Masteroppgave

NTNU
Norges teknisk-naturvitenskapelige universitet
Fakultet for naturvitenskap
Institutt for kjemi

Magnus Aleksander Henriksen

Molecular Dynamics Analysis of Enantioselective Mechanism of Chiral Thieno-, Furo-, and Pyrrolopyrimidine Kinase Inhibitors of EGFR

Masteroppgave i Industriell Kjemi- og bioteknologi
Veileder: Ida-Marie Høyvik
Juli 2020

NTNU
Kunnskap for en bedre verden

Magnus Aleksander Henriksen

Molecular Dynamics Analysis of Enantioselective Mechanism of Chiral Thieno-, Furo-, and Pyrrolopyrimidine Kinase Inhibitors of EGFR

Masteroppgave i Industriell Kjemi- og bioteknologi
Veileder: Ida-Marie Høyvik
Juli 2020

Norges teknisk-naturvitenskapelige universitet
Fakultet for naturvitenskap
Institutt for kjemi

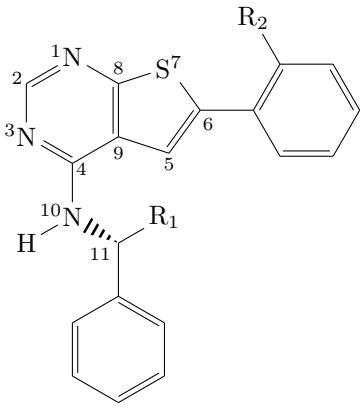
NTNU
Kunnskap for en bedre verden

Abstract

Epidermal growth factor receptor (EGFR) inhibitors interrupt EGFR-dependent cellular signaling pathways that lead to accelerated cancerous tumor growth and proliferation, and are actively developed for treatment of various types of non-small cell cancer. Here, we continue our investigation of an empirical chirality-potency relationship between the R/S enantiomers of thieno-, pyrrolo- and furopyrmidines when acting as Type I Epidermal Growth Factor Receptor Tyrosine Kinase (EGFR-TK) inhibitors, with the aim of providing a mechanism which relates molecular chirality to empirical measurements of inhibition.

Based on long Molecular Synamics simulations (1 µs) of ligand-in-receptor complexes between the active EGFR intracellular domain and inhibitors, we present qualitative evidence that the primary differentiator of potency is a combination of 3 stereo-specific interactions: a water-mediated hydrogen bond to Threonine-854, a pi-cation interaction with Lysine-745, and for methanol-containing inhibitors, a hydrogen bond to either Lysine-745 or Aspartate-855. These interactions are shown to occur more frequently for the high-potency enantiomers in our simulation. The water-bridge is a new addition which we previously couldn't have modeled in our gas-state simulation.

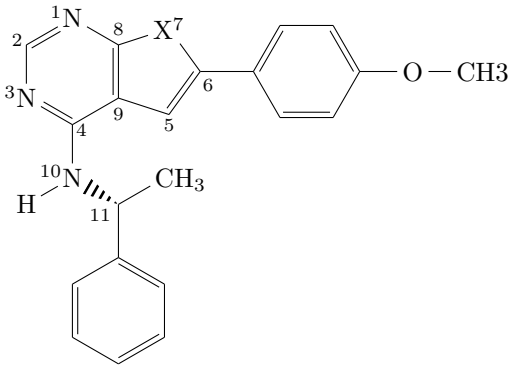
Our findings provide important insight for the design of EGFR inhibitors. More broadly, the results raise further questions about the role of water in ligand-receptor bonding, and add to a growing list of evidence that modeling of water is crucial in estimating the binding affinity of small molecule inhibitors.



1a: $R_1 = CH_3, R_2 = OH$

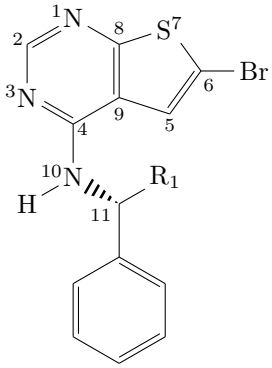
1b: $R_1 = CH_2OH, R_2 = OH$

1c: $R_1 = CH_2OH, R_2 = OCH_3$



2a: X=O

2b: X=NH



3a: $R_1 = CH_3$

3b: $R_1 = CH_2OH$

Chiral small-molecule receptor tyrosine kinase inhibitors studied in this thesis, here in their high-potency enantiomer.

Contents

1 Introduction **2**

2 Theoretical Background **8**

2.1 Biological function of Epidermal Growth Factor Receptor 8

2.1.1 Structure of intracellular kinase domain of EGFR 8

2.2 Enzymatic assay and binding free energies [complete] 10

2.2.1 MM/GB-SA binding free energies 11

2.2.2 Protein-Ligand Simulation Interaction Analysis 12

3 Method **14**

3.1 Practical considerations of Molecular Dynamics 14

3.2 Model system preparation 15

3.3 Simulation details 16

3.4 Data post-processing and analysis 17

3.4.1 Representative binding pose geometries 18

4 Results and Discussion **20**

4.1 MM/GB-SA Binding energy 20

4.2 Stability of the protein 20

4.3 Average inhibitor conformation - RMSD and cluster analysis 26

4.4 Residue-ligand interaction analysis 32

4.5 Summary of results 39

4.6 Retrospective of earlier study 40

4.7 Further work 43

5 Conclusion **45**

List of Figures **46**

List of Tables **48**

Bibliography **49**

Abbreviations

ATP	Adenosine triphosphate
ASP	Asparartate
EGFR	Epidermal Growth Factor Receptor
EGF	Epidermal Growth Factor
GLU	Glutamine
LIE	Ligand Interaction Energy
LYS	Lysine
MC	Monte Carlo simulations
MD	Molecular Dynamics simulations
MM	Molecular Mechanics
MM/GB-SA	Molecular Mechanics/Generalized Born Surface Area energy calculation
OPLS	Optimized Potentials for Liquid Simulations (force field)
QM/MM	Hybrid quantum mechanical and molecular mechanical method
RTK	Receptor tyrosine kinase
RMSD	Root Mean Square Deviation
THR	Threonine
XRD	X-Ray Diffraction

1 Introduction

Epidermal growth factor receptor (EGFR) is a transmembrane cell surface protein that is a receptor for the Epidermal Growth Factor (EGF) family of extracellular protein ligands. EGFR is part of the ErbB family of four receptor tyrosine kinases (EGFR/Her1, Neu/Her2, Her3, and Her4) whose primary biological function is to regulate cell growth [1]. However, they have also been shown to play a large role in several types of non-small cell cancer, including breast, lung, esophageal, and head and neck cancers [2,3]. Under certain conditions, as a result of over-expression, mutation, or co-expression of the growth factors and the receptor, these receptors can become hyperactivated; the result of this is uncontrolled cell proliferation. [4] Due to their multi-dimensional role in the progression of cancer, EGFR and its family members have emerged as popular targets for anti-cancer therapy [5], particularly small-molecule kinase inhibitors [6].

The first such inhibitor to be made commercially available was Erlotinib (under the brand name of Tarceva) in 2004, [7] and later FDA-approved kinase inhibitors include Gefitinib and Lapatinib [8]. These inhibitors compete directly with ATP and bind to EGFR in its place, preventing phosphorylation of target tyrosine residues and stopping the signal cascade. However, prolonged administration of 1st generation EGFR-TK inhibitors often leads to patients becoming immune to the drugs due to mutations in EFGR [9], necessitating further developments in mutation-resistant inhibitors.

In the last decade we have been developing ATP-competitive EFGR-inhibitors based on thieno- [10], pyrrolo- [11] and furopyrimidines [12], structurally inspired by the pyrrolopyrimidine AEE788, an EGFR inhibitor first elucidated in 2004 [13]. A structure-activity relationship (SAR) study in 2015 [14] synthesized and evaluated 44 such small molecule kinase inhibitors against EGFR, many of which had chiral substitutes. For all chiral compounds, the chirality was observed to have a surprisingly strong effect upon the biological activity of the inhibitor; when separated into enantiopure solutions, several enantiomers were reported to have a thousandfold difference in potency between the most active and least active enantiomer, with the racemic mixture following the trend of the most active enantiomer. This means that the active enantiomer competes strongly with the native ATP ligand in binding to EGFR, while the other enantiomer is much less competitive.

A chiral compound is one which is not superposable on its mirror image through any combination of translation and rotation. [15] In the case of the EGFR inhibitors, the chirality results from an amine substituent on the pyrrolpyrimidine scaffold whose carbon chain contains a chiral carbon bound to the amine, a phenyl group, a hydrogen, and either a methyl or a methanol group. In a symmetric environment, such as in gas or solvated in water, the energy difference between the mirror images is exactly zero. In proteins however, proteinogenic amino acids (save glycine) have at least one chiral center at C_α. Threonine and isoleucine have an additional chiral center at C_β. Further, only one of the two enantiomers is widely used in nature: according to the D-/L-naming convention, most naturally occurring amino acids are found in the L-configuration. Since proteins contain hundreds of amino acids, they are highly asymmetric. Thus,

it is not unexpected to see differences in binding affinity for chiral inhibitors bound to an enzyme like EGFR. However, in our case, two of the groups on the chiral center, hydrogen and methyl/methanol, are much smaller than the two ring structures, and the binding pocket appears to have sufficient space for them both. How is it that the configuration of methyl, a small, non-polar, functional group is able to produce such a large effect?

This is the central problem of this thesis:

> What is the structural mechanism that explains why the potency of 1-phenylethylamine-substituted furo-, thieno- and pyrrolopyrimidines as EGFR inhibitors depends so strongly on their stereoisomery?

Developing a theory that relates how the microscopic change in structure affects the macroscopic activity is a highly non-trivial task. The potential search space for explanations is enormous due to the vast amount of atomic, molecular and supramolecular interactions that take place in a biomolecule. Expressing, evaluating and validating all explanations would be the work of several lifetimes. Fortunately for the scope of this thesis, we have two factors that drastically cut down on the amount of work. The first is that EGFR is a very well studied enzyme, with a large amount of papers written about the protein's structure[16], dynamics[17], mechanisms[18] and interactions with other inhibitors[19], and crucially, three-dimensional protein structures submitted to the Protein Databank (PDB). The second is access to advanced computational modelling tools and the computational resources to use them, which allows us to both model atomic interactions directly and to do so much faster than is possible in a laboratory setting.

In the experimental studies of our inhibitors, potency was measured by IC_{50}. This is the concentration at which the natural reaction catalyzed by the enzyme (a tyrosine-phosphorylation) reacts at 50% of the non-inhibited rate. This is one of several standard ways of reporting the *binding affinity* of the inhibitor to the receptor - how strongly the inhibitor interacts with the receptor[20]. If the system is at equilibrium, the IC_{50} value can be directly related to the change in binding free energy[21] - which can be calculated computationally, provided we use realistic model systems and accurate energy calculations. By looking at which atomic interactions contribute to the differences in binding free energy, we would therefore have a top down approach to finding a theoretical explanation. Unfortunately, this has already proven itself to be a rather difficult task to accomplish in our earlier work[22].

Previously, we have used the Protein Database to create 3D models of the receptor-inhibitor complexes of our kinase inhibitors docked to EGFR. These 3D models were used for calculating docking scores using quick and simplified docking methods such as Glide and Induced Fit Docking,[23] which are intended for high-throughput virtual screening of compounds using a rigid receptor model (the familiar lock-and-key model of enzymatic reactions). Unfortunately these methods were found to be too inaccurate for the

task in terms of how the docked ligand-receptor system behaved; neither score nor structure showed clear differences between the enantiomers [24]. With that in mind, we attempted to model the system using a hybrid quantum mechanics and molecular mechanics method, arguing that the flexibility of the receptor as well as the quantum mechanical interactions between the ligand and the binding pocket might be the reason for the enantioselectivity. Here our focus was to quantify how strongly the inhibitors interacted with various residues in the binding pocket; Again we found no clear indications for one enantiomer being more stable than the other. Additionally, we had considerable difficulty computing more accurate energies and minimized energy structures due to computational cost. [22]

Heading back to the drawing board, we considered further approaches and addressed deficiencies in our models. One factor we hadn't considered yet was the impact of the surrounding solvent. It is well known that structures in the crystal and in solution differ in several important respects, such as radius of gyration, solvent accessible surface, intramolecular hydrogen bonds, and orientation of surface side chains. Indeed, recent papers modelling enzyme-ligand bonding mechanisms highlight the importance of properly modeling water [25,26]. The fast docking methods used implicit solvent surfaces along with three structural water molecules that came from the PDB crystal structure, while the QM/MM experiment modeled only the gas-state geometry. We haven't yet modeled the dynamic movement of the solvent in and around the inhibitor-receptor complex - a deficiency we now seek to address.

Initially we looked into using a polarizeable continuum model, [27] which models solvent effects by treating the solvent as a continuum that surrounds the solute; however, while studies indicate these can reproduce hydrogen bonding with the solvent, or at least the energetic contribution from such, a purely implicit solvent model struggles to reproduce effects that occur due to buried waters and hydrogen bonding networks. [28] These effects are highly likely to be important for us; the fact that the PDB crystal already contains three structural water molecules is a strong indication that water bridging - polar interactions between the ligand and the receptor mediated by waters - is an important phenomenon, something that can be attested by other studies of EGFR-inhibitors identifying a water bridge to Threonine-854 as an important stabilizing effect for pyrrolopyrimidine-based compounds [29]. As such, we saw the need for our models to incorporate explicit water molecules, so that such bridges can be modeled properly.

Adding explicit solvent also raises a second issue - at normal biological temperatures of 300K, water is a highly disordered liquid, not a solid. In order for water to behave like a liquid in our simulation, we need to account for the dynamic motions of the explicit water molecules which occur due temperature, and therefore we need to account for kinetic energy - which almost immediately implies we have to abandon the notion of a single energy-minimized structure due to the stochastic nature of temperature.

This in turn means quantum mechanical descriptions of the system states are infeasibly expensive - we had already run into problems with making *one* gas-state QM/MM geometry for each inhibitor - so we decided to drop the quantum mechanical parts of our computation completely. Instead we decide to

adopt purely molecular mechanics based approaches. Molecular mechnics, or MM, are based around force fields; approximations of the interatomic forces which are adjusted to reproduce results from quantum mechanical calculations and, typically, to certain empirical measurements. For example, a typical force field incorporates Coulombic terms describing electrostatic interactions between atoms, spring-like terms that model the preferred bond lengths and bond angles, terms describing Van der Waahls forces between atoms, and terms that amount to empirical corrections. Such force fields are inherently approximate to a much greater degree than quantum mechanical calculations. Comparison of simulations to a variety of experimental data indicates that force fields have improved substantially over the past decade, particularly when it comes to approximating protein structures [30]; however, they still suffer from various inaccuracies, such as being unable to model breaking and formation of covalent bonds [31].

In return for the MM approximation, the cost of calculating interatomic forces decreases drastically, to the point that using the same resources and time as before, we may calculate several thousands of structures. In molecular modelling, and particularily in modeling of biochemical systems, there are two principal methods of doing such sampling: Monte Carlo (MC) stochastic sampling, and Molecular Dynamics (MD) simulations. MC uses a stochastic acceptance method to ensure the final samples are a Boltzmann distribution of the relevant system, while MD solves Newton's equations of motion to evolve the model system forward in time. In our case, the choice naturally falls to Molecular Dynamics, because we are interested in capturing the dynamics of the solvent and ligand interactions, something which Monte Carlo methods are less suited for.

Within the field of biochemical modelling, Molecular Dynamics is an increasingly used method to model protein-ligand complexes such as ours. According to a recent review of Molecular Dynamics applied to biochemical modeling, the number of yearly studies involving MD in top journals surpassed 1000 in 2017 [31]. Our problem domain is a very common one: use MD to estimate the binding free energy of a ligand to a receptor and investigate what influences this energy. Unfortunately, there is no one universally applied method to this problem, due to the sheer size of biomolecular systems making accurate prediction a very slow process; rather, methods for solving this problem are continually being developed on various scales of accuracy vs. time (ranging from ligand screening which can evaluate thousands of ligands per day to long MD simulations which spend days computing one ligand in particular), and literature contains extensive reviews on these various methods [32–37].

On the computationally expensive side, there are the methods of Free Energy Perturbation (FEP) [38] and Thermodynamic Integration (TI) [39], called "alchemical" methods [40], as they involve gradually perturbing the system by transforming one molecule into another and calculating the resulting energy difference (in the limit of ligand dissassociation, the molecule is gradually replaced by nothing). We would certainly have liked to apply these methods to our system, since in theory, our alchemical transformation would be a very simple exchange of one methyl/methanol group and one hydrogen. Unfortunately, due to financial and technological reasons, such methods were beyond what we could compute at this time.

Next to these alchemical transformations, there are quite a number of so called "enhanced sampling" methods. In principle, a straightforward MD simulation, when based on a reasonably accurate force field and including solvation effects, should be able to simulate the systems of interest completely. In practice, the time required for a ligand to unbind on its own, or for a protein to undergo a conformational change, is much greater than what is feasable to simulate. Enhanced sampling methods are meant to solve this, by speeding up exploration of phase space or directing the simulation to explore particular events. Such methods include umbrella sampling,[41] stochastic tunneling,[42] metadynamics,[43][1] parallell tempering,[44] steered molecular dynamics,[45] taboo search,[46] and multicanonical MD,[47] to name but a few. In the context of relative binding free energy, they are commonly used to sample the path from bound to free ligand and compute the energy along the way. A thorough review of all of these are beyond what we consider the scope of this thesis, but there are many reviews in the literature of various enhanced sampling methods applied to biological systems.[48,49]

Finally, there are the more straightforward MD approaches which do not involve simulating a path between bound and free ligand, but rather evaluates the free energy of the free and bound systems separately, so-called end-point methods. One example is the Linear Interaction Analysis method[50,51] which is an application linear response theory to free energy approximation where the free energy is estimated as the difference between the sum of forces on the bound and free ligand - this methods relies only on equilibrium simulations of the ligand bound to the complex and the ligand in free solution. Another pair of commonly used end point methods are MM-PBSA (Molecular Mechanics with Poisson-Boltzmann Surface Area) and MM-GBSA (Molecular Mechanics with Generalized Born Surface Area). Both of these methods applied to binding free energy entail the same kind of calculation: simulate the protein-ligand complex in explicit solvent (and optionally, free ligand and free protein), post-process the resulting trajectory by removing this explicit solvent, and use an implicit surface model to estimate the average solvation energy in addition to the molecular mechanics energy. They differ in how they approximate this implicit surface, with MM-PBSA numerically solving the Poisson-Boltzmann equation while MM-GBSA uses a generalized approximation that is less computationally expensive[52,53].

With the sheer variety of options available to us, we decided to stop and re-evaluate our approach with regards to the overall goal. As mentioned, the previous QM/MM study was concerned with quantifying the interaction strength between the inhibitor and the ligand. However, that presupposes that the molecular system studied is more or less the same as the real world system and that the difficulty is in computing accurate energies. Our previous findings show that this is not the case, and that we need to know more about how the protein-ligand complex actually behaves near equilibrium - which parts of the binding pocket are solvent accessible, which aren't? Which parts of the protein are flexible, which parts are more rigid due to strong interactions? How does the binding pocket vary - is it a cramped space with

[1] We mention that we did attempt to apply metadynamics to our model system over the course of the thesis work, but we abandoned those attempts due to time constraints and unconvincing results combined with uncertainty and inexperience with the method.

little room for the inhibitor to maneuver, or is the protein flexible enough that the inhibitor may adopt several conformations with similar energy minima? Which residues tend to interact with our inhibitors, how strongly, and how often? While many of these questions have specific methods designed to answer them that offer good trade-offs between speed and accuracy, the nature of our thesis problem means we do not, *a priori*, know *which* of these quantities are important - answering that had, after all, been the point of computing the binding free energy in the first place.

Therefore, we decided to to take one of the simplest and straightforward approaches: regular equilibrium MD simulations. Rather than focus on calculating the binding free energy accurately and then from that estimate which effects and interactions contribute to the difference in relative binding affinity between our enantiomers, our goal will be to inspect the equilibrium state of our inhibitors bound to EGFR in order to get a qualitative understanding of the molecular dynamics. Due to the inclusion of kinetic energy terms and simulated thermostats, MD simulations are able to equilibriate our system in a heated environment, rather than the absolute zero of our previous gas state quantum mechanics. By expanding the system size to include a large amount of explicit solvent molecules, we hope to capture solvent-ligand interactions in the binding pocket as well as stabilize the receptor geometry. Finally, we may simulate the inhibitor-EGFR complexes over a medium long time period of one microsecond, which we estimate will be sufficient to capture most dynamical events short of conformational changes in the protein.

The rest of this thesis is structured as follows: In the Theory section we explain the biochemical background necessary to understand the function of EGFR, as well as the theoretical basis for our simulation analysis. In the Methods section we explain how our experiments were set up and run, and we explain the rationale behind our practical decisions regarding the simulations. The Results and Discussion section presents summaries of the data gathered from these simulations, and pulls it together to answer our research question, along with commentary on application to further work. Finally, our Conclusion contains a summary of our findings.

2 Theoretical Background

2.1 Biological function of Epidermal Growth Factor Receptor

As mentioned in the introduction, EGFR is a transmembrane protein receptor for the EGF family of extracelullar protein ligands. Binding of an EGF ligand to the extracellular domain of EGFR triggers ligand-induced dimerization of the receptor [18], resulting in an assymmetric conformation of the paired intracellular EGFR domains. This activates the intracellular tyrosine kinase domain, which in turn autophosphorylates several tyrosine residues on EGFR itself, fully activating the enzymatic domain and stimulating binding of other key signal-transducing intracellular proteins such as GRB2 (Growth Factor Receptor-bound protein 2), causing a signalling cascade which eventually results in DNA synthesis and cell proliferation [18] - the details of this cascade are too broad to cover within the scope of this thesis, but a summary is shown in figure 2.1.

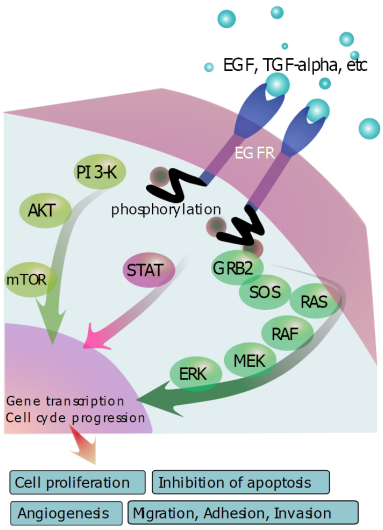


Figure 2.1: Outline of signal cascade path from activation of EGFR by growth factors. Public domain image taken from Wikipedia (Jan 2020)

In order to trigger the autophosphorylation and subsequent signal cascade, the tyrosine kinase domain requires ATP to donate the required phosphate groups. Small molecule tyrosine kinase inhibitors like erlotinib, AEE788 and our own series are reversible competitive antagonists to ATP. [54] binding to the tyrosine kinase domain in its place and preventing phosphorylation (i.e. they are type I tyrosine kinase inhibitors [55]). It is this kinase domain we model in the present thesis.

2.1.1 Structure of intracellular kinase domain of EGFR

The EGFR protein kinase domain is structured in two lobes (see Figure 2.2, showing the co-crystallized structure of EGFR with the inhibitor AEE788 (PDB:2J6M) [56]). The smaller N-terminus (red) lobe contains many β-sheets, while the larger C-terminus lobe is rich in α-helices. These lobes are connected by a hinge-region (yellow) which also forms the active site for ATP. Central to the structure is the activation loop (A-loop, green) running from Asp855 to Val726. A-loops are common in several protein kinases, where phosphorylation of an A-loop tyrosine acts as a common switch for activity. This has not been found to be the case for ERbB, however. In the inactive autoinhibited conformation, the DFG motif preceding the A-loop adopts a short helix structure (termed DFG-out), whereas growth factor-induced dimerization causes it to unwind (DFG-in), opening the active seat. Another regulatory mechanism is the orientation of the C-helix (orange) on the N-lobe; In the active form, a salt bridge can be observed between the phosphate groups of ATP and the C-lobe. In inactive kinases, however, this lobe moves out of position, resulting in the loss of this salt bridge interaction. [12,19]

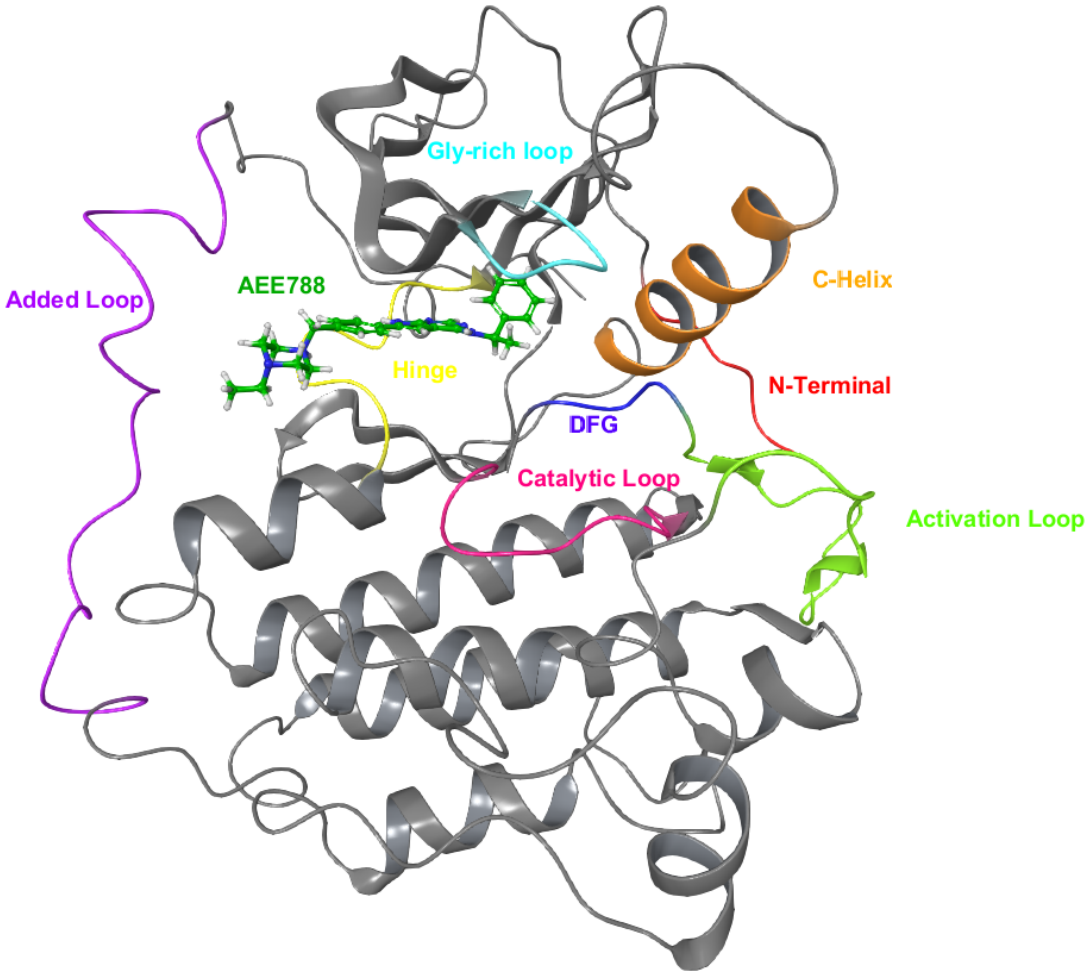


Figure 2.2: Crystal structure of EGFR inhibited by AEE788 (PDB:2J6M) [56] with key protein kinase features highlighted.

2.2 Enzymatic assay and binding free energies [complete]

Our empirical data on the potency of our inhibitors is based on a standard enzymatic assay performed by Invitrogen (LifeTechnology) using their Z'-LYTE®assay technology in an earlier paper[14]. In this assay, the conversion of a small-peptide molecule from unphosphorylated to phosphorylated state by the target enzyme (EGFR protein) is measured in absence and presence of the target inhibitor over the course of an hour, and inhibition is reported as interpolated IC_{50} [nM], the concentration of the inhibitor at which the enzymatic reaction reacts at 50% of the uninhibited rate. Thus the empirical data measures how competitive a given inhibitor is with the natural ATP ligand, with the most competitive inhibitors needing a very small dose to decrease enzymatic activity by half. In our case, both enantiopure and racemic mixtures of each inhibitor was made; generally, the reported IC_{50} was low for one enantiopure solution, slightly higher for the racemic mixture, and well above the measurement range for the other enantiopure solution. The experimental IC_{50} values for the enantiopure solutions is shown in Table 2.1, with the values of >1000 implying that the concentration was outside measurement range.

It is important to note that, in this thesis, we are not actually interested in a discrepancy in drug effectiveness in the human cell; we are studying a difference in empirical IC_{50} measured by an enzyme assay. Even in this environment the affinity of a ligand for its receptor does not, *per se*, define the effectiveness and duration of biological action. Rather, it is the lifetime of the binary receptor-ligand complex that in large part dictates the effect. However, if we assume the measured system is in a rapid equilibrium steady-state, then we can relate the residence time to binding affinity and to IC_{50}[20]. This can be done by using the Cheng-Prusoff equation[21] for binding free energy (ΔG) of an inhibitor:

$$\Delta G_{bind} = RT \log\left(\frac{IC_{50}}{1 + [S]/K_m}\right) \tag{1}$$

This equation holds under the following conditions, according to Cheng and Prusoff: (1) the reaction in the absence of the inhibitor follows a simple Michaelis-Menten equation; (2) the rate of the reaction depends on the amount of the enzyme-substrate complex; (3) a rapid equilibrium steady state method is used; and (4) only reversible inhibitors are discussed. Our inhibitors are reversible, we are already assuming a rapid equilibrium steady state, and the natural phosphorylation reaction of erlotinib has been established by the enzymatic assay to follow the Michealis-Menten for the purpose of the assay, so the equation should hold.

By subtracting the binding free energy of the active stereoisomer from that of its corresponding enantiomer, we obtain an equation for the difference in binding free energy ($\Delta\Delta G$) between two structurally similar inhibitors:

$$\Delta\Delta G = \Delta G^R_{bind} - \Delta G^S_{bind} = RT \log\left(\frac{IC_{50}(\text{most active})}{IC_{50}(\text{least active})}\right) \tag{2}$$

This result is also reported in Table 2.1 and tells us the size of the energy difference we should observe in the simulation.

Table 2.1: Experimental IC_{50} values and computed difference in relative binding Gibbs free energy.

Compound	IC_{50} (S) [nM]	IC_{50} (R) [nM]	$\Delta\Delta G$ [kcal/mol]
1a[57]	>1000	35	-2.1 >
1b[57]	3	>1000	-3.6 >
1c[14]	1.5	629	-3.7
2a[14]	>1000	5.3	-3.2 >
2b[11]	77	4.7	-1.7
3a[14]	>1000	38	-2.0 >
3b[14]	36	>1000	-2.1 >

Ideally, if we were to calculate these two binding free energies separately in an accurate manner based on simulation and their difference agrees with the values in Table 2.1 to a statistically significant manner, we would have come a great way towards explaining the empirical difference. We would be able to compare the contributions to the binding free energy and look for notable discrepancies, for example if one particular inhibitor-residue interaction has a significantly larger energy contribution to one binding free energy than the other; or perhaps, we might find that the contribution from, say, the protein flexibility is negligible.

Unfortunately, binding energy differences of 2 to 3 kcal/mol is just on the edge of the accuracy of conventional molecular mechanics methods (FEP binding affinity calculations using the OPLS3e force field report a root mean square error of about 1 kcal/mol[58]) - so even if the simulation is rather accurate in terms of predicted equilibrium state ensembles, the uncertainty in the subsequent energy terms may be too great for quantitive analysis, in particular since our choice of simulation setup is only suited for approximate end-point calculations such as MM/GB-SA. Therefore, in addition to calculating the binding energy by MM/GB-SA, we also devote some time to exploring other ways of obtaining useful data from the simulation trajectories.

2.2.1 MM/GB-SA binding free energies

During post-processing of our Molecular Dynamics trajectories, MM/GB-SA binding free energies will calculated by using the thermal_mmgbsa.py script provided by Schrodinger with some modifications. Normally, the MM/GB-SA protocol removes all explicit water molecules and replaces them by an implicit solvent surface (the Generalized Born Surface Area part of the protocol). This approach - simulating the system with explicit solvent and calculating binding free energy with an implicit solvent - is common in literature, but has been shown to be somewhat erroneous[59] in terms of correlation between predicted binding free energy and experimental binding free energy. Instead, we opt to make some adjustments by choosing to keep structural water molecules between protein and its binding inhibitors, that is, water molecules which mediate a polar interaction between the protein and the inhibitors; an approach that has

shown to have some merit in terms of improving correlation between experimental and calculated binding free energy.[60] due to water bridges being largely neglected by the Generalized Born Surface Area.

The script then calculates the energy of three systems: the protein-ligand complex, the protein alone, and the ligand alone. This is known as a single-trajectory approach, compared to the three trajectory approach where one simulates and calculates energies for the complex, the protein, and the ligand separately [31]. We have a reason for not performing the extra free-molecule simulations: beyond the increased computational cost, we are not, in fact, interested in the free energy of binding of each inhibitor, but the pair-wise difference between the two enantiomers of each inhibitor. Therefore we opt not to relax the separate protein and ligand systems, since these would have the same energy for both enantiomers anyway; the protein because its non-liganded geometry is always the same, and the ligand because the ΔG of changing chirality in a symmetric environment is strictly zero. As a consequence, the MM/GB-SA binding free energy, ΔG_{bind}, calculated by the script will not include energy terms arising from conformational changes in either the ligand or the protein, but only the difference in energy due to interactions between the ligand and the protein-water structure:

$$\Delta G_{bind} = G_{complex+waterinterface} - G_{protein+waterinterface} - G_{ligand} \tag{3}$$

where the Gibbs energy G of each structure is the sum of a gas-phase energy term (E_{MM}), a solvation free energy (G_{solv}) and an entropy term (TS) calculated via the Prime-MM/GB-SA driver:

$$G = E_{MM} + G_{solv} - TS \tag{4}$$

We can then exploit the fact that, since the relaxed protein and the relaxed ligand should in theory have the same energy, the only energy term that is actually different between the enantiomers is the energy of their respective complexes, $G_{complex+waterinterface}$. Thus we can express the difference in binding free energy between the R and S enantiomers as

$$\Delta\Delta G_{R \hookleftarrow S} = G^{R}_{complex+waterinterface} - G^{S}_{complex+waterinterface} \tag{5}$$

2.2.2 Protein-Ligand Simulation Interaction Analysis

For analysis of the ligand, we will use the Protein-Ligand Simulation Interaction Analysis (SIA), a Desmond toolchain meant to analyse trajectories of protein-ligand simulations, producing a wide variety of useful statistics about the simulation, such as root mean square deviation from reference structure, mean fluctuation of the ligand and protein atoms, the protein's secondary structure, and timeline plots torsional angles of the ligand.

One of the more important results from the SIA tool is a protein-ligand contact analysis of the trajectory. This generates a time series where, at each sampled timestep in the trajectory, the tool calculates which

residues interact with the ligand, and what kind of interaction this is. One residue may have several - possibly different - interactions with the ligand at the same type at the same time, for example Arginine forming separate hydrogen bonds with each of its nitrogens. The interaction types and their conditions are described in the Desmond [61] manual and are reproduced here:

- Hydrogen bonds are defined by distances and angles of the D−H ⋯ A−X atom arrangement: a D−A distance less than 2.5 Å, a D−H−A angle greater than 120°, and a H−A−X angle greater than 90°.

- Pi-pi stacking occurs between two aromatic groups stacked face-to-face or face-to-edge.

- Pi-cation bonds occur between aromatic and charged groups within 4.5Å

- General hydrophobic interactions occur when hydrophobic side chain is within 3.6 Å of a ligand aromatic or aliphatic carbon.

- Ionic interactions occur between oppositely charged atoms on the ligand and the protein that are within 3.7Å

- Water bridges involve hydrogen bonding via a water bridge molecule, broken down into protein donor and protein acceptor. The geometric criteria are a D−A distance less than 2.7 Å, a D−H−A angle greater than 110°, and a H−A−X angle greater than 80°.

Comparing how frequently a given residue interacts with the inhibitors may give us insight into whether some residues have a stereoselective effect, where they prefer interacting with one residue over the other. A related method is Ligand Interaction Energy (LIE) [62], which uses a slightly more rigorous approach of calculating the difference in Van der Waahls and electrostatic energy for the ligand in solution and the ligand in the binding pocket in order to estimate the binding free energy. We do not use this method in this thesis, but the ligand interaction frequency described above is motivated by the same underlying physical motivation and can be viewed as a cruder approximation of LIE.

3 Method

3.1 Practical considerations of Molecular Dynamics

The basic idea behind an MD simulation is straightforward. Given the positions of all the atoms in a molecular system, one can calculate the force exerted on each atom by all the other atoms. One can thus use Newton's laws of motion to predict the spatial position of each atom as a function of time. In particular, one steps through time, repeatedly calculating the forces on each atom and then using those forces to update the position and velocity of each atom. The resulting trajectory is, in essence, a three-dimensional movie that describes the atomic-level configuration of the system at every point during the simulated time interval.

The actual implementation, design and performance of MD simulations requires some practical considerations: First, which computing hardware should we use? In our case this is rather simple, because we have access to the Idun supercomputer courtesy of NTNU, which contains both high-end GPUs and massively parallel CPUs. We'll primarily be using GPUs due to our choice in the second question: Which software to use? Common choices include GROMACS,[63] NAMD,[64] AMBER,[65] CHARMM,[66] Desmond,[67] and OpenMM[68]; Because our previous work has primarily been performed within the (proprietary) molecular modelling suite designed by Schrödinger Inc.[23] we decided to use their high-performance MD implementation, Desmond - which runs on a single GPU per simulation, hence why we will be using GPU hardware. Third, which force field should we use? In the field of biochemical modelling, three force fields tend to see widespread use: CHARMM.[69] AMBER.[70] and OPLS.[71] and their successors.[31] In our case, Desmond provides access to the extensively optimized OPLS3e[58] force field, which has one of the broadest ligand parameter data sets of these force fields, though third-party evaluations of this force field are rare due to its proprietary nature. In particular, the parameter set for OPLS3e includes data for pyrrolopyrimidines, which makes it particularly well suited for our inhibitors.[72]

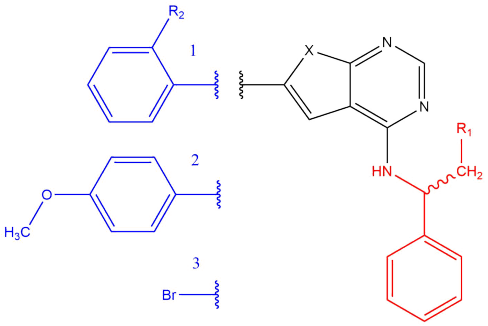
Once these three choices were made, in particular choosing Desmond as our Molecular Dynamics software, our remaining choices are more restricted based on what is implemented in Desmond. For the time integrator algorithm - the part of the simulation that steps the system forward in time once forces have been calculated - we use the only available: the r-RESPA (reversible REference System Propagation Algorithm) integrator, which splits the forces into short and long range calculations which can be updated on different time scales, and uses the Liouville formulation of mechanics rather than the Newtonian formulation.[73]

For our protein-ligand complexes, we are interested in simulating them at constant pressure and temperature (the NPT ensemble) rather than the default constant energy and volume (NVE ensemble), since biological systems at the protein scale operate more like the former than the latter; even if the cell tends to have a (near) constant volume, we are simulating only a fraction of the cell volume, where pressure

and temperature is more important. Algorithms for controlling pressure and temperature are termed barostats and thermostats, respectively.[74] In Desmond, our choices of thermostat are limited (in the GUI, at least) to the Nose-Hoover chain thermostat[75,76] and the Dissipative Particle Dynamics (DPD) thermostat, while the choice of barostats are limited to the Martyna-Tobias-Klein (MTK) extension of Nose-Hoover chains[77] or the Langevin barostat. The documentation recommends to only use DPD for coarse-grained molecular dynamics simulations, while we are interested in all-atom simulations, so we choose the Nose-Hoover chain thermostat, and it is then natural to use the MTK barostat since it extends Nose-Hoover. We do note that there is also the possibility of employing the Berendsen thermo/barostat and the Langevin thermo/barostat by modifying command line input to Desmond; however, outside of their use in the default relaxation schemes, we decided to refrain from using these as they are both less accurate than Nose-Hoover (in terms of physical basis).[78]

3.2 Model system preparation

Initially we performed experiments using only one inhibitor pair, the R and S enantiomers of 2-[4-(1-phenylethylamino)thieno[2,3-d]pyrimidine-6-yl]phenol (**1a**), based on the reasoning that simulating additional compounds was too time consuming when we weren't sure the method would yield interpretable results, since this had happened in our previous QM/MM study. We expanded the experiment to include other inhibitors for which we have experimental IC_{50} values (shown in 3.1) once results from Molecular Dynamics simulations showed a some interesting differences between the R and S enantiomer of compound 1a. For this reason, there were some minor differences in the set up of **1a** and the other inhibitors.



1a: X=S, R_1 = H, R_2 = OH
1b: X=S, R_1 = H, R_2 = OH
1c: X=S, R_1 = OH, R_2 = OCH_3
2a: X=O, R_1 = H
2b: X=NH, R_1 = H
3a: X=S, R_1 = H
3b: X=S, R_1 = OH

Figure 3.1: Model compounds used in this study. Fragment A (aniline) in red. Fragment B (Phenyl or bromine) in blue. Core scaffold in black.

Of the seven inhibitor pairs we simulated in this thesis, compounds **1a-c** are thienopyrimidines with an asymmetric polar *ortho*-substitute on fragment B (in blue). Compounds **2a-b** use a different heteroatom at X, allowing for comparison between furo- thieno- and pyrrolopyrimidines. Finally, compounds **3a-b** use the truncated Bromine structure investigated by Bugge et al.[14]. This naming convention was inherited from the empirical studies performed previously.[10,24] In this study we found it especially important to draw attention to the difference between compounds which have a methanol substitute at R_3 (**1b,1c** and **3b**) and the ones which have a methyl group (**1a, 2a, 2b** and **3a**); this is because the binding modes found by MD

showed distinct behaviour for the methanol compounds, since the hydroxy group can form hydrogen bonds which the methyl cannot.

For the preparation of the protein-ligand complex system, we took advantage of our earlier work [22,24] and reused the crystal structures prepared then; the details of this preparation is summarised here. The protein-ligand complex's geometry was prepared by starting with the co-crystallized EGFR-AEE788 complex (PDB code 2J6M) as elucidated by Yun et al.. This structure represents the intracellullar kinase domain of an active but inhibited EGFR monomer, which includes residues 696 to 1020.[2] All solvent molecules were removed and missing sequences, protonation of amino acids, and H-bond assignment was added via Maestro's Protein Preparation Wizard. In particular, the PDB crystal structure was missing the entire residue sequence from 984 through 1004, necessitating wholesale addition of the chain using Prime.

The AEE788 inhibitor was then replaced by our own inhibitors through docking with Glide, which uses a heuristic dock-and-score method in a rigid receptor; The waters were then removed completely and the complex energy minimized at absolute zero in QSite, a hybrid quantum mechanics/molecular mechanics module, using the OPLS_2005 force field for the protein and the quantum mechanical DFT-B3LYP method for the ligand, using a 6-31G+* basis set. It was these free gas complexes we used as the starting structures for our simulations in this thesis.

In order to perform molecular dynamics simulations in a solvated environemnt, we constructed an orthorombic simulation box around the protein-ligand complex using with a distance of 10 angstrom to the existing geometry in all three direction, and then populated the remaining volume with water molecules, using a TIP4P solvent model. We neutralized the system by adding counter-ions (8 Na^+ ions). This was all done by the Desmond System Builder module. An example of the resulting simulation box is shown in Figure 3.2 for compound *S*-**1c**.

3.3 Simulation details

All Molecular Dynamics simulations were performed using the Desmond molecular dynamics software [61] and the OPLS3e force field [58] using periodic boundary conditions.

The prepared model systems were relaxed using Desmond's default five-step NPT relaxation scheme; First, the system is simulated in a Brownian Dynamics NVT ensemble at 10K for 100 ps with restraints on heavy atoms, then in the same conditions in a Langevin NVT ensemble for 12 ps, followed by a Langevin NPT ensemble at 1 atm pressure for 12 ps. The next step increases temperature to 300 K and applies a Langevin NPT ensemble for 12 ps, and finally the restraints on the heavy atoms are lifted and the system is relaxed at 300K, 1 atm in a Langevin NPT ensemble for 24 ps.

[2] This thesis employs the PDB numbering scheme, which does not include the 24-residue membrane targeting signaling sequence.

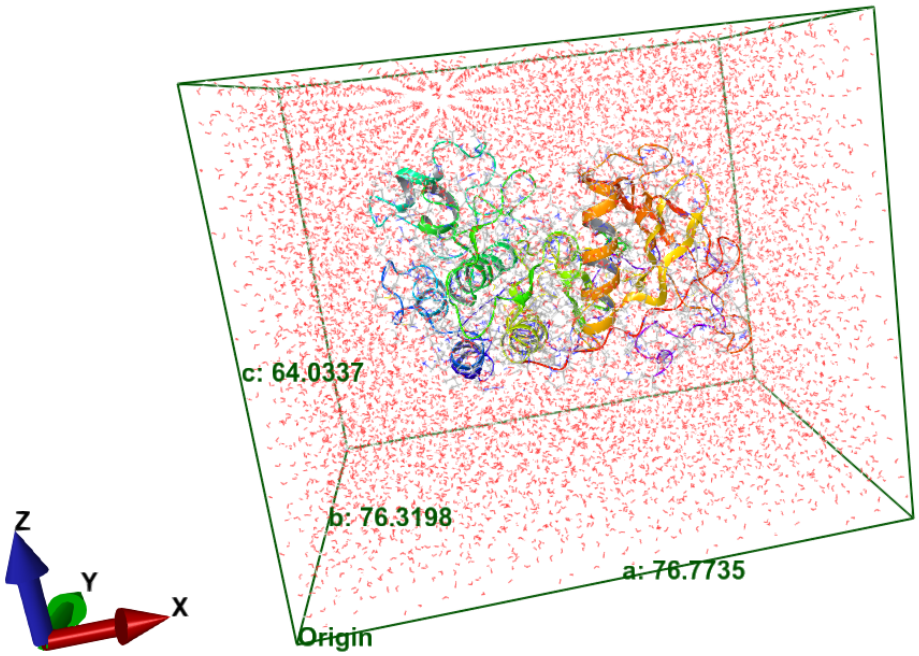


Figure 3.2: The simulation system for the Molecular Dynamics simulations was prepared using Desmond System Builder. Displayed here is the resulting simulation box for compound *S*-**1c**

After relaxation, the model systems were simulated for 1000 ns in an NPT ensemble ($P = 1.0135\,\text{bar}, T = 310\,\text{K}$) using a Nose-Hoover chain thermostat coupled to a Martina-Tobias-Klein barostat.[77] The simulations used the reversible reference system propagation algorithm (r-RESPA)[73] time step integrator with a far-time step of 6 femtoseconds, a near-time step of 2 femtoseconds, and the near-far intersection at 9 angstrom. Macroscopic properties were sampled every 1.2 picoseconds, while the trajectory was sampled every 500 picoseconds (ie. every 100 000 time steps), resulting in a trajector consisting of 2000 snapshots. The exception was our first two simulations, those of *S*-**1a** and *R*-**1a**, due to these being our trial systems; these were run for approximately 1600 ns and only sampled snapshots every nanosecond.

3.4 Data post-processing and analysis

Once the simulations had been completed, we were left with 14 very rich sets of data - macroscopic thermodynamical properties like temperature, potential energy, and pressure had been sampled every picosecond, while the system trajectory (i.e. the position of each atom) of each production run consists of 2000 (or about 1600 in the case of **1a**) distinct snapshots taken at regular intervals[3]. To simplify the task of extracting meaningful data from these trajectories, we relied heavily on automated analytical tools in Desmond, as well as visual inspection of the trajectories in Maestro.[23]

MM/GB-SA energies were calculated for every tenth trajectory frame (every 5 ns) using the thermal_-

[3] In theory we could have sampled velocity as well, but the resulting data files proved problematic to work with due to sheer size

mmgbsa.py script as described in Theoretical background, including the modification that we include interstitial water molecules. Specifically, for each MD snapshot, all water molecules whose oxygen atoms were within 3.5 Å of a protein heavy atom and a ligand heavy atom - i.e., were close to both protein and ligand - were kept by the script. Usually one to seven water molecules for each snapshot met that criterion, and these water molecules were considered as a part of the protein structure for the purposes of the MM/GBSA calculations.

The trajectories were analysed using the Simulation Interaction Analysis tool and both raw data and accompanying reports were exported. In terms of input, the tool requires a simulation trajectory, the atom specification of the protein, the atom specification of the ligand, and a reference structure to which all other snapshots are compared; we use the first frame in each trajectory as the reference. In the Results section of the present thesis, we include only those plots that are relevant for our discussion; some of these plots were made by ourselves by importing the output data into R, with data processing performed using tidyverse and plotting done with ggplot2.[79]

3.4.1 Representative binding pose geometries

A considerable amount of our analysis will be based on visual inspection in the Maestro GUI of the generated snapshots from each trajectory. Presenting such insights in a static research paper is a considerable challenge, as the medium precludes inserting three dimensional figures as well as animations. Additionally, relying on the human eye to discern whether structures are significantly different is an error-prone approach. In order to identify different binding modes and how the ligand conformation changes throughout the simulation, we combine plots of root mean square deviation (RMSD) of the ligand from a reference state with a clustering method based on this RMSD using the the trj_cluster.py script provided by Schrodinger, which uses affinity propagation to cluster the frames of a trajectory based on the RMSD of an input atom specification (in our case, the ligand atoms)

The timeline plot of RMSD allows us to identify if the ligand is oscillating about a thermal average structure or is changing conformation significantly; this plot is calculated as part of the SIA tool chain. The clustering method, on the other hand, allows us to identify which parts of trajectory have the same conformation and find a representative binding mode we can use in this report. In brief, each frame is viewed as a possible candidate "exemplar", i.e. a distinct binding mode, and the likelihood of that is calculated by subtracting how well it represents neighbouring frames from how well itself is represented by another frame. A more detailed description of this algorithm is beyond the scope of this thesis - details are described elsewhere.[80] The end result is a number of clusters, each having a representative "exemplar" frame which best represents the average binding mode of the ligand within the cluster.

In our case, we found that the number of clusters varied between 50 and 100 depending on the trajectory. Since 50 images per trajectory for 14 trajectories is still too large to reasonably fit in this report, we chose

the ten densest clusters, and visually inspected them while consulting the RMSD plots to identify if any of these exemplar clusters could further be represented by each other. Once this had been done, we were finally left with one to three representative frames that show the binding mode(s) generally adopted by each inhibitor throughout the simulation.

4 Results and Discussion

We performed long-timescale (1 μs) unbiased MD simulations of our inhibitors - one trajectory per enantiomer, across seven pairs, for a total of fourteen simulations. Each simulation consist of the inhibitor situated in the active seat of the intracellular TK domain of EGFR, with the complex surrounded by a box of about 10 000 T4P water molecules with periodic boundary conditions. Here we present the results from these simulations in a systematic manner, describing their stability, evolution, and the conformational state of the ligand, as well as the average receptor-ligand interactions and the computed MM/GB-SA energies. At the same time we discuss the implicaitons of these data and use them to build toward a plausible answer to our thesis problem. We then discuss this answer in a broader context, how our method performed, how well our assumptions hold up, and sources of error.

4.1 MM/GB-SA Binding energy

Figure 4.1 shows the calculated Gibbs free binding energy ΔG_{bind} as calculated by the MM/GB-SA approach described earlier, for every tenth frame of each simulation. Figure 4.2 shows a component of this binding free energy, the Gibbs Energy of the protein-ligand complex. In theory, the time-averaged difference in the latter should be equal to the logarithm of the ratio of IC_{50} values; unfortunately, from these plots, it becomes abundantly clear that the standard deviation of this time-averaged quantity is *much* greater than the calculated difference in $\Delta\Delta G$ from empirical data shown in Table 2.1. This is true even before we consider uncertainties such as incomplete simulation (violation of the ergodic hypothesis) and sensitivity to initial conditions. It is thus clear to us that the calculation of the relative free energy of binding via MM/GB-SA applied to our simulations is unable to clearly differentiate between high-potency and low-potency enantiomers. Fortunately, the purpose of our simulation was not, this time, to calculate the interaction energy exactly, but to generate a set of equilibrium conformations and investigate these for interesting differences between each enantiomer, which we shall now do.

4.2 Stability of the protein

Figures 4.3 and 4.4 show the root mean square deviation (in Angstrom) of the protein and the ligand from their initial (t=0) configuration, as calculated by the Simulation Interaction Analysis tool. The Protein RMSD is measured with regards to the frame of reference of its backbone peptide chain. It intentionally does not capture the deviations of the side chains, so that the plot captures larger conformational changes in the backbone rather than the frequent fluctuations of the side chains. In contrast, the ligand RMSD is measured with respect to the ligand heavy atoms, but realigned to the center of mass of the ligand - which means it shows the internal conformational fluctuations of the ligand, but not its orientation or distance to the binding pocket. The reader can be assured that we did not observe any unbinding events during the simulation.

The sharp increase in RMSD beyond the initial time step common to all production is due to a reconfig-

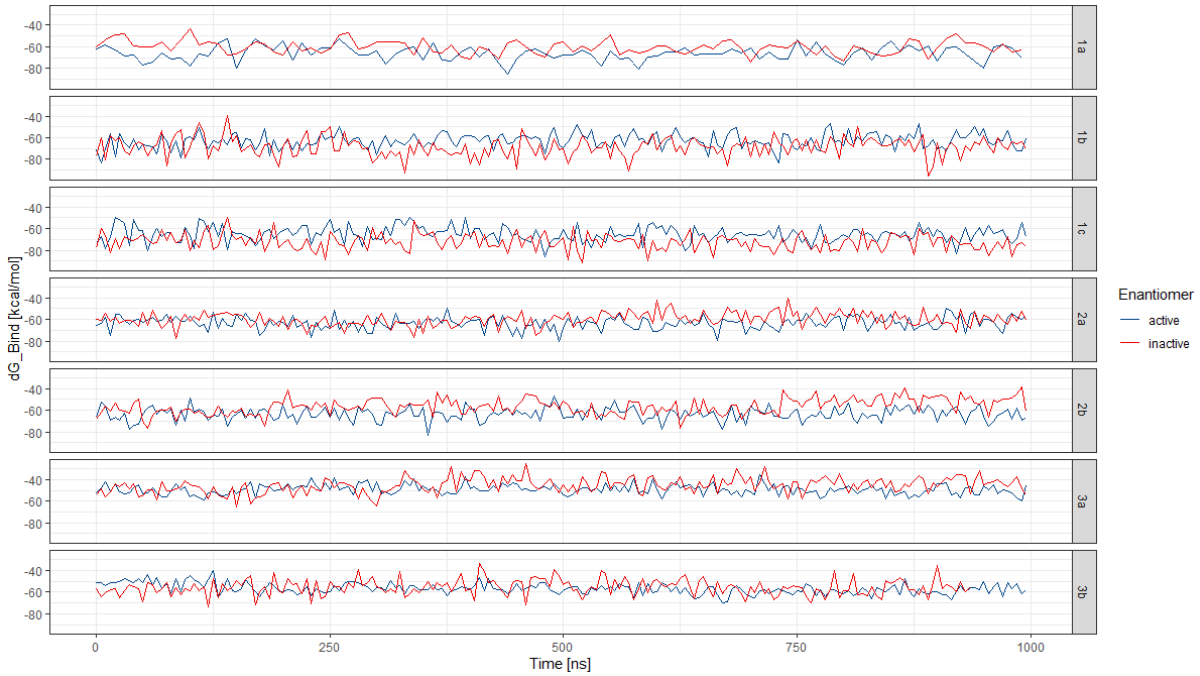


Figure 4.1: Timeline plot of the binding free energy of each inhibitors calculated for every tenth frame using the MM/GB-SA method.

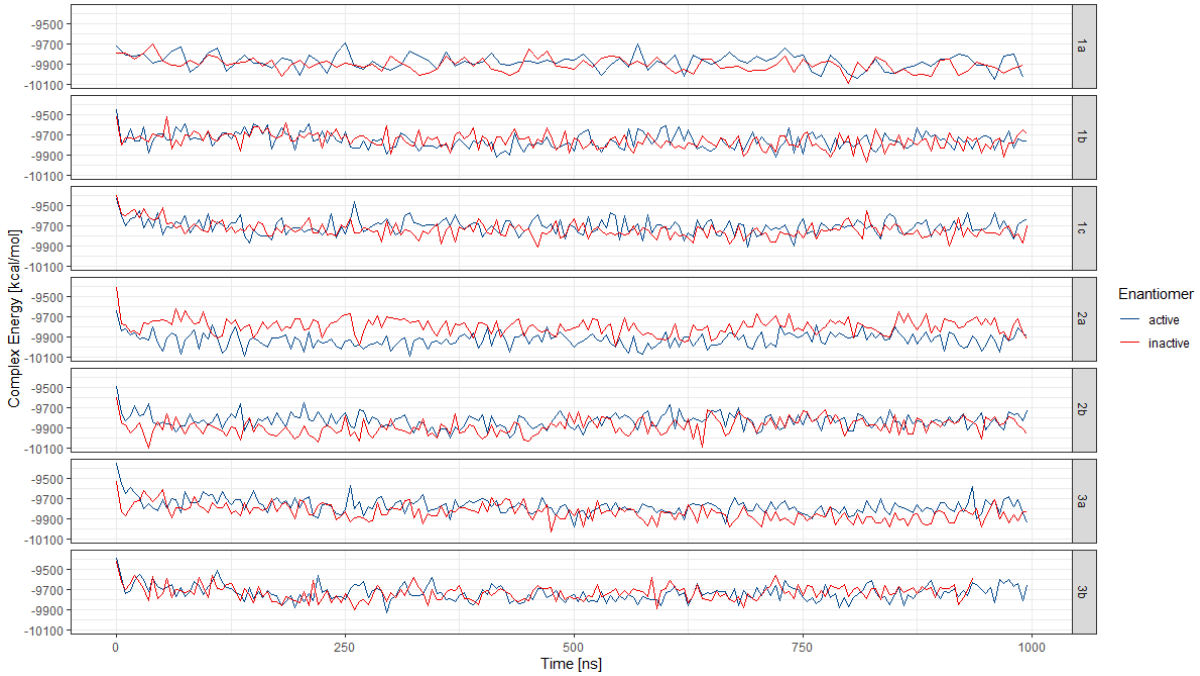
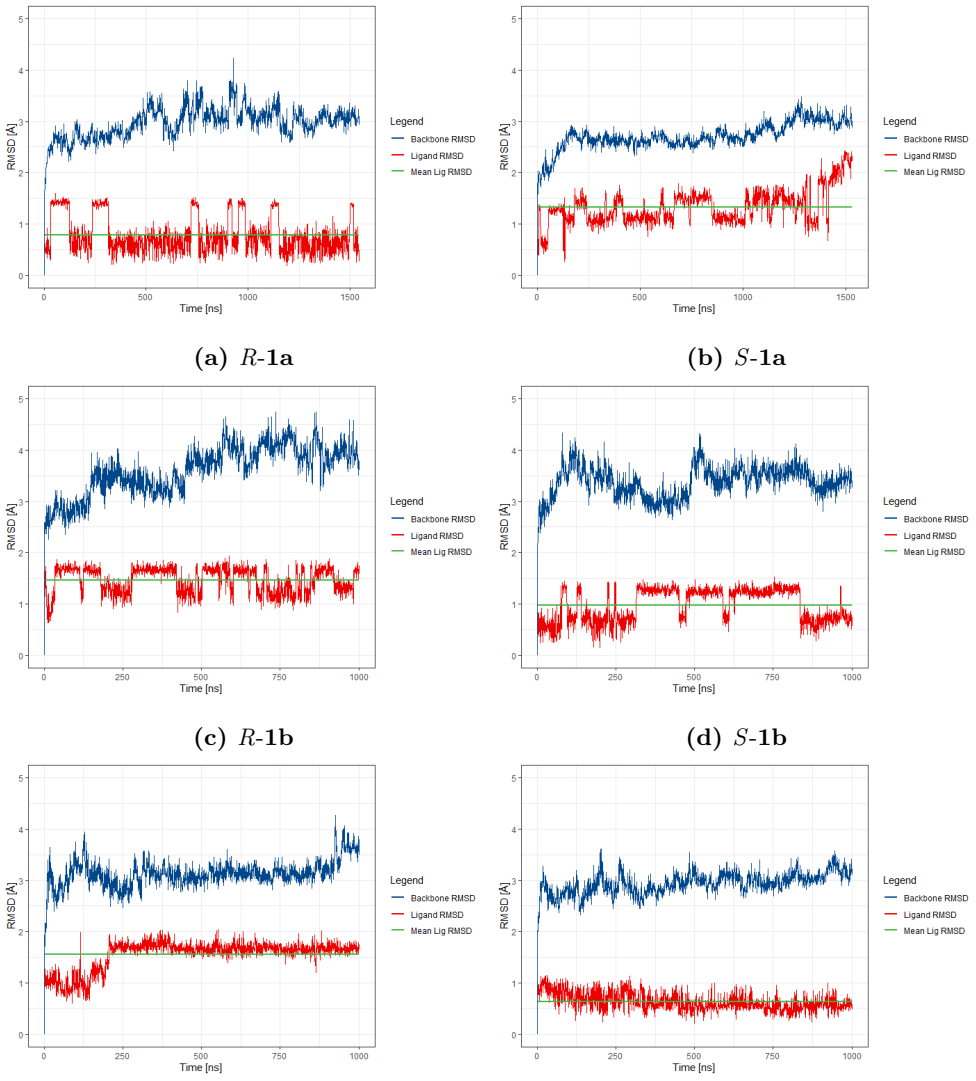


Figure 4.2: Timeline plot of the energy of the protein-ligand complex calculated by MM/GB-SA. In theory, the difference between the average of this energy for each enantiomer is proportional to the difference in empirical IC_{50} values, but noise makes this difference rather diffuse. Compound **1a** was sampled for longer but at a lower frequency. It is clear that any differences in binding energy based on an average of these will have too great variation to be statistically significant.

uration of the protein complex due to an unfinished relaxation procedure - the sum of hundreds of small adjustments across a macro-molecule consisting of 300 residues in response to the comparatively sudden presence of dynamic solvent as well as the change in force field. Beyond this initial reconfiguration, however, fluctuations in the protein RMSD on the order of 1 to 4 Angstrom are perfectly normal for a protein of our size.

Visually inspecting the protein backbone during the trajectories, we find that most parts of the protein tends to oscillate around a stable equilibrium, in agreement with the RMSD plots. There is one exception to this; we find that the loop added by Prime during model system preparation (res 984 to 1004) shows considerable fluctuations both within each simulation and across different simulation runs. Some runs in particular produced a stable helix conformation for this loop (see Figure 4.5 for an example). These fluctuations are not unexpected - the very fact that the XRD crystal structure lacked this loop implies that it fluctuates too much for the XRD to get a good resolution of it. Additionally, while Prime is a decent tool for making adding short sequences and estimating their conformation, its accuracy decreases proportionally when it has to complete longer sequences - in particular, since the resulting loop was not already in a helix or sheet conformation, it is natural that the loop tries to adopt one of these more stable conformations over the course of the simulation. The fact that it adopts a helix conformation in our MD runs does not necessarily imply that it does so in reality; it merely shows that the OPLS3e force field makes it favor the α-helix conformation over the β-sheet conformation. The real protein folding is a much harder problem to solve, and one we'd hoped to avoid, ideally. In terms of our simulations it causes some complications, because the Prime loop contains some residues which, for the right folding, turn out to interact rather frequently with some of the inhibitors - as seen in Figure 4.13 later in this section.

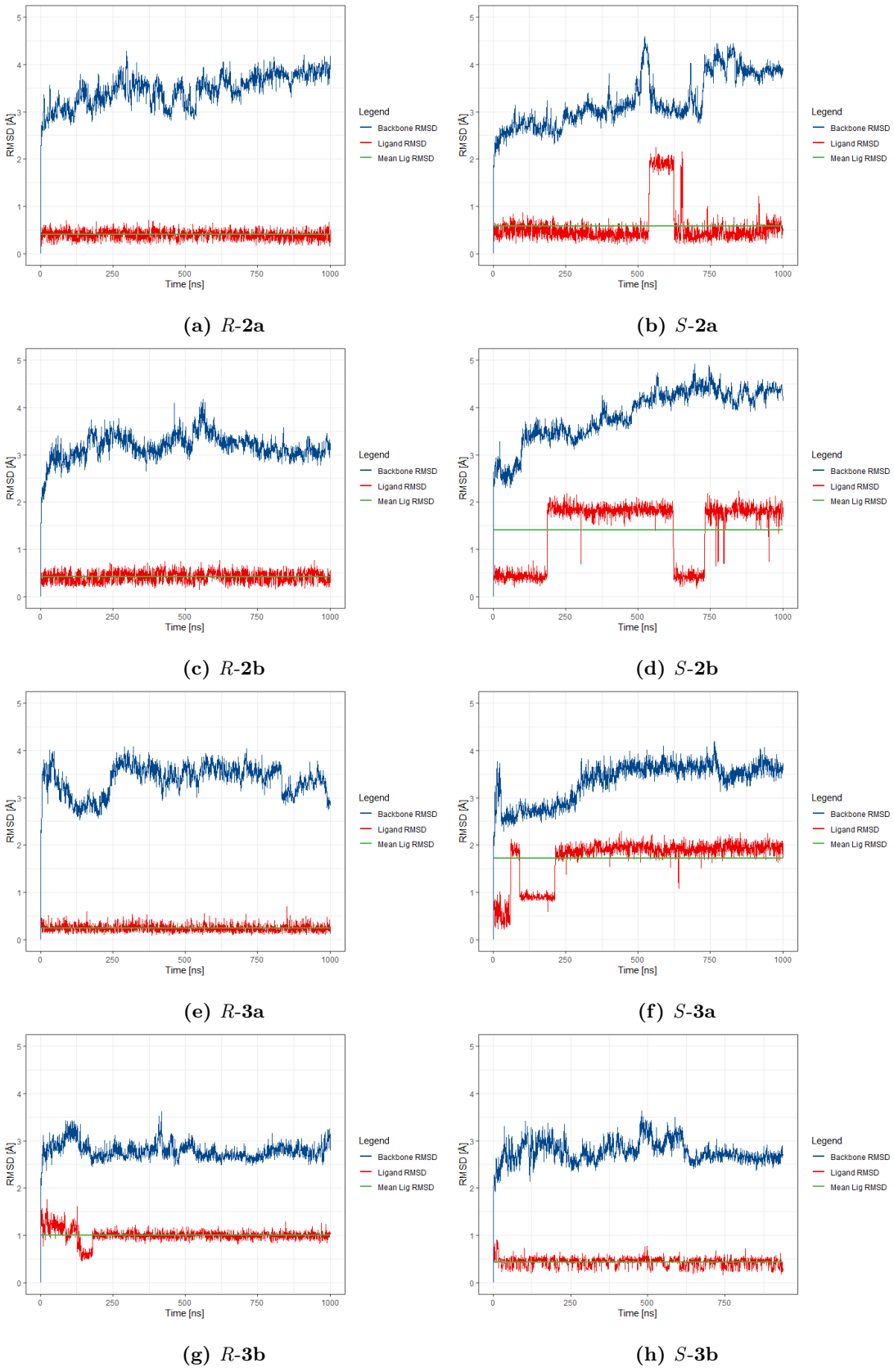


(a) *R*-**1a** **(b)** *S*-**1a**

(c) *R*-**1b** **(d)** *S*-**1b**

(e) *R*-**1c** **(f)** *S*-**1c**

Figure 4.3: Root mean square deviation (RMSD) of protein atoms (blue) and ligand atoms (red) of compounds **1a-c** relative to starting geometry for the protein backbone and ligand, as well as the mean (time-averaged) RMSD of the ligand.



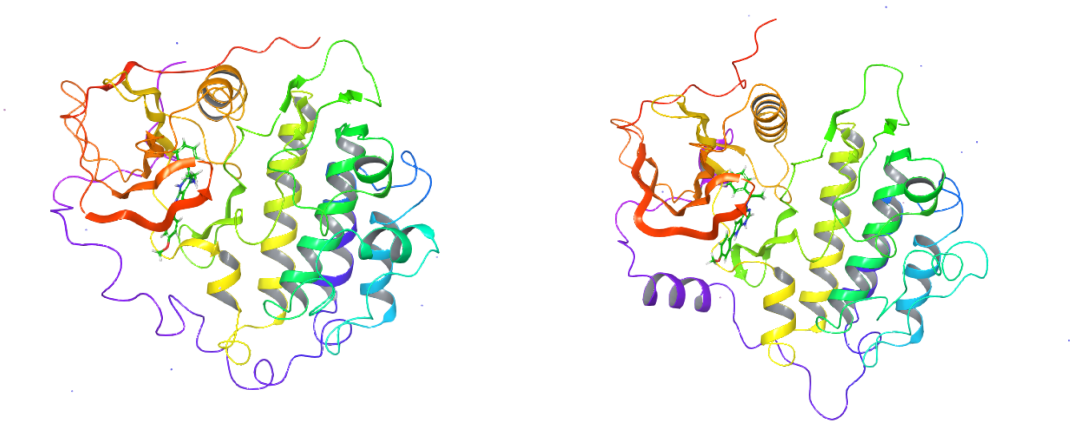
(a) *R*-**2a** **(b)** *S*-**2a**

(c) *R*-**2b** **(d)** *S*-**2b**

(e) *R*-**3a** **(f)** *S*-**3a**

(g) *R*-**3b** **(h)** *S*-**3b**

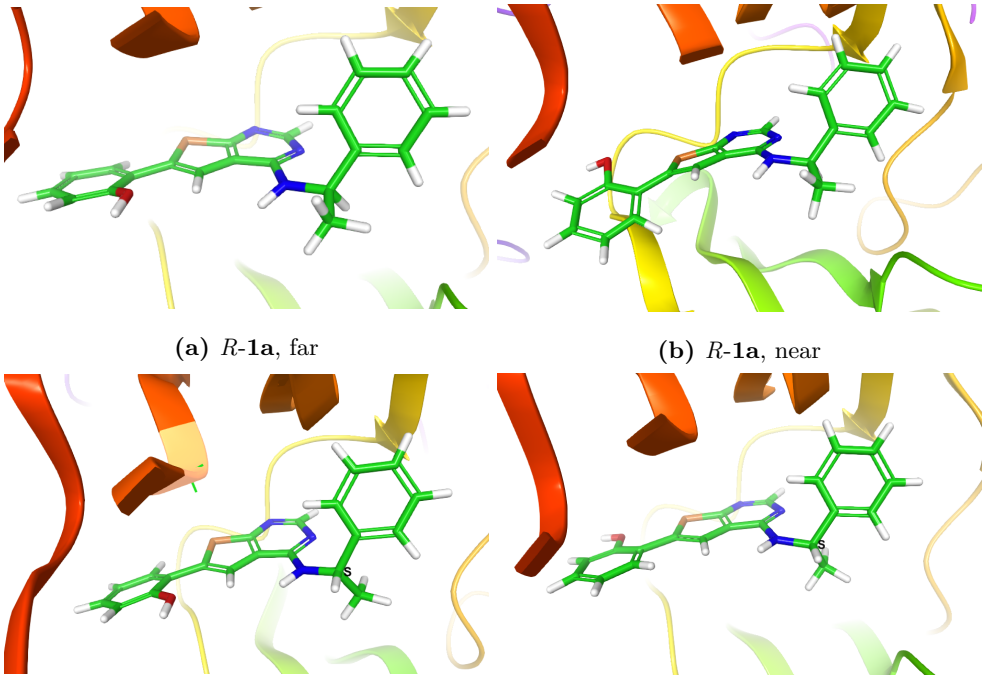
Figure 4.4: Root mean square deviation of compounds **2a-b** and **3a-b** relative to starting geometry for the protein backbone and ligand, as well as the mean (time-averaged) RMSD of the ligand.



(a) Initial state

(b) Purple strand folded to α-helix

Figure 4.5: The protein loop added by Prime (residues 990 to 1021) tended to fold in different ways across simulations. Depicted here is the start and end frames of the simulation of *S*-**3a**, where we saw it fold into an α-helix.



(a) *R*-**1a**, far **(b)** *R*-**1a**, near

(c) *S*-**1a**, far **(d)** *S*-**1a**, near

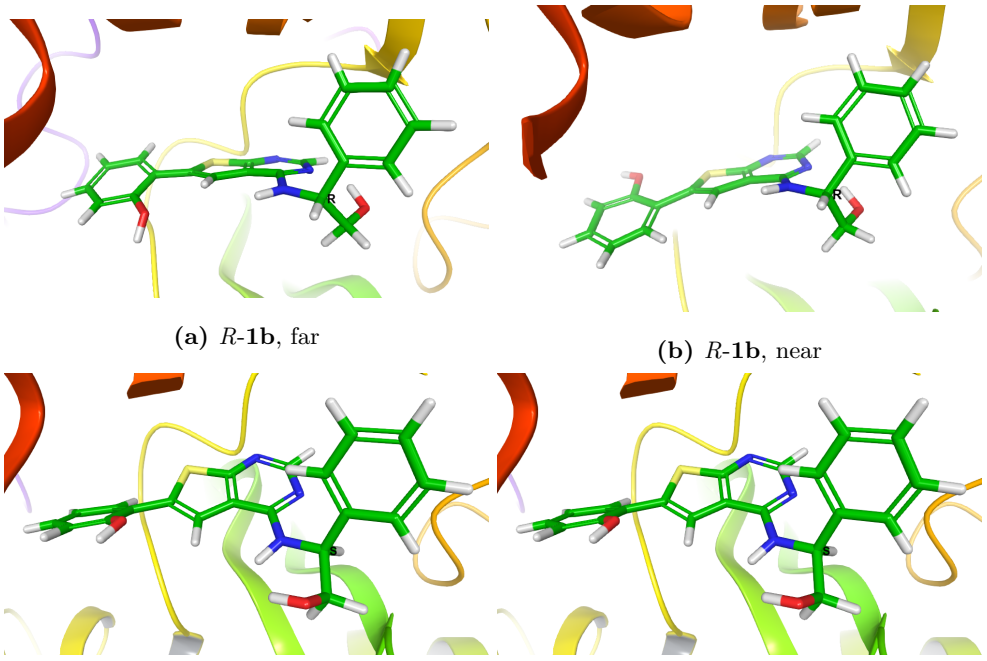
Figure 4.6: Representative conformations for compound **1a**. Both enantiomers have two significantly different conformations each in terms of RMSD, stemming from the rotation of the tail phenol fragment, and whether the hydroxy group lies near the heteroatom or on the far side of it. Otherwise the conformation is similar to most other inhibitors, with the phenylamine pointing "up" relative to the scaffold. Note the orientation of the chiral methyl group.

4.3 Average inhibitor conformation - RMSD and cluster analysis

Images of molecular geometry follow the conventional CPK colouring scheme with blue nitrogen, red oxygen, white hydrogen, and grey carbons. The exception are the ligand carbon atoms, which are coloured green to distinguish the ligand and protein.

From the plot of Ligand RMSD in figures 4.3 and 4.4 it is clear that the ligands tend to fluctuate around an average structure, with occasional large changes in RMSD indicating that they adopt a different conformation. Since the RMSD is measured entirely with respect to the ligand (i.e. it does not capture reorientation with respect to the protein), and since the ligand only forms noncovalent bonds to the receptor and the solvent, we can investigate what these binding modes look like without needing to keep track of the surrounding water molecules or receptor residues. As explained in Methods, presenting the entire trajectory in this report is infeasible; we therefore used a cluster analysis script to find good representative snapshots of the trajectory that can explain what the various inhibitor conformations look like. For the sake of simplicity, we first present the ligand conformations themselves without considering the binding site residues or the solvent.

Inspecting the RMSD plot of *R*-**1a**, *S*-**1a**, *R*-**1b**, and *S*-**1b**, it is clear they all oscillate between two



(a) *R*-**1b**, far

(b) *R*-**1b**, near

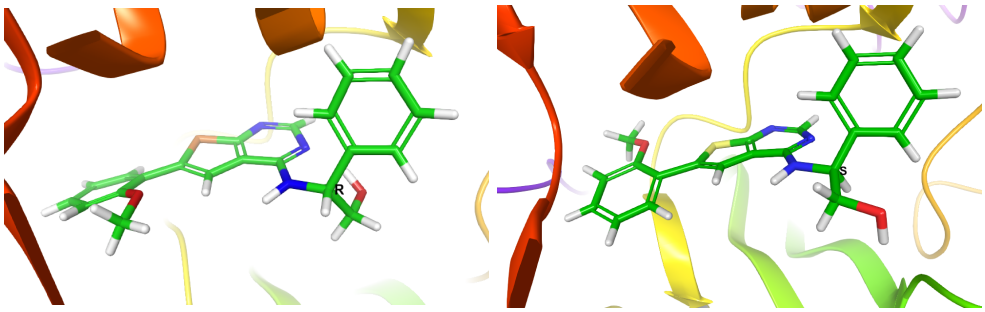
(c) *S*-**1b**, far

(d) *S*-**1b**, near

Figure 4.7: Representative conformations for compound **1b**. Like **1a**, both enantiomers have two significantly different conformations each in terms of RMSD, stemming from the rotation of the tail phenol fragment, and whether the hydroxy group lies near the heteroatom or on the far side of it. Notice how the R enantiomer forms an internal hydrogen bond between the chiral methanol group and the N3 nitrogen.

general conformations. The cluster analysis confirms what we observe visually - these two binding modes are not due to changes in the amine conformation containing the stereocenter, but are a result of the tail fragment of these inhibitors being a *ortho*-substituted phenol. In all four simulations, this substituted phenyl lies in the same plane as the double ring scaffold, but it occasionally flips 180°, showing that there are two roughly equivalent energy minima with a torsional barrier between them. The respective ligand conformations are in figures 4.6 and 4.7. It is clear that the phenol lies near-coplanar with the scaffold most of the time, and that the only difference is whether the substitute is on the near or far side of the heteroatom - the amine substitute doesn't contribute significantly to the RMSD beyond noise.

Even though **1c** also contains an asymmetrically substituted phenyl in its tail, the RMSD plots do not show the same oscillation. Instead, *S*-**1c** shows larger than normal fluctuations about its average conformation, while *R*-**1c** spends the first 200 seconds slowly converging to its average conformation. The methoxy group of *R*-**1c** is initially on the near side of the heteroatom, but eventually (after about 200 ns) flips to the far side. In contrast the methoxy group of *S*-**1c** stays on the near side of the heteroatom throughout the simulation. The representative conformation for compounds *S*-**1c** and *R*-**1c** are shown in Figure 4.8. The reason they have opposite tail angle is not necessarily due to any intrinsic effects of inhibitors themselves, but is a consequence of their different initial conformations - this is an error that



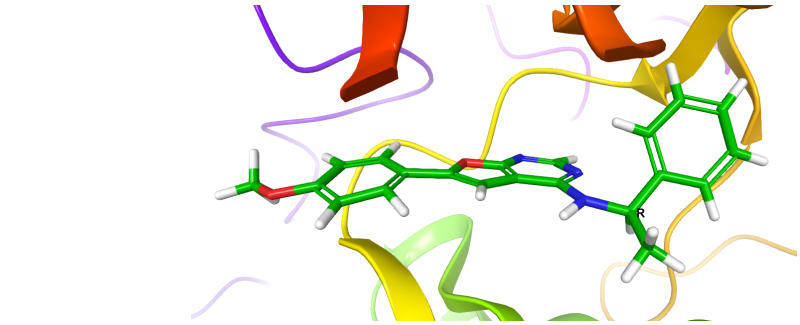
(a) *S*-**1c** **(b)** *S*-**1c**

Figure 4.8: Representative conformations for compound **1c**. Despite having an asymmetrically substituted tail like **1a** and **1b**, neither *R*-**1c** nor *S*-**1c** shows any tendencies to "flip" the tail, and so they only have one conformation each. Notice how the R enantiomer forms an internal hydrogen bond between the chiral methanol group and the N3 nitrogen.

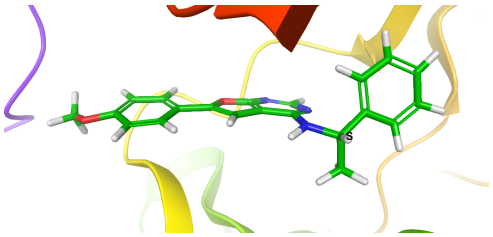
we unfortunately did not catch until after the simulations had been completed.

However, even though *S*-**1c** and *R*-**1c** start in different conformations, if they were to behave like **1a** and **1b**, they should flip back and forth between the conformations; they clearly do not. While the chemical environment of this fragment certainly plays a role in damping these rotations, the lack of oscillations in the tail can also be explained by considering the molecule as a rigid rotor - **1c**'s methoxy group is roughly twice as heavy as the hydroxy group of **1a** and **1b** while also being longer, which increases the methoxyphenyl's moment of inertia about the C6-C19 bond significantly compared to the phenol. This in turn heightens the energy barrier between near-heteroatom and far-heteroatom conformations, meaning oscillations between them are much less frequent.

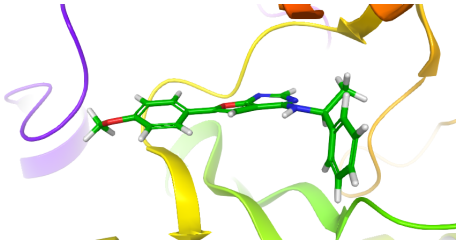
Unlike **1a-c**, compounds **2a-b** and **3a-b** do not have an asymmetric tail substitute. However, we still see conformational changes for *S*-**2a**, *S*-**2b**, and *S*-**3a** in their RMSD plots (Figure 4.4). By inspecting the conformations themselves, we see that the two conformations differ in the orientation of the phenylamine ring in relation to the scaffold. We see that the phenyl points either up or down relative to the scaffold; with the exception of these three enantiomers, the inhibitors adopt the phenyl "up" conformation throughout the entire simulation. Even for *S*-**2a** and *S*-**3a**, looking back to the RMSD plots (Figure 4.4b), we see that they only take the "down" conformation for a short period equal to roughly 10% of the simulation time (550 ns to 650 ns for *S*-**2a**, 100 ns to 200 ns for *S*-**3a**). Only S4a spends a substantial amount of time in the phenyl down conformation, about 30% of simulation time; it is also notable for changing conformation multiple times. There is certainly an element of chaos involved; the simulations are sensitive to initial conditions and the conformational changes appear to occur only once or twice per microsecond, so we cannot for certain say that these conformational changes are impossible in the other enantiomers; however, we find it very likely that there are some structural elements at play here. For now, though, we note that the phenyl up conformation is so common that we will primarily investigate that one.



(a) *R*-**2a**, phenyl up



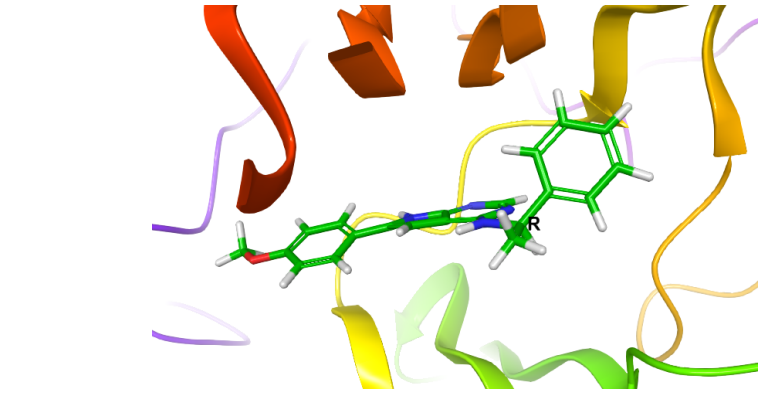
(b) *S*-**2a**, phenyl up



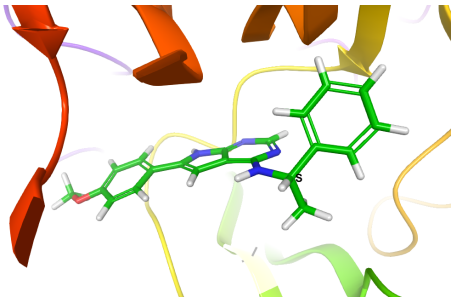
(c) *S*-**2a**, phenyl down

Figure 4.9: Representative conformations for both enantiomers of compound **2a**. During most of the simulation, both enantiomers have very similar conformations; however, for a short period (from 500ns to 700ns) *S*-**2a** adopts a rather different conformation in which the phenyl-amine substituent rotates about 120°.

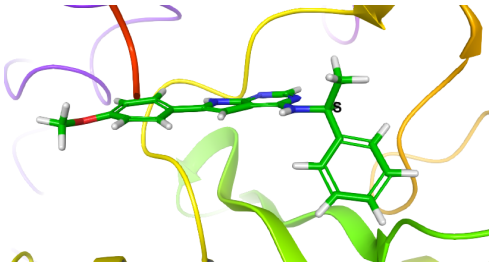
We have now covered what the average conformations of the inhibitors look like and how they change - in particular, we note that all inhibitors have an almost identical conformation, similar to that of AEE788 as described by Yun et al.. The amount of rotatable bonds is generally small compared to the number of atoms, owing to the rigid aromatic ring structures in these inhibitors. Before we move on to considering the influence of the binding pocket, we note that there is a significant structural difference to be observed between methyl-substituted inhibitors **1a**, **2a**, **2b** and **3a**) and methanol substituted compounds **1b**, **1c**, and **3b**). For the methyl compounds, no difference is apparent between the high-potency R enantiomer and the low-potency S enantiomer, outside of some increased steric repulsion between the methyl and the scaffol in the S conformation due to proximity. In the case of the methanol compounds, however, we see that the low-potency R enantiomer (the CIP priority switches due to the oxygen) adopts a similar orientation of the methanol group, but in this case the group forms an internal hydrogen bond to the N3 nitrogen on the scaffold. The fact that this stabilization occurs in the *low-potency* enantiomer was a cause of consternation to us until this thesis, as we had seen a similar behaviour in our QM/MM simulation. In this thesis, however, we find an answer that explains why the internal hydrogen bond actually results in a *less* favourable binding mode than the methanol "out" conformation that occurs in the high-potency S enantiomer. In order to fully explain this, though, we will need to finally involve the binding pocket residues and the effect of the solvent.



(a) *R*-**2b**, phenyl up

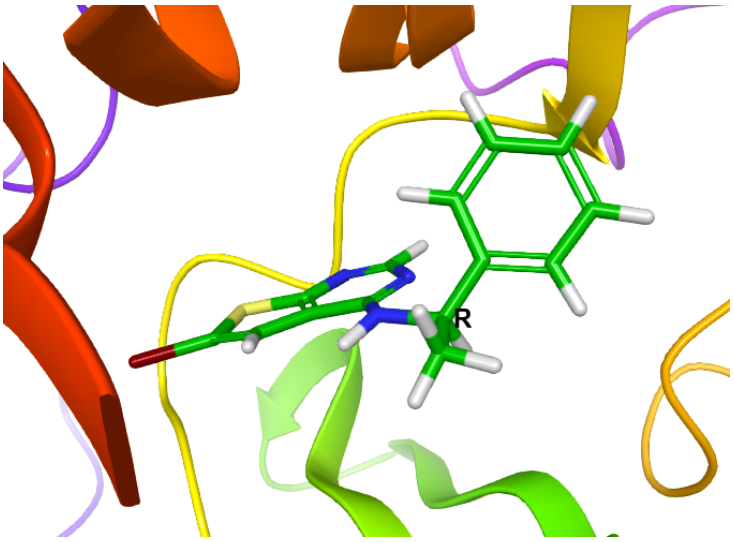


(b) *S*-**2b**, phenyl up

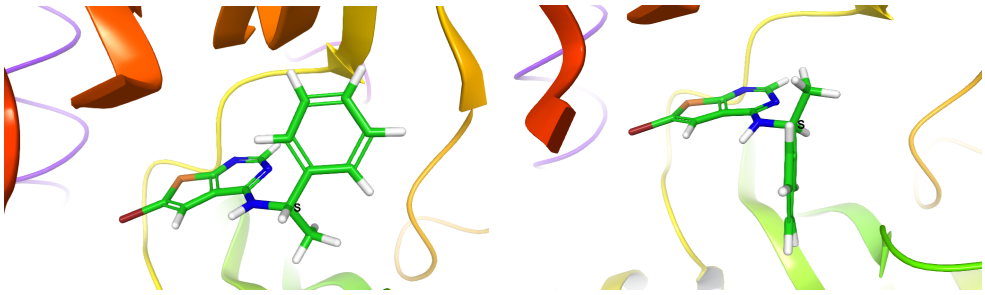


(c) *S*-**2b**, phenyl down

Figure 4.10: Representative conformations for both enantiomers of compound **2b**. While the R enantiomer stays in the phenyl up conformation throughout the entire simulation, the S enantiomer changes conformation thrice, spending about 25% of simulation time in the phenyl down position.



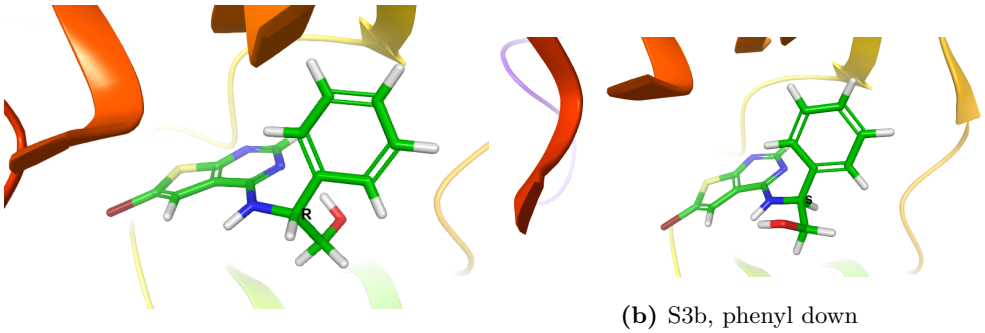
(a) *R*-**3a**, phenyl up



(b) *S*-**3a**, phenyl up

(c) *S*-**3a**, phenyl down

Figure 4.11: Representative conformations for both enantiomers of compound **3a**.



(b) S3b, phenyl down

(a) *S*-**3b**, phenyl up

Figure 4.12: Representative conformations for both enantiomers of compound **3b**. Notice how the R enantiomer forms an internal hydrogen bond between the chiral methanol group and the N3 nitrogen.

4.4 Residue-ligand interaction analysis

The protein-ligand interaction histograms (Figure 4.13) shows how frequently each residue interacted with the inhibitors as a fraction of total simulation time and what kind of interactions it had, restricted to those residuals with an interaction frequency higher than 30% with at least *one* inhibitor. Figure 4.14 shows these interactions as a timeline, indicating exactly when these interactions happen during the simulation. Finally, Table 4.1 lists the actual values of the interaction frequency. These three figures are all different means of showing the same idea - we can look at how frequently each residue interacts with each inhibitor and correlate that to the structural differences between each inhibitor as well as to the binding mode of these inhibitors.

Our first group of residues are the three hydrophobic residues Ala743, Leu718 and Leu844. From the interaction frequency graphs it is clear that these three interact in roughly the same manner across all compounds. If we look at the shape, orientation, and location of these three residuals, we find that

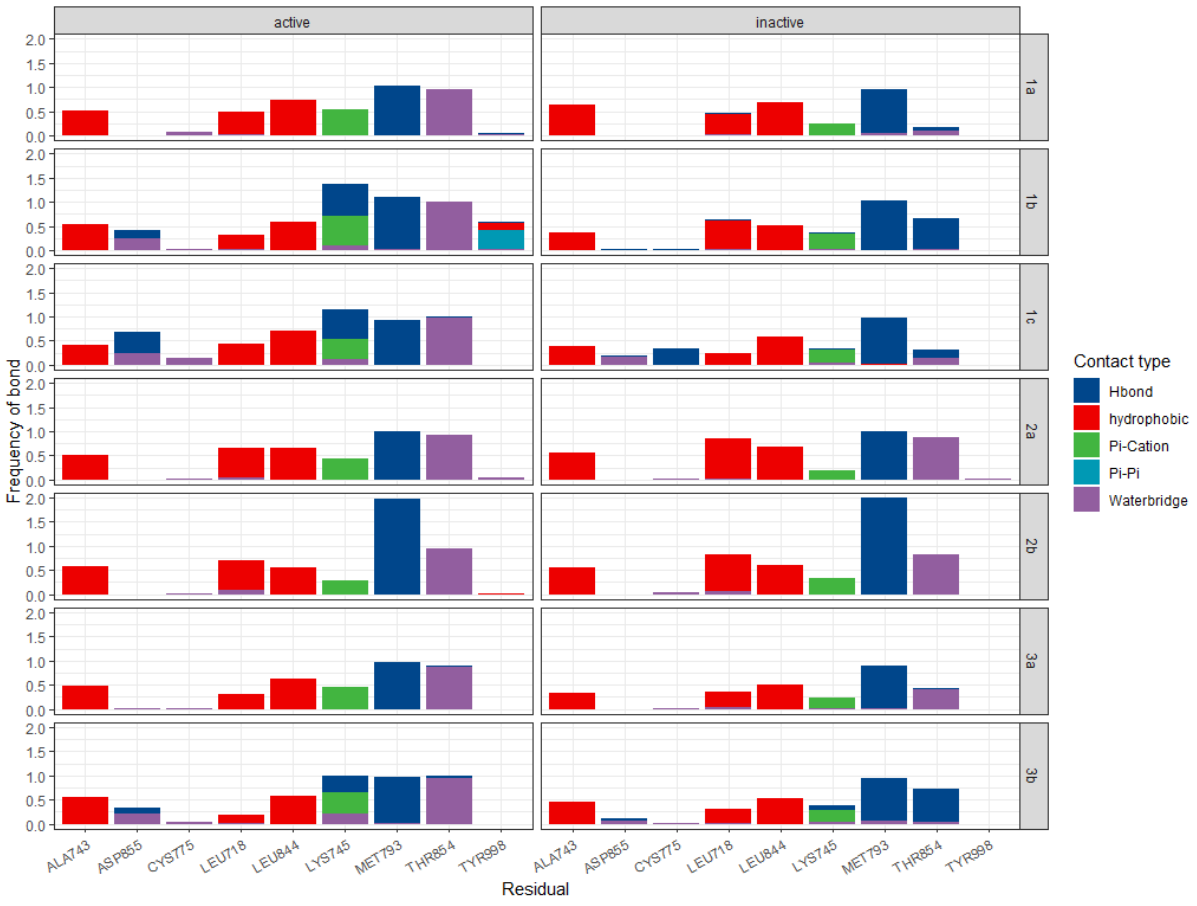


Figure 4.13: These histograms show how frequently each compound interacted with a given residue over the course of their respective simulation. Values over 1.0 are possible as some protein residue may make multiple contacts of same subtype with the ligand, as is the case for the two distinct hydrogen bonds between **2b** and Met793.

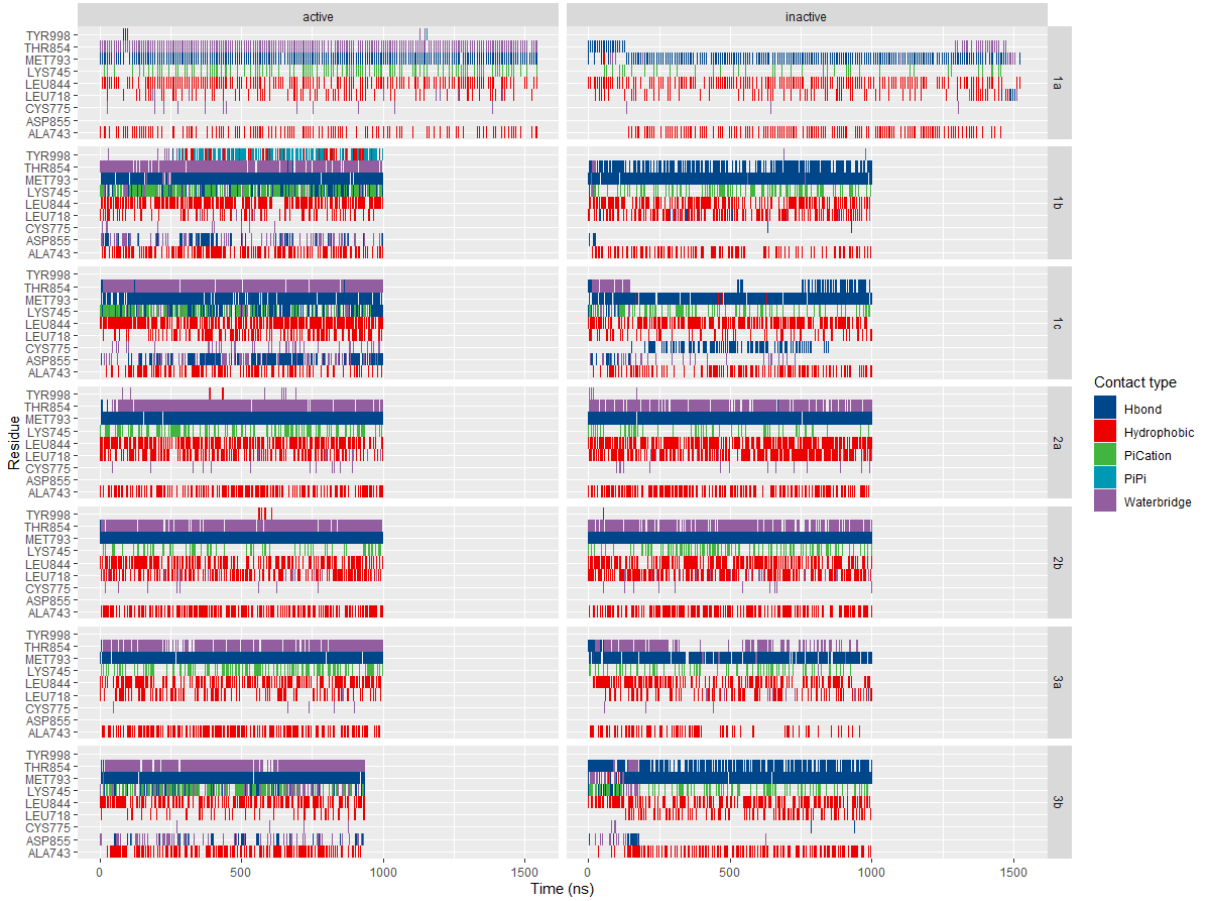


Figure 4.14: Histograms of interaction frequency as a fraction of simulation time for each inhibitor and any residue with a frequency higher than 30%. A blank (grey) spot indicates no interaction was registered at this time step, while a colour indicates which kind of interaction took place. We also want to make clear that the interaction density of **1a** is less contiguous without it necessarily having half the interactions - it was merely sampled at half the rate of the other simulations.

they are nonpolar residues which lie above and below the scaffold and tail fragment. If we orient the double ring to lie in the plane, Leu844 lies directly below while Leu718 and Ala743 lies directly above the double ring, forming a narrow hydrophobic cleft, see Figure 4.15. This cleft formation occurs for both the phenyl up and phenyl down conformations of *S*-**2a**, *S*-**2b** and *S*-**3a**. Leu718 additionally lies above the tail fragment, providing an explanation for why the tail prefers lying coplanar to the scaffold and why, as we explained earlier, there is an energy barrier between the two torsional energy minimums of compounds **1a-c**.

There is one other residue which interacts in the same manner for both enantiomers for all pairs: Methionine-793. In particular, the interaction with this hinge residue only involves its backbone peptide group - the side chain faces away from the binding pocket. The amine group of the backbone forms a hydrogen bond to the N1 nitrogen of the scaffol, with an additional hydrogen bond formed between the carbonyl part of the backbone and the NH group of the pyrrolopyrimidine scaffold of **2b** explaining why

Table 4.1: Interaction frequency between the inhibitors and surrounding residues organized by interaction type.

Molecule	Enantiomer	ALA743	ASP855		CYS775		LEU718			LEU844	LYS745			MET793		THR854		TYR998			
		hydrophobic	Hbond	Waterbridge	Hbond	Waterbridge	Hbond	hydrophobic	Waterbridge	hydrophobic	Hbond	Pi-Cation	Waterbridge	Hbond	Waterbridge	Hbond	Waterbridge	Hbond	hydrophobic	Pi-Pi	Waterbridge
1a	active	0.5402	-	0.0009	-	0.0332	0.0096	0.4554	0.0288	0.7150	-	0.5542	-	1.0079	-	-	0.9598	-	0.0070	0.0026	0.0096
	inactive	0.6841	-	-	-	0.0071	0.0327	0.4389	0.0310	0.6761	-	0.2257	0.0009	0.9425	0.0230	-	0.1257	-	-	-	0.0009
1b	active	0.5403	0.1949	0.2386	-	0.0287	0.0050	0.2748	0.0156	0.5709	0.6814	0.6002	0.0800	1.0737	0.0225	0.0037	1.0081	0.0112	0.1986	0.4841	0.0125
	inactive	0.3467	-	0.0037	0.0125	-	0.0275	0.6071	0.0200	0.4972	-	0.3691	-	1.0356	0.0087	0.6483	0.0019	-	0.0019	0.0012	0.0012
1c	active	0.4129	0.5147	0.2492	-	0.1393	-	0.4853	-	0.6877	0.6065	0.3785	0.0868	0.9244	-	0.0119	0.9838	-	-	-	-
	inactive	0.3973	-	0.1081	0.3991	0.0025	-	0.2180	0.0019	0.5971	-	0.2623	-	0.9744	-	0.1949	0.0031	-	-	-	-
2a	active	0.5122	-	-	-	0.0281	-	0.6034	0.0425	0.6359	-	0.4372	-	0.9969	-	-	0.9656	-	0.0050	-	0.0369
	inactive	0.5515	-	-	-	0.0294	-	0.8495	0.0212	0.6790	-	0.1711	-	0.9988	-	0.0025	0.8695	-	-	-	-
2b	active	0.5878	-	-	-	0.0187	-	0.6352	0.0856	0.5134	-	0.2692	-	1.9888	-	-	0.9525	-	0.0137	-	-
	inactive	0.5503	-	-	-	0.0331	-	0.7283	0.0768	0.5984	-	0.3660	-	1.9963	-	-	0.8239	-	-	-	-
3a	active	0.4766	-	0.0031	-	0.0150	-	0.2986	0.0019	0.6327	-	0.4603	-	0.9756	-	-	0.8882	-	-	-	-
	inactive	0.2911	-	-	-	0.0100	0.0012	0.3348	0.0562	0.4491	-	0.2255	0.0044	0.9019	0.0275	-	0.3285	-	-	-	-
3b	active	0.5410	0.1153	0.1776	-	0.0258	-	0.1817	0.0014	0.5776	0.3315	0.4407	0.2115	0.9824	-	0.0515	0.9532	-	-	-	-
3b	inactive	0.4922	-	0.0094	0.0156	0.0006	-	0.3548	0.0019	0.5203	-	0.2511	-	0.9925	-	0.6846	0.0006	-	-	-	-

it has double the interaction frequency of the other inhibitors. This bond was already known both from our earlier QM/MM study,[22] the docking study,[24] and the binding mode of AEE788 found in 2007.[56] It is reassuring that this bond reappears in our simulation, since it is considered one of the main reasons why thieno-, furo- and pyrrolopyrimidines are effective scaffolds for EGFR inhibitors in the first place.[12] The additional bond formed to **2b** additionally explains why this inhibitor has a lower IC_{50} value than the furo- and thienopyrimidines.

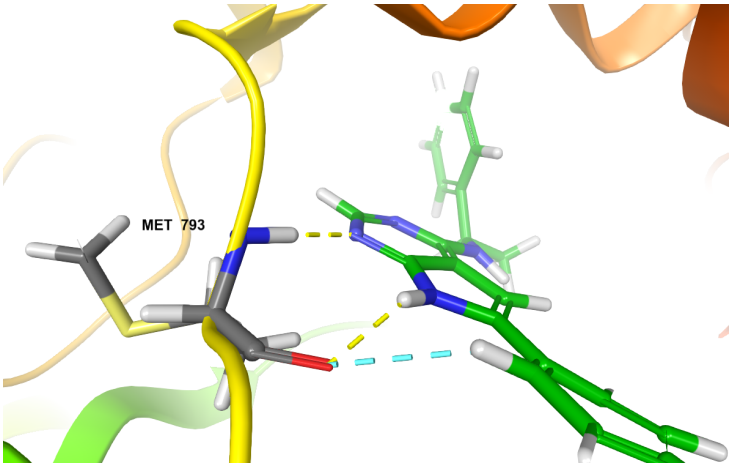


Figure 4.16: All investigated inhibitors bond strongly to the backbone of Metionine793 lying in the hinge region, forming a bond from the N1 nitrogen on the scaffold to the amine group of the backbone. Depicted is *R*-**2b**, which forms an extra bond due to its NH heteroatom. Dashed yellow lines are hydrogen bonds; dashed teal lines indicate possible aromatic hydrogen bonds.

One residue interacts frequently with only one inhibitor: Tyrosine-998, which only interacts with *S*-**1b**. This is because, as mentioned, the folding of the Prime loop, of which Tyr998 is a part of, is rather chaotic and so whether any inhibitors interact with any residues on this loop at all depends on the initial conditions of the simulation, and should ideally be determined by running multiple simulations with different starting points. It also poses a challenge for any energy calculation based on ligand-residue interactions (or indeed, any energy calculation at all which involves the Prime loop) because of this chaotic behaviour. Additionally, since this residue only interacts with the tail of *S*-**1b** (and not the

scaffold or the amine substitute), it is unlikely to be a significant contributor to the stabilization of one enantiomer above another. For these reasons, we disregard this interaction in our further analysis; it may certainly be important for further optimization of the ligand, but in this thesis, we estimate it to not have much real impact in terms of stabilizing one enantiomer more than the other, even if it does so in this simulation.

Having discussed the residues which are consistent across simulations and discarded one that isn't, we are left with the residues which behave *differently* across simulations - the ones which finally give us a clue towards explaining the stereoisomeric difference in potency. There are three of these residues: Asp855, which forms hydrogen bonds and water bridges; Lys745, which forms pi-cation bonds and hydrogen bonds; and Thr854, which forms water bridges and sometimes hydrogen bonds. Immediately we note that there is a significant difference in interaction behaviour between methanol-substituted inhibitors and methyl substituted inhibitors. Only the three methanol-containing inhibitors **1b**, **1c** and **3b**) show any interactions with Asp855 at all, and only these form a hydrogen bond to Lys745; the four methyl-containing compounds **1a**, **2a**, **2b** and **3a**) do not have these interactions. We therefore consider the methanol- and methyl- substituted compounds separately here.

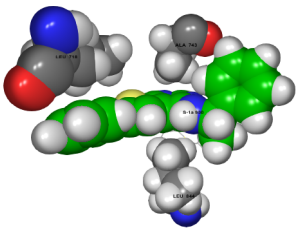
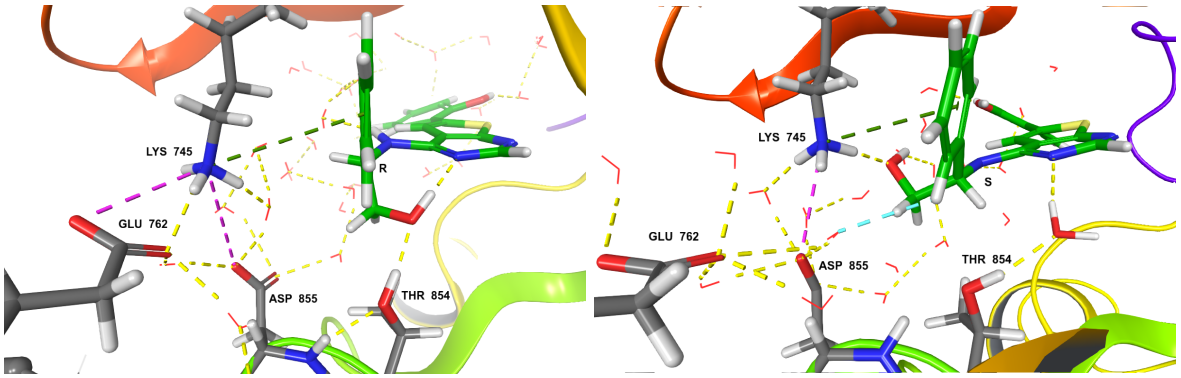


Figure 4.15: Space-filling CPK-model of compound *S*-**1a** (green) and the three hydrophobic residues of Ala743, Leu718 and Leu844 from frame 695 of the MD simulation of *S*-**1a** bound to EGFR, showing that the three hydrophobic residues form a cleft surrounding the inhibitor.

Focusing first on the methanol containing compounds, we notice that there are clearly systematic differences between the high-potency S enantiomer and the low-potency R enantiomer in how they interact with the aforementioned residues based on the interatcion frequency: First, only the S-enantiomer forms a hydrogen bond to Lys745, even though both form a pi-cation bond; Second, S-enantiomer's interaction with Asp855 is markedly stronger than the R-enantiomer's; and finally, the S-enantiomer forms a water bridge to Thr854, while the R-enantiomer forms a hydrogen bond (in the case of *R*-**1b** and *R*-**3b**) or has a smaller interaction consisitng of both hydrogen bonding and water bridges (R1c). In this case, we think a picture says more than a thousand words - Figure 4.17 shows the binding mode of the S and R enantiomers of **1b**, both showing the location and orientation of the aforementioned residues, and most importantly: the water molecules.

First, consider the bonds to Lys745 and Asp855. At pH 7, our preprocessing estimated that Lysine is a protonated amine and Asp855 a deprotonated carbooxylic acid; therefore, they form an ionic bond between each other. Further, since Lys745 is a cation, it forms a pi-cation bond to the phenyl ring

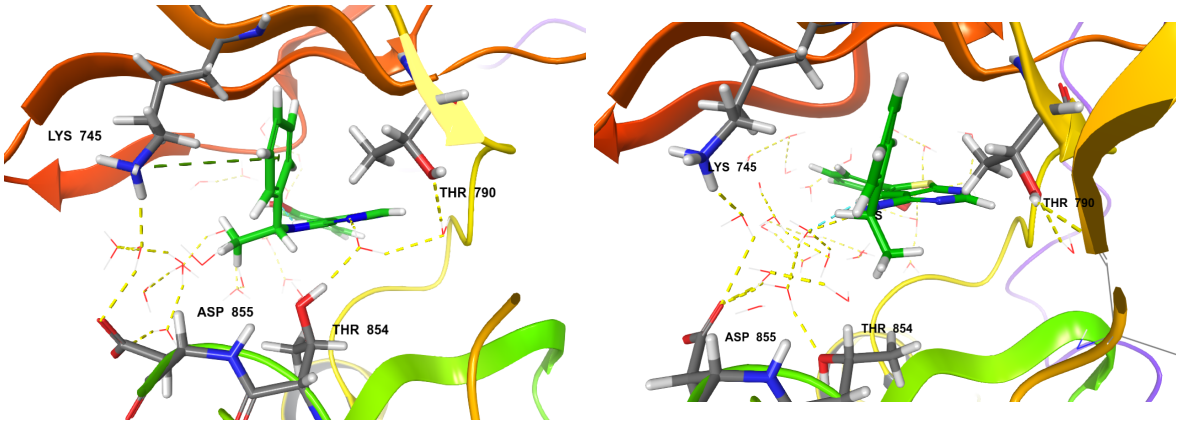


(a) *R*-**1b** **(b)** *S*-**1b**

Figure 4.17: The high potency S-enantiomer of the methanol compounds form additional hydrogen bonds to Lys745 and Asp855 via the methanol while retaining a bond to Thr854 mediated by water, in contrast to the low-potency R-enantiomer which bonds to Thr854 via a direct hydrogen bond from methanol. Shown are representative binding modes of the R/S-enantiomers of **1b** and their interactions with Lys745, Asp855, and Thr854, and the surrounding solvent. The bridging water molecule is thickened for emphasis. Dashed lines show non-covalent bonds: yellow for hydrogen bonds, green for pi-cation bonds, and purple for ionic bonds.

for both enantiomers. In addition, however, the S-enantiomer methanol forms a strong hydrogen-ionic bond to this residue - except for when it forms a hydrogen bond to Asp855 or one of their surrounding solvent molecules. These solvent molecules form a rather interesting network of hydrogen bonds between various polar/ionic residues, the solvent, and the inhibitor itself - in the broader context of EGFR, the formation of a salt bridge between Lys745 and Glu762 is one of the primary interactions that activate the protein. However, in our narrow scope of investigating the stereoselectivity of EGFR towards our inhibitors, *we theorize that these two alternating non-covalent bonds to Lys745 and Asp855 explain why the S-enantiomer of **1b**, **1c** and **3b** has a higher binding affinity than the R-enantiomer.* Still, one question remains: what about the aforementioned water bridge to Thr854?

Earlier we noted that the R-enantiomers of the methanol-substituted compounds tended to form internal hydrogen bonds between the methanol and the N3 nitrogen on the scaffold. When we include Thr854, it becomes obvious that the methanol actually forms two hydrogen bonds in the low-potency enantiomer: one internal bond to the N3 nitrogen, and one external bond to the hydroxy group of Thr854. More importantly, however, we see what instead happens when there is only a hydrogen instead of a methanol in this area: A water bridge forms between N3 and Thr854. This means that even without the methanol, the bond to Thr854 can be maintained in almost exactly the same manner, and with greater flexibility due to the free movement of solvent. Further this structural water is incredibly stable; normally, water diffuses in and out of the binding pocket continuously, meaning that labeling any water molecule in order to calculate its energy becomes a futile task. Not so with this water bridge, however; once a water molecule formed this bridge, it often stayed in that area for several hundred nanoseconds. In fact, this



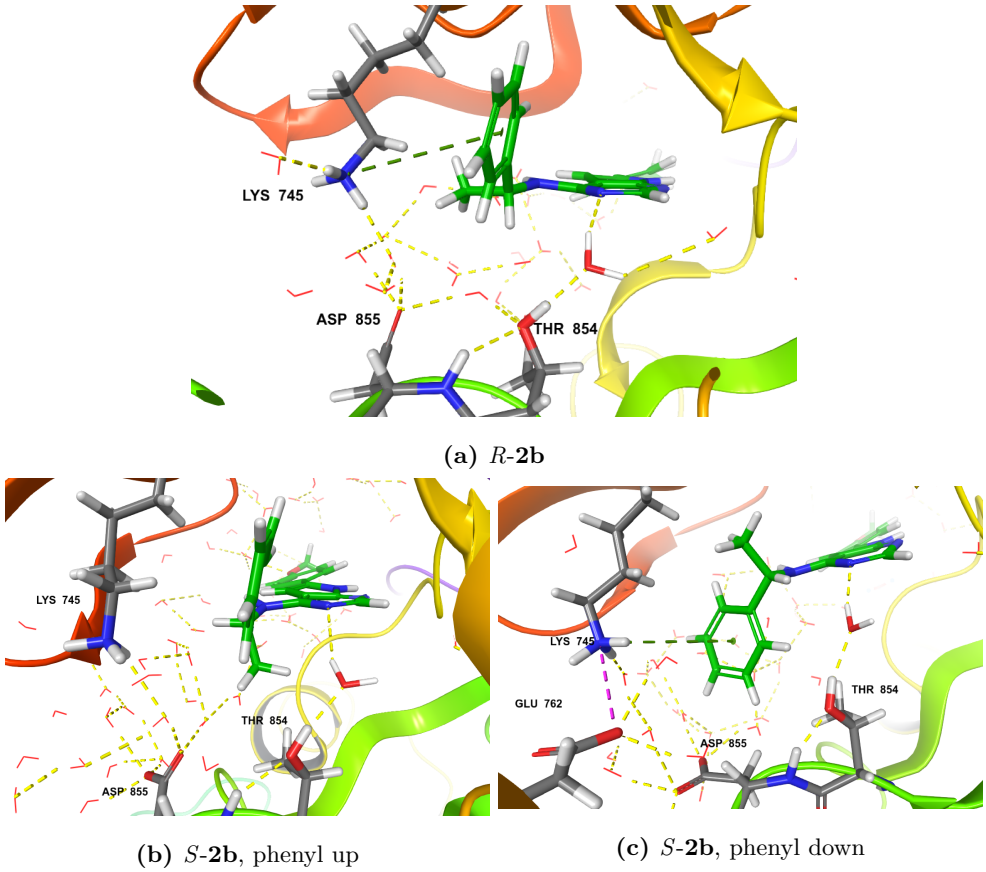
(a) *R*-**1a** **(b)** *S*-**1a**

Figure 4.18: The high potency R-enantiomer of the methyl thienopyrimidines **1a** and **3a**) forms a single bond that the low-potency S-enantiomer doesn't: a water-mediated hydrogen bond to Thr854. Shown are representative binding modes of the R/S-enantiomers of **1a** and their (non-)interactions with Lys745, Asp855, and Thr854, and the surrounding solvent. Dashed lines show non-covalent bonds: yellow for hydrogen bonds and green for pi-cation bonds.

water bridge happened for *all* high-potency enantiomers in our simulations, while only two low-potency enantiomers formed it consistently, namely the methyl-substituted furo- and thienopyrimidines **2a** and **2b**.

Speaking of which, do the methyl-substituted compounds show similar behaviour? Clearly, because of their structure, they cannot have the internal hydrogen bond, nor the polar bonds to Asp855 or Lys745, which is confirmed by interaction analysis. What remains is the pi-cation bond between the phenyl and Lys745 and the water bridge to Thr854. As mentioned, we see a difference between the thienopyrimidines **1a** and **3a** and the pyrimidines with different heteroatoms, **2a** and **2b**. The former establish the water bridge much more frequently in their high-potency R enantiomer than in their low potency S-enantiomer. Inspecting their trajectories, it is not hard to explain why this is - the methyl is a larger group than the hydrogen it replaces, which present a steric hindrance towards forming the water bridge, while at the same time, steric repulsion between the methyl and Thr854 forces the latter away from the ligand. For these two thienopyrimidines, then, we theorize that *the displacement of the water bridge to Thr854 due to steric hindrance from methyl in S-**1a** and S-**3a** causes these enantiomers to be less stable, explaining why the R-enantiomers have a higher binding affinity.*

What then, of the furo- and pyrrolopyrimidine based **2a** and **2b**? These do not displace this water bridge in their low-potency conformation. Instead, the structures are perturbed slightly compared to the thienopyrimidines such that the methyl lies just outside the space occupied by the water bridge, while phenyl is slightly farther away from Lys745. In the case of the phenyl-down conformation, we see a similar behaviour, with the phenyl still having a pi-cation interaction with Lys745 and the water bridge being



(a) *R*-**2b**

(b) *S*-**2b**, phenyl up

(c) *S*-**2b**, phenyl down

Figure 4.19: Unlike the methyl substituted thienopyrimidines, pyrrolopyrimidine **2b** (and furopyrimidine **2a**, not shown) does not displace the water bridge to Thr854; in fact, this bridge is maintained in both the phenyl up and phenyl down conformation.

maintained. The interaction frequency plots do not hint at any particular discrepancy either. Why is it that our simulations allow these two inhibitors to maintain the water bridge while **1a** and **3a** cannot?

So far, we have treated the inhibitors like the scaffold is in approximately the same position with respect to the binding pocket for all of them. However, if we actually measure the distance between Met793 and the heteroatom for our exemplar geometries, we find that the thienopyrimidines are farther away from the hinge compared to **2a** and **2b**. Unfortunately, we uncovered this rather late in our work, and so we did not have enough time to perform a more rigorous analysis of our trajectory to verify that the thienopyrimidines are *consistently* farther away from the hinge than **2a** and **2b**. We think it likely, because there is a casual relationship in play: compared to oxygen and nitrogen, sulphur has a larger radius, often longer bond lengths, and weaker non-covalent interactions, particularly in when parametrized in force fields.[81] Thus, the equilibrium distance is farther away due to weaker attractive interactions compared to the furo- and pyrrolpyrimidines. In turn, we hypothesize that because **2a** and **2b** are generally closer to the hinge, the steric strain on the phenyl (which is on the opposite side from the hinge) is decreased, allowing the amine moeity a greater degree of freedom to move methyl out of the way of the water and Thr854. This is why they show few differences between the two enantiomers in our simulation; they

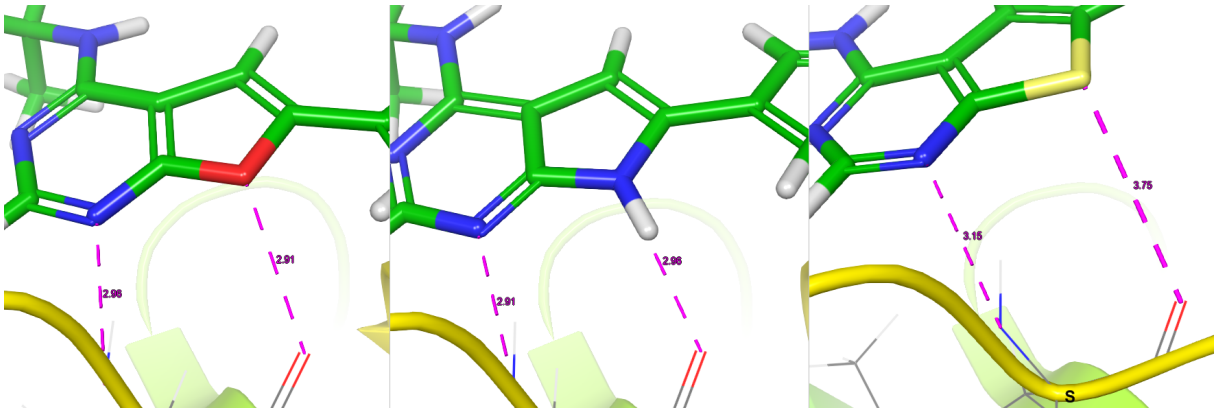


Figure 4.20: In our exemplar structures, the thienopyrimidine scaffold is often farther away from the hinge compared to the furo- and pyrrolopyrimidine scaffold. Depicted from left to right are the scaffold and hinge region of *S*-**2a**, *S*-**2b** and *S*-**1a** with measurements showing the interatomic distance between the N3 nitrogen of the inhibitor and the amine nitrogen of Met793, as well as the distance between the heteroatom and the carbonyl group of Met793.

simply are able to accommodate the water bridge better due to the scaffold's close proximity to Met793. This does not mean there is no repulsion in place, of course; only that its impact is lessened.

Here, the weakness of using the interaction frequency to reason about the molecular behaviour becomes clear; because the counting of an interaction is a binary check, we do not get a clear estimate of how strong a particular interaction is nor how much energy this interaction contributes. Our previous arguments surrounding the methanol-substituted compounds are possible because it is clear to us that the extra interactions are not just rarer, but actually inaccessible to the wrong enantiomer, even if the force field does not precisely model the strength of these interactions; in the case of the methyl-substituted compounds, the difference becomes one of degree rather than possibility, meaning that a proper analysis of this behaviour should employ methods that account for how interactions vary with distance. In the future, perhaps, we may want to once again model a portion of the system quantum mechanically in order to more accurately estimate the strength of these various interactions and the strain imposed upon the ligand, now that we have a clearer picture of the interactions that occur within the binding pocket. Unfortunately, this also means that *our simulations are unable to determine whether the enantiomers **2a** and **2b** can be differentiated by a similar set of interaction behaviours as those of **1a** and **3a***, even if we consider it likely that the water bridge to Thr854 is a key factor in all of these inhibitors.

4.5 Summary of results

In the preceding section, we have presented analyses of data generated by Molecular Dynamics simulations of a set of EGFR inhibitors bound to the active seat of EGFR, and gradually pieced together a picture of how the inhibitors interacts with the binding pocket. The final paragraphs outline our answer to the thesis question: a possible explanation for why one enantiomer is more stable than the other, with

separate behaviour for the methanol-substituted inhibitors **1b**, **1c** and **3b**) and the methyl-substituted thienopyrimidine inhibitors **1a** and **3a**). In common for both of these is a water-mediated hydrogen bridge between the N3 nitrogen on the scaffold and Thr854, which is present in the high-potency conformation and either displaced (for methyl-substituted inhibitors) or replaced (by methanol) in the low-potency conformation. Additionally, the high-potency methanol-substituted inhibitors have favourable hydrogen bond interactions to Lys745 and Asp855 which are not possible in the low-potency enantiomers. As another way of illustrating this, we provide the 2D chemical diagram in Figure 4.21. We hit one snag, that being the furopyrimidine **2a** and the pyrrolopyrimidine **2b**, neither of which show any differences between their enantiomers that we could determine from our simulations, unfortunately.

4.6 Retrospective of earlier study

In light of our new theory about the binding site interactions, we reviewed one of our old studies, the Glide docking study first used to model our inhibitors. In this study, the 2J6M was allowed to keep its native three structural waters; however, the methodology of Glide naturally freezes these water molecules in place, since it considers them a part of the receptor, which itself is completely rigid. Further, it does not model solvent at all. This is natural considering its purpose as a massive screening tool meant to sort through thousands of ligands, evaluate possible docking poses, score these, and return a ranked set of the best scoring ligand poses. However, even with these massive approximations, the calculated geometries turn out surprisingly accurate - at least for the high-potency ligand. Figure 4.22a shows a slightly modified result of Glide, where the low-potency *S*-**1a** has been superimposed upon the high potency *R*-**1a** based on their scaffold. The image clearly shows that while both conformations maintain the pi-cation bond to Lys745 and the water bridge to Thr854, *S*-**1a** would experience significant steric clash with Thr854 in this position. This is a consequence of the rigid water - the displacement of the water molecule seen in our MD simulations cannot happen, so instead, the ligand conformation adopts to allow the water bridge to continue.

If we instead look at Figure 4.22b, we see the actual location of the S-1a inhibitor pose (purple) and the aforementioned superimposed-on-*R*-**1a** position. It is clear that in order to decrease steric repulsion from Thr854, Glide moved S-1a away, which weakened both the hydrogen bond to Met793 and the pi-cation bond to Lys745 in the process. Overall, we find that the biggest error committed by Glide in terms of optimal ligand pose is its inability to displace or remove the structural waters. This results in accurate poses for the high-potency ligands, but significantly different poses for the low-potency enantiomers. Intuitively, this should mean that the internal energy as well as the Van der Waals repulsion of the low-potency enantiomer should be higher; looking at the calculated energy terms, that intuition appears to be entirely correct. While the final GlideScore only shows weak correlation to actual measured activity, if we instead sum up the Van der Waals, Coulomb, and Internal Energy terms computed by Glide, we get roughly the expected energy difference (2-4 kcal/mol) not only for 1a, but for 2a and 3a as well. The only methyl-substituted inhibitor this doesn't work on is 2b (Figure 4.23), which we believe is due to the

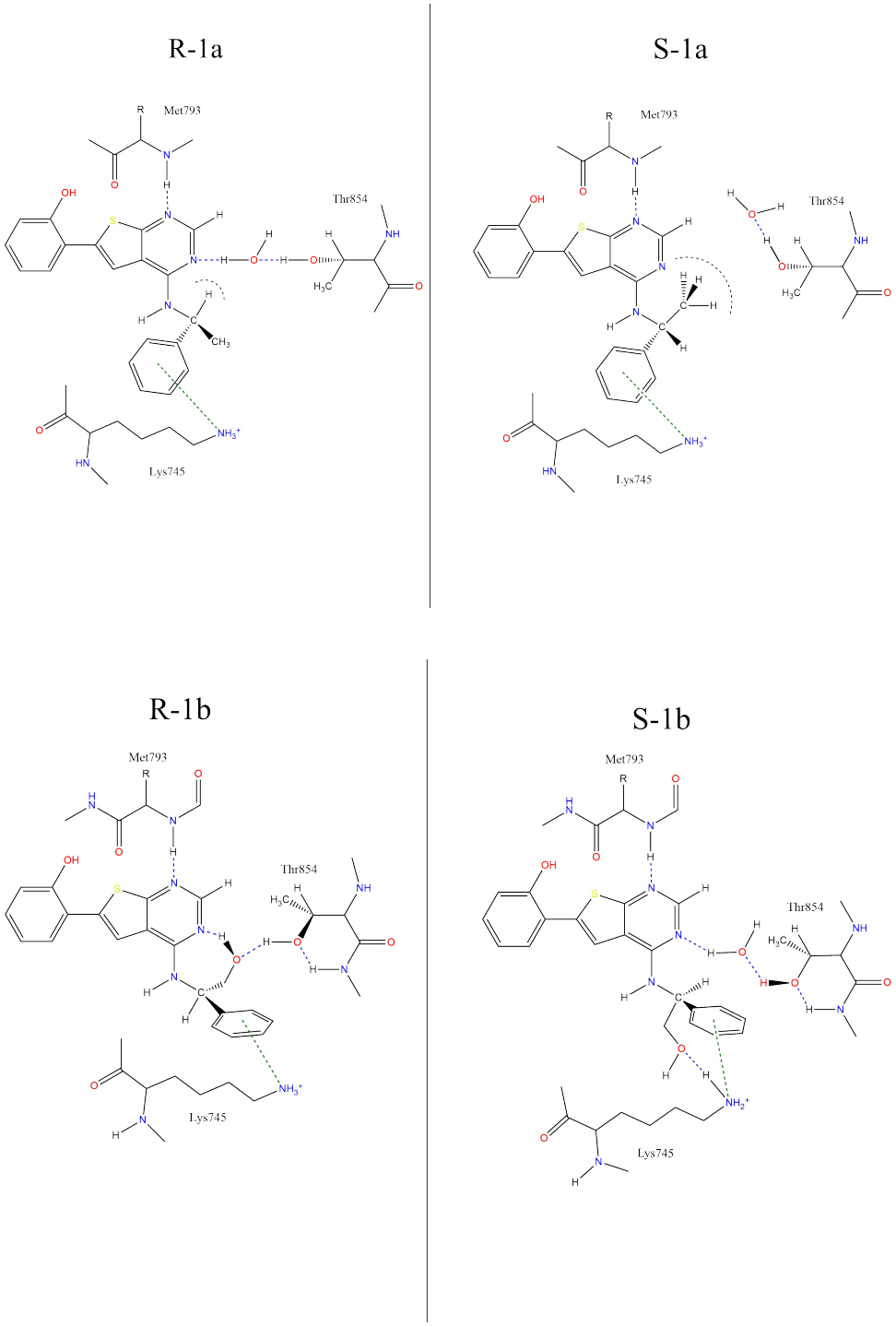
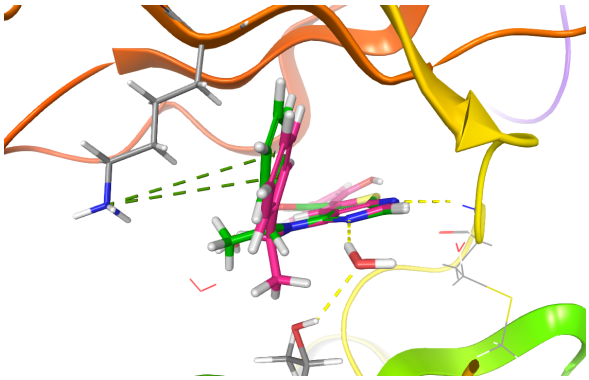
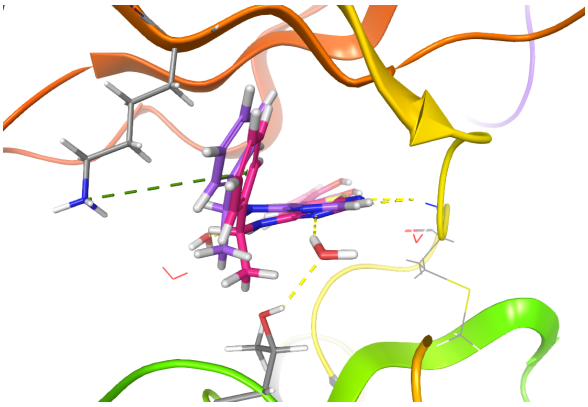


Figure 4.21: Simplified 2D schematic showing how chirality affects non-covalent interactions. Top: Steric clash with a methyl group (shown as dashed lines) prevents a water bridge to Thr854 in the *S*-enantiomer of methyl-substituted thienopyrimidine inhibitors, here **1a**, but not in the *R*-enantiomer. Bottom: Additional bond from methanol to Lys745 causes the *S* enantiomer to be more stable than the *R* enantiomer in methanol-substituted thienopyrimidines, here **1b**.



(a) Top Glide pose of R-1a (green), with the top pose of S-1a (pink) superimposed on the scaffold of R-1a. It is clear that this hypothetical pose would experience significant steric clash between the S-1a methyl and Thr854.



(b) Top pose for S-1a found by Glide, in two locations: the correct position as found by Glide (purple), and the position it takes when superimposed on the scaffold of R-1a.

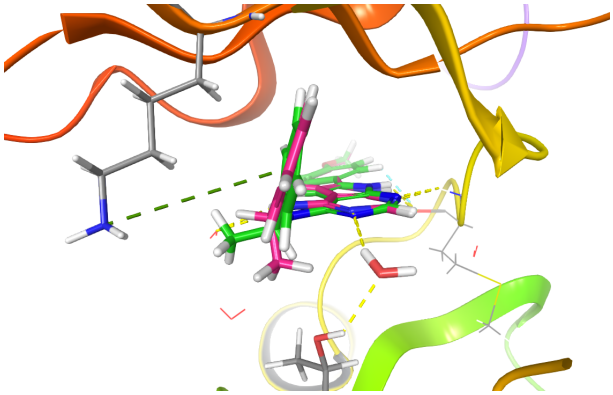
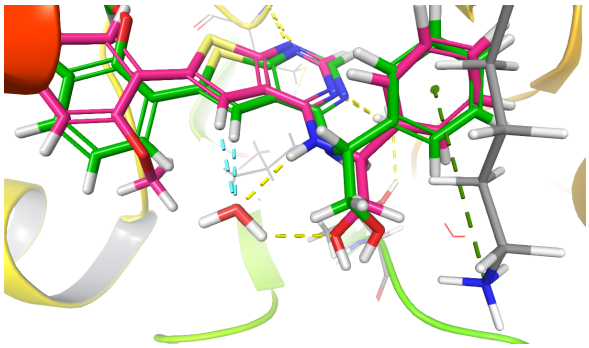


Figure 4.23: Top Glide poses of R-2b (green) and S-2b (pink)

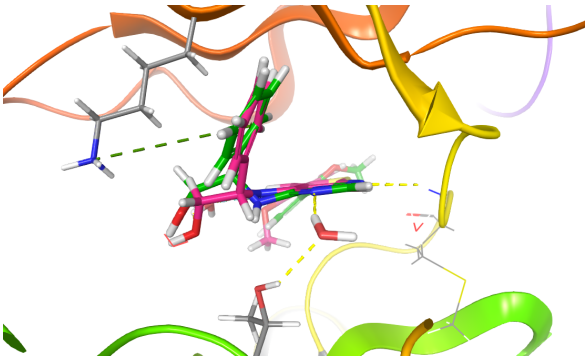
extra hydrogen bond to Met793 being overestimated.

In the case of the methanol-substituted compounds, we find that Glide does not predict the bond to Lys745 at all (Figure 4.24b); however, we also see that this is not, necessarily, due to some peculiarities about Lys745, but simply that one of the structural waters is perfectly positioned to bond to the S-methanol (Figure 4.24a) - in the MD simulations, this area is generally saturated with water molecules, so there's no single molecule which consistently forms this bond. We also looked at the later poses found by GlideXP and MM/GB-SA, and find that they suffer from the same problem of being unable to displace or remove the water molecules. These considerations make it clear that Glide is actually pretty adept at predicting binding poses, but that it suffers when structural waters are included when they shouldn't be (and likely vice versa).

This supports our own findings that modelling of *flexible* water molecules is crucial to obtaining correct binding modes for EGFR inhibitors based on thieno-, furo-, and pyrrolopyrimidine scaffolds. Without



(a) Best Glide poses for 1c from a different angle, showing how the structural water forms a hydrogen bond to the methanol group.



(b) Top glide poses of R-1c (green) and S-1c (pink)

the water, the interaction with Thr854 disappears, removing one of the key stereo-selective interactions as was the case for the QM/MM study; With rigid water, the low-potency enantiomers adopt unusual binding poses due to the steric repulsion from the rigid water. One approach that doesn't involve long MD simulations would be to estimate the location of water molecules in the binding pocket and then include the cost of desolvating these in the energy calculation; programs such WaterMap [82] and HydraMap [83] were made explicitly for this task.

4.7 Further work

Our theory could well do with more rigorous validation through both computational modelling and laboratory experiments. Since we opted to run long single-trajectory simulations, we made headway in uncovering dynamic processes that happen on time scales under a microsecond, such as the ligand conformation transitions revealed by the RMSD plot, but we did not anticipate that the protein sequence added by Prime would be as weakly folded as it was. Using ensemble methods to repeatedly simulate our model systems with slightly different initial conditions would help quantify model uncertainty and dependence on initial conditions, furnishing our work with some more statistically well founded data now that we have a clear hypothesis to investigate. In a similar vein, physical experiments such as X-Ray Diffraction of EGFR co-crystallized with some of our inhibitors may be used to validate and/or disprove our hypothesis about the water bridge being displaced in the low-potency enantiomer.

A weakness with our method is the lack of energy quantification, particularly with respect to residue interaction strength. Our current methodology is limited in scope to what are essentially binary evaluations of interactions ("is the interaction present, or not?"), unlike real interactions whose strength varies non-linearly with distance. An obvious next step is to compute the interaction strength quantitatively, as we tried in our previous project, but now with a clearer understanding of which residues interact in which way - e.g. the strength of the water bridge vis-a-vis the internal hydrogen bond can be settled more accurately using quantum mechanics on a reasonable scale, since the ligand-water-residue system contains at most 70 atoms.

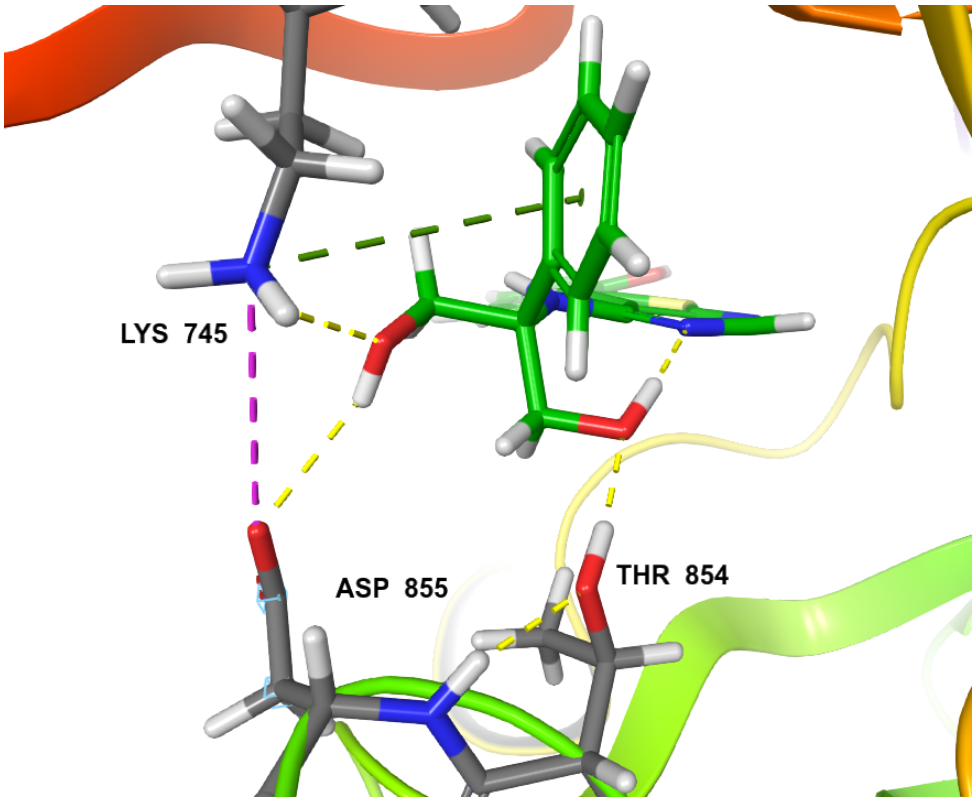


Figure 4.25: A mockup of a possible binding mode of a dimethanol compound

For the experimentalist, our work presents further insights on molecular behaviour that may be used to guide further development of EGFR inhibitors. The primary takeaway is that the cramped conformation of the binding pocket causes the carbon-amine substitute to curve back towards the scaffold, necessitating consideration of spatial distance between distantly-bonded atoms; the secondary takeaway is that the scaffold has a propensity for forming water-mediated hydrogen bonds, so one cannot just consider the ligand and the binding pocket alone. Additionally, we find that the ligand is much more solvent-exposed than was previously estimated by Prime. Our simulations indicate that methanol-substituted compounds are more likely to have higher binding affinity owing to the additional polar bonds these are able to make. Our findings indicate the possibility of creating compounds that replace the Thr854 water bridge with e.g. a hydroxy group, as the low-potency enantiomers were found to do, while also forming the favourable interactions of the high-potency enantiomer. A simple candidate, at least in terms of structural changes, is one where both small functional groups are methanol; we illustrate the imagined binding mode in Figure 4.25. This compound would both have the favourable bonds to Lys745 and Asp855, while also replacing the water bridge with an internal hydrogen bond.

5 Conclusion

In this thesis, we have used Molecular Dynamics simulations to investigate the underlying mechanisms of chirality-dependent inhibitory potency of a family of EGFR inhibitors. We performed long equilibrium simulations of inhibitors in the binding pocket of the intracellular kinase domain of EGFR, in order to learn more about their equilibrium binding mode and the residue interactions that occur. In order to do so, we perform interaction analyses of the trajectory and make heavy use of visualization of the trajectory to glean information about the system. We also tried to calculate the difference in relative binding free energy using these simulations as a one-trajectory MM/GB-SA calculation, but the resulting binding free energy had too high uncertainty to be considered significant.

We find that most inhibitors adopts a common binding mode most of the time. This binding mode coincides generally with the binding mode of AEE788, as described by Yun et al.. The furo/pyrrol/thieno-pyrimidine ring (the scaffold) is oriented with the 1-N in the back of the ATP-binding pocket, where it bonds with the main chain amide of Met793. Met793 lies in the hinge region of the kinase, which connect the N and C lobes. The scaffold is further sandwiched between the hydrophobic residues of Ala743 and Leu844, see figure 4.15. The 4-phenylethylamine moiety extends "up" into the hydrophobic pocket defined by Thr790, Leu788, Lys745, and Met766, where the phenyl ring forms approximately a 70°angle to the pyrimidine ring. The 6-phenyl substituent is sandwiched between Leu718 above and Gly796 below.

Our key finding is that in this binding mode, the difference in inhibitory potency between the high-potency and low-potency enantiomers can be explained by the stereoselective formation of a water bridge between the N3 nitrogen and the hydroxyl of Thr854; In the low-potency enantiomer this bride is sterically hindered, being either displaced by a methyl group or replaced by a methanol group. The methanol-substituted compounds additionally form polar bonds to Lys745 and Asp855. We propose that the increased stability due to these extra bonds are the underlying cause of the chirality-dependent inhibiton of EGFR. These findings provide new insights into the behaviour of the EGFR binding pocket and highlight the importance of solvent modelling.

Acknowledgment

Thanks to supervisor Ida-Marie Høyvik and co-supervisor Eirik Sundby for their assistance throughout the whole project. Thanks to Marcus Lexander for assistance with scripting and debugging. Thanks to Peder Langsholt Holmqvist for assistance with parsing and plotting data as well as proofreading. Thanks to Signe Onstad Saevareid, Oda Helene Berg and Martin-Kristofer Helgeland-Rossavik for insights in and proofreading of biochemistry. The computations were performed on resources provided by the NTNU IDUN/EPIC cluster as well as the Stallo cluster through UNINETT Sigma2 with allocation from project nn9409k.

List of Figures

2.1 Outline of signal cascade path from activation of EGFR by growth factors. Public domain image taken from Wikipedia (Jan 2020) 8
2.2 Crystal structure of EGFR inhibited by AEE788 (PDB:2J6M) [56] with key protein kinase features highlighted. 9
3.1 Model compounds used in this study. Fragment A (aniline) in red. Fragment B (Phenyl or bromine) in blue. Core scaffold in black. 15
3.2 The simulation system for the Molecular Dynamics simulations was prepared using Desmond System Builder. Displayed here is the resulting simulation box for compound *S*-**1c** 17
4.1 Timeline plot of the binding free energy of each inhibitors calculated for every tenth frame using the MM/GB-SA method. 21
4.2 Timeline plot of the energy of the protein-ligand complex calculated by MM/GB-SA. In theory, the difference between the average of this energy for each enantiomer is proportional to the difference in empirical IC_{50} values, but noise makes this difference rather diffuse. Compound **1a** was sampled for longer but at a lower frequency. It is clear that any differences in binding energy based on an average of these will have too great variation to be statistically significant. 21
4.3 Root mean square deviation (RMSD) of protein atoms (blue) and ligand atoms (red) of compounds **1a-c** relative to starting geometry for the protein backbone and ligand, as well as the mean (time-averaged) RMSD of the ligand. 23
4.4 Root mean square deviation of compounds **2a-b** and **3a-b** relative to starting geometry for the protein backbone and ligand, as well as the mean (time-averaged) RMSD of the ligand. 24
4.5 The protein loop added by Prime (residues 990 to 1021) tended to fold in different ways across simulations. Depicted here is the start and end frames of the simulation of *S*-**3a**, where we saw it fold into an α-helix. 25
4.6 Representative conformations for compound **1a**. Both enantiomers have two significantly different conformations each in terms of RMSD, stemming from the rotation of the tail phenol fragment, and whether the hydroxy group lies near the heteroatom or on the far side of it. Otherwise the conformation is similar to most other inhibitors, with the phenylamine pointing "up" relative to the scaffold. Note the orientation of the chiral methyl group. . . 26
4.7 Representative conformations for compound **1b**. Like **1a**, both enantiomers have two significantly different conformations each in terms of RMSD, stemming from the rotation of the tail phenol fragment, and whether the hydroxy group lies near the heteroatom or on the far side of it. Notice how the R enantiomer forms an internal hydrogen bond between the chiral methanol group and the N3 nitrogen. 27

4.8 Representative conformations for compound **1c**. Despite having an asymmetrically substituted tail like **1a** and **1b**, neither *R*-**1c** nor *S*-**1c** shows any tendencies to "flip" the tail, and so they only have one conformation each. Notice how the R enantiomer forms an internal hydrogen bond between the chiral methanol group and the N3 nitrogen. 28

4.9 Representative conformations for both enantiomers of compound **2a**. During most of the simulation, both enantiomers have very similar conformations; however, for a short period (from 500ns to 700ns) *S*-**2a** adopts a rather different conformation in which the phenylamine substituent rotates about 120°. 29

4.10 Representative conformations for both enantiomers of compound **2b**. While the R enantiomer stays in the phenyl up conformation throughout the entire simulation, the S enantiomer changes conformation thrice, spending about 25% of simulation time in the phenyl down position. 30

4.11 Representative conformations for both enantiomers of compound **3a**. 31

4.12 Representative conformations for both enantiomers of compound **3b**. Notice how the R enantiomer forms an internal hydrogen bond between the chiral methanol group and the N3 nitrogen. 31

4.13 These histograms show how frequently each compound interacted with a given residue over the course of their respective simulation. Values over 1.0 are possible as some protein residue may make multiple contacts of same subtype with the ligand, as is the case for the two distinct hydrogen bonds between **2b** and Met793. 32

4.14 Histograms of interaction frequency as a fraction of simulation time for each inhibitor and any residue with a frequency higher than 30%. A blank (grey) spot indicates no interaction was registered at this time step, while a colour indicates which kind of interaction took place. We also want to make clear that the interaction density of **1a** is less contiguous without it necessarily having half the interactions - it was merely sampled at half the rate of the other simulations. 33

4.16 All investigated inhibitors bond strongly to the backbone of Metionine793 lying in the hinge region, forming a bond from the N1 nitrogen on the scaffold to the amine group of the backbone. Depicted is *R*-**2b**, which forms an extra bond due to its NH heteroatom. Dashed yellow lines are hydrogen bonds; dashed teal lines indicate possible aromatic hydrogen bonds. 34

4.15 Space-filling CPK-model of compound *S*-**1a** (green) and the three hydrophobic residues of Ala743, Leu718 and Leu844 from frame 695 of the MD simulation of *S*-**1a** bound to EGFR, showing that the three hydrophobic residues form a cleft surrounding the inhibitor. 35

4.17 The high potency S-enantiomer of the methanol compounds form additional hydrogen bonds to Lys745 and Asp855 via the methanol while retaining a bond to Thr854 mediated by water, in contrast to the low-potency R-enantiomer which bonds to Thr854 via a direct hydrogen bond from methanol. Shown are representative binding modes of the R/S-enantiomers of **1b** and their interactions with Lys745, Asp855, and Thr854, and the surrounding solvent. The bridging water molecule is thickened for emphasis. Dashed lines show non-covalent bonds: yellow for hydrogen bonds, green for pi-cation bonds, and purple for ionic bonds. 36

4.18 The high potency R-enantiomer of the methyl thienopyrimidines **1a** and **3a**) forms a single bond that the low-potency S-enantiomer doesn't: a water-mediated hydrogen bond to Thr854. Shown are representative binding modes of the R/S-enantiomers of **1a** and their (non-)interactions with Lys745, Asp855, and Thr854, and the surrounding solvent. Dashed lines show non-covalent bonds: yellow for hydrogen bonds and green for pi-cation bonds. . 37

4.19 Unlike the methyl substituted thienopyrimidines, pyrrolopyrimidine **2b** (and furopyrimidine **2a**, not shown) does not displace the water bridge to Thr854; in fact, this bridge is maintained in both the phenyl up and phenyl down conformation. 38

4.20 In our exemplar structures, the thienopyrimidine scaffold is often farther away from the hinge compared to the furo- and pyrrolopyrimidine scaffold. Depicted from left to right are the scaffold and hinge region of *S*-**2a**, *S*-**2b** and *S*-**1a** with measurements showing the interatomic distance between the N3 nitrogen of the inhibitor and the amine nitrogen of Met793, as well as the distance between the heteroatom and the carbonyl group of Met793. 39

4.21 Simplified 2D schematic showing how chirality affects non-covalent interactions. Top: Steric clash with a methyl group (shown as dashed lines) prevents a water bridge to Thr854 in the *S*-enantiomer of methyl-substituted thienopyrimidine inhibitors, here **1a**, but not in the *R*-enantiomer. Bottom: Additional bond from methanol to Lys745 causes the *S* enantiomer to be more stable than the *R* enantiomer in methanol-substituted thienopyrimidines, here **1b**. 41

4.23 Top Glide poses of R-2b (green) and S-2b (pink) 42

4.25 A mockup of a possible binding mode of a dimethanol compound 44

List of Tables

2.1 Experimental IC_{50} values and computed difference in relative binding Gibbs free energy. . 11

4.1 Interaction frequency between the inhibitors and surrounding residues organized by interaction type. 34

Bibliography

[1] Linggi, B.; Carpenter, G. ErbB receptors: new insights on mechanisms and biology. *Trends in Cell Biology* **2006**, *16*, 649 – 656.

[2] Salomon, D. S.; Brandt, R.; Ciardiello, F.; Normanno, N. Epidermal growth factor-related peptides and their receptors in human malignancies. *Critical Reviews in Oncology/Hematology* **1995**, *19*, 183 – 232.

[3] Seshacharyulu, P.; Ponnusamy, M. P.; Haridas, D.; Jain, M.; Ganti, A. K.; Batra, S. K. Targeting the EGFR signaling pathway in cancer therapy. *Expert Opinion on Therapeutic Targets* **2012**, *16*, 15–31.

[4] Zwick, E.; Bange, J.; Ullrich, A. Receptor tyrosine kinases as targets for anticancer drugs. *Trends in Molecular Medicine* **2002**, *8*, 17 – 23.

[5] Grandis, J. R.; Sok, J. C. Signaling through the epidermal growth factor receptor during the development of malignancy. *Pharmacology and Therapeutics* **2004**, *102*, 37 – 46.

[6] Jamal-Hanjani, M.; Spicer, J. Epidermal Growth Factor Receptor Tyrosine Kinase Inhibitors in the Treatment of Epidermal Growth Factor Receptor–Mutant Non–Small Cell Lung Cancer Metastatic to the Brain. *Clinical Cancer Research* **2012**, *18*, 938–944.

[7] Food,; Administration, D. Drug Approval Package: Tarceva (Erlotinib) NDA #021743. 2004; `https://www.accessdata.fda.gov/drugsatfda_docs/nda/2004/21-743_Tarceva.cfm`.

[8] Kalman, B.; Szep, E.; Garzuly, F.; Post, D. E. Epidermal Growth Factor Receptor as a Therapeutic Target in Glioblastoma. *Neuromol Med* **2013**, *15*, 420–434.

[9] Kobayashi, S.; Boggon, T. J.; Dayaram, T.; Jänne, P. A.; Kocher, O.; Meyerson, M.; Johnson, B. E.; Eck, M. J.; Tenen, D. G.; Halmos, B. EGFR Mutation and Resistance of Non–Small-Cell Lung Cancer to Gefitinib. *New England Journal of Medicine* **2005**, *352*, 786–792, PMID: 15728811.

[10] Bugge, S.; Buene, A. F.; Jurisch-Yaksi, N.; Moen, I. U.; Skjønsfjell, E. M.; Sundby, E.; Hoff, B. H. Extended structure-activity study of thienopyrimidine-based EGFR inhibitors with evaluation of drug-like properties. *European Journal of Medicinal Chemistry* **2016**, *107*.

[11] Han, J.; Henriksen, S.; Nørsett, K. G.; Sundby, E.; Hoff, B. H. Balancing potency, metabolic stability and permeability in pyrrolopyrimidine-based EGFR inhibitors. *European Journal of Medicinal Chemistry* **2016**, *124*.

[12] Han, J. Investigation of Pyrrolo- and Furopyrimidines as Epidermal Growth Factor Receptor Tyrosine Kinase Inhibitors. Ph.D. thesis, Norwegian University of Science and Technology, 2016.

[13] Traxler, P.; Allegrini, P. R.; Brandt, R.; Brueggen, J.; Cozens, R.; Fabbro, D.; Grosios, K.; Lane, H. A.; McSheehy, P.; Mestan, J.; Meyer, T.; Tang, C.; Wartmann, M.; Wood, J.; Caravatti, G. AEE788: A Dual Family Epidermal Growth Factor Receptor/ErbB2 and Vascular Endothelial Growth Factor Receptor Tyrosine Kinase Inhibitor with Antitumor and Antiangiogenic Activity. *Cancer Research* **2004**, *64*, 4931–4941.

[14] Bugge, S.; Moen, I. U.; Sylte, K.-O. K.; Sundby, E.; Hoff, B. H. Truncated structures used in search for new lead compounds and in a retrospective analysis of thienopyrimidine-based EGFR inhibitors. *European Journal of Medicinal Chemistry* **2015**, *94*.

[15] IUPAC, In *Compendium of Chemical Terminology*, 2nd ed.; McNaught, A. D., Wilkinson, A., Eds.; Blackwell Scientific Publications: Oxford, 1997.

[16] Stamos, J.; Sliwkowski, M. X.; Eigenbrot, C. Structure of the Epidermal Growth Factor Receptor Kinase Domain Alone and in Complex with a 4-Anilinoquinazoline Inhibitor. *Journal of Biological Chemistry* **2002**, *277*, 46265–46272.

[17] Shan, Y.; Eastwood, M.; Zhang, X.; Kim, E.; Arkhipov, A.; Dror, R.; Jumper, J.; Kuriyan, J.; Shaw, D. Oncogenic Mutations Counteract Intrinsic Disorder in the EGFR Kinase and Promote Receptor Dimerization. *Cell* **2012**, *149*, 860 – 870.

[18] Dawson, J. P.; Berger, M. B.; Lin, C.-C.; Schlessinger, J.; Lemmon, M. A.; Ferguson, K. M. Epidermal Growth Factor Receptor Dimerization and Activation Require Ligand-Induced Conformational Changes in the Dimer Interface. *Molecular and Cellular Biology* **2005**, *25*, 7734–7742.

[19] Bose, R.; Zhang, X. The ErbB kinase domain: Structural perspectives into kinase activation and inhibition. *Experimental Cell Research* **2009**, *315*, 649 – 658, Invited Reviews: ErbB Receptors.

[20] Tummino, P. J.; Copeland, R. A. Residence Time of Receptor-Ligand Complexes and Its Effect on Biological Function. *Biochemistry* **2008**, *47*, 5481–5492, PMID: 18412369.

[21] Yung-Chi, C.; Prusoff, W. H. Relationship between the inhibition constant (KI) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. *Biochemical Pharmacology* **1973**, *22*, 3099 – 3108.

[22] Henriksen, M. A.; Høyvik, I.-M.; Sundby, E. Project report, NTNU.

[23] 2020-1, S. R. Induced Fit Docking Protocol 2020-1, Glide Version 8.6, Prime Version 3.4, Maestro Version 12.3, Jaguar Version 10.7, Desmond Version 6.1. 2020.

[24] Trinh, T.; Sundby, E.; Bugge, S.; Hoff, B. H. Paper from NTNU.

[25] Lockett, M. R.; Lange, H.; Breiten, B.; Heroux, A.; Sherman, W.; Rappoport, D.; Yau, P. O.; Snyder, P. W.; Whitesides, G. M. The Binding of Benzoarylsulfonamide Ligands to Human Carbonic Anhydrase is Insensitive to Formal Fluorination of the Ligand. *Angewandte Chemie International Edition* **2013**, *52*, 7714–7717.

[26] Ruvinsky, A. M.; Aloni, I.; Cappel, D.; Higgs, C.; Marshall, K.; Rotkiewicz, P.; Repasky, M.; Feher, V. A.; Feyfant, E.; Hessler, G.; Matter, H. The Role of Bridging Water and Hydrogen Bonding as Key Determinants of Noncovalent Protein–Carbohydrate Recognition. *ChemMedChem* **2018**, *13*, 2684–2693.

[27] Tomasi, J.; Persico, M. Molecular Interactions in Solution: An Overview of Methods Based on Continuous Distributions of the Solvent. *Chemical Reviews* **1994**, *94*, 2027–2094.

[28] Manzoni, V.; Lyra, M. L.; Coutinho, K.; Canuto, S. Comparison of polarizable continuum model and quantum mechanics/molecular mechanics solute electronic polarization: Study of the optical and magnetic properties of diazines in water. *The Journal of Chemical Physics* **2011**, *135*, 144103.

[29] Wissner, A. et al. 4-Anilino-6,7-dialkoxyquinoline-3-carbonitrile Inhibitors of Epidermal Growth Factor Receptor Kinase and Their Bioisosteric Relationship to the 4-Anilino-6,7-dialkoxyquinazoline Inhibitors. *Journal of Medicinal Chemistry* **2000**, *43*, 3244–3256.

[30] Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. Systematic Validation of Protein Force Fields against Experimental Data. *PLOS ONE* **2012**, *7*, 1–6.

[31] Hollingsworth, S. A.; Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **2018**, *99*, 1129 – 1143.

[32] Warshel, A.; Tao, H.; Fothergill, M.; Chu, Z.-T. Effective Methods for Estimation of Binding Energies in Computer-Aided Drug Design. *Israel Journal of Chemistry* **1994**, *34*, 253–256.

[33] Kollman, P. Free energy calculations: Applications to chemical and biochemical phenomena. *Chemical Reviews* **1993**, *93*, 2395–2417.

[34] Straatsma, T. P.; McCammon, J. A. Computational Alchemy. *Annual Review of Physical Chemistry* **1992**, *43*, 407–435.

[35] Beveridge, D. L.; DiCapua, F. M. Free Energy Via Molecular Simulation: Applications to Chemical and Biomolecular Systems. *Annual Review of Biophysics and Biophysical Chemistry* **1989**, *18*, 431–492, PMID: 2660832.

[36] Simonson, T.; Archontis, G.; Karplus, M. Free Energy Simulations Come of Age: Protein-Ligand Recognition. *Accounts of Chemical Research* **2002**, *35*, 430–437, PMID: 12069628.

[37] Deng, Y.; Roux, B. Computations of Standard Binding Free Energies with Molecular Dynamics Simulations. *The Journal of Physical Chemistry B* **2009**, *113*, 2234–2246, PMID: 19146384.

[38] Bash, P.; Singh, U.; Langridge, R.; Kollman, P. Free energy calculations by computer simulation. *Science* **1987**, *236*, 564–568.

[39] Schlitter, J.; Klähn, M. A new concise expression for the free energy of a reaction coordinate. *The Journal of Chemical Physics* **2003**, *118*, 2057–2060.

[40] Kästner, J.; Senn, H. M.; Thiel, S.; Otte, N.; Thiel, W. QM/MM Free-Energy Perturbation Compared to Thermodynamic Integration and Umbrella Sampling: Application to an Enzymatic Reaction. *Journal of Chemical Theory and Computation* **2006**, *2*, 452–461, PMID: 26626532.

[41] Torrie, G.; Valleau, J. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics* **1977**, *23*, 187 – 199.

[42] Merlitz, H.; Burghardt, B.; Wenzel, W. Application of the stochastic tunneling method to high throughput database screening. *Chemical Physics Letters* **2003**, *370*, 68 – 73.

[43] Gervasio, F. L.; Laio, A.; Parrinello, M. Flexible Docking in Solution Using Metadynamics. *Journal of the American Chemical Society* **2005**, *127*, 2600–2607, PMID: 15725015.

[44] Earl, D. J.; Deem, M. W. Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3910–3916.

[45] Isralewitz, B.; Gao, M.; Schulten, K. Steered molecular dynamics and mechanical functions of proteins. *Current Opinion in Structural Biology* **2001**, *11*, 224 – 230.

[46] Cvijović, D.; Klinowski, J. Taboo Search: An Approach to the Multiple Minima Problem. *Science* **1995**, *267*, 664–666.

[47] Nakajima, N.; Higo, J.; Kidera, A.; Nakamura, H. Flexible docking of a ligand peptide to a receptor protein by multicanonical molecular dynamics simulation. *Chemical Physics Letters* **1997**, *278*, 297 – 301.

[48] Abrams, C.; Bussi, G. Enhanced Sampling in Molecular Dynamics Using Metadynamics, Replica-Exchange, and Temperature-Acceleration. *Entropy* **2014**, *16*, 163–199.

[49] Bernardi, R. C.; Melo, M. C.; Schulten, K. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochimica et Biophysica Acta (BBA) - General Subjects* **2015**, *1850*, 872 – 877, Recent developments of molecular dynamics.

[50] Wang, J.; Dixon, R.; Kollman, P. A. Ranking ligand binding affinities with avidin: a molecular dynamics-based interaction energy study. *Proteins: Structure, Function, and Bioinformatics* **1999**, *34*, 69–81.

[51] Åqvist, J.; Medina, C.; Samuelsson, J.-E. A new method for predicting binding affinity in computer-aided drug design. *Protein Engineering, Design and Selection* **1994**, *7*, 385–391.

[52] Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Accounts of Chemical Research* **2000**, *33*, 889–897, PMID: 11123888.

[53] Rifai, E. A.; van Dijk, M.; Vermeulen, N. P. E.; Yanuar, A.; Geerke, D. P. A Comparative Linear Interaction Energy and MM/PBSA Study on SIRT1–Ligand Binding Free Energy Calculation. *Journal of Chemical Information and Modeling* **2019**, *59*, 4018–4033, PMID: 31461271.

[54] Pollack, V. A.; Savage, D. M.; Baker, D. A.; Tsaparikos, K. E.; Sloan, D. E.; Moyer, J. D.; Barbacci, E. G.; Pustilnik, L. R.; Smolarek, T. A.; Davis, J. A.; Vaidya, M. P.; Arnold, L. D.; Doty, J. L.; Iwata, K. K.; Morin, M. J. Inhibition of Epidermal Growth Factor Receptor-Associated Tyrosine Phosphorylation in Human Carcinomas with CP-358,774: Dynamics of Receptor Inhibition In Situ and Antitumor Effects in Athymic Mice. *Journal of Pharmacology and Experimental Therapeutics* **1999**, *291*, 739–748.

[55] Lamba, V.; Ghosh, I. New directions in targeting protein kinases: focusing upon true allosteric and bivalent inhibitors. *Current pharmaceutical design* **2012**, *18 20*, 2936–45.

[56] Yun, C.-H.; Boggon, T. J.; Li, Y.; Woo, M. S.; Greulich, H.; Meyerson, M.; Eck, M. J. Structures of Lung Cancer-Derived EGFR Mutants and Inhibitor Complexes: Mechanism of Activation and Insights into Differential Inhibitor Sensitivity. *Cancer Cell* **2007**, *11*, 217 – 227.

[57] Bugge, S.; Kaspersen, S. J.; Larsen, S.; Nonstad, U.; Bjørløy, G.; Sundby, E.; Hoff, B. H. Structure-activity study leading to identification of a highly active thienopyrimidine based EGFR inhibitor. *European Journal of Medicinal Chemistry* **2014**, *75*.

[58] Harder, E. et al. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *Journal of Chemical Theory and Computation* **2016**, *12*, 281–296.

[59] Godschalk, F.; Genheden, S.; Söderhjelm, P.; Ryde, U. Comparison of MM/GBSA calculations based on explicit and implicit solvent simulations. *Phys. Chem. Chem. Phys.* **2013**, *15*, 7731–7739.

[60] Zhu, Y.-L.; Beroza, P.; Artis, D. R. Including Explicit Water Molecules as Part of the Protein Structure in MM/PBSA Calculations. *Journal of Chemical Information and Modeling* **2014**, *54*, 462–469, PMID: 24432790.

[61] Desmond Molecular Dynamics System, version 2.3.

[62] Su, Y.; Gallicchio, E.; Das, K.; Arnold, E.; Levy, R. M. Linear Interaction Energy (LIE) Models for Ligand Binding in Implicit Solvent: Theory and Application to the Binding of NNRTIs to HIV-1 Reverse Transcriptase. *Journal of Chemical Theory and Computation* **2007**, *3*, 256–277, PMID: 26627170.

[63] Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1-2*, 19 – 25.

[64] Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry* **2005**, *26*, 1781–1802.

[65] Case, D. et al. AMBER 2020. **2020**,

[66] Brooks, B. R. et al. CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry* **2009**, *30*, 1545–1614.

[67] Bowers, K. J.; Chow, D. E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Shan, Y.; Shaw, D. E. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. SC '06: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing. 2006; pp 43–43.

[68] Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology* **2017**, *13*, 1–17.

[69] Brooks, B. R. et al. CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry* **2009**, *30*, 1545–1614.

[70] Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. An overview of the Amber biomolecular simulation package. *WIREs Computational Molecular Science* **2013**, *3*, 198–210.

[71] Jorgensen, W. L.; Tirado-Rives, J. 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* **1988**, *110*, 1657–1666.

[72] Roos, K.; Wu, C.; Damm, W.; Reboul, M.; Stevenson, J. M.; Lu, C.; Dahlgren, M. K.; Mondal, S.; Chen, W.; Wang, L.; Abel, R.; Friesner, R. A.; Harder, E. D. OPLS3e: Extending Force Field Coverage for Drug-Like Small Molecules. *Journal of Chemical Theory and Computation* **2019**, *15*, 1863–1874, PMID: 30768902.

[73] Tuckerman, M.; Berne, B. J.; Martyna, G. J. Reversible multiple time scale molecular dynamics. *The Journal of Chemical Physics* **1992**, *97*, 1990–2001.

[74] Frenkel, D.; Smit, B. *Understanding Molecular Simulation: from algorithms to applications*, 1st ed.; Academic Press, 1996.

[75] Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of Chemical Physics* **1984**, *81*, 511–519.

[76] Hoover, W. Canonical Dynamics: Equilibrium Phase-Space Distributions. *Phys. Rev. A: At., Mol., Opt. Phys.* **1985**, *31*, 1695.

[77] Martyna, G. J.; Tobias, D. J.; Klein, M. L. Constant pressure molecular dynamics algorithms. *The Journal of Chemical Physics* **1994**, *101*, 4177–4189.

[78] Hünenberger, P. H. In *Advanced Computer Simulation: Approaches for Soft Matter Sciences I*; Dr. Holm, C., Prof. Dr. Kremer, K., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2005; pp 105–149.

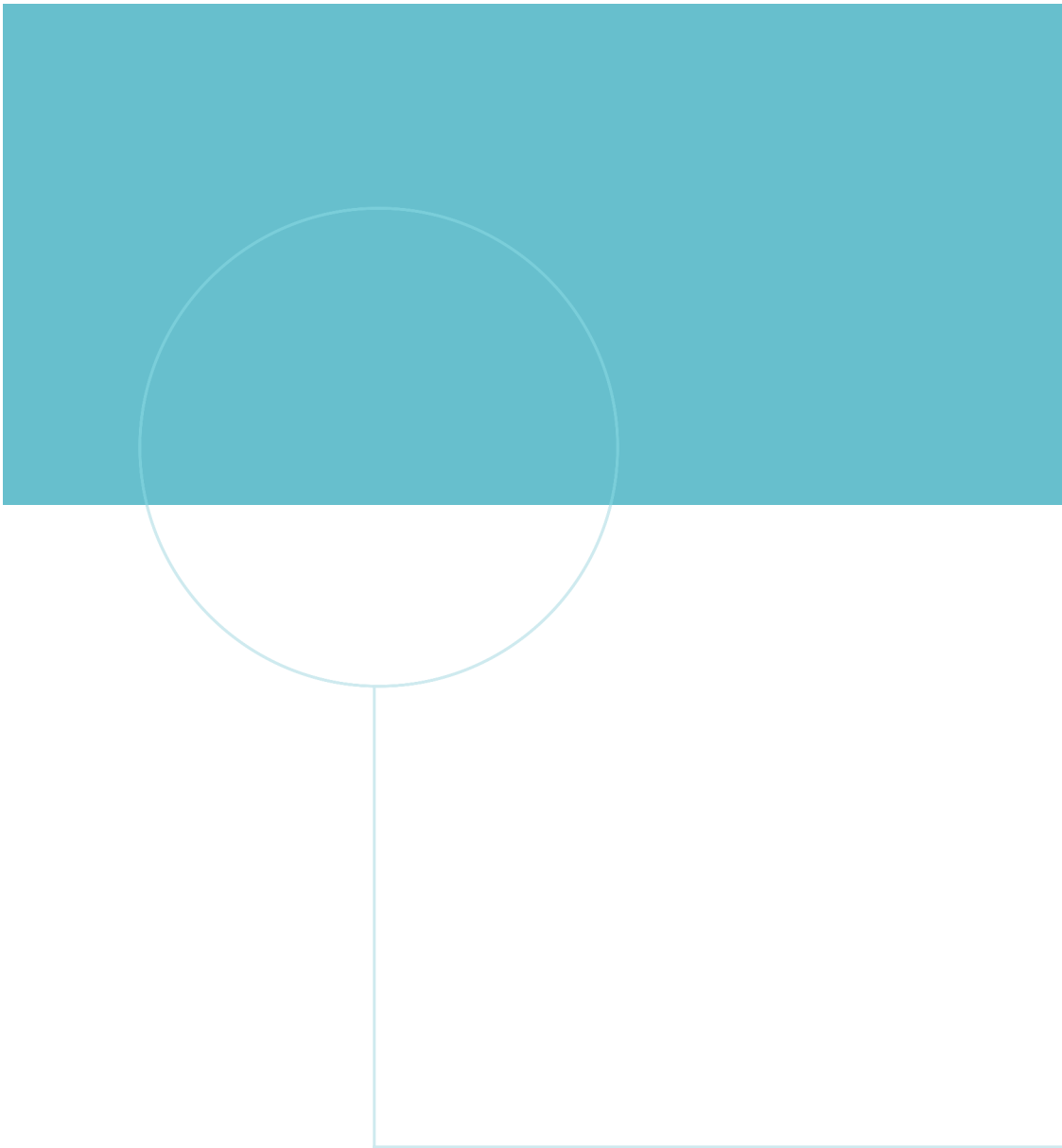
[79] R Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria, 2020.

[80] Frey, B. J.; Dueck, D. Clustering by Passing Messages Between Data Points. *Science* **2007**, *315*, 972–976.

[81] Yan, X. C.; Robertson, M. J.; Tirado-Rives, J.; Jorgensen, W. L. Improved Description of Sulfur Charge Anisotropy in OPLS Force Fields: Model Development and Parameterization. *The Journal of Physical Chemistry B* **2017**, *121*, 6626–6636, PMID: 28627890.

[82] Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A. Motifs for molecular recognition exploiting hydrophobic enclosure in protein–ligand binding. *Proceedings of the National Academy of Sciences* **2007**, *104*, 808–813.

[83] Li, Y.; Gao, Y.; Holloway, M. K.; Wang, R. Prediction of the Favorable Hydration Sites in a Protein Binding Pocket and Its Application to Scoring Function Formulation. *Journal of Chemical Information and Modeling* **0**, *0*, null, PMID: 32401510.



NTNU

Kunnskap for en bedre verden