Application of molecular mechanics polarization to fragment based drug design

David Adrian Woodlock

A thesis submitted for the degree of Doctor of Philosophy

Department of the School Biological Sciences

University of Essex

September 2015

Table of Contents

Acknowledgements

There are many people that I have to thank for making these past 4 years one to remember and who were instrumental in helping me finish this final document.

First, I'd like to thank Prof. Christopher Reynolds for bringing me into this field and helping me each step of the way. Without his guidance, I doubt I'd have reached the end of my years here. I enjoyed working in his lab and amongst his other Ph.D. students.

I'd like to express my gratitude to David Reha who introduced me to the wonders of scripting which proved invaluable for the years to come. He helped me with my 3rd-year project and cultivated my interest in the field I am now in.

I'd like to thank all the colleagues I've had in the lab over the years, those people who kept me company over the long nights pouring over computers; David Crompton, Hani Mohammed, Kevin Smith and Bruck Taddese. I doubt my years in the lab would've been the same without you all.

I'd like to thank my parents Clive and Jennifer Woodlock for supporting me throughout my degree. I'd like to thank my sister Emma Woodlock for the encouragement and I'd like to thank my significant other Alana Rea, who stuck with me despite the long nights of studying and writing that I no doubt kept her up through.

I'd lastly like to thank my trusty computer in the lab, bs1060. Every time I came back from an extended period it'd break down. Despite that, it lasted me through all the years of this thesis. I am sure as its final act of insubordination it'll combust after I leave.

Abstract

Polarization is a term that is often excluded from almost all virtual screening. Polarizability helps explain interactions between nonpolar atoms and electrically charged species. When studying fragments in FBDD these minor interactions could have large effect in changing how well a ligand will bind to its target. After including the polarizability terms in docking a validation set of ligands (Favia et al., 2011) with GLIDE, it improved the results the amount good docked poses (< 2 Å RMSD) by up to 12%. However some ligands were bound in incorrect poses. Further investigation was carried out with MD to observe if given enough time ligands bound in an incorrect pose would return to the binding site. In the first stages of investigating MD we ascertained if we could use GPUs to simulate larger systems and faster. After some performing some MD simulations in GROMACs we found that GPUs were an improved option and thus continued the simulation work with ACEMD which allowed multiple GPUs in tandem. After running the MD simulations for 200ns with atomic charges generated from the polarization the results we found were quite interesting. Some ligands would be trapped in their binding site but would fluctuate quite readily such as 2GVV. Some ligands showed that despite low RMSD they would be ejected from the binding site. In some cases the ligands would then attempt to return to their binding site. Ligands such as in 2CIX would show binding based on the breathing movement of the protein. Some ligands such as 1F5F or 1F8E bound tightly to their binding site during the MD, these ligands also enjoyed improved docking polarization with 0.1 – 1.0 Å improvement. These could be carried forward to become good candidates for experimental testing. Polarization is shown to have an overall positive effect improving binding data and if implemented with simple methods would have little opportunity cost to be added to modern FBDD methods.

Chapter 1

Introduction and applications of FBDD

1.1 Introduction

Over the past two decades with the advancement in technology, the use of simulations using computer programs simulations in the Bioinformatics field has been growing greatly as according to Perola et al. (2004) and Law et al. (2009). Computers have gained more processing speed and are able to accommodate larger and more complex calculations than ever before. One such field that has seen growth in computation is Fragment Based Drug Discovery (FBDD); this is a target based-approach to drug design and discovery. The central theme to FBDD is to optimise each unique interaction along the way as the small fragments bind to the active site of the protein, as discussed by Hajduk and Greer (2007). Starting with a single small fragment with a high affinity, each interaction is maximised then more fragments or groups are added. Eventually, multiple suitable fragments are combined into a single lead compound with drug-like properties, as described for example by Lipinski's rule of 5, or preferably the rule of 3 for lead compounds as shown in Congreve et al. (2008).

1.1.1 Aims

Here we aim to improve computational docking of ligands by including polarization. Currently full polarization is lacking in almost all virtual screening in Fragment Based Drug Design experiments be they docking or QM/MM studies as noted by Jorgensen (2007). The calculations in chapter 2 seek to account for these missing polarization terms to improve the docking results that can be obtained by fragments.

1.2 Fragment based drug design

1.2.1 Brief History

The origins of FBDD started with work on binding energies, e.g. as performed by Jencks (1981). Jenks used empirical approaches to derive a Gibbs free energy term for the intrinsic binding energy for a ligand binding to a protein. These early works did not yet use the powers of computers but were instead establishing equations that could be used. A few years later, Goodford (1984) implemented these similar calculations into a computational environment. Goodford (1985) implemented a scheme for determining the interaction energy of very small fragments (atomic and molecular probes that represented functional groups) on a 3D grid around a protein. Within the GRID program, the interaction energies were plotted as contour maps and facilitated the discovery of favourable binding sites for discrete functional groups (and hence ligands) around the protein.

Around the same time Abraham et al. (1984) were beginning to design, by molecular modelling, new compounds to bind active sites, which is the basis of FBDD. This was a follow on from Abraham et al. (1983) tests of testing potential antisickling agents using X-ray studies. In their tests on the Val-6 beta mutation site of sickle cell haemoglobin they observed in approved drug libraries that an aromatic ring containing two halogens attached to a benzyloxy or phenoxyacetic acid were required to generate the desired effect. In this case it was an antigelling agent. However they used this data to perform tests on different molecules based on this common structure found amongst the approved drugs. Similarly in FBDD as seen later there are fragments that provide the necessary binding and Abraham et al. (1984) used these fragments to formulate a complete drug by fusing the fragments together to target binding sites within proteins. In this paper Abraham et al. (1984) have thus have fabricated multiple lead targets based up the central benzene ring.

This process of fusing fragments together to formulate a complete drug is called linking. However the programs that could perform these links came a few years later. This started with Lauri and Bartlett (1994) with the CAVEAT program and Eisen et al. (1994) with the HOOK program. Both programs had the same goal, namely making molecular skeletons from a database of ligands that suit the binding site. Both these programs were meant as assisting tools to a chemist as they were visual tools that allowed the user to make the molecular scaffold, albeit in a somewhat automated manner. Upon generating a structure these programs would calculate how favourable the binding of the ligand is to the active site. These were the preliminary steps to what has become a modern technique. There are still good uses for designing a ligand by hand but modern methods can help predict better structures that can then be adjusted by hand for an improved fit.

Newer techniques began to surface in the 90s such as Mattos and Ringe (1996) use of X-ray crystal structure to address the interaction of functional groups with proteins through the use of organic solvents. The technique they used is one called multiple solvent crystal structure or MSCS. Much of the previous work relies on the biological functions of the ligand to the protein such as the Gibbs free energy of binding to solve the binding. MSCS began to implement modelling techniques from processes such as GRID to develop a map of the entire binding surface of the protein. They repeatedly solved the crystal structure of their target protein in several different organic solvent to mimic different functional groups. This allows them to map out all the binding sites in a protein as opposed to just the binding site of a known substrate. This method opened up new techniques to study the protein as a whole to design inhibitors and substrates for different sites that were mapped out in the protein. Combining this with previous techniques such as those found in CAVEAT or HOOK, it is possible to design drugs for different sites of a protein.

Later these methods were targeted by big pharma to locate lead targets that failed high throughput screening (HTS), the common technique at the time to discover drug lead targets. This led to the rise of many small pharma companies in the industry, such as Astex (2015) in 1999, Vertex (2015) in 1989 and Plexxikon (2015) in 2001 to commence focusing on the FBDD process.

1.2.2 Modern FBDD

FBDD in recent years has quickly become a suitable substitute to older methods of drug target screening due to favouring more structure-based approaches as reported by Law et al. (2009). It has been recognised that fragments can quickly and efficiently be built into suitable lead compounds using FBDD. Computers have played an integral role in FBDD, as over the past 25 years computers have constantly expanded the capabilities of this field, as summarised by Martin et al. (2012). As stated by Congreve et al. (2008) and Orita et al. (2009), FBDD has two major ideas that set it apart from older methods such as High throughput screening (HTS). The first is using computational analysis of the chemical space near the active site. It is far more efficient to use small sized fragment sized ligands rather than the traditional libraries of large molecules (which also need to contain far more molecules). The key idea behind this efficiency is that chemical space of fragment is far smaller at approximately 10^7 compounds for up to 12 non-hydrogen atoms. In contrast, the chemical space of a drug-like compound of up to 30 atoms is approximately 10^{60} compounds. Due to the difference in the sizes of the two sets of chemical space, screening in FBDD is eminently tractable as it is quite feasible to screen 10^6 – 10^7 compounds. However, screening 10^{60} compounds in a traditional screen of drug-like compounds is impossible, requiring the need to design specific targeted libraries and the inherent risk of missing hits.

The second idea is that due to the small nature of fragments when they bind to the protein they will bind with less affinity since they cannot make as many interactions as the larger drug-like molecules; however the binding efficiency for each individual atom is potentially the same in both methods, meaning that the fragments, despite their low binding energy, can still make a number of high-quality interactions. Fragment-based Quantum Mechanical (QM) calculations on fragments are easier to carry out as shown in Hesterkamp and Whittaker (2008), due to their small size. In addition, at these low affinities it becomes difficult to screen and therefore screening can become more expensive as it requires sophisticated methods. This is where evolving computational techniques such as FBDD can help cut down the number of compounds that need to be screened experimentally, thus lowering the cost of the overall research, as reported by Pors (2011). It is possible to use these small fragments as building blocks to develop a drug-like molecule by fusing them together with small linker fragment to maintain structural and conformational integrity.

This process is called Fragment based lead discovery, as reviewed by Erlanson (2006); the basic concept is also shown in Figure 1.1 below. The process starts with an initial fragment hit from fragment based screening which is then optimised by evolution or by linking to another fragment. Growing a molecule is shown in Figure 1.1. This starts with the initial hit, and then the molecule is expanded around that initial hit to fill up the space to form more interactions with the rest of the active site. In contrast, in linking, the initial hit is combined with another initial hit, by using a linker molecule specifically designed/chosen for the site in question. Although using linkers is on the increase, their use is inefficient. In some cases linking will generate needlessly large molecules that have lost some of their binding affinity to the site in the process. The two methods can be used individually or in conjunction with each other to optimise a final molecule that binds to the site.

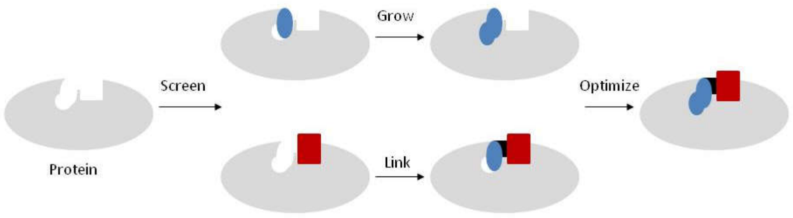


Figure 1.1: Basic concept to fragment-based lead discovery, adapted from Erlanson (2006).

Erlanson (2006) lists some examples such as Wood et al. (2005) who used fragment based methods to discover and optimise protease inhibitors. Despite the examples shown, not all attempts of using FBDD increased the efficiency of the final molecule when compared to the initial hit. However, a significant portion still did increase the ligand efficiency by use of growing. Whereas using linking there was never a case where the final molecule had greater ligand efficiency when compared to both the initial hits used to form the final linked molecule. FBDD is in its infancy but it is starting to show promise within the pharmaceutical industry. FBDD should hopefully reach its potential to increase the speed of lead targets found but also improve the number of ligands that may tightly bind to their proteins drug target.

1.2.3 Ligand Efficiency

Ligand Efficiency (LE) is the measure of the free energy of binding of a ligand to its target protein per number of non-hydrogen atoms. The binding free energy in the equations 1.1 and 1.2 is shown as G, often defined as G = -RTlnK. The Gibbs free energy change due to mutation of a structure can be used to characterize the stability of that structure. This Gibbs free energy change is defined as G . Equation 1.1 from Kuntz et al. (1999) used this term to look at the best ligands for macromolecular targets based on their stability.

$$G_{binding} = \quad G_{hydrogen\,bond} + \quad G_{hydrophobic} + G_{vdw} + G_{electrostatic} \quad (1.1)$$

Hopkins et al. (2004) improved upon the initial equation 1.1 and calculations put forward by Kuntz et al. (1999) to make equation 1.2 given below to define Ligand Efficiency. Non-hydrogen atoms are otherwise defined as heavy atoms.

$$LE = \quad g = (\ G)/N \qquad (1.2)$$

Where LE is ligand efficiency, g (Delta small g) is the free energy per atom and N is the number of non-hydrogen atoms (Abad-Zapatero 2007). LE is a useful tool to any researcher designing drugs through the use of lead-based design. Ligand efficiency is used at each step of the design, either growing or linking a ligand, to ensure that it remains with the target range, typically < 0.3. As shown in Congreve et al. (2008), this helps prevent the researcher being blindsided and only focusing on the potency (defined here as the binding affinity) of the drug and not neglect its physiochemical properties of binding.

Abad-Zapatero (2007) reports that other terms similar to LE have been in development during recent years, which are based on the LE equation. The terms were designed due to the variation of molecular weight in a compound which can also change its potential size and surface area for interaction. Thus, the following indices were developed using molecular weight as measured in kDa: Percent Efficiency Index (PEI), Binding Efficiency Index (BEI) and Surface Efficiency Index (SEI). The PEI is an indication of inhibition brought about by the ligand, presented as a fraction. The BEI is a measure of the potency in relation to the molecular weight of the ligand. The SEI is a measure of potency gained that is related to the polar surface area (PSA) of the ligand. These three indicators can be used as a numerical reference for considerations when formulating a drug. These indicators show representations for important variables, potency, molecular weight and surface area. When used in tandem with older rules, such as Lipinski's rule of five and other drug indicators, a researcher can estimate the confines he has to work with when developing the drug from the initial hit.

The above indicators have been useful in a myriad of techniques used in FBDD to quantify how potentially useful compounds are for the drug design purposes.

1.2.4 Fragment Libraries

In FBDD, in order to begin study on a protein or other biomacromolecule a fragment library is required. This allows a base set of ligands from which to work. According to Congreve et al. (2008) there are a good number of considerations that constitute a good or useful fragment library. The range of physiochemical properties for the fragments included, quality control, assessment of molecular diversity, chemical tractability of the fragments, which chemical functionalities are allowed, druglikeness of the fragments with precedence set by studies in oral drugs and natural products and sampling sets of privileged medical scaffolds.

Also to take into consideration is the complexity of a ligand (Hann et al., 2001), and if there is an optimal level of complexity that a library should consider. As shown in Fattori (2004) if there is too much complexity in a library then overall the there will be fewer hits from which to create a lead. This of course reinforces the idea that fragments need to be small and when designing a drug, it should be built it up either from several fragments or by growing it out as shown earlier in section 1.2. However, at the same time a ligand can’t be too small, as a ligand must be large enough to act as a molecular anchor. This is to say that it has enough interactions (electrostatic, H-bonds etc) so that it binds with enough affinity, thus taking into account the ligand efficiency which was discussed in section 2.1. With regards to an optimal size Congreve et al. (2008) believes that fragments should be between 100-250 Da, which lowers the chance of steric clashes while maintaining a high ligand efficiency.

1.2.5 Deconstructing leads into Fragments

One other method in FBDD to find a fragment is to reverse engineer one from lead compounds. Reverse engineering can derive a fragment from the larger lead like structure where it's possible to take parts of the lead target apart and by use of these pieces to discover structural information for the binding site, as shown in Hajduk (2006). In 2006, Hajduk (2006) deconstructed several inhibitors and found that the fragments all had a similar ligand efficiency. However, in experimental tests this might not always be true despite Hajduk's similar ligand efficiency. Babaoglu and Shoichet (2006) deconstructed a β-lactamase inhibitor into four fragments. Only one of their fragments bound in the same way as the inhibitor would. According to Congreve et al. (2008) this has caused much debate in the FBDD community in how much weight binding mode would have for optimisation and subsequent growth of a fragment.

One example of deconstructing being a benefit is the work by Liu et al. (2001). In their study they found a series of *p*-arylthio cinnamides that could act as antagonists to a reaction between leukocyte function-associated antigen-1 (LFA-1) and intracellular adhesion molecule-1 (ICAM-1). They used one of the ligands that bound to an allosteric site, a known diaryl sulphide. They deconstructed it, splitting it at its sulphide bond. They found that this fragment could still bind well and used NMR screening with the fragment present to find an alternative fragment that could be linked on to the structure they had to improve binding. They found two different compounds which they appended to the main fragment. One of the new drug targets only improved the binding affinity of the drug marginally form 0.044 µM to 0.040 µM. The other structure improved the binding to 0.020 µM. More importantly the presence of the sulphide portion of the molecule caused the drug target to have no oral bioavailability as shown in Liu et al. (2001) but the new molecule that had stronger binding affinity had an oral half-life of 4.7hrs. This means they could make a drug that could be ingested out of this new molecule. So by using known ligands as a scaffold it is possible to change parts of the ligand by deconstructing them to try to give the target more favourable properties.

1.2.6 Interactions of ligands with proteins and the importance of structure

According to Murray et al. (2012), small molecules i.e. fragments, need to pass a large entropic barrier in order to bind. This barrier is actually larger proportionally for the fragment than it is for a complete drug as shown in Murray and Verdonk (2002). This means that high quality bonds are required to bind fragments. This can be seen by inspecting fragments visually when in a binding pocket. Many of them have strong hydrogen bonds or powerful electrostatic interactions. This arises from the complementary structure of the fragment to its binding site. Not having key components in a ligand for a particular binding site will cause a lot lower binding affinity. For example a carboxylate group has the potential to form strong hydrogen bonds with a protonated amino acid such as arginine. If the binding site has an arginine nestled in its structure then a carboxylate group on a fragment can serve as an anchor. Subsequently this anchor could be used to grow out the proceeding drug.

Congreve et al. (2008) performed a review of several binding sites and fragments attached to them to better show how a complementary fragment forming multiple hydrogen bonds to the binding site is important for a ligand. In figure 1.2 we can see one of the ligands from Favia et al. (2011) used in section 2.9.

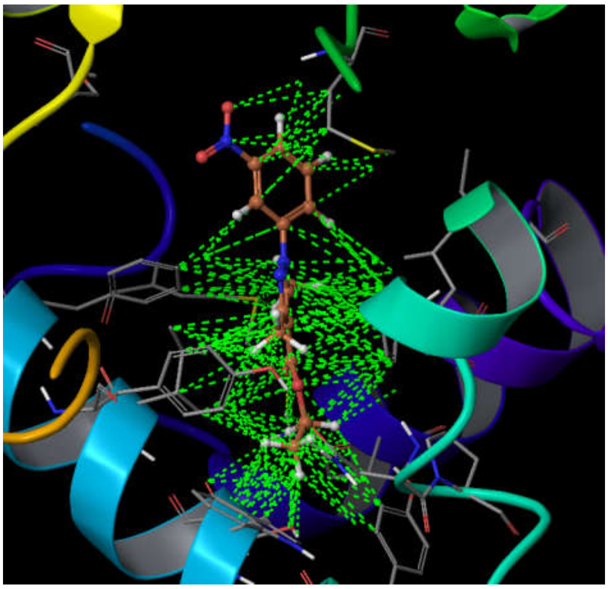


Figure 1.2: Ligand-receptor interactions within a binding pocket. In the centre of the figure is a fragment. The coil structures are the alpha helices of the receptor protein. The green dotted lines represent the Van der Waal and electrostatic interactions between the ligand and receptor.

The ligand in figure 1.2 has a strong anchor point. This can be seen by the many interactions with the binding pocket shown by the green dotted lines. These can be electrostatic, hydrophobic or hydrogen bonds. These interactions are important to the overall structure of the ligand as the more interactions present the tighter the ligand will bind. In the case of the ligand in figure 1.2 the bottom half of the ligand acts as the anchor which will fluctuate infrequently, whereas the top half has relatively few interactions so it will fluctuate rapidly. This means the top half is a candidate to grow the ligand using FBDD.

1.2.7 Thermodynamic equilibrium

As explained by Kuriyan et al. (2012), the equilibrium of a ligand binding to a protein is expressed as shown in equation 1.1.

$$[P] + [L] \quad [PL] \tag{1.3}$$

Where [P] is concentration of the protein, [L] is the concentration of the ligand and [PL] is the concentration of the ligand-protein complex. Equation 1.1 is where at the point where the concentrations of reaction do not change, there is an equilibrium between the rate of formation of the complex (products) and the formation protein and ligand (reactants).

The equilibrium of this reaction can be expressed as a constant as shown in equation 1.4.

$$K = \frac{[PL]}{[P]\,[L]} \tag{1.4}$$

Since this is a reaction of the ligand binding to the protein, this is a binding constant. The free energy of binding for a reaction is expressed in section 1.2.3. In binding studies it is more common place to measure the dissocation of a ligand, which is simply the inverse of the association constant as shown in equation 1.5.

$$K_i = \frac{[P]\,[L]}{[PL]} \tag{1.5}$$

Where K_i is the equilibrium dissociation constant of inhibition. As shown in Neubig et al. (2003), the K_i is usually determined by using a radioligand binding study to measure the inhibition of binding against the binding of a reference ligand in equilibrium conditions. Another binding term often used is IC_{50} which is an assay term that denotes an inhibitor concentration needed to reduce a protein or enzyme activity to 50% of its capacity. This however can be affected by concentrations of substrate, inhibitor and protein.

1.2.8 Other experimental and computational approaches

There are other approaches to FBDD than just HTS. One such method experimentally is High-Content Screening (HCS) as reported by Abraham et al. (2004).

HCS. Unlike HTS, HCS uses cell-based system in the screening process instead of just the target proteins. The screening technology is based around automated digital microscopy and flow cytometry, in combination with computer systems to quickly analyse and store the data. HCS exposes cells to a potential drug and changes in the cell morphology and protein production are measured using the aforementioned microscopy and flow cytometry to determine the impact of that drug.

FEP. Another method that has been in use for many years is the Free Energy Perturbation (FEP) method. This method can be summarized as measuring the free energy difference between the initial state and final state of a system to their average energy difference. This was developed by Zwanzig (1954) and is often expressed as shown in equation 1.6.

$$F(A \rightarrow B) = F_B \quad F_A = \quad k_B T ln \ \exp \left(\ \frac{E_B - E_A}{k_B T} \right) \ _A \qquad (1.6)$$

Where F is the free energy, k***b*** is the Boltzmann constant, T is the temperature of the system and E is the energy of the initial or reference state. This free energy difference can be used to measure the binding strength of a ligand as seen in Jorgensen and Thomas (2008). However, this measure free energy differences between say two ligands, and in order for the average to converge to the correct free energy, it is important that the ligands are not too different and that they bind in a similar mode. These restrictions limit the use of this method in practical drug design.

MMPBSA. Another method used that involves MD simulation is the Molecular Mechanics-Poisson-Boltzmann Surface Area (MMPBSA) method as developed by Kollman et al. (2000). According to Zoete et al. (2010) this method uses the free energy of binding (G_{bind}) written as a sum of its gas phase contribution (H_{bind}^{gas}), desolvation free energy (G_{desolv}) of the system upon binding and its entropic contribution ($T \ S$), this is shown in equation 1.7.

$$G_{bind} = \quad H_{bind}^{gas} + \quad G_{desolv} \quad T \ S \qquad (1.7)$$

To find each term the method runs a single trajectory but evaluating the energy for each portion separately. As seen in Zoete et al. (2010) evaluting the energy of the complex, solely the protein and solely the ligand. G_{desolv} is the difference between the solvation free energy of the complex and its isolated parts, so this is calculated from the energy difference from the trajectory. H_{bind}^{gas} contains difference in energy between the interactions of the complex and the isolated parts. This includes the van der Waal, electrostatic and intramolecular forces. Lastly, the $T \quad S$ term is caclculated using standard equations of statical mechanics. The accuracy of this method is like most MD methods dependent on the accuracy of sampling. The MMPBSA also largely depends upon the Surface Area terms used as seen in Miller III et al. (2012).

1.3 Conclusion

FBDD, while still a budding field has progressed in leaps and bounds, in both experimental aspects and also the associated computational aspects. Many of the calculations that previously might have taken weeks now only take hours with modern computers. This will allow the field to branch out into new techniques, by allowing larger systems and full proteins to be modelled and studied. Virtual screening has come a long way in 30 years but is continuing to show that it can be used as an alternative and as an adjustment to high throughput screening. Many of the techniques used by modern FBDD can be performed efficiently and with finesse in ligand construction, providing new drug candidates each year. However with any growing field there are portions of it that still need expanding. In this thesis we will investigate the pressing quest of lack of polarization in FBDD virtual screening. We will examine how it can be used in both docking and in molecular dynamics to attain more accurate binding and to examine how it can give us insight into the behaviour of ligands.

1.4 References

ABAD-ZAPATERO, C. 2007. Ligand efficiency indices for effective drug discovery.

ABRAHAM, D., KENNEDY, P., MEHANNA, A., PATWA, D. & WILLIAMS, F. 1984. Design, synthesis, and testing of potential antisickling agents. 4. Structure-activity relationships of benzyloxy and phenoxy acids. *Journal of medicinal chemistry,* 27**,** 967-978.

ABRAHAM, D. J., PERUTZ, M. F. & PHILLIPS, S. 1983. Physiological and x-ray studies of potential antisickling agents. *Proceedings of the National Academy of Sciences,* 80**,** 324-328.

ABRAHAM, V. C., TAYLOR, D. L. & HASKINS, J. R. 2004. High content screening applied to large-scale cell biology. *Trends in biotechnology,* 22**,** 15-22.

ASTEX. 2015. *Astex Pharmaceuticals* [Online]. Available: http://www.astx.com/ [Accessed 22 Sep 2015 2015].

BABAOGLU, K. & SHOICHET, B. K. 2006. Deconstructing fragment-based inhibitor discovery. *Nature chemical biology,* 2**,** 720-723.

CONGREVE, M., CHESSARI, G., TISI, D. & WOODHEAD, A. J. 2008. Recent developments in fragment-based drug discovery. *Journal of medicinal chemistry,* 51**,** 3661-3680.

EISEN, M. B., WILEY, D. C., KARPLUS, M. & HUBBARD, R. E. 1994. HOOK: a program for finding novel molecular architectures that satisfy the chemical and steric requirements of a macromolecule binding site. *Proteins: Structure, Function, and Bioinformatics,* 19**,** 199-221.

ERLANSON, D. A. 2006. Fragment-based lead discovery: a chemical update. *Current opinion in biotechnology,* 17**,** 643-652.

FATTORI, D. 2004. Molecular recognition: the fragment approach in lead generation. *Drug discovery today,* 9**,** 229-238.

FAVIA, A. D., BOTTEGONI, G., NOBELI, I., BISIGNANO, P. & CAVALLI, A. 2011. SERAPhiC: A benchmark for in silico fragment-based drug design. *Journal of chemical information and modeling,* 51**,** 2882-2896.

GOODFORD, P. J. 1984. Drug design by the method of receptor fit. *Journal of medicinal chemistry,* 27**,** 557-564.

GOODFORD, P. J. 1985. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of medicinal chemistry,* 28**,** 849-857.

HAJDUK, P. J. 2006. Fragment-based drug design: how big is too big? *Journal of medicinal chemistry,* 49**,** 6972-6976.

HAJDUK, P. J. & GREER, J. 2007. A decade of fragment-based drug design: strategic advances and lessons learned. *Nature reviews Drug discovery,* 6**,** 211-219.

HANN, M. M., LEACH, A. R. & HARPER, G. 2001. Molecular complexity and its impact on the probability of finding leads for drug discovery. *Journal of chemical information and computer sciences,* 41**,** 856-864.

HESTERKAMP, T. & WHITTAKER, M. 2008. Fragment-based activity space: smaller is better. *Current opinion in chemical biology,* 12**,** 260-268.

HOPKINS, A. L., GROOM, C. R. & ALEX, A. 2004. Ligand efficiency: a useful metric for lead selection. *Drug discovery today,* 9**,** 430-431.

JENCKS, W. P. 1981. On the attribution and additivity of binding energies. *Proceedings of the National Academy of Sciences,* 78**,** 4046-4050.

JORGENSEN, W. L. 2007. Special issue on polarization. *Journal of Chemical Theory and Computation,* 3**,** 1877-1877.

JORGENSEN, W. L. & THOMAS, L. L. 2008. Perspective on free-energy perturbation calculations for chemical equilibria. *Journal of chemical theory and computation,* 4**,** 869-876.

KOLLMAN, P. A., MASSOVA, I., REYES, C., KUHN, B., HUO, S., CHONG, L., LEE, M., LEE, T., DUAN, Y. & WANG, W. 2000. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Accounts of chemical research,* 33**,** 889-897.

KUNTZ, I., CHEN, K., SHARP, K. & KOLLMAN, P. 1999. The maximal affinity of ligands. *Proceedings of the National Academy of Sciences,* 96**,** 9997-10002.

KURIYAN, J., KONFORTI, B. & WEMMER, D. 2012. *The molecules of life: Physical and chemical principles*, Garland Science.

LAURI, G. & BARTLETT, P. A. 1994. CAVEAT: a program to facilitate the design of organic molecules. *Journal of computer-aided molecular design,* 8**,** 51-66.

LAW, R., BARKER, O., BARKER, J. J., HESTERKAMP, T., GODEMANN, R., ANDERSEN, O., FRYATT, T., COURTNEY, S., HALLETT, D. & WHITTAKER, M. 2009. The multiple roles of computational chemistry in fragment-based drug design. *Journal of computer-aided molecular design,* 23**,** 459-473.

LIU, G., HUTH, J. R., OLEJNICZAK, E. T., MENDOZA, R., DEVRIES, P., LEITZA, S., REILLY, E. B., OKASINSKI, G. F., FESIK, S. W. & VON GELDERN, T. W. 2001. Novel p-arylthio cinnamides as antagonists of leukocyte function-associated antigen-1/intracellular adhesion molecule-1 interaction. 2. Mechanism of inhibition and structure-based improvement of pharmaceutical properties. *Journal of medicinal chemistry,* 44**,** 1202-1210.

MARTIN, E., ERTL, P., HUNT, P., DUCA, J. & LEWIS, R. 2012. Gazing into the crystal ball; the future of computer-aided drug design. *Journal of computer-aided molecular design,* 26**,** 77-79.

MATTOS, C. & RINGE, D. 1996. Locating and characterizing binding sites on proteins. *Nature biotechnology,* 14**,** 595-599.

MILLER III, B. R., MCGEE JR, T. D., SWAILS, J. M., HOMEYER, N., GOHLKE, H. & ROITBERG, A. E. 2012. MMPBSA. py: an efficient program for end-state free energy calculations. *Journal of Chemical Theory and Computation,* 8**,** 3314-3321.

MURRAY, C. W. & VERDONK, M. L. 2002. The consequences of translational and rotational entropy lost by small molecules on binding to proteins. *Journal of computer-aided molecular design,* 16**,** 741-753.

MURRAY, C. W., VERDONK, M. L. & REES, D. C. 2012. Experiences in fragment-based drug discovery. *Trends in pharmacological sciences,* 33**,** 224-232.

NEUBIG, R. R., SPEDDING, M., KENAKIN, T. & CHRISTOPOULOS, A. 2003. International Union of Pharmacology Committee on Receptor Nomenclature and Drug Classification. XXXVIII. Update on terms and symbols in quantitative pharmacology. *Pharmacological Reviews,* 55**,** 597-606.

ORITA, M., WARIZAYA, M., AMANO, Y., OHNO, K. & NIIMI, T. 2009. Advances in fragment-based drug discovery platforms. *Expert opinion on drug discovery,* 4**,** 1125-1144.

PEROLA, E., WALTERS, W. P. & CHARIFSON, P. S. 2004. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins: Structure, Function, and Bioinformatics,* 56**,** 235-249.

PLEXXIKON. 2015. *Plexxikon Pharmaceuticals* [Online]. [Accessed 22 Sep 2015 2015].

PORS, K. 2011. *Drug Discovery Into the 21st Century*, INTECH Open Access Publisher.

VERTEX. 2015. *Vertex Pharmaceuticals* [Online]. Available: http://www.vrtx.com/ [Accessed 22 Sep 2015 2015].

WOOD, W. J., PATTERSON, A. W., TSURUOKA, H., JAIN, R. K. & ELLMAN, J. A. 2005. Substrate activity screening: a fragment-based method for the rapid identification of nonpeptidic protease inhibitors. *Journal of the American Chemical Society,* 127**,** 15521-15527.

ZOETE, V., IRVING, M. & MICHIELIN, O. 2010. MM–GBSA binding free energy decomposition and T cell receptor engineering. *Journal of Molecular Recognition,* 23**,** 142-152.

ZWANZIG, R. W. 1954. High-temperature equation of state by a perturbation method. I. nonpolar gases. *The Journal of Chemical Physics,* 22**,** 1420-1426.

Chapter 2

Exploring the use of polarization in docking

2.1 Introduction

Docking attempts to determine the ideal pose or orientation of a molecule, usually a small molecule ligand, as it is bound to another molecule, which is usually a protein. In this part of the project we look at how ligands, or fragments, bind to large biomacromolecules that are proteins. Docking is useful for predicting binding affinities of ligands to the protein. Docking is also useful for beginning the process of fragment-based drug discovery. Here, the rationale is that to begin either linking or growing a fragment hit into a lead compound, it is helpful to know the docked pose of the ligand that is being used as a base for the lead optimization. With regards to hits once you know a ligand docks it might help decide which kind of ligands to screen next based on the binding pocket and the chemical properties of the ligands.

There are a number of widely used approaches to docking, as implemented in a number of well-used programs, e.g. AUTODOCK (Trott and Olson, 2010), GLIDE (Eldridge et al., 1997), GOLD (Jones et al., 1997), FLEXX (Schellhammer and Rarey, 2004), FRED (McGann, 2011), SURFLEX (Jain, 2003) and QXP (McMartin and Bohacek, 1997). All of these programs perform reasonably well, though the performance of the program tends to be system-specific as seen in Warren et al. (2006) . Our hypothesis is that polarization will improve docking, and this limits the choice of program that can be used to test this hypothesis. For example, GOLD does not use electrostatics and so cannot be used here. AUTODOCK does use electrostatics, but the implementation is complex and it is difficult to modify the electrostatics as seen in Illingworth et al. (2008). Consequently, the docking in this chapter utilizes the program GLIDE (Eldridge et al., 1997), (Friesner et al., 2004), (Halgren et al., 2004), (Friesner et al., 2006), partly because it is relatively straight forward to modify the atomic charges and partly because GLIDE performs well on a number of different systems as seen in Repasky et al. (2007) and its implementation within the Maestro GUI facilitates comparison against other methods e.g. QM/MM. GLIDE or Grid-based ligand docking with energetics is a relatively new docking methodology found in software such as FirstDiscovery (Schrödinger, 2015) or Maestro, as reported by Halgren et al. (2004); grid-based docking was also used by ligandfit (Venkatachalam et al., 2003). GLIDE itself has the advantage of both being able to maintain a sufficient processing speed scanning large libraries while performing an exhaustive search of all the conformational positions for the ligand in the space provided within its receptor docking grid.

2.2 Docking using GLIDE

The main method that GLIDE uses to perform docking is through the use of a set of hierarchical filters. It performs four major steps, a site-point search, a series of refinements, grid minimization + Monte Carlo methods then moves to the final scoring function as shown in Figure 2.1, taken from Halgren et al. (2004).

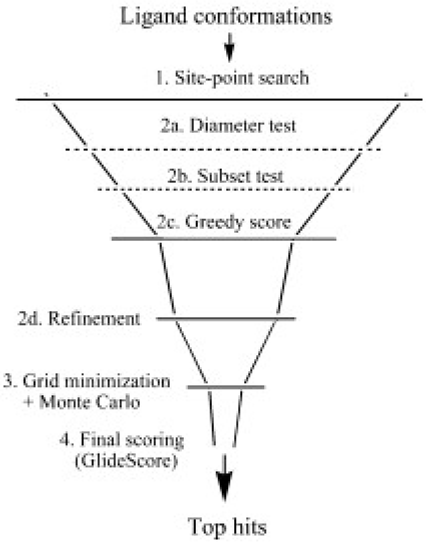


Figure 2.1. Glide docking hierarchy

A receptor based grid is first generated that maps out the region of the receptor where the ligand is likely to bind. A grid involves the pre-calculations of part of the Coulombic and Lennard-Jones, energies on the grid, due to the target receptor, similar to the methods shown in Goodford (1985). A grid helps to speed up calculations for finding correct conformations because of the pre-computed values when it is set up. In the majority of cases the target receptor is a protein. Each point within the grid that is either too close to the receptor or touching it is flagged as a point for the conformations to avoid. Afterwards a set of initial ligand conformations are generated, as reported by Halgren et al. (2004). This part focuses on the ligand torsion-angle space and takes a set of the best scored poses based on the energetic minima for these conformations. Then the poses that have been selected are energetically minimised within the receptor grid according to an energy function set by the user in tandem with using a distance-dependent dielectric model. In most cases, of this project this energy function is one of the OPLS force fields, generally the OPLS 2015 force field, as seen in Robertson et al. (2015). A scoring function is an approximation used in many methods to predict the strength of the binding affinity for a ligand to the binding site as reported by Jain (2006) The lowest energy poses after this step are subject to a Monte Carlo procedure to examine the torsional minima. In the cases where GLIDE is calculating a redocking based upon an original crystal structure as opposed to using a compound library at this point, the program will calculate the Root mean square deviation (RMSD) value based upon the original position of the ligand in the crystal structure; the smaller this number the less the ligand has been displaced from its original positions. The RMSD has helped determine the accuracy of the program compared to similar docking programs, as GLIDE consistently attains RMSD results of 2 Å or lower as shown in Kontoylanni *et al.* (2004). The RMSD script in GLIDE can give erroneously high values by equivalent atoms or shapes swapping places for example in symmetric molecules. However we use an in-house script that can be found in the appendix 2.1 that does away with this issue.

2.2.1 Scoring Function

As stated previously, scoring functions are used by many Ligand docking programs as an approximation of how to rank ligand poses as they are generated. Generally a search algorithm will produce far too many results to be useful. Approximately 10^9 poses could be generated according to Halperin et al. (2002). Thus scoring functions are utilised to discriminate to find the plausibly right answer, which is usually a pose with a lower RMSD, though that might not always be the case. Dependent on the algorithm and scoring function used, the results for a search can vary widely as scoring algorithms that can produce false positives. This means that the ligand docked with a high RMSD is given a low rank. This can be a problem as there isn't currently a reliable way to discern the difference between false positives and the correct docked posed for a given receptor. If the binding site and ligand is known from a source, such as the X-ray crystal structure, it is possible to notice that the ligand has been docked incorrectly forming this false positive. However if the binding site is unknown then the problems of false positives becomes a major hindrance. There is no way to discern the difference between the false positive and the right answer as stated before. Thus when the ligand is docked to an unknown receptor there is no way of confirming if the poses generated are correct. The way forward therefore is to seek to improve the scoring function, hence our interest in including polarization of both the ligand and the protein.

GLIDE's scoring functions is based upon the ChemScore function of Eldridge et al. (1997), as shown in equation 2.1.

$$G = C_0 + C_{lipo} \sum f(r_{lr}) + C_{hbond} \sum g(\ \ r)\ \ (\ \ \alpha) + C_{metal} \sum f(r_{lm}) + C_{rotb} H_{rotb} \qquad (2.1)$$

Where C_0 is a constant, according to the review presented by (Friesner et al., 2004), the sum in the second term (C_{lipo}) refers to the interaction between ligand-atom and receptor-atom pairs that are defined as lipopholic, this is a hydrophobic effect and is an entropy term, GLIDE uses this term without changing it The third term's (C_{hbond}) sum takes into account all ligand-receptor hydrogen-bonding interactions, GLIDE expands this term weighting the components that depend on whether the donor or acceptor are both neutral or if one or both are charged. The fourth term’s (C_{metal}) takes into account all metal-ligand bonding interactions, GLIDE changes this term to consider only the anionic accepter atoms, it counts the single best interaction when two or more metals are found and GLIDE assess thenet charge on the metal ion in the unligated protein. If it is a positive value the metal ligand is incorporated into scoring if it is negative it is suppressed. The fifth term (C_{rotb} and H_{rotb}) is a penalty scored for freezing rotatable bonds (which is an entropy term), this term is used by GLIDE unchanged from the Chemscore. The terms f, g and h will all denote either a full score (1.00) or a partial score (0.00 - 1.00). The full score is given to a pose that has distances or angles that lie within nominal limits whereas the partial score is given to those that lie outside those boundaries but have not passed a threshold limit. The term (*r*) being the distance or angles in those cases. To include solvation effects GLIDE docks explicit waters into the binding site for each ligand pose and uses an empirical scoring term that measures the exposure of various functional groups to the explicit waters.

GLIDE has two main forms of Glidescore that alter and extend the ChemScore function. They are defined as two docking modes of Standard-Precision (SP) and Extra-Precision (XP), the majority of tests performed in this project are under the SP docking procedure. The SP scoring function is more forgiving in its parameters than that of the XP function and is thus a soft scoring function, as presented in Friesner et al. (2004). The advantage of using this scoring function over the XP version is that it is more proficient at identifying ligands that have decent susceptibility to bind to the receptor. This allows for some flexibility in the glide pose and takes into account a possibly lower resolution receptor or ligand. Thus, it is far superior at general screening. XP scoring functions place some additional penalties into the function which forces only poses or ligands that have a favourable conformation to score. This has its advantages: mostly it has the potential to reduce the potential of false positives. However it is best used with a small subset of compounds that are under study for lead-optimisation and other methods. Due to the harsh parameters set for XP the ligand and receptor need to be of a very high resolution. This was the case for some of the complexes used (e.g. 1S5N, 1PWM and 1UWC (Favia et al., 2011)), but not all, so for consistency Glide SP was used.

2.2.2 OPLS-AA Force fields

GLIDE utilises force fields as a standard during refinement for scoring. During the grid minimisation step (step 3, figure 2.1) GLIDE performs a torsionally flexible energy optimisation on an OPLS-AA grid for the last few poses left after initial refinement (Friesner et al., 2004). OPLS force fields, in short, are a set of parameters that mimic experimental thermodynamic and structural data of atoms in fluid as seen in Jorgensen and Tirado-Rives (1988). An OPLS force field packages a description or set of values for the potential interactions that an organic liquid would yield. These potential interactions cover bond angles and stretches as well as intermolecular and intramolecular forces; the equation is similar to that of the AMBER and CHARMM force fields – see chapter 4.2.

OPLS force fields continue to change and include more parameters and conformations as the field evolves. In the works of Damm et al. (1997), the force fields were expanded to include carbohydrates. As techniques become more refined so too does OPLS. Kaminski et al. (2001) re-evaluated the OPLS force field for peptides.

The computational efficiency and accuracy of the OPLS force field has made it popular for simulating biomolecules. To keep it up with the times, scientists continue to update the parameters for each new challenge. Recently Siu et al. (2012) started optimising OPLS parameters for use with long hydrocarbon chains. The OPLS parameters were originally grounded in using short alkanes thus were not good for long chains. However Siu et al. (2012) have used gas-phase ab initio energy profiles as an addition to OPLS. By having this as an additional parameter, the OPLS yields improved hydrocarbon diffusion coefficients.

2.2.3 The docking problem

One of the biggest hindrances in Ligand docking has been around since the very beginning that of rigidity of the docking process. Originally, both the ligand and the receptor were both frozen rigidly for the process, as stated by Lorber and Shoichet (1998). This created the problem where the ligand would often not dock in the correct conformation or in the binding site, as the ligands could not flex around the binding site, although this problem only affected flexible ligands. However, since the early days, programs like GLIDE have overcome the rigid ligand problem. These programs use the strength of computer speeds so it has become less of a problem to calculate all the conformations a ligand might have as it is relatively small and thus has a manageable number of poses. However the problem still remains that there is a rigid receptor in place. For rigid receptors it must be assumed that the ligand conformations must be tested near to the experimentally observed conformations or the test will not function. Thus, this is a problem that has yet to be fully addressed. However, as reported by Verdonk et al. (2005), some programs are starting to accommodate this problem, at least modestly. GOLD, like GLIDE, includes full ligand flexibility but it also allows rotational flexibility for hydrogen atoms based around the receptor; this was also a feature of ligandfit. Another problem with docking is the use of the scoring function. As the scoring function in docking programs tends towards neglecting solvation effects, including them in some implicit way, or using solvent models as a snap shot (i.e. including a small number of explicit water molecules from the X-ray structure). A snap shot is where a structure is generated *in vacuo* with solvent molecules in place and it is then ranked with a scoring function. The consequence of this is that the docked ligand will be funnelled onto the correct place by the water molecules.

However, there have been studies to sidestep the receptor rigidity problem. Techniques such as soft docking and partial side-chain flexibility are amongst those used today according to Cavasotto et al. (2005). Using multiple receptor conformations (MRCs) is one of the best choices to reduce the problem currently, as presented by (Carlson, 2002). The advantage to this method over others is that the structural space of the binding pocket can be represented even in the case of loop displacements. This works by making an ensemble of multiple conformations, hence the name, of different discrete conformations of separate structures sharing the same backbone trace. Alternatively, for example, multiple snapshots from MD simulations can be used (Carlson, 2002). However, this introduces additional variables into the docking assessment and so this was not used here.

2.2.4 The presence of water molecules in Ligand docking

Despite the relative accuracy of GLIDE, there are still problems within the field of Ligand docking, as shown in Verdonk et al. (2005). One of these major problems is the prediction of the effect of water molecules on ligand to protein interactions. Either water molecules can have no effect on the docking and thus do not need to be present as this only causes spatial and conformational problems and the water is normally displaced, or in some cases water molecules near the active site will form hydrogen bonds with the ligand thus helping to mediate the docking process. Therefore in some cases, the presence of water is important to the docking process, such as in the HIV-1 protease system (Lam et al., 1994). However, another problem arising from having water molecules present for the docking process is the fact that a rigid crystal structure is used in all calculations; this includes the water molecules present. Amongst the other problems the rigid paradigm causes, this can develop a slight problem where when the receptor grid is generated there is an empty binding pocket where the ligand would normally be situated. This in nature is usually filled with water that will be displaced by the ligand, however now we have an energetically favourable location for the ligand to bind too with a slight bias towards it as there is more space for the ligand within the structure. In essence it would not generate false positives but it is not be an accurate test of docking as it would be artificially biased.

Verdonk et al. (2005) discuss that a key missing concept to current approaches is that in nature when a ligand displaces a water molecule to bind to the receptor site it gains rigid-body translational and rotational entropy. This is currently not factored into scoring functions and should be used possibly as a term that would reward a ligand for performing this displacement. This can help the scoring function have a higher chance of generating the correct answer when water is present.

2.2.5 Series of test molecules

The SERAPhiC benchmark set, as presented in Favia et al. (2011) was chosen for this study; these proteins were chosen as they were part of a validation set. In this case, a majority of the proteins have a good RMSD associated with the docking tests in the paper, as a rule of thumb this is about 2.0 Å. Each entry in the SEARPhiC set needed to: have a resolution of ≤ 2.5 Å, date of disposition ≥ year 2000, presence of an article describing the crystal structure, a macromolecule of ≥ 200 amino acids and have at least one ligand of between 78 and 300 daltons with at least 6 heavy atoms. Each protein was prepared using Maestro's protein preparation package following the standard procedure, except all the waters were removed from around the structure.

Each of the fragments in the docking set forms non-covalent bonds with their respective proteins. The fragments in the SERAPhiC set all modulate protein activity by either being an inhibitor or a substrate.

In the SERAPhiC set Favia et al. (2011) performed docking similar to the methods presented in this chapter. The docking program adopted by Favia et al. (2011) was Molsoft using Internal Coordinate Mechanics (ICM) forcefields. Molsoft for the SERAPhiC set used a Biased Probability Monte Carlo Method, as shown in Totrov and Abagyan (2001). Each ligand was considered with and without water molecules. There were 3 grid sizes generated for each fragment; 3.5 Å, 5 Å and 7 Å. The analysis performed was to determine if the docking used with the ICM engine could find the native pose of the fragment in the top 10 hits and if the best docking score attained was the native pose. Favia et al. (2011) referred to this as the soft and hard successes respectively. Using their methods soft success was approximately 90% of all structures and hard success was approximately 60% of all structures.

The SERAPhiC set was also analysed using BINANA which is an algorithm developed by Durrant and McCammon (2011). This algorithm summarises the interactions between a fragment and protein within their complexes. The data for this can be found in the paper by Favia et al. (2011). The resolution of each model can be found in appendix 2.2.

2.2.6 Molecule preparation

The standard procedure in Maestro's protein preparation wizard is as follows: assign bond orders, add hydrogens, create zero-order bonds to metals, create disulfide bonds, convert selenomethionines to methionines, Cap N and C termini with ACE and NH3 (small fragments were not capped). We had no selenomethionines in the protein but the setting was left on as default.

Each protein was then minimized. The waters were removed by setting the maximum distance that waters were kept to 0 Å. Once the protein has been prepared then a docking receptor grid is generated using the original position of the ligand from the crystal structure as the centre of the grid. The ligand is removed from the grid to create an open pocket. The size of the docking grids was set to be larger than normal, to 36 Å × 36 Å × 36 Å. This was to reduce bias to the binding site so that the ligand might dock in other pockets within the grid and not just to the centre. GLIDE has a slight inbuilt bias towards the centre of the grid; making the box larger offsets this bias. The ligand is then docked into this new docking receptor grid and the best 15 ligand poses were generated for each protein. The criteria for the best poses was set by glidescore, as detailed in Friesner et al. (2004). This often coincides with the lowest RMSD though this was not always the case. In the cases where there was more than one identical chain, a docking run was performed with only one chain present, usually chain A, with their respective ligand, but all active site from the chain were left empty as per the other runs.

2.3 QM/MM

The concept behind the QM/MM method is to take into account QM calculations such as the processes behind bond-breaking/forming, charge transfer and electronic excitation of atoms while also including a MM force field that is based on small inter and intra molecular forces, such as Van der Waal and electrostatic interactions as presented by Lin and Truhlar (2007). The QM/MM method aims to include the environment around the active site of the protein.

The QM portion of the method takes the localised region around the active site and calculates the QM forces in this region. The QM region will also include its surroundings, the protein environment also known as the secondary subsystem (SS) at the MM level. The key amino acids in the active site and its neighbours are defined as the primary system (PS); here the PS was the ligand. QM/MM methods can be treated in two different ways by changing how the electrostatic interactions take place between PS and SS, either by mechanical or electrostatic embedding. According to Bakowies and Thiel (1996), mechanical embedding handles the interaction between the PS and SS only at the MM level, which is a simpler approach. The electrostatic embedding calculates the QM calculations for the PS in presence of the SS. One electron operators are included to describe the electrostatic interactions between the two systems; these operators enter the QM Hamiltonian. This is the sum of all kinetic energies for the atoms in the system and the potential energy for any atoms currently associated with the system; this is the approach taken here.

Methods have been suggested by Illingworth et al. (2008) that takes the QM/MM approach and combines it with the polarization of both the ligand and its target. As often polarisation is a key term that isn't present for most molecular mechanics studies, it is an area of interest. This project hopes to expand on this research and assess new techniques using the hybrid QM/MM and polarization to attain more accurate results for blind docking.

2.3.1 QM/MM using Jaguar

For a given docking run, each of the 15 ligand poses were then merged with the rest of the protein structure, creating a separate file for each. These merged files were a separate input files for each pose that included the posed ligand and the protein. A Qsite input file was generated for each of the 15 combined structures. Each test was performed as a single point QM/MM calculation with the ligand as QM and the protein as MM, for the purposes of calculating the QM potential derived charges. Each test also had 1000 cycles set for MM optimization. This was performed for general geometric minimization of the structure. At this point, instead of using Qsite normally with the above settings, in-house polarisation scripts to polarise the ligand and the surrounding binding pocket were used instead. The in-house scripts can be found in the appendix 2.3. This generated the polarized charges for the ligand and for the protein, which could be used in the re-docking (see below).

2.3.2 Methods for modelling Polarization

Polarizability helps explain interactions between nonpolar atoms and electrically charged species, such as ions or polar molecules. These charged species will have dipole moments. When a neutral nonpolar atom is subjected to an electric field, its electron cloud can be distorted. Usually nonpolar atoms have roughly symmetric arrangement of electron clouds. The ease of this distortion is polarizability as according to Cornah (2013). There are several ways for modelling polarization, as described in sections 2.2.3.1 – 2.2.4.

2.3.3.1 Quantum mechanics

During the SCF procedure, the atomic charge distribution that emerges can be viewed as a consequence of the optimized linear combination of atomic orbitals (LCAO) coefficients. When the wavefunction is perturbed during a QM/MM calculation, these coefficients must be re-optimized. Ligand polarization is therefore automatically included in QM/MM calculations (Feynman and Hibbs, 1965).

2.3.3.2 Induced dipole method

This model uses fixed atomic partial charges, as found in non-polarziable force fields, then adds a set of inducible point dipoles as shown in Friedrich and Herschbach (1999). Each induced dipole, μ_i, at site i, is then determined by the electric field, E_i, at that site according to equation 2.2, where α_i is the isotropic polarizability of site i.

Electrostatic interactions cannot be fully calculated via Coulomb potential in the region of the dipole as the Coulomb potential evaluates charge-charge interactions. New terms must be added to take into consideration the dipole-charge or dipole-dipole interactions as shown in Maple et al. (2005), and these considerably increase the complexity and cost of the calculations.

$$\mu_i = \alpha_i E_i \qquad (2.2)$$

2.3.3.3 Drude oscillator

In the Drude oscillator model, the polarizability is modelled by adding massless charged particles attached to the polarizable atoms. The massless charged particles are Drude particles. In current models the polarizable atoms do not include hydrogens. These charged particles are attached via a harmonic spring. The polarizable atom then has its charged spread across its core and the massless Drude particle (Lopes et al., 2009).

2.3.3.4 Charge equilibration

The Charge equilibration model predicts charges of large molecules by using their geometry and experimental charge properties as shown in Rappe and Goddard III (1991). The formalism of this model assumes that the chemical potential of a molecule is equilibrated via the redistribution of charge density over the molecule. This approach allows the charges to respond to changes in environment. Due to the changing charges this is also referred to as the fluctuating charge method, as shown in Baker (2015). To determine the polarizabilities, the partial charge of each atom is placed at the atomic nucleus. The electrostatic interactions are calculated using the normal Coloumb potential. To achieve the effect of changing charges according to the environment, this method assigns fictitious masses to each of the charges, therefore each of the nuclei, and treats them as new degrees of freedom when calculating the equations of motion, as shown in Rappe and Goddard III (1991). The charges continually flow between the atoms until their electronegativities are equalized. However this method cannot represent polarization that is not in the direction of the bonds. This makes it difficult to polarize benzene against a species perpendicular to it, as stated in Baker and Grant (2006) and Winn et al. (1997).

2.3.4 The induced charge method

2.3.4.1 Theoretical basis of the induced charge method

In the induced charge method, an induced dipole is approximated by the charges on neighbouring atoms and the central atom itself. This can be shown in Fig. 2.2.

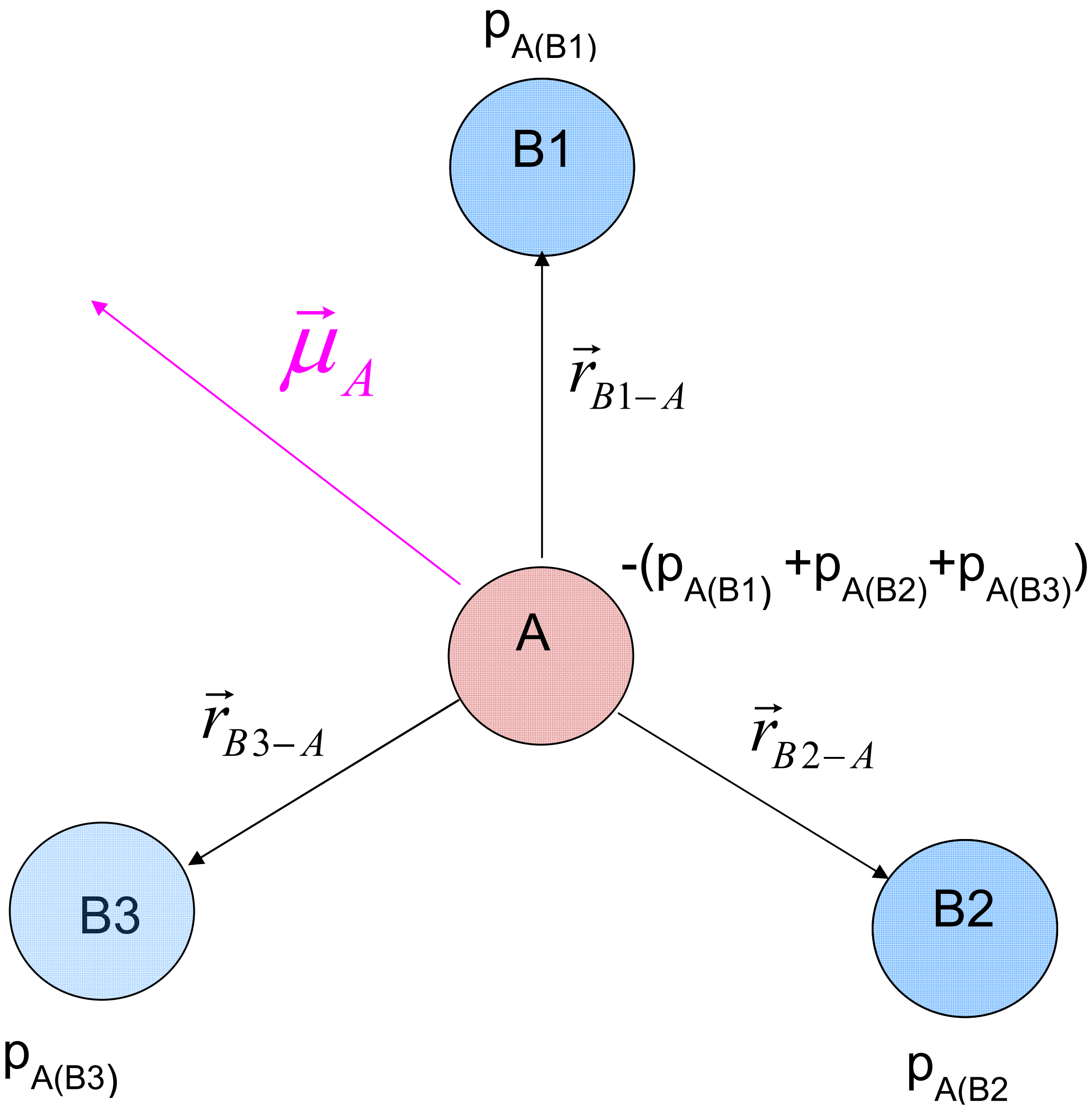


Figure. 2.2. Key descriptors relevant to the induced charge method of approximating polarization.

$$\vec{\mu}_A = \chi_A \cdot \vec{p}_A \tag{2.3}$$

All equations in this section are from Ferenczy and Reynolds (2001). Here, μ_A is the induced dipole in equation 2.3, this is expanded by using the equations 2.4 and 2.5, where χ_A is the vector of the distances between the neighbouring atoms (B) and the central atom (A) shown in equation 2.4.

$$\chi_A = (\vec{r}_{B1-A}, \vec{r}_{B2-A}, \vec{r}_{B3-A}) \tag{2.4}$$

In equation 2.5, $\bar{P}_A$(Bx) is the partial induced charges of each of the respective neighbouring atoms.

$$\vec{p}_A = \begin{pmatrix} p_{A(B1)} \\ p_{A(B2)} \\ p_{A(B3)} \end{pmatrix} \tag{2.5}$$

Using equation 2.2 from the induced dipole method, we can modify it to generate equation 2.6.

$$\vec{p}_A = \alpha_A \cdot (\chi_A^+ \chi_A)^{-1} \chi_A^+ \cdot \vec{F}_A \tag{2.6}$$

Then using Taylor expansion for electrostatic potential we can write it as equation 2.7.where ϕ is defined as the electrostatic potential.

$$\Phi_{B1} = \Phi_A - \vec{r}_{B1-A} \cdot \vec{F}_A + \dots \tag{2.7}$$

Truncating equation 2.7 we can write equation 2.8.

$$\chi_A^+ \cdot \vec{F}_A = \vec{\Delta}(\Phi_A) \tag{2.8}$$

Where $\Delta(\phi_{A)}$ is the difference vector of electrostatic potential so we can write out a final formula of equation 2.9

$$\vec{p}_A = \alpha_A \cdot (\chi_A^{+} \chi_A)^{-1} \cdot \vec{\Delta}(\Phi_A) \tag{2.9}$$

Thus, calculation of the MM induced charges required the QM calculation of the electrostatic potential at the MM atoms (using Jaguar); the electrostatic field was not required. The sum of induced charges on a given atom can be written as q^{IND}(A). This is the sum of negative partial induced charges of atom A in the above diagram and the partial induced charges of the neighbouring atoms which can be summarized as:

$$q_A^{ind} = -\sum_{A' \in A} p_{AA'} + \sum_{K} \sum_{K' \in K} p_{KK'} \delta_{AK'} \tag{2.10}$$

The total atomic polarized charge is a sum of the permanent and induced charges which can be written as:

$$q_A^{tot} = q_A^{per} + q_A^{ind} \tag{2.11}$$

The Electrostatic energy can be then calculated using Coulomb following:

$$E^{ele} = \frac{1}{2} \sum_{I,J} q_I^{tot} q_J^{tot} \frac{1}{r_{IJ}} + E^{self} \tag{2.12}$$

Where:

$$E^{self} = \frac{1}{2} \sum_{I} \frac{1}{\alpha_I} \vec{p}_I^{\,+} \chi_I^{+} \chi_I \vec{p}_I \tag{2.13}$$

2.3.4.2 Application of the induced charge method to GLIDE docking

For the docking, we wish to be able to polarize just the ligand or the ligand and the enzyme target. Ideally, the ligand would be polarized in its correct pose, but in an actual docking investigation as part of a drug design programme, the correct pose would not be known prior. For this reason, we have made the assumption that the best docked pose is a sufficiently good approximation to the correct pose. Consequently, we have polarized the ligand and enzyme according to the geometry of the best pose and then re-docked the ligand to the original crystal structure. Then their RMSDs of these new poses were compared to the original docked poses.

Using the method within the in-house scripts, a set of output files were generated with the new polarized charges for the QM region, which was set as the ligand; the MM region of the QM/MM system was the rest of the system, i.e. protein. The basis set used for the QM region was 6-31G*. The method used was B3LYP. This in house script changed the polarisation charges for all the atoms in the region, both from the ligand and protein. The ligand charges were derived from the perturbed wavefunction of the ligand within the protein, as potential derived charges, while determination of the QM potential at the MM atoms enabled the calculation of the induced charges for the enzyme (via equation 2.9). The total enzyme charges are the induced charges added to the base charges. Once the charges were extracted from the file they were reinserted into the original structure, replacing the atomic charges of the atoms for the ligand and the protein. A set of atomic charges were generated for each of the 15 ligand poses of each complex for the ligand and protein. Chart 2.1

A flowchart indicating the steps required in protein and ligand preparation and the subsequent polarized docking, as described in sections 2.2.6 and 2.3.4.2.

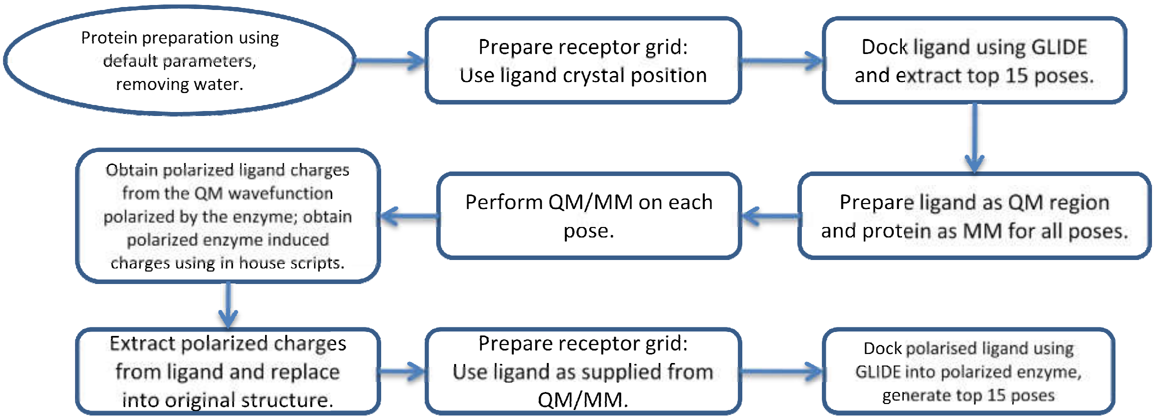

Protein preparation using default parameters, removing water.
Prepare receptor grid: Use ligand crystal position
Dock ligand using GLIDE and extract top 15 poses.
Prepare ligand as QM region and protein as MM for all poses.
Perform QM/MM on each pose.
Obtain polarized ligand charges from the QM wavefunction polarized by the enzyme; obtain polarized enzyme induced charges using in house scripts.
Extract polarized charges from ligand and replace into original structure.
Prepare receptor grid: Use ligand as supplied from QM/MM.
Dock polarised ligand using GLIDE into polarized enzyme, generate top 15 poses

Chart 2.1: Summary of protein preparation, docking and generating polarization charges through the QM/MM method.

2.3.4.3 Generating new docking grids based upon new atomic charges from ligands

With the new atomic charges in place, each of the 15 different ligand poses was re-imported into Maestro; each ligand has the same set of charges, namely the set derived from the top docked pose. A receptor grid was once again generated for the top pose of each structure using the ligands in place as the centre of the grid as per the beginning of this method (section 2.3.4.2). The GLIDE docking procedure is repeated, but with the new polarized charges. For each structure, 15 new poses were generated from the grid based upon the new charges from the QM/MM. These poses were then compared to the original set to observe if docking had improved with the inclusion of the polarisation charges. This was performed by comparing the

RMSD against the crystal structure for each of the original docked posed versus the RMSD of each of the polarized poses against the crystal structure. The in-house scripts used to implement the induced charge method in GLIDE docking are shown in Appendix 2.2.

2.4 Results

2.4.1 Initial Docking

Table 2.1 shows the RMSD for each docked ligand before and after generating a QM/MM region around the binding site and running a QM/MM optimization with ligand and protein polarization. Thus, the basic Glide docking results are given in the 'before' column, while the polarized docking results are given in the 'after' column. This is a condensed table of actual results as there are 15 QM/MM possible regions generated based on the 15 initial docking poses for each ligand. The QM/MM regions here are those generated from the first pose of each initial dock.

Table 2.1: Table of RMSD values of GLIDE docking before standard docking and after polarized docking using QM/MM polarisation taking the top docked posed according to Glidescore as starting geometry. Lowest RMSD for each column is shown in ***bold italics***. For some proteins, fewer than 15 poses were generated. The brackets next to a name denotes chain.

PDB	1MLW		1Y2K	IC50 = 21 nM	1FSG (C)	IC50= 9.5μM	1FSG (A)	IC50= 9.5μM	1TKU (A)		1TKU (B)		1S5N	
Pose No.	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM
1	3.331	3.360	2.336	3.803	***0.142***	***0.155***	***0.146***	***0.085***	3.336	3.354	3.132	***1.911***	2.857	2.877
2	3.179	3.575	2.480	5.077	3.384	3.385	3.385	3.379	***2.526***	3.178	3.022	2.044	2.253	2.626
3	3.572	2.863	***2.328***	***2.468***	7.376	6.812	5.770	5.146	3.331	***2.528***	***2.583***	2.969	2.105	2.641
4	2.857	2.580	2.333	6.967	5.970	6.062	6.510	5.728	3.325	3.076	3.203	2.081	3.053	2.711
5	1.633	2.070	2.746	7.063	6.677	6.096	7.540	6.70	3.229	3.256	3.209	1.896	2.507	2.995
6	1.964	3.004	6.931	7.218	5.597	5.154	5.100	5.811	3.049	3.315	2.964	0.731	4.332	1.890
7	2.050	2.679	7.043	6.988	6.679	7.237	5.721	7.568	2.787	3.342	3.316	***0.454***	2.926	3.122
8	3.064	***1.832***	7.002	6.977	5.174	6.994	6.058	6.479	3.277	3.096	3.028	0.563	2.668	4.534
9	1.715	2.150	7.010	3.498	6.977	7.471	4.950	6.703	3.248	3.055	2.888	1.096	3.373	2.531
10	2.214	2.217	7.258	7.178	7.798	7.344	7.516	8.209	2.965	3.094	2.856	1.893	2.379	2.153
11	2.275	2.578	7.220	7.277	7.735	6.231	4.974	7.732	2.984	3.132	2.855	1.573	2.566	***1.595***
12	2.02	2.298	3.530	4.979	7.091	7.807	7.594	6.034	2.934	2.929	2.9	1.279	2.046	1.867
13	***1.621***	2.764	7.511	4.957	6.022	5.299	5.682	7.833	3.184	3.052	2.989	1.791	2.289	3.329
14	2.285	2.377	3.531		7.034	5.694	6.962	7.064	2.917	3.022	2.915	1.455	2.236	2.507
15	2.598	2.767	3.601		5.115	4.857	5.375	5.443	2.981	2.861	3.003	1.458	***1.572***	2.454

1MLW: tryptophan 5-monooxygenase (Wang et al., 2002). **1Y2K:** camp-specific-3',5'-cyclic phosphodiesterase (Card et al., 2005) **1FSG:** hypoxantine-guanine phosphoribosyltransferase (Héroux et al., 2000) **1TKU:** 3,4-dihydroxy-2-butanone 4-phosphate synthase (Echt et al., 2004) **1S5N:** xylose isomerase (Fenn et al., 2004)

PDB	1R5Y	K i= 0.35μM	1F8E	K i= 15μM	1PWM	IC50= 935 nM	1SQN	K d= 0.4 nM	1UWC		2BRT		1F5F	IC50= 3.8 nM
Pose No.	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM
1	***0.112***	***0.162***	1.034	***0.173***	***0.328***	***0.222***	***0.125***	***0.152***	***0.276***	***0.377***	***0.489***	***0.787***	***0.167***	***0.051***
2	0.148	2.715	1.023	0.715	4.542	4.570	0.957	0.511	5.102	1.253	6.727	1.816	6.284	6.296
3	2.897	4.268	***0.488***	1.208	4.564	4.640	2.255	1.415	2.594	2.607	1.834	5.239	6.338	6.342
4	4.244	4.383	0.488	1.349	4.66	4.738	2.585	2.348	1.150	0.592	1.94	2.172	6.634	6.651
5	4.926	4.914	1.147	1.328	4.518	4.476	6.591	6.489	0.629	5.090	6.708	6.634	2.439	2.737
6	2.927	3.005	1.189	1.443	4.427	4.417	6.525	6.600	1.006	2.357	2.205	6.672	2.655	2.639
7	4.928	4.341	1.44	0.933	4.749	5.372	6.547	2.591	2.234	2.358	6.718	3.031	2.809	2.810
8	4.459	4.675	1.287	0.868	4.473	6.679	2.907	6.535	5.978	4.775	1.953	6.770	6.503	2.551
9	4.31	4.277	1.504	4.538	6.678	5.346	6.433	6.444	0.867	0.898	2.007	3.283	6.667	6.498
10	4.12	4.093	0.674	1.347	6.906	4.886	2.645	2.282	2.574	6.065	6.544	3.627	2.886	2.530
11	4.718	7.903	4.341	4.576	6.983	4.889	6.648	6.719	2.607	5.055	1.134	0.936	6.572	2.889
12	4.872	8.458	4.562	4.054	7.499	4.239	6.475	6.526	2.427	2.698	0.883	6.467	6.420	6.682
13	8.022		2.488		5.862	5.875			5.134		6.12	6.769	3.473	3.513
14	6.968		4.405		4.97	4.222			5.598		4.55	4.550	2.728	6.428
15	7.563		4.318		4.349						4.55	6.823		

1R5Y: queine trna-ribosyltransferase (Brenk et al., 2004) **1F8E:** neuramidase (Smith et al., 2001) **1PWM:** aldose reductase (El-Kabbani et al., 2004) **1SQN:** progesterone receptor (Madauss et al., 2004) **1UWC:** feruloyl esterase (McAuley et al., 2004) **2BRT:** leucoanthocyanidin dioxygenase (Welford et al., 2005) **1F5F:** sex-hormone binding globulin (Avvakumov et al., 2000)

PDB	1UI0 (B)	K i= 88 nM	1UI0 (A)	K i= 88 nM	1YNH		2B0M	K i= 2.8μM	2BL9	K i = 0.16 nM	2CIX	K d= 33 mM	2CIX redock	
Pose No.	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM
1	2.810	2.041	***0.259***	***0.336***	0.480	***0.164***	***0.332***	***0.306***	***0.133***	***0.187***	9.278	9.236	2.135	2.929
2	8.568	7.716	2.86	2.877	***0.444***	0.566	1.436	1.002	0.642	0.660	7.915	9.347	1.748	1.609
3	2.987	3.169	2.468	3.138	0.766	1.085	3.087	0.660	2.738	1.033	2.064	2.777	**0.257**	2.129
4	8.344	2.292	3.056	2.399	0.459	1.165	3.14	1.242	2.432	2.647	2.774	***1.336***	2.409	1.902
5	3.176	***0.440***	2.967	6.016	0.575	0.567	0.716	0.840	3.341	2.590	1.458	2.100	1.997	2.548
6	2.837	2.456	6.074	2.583	0.834	0.670	1.692	3.088	5.609	2.916	***1.349***	1.791	0.852	**0.511**
7	***0.170***	1.765	2.601	2.602	0.625	0.921	3.221	3.143	3.843	2.430	1.906	1.905	1.899	1.817
8	2.216	2.235	3.064	3.102	1.022	0.544	2.769	3.231	3.745	5.634	1.637	8.562	2.839	2.222
9	1.117	3.086	1.143	2.743	1.092	0.823	3.730	3.545	6.269	5.616	1.822	1.823	7.032	7.142
10	3.220	2.447	5.178	5.842			2.251	2.254	4.897	3.689	7.641	1.901	6.944	8.135
11	2.863	2.893	5.877	2.308			2.339	2.341	5.989	3.466	3.329	7.707	6.828	6.853
12	2.615	2.839	2.922	5.182					5.145	6.051			5.824	7.297
13	2.268	2.999	5.573	5.396					6.173	6.128				
14	2.509	3.257	4.037	3.793					6.662					
15	2.674	2.731	8.002	3.083										

1UI0: uracil-dna glycosylase (Marcyjaniak et al., 2004) **1YNH:** succynilarginine dihydrolase (Tocilj et al., 2005) **2B0M:** dihydroortate dehydrogenase (Hurt et al., 2006) **2BL9:** dihydrofolate reductase-thymidylate synthase (Kongsaeree et al., 2005) **2CIX:** chloroperoxidase (Kühnel et al., 2006)

PDB	1W1A		2FDV	K i= 0.8µM	1M3U		2AIE	IC50= 2.2	2BKX		1M2X	K i= 70 100µM	1YKI	
Pose No.	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM
1	***0.149***	0.643	1.168	1.621	***0.217***	6.342	1.608	1.597	1.964	***1.485***	2.607	***0.182***	8.773	1.870
2	0.645	***0.135***	1.179	1.198	0.855	6.333	1.528	1.509	1.506	2.519	***0.439***	0.681	10.400	2.845
3	2.72	2.648	1.189	1.175	1.600	7.044	1.477	1.458	1.880	2.346	1.562	1.928	10.113	2.558
4	4.875	1.587	***0.276***	0.832	1.635	6.426	0.902	1.040	2.167	2.468	1.578	6.349	14.332	5.201
5	2.578	1.723	0.835	***0.261***	1.612	5.987	1.339	***0.898***	1.715	2.322	7.499	2.684	11.929	4.046
6	2.607	1.743	1.254	1.225	0.749	6.296	1.525	3.025	2.030	2.545	4.430	6.067	9.358	3.728
7	5.541	2.524	1.255	0.796	1.842	7.012	2.264	1.521	***0.811***	2.756	7.548	2.356	***6.470***	***0.481***
8	1.656	2.711	1.561	0.937	0.795	6.554	1.442	1.645	2.163	2.155	0.986	5.703	12.457	10.162
9	5.41	1.761	0.968	1.535	2.116	7.181	2.223	2.129	2.854	2.484	7.577	5.036	9.840	2.728
10	5.728	4.868	0.922	1.499	1.799	8.348	***0.372***	2.782	2.191	2.146	7.555	5.326	7.092	6.323
11	5.298	4.919	1.812	1.347	0.894	8.511	2.243	2.32	2.166	2.138			12.328	6.853
12	5.381	5.488	2.492	1.704	1.792	9.919	1.901	1.988	1.598	2.435				
13	5.512	5.487			1.698	7.087	2.442	2.848	1.003	1.795				
14	5.509	5.675			2.321	8.107	2.696		1.667	2.614				
15							2.224		1.894	2.105				

1W1A: polysaccharide deacetylase (Blair and van Aalten, 2004) **2FDV:** cytochrome p450 (Yano et al., 2006) **1M3U:** ketopantoate transferase (von Delft et al., 2003) **2AIE:** peptide deformylase (Smith et al., 2003) **2BKX:** glucosamine-6-phosphate deaminase (Vincent et al., 2005) **1M2X:** metallo-beta-lactamase (García-Sáez et al., 2003) **1YKI:** oxygen insensitive nad(p)h nitroreductase (Race et al., 2005)

PDB	2UY5	K i= 3.2μM	2RDR		2J5S		2Q6M	K d= 510 nM	2I5X		1OFZ	K d= 24.1μM	1WOG	
Pose No.	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM
1	***0.151***	***0.172***	***0.449***	***0.161***	1.169	1.042	1.902	***0.566***	1.267	1.292	**0.270**	**0.163**	4.567	**0.649**
2	0.939	0.940	4.339	4.451	***0.258***	2.338	1.928	2.040	1.314	1.267	0.866	0.865	**4.181**	0.861
3	1.143	6.838	1.374	1.429	0.756	***0.274***	1.896	1.995	1.308	1.305	0.871	2.887	4.42	4.721
4	2.497	6.894	4.606	1.071	0.708	2.254	0.607	1.452	1.052	1.077	3.054	4.053	4.722	1.366
5	6.931	2.531	1.253	4.471	2.308	2.055	***0.321***	2.251	1.255	1.253	2.792	3.426	5.343	1.184
6	6.972	5.459	4.554	4.301	2.021	0.883	1.988	0.689	1.241	1.193	4.074	3.459	4.479	4.324
7	5.516	5.914	1.303	0.683	0.996	0.897	2.330	1.349	1.159	1.448	3.623	2.869	5.263	1.398
8	5.535	5.193	0.728	0.839	2.244	0.700	1.520	5.652	0.697	1.697	3.337	2.411	4.794	2.784
9	6.190	5.861	4.411	4.440	2.049	1.849	5.522	6.009	***0.345***	0.694	3.418	4.096	4.763	1.112
10	5.785	6.065	4.555	1.723	1.968	2.249	6.020	5.895	1.387	***0.342***	3.367	2.895	4.868	0.709
11	5.090	5.912	4.431	2.065	1.791	2.908	5.891	5.524			1.706	3.285	4.750	0.893
12	5.058	4.767	4.468	4.304			5.517	3.511			2.009		5.066	1.219
13	3.850	5.933	4.425	7.340			3.540	3.506			2.392		5.366	1.784
14							5.525	3.314			3.997		5.114	
15											4.087			

2UY5: endochitinase (Hurtado-Guerrero and van Aalten, 2007) **2RDR:** 1-deoxypentalenic acid 11-beta hydroxylase fe(ii)/alpha-ketoglutarate dependent hydroxylase (You et al., 2007) **2J5S:** beta-diketone hydrolase (hydrolase) (Bennett et al., 2007) **2Q6M:** cholix toxin (Jørgensen et al., 2008) **2I5X:** receptor-type tyrosine-protein phosphatase beta (Evdokimov et al., 2006) **1OFZ:** fucose specific lectin (Wimmerova et al., 2003) **1WOG:** agmatinase (Ahn et al., 2004)

PDB	2GG7	IC50= 1.75μM	2GVV	K i= 125μM	2ZVJ	IC50= 1.8μM	1YV5	IC50 = 5.7 0	3DSX	K d= 1.4 mM	2FF2	K i = 6.2 nM
Pose No.	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM	Before QM/MM	After QM/MM
1	***0.303***	***0.278***	***0.570***	0.812	4.024	***0.101***	1.255	***0.609***	2.695	2.849	2.017	***0.230***
2	0.894	1.903	0.894	0.653	4.021	0.646	***0.381***	1.486	2.760	***2.475***	1.021	3.092
3	1.428	0.739	0.879	***0.271***	0.805	4.163	1.439	1.304	5.799	6.080	***0.902***	3.231
4	2.387	2.118	0.861	1.113	***0.176***	1.018	4.793	1.657	***2.503***	2.399	2.052	2.021
5	4.363	3.237	1.269	1.014	3.521	4.069		1.010	7.584	5.126	2.462	2.379
6	6.525	2.189	0.988	1.090	3.581	4.465		1.212	5.773	3.673	2.449	2.150
7	6.594	2.393	0.935	1.233	4.336	1.591		1.081	5.544	4.826	3.118	3.170
8	6.937	2.494	0.835	0.960	1.449	1.684		1.473	5.769	7.347	3.122	2.151
9	7.056	3.097	0.526	0.795	2.698	3.337		1.007	4.985	5.830	2.762	3.149
10	6.653	2.617	1.116	1.142	2.321	4.674		4.822	5.220	2.997	2.678	3.503
11	6.971	4.791	1.008		2.426	3.396		4.709	9.745	4.223	2.015	
12	6.858	3.282	1.215					4.908	5.961	2.853	2.652	
13	7.183								2.863	5.848	2.284	
14	7.622								4.528	6.409	2.560	
15										9.522		

2GG7: methionine aminopeptidase (Evdokimov et al., 2007) **2GVV:** phosphotriesterase (Blum et al., 2006) **2ZVJ:** catecholo-methyltransferase (Tsuji et al., 2009) **1YV5:** farnesyl pyrophosphate synthetase (Kavanagh et al., 2006) **3DSX:** geranylgeranyl transferase type-2 subunit alpha (Guo et al., 2008) **2FF2:** iag-nucleoside hydrolase (Versées et al., 2006)

The overall results of the docking are summarized in Table 2.2. Standard glide has a 39% error rate, as it failed to dock top ranked pose for 16 out of 41 molecules to within 2 Å, while the polarized docking reduced the error by 39% down to 27%. An even more significant reduction in the error was observed if the criterion was to dock any of the poses to with 2 Å; the reduction was from 23% to 13%. In polarized docking, the top pose was more likely to be the best pose: the top pose was the best pose in 21 case for polarized docking and only 19 cases for standard Glide. In only one case did the standard Glide give a correct top pose while polarized docking failed. In contrast, polarized docking gave a correct pose in 6 cases where standard Glide failed. In only 2 of these 7 cases was there a failure to find a good pose with an RMSD of less than 2 Å, some if the aim is to find a correct pose of any rank, then there are certainly grounds for supplementing standard Glide with polarized docking.

Table 2.2 Summary of Table 2.1.

	Before (no polarization)	# incorrect	After (polarized docking)	# incorrect
N[a]	41		41	
# < 2 A (top pose)	25	16	30	11
# < 2 A (all poses)	33	9	36	5
Top pose lowest	18	23	25	16
Top pose lowest and < 2 Å	18	23	25	16
best (all poses)	19	22	21	20
Best (top pose)	19	22	21	20
Only before correct (top pose < 2 Å)	1			

Only after correct (top pose < 2 Å)		6	6
After and Before incorrect (top pose) / all poses	7 / 2		
After or before incorrect (top pose) / all poses	6/4		

[a] including the 2CIX redock

Thus, results such as 1WOG and 2ZVJ have improved over those obtained from standard Glide. For example, there is a significant improvement of the RMSD for 1WOG from 4.57 to 0.65. There is still one result where the regular glide dock is better than after the QM/MM calculations. The RMSD for 1M3U was 0.22 for standard Glide and 6.34 after polarized docking.

The industry standard for a decent dock is an RMSD under 2 Å; thus for comparison of methods a good result for either method should have an RMSD of under 2 Å.

2CIX first gave us some problems with docking to get a good RMSD. It was later re-docked several times by varying the size of the GRID before moving on to polarized docking to see if we could get an improvement. The best we could get is shown in table 2.1 as the last column.

2.4.2 Polarized docking (re-docking)

Table 2.3 presents the results of polarized docking in which both the ligand and the protein are mutually polarized according to the top ranked pose from the initial docking given in Table 2.1. Table 2.3 shows the glide score ranks of the first results under 2 Å for the proteins that had better or worse RMSDs with polarized docking. It shows that in some cases such as 1TKU chain B, 1F8E or 1UI0 chain B that when taking the best glide score even when there is better RMSD results in the lower rankings we can still obtain a good result following re-docking as an additional step after the QM/MM calculations.

13 of the proteins tested have a markedly better RMSD. 1 is worse and 24 have not changed significantly. 1S5N fails on both methods. This is where the top glidescore poses were not below the 2 Å industry standard. 1S5N's RMSD for both methods was above 2 Å.

As stated in section 2.3.1, of the 38 results, 25 gave an RMSD below 2 Å and 24 gave an RMSD below 1.5 Å for the basic docking.

After the QM/MM results, these figures were 31 and 29 respectively. In 24 of these cases the RMSD improved after QM/MM method. In 25 cases, the lowest RMS (over all poses) was for the QM/MM results. In 6 cases, the QM/MM method brought the results into the desired range of RMSD < 2 Å, while in one case, the good docking results were spoiled by the QM/MM method.

Table 2.2: Table of ranks of 1st results of initial docks or re-docks under 2 Å. Where there is a difference in rank, the best result is in **bold**.

Re-dock after QM/MM better	Rank of 1st significant result on initial dock (< 2 Å)	Rank of 1st significant result on re- dock (< 2 Å)
1FSG chain A	1	1
1TKU chain A	**2**	3
1TKU chain B	3	**1**
1F8E	1	1
1PWM	1	1
1UI0 chain A	1	1
1UI0 chain B	7	**5**
1F5F	1	1
1M2X	2	**1**
1YKI	11	**1**
1YNH	1	1
2B0M	1	1
2RDR	1	1
2Q6M	1	1
2BL9	1	1
2AIE	1	1
2BKX	1	1
1M2X	2	**1**
2J5S	1	1
1OFZ	1	1
1WOG	16+	**1**
2GG7	1	1
2ZVJ	3	**1**
1YV5	1	1
2FF2	2	1
Both Fail		
1S5N	15	**5**
1Y2K	16+	16+
2CIX	5	**4**
Initial Glide		
1M3U	**1**	15+
1MLW	**5**	8
1FSG chain C	1	1

1R5Y	1	1
1SQN	1	1
1UWC	1	1
2BRT	1	1
1W1A	1	1
2FDV	1	1
2UY5	1	1
2I5X	1	1
2GVV	1	1
2CIX (redock)	2	2
3DSX	4	2

Table 2.3: Table of atomic charges before and after polarization with associated atom for ligands in 1W0G, 1YV5 and 2GG7.

	Before	After			Before	After			Before	After	
Protein	polarization	Polarization	Atom	Protein	polarization	Polarization	Atom	Protein	polarization	Polarization	Atom
1W0G	-0.90000	-0.73233	N	1YV5	-1.11997	-0.72225	O	2GG7	-0.80000	0.71130	O
	0.06000	0.50270	C		1.55780	0.41361	P		0.71500	0.71215	C
	-0.12000	-0.41183	C		-1.11997	-1.14608	O		-0.80000	-0.80497	O
	-0.12000	0.28799	C		-1.11997	-0.81983	O		-0.11500	-0.03869	C
	-0.12000	-0.19768	C		-0.13080	1.62935	C		-0.11500	-0.32094	C
	-0.08000	-0.01248	C		-0.68300	-1.12335	O		-0.11500	-0.02051	C
	0.31340	0.25927	N		1.55780	0.49767	P		-0.11500	-0.44919	C
	-0.67240	-1.00933	H		-1.11997	-0.63453	O		-0.11500	-0.15878	C
	0.36000	0.52811	H		-1.11997	-1.01376	O		0.03800	0.35427	C
	0.36000	0.46137	H		-1.11997	-0.89681	O		-0.03800	-0.29951	N
	0.06000	0.08927	H		0.04300	-0.71285	C		-0.03800	-0.24570	N
	0.06000	-0.05856	H		-0.45500	0.23031	C		-0.13920	0.11339	C
	0.06000	0.05183	H		0.22700	-0.01071	C		0.71000	0.20783	C
	0.06000	0.15589	H		-0.44700	-0.48043	C		-0.92000	-0.76951	N
	0.06000	-0.16328	H		0.47300	0.33497	C		-0.14900	-0.12707	N
	0.06000	-0.01769	H		-0.67800	-0.82029	N		0.53800	0.43588	C
	0.06000	-0.00763	H		0.47300	0.37390	C		-0.53100	-0.16269	N
	0.06000	-0.06436	H		0.41800	0.63845	H		-0.92000	-1.12764	N
	0.06000	0.02397	H		0.06000	0.11252	H		0.11500	0.05601	H
	0.06000	0.13861	H		0.06000	0.07571	H		0.11500	0.04615	H
	0.40200	-0.10243	H		0.06500	0.17730	H		0.11500	0.12101	H
	0.28270	0.27860	H		0.15500	0.05251	H		0.11500	0.03814	H
					0.01200	-0.12935	H		0.36000	0.48859	H
					0.01200	-0.02606	H		0.36000	0.26504	H
									0.34100	0.49600	H
									0.36000	0.50829	H
									0.36000	0.39378	H

Table 2.3 shows a sample of three of the ligands from the SERAPhiC set before and after polarization with the atom associated with it next to the values. After the polarization for the ligand in 1W0G six of the hydrogen atoms gain a negative charge about as strong as the positive charge they had before. Two of the carbons have larger negative charges than before, -0.1200 increased to -0.41183 and -0.1200 increased -0.19768.

1YV5 and 2GG7 have several large changes in their carbon charges. 1YV5 has a carbon changing from -0.13080 to 1.62935 gaining a much larger positive charge. 2GG7 has a carbon changing from 0.71000 to 0.20783 losing a lot of its positive charge. Several atoms involved in hydrogen bonding also gain a larger negative charge from -0.67800 to -0.82029 in a nitrogen from 1YV5 and -0.92000 to -1.12764 in a nitrogen from 2GG7. The largest changes across the ligands is the change in their polar atoms such as 1.55780 to 0.49767 in P for 1YV5. Across all three the rest of the ligand goes through small subtle charge changes.

2.5 Discussion

GLIDE is a good docking program and is great for docking ligand libraries very quickly. Many reviews highly rate the program such as Warren et al. (2006), Perola et al. (2004), Cross et al. (2009) and Abagyan and Totrov (2001). Making marked improvements to the docking algorithms can be difficult. Despite this, we were able to attain some good improvements using our induced charge polarization. However our basic Glide docks were not perfect. In some cases the Favia et al. (2011) were able to get better results than us for the top docking poses and sometimes we attained better results than them. This could be due to the differences in set up between our methods compared to theirs, especially when it comes to the grid size and protonation sites on histidine. However all these results come from crystal structure so each ligand should ideally bind to the binding site. The negative results show that even this simple experiment of docking a ligand back into its own crystal structure is not trivial. Consequently, we looked for other techniques that could improve the results.

The QM/MM results can show more defined energy interactions of the binding. Since QM energy is potentially better in the sense that the QM energy is well defined by our method and basis set. The MM energy is also well defined as the OPLS force field is well defined for proteins from the literature (Damm et al., 1997).

The methods follow the trend where the docking process tends to improve if a QM/MM region with polarization is generated. However in some cases like 2CIX the RMSD change was 9.28 to 9.24, and so this result is equally wrong by both methods. However in the case of 2CIX we re-did the initial dock several times attaining variable results each time. The last column of table 2.1 shows the best results we attained. Even then however it was still not a remarkable improvement. 2CIX possibly requires something else to help it bind to its site. We left out water from our method so that might be what is causing the issue. Some water is tightly bound in some of the binding sites from the crystal models. These waters that are found there might be help facilitate binding of the ligand (Smith, 2015). We don’t usually want to explicitly add these water molecules into docking as the ligand would displace the water; however there are methods for predicting water molecules in

binding sites and for predicting which one as ‘happy’, and so unlikely to be displaced and which ones are ‘unhappy’ so are likely to be displaced. (Goodford, 1985), (Mason et al., 2013), (Wang et al., 2011).
There are some results with quite a substantial difference in RMSD from the initial dock such as 1M2X with a change of 2.5 Å or 2Q6M with a change of 1.4 Å. This is quite a significant difference; in both cases these represented improvements.

There were several proteins such as 2CIX and 1M2X that were investigated further using MD methods to investigate why there was either a great difference in their RMSD or why it didn’t improve as much as expected.

There were some ligands such as in 1MLW where after polarization it found a new binding site. All the new poses were bound near the same region as evidenced by the graphics of the poses and the RMSDs. During sampling 1MLW could have found a new energy well where it can bind. This could suggest that there is a different position that 1MLW can bind when polarization is present that it wouldn’t have found before.

Further works that could be planned other than what is described in sections 3.8 and 4.2 is to potentially performing tests with actively bound water in a few of the ligands that showed higher RMSDs. In some cases, as stated before, there is water tightly bound in an active site of a protein that facilitates the binding of the ligand or forces the ligand to take the correct binding site. All the tests in this section were performed with no water present. The benefit to this is that the ligand can dock to other binding sites so we can analyze if the binding site in the crystal is one of the better sites. However this also eschews the fact that water is in the binding site in some cases and that this could explain the abnormal RMSDs in some of the proteins such as 1M3U.

The avenue we followed to help deal with some of the problems we found was to initiate MD simulations on our polarized ligands. Some of the problems might have been coming from the absence of water as stated before. However, when docking, if the crystal waters are left in there is a hole left behind by the missing ligand. When the dock is then performed with a large grid, as here, it will funnel the results into the open space left by the waters as the waters are frozen in place for the calculation. In molecular dynamics however the water when explicit moves around and can either displace the ligand or be displaced. This can help more visibly show the binding of the ligands. Molecular dynamics can also show if our well docked poses hold up in a fully moving water environment. The water has enough energy to be able to push the ligand. Thus, if the ligand is bound as well as the results show it should stay in place. However, we may possibly see that some of the ligands might be pushed away, thus showing that the binding strength is not as strong as we first thought. MD might also show if ligands bound in incorrect poses can find their binding site if they given enough time.

2.6 References

ABAGYAN, R. & TOTROV, M. 2001. High-throughput docking for lead generation. *Current opinion in chemical biology,* 5**,** 375-382.

AHN, H. J., KIM, K. H., LEE, J., HA, J.-Y., LEE, H. H., KIM, D., YOON, H.-J., KWON, A.-R. & SUH, S. W. 2004. Crystal structure of agmatinase reveals structural conservation and inhibition mechanism of the ureohydrolase superfamily. *Journal of Biological Chemistry,* 279**,** 50505-50513.

AVVAKUMOV, G. V., MULLER, Y. A. & HAMMOND, G. L. 2000. Steroid-binding specificity of human sex hormone-binding globulin is influenced by occupancy of a zinc-binding site. *Journal of Biological Chemistry,* 275**,** 25920-25925.

BAKER, C. M. 2015. Polarizable force fields for molecular dynamics simulations of biomolecules. *Wiley Interdisciplinary Reviews: Computational Molecular Science,* 5**,** 241-254.

BAKER, C. M. & GRANT, G. H. 2006. The structure of liquid benzene. *Journal of Chemical Theory and Computation,* 2**,** 947-955.

BAKOWIES, D. & THIEL, W. 1996. Hybrid models for combined quantum mechanical and molecular mechanical approaches. *The Journal of Physical Chemistry,* 100**,** 10580-10594.

BENNETT, J. P., WHITTINGHAM, J. L., BRZOZOWSKI, A. M., LEONARD, P. M. & GROGAN, G. 2007. Structural characterization of a β-diketone hydrolase from the cyanobacterium Anabaena sp. PCC 7120 in native and product-bound forms, a coenzyme A-independent member of the crotonase suprafamily. *Biochemistry,* 46**,** 137-144.

BLAIR, D. E. & VAN AALTEN, D. M. 2004. Structures of Bacillus subtilis PdaA, a family 4 carbohydrate esterase, and a complex with N-acetyl-glucosamine. *FEBS letters,* 570**,** 13-19.

BLUM, M.-M., LÖHR, F., RICHARDT, A., RÜTERJANS, H. & CHEN, J. C.-H. 2006. Binding of a designed substrate analogue to diisopropyl fluorophosphatase: implications for the phosphotriesterase mechanism. *Journal of the American Chemical Society,* 128**,** 12750-12757.

BRENK, R., MEYER, E., REUTER, K., STUBBS, M. T., GARCIA, G. A., DIEDERICH, F. & KLEBE, G. 2004. Crystallographic study of inhibitors of tRNA-guanine transglycosylase suggests a new structure-based pharmacophore for virtual screening. *Journal of molecular biology,* 338**,** 55-75.

CARD, G. L., BLASDEL, L., ENGLAND, B. P., ZHANG, C., SUZUKI, Y., GILLETTE, S., FONG, D., IBRAHIM, P. N., ARTIS, D. R. & BOLLAG, G. 2005. A family of phosphodiesterase inhibitors discovered by cocrystallography and scaffold-based drug design. *Nature biotechnology,* 23**,** 201-207.

CARLSON, H. A. 2002. Protein flexibility is an important component of structure-based drug discovery. *Current pharmaceutical design,* 8**,** 1571-1578.

CAVASOTTO, C. N., KOVACS, J. A. & ABAGYAN, R. A. 2005. Representing receptor flexibility in ligand docking through relevant normal modes. *Journal of the American Chemical Society,* 127**,** 9632-9640.

CORNAH, B. E. 2013. *Influence of Molecular Polarizability on the Active Layer of Organic Photovoltaic Cells: A case study of Poly (3-hexylthiophene).* African University of Science and Technology.

CROSS, J. B., THOMPSON, D. C., RAI, B. K., BABER, J. C., FAN, K. Y., HU, Y. & HUMBLET, C. 2009. Comparison of several molecular docking programs: pose prediction and virtual screening accuracy. *Journal of chemical information and modeling,* 49**,** 1455-1474.

DAMM, W., FRONTERA, A., TIRADO–RIVES, J. & JORGENSEN, W. L. 1997. OPLS all-atom force field for carbohydrates. *Journal of Computational Chemistry,* 18**,** 1955-1970.

DURRANT, J. D. & MCCAMMON, J. A. 2011. BINANA: A novel algorithm for ligand-binding characterization. *Journal of Molecular Graphics and Modelling,* 29**,** 888-893.

ECHT, S., BAUER, S., STEINBACHER, S., HUBER, R., BACHER, A. & FISCHER, M. 2004. Potential anti-infective targets in pathogenic yeasts: structure and properties of 3, 4-dihydroxy-2-butanone 4-phosphate synthase of Candida albicans. *Journal of molecular biology,* 341**,** 1085-1096.

EL-KABBANI, O., DARMANIN, C., SCHNEIDER, T. R., HAZEMANN, I., RUIZ, F., OKA, M., JOACHIMIAK, A., SCHULZE-BRIESE, C., TOMIZAKI, T. & MITSCHLER, A. 2004. Ultrahigh resolution drug design. II. Atomic resolution structures of human aldose reductase holoenzyme complexed with Fidarestat and Minalrestat: implications for the binding of cyclic imide inhibitors. *Proteins: Structure, Function, and Bioinformatics,* 55**,** 805-813.

ELDRIDGE, M. D., MURRAY, C. W., AUTON, T. R., PAOLINI, G. V. & MEE, R. P. 1997. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *Journal of computer-aided molecular design,* 11**,** 425-445.

EVDOKIMOV, A. G., POKROSS, M., WALTER, R., MEKEL, M., COX, B., LI, C., BECHARD, R., GENBAUFFE, F., ANDREWS, R. & DIVEN, C. 2006. Engineering the catalytic domain of human protein tyrosine phosphatase β for structure-based drug discovery. *Acta Crystallographica Section D: Biological Crystallography,* 62**,** 1435-1445.

EVDOKIMOV, A. G., POKROSS, M., WALTER, R. L., MEKEL, M., BARNETT, B. L., AMBURGEY, J., SEIBEL, W. L., SOPER, S. J., DJUNG, J. F. & FAIRWEATHER, N. 2007. Serendipitous discovery of novel bacterial methionine aminopeptidase inhibitors. *PROTEINS: Structure, Function, and Bioinformatics,* 66**,** 538-546.

FAVIA, A. D., BOTTEGONI, G., NOBELI, I., BISIGNANO, P. & CAVALLI, A. 2011. SERAPhiC: A benchmark for in silico fragment-based drug design. *Journal of chemical information and modeling,* 51**,** 2882-2896.

FENN, T. D., RINGE, D. & PETSKO, G. A. 2004. Xylose isomerase in substrate and inhibitor michaelis states: atomic resolution studies of a metal-mediated hydride shift. *Biochemistry,* 43**,** 6464-6474.

FERENCZY, G. G. & REYNOLDS, C. A. 2001. Modeling polarization through induced atomic charges. *The Journal of Physical Chemistry A,* 105**,** 11470-11479.

FEYNMAN, R. P. & HIBBS, A. R. 1965. *Quantum mechanics and path integrals*, McGraw-Hill New York.

FRIEDRICH, B. & HERSCHBACH, D. 1999. Enhanced orientation of polar molecules by combined electrostatic and nonresonant induced dipole forces. *Journal of Chemical Physics,* 111.

FRIESNER, R. A., BANKS, J. L., MURPHY, R. B., HALGREN, T. A., KLICIC, J. J., MAINZ, D. T., REPASKY, M. P., KNOLL, E. H., SHELLEY, M. & PERRY, J. K. 2004. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of medicinal chemistry,* 47**,** 1739-1749.

FRIESNER, R. A., MURPHY, R. B., REPASKY, M. P., FRYE, L. L., GREENWOOD, J. R., HALGREN, T. A., SANSCHAGRIN, P. C. & MAINZ, D. T. 2006. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *Journal of medicinal chemistry,* 49**,** 6177-6196.

GARCÍA-SÁEZ, I., HOPKINS, J., PAPAMICAEL, C., FRANCESCHINI, N., AMICOSANTE, G., ROSSOLINI, G. M., GALLENI, M., FRÈRE, J.-M. & DIDEBERG, O. 2003. The 1.5-Å structure of Chryseobacterium meningosepticum zinc β-lactamase in complex with the inhibitor, D-captopril. *Journal of Biological Chemistry,* 278**,** 23868-23873.

GOODFORD, P. J. 1985. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of medicinal chemistry,* 28**,** 849-857.

GUO, Z., WU, Y. W., DAS, D., DELON, C., CRAMER, J., YU, S., THUNS, S., LUPILOVA, N., WALDMANN, H. & BRUNSVELD, L. 2008. Structures of RabGGTase–substrate/product complexes provide insights into the evolution of protein prenylation. *The EMBO journal,* 27**,** 2444-2456.

HALGREN, T. A., MURPHY, R. B., FRIESNER, R. A., BEARD, H. S., FRYE, L. L., POLLARD, W. T. & BANKS, J. L. 2004. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *Journal of medicinal chemistry,* 47**,** 1750-1759.

HALPERIN, I., MA, B., WOLFSON, H. & NUSSINOV, R. 2002. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Structure, Function, and Bioinformatics,* 47**,** 409-443.

HÉROUX, A., WHITE, E. L., ROSS, L. J., KUZIN, A. P. & BORHANI, D. W. 2000. Substrate deformation in a hypoxanthine-guanine phosphoribosyltransferase ternary complex: the structural basis for catalysis. *Structure,* 8**,** 1309-1318.

HURT, D. E., SUTTON, A. E. & CLARDY, J. 2006. Brequinar derivatives and species-specific drug design for dihydroorotate dehydrogenase. *Bioorganic & medicinal chemistry letters,* 16**,** 1610-1615.

HURTADO-GUERRERO, R. & VAN AALTEN, D. M. 2007. Structure of Saccharomyces cerevisiae chitinase 1 and screening-based discovery of potent inhibitors. *Chemistry & biology,* 14**,** 589-599.

ILLINGWORTH, C. J., MORRIS, G. M., PARKES, K. E., SNELL, C. R. & REYNOLDS, C. A. 2008. Assessing the role of polarization in docking. *The Journal of Physical Chemistry A,* 112**,** 12157-12163.

JAIN, A. N. 2003. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *Journal of medicinal chemistry,* 46**,** 499-511.

JAIN, A. N. 2006. Scoring functions for protein-ligand docking. *Current Protein and Peptide Science,* 7**,** 407-420.

JONES, G., WILLETT, P., GLEN, R. C., LEACH, A. R. & TAYLOR, R. 1997. Development and validation of a genetic algorithm for flexible docking. *Journal of molecular biology,* 267**,** 727-748.

JØRGENSEN, R., PURDY, A. E., FIELDHOUSE, R. J., KIMBER, M. S., BARTLETT, D. H. & MERRILL, A. R. 2008. Cholix toxin, a novel ADP-ribosylating factor from Vibrio cholerae. *Journal of Biological Chemistry,* 283**,** 10671-10678.

JORGENSEN, W. L. & TIRADO-RIVES, J. 1988. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society,* 110**,** 1657-1666.

KAMINSKI, G. A., FRIESNER, R. A., TIRADO-RIVES, J. & JORGENSEN, W. L. 2001. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *The Journal of Physical Chemistry B,* 105**,** 6474-6487.

KAVANAGH, K. L., GUO, K., DUNFORD, J. E., WU, X., KNAPP, S., EBETINO, F. H., ROGERS, M. J., RUSSELL, R. G. G. & OPPERMANN, U. 2006. The molecular mechanism of nitrogen-containing bisphosphonates as antiosteoporosis drugs. *Proceedings of the National Academy of Sciences,* 103**,** 7829-7834.

KONGSAEREE, P., KHONGSUK, P., LEARTSAKULPANICH, U., CHITNUMSUB, P., TARNCHOMPOO, B., WALKINSHAW, M. D. & YUTHAVONG, Y. 2005. Crystal structure of dihydrofolate reductase from Plasmodium vivax: pyrimethamine displacement linked with mutation-induced resistance. *Proceedings of the National Academy of Sciences of the United States of America,* 102**,** 13046-13051.

KÜHNEL, K., BLANKENFELDT, W., TERNER, J. & SCHLICHTING, I. 2006. Crystal structures of chloroperoxidase with its bound substrates and complexed with formate, acetate, and nitrate. *Journal of Biological Chemistry,* 281**,** 23990-23998.

LAM, P., JADHAV, P., EYERMANN, C. J., HODGE, C. N., RU, Y., BACHELER, L. T., MEEK, J. L., OTTO, M. J., RAYNER, M. M. & WONG, Y. N. 1994. Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. *Science,* 263**,** 380-384.

LIN, H. & TRUHLAR, D. G. 2007. QM/MM: what have we learned, where are we, and where do we go from here? *Theoretical Chemistry Accounts,* 117**,** 185-199.

LOPES, P. E., LAMOUREUX, G. & MACKERELL, A. D. 2009. Polarizable empirical force field for nitrogen-containing heteroaromatic compounds based on the classical Drude oscillator. *Journal of computational chemistry,* 30**,** 1821-1838.

LORBER, D. M. & SHOICHET, B. K. 1998. Flexible ligand docking using conformational ensembles. *Protein Science,* 7**,** 938-950.

MADAUSS, K. P., DENG, S.-J., AUSTIN, R. J., LAMBERT, M. H., MCLAY, I., PRITCHARD, J., SHORT, S. A., STEWART, E. L., UINGS, I. J. & WILLIAMS, S. P. 2004. Progesterone receptor ligand binding pocket flexibility: crystal structures of the norethindrone and mometasone furoate complexes. *Journal of medicinal chemistry,* 47**,** 3381-3387.

MAPLE, J. R., CAO, Y., DAMM, W., HALGREN, T. A., KAMINSKI, G. A., ZHANG, L. Y. & FRIESNER, R. A. 2005. A polarizable force field and continuum solvation methodology for modeling of protein-ligand interactions. *Journal of Chemical Theory and Computation,* 1**,** 694-715.

MARCYJANIAK, M., ODINTSOV, S. G., SABALA, I. & BOCHTLER, M. 2004. Peptidoglycan amidase MepA is a LAS metallopeptidase. *Journal of Biological Chemistry,* 279**,** 43982-43989.

MASON, J. S., BORTOLATO, A., WEISS, D. R., DEFLORIAN, F., TEHAN, B. & MARSHALL, F. H. 2013. High end GPCR design: crafted ligand design and druggability analysis using protein structure, lipophilic hotspots and explicit water networks. *Silico Pharmacol,* 1**,** 23.

MCAULEY, K. E., SVENDSEN, A., PATKAR, S. A. & WILSON, K. S. 2004. Structure of a feruloyl esterase from Aspergillus niger. *Acta Crystallographica Section D: Biological Crystallography,* 60**,** 878-887.

MCGANN, M. 2011. FRED pose prediction and virtual screening accuracy. *Journal of chemical information and modeling,* 51**,** 578-596.

MCMARTIN, C. & BOHACEK, R. S. 1997. QXP: powerful, rapid computer algorithms for structure-based drug design. *Journal of computer-aided molecular design,* 11**,** 333-344.

PEROLA, E., WALTERS, W. P. & CHARIFSON, P. S. 2004. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins: Structure, Function, and Bioinformatics,* 56**,** 235-249.

RACE, P. R., LOVERING, A. L., GREEN, R. M., OSSOR, A., WHITE, S. A., SEARLE, P. F., WRIGHTON, C. J. & HYDE, E. I. 2005. Structural and Mechanistic Studies of Escherichia coli Nitroreductase with the Antibiotic Nitrofurazone REVERSED BINDING ORIENTATIONS IN DIFFERENT REDOX STATES OF THE ENZYME. *Journal of Biological Chemistry,* 280**,** 13256-13264.

RAPPE, A. K. & GODDARD III, W. A. 1991. Charge equilibration for molecular dynamics simulations. *The Journal of Physical Chemistry,* 95**,** 3358-3363.

REPASKY, M. P., SHELLEY, M. & FRIESNER, R. A. 2007. Flexible ligand docking with Glide. *Current Protocols in Bioinformatics***,** 8.12. 1-8.12. 36.

ROBERTSON, M. J., TIRADO-RIVES, J. & JORGENSEN, W. L. 2015. Improved Peptide and Protein Torsional Energetics with the OPLS-AA Force Field. *Journal of chemical theory and computation,* 11**,** 3499-3509.

SCHELLHAMMER, I. & RAREY, M. 2004. FlexX-Scan: Fast, structure-based virtual screening. *PROTEINS: Structure, Function, and Bioinformatics,* 57**,** 504-517.

SCHRÖDINGER 2015. Schrödinger Release 2015-3: Maestro, version 10.3. LLC, New York, NY.

SIU, S. W., PLUHACKOVA, K. & BÖCKMANN, R. A. 2012. Optimization of the OPLS-AA force field for long hydrocarbons. *Journal of Chemical Theory and Computation,* 8**,** 1459-1470.

SMITH, B. J., COLMAN, P. M., VON ITZSTEIN, M., DANYLEC, B. & VARGHESE, J. N. 2001. Analysis of inhibitor binding in influenza virus neuraminidase. *Protein Science,* 10**,** 689-696.

SMITH, K. 2015. *RE: PhD thesis.*

SMITH, K. J., PETIT, C. M., AUBART, K., SMYTH, M., MCMANUS, E., JONES, J., FOSBERRY, A., LEWIS, C., LONETTO, M. & CHRISTENSEN, S. B. 2003. Structural variation and inhibitor binding in polypeptide deformylase from four different bacterial species. *Protein science,* 12**,** 349-360.

TOCILJ, A., SCHRAG, J. D., LI, Y., SCHNEIDER, B. L., REITZER, L., MATTE, A. & CYGLER, M. 2005. Crystal structure of N-succinylarginine dihydrolase AstB, bound to substrate and product, an enzyme from the arginine catabolic pathway of Escherichia coli. *Journal of Biological Chemistry,* 280**,** 15800-15808.

TOTROV, M. & ABAGYAN, R. 2001. Protein-ligand docking as an energy optimization problem. *Drug-receptor thermodynamics: Introduction and applications,* 1**,** 603-624.

TROTT, O. & OLSON, A. J. 2010. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry,* 31**,** 455-461.

TSUJI, E., OKAZAKI, K. & TAKEDA, K. 2009. Crystal structures of rat catechol-O-methyltransferase complexed with coumarine-based inhibitor. *Biochemical and biophysical research communications,* 378**,** 494-497.

VENKATACHALAM, C. M., JIANG, X., OLDFIELD, T. & WALDMAN, M. 2003. LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *Journal of Molecular Graphics and Modelling,* 21**,** 289-307.

VERDONK, M. L., CHESSARI, G., COLE, J. C., HARTSHORN, M. J., MURRAY, C. W., NISSINK, J. W. M., TAYLOR, R. D. & TAYLOR, R. 2005. Modeling water molecules in protein-ligand docking using GOLD. *Journal of medicinal chemistry,* 48**,** 6504-6515.

VERSÉES, W., BARLOW, J. & STEYAERT, J. 2006. Transition-state complex of the purine-specific nucleoside hydrolase of T. vivax: enzyme conformational changes and implications for catalysis. *Journal of molecular biology,* 359**,** 331-346.

VINCENT, F., DAVIES, G. J. & BRANNIGAN, J. A. 2005. Structure and Kinetics of a Monomeric Glucosamine 6-Phosphate Deaminase MISSING LINK OF THE NagB SUPERFAMILY? *Journal of Biological Chemistry,* 280**,** 19649-19655.

VON DELFT, F., INOUE, T., SALDANHA, S. A., OTTENHOF, H. H., SCHMITZBERGER, F., BIRCH, L. M., DHANARAJ, V., WITTY, M., SMITH, A. G. & BLUNDELL, T. L. 2003. Structure of E. coli ketopantoate hydroxymethyl transferase complexed with ketopantoate and Mg 2+, solved by locating 160 selenomethionine sites. *Structure,* 11**,** 985-996.

WANG, L., BERNE, B. & FRIESNER, R. 2011. Ligand binding to protein-binding pockets with wet and dry regions. *Proceedings of the National Academy of Sciences,* 108**,** 1326-1330.

WANG, L., ERLANDSEN, H., HAAVIK, J., KNAPPSKOG, P. M. & STEVENS, R. C. 2002. Three-dimensional structure of human tryptophan hydroxylase and its implications for the biosynthesis of the neurotransmitters serotonin and melatonin. *Biochemistry,* 41**,** 12569-12574.

WARREN, G. L., ANDREWS, C. W., CAPELLI, A.-M., CLARKE, B., LALONDE, J., LAMBERT, M. H., LINDVALL, M., NEVINS, N., SEMUS, S. F. & SENGER, S. 2006. A critical assessment of docking programs and scoring functions. *Journal of medicinal chemistry,* 49**,** 5912-5931.

WELFORD, R. W., CLIFTON, I. J., TURNBULL, J. J., WILSON, S. C. & SCHOFIELD, C. J. 2005. Structural and mechanistic studies on anthocyanidin synthase catalysed oxidation of flavanone substrates: the effect of C-2 stereochemistry on product selectivity and mechanism. *Organic & biomolecular chemistry,* 3**,** 3117-3126.

WIMMEROVA, M., MITCHELL, E., SANCHEZ, J.-F., GAUTIER, C. & IMBERTY, A. 2003. Crystal Structure of Fungal Lectin SIX-BLADED β-PROPELLER FOLD AND NOVEL FUCOSE RECOGNITION MODE FOR ALEURIA AURANTIA LECTIN. *Journal of Biological Chemistry,* 278**,** 27059-27067.

WINN, P. J., FERENCZY, G. G. & REYNOLDS, C. A. 1997. Toward improved force fields. 1. Multipole-derived atomic charges. *The Journal of Physical Chemistry A,* 101**,** 5437-5445.

YANO, J. K., DENTON, T. T., CERNY, M. A., ZHANG, X., JOHNSON, E. F. & CASHMAN, J. R. 2006. Synthetic inhibitors of cytochrome P-450 2A6: inhibitory activity, difference spectra, mechanism of inhibition, and protein cocrystallization. *Journal of medicinal chemistry,* 49**,** 6987-7001.

YOU, Z., OMURA, S., IKEDA, H., CANE, D. E. & JOGL, G. 2007. Crystal structure of the non-heme iron dioxygenase PtlH in pentalenolactone biosynthesis. *Journal of Biological Chemistry,* 282**,** 36552-36560.

Chapter 3

Investigating GPUs in Molecular Dynamics Simulation

3.1 Introduction

Molecular Dynamics (MD) is a useful tool for understanding the physical bases of structures and their function. Molecular dynamics at its core is a computer simulation of a system to represent the physical movement of atoms over time within given dimensions and according to Newton's laws. MD can be used to calculate the interactions between these particles in motion; the trajectories of these atoms are determined by numerically solving Newton's equations of motion. There are multiple programs that use various algorithms to solve the equations, such as ACEMD (Harvey et al., 2009), AMBER (Case et al., 2015), CHARMM (Karplus, 1983), GROMACS (Berendsen et al., 1995), NAMD (Phillips et al., 2005), and Tinker (Ponder, 2004).

At the initial stage of the project we utilised GROMACS as it was open-source. GROMACS' main advantage was the speed of calculation, and in recent years has had added modules that utilise GPUs. This was the most cost effective, source code was available and also it did not need a proprietary (i.e. commercially purchased) GUI to run. Therefore any issues could be solved by altering the code in house. Moreover, GROMACS has a community through which newly added code is constantly added to a database providing new functions and modules to the program. Due to this community, many guides and help are available to users.

3.1.1 Boltzmann sampling in MD

MD in modern usage concerns large macromolecules, due to their size there is a considerable amount of degrees of freedom in any statistical simulation. There needs to be a compact description of the thermodynamic properties of a system. In the context of Boltzmann sampling of phase space a simulation might lodged on one side of a high energy gradient. Even if a simulation was run for several ms the simulation could remain on one side of an energy gradient and be unable to sample the other side, this is one of the shortcomings of MD. There have been methods to adapt for this issue such as blue-moon sampling where part of the system is constrained. The part of the system is changed for each run, essentially decomposing the free energy gradients into separate components to test and later averaging out the sampling.

3.2 Molecular Mechanics (MM) and the MM force field

Classical mechanics is used to model molecular systems in molecular mechanics-based approaches. The main aim in molecular mechanics is to define the energy within a molecule. It is possible to use this information to adjust the energy by changing the bond lengths and angles to minimise a structure.

Each atom in MM is defined as a single particle with an assigned net charge, radius and polarizability. As shown in Hehre (2003), bonded interactions are considered as a set of springs, as the interactions include the stretching and compressing of bonds in motion. Conventionally in MM this is called the steric energy. This is based upon the equation 3.1, taken from Hehre (2003).

$$E_{\text{steric energy}} = E_{\text{str}} + E_{\text{bend}} + E_{\text{str-bend}} + E_{\text{oop}} + E_{\text{tor}} + E_{\text{VdW}} + E_{\text{qq}} \qquad (3.1)$$

This includes the stretching (E_{str}) and bending of bonds past their equilibrium (E_{bend}). It also includes the bonds out of plane movement (E_{oop}), torsion interactions (E_{tor}), Van der Waals (E_{VdW}) and electrostatic interactions (E_{qq}). Figure 3.1, shows the direction of each of the interactions.

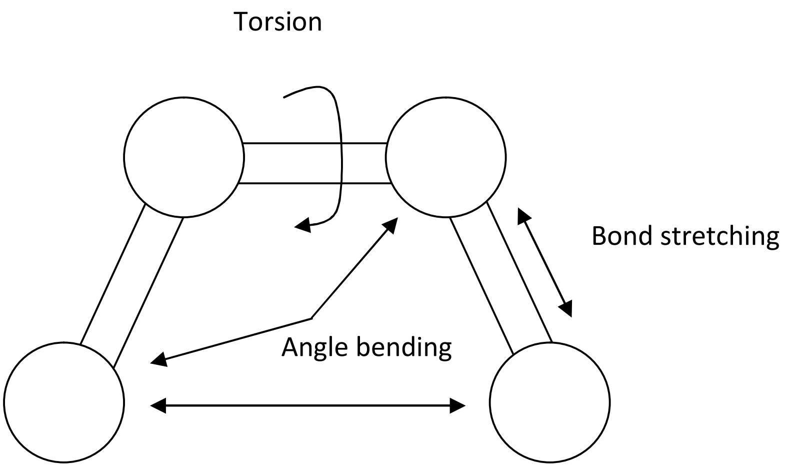


Figure 3.1. Anatomy of MM force field interactions

The bonding portions of the equation are based on Hooke's Law for a spring. The non-bonded interactions are the VdW (E_{VdW}) and electrostatic interactions (E_{qq}). Non-bonded interactions are between atoms that are more than two bonds apart as shown in Figure 3.1. The 1-4 interactions usually take the same form as those shown in Figure 3.2, but may have different scaling factors in from of the terms; these 1-4 interactions contribute to the torsional energy.

$$E = \underbrace{\sum_i \sum_j \frac{-A_{ij}}{{r_{ij}}^6} + \frac{B_{ij}}{{r_{ij}}^{12}}}_{\text{VdW term}} + \underbrace{\sum_i \sum_j \frac{c\, q_i q_j}{4\pi\varepsilon_r\, r_{ij}}}_{\text{Electrostatic term}} \qquad (3.2)$$

Equation 3.2 shows the expanded VdW term and the Electrostatic term taken from Eliav (2008). VdW attraction occurs when the atoms are close, if they are more than a few Angstroms apart this force rapidly diminishes. The VdW forces are expressed as a Lennard-Jones potential. The term r_{ij} refers to the distance between the two nuclei. A and B are constant values based upon the atom identity. The A and B parameters can be calculated by QM and can also be obtained from atomic polarizabilities or by fitting to experimental data, e.g. crystallographic data. When expressed on a plot, the energy, E, can be varies with distance as shown in Figure 3.2. The first two terms in equation 3.2 can also be expressed as shown in equation 3.3, giving a relationship between the A and B parameters:

$$E = 4\,\varepsilon[(\sigma/r)^{12} - (\sigma/r)^{6}] \qquad (3.3)$$

where σ = r at E = 0 and ε is the depth of the energy minima.

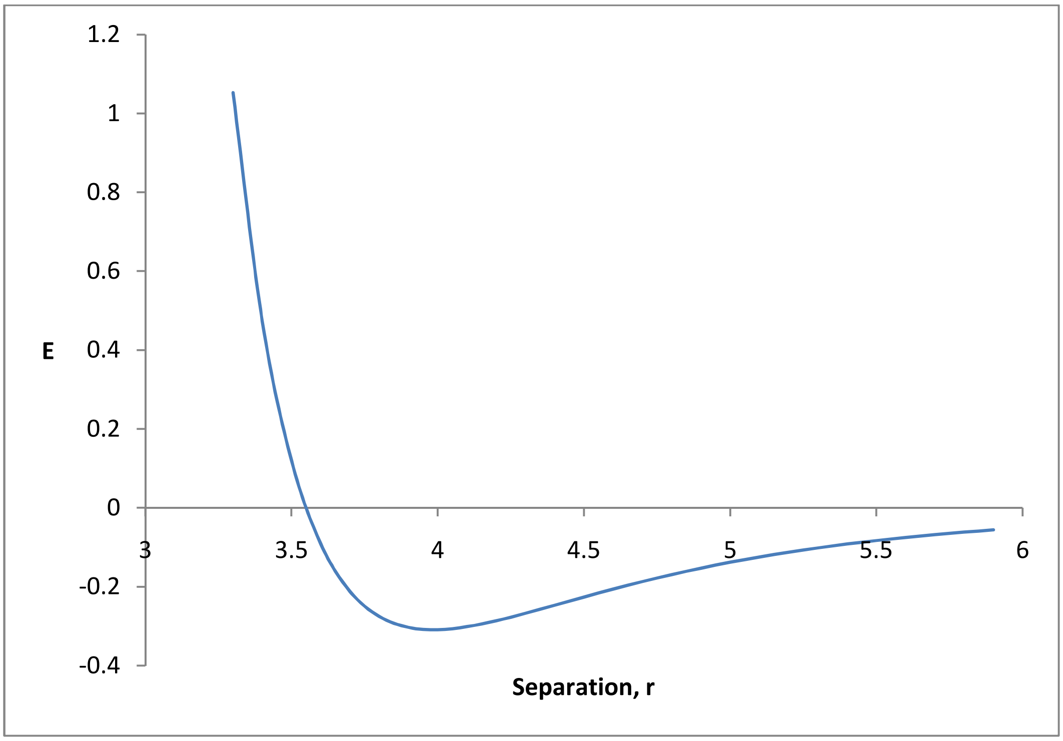


Figure 3.2. Plot of Lennerd-Jones potential between two carbons

Summarized, figure 3.2 shows that when the atoms are close together they repel each other with a large force, as shown in the top left section. Once they are an optimum distance apart the energy is lower and in an ideal state shown by the trough in the middle of the figure. As the atoms move further away, the less attractive force there is between them.

The electrostatic interactions are based on Coloumbic potential, shown in equation 3.1. This is a function of the charge based on their distance, once again defined as r_{ij} and each atom's partial charge defined as q_i. ε_r is the relative dielectric constant, this is usually set to 1 in the gas phase or if all particles are present. The partial charges can be calculated by high level quantum mechanical calculations on small molecules or peptide fragments, as shown in Shattuck (2008) and Shirts et al. (2003). Some programs will assign charges using rules or templates based on previous literature, especially for macromolecules.

The steric energy of each bond pair is the function of a force field. All the constants in the previous equations 3.1, 3.2 and 3.3 can be calculated by QM. However in practice it is necessary to derive them empirically. Usually this takes the form of all-atom models where the force field terms for all atoms are parameterized. This allows for the calculation of the energy function of more complex molecules such as proteins.

Non-bonded interactions are typically as shown in figure 3.1. These are longer range interactions than the bonding interactions and take up the majority of the computational time. The number of non-bonded interactions increases the larger the molecule. To compensate for this increase in computational cost, many programs use cut-offs, as shown in Plimpton (1995). This is where when calculating the force field energy terms, a program will cease to calculate non-bonding interactions beyond a given cut-off point. For example the typical cut-off length for VdW forces is about 10 Å.

3.3 Equations of motion in Molecular Dynamics

The classical MD simulations require numerical integration of Newton's equations of motions for the particles in a system. According to Rapaport (1995), the forces are determined from derivatives of the potential functions. The potential energy is a function of the atomic positions of all atoms in a system according to Stote et al. (1999); however these must be solved numerically as there is no analytical solution to the equations of motion.

Molecular dynamics uses different algorithms to integrate the equations of motions into molecular dynamics. When considering an algorithm the following criteria should be adhered to; the algorithm should conserve energy and momentum, compute efficiently and should permit a long time step for integration. Computing efficiently is one of the most important parts of any algorithm as evaluation of these forces in motion is the most time consuming component of any MD calculation, as shown in Stote et al. (1999).

Some common algorithms include the leap-frog and Verlet methods which are low order methods. These are easier to implement and are more stable than predictorcorrector methods and take far less memory requirements on computers to be used as shown in Rapaport (1995) and Berendsen and Van Gunsteren (1984).

All integration algorithms assume the positions, velocities and accelerations by a Taylor series expansion according to Stote et al. (1999). This can be shown as in equation 3.4 also shown in Stote et al. (1999).

$$r(t+\delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 + \ldots$$

$$v(t+\delta t) = v(t) + a(t)\delta t + \frac{1}{2}b(t)\delta t^2 + \ldots$$

$$a(t+\delta t) = a(t) + b(t)\delta t + \ldots \qquad (3.4)$$

Where ***r*** is the position of a particle, ***v*** is the velocity, ***a*** is the acceleration, ***b*** is the second derivative of v(t) with respect to t. Each of the low order methods is derived from the expansions shown in equation 3.4.

3.3.1 Verlet

The Verlet algorithm uses positions and accelerations at a point in time ***t*** and the new positions at time ***t-δt*** to calculate the new positions of the particles at time ***t+δt.*** This is derived from the equation expansion in 3.4.

$$r(t+\delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2$$

$$r(t-\delta t) = r(t) - v(t)\delta t + \frac{1}{2}a(t)\delta t^2 \qquad (3.5)$$

When combining equations 3.5, the resulting equation is 3.6.

$$r(t+\delta t) = 2r(t) - r(t-\delta t) + a(t)\delta t^2 \qquad (3.6)$$

Verlet according to Stote et al. (1999), Rapaport (1995) and Verlet (1980) retain the advantages of being low order but only offer moderate precision.

3.3.2 Leap-frog

Leap-frog algorithm according to Stote et al. (1999) has the velocities first calculated at time ***t+1/2δt***. The velocities are then used to calculate the positions at the time ***t+δt***. In this way the algorithm leaps the velocities over the position, then the position leaps over the velocity hence giving the name leap-frog. The velocities are explicitly calculated in this method however because of the leaping nature of the equation it does not calculate this at the same time as the positions. This is shown in the equation 3.7.

$$v(t) = \frac{1}{2}\left[v\left(t-\frac{1}{2}\delta t\right) + v\left(t+\frac{1}{2}\delta t\right)\right] \qquad (3.7)$$

3.3.3 ‘Velocity’ Verlet

The ‘Velocity’ Verlet algorithm is algebraically equivalent to the leap-frog method according to Rapaport (1995). However it yields positions, velocity and accelerations at a point in time. This method doesn’t compromise precision. The method uses the same equation shown in 3.5 but rearranges it so it is a function of velocity hence the name Velocity Verlet. This is shown in the equation 3.8.

$$v(t+\delta t) = v(t) + \frac{1}{2}\left[a(t) + a(t+\delta t)\right]\delta t \tag{3.8}$$

3.4 Periodic boundary conditions

As seen in Rapaport (1995), de Leeuw et al. (1980) and (Makov and Payne, 1995), when calculating the behaviour of a system, finite systems behave very differently to functionally infinite systems, i.e. a massive system such as cell. Regardless of how big a simulated system is, it is considered to be a lot smaller than a macroscopic system. A simulated system is relatively small within the walls of the system, the computed edges of the box, within which the calculations are taking place. In such a small box the proportion of particles that are near the edge of the system is a lot greater than in an infinite system. This would mean that within the box of the system, the majority of the behaviour would be dominated by surface effects. Electrostatic effects are long-range and so these also require periodic boundary conditions.

To overcome these issues of size and surface effects, MD simulations use periodic boundary conditions. This is where the system that is enclosed in the box is functionally replicated to infinity in all 3 Cartesian directions, filling in the space. If a particle would leave the edge of the boundary conditions instead of rebounding as if within a non-periodic box it will re-enter the box from the opposite edge; an example of this in effect is shown in figure 3.3. This diminishes the surface effects and makes it so the position of the box plays no role in the calculation of the trajectory.

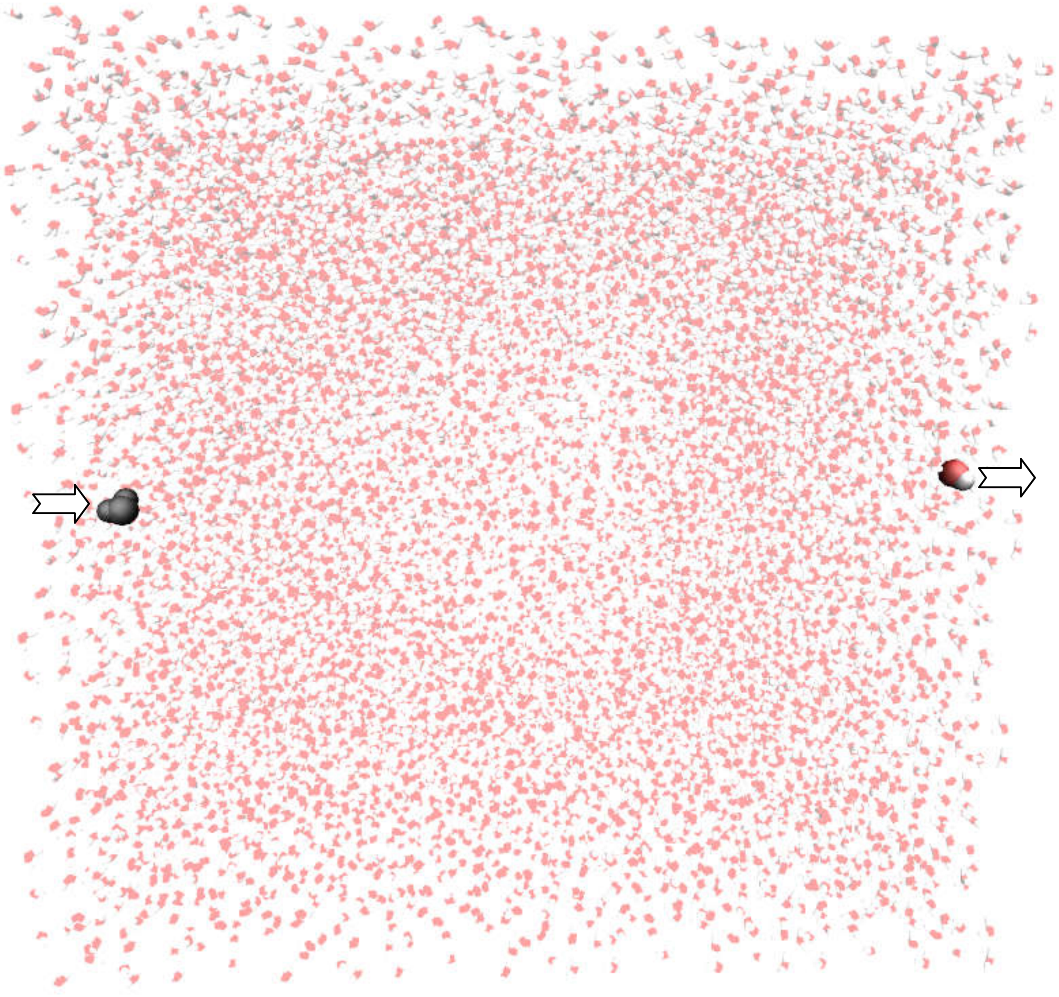


Figure 3.3. An example of water passing through the periodic boundary

3.5 Trajectories

In summation, MD simulations take place using atomic motions governed by classical mechanics. The force fields define the molecular surfaces and the steric bond interactions. This culminates in a trajectory where the process can calculate the dynamics of a system over the course of a relatively long time. The trajectory is a summation of the dynamics for a multi-bodied system, while taking into account full dimensionality of the molecules within.

The trajectories in MD however useMolecular Mechanics in classical mechanics, as shown in equations 3.6, 3.7 and 3.8. However MM as a whole has some well documented problems, shown in Moskowitz et al. (1988), Hu et al. (2003) Allinger et al. (1989), and Rappé et al. (1992). The limitation of these trajectories comes where the quantum mechanical behaviours are largely ignored and there is a singular potential governing the motion. Many methods are attempting to incorporate different quantum mechanical effects to attain more accurate results, such as processes in condensed matter as seen in Webster et al. (1991), treatment of multi-electronic states by Amarouche et al. (1989) and heavy ion collisions shown in (Aichelin, 1991). Quantum mechanical effects such as electronic transitions apply to biological processes such as carrier recombination and photochemistry, as seen in Tully (1990) and Weller et al. (1986).

Within a trajectory, the algorithms can be run in parallel as seen in (Hess et al., 2008) and Hess (2008). This is to improve the speed of calculation. When a computer uses more than one core or node it runs multiple algorithms in parallel. Some calculations such as updating particle position can be parallelized without any communication; however most calculations, such as those shown in equation 3.1 for the force fields, require parallel algorithms. In addition, there are parallel constraint algorithms such as P-LINCs by Hess (2008) which is used by GROMACs to remove the fastest degrees of freedom in the equation. This speeds up the calculation by 2 to 4 times, allowing an increase in the timestep.

The timestep's size is usually constrained by the vibrational motion of atoms in a solid or liquid (Plimpton (1995)). This limits the time scale traditionally to fs, so between each calculation there are a few fs; this is what defines the timestep, e.g. δt in equation 3.8.

MD programs can use rectangular periodic boundary conditions as seen in section 3.4 with a sliding scale of different pressures and temperatures to impose on the system. Molecular interactions can be handled in a number of ways, namely; Coulomb and Lennard-Jones or Buckingham potentials. Using this we can make NVT and NPT ensembles. As shown in McDonald (1972), Wood (1968) and Rushbrooke et al. (1968), NVT ensembles are isothermal and constant volume, so the periodic boundary will not change during a trajectory with this ensemble but as per Boyle's law, pressure has to change so MD programs will change the pressure to compensate. NPT ensembles on the other hand are isobaric and isothermal so during a trajectory the periodic boundary will change in size to compensate for the constant pressure.

3.6 Simulating the Water Bath

GROMACS has two major settings to simulate the water bath surrounding a macromolecule, either implicit or explicit water.

3.6.1 Implicit Solvent Methods

Sometimes known as continuum solvation, this is where a water bath is represented as a continuous medium as opposed to individual molecules of water as in explicit solvation methods. Other methods include coarse-graining, where water is represented as large subcomponents of force fields.

The potential of mean force of the implicit solvent is an approximation of the average behaviour and movement of a large body of liquid; in the case of our docking problems, this is water. In GROMACS, the implicit solvent methods use the Generalized Born model augmented with hydrophobic accessible surface area, abbreviated as GBSA. In the GBSA method the total solvation free energy, $G_{solv,}$ is given as a sum of a solvent-solvent cavity, G_{cav}, a solute-solvent Van der Waals term, G_{VdW}, and a solute-solvent electrostatics polarisation term $G_{pol,}$ as in equation (3.9) shown in P.Koehl and M.Levitt (2002) and Karplus and McCammon (2002).

$$G_{solv} = G_{cav} + G_{VdW} + G_{pol} \tag{3.9}$$

When performing a simulation using implicit solvent method, the macromolecules should ideally be reparameterised appropriately for GBSA force fields; compatible parameters must also be applied to the ligand. However these GBSA parameters found in the implicit parameters file are OPLS force fields that are close to the structure of the ligand though are not the same; as shown in Bjelkmar et al. (2010), they are substitutes based on work by Qiu et al. (1997). Many of the ideal OPLS force fields are not compatible with implicit methods; we took a large amount of time to sift through the available parameterization to find force fields compatible with the chosen end groups of the protein and the ligand. In terms of speed however, the minutiae of parameterizing each ligand and then running the simulation was faster than running an explicit solvent simulation.

The major advantage of this method is that the simulation does not have to calculate the movement of each individual explicit water molecule; this results in considerable savings on computational time.

3.6.2 Explicit Solvent Methods

Explicit solvent relies on using discrete solvent molecules and describing their interactions fully. A water bath for a GROMACs simulation is usually composed of thousands of water molecules. The MD simulation then treats each of these as a separate body and calculates their interactions individually. This increases the computational time considerably (see Chapter 4). For this work, the TIP4P water molecule by Abascal and Vega (2005) was usually used.

3.7 GROMACS

GROMACS stand for GROningen MAchine for Chemical Simulation; it is a parallel message-passing implementation for Molecular Dynamics (MD) for macromolecules in aqueous environments. The program is also able to assign special forces to groups of particles which are of particular interest. As stated in previous chapters, we have an interest in polarisation. We are thus attempting to utilise GROMACS to add in polarisation to a MD environment. GROMACs doesn't handle explicit polarisation as it uses classical force fields such as CHARMM or OPLS as shown in Van Der Spoel et al. (2005) and Yu and van Gunsteren (2005). In our method in section 2.3.4.3 we generated new atomic charges based on the polarization. We will be utilising these atomic charges with the OPLS force field for molecular dynamics.

GROMACS can do these MD simulations through the use of parallelization. Parallelization is the use of multiple computers or cores working in tandem upon the same system to gain a higher throughput for the calculations. GROMACS is using sequential code and multithreading it between each of the cores to use all of them simultaneously. This can greatly increase the throughput of the process by utilising multiple cores to complete the simulation. GROMACS is specialised in the use of CPU architecture and in recent years the developers have created GPU modules; we will touch on this later.

GROMACS was utilized to investigate whether polarized charges facilitated the correct binding of the ligand to the target binding site. Since the binding pocket and ligand were polarized, the positions with the lowest RMSD or top glide score should in theory be bound relatively tightly by the stronger electrostatic forces present during a molecular simulation. As discussed in section 2.2.1 due to how Glidescore calculates poses it can give a better rank to poses with a high RMSD, the top Glide score and lowest RMSD are not always the same. Thus, the aim was to perform dynamic simulations for both of these results for each molecule to see if a correctly polarized ligand is more likely to remain in a true binding site rather than one erroneously selected by GLIDE.

3.8 MD simulations of the protein-ligand complex

For each protein in the polarization set (described in chapter 2.3), both the lowest RMSD and top glide score poses were exported from Maestro and converted into pdb format. Before exporting the structures, the ACE and NME caps on the proteins from chapter 2.3 should be removed and hydrogens added in their place. The ligands bound to each protein were exported in separate files. This is due to GROMACS' inability to convert ligands into OPLS format topology automatically. Instead the ligands were converted to OPLS format topology by using the PRODRG server.

PRODRG is an online service provided by SchuÈttelkopf and Van Aalten (2004). This service can take a description of a small molecule based on its pdb coordinates and generate topologies for use with MD or other programs. In our case, this is GROMACs. However it can be used for other programs such as Autodock (Goodsell et al., 1996), HEX (Ritchie, 2003) or REFMAC (Murshudov et al., 2011).

The proteins were then converted to OPLS format topology by using the pdb2gmx GROMACS utility program which outputs a structure file and a topology file for each protein. OPLS forcefields were chosen for the topology to stay consistent with the Maestro-based portion of the method. TIP4P water (Abascal and Vega, 2005) was used in the box as this water type functions better with OPLS forcefields as shown in Bjelkmar et al. (2010) and van der Spoel and Lindahl (2003) since OPLS-AA force fields were developed almost completely using TIP4P. After this, the ligand structure file was appended to the protein structure file. Several settings in the topology file were manually changed to include the newly appended ligand.

Using genbox, a virtual box containing the protein-ligand complex and a TIP4P water box was generated. Genbox is a GROMACS utility program that randomly inserts water molecules into a set box size until the box is filled. This structured was then minimized. The GROMACS simulation was then performed on the minimized structure.

The typical protocol for each complex had a time step of 2 fs, cut-off of 1.5 Å for VdW forces, the temperature was set to 300 K, pressure was set to the default of 1.01325 bar and each production run was 10 ns. The pH of each simulation was 7 and all the crystal structures resolutions were 2.5 Å or better. A typical input file is shown in appendix 3.1.

All the frames generated from the simulation were then imported into VMD where they were superimposed on top of each other. This was done as the protein could potentially move around the box due to translation of the whole system during the simulation. Consequently, to calculate the RMSD, the structure was required to be relatively stationary. An RMSD analysis tool was then used to calculate the RMSD of the ligand. Root mean square deviation (RMSD) is used as a measurement to structural similarities between two structures, as shown in Maiorov and Crippen (1994). RMSD is a measure of the distance between each atom of two superimposed proteins. For the methods shown here, the RMSD is calculated using two different positions of the protein. The RMSD shows the distance between the atoms of the protein for each frame of the simulation against the crystal structure or the first frame of the simulation.

3.8.1 Problems arising with pdb to maestro format conversion

At the initial stages of the methods shown in 2.3.4.2-2.3.4.3 all the pdb files of proteins were converted into the maestro format (.mae) so that maestro would natively interpret the files for the completion of the method. Simple problems arose through the different use of nomenclature of amino acids between the standard format and maestro. This is especially true for protein caps as each programwe used had a different name for the same cap. However the formatting issues were remedied through use of text editors, where each file was examined and the corresponding amino acids were altered to the maestro format. Without these changes maestro would have excluded those particular amino acid groups from the protein. The lack of a correct structure could cause issues in the docking calculations.

As we altered the amino acid format for maestro, we caused issues in later methods. The files were exported from maestro but they would not be converted accurately back into .pdb format. The files converted from .mae to .pdb files no longer contained their formal charge column that was native to the maestro format. This was remedied by either duplicating each atom in the file, giving it a separate ANISOU row to store the charge data for GROMACS or have a separate charge file, denoted as .chg, that could map the formal charges for ACEMD.

3.8.2 Nomenclature problems and the order of hydrogens or carbons.

The files of the protein-ligand complexes that were extracted from stage 2.3.4.2 of the methods had all the formal charges after polarisation so these were the files that we needed to bring forward to the rest of the methods. As previously stated, there was some nomenclature problem converting from .pdb to .mae. This also happened in reverse for a different reason. Maestro has different labels for the carbons in amino acids, usually the order in which the carbons appear or the letter code would be different. For example Maestro might label a string of carbons as CH1, CH2 and CH3 whereas the pdb standard for the same set of carbons is C2, C3 and C4. We once again used text editors to change the nomenclature so that GROMACS and AMBER would be able to read the files. Caution was exercised, as taking the previous example we could not change C4 into C1. Instead we had to inspect each amino acid and either shift the numbers along (e.g changing C2, C3, C4 to C1, C2, C3) or change the letter code depending on the issue. Another problem that arose was that when converting back to pdb the protonation state of histidine would be lost. However this was a simple, albeit tedious, matter of checking each histidine missing a protonation state against the literature and labelling it correctly. Fortunately, as maestro used a conversion algorithm each instance of a nomenclature problem was fairly consistent, so it became easier to correct with each subsequent protein.

A similar issue arose due to how maestro labels hydrogens. It is difficult to identify what name it gave to which hydrogen and how they differ for each instance. The work would be fairly laborious to change each one in turn. However, both MD packages we used could generate missing hydrogens and label them appropriately for the program. This problem was thus remedied by deleting hydrogens that were mislabelled; each program would then replace the missing hydrogens.

3.9 A first investigation into the use of GPUs for protein-ligand simulations

Recent versions of GROMACS have allowed the use of parallel processing using GPU cores instead of CPUs. This is more advantageous for speed as GPUs have far more cores than CPUs. Typical configuration of our system was a pair of SLI linked GeForce 560Ti cards which have 960 cores as seen on Nvidia (2015) which is more than the intel 8core CPUs in our typical configuration. A GPU is usually contained on a graphics card. As opposed to a CPU it is possible to link multiple GPUs set up in tandem on a workstation. We bought a pair of the previously mentioned GPUs and installed them. Then we set up GROMACs to work off these cards with MPI. We decided to test whether we would get similar results for each of the methods: implicit solvent on CPUs, explicit solvent on CPUs, implicit solvent on GPUs and explicit on GPUs. We measured how long they would take in comparison to each other. We took a small subset of the SeraPHic collections of molecules (Favia et al., 2011) to compare the effective speeds of a MD simulation on a GPU compared to a CPU. Each simulation was using the same method as described in section 3.8. We compared the speeds of both implicit and explicit methods. The CPU was a 4 core Intel Pentium chip 2600k running at 4.0Ghz (8 cores with hyperthreading) and used 8Gb of RAM. The GPUs were a pair of GeForce 560Ti, using the same 8Gb of RAM.

3.10 Results – GROMACS

3.10.1 GPU comparisons to CPU speeds

Early on when learning to use GROMACs we attempted to use GPUs as a proof of concept as stated earlier in section 3.9. GROMACs introduced several modules to accommodate specific graphics cards. We started by assessing the speeds (ns/day) to see if we could obtain faster results with the simulations performed on GPUs. The timing results for 5 proteins tested are given in Figure 3.4.

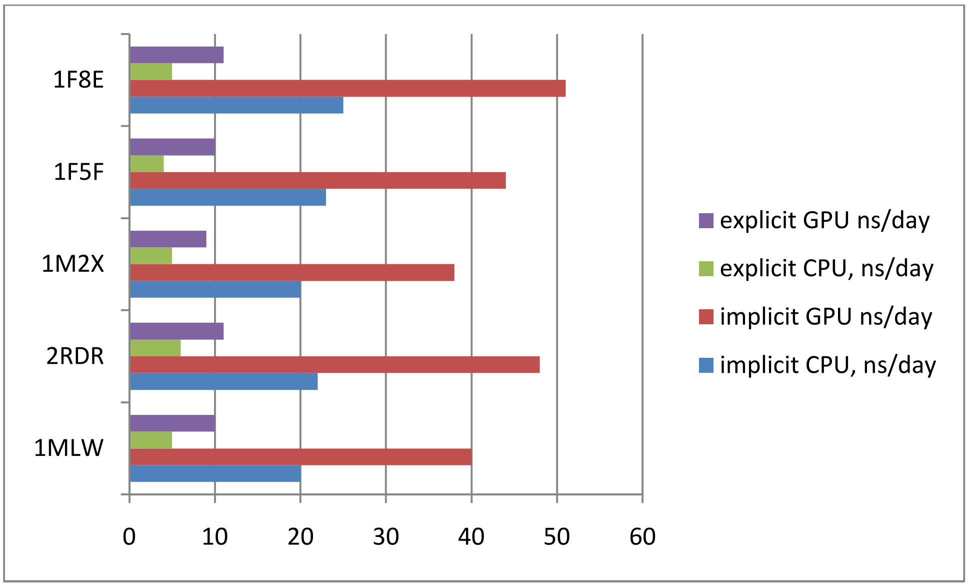


Figure 3.4. Plot of comparisons of speed between GPUs and CPUs tested with either GPUs or CPUs.

Figure 3.4 shows the timing results a small subset of ligands for both Implicit and explicit solvent methods with the speed of calculation for each. Implicit solvent does not need to calculate sets of all-atom water such as TIP3P over every step. This explains the 4-fold increase in speed regardless of whether CPUs or GPUs are processing the data. These speed tests were performed before the main experiments in section 3.8. This was to decide if the speed was worth the considerable effort to setup calculations on the GPUs, as additional preparatory work was required to accommodate the graphics cards. Most programs will run on CPUs regardless of the code, whereas CUDA requires a specific setup on the different iterations of graphics card, shown in Nvidia (2008). As shown in Figure 3.4 the speed of the GPUs is almost double of the CPUs, despite the computer that was used only had a single graphics card. From this modest result, it was concluded that the additional efforts required to set up the GPU simulations were worthwhile, particularly given the prospect of more powerful GPUs in the future. The remaining simulations described in this chapter were therefore run on GPUs to explore whether they were capable of generating results relevant to FBDD.

3.10.2 MD simulation results with RMSD based on crystal pose

Figure 3.5a-r, show a nanosecond by nanosecond RMSD analysis of the results of the GROMACs simulations of 18 different ligands binding to their protein targets. The simulation is of the top ranked pose; the RMSD is relative to the corresponding X-ray crystal structure. Table 3.1 shows the original RMSDs from the polarization docking methods. In addition, some of the figures (3.6a, 3.6b and 3.6c) also show the results of simulation of the docked pose with the lowest RMSD. The lowest RMSD is shown as a red line and the legend shows which pose number had the lowest RMSD, as given by Glide.

Unlike the docking results seen in 2.4 we can observe that it is difficult to maintain a RMSD of 2Å across the MD simulation. In some cases such as 1MLW 13 in figure 3.6a the binding can stay around that low value. However, the majority of RMSDs over the MD simulations rest at about 4-6Å. This is caused by the fluctuations of movement that the ligand is experiencing as it rests in the binding pocket. Since these ligands are all fragments they are being pushed around by water molecules as they move in and out of the binding site. This is different to docking where the ligand is static and we get specific poses. The MD data can show the ligands finding these poses for several ps, typically between 4-5 ps, but then move back out again. This is the nature of the results for MD so overall we would not expect as low RMSD numbers.

The first thing that is apparent from the simulation is that the very low RMSDs obtained in docking (See Table 2.1) are not replicated in the simulations with the same frequency. Neuramidase (1F8E) is an exception to this as the RMSD stays at about 1 Å throughout the simulation, while agmatinase (1W0G) shows more typical results with the RMSD hovering around 2-4 Å. However, not all ligands can be considered to remain bound throughout the simulation.

Table 3.1 The RMSDs of the top ranked pose for the docking of the ligand to its protein target; the system is denoted by its PDB code. Where a lower RMSD but lower ranked pose is available, the lower RMSD is also given along with the pose number in parentheses.

PDB	RMSD / Å	Lowest RMSD (and pose number)
1FSF	0.17	-
1F8E	0.17	-
1M2X	0.18	-
1M3U[a]	6.34	6.333 (2)
1MLW	3.36	1.83 (13)
1OFZ	0.57	-
1PWM[a]	0.22	-
1TKU_A	0.54	-
1UWC[a]	0.38	-
1W1A	0.72	-
1WOG	2.02	1.25 (2)
1YV5	0.6	-
1Y2K[a]	3.8	2.4 (3)
2AIE	1.60	0.90 (4)
2BRT	0.79	-
2BL9	0.19	-
2CIX[b]	2.9	-
2FF2		
2GG7	0.95	-
2GVV	0.75	-
2RDR	1.29	-
2J5S	1.04	-
2Q6M	0.57	-
2ZVJ	0.89	-
3DSX[a]	2.85	2.47 (2)

[a] These protein complexes were used in simulations only in chapter 4.

[b] This protein was tested using the redock data from chapter 2.

The figures generally contain 3 shapes of graphs, as typified by 1YV5, 1F5F and 2J5S (figures 3.5i, 3.5a and 3.5r). Figure 3.5i 1YV5 shows that in the first few frames, the RMSD jumps to 4-6 Å then the RMSD fluctuates between these values. This first type of shape indicates a relatively stable ligand that stays close to the binding site and just moves about in the binding site possible rolling or shifting slightly as these are fragments, only part of the fragment is likely to bind. This means that there is some movement that arises because there are only one or two strong interactions instead of more, which would be expected if the whole structure of a lead-like compound was locked in place.

In the second type of shape, as typified by Figure 3.5a (1F5F), the RMSD steadily increases; sometimes the increase is very sharp like the one seen in protein chloroperoxidase (2CIX) where the RMSD continues to climb to 10-15 Å (figure 3.5l). For the third type of shape, the RMSD goes up then down. There is just one example of this as in beta-diketone hydrolase (2J5S, figure 3.5r). This signifies the ligand leaving the binding site then returning to a very similar binding pose.

The hypothesis to this part of the work was that the results with low RMSD, if correctly docked, would yield graphs similar to the first shape (or 3rd) and those with a high RMSD would yield the second shape. This wasn't always the case: 1M2X and 1TKU had the ligand leave the binding site steadily, despite their low RMSD values in the docking. 1M2X and 1TKU were also the top-scoring glide poses. However the other low RMSD poses such as 1OFZ (figure 3.5g) and 2FF2 (figure 3.5m) yielded results like the first shape (figure 3.5i c.f. 1YV5). Interestingly though, if we look at 1MLW the top glide score pose stayed within the binding site whereas the 13th was bound much tighter and fluctuated less despite the lower Glidescore (highest score is best). 1WOG and 2AIE show that both the top scoring and the lowest RMSD stay in the binding site. However, for 2AIE the lowest RMSD showed better results, as would be expected.

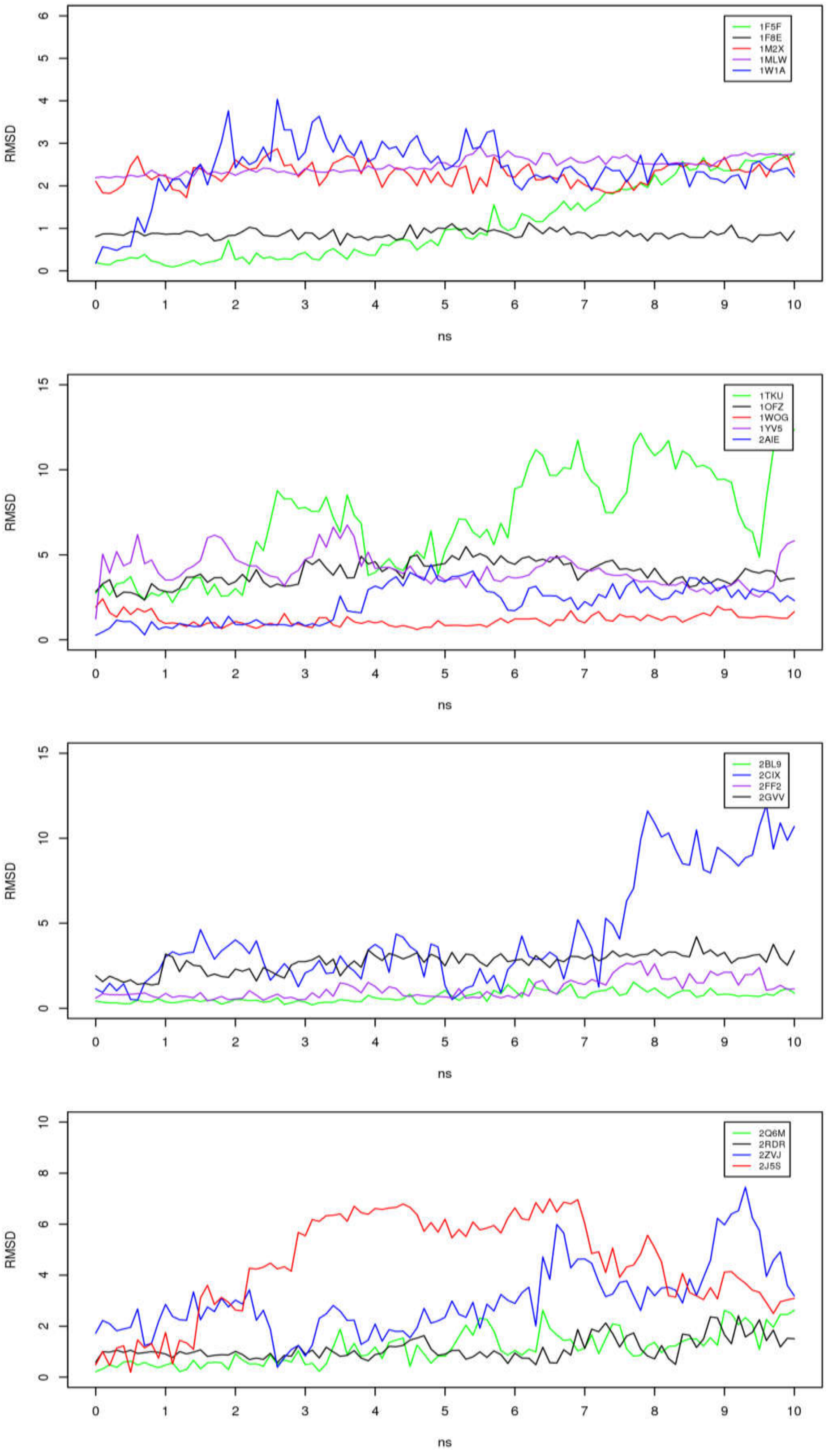


Figure 3.5a-r. RMSD (in Å) of the GROMACs data set against the crystal structure. In the text lines are referred to as 3.4: a: 1F5F, b: 1F8E, c: 1M2X, d: 1MLW, e: 1W1A, f: 1TKU, g: 1OFZ, h: 1WOG, i: 1YV5, j: 2AIE, k: 2BL9, l: 2CIX, m: 2FF2, n: 2GVV, o: 2Q6M, p: 2RDR, q: 2ZVJ, r: 2J5S

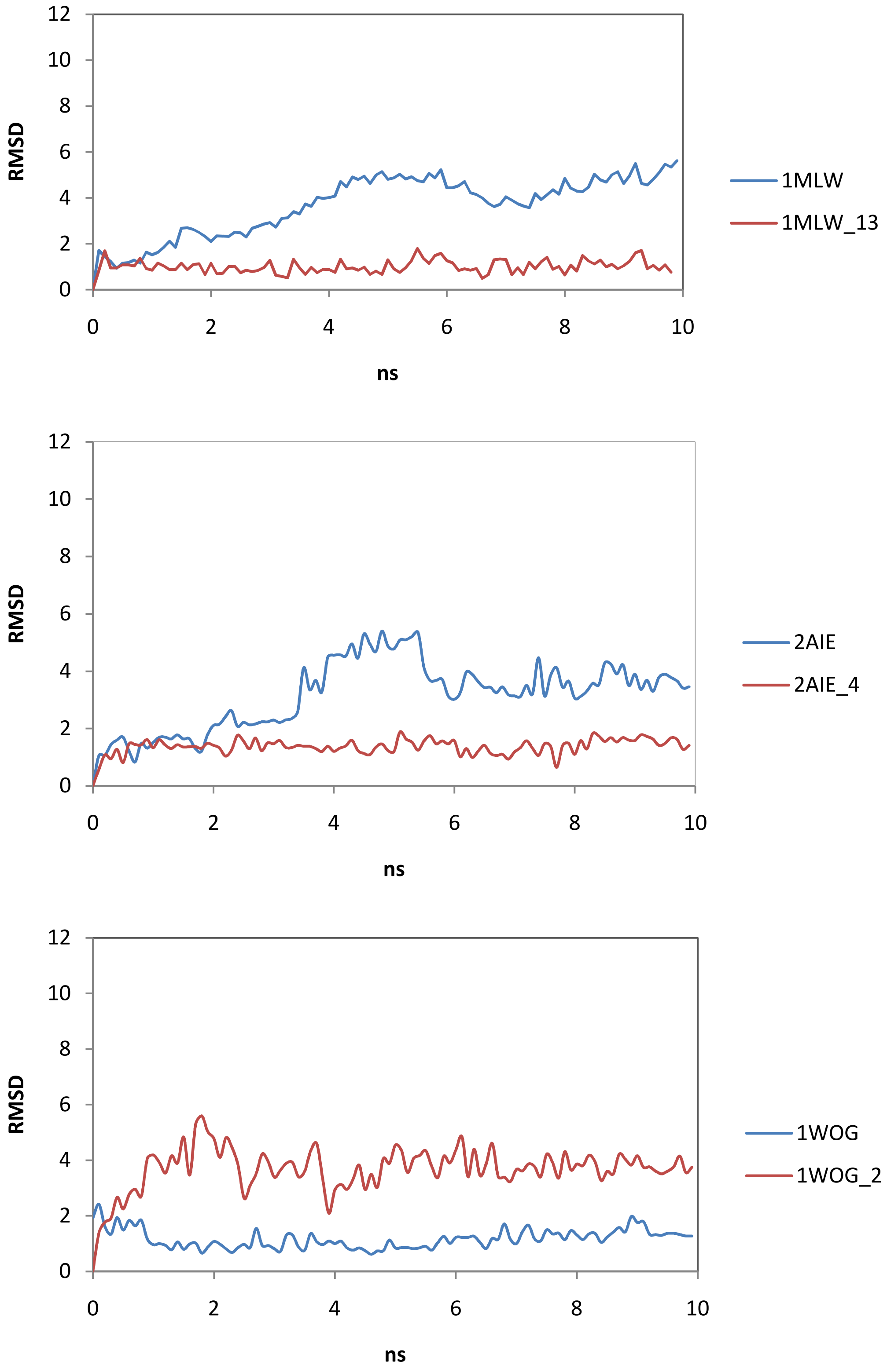


Figure 3.6a-c. RMSD (in Å) of simulation for top pose and pose with the lowest RMSD (in red) for a: 1MLW, b: 2AIE, c: 1WOG

In some cases such as 1F5F and 2CIX the ligand was ejected after sometime. 1F5F has a fairly strong binding affinity with an IC_{50} = 3.8 nM as in table 2.1. This could explain why there is little fluctuation despite binding in the wrong position, in that if it can form strong bonds to its binding site it could form strong bonds elsewhere.

Conversely 2CIX has a much lower binding affinity with a K_d of 33 mM as in table 2.1. This weaker binding could explain how water can dislodge the ligand from its binding pocket: Fig 3.5l shows that the RMSD starts low and gradually increases. If we look at some of the more successful simulations such as 2Q6M and 2BL9 where the RMSD stays within the 1-2Å region the binding affinity is stronger. Their affinities values are K_d = 510 nM for 2Q6M and K_i = 0.16 nM for 2BL9 as in table 2.1. So, some of the MD data reinforces strong affinity with low fluctuations. Even if, as shown for 1OFZ, it obtains the wrong position, the binding may still remain strong.

3.10.3 MD simulation results with RMSD based on starting frame

Another way we can look at the results is instead of comparing the RMSD to the crystal structure we can instead take the first frame of movement for the ligand and calculate the RMSD from that position. This first frame is where the ligands were docked in section 2.3.4.2. In figure 3.7a-r we can see a nanosecond by nanosecond analysis of the same data in figure 3.5a-r except based on the first frame of the MD simulation. For correctly docked ligands, the analysis in figure 3.5a-r and figure 3.7a-r will be similar, but for an incorrectly docked ligand these could be quite different.

This can help paint a clearer picture as to whether a ligand was ejected or not. For example for beta-diketone hydrolase (2J5S, Figure 3.7r) we see a near constant value of 8 Å, this means it quickly left its starting position – even before the RMSD for a pose close to the starting point could be determined. Even when it came back following the type of behaviour as shown in figure 3.5r it has moved far away from where it was initially docked. Most of the ligands follow a similar shape, staying within 1-2 Å of fluctuations, showing that even if they initially moved from where they initially docked, they nevertheless tend to settle in a neighbouring pocket. In some cases such as 2CIX it does not stay settled and moves away, 1TKU exhibits similar behaviour. 2ZVJ exhibit similar RMSD however rotates a lot changing positions within the pocket.

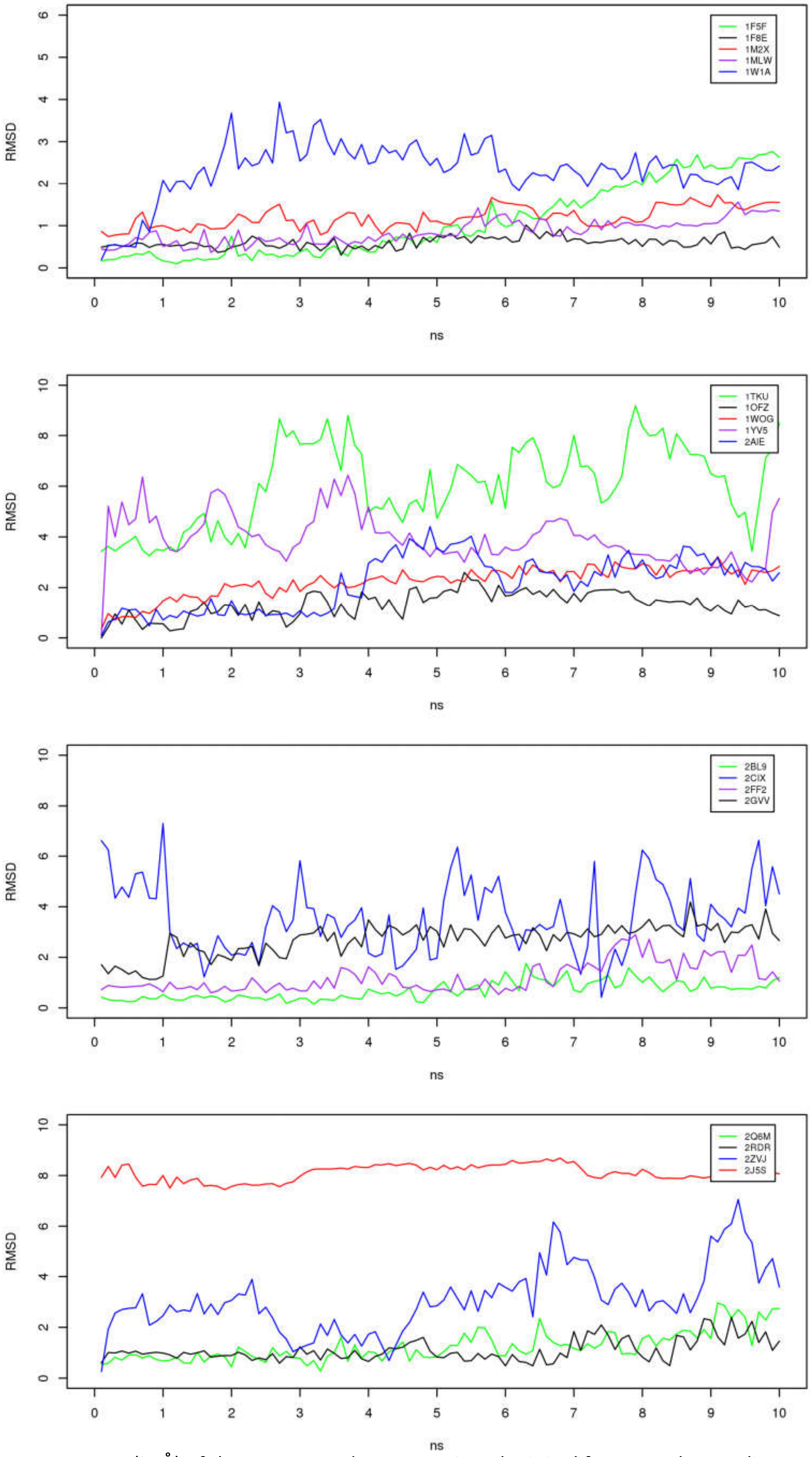
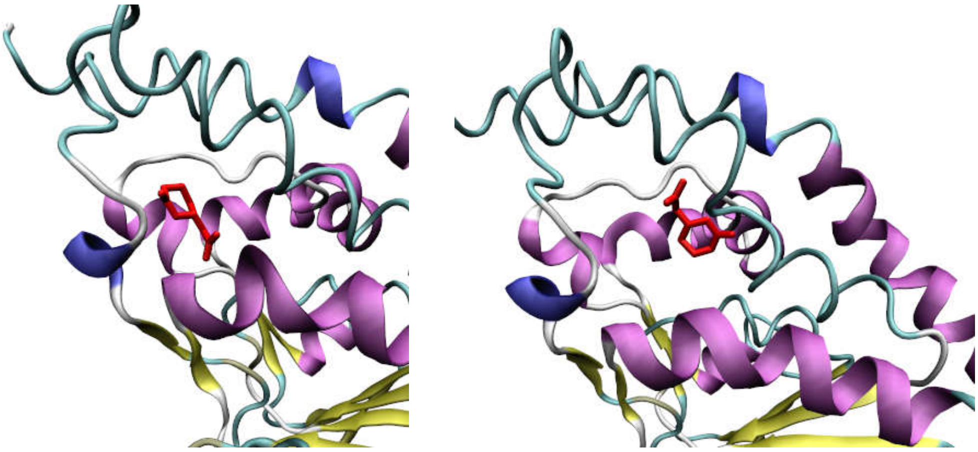


Figure 3.7a-r. RMSD (in Å) of the GROMACs data set against the initial frame. In the text lines are referred to as 3.4: a: 1F5F, b: 1F8E, c: 1M2X, d: 1MLW, e: 1W1A, f: 1TKU, g: 1OFZ, h: 1WOG, i: 1YV5, j: 2AIE, k: 2BL9, l: 2CIX, m: 2FF2, n: 2GVV, o: 2Q6M, p: 2RDR, q: 2ZVJ, r: 2J5S

3.10.4 Examining further

There are several of the protein-ligand complexes that exhibited unexpected behaviour. In figure 3.9a-f we examine the RMSD of these ligands based on their average position. For 2J5S the average is calculated without the first 2 ns as its movement began with a large shift away from its initial position so that its RMSD rose to ~ 8 Å; this movement occurred during the 2 ns of equilibration.

2J5S fluctuates surprisingly little compared to its average position; this means it kept within a close region from its binding site despite the erratic movement shown in figure 3.5r, which shows the RMSD relative to the X-ray position. Figures 3.8a-b shows the movement of 2J5S. 3.8a is its initial position and figure 3.8b is its final position. The ligand leaves the initial section of the binding site which is the pose closer to the X-ray structure and continues to rotate in a gap near the binding site; this constant movement and likely lower binding affinity is preventing it from moving back into position. It latches onto the other side of the binding site shown in figure 3.8b



Figures: 3.8a-b. A graphical representation of binding site of 2J5S; a on the left is the initial pose, b is the final pose. The ligand is in red.

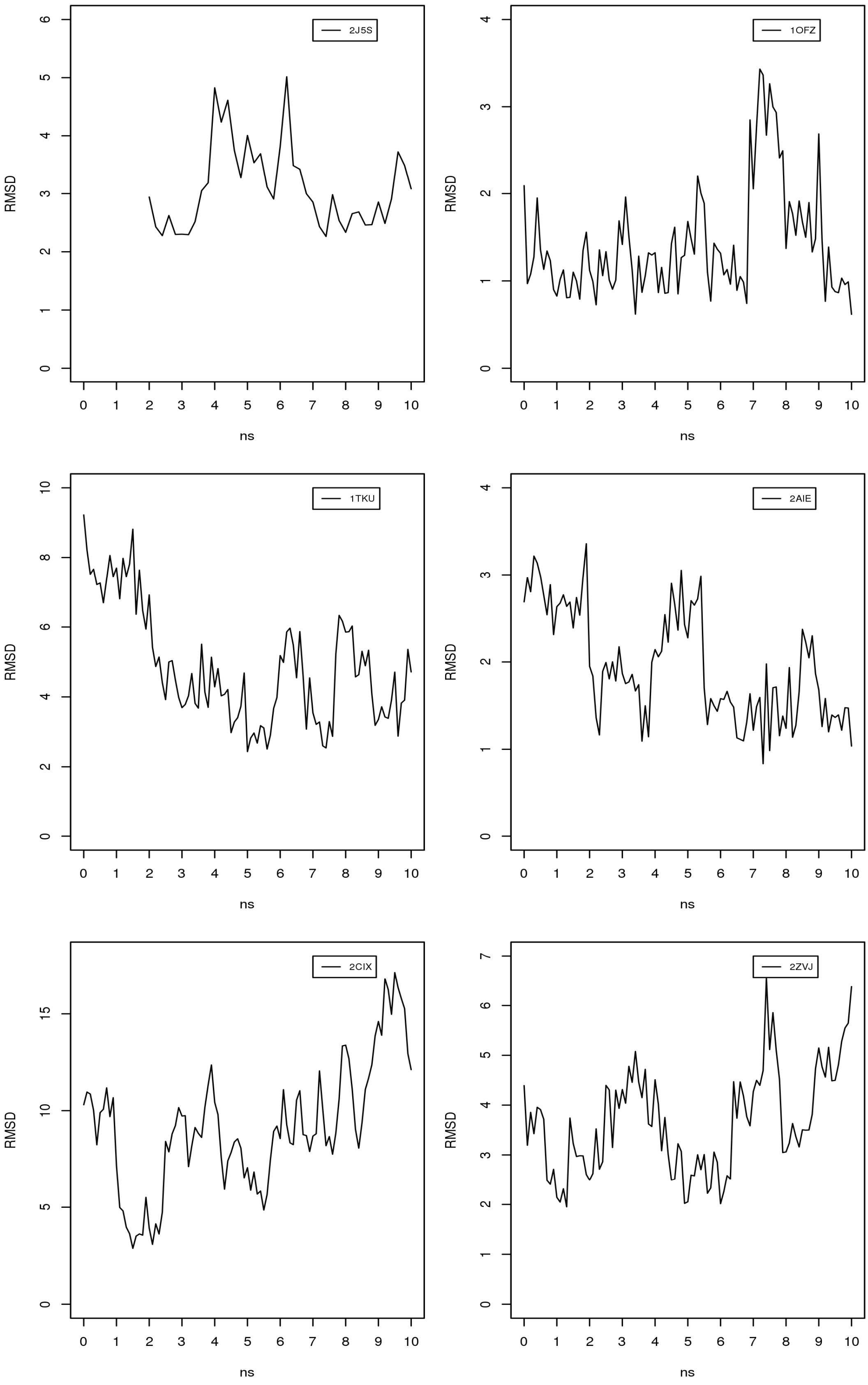


Figure 3.9a-f: RMSD (in Å) of MD simulation based on the average RMSD, a: 25JS, b: 1OFZ, c: 1TKU, d: 2AIE, e: 2CIX, f: 2ZVJ

For 2AIE, as with 2J5S, there is not a significant fluctuation compared to its average pose. It moves into a slightly different point in the pocket as shown in 3.10b. The hydrophobic properties of this pocket appear to be similar to where the crystal structure is bound, so this could be a different configuration. This pose is slowly pushed a little bit away from the pocket and is shown in figure 3.10b to be relatively stable but not as tightly bound as in 3.10a.

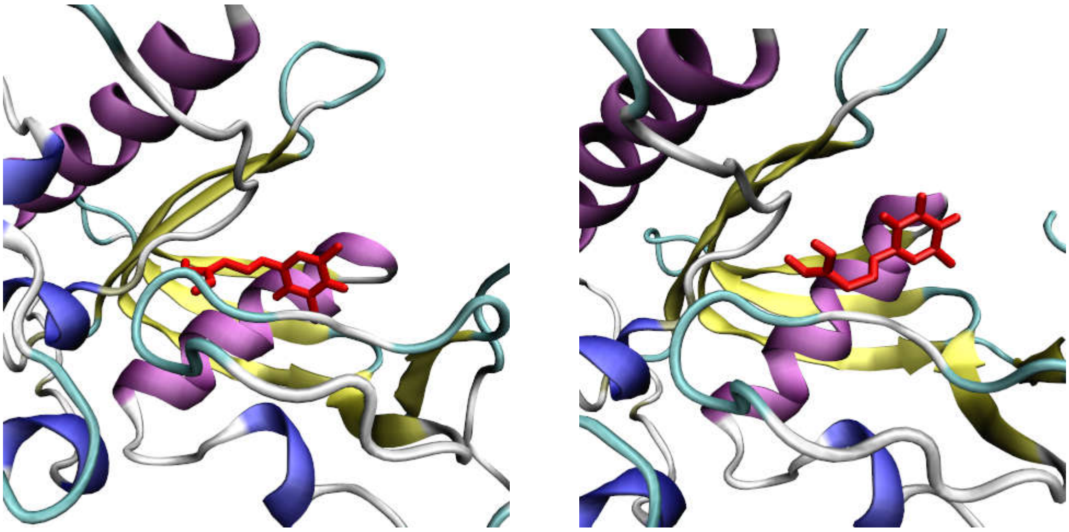


Figure 3.10a-b. A graphical representation of binding site of 2AIE; a on the left is the initial pose, b is the final pose. The ligand is in red.

For Chloroperoxidase (2CIX), the ligand is ejected completely and the average data shows this. This behaviour is unfortunately not that interesting 10 ns into the run. The average position is far away from the correct position compared to the crystal structure and hasn't found a suitable alternative binding site either.

Catecholo-methyltransferase (2ZVJ), like 2CIX, shows some erratic movement, as shown in figure 3.11 the ligand has not been completely ejected. The IC_{50} of the ligand is 1.8 μM according to Tsuji et al. (2009). Figure 3.11 shows that the ligand is not tightly binding to anywhere in its pocket but is being thrown around inside. The original pose is found in blue whereas the majority of the simulation it is at the position in red. The fragment is big enough that it stays interacting with the pocket as it moves away and still manages to stay on the other end of the binding site. The original pose is closer to the conformation in the X-ray structure.

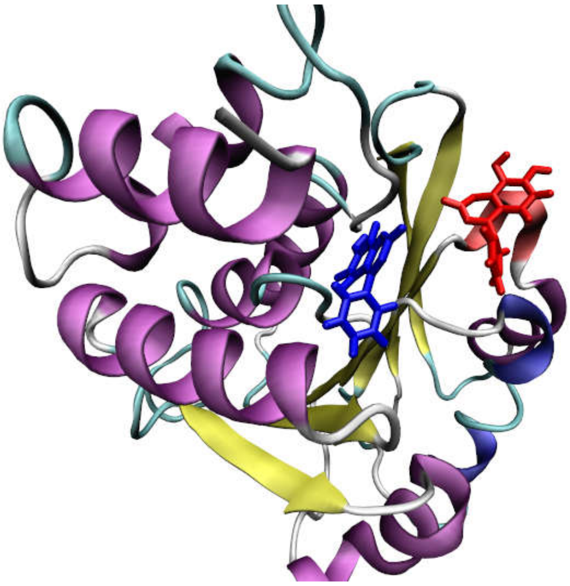


Figure 3.11. A graphical representation of the binding site of 2ZVJ. The original pose is in blue and the average position of the ligand over the course of the simulation is in red.

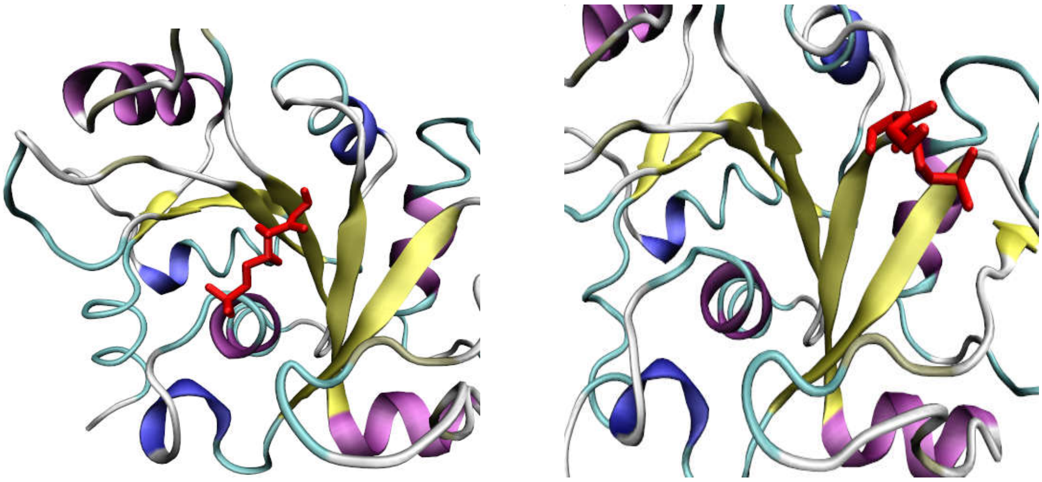


Figure 3.12a-b. A graphical representation of the binding site of 1TKU; a on the left is the initial pose, b is the final pose. The ligand is in red.

1TKU, according to figure 3.12, shows similar behaviour to 2J5S in that the ligand fluctuates in the binding site. However in the case of 1TKU it leaves the initial binding position which is the pose that is closer to the X ray structure immediately then finds a new position at the other side of the pocket where it spends most of the simulation trying to find a resting position within that pocket.

3.11 Discussion

The use of even modest GPUs on simple workstations has indicated that GPU based simulations offer the potential of faster simulations than traditional CPUs, and hence raises the prospect of using such simulations in FBDD. Until recently, the use of GPUs would have involved considerable obstacles, not least because of the need to port the code. However, the implementation of suitable code into programs such as GROMACS, AMBER and ACEMD has transformed this situation, enabling us to explore various protein-fragment complexes using explicit atomistic molecular dynamics simulations.

The hypothesis was that for an incorrectly docked ligand, e.g. 1MLW and 2CIX in Table 3.1 where the RMSD of the top docked GLIDE pose was 3.36 Å and 2.9 Å, the ligand would not be stable, and would leave the binding site, which would be shown by an increasing RMSD to the docked pose. This is not fully the case, as 1MLW maintains an RMSD of ~ 1 Å, suggesting it is correctly docked. 2CIX on the other hand presents an oscillating RMSD, which does suggest that the incorrect docking has yielded an unstable pose.

A number of correctly docked ligands do show the expected behaviour, and maintain a relatively low RMSD (< 4 Å) to both the initial pose and to the X-ray pose, which may not necessarily be known in a de novo drug design program. Examples of this behaviour include 1F5F, 1F8E, 1M2X, 1W1A, 1WOG, 1OFZ, 2AIE, 2BLS, 2FF2, 2GVV and possibly 1YV5. However, whereas in Table 2.1 the RMSDs are generally below 2 Å, the RMSDs obtained from the simulations are higher, partly because these are dynamic systems and so it is inevitable that the breaking and reforming of hydrogen bonds will yield a higher RMSD. This is particularly true for compounds with a low experimental binding affinity such as 1TKU where the RMSD to the X-ray pose was ~ 15 Å, despite a correctly docked pose, but compounds with a nanomolar binding affinity such as 2Q6M and 2BL9 showed low fluctuations within the correct binding site. Thus, it is possible that the hypothesis is more valid for tightly binding fragments.

1MLW is the compound that most strongly challenges the hypothesis as it is docked incorrectly, but nevertheless has a low RMSD of 1 Å to the initial MD structure. Analysis of the graphics has shown that this has docked to an alternative nearby binding site. The implications of this are that molecular dynamics could be used to find alternative binding sites, or alternative binding modes within the main binding site. Such second sites could be useful in a linking strategy of FBDD. The ability to identify alternative binding sites is important as there is currently much effort within the pharmaceutical industry to utilize alternative binding sites so as to modify rather than to totally block the natural enzyme or receptor activity.

Taking the example of 1OFZ we can perhaps find something interesting about the ligand using MD. On some initial trials which failed to run to completion (results not shown) it would move to another nearby binding site. 1OFZ has many alternative sites as seen in Wimmerova et al. (2003). As reported by Hardy and Wells (2004), many drugs are designed to target these alternative sites. By targeting these sites it is possible to make a drug-like target that isn't directly competing with the active site or the inhibitors. In some circumstances, this could lead to a much more successful drug with fewer side effects.

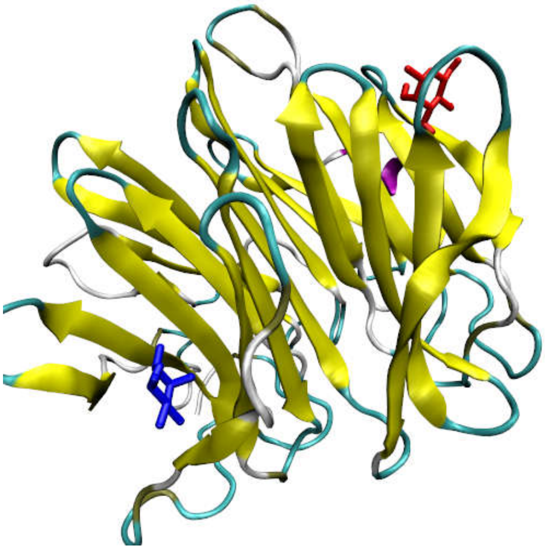


Figure 3.13. A graphical of the bound ligand for 1OFZ shown in blue and an alternative site shown in red.

1OFZ as shown in Fig 3.13 The bound position it found during one of the initial runs is shown in blue and the position it was in for most of the simulation is shown in red.

As stated before (Section 3.10), it managed to bind quite tightly to this binding site as much as its normal one, showing that despite where it was initially placed by the docking it is able to find a new site and bind there.

The results support the hypothesis to some extent, but as will be shown in the next chapter, the 2CIX results are somewhat limited by the short 10 ns simulations. This limited time can show different results when compared to the longer simulations performed with ACEMD later.

One downside to using the charges as we have in this method is that as the ligand moves to a different environment the polarization may not be valid, and that perhaps improvements could have been made by re-evaluating this every time there was a significant change in RMSD however the methods we have used in this chapter and chapter 4 do not account for a change in charge during a simulation.

All of the preceding results were thus generated using the GPU methods encoded in GROMACs. After obtaining the data for these relatively short explicit simulations we were able to expand the work by using more new powerful equipment. This prompted our move to ACEMD and using more than one graphics card in tandem to process not only bigger systems but also to run the simulations over a longer period of time; this will be discussed in the next chapter.

3.12 References

ABASCAL, J. L. & VEGA, C. 2005. A general purpose model for the condensed phases of water: TIP4P/2005. *The Journal of chemical physics,* 123**,** 234505.

AICHELIN, J. 1991. "Quantum" molecular dynamics—a dynamical microscopic n-body approach to investigate fragment formation and the nuclear equation of state in heavy ion collisions. *Physics Reports,* 202**,** 233-360.

ALLINGER, N. L., YUH, Y. H. & LII, J. H. 1989. Molecular mechanics. The MM3 force field for hydrocarbons. 1. *Journal of the American Chemical Society,* 111**,** 8551-8566.

AMAROUCHE, M., GADEA, F. & DURUP, J. 1989. A proposal for the theoretical treatment of multi-electronic-state molecular dynamics: Hemiquantal dynamics with the whole dim basis (HWD). A test on the evolution of excited Ar 3+ cluster ions. *Chemical Physics,* 130**,** 145-157.

BERENDSEN, H. & VAN GUNSTEREN, W. 1984. Molecular dynamics simulations: Techniques and approaches. *Molecular Liquids.* Springer.

BERENDSEN, H. J., VAN DER SPOEL, D. & VAN DRUNEN, R. 1995. GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications,* 91**,** 43-56.

BJELKMAR, P., LARSSON, P., CUENDET, M. A., HESS, B. & LINDAHL, E. 2010. Implementation of the CHARMM force field in GROMACS: Analysis of protein stability effects from correction maps, virtual interaction sites, and water models. *Journal of Chemical Theory and Computation,* 6**,** 459-466.

CASE, D. A., J.T. BERRYMAN, R.M. BETZ, D.S. CERUTTI, T.E. CHEATHAM, I., T.A. DARDEN, R.E. DUKE, T.J. GIESE, H. GOHLKE, A.W. GOETZ, N. HOMEYER, S. IZADI, P. JANOWSKI, J. KAUS, A. KOVALENKO, T.S. LEE, S. LEGRAND, P. LI, T. LUCHKO, R. LUO, B. MADEJ, K.M. MERZ, G. MONARD, P. NEEDHAM, H. NGUYEN, H.T. NGUYEN, I. OMELYAN, A. ONUFRIEV, D.R. ROE, A. ROITBERG, R. SALOMON-FERRER, C.L. SIMMERLING, W. SMITH, J. SWAILS, R.C. WALKER, J. WANG, R.M. WOLF, X. WU, YORK, D. M. & KOLLMAN, A. P. A. 2015. AMBER 2015. University of California, San Francisco.

DE LEEUW, S. W., PERRAM, J. W. & SMITH, E. R. Simulation of electrostatic systems in periodic boundary conditions. I. Lattice sums and dielectric constants. Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 1980. The Royal Society, 27-56.

ELIAV, E. 2008. *Elementary introduction to Molecular Mechanics and Dynamics* [Online]. Tel Aviv University. [Accessed 22 September 2015.

FAVIA, A. D., BOTTEGONI, G., NOBELI, I., BISIGNANO, P. & CAVALLI, A. 2011. SERAPhiC: A benchmark for in silico fragment-based drug design. *Journal of chemical information and modeling,* 51**,** 2882-2896.

GOODSELL, D. S., MORRIS, G. M. & OLSON, A. J. 1996. Automated docking of flexible ligands: applications of AutoDock. *Journal of Molecular Recognition,* 9**,** 1-5.

HARDY, J. A. & WELLS, J. A. 2004. Searching for new allosteric sites in enzymes. *Current opinion in structural biology,* 14**,** 706-715.

HARVEY, M. J., GIUPPONI, G. & FABRITIIS, G. D. 2009. ACEMD: accelerating biomolecular dynamics in the microsecond time scale. *Journal of chemical theory and computation,* 5**,** 1632-1639.

HEHRE, W. J. 2003. *A Guide to Molecular Mechanics and Quantum Chemical Calculations,* Wavefunction, Inc., Irvine, CA.

HESS, B. 2008. P-LINCS: A parallel linear constraint solver for molecular simulation. *Journal of Chemical Theory and Computation,* 4**,** 116-122.

HESS, B., KUTZNER, C., VAN DER SPOEL, D. & LINDAHL, E. 2008. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of chemical theory and computation,* 4**,** 435-447.

HU, H., ELSTNER, M. & HERMANS, J. 2003. Comparison of a QM/MM force field and molecular mechanics force fields in simulations of alanine and glycine "dipeptides"(Ace-Ala-Nme and Ace-Gly-Nme) in water in relation to the problem of modeling the unfolded peptide backbone in solution. *Proteins: Structure, Function, and Bioinformatics,* 50**,** 451-463.

KARPLUS, M. 1983. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem,* 4**,** 187217.

KARPLUS, M. & MCCAMMON, J. A. 2002. Molecular dynamics simulations of biomolecules. *Nature Structural & Molecular Biology,* 9**,** 646-652.

MAIOROV, V. N. & CRIPPEN, G. M. 1994. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *Journal of molecular biology,* 235**,** 625-634.

MAKOV, G. & PAYNE, M. 1995. Periodic boundary conditions in ab initio calculations. *Physical Review B,* 51**,** 4014.

MCDONALD, I. 1972. NpT-ensemble Monte Carlo calculations for binary liquid mixtures. *Molecular Physics,* 23**,** 41-58.

MOSKOWITZ, J. W., SCHMIDT, K., WILSON, S. & CUI, W. 1988. The application of Simulated Annealing to problems of molecular mechanics. *International Journal of Quantum Chemistry,* 34**,** 611-617.

MURSHUDOV, G. N., SKUBÁK, P., LEBEDEV, A. A., PANNU, N. S., STEINER, R. A., NICHOLLS, R. A., WINN, M. D., LONG, F. & VAGIN, A. A. 2011. REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallographica Section D: Biological Crystallography,* 67**,** 355-367.

NVIDIA. 2015. *Nvidia* [Online]. Available: http://www.nvidia.co.uk/object/geforce-desktop-graphics-cards-uk.html [Accessed 22 Sep 2015 2015].

NVIDIA, C. 2008. Programming guide.

P.KOEHL & M.LEVITT. 2002. *Implicit and Explicit solvent models* [Online]. Available: http://csb.stanford.edu/~koehl/ProShape/born.php [Accessed 22 Sep 2015 2015].

PHILLIPS, J. C., BRAUN, R., WANG, W., GUMBART, J., TAJKHORSHID, E., VILLA, E., CHIPOT, C., SKEEL, R. D., KALE, L. & SCHULTEN, K. 2005. Scalable molecular dynamics with NAMD. *Journal of computational chemistry,* 26**,** 1781-1802.

PLIMPTON, S. 1995. Fast parallel algorithms for short-range molecular dynamics. *Journal of computational physics,* 117**,** 1-19.

PONDER, J. W. 2004. TINKER: Software tools for molecular design. *Washington University School of Medicine, Saint Louis, MO,* 3.

QIU, D., SHENKIN, P. S., HOLLINGER, F. P. & STILL, W. C. 1997. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. *The Journal of Physical Chemistry A,* 101**,** 3005-3014.

RAPAPORT, D. C. 1995. *The Art of Molecular Dynamics Simulation*, Cambridge University Press, Cambridge.

RAPPÉ, A. K., CASEWIT, C. J., COLWELL, K., GODDARD III, W. & SKIFF, W. 1992. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society,* 114**,** 10024-10035.

RITCHIE, D. W. 2003. Evaluation of protein docking predictions using Hex 3.1 in CAPRI rounds 1 and 2. *Proteins: Structure, Function, and Bioinformatics,* 52**,** 98-106.

RUSHBROOKE, G., ROWLINSON, J. S. & TEMPERLEY, H. 1968. *Physics of simple liquids*, Wiley Interscience Division.

SCHUÈTTELKOPF, A. W. & VAN AALTEN, D. M. 2004. PRODRG: a tool for high-throughput crystallography of protein–ligand complexes. *Acta Crystallographica Section D: Biological Crystallography,* 60**,** 1355-1363.

SHATTUCK, T. W. 2008. *Colby College Molecular Mechanics Tutorial* [Online]. http://www.colby.edu/chemistry/CompChem/MMtutor.pdf: Department of Chemistry, Colby College, Waterville, Maine 04901. [Accessed 22 September 2015 2015].

SHIRTS, M. R., PITERA, J. W., SWOPE, W. C. & PANDE, V. S. 2003. Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. *The Journal of chemical physics,* 119**,** 5740-5761.

STOTE, R., DEJAEGERE, A., KUZNETSOV, D. & FALQUET, L. 1999. *Molecular Dynamics Simulations HARMM tutorial* [Online]. Available: http://www.ch.embnet.org/MD_tutorial/ [Accessed 22 September 2015 2015].

TSUJI, E., OKAZAKI, K. & TAKEDA, K. 2009. Crystal structures of rat catechol-O-methyltransferase complexed with coumarine-based inhibitor. *Biochemical and biophysical research communications,* 378**,** 494-497.

TULLY, J. C. 1990. Molecular dynamics with electronic transitions. *The Journal of Chemical Physics,* 93**,** 1061-1071.

VAN DER SPOEL, D. & LINDAHL, E. 2003. Brute-force molecular dynamics simulations of villin headpiece: comparison with NMR parameters. *The Journal of Physical Chemistry B,* 107**,** 11178-11187.

VAN DER SPOEL, D., LINDAHL, E., HESS, B., GROENHOF, G., MARK, A. E. & BERENDSEN, H. J. 2005. GROMACS: fast, flexible, and free. *Journal of computational chemistry,* 26**,** 1701-1718.

VERLET, L. 1980. Integral equations for classical fluids: I. The hard sphere case. *Molecular Physics,* 41**,** 183-190.

WEBSTER, F., ROSSKY, P. & FRIESNER, R. 1991. Nonadiabatic processes in condensed matter: semi-classical theory and implementation. *Computer Physics Communications,* 63**,** 494-522.

WELLER, H., SCHMIDT, H., KOCH, U., FOJTIK, A., BARAL, S., HENGLEIN, A., KUNATH, W., WEISS, K. & DIEMAN, E. 1986. Photochemistry of colloidal semiconductors. Onset of light absorption as a function of size of extremely small CdS particles. *Chemical physics letters,* 124**,** 557-560.

WIMMEROVA, M., MITCHELL, E., SANCHEZ, J.-F., GAUTIER, C. & IMBERTY, A. 2003. Crystal Structure of Fungal Lectin SIX-BLADED β-PROPELLER FOLD AND NOVEL FUCOSE RECOGNITION MODE FOR ALEURIA AURANTIA LECTIN. *Journal of Biological Chemistry,* 278**,** 27059-27067.

WOOD, W. 1968. Monte Carlo Calculations for Hard Disks in the Isothermal-Isobaric Ensemble. *The Journal of Chemical Physics,* 48**,** 415-434.

YU, H. & VAN GUNSTEREN, W. F. 2005. Accounting for polarization in molecular simulation. *Computer physics communications,* 172**,** 69-85.

Chapter 4

Protein-fragment complex simulations using ACEMD

4.1 An introduction to ACEMD

ACEMD is a high performance MD code designed specifically for NVIDIA GPUs (Harvey et al., 2009). It has similarities to all the main MD codes such as GROMACs as it uses parallelisation. Due to the high performance of GPUs, ACEMD can reach over 40 ns/day in calculations with over 23000 atoms in a single system, as seen in Harvey et al. (2009); with more modern GPUs the performance is even higher. GPUs have many processing units each with their own cache and control units dedicated to their use as shown in (Gupta and Babu, 2011). By the use of each of these many processing units that work in parallel, a modern GPU can work with thousands of cores as stated in the specifications found in Nvidia (2015). This can be pictured as in figure 4.01, taken from Gupta and Babu (2011). ACEMD uses CHARMM, AMBER or OPLS force fields to perform its calculations as seen in Harvey et al. (2009). Due to this speed advantage there is more incentive to use an all-atom system of explicit water molecules to generate more accurate data rather than using the potentially less accurate implicit solvent methods.

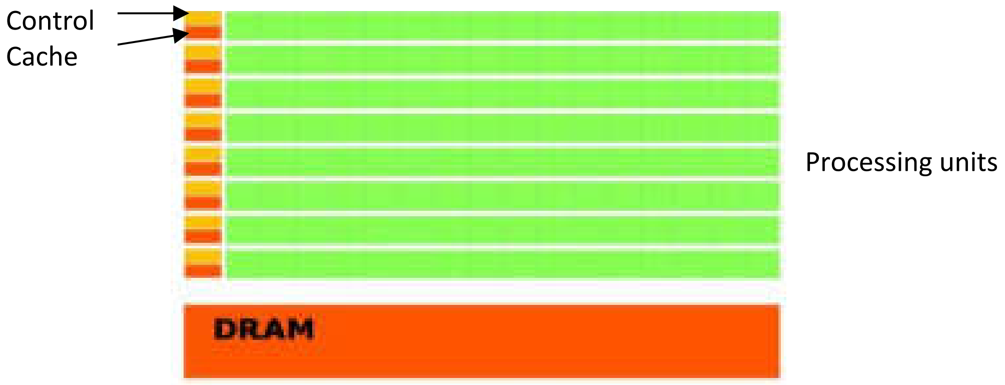


Figure 4.1. General architecture of a GPU

ACEMD's advantage in speed comes with a price, namely that the code used must usually be in CUDA due to the use of NVIDIA GPUs, as seen in Nvidia (2008). Moreover, CUDA is undergoing constant changes with new graphics cards and updated versions of the NVIDIA drivers are also required (Kirk, 2007). This causes further development time for the methods when upgrades are needed compared to higher order code written for a CPU which usually requires no major alterations from version to version, as seen for C++ in Gregor et al. (2006). However, in more recent years the Khronos group has published papers that demonstrate the use of OpenCL instead of CUDA for MD codes as seen in Harvey and De Fabritiis (2011). The advantage this could present is that OpenCL is open source and not propriety of NVIDIA. OpenCL can be used across any form of GPU, and the language offers similarities to traditional CPU programming as it doesn't require many alterations to use as versions of the CUDA language are upgraded (Shams and Kennedy, 2007).

For consistency with the GROMACs work, we used the AMBER force field to perform MD simulations on the proteins described in the SeraPhic paper by Favia et al.

(2011). However we took the QM/MM polarized ligand charges and integrated them into the ligand. This is similar to the tests carried out using GROMACs. However they were performed at a much more rapid rate due to the aforementioned power of GPUs, as installed in a set of newly acquired workstations specifically designed to run ACEMD. These workstations featured a quad core i7 processor and 1 or 4 GT780 Nvidia commodity GPUs that were heavily tested to ensure that they were of an appropriate quality for running MD simulations.

4.2 Force fields

As stated earlier, ACEMD is optimised to use two different force fields; CHARMM and AMBER. As described in section 3.2 these are MM force fields and they utilise similar equations to implement the energy and derivative calculations. Below (Equations 4.1, 4.2) is a comparison of the equations of these classical force fields as shown in Case et al. (2015). Due to their similarities, ACEMD can use them interchangeably with some re-parameterization if the user chooses to convert from one force field to the other.

$$
\begin{aligned}
V_{\mathrm{AMBER}} \quad = \quad & \sum_{\text{bonds}} k(r-r_{eq})^2 + \sum_{\text{angles}} k(\theta-\theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2}[1+cos(n\phi-\gamma)] \; / \\
& + \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}}\right] + \sum_{i<j}\left[\frac{q_i q_j}{\varepsilon R_{ij}}\right]
\end{aligned}
\tag{4.1}
$$

$$
\begin{aligned}
V_{\mathrm{CHARMM}} \quad = \quad & \sum_{\text{bonds}} k_b(b-b_0)^2 + \sum_{\text{angles}} k_\theta(\theta-\theta_0)^2 + \sum_{\text{dihedrals}} k_\phi[1+cos(n\phi-\delta)] \\
& + \sum_{\text{Urey-Bradley}} k_u(u-u_0)^2 + \sum_{\text{impropers}} k(\omega-\omega_0)^2 + \sum_{\phi,\psi} V_{CMAP} \\
& + \sum_{\text{nonbonded}} \varepsilon\left[\left(\frac{R_{\mathrm{min}_{ij}}}{r_{ij}}\right)^{12} - \left(\frac{R_{\mathrm{min}_{ij}}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{\varepsilon r_{ij}}
\end{aligned}
\tag{4.2}
$$

Both force field equations share the same sum of energy terms over the bonds, angles and dihedrals. The non-bonded forms have different scaling Case et al. (2015), but the major difference between the two force fields is that CHARMM uses 3 additional bonded terms. The two body Urey-Bradley term (MacKerell et al., 2004) that extends to all 1-3 interactions and the four body quadratic improper term from MacKerell et al. (2000). Lastly, there is a cross term (CMAP), which is a function of two sequential backbone dihedrals (Crowley et al., 2009) which was introduced to improve the accuracy of the force field for treating alpha helices.

4.2.1 CHARMM

The Chemistry at HARvard Macromolecular Mechanics (CHARMM) force fields are specific force fields developed for use with biological Macromolecules, unlike force fields such as MMF94 (Halgren, 1996) that are more general purpose. The CHARMM program is currently able to support both serial and parallel architectures for MD

Brooks et al. (2009) and functions well in hybrid models similar to the QM/MM methods described previously in section 2.3

In the CHARMM force field, an atom is considered to be a charged point with no directional properties and without internal degrees of freedom Karplus (1983). In the earlier AMBER and CHARMM force fields, non-polar hydrogens were simplified and combined with neighbouring heavy atoms to which they are bonded with in an extended atom model. They can also be set apart in an all atom model as shown in Zimrnermann (2003). The main advantage of the extended atom model is that i the force field describes fewer atoms and therefore has fewer terms; it is thus easier to calculate Since there are large numbers of non-polar hydrogen atoms in biological macromolecules, this results in fewer internal degrees of freedom that need to be taken into consideration. However, this is a crude approximation (Karplus, 1983).

As shown in equation 4.3, the CHARMM energy function is based on separable internal coordinate and pairwise interaction terms. Expressed as:

$$E = E_b + E_\theta + E_\phi + E_\omega + E_{vDW} + E_{el} + E_{hb} + E_{er} + E_{e\phi} \qquad (4.3)$$

where the individual energy terms in equation 4.3 are Bond potentials (E_b), Bond angle potentials (E_θ), Dihedral angle potential (E_ϕ), Improper torsions (E_ω), Van der Waal interactions (E_{vDW}), electrostatic interactions (E_{el}), Hydrogen bonding (E_{hb}), constraints for atom harmonics (E_{er}) and dihedrals ($E_{e\phi}$). When expanded it forms the equation 4.3 shown earlier in equation 4.2. These terms are described further in reference Karplus (1983).

There are two ways to generate CHARMM parameters for a given system. The first is to use an in-house version of the CHARMM program, it is a proprietary software so if it is used this requires a purchase of a licence (Brooks et al., 2009) and the second is to utilise the CHARMM-GUI (Jo et al., 2008). The CHARMM-GUI is useful for generating force fields based on PDB files from either saved files or from the RSCB PDB database (H.M. Berman et al., 2000). However, the database does not contain the force field parameters for ligands and can require more preparation than is necessary. In contrast, AMBER is more appropriate for simulations involving protein-ligand complexes due to facilities for generating ligand parameters, as will be discussed in section 4.2.2.

4.2.2 AMBER

The Assisted Model Building with Energy Refinement (AMBER) software was developed originally for biological macromolecules (as was CHARMM). MD simulations involving ligands use the GAFF or General Amber Force Field for their calculations as seen in Wang et al. (2004). The philosophy is to simplify the parameterisation process by having a general force field for common molecular fragments such as amino acid groups. It uses a small number of atom types, namely those found in bio-macromolecules and ions, though GAFF incorporates both empirical and heuristic models to estimate force constants and partial atomic charges as seen in Case et al. (2005).

One of GAFFs advantages is the ability to parameterise organic molecules that do not follow normal protein structure, such as ligands. GAFF has sets of basic atoms types with different hybridsations that can be used to assign force fields to atoms (Wang et al., 2004). Amber can use a program called antechamber (Wang et al., 2006) that can automatically assign the GAFF parameters to an unknown molecule such as a ligand. It also requires only the atomic numbers and bond connectivity of the basic molecule in order to use its algorithms.

Due to this advantage, we used the AMBER force field since all of the validation proteins had by definition a ligand in their binding site. We also used the ff14SB protein force field, which is the most recent version of the AMBER force field for proteins (Maier et al., 2015). CHARMM was also considered but due to the ease of working with unknown molecules AMBER was a better fit to our purposes.

4.3 Methods for simulating protein-ligand complexes

While GROMACs and ACEMD share similar protocols, there are distinct differences between the two programs. The major operation change was to convert from the OPLS force field, used by maestro and GROMACs , to the AMBER force field.

4.3.1 Protein-ligand complex Preparation

AMBER version 14 has the capability to automatically process a protein to generate the AMBER force field through LEAP. However, as described earlier the ligand needs to be assigned its force fields separately. Consequently, the initial step was to separate the protein and the ligand into separate files.

The ligand was prepared using antechamber. However, instead of letting antechamber assign new charges, bespoke charges were assigned using a specific charge file. This specific charge file was taken from previous work on the ligand so that antechamber would obtain the polarised charges and assign them to the ligand. Antechamber would then assign GAFF force field parameters based on atom type before creating a prepi file, an input file containing the ligand topology, input co-ordinates and charges that could be read by LEAP. It also creates an frcmod file which lists the missing force field parameters taken from the GAFF force field. After the ligand was parameterized it was recombined with the parameterized protein to be read through LEAP. During this stage minor edits were made to some of the nomenclature in the protein files as the names used for capped amino acids is different between the two force fields.

We prepared the complex in an explicit water bath. The basic script that converts the protein to use the AMBER force field is given in Appendix 4.1.

Also at this stage, a constraint file was generated. This is for the next stage of minimisation where the waters and non-backbone molecules are minimised. The constraint file is a simple pdb file where the beta column is filled with a 1 for each carbon of the back bone. This tells ACEMD that these are the constrained atoms; the size of the number is proportional to the harmonic potential that constrains the atoms – a value of 1 was found to be suitable.

4.3.2 Minimisation and MD

The refinement of the protein prior to an unconstrained production simulation was performed in 4 stages. Firstly, a simple energy minimisation for 5000 steps was performed before equilibrating the molecule with NVT for 1 ns with constraints, and then changing the ensemble to NPT with constraints for 10 ns followed by keeping the ensemble as NPT without constraints for 10 ns. The time step was set to 4 fs, this large step was made possible by increasing the mass of the hydrogen atoms to 4 da as shown in Pomès and McCammon (1990). The temperature was set to 300 K. Pressure was set to the default of 1.01325 bar. The production run for each protein was 200 ns. A typical script for carrying out these operations is given in Appendix 4.2.

4.3.3 Constraint Scaling Methods

According to the work of Ryckaert et al. (1977), polyatomic molecules in MD have fast internal vibrations that are usually decoupled from rotational and translational motion. These vibrations are bond vibrations as seen in Hess et al. (1997). The time step becomes limited due to these bond vibrations. By adding constraints, the time scale can be increased by a factor of four according to Hess et al. (1997). These can be frozen by placing rigid bonds on the skeleton of the molecule; this is a form of constraint. In classical models this is treated by Lagrange-Hamilton formalism; however, in modern systems with large macromolecules the number of degrees of freedom is very large so different constraints need to be used.

It is possible to use holonomic constraints as shown in Ryckaert et al. (1977). The most wildly used constraint algorithms are SHAKE and LINCS for large molecules, as seen in Kräutler et al. (2001) and Hess et al. (1997). These propose to solve the non-linear problem on resetting coupled constraints of bonds after an unconstrained update. SHAKE is an iterative method which first sequentially sets all the bonds to the correct length. To attain the desired accuracy SHAKE will repeat this iteration since bonds are coupled. However due to the iterative nature of SHAKE it is hard to parallelize the process. LINCS on the other hand resets the constraints on each step of the calculation instead of using the derivatives of each of the iterations. LINCS uses a leap-frog algorithm to calculate the constraints. However this has a drawback in that it does not set the real bond lengths but instead projects it using the leap-frog calculation. This causes the bond lengths to increase slightly but there are corrective algorithms to help alleviate this, as shown in Hess et al. (1997). LINCS is also designed to be used by modern computers and uses parallelization in its calculations.

Constraining the molecule as stated before, will semi-freeze the structure so that during equilibration it can find the energy minima. Water is also free to move while the macromolecule is equilibrating; in this process the water can move to reduce steric strain on the macromolecule, simply by reducing the repulsive forces. However in production runs there should be no constraints as in nature the molecules are not locked to a set position in Cartesian space.

In addition, in the NPT and NVT ensembles, sometimes the protein had some trapped water molecules nestled within its structure. Removing the constraints all in one go would result in a spike of energy as these waters attempted to escape, thus crashing the run. Thus we utilised a scaling method where the constraints were slowly eased off at a slow rate. A typical TCL script to implement this method is shown in appendix 4.3. This script also gives the default simulation parameters such as non-bonded cut-off etc.

4.3.4 Correct usage of water bath sizes in ACEMD simulations

As shown in sections 4.3.1 and 4.3.2 water baths needed to be generated for each of the MD simulations. In the case of ACEMD, if this was prepared incorrectly the periodic boundary would not be kept well. When preparing a water bath, ACEMD would randomly generate water in a pre-determined box for each protein. The exact size of this box was not carried over to the simulation.

ACEMD could guess where the edge of a water bath. However, if it was off by a fraction of an Å then water molecules would form large pockets of empty space at the corners of the water bath. The water would conglomerate nearer the centre and begin to stretch out. Sometimes, this behaviour could cause the protein in the middle of the water bath to reach the pocket of empty space; this caused the protein chain to move differently to how we would expect it to move. This problem was easy to encounter if measurements of the box were slightly off. When preparing each water bath extra precaution was taken so that the size of simulation was the exact size of the water bath or slightly smaller.

In membrane simulations in chapter 5, if this problem was encountered the entire membrane would unravel. The membrane would deform and attempt to form lipid micelles. This would cause the protein to lose some structural stability as the protein needed the lipid bilayer to maintain its shape.

4.3.5 Swan: OpenCL errors

A common error that can arise from ACEMD simulation is a swan error. This is referring to the Swan conversion program (Harvey and De Fabritiis, 2011) which ports CUDA programs into OpenCL. ACEMD was written in the CUDA language however Swan can be used to change it into OpenCL so the program can be read on most any card architecture and not just NVIDIA.

The errors were caused for a variety of reasons. The most common cause was the equilibration was performed incorrectly, either by placing a box too small which would increase the energy of the system to an incorrectly high threshold or the constraints weren't eased off slowly enough. If the constraints weren't eased off slowly enough this would cause a water to be trapped and build up a lot of energy, usually to an incalculable number that Swan wouldn't recognize and subsequently would not convert that part of the simulation into OpenCL which would cause the program to crash.

Swan errors could also occur if ACEMD did not recognize a portion of the input file(s). Thus the front end of the program that would attempt to recognize this anomaly would put nonsense into the conversion algorithm so it would be nonsense when it was converted to OpenCL to be read by the card, this would also result in a crash. This issue could be caused by numerous issues such as formatting one of the input files incorrectly or having a bad initial structure of a ligand or protein. These errors are remedied by carefully inspecting each file before taking it to the next stage of the process avoiding the potential crashes to the program.

4.4 Results – ACEMD simulation of SeraPhic validation set

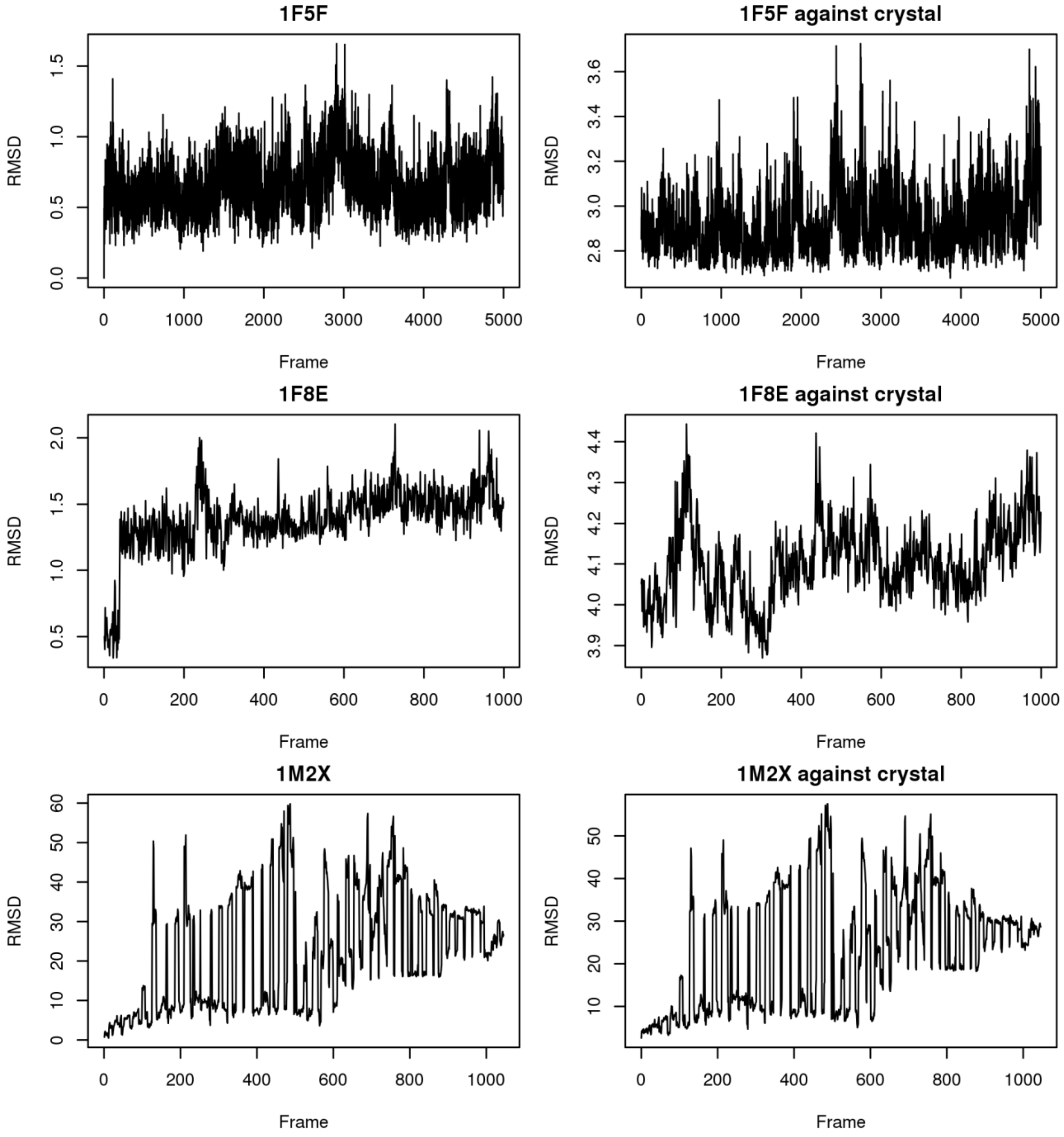


Figure 4.2. RMSD (in Å) for 1F5F, 1F8E and 1M2X - against the initial docked pose and against the crystal structure. The total length of the simulation is 200 ns.

With the longer simulation times in ACEMD when compared to our GROMACs MD simulations (Chapter 3) we can see more concrete patterns for ligand movement. Fig. 4.2 shows a side by side comparison of each ligand's RMSD from their initial docked position (frame zero) on the left and the RMSD against the ligand as it is bound in the crystal structure on the right. The ligands do present varying results in terms of the variation of RMSD with time but the majority have found a similar position to the crystal structure within the binding pocket and remain with the pocket. This is shown when the RMSD is low for both the frame zero and for the comparison against the crystal structure. A good example of this is 1F5F. However as we compare all of them side by side we can see some interesting behaviours come to light.

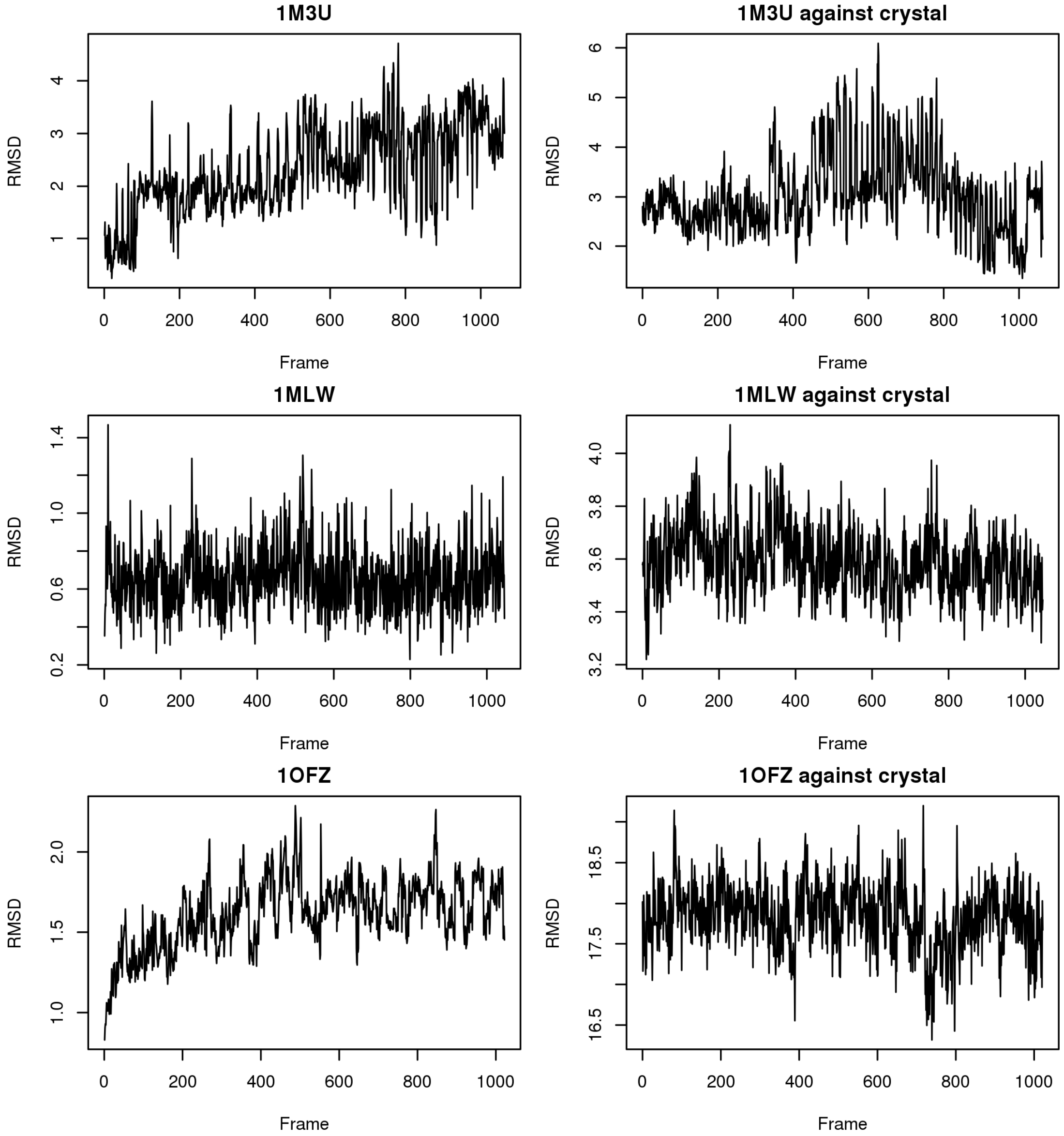


Figure 4.3. RMSD (in Å) for 1M3U, 1MLW and 1OFZ against the initial docked pose and against the crystal structure. The total length of the simulation is 200 ns.

In figure 4.2, 1M2X has a very high RMSD, both against the crystal and first frame of the simulation. The ligand of 1M2X quickly leaves the binding pocket and moves around the outside of the protein. In figure 4.3 we can see that 1M3U when compared to its docked pose slowly leaves that position and moves to a different part of the binding pocket. As shown in figure 4.4 it moves from adjacent to PHE 216 of the pocket over towards LYS 109. When we compare this movement to the position in the crystal structure it is correcting itself and repositioning itself closer to the crystal structure over time. The red ligand in figure 4.4 is the crystal structure position for reference.

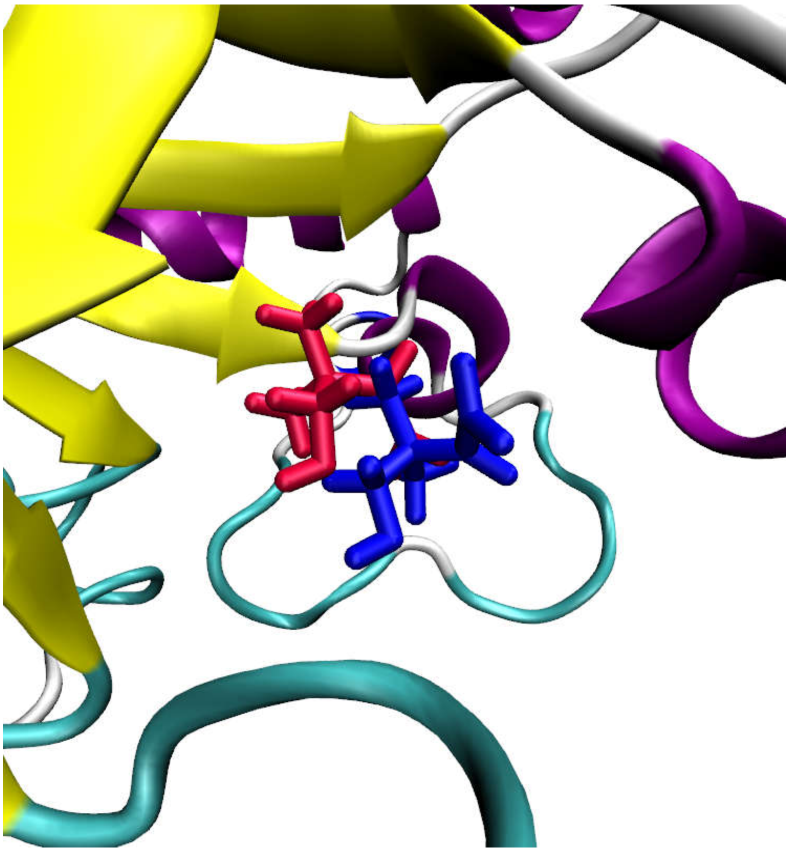


Figure 4.4. A graphical representation of the binding site of 1M3U. The red ligand is the crystal pose, the blue ligand is the average simulation pose.

Again, in Figure 4.3 we can see that both 1MLW and 1OFZ have found good ‘docking’ positions, defined similarly in section 3.10 as RMSD less than 4 Å, against the first step of trajectory as they do not fluctuate much. As by comparison to their initial dock (which is usually similar to frame zero of the simulation, if equilibration has a small effect) they do not leave the pocket they docked to; they have a low RMSD of around 0.8 and 1.5 Å respectively during the majority of the run.

1MLW however docked incorrectly according to the initial docking results of the GLIDE docking. Further, the idea that the ligand in 1MLW is incorrectly docked is corroborated by the comparison to the crystal structure where their RMSDs are particularly high to start with but it gradually improves. The initial pose used for the simulation was however the best pose according to their GLIDE score despite not being the best RMSD pose. There is no binding information for 1MLW. This means that the glide docking has found a better pose for it and the polarisation has not changed that. This could mean that the binding pocket in the crystal structure is not a particularly good target or that when the ligand is binding to the structure it binds to other alternative sites as seen similarly in section 3.10. 1OFZ according to the SERAphic paper (Favia et al., 2011) also has a good experimental binding affinity of 24.1 µM , which gives it a consistently good RMSD as seen in figure 4.3.

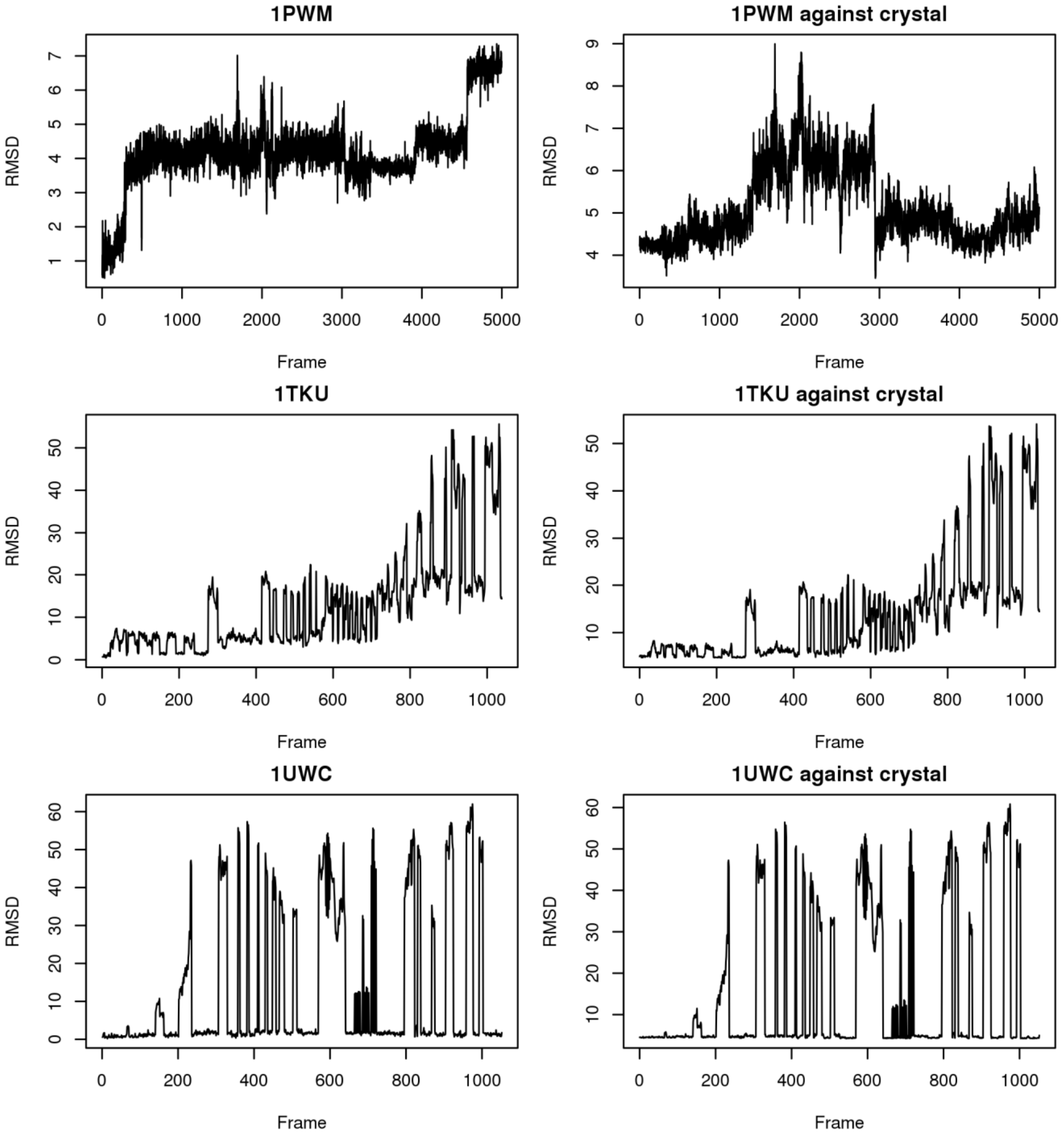


Figure 4.5. RMSD (in Å) for 1PWM, 1TKU and 1UWC against the initial docked pose and against the crystal structure. The total length of the simulation is 200 ns.

In figure 4.5 we again see some interesting behaviour due to the simulation with 1PWM. According to the RMSD results against frame zero, the ligand begins to leave its docked pose and continues to go further away from the docked position. Although when we compare it to the RMSD results against the crystal structure, shown in figure 4.5, we see that 1PWM is initially docked nearby, partly in the correct position, it then leaves the binding pocket and attempts to correct itself by returning close to the crystal structures position. So, despite the good GLIDE score, the ligand leaves the binding site but then the simulation attempts to return it to the correct position.

The ligand in 1TKU starts off bound in the right position according to the crystal structure and is ejected from the binding pocket. This is caused by oscillation of the ligand within the binding site. The ligand appears to be bound weakly and builds up enough energy to leave the binding site. There is no binding data on the ligand. To some extent the ligand attempts returns to the binding site, but is finally fully ejected. 1UWC shows an interesting motion of the ligand in both graphs. When looking at the graphics of the simulation, we see that there is actually a breathing movement to the protein. The ligand does leave the binding pocket for small bursts of time but due to the motion of the protein and how it binds to the sides of protein, it is pushed back into the pocket by intermolecular forces. This protein also had a problem in simulation that we were not able to correct. As 1UWC breathed it pushed the ligand to the edge of the periodic boundary however during this simulation ACEMD was set to wrap the boundary box before trajectory completion in error. This means the ligand appears to be on the opposite side of the system despite periodic boundary conditions being in effect. Thus it has given erroneous spikes where the RMSD appears to be 50 Å; a similar effect is also seen for 1M2X (Fig 4.2) and 1TKU (Fig 4.5). In this case the lower RMSDs are more indicative of the true value.

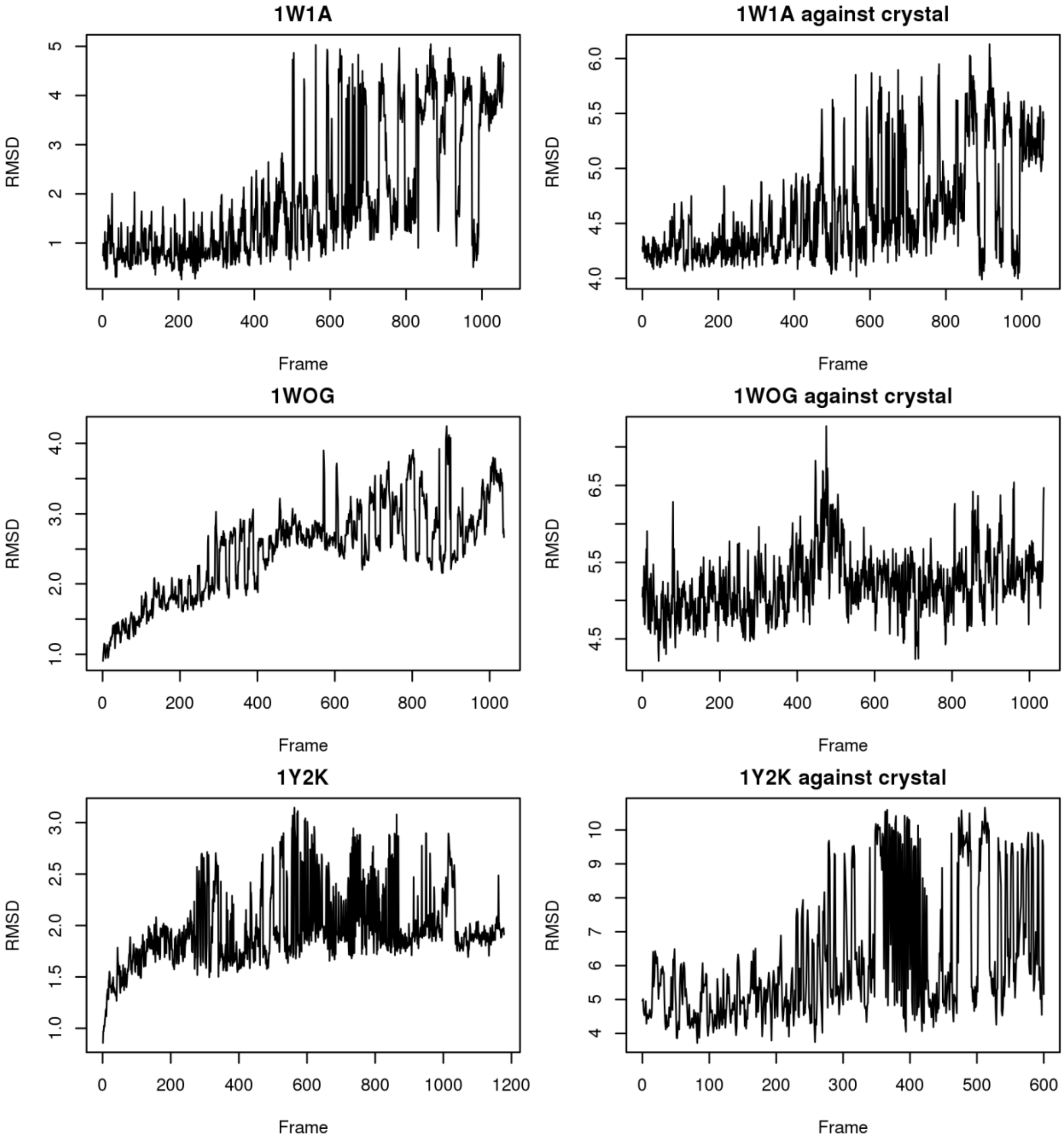


Figure 4.6. RMSD (in Å) for 1W1A, 1WOG and 1Y2K against the initial docked pose and RMSD based on the crystal structure. The total length of the simulation is 200 ns.

Figure 4.6 shows that 1W1A has found a slightly different pocket to that in the crystal structure. Starting at 4.0 Å for 1W1A, the ligand begins to leave the docked position. This is the position it found with the highest glidescore but in this case it is not well suited to binding this ligand in place. The oscillations observed are similar to those observed for 1TKU and 1UWC, as such the ligand returns for small periods of time. There isn’t any binding data for 1W1A.

1WOG starts as a quite well-bound structure and then fluctuates within its binding site, the RMSD fluctuates between 1 Å and 3 Å. However this is still within the pocket. This is one of the simple cases of a correct dock.

1Y2K is another ligand that exhibits some interesting behaviour. When compared to frame zero, it flips over according to the graphics and then fluctuates in that state. However, upon examining what it does compared to the crystal structure it is docked in a different pose then continues to move away and comes back in a constant motion. Looking at the graphics of the simulation in figure 4.7, this is due to part of the ligand anchoring itself to the binding pocket while the rest of the ligand then moves around in the binding site, shifting and rotating around this anchor. So, despite this high RMSD fluctuation the ligand is still bound well enough for a fragment. The benzene ring on the top part of the ligand is the fluctuating portion.

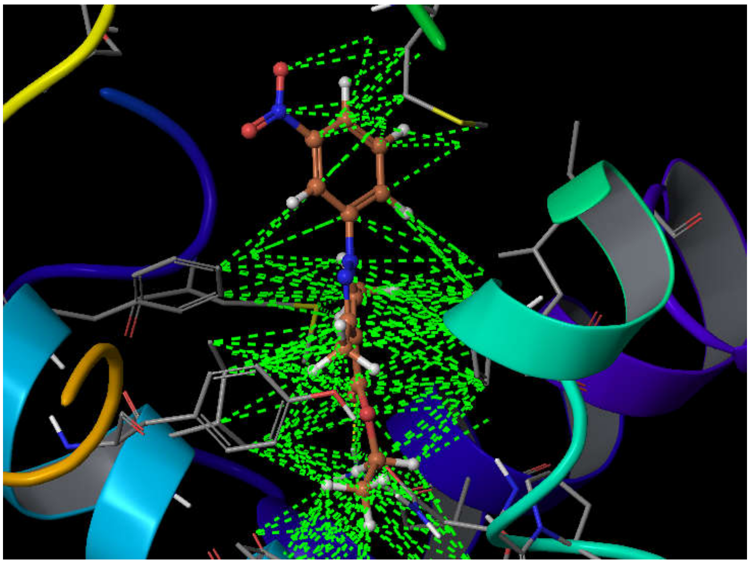


Figure. 4.7 The ligand (orange) of 1Y2K (ribbons). Receptor - ligand interactions are shown as green dotted lines.

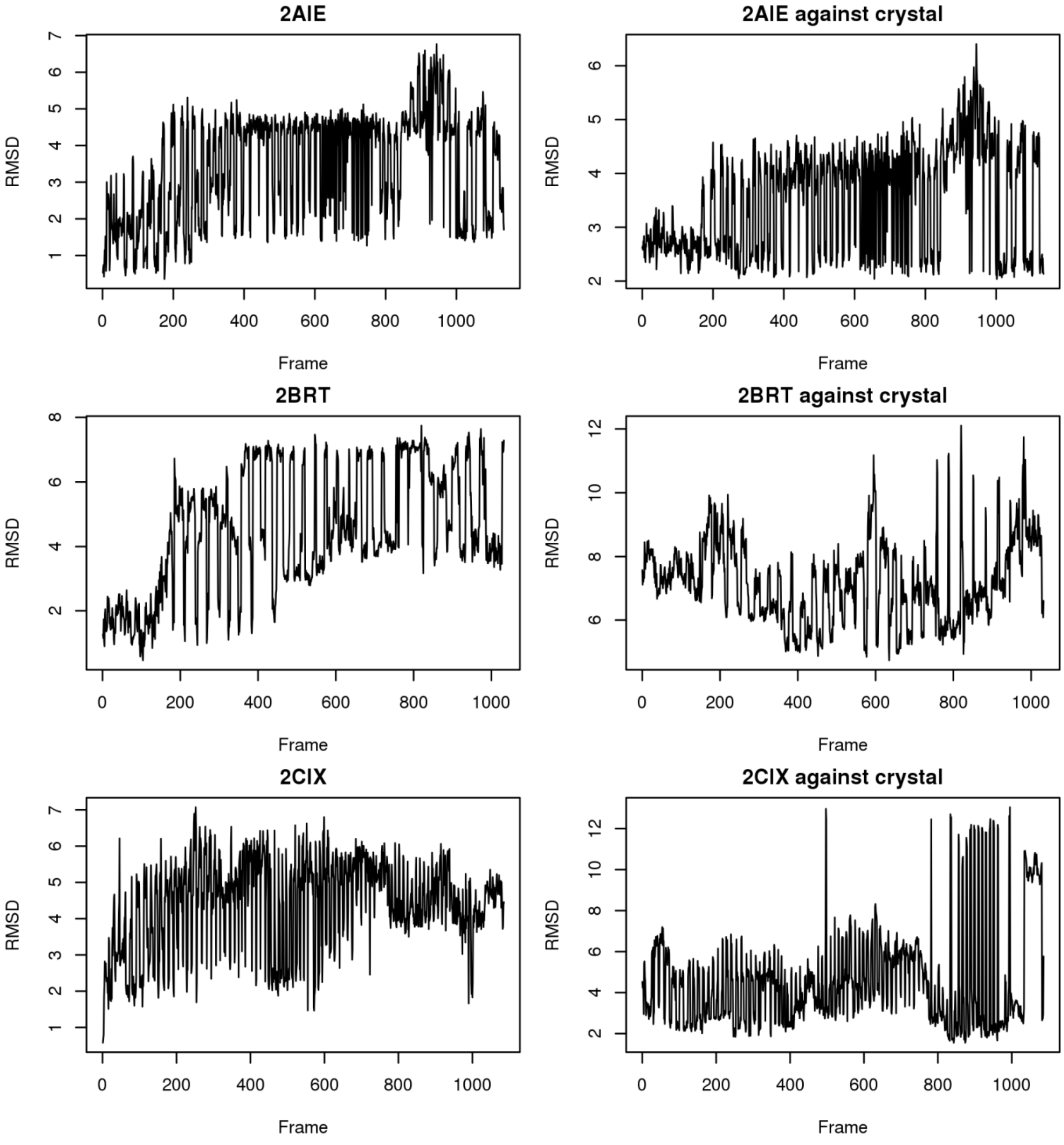


Figure 4.8. RMSD (in Å) for 2AIE, 2BRT and 2CIX against the initial docked pose and RMSD based on the crystal structure. The total length of the simulation is 200 ns. 2CIX redock was used in simulation.

In Figure 4.8 we see for 2AIE that in both the crystal structure-based RMSD and the frame zero-based RMSD, there is a fluctuation of about 2 Å. This is similar to 1Y2K, where part of the ligand has bound well whereas the rest of it is oscillating within the place it is bound in, as shown in figure 4.9. This again is similar to 1Y2K where since the ligand is a fragment and only part of the ligand is bound instead of the whole molecule like a drug. This could be useful knowledge when building linkers or seeking to identify a decent point to grow such a fragment into a lead compound.

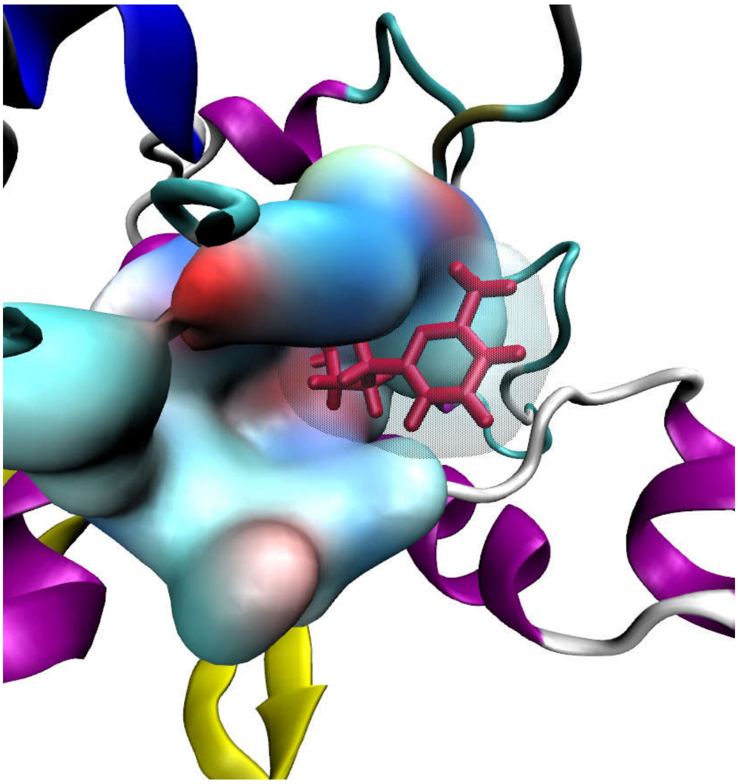


Figure 4.9. A graphical representation of the 2AIE binding site. The residues around the ligand are shown as a molecular surface.

2BRT is one of the ligands showing some more promise. According to frame zero it leaves where it’s bound and moves further and further away. However when compared to the crystal structure it actually begins to get closer towards the crystal binding site. It doesn’t quite get there in over 200 ns of MD simulation but looking at the graphics in the simulation we see that it attempts to get closer than its initial position.

2CIX exhibits some very sharp fluctuations within its simulation and appears to be leaving the binding site continuously. Due to its small size we can't say that only part of the ligand is fluctuating as the ligand is only a small fructose ring. Looking at the simulation, what is happening is that protein is showing a breathing moment around the binding site. The site is actually very small and a tight fit even for this small ligand. The ligand appears to bind to a fold in the protein where it binds to Val182 and Glu183. The ligand is then pushed in when the protein 'breathes in' then is ejected again when the protein 'breathes out'. This motion happens quite consistently and the ligand continually returns to its low RMSD. This motion is shown in the PCA figure 4.10, the arrows show which parts of the ligand are moving significantly. We can see to the right of the ligand the breathing portion which opens up to let the ligand out. The constant leaving and returning the ligand exhibits shows that it can consistently bind to the site with a high affinity and in nature would likely begin a reaction to bind fully instead of being ejected out again.

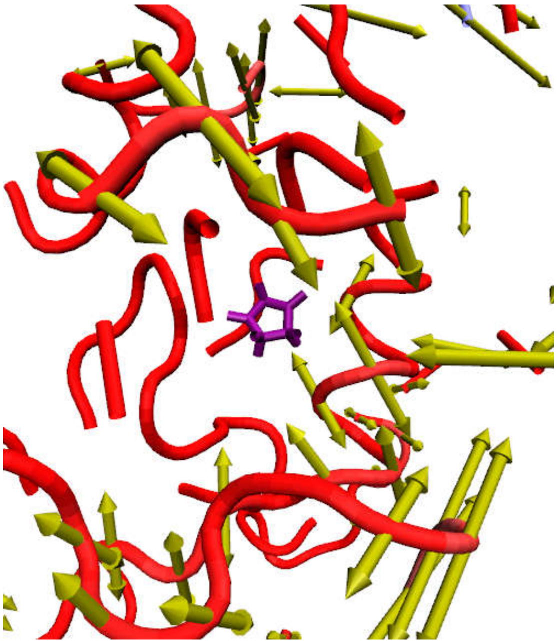


Figure 4.10. A graphical representation of a PCA for 2CIX.

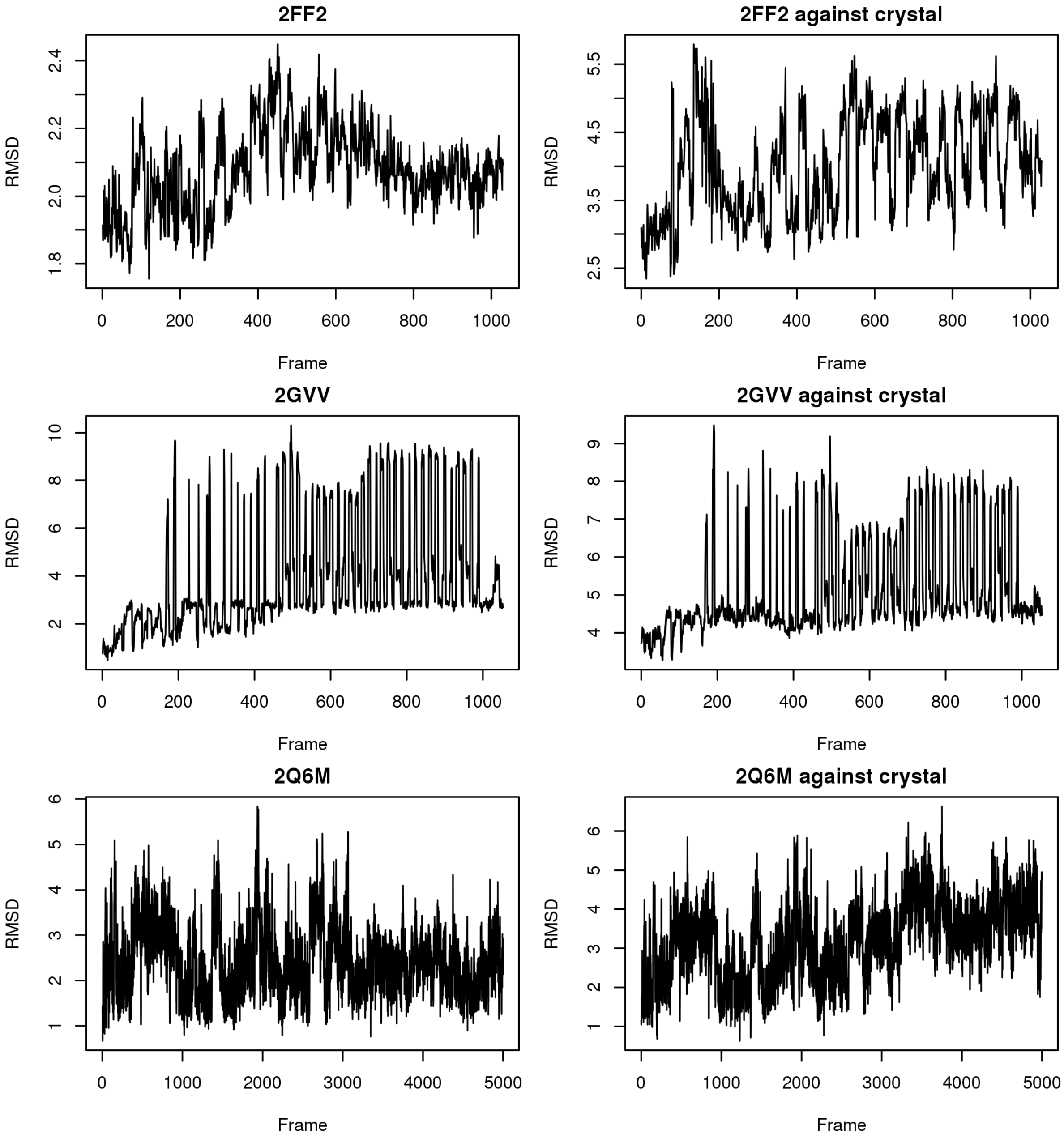


Figure. 4.11. RMSD (in Å) for 2FF2, 2GVV and 2Q6M against the initial docked pose and RMSD based on the crystal structure. The total length of the simulation is 200 ns.

In Figure 4.11, 2FF2 in both frame zero and against the crystal structure fluctuates around 2 Å within the binding site, thus it is bound but relatively weakly.

2GVV has very sharp spikes, similar to what was seen in 2CIX but exhibits behaviour similar to 1Y2K. When looking at the graphics of the simulation in figure 4.12, the ligand in 2GVV has an anchor point then fluctuates between 2 poses with the chain on one side then on the opposite side. It is doing so quite suddenly which explains the spikes.

The ligand in 2Q6M begins in its binding site and begins to leave, rising to an RMSD of ~ 6 Å and then returns, with an RMSD of ~ 2 Å. It had been bound close to the crystal structure in its initial pose, which is why there is barely any difference between the two plots for 2Q6M in figure 4.10. For 2Q6M K_d = 510 nM as seen in table 2.1 which could explain why it is bound in a small area.

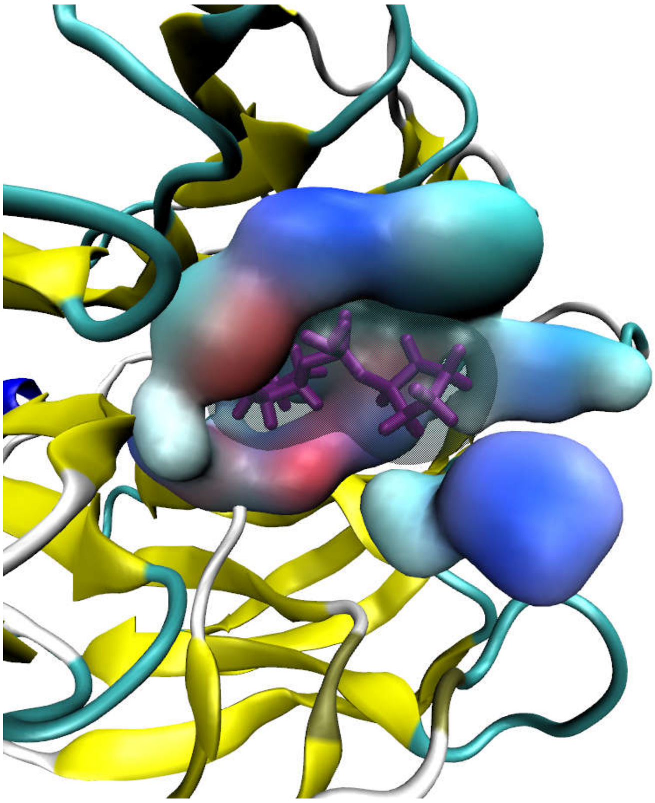


Fig. 4.12. A graphical representation of the binding site of 2GVV. The ligand is in purple, with a transparent molecular surface; neighbouring protein residues are denoted by a coloured molecular surface.

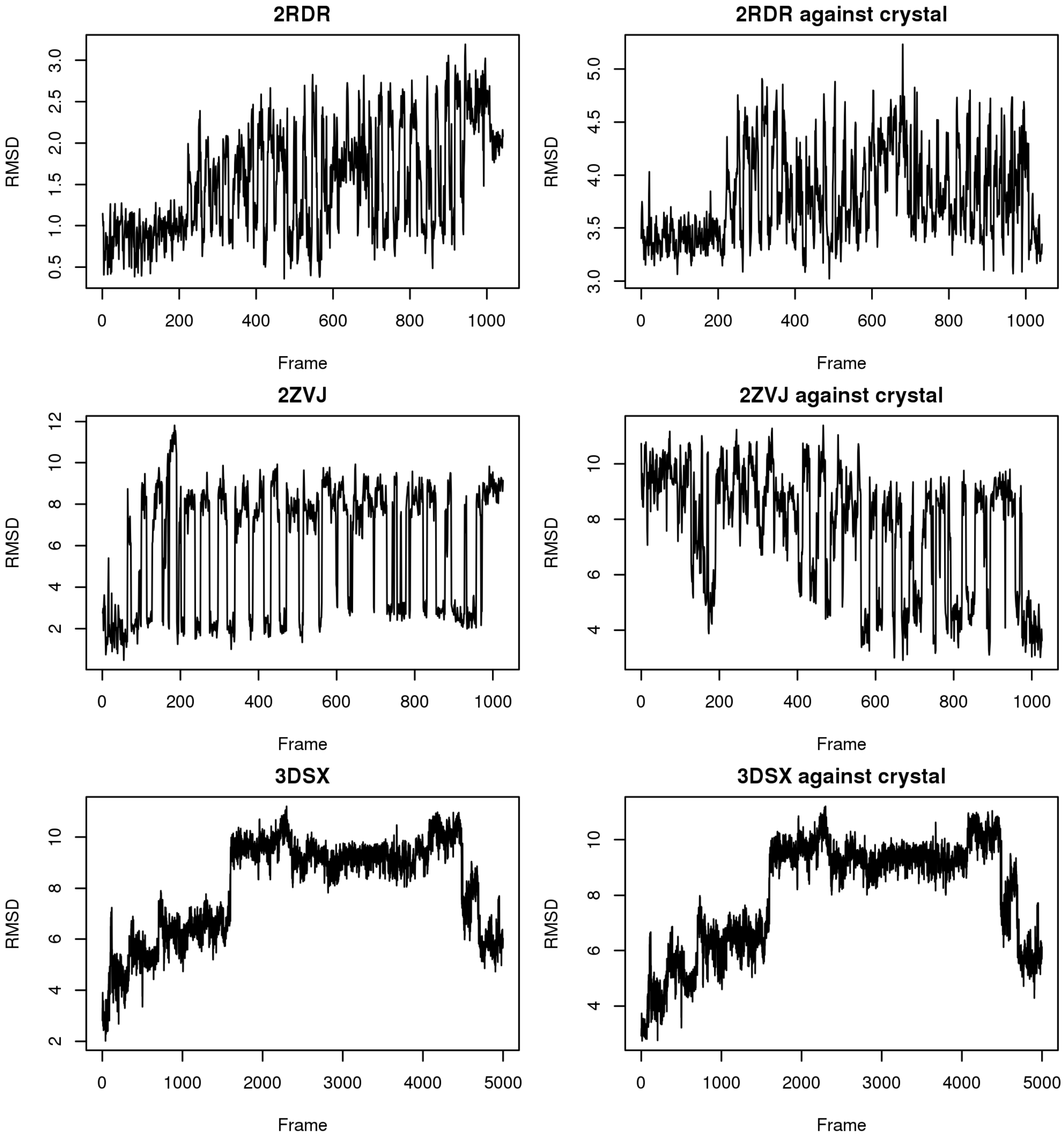


Figure 4.13. RMSD (in Å) for 2RDR, 2ZVJ and 3DSX against the initial docked pose and RMSD based on the crystal structure. The total length of the simulation is 200 ns.

Figure 4.13 shows more examples of the ligands oscillating into and out of their crystal structure poses, despite the high glidescore they were given in the original docking and polarized re-docking, Table 2.1. 2RDR appears to be leaving the binding site according to the RMSD relative to frame zero but it is moving away from the site to get closer towards its crystal structure binding site.

2ZVJ is similar in that regard where it attempts to leave the initial (top glide pose / step 0) binding site, it is oscillating quickly to build up momentum to leave, as reflected in the many dips and troughs in RMSD. It then finally leaves. What is happening better shown against the crystal structure in figure 4.13: is it is leaving a binding site that is very far away from the crystal structure pose. The ligand breaks away from its top glide pose position then moves towards its experimental binding site in the crystal structure. It hasn't made it to a low RMSD of 2 Å. However, it is attempting to rearrange itself in the binding site at the end of the simulation.

3DSX is leaves the binding site only to return to the same one in a slightly different part of the pocket. The ligand for 3DSX has a K_d of 1.4 mM. It leaves the binding site and moves around on the outside of the protein before being pushed back in and rebinding to the pocket.

4.5 Discussion

As in chapter 3 (GROMACS simulations), there were several ligands that continued to show expected behaviour by binding strongly and staying with relatively low RMSD (<4 Å) such as 1F5F, 1F8E, 2RDR and 1WOG, continuing the trend that tightly bound fragments will stay bound tightly. However, whereas in Table 2.1 the RMSDs are generally below 2 Å, the RMSDs obtained from the simulations are higher, partly because these are dynamic systems and so it is inevitable that the breaking and reforming of hydrogen bonds will yield a higher RMSD. Fragments showing this type of behaviour under MD may be the best candidates for experimental testing, or if the experimental screening is positive, for optimization into a lead compound.

Some ligands went through a rapid oscillating motion showing one of two things. Either the ligand is moving around rapidly it its binding site or part of the ligand is anchored down while part of the ligand is oscillating freely causing a higher RMSD such as 2GVV (Figure 4.11).

Other ligands were docked with a low RMSD were then ejected from the binding site, then returning to the binding site later in the simulation. Examples include 2RDR and 3DSX (Figure 4.13). This potentially in part shows the ligand kinetics as the dissociation constant is an equilibrium constant that measures the propensity of the products (protein complex) to dissociate reversibly into its reactants (ligand and protein). This can be shown in MD as the ligand moving to infinity, or at least moving a reasonable distance away where the attractive forces are low or insignificant. . Some of the ligands become trapped in their pocket by the conformation of the protein as seen in 1M3U (Fig 4.4). The flap that is keeping the ligand inside the binding pocket feeds into the ligand kinetics as this mechanism shows how it increases the on rate. Thus the MD simulations have the potential to show specific mechanisms of ligand binding in proteins.

Some ligands eventually built up enough energy to move to a new binding site and fluctuate in that region such as 2CIX which ejects part way through the simulation to find a new site. So it's possible that we can use the MD methods to find other positions for the ligands to bind as the ligand was fluctuating mostly in this new area.

The longer simulations made possible by the more powerful GPUs has allowed a better look at the behaviours of several of the ligands. 2CIX for example moved out of its binding site and then returned as the protein breathed. This return could not be seen in the very short 10 ns MD runs which we performed with GROMACs (Chapter 3, Figure 3.5). However, the oscillations in Figure 4.8 suggest that even with a reasonably low RMSD after the redock, because of its poor binding and the breathing nature of the protein it will probably continue to leave that binding site.

1MLW, as shown before (Figure 4.3), seems content in its new binding site, fluctuating little and not leaving it, suggesting further that it was correctly docked (into an allosteric binding site) despite the higher RMSD at the initial phase of docking.

With this set we also have another poorly docked ligand in 1M3U. The ligand in 1M3U has found a different conformation in the binding pocket in which to bind after simulation. After equilibration it moved closer to its crystal structure position, giving an RMSD of 3 Å instead of an RMSD of 6 Å, as in the initial docking. However, during the simulation it continued to move to where it was originally docked without polarization (Table 2.1, ‘before’ column, RMSD ~ 2 Å).

Ligands in proteins such as 1PWM, despite the low RMSDs from docking (Chapter 2 / Table 2.1), could be ejected and slowly leave the binding site during the MD. 1TKU exhibits similar behaviour but it was not docked that well, with the RMSD of initial and polarized docking both at around 3 Å. The similar behaviour between these two show that both ligands can be ejected despite obtaining a good conformation in the binding site, showing possibly that strong binding affinity is also needed to serve as a good anchor (K_D for 1TKU is unknown). The polarization results could suggest that 1PWM, despite being able to get a good conformation (low RMSD), actually doesn’t have a good charge complementarity in that binding area, possibly because the K_D is low; IC_{50} = 935 nM (Table 2.1).

However, sometimes the RMSDs can be misleading and extra care should be taken to look at the trajectory in motion, for example with 2AIE and 2GVV. In the shorter simulations, both were found to have tighter binding and did not fluctuate much. However when looking at their RMSDs, it appears as the ligands act poorly by not staying bound. Part of the ligand is still bound to the correct binding site but due to the length of the ligand in both of these cases, only part of the fragment finds a strong anchor point and the rest continues to fluctuate in the binding site. This is interesting information though and can be used when growing the ligand out in an FBDD programme. If we examine the ligand with the MD and find that one side of a ligand has a clear anchor point and the other does not, we can grow the ligand out from the section that is fluctuating a lot. We could also possibly try to find a different part of the pocket that we could grow the ligand out into and perform docking experiments to find a new fragment that has a good conformation in that site, and then combine the two. In either FBDD approach, be it growing or combining, we can then do a retest of the fragment in the binding area to see if the fluctuations are far less. If the new end is bound tightly then the new ligand should have fewer fluctuations.

4.6 References

BROOKS, B. R., BROOKS, C. L., MACKERELL, A. D., NILSSON, L., PETRELLA, R. J., ROUX, B., WON, Y., ARCHONTIS, G., BARTELS, C. & BORESCH, S. 2009. CHARMM: the biomolecular simulation program. *Journal of computational chemistry,* 30**,** 1545-1614.

CASE, D. A., CHEATHAM, T. E., DARDEN, T., GOHLKE, H., LUO, R., MERZ, K. M., ONUFRIEV, A., SIMMERLING, C., WANG, B. & WOODS, R. J. 2005. The Amber biomolecular simulation programs. *Journal of computational chemistry,* 26**,** 1668-1688.

CASE, D. A., J.T. BERRYMAN, R.M. BETZ, D.S. CERUTTI, T.E. CHEATHAM, I., T.A. DARDEN, R.E. DUKE, T.J. GIESE, H. GOHLKE, A.W. GOETZ, N. HOMEYER, S. IZADI, P. JANOWSKI, J. KAUS, A. KOVALENKO, T.S. LEE, S. LEGRAND, P. LI, T. LUCHKO, R. LUO, B. MADEJ, K.M. MERZ, G. MONARD, P. NEEDHAM, H. NGUYEN, H.T. NGUYEN, I. OMELYAN, A. ONUFRIEV, D.R. ROE, A. ROITBERG, R. SALOMON-FERRER, C.L. SIMMERLING, W. SMITH, J. SWAILS, R.C. WALKER, J. WANG, R.M. WOLF, X. WU, YORK, D. M. & KOLLMAN, A. P. A. 2015. AMBER 2015. University of California, San Francisco.

CROWLEY, M. F., WILLIAMSON, M. J. & WALKER, R. C. 2009. CHAMBER: Comprehensive support for CHARMM force fields within the AMBER software. *International Journal of Quantum Chemistry,* 109**,** 3767-3772.

FAVIA, A. D., BOTTEGONI, G., NOBELI, I., BISIGNANO, P. & CAVALLI, A. 2011. SERAPhiC: A benchmark for in silico fragment-based drug design. *Journal of chemical information and modeling,* 51**,** 2882-2896.

GREGOR, D., WILLCOCK, J. & LUMSDAINE, A. 2006. Concepts for the C++ 0x Standard Library: Utilities.

GUPTA, S. & BABU, M. R. 2011. Performance Analysis of GPU compared to Single-core and Multi-core CPU for Natural Language Applications. *International Journal of Advanced Computer Science and Applications,* 2.

H.M. BERMAN, J. WESTBROOK, Z. FENG, G. GILLILAND, T.N. BHAT, H. WEISSIG, I.N. SHINDYALOV & BOURNE, P. E. 2000. *The Protein Data Bank* [Online]. Nucleic Acids Research, 28: 235-242. Available: www.rcsb.org [Accessed 22 Sep 2015 2015].

HALGREN, T. A. 1996. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of computational chemistry,* 17**,** 490-519.

HARVEY, M. J. & DE FABRITIIS, G. 2011. Swan: A tool for porting CUDA programs to OpenCL. *Computer Physics Communications,* 182**,** 1093-1099.

HARVEY, M. J., GIUPPONI, G. & FABRITIIS, G. D. 2009. ACEMD: accelerating biomolecular dynamics in the microsecond time scale. *Journal of chemical theory and computation,* 5**,** 1632-1639.

HESS, B., BEKKER, H., BERENDSEN, H. J. & FRAAIJE, J. G. 1997. LINCS: a linear constraint solver for molecular simulations. *Journal of computational chemistry,* 18**,** 1463-1472.

JO, S., KIM, T., IYER, V. G. & IM, W. 2008. CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry,* 29**,** 1859-1865.

KARPLUS, M. 1983. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem,* 4**,** 187217.

KIRK, D. NVIDIA CUDA software and GPU parallel computing architecture. ISMM, 2007. 103-104.

KRÄUTLER, V., VAN GUNSTEREN, W. F. & HÜNENBERGER, P. H. 2001. A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *Journal of computational chemistry,* 22**,** 501-508.

MACKERELL, A. D., BANAVALI, N. & FOLOPPE, N. 2000. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers,* 56**,** 257-265.

MACKERELL, A. D., FEIG, M. & BROOKS, C. L. 2004. Improved treatment of the protein backbone in empirical force fields. *Journal of the American Chemical Society,* 126**,** 698-699.

MAIER, J. A., MARTINEZ, C., KASAVAJHALA, K., WICKSTROM, L., HAUSER, K. E. & SIMMERLING, C. 2015. ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of Chemical Theory and Computation,* 11**,** 3696-3713.

NVIDIA. 2015. *Nvidia* [Online]. Available: http://www.nvidia.co.uk/object/geforce-desktop-graphics-cards-uk.html [Accessed 22 Sep 2015 2015].

NVIDIA, C. 2008. Programming guide.

POMÈS, R. & MCCAMMON, J. A. 1990. Mass and step length optimization for the calculation of equilibrium properties by molecular dynamics simulation. *Chemical Physics Letters,* 166**,** 425-428.

RYCKAERT, J.-P., CICCOTTI, G. & BERENDSEN, H. J. 1977. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics,* 23**,** 327-341.

SHAMS, R. & KENNEDY, R. Efficient histogram algorithms for NVIDIA CUDA compatible devices. Proc. Int. Conf. on Signal Processing and Communications Systems (ICSPCS), 2007. Citeseer, 418-422.

WANG, J., WANG, W., KOLLMAN, P. A. & CASE, D. A. 2006. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of molecular graphics and modelling,* 25**,** 247-260.

WANG, J., WOLF, R. M., CALDWELL, J. W., KOLLMAN, P. A. & CASE, D. A. 2004. Development and testing of a general amber force field. *Journal of computational chemistry,* 25**,** 1157-1174.

ZIMRNERMANN, K.-H. 2003. An introduction to protein informatics. *Identity,* 37**,** 30.

Chapter 5

Membrane simulations for wildtype thermostabilized β_1-adrenergic receptor

5.1 Introduction

In a seminal paper Serrano-Vega et al. (2008) described conformational thermostabilization of a G-protein coupled receptor. This facilitated crystallization of a β_1-adrenergic receptor (βAR), the structure of which was found by Warne et al. (2008). Serrano-Vega et al. (2008) indicated that thermostabilization may greatly increase the ease with which X-ray crystallographic structures of integral membrane drug targets such as G-protein coupled receptors (GPCRs) can be obtained. Such structures are highly desirable in a FBDD programme. Despite much success since as described by Congreve et al. (2011), the process of obtaining structures of these important medicinal chemistry targets is not necessarily straightforward. For example, the mutations identified by thermostabilization experiments shown in Shibata et al. (2009) were not the ones that yielded an X-ray crystal structure; the resultant structure did not therefore have the motifs or activity that was originally expected, as seen in White et al. (2012).

A serious limitation to rational drug design for GPCR targets is the lack of 3D structural information. The crystallographic determination of GPCR structures remains difficult because of low expression levels as shown in Chelikani et al. (2006), non-homogeneous modifications (such as glycosylation) as seen in Reeves et al. (2002), folding problems in bacteria demonstrated by the work of Baneres et al.

(2003), instability in detergents and multiple conformational states shown in Schwartz et al. (2006). Thus, Serrano-Vega et al. (2008) devised a generic strategy for producing detergent-stable eukaryotic integral membrane proteins. Their comprehensive mutagenesis study involved over 300 mutations, including residues in all seven transmembrane domains, yielding 18 mutants with increased stability. Promising mutations were combined to generate βAR-m23 in which the Tm (the temperature at which a reference ligand can bind to 50% of receptors) had risen to 53°C from 32°C for the β_1-AR wild-type (βAR-wt). The X-ray structure of this stabilized form of the receptor in complex with the inverse agonist cyanopindolol was subsequently determined as seen in the results of the paper in Warne et al. (2008). However, neither the resultant βAR-m23 crystal structure nor the analysis of the equivalent mutated positions in the rhodopsin structure gave any indication as to the origin of the enhanced stability as seen in Warne et al. (2008) and Serrano-Vega et al. (2008). Moreover, since similar stabilizing mutations in the adenosine, chemokine and neurotensin receptors as seen in the papers by Shibata et al. (2009), Lebon et al. (2011a), Lebon et al. (2011b) and Wu et al. (2010) did not occur at equivalent positions, it seems that in its current form, conformational thermostabilization is difficult to apply in a rational way.

Other researchers such as Balaraman et al. (2010) and Chen et al. (2012) have studied packing effects in GPCR stability, RMSF and connectivity analysis as seen in Simpson (2011). Here we consider RMSD and root mean square fluctuation (RMSF) analysis as a measure of thermostabilization.

This study is to investigate whether computational methods can be used to understand and model the stability of mutant receptors. The hypothesis is that by using these computational methods, if the results are as expected, it should be possible to gain understanding and prioritize and study proposed mutants prior to experimental assessment.

5.2 Methods

The mutant of a β_1-adrenergic receptor, pdb code 2VT4, found by Serrano-Vega et al. (2008), was shown to be thermodynamically stabilized. The topology for the wild type was found in Yarden et al. (1986). According to Serrano-Vega et al. (2008), this thermostability is attained by the following mutations: R68$^{1.59}$S, M90$^{2.53}$V, Y227$^{5.58}$A, A282$^{6.27}$L, F327$^{7.37}$A and F338$^{7.48}$M. Where 68 is the beta(1)-AR residue number and 1.59 is the Weinstein and Ballasteros universal number (Ballesteros and Weinstein, 1995) in which 1 refers to the helix number, and position 59 is the most conserved residue in the helix. Here, we investigated these findings with molecular dynamics. The aim was to observe whether during an MD simulation the mutant was more stable than the wildtype. We measured the RMSD (as a measure of stability) over the course of 200 ns and we also measured the RMSF across both proteins. Afterwards, we performed tests to ascertain if temperature had any effect on the RMSF of each protein and if there was any temperature where one protein was more stable than the other. In addition, we investigated the stability with and without the ligand. The hypothesis is that the RMSD and RMSF data should correlate with Serrano-Vega et al. (2008) where the mutant should be more stable than the wildtype – provided that RMSD and RMSF are related to thermostability.

5.2.1 Preparation of the membrane around the proteins

The two proteins were prepared separately, namely the wildtype and the mutant. The mutant was taken from the RSCB; The wildtype was developed by comparative modelling using Modeller 9.12 as the best structure from 100 alternative models (Eswar et al., 2007). Each protein was prepared in the identical process shown in section 4.3.

After the basic preparation and minimisation, we used the desmond membrane package from maestro to generate a POPC lipid bilayer. The proteins were duplicated and the ligand DHA was docked to the protein. At this point, there were 4 prepared proteins with membranes; wildtype with and without ligand, mutant with and without ligand. Each of these proteins were exported from maestro and converted into the amber format. A water bath and constraint file was generated in the same manner as in section 4.3. A set of typical input files is found in appendix 5.1.

5.2.2 MD simulation

Full details of the input files that specify the details of the MD simulations can be found in Appendix 5.2. The typical parameters are summarized below.

There were 6 production runs which are as follows: Wildtype, Wildtype with ligand, Wildtype at various temperatures, Mutant, Mutant with ligand, Mutant at various temperatures.

All proteins were minimised for 5000 steps each. They were then equilibrated using NVT for 2 ns and NPT with constraints for 20 ns followed by 20 ns without constraints. The time step was set to 4 fs; this large step was made possible by increasing the mass of the hydrogen atoms to 4 da. The temperature was set to 300 K. Pressure was set to the default of 1.01325 bar. The production run for each protein was 200 ns.

For the temperature production runs on each protein a tcl script was appended to the input file. The tcl script increased the temperature every 40 ns. The temperature started at 300 K and at every 40 ns was increased by 8 K.

5.2.3 Converting membranes for use in ACEMD

As discussed in section 5.2.1, ACEMD has no native program that can generate a membrane. Options were available for generating a membrane through the CHARMM-GUI or by using Maestro. The CHARMM-GUI seemed the simpler option as CHARMM and amber shared similar formats. However, the mutated 2VT4 protein was generated through modelling and was not completely standardized. Due to this, the CHARMM-GUI did not recognize any of our files correctly and would generate membranes without a protein present. Maestro however could read the protein and generate a membrane for each protein, so this option was chosen. Maestro again has its own format and had to be converted to be recognized by amber. The membrane had to be converted using old conversion scripts found in the appendix, the scripts were made by Ian Gould (Imperial College London) and Ross Walker (UCSD).

5.3 Results

The βAR-m23 and βAR-wt receptor structure is shown in Fig. 5.1

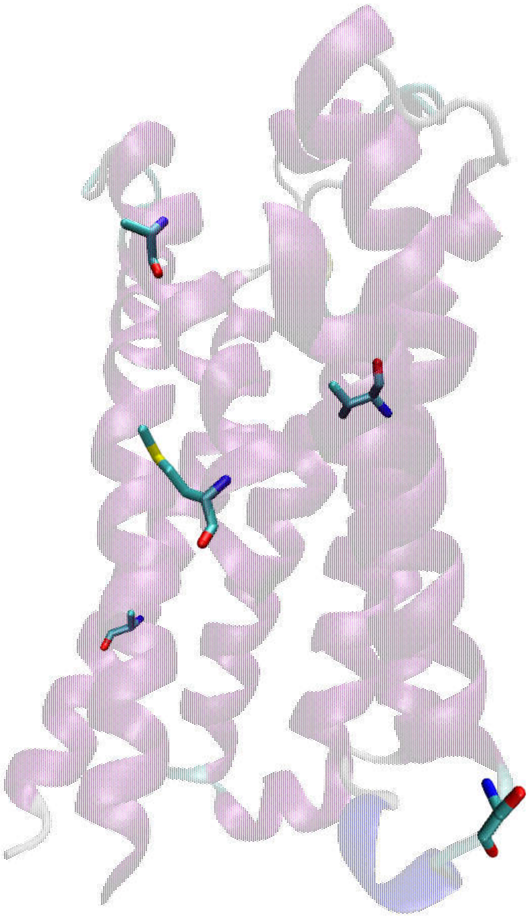


Figure 5.1. The β_1-AR model; the mutated amino acids are shown as sticks.

The RMSDs of βAR-m23 and βAR-wt over the course of a 200 ns simulation and at different temperatures are shown in Figure 5.2. Each protein appears to be reasonably stable with just modest fluctuation in RMSD. βAR-m23 (blue) however across the different runs has a consistently lower RMSD than βAR-wt, which is in line with expectations, showing more thermostability with the 2VT4 mutant than with WT. One of the reported effects of thermostabilization is to freeze the protein in a particular state, which may be an inactive form, if an inverse agonist was used as the ligand in the stabilization process, or it may be an active form if an agonist was used as shown in Serrano-Vega et al. (2008). In this case, the ligand was an inverse agonist, so while βAR-wt may sample active and inactive conformations, βAR-m23 should only sample inactive conformations and so should have a lower RMSD. The expected result in Figure 5.2a-e suggests that average RMSD is a good indicator of receptor stability. Consequently, we have also analysed the RMSF over all atoms of the receptors in the absence of ligand; these results are given in Figure 5.3.

Figures 5.2a-e shows that the RMSD of the average structure for the βAR-m23 mutant at each temperature is lower, with the exception of 324 K. At 324 K the wildtype has a lower RMSD for ~20 ns before 2VT4 once again has the lower RMSD. This lower RMSD for the βAR-m23 suggests that it is more thermodynamically stable, which matches the findings of the experimental data.

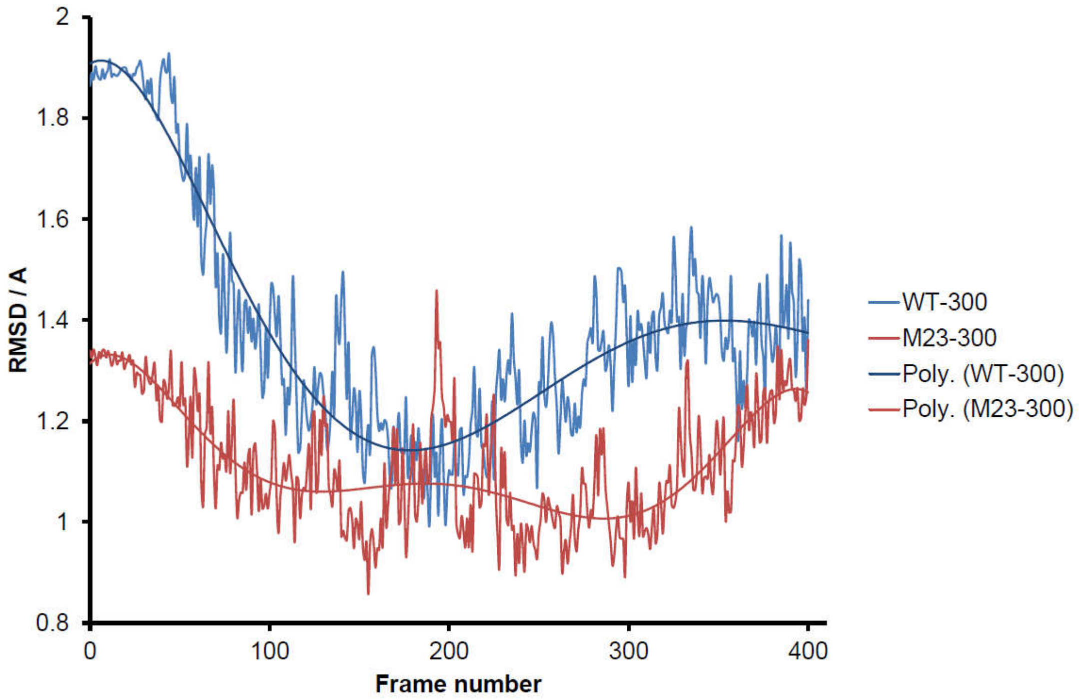
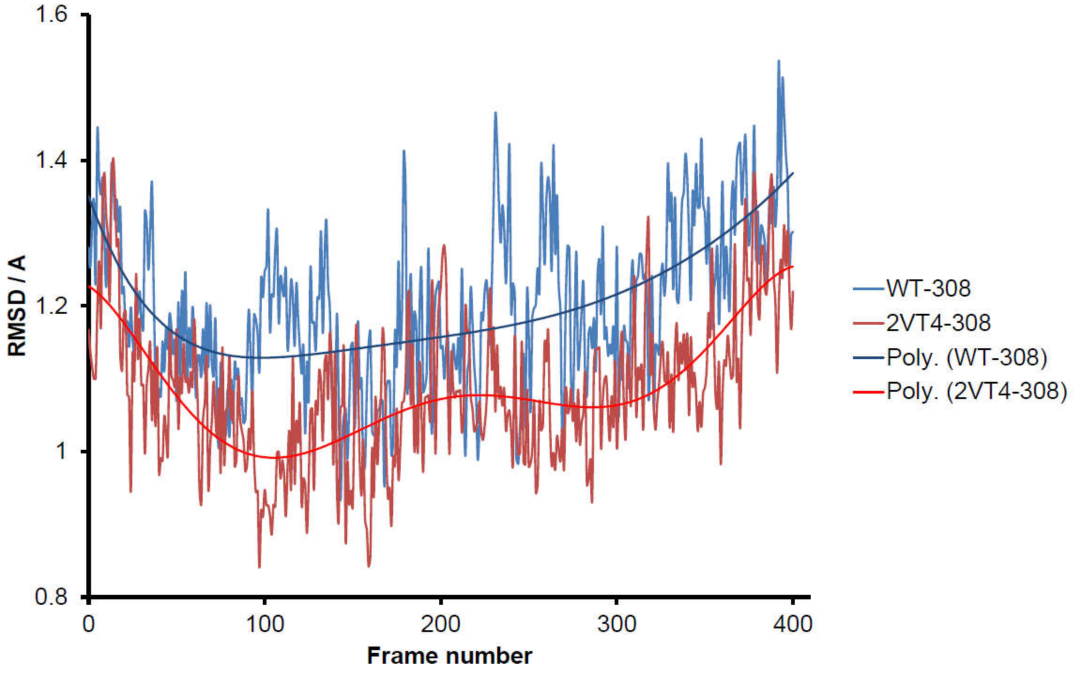


Figure 5.2a the RMSD (in Å) against the average structure of βAR-m23 (red) and βAR-wt (wildtype) over the full 200 ns of the production run. The RMSD is for the backbone of the protein excluding any hydrogens; this figure is for 300k. Each frame is 200 ps. A 6th order polynomial was fitted to the RMSD data, denoted Poly.(WT-300 and 2VT4-300).



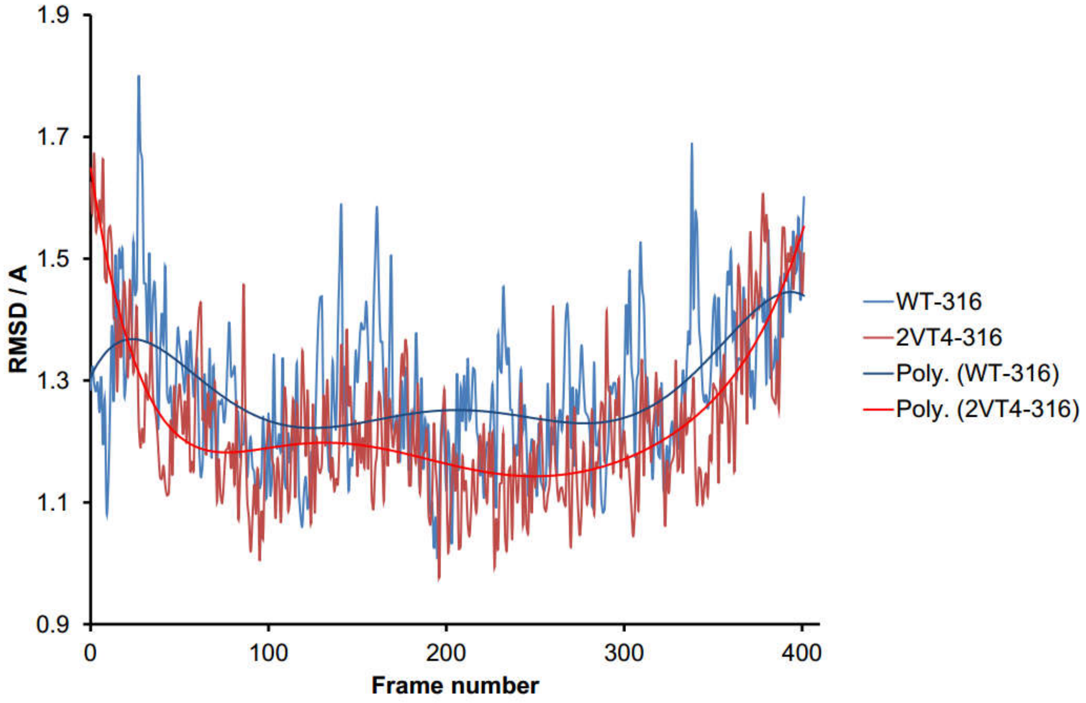


Figure 5.2b(top) and 5.2c(bottom) the RMSD (in Å) against the average structure of βAR-m23 (red) and βAR-wt (wildtype) over the full 200 ns of the production run. The RMSD is for the backbone of the protein excluding any hydrogens; figure 5.2b is for 308k and figure 5.2c is for 316K. Each frame is 200 ps. A 6th order polynomial was fitted to the RMSD data, denoted Poly.(WT-308 or 316 and 2VT4-308 or 316).

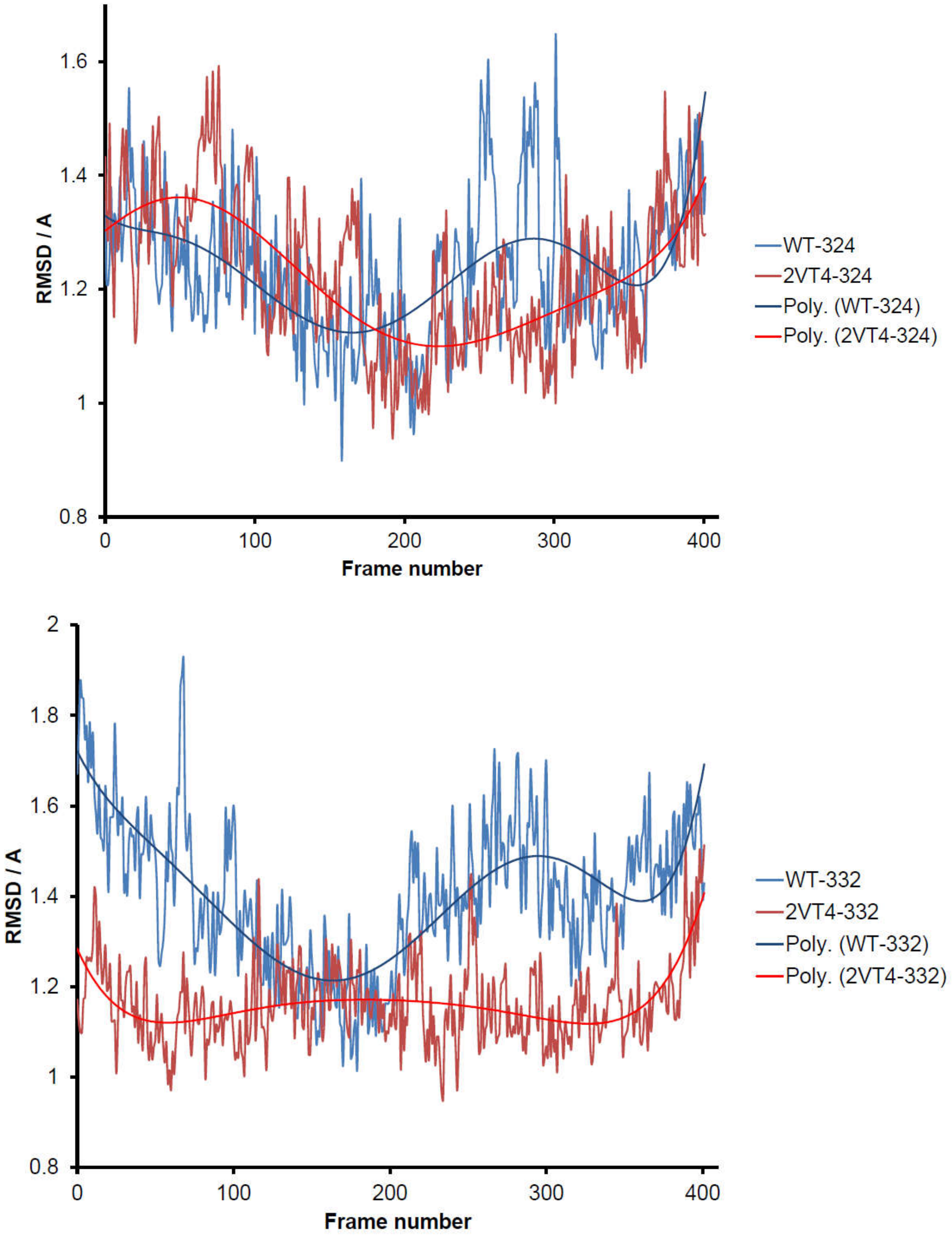


Figure 5.2d(top) and 5.2e (bottom) the RMSD (in Å) against the average structure of βAR-m23 (red) and βAR-wt (wildtype) over the full 200 ns of the production run. The RMSD is for the backbone of the protein excluding any hydrogens;figure 5.2d is for 324k and figure 5.2c is for 332K . Each frame is 200 ps. A 6th order polynomial was fitted to the RMSD data, denoted Poly.(WT-324 or 332 and 2VT4-324 or 332).

Figure 5.3 shows that the RMSF is largely negative, suggesting that the βAR-m23 fluctuates more than βAR-wt. The point where the βAR-m23 appears to be more stable is at the ends of each helix. This is seen in spikes of positive values along the figure. For the other regions, the wildtype is more stable. So the central part of the bundle appears to be more stable in the wildtype, while the ends of the helix are more stable in the thermostabilized receptor. Since during activation, it is the ends of the helices that move most, the lower RMSF in this region may be an indicator of increased thermostability.

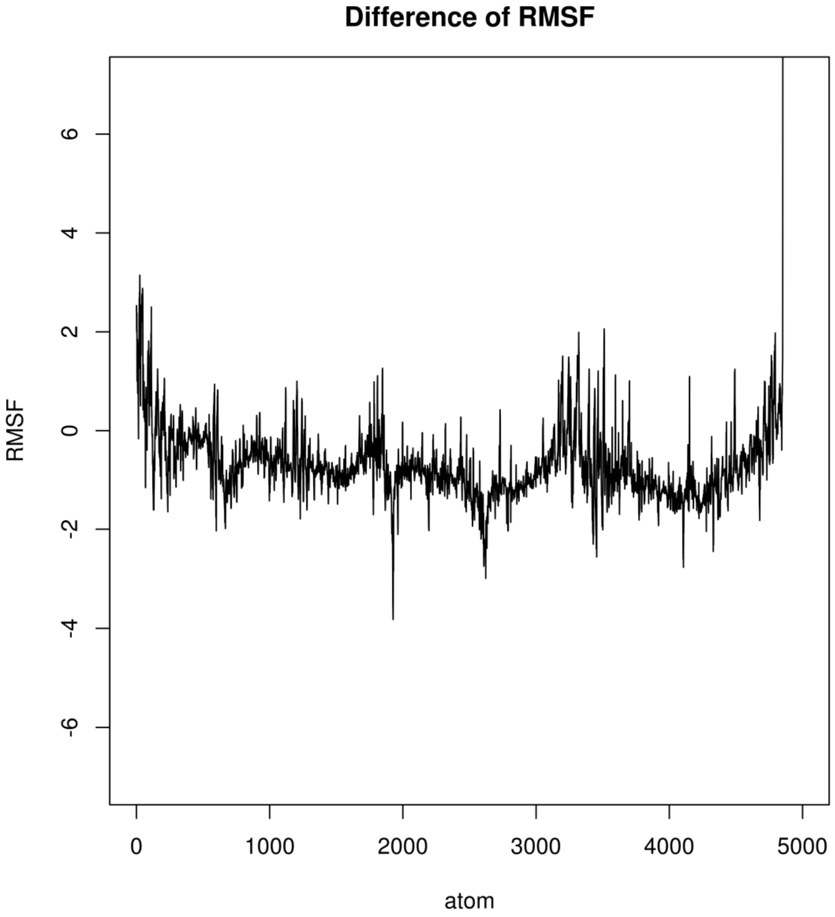


Figure 5.3 The difference of RMSF across the all atoms between the βAR-m23 mutant and the βAR-wt wildtype. A positive value indicates that the RMSF is higher

in the wildtype. (The tail end of the structure is different between the two receptors which is why there is a spike of RMSF at the very end.)

The wildtype was used as the base and the RMSF of the mutant was subtracted from it at each point. This means that a positive value would show that the conformation of the mutant was fluctuating less as its RMSF was lower than that for the wildtype, meaning the mutant was more stable, whereas a negative value would show that the wildtype has a lower RMSF at those values showing that it was fluctuating less. .

The results shown in Figure 5.3 were obtained in the absence of ligand, and so the corresponding results in the presence (and absence) of ligand are shown in Figure 5.4, over the Cα atoms. On the right of Figure 5.4 we can see that the RMSF concurs with the previous data in 5.3 with βAR-m23 (in blue) as it has a higher RMSF, continuing the trend of being less stable in simulation. With the inclusion of the ligand, both proteins exhibit more stability. Despite their stability being very close, Figure 5.4 shows that the mutant is still slightly less stable as it exhibits a higher RMSF.

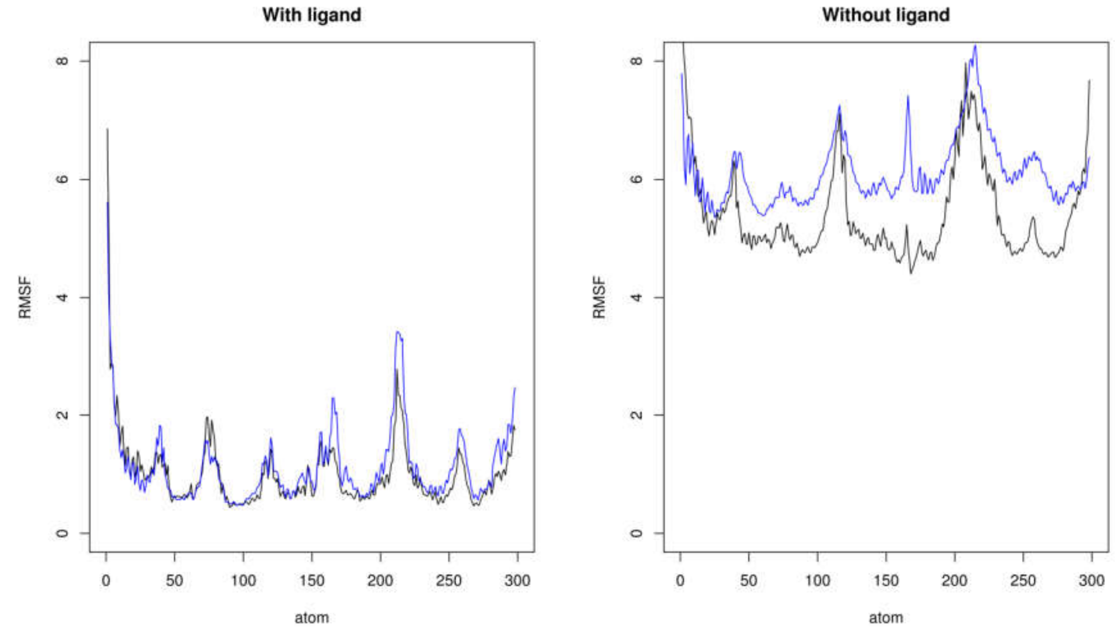


Figure 5.4. The effects of the presence (left) and absence (right) of ligand on the RMSF of the alpha carbons of βAR-m23 (blue) and βAR-wt (black).

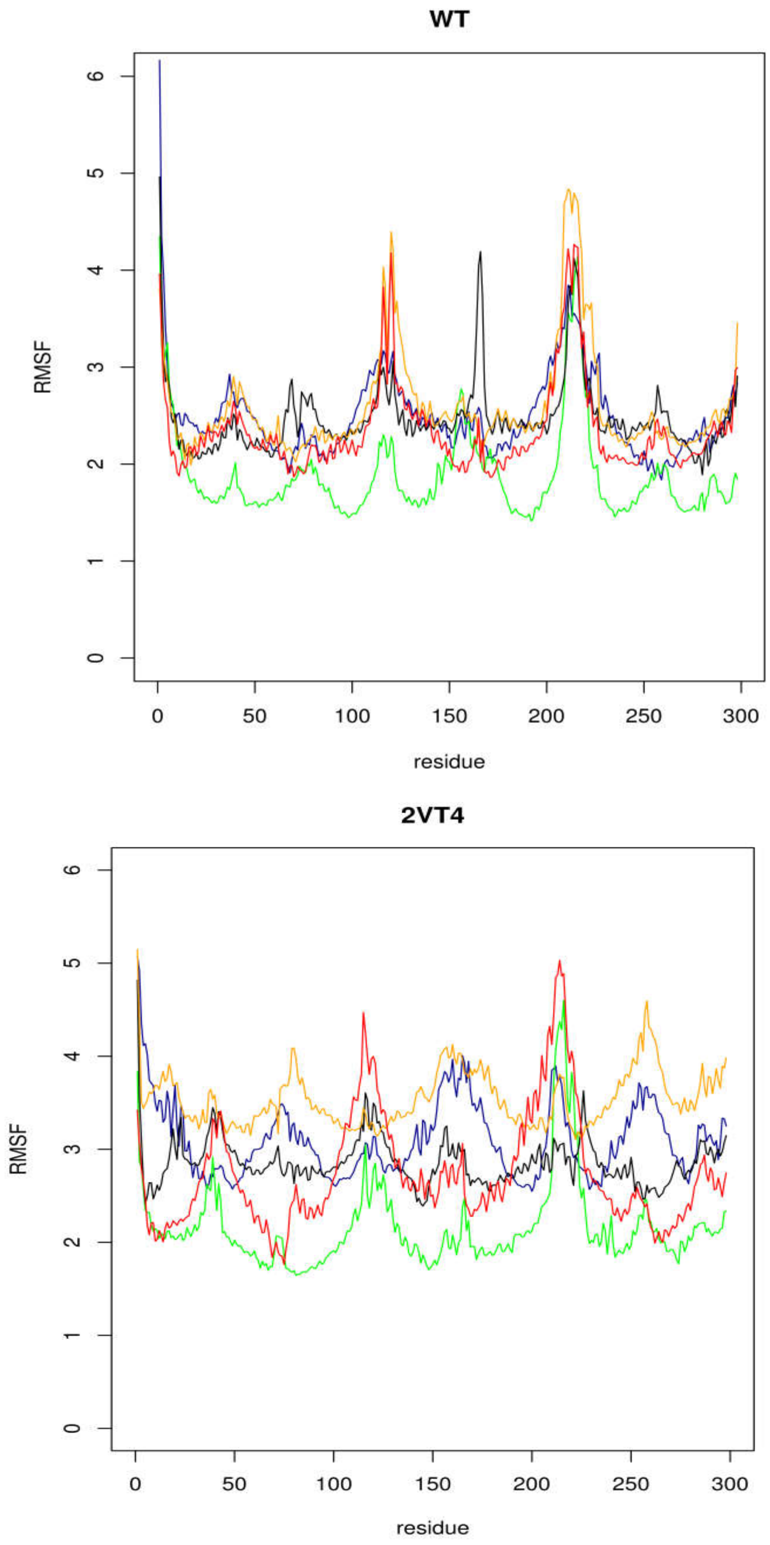


Figure 5.5. The effect of temperature on RMSF for the wild-type and mutant receptor. The temperatures are, 300 K (blue), 308 K (green), 316 K (black), 324 K (orange), 332 K (red).

The effect of temperature on RMSF is shown in Figure 5.5. This shows some interesting results. There is a lot more of an effect of temperature on the mutant than there is on the wildtype. The RMSF drops to the most stable at 308 K for both proteins, which is relatively close to body temperature; this is likely to be close to the optimal temperature for this receptor. As the temperature increases for the wildtype it stays relatively stable, fluctuating around 2.5 Å RMSF with spikes for the intracellular and extracellular loops that join the transmembrane helices. The βAR-m23 mutant (2VT4) is least stable at 324 K but drops back to a 2.5 Å RMSF, with exception to the non-transmembrane regions at 332 K. The result for the mutant appears to be erratic but again, is consistently less stable than the Wildtype.

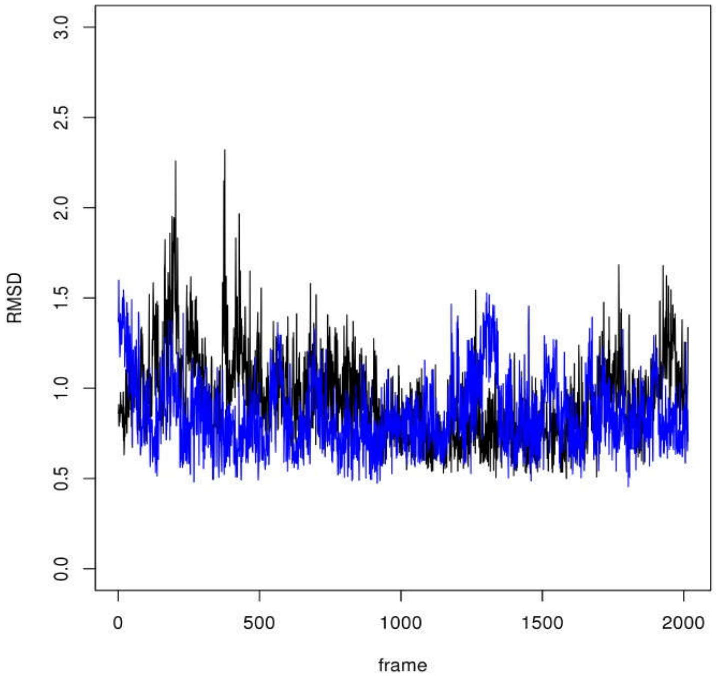


Figure 5.6: Ligand RMSD (in Å) for the receptor:ligand complex over the full 200 ns of the production run. βAR-m23 is shown in blue and βAR-wt is shown in black. The RMSD is for the bound ligand; the reference point is the average structure. Each frame is equivalent to 200 ps.

The RMSD of the ligands for βAR-m23 and βAR-wt over the full 200 ns of the production run is shown in figure 5.6. The RMSD for the ligand in the βAR-m23 mutant is lower than that of the wildtype, fluctuating at about 0.8 Å, compared to the wildtype's 1-1.2 Å. This shows that the ligand in the βAR-wt is not bound as tightly as in the mutant.

This is further corroborated by figures 5.7a-e which shows how many hydrogen bond interactions there are between an important interaction between Arg^{135} and Glu^{285} in helices 3 and 6 over the course of the simulation. This bond helps the overall stability between helices 3 and 6. Each trough shows that there are no hydrogen bonds between the two amino acids. In figure 5.7a the βAR-m23 is forming more hydrogen bonds at 300 K and has fewer disconnects than the wildtype. This pattern stays the same for figures 5.7b-5.7d at 308, 316 and 324 K. In figure 5.7e at 332 K something interesting happens. The wildtype is unable to keep the Arg-Glu salt bridge for 10 ns whilst the βAR-m23 structure manages to maintain the interaction. This further supports the idea that 2VT4 is more thermodynamically stable. As seen in figure 5.8, the RMSF of the ligand for βAR-m23 mutant (2VT4) is lower. This shows that the structure of the ligand is more stable in the binding site but is a little less tightly bound the ligand in the wildtype which continues the trend shown in figure 5.6. A principal component analysis (PCA) was also performed for the each of the biomacromolecules shown in figure 5.9a-b. The green arrows attached to the helices show which parts are fluctuating greater than 1 Å. It is difficult to see from the figure but the central helices for 2VT4 are fluctuating slightly more than the wildtype.

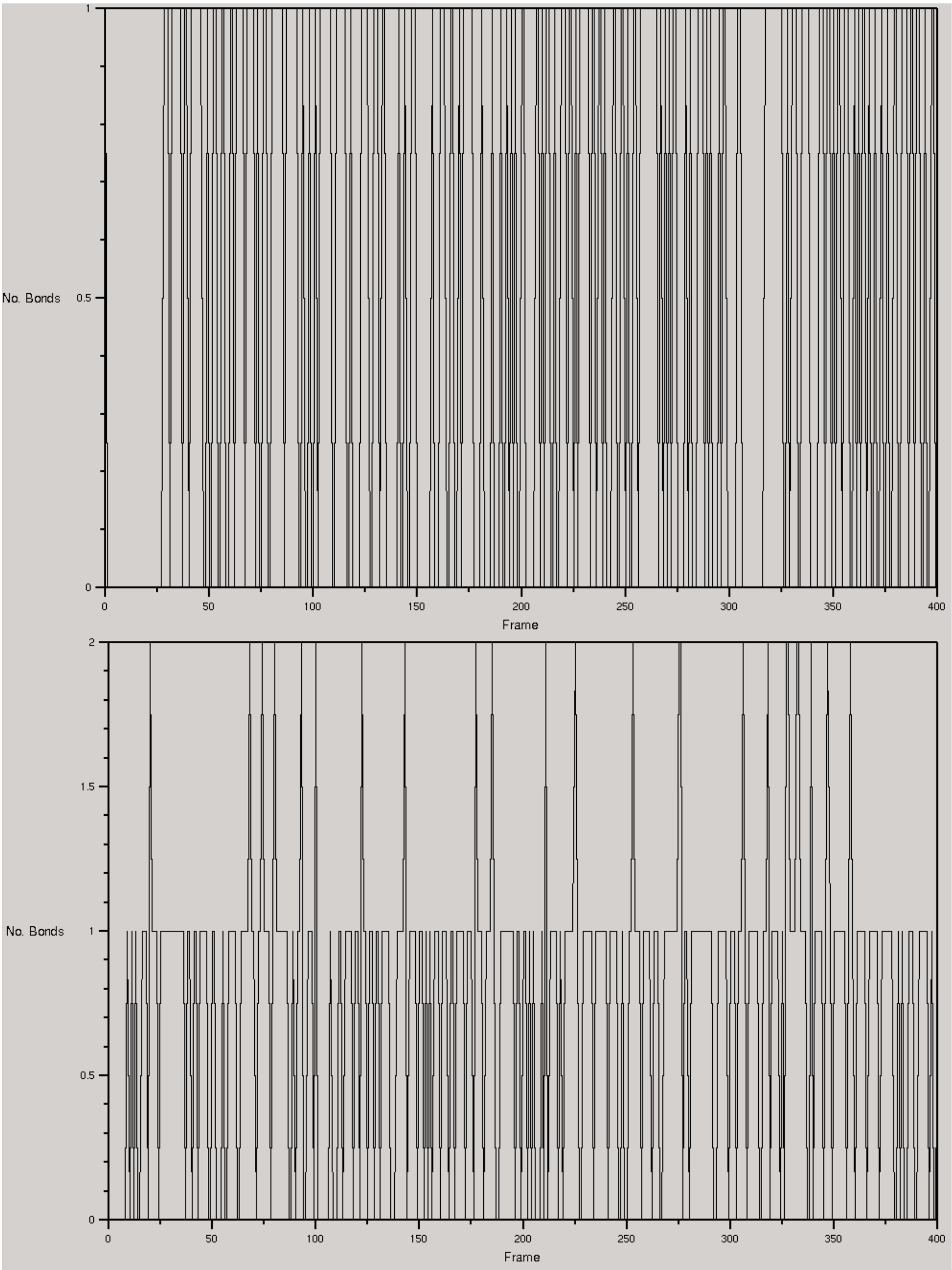


Figure 5.7a. Hydrogen bonds between Arg^{135} and Glu^{285}in helices 3 and 6 at 300 K. 5.7.a1 (Top) is βAR-wt, 5.7.a2 (bottom) is βAR-m23. Each frame is 200 ps.

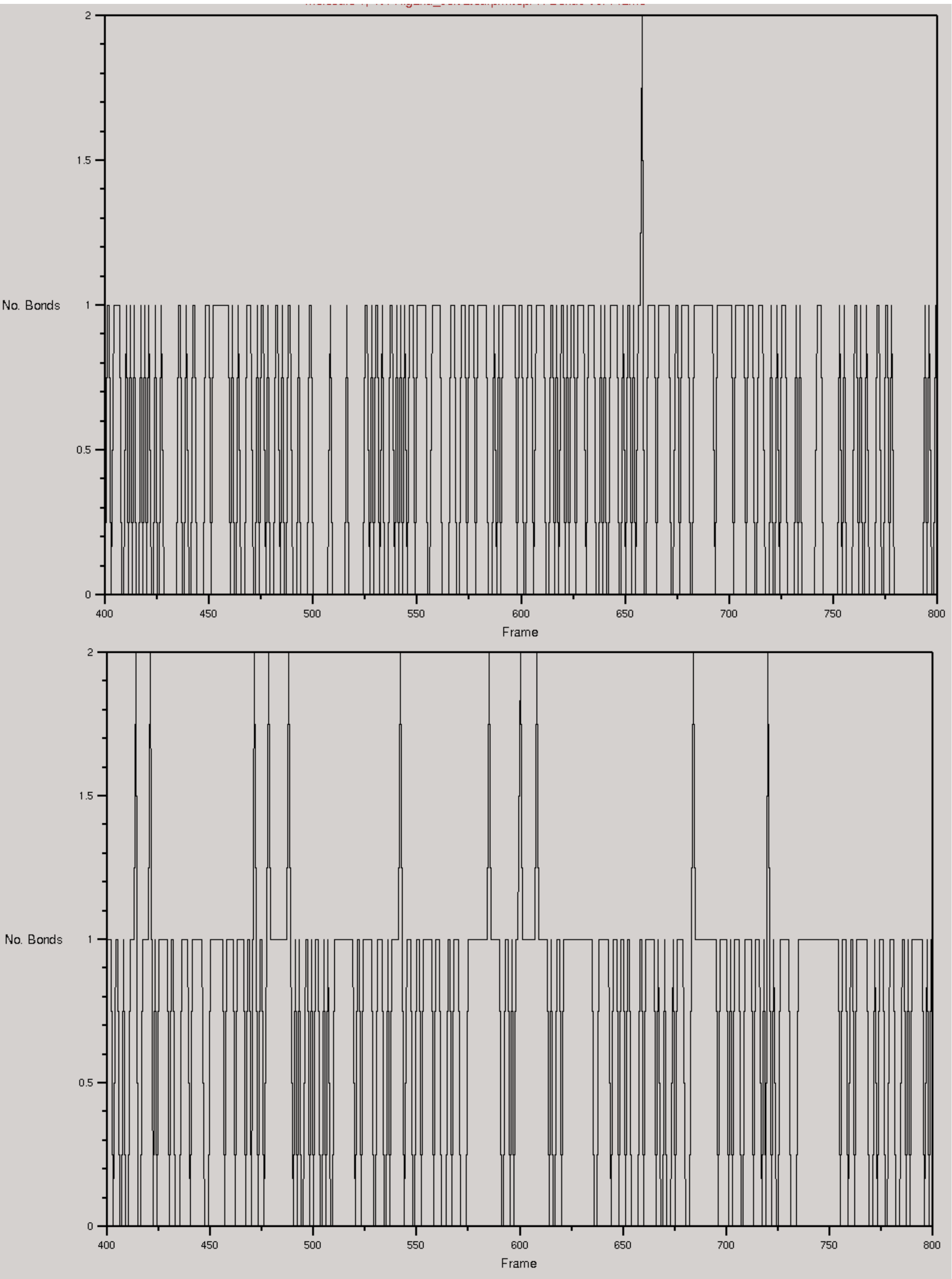


Figure 5.7b. Hydrogen bonds between Arg^{135} and Glu^{285} in helices 3 and 6 at 308 K. 5.7.b1 (Top) is βAR-wt, 5.7.b2 (bottom) is the βAR-m23. Each frame is 200 ps.

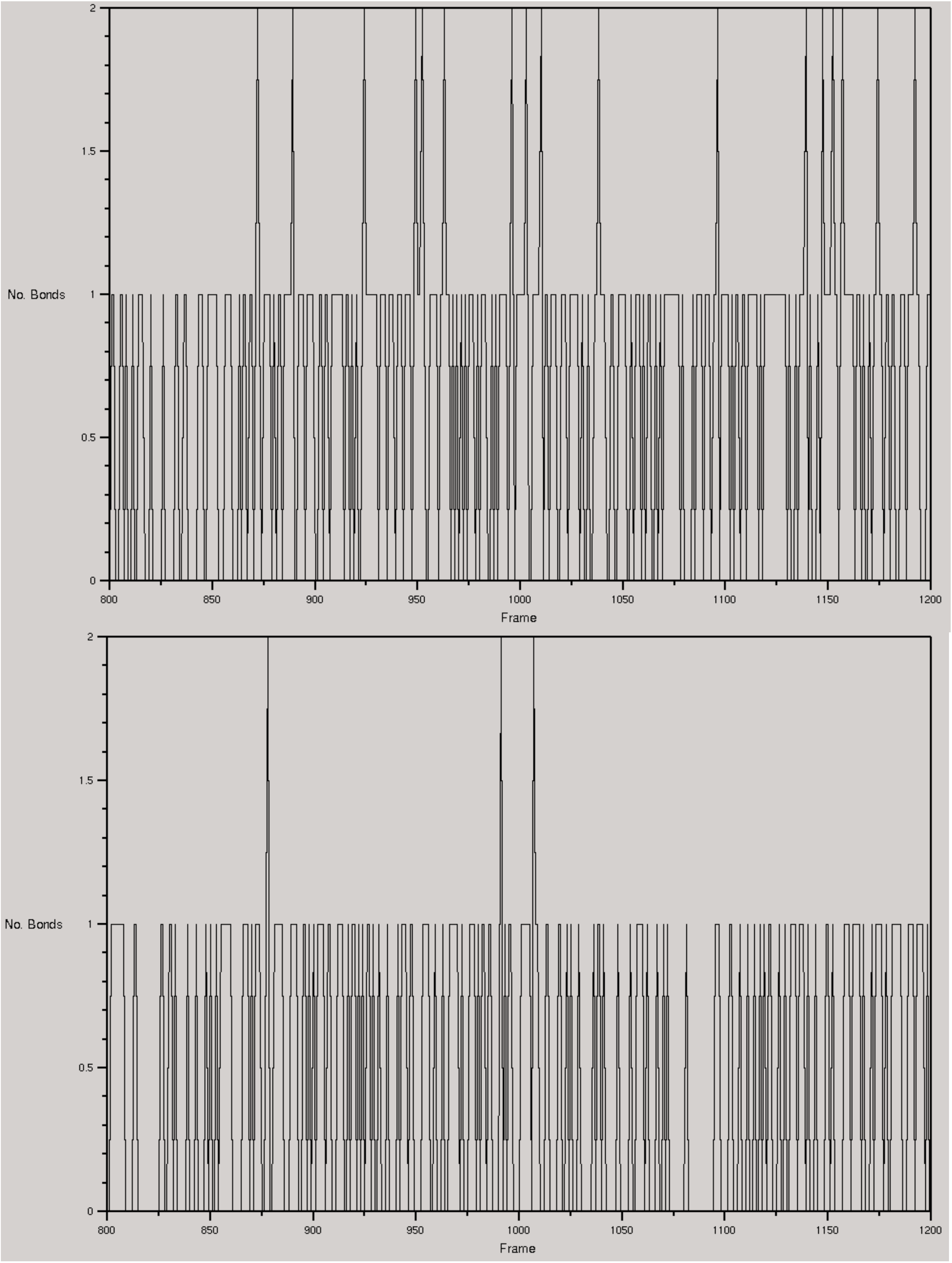


Figure 5.7c. Hydrogen bonds between Arg^{135} and Glu^{285} in helices 3 and 6 at 316 K. 5.7.c1 (Top) is βAR-wt, 5.7.c2 (bottom) is βAR-m23. Each frame is 200 ps.

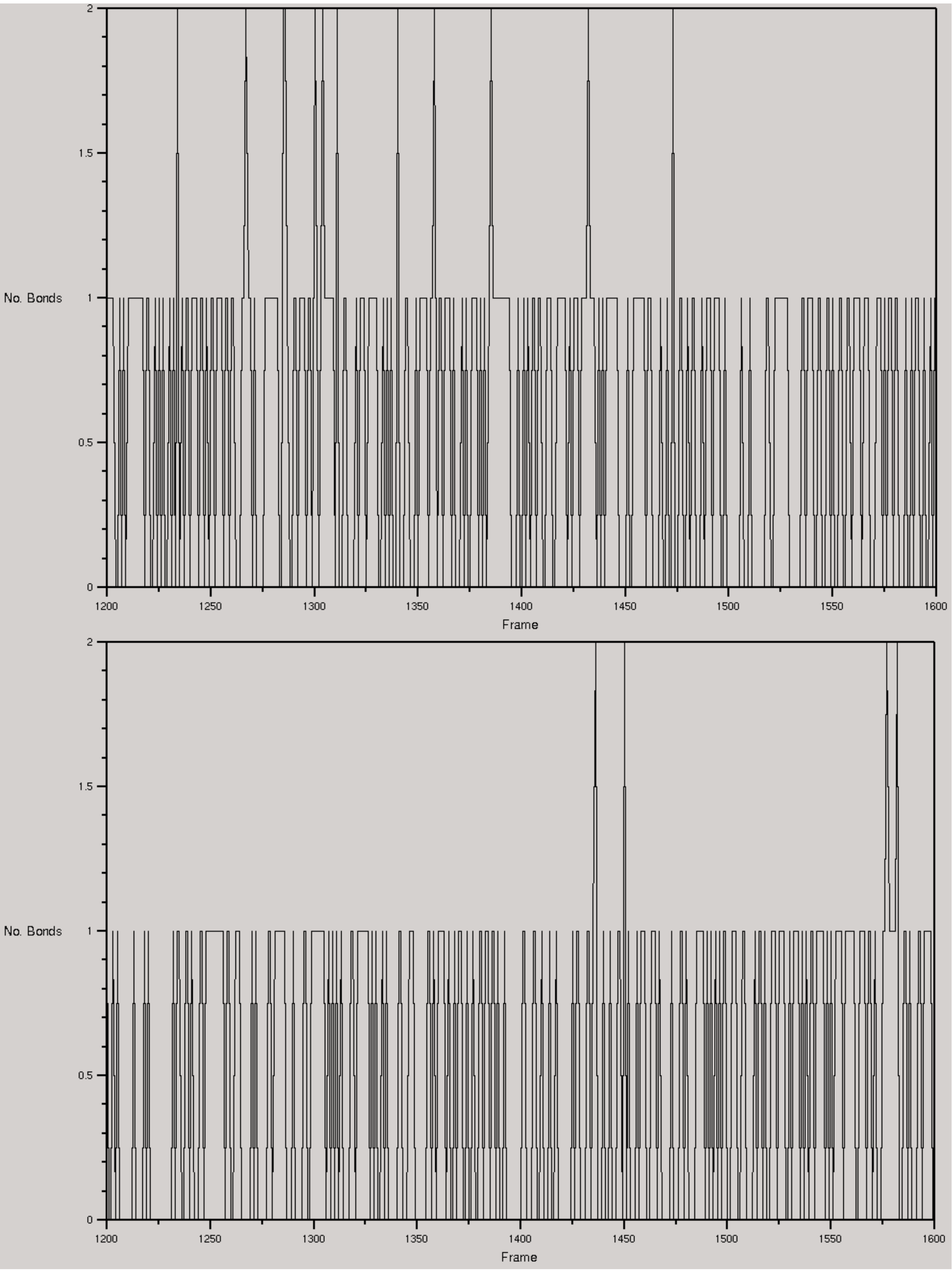


Figure 5.7d. Hydrogen bonds between Arg[135] and Glu[285] in helices 3 and 6 at 324 K. 5.7.d1 (Top) is βAR-wt, 5.7.d2 (bottom) is βAR-m23. Each frame is 200 ps.

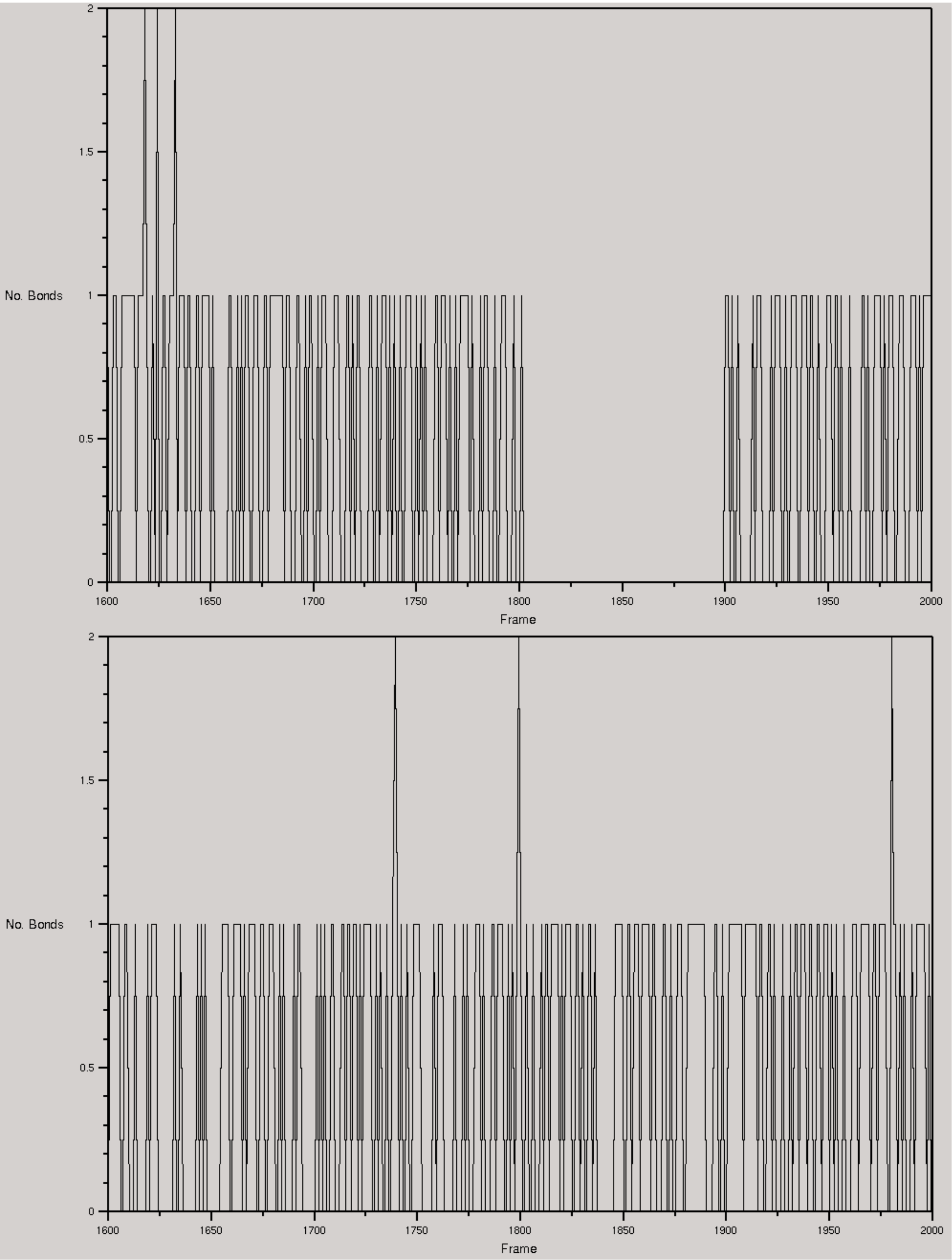


Figure 5.7e. Hydrogen bonds between Arg[135] and Glu[285] in helices 3 and 6 at 332 K. 5.7.e1 (Top) is βAR-wt, 5.7.e2 (bottom) is βAR-m23. Each frame is 200 ps.

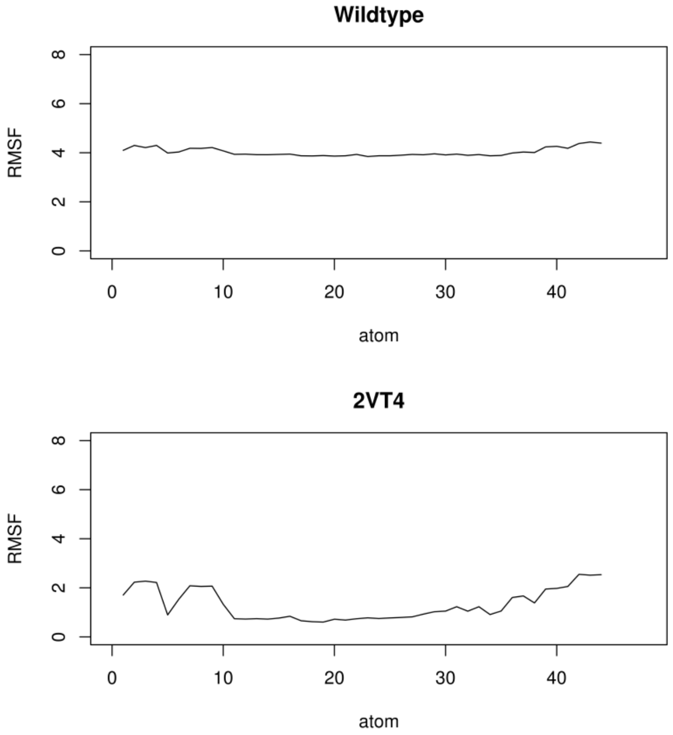


Figure 5.8. Ligand RMSF over the 200 ns simulation of the ligands in complex with (A) βAR-m23 (2VT4) and (B) βAR-wt (Wildtype).

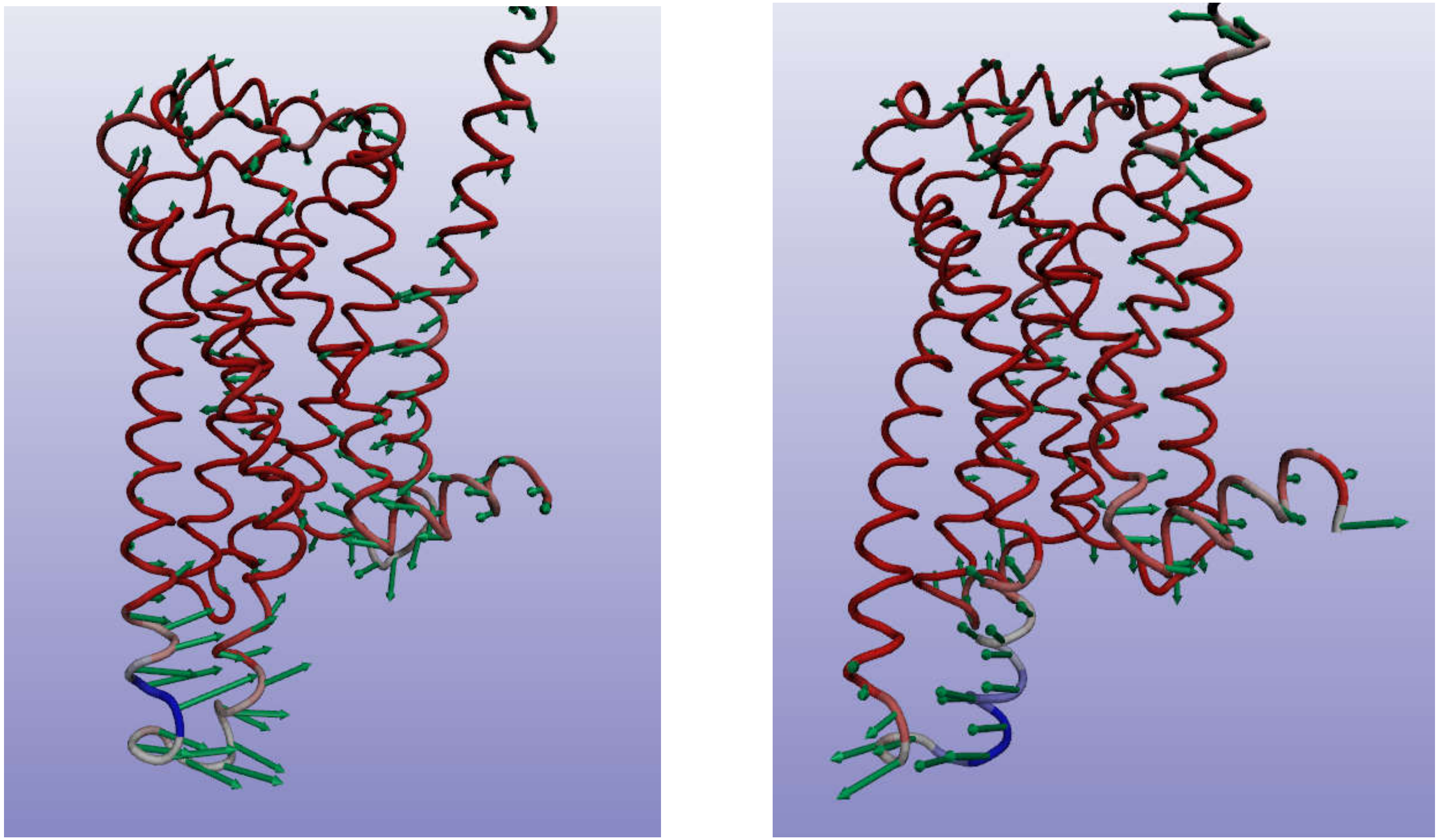


Figure 5.9a-b: Graphical representations of PCA for (a) βAR-m23 (2VT4) on the left and (b) βAR-wt on the right. The green arrows represent the movement for that portion of the protein.

5.4 Discussion

The results complement data from the study shown in Serrano-Vega et al. (2008). In the study, Serrano-Vega et al. measured DHA binding at different temperatures and with different detergents. Their results showed that the mutant had more DHA bound throughout most temperatures between 6 ^{o}C and 40^{o}C (279-313 K). Whilst our data does not show how much is bound, the RMSD and RMSF of the ligand in figures 5.6 and 5.8 show that the βAR-m23 structure fluctuates less and maintains its conformation more than the wildtype. This means that our data is showing that the ligand binds more tightly in βAR-m23 in than the wildtype which is consistent with the literature.

The paper also inferred that due to this stronger binding, βAR-m23 is more thermostable for use in crystallography, since X-ray crystallography ideally requires binding of ligands to stabilize the receptor structure and hence to create good crystals that will diffract. Serrano-Vega et al. (2008) go onto say that this stability is used to prevent conformational change, making it easier to crystallize. Our data corroborates this, showing that the average RMSD of the βAR-m23 is consistently less than the wildtype. RMSF data and PCA in figures 5.5 and 5.9 show however that the βAR-m23 mutant fluctuates more as a whole, but differently. This is due to the greater movements of the extra-membranous regions of the protein that fluctuate more in the mutant than the wildtype, shown in the PCA as longer arrows in those regions. This therefore will increase the RMSF as a whole for the βAR-m23. The mutations might have had an effect on the stability on these extra-membranous regions of the structure as two of the mutations are in this region, as shown in figure

5.2. This however, has a lesser effect on the binding region as the data shows the intra-membranous region is more stable in the βAR-m23 mutant.

The ligand results also appear to be consistent what with Serrano-Vega et al. (2008) have found. As stated previously, there is more bound ligand in the mutant than the wildtype. However in the results in figures 5.7a-e and 5.8 the helices (3 & 6) around the binding site open up more readily in the wildtype than in the βAR-m23 mutant shown in the results on hydrogen bonds. The ligand is bound more tightly in βAR-m23 according to the RMSD studies as well.

5.5 References

BALARAMAN, G. S., BHATTACHARYA, S. & VAIDEHI, N. 2010. Structural insights into conformational stability of wild-type and mutant β 1-adrenergic receptor. *Biophysical journal,* 99**,** 568-577.

BALLESTEROS, J. A. & WEINSTEIN, H. 1995. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors.

BANERES, J.-L., MARTIN, A., HULLOT, P., GIRARD, J.-P., ROSSI, J.-C. & PARELLO, J. 2003. Structure-based analysis of GPCR function: conformational adaptation of both agonist and receptor upon leukotriene B 4 binding to recombinant BLT1. *Journal of molecular biology,* 329**,** 801-814.

CHELIKANI, P., REEVES, P. J., RAJBHANDARY, U. L. & KHORANA, H. G. 2006. The synthesis and high ‐ level expression of a β 2 ‐ adrenergic receptor gene in a tetracycline ‐ inducible stable mammalian cell line. *Protein Science,* 15**,** 1433-1440.

CHEN, K.-Y. M., ZHOU, F., FRYSZCZYN, B. G. & BARTH, P. 2012. Naturally evolved G protein-coupled receptors adopt metastable conformations. *Proceedings of the National Academy of Sciences,* 109**,** 13284-13289.

CONGREVE, M., LANGMEAD, C. J., MASON, J. S. & MARSHALL, F. H. 2011. Progress in structure based drug design for G protein-coupled receptors. *Journal of medicinal chemistry,* 54**,** 4283-4311.

ESWAR, N., WEBB, B., MARTI-RENOM, M., MADHUSUDHAN, M., ERAMIAN, D., SHEN, M., PIEPER, U. & SALI, A. 2007. Comparative protein structure modeling using MODELLER Curr. Protoc. *Protein Sci*.

LEBON, G., BENNETT, K., JAZAYERI, A. & TATE, C. G. 2011a. Thermostabilisation of an agonist-bound conformation of the human adenosine A 2A receptor. *Journal of molecular biology,* 409**,** 298-310.

LEBON, G., WARNE, T., EDWARDS, P. C., BENNETT, K., LANGMEAD, C. J., LESLIE, A. G. & TATE, C. G. 2011b. Agonist-bound adenosine A2A receptor structures reveal common features of GPCR activation. *Nature,* 474**,** 521-525.

REEVES, P. J., CALLEWAERT, N., CONTRERAS, R. & KHORANA, H. G. 2002. Structure and function in rhodopsin: high-level expression of rhodopsin with restricted and homogeneous N-glycosylation by a tetracycline-inducible N-acetylglucosaminyltransferase I-negative HEK293S stable mammalian cell line. *Proceedings of the National Academy of Sciences,* 99**,** 13419-13424.

SCHWARTZ, T. W., FRIMURER, T. M., HOLST, B., ROSENKILDE, M. M. & ELLING, C. E. 2006. Molecular mechanism of 7TM receptor activation-a global toggle switch model. *Annu. Rev. Pharmacol. Toxicol.,* 46**,** 481-519.

SERRANO-VEGA, M. J., MAGNANI, F., SHIBATA, Y. & TATE, C. G. 2008. Conformational thermostabilization of the β1-adrenergic receptor in a detergent-resistant form. *Proc Natl Acad Sci U S A,* 105**,** 877-82.

SHIBATA, Y., WHITE, J. F., SERRANO-VEGA, M. J., MAGNANI, F., ALOIA, A. L., GRISSHAMMER, R. & TATE, C. G. 2009. Thermostabilization of the neurotensin receptor NTS1. *Journal of molecular biology,* 390**,** 262-277.

SIMPSON, L. M. 2011. *PhD Topic.* University of Essex.

WARNE, T., SERRANO-VEGA, M. J., BAKER, J. G., MOUKHAMETZIANOV, R., EDWARDS, P. C., HENDERSON, R., LESLIE, A. G., TATE, C. G. & SCHERTLER, G. F. 2008. Structure of a &bgr; 1-adrenergic G-protein-coupled receptor. *Nature,* 454**,** 486-491.

WHITE, J. F., NOINAJ, N., SHIBATA, Y., LOVE, J., KLOSS, B., XU, F., GVOZDENOVIC-JEREMIC, J., SHAH, P., SHILOACH, J. & TATE, C. G. 2012. Structure of the agonist-bound neurotensin receptor. *Nature,* 490**,** 508-513.

WU, B., CHIEN, E. Y., MOL, C. D., FENALTI, G., LIU, W., KATRITCH, V., ABAGYAN, R., BROOUN, A., WELLS, P. & BI, F. C. 2010. Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. *Science,* 330**,** 1066-1071.

YARDEN, Y., RODRIGUEZ, H., WONG, S. K., BRANDT, D. R., MAY, D. C., BURNIER, J., HARKINS, R. N., CHEN, E. Y., RAMACHANDRAN, J. & ULLRICH, A. 1986. The avian beta-adrenergic receptor: primary structure and membrane topology. *Proc Natl Acad Sci U S A,* 83**,** 6795-9.

Chapter 6

Concluding remarks

6.1. Soluble proteins

As shown in the results presented here, inclusion of polarization has an effect on improving the docking results of ligands to the protein, as signified by the RMSD values of the molecules in the validation sets; this effect can range from minor to major. Thus when polarization was included, 13 out of 38 tested proteins had better RMSD values. The docking process was optimised to generate polarized charges that could replace the standard charges found on the ligands in the validation set; this process involved polarizing the ligand according to the top pose generated by docking the ligand with standard unpolarised charges. Inevitably, because of the nature of the docking process, 6 of the ligands docked into the wrong positions, with errors up to ~ 10 Å in the RMSD. Nevertheless, during MD simulations of incorrectly docked ligands, each of these ligands showed signs of moving closer to their correct binding sites, sometimes returning to a binding position with an RMSD of less than 1 Å (9/21 returned to within 3 Å, while 2/21 returned to within 1 Å). Polarization can therefore impact some results while leaving some unchanged; In only 1 case did polarization make the results worse.

Thus, we have shown that inclusion of polarization has a clear positive effect on these calculations. However, it is important to discuss the key question as to whether polarization can be routinely included in docking and simulation to improve docking and simulation studies relevant to drug design.

The question comes down to opportunity, ease of implementation and cost. So it seems that inclusion of polarization is a positive step forwards. The time added to the preparation of each ligand was minimal even with the use of in-house scripts that weren’t part of the docking programs. If the process was fully incorporated into the docking program very little additional time would be required to improve the results. This is probably the most significant next step for carrying this polarized docking work forward. Ideally, this would be carried forward in GLIDE but could also be carried forward into other docking programs such as Autodock (Morris et al., 2009).

Blind docking and cross-docking in drug design: Including molecular dynamics as an extra stage does add a lot more computational time to the process and at the current time this might not be the most useful additional step to a screening protocol. There is therefore scope for discussion on the best way to use MD in this situation. Firstly, it could be used to assess whether a ligand is docked correctly – if it drifts away from its binding site then perhaps it was not docked correctly. Alternatively, for a set of fragments in a FBDD programme where it is not necessarily known how the fragment binds, it could be included in a two stage process similar to that used in current methods of HTS. First, attempt to dock a ligand library to a targeted protein through use of programs such as GLIDE. Find a small subset of ligands that have high binding affinity. Secondly, for this subset of ligands, new charges based upon polarization could be generated. Then having generated these new polarized ligand/protein charges, MD simulations could be performed on the complexes. The ligands should be able to converge near to the correct binding site. This process would allow comparison of the location of all the ligands as they should be converge towards the binding site. This could help pinpoint alternative or allosteric binding sites by observing the region in where the ligands are attempting to bind. Finally, the MD simulations give insight into the dynamic nature of fragment binding, which cannot necessarily be obtained from docking alone or from X-ray crystallography. Some ligands bind as expected, forming well-defined interactions, while other ligands are more dynamic and oscillate between alternative binding sites in a process where interactions are made and broken. This in particular may be useful for indicating how fragment interactions could be improved.

For our MD simulations the charges stayed static through the entire simulation. Future research should be into how we could improve the charges to update for each frame of a simulation. If these charges were based on the QM/MM induced charge method we used, the QM methods would be updating at each step of the simulation or updating every time the geometry of the system changed by a set amount. This would become very resource intensive as it stands for current programs and systems. We could wait until the systems improve to a point we could do this but it could take decades if we are to use CPUs as they have hit a clock speed cliff in this past decade. The solution to that problem is adding more cores. However, that is a crude solution, with the current advancements in the field of FBDD with regards to GPU use it is possible that the complexity allowance in systems will continue to rise.

With QM/MM hybridization methods the problem will always be the resource intensive QM portion. New methods are emerging where the protein is divided into different regions that are modelled using different water models. Within 5 Å of the point of interest in the protein, the water is explicit, between 5-10 Å the water is coarse-grained and beyond 10 Å the water is modelled implicitly. This is an efficient compromise for calculating the energy and the positions of molecules in the system because the calculations take into account fewer terms the further one goes from the point of interest. Using the ideas from these studies we could propose a system that changes the model of the protein based on distance from the point of interest, e.g. the ligand. Currently in the QM/MM method, the protein is treated by MM and explicitly using all atom models. However over the years in the literature there have been coarse-grained models for proteins as well (Tozzini, 2005). For coarse grain if the protein in a system was treated as a Gō-like model it depicts each amino acid as a separate bead. This was originally used for models to research folding. However it could equally be used in MD. By following this model for amino acids say 5 - 10 Å away from the ligand, it could become far simpler and easier to calculate the dynamics of the system. However the region closest to the ligand will still be explicit. Therefore at this point it could be possible to calculate the QM more easily. This will cut down on the computational time needed to calculate the QM at each step. There however is one additional problem to this. That is the water in the system. When these ligands were polarized they were not in the presence of water and in the presence of a static protein so the computational power that was needed is small. Since water is a polar molecule, when it is added into the system it will affect the polarization of each step. Due to how ubiquitous water is in a system there will be many interactions with the ligand in that system. As was done it chapter 2 we could change the water to an implicit model to improve the computational time. However during a simulation this means that we are losing the polarization that would happen due to the explicit waters' presence. There are many challenges ahead but we believe that there will eventually be a programming solution for determining the charges of a dynamic system.

6.2. Membrane simulations: In chapter 5, the MD simulations for the transmembrane protein β_1-AR were largely successful and corroborated with experimental data in the literature. The methods we use to ascertain this are similar to the RMSD-based methods used to assess binding in Chapters 2-4. By showing MD simulation data (Figures 5.2 – 5.6) that matches the experimental data on the protein stability in the paper by Serrano-Vega et al. (2008)we have provided further evidence to help validate our findings.

However, to further improve on these simulations we could continue to increase the sampling time to 200 ns for each temperature. This should allow for any possible errors in sampling to be greatly reduced.

These simulations also did not include our polarization terms as the aim was mostly to ascertain the stability of the mutant protein and the wildtype. However we could also include the polarization of the DHA ligand. As seen in figure 5.4, the RMSF of the protein is far lower when it the ligand is bound. This is due to the breathing motion of the helices 3 and 6; with no ligand bound helixes 3 and 6 fluctuate more freely. Thus, the interaction between ligand and protein is used to improve its stability. By including polarization we should be able to further improve the computational method for assessing protein stability. We would follow the same procedure described in Chapter 2 to generate atomic charges based on our induced charge method. Then redock the ligand to the binding site and importing the new atomic charges into the MD before performing a production run.

Serrano-Vega et al. (2008)described assays at high temperatures. The highest temperature we used is 332 K.. However, we could also perform extreme temperature simulations in ACEMD. We could then also ascertain if the thermostability present at the lower temperatures for βAR-m23 mutant stays consistent into the higher temperatures such as 360 K which is where the experimental data tests the half-life of the stable mutant.

References

MORRIS, G. M., HUEY, R., LINDSTROM, W., SANNER, M. F., BELEW, R. K., GOODSELL, D. S. & OLSON, A. J. 2009. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry,* 30**,** 2785-2791.

SERRANO-VEGA, M. J., MAGNANI, F., SHIBATA, Y. & TATE, C. G. 2008. Conformational thermostabilization of the β1-adrenergic receptor in a detergent-resistant form. *Proc Natl Acad Sci U S A,* 105**,** 877-82.

TOZZINI, V. 2005. Coarse-grained models for proteins. *Current opinion in structural biology,* 15**,** 144-150.

Appendix

2.1 Script to take into account atoms of equivalent value in RMSD calculations

```
#!/usr/bin/perl
$symetry="symetry";
open(AGR,">agree");
$arg=@ARGV;
if($arg<2){
      die "Need residuum names for QM and list file containing names
of mae files \n";
};
if(-s "$symetry"){
      $sym=1;
      print STDERR "Symetry file will be aplied\n";
}else{
      $sym=0;
      print STDERR "No symetry file will be aplied\n";
};

#Read template name
for($i=0;$i<$arg-1;$i++){
      if(($lig[$i],$lnum[$i])=$ARGV[$i]=~/^(.+)%(.*)/){
            unless($lnum[$i]=~/^[-]?\d+$/){
                  print STDERR "Warning: $lnum[$i] is not a number.
Using all $lig[$i]\n";
                  $lnum[$i]="";
            };
      }else{
            $lig[$i]=$ARGV[$i];
            $lnum[$i]="";
      };
};
#Read the list of mae files
open(LST,"$ARGV[$arg-1]") or die "File  $ARGV[$arg-1] does noi
eixts\n";
$i=0;
while($line=<LST>){
      chomp($line);
      if(-e $line){
            $filename[$i]=$line;
            $i++;
      }else{
            print STDERR "Filename $line does not exsists, it will be
ignored\n";
      };
};
$nposes=$i;
for($prvni=0;$prvni<$nposes-1;$prvni++){
open(MAE1,"$filename[$prvni]") or die "File $filename[$prvni] does
not exist\n";
#Read input file
$i=$hi=$mi=0;
while($line=<MAE1>){
```

```
if(($indx1[$i],$x,$y,$z,$rn,$ch,$r,$a,$el)=$line=~/^\s*(\d+)\s+\d+\s
+([-+]?\d+\.\d+)\s+([-+]?\d+\.\d+)\s+([-+]?\d+\.\d+)\s+([-
+]?\d+)\s+\"[^"]*\"\s+\w+\s+(?:(?:\"[^"]*\")|\w+)\s+\d+\s+([-
+]?\d+\.\d+)\s+[-
+]?\d+\.\d+\s+\"\s*(\w+)\s*\"\s+\"\s*([\w']+)\s*\"\s+\"[^"]*\"\s+(\d
+)\s+/){
            $flag=0;
            for($ii=0;$ii<$arg-1;$ii++){
                  if($r eq $lig[$ii]) {
                        if($lnum[$ii] eq "" or $lnum[$ii]==$rn){
                              $flag=1;
                                    $last;
                        };
                  };
            };
                  if($flag==1) {
                          $nm1[$i]=$a;
                          $xx1[$i]=$x;
                          $yy1[$i]=$y;
                          $zz1[$i]=$z;
                  print "$a $x $y $z\n";
                  unless($el==1){
                        $nm1h[$hi]=$a;
                        $xx1h[$hi]=$x;
                        $yy1h[$hi]=$y;
                        $zz1h[$hi]=$z;
                        $hi++;
                  };
            $i++;
                 }else{
                  unless($el==1){
                        $xxm1[$mi]=$x;;
                        $yym1[$mi]=$y;;
                        $zzm1[$mi]=$z;;
                        $mi++;
                  };
            };
      };
};
$cpa1=$i;
$cpa1h=$hi;
if($cpa1h==0){
      die "There is zero QM havy atoms\n";
};
$cpm1=$mi;
for($druhy=$prvni+1;$druhy<$nposes;$druhy++){
open(MAE2,"$filename[$druhy]") or die "File $filename[$druhy] does
not exist\n";
$i=$hi=$mi=0;
while($line=<MAE2>){

if(($indx2[$i],$x,$y,$z,$rn,$ch,$r,$a,$el)=$line=~/^\s*(\d+)\s+\d+\s
+([-+]?\d+\.\d+)\s+([-+]?\d+\.\d+)\s+([-+]?\d+\.\d+)\s+([-
+]?\d+)\s+\"[^"]*\"\s+\w+\s+(?:(?:\"[^"]*\")|\w+)\s+\d+\s+([-
+]?\d+\.\d+)\s+[-
+]?\d+\.\d+\s+\"\s*(\w+)\s*\"\s+\"\s*([\w']+)\s*\"\s+\"[^"]*\"\s+(\d
+)\s+/){
        $flag=0;
        for($ii=0;$ii<$arg-1;$ii++){
            if($r eq $lig[$ii]) {
                if($lnum[$ii] eq "" or $lnum[$ii]==$rn){
                    $flag=1;
                        $last;
                };
            };
        };
            if($flag==1) {
                  $nm2[$i]=$a;
                  $xx2[$i]=$x;
                  $yy2[$i]=$y;
                  $zz2[$i]=$z;
            print "$a $x $y $z\n";
            unless($el==1){
                $nm2h[$hi]=$a;
                $xx2h[$hi]=$x;
                $yy2h[$hi]=$y;
                $zz2h[$hi]=$z;
                $hi++;
            };
        $i++;
           }else{
            unless($el==1){
                $xxm2[$mi]=$x;;
                $yym2[$mi]=$y;;
                $zzm2[$mi]=$z;;
                $mi++;
            };
        };
    };
};
$cpa2=$i;
$cpa2h=$hi;
$cpm2=$mi;
if($cpa1 != $cpa2){
        die "Number of QM atoms file $ARGV[$arg-2] is $cpa1 and
in file $ARGV[$arg-1] is $cpa2 disagreement!!!!\n";
};
if($cpa1h != $cpa2h){
        die "Number of the heavy QM atoms in file $ARGV[$arg-2]
is $cpa1h and in file $ARGV[$arg-1] is $cpa2h disagreement!!!!\n";
};
if($cpm1 != $cpm2){
        die "Number of the heavy MM atoms in file $ARGV[$arg-2]
is $cpm1 and in file $ARGV[$arg-1] is $cpm2 disagreement!!!!\n";
};
print STDERR "There are $cpa1 atoms including $cpa1h heavy atoms\n";


if($sym==0){
    for($i=0;$i<$cpa1h;$i++){
           $rms=$rms+($xx2h[$i]-$xx1h[$i])**2+($yy2h[$i]-
$yy1h[$i])**2+($zz2h[$i]-$zz1h[$i])**2;
     };
     $rmsd=sqrt($rms/$cpa1h);
     $bestrmsd=$rmsd;
     print STDERR "$cpa1h QM heavy atoms with RMSD=$rmsd\n";
}else{
#symetry operations
open (SYM,"$symetry") or die "File $symetry does not exist\n";
$i=0;
while($line=<SYM>){
        $line=~s/^\s*//;
        @tmp=split(/\s+/,$line);
        $tmpn=@tmp;
        $nsyat[$i]=int($tmpn/2);
        for($k=0;$k<$nsyat[$i];$k++){
       $flag1=$flag2=0;
          for($l=0;$l<$cpa1h;$l++){
             if($nm1h[$l] eq $tmp[2*$k]){
           if($flag1==1){
                die "Duplicit marking of heavy atom $nm1h[$l] \n";
           };
           $flag1=1;
           $syat1[$i][$k]=$l;
          };
             if($nm1h[$l] eq $tmp[2*$k+1]){
           if($flag2==1){
                die "Duplicit marking of heavy atom $nm1h[$l] \n";
           };
           $flag2=1;
           $syat2[$i][$k]=$l;
             };
          };
        };
     $#tmp=-1;
        $i++;
};
$cpso=$i;
for($i=0;$i<2**$cpso;$i++){
     for($l=0;$l<$cpa1h;$l++){
           $tpx[$l]=$xx1h[$l];
           $tpy[$l]=$yy1h[$l];
           $tpz[$l]=$zz1h[$l];
     };
     $b=dec2bin($i);
     $symper[$i]=substr($b,-$cpso);
        for($sm=0;$sm<$cpso;$sm++){
          if(substr($symper[$i],$sm,1) eq "1"){
           for($k=0;$k<$nsyat[$sm];$k++){

     ($tpx[$syat1[$sm][$k]],$tpx[$syat2[$sm][$k]])=($tpx[$syat2[$sm
][$k]],$tpx[$syat1[$sm][$k]]);

     ($tpy[$syat1[$sm][$k]],$tpy[$syat2[$sm][$k]])=($tpy[$syat2[$sm
][$k]],$tpy[$syat1[$sm][$k]]);
```

```
    ($tpz[$syat1[$sm][$k]],$tpz[$syat2[$sm][$k]])=($tpz[$syat2[$sm
][$k]],$tpz[$syat1[$sm][$k]]);
        };
     };
      };
    $rms=0;
     for($l=0;$l<$cpa1h;$l++){
        $rms=$rms+($xx2h[$l]-$tpx[$l])**2+($yy2h[$l]-
$tpy[$l])**2+($zz2h[$l]-$tpz[$l])**2;
    };
    $rmss[$i]=sqrt($rms/$cpa1h);
      print STDERR "RMS for symetrical operation $symper[$i] is
$rmss[$i]\n";

};
$bestrms=0;
for($i=1;$i<2**$cpso;$i++){
    if($rmss[$i] < $rmss[$bestrms]){
        $bestrms=$i;
    };
};
$bestrmsd=$rmss[$bestrms];
print STDERR "The best rms of all symetrical operations is
$rmss[$bestrms] of symetrical operation $symper[$bestrms]\n";
};
printf AGR "%s %s
%8.4f\n",$filename[$prvni],$filename[$druhy],$bestrmsd;
close(MAE2);
};
close(MAE1);
};


#--------------------------------------------------------------------
--------------------------------
sub dec2bin{
    my $str=unpack("B32", pack("N",shift));
    return $str;
};
```

2.2 In house polarization script for polarization of docked pose

```
#!/usr/bin/perl
$dir="/software_repository/polarization_reha/ian_script";
$covergence=0.1;
$arg=@ARGV;
$schrod=$ENV{'SCHRODINGER'};
if($arg<2){
      die "Need residuum names for QM and mae file\n";
};
$xyz=$ARGV[$arg-1];
$VYB="";
for($i=0;$i<$arg-1;$i++){
      $VYB=sprintf("%s %s",$VYB,$ARGV[$i]);
};
unless(($infile)=$xyz=~/^([\w\.]+).mae$/){
      die "Improper name for mae file($xyz)\n";
};
unless(-e "$infile.in") {
      die "$infile.in does not exist\n";
};
$infilep=$infile;
$infile=sprintf("%s_%02d",$infilep,0);
$infilej=sprintf("%sJ%02d",$infilep,0);
print "$infilep $infile\n";
mkdir "00";
chdir "00";
`cp ../$infilep.in $infile.fin.in`;
`cp ../$xyz $infile.fin.mae`;
if(-f "atomsPP"){
      `cp ../atomsPP atomsPF`;
}else{
      `touch atomsPF`;
};
print "Copied\n";
$opst=1;
for($z=0;$z<5;$z++){
      $opstf=sprintf("%02d",$opst);
      $opstpf=sprintf("%02d",$opst-1);
      print "Cycle $opst\n";
      mkdir "../$opstf";
      chdir "../$opstf";
      print "Created changes\n";
      system "pwd";
      open(IFL,"../$opstpf/$infile.fin.in") or die "cannot open the
file ../$opstpf/$infile.fin.in\n";
      open(IFLO,">$infilej.in");
      print IFLO "GPTSFILE: $infile.pts\n";
      print "Read init\n";
      while($line=<IFL>){
            if($line=~/GPTSFILE:/){
                  next;
            };
            if($line=~/MAEFILE:/){
                  next;
```

```
        };
        print IFLO "$line";
        if($line=~/\&gen/){
            last;
        };
    };
    print IFLO "gcharge=-6\nip172=2\n";
    $igeopt=0;
    while($line=<IFL>){
        if($line=~/gcharge=-6/ or $line=~/ip172=2/){
            next;
        };
        if($line=~/mmqm=1/){
            next;
        };
        if($line=~/igeopt/){
            unless($line=~/igeopt=0/){
                $igopt=$line;
                $igeopt=1;
                print IFLO "igeopt=0\n";
            }else{
                print IFLO  $line;
            };
        }else{
            print IFLO $line;
        };
        if($line=~/\&$/){
            last;
        };

    };
    close(IFLO);
    close(IFL);
    `cp ../$opstpf/$infile.fin.mae $infile.mae`;
    `cp ../$opstpf/atomsPF atomsPP`;
    `$dir/extractMM_filed_mae2pts $VYB $infile.mae >$infile.pts`;
    `$dir/extractQMmae2jag $VYB $infile.mae >>$infilej.in`;
    `$dir/extractMMmae2jag $VYB atomsPP $infile.mae
>>$infilej.in`;
    `$dir/extractGUESSjag ../$opstpf/$infile.fin.in
>>$infilej.in`;
    print "$schrod/jaguar run -WAIT $infilej.in\n";
#   exit;
    `$schrod/jaguar run -WAIT $infilej.in`;
    print "After schrod\n";
    `$dir/readESP_J $infilej.resp >field0`;
    print "$dir/polasignMMmae_def $VYB $infile.mae >XYZS\n";
    `$dir/polasignMMmae_def $VYB $infile.mae >XYZS`;
    $fiel=`wc -l field0`;
    $ccc=`wc -l XYZS`;
    unless($fiel==$ccc){
        die "Disagreement between field0 and XYZS";
    };
    `paste -d ' ' field0 XYZS >FCS`;
    `$dir/calc_indpolF FCS`;
    `cp ene2 ene2P0`;
    $enex=`cat ene2`;
    ($enep)=$enex=~/^\s*([-+]?\d+\.\d+)/;
    print "Energy $enep\n";
    `cp atoms2 atomsP00`;
    `cp com comP00`;
    `cp centr centrP00`;
    unless(-e "$infilej.01.in"){
        die "File $infilej.01.in does not exist, probably qsite
run crushed\n";
    };
    $bstd=1;
    $num=1;
    while ( $bstd > $covergence){
        $numa=$num-1;
        $numf=sprintf("%02d",$num);
        $numaf=sprintf("%02d",$num-1);
        $numff=sprintf("%02d",$num+1);
        `cp $infilej.$numf.in $infilej.$numf.in.tmp`;
        open(IFA,"$infilej.$numf.in.tmp");
        open(IFAO,">$infilej.$numf.in");
        while($line=<IFA>){
            if($line=~/&pointch/){
                last;
            };
            print IFAO $line;
        };
        close(IFAO);
        `$dir/extractMMmae2jag $VYB atomsP$numaf $infile.mae
>>$infilej.$numf.in`;
        open(IFAO,">>$infilej.$numf.in");
        while($line=<IFA>){
            if($line=~/&\s*$/){
                last;
            };
        };
        while($line=<IFA>){
            print IFAO $line;
        };
        close(IFAO);
        close(IFA);
        print "$schrod/jaguar run -WAIT $infilej.$numf.in\n";
        `$schrod/jaguar run -WAIT $infilej.$numf.in`;
        `$dir/readESP_J $infilej.$numf.resp >field$numf`;
        $fiel=`wc -l field$numf`;
        $ccc=`wc -l centrP$numaf`;
        unless($fiel==$ccc){
            die "Disagreement between field$numf and
centrP$numaf";
        };
        `paste -d ' ' field$numf centrP$numaf >FCS$numf`;
        `$dir/calc_indpolF FCS$numf`;
        `cp ene2 ene2P$numf`;
        $enex=`cat ene2`;
        ($ene)=$enex=~/^\s*([-+]?\d+\.\d+)/;
        `cp atoms2 atomsP$numf`;
        `cp com comP$numf`;
```

```
          `cp centr centrP$numf`;
          unless(-e "$infilej.$numff.in"){
               die "File $infilej.$numff.in does not exist,
probably qsite run crushed\n";
          };
          $bstd=abs(($ene-$enep))/$ene;
          print "Energy $ene\n";
          print  "Difference $bstd\n";
          $enep=$ene;
          $num++;
     };
     `cp atomsP$numf atomsPF`;
     `$dir/readinmae_indchg $VYB atomsP$numf
../$opstpf/$infile.fin.mae >$infile.mae`;
     open(IFL,"../$opstpf/$infile.fin.in") or die "cannot open the
file ../$opstpf/$infile.in\n";
     open(IFLO,">$infile.in");
     while($line=<IFL>){
          if($line=~/MAEFILE/){
               print IFLO "MAEFILE: $infile.mae\n";
               next;
          };
          print IFLO $line;
          if($line=~/\&mmkey/){
               last;
          };
     };
       print IFLO "use_mae_charges=YES\n";
     while($line=<IFL>){
          if($line=~/&guess/){
               last;
          };
          unless($line=~/use_mae_charges/){
               print IFLO $line;
          };
     };
     close(IFL);
     close(IFLO);
     `$dir/extractGUESSjag $infilej.$numff.in >>$infile.in`;
     print "$dir/extractGUESSjag $infilej.$numff.in
>>$infile.in\n";
     `$schrod/qsite -WAIT $infile.in`;
     print "$igeopt\n";
     if($igeopt!=0){
          print "Optimization\n";
          unless(-e "$infile.01.in"){
               die "File $infile.01.in does not exist, probably
qsite run crushed\n";
          };
          $infilef=sprintf("%s_%02d",$infilep,$opst);
          `cp $infile.01.in $infilef.fin.in`;
          `$dir/readoutmae_indchg $VYB atomsPF $infile.01.mae
>$infilef.fin.mae`;
     }else{
          last;
     };
```

```
	$infile=$infilef;
	$infilej=sprintf("%sJ%02d",$infilep,$opst);
	$opst++;
};
```

3.1 Gromacs input file

```
title		= mini
cpp		= /lib/cpp
define		= -DFLEX_TIP4P
constraints	= none
integrator = steep
dt		= 0.002
nsteps		= 50000
nstlist		= 10
ns_type		= grid
rlist		= 0.9
coulombtype	= PME
rcoulomb	= 0.9
rvdw		= 1.0
fourierspacing	= 0.12
fourier_nx = 0
fourier_ny = 0
fourier_nz = 0
pme_order	= 4
ewald_rtol = 1e-5
optimize_fft	= yes

emtol		= 1000.0
emstep		= 0.01
```

4.1 Script to generate AMBER force field

```
source leaprc.gaff
source leaprc.ff14SB

# Load special FF files for phoshpotyrosine
loadamberparams 16D.frcmod
#loadoff par_files/Y2P.off

# Note: .off and .lib files are the same

#set default disulfide auto

loadamberprep 16D.prepi
mol1 = loadpdb 1WOG.pdb
mol2 = loadpdb NEWPDB.PDB
fullmol = combine{mol1, mol2}
```

```
solvatebox fullmol TIP3PBOX 12.0
#addions fullmol Na+ 0
#addions fullmol Cl- 0
saveamberparm fullmol 1WOG_solvated.prmtop 1WOG_solvated.inpcrd
savepdb fullmol 1WOG_solvate.pdb

quit
```

4.2 ACEMD input file

```
# Configure time variables
set steps_min   500     ; # Number of steps to minimize
set steps_nvt   25   ; # Number of steps for NVT
set steps_npt1  250   ; # Number of steps for NPT with constraints
set steps_npt2  250 ; # Number of steps for NPT without constraints
set numSteps    [expr $steps_nvt + $steps_npt1 + $steps_npt2]    ; #
Total number of steps for the simulation.

# Set reusable variables
set inputname   input
set outputname  nve
set structure   1F8E_solvated
set parameters  parameters
set temperature 300
set logfreq     1000

# Set inputs
#structure       $structure.psf
#coordinates     $structure.pdb
#bincoordinates  $structure.coor
#binvelocities   $structure.vel
#parameters      $parameters

# Set outputs
energyfreq      $logfreq
restart         on
restartfreq     5000
restartname     $outputname.restart
outputname      $outputname
dcdfreq         25000
dcdfile         $outputname.dcd

# Set box dimensions, manually or via extendedsystem
celldimension 82 90 91
#extendedsystem  $inputname.xsc

# Configure holonomic restraints
rigidbonds      all

# Configure integration
timestep        4
hydrogenscale   4
```

```
# Configure electrostatics
pme             on
pmegridspacing  1.0
#pmefreq         2
cutoff          9
switching       on
switchdist      7.5
exclude         scaled1-4
1-4scaling      1.0
fullelectfrequency 2

# Configure positional restraints, if any
constraints     on
consref         $structure.restrain.pdb
constraintscaling   1.0

# Configure thermostat
langevin        on
langevintemp    $temperature
langevindamping 1

# Configure barostat
berendsenpressure   on
berendsenpressuretarget 1.01325
berendsenpressurerelaxationtime  800
#useconstantratio on ; # For use with membrane systems

# Amber settings
amber on
coordinates $structure.pdb
parmfile    $structure.prmtop
1-4scaling      0.833333333333

# Run minimization
minimize $steps_min
# Run simulation
run $numSteps
```

4.3 TCL script to ease off constraints

```
tclforces on
tclforcesfreq   1

# Relaxation variables
set scaling_original 1.0
set scalefreq 25 ; # Must be a factor of $steps_npt1
set downscale 0.8

set t [getstep]
set scalefactor [expr floor($t/$scalefreq)]
set scalemax [expr ceil($steps_npt1/$scalefreq)]
set scaling [expr $scaling_original*pow($downscale, $scalefactor)]
constraintscaling $scaling
```

```
proc calcforces_init {} {
  global steps_nvt steps_min
  global steps_nvt steps_npt1 steps_npt2
  global scaling_original scaling scalefactor scalefreq scalemax
downscale

  set t [ getstep ]
  if { $t == 0 } {
    puts "Running $steps_min steps of energy minimization"
  } else {
    puts "Restarting"
    if { $t < $steps_nvt } {
      puts "Running $steps_nvt steps of NVT with constraints"
      berendsenpressure off
      tclforcesfreq [expr $steps_nvt ]
    } elseif { $t < [expr $steps_nvt + $steps_npt1 ] } {
      if { $steps_npt1 % $scalefreq == 0 } {
        tclforcesfreq $scalefreq
        set scalefactor [expr floor(($t-$steps_nvt)/$scalefreq)]
        set scaling [expr $scaling_original*pow($downscale,
$scalefactor)]
        constraintscaling $scaling
        puts "STEP:     $t"
        puts "SCALING:  $scaling"
        puts "SCALEFACTOR: $scalefactor"
      } else {
        # Dont do force scaling if $steps_npt1 not a multiple of
$scalefreq
        set scalefactor $scalemax
        tclforcesfreq [expr $steps_nvt + $steps_npt1 ]
        puts "Force scaling will not be performed because
\$scalefreq not a factor of \$steps_npt1"
      }
    } else {
      puts "Running $steps_npt2 steps of NPT without constraints"
      constraintsclaing 0.
    }
  }
}

proc calcforces {} {
  global steps_nvt steps_npt1 steps_npt2
  global scaling_original scaling scalefactor scalefreq scalemax
downscale
  set t [ getstep ]

  if { $t == 1 } {
    puts "Running $steps_nvt steps of NVT with constraints"
    berendsenpressure off
    tclforcesfreq $steps_nvt
  }
  if { $t == $steps_nvt } {
    # turn barostat on
    puts "Running $steps_npt1 steps of NPT with constraints"
    berendsenpressure on
    if { $steps_npt1 % $scalefreq == 0 } {
      tclforcesfreq $scalefreq
      set scalefactor [expr floor(($t-$steps_nvt)/$scalefreq)]
      set scaling [expr $scaling_original*pow($downscale,
$scalefactor)]
      constraintscaling $scaling
      puts "STEP:     $t"
      puts "SCALING:  $scaling"
      puts "SCALEFACTOR: $scalefactor"
    } else {
      # Dont do force scaling if $steps_npt1 not a multiple of
$scalefreq
      set scalefactor $scalemax
      tclforcesfreq [expr $steps_nvt + $steps_npt1 ]
      puts "Force scaling will not be performed because \$scalefreq
not a factor of \$steps_npt1"
    }
  }
  #if { $t > $steps_nvt && [expr $steps_nvt +
$scalefreq*$scalefactor] < [expr $steps_nvt + $steps_npt1] } {}
  if { $t > $steps_nvt && [expr $scalefreq*$scalefactor] < [expr
$steps_npt1] } {
    # Gradually reduce constraint scaling
    set scalefactor [expr floor(($t-$steps_nvt)/$scalefreq)]
    set scaling [expr $scaling_original*pow($downscale,
$scalefactor)]
    constraintscaling $scaling
    puts "STEP:     $t"
    puts "SCALING:  $scaling"
    puts "SCALEFACTOR: $scalefactor"
  }
  if { $t == [expr $steps_nvt + $steps_npt1] } {
    # turn constraints off
    constraintscaling  0.
    puts "Running $steps_npt2 steps of NPT without constraints"
  }

}

proc calcforces_endstep { } { }
```

5.1 ACMED input file for membrane simulations

```
# Configure time variables
set steps_min   500     ; # Number of steps to minimize
set steps_nvt   250000  ; # Number of steps for NVT
set steps_npt1  2500000 ; # Number of steps for NPT with constraints
set steps_npt2  2500000 ; # Number of steps for NPT without
constraints
set run1        5000000 ; # temp1
set run2        5000000 ; # temp2
set run3        5000000 ; # temp3
set run4        5000000 ; # temp4
set run5        5000000 ; # temp5
#set run6        2500000 ; # temp6
#set run7        2500000 ; # temp7
#set run8        2500000 ; # temp8
#set run9        2500000 ; # temp9
#set run10       2500000 ; # temp10
set numSteps    [expr $steps_nvt + $steps_npt1 + $steps_npt2 + $run1
+ $run2 + $run3 + $run4 + $run5 ]  ; # Total number of steps for the
simulation.

# Set reusable variables
set inputname   input
set outputname  2VT4_x2_temp_output
set structure   2VT4+ligand_solvated
set parameters  parameters
set temperature 289
set logfreq     15000

# Set inputs
#structure       $structure.psf
coordinates     $structure.pdb
#bincoordinates  $structure.coor
#binvelocities   $structure.vel
parameters      $parameters

# Set outputs
energyfreq      $logfreq
restart         on
restartfreq     30000
restartname     $outputname.restart
outputname      $outputname
dcdfreq         15000
dcdfile         $outputname.dcd

# Set box dimensions, manually or via extendedsystem
#celldimension
extendedsystem  input.xsc

# Configure holonomic restraints
rigidbonds      all

# Configure integration
timestep        4
hydrogenscale   4

# Configure electrostatics
pme             on
pmegridspacing  0.16
#pmefreq         2
cutoff          11
switching       on
switchdist      7.5
exclude         scaled1-4
```

```
1-4scaling      1.0
fullelectfrequency 2
pmegridsizex    100
pmegridsizey    100
pmegridsizez    100

# Configure positional restraints, if any
constraints     on
consref         $structure.restrained.pdb
constraintscaling   1.0

# Configure thermostat
langevin        on
langevintemp    $temperature
langevindamping 1

# Configure barostat
berendsenpressure   on
berendsenpressuretarget 1.01325
berendsenpressurerelaxationtime  800
useconstantratio on ; # For use with membrane systems

# Amber settings
amber on
coordinates $structure.pdb
parmfile    $structure.prmtop
1-4scaling      0.833333333333

# Run minimization
minimize $steps_min
# Run simulation
run $numSteps
```

5.2 TCL script to change temperature during simulation

```
tclforces on
tclforcesfreq   1

# Relaxation variables
set scaling_original 1.0
set scalefreq 25000 ; # Must be a factor of $steps_npt1
set downscale 0.8

set t [getstep]
set scalefactor [expr floor($t/$scalefreq)]
set scalemax [expr ceil($steps_npt1/$scalefreq)]
set scaling [expr $scaling_original*pow($downscale, $scalefactor)]
constraintscaling $scaling

proc calcforces_init {} {
  global steps_nvt steps_min
  global steps_nvt steps_npt1 steps_npt2
  global scaling_original scaling scalefactor scalefreq scalemax
downscale

  set t [ getstep ]
  if { $t == 0 } {
    puts "Running $steps_min steps of energy minimization"
  } else {
    puts "Restarting"
    if { $t < $steps_nvt } {
      puts "Running $steps_nvt steps of NVT with constraints"
      berendsenpressure off
      tclforcesfreq [expr $steps_nvt ]
    } elseif { $t < [expr $steps_nvt + $steps_npt1 ] } {
      if { $steps_npt1 % $scalefreq == 0 } {
        tclforcesfreq $scalefreq
        set scalefactor [expr floor(($t-$steps_nvt)/$scalefreq)]
        set scaling [expr $scaling_original*pow($downscale,
$scalefactor)]
        constraintscaling $scaling
        puts "STEP:     $t"
        puts "SCALING:  $scaling"
        puts "SCALEFACTOR: $scalefactor"
      } else {
        # Dont do force scaling if $steps_npt1 not a multiple of
$scalefreq
        set scalefactor $scalemax
        tclforcesfreq [expr $steps_nvt + $steps_npt1 ]
        puts "Force scaling will not be performed because
\$scalefreq not a factor of \$steps_npt1"
      }
    } else {
      puts "Running $steps_npt2 steps of NPT without constraints"
      constraintscaling 0.
    }
  }
}

proc calcforces {} {
  global steps_nvt steps_npt1 steps_npt2 run1 run2 run3 run4 run5
run6 run7 run8 run9
  global scaling_original scaling scalefactor scalefreq scalemax
downscale
  set t [ getstep ]
  set outputname 2VT4_step

  if { $t == 1 } {
    puts "Running $steps_nvt steps of NVT with constraints"
    berendsenpressure off
    tclforcesfreq $steps_nvt
  }
  if { $t == $steps_nvt } {
    # turn barostat on
    puts "Running $steps_npt1 steps of NPT with constraints"
    berendsenpressure on
    if { $steps_npt1 % $scalefreq == 0 } {
      tclforcesfreq $scalefreq
      set scalefactor [expr floor(($t-$steps_nvt)/$scalefreq)]
      set scaling [expr $scaling_original*pow($downscale,
$scalefactor)]
      constraintscaling $scaling
      puts "STEP:     $t"
      puts "SCALING:  $scaling"
      puts "SCALEFACTOR: $scalefactor"
    } else {
      # Dont do force scaling if $steps_npt1 not a multiple of
$scalefreq
      set scalefactor $scalemax
      tclforcesfreq [expr $steps_nvt + $steps_npt1 ]
      puts "Force scaling will not be performed because \$scalefreq
not a factor of \$steps_npt1"
    }
  }
  #if { $t > $steps_nvt && [expr $steps_nvt +
$scalefreq*$scalefactor] < [expr $steps_nvt + $steps_npt1] } {}
  if { $t > $steps_nvt && [expr $scalefreq*$scalefactor] < [expr
$steps_npt1] } {
    # Gradually reduce constraint scaling
    set scalefactor [expr floor(($t-$steps_nvt)/$scalefreq)]
    set scaling [expr $scaling_original*pow($downscale,
$scalefactor)]
    constraintscaling $scaling
    puts "STEP:     $t"
    puts "SCALING:  $scaling"
    puts "SCALEFACTOR: $scalefactor"
  }
  if { $t == [expr $steps_nvt + $steps_npt1] } {
    # turn constraints off
    constraintscaling  0.
    puts "Running $steps_npt2 steps of NPT without constraints"
  }

if { $t == $steps_nvt + $steps_npt1 + $steps_npt2 } {
   puts "Production run @300 for 40ns"
   berendsenpressure off
   langevintemp 300
   langevindamping 0.1
   constraintscaling 0.
   restartname     $outputname.300.restart
  }

if { $t == $steps_nvt + $steps_npt1 + $steps_npt2 + $run1 } {
   puts "Production run @308 for 40ns"
   berendsenpressure off
   langevintemp 308
   langevindamping 0.1
   constraintscaling 0.
   restartname     $outputname.308.restart
  }

if { $t == $steps_nvt + $steps_npt1 + $steps_npt2 + $run1 + $run2 }
{
   puts "Production run @316 for 40ns"
   berendsenpressure off
   langevintemp 316
   langevindamping 0.1
   constraintscaling 0.
   restartname     $outputname.316.restart
  }

if { $t == $steps_nvt + $steps_npt1 + $steps_npt2 + $run1 + $run2 +
$run3 } {
   puts "Production run @324 for 40ns"
   berendsenpressure off
   langevintemp 324
   langevindamping 0.1
   constraintscaling 0.
   restartname     $outputname.324.restart
  }
if { $t == $steps_nvt + $steps_npt1 + $steps_npt2 + $run1 + $run2 +
$run3 + $run4 } {
   puts "Production run @332 for 40ns"
   berendsenpressure off
   langevintemp 332
   langevindamping 0.1
   constraintscaling 0.
   restartname     $outputname.332.restart
  }

if { $t == $steps_nvt + $steps_npt1 + $steps_npt2 + $run1 + $run2 +
$run3 + $run4 + $run5 } {
   puts "Production run @340 for 40ns"
   berendsenpressure off
   langevintemp 340
   langevindamping 0.1
   constraintscaling 0.
   restartname     $outputname.340.restart
  }
}

proc calcforces_endstep { } { }
```