedure is repeated and the new positions are used to calculate the positions and velocities for the next step.

Many algorithms exist for integrating the equations in the finite difference method; all of them use Taylor expansions to approximate the positions, velocity, acceleration and other dynamic properties.

$$r(t+\delta t) = r(t) + \delta t v(t) + \frac{1}{2}\delta t^2 a(t) + \frac{1}{6}\delta t^3 b(t) + \frac{1}{24}\delta t^4 c(t) + \cdots \quad (13)$$

$$v(t+\delta t) = v(t) + \delta t a(t) + \frac{1}{2}\delta t^2 b(t) + \frac{1}{6}\delta t^3 c(t) + \cdots \quad (14)$$

$$a(t+\delta t) = a(t) + \delta t b(t) + \frac{1}{2}\delta t^2 c(t) + \cdots \quad (15)$$

$$b(t+\delta t) = b(t) + \delta t c(t) + \cdots \quad (16)$$

r is the position and v is the velocity (first derivative), a is the acceleration (second derivative) and b is the third derivative and so on.

Time step

The time step is an integral part of MD simulations using finite difference methods. Choosing the time step is a compromise between speed and stability. A large time step allows for quicker conformational sampling, yet it might cause instabilities and inaccurate trajectories. A general rule of thumb is that the time step should be roughly a tenth of the highest-frequency vibrations in the system. A C-H bond vibrates with a period of about 10 fs, and an appropriate time step in such a case will be about 1.0 fs. Certain constraints are often applied to the fastest period vibrations, one such method is the SHAKE procedure proposed by Ryckaert, Ciccotti and Berendsen [25].

1.3.3 Rigorous free energy methods

Free energy perturbation (FEP) and thermodynamic integration (TI) are known as rigorous approaches and calculate the relative binding free energy between two equilibrium states. The free energy between two states can be expressed by Zwanzig's formula [26]:

$$\Delta G = G_B - G_A = -\beta^{-1} \ln \left\langle \exp\left(-\beta \Delta V\right) \right\rangle_A \quad (17)$$

where $\beta = 1/kT$, $\langle\ \rangle_A$ is the MC or MD ensemble average of $\Delta V = V_B - V_A$ sampled using the V_A potential. The equation requires constant temperature and pressure during the sampling, while another requirement is that configurations from V_A should

also have a chance of occurring in V_B, which means V_A and V_B have to be somewhat similar.

The FEP approach is theoretically exact, but the drawbacks are often significant. The systems V_A and V_B have to be similar and this limits diversity of ligands the FEP method can be applied to. Many protein-protein interactions involving P1-mutations involve changes that are too large for the FEP method to handle. Secondly the FEP approach spends most of the computer time exploring intermediate states which are not physically relevant, resulting in long and costly calculations. These two drawbacks led to the development of the LIE method, which will be discussed later.

A multistep approach is usually employed to solve the previous equation by constructing intermediate potential energy functions as a linear combination of initial and final potentials, V_A and V_B:

$$V_m = (1-\lambda_m)V_A + \lambda_m V_B \tag{18}$$

where λ_m assumes values from 0 to 1, but is in practice divided into a number of discrete points which represents a potential energy function. Through summation of the energy states the total free energy change can be found:

$$\Delta G = G_B - G_A = -\beta^{-1}\sum_{m=1}^{n-1} \ln \left\langle \exp\left[-\beta\left(V_{m+1} - V_m\right)\right]\right\rangle_m \tag{19}$$

Combining these two equations leads to:

$$\Delta G = \int_0^1 \left\langle \frac{\partial V(\lambda)}{\partial \lambda} \right\rangle_\lambda d\lambda \tag{20}$$

which is often referred to as the thermodynamic integration formula and is sometimes simplified as:

$$\Delta G = \int_0^1 \left\langle \Delta V \right\rangle_\lambda d\lambda \tag{21}$$

A more detailed overview of FEP and TI is given by Brandsdal et al. [27].

1.3.4 Linear interaction energy method

The linear interaction energy (LIE) method was proposed by Åqvist et al. in 1994 [28] to predict the binding affinity of drug-sized ligands to proteins through the use of MC or MD simulations. It was originally tested on a set of endothiopepsin inhibitors and accurately reproduced experimental binding energies [28]. The idea was to consider

only the physically relevant states of the ligand, i.e. bound and unbound and calculate the absolute binding energy as the difference in free energy of solvation between these states. TI and FEP approaches are inefficient since most of the simulation time is spent investigating intermediate states, whereas the LIE method requires only two simulations.

The LIE model relies on scaling of the electrostatic and vdW energies obtained from MC or MD simulations and the binding free energy is expressed as:

$$\Delta G_{bind} = \alpha \Delta \left\langle V_{l-s}^{vdW} \right\rangle + \beta \Delta \left\langle V_{l-s}^{el} \right\rangle + \gamma \tag{22}$$

$\langle\ \rangle$ denotes MC or MD averages of the vdW and electrostatic contributions. l-s is the interaction between the ligand and the surroundings and Δ refers to the difference in the averages from bound and free state. α, β and γ are the LIE coefficients. For small drug like ligands $\alpha = 0.18$ have produced good results, linear response give $\beta = ½$ although later work has revised the β value depending on the chemical nature of the ligand [29]. The final variable γ has been treated as 0 and is necessary if absolute binding energies are to be reproduced in some systems.

A key concept in the development of the LIE method was the linear response approximation. The linear response approximation was employed in the original version of the LIE method to approximate the electrostatic contribution to the binding free energy. For a given environment the linear response gives the electrostatic component as:

$$\Delta G_{el}^{i} = \frac{1}{2}\left\{ \left\langle V_{l-s}^{el} \right\rangle_{on} + \left\langle V_{l-s}^{el} \right\rangle_{off} \right\} \tag{23}$$

the $\left\langle V_{l-s}^{el} \right\rangle$ term is MC or MD averages, while the "on" and "off" terms refer to whether or not the electrostatic interactions between the ligand and its surroundings are turned on or off. To simplify the LIE method, $\left\langle V_{l-s}^{el} \right\rangle_{off}$ term is neglected. This has proven to be a good approximation in water as the molecules are more or less randomly oriented, and this term largely cancels itself out [29].

The non-polar energy requires a different approach. Here the idea is to measure the non-electrostatic interactions between the ligand and its surroundings in both the bound and unbound states and scale the contribution by an empirical factor. In a typical MD simulation this energy is obtained from a Lennard-Jones potential. Although "hydrophobic" energy is not directly described by this potential, two

observations make this a sound approach. Solvation free energies for typical non-polar compounds are experimentally found to scale linearly with solute size measures such as accessible surface area, and MD simulations have shown that the average van der Waals interaction energies scale approximately linearly with solute size in both polar and non-polar solvents [28, 29]. Combining these observations suggests it is possible to use average van der Waals energies to estimate the non-polar binding contribution. This is because the "hydrophobic" energies are scaled in approximately the same way as the van der Waals energy [27].

1.3.5 MM-PBSA

MM-PBSA (Molecular Mechanics/Poisson-Boltzmann/Surface Area) is a method for estimating the free energy in molecular complexes [30, 31]. It was initially used to study DNA and RNA fragments but has also been used to calculate binding free energies between proteins and small ligands [32] as well as protein-protein complexes [33]. MM-PBSA uses a continuum solvent model and estimates the binding free energy according to:

$$\langle G \rangle = \langle E_{MM} \rangle + \langle E_{PBSA} \rangle - T \langle S_{MM} \rangle \tag{24}$$

$\langle E_{MM} \rangle$ is the average energy obtained from a typical molecular mechanics force-field with contributions from bond stretching, angle bending, torsions, electrostatic and van der Waals terms without any non-bonded cut-offs. $\langle E_{PBSA} \rangle$ is the solvation free energy and consist off polar and non-polar contributions. The polar contribution is typically calculated from the Poisson-Boltzmann equation [34], while the non-polar solvation free energy is estimated from a solvent-accessible-surface-area term [35]. $-T \langle S_{MM} \rangle$ is the solute entropy and can be estimated using normal mode analysis [30]. There are two ways to estimate the binding free energy. The first is to examine the MM-PBSA equation (Equation 25) for the complex, receptor and ligand, and calculate the binding free energy according to:

$$\Delta G_{bind} = \langle G_{complex} \rangle - \langle G_{receptor} \rangle - \langle G_{ligand} \rangle \tag{25}$$

This approach is not suitable for protein-ligand or protein-protein systems because the $\langle E_{MM} \rangle$ term does not converge within a within a reasonable computing time. The second approach is to solve the MM-PBSA equation using MD snapshots from the

simulation of the complex only and contributions from the receptor or the ligand are calculated by removing the other molecule from the trajectory. This approach assumes the receptor and ligand will explore similar conformations both in the complex and free in the solution. This is not always the case and studies using the LIE method have revealed that the ligand often adopts different conformations while free in the solution compared to in a complex [27].

1.4 Protein-protein interactions and the LIE method

The physical principles for protein-protein interactions are the same as for protein-ligand interactions, it can however be expected to have more complex enthalpic and entropic contributions. In fact, this makes computer simulations for determining absolute binding energies in protein-protein complexes extremely difficult. Because of extremely large interaction energies, approaches such as the LIE method would require extremely long simulations in order to obtain stable averages and convergent energies. Such an approach is therefore clearly inefficient. Accurate estimations of absolute protein-protein associations are not feasible by today's methodologies.

Protein-protein interfaces are usually composed of hot spots where a few residues make up for almost all the binding energy [36]. The interface between serine proteases and their canonical inhibitors is composed of hot spot residues. In the case of the complex between bovine pancreatic trypsin inhibitor (BPTI) and trypsin, 70% of the binding energy comes from the P1 residue [37]. Another way to calculate binding energies for protein-protein complexes is possible. Instead of absolute binding energies, one can calculate the relative binding energy for some of the hot spot residues. Although this approach cannot give the absolute binding energies, relative energies are useful to examine effects of point mutations. Analysis of crystal structures of different P1 variants of BPTI with trypsin shows that the secondary interactions are almost identical regardless of the P1 side chain [38]. The P1-Gly variant is suitable as reference state, as glycine does not have a side chain that enters the S1 site and the binding energy of the complex will therefore come from the secondary interaction sites only.

When using the LIE method to study the relative binding energies of point mutations, the mutated residues are treated as ligand in the LIE framework. The rest of the protein and inhibitor is treated as the surroundings. Several P1 mutational

studies of complexes have been done using the LIE method in this manner. This strategy has proven to be useful since the ligand-surrounding interactions converge rapidly in MD simulations.

The LIE parameters in this kind of protein-protein complex studies have been found to be quite different than in protein-ligand simulations [39]. β assumes values ranging from 0.33 to 0.50 as noted earlier. The reason for this lies in relaxing the linear response approximation, and is related to the number of hydrogen bonds the residue can form. Contrary to protein-ligand simulations where $\alpha = 0.18$, protein-inhibitor simulations have accurately reproduced experimental results with $\alpha \approx 0.5$. γ is not needed to calculate relative binding energies but is necessary for the absolute binding free energy [39].

1.5 Aims of the study

The purpose of the project is to study the effect of point mutations at protein-protein interfaces through the use of computer based models to accurately predict absolute and relative binding energies.

1. Use an available crystal structure of Streptomyces griseus proteinase B (SGPB) in complex with the third domain of the turkey ovomucoid inhibitor (OMTKY3), and perform point mutations on the P1 residue.
2. Create a model of porcine pancreatic elastase (PPE) in complex with the third domain of the turkey ovomucoid inhibitor by docking PPE with OMTKY3, and perform point mutations on the P1 residue.
3. Subject the models to molecular dynamics simulations and use the linear interaction energy method to calculate binding energies.
4. Investigate to what degree the linear interaction energy method reproduces experimental binding energies and evaluate the applicability of the LIE method.
5. Suggest and examine improvements of the methodology.

2 Methods

2.1 Model building

The starting point for the SGPB-OMTKY3 model was obtained from the 1SGQ [3] protein data bank (pdb) structure. The pdb file was first edited to the format required by the Q [40] program package. This involved removing excess crystallographic information from the file, so that only the atomic coordinates of the protein remained and removing multiple configurations of side chains refined with alternate conformations. Crystallographic water molecules and ions were removed so that only the complex remained.

There is presently no crystal structure of PPE in complex with OMTKY3 available, and it was therefore necessary to build a model of this complex. The 1QNJ [41] structure was used as template with a 1.1 Å resolution of the native PPE structure. The OMTKY3 structure was obtained from 1SGQ [3] which contains SGPB in complex with OMTKY3. Because PPE and SGPB are serine proteases and OMTKY3 is a canonical inhibitor, the binding interactions are expected to be similar. Manual docking was carried out using PyMol [6] aligning the catalytic triad of PPE onto SGPB-OMTKY3. Visual inspection revealed a good match around the binding site, requiring no further alignment. Certain close contacts remained, in particular the 145-150 loop and the 215-220 loop from PPE caused a number of close contacts. The loops were shifted using the modelling program O [42] removing high energy contacts. Finally the PPE-OMTKY3 system was subjected to a 0.1 ns MD simulation using the Amber95 force-field [23], to further remove any remaining steric clashes and optimize bond angles etc., which had been strained when the loops were shifted. The result was a working model of the PPE-OMTKY3 complex.

P1 mutation and minimization

Point mutations of the P1 residue were carried out using the Maestro 9.1 software package [43] and a total of 28 P1 variants were constructed for each complex. Although only 20 amino acids exist naturally, several configurations exist for some of the amino acids, such as glutamic acid which can exist in either charged or uncharged form. To optimize the geometry after mutating the P1 residue each complex was run through an energy minimization using MacroModel [44] and the OPLS2005 [24, 45]

force field. The OMTKY3-Water model was made by manually removing the PPE atoms from the PPE-OMTKY3 complex.

2.2 Molecular dynamics simulations

MD simulations and free energy calculations were carried out using the program package Q [40]. The Q package consists of four different programs, Qprep, Qdyn, Qfep and Qcalc. Each program has its own role in the simulation timeline as seen in Figure 2.1. Qprep is used to generate topologies for the system which contains all the necessary information for the MD simulation. The finished topologies are then read by Qdyn which is used to run the simulations themselves. Analysis and extraction of energies and trajectories are done using Qfep and Qcalc.

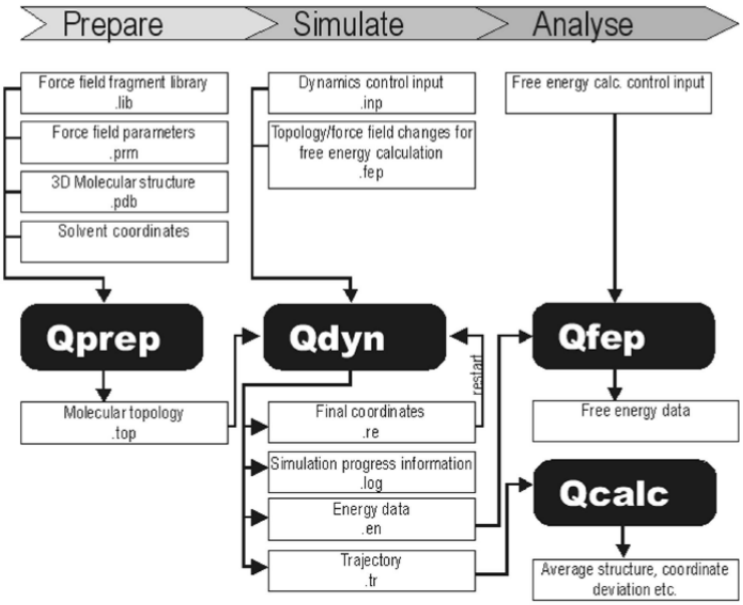

Figure 2.1: Flowchart of the program structure of Q, displaying typical input and output and common file extensions.

The final structures for all P1 variants of the SGPB-OMTKY3, PPE-OMTKY3 and OMTKY3-Water systems were subjected to MD simulations with the program package Q [40] using the OPLS-AA force field [24]. The atomic structure for each P1 variant was loaded into Qprep5 along with the force-field to generate the topology. Qdyn5 was then used to run the simulations themselves. All simulations were carried out using a 20 Å sphere with the C_α-atom of the P1 residue as the centre.

Atoms in the outermost 3 Å of the sphere were weakly restrained to their crystallographic coordinates by a harmonic potential of 5 kcal/mol Å^2 and atoms outside the 20 Å sphere was strongly restrained by a harmonic potential of 200 kcal/mol Å^2. Both the equilibration phase and the production phase used the same set of cut-offs. The P1 residue was allowed to interact with the entire system without any truncations. Non-bonded forces were explicitly calculated within 10 Å, while contributions outside this range were estimated using the local reaction field (LRF) method [46]. The LRF method uses Taylor expansion to approximate the long range interactions and is computationally faster than the calculating the forces explicitly. The nonbonded pair list was updated every 25 steps. SHAKE [25] was used to restrain the bonds and angles of the solvent molecules during the entire MD simulation. The first part of the MD simulation was an equilibration phase where the complex was heated from 1 K to 300 K using a stepwise scheme and time step of 1.0 fs for the first 16 000 steps and 1.5 fs during the latter 60 000 steps, for a total equilibration time of just over 0.1 ns. The production phase was 3 000 000 steps at 300 K with a time step of 1.5 fs. A total of three parallel series was run for each P1-variation for a total of 9 000 000 production steps equalling 13.5 ns of simulation time. The simulations were done on the supercomputer Stallo [47] using the multiprocessor version of Qdyn5 set up to utilize 4 threads.

The integrity of the finished MD simulations was verified by assessing the root mean square deviation (RMSD) of the atomic positions obtained from the trajectories according to:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N_{atoms}} d_i^2}{N_{atoms}}} \tag{26}$$

where d_i is the distance between current and initial atomic coordinates.

Treatment of ionisable residues

Ionisable residues (Asp, Glu, Arg, Lys and His), within the simulation sphere were by default treated as charged. Some modification to the scheme was necessary to avoid potential artificial energy contributions, known as Born terms [48]. Due to the way long range electrostatics are calculated, contributions outside the simulation cut offs are neglected. This energy contribution is dominated by the Born term:

$$\Delta G_{Born} = -\left(\frac{Q^2}{2r}\right)\left(1-\frac{1}{\varepsilon}\right) \quad (27)$$

where Q is the charge of the system, r is the radius of the simulation sphere and ε is the dielectric constant of the surroundings. In order to avoid Born terms from entering the into the calculated free energy the system must have the same net charge, the same size as described by the simulation sphere and surroundings with the same dielectric constant.

Two solutions have been suggested to solve this problem. One way is to include counter ions to ensure the net charge is the same within the simulation series. Although this approach technically solves the Born term problem, the mobility of the counter ions hampers the convergence of the simulation. The preferred method is to neutralize residues as close to the edge of the boundary as possible, in order to avoid artificial effects related to the solvation of charges.

In these simulation series, SGPB-OMTKY3, PPE-OMTKY3 and OMTKY3-Water, the net charge was set to zero. This was done by the latter method of neutralizing residues close to the edge of the simulation sphere. The different protein systems required different charge distributions. It was decided to use a net charge of zero and this more or less dictated which amino acids should be charged. Overviews of the chargeable residues within the simulation sphere for the different systems are found in Table 2.1, Table 2.2 and Table 2.3

Table 2.1: List of chargeable residues within 17Å for SGPB-OMTKY3.

Residue	Number	Protein	Distance (Å)*	Charge
Arg	41	SGPB	9.3	+1
Asp	60	SGPB	14.2	-1
Arg	81	SGPB	15.9	+1
Asp	102	SGPB	8.9	-1
Arg	138	SGPB	9.8	+1
Arg	139	SGPB	16.8	+1
Asp	175	SGPB	14.5	-1
Arg	182	SGPB	16.8	+1
Glu	192A	SGPB	10.2	-1
Asp	194	SGPB	7.6	-1
Lys	13	OMTKY3	13.8	0
Glu	19	OMTKY3	5.2	-1
Arg	21	OMTKY3	11.6	+1
Lys	29	OMTKY3	14.6	0
Lys	34	OMTKY3	14.4	0
Glu	43	OMTKY3	16.2	0
Lys	55	OMTKY3	16.5	0

* Distance measured from simulation centre to charged atoms at start of simulation.

Table 2.2: List of chargeable residues within 17Å for PPE-OMTKY3.

Residue	Number	Protein	Distance (Å)*	Charge
Val	16	PPE	9.3	+1
Asp	60	PPE	12.1	-1
Arg	61	PPE	9.8	+1
Asp	97	PPE	16.9	-1
Asp	98	PPE	12.3	-1
Asp	102	PPE	8.5	-1
Asp	194	PPE	7.5	-1
Arg	217A	PPE	12.5	+1
Lys	224	PPE	13.2	0
Lys	13	OMTKY3	13.2	+1
Glu	19	OMTKY3	5.3	-1
Arg	21	OMTKY3	9.8	+1
Lys	34	OMTKY3	14.5	+1
Glu	43	OMTKY3	16.9	0
Lys	55	OMTKY3	15.5	0

* Distance measured from simulation centre to charged atoms at start of simulation.

Table 2.3: List of chargeable residues within 17Å for OMTKY3-Water.

Residue	Number	Protein	Distance (Å)*	Charge
Lys	13	OMTKY3	13.2	0
Glu	19	OMTKY3	5.3	-1
Arg	21	OMTKY3	9.8	+1
Lys	34	OMTKY3	14.5	0
Lys	55	OMTKY3	15.5	0

* Distance measured from simulation centre to charged atoms at start of simulation.

2.3 LIE model and analysis

The linear interaction energy method was used to obtain absolute binding energies for the SGPB-OMTKY3 and PPE-OMTKY3 complexes. In the LIE framework the P1 residue is treated as a ligand and the binding energy is calculated as the difference in P1-surroundings interaction energies between bound and unbound states. Qfep5 was used to obtain the ligand-surroundings interaction energies, where the P1 residue is the ligand. Since multiple simulations were used for each P1 variant, the average interaction energy was calculated. The binding energy was then estimated according to the LIE equation: $\Delta G_{bind} = \alpha\Delta\left\langle V_{l-s}^{vdW}\right\rangle + \beta\Delta\left\langle V_{l-s}^{el}\right\rangle + \gamma$

An initial parameterisation of α = 0.58 was chosen and γ was adjusted to allow the P1-Gly variants to reproduce the experimental binding energies exactly, and β is fixed depending on the nature of the P1 residue. The free LIE variables, α and γ, were then optimized to minimize the RMSD of the error, measured as the difference in calculated and experimental binding energies. To evaluate the reliability of the simulations and the calculated binding energies, the energies and trajectories were examined using Qfep5 and Qcalc5. PyMol [6] was used to visualize structures and generate illustrations of both MD snapshots and MD average structures.

3 Results and discussion

3.1 Construction of molecular complexes

As mentioned, there is presently no experimental structure of OMTKY3 in complex with PPE. Secondary interactions between serine proteinases and their canonical inhibitors are almost identical, and it was therefore decided to attempt a manual alignment between PPE and OMTKY3 using the SGPB-OMTKY3 complex as a template. The catalytic triad of PPE was aligned using PyMol and an RMSD of 0.245 Å was obtained for the catalytic residues. Figure 3.1 shows the result of the structural alignment and the similarities between the two enzymes are apparent.

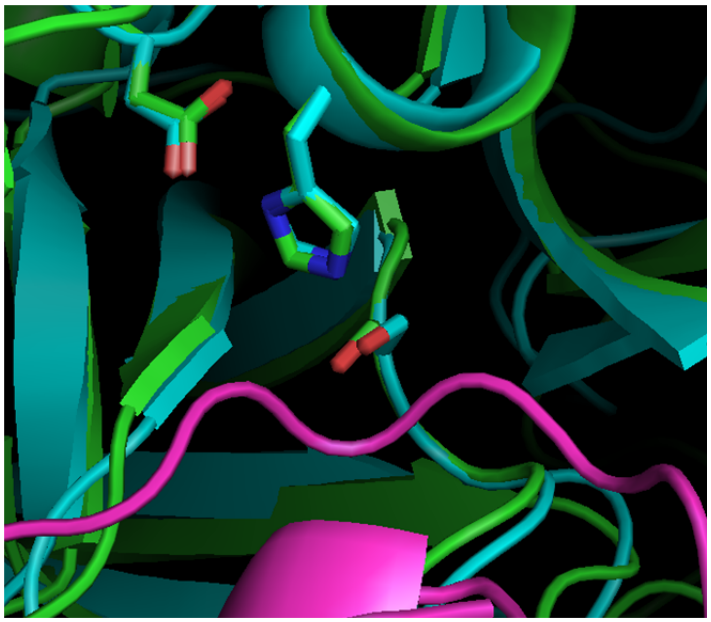

Figure 3.1: Aligning the catalytic triad (drawn as sticks) resulted in a high degree of overlap between the SGPB and PPE enzymes, showing PPE (green), SGPB (cyan) and OMTKY3 (pink).

The next step was to produce a working model of the PPE-OMTKY3 complex, and the structure was therefore refined. Although surface plots show a good match around the primary binding site (Figure 3.2) a number of steric clashes remained. The clashes were primarily caused by the 145-150 and the 215-220 loops, and most notably by Arg217 as seen in the surface plot in Figure 3.3. Manual shifting of the loops using O removed the worst contacts. In order to further relax the model, a short MD simulation of 0.1 ns was carried out, and the last structure from this simulation was then used as a template for further work on the PPE-OMTKY3

complex. Figure 3.4 show the effect of the MD simulation and the effect this has on the surface plot.

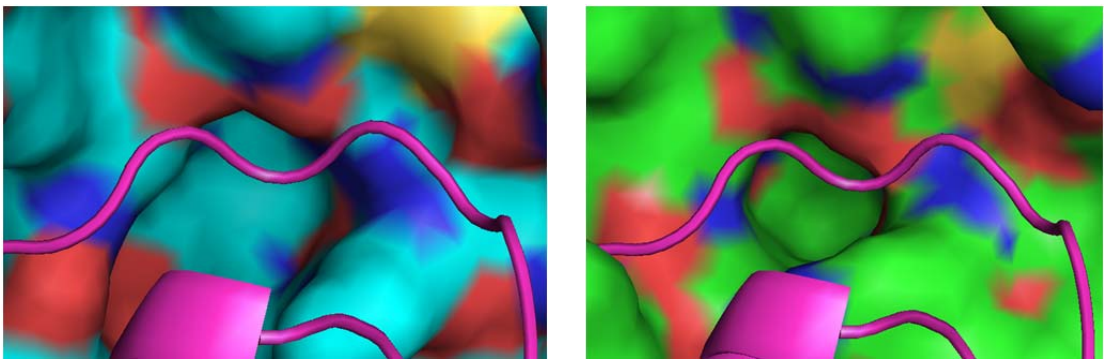

Figure 3.2: (Left) Surface plot of the SGPB binding site, showing the OMTKY3 binding loop. (Right) Surface plot of the PPE binding site, showing the OMTKY3 binding loop.

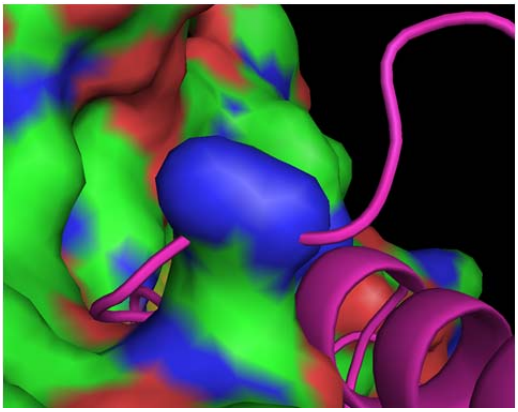

Figure 3.3: Steric clash between PPE and OMTKY3 immediately after alignment. The OMTKY3 main chain is seen passing through the Arg217 side chain.

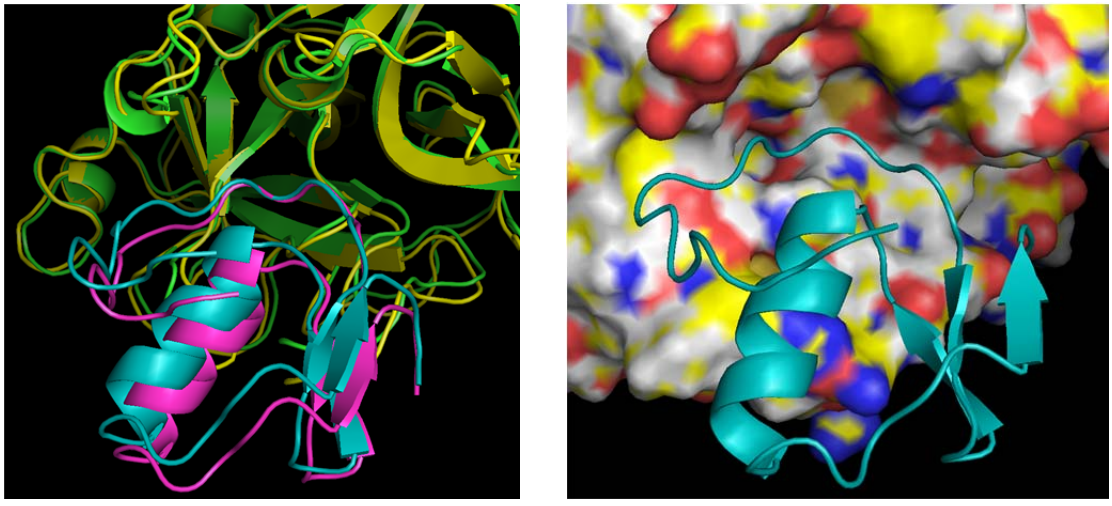

Figure 3.4: (Left) Comparison between the initial alignment and refined model of the PPE-OMTKY3 complex. The initial model is coloured in green and pink, and the refined model is yellow and teal. (Right) Surface plot of the refined PPE-OMTKY3 model, showing the ribbon structure of the inhibitor now adapted to the surface of the PPE enzyme.

3.2 Point mutation and energy minimization

Both PyMol and Maestro were used to construct the P1 mutants. For most P1 variants it was possible to obtain models with PyMol, however, when attempting to fit large residues into the narrow S1 site on PPE steric clashes were significant. The Maestro software package allowed for easy mutation as well as good fitting of the P1 side chain, and the subsequent energy minimization run using MacroModel [44] removed steric clashes and optimized the geometry. The resulting structures were successfully subjected to MD simulations.

The pdb files generated by Maestro could not be read by Q due to different naming schemes for atoms, in particular hydrogen atoms. Perl scripts were made to change the names first from Q-format to Maestro-format and then back to Q-format after completing the P1 mutation and energy minimization. The resulting structures were subjected to MD simulations and free energy calculations using the LIE method.

3.3 Molecular dynamics simulations

MD simulations were carried out for both the SGPB-OMTKY3 complex and the PPE-OMTKY3 complex, yielding a total of 56 systems, differing only in the P1 position. All 20 commonly occurring amino acids were simulated as well as different protonation states where applicable, increasing the number to 28 per complex and 56 in total. Each calculation consisted of three parallel MD simulations using the same protocol, consisting of an equilibration phase and a production phase. The parallels used different random seed for the equilibration phase. In the equilibration phase the system was heated stepwise from 0 to 300 K and the production phase was run at 300 K over 3 000 000 steps with a time step of 1.5 fs, for a total simulation time of 4.5 ns.

An identical setup was used for the OMTKY3-Water simulations, in total 28 simulations. The OMTKY3-Water models were made by removing PPE from the PPE-OMTKY3 complex and were then subjected to the same MD scheme as the protein-protein complexes. Thus 84 simulations were run in three parallels and the final energies were collected from averages over all three series.

Choice of time step

To reduce the computational cost of the planned MD simulation a time step of 2.0 fs was originally chosen. Compared to a time step of 1.5 fs, this reduces the number of steps required for a given simulation time by 25 % and would in the long run save thousands of computing hours. The MD simulations started at 1 K in order to relax the molecular system and allowing it to adapt to the chosen force field. Hot atoms were observed in the beginning of the MD simulations due to atoms coming too close and vdW repulsion moves the atoms apart again. In itself this is quite common as most crystal structures include close contacts that triggers repulsions during the first stages of equilibration and then disappears as the simulation progresses. In particular, crystal structures are refined with different force field parameters than what is normally used in traditional MD simulations, and the systems are normally relaxed prior to MD to accommodate new parameters.

However this can also happen at a later stage if the time step is too large, as a result of instabilities in the simulations. As MD integrates Newton's laws and assigns each atom with a speed and acceleration according to the Taylor expansion it does so for the entire time step, it is possible for two atoms to move closer than they could naturally. If the atoms are too close the repulsion forces, which usually scale at r^{-12} in most force-fields, will throw the atoms apart violently and can cause the simulation to "blow up".

Test simulations carried out on the PPE-OMTKY3 complex as well as OMTKY3 in water yielded mixed results. Some of the simulations completed successfully while others crashed during the equilibration phase. Closer examination of the simulations revealed that this was caused by SHAKE failure. Normally, SHAKE failure is a symptom of something wrong in the simulations and not the SHAKE algorithm itself. SHAKE was used only on hydrogen atoms and on solvent molecules, and an update frequency of 25 was used for the non-bonded pair list. To solve this various changes were tested. The simulation sphere was increased but without consistent effect. Different random seeds were tested with the same result. Since all the hot atoms were hydrogen, removing them and then let Q reinsert them into a hopefully less strained configuration was also tested with the aforementioned result. As neither approach could produce stable MD simulations the time step was reduced to 1.5 fs. The new simulation was set up to run 3 000 000 steps with a time

step of 1.5 fs, in total 4.5 ns of simulation time per parallel. When reducing the time step to 1.5 fs all simulations finished successfully.

3.3.1 Molecular dynamics quality

Before analysing the energetics associated with point mutations, the quality of the MD simulations must be sufficient. In general, the quality of MD simulations can be addressed by examining the physical properties of the simulations (e.g. temperature, total energy, pressure, volume etc.) and the quality of the simulated structures. The variance in the temperature and the total energy was examined for all simulations, and are stable throughout the production phase. The temperature was set to 300 K, but was found to vary within ±5 K. This indicates that the bath coupling is correctly set, as too weak or strong coupling can give artificial trajectories. Another property examined was the root-mean-squared deviation with respect to the initial structure. The RMSD is in the range 0.6 to 1.0 Å excluding hydrogen atoms and solvent molecules, confirms that the simulations maintain the overall structure. Figure 3.5 and Figure 3.6 show plots of the RMSD as a function of simulation time for the P1-Tyr mutant of the SGPB-OMTKY3 and PPE-OMTKY3 complexes obtained from the first parallel. RMSD plots from the other P1 mutants show similar results.

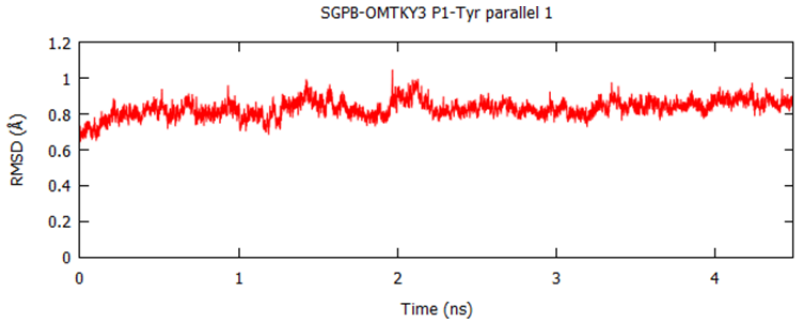

Figure 3.5: RMSD plot for first parallel of the P1-Tyr variant from the SGPB-OMTKY3 complex, in relation to the starting structure.

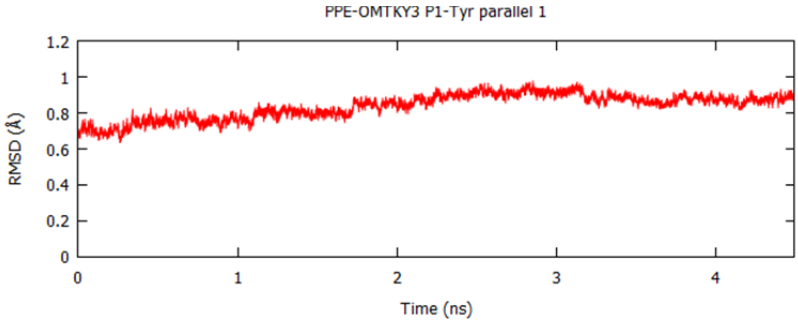

Figure 3.6: RMSD plot for first parallel of the P1-Tyr variant from the PPE-OMTKY3 complex, in relation to the starting structure.

3.4 LIE calculations

Complexes between serine proteinases and their canonical inhibitors have almost identical secondary interactions regardless of the P1 residue [3, 49]. In the LIE method the P1 residue is treated as a ligand and the binding energy is calculated as the difference between bound and unbound states with the remainder of the complex forming the surroundings. The LIE equation use three parameters to approximate the binding energy, α scales the vdW energy, β scales the electrostatic energy and the final term γ corresponds to the binding energy contributions from the secondary interactions and is only necessary to reproduce absolute binding energies. A key assumption in the LIE method is that the secondary interactions remain the same and for serine proteinases this is generally true as noted earlier. The constant term γ is therefore the same for all P1 variants, but varies between different proteinases.

The LIE variables were initially estimated, α was set to 0.58 as this has proved to be a good value [39]. The γ variable corresponds to the binding energy contribution from the secondary interactions. P1-Gly with only a single hydrogen as its side chain has virtually no primary interactions, all binding energy contributions come from the secondary interactions. Based on this, γ was adjusted to allow the calculated binding energy for P1-Gly to exactly reproduce the experimental binding data. Using this estimate the most favourable configurations for the P1 residues were chosen and the LIE variable estimates were optimized for the new selection of P1 variants and the result is displayed in Table 3.1.

In the initial LIE equation the β variable was set to 0.50, but FEP investigations of the various amino acids have calculated different β values, 0.37, 0.43 or 0.50 depending on the chemical nature of the side chain [50]. Deviations from the linear response value of 0.50 are primarily caused by functional groups capable of extensive hydrogen bonding. Hydroxyl groups were found to deviate most from the value of 0.50.

Certain amino acids can undergo protonation or deprotonation depending on the pH. The experimental association energies were obtained at pH = 8.3 [4] and at this value P1-Arg, P1-Lys, P1-Asp and P1-Glu can be expected to be ionized when the inhibitor is free. Since both simulations require the inhibitor to be of the same state, the binding energy must be corrected free energy required to protonate/deprotonate chargeable residues. This is calculated according to:

$$\Delta\Delta G_{bind}^{pH} = 1.35\left|pH - pK_a\right| \quad (28)$$

where pH = 8.3 and pK_a of the amino acid in question is used. This result in free energy penalty upon protonation of 5.8, 5.1, 5.7, 3.2 and 3.1 kcal/mol for Asp, Glu, Arg, Lys and His respectively.

Qfep5 was used to obtain vdW and electrostatic molecular mechanics averages between the ligand and the surroundings. In the LIE framework the P1 residue is treated as a ligand while the rest of the protein is treated as the surroundings. The energies obtained from the SGPB-OMTKY3, PPE-OMTKY3 and OMTKY3-water series are listed in Table 3.2 and Table 3.3 and the LIE equation is used to calculate the binding energy according to $\Delta G_{bind} = \alpha\Delta\left\langle V_{l-s}^{vdW}\right\rangle + \beta\Delta\left\langle V_{l-s}^{el}\right\rangle + \gamma$. In addition the experimental binding energies from Lu et al. [4] are shown for comparison. Scatter diagrams of calculated and experimental binding free energies are found in Figure 3.7 and Figure 3.8.

Table 3.1: Optimized LIE variables for SGPB-OMTKY3 and PPE-OMTKY3 complexes.

Enzyme	α	β	γ	RMS (kcal/mol)#	$\langle\lvert Error\rvert\rangle$ (kcal/mol)#
SGPB*	0.45	FEP	-8.7	1.27	1.16
PPE*	0.55	FEP	-9.2	1.18	1.06

* Only one P1 variant for each amino acid included in optimization and outliers with error > 2.0 kcal/mol are not included in the optimization.

RMS and $\langle\lvert Error\rvert\rangle$ is calculated for the P1 variants included in the optimization.

Table 3.2: Ligand-surroundings interaction MD average energies for P1 variants of OMTKY3 in water and bound to SGPB.

	SGPB-OMTKY3		OMTKY3-Water			
P1 residue	$\langle V_{l-s}^{el} \rangle$	$\langle V_{l-s}^{vdw} \rangle$	$\langle V_{l-s}^{el} \rangle$	$\langle V_{l-s}^{vdw} \rangle$	ΔG_{bind}^{calc} *	$\Delta G_{bind}^{\exp}$ **
Gly	-71.6	-6.2	-70.7	-4.0	-10.1	-9.5
Ala	-72.1	-9.2	-70.4	-6.0	-10.9	-11.5
Arn[n]	-95.8	-23.0	-102.2	-11.8	-5.3	-11.1
Arg[c]	-177.3	-17.8	-197.7	-5.6	-4.1	-11.1
Asn	-89.9	-15.8	-89.8	-7.6	-12.5	-11.1
Asp[c]	-201.4	-8.1	-218.2	-0.1	-3.9	-8.8
AspH1[a]	-83.6	-17.3	-92.8	-7.4	-4.0	-8.8
AspH2[a]	-83.1	-17.3	-83.2	-8.0	-7.0	-8.8
Cys	-77.1	-12.8	-72.2	-7.9	-13.1	-14.4
Gln	-92.6	-18.6	-92.1	-10.3	-12.7	-11.9
Glu[c]	-200.7	-11.6	-220.5	-1.7	-3.2	-8.5
GluH1[a]	-85.3	-20.1	-93.2	-10.6	-5.0	-8.5
GluH2[a]	-85.7	-18.8	-93.4	-10.1	-4.7	-8.5
Hid[b]	-88.7	-21.6	-91.6	-12.0	-11.9	-12.7
Hie[b]	-87.4	-20.7	-91.8	-11.8	-10.8	-12.7
Hip[c]	-166.8	-16.5	-180.3	-7.0	-3.1	-12.7
Ile	-71.1	-20.2	-69.1	-12.7	-13.0	-10.0
Leu	-72.5	-19.9	-69.8	-12.4	-13.3	-14.4
Lyn[n]	-83.5	-20.2	-83.2	-11.3	-9.7	-11.3
Lys[c]	-197.0	-13.9	-208.6	-5.3	-6.8	-11.3
Met	-77.1	-19.9	-77.1	-12.5	-12.1	-14.0
Phe	-74.6	-24.5	-74.0	-15.2	-13.2	-13.1
Pro	-49.3	-14.2	-47.4	-8.9	-12.0	-6.1
Ser	-86.4	-9.0	-85.1	-4.8	-11.1	-10.3
Thr	-84.1	-13.0	-81.6	-7.2	-12.2	-11.3
Trp	-82.1	-30.2	-81.9	-18.4	-14.2	-12.6
Tyr	-85.2	-25.0	-84.8	-14.6	-13.5	-12.8
Val	-72.1	-16.9	-68.2	-10.3	-13.3	-11.4

Energies listed in kcal/mol

* Energies calculated using the LIE equation: $\Delta G_{bind}^{calc} = \alpha\Delta < V_{l-s}^{vdW} > + \beta\Delta < V_{l-s}^{el} > + \gamma$

** Experimental binding affinities from Lu et al. [4].

Binding energy corrected with free energy required to protonate/deprotonate the side chain based on: $\Delta\Delta G_{bind}^{pKa} = 1.35(pH - pKa)$

[a] Refers to either variant of protonated, neutral, aspartic or glutamic acid.

[b] Hid and Hie refers to protonation state of neutral histidine.

[c] Refers to charged variant of P1 residue (Asp, Glu, Arg, Lys or His).

[n] Refers to protonated, neutral, variant of arginine or lysine.

Table 3.3: Ligand-surroundings interaction MD average energies for P1 variants of OMTKY3 in water and bound to PPE.

	PPE-OMTKY3		OMTKY3-Water			
P1 residue	$\langle V_{l-s}^{el} \rangle$	$\langle V_{l-s}^{vdw} \rangle$	$\langle V_{l-s}^{el} \rangle$	$\langle V_{l-s}^{vdw} \rangle$	ΔG_{bind}^{calc} *	$\Delta G_{bind}^{\exp}$ **
Gly	-72.2	-7.04	-70.7	-4.0	-11.5	-12.0
Ala	-73.1	-10.9	-70.4	-6.0	-13.1	-14.2
Arn[n]	-85.7	-22.8	-102.2	-11.8	-2.5	-4.9
Arg[c]	-145.2	-18.0	-197.7	-5.6	10.2	-4.9
Asn	-85.7	-18.5	-89.8	-7.6	-13.5	-10.5
Asp[c]	-186.6	-10.4	-218.2	-0.1	0.9	-6.5
AspH1[a]	-90.2	-17.1	-92.8	-7.4	-7.8	-6.5
AspH2[a]	-72.9	-18.9	-82.9	-7.8	-5.8	-6.5
Cys	-76.7	-13.9	-72.2	-7.9	-14.4	-13.9
Gln	-81.0	-21.3	-92.7	-10.4	-10.2	-10.2
Glu[c]	-167.1	-15.6	-220.5	-1.7	9.9	-6.6
GluH1[a]	-74.2	-22.0	-93.2	-10.6	-3.4	-6.6
GluH2[a]	-73.8	-21.5	-93.4	-10.1	-3.2	-6.6
Hid[b]	-86.0	-19.7	-91.6	-12.0	-11.1	-6.1
Hie[b]	-77.1	-22.7	-91.8	-11.8	-8.9	-6.1
Hip[c]	-127.0	-17.1	-180.3	-7.0	15.0	-6.1
Ile	-73.4	-19.4	-69.1	-12.7	-14.8	-13.1
Leu	-74.4	-19.4	-69.8	-12.4	-15.0	-14.2
Lyn[n]	-75.1	-21.2	-83.2	-11.3	-8.0	-6.3
Lys[c]	-152.3	-15.9	-208.6	-5.3	13.1	-6.3
Met	-73.9	-21.3	-77.1	-12.5	-12.7	-13.6
Phe	-70.3	-22.1	-74.0	-15.2	-11.5	-6.3
Pro	-50.3	-13.7	-47.4	-8.9	-13.1	-7.7
Ser	-79.1	-11.8	-85.1	-4.8	-10.8	-12.0
Thr	-83.1	-12.9	-81.6	-7.2	-12.9	-14.0
Trp	-76.7	-30.9	-81.9	-18.4	-13.9	-5.9
Tyr	-79.0	-24.5	-84.8	-14.6	-12.5	-5.2
Val	-74.8	-15.6	-68.2	-10.3	-14.9	-13.3

Energies listed in kcal/mol

* Energies calculated using the LIE equation:

$$\Delta G_{bind}^{calc} = \alpha \Delta < V_{l-s}^{vdW} > + \beta \Delta < V_{l-s}^{el} > + \gamma$$

** Experimental binding affinities.

Binding energy corrected with free energy required to protonate/deprotonate the side chain based on: $\Delta\Delta G_{bind}^{pKa} = 1.35(pH - pKa)$

[a] Refers to either variant of protonated, neutral, aspartic or glutamic acid.

[b] Hid and Hie refers to protonation state of neutral histidine.

[c] Refers to charged variant of P1 residue (Asp, Glu, Arg, Lys or His).

[n] Refers to protonated, neutral, variant of arginine or lysine.

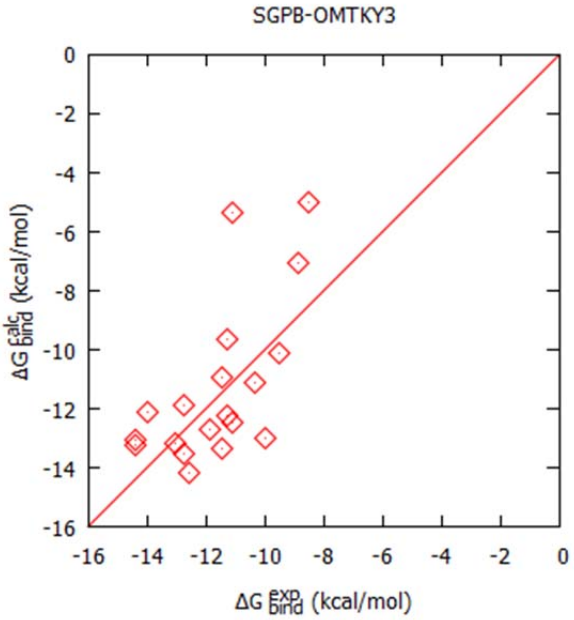

Figure 3.7: Scatter diagram of calculated and experimental binding energies for SGPB-OMTKY3, using the optimized LIE variables. Only the best P1 variants are drawn, while P-Pro is omitted.

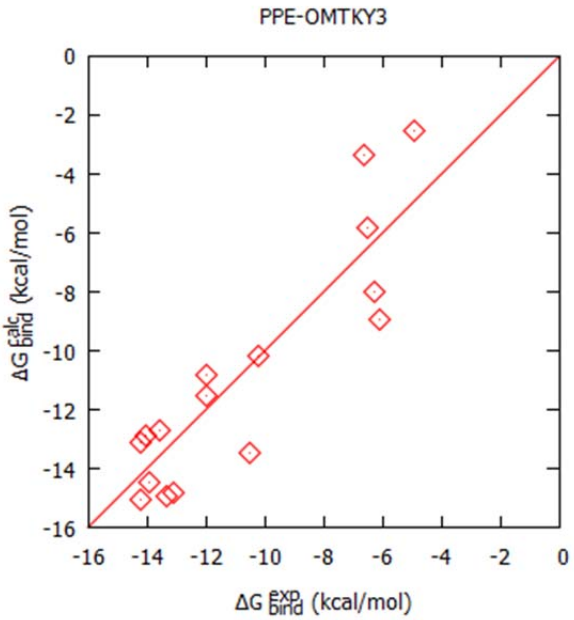

Figure 3.8: Scatter diagram of calculated and experimental binding energies for PPE-OMTKY3, using the optimized LIE variables. Only the best P1 variants are drawn, while P1-Phe, P-Pro, P1-Trp, P1-Tyr are omitted.

The chosen LIE model reproduces the experimental binding energies within acceptable accuracy for most P1 variants. Where multiple configurations of the P1 residue are available only the one who reproduces the experimental binding energies best were used during the optimization of the LIE variables. Certain P1 variants, in particular the charged residues like protonated Arg are calculated to bind much worse than what is found experimentally. The most obvious explanation is that P1-Arg is uncharged. As such the binding energy obtained from unfavourable configurations are ignored and only one result is chosen for each P1 variant with multiple configurations. In previous protein-protein simulations using the LIE method P1-Pro has been troublesome [51]. The restrictions to binding angles imposed by proline's side chain have been difficult to model accurately and the binding energy obtained from P1-Pro is omitted when optimizing the LIE variables.

After the LIE optimization the mean unsigned error for the SGPB-OMTKY3 complex was 1.82 kcal/mol and for PPE-OMTKY3 it was 2.45 kcal/mol. This is somewhat higher than comparative LIE simulations [51], which report mean unsigned errors of 0.69, 0.43 and 1.01 kcal/mol. However these figures includes several outliers, and if P1-Pro is excluded the figure drops to 1.61 kcal/mol for SGPB-OMTKY3. P1-Pro and the aromatic residues (Phe, Trp and Tyr) are problematic in the PPE-OMTKY3 simulations and when these are dropped the new figure is 1.45 kcal/mol.

One of possible source for the discrepancy between theoretical and experimental binding energies can be related to convergence of the interactions energies between the P1 residue and its surroundings. Figure 3.9 show how the electrostatic and the van der Waals interaction energies vary as a function of simulation time. Inspection of P1–surroundings interaction energies reveal no sudden jumps and that both the electrostatic and van der Waals component rapidly reach a stabile plateau. Three independent MD simulations were carried out and comparisons of the individual free energies are very similar, indicating stable structures and energetics.

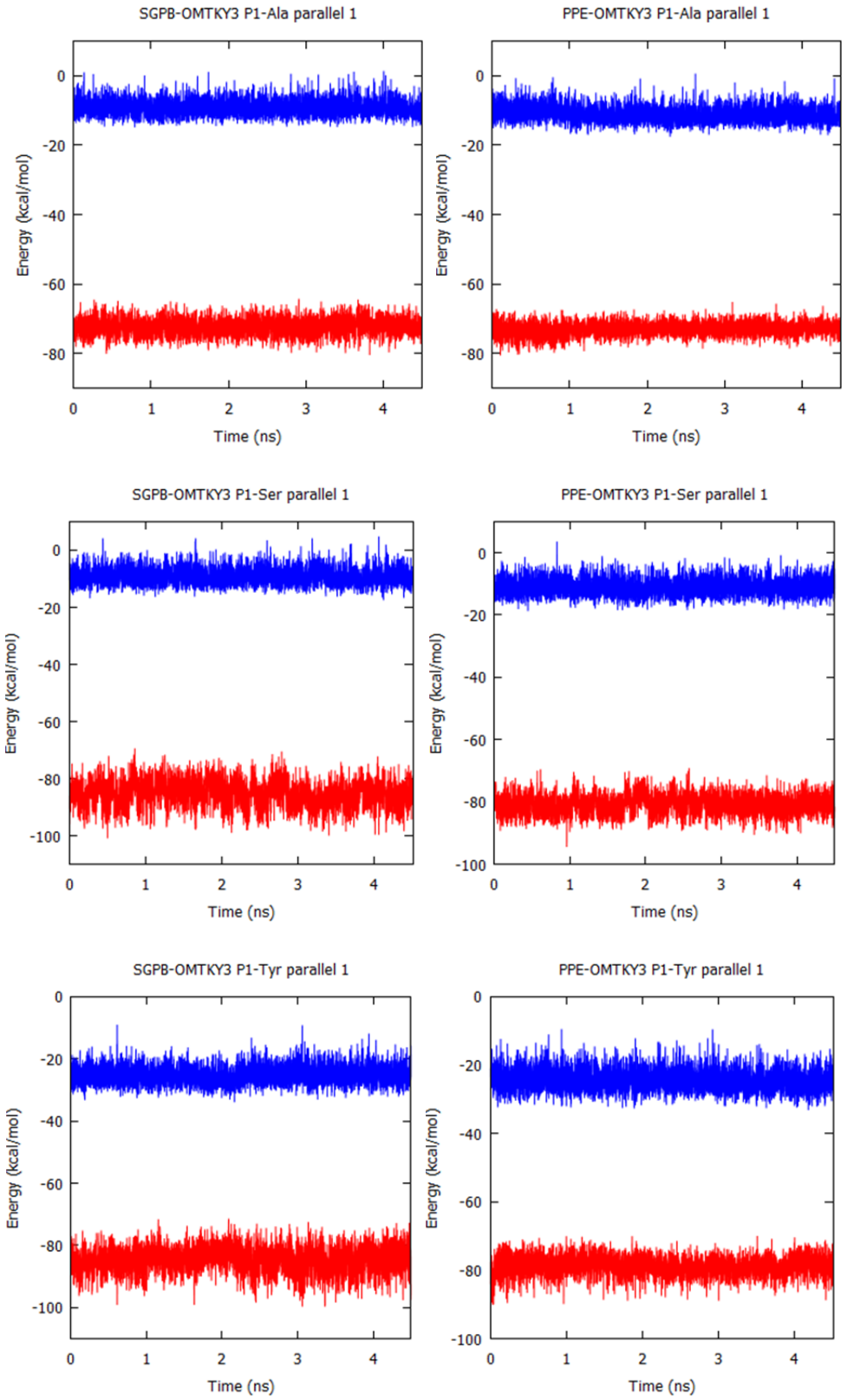

Figure 3.9: P1-surroundings interaction energy plots for a selection of SGPB-OMTKY3 complexes (left) and PPE-OMTKY3 complexes (right). Electrostatic energy (red) and vdW energy (blue), in kcal/mol, are plotted against time measured in nanoseconds.

3.5 Effect of multiple simulations

To evaluate the effect of multiple simulations the LIE binding energy was calculated for each parallel. Most of the simulations result in fairly similar LIE binding energies, but some parallels result in large deviations (Table 3.4 and Table 3.5). Structural analysis indicates this is most likely caused by conformational changes during the equilibration phase but sometimes conformational changes occur during the production phase. Specific details are discussed under the appropriate P1 heading.

Table 3.4: Comparison of calculated binding energies for each parallel in SGPB-OMTKY3.

P1	ΔG_{bind}^{calc} *MD*−1	ΔG_{bind}^{calc} *MD*−2	ΔG_{bind}^{calc} *MD*−3
Gly	-9.9	-10.2	-10.2
Ala	-11.2	-10.8	-10.9
Arn[n]	-5.6	-5.4	-5.1
Arg[c]	-7.4	-5.3	0.5
Asn	-11.7	-13.0	-12.7
Asp[c]	-4.9	-2.3	-4.6
AspH1[a]	-3.9	-3.6	-4.4
AspH2[a]	-7.1	-7.6	-6.4
Cys	-13.2	-13.3	-12.7
Gln	-13.1	-12.8	-12.1
Glu[c]	-5.2	-2.8	-1.7
GluH1[a]	-4.4	-4.4	-6.2
GluH2[a]	-5.2	-5.1	-3.8
Hid[b]	-11.9	-12.4	-11.3
Hie[b]	-10.5	-11.2	-10.9
Hip[c]	-3.9	-2.9	-2.6
Ile	-13.3	-12.5	-13.2
Leu	-13.7	-12.9	-13.2
Lyn[n]	-9.9	-9.2	-9.9
Lys[c]	-7.0	-6.6	-6.8
Met	-12.4	-11.4	-12.5
Phe	-13.6	-13.3	-12.7
Pro	-11.7	-12.3	-11.9
Ser	-10.9	-11.2	-11.4
Thr	-12.5	-12.3	-11.9
Trp	-14.1	-14.3	-14.0
Tyr	-13.6	-13.5	-13.5
Val	-13.7	-12.9	-13.4

Energies listed in kcal/mol

Energies calculated using the LIE equation: $\Delta G_{bind}^{LIE} = \alpha\Delta < V_{l-s}^{vdW} > + \beta\Delta < V_{l-s}^{el} > + \gamma$

MD-1, MD-2 and MD-3 indicate from which parallel the binding free energy was obtained.

[a] Refers to either variant of protonated, neutral, aspartic or glutamic acid.

[b] Hid and Hie refers to protonation state of neutral histidine.

[c] Refers to charged variant of P1 residue (Asp, Glu, Arg, Lys or His).

[n] Refers to protonated, neutral, variant of arginine or lysine.

Table 3.5: Comparison of calculated binding energies for each parallel in PPE-OMTKY3.

P1	ΔG_{bind}^{calc} MD−1	ΔG_{bind}^{calc} MD−2	ΔG_{bind}^{calc} MD−3
Gly	-11.4	-11.6	-11.7
Ala	-13.3	-12.8	-13.1
Arn[n]	-1.4	-3.1	-2.4
Arg[c]	11.1	10.2	10.1
Asn	-12.2	-13.4	-14.2
Asp[c]	0.4	0.6	2.3
AspH1[a]	-8.1	-6.6	-8.3
AspH2[a]	-5.4	-5.9	-9.5*
Cys	-14.5	-14.0	-14.6
Gln	-12.8*	-10.0	-10.0
Glu[c]	9.8	8.2	12.4
GluH1[a]	-3.9	-3.1	-2.4
GluH2[a]	-3.5	-3.3	-2.1
Hid[b]	-11.3	-10.6	-11.0
Hie[b]	-8.6	-8.3	-9.3
Hip[c]	14.6	14.8	16.1
Ile	-15.5	-13.8	-14.9
Leu	-15.1	-14.9	-14.9
Lyn[n]	-9.4	-6.3	-7.7
Lys[c]	12.9	12.3	14.6
Met	-13.2	-11.2	-13.3
Phe	-11.9	-11.3	-11.0
Pro	-13.1	-13.3	-12.9
Ser	-11.1	-10.6	-10.7
Thr	-13.3	-12.9	-12.4
Trp	-13.1	-12.8	-15.0
Tyr	-12.4	-11.9	-12.7
Val	-15.3	-14.8	-14.7

Energies listed in kcal/mol

Energies calculated using the LIE equation: $\Delta G_{bind}^{LIE} = \alpha\Delta < V_{l-s}^{vdW} > + \beta\Delta < V_{l-s}^{el} > + \gamma$

MD-1, MD-2 and MD-3 indicate from which parallel the binding free energy was obtained.

[a] Refers to either variant of protonated, neutral, aspartic or glutamic acid.

[b] Hid and Hie refers to protonation state of neutral histidine.

[c] Refers to charged variant of P1 residue (Asp, Glu, Arg, Lys or His).

[n] Refers to protonated, neutral, variant of arginine or lysine.

* Indicates anomalous parallel, energy contribution excluded from average.

3.6 SGPB-OMTKY3

The optimized LIE model is very accurate (error less than 1.0 kcal/mol) for 8 of the 20 P1 mutants when choosing the configuration with the least error, and fairly accurate (error less than 2.0 kcal/mol) for another 8 mutants. Since P1-Pro is omitted, due to the restrictions imposed by the side chain, this leaves 3 P1 mutants as significant outliers. The majority of the simulations show only small differences in binding energy between the parallels. Notable exceptions are arginine, aspartic acid and glutamic acid, which show high degree of variation between some of parallels and are discussed in greater detail under the appropriate heading.

Crystal structures are available for 18 of the 20 P1 amino acids. Even though the simulations are based on models made from the P1-Gly variant, the crystal structures provide useful insight into actual binding actions between the enzyme and inhibitor. Early in the project it was decided to create models rather than adopt individual crystal structures for each P1 variation. Partly because crystal structures were not available for all P1 variants of SGPB-OMTKY3, but also to evaluate the LIE model and determine if the model would yield accurate results without relying on x-ray diffraction. This also provides a reference point for studying the PPE-OMTKY3 system, for which there are no crystal structures.

P1-Asp and P1-Glu

The side chains aspartic and glutamic acid are similar in many respects, and the binding free energies for both are under estimated by the chosen LIE model. Both side chains end with a carboxylic acid unit, and the possible protonation states are identical for either amino acid. The side chain can be charged (deprotonated) or one of the oxygen atoms can be protonated.

Binding free energies obtained from the simulations show the P1-AspH2 variant best reproduces the experimental binding energies. The charged P1-Asp is predicted to be unfavourable and is so for all parallels, and the P1-AspH1 mutant is significantly underestimated as well. Apart from one parallel involving the charged P1-Asp, which is unfavourable anyway, the simulations yield similar energies.

Analysis of the trajectories reveals the presence of a hydrogen bond between the P1 side chain and Ser214. Examination of the crystal structure of P1-Asp (1SGD) shows that the P1 side chain adopts a different conformation when compared to the

dominating one found in the simulations. Rather than H-bond with Ser214 the P1 side chain is oriented outwards and faces crystallographic water molecules. The resolution, 1.8 Å, is not good enough to identify hydrogen atoms and the actual protonation state for the P1 side chain is impossible to ascertain. Nevertheless H-bonds likely exist between the P1 side chain and the water molecules, but this does not offer new insight into the protonation state.

Water molecules are similarly positioned during the simulation and are available to form of H-bonds, but the P1 side chain is oriented inward rather than to the water molecules. The formation of an H-bond with Ser214 is likely favourable for the calculated binding energy but this provokes changes in the secondary interactions whose energy contributions are not factored into the LIE equation.

P1-Glu is underestimated as well and examination of the crystal structure (1SGE) show that the P1 side chain is oriented towards the crystallographic water. The experimental binding energy is best reproduced by the P1-GluH1 protonation state and structural analysis show interaction with water molecules during a part of the simulation. The third parallel of the P1-GluH1 mutant show the formation of a hydrogen bond with Ser214 in the latter part of the simulation, and this is clearly visible in Figure 3.10.

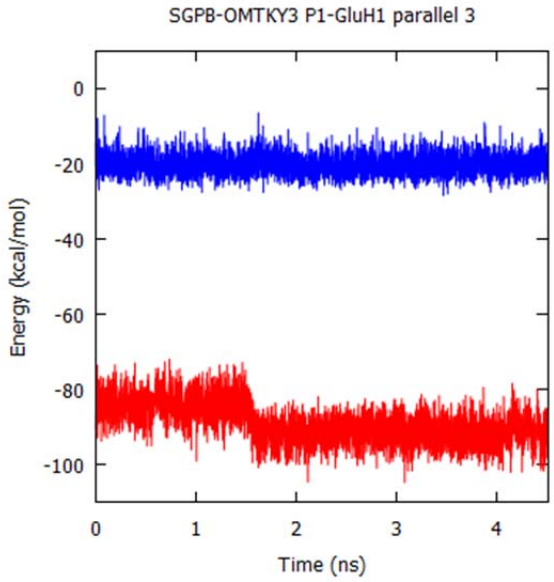

Figure 3.10: Interaction energy plot for the third parallel of the P1-GluH1 mutant from the SGPB-OMTKY3 complex. vdW energy is blue and electrostatic energy is red.

As with P1-Asp it is not clear if the H-bond in question should form. It does however provide a strong bond and it affects the LIE equation favourably. Nevertheless it also induces changes in the secondary interactions that are not included in the LIE equation and as such is suspect.

It is possible the experimental binding energies could be reproduced if greater care were taken during the construction of the models in respect to the orientation of the P1 side chain. However for such a technique to work available crystal structures are required but defeats the purpose of creating a method that can reliably predict binding energies based on models alone.

Two sets of crystal structures are available for the Asp and Glu variants, one at pH = 6.5 (1SGD, 1SGE) and the other at pH = 10.7 (2SGD, 2SGE) for Asp and Glu respectively. The high pH structures show the presence of a potassium ion close to the carboxylic unit. As no ion is present in the simulation this could possibly explain the difference in calculated and observed binding energies. However, MM-PBSA simulations on the SGPB-OMTKY3 complex conducted by Fujinaga et al. achieved best results without the K^+ ion [52].

To summarize, the P1-AspH2 reproduces the experimental binding energy to some degree, but this may be due to a cancellation of errors. The P1-Glu simulations fail to reproduce the experimental binding energies and obtain errors ranging from 3.6-5.3 kcal/mol.

P1-Arg

Accommodation of P1-Arg is underestimated by 7.0 and 5.8 kcal/mol for the charged and uncharged mutants respectively. The starting orientation of the P1 side chain is different from the crystal structure (2NU2), but trajectories from the simulations show that the P1 side chain adopt a similar conformation for both the charged and uncharged side chain. A number of protonation states are possible for neutral Arg, as deprotonation can occur at either $-NH_2$ group. The true protonation state is unknown, and although crystal structures are available the resolution is not good enough to identify hydrogen atoms. Previous LIE studies have concluded that chargeable P1 residues (His, Arg, Lys, Asp, Glu) bind predominantly in uncharged state [39]. For an accurate estimation of the binding energy the different protonation states will likely have to be considered.

Another issue is the low agreement among the charged P1-Arg simulations, in particular the third parallel. Structural investigation shows that the P1 side chain adopts a different conformation during the equilibration. The interaction energies show stable sampling, but unfavourable energies. In fact the high variation in calculated binding energies for the charged P1-Arg questions the validity of the energetics and prompts further simulations to reach a consensus.

P1-Ile

The LIE model calculates the binding free energy for P1-Ile to be -13.0 kcal/mol, and thus overestimates the energy by 3.0 kcal/mol. The parallel simulations yield consistent binding energies, and P1-surroundings interaction energies are stable. Examinations of the trajectories and comparison with the crystal structure (1CSO) show the P1 side chain adopting similar configurations during the simulations for parallel 1 and parallel 3. The main chain H-bond between the P1 residue and the enzyme is unchanged in all simulations with a distance similar to the crystal structure. No specific structural details can be pointed to as the cause of the discrepancy, but the over estimation is likely caused by too strong vdW interactions. This is however difficult to verify without further studies.

3.7 PPE-OMTKY3

Elastase has a much smaller binding pocket than SGPB and chymotrypsin. The binding site in elastase is closed by Val216 and Thr226, whereas these residues are glycine in chymotrypsin. This prevents larger side chains from entering the binding site, which is reflected in the experimental binding energies (Table 3.3). 12 of 20 possible amino acids at the P1 position are calculated within 2.0 kcal/mol of their experimental binding energies. Of the 8 outliers, 4 have previously proven to be problematic (Pro, Phe, Trp, and Tyr). In several cases H-bonds with Ser214 are shown to be the problem and this is further discussed under the appropriate P1 headings.

Aromatic residues and P1-Trp

The binding affinities of the aromatic P1-variants are vastly over estimated, LIE calculations estimate the free energy of binding of P1 Phe, Trp and Tyr to -11.5, -13.9, -12.5, respectively, with corresponding experimental values of -6.3, -5.9 and -5.2 kcal/mol. The binding affinities of the aromatic residues are thus overestimated by up to 8.0 kcal/mol. Phe, Trp and Tyr are experimentally among the worst binders. Similar effects are observed in the human leukocyte elastase (HLE) [51]. In the case of HLE the cause is suspected to be changes in the secondary interactions from trying to fit a large residue into the small S1 site of elastase. In the LIE method the P1 residue is treated as a ligand and secondary interactions are assumed to be the same. When a large P1 residue is docked into a narrow pocket the steric clashes will interact with the S1 site and the pocket will adapt and expand. This will likely change the secondary interactions and as noted earlier the LIE method requires the secondary interactions to be identical. Whenever the secondary interactions change the LIE method will fail to predict the effect. This problem is explored further in the preorganization energy section.

The third P1-Trp parallel is ~2 kcal/mol overestimated compared to the other two. Investigation of the trajectory of the P1-Trp complex simulations reveals that main chain-main chain hydrogen bonds at the P1, P2 and P3 sites are broken. The structural change is illustrated in Figure 3.11 which shows the secondary interactions between the protein and inhibitor main chains from the P1-Trp and P1-Ala mutants. Figure 3.12 show the changes in the surface of PPE from the P1-Trp and P1-Ala simulations respectively. In the other two parallels the P1 side chain is positioned inside the binding pocket.

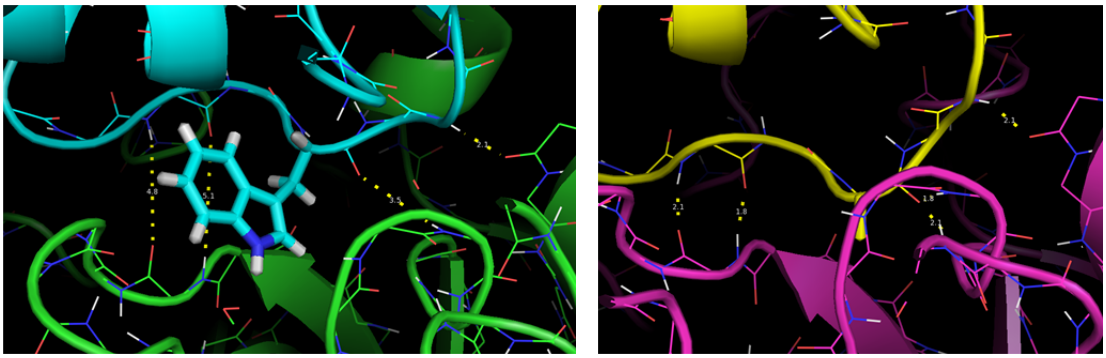

Figure 3.11: (Left)MD average trajectory of P1-Trp showing broken hydrogen bonds at secondary binding sites, only P2-S2 interaction is intact. (Right) MD average trajectory of P1-Ala, showing intact hydrogen bonds at secondary binding sites.

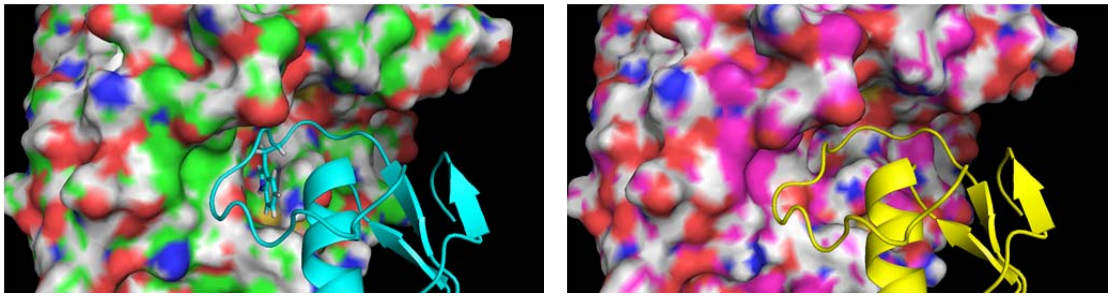

Figure 3.12: (Left) Surface of PPE bound to P1-Trp. The tryptophan side chain is seen carving out space for itself. (Right) Surface of PPE bound to P1-Ala shown for comparison.

P1-Arg

The LIE equation underestimates both variants of arginine, although only the neutral arginine is predicted to actually bind. Experimentally Arg is the worst P1 residue in respect to binding energy and this is also replicated in the simulations. The 2006 study by Almlöf et al. into HLE-OMTKY3 overestimates Arg by 3.7 kcal/mol and suggests this is caused by different protonation states [39]. The parallels show similar energies, although differing by somewhat they consistently underestimate the binding energy. Since crystal structures are not available it is difficult to say which side chain orientation is correct or whether it is reproduced during the simulations.

P1-Asn

The accommodation of P1-Asn is overestimated by 3.0 kcal/mol. Analysis of the MD trajectories show that the P1 side chain forms a hydrogen bond with the carbonyl oxygen of Ser214. This change can be seen in Figure 3.13 and Figure 3.14 which show the starting structure compared to a snapshot from the MD simulation. Examination of the non-bonded interaction energies from the trajectories confirms there is a strong electrostatic interaction between P1-Asn and Ser214.

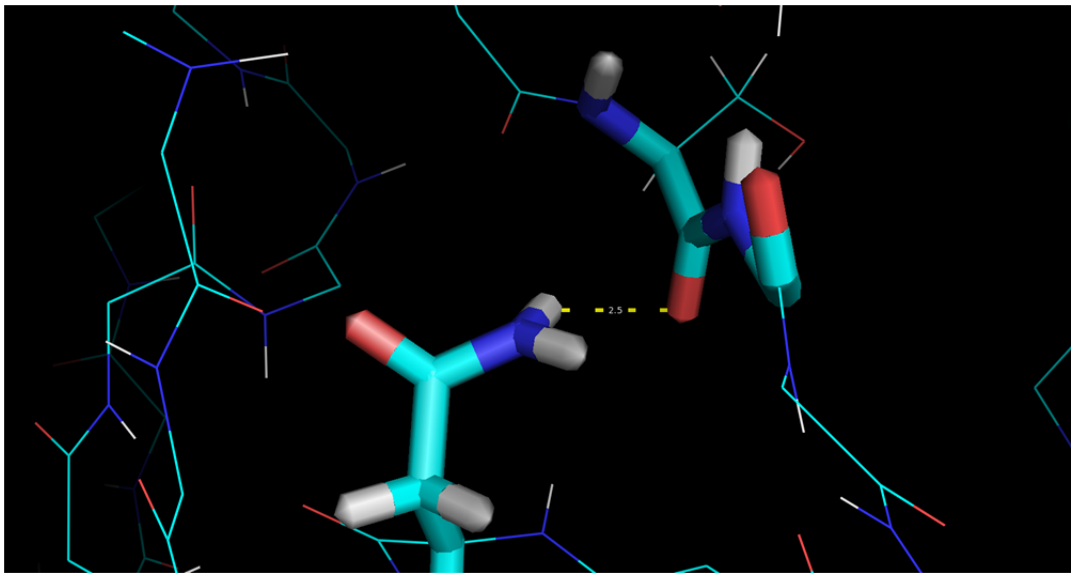

Figure 3.13: Starting structure of the PPE-OMTKY3 P1-Asn complex, showing the distance (2.5 Å) between the Asn and the carbonyl oxygen in the binding pocket.

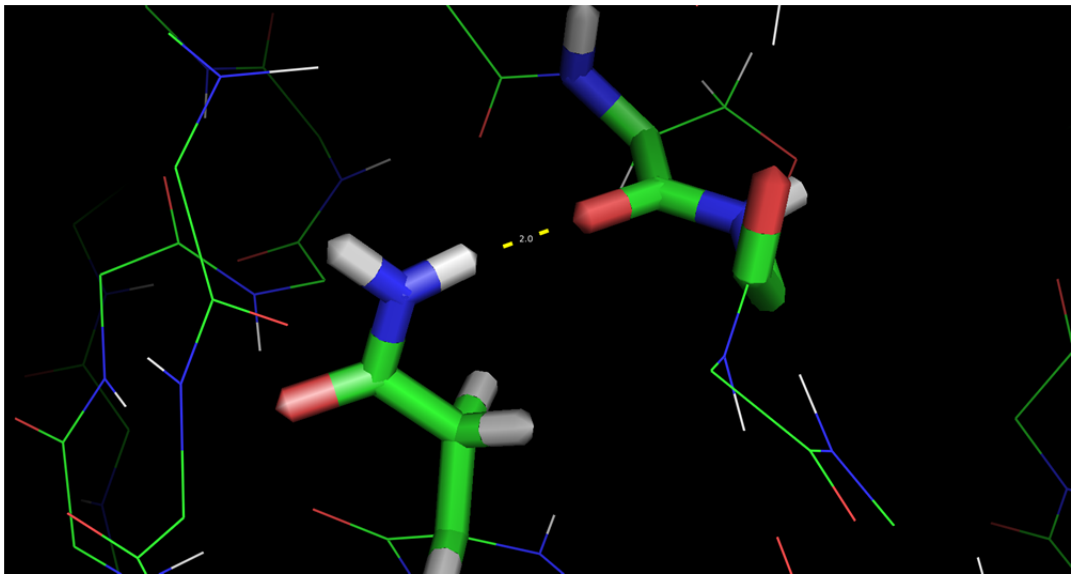

Figure 3.14: Snapshot of the PPE-OMTKY3 P1-Asn simulation, showing the distance (2.0 Å) between the Asn and the carbonyl oxygen in the binding pocket. The distance and geometry is typical for hydrogen bonds.

P1-Gln

The binding energy of P1-Gln is overestimated by 1.6 kcal/mol, this is attributed to one of the P1-Gln parallels which have a significantly different binding energy compared to the other two, -13.0 kcal/mol against -10.2 kcal/mol for the two others. Analysis of the trajectories shows an increase in favourable non-bonded interactions between the P1-residue and several residues in the enzyme, as well as a decrease in unfavourable non-bonded interactions. The differences can for the most part be ascribed to electrostatic interaction with His57 (favourable), Asp194 (unfavourable) and Ser214 (favourable). The large difference in binding energy, ~3 kcal/mol, questions the validity of the anomalous parallel.

Plotting the interaction energies during the simulation reveal that the first parallel is trapped in a local minimum during much of the simulation (Figure 3.15). As it does not sample the energies properly contributions from the first parallel is removed from the energy average. The second and third parallel reproduce the experimental binding energy with very good accuracy.

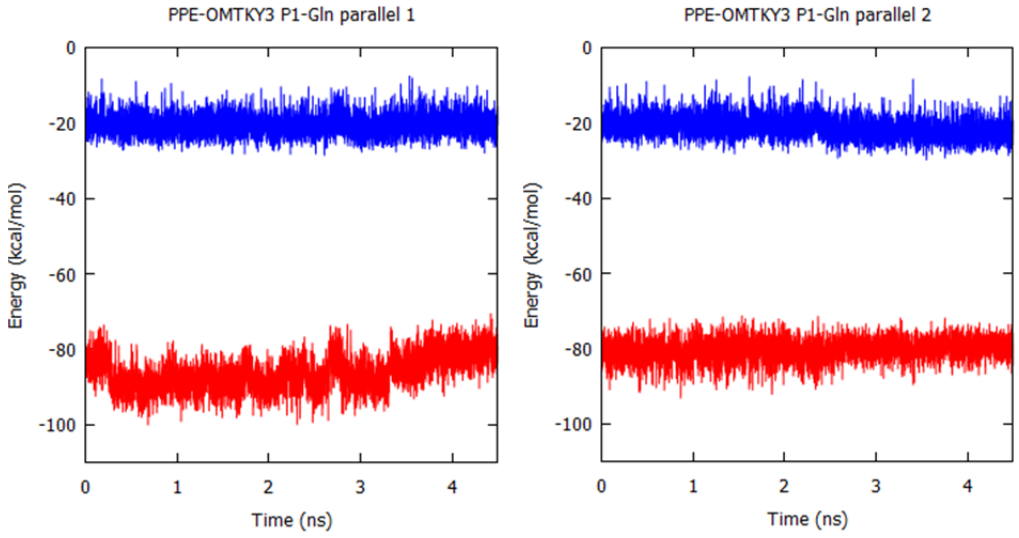

Figure 3.15: Interaction energy plot for first (left) and second parallel (right) of the P1-Gln variant of PPE-OMTKY3. The first parallel is trapped in a local minimum for a large part of the simulation. Electrostatic energy (red) and vdW energy (blue).

P1-Asp

The P1-Asp models are made up of the three different protonation states, each subjected to three parallel MD-simulations. Examining the LIE binding energies the AspH2 mutant reproduces the experimental association energy closest, while a charged side chain is energetically unfavourable. Which protonated configuration is correct, is difficult to determine as there is no crystal structure available of the PPE-OMTKY3 complex. Although an equilibrium between the protonation states can be expected.

The LIE method works best for AspH2 and examination of the MD stability reveal somewhat stable energies for the first and second parallel. Some conformational rearrangement is mirrored in the energies, but the energies are fairly stable as far as the average is concerned. The last parallel adopts a different conformation, in which the electrostatic interaction energy measured between the P1 residue and the surroundings is greatly increased (Figure 3.16). This is reflected in the LIE energy which is overestimated by ~4 kcal/mol. For the most part this is attributed to a hydrogen bond between the –OH in the P1 side chain and Ser214. In addition changes in the enzyme's hydrogen bond network are observed; Ser195 H-bonds to His57 but this bond assume different conformations and is also broken for a part of the simulation.

Overall the validity of the third parallel is in question. It undergoes conformational change and forms a hydrogen bond not found in the other two parallels. It is unknown whether the H-bond in question should form, because there is no available crystal structure to compare with. But if the LIE method is accurate in predicting the binding energy, the energy data for the first and second parallel indicate it should not form. If the energy contribution from the suspect third parallel is excluded the LIE model is in excellent agreement with the experimental binding energy.

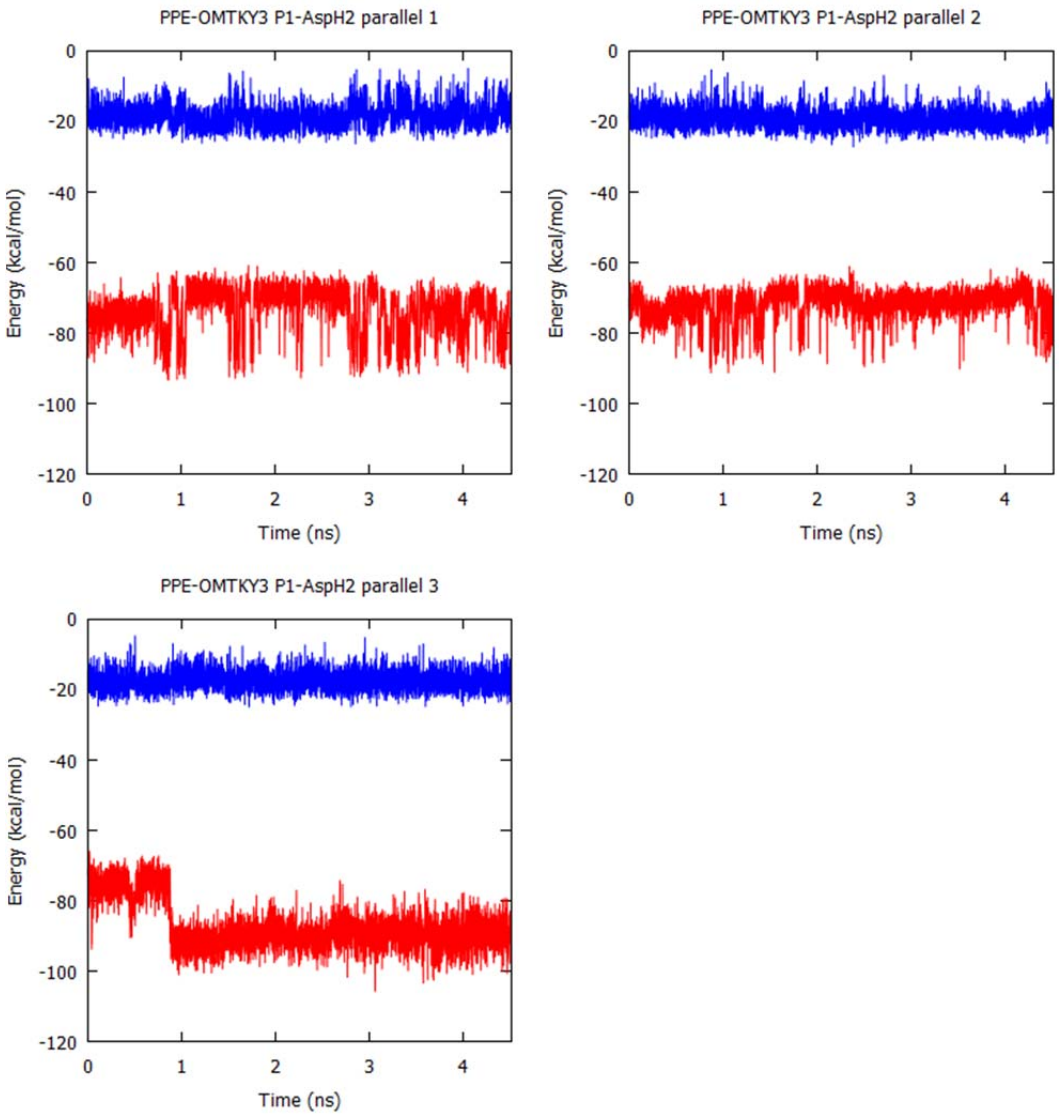

Figure 3.16: Ligand-surroundings interaction energies for P1-AspH2 variant of PPE-OMTKY3 for all three parallels. Electrostatic energy (red) and vdW energy (blue).

P1-Glu

Glutamic acid is either charged or protonated at either carboxylic oxygen. Experimentally P1-Glu binds weakly, and this is replicated in the simulations. The charged variant is predicted not to bind at all, while the two uncharged are predicted to bind poorly, but the binding energies are underestimated by roughly 3 kcal/mol for both protonation states. Some differences in binding energies are observed among the parallels but nothing that compromises the results of the simulations.

The P1-GluH1 variant matches the experimental binding energy best with an error of 3.2 kcal/mol, just slightly lower than P1-GluH2 with an error of 3.4 kcal/mol. There are possible hydrogen bonding partners in the binding pocket, but the MD trajectories show no evidence of H-bond formation. The energy difference is the same

as typically seen in H-bonds and the error can be a result of hydrogen bonding networks that are not properly reproduced in the simulations.

P1-His

The calculated binding free energy overestimates P1-His by 5.0 kcal/mol and 2.8 kcal/mol for the Hid and Hie variants respectively. The charged histidine variant is calculated to be extremely unfavourable with $\Delta G >> 0$ and is as such unlikely to occur. An equilibrium among the protonation states is possible, but with a calculated $\Delta G = 15.0$ kcal/mol the charged histidine would not be expected to bind at all. Experimentally P1-His is found to bind poorly but the simulations predict the binding to be stronger, with the Hie variant closest with an error of 2.8 kcal/mol. Investigation of the Hid and Hie structures suggest two possible explanations for the overestimation.

As with the aromatic P1 variants histidine incorporates a large side chain which will expand the binding site and weaken the secondary interactions. These energy contributions are not included in the LIE equation and the calculated binding energy will be overly favourable. Surface plots of the average structure from the Hie variant compared with P1-Ala show an enlarged binding pocket (Figure 3.17).

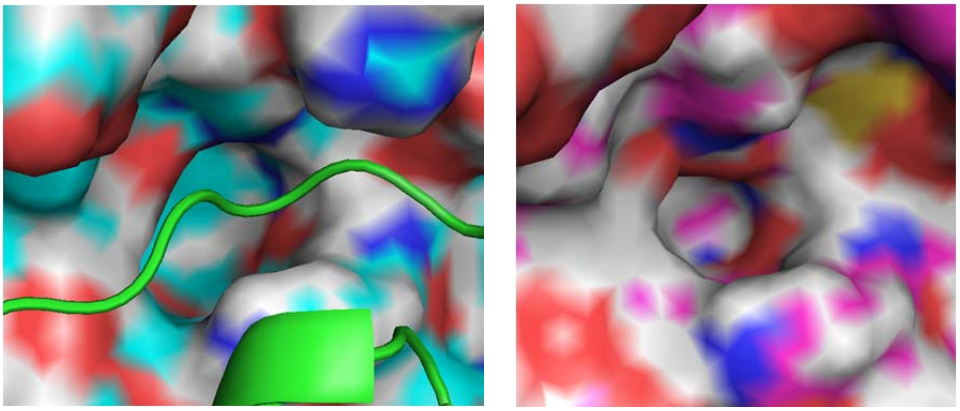

Figure 3.17: (Left) Surface plot of P1-Hie (MD-average structure, parallel 1), showing enlarged binding pocket in relation to the binding loop. (Right) Surface plot of P1-Ala, shown for comparison.

Investigation of the non-bonded interaction energies measured between the P1 residue and the surroundings show strong vdW and electrostatic interactions with residue 190-194 and 213-216. If these interactions are stronger than they are supposed to in the simulations, it would explain why the binding energy is overestimation. It

can also be a combination of strengthened binding pocket interactions combined with neglecting the energy contribution associated with enlarging the binding site that causes the overestimation.

In the study by Almlöf et al. the experimental binding energy of P1-His from HLE-OMTKY3 is reproduced with high accuracy [39]. This suggests it is possible withing the current LIE model to calculate the binding energy accurately for histidine bound to elastases. The current simulations however fail to yield an accurate result, and further simulation possibly using different orientations of the P1 side chain could reproduce the experimental binding energy better.

3.8 Preorganization energy

Several studies have examined the validity of the linear response approximation [27]. Linear response gives: $\Delta G_{el}^{i} = \frac{1}{2}\left\{\left\langle V_{l-s}^{el}\right\rangle_{on} + \left\langle V_{l-s}^{el}\right\rangle_{off}\right\}$

As noted earlier the $\left\langle V_{l-s}^{el}\right\rangle$ term denotes MD or MC averages, while "on" and "off" refer to whether electrostatic interactions between the ligand and surroundings are turned on or off. The $\left\langle V_{l-s}^{el}\right\rangle_{off}$ term is often referred to as the preorganization term or preorganization energy, and is approximated to zero in the LIE equation. When the electrostatic interactions are turned off the solvent molecules relax and orient themselves randomly in relation to the solute, thus cancelling out energy contributions from the $\left\langle V_{l-s}^{el}\right\rangle_{off}$ term [29]. While this is found to be true in water, the surroundings are vastly different in protein-protein complexes. When a canonical inhibitor such as OMTKY3 is bound to serine proteinases the secondary interactions keep the complex together and the close contacts restrict the relaxation in both enzyme and inhibitor. Under these conditions the $\left\langle V_{l-s}^{el}\right\rangle_{off}$ term is most likely not zero. Examinations of chymotrypsin in complex with OMTKY3 have found non-zero preorganization terms [39]. To calculate the $\left\langle V_{l-s}^{el}\right\rangle_{off}$ term another simulation is required and is further complicated by slow convergence of the preorganization term.

Another finding by Almlöf and co-workers [39] is that the preorganization term increases with the size of the ligand, in this case the P1 residue. And perhaps more importantly that it is possible to incorporate the preorganization term into the α variable for some systems, instead of calculating it explicitly.

4 Concluding remarks

Evaluation of molecular dynamics simulations and LIE calculations

Molecular dynamics simulations and free energy calculations have been used to study the effect of mutations of the P1 residue for complexes between SGPB and PPE with OMTKY3. It is found that the method is generally able to predict the effect quite reliable. However, some residues appear to be more problematic than others. In the SGPB with OMTKY3 complex glutamic acid at the P1 position is underestimated by 3.6 kcal/mol and P1 asparagine is overestimated by 3.0 kcal/mol in the PPE with OMTKY3 complex. Analysis of the structural data reveals that subtle differences in primarily hydrogen bonding interactions are the source of the discrepancy. Indeed, in some cases conformational sampling becomes trapped in local minima, yielding overestimation of the free energy of binding. Three independent simulations were carried out for each possible amino acid at the P1 position, and in some cases the three parallels are seriously different (up to 8 kcal/mol). This indicates that the sampling of the phase space is not sufficient and to obtain converged binding free energies more simulations should ideally been carried out. The statistical mechanical nature of the methodology used here raises the question of conducting more independent simulations rather than few long simulations to assess the convergence and accuracy. It is more likely that 100 independent 1 ns simulations cover a larger fraction of phase space compared to two 50 ns simulations. This is an exercise left for future work.

Ionisable residues

The effect of ionisable amino acids at the interface is generally found to be problematic to model. However, it appears that most if not all are neutral in the bound state, consistent with the fact that burial of charges is energetically unfavourable. The major challenge with modelling ionisable P1 variants is due to the nature of force fields where one has to define only one state, neutral or charged. In reality, an equilibrium exists between these two states which is difficult to capture by using a single-state model. Modelling of ionic P1 variants is further complicated by the potential presence of counter-ions in at the protein-protein interface. It is difficult to know the position of such ions unless they contain a significantly different number of

electrons than water, so that they can be detected in the electron density of crystal structures. Nonetheless, we find that the ionisable P1 variants are preferred to be in their neutral state when bound to PPE and SGPB.

Improving the linear interaction energy method

Modelling the effect of large aromatic P1 variants at the protein-protein interface has been very difficult, particularly for PPE. In this case, the free energies of binding are overestimated by 5-8 kcal/mol. This is at first somewhat surprising, but can be understood when examining the equilibration phase. As mentioned, the S1 site of PPE is too small to accommodate these side chains, but model construction phase they are placed inside or at the entrance of the S1 site. Then, during the equilibration phase the S1 site adjusts and becomes sufficiently large to accommodate these side chains. It is thus not a deficiency of the LIE method, but rather the fact that we force something to stay in the binding pocket which is not supposed to be there. An alternative technique to study mutations, which not been pursued in this work, is to use free energy perturbation with dual topologies and soft-core potentials.

Preorganization energy contributions have been listed as a source of error in the LIE method. Although it is not investigated in this project, it can to some extent be corrected by the parameterization. To improve the parameterization two sets of LIE variables can be used; one in bound state and another set when the ligand is free in solution. This can better compensate for the difference in preorganization of charges from bound to unbound states than a single variable.

Ideas for the future

This thesis has focused exclusively on the LIE method using explicit solvent models. However continuum based solvent models can also be used to give a new insight into problematic models or simply extend the knowledge of current models.

Since no crystal structure is available for PPE with OMTKY3 this is another possible task. A crystal structure can verify and possible identify problems related to the MD simulations, in particular related orientation of large amino acids in the S1 pocket.

Summary

This project demonstrates the ability of the LIE method to predict the binding free energy of different P1 variants using models based on a small number of crystal structures only. In about 85 % of the cases the method was able to differentiate between strong and weak binders and 70 % of the values were within 2.0 kcal/mol of the experimental binding free energy. The method is as such applicable in tasks such as aiding in drug design and screening of binding candidates before experiments are carried out.

5 References

1. Barrett, A.J., N.D. Rawlings, and J.F. Woessner, *Handbook of proteolytic enzymes,* 1998, San Diego: Academic Press. xxix.
2. Page, M.J. and E. Di Cera, *Serine peptidases: classification, structure and function.* Cell Mol Life Sci, 2008. **65**(7-8): p. 1220.
3. Huang, K., W. Lu, S. Anderson, M. Laskowski, Jr., and M.N. James, *Water molecules participate in proteinase-inhibitor interactions: crystal structures of Leu18, Ala18, and Gly18 variants of turkey ovomucoid inhibitor third domain complexed with Streptomyces griseus proteinase B.* Protein Sci, 1995. **4**(10): p. 1985.
4. Lu, W., I. Apostol, M.A. Qasim, N. Warne, R. Wynn, W.L. Zhang, S. Anderson, Y.W. Chiang, E. Ogin, I. Rothberg, K. Ryan, and M. Laskowski, Jr., *Binding of amino acid side-chains to S1 cavities of serine proteinases.* J Mol Biol, 1997. **266**(2): p. 441.
5. Schechter, I. and A. Berger, *On the size of the active site in proteases. I. Papain.* Biochem Biophys Res Commun, 1967. **27**(2): p. 157.
6. The PyMOL Molecular Graphics System, Version 1.4.1, Schrödinger, LLC.
7. Brändén, C.-I. and J. Tooze, *Introduction to protein structure*. 2nd ed, 1999, New York: Garland Pub. xiv.
8. Isaksen, G.V., *Flexible membrane active antimicrobial tripeptides with stability towards chymotryptic degradation*, *Department of Chemistry,* 2010, University of Tromsø: Tromsø.
9. Laskowski, M., Jr. and I. Kato, *Protein inhibitors of proteinases.* Annu Rev Biochem, 1980. **49**: p. 593.
10. Travis, J. and G.S. Salvesen, *Human plasma proteinase inhibitors.* Annu Rev Biochem, 1983. **52**: p. 655.
11. Bode, W. and R. Huber, *Natural protein proteinase inhibitors and their interaction with proteinases.* Eur J Biochem, 1992. **204**(2): p. 433.
12. Laskowski, M. and M.A. Qasim, *What can the structures of enzyme-inhibitor complexes tell us about the structures of enzyme substrate complexes?* Biochim Biophys Acta, 2000. **1477**(1-2): p. 324.
13. Ardelt, W. and M. Laskowski, Jr., *Effect of single amino acid replacements on the thermodynamics of the reactive site peptide bond hydrolysis in ovomucoid third domain.* J Mol Biol, 1991. **220**(4): p. 1041.
14. Krowarsch, D., T. Cierpicki, F. Jelen, and J. Otlewski, *Canonical protein inhibitors of serine proteases.* Cell Mol Life Sci, 2003. **60**(11): p. 2427.
15. Zhang, Q., M. Sanner, and A.J. Olson, *Shape complementarity of protein-protein complexes at multiple resolutions.* Proteins, 2009. **75**(2): p. 453.
16. London, F., *On the Theory and Systematic of Molecular Forces.* Z Phys, 1930. **63**(3-4): p. 245.
17. Leach, A.R., *Molecular modelling : principles and applications*. 2nd ed, 2001, Harlow, England ; New York: Prentice Hall. xxiv.
18. E. Arunan, G.R.D., R. A. Klein, J. Sadlej, S. Scheiner, I. Alkorta, D. C. Clary, R. H. Crabtree, J. J. Dannenberg, P. Hobza, H. G. Kjaergaard, A. C. Legon, B. Mennucci, D. J. Nesbitt., *Definition of the hydrogen bond.* Pure Appl. Chem., 2011. **83**(8): p. 1637.
19. E. Arunan, G.R.D., R. A. Klein, J. Sadlej, S. Scheiner, I. Alkorta, D. C. Clary, R. H. Crabtree, J. J. Dannenberg, P. Hobza, H. G. Kjaergaard, A. C. Legon, B.

Mennucci, D. J. Nesbitt., *Defining the hydrogen bond: An account.* Pure Appl. Chem., 2011. **83**(8): p. 1619.
20. Fersht, A.R., *The Hydrogen-Bond in Molecular Recognition.* Trends Biochem Sci, 1987. **12**(8): p. 301.
21. Fersht, A.R., *Basis of Biological Specificity.* Trends Biochem Sci, 1984. **9**(4): p. 145.
22. Jones, S. and J.M. Thornton, *Principles of protein-protein interactions.* Proc Natl Acad Sci U S A, 1996. **93**(1): p. 13.
23. Cornell, W.D., P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, and P.A. Kollman, *A second generation force field for the simulation of proteins, nucleic acids, and organic molecules (vol 117, pg 5179, 1995).* J Am Chem Soc, 1996. **118**(9): p. 2309.
24. Jorgensen, W.L., D.S. Maxwell, and J. TiradoRives, *Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids.* J Am Chem Soc, 1996. **118**(45): p. 11225.
25. Ryckaert, J.P., G. Ciccotti, and H.J.C. Berendsen, *Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes.* J Comput Phys, 1977. **23**(3): p. 327.
26. Zwanzig, R.W., *High-Temperature Equation of State by a Perturbation Method .1. Nonpolar Gases.* J Chem Phys, 1954. **22**(8): p. 1420.
27. Brandsdal, B.O., F. Osterberg, M. Almlof, I. Feierberg, V.B. Luzhkov, and J. Aqvist, *Free energy calculations and ligand binding.* Adv Protein Chem, 2003. **66**: p. 123.
28. Aqvist, J., C. Medina, and J.E. Samuelsson, *A new method for predicting binding affinity in computer-aided drug design.* Protein Eng, 1994. **7**(3): p. 385.
29. Aqvist, J. and T. Hansson, *On the validity of electrostatic linear response in polar solvents.* J Phys Chem-Us, 1996. **100**(22): p. 9512.
30. Srinivasan, J., T.E. Cheatham, P. Cieplak, P.A. Kollman, and D.A. Case, *Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate - DNA helices.* J Am Chem Soc, 1998. **120**(37): p. 9401.
31. Kollman, P.A., I. Massova, C. Reyes, B. Kuhn, S.H. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D.A. Case, and T.E. Cheatham, *Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models.* Accounts Chem Res, 2000. **33**(12): p. 889.
32. Jain, A.N., *Scoring noncovalent protein-ligand interactions: A continuous differentiable function tuned to compute binding affinities.* J Comput Aid Mol Des, 1996. **10**(5): p. 427.
33. Pearl, L.H., *Structure and function in the uracil-DNA glycosylase superfamily.* Mutat Res-DNA Repair, 2000. **460**(3-4): p. 165.
34. Honig, B. and A. Nicholls, *Classical Electrostatics in Biology and Chemistry.* Science, 1995. **268**(5214): p. 1144.
35. Sitkoff, D., K.A. Sharp, and B. Honig, *Accurate Calculation of Hydration Free-Energies Using Macroscopic Solvent Models.* J Phys Chem-Us, 1994. **98**(7): p. 1978.
36. Bogan, A.A. and K.S. Thorn, *Anatomy of hot spots in protein interfaces.* J Mol Biol, 1998. **280**(1): p. 1.

37. Krowarsch, D., M. Dadlez, O. Buczek, I. Krokoszynska, A.O. Smalas, and J. Otlewski, *Interscaffolding additivity: binding of P1 variants of bovine pancreatic trypsin inhibitor to four serine proteases.* J Mol Biol, 1999. **289**(1): p. 175.
38. Helland, R., J. Otlewski, O. Sundheim, M. Dadlez, and A.O. Smalas, *The crystal structures of the complexes between bovine beta-trypsin and ten P1 variants of BPTI.* J Mol Biol, 1999. **287**(5): p. 923.
39. Almlof, M., J. Aqvist, A.O. Smalas, and B.O. Brandsdal, *Probing the effect of point mutations at protein-protein interfaces with free energy calculations.* Biophys J, 2006. **90**(2): p. 433.
40. Aqvist, J., J. Marelius, K. Kolmodin, and I. Feierberg, *Q: A molecular dynamics program for free energy calculations and empirical valence bond simulations in biomolecular systems.* J Mol Graph Model, 1998. **16**(4-6): p. 213.
41. Wurtele, M., M. Hahn, K. Hilpert, and W. Hohne, *Atomic resolution structure of native porcine pancreatic elastase at 1.1 A.* Acta Crystallogr D Biol Crystallogr, 2000. **56**(Pt 4): p. 520.
42. Jones, T.A., J.Y. Zou, S.W. Cowan, and M. Kjeldgaard, *Improved methods for building protein models in electron density maps and the location of errors in these models.* Acta Crystallogr A, 1991. **47 (Pt 2)**: p. 110.
43. Maestro, version 9.1, Schrödinger, LLC, New York, NY, 2010.
44. MacroModel, version 9.8, Schrödinger, LLC, New York, NY, 2010.
45. Kaminski, G.A., R.A. Friesner, J. Tirado-Rives, and W.L. Jorgensen, *Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides.* J Phys Chem B, 2001. **105**(28): p. 6474.
46. Lee, F.S. and A. Warshel, *A Local Reaction Field Method for Fast Evaluation of Long-Range Electrostatic Interactions in Molecular Simulations.* J Chem Phys, 1992. **97**(5): p. 3100.
47. NOTUR. Stallo. http://www.notur.no/hardware/stallo/
48. Aqvist, J., *Calculation of absolute binding free energies for charged ligands and effects of long-range electrostatic interactions.* J Comput Chem, 1996. **17**(14): p. 1587.
49. Helland, R., G.I. Berglund, J. Otlewski, W. Apostoluk, O.A. Andersen, N.P. Willassen, and A.O. Smalas, *High-resolution structures of three new trypsin-squash-inhibitor complexes: a detailed comparison with other trypsins and their complexes.* Acta Crystallogr D, 1999. **55**: p. 139.
50. Hansson, T., J. Marelius, and J. Aqvist, *Ligand binding affinity prediction by linear interaction energy methods.* J Comput Aid Mol Des, 1998. **12**(1): p. 27.
51. Almlof, M., J. Aqvist, A.O. Smalas, and B.O. Brandsdal, *Probing the effect of point mutations at protein-protein interfaces with free energy calculations.* Biophys J, 2006. **90**(2): p. 433.
52. Fujinaga, M., K. Huang, K.S. Bateman, and M.N. James, *Computational analysis of the binding of P1 variants of domain 3 of turkey ovomucoid inhibitor to Streptomyces griseus protease B.* J Mol Biol, 1998. **284**(5): p. 1683.