University of
Nottingham
UK | CHINA | MALAYSIA

Molecular Simulation of Protein-Ligand Complexes

Ellen E. Guest, MSci.

Thesis submitted to the University of Nottingham for the degree of doctor of Philosophy

2021

Abstract

Computational methods provide important contributions to modern drug discovery projects. In this thesis, we discuss the insights into protein-ligand interactions afforded by methods such as molecular docking, molecular dynamics (MD) and alchemical free energy calculations, which expedite the process of lead compound design and optimisation. These methods are applied to two case studies of biomolecular systems of therapeutic interest. The targets of the studies are the integrin αvβ6 and the bromodomain-containing protein 4 (BRD4). As the accuracy of molecular mechanics based methods relies on the quality of the force field in which the potential energy is calculated from, we focus on developing force field parameters for a series of small molecule inhibitors of αvβ6. Parameters are then applied to MD and relative free energy perturbation (FEP) simulations. MD simulations highlight the importance of hydrogen bonds, metal chelate interactions and cation-π interactions between the compounds and αvβ6. FEP simulations predict relative binding affinities for these compounds with an average accuracy of 1.5 kcal mol^{-1}, when compared to experiment.

Initial protein structure and the inclusion of crystallographic water molecules can have an impact on the accuracy of computational predictions. To aid the selection of X-ray crystal structure of the first bromodomain (BD1) of BRD4 for the starting point of any *in silico* study, an analysis of the structures available in the Protein Data Bank was performed. To validate this analysis, molecular docking and absolute FEP simulations were employed. Docking showed that 82% of ligand poses were better predicted when including a network of water molecules in the binding site of BRD4-BD1. We also investigate how the alchemical perturbation methods, relative FEP and multisite lambda dynamics (MSλD) compare in the prediction of relative binding affinities when targeting BRD4-BD1. Although the accuracy of the two methods was very similar, an average of 0.6 kcal mol^{-1} from experiment for both, the computational demand of MSλD is significantly less, with 90% less simulation time required. Therefore, this study provides a foundation for the investigation of novel inhibitors of BRD4 using MSλD. Overall, the work presented in this thesis demonstrates the application of molecular docking, MD and predictions of binding free energy in drug discovery.

Publications

1. **Ellen E. Guest**, Luis F. Cervantes, Stephen D. Pickett, Charles L. Brooks III and Jonathan D. Hirst, "Alchemical Free Energy Methods Applied to Complexes of the First Bromodomain of BRD4", *J. Chem. Inf. Model*, 2022, **62**, 1458-1470.

2. **Ellen E. Guest**, Stephen D. Pickett and Jonathan D. Hirst, "Structural variation of protein-ligand complexes of the first bromodomain of BRD4", *Org. Biomol. Chem.*, 2021, **19**, 5632.

3. Francesco Segatta, David M. Rogers, Naomi T. Dyer, **Ellen E. Guest**, Zhuo Li, Hainam Do, Artur Nenov, Marco Garavelli and Jonathan D. Hirst, "Near-ultraviolet circular dichroism and two-dimensional spectroscopy of polypeptides", *Molecules*, 2021, **26**, 396.

4. **Ellen E. Guest**, Steven A. Oatley, Simon J. F. Macdonald and Jonathan D. Hirst, "Molecular simulation of αvβ6integrin inhibitors", *J. Chem. Inf. Model*, 2020, **11**, 5487-5498.

5. Sarah B. Jasim, Zhuo Li, **Ellen E. Guest** and Jonathan D. Hirst, "Dichrocalc: Improvements in computing protein circular dichroism spectroscopy in the near-ultraviolet", *J. Mol. Biol*, 2018, **15**, 2196-2202.

Acknowledgements

First, I would like to thank Prof. Jonathan Hirst. You have been an incredible supervisor and mentor to me since my Master's project in 2016. Thank you for all of your guidance, encouragement and for introducing me to the world of computational chemistry. Your support has led me to publish articles and present at conferences, something that I never thought would be achievable.

I am grateful to GSK for the opportunities they have given me during my PhD. Thank you to my second supervisors along the way, Simon Macdonald and Stephen Pickett, for many useful discussions. Thank you to the whole Theme 1 Prosperity Partnership team. Despite only meeting face to face a handful of times before COVID-19, I have valued our online meetings and the team spirit in which we have tackled the project. In particular, thank you to the computational chemistry team, Dr. Arnaldo Fernandes Da silva filho, Alexe Haywood and Dr. David Rogers. A huge thank you, as well, to Prof. Charles Brooks III and Luis Cervantes at the University of Michigan for all of your help with the MSλD calculations.

I am so grateful to everyone in the computational chemistry department. A big thank you to Steve Oatley, Stephen Mason, Steve Skowron, Adam Fouda and Josh Baptiste. You have all been there for me since day one in the comp chem offices. Steve O, thank you for your endless help and wisdom on all things CADD. Stephen (Darren) and Steve S, thank you for keeping me sane (ish) and the constant entertainment (and distractions), especially when it included salsa dancing around the office. Thank you to

Abi Miller and Alexe as well, for being my comp chem girls.

A massive thank you goes to Pritesh Tailor. You basically taught me how to be a scientist and, most importantly, how to put a decent presentation together. I would not be where I am today without the help you gave me.

My PhD experience wouldn't have been the same without my involvement with the Women in Chemistry group at UoN or the Pint of Science festival. Being surrounded by such inspirational groups has been a great joy and I am grateful to everyone involved.

Finally, I would like to say thank you to all of my loved ones. Grace Belshaw, Lizzie Killalea and Chloe Peach - I am so proud of us! Thank you for being amazing friends during the whole of my Nottingham experience. Dr. Peach, I think we would all agree that you are such an inspiration, your energy and positivity has been invaluable. Lizzie, you have supported me throughout undergrad and the PhD years, I am so grateful to you. Graco, thank you for being literally the best, I can't wait for our adventures, and your rational advice, to continue at the EPSRC. A very special mention goes to Max Astle (sorry - Dr Max Astle). Not only are you an incredible role model and scientist, you are also a pretty decent squash partner.

To my mom, dad and brother, you are all amazing. Thank you for all of your help, faith and encouragement. Mom, I think it is time that I stop claiming that I'm not very good at writing. Thank you for all of the love and support that all three of you have given me, which has ultimately led to this thesis.

List of Abbreviations and Symbols

ΔG Gibbs free energy.

ΔG_{bind} Gibbs free energy of ligand binding.

τ Residence time.

r_s Spearman correlation.

$\boldsymbol{k_{off}}$ Ligand dissociation rate constant.

$\boldsymbol{k_{on}}$ Ligand association rate constant.

ABFE Absolute binding free energy.

ADMIDAS Adjacent to MIDAS.

ALF Adaptive landscape flattening.

BAR Bennett acceptance ratio.

BD1 Bromodomain 1.

BET Bromodomain and extra-terminal domain.

BPTI Bovine pancreatic trypsin inhibitor.

BRD4 Bromodomain-containing protein 4.

BRDT Bromodomain testis-specific protein.

CADD Computer-aided drug design.

CCDC Cambridge Crystallographic Data Centre.

COVID-19 Coronavirus Disease of 2019.

DFT Density functional theory.

DHFR Dihydrofolate reductase.

ESMACS Enhanced sampling of molecular dynamics with approximation of continuum solvent.

FEP Free energy perturbation.

GPCRs G protein-coupled receptors.

GPU Graphics processing unit.

GSK GlaxoSmithKline.

HTS High throughput screening.

IC$_{50}$ Half-maximal inhibitory concentration.

IPF Idiopathic pulmonary fibrosis.

MBAR Multistate Bennett's acceptance ratio.

MCS Maximum common substructure.

MCSA Monte-Carlo Simulated Annealing.

MD Molecular dynamics.

MIDAS Metal ion-dependent adhesion site.

MM Molecular mechanics.

MM-GBSA MM-generalised Born surface area.

MM-PBSA MM Poisson-Boltzman surface area.

MSλD Multi-site λ-dynamics.

NMR Nuclear magnetic resonance.

PBC Periodic boundary conditions.

PDB Protein Data Bank.

PDE2A Phosphodiesterase-2A.

PES Potential energy surface.

PME Particle mesh Ewald.

QM Quantum mechanics.

QSAR Quantitative structure-activity relationship.

RBFE Relative binding free energy.

RMSD Root-mean-square deviation.

SAR Structure activity relationship.

SBDD Structure-based drug design.

SMILES Simplified molecular-input line-entry system.

SMIRNOFF SMIRKS Native Open Force Field.

SPC Simple point charge.

TGF-β1 Transforming growth factor β1.

THQ Tetrahydroquinoline.

TI Thermodynamic integration.

TIES Thermodynamic integration with enhanced sampling.

TIPS Transferable intermolecular potential surface.

UB Urey-Bradley function.

UoN University of Nottingham.

vdW van der Waals.

VMD Visual Molecular Dynamics.

VS Virtual screening.

Contents

Chapter 1

Drug Discovery

1.1 The drug development process

Modern drug discovery has a hugely positive impact on the quality and longevity of human life.[1,2] A topical example includes dexamethasone, the first drug to show life-saving efficacy in patients infected with COVID-19.[3] The repurposing of dexamethasone, a drug to treat skin diseases and severe allergies among other things,[4] has been shown to reduce the number of COVID-19 related deaths by 35% in patients who require mechanical ventilation.[5] However, developing a new pharmaceutical product is a time consuming and expensive task. Figure 1.1 shows the stages of the drug discovery process. In general, the stages become more resource and expense intensive as they progress. So clear criteria must be met before progression along the discovery pipeline.[6] In the discovery stage, a gene or protein that plays a significant role in a disease is identified; this is the target. Following evaluation of a target for its therapeutic potential and druggability, the

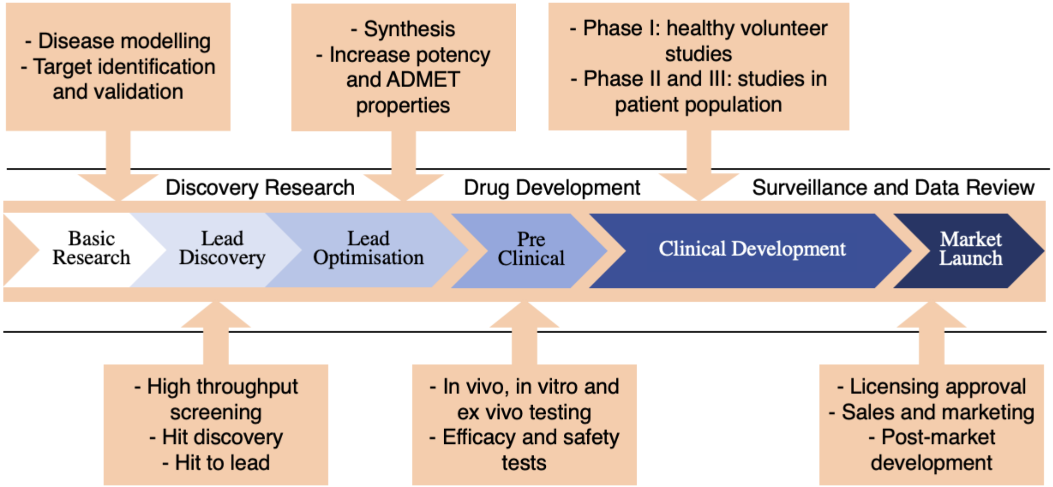
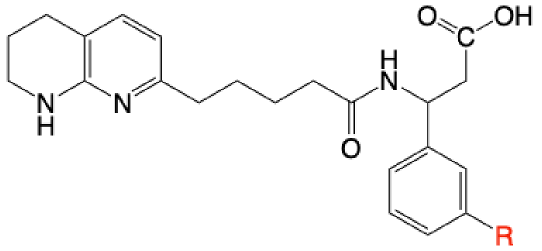

- Disease modelling
- Target identification and validation
- Synthesis
- Increase potency and ADMET properties
- Phase I: healthy volunteer studies
- Phase II and III: studies in patient population
Discovery Research
Drug Development
Surveillance and Data Review
Basic Research
Lead Discovery
Lead Optimisation
Pre Clinical
Clinical Development
Market Launch
- High throughput screening
- Hit discovery
- Hit to lead
- In vivo, in vitro and ex vivo testing
- Efficacy and safety tests
- Licensing approval
- Sales and marketing
- Post-market development

Figure 1.1: A typical drug discovery and development pipeline.

hit discovery process involves assay development and screening to search for a compound that binds to the target and has the desired effect. High throughput screening (HTS) is utilised to automate the screening of large chemical libraries for activity against the target. In the hit to lead process, a small number of compounds from HTS are taken forward for evaluation and these become lead compounds. Establishing which chemical series have the potential to become a drug candidate is an important decision as considerable amounts of synthetic resources are needed in lead optimisation. Lead optimisation involves small modifications being made to the compound to maximise potency against the target and optimise properties for the route of administration and target location. Figure 1.2 shows an example of lead optimisation of a scaffold to inhibit integrin $\alpha v\beta 6$, a protein linked to the initiation of idiopathic pulmonary fibrosis (IPF).[7] The R substituent on the aryl ring is modified to increase the potency of the compound towards the target.[8] Potency is measured by pIC_{50}, which is the negative log of the half-maximal inhibitory concentration (IC_{50}) and represents the concentration of



R	pIC_{50}
H	5.7
F	6.1
Cl	6.6
CH_3	6.4
OMe	6.5
(S)-CF_3	7.1

Figure 1.2: Activity[8] of aryl substituted derivatives of a lead scaffold in $\alpha v\beta 6$ integrin cell adhesion assays. All compounds are racemic unless specified.

the drug that is required to achieve 50% reduction in activity of the target. A pIC_{50} of $\geq$ 8 is desirable in lead discovery. However, it is important to consider the associated error of 0.3 for experimental pIC_{50} measurements, meaning that a least four of the compounds in Figure 1.2 are equivalent in potency.

There are opportunities to be smarter about the selection of compounds that are synthesised and tested before and during lead optimisation. The search for maximum potency, facilitated by HTS, often neglects the need for optimal physicochemical properties to make a compound drug-like.[9] Lipinski's rule of five[10] sets out a series of characteristics for an oral drug molecule and should be applied in the selection of lead compounds. Despite this, a recent appraisal of these rules considers that they should be used only as guidance, with room for compromise.[9] This is because some properties, such as molecular weight may not play such an important role, while lipophilicity control is the most important principle. Computational methods such as free energy calculations present a way to minimise the number of compounds made in the laboratory, while also giving synthetic chemists the confidence to embark on novel and often challenging syntheses.

Once a suitable drug candidate is designed, preclinical research is conducted to test the efficacy and safety of the compound before it is tested in people. On average, it takes 4.5 years to get to this stage.[11] The difficulty and cost of synthesis can often become a challenge here, as scaling up the synthesis can have an impact on the commercial viability of the project. If a compound passes preclinical trials, it enters clinical trials. Phase I of clinical trials involves healthy volunteer studies, where the pharmacokinetics, absorption and metabolic effects on the body are tested, as well as testing for a safe dosage range and side effects. Phases II and III involve studies in patient populations, where the drug is administered to up to thousands of volunteers who have the disease or condition. Once the full story of a compound shows evidence that it is safe and effective for its intended use, the compound is submitted for licensing approval, and once approved, it becomes available on the market. Over the last decade, the time taken for a successful drug discovery project was, on average, 8.7 years (±3.8) and cost around £1.15 billion.[12,13] It is estimated that five out of 40,000 compounds tested in preclinical trials reach human testing and only one in five compounds that reach clinical trials are approved.[14] Furthermore, only three out of ten drugs that make it to the clinic recover their capital investment.[15] Therefore, new strategies must be employed to improve this process. Computational methods present a solution for making the timeline of drug discovery smarter, quicker and cheaper. Each stage of the drug discovery process presents its own challenges; within this thesis, we focus on the discovery stages.

1.2 Computer Aided Drug Discovery

In the first section we have described how the search for a novel compound that balances biological activity and drug like properties is a challenging, time consuming and expensive task. To expedite and facilitate this process, computational approaches are now common place in both searching for a starting point in hit discovery and making rational decisions about chemical modifications to improve a compound's profile in lead optimisation. The term computer-aided drug design (CADD) has been adopted for the use of computers in drug discovery. To appreciate CADD and the modelling of protein-ligand interactions, it is relevant to cover the concept of structure-based design.

1.2.1 Structure-based drug design

Biomolecules such as proteins play a critical role in disease progression by communicating through protein-protein interactions or protein-nucleic interactions. Signalling events or changes in metabolic processes, as a result of these interactions, can lead to disease.[16] Therefore the design of a compound that competitively binds to an active site within a target, to stimulate or block its activity, is necessary. X-ray crystal structures and NMR structures of biomolecules not only provide insight on the mechanism of action of how they function, they also give understanding of the specific interactions to target for efficient binding. This enables the design of small changes that can be made to a compound that lead to increased activity and selectivity. Improving the physicochemical properties without compromising these features is also important. Selectivity can also be tuned by using structural information of anti-targets. Atomistic detail of other relevant biomolecules aids the design of a compound to reduce off-target interactions and improve the safety and pharmacokinetic profile of the compound.[17] The use of protein structure to design competitive binders has become known as structure-based drug design (SBDD) and is an established analysis used in drug discovery.[6]

The Protein Data Bank (PDB) contains a wide variety of macromolecular structure data.[18] Since it was established in 1971, the number of structures available in the PDB has grown each year (Figure 1.3). Structures include apoproteins, proteins bound with endogenous substrates and biomolecules that have more than one structure available, each bound with different inhibitors.[19,20] The first protein crystal structures around the 1960s, myoglobin, haemoglobin and lysozome, showed that it is possible to understand protein function through structure and lay the foundation for SBDD.[21–23] An early example of SBDD involves the work of Matthews et al.,[24,25] where new inhibitors of dihydrofolate reductase (DHFR) were found based on the structural understanding of protein-ligand interactions between DHFR and methotrexate. Since then, the development and advancement of CADD means that modern SBDD projects almost always involve the use of chem-informatics. For example, the improving predictive power of free energy calculations aids the calculation of binding affinities between protein and ligands.[26,27] These methods allow lead optimisation type questions to be answered, without synthetic expense. A number of computational methods that are used in contemporary drug discovery projects are explored in the

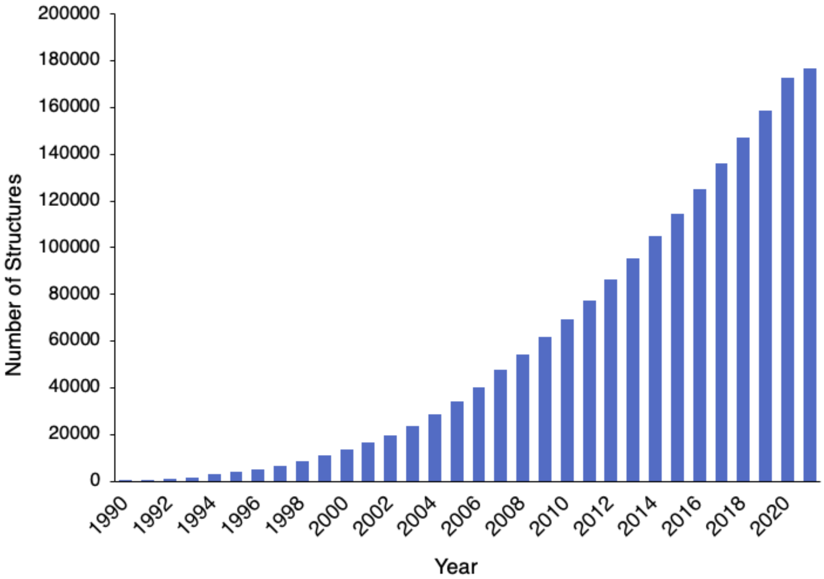


Figure 1.3: Yearly growth of the number of structures available in the RCSB Protein Data Bank since 1990. [Data taken from http://www.rcsb.org (April 2021).]

following sections.

1.2.2 Molecular docking

Although HTS presents an efficient way to explore chemical space, especially with recent improvements including target-focused libraries[28] and the recognition of privileged scaffolds,[29] it is an expensive technique and often inaccessible in academia without the help of an industrial collaborator. In contrast, virtual screening (VS) is a more accessible way in which libraries of small molecules can be searched for compounds that bind to a therapeutic target. VS is the in silico evaluation of large compound libraries to rank their viability to bind to the target and meet the physicochemical requirements necessary for the project. Molecular docking is the most widely used validation method for structure-based VS. Starting from a structure of the target, which could be an experimental structure or obtained through homology modelling, compound binding is simulated and a scoring function is used to estimate the binding affinity. This method assesses which of the large database of compounds will bind favourably to the target and can generate new ideas for possible interactions. Figure 1.4 demonstrates the molecular docking of an inhibitor of the bromodomain 1 (BD1) of bromodomain-containing protein 4 (BRD4). The method consists of two components, a search algorithm and a scoring function.[30] The search algorithm is responsible for searching different poses and conformations of a ligand within the active site. The scoring function provides a quantitative estimation of the binding energetics, delineates correct poses from incorrect poses and ranks different compounds, which is a useful tool for selecting compounds for further exploration. However, the accuracy of scoring functions remains a challenge in molecular docking. It is estimated that the poses of compounds that are known to bind to a target are successfully predicted 80% of the time.[6] In contrast, limitations to the scoring function prohibits the relative ranking of different molecules, leading to many false positives when trying to identify which ligands bind into a particular binding site. A more comprehensive explanation of the components of molecular docking is provided in the next chapter.

Early procedures employed a rigid docking method, where fixed structures of the ligand and receptor were used. Alberg et al.[32] designed a novel bridging compound (TCsA), which binds to cyclophilin A and FK506-binding protein 12 simultaneously, through searching in a six-dimensional rotational and translational space so the ligand would fit in the binding site.

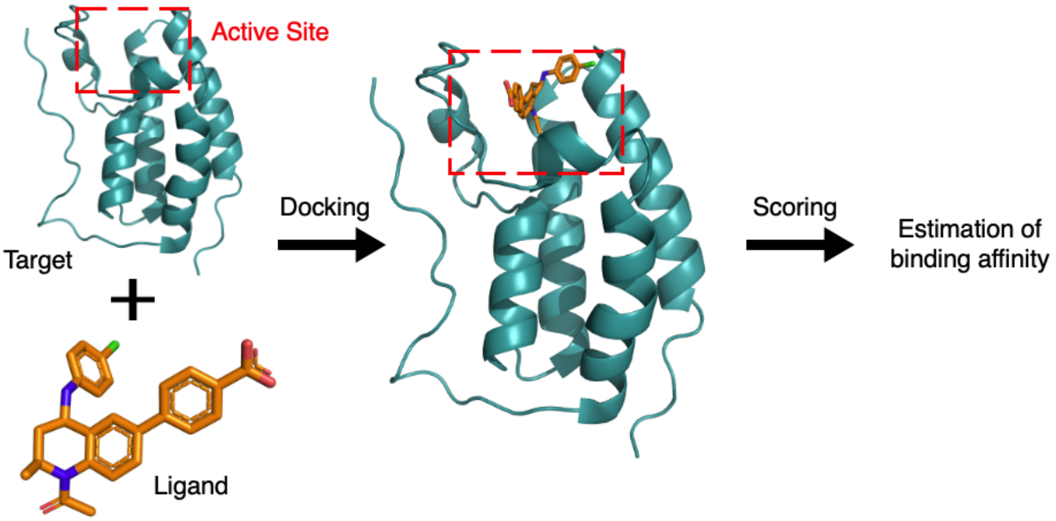


Figure 1.4: An example of molecular docking. I-BET726[31] is docked into its original crystal structure of BRD4-BD1 (PDB: 4BJX).

TCsA was developed as a lead compound for an immunosuppressive agent and provided an early understanding of the principles of SBDD, demonstrating the merits of rigid docking. However, this method does not take into account ligand or protein flexibility. Proteins are highly dynamic and possess inherent flexibility so that they can adapt to form interactions and achieve their function. Therefore, to understand how a compound binds, it is important to recognise the different conformations of an active site. Furthermore, the entropy loss and changes in internal energy of a flexible ligand upon binding affect the binding affinity, which is not reflected in the docking score of a rigid docking protocol.[33] These issues were first recognised by Koshland's "induced fit" theory[34,35] which stated that the ligand and receptor should be treated as flexible during docking as the protein is continually reshaped by interactions with the ligand as it binds. To accommodate this, many molecular docking programs have been developed over the years and flexible ligand and flexible receptor docking procedures are now common practice.

Flexible ligand docking involves the use of a flexible ligand and a rigid receptor. This is more accurate than rigid docking and is implemented in many molecular docking programs such as Flex X,[36] OpenEye FRED[37] and Dock.[38] However, side chain flexibility also plays an important role in ligand binding, as changes in side chain conformation allow the receptor to alter its binding site according to the orientation of the ligand.[33] One way to account for different protein conformations is to use an ensemble of target structures that reflect different binding site conformations. These could be experimental or modelled structures. Using this method, Dayam and Neamati[39] successfully predicted the bioactive conformations of S-1360, which was one of the first HIV-integrase inhibitors to enter clinical trials. Molecular docking programs have also been developed to model side chain flexibility when only using one crystal structure. These programs include Gold,[40] Glide,[41] Autodock Vina[42] and MedusaDock.[43] There are many computer-based drug discovery projects that have successfully used these programs.[44,45] One example includes work by Kumari et al.[46] who used quantitative structure-activity relationship (QSAR) models and molecular docking to identify the structural requirements for the inhibition of PfM18AAP, an important drug target for the treatment of malaria. Although flexible receptor docking can improve the accuracy of docking, it is more computationally demanding than rigid docking.[47] However, with the development of graphics processing unit (GPU)-accelerated docking,[48] it is becoming a more viable and attractive option for VS and early drug development projects.

Protein flexibility is not the only recent development that has resulted in the increased accuracy, reliability and efficiency of molecular docking. Other areas of improvements include the consideration of solvent, fragment docking, nonlinear scoring functions and machine-learning approaches.[49–52] Crystallographic water molecules can be important in molecular docking. Water molecules can form strong bonds to the receptor, especially in its active site. A challenge remains in knowing if it is thermodynamically favourable to displace water molecules or if the molecules improve the stability of ligand binding. The displacement of a crystallographic water molecule is associated with a favourable gain in entropy, with the release of a well-ordered molecule into the bulk solvent. However, the process can also cause a loss in enthalpy.[53] Furthermore, water molecules can increase binding affinity through mediated hydrogen bonds with the ligand and water networks throughout receptor cavities. Analysis, performed by Klebe,[54] revealed that in approximately 65% of several thousand complex crystal structures, at least one crystallographic water molecule is involved in ligand binding. Additionally, a systematic study[55] on the inclusion of bound water molecules into the docking program AutoDock4 showed an improvement in docking performance across 18 different protein-ligand systems. A further improvement was seen when water and side chain flexibility were considered. This study illustrates the importance of a proper treatment of water molecules in molecular docking.

Artificial intelligence is a technology that encompasses a set of computational algorithms that allow machines and computers to simulate human cognitive abilities such as learning and problem-solving. Machine learning and deep learning are two sub-fields of artificial intelligence that show immense promise in enhancing the efficiency and accuracy of molecular docking for VS.[56] Kinnings et al.[57] improved the accuracy of docking scores by addressing the incorrect assumption that individual interactions contribute towards binding affinity in an additive manner. A machine learning approach was used to capture the nonlinear cooperative features of noncovalent interactions. Furthermore, a recent study by Gentile et al.[58] uses deep learning (Figure 1.5) as a way to rapidly dock billions of compounds, while still maintaining accuracy. The approach uses QSAR models trained on a small subset of docking scores to predict the scores for a much larger database of compounds.

Molecular docking is a helpful tool in areas of drug development besides VS. For example, in drug repurposing, which is an efficient strategy for identifying new uses for existing drugs that are outside the scope of their original therapeutic use.[59] There are many examples of how molecular docking has aided the understanding of how a drug can be repurposed for the treatment of COVID-19.[60–63] Molecular docking is also commonly used as a basis for further computational investigations. In the absence of a protein-ligand crystal structure, molecular docking is often used as a precursor to molecular dynamics (MD) simulations and calculations of free energy, which further develop our understanding of the dynamic nature of protein-ligand interactions.

1.2.3 Molecular dynamics simulations

MD simulations give additional insight into the structural, dynamic and thermodynamic properties of a molecular system, which cannot be gained

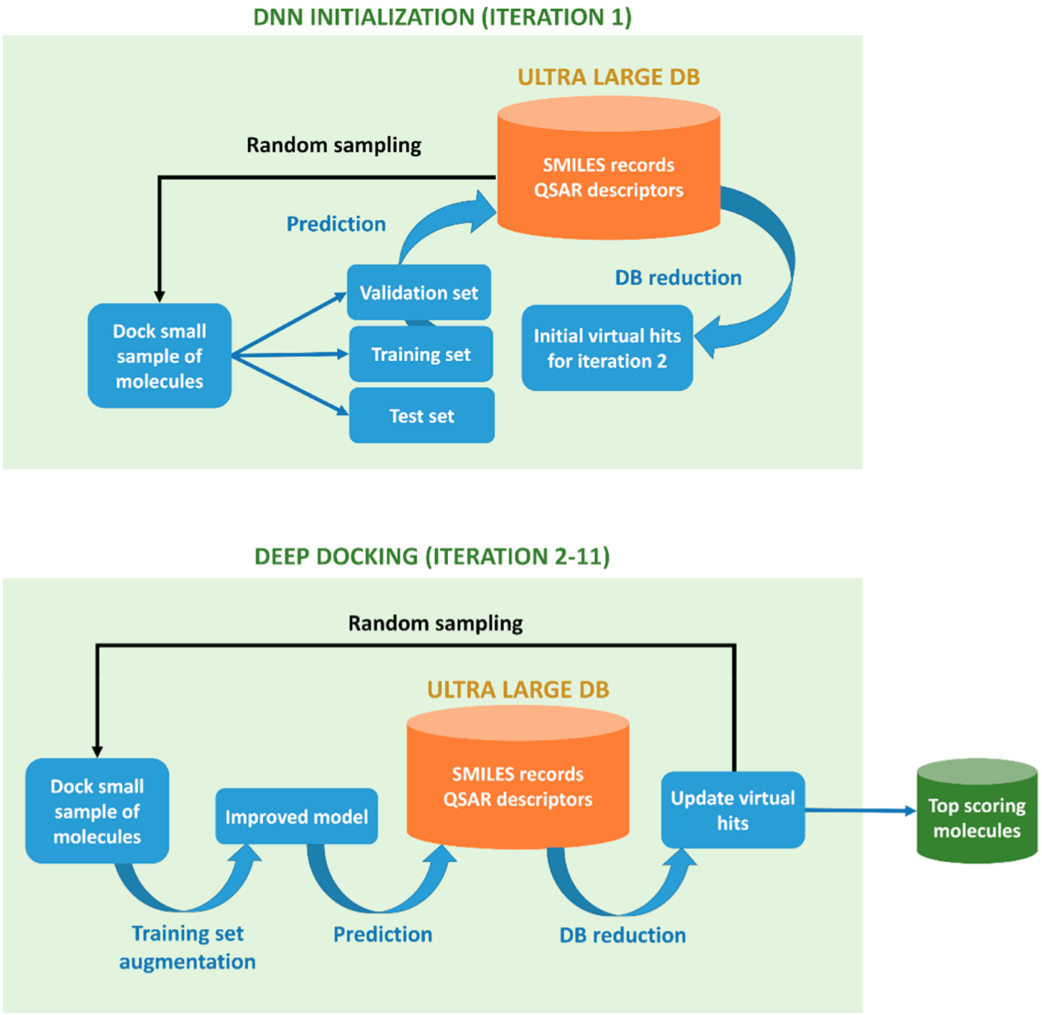


Figure 1.5: Deep learning approach to molecular developed by Gentile et al. For the first iteration of the model, a small number of compounds are extracted from a large database and docked to the target. The docking scores from the sample compounds are used to build a QSAR deep model. Virtual hits generated from this model are then used to start iteration two. From iteration two onward, the deep learning model gradually improves by augmenting the training set with randomly sampled virtual hits from the previous iteration. [Reproduced with permission from reference 58.]

from a static X-ray crystal structure or docked model.[64] In this context, a molecular system usually consists of a solute, typically a protein, surrounded by a solvent such as water. Newton's classical laws of motion are solved to show how atom positions vary with time, where the forces acting on the atoms are calculated using empirical potential energy functions (force fields). MD trajectories can show the small fluctuations of protein conformation within an energy minimum, as well as larger conformational changes as structures cross energy barriers to other local minima. MD simulations of protein-ligand complexes can provide important information on the dynamic character of an active site and the mechanisms responsible for ligand recognition, therefore guiding the choice of the best compounds for further drug development. This method also facilitates the evaluation of binding energetics and kinetics of protein-ligand interactions. Binding kinetics can have important pharmacological implications as ligand binding/unbinding rates have been found to be an accurate predictor of drug activity, in comparison with selectivity.[65,66]

The earliest example of a simulation of a biomolecule was published in 1977 where McCammon et al.[67] explored the dynamics of bovine pancreatic trypsin inhibitor (BPTI). Figure 1.6 shows the structure of BPTI, a protein consisting of 58 residues. In the figure, a solvated system is illustrated. However, the first BPTI simulation was performed in vacuum. Despite this, and it lasting for only 9.2 ps, this study is seen as instrumental in our understanding of proteins as flexible structures.[68] Since then, MD simulations have become a popular method to model proteins and a large number of software packages for MD of biomolecules have been developed, for example, CHARMM,[69] GROMACS,[70] AMBER[71] and NAMD.[72] MD simulations are significantly more computationally demanding compared to molecular docking. However, the recent development of computer hardware and software, especially GPU acceleration,[73–75] means it is now possible to model systems of considerable sizes and to achieve microseconds of simulation in

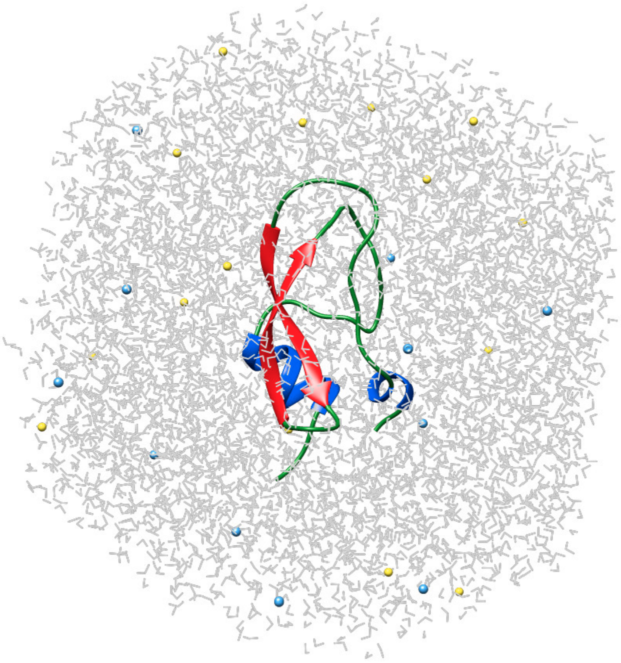


Figure 1.6: Simulation box of solvated BPTI. Yellow and blue spheres correspond to KCl counter ions. Water molecules are shown as lines.

realistic timescales. A recent example includes a study by Jung et al.,[76] who performed one of the largest and first atomic-scale simulations of an entire gene, which is composed of one billion atoms.

Although crystallographic structures are a good way to investigate protein-ligand complexes, MD simulations provide a more thorough evaluation of the interactions of a hit compound with a target, which leads to the rational design of more effective inhibitors. For example, selective inhibitors of phosphodiesterase-2A (PDE2A) are potential therapeutic targets for the treatment of Alzheimer's disease and pulmonary hypertension. Zhang et al.[77] used a combination of VS, molecular docking and MD simulations to design a novel compound, which inhibits PDE2A with high affinity. MD simulations showed additional interactions formed by the compound with the active site, compared to common interactions formed by previous inhibitors, which guided the structural modification of the hit compound. A further example of how understanding the binding site of a target, through MD, can lead to new strategies for inhibitor design is reported by Durrant et al.,[78] who performed the first all-atom simulation of the influenza virus. MD simulations were used to quantify the kinetics of the transition between the open and closed conformation of the active site, providing important insight into how to develop anti-influenza therapeutics.

To draw meaningful conclusions from MD simulations, it is important to have an adequate sampling of conformational space. Enhanced sampling methods, such as replica-exchange,[79] metadynamics[80] and simulated annealing,[81] enable the simulation of time consuming processes, which are not always achievable to model using standard MD. For example, the escape from non-relevant conformations with high barriers along the potential energy surface (PES), protein folding and ligand binding events.[82] The ability to model ligand binding and unbinding, provides information on the binding kinetics. Binding kinetics are characterised by the association rate constant (k_{on}) and the dissociation rate constant (k_{off}), or residence time (τ).[83] High residence times mean the compound interacts with the target for a long time and therefore has increased physiological effects. Furthermore, compounds with high k_{on} rates are effective competitors as they have the potential to bind faster than other ligands. However, it should be noted that this is only the case under certain circumstances when in vivo. There can be additional factors to consider such as diffusion rates and the binding mechanism.[84] Nevertheless, the ability to design compounds with desirable k_{on} and k_{off} rates, can lead to the design of more effective and safer drugs. An example of calculating binding rates through enhanced sampling is a study by Gobbo et al.[85] The combination of an electrostatic-like collective variable with adiabatic bias MD achieved a good agreement between computational and experimental measures for a series of glycogen synthase kinase 3 beta inhibitors.

A limitation of MD simulations is that they cannot be used to model the finer details of chemical reactions, as they cannot properly handle bond-forming and breaking events. This can hinder the study of enzyme activity, for example, as enzymes create chemical reactions in the body such as destroying toxins and breaking down food particles during digestion. However, changes in bonding can be studied using quantum mechanics (QM). QM is considerably more computationally demanding than molecular mechanics (MM) methods such as MD and it is unrealistic to model an entire biomolecule such as a protein or enzyme at the quantum level. Therefore, hybrid QM/MM methods present a solution. In QM/MM, the active site, including all residues and chemical groups that play a role in the reaction, are modelled by QM. The remaining protein and solvent are modelled by MM, most commonly using MD.[86] Figure 1.7 shows an example of a QM/MM setup.[87] QM/MM can also aid the development of drugs. *Staphylococcus aureus* is a bacterium that can cause several life-threatening diseases such as pneumonia, meningitis and toxic shock syndrome.[88] Its antibiotic resistance is linked to a gene, fmtA. Dalal et al.[89] used a combination of VS, molecular docking, MD, MM-generalised Born surface area (MM-GBSA) and QM/MM methods to develop compounds that target fmtA. MM-GBSA (more on this method in the next section) and thermodynamic results from

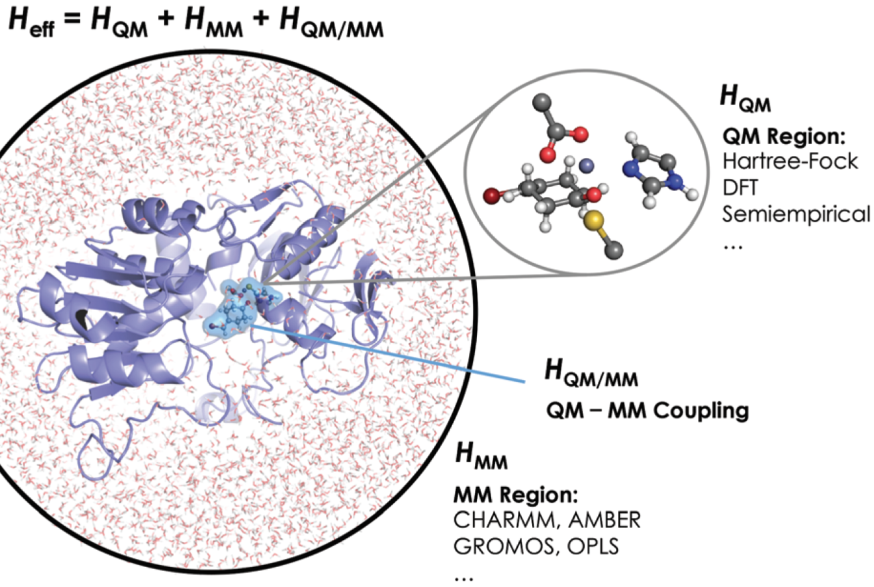


Figure 1.7: An example representation of a QM/MM biocatalytic system. [Reproduced with permission from reference 88.]

QM/MM revealed active site residues to target to form stable protein-ligand complexes.

MD simulations facilitate the calculation of protein-ligand binding free energies. These types of calculations have a considerable impact on the hit to lead and lead optimisation stages of drug development, as they provide more accurate estimations of compound activity compared to molecular docking. Free energy calculations are discussed in detail in the next section.

1.2.4 Free energy calculations

Designing a compound that binds competitively and strongly is crucial in drug development. The amount and types of interactions between a protein and a ligand are a key component of activity, and these interactions can be quantified by the free energy of binding (ΔG_{bind}). Conventional free energy changes (ΔG) describe the thermodynamic and kinetic properties of a system and are representations of the energy released or required for a chemical process. Therefore, ΔG can be used as a measure of the stability of a system and ΔG_{bind} relates specifically to the stability of a protein-ligand complex or ligand binding affinity. Free energy calculations estimate ΔG_{bind} based on the principles of statistical thermodynamics. These calculations are often based on MD simulations and are computationally more expensive than traditional scoring methods in molecular docking. However, unlike docking, free energy calculations account for the energetic and entropic effects of ligand binding and produce significantly more accurate results.

There are many different methods, which range in computational expense, to approximate ligand ΔG_{bind}.[64,90] Less computational demanding methods include the MM Poisson-Boltzmann Surface Area (MM-PBSA)[91] and MM-GBSA approaches.[92] In MM-PB(GB)SA methods, only the bound and unbound states of the system are used to estimate binding free energy, compared to using information on the pathway connecting the two states. Therefore, these are classed as end-point methods. The balance between computational efficiency and accuracy make MM-PB(GB)SA attractive methods in SBDD.[93] Early examples of MM-PB(GB)SA date back to the early 2000s.[91,92,94] However, a more recent example is reported by Arba et al.[95] who, based on molecular docking, MD and MM-PBSA, developed a compound with high affinity for CDK2, a target for anticancer treatments.

More rigorous, and often more accurate, free energy methods include alchemical approaches such as free energy perturbation (FEP),[96,97] Bennett's Acceptance Ratio (BAR)[98] and thermodynamic integration (TI).[99] These methods follow the path from the initial state to the final state of the system through alchemical changes of the energy function during an MD simulation. FEP is one of the earliest alchemical methods, with the first example in 1985 describing the calculation of the relative hydration free energy of ethane and methanol.[100] Since then, FEP has been employed for a number of uses, including the calculation of solvation free energies and most notably, ligand binding affinities.[101] There are two different strategies that are used within FEP, absolute binding free energy (ABFE) and relative binding free energy (RBFE) calculations. ABFE calculations make use of an alchemical process where the ligand is nonphysically “removed" from solution and “inserted" into the protein's binding site.[102] The free energy change during this process is then calculated. However, ABFE are often challenging and computationally demanding to carry out and so RBFE are often favoured. Figure 1.8 shows the thermodynamic cycle for a RBFE calculation. An alchemical transformation between two structurally related ligands is performed to calculate the ΔG along the two vertical legs. The difference between these two values then yields the relative difference in binding between the two compounds. The ability to accurately compare the binding of two similar compounds makes this method especially applicable in the hit to lead and lead optimisation stages of drug development. A full theoretical description of these methods is given in the next chapter of this thesis.

Over the recent years, FEP calculations have been applied to a large number of drug discovery projects.[26,101,103] Many projects are retrospective in matching predicted activities with experimental values and are important in laying the foundation for future calculations. For example, Deflorian et al.[104] established a reliable set up for FEP calulcations for inhibitors of G protein-coupled receptors (GPCRs). GPCRs are one of the most important drug target classes but are notoriously challenging to model compared to globular proteins. By recognising the importance of key water molecules, amino acid ionisation states and equilibration protocols, a successful FEP protocol was developed. However, not all studies are retrospective. One example involves a collaboration between Schrodinger Inc. (a computational chemistry company) and Nimbus Therapeutics (a bio-technology company).[105] Their target was Tyk2, a member of the JAK family of kinases, which is implicated in a number of autoimmune diseases, such as psoriasis, inflammatory bowel disease, and rheumatoid arthritis. Current approved treatments are pan-inhibitors of the JAK family and can lead to significant side effects such as anemia and reduced immune function. MM-GBSA and FEP calculations were utilised to prioritise chemistry decisions in the lead optimisation of a highly selective inhibitor of Tyk2. This study is instrumental in showing that free energy methods can be fast enough to have an impact and improve efficiency of active drug discovery projects.

Despite RBFE being less demanding compared to calculations of ABFE, high computational cost is still a limitation. To obtain accurate results, it is essential to have sufficient sampling of free energies along the entire alchemical reaction path, requiring long simulation times. Additionally, for each pairwise set of compounds, it is necessary to perform a separate calculation, which often calls for manual intervention for the set up of each. Lambda dynamics calculations present a solution to these issues.[106,107] These types of free energy calculations predict the relative ΔG_{bind} energies for large sets of compounds in a small number of simulations. An extension of lambda

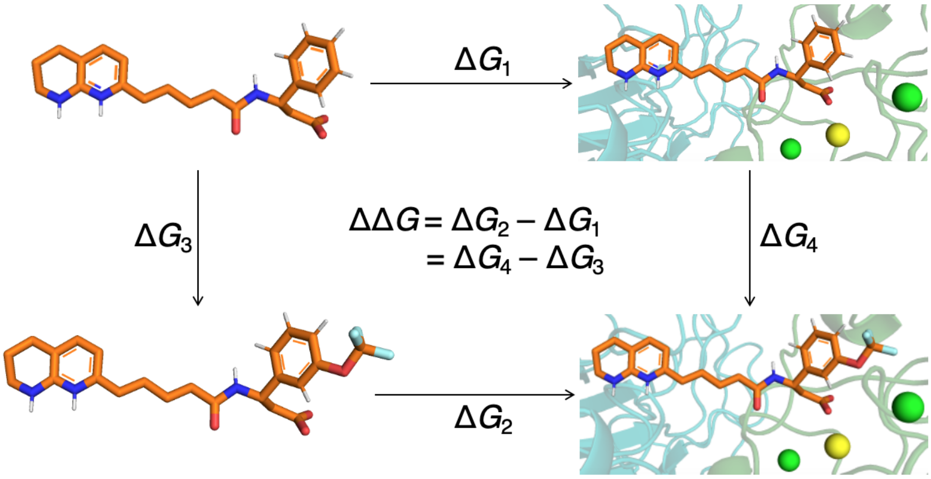


Figure 1.8: Thermodynamic cycle describing the binding of two compounds. ΔG_1 and ΔG_2 represent the free energy of binding of two ligands. ΔG_3 and ΔG_4 describe the free energy change of an alchemical transformation of one ligand into the other, as the free ligand in solution and bound to the receptor, respectively. The relative free energy of binding of two ligands is the difference between the alchemical free energy changes.

dynamics is multisite lambda dynamics (MSλD),[108] which allows for the chemical perturbation at multiple sites of a compound scaffold. A recent MSλD study successfully predicted, with a high degree of accuracy despite being large perturbations, relative ΔG_{bind} of 21 compounds in a single MD simulation.[109] These types of calculations have the potential to explore large chemical spaces, compared to traditional free energy methods, and enable rapid insights into compound modifications for the optimisation of potential drugs.

Binding free energy calculations still present a series of challenges, which limits their mainstream use in drug discovery projects.[103,110] The calculations are highly dependent on force field accuracy, reliability of the starting configuration of the target or complex and the technical challenges of setup and analysis. Despite this, alchemical free energy methods remain the most accurate types of binding affinity calculations and could revolutionise the hit to lead and lead optimisation stages of drug discovery through the effective prediction of affinity and selectivity.

1.3 Summary

The concepts of drug discovery and structure-based design have been introduced throughout this chapter. The discovery of new drugs is crucial for the quality and longevity of life, and over the recent decades, computational methods have increasingly facilitated the discovery of novel drugs and methodologies for their design. This thesis describes the application of CADD to drug discovery projects, in a collaboration between the University of Nottingham (UoN) and GlaxoSmithKline (GSK).

1.3.1 Outline of thesis

In the next chapter, the basic principles of molecular modelling are outlined. This chapter covers the background and methodology used throughout the work. The results chapters then presented in this thesis have a focus on two case studies of biomolecular systems with therapeutic interest. The first protein of interest is the integrin αvβ6, which is linked to the progression of the disease IPF. Chapter 3 describes the development of CHARMM force field parameters for small molecule inhibitors of αvβ6. This work is significant as it allows for complexes to be modelled accurately and reliably. In Chapter 4, we utilise these force field parameters by performing MD and FEP simulations on αvβ6-inhibitor complexes. These simulations provide an understanding of key binding site interactions and guidance for effective future computational studies on $\alpha v \beta 6$-ligand systems. This work is part of a wider drug discovery project between UoN and GSK.[8,111] The integrin project between these collaborators was established in 2011 and involves research carried out by 4th year MSci chemistry students. Each year (up to 2019), a cohort of ten MSci students design and synthesize compounds for the inhibition of $\alpha v \beta 6$, with compounds tested at GSK. The computational aspect of this project aimed to give additional insight to the system, which can then be used to guide compound design.[27,112]

Chapters 5 and 6 are based on work on BRD4-BD1, a protein that plays a key role in several diseases, especially cancers. Due to this involvement, BRD4 has been extensively studied and there are a large number of X-ray crystal structures available. Chapter 5 presents a thorough analysis of these structures and describes how the findings can be implemented in methodologies such as molecular docking and FEP.[113] In Chapter 6, different types of free energy calculations are explored, with BRD4 serving as a case study. The time and accuracy advantages of MSλD simulations, compared to traditional FEP, are demonstrated.

Chapter 2

Methods in Molecular Modelling

2.1 Introduction to Molecular Mechanics

Computational methods used to understand the behaviour of a molecular system can be split into two fundamental approaches, QM and MM. Modern QM approaches can calculate molecular properties, often with high accuracy, by characterising both the nuclei and electrons of an atom. Systematic approximations for solving the molecular Schrödinger equation can achieve energy estimations with an accuracy of greater than 99%, which can lead to chemical predictions that are accurate to a fraction of a kcal mol^{-1}.[114] However, QM calculations require considerable amounts of computing power and are generally not possible for large systems such as proteins and other biomolecules. In drug discovery, QM methods are most useful for studying the properties of isolated drug-like compounds or small active site regions of a receptor. In comparison, MM approaches are less computationally expensive and therefore most commonly used for modelling biomolecules. In this chapter, the methodologies of a selection of MM based methods are described. This provides a background to the techniques, which have been applied and are presented in this thesis.

MM calculations adopt the Born-Oppenheimer approximation,[115] which describes the energy of a molecule in terms of its nuclear positions. The rapid motion of the electrons is averaged out and assumed to be at an equilibrium surrounding the nuclei. This allows the atoms to be treated with fixed-point charges and molecular energy is calculated from the sum of bond, angle and non-bonded interaction contributions. Potential energy functions, which approximate the energy of a system at a given configuration are also known as force fields.

2.1.1 Empirical force field models

Force fields calculate the total potential energy (V_{total}) of a molecule based on the distortion of its bond lengths, bond angles and dihedral angles from their equilibrium values, along with non-bonded interactions, which are the sum of its van der Waals (vdW) and Coulombic interactions. A typical force field takes the general form:

$$V_{total} = V_{bond} + V_{angle} + V_{dihedral} + V_{vdW} + V_{elec} \tag{2.1}$$

The first three terms correspond to intramolecular properties, arising from the stretching of bonds between atom pairs, the bending of bond angles and the rotation around a dihedral, respectively. The following equations describe how each contribution is derived:

$$V_{bond} = \sum_{bonds} \frac{1}{2} k_b (r - r_0)^2 \tag{2.2}$$

$$V_{angle} = \sum_{angles} \frac{1}{2} k_\theta (\theta - \theta_0)^2 \tag{2.3}$$

$$V_{dihedral} = \sum_{dihedrals} \frac{1}{2} k_\phi [1 + cos(n\phi - \delta)] \tag{2.4}$$

k_b and k_θ are force constants and r_0 and θ_0 are equilibrium values so that bond lengths and angles are treated using a harmonic potential. The energy associated with the stretching of a bond (V_{bond}) between two covalently bonded atoms or bending of an angle (V_{angle}) between three consecutive atoms is calculated from the force needed to distort them from their minimum energy positions. The dihedral energy contribution ($V_{dihedral}$) arises from the rotation of bonds and the presence of steric barriers between four atoms that are separated by three covalent bonds. The potential energy for a dihedral angle, ϕ, includes terms for the force constant, k_ϕ, the multiplicity, n and the phase, δ. The multiplicity indicates the number of cycles per 360° rotation of the dihedral angle, while the phase describes the location of minima on the PES. A graphical representation of the force field terms is also provided in Figure 2.1.

Non-bonded interactions are treated using the Lennard-Jones 12-6 potential[116] for vdW interactions and the Coulomb potential for the electrostatics:

$$V_{vdW} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} 4\epsilon_{ij} [(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^6] \tag{2.5}$$

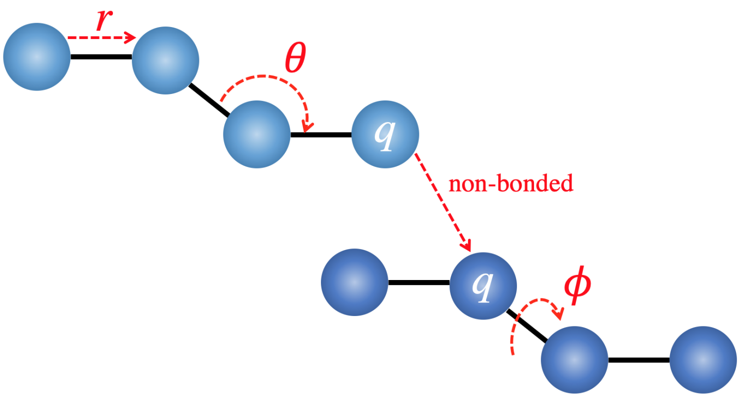


Figure 2.1: Schematic of the bonding, r, angle, θ, dihedral, ϕ, and non-bonded terms in a molecular mechanics force field.

$$V_{elec} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{q_i q_j}{4\pi\epsilon\epsilon_0 r_{ij}} \tag{2.6}$$

The Lennard-Jones function (Equation 2.5) consists of a short range repulsion term, $(1/r_{ij})^{12}$, where two atoms repel each other at a very close distance, and r_{ij} represents the separation between atoms i and j. This is modelled on the phenomenon of Pauli repulsion,[117] which prevents the overlapping of electronic orbitals. An attractive term, $(1/r_{ij})^6$ is also included, which models the vdW forces of attraction, or dispersion forces, between instantaneous and induced dipoles in atoms that arise from electronic fluctuations in molecules. ϵ_{ij} is the potential well depth, where $-\epsilon$ is the minimum energy between atom pairs i and j at the bond distance $r = 2^{1/6}\sigma$ and σ is the interatomic distance at which the potential energy equals zero. The Coulomb potential (Equation 2.6) models the electrostatic attraction and repulsion between atoms carrying an unequal charge distribution, where q_i and q_j are the partial charges of atom i and j respectively and r_{ij} is their separation. ϵ and ϵ_0 represent the relative dielectric constant and permittivity of vacuum, respectively.

Terms in a force field, which are not directly obtained from molecular coordinates, are parameters that are developed based on unique atom types. For example, carbon-carbon single bond lengths generally fall in the range of 1.45 to 1.55 Å and so r_0 in the V_{bonds} term is set to a value within this range. In contrast, a carbon-carbon double bond is shorter and will therefore be assigned a different r_0 value, despite also involving two carbon atoms. Therefore, atom types are generally assigned based on the chemical environment of the atom. The majority of MM software packages include atom typing algorithms, to carry out this task. The equilibrium values and force constants in these sets of parameters are ideally determined through experimental measurements. However, if experimental data cannot be obtained efficiently, these parameters are often derived from QM calculations on small representative structures. The accuracy of MM models is heavily dependent on the quality of the force field parameters and how well they can be generalised from one molecule to another. Therefore, it is sometimes necessary to perform parameter optimisation for the system or compound of interest.

There are various force fields, which have been developed for different types of molecules. Generally, they consist of some variation of the energy potential described in the previous paragraphs. Often they contain additional terms, such as hydrogen bonding or improper torsions, depending on their intended use. They may also go beyond the pairwise terms, to include, for example, three-body terms. Commonly used force fields for proteins include AMBER,[118] CHARMM,[119] GROMOS[120] and OPLS-AA.[121] The force field parameters, which describe a protein, are developed based on individual amino acids. For example, parameterisation is performed for a single alanine amino acid. For a specific protein or system, a database of parameters is then compiled based on protein sequence. Force field parameters for small drug-like compounds are provided separately. Common force fields for these organic compounds include the General AMBER Force Field (GAFF),[122] CHARMM General Force Field (CGenFF),[123] Merck Molecular Force Field (MMFF),[124–128] OPLS3[129] and GROMOS96.[130–132] The force field employed (and extended) in work presented in this thesis is the CHARMM force field for proteins and CGenFF for small molecules.

2.2 Protein-ligand Docking

Molecular docking is used to predict the most stable structures of protein-ligand complexes according to how the two 'fit' together to result in favourable steric and electrostatic interactions. Hence, docking can provide valuable insights when designing a drug compound, by estimating ligand binding affinities and modelling the interactions that take place between a ligand and the active site of a protein. These types of calculations are often very fast and are therefore efficient for HTS of large compound libraries. Furthermore, molecular docking is often used as a precursor to MD simulations, when an X-ray crystal structure of the complex is unavailable.

The first stage in molecular docking is the identification of a binding site. The coordinates of a receptor are often obtained from a crystal structure and the binding site can be identified through studying a crystal structure, where the protein is bound to a different ligand. The accuracy of docking is significantly improved when the binding site is already known.[33] However, comparing a target protein with a family of proteins, where their binding sites are known and they share a similar function can also be sufficient. There are also several programs that can identify potential binding sites through searching for cavities in protein structures. These include GRID,[133] POCKET,[134] SurfNet,[135] PASS[136] and MMC.[137] The quality of the receptor structure can also affect the effectiveness of molecular docking. Therefore, it is important to choose a reliable crystal structure with good resolution (< 2.5 Å).

Beyond binding site identification, molecular docking consists of two components. The conformational search algorithm identifies the possible poses and conformations of a ligand within an active site. A scoring function then assesses the affinity of each of these configurations. Docking programs perform these tasks in an iterative process, until the scoring function identifies a minimum energy conformation.

2.2.1 Search algorithms

In the search stage, torsional, translational and rotational degrees of freedom are modified, to generate different conformations of a ligand.[138] It would be too computationally expensive to sample every possible iteration of these binding modes. Therefore, sampling methods have been developed. These are typically performed through systematic or stochastic search algorithms (Figure 2.2).

Systematic methods sample the search space at predefined intervals, gradually changing the conformational parameters of the ligand.[139] The algorithm continues until an energy minimum is reached (Figure 2.2B). To

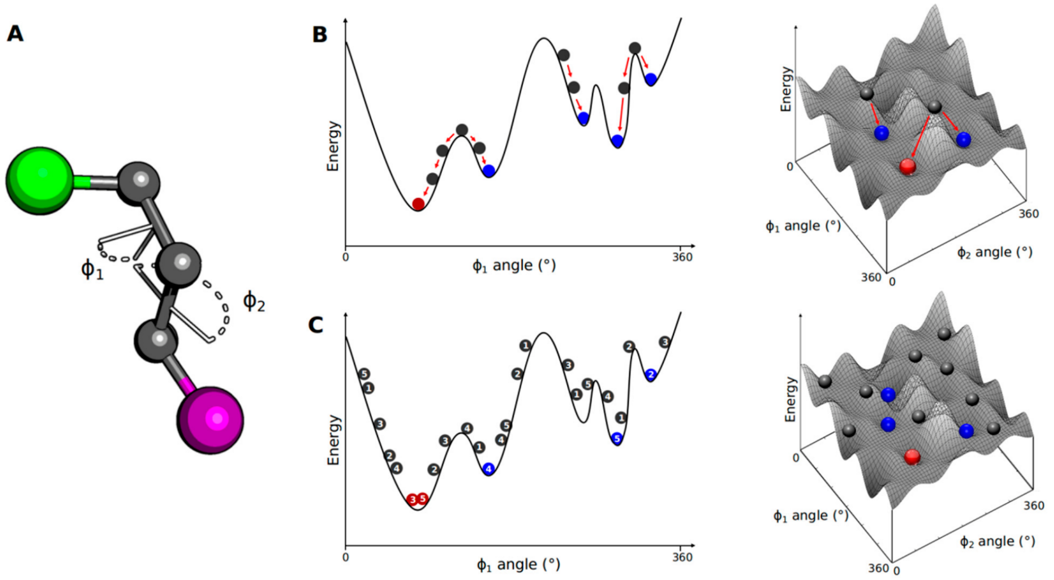


Figure 2.2: (A) Two dihedral angles, ϕ_1 and ϕ_2, define the possible conformations of a molecule. (B) The red (global energy minimum), blue (local minima) and black circles represent conformations generated by a systematic search algorithm, along a PES. The first graph represents the energy variation due to the rotation around ϕ_1, where ϕ_2 is kept frozen. (C) The circles along the PES represent the conformation generated by a stochastic search algorithm.[138] [Reproduced with permission from reference 139.]

avoid converging to a local minimum, compared to the global minimum, these types of methods are more effective when starting from multiple starting conformations of the ligand, which are iterated through simultaneously.[140] Stochastic methods search the conformational space by making random modifications to the ligand. The algorithm generates ensembles of conformations and populates a wide range of the energy landscape (Figure 2.2C).[138] The choice of search method is often dependent on the type of problem being addressed. For example, how much speed is necessary compared to a more comprehensive search of conformational space.

Matching algorithms[141–143] are a type of systematic search. They are based on matching a shape map of a ligand, using ligand pharmacophores and chemical information such as the position of hydrogen bond donors and acceptors, to the binding site of a receptor. These algorithms have the advantage of speed, so are practical for large chemical libraries such as in HTS. However, a limited accountability of ligand flexibility is achieved. A systematic search method that gives a better representation of ligand flexibility is incremental construction.[144–146] In this method, a ligand is broken into several fragments and then gradually built back together in the binding site. As the fragments are added, different orientations are explored, which accounts for the flexibility of the ligand. Also, as a conformational search is only performed for the fragment that is being added, there is a reduction in the degrees of freedom, which reduces the number of possible combinations of internal parameters. This prevents a combinatorial explosion.[138]

Monte Carlo[147,148] and genetic algorithms[40,149] are two types of stochastic methods. In Monte Carlo methods, ligand poses are generated through bond rotations or rigid-body translations. The conformations are then evaluated by an energy function. If the conformation passes an energy criterion, it is accepted and further modified to generate the next conformation. This procedure is repeated until the required number of conformations is obtained. An advantage of using Monte Carlo search algorithms is that the changes made to the ligand can be large, meaning energy barriers can be crossed, increasing the chance of finding the global energy minimum.[150]

Genetic algorithms apply the concepts of the theory of evolution and natural selection in generating ligand conformations. The degrees of freedom are encoded as lists of values called genes, which then make up a chromosome and represents the full structure of the ligand. This starting chromosome is used to make mutations and the crossover of genes to generate a population of chromosomes. This results in new structures, which are assessed using a scoring function. If the score is sufficient, these structures are used for the next generation of chromosomes. Genetic algorithms present an efficient way to sample wide areas of conformational space in a small number of conformations.[138,150]

Docking studies presented in this thesis use the program OpenEye OMEGA[151] for the generation of ligand conformations. OMEGA is a systematic search algorithm which consists of three components. The first involves the assembly of an initial 3D structure from a fragment library. This library has been constructed by fragmenting a large collection of commercially available compounds into ring systems and small linear linkers. Fragments are optimised using a modified version of MMFF (MMFF94[125,126]) and distinct conformations with the lowest energies are chosen. From the user input of the ligand (usually in the form of SMILES strings), fragments are assembled to form an initial conformation. Next, every rotatable bond in this initial conformation is compared to a torsion library, which is a knowledge-based list of angles and rules to reduce the potential energy. This stage generates a large set of conformations. Finally, these conformations are sampled by geometric and energy filters so that (a) all conformers have a score less than ten units higher than the lowest energy conformations or (b) 200 mutually unique conformers are generated. A common way to test the effectiveness of conformer generation in docking is to re-dock a ligand into its original protein crystal structure and compare the docked pose to the crystallographic pose. When this validation is carried out on a set of 197 challenging ligands, OMEGA has been found to perform very well in reproducing crystallographic conformations.[151]

2.2.2 Scoring functions

Once ligand conformations are generated and docked into the active site of a receptor, a scoring function is used to assign a score to the pose. These scores serve as estimations of the binding affinity and can be used (with caution) to rank compounds in drug discovery projects. Scoring functions can be empirical, force field based or knowledge-based.[152]

Empirical functions estimate binding affinity by breaking it down into several components, which account for hydrogen-bonding, ionic interactions, hydrophobic interactions and entropic effects. Based on a training set of compounds with known binding affinities, these components are each weighted and summed to give a total energy score. Although the simplicity of empirical functions mean that they provide quick evaluations of ligand poses, they are limited by the transferability and quality of the test set of compounds, used to develop the model.[153–155]

Force field based scoring functions[156–158] provide a more system specific assessment of a docked pose, compared to empirical functions. Binding affinity is estimated by the sum of bonded and non-bonded terms, as previously described in Equation 2.1. However, these types of scoring functions have a slower computational speed, so are less suitable for evaluating large sets of compounds. Additionally, functions to account for hydrogen bonds, entropy contributions and solvent effects should be included to improve their accuracy.[159–161]

Knowledge-based scoring functions[162–164] use statistical analysis to extract pairwise energy potentials from known crystal structure protein-ligand complexes. Based on the assumption that favourable interactions will frequently occur in known structures, a score is calculated by favouring these types of interactions and penalising repulsive protein-ligand interactions. An advantage of knowledge-based scoring functions is their ability to model uncommon interactions, such as sulfur aromatic or cation-π, which are difficult to account for in empirical and force field based functions.[150]

Molecular docking studies described in this thesis use the OpenEye FRED program[37] for docking. This program uses Chemgauss4 as a scoring function. Chemgauss4 recognises the shape, protein-ligand hydrogen bonds, hydrogen bonding with solvent and metal-chelate interactions of a ligand within an active site. Shape interactions are based on the vdW radii of heavy atoms. A penalty score is assigned in the event of the distance between two atoms being within the sum of their vdW radii. Otherwise, a score is assigned that is proportional to the number of protein heavy atoms within 1.25 and 2.5 the sum of the vdW radii of the ligand atoms. Protein-ligand hydrogen bonding scores are dependent on (a) how far the hydrogen bond donor is from the ideal position based on the position of the acceptor atom and (b) how far the hydrogen bond acceptor is from the ideal position based on the position of the donor atom. For one particular hydrogen bond, the score is the product of two Gaussian functions of these distances, multiplied by the strength of the hydrogen bonding groups involved. The total score is the sum of all protein-ligand hydrogen bonds. The scoring function also penalises the breaking of solvent hydrogen bonds as the ligand docks into the active site. For metal chelating groups, a fixed score is assigned for each protein metal that is within 1.0 Å of any chelating position on the ligand.

Limitations

Most docking programs are accurate predictors of ligand binding poses, especially when the active site of the protein is already known. However, there are often system dependent practical considerations, such as the treatment of crystallographic water molecules. It should be carefully decided whether to include water molecules, so that they can facilitate binding through bridging protein-ligand interactions, or whether the entropy gain of displacing an active site water molecule is more favourable. The main limitation in molecular docking, however, lies in the accuracy of the scoring function. Although docking can be sufficient as a binary predictor of which compounds bind and which do not, the ability of molecular docking to correctly rank ligands in the order of their binding affinity is a challenge. Furthermore, improvements to the accuracy of scoring functions are constrained by the need for fast evaluations of large numbers of compounds. Often there needs to be a compromise between accurate predictions and speed. In projects where accurate predictions of binding free energy are necessary, more rigorous thermodynamic techniques such as MM-PB(GB)SA are required. FEP methods also present a solution to obtaining accurate predictions. However, these are significantly more computationally expensive and are not feasible at the required timescales for HTS.

Modern docking programs often account for both ligand and protein flexibility. However, protein flexibility is generally limited to binding site side chain rearrangements. As larger scale protein conformational changes are often important upon ligand binding, higher level sampling methods are required, such as MD simulations. MD simulations can also provide information on the stability of binding site interactions, based on the frequency they are maintained throughout a simulation. This allows us to understand which interactions are most important for binding and should be targeted when designing drug compounds.

2.3 Molecular Dynamics Simulations

In solution, proteins are flexible and the dynamics of their side chains gives insight to their function. Although X-ray crystal structures can provide an atomic level of resolution for a protein, the averaged measurements of protein conformations from single crystals do not reflect well the mobility of a protein in solution. MD simulations are a more comprehensive way to model the conformational dynamics of protein and protein-ligand complexes. Atomic motion is simulated along the PES of a system, which determines the relative stability of different conformations. Energy is supplied to the system using a constant temperature. Depending on the temperature specified, the conformations that are sampled are often around a local or global minimum, as shown in Figure 2.3.

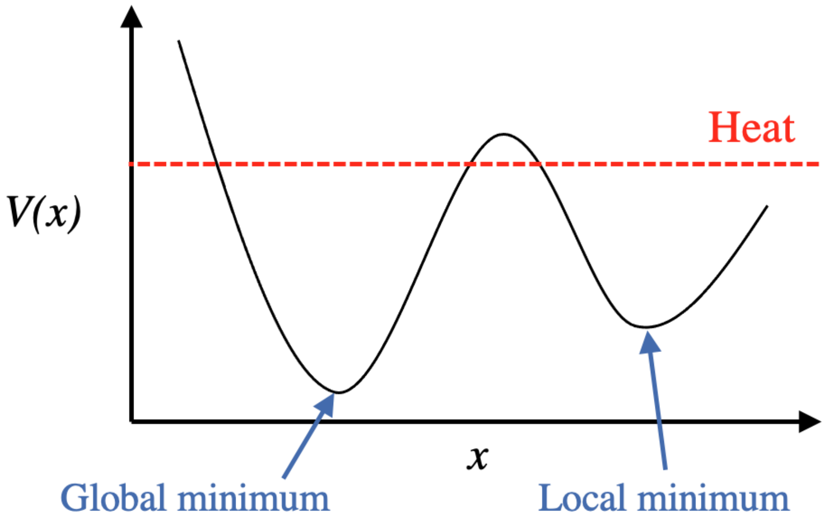


Figure 2.3: A 2D representation of a PES. The y axis is potential energy, while the x axis represents the coordinates of a protein conformation. Protein structure is at an energy minimum or at the top of an energy barrier when $\frac{dV(x)}{dx} = 0$. Energy is provided to the system in the form of heat, as shown by the dotted red line. Energy barriers, which have a higher energy than the heat provided, cannot be crossed.

2.3.1 Conformational sampling

In MD simulations, the forces acting on atoms are used to calculate the dynamics of the system. This is done by solving Newton's equations of motion.

$$F_i(t) = m_i a_i = \frac{-dV(r(t))}{r_i} \tag{2.7}$$

$F_i(t)$ is the force exerted on atom i at time t, where $i = 1, 2, ..., N$ and N is the total number of atoms in the system. m is the mass and a is acceleration. Force is also equivalent to the gradient of potential energy, V, with respect to atom position, r. This means that from the potential energy of each atom (calculated using a force field), it is possible to calculate the acceleration of each atom.

To obtain trajectories $(r_i(t))$ of all N atoms in a system as a function of time, integration algorithms are used. These are approximated by the Taylor series:

$$r(t+\delta t) = r(t) + r'(t)\delta t + r''(t)\frac{\delta t^2}{2} + r'''(t)\frac{\delta t^3}{6} + ... \tag{2.8}$$

The Verlet algorithm,[165] one possible integration algorithm, uses up to the third term in the expansion. As velocity is the first derivative of position with respect to time and acceleration is the second derivative, this gives the following expansions:

$$r(t+\delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 \tag{2.9}$$

$$v(t+\delta t) = v(t) + a(t)\delta t + \frac{1}{2}a'(t)\delta t^2 \tag{2.10}$$

where v is the velocity. In the Verlet algorithm, expansions from t to $t + \delta t$ and $t - \delta t$ are combined to give:

$$r(t+\delta t) = 2r(t) - r(t-\delta t) + a(t)\delta t^2 \tag{2.11}$$

In this basic Verlet algorithm, velocities are not calculated explicitly and so they are calculated using Equation 2.12:

$$v(t) \approx \frac{r(t+\delta t) - r(t-\delta t)}{2\delta t} \tag{2.12}$$

This means, given two sets of atomic coordinates and a set of velocities, the time-dependant behaviour of a system can be calculated. However, at the start of a MD simulation, two sets of atomic coordinates are not yet known. Furthermore, atomic positions for three consecutive time steps need to be stored if velocities are to also be calculated. Therefore, there are better tools to use in practice. Popular algorithms that overcome these issues include the leapfrog algorithm[166] and the Velocity Verlet algorithm.[167]

The time step, δt, is an important consideration. Although a smaller time step gives a better approximate integration, if it is too small the computational cost is too high to achieve sufficient exploration of configurational space. Conversely, large time steps result in unstable simulations with large errors during the integration of motion. For a flexible molecule such as a protein, δt should be no greater than $1/10$ the time of the shortest period of motion, which is the stretching vibration in molecules ($\sim 10^{-15}$ s). The SHAKE algorithm[168] is a common application for bond distance constraints in MD simulations. In simulations of biomolecules, the highest frequency vibration is those involving hydrogen atoms. Therefore, the SHAKE algorithm is often applied to constrain any bonds that contain hydrogen atoms, while still allowing all other atoms to move and vibrate. This allows for the time step to be increased and results in a lower computational cost for a simulation of a given length. A typical time step for a protein MD simulation is 2 fs. However, this limits the simulation length of a MD simulation to nanoseconds or microseconds, whereas biological processes often happen on the microsecond to second timescale.[169] Therefore, starting conformations are important as MD simulations tend to sample the local configurational space from where they are initiated. For better sampling, it is good practice to run multiple simulations with different initial velocities or to use enhanced sampling methods.

Once a MD simulation has been run for a specified length of time, a trajectory is produced. This is a combination of frames, which show snapshots of the protein conformation as time progresses. Trajectories can be observed by using visualisers such as Visual Molecular Dynamics (VMD),[170] PyMOL[171] and UCSF Chimera,[172] providing a qualitative view of protein dynamics. Many MD simulation packages such as CHARMM,[69] NAMD,[173] AMBER[174] and GROMACS[175] also provide analytical tools to obtain quantitative measures, such as the root-mean-square deviation (RMSD) of the protein backbone during the simulation, compared to its average or starting position.

2.3.2 Ensemble averages

In MD, a system is prepared in a given state, and then allowed to relax towards equilibrium. At equilibrium, physical averages of the system can then be theoretically determined. The microscopic state of a system describes the positions, velocities and momenta of the atoms it contains. Phase space is then explored during MD simulations, where phase space is the space of all possible microscopic states of the system. The thermodynamic, or macroscopic, state of the system is defined by the temperature, pressure and number of particles. Thus, an ensemble is a collection of all possible microscopic states within a given thermodynamic state. MD simulations generate different microscopic states, i.e. configurations of a protein, which belong to the same ensemble. Different ensembles have the following characteristics:

- **Microcanonical ensemble:** The simplest thermodynamic state, where the system is sampled with a constant number of particles, volume and energy (NVE). However, this ensemble can be unrealistic as it does not involve any interaction or heat exchange with the environment. It is most suitable for investigating time-dependent phenomena such as the vibrational frequencies of a complex system.

- **Canonical ensemble:** This state better matches the conditions of a simulation to experiment by keeping the number of particles, volume and temperature constant and allowing the energy to change (NVT). This ensemble is most suitable for investigating finite temperature phenomena in an isolated system, such as diffusion.
- **Isobaric-isothermal ensemble:** Matches the conditions of a simulation with experiment by keeping a constant number of particles, pressure and temperature (NPT). This ensemble resembles the conditions for chemical reactions as it is suitable for non-isolated systems where the volume can change.
- **Grand canonical ensemble:** The chemical potential, volume and temperature of a system is kept constant (μVT).

To understand any physical property, A, its average over the contributions of all states is required. For a system in thermal equilibrium with a heat bath at a fixed temperature (NVT, NPT, μVT), this ensemble average is given by:

$$\langle A \rangle = \sum_{i=1}^{states} \rho_i A_i \tag{2.13}$$

where ρ is the probability of the system being in state i, which is given by the Boltzmann function,

$$\rho_i = \frac{1}{q} \exp{(\frac{-E_i}{k_b T})} \tag{2.14}$$

$$q = \int dr^N \exp(\frac{-E(r^N)}{k_b T}) \tag{2.15}$$

where q is the partition function, E is the energy of state i, k_b is the Boltzmann constant, T is temperature and r^N is the position of all particles, N. From this, we can interpret that lower energy states are favoured, while higher energy states have an exponentially decreasing probability of being sampled.

To calculate the full ensemble average for A, and integrate over all possible states, an MD simulation would have to pass through all possible configurations of a system, which is not feasible. Therefore, MD uses a time average over the course of the MD simulation.

$$\langle A \rangle = \langle A \rangle_{time} \approx \int_0^t A(t)dt \tag{2.16}$$

2.3.3 Practical considerations

There are factors, additional to the force field, time step and software, to be considered when setting up an MD simulation. For the majority of *in silico* studies of biomolecular systems, it is desirable to simulate the protein as close to experimental or physiological conditions as possible. Therefore, it is necessary to also simulate a solvent environment. Furthermore, as it is only possible to simulate a finite number of atoms, boundary effects should also be accounted for in the form of periodic boundary conditions (PBC). Energy minimisation and equilibration are important to correct any bad atom connections in the crystal structure and to ensure an appropriate area of conformational space is being explored. These considerations are made prior to the main MD simulation, which we term the data collection or production stage.

Solvation

As proteins function in an aqueous environment, the effects of a solvent environment must be taken into account to achieve a realistic model. Solvent plays an important role in the physiological function of a biomolecule. It gives rise to the hydrophobic effect,[176] an entropic effect which brings together non-polar regions of a protein. Water also provides hydrogen bond donor and acceptors, which can stabilise tertiary structures and facilitate ligand binding. Modelling the behaviour of solvent also allows for the cost of water displacement to be accounted for in binding events.

There are different levels of chemical accuracy and computational cost that can be achieved by different solvent models. The simplest of methods is the implicit solvent model. As water dynamics are typically much faster than protein rearrangements, the averaged behaviour of many water molecules can be treated as a continuum environment. The electrostatic components are calculated according to the Poisson-Boltzmann equation[177] or the Generalised Born equation.[178] Solvent can also be modelled using a coarse grained approach, where each water molecule is treated as just one atom with a surrounding potential. These two approaches mimic the behaviour of bulk water well and have a low computational cost. However, they do not capture individual solute-solvent interactions, which can be important for understanding ligand binding. Explicit solvent models, which include water molecules explicitly as three atoms, obtain a higher level of accuracy although they have a higher computational cost. In most explicit models, a fixed point charge is assigned to each atom. However, it is possible to go a level higher and account for polarisation on each of the atoms.[179] In this thesis, simulating protein-ligand interactions and calculating binding free energies are of the most interest. Therefore, explicit solvent models are used throughout the work.

For computational efficiency the bond lengths and angles of water molecules are often kept rigid in explicit models. Examples of rigid water molecule models include the transferable intermolecular potential surface (TIPS) model,[180] the simple point charge (SPC) model[181] and the transferable intermolecular potential 3P (TIP3P) model.[182] Figure 2.4 shows the different site models that can be applied. In a 3-site model, the water molecule has three interaction sites, corresponding to each of the atoms. In a 4-site model, only the vdW forces are accounted for on the oxygen atom and a mass-less charge that is associated with the oxygen atom is appended along the bisector of the H-O-H bond angle. In a 5-site model negative charges are located on the electron lone-pair positions on the oxygen atom. A 6-site model is a combination of the 4- and 5-site models. In the work presented in this thesis, the TIP3P (3-site) model is used.

In a simulation of a protein, when replicating physiological conditions, it is necessary to include amino acids that have protonation states according to physiological pH (~ 7.4). This means lysine, arginine and histidine residues should be protonated and have a positive charge and aspartate and glutamate residues should be deprotonated and have a negative charge. Therefore, when constructing a solvent environment, counter ions should be

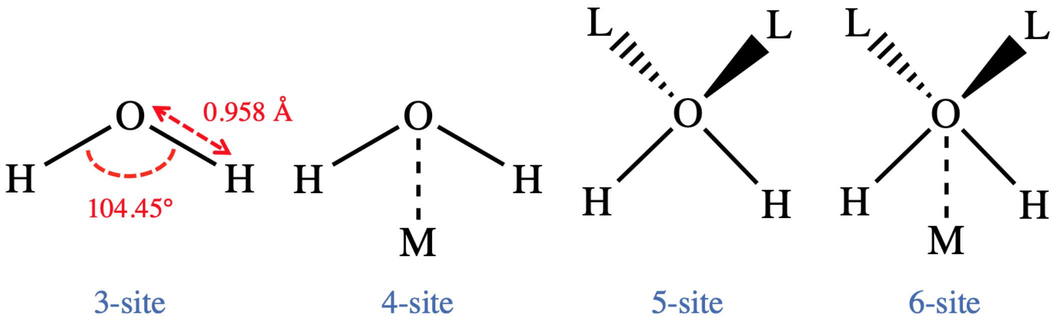


Figure 2.4: The TIPS water models.[182] 'M' represents a mass-less charge associated with with the oxygen atom and the 'L' characters represent the electron lone pairs on the oxygen atom.

added to give a net neutral charge. Alternatively, ions could be added at a concentration to match experimental conditions.

To control the flow of atoms that are moving in a system and due to the limited number of atoms that can be simulated, PBC are typically applied to explicit solvent systems.

Periodic boundary conditions

In MD simulations, the solute is solvated with a finite number of water molecules. Therefore, without boundary conditions, the edge atoms would face vacuum, which is not a good physical description of a biological system. PBC present a solution, which involves extending the system periodically in all three directions to represent a pseudo-infinite system (Figure 2.5). When using PBC, Newton's equations of motion are solved for one primary cell and the same movements, momenta and interactions are applied to the identical atoms in the replica cells. If an atom or molecule leaves the primary cell, the same atom or molecule enters from the opposite replica cell. Therefore, satisfying the requirement for a constant number of particles, as needed for

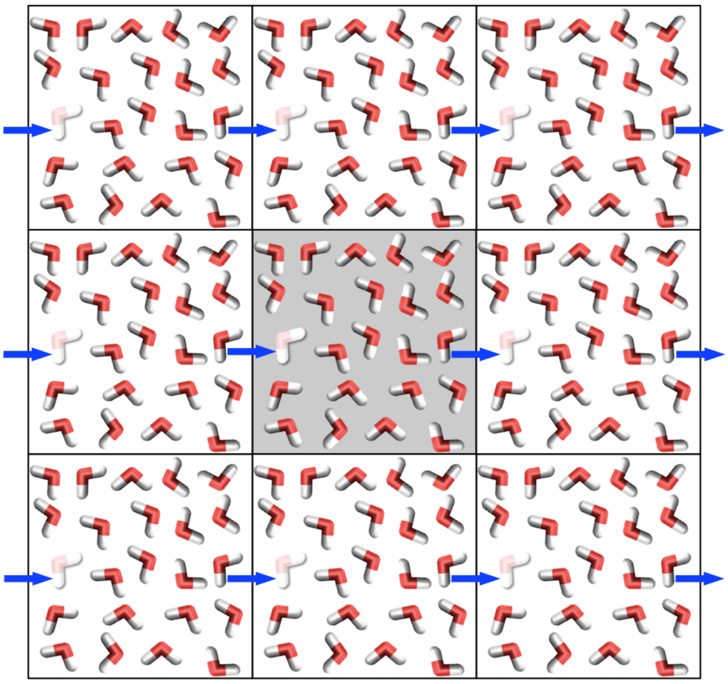


Figure 2.5: A 2D schematic of a periodic system with the primary cell shaded in grey.[183] The blue arrows show how when one water molecule moves across the boundary of a cell, all corresponding images of the molecule also move. [Figure adapted from reference 182.]

NVT, NPT and NVE ensembles.

In systems where PBC are applied, only the interactions with surrounding atoms within a cutoff are considered. This saves computational time and ensures that the solute does not interact with a periodic replica of itself. A switching function is applied so that the contribution from vdW interactions is set to zero as the distance between two atoms approaches the cutoff. As vdW interactions quickly approach zero as the distance increases anyway, the cutoff does not introduce too many errors to the potential energy.[69] However, Coulomb interactions decay to zero slowly as the distance between two atoms increases, which introduces discontinuities in the potential energy at distances around the cutoff. The Coulomb energy between N atoms in all simulation cells is:

$$V = \frac{1}{2} \sum_{\boldsymbol{n}} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{q_i q_j}{4\pi\epsilon_0 |r_i - r_j| + \boldsymbol{n}} \tag{2.17}$$

where $\boldsymbol{n}$ is the cell vector and q_i, q_j, r_i and r_j are the partial charges and positions of atoms i and j, respectively. To solve this equation for long range electrostatic interactions that are separated by a distance larger than the cutoff, the Ewald summation is used.[184] In the Ewald summation, a charge distribution of opposite sign is placed around a point charge. This means the interaction between the charges have short range character, making them easier to compute. To counteract the opposite sign distributions, the same charge distributions are introduced in reciprocal space. The particle mesh Ewald (PME) method[185] is a fast and efficient way to approximate the Ewald summation, and is used in the majority of simulation packages to compute long range interactions.

PBC can be constructed with various geometries. Figure 2.6 shows the shapes of periodic cells that are often used in protein MD simulations. The choice of periodic cell can be made based on the shape of the protein. For example, a cubic or truncated octahedra could be the most appropriate for a globular protein. Truncated octahedra are often preferred as they allow the least number of solvent molecules in the system and therefore speed up the simulation.

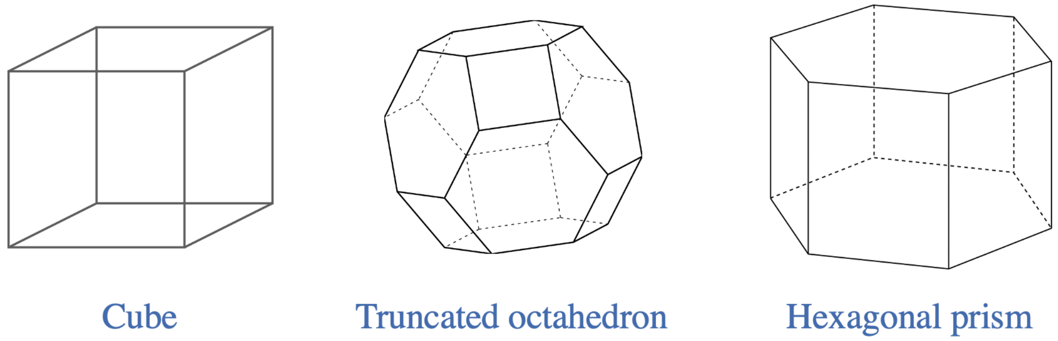


Figure 2.6: Examples of periodic cells used in simulations.

Energy Minimisation

Energy minimisation is used to prepare a system for MD. As energy minimisation algorithms search the PES for local minima, which correspond to stable arrangements of the atoms, they are useful to relieve any unfavourable interactions that may be present in the starting X-ray crystal structure. Although there are different types of energy minimisations algorithms available,[186] this section focuses on those that use the derivative of the potential energy with respect to the Cartesian coordinates of the system atoms.

Derivative algorithms operate as an iterative process. Atomic positions are progressively changed towards a minimum energy configuration, where the first derivative of the energy function is zero with respect to the coordinates (Figure 2.3). Minimisation is continued until a defined set of iterations has been performed or till the potential energy converges to a minimum value. Steepest descent[187] and conjugate gradient[188] methods are examples of first order minimisation algorithms. Figure 2.7 shows a representation of steepest descent. The search starts at an arbitrary position (the initial coordinates) and moves in the direction where the energy decreases most quickly. The Newton-Raphson method[189] is an example of a second order derivative

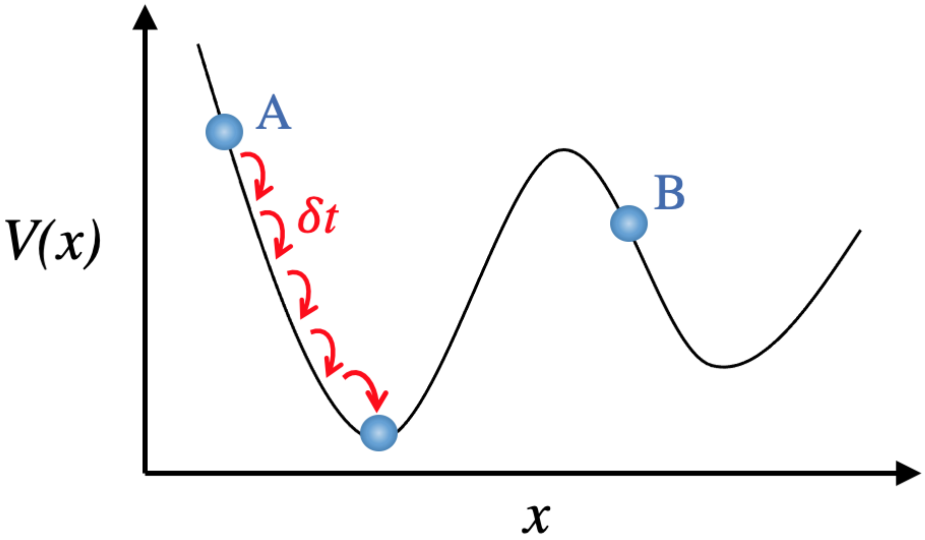


Figure 2.7: A 2D schematic of a steepest descent minimisation. Initial guess structure ‘A’ is minimised to an energy minimum conformation using a timestep δt. Structure ‘B’ would relax to the energy well to its right, despite the alternative energy well having a lower energy.

method, where a Taylor expansion is used around the initial set of coordinates. Although this method is more computationally expensive, it usually requires fewer steps, compared to first derivative methods, to reach a minimum configuration.

An important consideration when performing energy minimisations is that they follow a ‘downhill’ path along the PES. This means that algorithms can only identify the minimum energy point which is closest to the starting point, regardless of whether it is a local or global energy minimum. For example, point B on Figure 2.7 will reach the local minimum to its right, despite being one energy barrier away from the global minimum. For this reason, more sophisticated sampling methods such as MD simulations are used to achieve more detailed understanding of the PES of a protein, and energy minimisations cannot be used as a complete alternative.

Equilibration

As a system is heated, typically from 0 K to 298 K, the kinetic energy being added to the system must be transferred to potential energy. During this time, the solvent and protein atoms undergo a relaxation, which can last for up to nanoseconds before the system reaches a stable configuration. This period is typically discarded from any trajectory analysis and is called the equilibration stage. Equilibration is often conducted in two phases. The first phase is conducted under the NVT ensemble and is continued until the temperature fluctuates around a specified stable average. It is good practice to check that the protein system has reached a stable temperature before continuing with equilibration. Figure 2.8 shows an example of a protein system that has been heated and is stable around 298 K. The data comes from the NVT equilibration stage of a MD simulation of a BRD4 complex, as presented in Chapter 6 of this thesis. The second phase of equilibration is performed under the NPT ensemble. During this phase, the pressure of the system is also stabilised. To be confident that equilibration has been performed sufficiently, criteria such as the average velocity distribution and thermodynamic properties should be checked. To ensure the structure is stable, RMSD as a function of time is commonly calculated between the backbone atoms of the protein at each time step and their average position, as shown in Figure 2.9. When the RMSD reaches a plateau, this indicates that the structure is stable and the data collection stage of the MD simulation can begin.

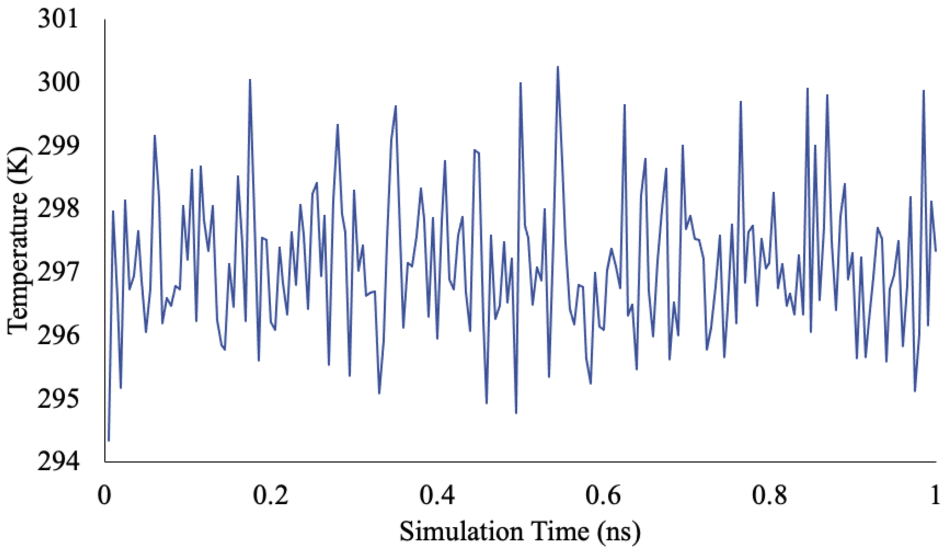


Figure 2.8: Temperature of a 1 ns MD simulation in the NPT ensemble. The temperature quickly reaches the target value of 298 K and remains stable over the remainder of the simulation.

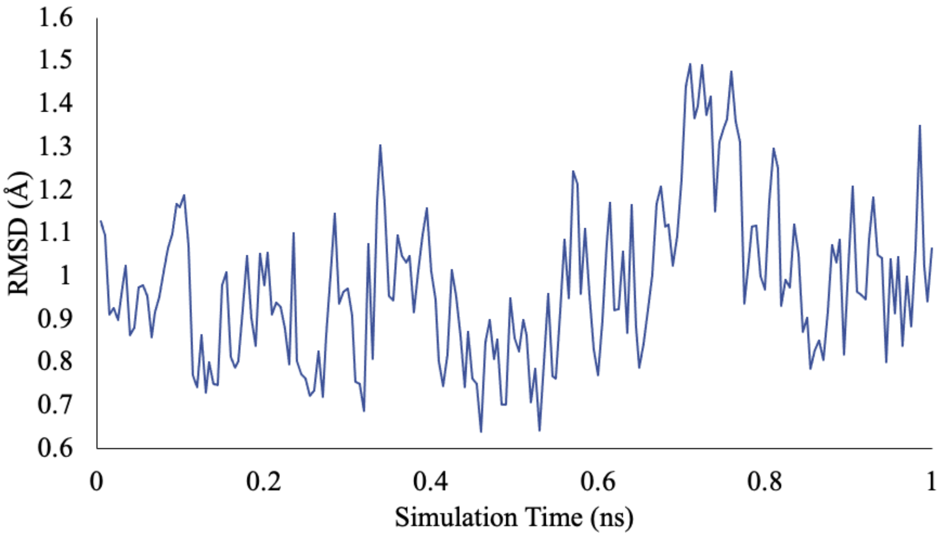


Figure 2.9: RMSD of the position of the backbone atoms of a protein with respect to their average position during a 1 ns MD simulation. A stable RMSD indicates an equilibrated system.

Enhanced sampling methods

Efficient sampling can often be a limitation of MD simulations. MD trajectories may not reach all relevant configurations during the timescale of a simulation or the protein may become trapped in local minima for its duration. Enhanced sampling methods, such as metadynamics,[80] replica exchange[79] and simulated annealing,[81] can be used to address this issue:

- **Metadynamics** was introduced by Laio and Parrinello in 2002.[80] In this method, a number of Gaussian potentials are placed in the potential energy wells. This enables a system to cross energy barriers and prevents it from visiting configurations, which have already been sampled.

- **Replica exchange MD** involves running multiple simulations at different temperatures in parallel, and then frequently exchanging conformations between each simulation. Higher temperatures provide more energy to the system, enabling it to cross high energy barriers along the PES. Successful replica exchange techniques rely on the choice of temperature range, length of each simulation and the number of replicas.[190]

- **Simulated annealing** involves initiating a simulation at a high temperature, where energy barriers can be crossed and an extensive exploration of configurational space is possible. The temperature is then cooled during the simulation, allowing the system to relax into a minimum energy conformation.[190]

2.4 Free Energy Perturbation Methods

Accurately predicting the binding affinity between a protein and a potential drug compound is one of the ultimate goals of CADD. There are various methods for the estimation of binding free energies, which are continuously advancing with the increasing availability of computational resources. Recent advancements include opportunities for performing free energy calculations on a cloud computing platform.[191] In this section, focus is placed on alchemical pathway methods, specifically FEP and lambda dynamics.

2.4.1 Absolute free energy perturbation

Gibbs free energy is a thermodynamic potential that measures the capacity of a thermodynamic system to do maximum or reversible work at a constant temperature and pressure. Protein-ligand binding occurs when the change in Gibbs free energy, ΔG, is negative. The magnitude of the negative value is analogous to ligand binding affinity, i.e., a compound with a more negative ΔG of binding forms tighter interaction with a receptor, compared to a compound with a less negative ΔG.

Zwanzig's equation for FEP[192] estimates the free energy difference between two states as:

$$\Delta G = G_1 - G_0 = -k_B T \ln \left\langle \exp(-\frac{U_1 - U_0}{k_B T}) \right\rangle_0 \tag{2.18}$$

where U_0 and U_1 are the potential energies of each state, k_B is the Boltzmann constant, T is temperature and $\langle\rangle_0$ represents the Boltzmann average at state 0. In FEP, the two end states are split into a series of intermediate states via a coupling parameter, λ. The potential energy of the states, U_λ, then changes from U_0 to U_1 as λ is incremented from zero to one:

$$U_\lambda = \lambda U_1 + (1 - \lambda)U_0 \qquad (0 \leq \lambda \leq 1) \tag{2.19}$$

The total free energy change between G_0 and G_1, is the summation of the ΔG between each pair of adjacent intermediate states. Through the linear mixing of force field parameters, MD simulations can be used as a sampling technique for generating representative conformations for each λ state and calculating U_λ.

As free energy is a state function, ΔG is defined by the initial and final states of a system, regardless of the pathway connecting them. Therefore, absolute FEP calculations take advantage of the thermodynamic cycle shown in Figure 2.10. This thermodynamic cycle is used to calculate the free energy of binding (ΔG_b^o) of a ligand (L) for a protein (P). It should be noted that absolute FEP does not refer to computing the absolute free energy, G, but to the difference in free energy between the bound state of the ligand and the state where it is free in solvent. The word 'absolute' is used to distinguish between relative FEP calculations, which are discussed under the next sub heading.

To calculate ΔG_b^o via the most direct path would require large amounts of complicated computation. Therefore, ΔG_b^o is calculated through a series of alchemical intermediate states (B, C, D and E in Figure 2.10). The first step involves decoupling the ligand from solution to calculate $G_{elec+vdW}^{solv}$. During a series of λ states, often called λ windows, the Coulombic interactions are turned off, followed by the vdW interactions. It is important to decouple

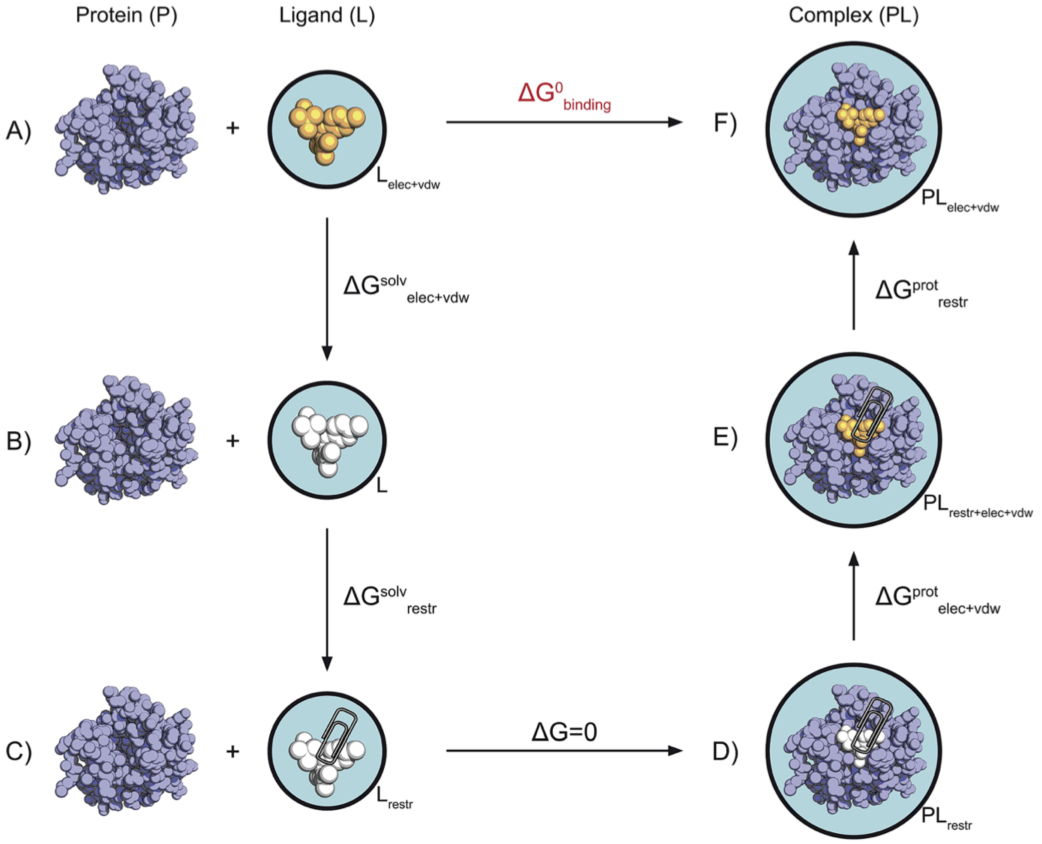


Figure 2.10: Thermodynamic cycle used to obtain absolute binding free energies.[193] The top horizontal ΔG value, highlighted in red, is obtained by calculating the vertical legs of the cycle. The ligand in solution (A) is transformed into a non interacting solute (B). The non-interacting ligand is then restrained (represented by paper clips) and ΔG^{solv}_{rest} is calculated analytically using a protocol described by Boresch et al.[194] The right vertical leg is calculated in reverse, where constraints are applied to the ligand in complex (E), followed by the decoupling of the ligand from the system (D). [Reproduced with permission from reference 192.]

the Coulombic interactions first to prevent errors caused by atoms becoming too close due to their lack of vdW surface. The free energy difference between states B and C is calculated analytically, according to an expression proposed by Boresch et al.[194] State C is equivalent to state D, where the ligand interactions within the protein binding site have been decoupled. Therefore, the ΔG between these states is zero. The next two legs of the thermodynamic cycle, $\Delta G^{prot}_{elec+vdW}$ and ΔG^{prot}_{restr}, are calculated in reverse. To keep the position and orientation of the ligand close to the bound pose once its interactions with the environment are decoupled, restraints are applied (represented with paper clips in Figure 2.10). Typically, one distance, two angles and three dihedral angle restraints between ligand and receptor atoms are introduced through a harmonic potential with a high force constant (~10 kcal^{-1} Å^2 deg^2). The contribution of these restraints to ΔG^{solv}_{restr} are evaluated by:

$$\Delta G^{solv}_{restr} = RT \ln \left[\frac{8\pi^2 V^0}{r_0^2 \sin\theta_{A,0} \sin\theta_{B,0}} \frac{(k_r k_{\theta A} k_{\theta B} k_{\phi A} k_{\phi B} k_{\phi C})^{\frac{1}{2}}}{(2\pi k_B T)^3} \right] \tag{2.20}$$

where R is the ideal gas constant, T is the temperature in Kelvin, V^0 is the volume corresponding to one molar standard state, or 1660 Å^3. r_0 is the reference bond distance and θ_A and θ_B are the reference angles. k_n are the force constants applied to the bond distance, angles and dihedral angles (ϕ_A, ϕ_B and ϕ_C). A series of λ windows are performed to gradually apply the restraints and then decouple the Coulombic and vdW interactions of the ligand from the binding site. The binding free energy, ΔG^o_b, can then be determined by:

$$\Delta G^o_b = -\Delta G^{prot}_{elec+vdW+restr} + \Delta G^{solv}_{elec+vdW} + \Delta G^{solv}_{restr\_on} \tag{2.21}$$

As two separate simulations are performed, uncertainties are calculated by the root mean square:

$$\Delta G^o_{b\_error} = \sqrt{\sigma_1^2 + \sigma_2^2} \tag{2.22}$$

The methodology described is adapted from a GROMACS absolute binding free energy tutorial, based on the work of Boyce et al.[195] and Aldeghi et al.[193]

2.4.2 Relative free energy perturbation

Relative FEP calculations are beneficial for evaluating the binding affinity of chemically similar compounds, such as in lead optimisation. A thermodynamic cycle is constructed so that the vertical legs involve making a simple modification to a scaffold, with the compound in the solvent phase on one side and the compound in complex with the receptor on the opposite side of the cycle (Figure 1.8). The change in free energy for each of these alchemical transformation is measured. Providing that the overall binding mode of the compound is conserved, it is possible to determine the relative difference in the free energy of binding, $\Delta\Delta G$, between the two compounds.

During an alchemical transformation, the force field parameters assigned to the 'disappearing' atoms on the ligand are slowly decoupled from the system, while the parameters for the 'appearing' atoms are introduced over a series of λ windows. Figure 2.11 shows a representation of this. The hydrogen derivative of the aryl ring is being turned off, while the methoxy aryl derivative is being introduced to the system. Typically, the perturbation being made to the ligand must not be too large, as accurate predictions require an overlap of phase space between neighbouring λ windows. There is no defined rule to the amount of atoms that can be perturbed. However, Cournia et al.[196] suggest that a change of a few atoms can be handled routinely, while larger perturbations are possible depending on the energy landscape of the system. To ensure a good overlap of phase space, the number of λ windows and the length of the simulations should be optimised.

The requirement for phase space overlap between incoming and outgo-

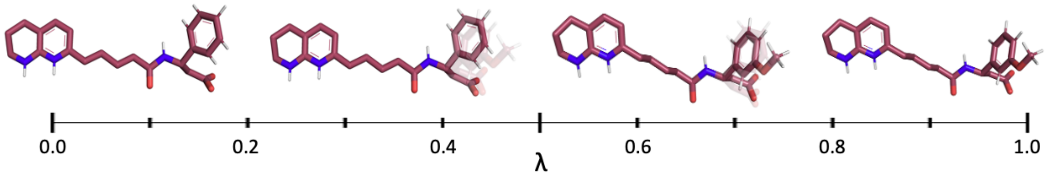


Figure 2.11: Alchemical transformation of a ligand with the progression of the λ variable used in relative FEP calculations. When λ=0 the 3-position of the aryl ring contains a hydrogen atom and when λ=1, there is a methoxy substituent.

ing atoms means that atom mapping is important. This refers to how the perturbed atoms are aligned with each other within the setup of the simulation. Maximum common substructure (MCS) alignment is a common approach, which maximises topology overlap, regardless of atom type and bond valences. Additionally, when performing multiple relative FEP simulations based on a single scaffold, it is sensible to choose a perturbation map that provides the smallest number of changes between compounds. Although the simplest way to compare binding free energies between a series of compounds based on a common scaffold may be to perform perturbations from a single reference compound, this is not always efficient. Therefore, arranging the compounds by similarity and performing a set of sequential perturbations can be the most constructive approach. A closed cycle is often created, so that energies sum to zero and it is possible to detect errors.

Single vs. dual topology approaches

When setting up a relative FEP simulation, there are two ways in which the ligand topology can be constructed. These are the single and dual topology approaches. In the single topology approach,[197] the topology is designed so that it contains sites that correspond to both molecules. For example, the top schematic in Figure 2.12 shows the perturbation of an -OH group to a -CH_3 group, using the single topology method. In the initial state, two hydrogen atoms, corresponding to the -CH_3 group, are introduced as dummy atoms. During the intermediate λ windows, parameters are scaled so that the oxygen atom and the dummy atoms become non-physical atoms, which then become a carbon atom and two hydrogen atoms in the final state. In contrast, atoms conserve their atom types in the dual topology approach.[198] As shown in the bottom schematic in Figure 2.12, both substituents exist in the topology as a whole. Progression along the λ windows involves the incoming carbon and hydrogen atoms changing from dummy atoms to fully interacting particles, and vice versa for the outgoing oxygen and hydrogen atoms.

Less atoms are perturbed in the single topology approach. However, dual topology approaches are simpler to construct and have the advantage of being able to sample the configurational space while being decoupled, aiding convergence.[199] Although more efficient, dual topology approaches can lead to “end-point catastrophes", where instabilities are created as the result of surrounding atoms clashing with the incoming or outgoing perturbed atoms. To prevent this, a soft-core potential can be used.[200,201] This eliminates the singularities at each end point by progressively scaling interactions of outgoing atoms and incoming atoms. The short range repulsive term in the standard Lennard-Jones potential is scaled to allow “soft" overlap of vdW spheres at regions surrounding incoming and outgoing atoms. The Coulombic term in the potential is also scaled to avoid abnormal electrostatic interactions between the softened atoms and their environment. To

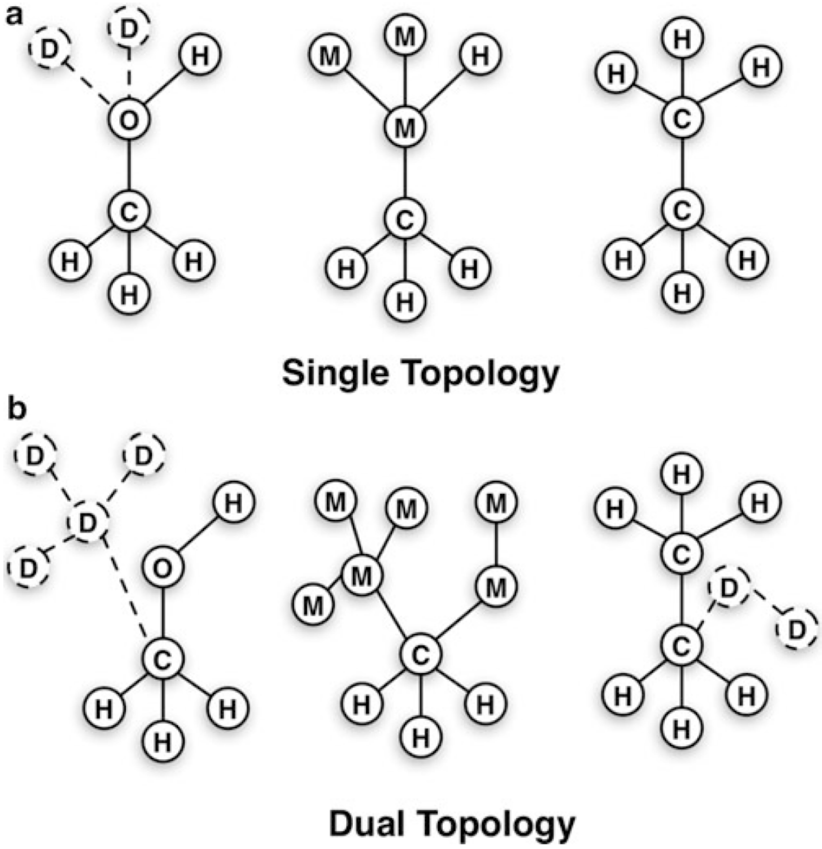


Figure 2.12: Single and dual topology approaches for constructing an alchemical path between ethane and ethanol. D represents a dummy atom, which has its non-bonded interactions decoupled from the system. M represents a non-physical intermediate atom, which exists at $\lambda \neq 0$ and $\lambda \neq 1$. [Reproduced with permission from reference 198.]

further prevent issues at the end points of the transformation, it is common to use smaller increments of λ as it approaches zero and one.

2.4.3 Energy evaluation

So far, only the Zwanzig formula (Equation 2.19) for approximating free energies has been discussed. The Bennett acceptance ratio (BAR) is another way in which free energy differences can be estimated.[98] This method uses data from sampling configurations in two states. For example, the forward calculation could be performed from λ=0 to λ=1, followed by the backward transformation from λ=1 to λ=0. Running the forward and backward transformation is also a useful way to check for convergence of the free energy, as the two ΔG values should be equal with opposite signs. Bennett derived this method from the following equation:

$$\Delta G = G_1 - G_0 = -k_B T \ln \frac{\langle w \exp(-\beta U_1) \rangle_0}{\langle w \exp(-\beta U_0) \rangle_1} \tag{2.23}$$

where β=1/$k_B T$. The weighting function, w, was obtained by minimising the variance of the free energy. The resulting free energy difference for the BAR method is given by Equation 2.24, where $G_1 - G_0$ is solved self consistently.

$$\Delta G = G_1 - G_0 = -k_B T \ln \frac{\left\langle \frac{1}{1+\exp(-\beta(U_1 - U_0) + \beta(G_1 - G_0))} \right\rangle_0}{\left\langle \frac{1}{1+\exp(\beta(U_1 - U_0) - \beta(G_1 - G_0))} \right\rangle_1} \tag{2.24}$$

This method significantly improves free energy estimations compared to traditional FEP.[202,203] Additionally, less overlap of configurational space between λ windows is required, meaning fewer windows are necessary and calculations are faster to perform. A variation of the BAR method is the multistate Bennett's acceptance ratio (MBAR).[204,205] This method combines simulation data from multiple states, compared to just two, which further improves free energy estimations.

Convergence and sufficient sampling are important properties for accurate energy estimations. Small changes made to a ligand can result in large rearrangements of protein binding sites. One way to measure for sufficient sampling is to calculate, for each λ window, the RMSD of the protein as a whole, or of the binding site, with respect to its initial conformation. A consistent RMSD within each window means that the protein structure is stable and that a converged free energy value is likely to be measured for that value of λ. An additional way to check for convergence is to plot the cumulative free energy differences over varying MD time ranges, as outlined by Klimovich et al.[206]

2.4.4 Lambda dynamics

Lambda dynamics,[106,207] or λ-dynamics, is an alternative alchemical free energy method to FEP. Much like FEP, λ-dynamics has recently become much more feasible, since its development by Brooks et al. in 1995, due to the advancement of computational resources such as computer clusters and GPU acceleration. Similar to relative FEP calculations, λ-dynamics performs best when applied to lead optimisation tasks where knowing the difference in binding affinity between small changes on a common scaffold is required. However, λ-dynamics has the ability to estimate the relative $\Delta\Delta G$ values of multiple different variations of a scaffold in one single simulation, negating the need to do a separate simulation for each pairwise set of compounds.

Recent advancements to λ-dynamics include MSλD and adaptive landscape flattening (ALF).[208,209] In MSλD, it is possible to simultaneously perform perturbations on more that one substitution site of a scaffold, which is more realistic of the types of changes that are made to a compound in typical SBDD projects. For example, Figure 2.13 shows a set of perturbations made to a tetrahydroquinolone scaffold, which is the framework for a set of

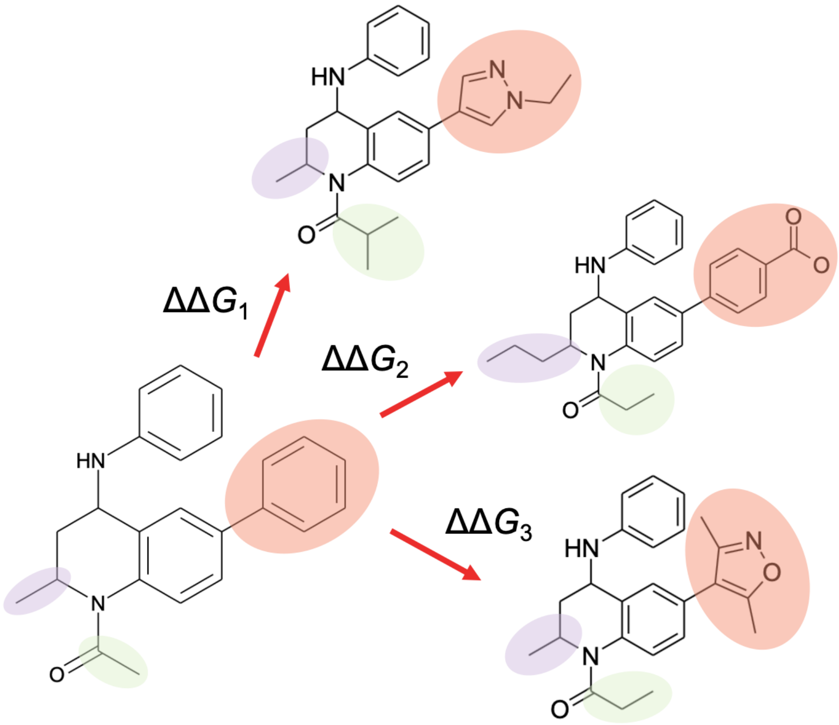


Figure 2.13: Substitutions made on three sites of a tetrahydroquinolone scaffold. Relative binding free energies $\Delta\Delta G_1$, $\Delta\Delta G_2$ and $\Delta\Delta G_3$ can be obtained from a single MSλD simulation.

BRD4 inhibitors (discussed in Chapter 6). In this example, there are three substitution sites on the scaffold, with four possible substituents on the orange site, three on the green site and two on the purple site. Values for $\Delta\Delta G_1$, $\Delta\Delta G_2$ and $\Delta\Delta G_3$ can be obtained from a single simulation. Additionally, $\Delta\Delta G$ values for the combinatorial set of substituents are obtained, meaning these types of calculations are significantly more efficient than traditional FEP calculations. Previous studies have shown MSλD calculations to be ~50 times faster than FEP calculations and obtain results within a similar accuracy.[208] In these studies, free energy values were calculated within 0.9 kcal mol^{-1} of experiment. ALF facilitates better sampling of physically meaningful states, meaning larger and more complex perturbations can be performed, while maintaining precision and accuracy.[209]

In λ-dynamics, λ is treated as a dynamic variable that propagates through the simulation, along with the coordinates. This is in contrast to traditional FEP, where several simulations are performed at fixed values of λ. Furthermore, the introduction of additional λ coordinates means that λ-dynamics can be performed between more than two systems of interest. The potential energy for MSλD with ALF is calculated using Equation 2.25.[208,209]

$$\begin{aligned} V(X,\{x\},\{\lambda\}) =& V_{env}(X) + \sum_{s=1}^{M}\sum_{i=1}^{Ns}\lambda_{si}(V(x_0,x_{si}) + V(x_{si},x_{si})) \\ &+ \sum_{s=1}^{M}\sum_{i=1}^{Ns}\sum_{t=s+1}^{M}\sum_{j=1}^{Nt}\lambda_{si}\lambda_{tj}V(x_{si},x_{tj}) + V_{Bias}(\{\lambda\}) \end{aligned} \tag{2.25}$$

In this equation, X represents the atoms that are present in the system for the entire simulation and x_0 is their coordinates. These include the solvent atoms, receptor atoms and the atoms in the common core of the ligand, which are not being perturbed. The coordinates of substituent i at site s are given by x_{si}, N_s is the total number of substituents that are on one site, while M is the total number of sites. The interaction potential of the environment atoms ($V_{env}(X)$) is not scaled by λ, whereas the interaction potentials of substituent i at site s interacting with the environment atoms ($V(x_0, x_{si})$) and itself ($V(x_{si}, x_{si})$) are scaled by λ variables. The interaction potential between substituents i and j at different sites ($V(x_{si}, x_{tj})$) is also scaled by the λ variables. The intramolecular properties are not scaled by λ to ensure that the geometry and basic connectivity of the ligand is preserved even when a substituent is in a non interacting state. The biasing potential, V_{Bias}, is discussed further on in this section.

A substituent, i, on a scaffold is fully interacting when λ_{si}=1 and λ_{sj}=0 for $i \neq j$. At this point, the system is at a physically meaningful end state. In practice, however, if a substituent has $\lambda \geq 0.8$, it is often counted. To ensure that only one substituent at a given site is interacting at a physically relevant end point, a set of constraints are employed.

$$\sum_{i=1}^{N_s} \lambda_{si} = 1 \qquad (0 \leq \lambda_{si} \leq 1) \tag{2.26}$$

These constraints are maintained by the implicit constraints shown in Equation 2.27.

$$\lambda_{si} = \frac{\exp(c \sin \theta_{si})}{\sum_{j=1}^{N_s} \exp(c \sin \theta_{sj})} \tag{2.27}$$

In MSλD, θ_{si} are the dynamic variables, which are treated as volume-less particles with mass m_θ. To ensure good sampling of end points and stability of the calculation, a value of $c = 5.5$ has been found to be optimal.[210] The dynamics of the system are generated from the extended Hamiltonian, given by:

$$H_0(X, \{x\}, \{\lambda\}) = T_x + T_\theta + V(X, \{x\}, \{\lambda(\theta)\}) \tag{2.28}$$

where T_x and T_θ are the kinetic energies of the atomic coordinates and the θ variables, respectively. Finally, the free energy difference between two derivatives is approximated using the probability of finding the system in each of the physically meaningful end states, as shown in Equation 2.29.[208]

$$\begin{aligned} \Delta\Delta G(\lambda_{\{si\}} \rightarrow \lambda_{\{sj\}}) \approx & - k_B T \ln \frac{P(\{\lambda_{sj}\} \geq \lambda_c)}{P(\{\lambda_{si}\} \geq \lambda_c)} \\ & - (V_{bias}(\{\lambda_{sj}\} = 1) - V_{Bias}(\{\lambda_{si}\} = 1)) \end{aligned} \tag{2.29}$$

As λ_c approaches 1, this equation becomes exact. However, a value of λ_c=0.8

is sufficient.

Using ALF, the biasing potential energy term ($V_{bias}(\{\lambda\})$) can be calculated using Equations 2.30 to 2.33.[209] This function is important, as to obtain accurate free energy results it is necessary to have sufficient sampling of all physically meaningful end states. In alchemical transformations, sampling can be limited by high energy barriers and so ALF is applied to calculate the biases needed to flatten the energy surface between end points. The fixed bias (V_{Fixed}) ensures that all end points have a similar free energy and can be sampled within the same simulation.[211] The potential $V_{DiagQuad}$ is a bias to flatten the quadratic barriers shown by λ-dynamics. An end point bias (V_{End}) is also used to overcome the deeper energy wells that exist at the end points. Variables, ϕ_{si}, ψ_{si}, $\omega_{si,sj}$ and α, within these equations are altered during the ALF phase of the MSλD simulation, until good sampling is observed for all physically meaningful end points.

$$V_{Bias} = V_{Fixed} + V_{DiagQuad} + V_{End} \tag{2.30}$$

$$V_{Fixed} = \sum_{s}^{M} \sum_{i}^{N_s} \phi_{si}\lambda_{si} \tag{2.31}$$

$$V_{DiagQuad} = \sum_{s}^{M} \sum_{i}^{N_s} \psi_{si}(\lambda_{si}^2 - \lambda_{si}) \tag{2.32}$$

$$V_{End} = \sum_{s}^{M} \sum_{i}^{N_s} \sum_{j \neq i}^{N_s} \frac{\omega_{si,sj}\lambda_{si}\lambda_{sj}}{\alpha + \lambda_{si}} \tag{2.33}$$

2.4.5 Limitations of alchemical free energy calculations

Aside from the limitations already discussed, such as end-point catastrophes, the requirement for phase overlap and sufficient sampling, there are a number of other aspects that should be considered for accurate and reliable estimations of protein-ligand binding free energy.[196]

Initial structures are a crucial part of the free energy calculation process. A recent study by Suruzhon et al.[212] found that the choice of crystal structure can have an impact of greater than 1 kcal mol^{-1}, which roughly translates to a pIC$_{50}$ of 0.7. Furthermore, sampling rare events such as ligand torsional motions can have an even greater impact on free energy values. This work affirms the need for extensive equilibration and averaging values from multiple repeat simulations to improve the uncertainty around the free energy. A reasonable initial binding pose of the ligand is also a key factor in free energy calculations. Without the use of enhanced sampling, capturing events with high-energy barriers such as changing binding modes or ligand torsions is unlikely. Therefore, it is important to use information from existing experimental data of similar known binders. The influence of binding pose and crystallographic water molecules on binding free energy is explored for a bromodomain-containing protein system in Chapter 5.[113]

Alchemical free energy calculations that involve a change in charge when going from the initial to the final state of a ligand are generally difficult to perform.[196] One obstacle is that the time required to rearrange the surrounding solvent network as a result of a change in net charge will need to be considerably longer and the overlap between adjacent λ windows will be poorer, compared to perturbations with a conserved net charge.[196] Additionally, transformations are unable to model protonation or electronic polarization changes that may occur when a ligand goes from solvent to a protein environment. Recently, King et al.[213] demonstrated the impact of protonation and polarization conditions on the predictive accuracy of free energy calculations and introduced a MBAR/PBSA approach to improve results. In Chapter 6, a MSλD protocol for accurately predicting changes in ligand charge is outlined.

Even with the most rigorous set up and extensive simulation times, the accuracy of free energy estimations fundamentally relies on the quality of the force field and force field parameters that describe the protein-ligand system. Without the right potential energy function, calculations will converge to the wrong answer.[196] Different biomolecular force fields can also arrive at different free energy values. Pérez-Benito et al.[214] performed FEP with two different software and force field methods. For one set of data, calculations using Schrödinger-Desmond FEP+[215] with the OPLSv3e force field[216] and calculations using GROMACS[175] with the AMBER force field[122] both arrived at the same error from experiment. However, for a second data set, the FEP+ method had an error of 1.17 kcal mol^{-1} from experiment, whereas the GROMACS method resulted in an error of 1.90 kcal mol^{-1}. This study demonstrates the importance of choosing the force field and parameter set that best describes the ligand of interest. Parameters can be validated by comparing energy minimised values of intramolecular ligand properties, such as bond lengths, with QM optimised values. Best practice often requires extensive validation and optimisation of ligand force field parameters, as outlined in

Chapter 3.[27]

2.4.6 Conclusions

The advance of computational power to perform MD and free energy calculations has transformed the possibilities of CADD over the last couple of decades. Starting from X-ray crystal structures or docked complexes, sampling methods aid the understanding of protein-ligand binding, which in turn provides guidance for compound design and lead optimisation. Despite the pitfalls associated with alchemical transformations, they are regarded as among the most rigorous techniques for predicting ligand binding affinity. λ-dynamics and MSλD methods for estimating relative binding free energies show promise as a way to overcome the time limited scalability of traditional FEP calculations. Their implementation in early drug discovery projects could provide rapid and reliable estimations of binding affinity for large compound series, saving huge amounts of synthetic effort. The following chapters outline how a variety of computational methods have been applied to drug discovery projects in a collaboration between UoN and GSK.

Chapter 3

Force Field Parameter Development for αv Integrin Inhibitors

3.1 Introduction

The urgent need for new treatments for the chronic lung disease IPF motivates research into antagonists of the RGD binding integrin αvβ6, a protein linked to the initiation and progression of the disease. In this chapter, we present the development of new force field parameters for a scaffold of a series of αvβ6 inhibitors.

3.1.1 Idiopathic pulmonary fibrosis

IPF, a chronic disease characterised by progressive scarring of the lungs, has a survival rate of two to four years upon diagnosis.[217] Despite being considered a rare disease, 5000 new cases are estimated each year in the UK and are reported more frequently than leukemia and brain and stomach cancers.[218–220] Symptoms of IPF include shortness of breath, weight loss and fatigue. These are the result of respiratory insufficiency, caused by a 'honeycombing effect' of the lungs (Figure 3.1).[221] Furthermore, the progression of the disease can require patients to need constant hospitalisation or hospice care.[222] There are currently two approved treatments of IPF but despite slowing down progression, each has significant side effects and are not curative.[223] Clearly, IPF is an important public health issue and the need for alternative treatment is urgent.

The exact mechanisms leading to the pathogenesis of IPF are unclear. However, aggregates of actively proliferating fibroblasts, termed fibroblastic foci, are a key feature in IPF pathology.[224] Lung biopsies indicate that there is a direct correlation between these foci, progression of the disease and shortened survival.[225] Consequently, it has been suggested that new therapies should target the regulation of fibroblast functions, rather than the inflammatory response.[226] The transforming growth factor β1 (TGF-β1) plays a key role in controlling these functions, by responding to tissue injury or infection and mediating tissue repair. However, increased activity of TGF-β1 can lead to tissue inflammation in uninjured areas and scar formation. Therefore, prevention of TGF-β1 activation is important for new therapies

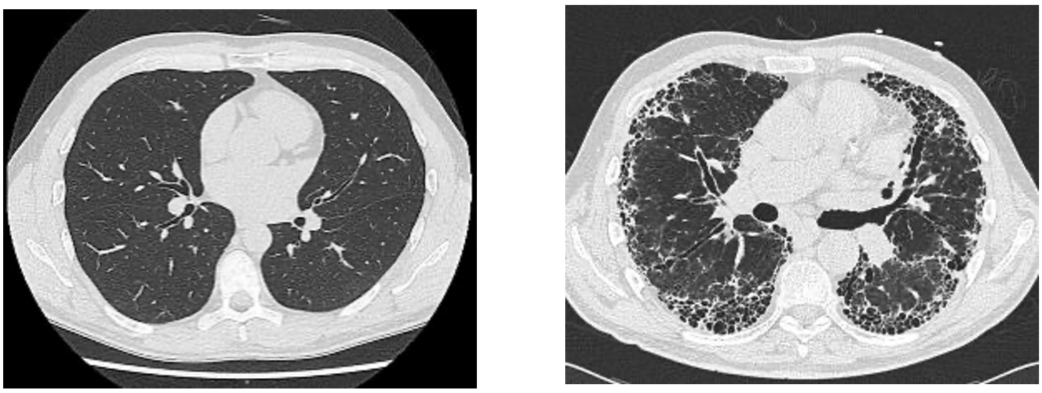


Figure 3.1: (Left) CT scan of a pair of healthy lungs. (Right) Typical CT scan from a patient with IPF. Scarring and honeycombing of the lungs is observed [Reproduced with permission from reference 220].

of IPF.[227,228] One mechanism of action of TGF-β1 involves the integrin αvβ6. Latent TGF-β1 is stored in excess in the extracellular matrix and binds to the extracellular head region of the transmembrane protein. A drug antagonist of the αvβ6 receptor could treat IPF through its active site binding in place of TGF-β1.

3.1.2 Lead compounds for αv antagonism

Latent TGF-β1 is activated upon binding to integrin αvβ6 through an Arg-Gly-Asp(RGD)LXX(I/L) motif (where X is any amino acid) in the prodomain. Key binding interactions are highlighted in Figure 3.2. The Arg^{RGD} side chain interacts with the carboxyl group on (αv)-Asp218 through bidentate hydrogen bonds. The carboxyl group on Asp^{RGD} coordinates with a Mg^{2+} ion in the metal ion-dependent adhesion site (MIDAS) of the β6 subunit. The Mg^{2+} ion is flanked by two Ca^{2+} ions, one of which is called the adjacent to MIDAS (ADMIDAS). Allosteric antagonists that mimic an RGD sequence, and therefore these interactions, are of particular interest.

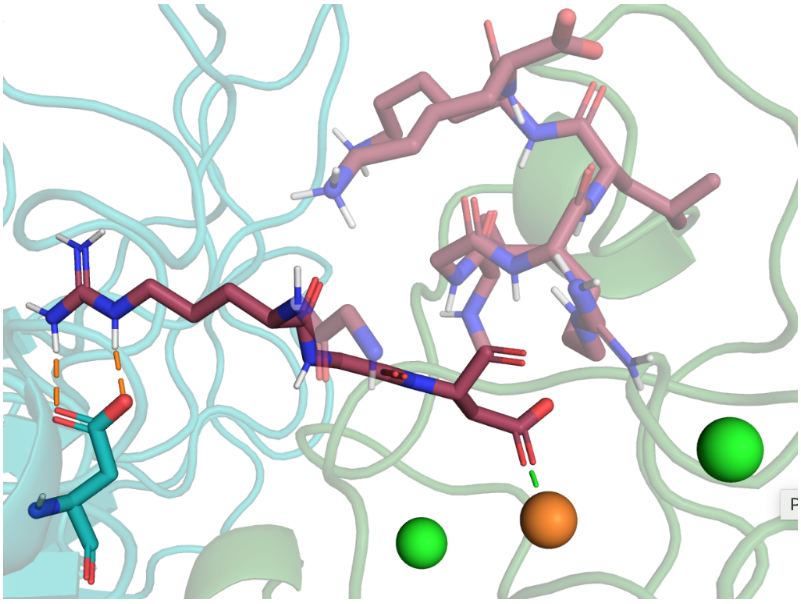


Figure 3.2: TGF-β1 bound in the active site of αvβ6, showing the αv subunit (faded teal), the β6 subunit (faded green), the backbone of the ligand (maroon) and heteroatoms (blue and red). The RGD sequence of TGF-β1 is shown in bold. The bidentate (αv)-Asp218 hydrogen bonding and the metal chelate interaction with the Mg^{2+} ion (orange) are indicated by dashed lines. Asp218 is part of the αv subunit and is shown in bold teal. The Ca^{2+} ions are shown in green.

Successful integrin antagonists to reach the market treat diseases such as multiple sclerosis, Crohn's disease and acute coronary syndrome.[229] However, to date there have been no approved drugs specifically targeting αv integrins. Despite this, there a number of αv antagonists to receive interest and reach clinical trials.[230] A well documented inhibitor of αvβ3 and αvβ5 is cilengitide (Figure 3.3), a cyclic RGD mimetic that reached phase three clinical trials. Cilengitide binds to these receptors, preventing signalling and inducing apoptosis of tumour cells.[231] Although this compound had a promising start, it did not meet regulatory approval, due to lack of activity and not significantly increasing overall survival in patients.[232] Nevertheless, the majority of integrin inhibitors currently in clinical development are RGD-binding and remain a key area of research.

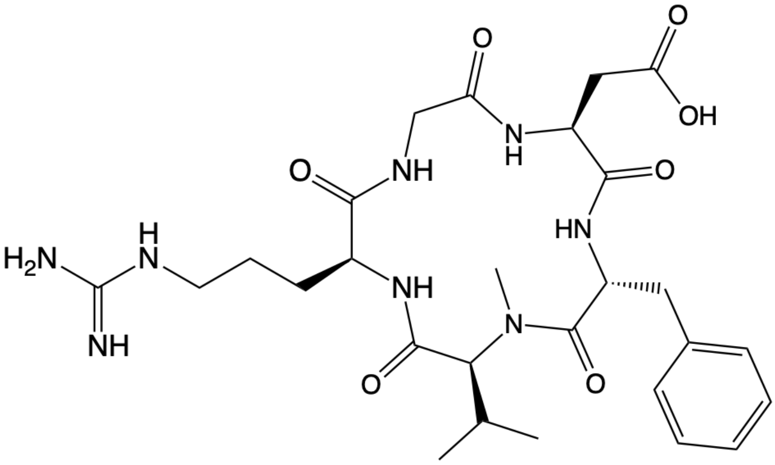


Figure 3.3: The structure of cilengitide, an αv inhibitor to reach stage three clinical trials.

A collaboration between GSK and UoN has presented structure activity relationship (SAR) studies of 36 novel compounds.[8] Analogues of a parent compound were synthesised and pIC_{50} values were used to measure activity and selectivity against αvβ3, αvβ5, αvβ6 and αvβ8. Figure 3.4 shows the structure of the compound's scaffold. A 1,2,3,4-tetrahydro 1,8-napthyridine group at one end of the compound mimics the Arg residue in the RGD tripeptide. This moiety is of particular interest in medicinal chemistry as 1,8-napthyridines and their derivatives are found in many natural substances with biological activities.[233] The carboxyl on the opposite end of the compound mimics the Asp residue and binds to the Mg^{2+} MIDAS ion. Substituents on the aryl group are varied, influencing the potency and selectivity of the compound.

SAR shows that subtle changes in aryl substituent can lead to substantial effects. This is highlighted by the profound change in activity of enantiomers when a CF_3 group is added at the 3-position of the aryl ring. The (*S*)-enantiomer has a pIC_{50} of 7.1, while the (*R*)-enantiomer has a value of

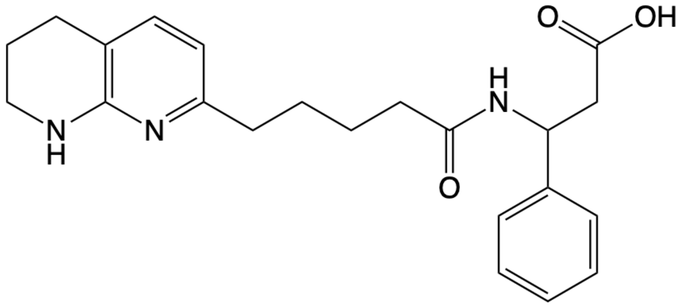


Figure 3.4: The structure of an RGD mimetic functioning as the parent compound of an αvβ6 antagonist series.

5.2. Other aryl substituents considered in the SAR include H, Cl, F, methyl, propyl, methoxy, nitrile, OCF_3, SO_2Me, phenyl and OCH_2O. Possible derivatives of the scaffold is not limited to those that have been previously synthesised. Taking into account constraints such as synthetic accessibility and the need for drug-like properties, there are still a huge number of viable candidates. This motivates the use of computational approaches to investigate the relationship between derivatives and their activity, to save cost and synthetic resources.

3.1.3 The CHARMM force field

The aim was to use MD and FEP simulations to model the binding interactions of αvβ6 with a series of inhibitors, based on the GSK/UoN scaffold discussed.[8,234] The potential energy function chosen for these calculations was the CHARMM force field,[119] as shown in Equation 3.1.

$$
\begin{aligned}
V = & \sum_{bonds} k_b(b-b_0)^2 + \sum_{UB} k_{UB}(r_{1,3}-r_{1,3;0})^2 + \sum_{angles} k_\theta(\theta-\theta_0)^2 \\
& + \sum_{impropers} k_\varphi(\varphi-\varphi_0)^2 + \sum_{dihedrals} k_\chi(1+cos(n\chi-\delta) \\
& + \sum_{nonbonded} \epsilon_{ij}[(\frac{R_{minij}}{r_{ij}})^{12} - (\frac{R_{minij}}{r_{ij}})^6] + \frac{q_iq_j}{4\pi r_{ij}\epsilon_{0ij}} \\
& + \sum_{residues} V_{CMAP}(\phi,\psi)
\end{aligned}
\qquad (3.1)
$$

The CHARMM force field contains functions typical to most biomolecular force fields (as discussed in Chapter 2). Bond lengths and angles are described by harmonic potentials, dihedral angles are treated by their force constant, k_χ, the multiplicity, n, and the phase, δ, and non-bonded interactions are calculated using the Lennard-Jones 12-6 potential[116] and the Coulomb potential. In the Lennard-Jones term, R_{minij} is the equilibrium position of two particles and relates to the van der Waals (vdW) radius by $R_{min} = \sqrt[6]{2}\sigma$. There are also some additional terms in the CHARMM force field. The Urey-Bradley (UB) function describes the 1-3 bond distances, where atoms 1, 2 and 3 are connected, using a force constant, k_{UB}, and equilibrium distance, $r_{1,3;0}$. Improper dihedral angles are also accounted for using a harmonic potential with a force constant, k_φ, and equilibrium value, φ_0. The final CMAP term serves as an energy estimation for the conformational flexibility of a peptide backbone.[235]

Parameter development

Terms in the force field not obtained directly from molecular coordinates are parameters that can be developed and optimised based on unique atom types. This allows different types of atoms and molecular connectivity to be treated using the same set of equations. The CHARMM force field has been developed beyond the treatment of proteins. The CHARMM General Force Field (CGenFF) is an extension of the force field, which contains parameters to describe small drug-like molecules.[123] However, not all atom types in the RGD mimetic of interest, specifically the 1,2,3,4-tetrahydro 1,8-naphthyridine group, exist within the CGenFF. Therefore, parameters consistent with the CHARMM force field have been developed to enable the computational study of the potential αvβ6 antagonists.

3.2 Materials and Methods

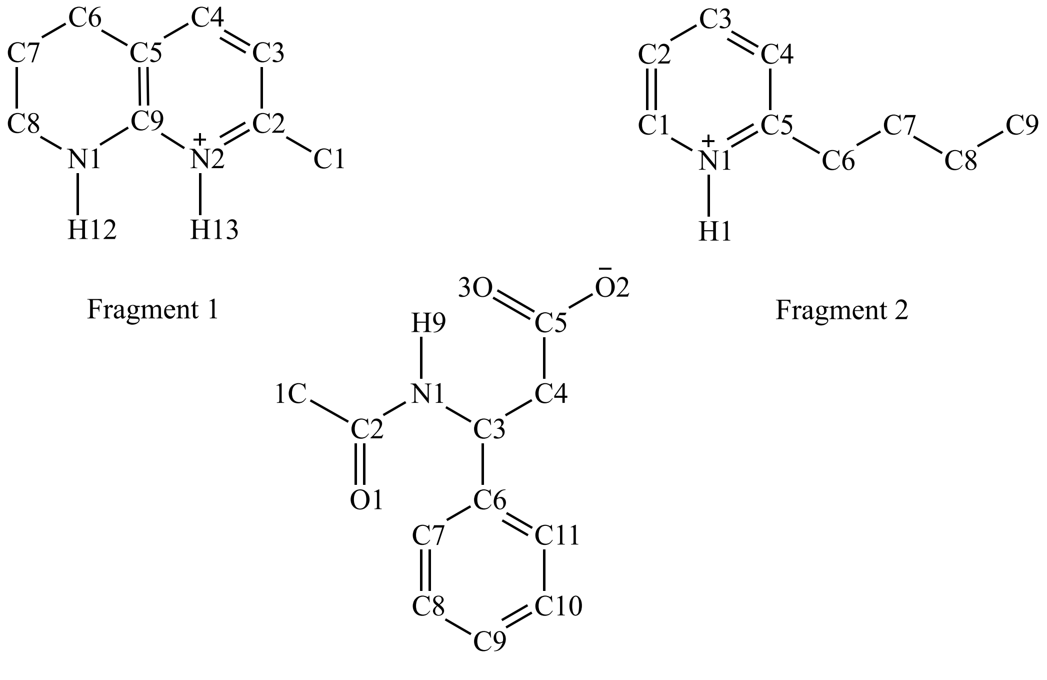
The key to developing compatible parameters is consistency with how the CHARMM force field has been developed. A systematic protocol for parameter development, outlined by Vanommeslaeghe et al.,[236,237] was followed to attain a sufficient level of accuracy in a timely manner. Once parameters were optimised, as outlined below, the process was repeated so that all convergence criteria were met. Two iterations, at least, of optimisation are typically required due to the sensitivity of non-bonding interactions to intramolecular properties.

As the GSK/UoN scaffold does not contain an extended peptide backbone, the CMAP term in the potential energy function was discarded. The UB and improper angle terms were also left undefined as these are typically only used as an additional energy correction when the other terms are not satisfactory.

3.2.1 Target fragments and initial guess

For computational efficiency, the scaffold was split into three fragments (Figure 3.5). Fragment 1 contained the 1,2,3,4-tetrahydro 1,8-napthyridine group, the focus of fragment 2 was the alkyl chain and fragment 3 contained the amide bond, carboxyl group and aryl ring. The deprotonated carboxyl group and positively charged naphthyridine fragment were important for modelling the key RGD interactions and the physiological form of the compound.

Initial guess parameters were obtained using a CGenFF atom typing program.[236,237] This program finds chemically similar groups already available in the CGenFF force field and returns a penalty score based on the recommended level of optimisation required for each set of parameters. For example, fragment 1 was matched with 1,8-napthyridine, which is pre-existing in CGenFF, and was assigned an overall penalty score of 223. For penalty scores between 10 and 50, basic validation is recommended, whereas extensive optimisation is necessary for scores above 50. Fragment 1 was the only fragment to return a score above 10. Therefore, optimisation efforts were concentrated on the partial charges and intramolecular parameters of the 1,2,3,4-tetrahydro 1,8-napthyridine group.



Fragment 1

Fragment 2

Fragment 3

Figure 3.5: Fragments 1, 2 and 3, derived from the scaffold of the potential αv antagonist. Atom labels correspond to those used in the parameterisation process.

3.2.2 Intermolecular parameters

Target data was generated by QM calculations, using the package QChem.[238] Following, Vanommeslaeghe et al., for each possible hydrogen bonding interaction, a complex was built of the MP2/6-31G* optimised fragment and a single water molecule in the TIP3P geometry.[182] The water molecule was initially placed in an “ideal" position for hydrogen bonding and the interaction distance was optimised at the HF/6-31G* level. From this optimised distance, an interaction energy can be determined. While higher levels of theory for QM calculations may give more accurate target data, their use slows down the parameterisation process and could lead to an imbalance

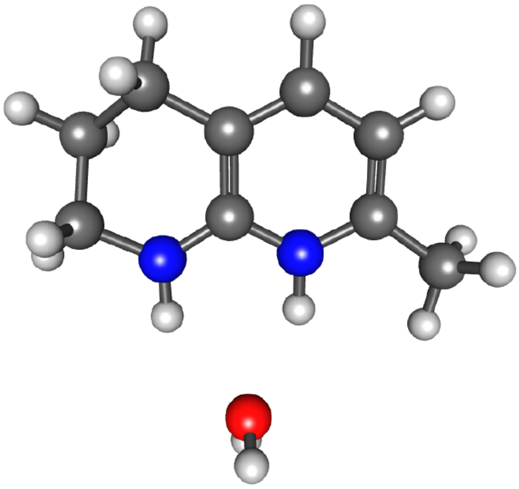


Figure 3.6: Bidentate interaction between a water molecule and the polar hydrogen atoms on fragment 1.

with parameters in other parts of the force field. Usually, this process is repeated for each hydrogen bonding atom. However, due to the bidentate interaction of fragment 1 with the receptor, a single water molecule was used to replicate this interaction (Figure 3.6).

Once QM values were determined, they were compared with the interaction distance and energy obtained from an energy minimisation calculation using the CHARMM force field and initial guess parameters. Partial charges were then optimised so that the model fragment-water interaction energies were within 0.2 kcal mol^{-1} and distances were within 0.1 Å of the QM target data. This was done by iteratively making small changes to the partial charges, while maintaining the overall net charge of the fragment.

3.2.3 Intramolecular parameters

Bond length and angle parameters were optimised so that values from MP2/6-31G* optimised fragments were replicated by energy minimisation calculations using the CHARMM force field. Penalty scores from the initial guess parameters indicated that, for fragment 1, one bond length and several bond angle parameters required attention. The penalty score for the N1-C9 bond length parameters was 145. Therefore, the force constant and equilibrium terms were modified until the CHARMM energy minimised bond length was within 0.03 Å of the QM optimised structure. Bond angle parameters were optimised in a similar way, so that angles were within 3° of the MP2/6-31G* optimised structure.

Optimisation of dihedral angle parameters used the MP2/6-31G* PES as target data. Due to the pseudo-planarity of fragment 1, the only parameters which required optimisation were those for the N1-C8-C7-C6 torsion. The initial method employed to generate the QM surface involved an optimisation with the dihedral angle constrained and using the resulting conformation for the subsequent dihedral optimisation. However, this produced irregularities in the PES, due to the ring flip as the dihedral angle passed through 0°. To obtain a smooth PES, geometry optimisations were performed using a planar initial structure for each new dihedral constraint. The potential energy term describing rotation around a dihedral angle in the CHARMM force field features parameters for the force constant, multiplicity and the phase. The same dihedral scan was carried out using the CHARMM force field, with these dihedral parameters set to zero. A Monte-Carlo Simulated Annealing (MCSA) protocol[239] with exponential cooling was used to minimise the root mean square error between the QM and MM energy profiles.

Aryl substituent parameters

To test how well the existing CHARMM force field parameters describe the compound derivatives, when an aryl substituent is attached, penalty scores were obtained for substituents H, F, CF_3, OCF_3, methyl and methoxy. From these scores, optimisation of partial charges on the CF_3 derivative was considered necessary. Dihedral angle parameters were also validated for the C-O-C-F torsion on the OCF_3 derivative. No further optimisation was needed for any other aryl substituent.

3.3 Results and Discussion

3.3.1 Partial charges

For the 1,2,3,4-tetrahydro 1,8-naphthyridine group (fragment 1 in 3.5), Table 3.1 shows the optimised interaction energies and distances for hydrogen bonding interactions with polar atoms on the antagonist at the HF/6-31G* level, and when using the CHARMM force field once partial charges were optimised. Interaction energies are within 0.2 kcal mol^{-1}, as recommended by parameter optimisation methodology.[240] Although the interaction distance between the fluorine atoms on the CF_3 derivative and the TIP3P oxygen atom is above the recommended 0.1 Å, similar disagreement has been acceptable in other major parameterisation efforts.[240]

Table 3.1: QM and MM interaction energies and distances of a single water molecule in a hydrogen bonding interaction with polar atoms on fragment 1 and the CF_3 aryl derivative. O represents the oxygen atom on the TIP3P water molecule. Distances are in Å and energies are given in kcal mol^{-1}.

		H12-O	H13-O	CF-O
HF/6-31G*	Distance	2.1	2.0	2.3
	Energy	-16.4	-16.2	-1.9
Optimised CHARMM Parameters	Distance	2.0	1.9	2.0
	Energy	-16.4	-16.4	-1.9
Difference	Distance	0.1	0.1	0.3
	Energy	0.0	0.2	0.0

3.3.2 Bond lengths, angles and dihedrals

Figure 3.7 shows the MP2/6-31G* bond lengths and partial charges assigned to the methylated 1,8-naphthyridine fragment. The similarity in bond lengths between the central carbon atom and adjacent nitrogen atoms suggests that the charge delocalisation extends across the ring to include both nitrogen atoms, rather than a localised aromatic pyridine-piperidine structure. This is reflected in the partial charges assigned to these atom types in the developed force field parameters.

Optimised values for the high penalty bond lengths and angles in fragment one are shown in Table 3.2. These values are also in close agreement with the bond lengths and angles in four naphthyridine containing molecules from the Cambridge Crystallographic Data Centre (CCDC) (Table 3.3).[241]

Due to the planarity of fragment 1, the only dihedral angle that needed optimising was the N1-C8-C7-C6 dihedral angle (Figure 3.8). Figure 3.9

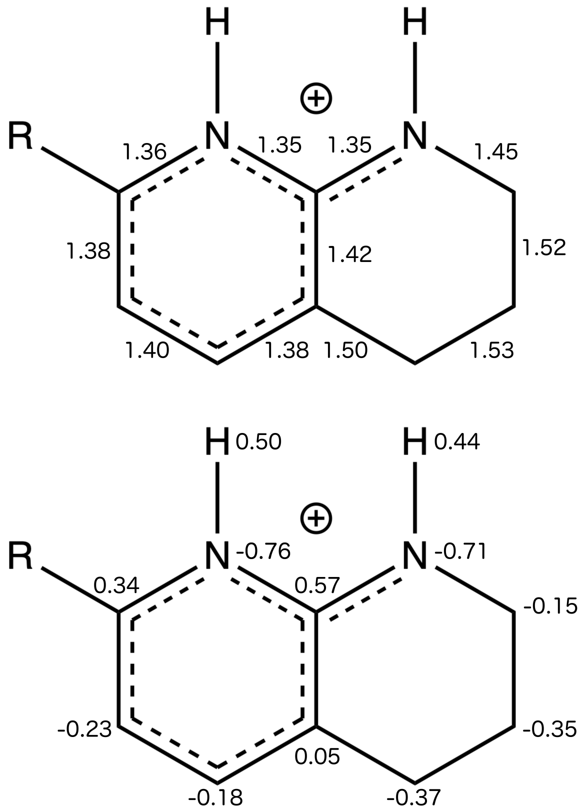


Figure 3.7: Bond lengths and atomic partial charges calculated at the MP2/6-31G* level for fragment 1. Bond lengths are shown in Å.

Table 3.2: Bond lengths and angles on fragment 1 from MP2/6-31G* geometry optimisation and energy minimisation using optimised CHARMM force field parameters. Bond lengths shown in Å and bond angles shown in degrees.

Bond/Angle	MP2/6-31G* Optimisation	Optimised CHARMM Parameters	Difference
N1 - C9	1.34	1.33	0.01
N1 - C9 - N2	119	118	1
C9 - N1 - H12	120	120	0
C9 - N1 - C8	122	119	3
C5 - C9 - N1	123	124	1

Table 3.3: Bond lengths and angles in molecules containing a dihydro 1,8-naphthyridine fragment. Atom indexing is taken from Figure 3.5. The structures are given by CCDC ID. Bond lengths are shown in Å and bond angles are shown in degrees.

Structure	N1-C9	N1-C9-N2	C9-N1-H12	C9-N1-C8	C5-C9-N1
MP2 Optimised	1.34	119	120	122	123
MM Optimised	1.33	118	120	119	124
COYFOM[242]	1.33	115	-	121	123
BEKWOE[243]	1.32	118	122	121	125
IPAGEN[244]	1.37	116	116	123	120
MERLNID[245]	1.36	116	121	118	121

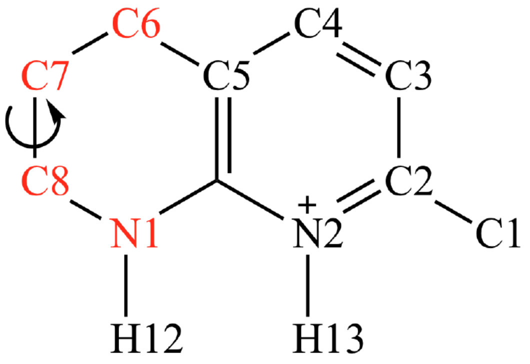


Figure 3.8: Fragment 1 with the dihedral angle requiring parameter optimisation highlighted.

shows the QM and the MM PES of the N1-C8-C7-C6 dihedral angle. There is a closer match between MM and QM surfaces once dihedral parameters, k, n and δ, are optimised using a MCSA protocol. The penalty score returned for the C-O-C-F dihedral parameters in the OCF_3 derivative was 98, implying necessity for optimisation. However, the QM and MM PES around the torsion indicated a sufficient match between the curves when using the initial guess parameters; therefore, no optimisation was performed.

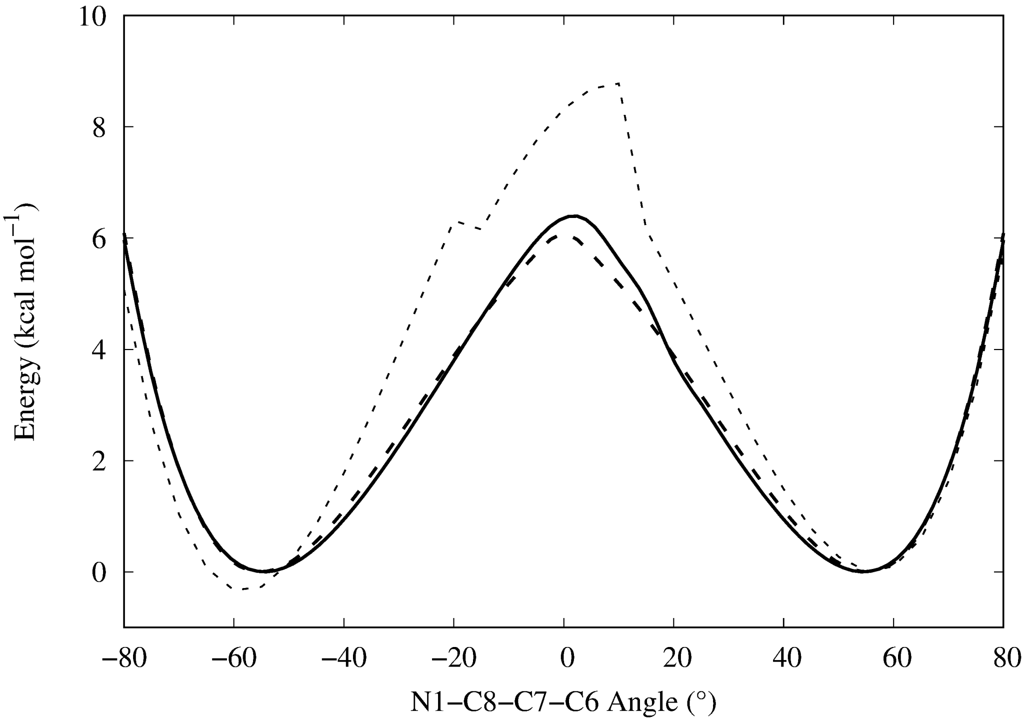


Figure 3.9: Dihedral scan around the N1-C8-C7-C6 angle showing the MP2/6-31G* surface (dashed line), the CHARMM initial guess (dotted line) and the CHARMM surface after the parameters were optimised (solid line).

3.4 Conclusions

Computational methods aid the investigation of potential antagonists of αvβ6, a therapeutic target for the treatment of IPF. Consequently, a new set of parameters compatible with the CHARMM force field has been developed for the scaffold of a series of αv inhibitors. In particular, parameters have been developed for 1,2,3,4-tetrahydro 1,8-naphthyridine. As naphthyridine is an important moiety in medicinal chemistry,[233] we expect these new parameters will find utility in future computational studies, beyond our own.

Starting from initial guess values, partial charges and intramolecular parameters have been optimised to match QM target data within the desired accuracy. Parameters for three fragments of the UoN/GSK scaffold can now be combined to accurately model the full compound using MM. Parameters can be found in the Supporting Information of reference 27. In the next chapter, these parameters are used and validated through MD and FEP simulations.

Chapter 4

Molecular Simulation of αv Integrin Inhibitors

4.1 Introduction

In the previous chapter, we outlined how the involvement of integrin αvβ6 in the progression of several tumour types and diseases, including IPF, makes it an important target for the development of drug compounds.[7,230,246] CHARMM force field parameters were optimised for an RGD mimetic, which serves as the scaffold for a series of αvβ6 inhibitors. In this chapter, we describe MD simulations of αvβ6 in complex with its natural ligand, TGF-β1, and each of the RGD mimetic derivatives. Furthermore, to calculate the difference in binding free energy between derivatives of the scaffold, relative FEP simulations have been performed on pairs of the RGD derivatives.

4.1.1 Integrins as receptors

Integrin αvβ6 is a transmembrane, heterodimeric protein. The αv subunit of extracellular αvβ6 comprises a lower and upper calf domain, the thigh domain and the β-propeller domain (Figure 4.1). Extracellular β6 is made up of a lower leg, upper leg, and head region. The head region consists of the plexin-semaphorin-integrin (PSI) domain, the hybrid domain and the β1 domain. Ligand binding occurs at the interface of the αv and β6 subunit, with interactions formed between the ligand and each of the head regions (Figure 4.1). The activity of αvβ6 is influenced by conformational changes in the multidomain subunits of the integrin. In its inactive bent-closed conformation, the subunits hinge at each Genu, so that the head regions fold towards the cell membrane. Activation involves a switchblade-like opening motion of the upper domains to give the extended-closed conformation, exposing the active site to ligand binding. A swing out motion of the hybrid domain away from the αv subunit induces the extended-open conformation. This results in a more closed binding site at the other end of the β1 domain.

The natural ligand of αvβ6, pro-TGF-β1 binds to αvβ6 through an RGD motif. Key interactions include a bidentate hydrogen bond and metal chelate interaction. As these interactions are consistent across all αv integrins bound to RGD ligands, we term this set of interactions as canonical interactions. The MIDAS Mg^{2+} and the ADMIDAS Ca^{2+} ions are also important contributors to ligand binding and conformational rearrangements of the receptor. The opening of the hybrid domain to give the extended-open conformation results in a 3 Å movement of the ADMIDAS towards the MIDAS.[247] This

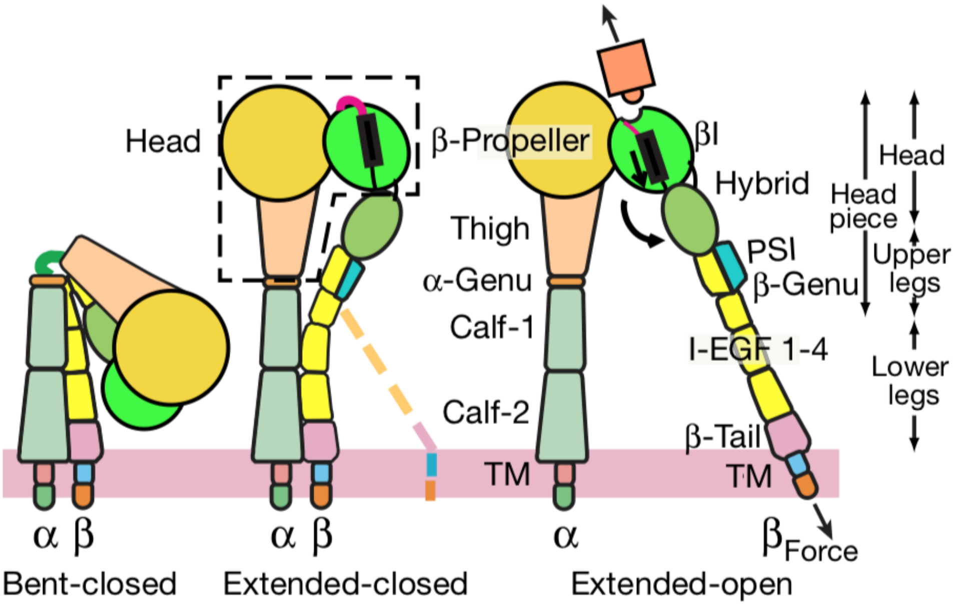


Figure 4.1: Schematic showing the different conformations of αvβ6. The domains of the integrin are labelled with the extracellular protein shown above the membrane (denoted as a pink strip); the ligand is shown in orange.[247] [Reproduced with permission from reference 246.]

concerted movement decreases the accessibility of the binding site in the extended-open conformation. The X-ray crystal structure with PDB code 4UM9[19] has been selected for our study. In this structure, αvβ6 is in complex with TGF-β1 and has a complete MIDAS and ADMIDAS occupancy. In addition to the canonical interactions displayed in the crystal structure, there are other binding site contacts. The RGD aspartate on the ligand forms hydrogen bonds with (β6)-Asn218 and (β6)-Ala126 in the receptor. The backbone of the ligand and (β6)-Thr221 also interact (Figure 4.2).

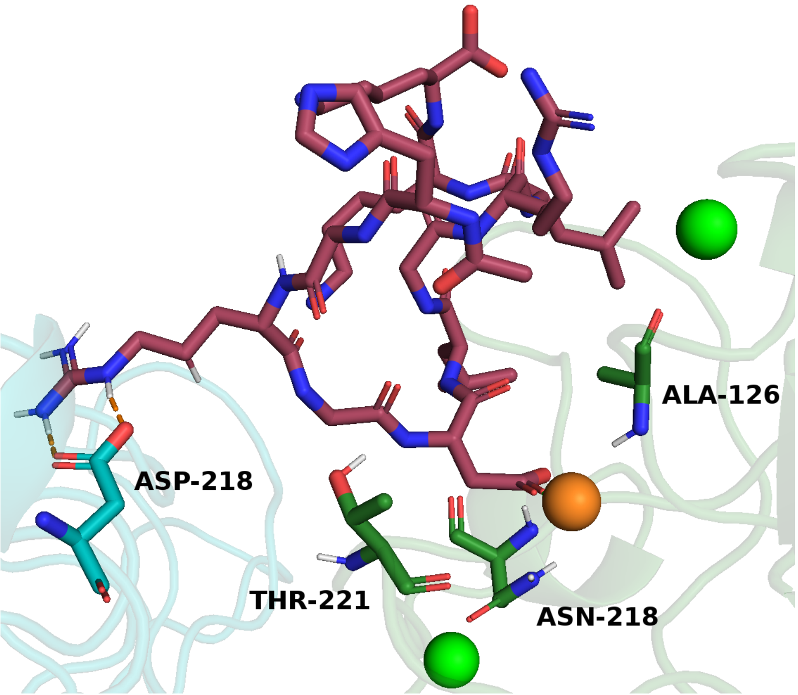


Figure 4.2: TGF-β1 (maroon backbone) bound in the active site of αvβ6 (PDB: 4UM9). Interacting binding site residues are shown as sticks. Blue residues correspond to the αv subunit and green to the β6 subunit. The Mg^{2+} MIDAS ion is shown as an orange sphere and the Ca^{2+} ions shown as green spheres.

4.1.2 Computational approaches for integrin inhibitor design

Identifying new clinical candidates which balance all necessary properties and that can be administered as a low dose medicine, is difficult. An important aspect, particularly for the clinical dose size, is the affinity of the modulator for its biological target. The more potent the molecule is whilst controlling its lipophilicity, the greater the chance of a lower clinical dose. The study presented in this chapter lays the foundation for computationally estimating the affinity of inhibitors of the αvβ6 integrin. Predicting ligand affinity in integrin drug discovery from docking studies has historically been difficult (although it is generally possible to rationalise the activity once the data is available).[248] Therefore, understanding which structural features of the inhibitor are most important for driving affinity from molecular simulations complements the empirical process that is generally used in integrin lead optimisation where compounds are made, tested and SAR are developed. Computational predictions become particularly important with longer syntheses, which use considerable resources and take months to complete.

In a previous computational study of αvβ6, Di Leva et al.[249] used MD simulations to identify additional non-canonical interactions and to develop a αvβ6 potent cyclic peptide from an RGD containing linear oligomer. This illustrates the utility of MD simulations for identifying potential areas for ligand development through a detailed description of how the dynamic changes of active site residues contribute to receptor-ligand binding. A computational study of αvβ3 - lig^{RGD} complexes found the Arg^{RGD} - (αv)-Asp218 interaction is maintained over 100 ns.[250] Over the simulation the distance between Asp^{RGD} and the Mg^{2+} ion in the MIDAS of β3 decreased, indicating these interactions are stable. Another MD simulation study[251] revealed, by varying isoDGR-containing cyclopeptides in complex with αvβ3, subtle differences in ligand interactions that affect the allosteric response of the receptor to ligand binding.

Our interest focuses on an RGD mimetic as the framework for a series of potential antagonists (Figure 3.4). Each compound varies at the 3-position of the aryl group, with substituents including H, F, CF_3, OCF_3, CH_3 and OCH_3. These compounds were taken from a class of compounds in an ongoing study of the SAR of αv integrin inhibitors.[8,234] Although other chemotypes have been explored in the literature,[248,252] our study focuses on a single scaffold. Scaffold substituents were chosen for this study as they are compatible with the CHARMM force field, obviating the need for extensive force field development for individual compounds. In previous work, the activity of each compound has been measured through a cell adhesion assay.[8] The compounds have pIC_{50} values in the range of 5.2 to 7.1. Due to the racemic nature of the compound, a single pIC_{50} value is assigned to both the (R)- and (S)-enantiomer of each compound, with the exception of CF_3. As the CF_3 analogue was prepared from commercially available (R) and (S) precursors, distinct pIC_{50} values can be assigned to each enantiomer. pIC_{50} is the negative $\log_{10}$ of the half maximal inhibitory concentration (IC_{50}), which we relate to ΔG using the following equation:

$$\Delta G = -RT \ln \left(IC_{50}\right) \tag{4.1}$$

where R is the gas constant and T is temperature.

There is growing interest in predicting relative binding energies using FEP simulations and integrating them into drug discovery workflows.[196] Chemically accurate *in silico* binding affinities can provide guidance when optimising lead compounds, saving synthetic resources and effort. There is a question, however, about the ease with which FEP can be applied to new systems. Nevertheless, with improved force fields and greater computational resources becoming more available, FEP simulations are increasingly attractive and tractable. The accuracy of FEP simulations rely on a quality force field, sufficient sampling and a well equilibrated system. This poses some practical considerations. For example, force field parameter optimisation is often necessary. For convergence of the predicted free energy change for a transformation, it is important that at each alchemical perturbation between two end states, the system is equilibrated to that intermediate state. Also, as perturbations need to be conservative, many windows may be needed to cope with the change between the two end states. As a result, these simulations can become computationally intensive and access to parallel computing systems is required for thorough sampling regimes at realistic timescales.

In this chapter, we build on previous calculations[112] using molecular docking. Starting from docked conformations, we use MD and FEP simulations to assess the effects of the different aryl substituents on active site interactions. MD simulations of αvβ6 in complex with its natural ligand, TGF-β1, enable us to investigate the dynamic and thermodynamic behaviour of ligand binding. We identify the contributions of active site residues to binding by monitoring how often they interact with the ligand. Furthermore, FEP simulations have been performed to calculate relative binding free energies.

4.2 Materials and Methods

4.2.1 Molecular docking

Coordinates were taken from an X-ray crystal structure of a αvβ6 dimer with the pro-domain of its natural ligand, TGF-β1, bound (PDB 4UM9[19]). To prepare for docking, chains C, D and F were extracted from the equilibrated structure with all water molecules and ions removed (equilibration methodology outlined below). Using receptor generation software as part of the OpenEye docking toolkit,[37,253] Chain F (TGF-β1) was assigned as ligand and thus did not interact with the docked molecules. A box centred on (β6)-Thr221 with sides of length 21.0 × 22.7 × 27.3 Å was situated to fully cover the TGF-β1 occupied binding site, giving a total receptor volume of 8680 Å^3. Constraints were then applied, ensuring a metal chelate interaction with MIDAS and hydrogen bond donors to both of the carboxyl oxygen atoms on the (αv)-Asp218 residue.

Compounds were protonated according to physiological pH. The compounds studied were prepared using OMEGA,[253] for both (*R*) and (*S*) enantiomers. Conformers were generated using a truncated form of the MMFF94s force field,[254] a variant that excludes both Coulomb interactions and the attractive part of vdW interactions. A maximum energy difference of 20 kcal mol^{-1} was allowed from the lowest energy conformer. These allowed molecules to explore additional conformational space. A maximum of 10,000 conformers per enantiomer was set and conformers within 0.5 Å of any others were considered duplicate and thus removed. Docking was performed using OpenEye FRED.[37] Compounds were docked using the high resolution setting with rotational and translational step sizes of 1 Å. Chemgauss4 was used to score the poses. The poses of each enantiomer were inspected for anomalies and the top scoring poses were chosen as the starting positions for MD simulations.

4.2.2 Molecular dynamics simulations

Coordinates of chains C and D were used as starting structures for subunits αv and β6 respectively. Simulations involving the natural ligand used chain F coordinates. All water molecules and metal ions throughout the crystal structure were included. It is particularly important to retain crystallographic water molecules for FEP simulations as they stabilise the system and improve the equilibration process.[196] In accordance with physiological pH, the zwitterion form of the RGD mimetic was used, with the 1,2,3,4-tetrahydro 1,8-naphthyridine protonated and the carboxyl group deprotonated. The protonated states of arginine and lysine residues were used and all aspartic and glutamic acids were deprotonated. Histidine residues were treated as neutral, with the nitrogen atom nearer the backbone protonated. Procedures to build hydrogen atoms, solvate the system and apply PBC were generated using the quick MD simulator module in CHARMM-GUI.[255] The system was solvated in a truncated octahedral periodic boundary cell with edge distances of 10 Å to construct an explicitly modelled solvent consisting of 16,886 TIP3P water molecules,[182] eight Mg^{2+} and four Cl^{-} ions, to give a net neutral charge (Figure 4.3). The concentration of the counter ions matched conditions used in cell adhesion assays performed on these complexes. To optimise the solvent positions, all heavy atoms were fixed, except for water molecules, during 50 steps of steepest descent (SD) and 50 steps of Adopted Basis Newton-Raphson (ABNR) minimisation. Protein and metal ion parameters were obtained from the C36 version of the CHARMM force field.[119] Parameters for metal ions were developed and validated by Beglov et al.[256] and are commonly employed in biomolecular studies.

Upon system setup, the NAMD software[173] was used for simulations of all complexes. Firstly, the solvated crystal structure, still containing TGF-β1, was minimised for 20 ps using a conjugate gradient and line search algo-

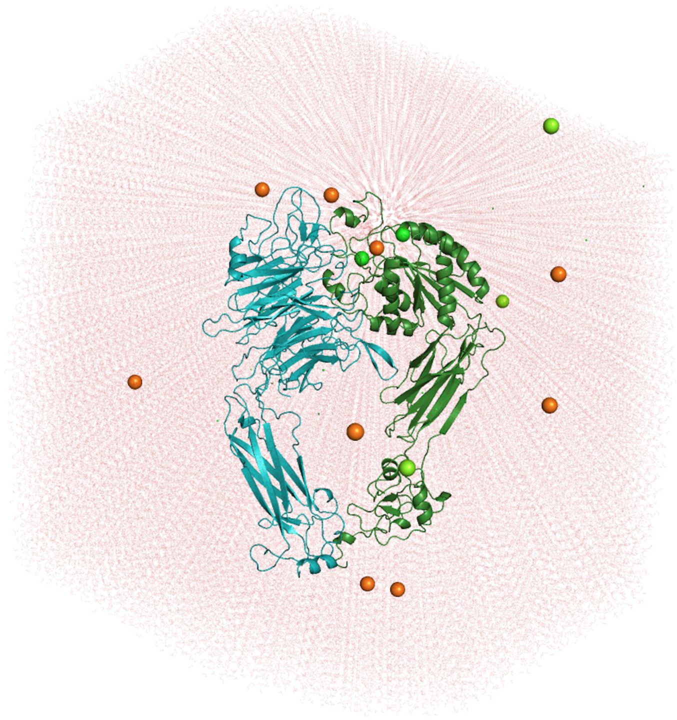


Figure 4.3: Complex of an RGD mimetic docked into the active site of αvβ6 (PDB 4UM9). The system has been solvated in a truncated octahedral box with TIP3P water molecules (red dots) and neutralising Mg^{2+} (orange) and Cl^- (green) ions.

rithm. Protein backbone and side chain restraints were applied using harmonic constraints with force constants of 10 kcal mol^{-1} $Å^{-2}$ and 5 kcal mol^{-1} $Å^{-2}$. The system was heated to 298 K in increments of 3 K every 1 ps using NAMD velocity reassignment. Backbone and side chain restraints were gradually switched off during a 2 ns equilibration period in the NVT ensemble. The coordinates of the equilibrated receptor were used for ligand docking.

To ensure canonical interactions were maintained during an additional 1 ns equilibration of the docked complexes, both carboxyl oxygen atoms on the aspartate mimetic were constrained to a distance of 2 Å from the Mg^{2+} ion. The polar hydrogen atoms on the protonated 1,2,3,4-tetrahydro

1,8-naphthyridine segment were constrained to a distance of 2 Å from the carboxyl oxygen atoms on (αv)-Asp218. An initial force constant of 10 kcal mol^{-1} Å^{-2} was used for all distance constraints. Force constants for distance constraints were steadily decreased to 2.5 kcal mol^{-1} Å^{-2} during an equilibration of 0.5 ns in the NVT ensemble and 0.5 ns in the NPT ensemble. This meant all canonical interactions were present at the start of production runs. Five independent 10 ns production runs were performed in the NPT ensemble for the (*R*)- and (*S*)-enantiomers of each derivative. Simulations of αvβ6 bound with TGF-β1 were also performed, resulting in a total of 65 simulations of 10 ns. Temperature was controlled using Langevin dynamics parameters, with a friction coefficient of 5 ps^{-1} for all equilibration and production runs. Constant pressure was maintained using the Langevin piston Nosé-Hoover method[257] with a target pressure of 1 atm. A cutoff distance of 12 Å was used for vdW pairs, with a switching function at a distance of 10 Å. The electrostatic potential energy was computed using the PME method.[258] The SHAKE algorithm[168] was used to fix all bond lengths involving hydrogen atoms and a timestep of 2 fs was used. Upon completion of production runs, all solvent molecules were removed except those within 10 Å of the ligand, in order to expedite the analysis. Trajectories were sampled every 20 ps, resulting in 500 frames for each replica.

4.2.3 Free energy perturbation simulations

FEP simulations, as discussed in Chapter 2, involve an alchemical transformation between two structurally related ligands.[259] A change is made to the system so that the potential energy is equivalent to the original potential energy with an additional “perturbing" potential energy term (V_{BA}):

$$U_B = U_A + V_{BA} \tag{4.2}$$

where U_A and U_B are the respective potential energies of state A and B, which represent the states where a different ligand is bound in each. States A and B should arise from the same conformational space and therefore, V_{BA} needs to be very small. To address this, the transformation of A to B is divided into several discrete simulations, called windows, which are connected using a coupling parameter, λ.

Figure 4.4 shows the thermodynamic cycle of a transformation of one ligand, L1, (state A) into another, L2 (state B). ΔG_1 and ΔG_2 are the binding free energies of each ligand, while ΔG_3 and ΔG_4 are the free energies of transforming one ligand into the other, as the free ligand and in complex with the receptor. At each end point of the simulation (L1 and L2) λ equals zero or one. At intervals between these end points, bonded and non-bonded parameters are scaled so that they are “switched off" for outgoing atoms and “switched on" for incoming atoms. There are two ways to alchemically mutate ligands. In our FEP simulations, we employ the dual topology approach. A soft-core potential is used to avoid “end-point catastrophes", where disappearing atoms can leave empty pockets and appearing atoms clash with the existing environment.[200,201] Electrostatic interactions of outgoing atoms are decoupled from the system over the λ range of 0 to 0.5, while the electrostatics for incoming atoms are coupled to the system over the λ range of 0.5 to 1.

$$\begin{tikzcd}
\mathrm{L1} \arrow[r, "\Delta G_1"] \arrow[d, "\Delta G_3"'] \arrow[dr, phantom, "{\begin{aligned}\Delta\Delta G &= \Delta G_2 - \Delta G_1 \\ &= \Delta G_4 - \Delta G_3\end{aligned}}"] & \mathrm{RL1} \arrow[d, "\Delta G_4"] \\
\mathrm{L2} \arrow[r, "\Delta G_2"'] & \mathrm{RL2}
\end{tikzcd}$$

Figure 4.4: A thermodynamic cycle describing the binding of two ligands, L1 and L2, to a receptor, R. The relative free energy of binding can be calculated from either the physical ($\Delta G_2 - \Delta G_1$) or alchemical ($\Delta G_4 - \Delta G_3$) legs of the cycle. In FEP calculations, the transformations of the alchemical pathways are modelled.

To calculate relative binding free energies of pairs of ligands, ten different alchemical transformations of aryl substituents were performed. For each perturbation, three replicas of the forwards and backwards transformation were performed, resulting in 60 FEP simulations. Ahead of system setup, dual topologies were constructed for both the free and bound ligands. To prevent system drift, caused by the charged ends of the free ligand coming together, the transformations of the free ligand in solvent started from structures close to an energy minimised conformation. The setup of systems and free energy simulations were performed using the procedures for the standard MD simulations described above. During FEP, λ was increased from 0.0 to 0.1 in 16 discrete steps, from 0.1 to 0.9 in steps of 0.02 and from 0.9 to 1.0 in 16 steps to give a total of 72 windows for each transformation. Equilibration was performed for 20 ps at the start of each window, followed by

100 ps of sampling.

4.3 Results and Discussion

4.3.1 Protein-ligand interactions of TGF-β1

Polar residues that are important for binding can be identified by investigating the dynamics of hydrogen bonds between the natural ligand of αvβ6, TGF-β1, and the receptor. From the crystal structure, it is clear that (αv)-Asp218, (β6)-Ala126, (β6)-Asn218 and (β6)-Thr221 form hydrogen bonds with TGF-β1. The fraction of time that these interactions were maintained over the 50 ns MD simulation was measured. A pair of atoms are considered to be hydrogen bonded if a polar hydrogen atom is within 2.5 Å of an oxygen, nitrogen or fluorine atom. A bidentate interaction between (αv)-Asp218 and the Arg residue of the RGD unit on the ligand was present for 82% of the simulation time. (β6)-Ala126, (β6)-Asn218 and (β6)-Thr221 were all hydrogen bonded with TGF-β1 for over 90% of the simulation time. The MD simulation indicated two hydrogen bonds formed between the Asp residue of the RGD binding tripeptide in the ligand and (β6)-Asn218. One interaction involved the amide backbone of the (β6)-Asn218 residue for 98% of the simulation time and the other involved the side chain of (β6)-Asn218 for 66% of the simulation time. The hydrogen bond with the side chain is not present in the starting structure and is only observed as the result of dynamics. Both interactions with (β6)-Asn218 are known among RGD ligands and have been previously recognised.[19,260,261] The metal chelate interaction

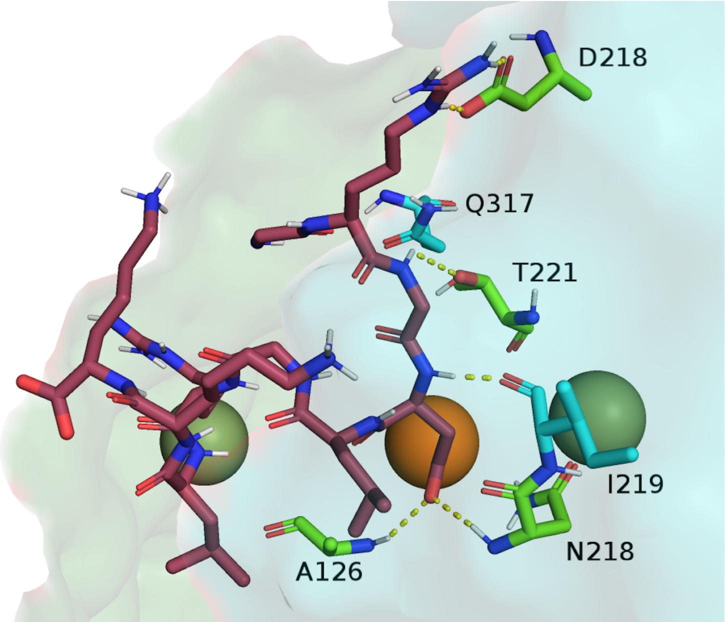


Figure 4.5: Residues that form interactions with TGF-β1 (maroon backbone) over a 50 ns MD simulation. Interacting residues in the crystal structure are shown (light green) as well as residues that do not interact in the crystal structure, but gain an interaction during the MD simulation (light blue). The Mg^{2+} MIDAS ion (orange) and Ca^{2+} ions (dark green) are also shown.

between Asp^{RGD} and the Mg^{2+} MIDAS ion was maintained throughout the entire simulation.

Figure 4.5 depicts the position of the binding site residues that form hydrogen bonds with TGF-β1, identified by the crystal structure and MD simulations. Residues (β6)-Ile219 and (β6)-Gln317 are also shown. These do not appear to interact in the crystal structure but are in close proximity with TGF-β1 for 78% and 7% of the MD simulation time, respectively. This highlights how MD simulations can identify active site residues that contribute to ligand binding, which are not observed in a static crystal structure. In an ongoing study, these receptor conformations generated by MD simulations are being used for docking as a way to extend the receptor conformational search space and identify further antagonists.

Hydrophobic interactions are also important for the binding of TGF-β1 to αvβ6. The 244-LGRLK-248 sequence directly following the RGD motif folds into an amphipathic α-helix, which fits into a β6-specific hydrophobic pocket, as previously reported.[19,249,262,263] Leu224 and Leu247 form lipophilic contacts with this pocket, which is formed by residues distinct to β6 when compared to other RGD binding members of the integrin family such as αvβ3 and αvβ5.[19] Contact between the amphipathic α-helix of TGF-β1 and the hydrophobic pocket of β6 was maintained throughout the MD simulations.

4.3.2 Protein-ligand interactions of RGD mimetics

By monitoring the dynamics of hydrogen bonding interactions between each derivative of the RGD mimetic and the receptor, we characterise the nature of ligand binding. The stability of the canonical interactions could reflect the potency of each ligand, as an active compound should maintain these interactions and remain bound in the active site. In principle, there could be a link between analogues with higher pIC_{50} values and better maintained canonical interactions. All derivatives of the pseudo-RGD compound remain bound throughout the MD simulations, through at least one component of the canonical interactions. Table 4.1 shows the pIC_{50} values of each analogue, which are a measure of activity[8] (with additional pharmacological data available in reference 8), and the fraction of time that the canonical interactions were maintained. The interaction frequency is the proportion of frames with the interaction present with respect to the total number of frames, averaged over five 10 ns simulations. The bidentate interaction is present in all derivatives of the RGD mimetic to a varying extent. However, there is no correspondence between the stability of the bidentate interaction and pIC_{50}.

The loss of the bidentate interaction is caused by a slight overall translation of the ligand in the binding site so that only one polar hydrogen atom on the naphthyridine fragment is within 2.5 Å of (αv)-Asp218. Although the bidentate interaction is lost for a considerable amount of time in some ligands, a monodentate interaction between the Arg mimetic and (αv)-Asp218 is commonly observed for all ligands (Table A.1). All analogues of the ligand form at least one metal chelate interaction with the Mg^{2+} ion throughout the MD simulations. The difference between interactions Mg^{2+} - O^1 and Mg^{2+} - O^2 is the oxygen atom on the carboxyl group that the Mg^{2+} interacts with. Given the proximity of the chiral centre to the MIDAS site, it might be expected that interaction frequencies for the less active enantiomer with MIDAS amino-acid residues will be lower as the calculations suggest. Given the conformational flexibility of the molecule however, what is less expected is the impact on the bidentate interaction frequency between (αv)-Asp218 and the 1,8-naphthyridine at the other end of the molecule which is much lower for the less active enantiomer compared to the more active enantiomer ((*R*)-CF_3 19% and (*S*)-CF_3 74%). Nevertheless, the interaction with the MIDAS ion is better maintained than the bidentate hydrogen bond in all cases. Therefore, we observe that the stability of the metal chelate interaction between Asp^{RGD} and the Mg^{2+} ion is more important than the bidentate hydrogen bonding to (αv)-Asp218 for ligand binding.

Table 4.1: Interaction frequency in % of simulation time that the bidentate, Mg^{2+} - O^1 and Mg^{2+} - O^2 interactions are maintained for each analogue of the RGD mimetic. Contact frequencies between atoms on the aryl rings and active site residues are also shown in columns 8-10. The experimental activity of each analogue is shown by its pIC_{50} value.[8] The molecular docking score (arbitrary units), generated using OpenEye FRED[37] and OpenEye OMEGA,[253] is also shown.

				InteractionFrequency (% Time)					
Substituent	Enantiomer	pIC_{50}	Docking Score	Bidentate Interaction	Mg^{2+} - O^1	Mg^{2+} - O^2	(β6)-Ser127	(β6)-Ala126	(β6)-Ile219
H	*S*	5.7	-9.99	89	100	0	96	8	20
	R		-9.19	96	100	100	24	24	0
F	*S*	6.1	-9.64	96	100	46	50	22	2
	R		-9.29	29	89	80	39	23	5
CH_3	*S*	6.4	-7.96	100	100	73	81	38	49
	R		-4.64	100	100	100	47	43	2
OCH_3	*S*	6.5	-9.38	65	100	100	100	62	0
	R		-6.57	100	100	100	84	3	0
OCF_3	*S*	6.7	-8.49	98	100	40	45	19	0
	R		-3.59	99	100	0	0	0	3
CF_3	*S*	7.1	-9.33	74	100	100	57	43	25
	R	5.2	-5.70	19	100	100	0	0	3

The position of the aryl ring on the bound ligand is indicated in Figure 4.6. The metal chelate interaction formed by the carbonyl and Mg^{2+} MIDAS ion points inwards into the binding pocket, forcing the aryl ring to become solvent exposed. Nearby residues are on the $\beta 1$-$\alpha 1$ and $\alpha 2$-$\alpha 3$ loops, specifically (β6)-Ala126, (β6)-Ser127, (β6)-Asn218, (β6)-Ile219, (β6)-Asp220 and (β6)-Thr221. The side chain of (β6)-Ala126 and the backbone of (β6)-Asn218 contribute to the β6 hydrophobic binding pocket.[19] All (*S*)-enantiomer derivatives of the ligand show an interaction with both groups (Table A.1). Further contacts within 2.5 Å of each substituted ring and any atom on the β6 unit were investigated. Table 4.1 shows the frequency of contacts with (β6)-Ala126, (β6)-Ser127 and (β6)-Ile219. From the proximity of the ring to (β6)-Ser127, we observe that the (*S*)-H, (*S*)-CH_3 and (*S*)-OCH_3 derivatives remain in close contact with the $\beta 1$-$\alpha 1$ loop. The remaining S-derivatives, (*S*)-CF_3, (*S*)-F and (*S*)-OCF_3, stay close to the $\beta 1$-$\alpha 1$ loop, but for less than 50% of the simulation time. As no other contacts are formed between these derivatives and the receptor for a significant amount of time, we suggest that solvent interactions with these more polar substituents influence the position of the aryl ring.

MD simulations of the natural ligand, TGF-β1, indicated that residues (β6)-Ala126, (β6)-Thr221 and (β6)-Asn218 in the receptor are all involved in binding. Therefore, hydrogen bonding interactions between these residues and the RGD mimetics were investigated. The carboxyl oxygen atoms on the (*S*)-forms of the H, F, CH_3 and OCF_3 derivatives interacted with (β6)-Ala126 for 67%, 45%, 10% and 60% of the simulation time, respectively. No stable interactions are identified between any derivatives of the ligand and (β6)-

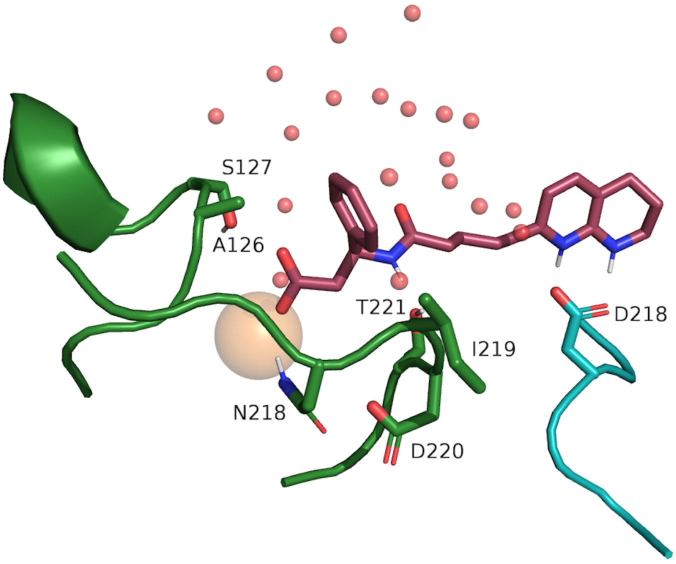


Figure 4.6: Position of the aryl ring on the (S)-RGD mimetic within the binding site. Nearby residues on the $\beta1$-$\alpha1$ (blue) and $\alpha2$-$\alpha3$ (green) are labelled. The oxygen atoms of water molecules within 7 Å of the aryl ring are also shown (red spheres).

Thr221. Although interactions between the ligands and residue (β6)-Thr221 are present in the docked structures, they are lost during equilibration of the systems and are not observed during dynamics. In the docked structures, (β6)-Thr221 interacts with the carbonyl oxygen atom on the amide backbone of the ligand. However, during equilibration, the carboxyl group rotates away from the receptor to become more solvent exposed.

4.3.3 Cation-π interactions in αvβ6-RGD systems

A cation-π interaction consists of an electrostatic attraction between an electron rich aryl ring and an electron deficient cation. This type of interaction is common in proteins; an early data mining study of a subset of the PDB found that over 70% of all Arg side chains are near an aromatic side chain.[264]

The crystal structure of αvβ6-TGF-β1 is no exception, with a face-to-face intermolecular cation-π interaction between Arg^{RGD} and residue (αv)-Tyr178 (Figure 4.7). A mixture of factors govern cation-π interaction orientation, such as competitive hydrogen bonding and the influence of solvent.[265] However, this parallel geometry is more common in proteins than T-shaped geometries, although T-shaped geometries are preferred in the gas-phase.[264,266] To monitor the persistence of the cation-π interaction during MD, the distance between the cationic Arg group (the guanidinium carbon in Arg) in TGF-β1 and the centre of the aromatic (αv)-Tyr178 ring was measured. A distance less than 5 Å, a common geometrical constraint used in previous studies, indicated an interaction. The average distance between the two residues was 4.41 Å with a standard deviation of 0.48 Å. The face-to-face geometry was also maintained throughout.

As the overarching aim is to design a compound that competitively binds with αvβ6, it would be sensible to take advantage of the position of (αv)-Tyr178 and replicate this cation-π interaction with the RGD mimetics. The cation in this case is the protonated 1,2,3,4-tetrahydro 1,8-naphthyridine group. Although the interaction between the naphthyridine group, an aromatic cation, and the aromatic side chain of (αv)-Tyr178 could be dominated by π-π interactions, a study by Tsuzuki et al.[267] found, through high level ab initio calculations, that interactions of benzene complexes with aromatic cations should be categorised as cation-π interactions as they are stabilised by large electrostatic and induction interactions. The distance between the positively charged nitrogen atom and the centre of the aromatic (αv)-Tyr178 ring was measured for each of the RGD derivatives. Table 4.2 shows the

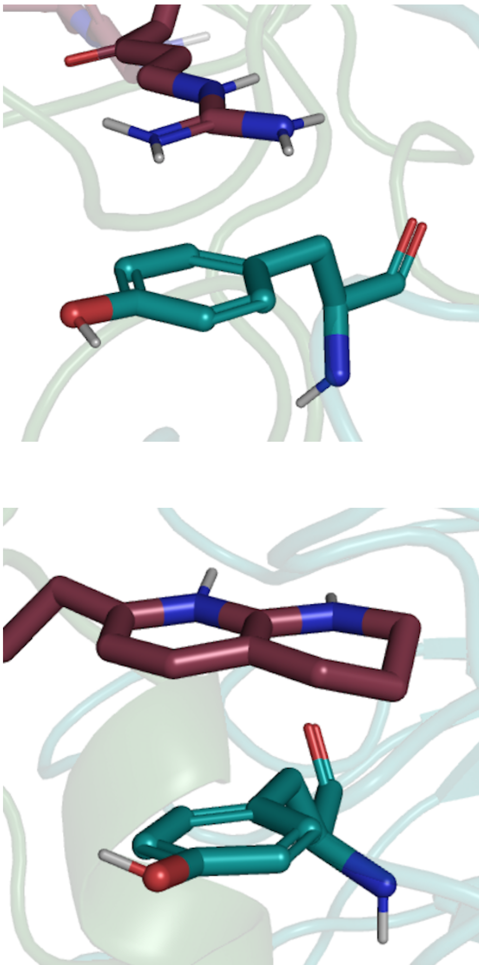


Figure 4.7: Cation-π interactions between the cationic nitrogen atoms on Arg^{RGD} (top) or the 1,2,3,4-tetrahydro 1,8-naphthyridine group of the (S)-CF_3 derivative (bottom) and the aromatic side chain of (αv)-Tyr178.

distances, averaged over 10 ns of MD. The majority of compounds had an average distance within the 5 Å cutoff. The cation-π interaction orientation sampled most was a face-to-face geometry of the (αv)-Tyr178 side chain with the naphthyridine group (Figure 4.7). Thus, the RGD mimetics replicate this type of interaction seen with TGF-β1, in addition to the canonical hydrogen bonding interactions.

Table 4.2: The distance between the protonated N atom on each RGD derivative and the centre of the aromatic ring on (αv)-Tyr178. Distances have been averaged over 10 ns of MD simulations, with the standard deviation shown.

Substituent	NH^+ - Tyr178 Distance (Å)
(*S*)-H	3.77 ± 0.33
(*R*)-H	5.01 ± 0.86
(*S*)-F	4.80 ± 0.41
(*R*)-F	4.78 ± 1.06
(*S*)-CH_3	4.63 ± 0.43
(*R*)-CH_3	4.13 ± 0.50
(*S*)-OCH_3	3.82 ± 0.47
(*R*)-OCH_3	4.21 ± 0.49
(*S*)-OCF_3	5.21 ± 0.44
(*R*)-OCF_3	5.33 ± 0.69
(*S*)-CF_3	4.19 ± 0.62
(*R*)-CF_3	4.33 ± 0.50

4.3.4 Relative free energy perturbation calculation

To assess the convergence of the alchemical calculations, a method outlined by Klimovich et al.[206] was used. In the example (Figure 4.8) the transformation of a hydrogen substituent to an OCF_3 substituent, with the ligand in the bound state is considered. This transformation was chosen as it is the most diverse in terms of the number and types of atoms perturbed and therefore, it might be most likely to have convergence issues. The free energy change is calculated using an increasing fraction of the simulation data (i.e. 0-0.1, 0.0-0.2). Also plotted in Figure 4.8 is the free energy change calculated with the reverse proportion of the data (i.e. 0.9-1.0, 0.8-1.0). Both sets

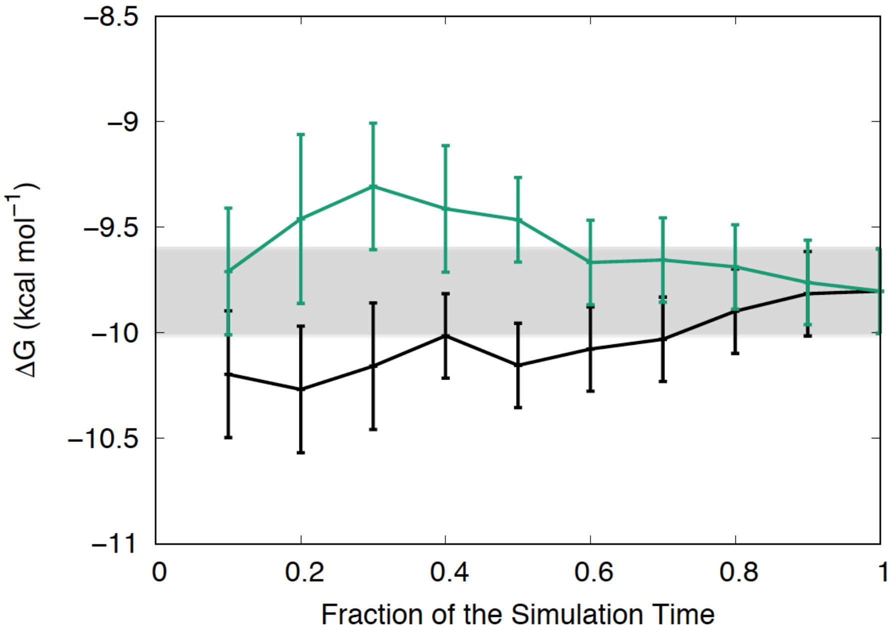


Figure 4.8: Convergence assessment of the transformation of the hydrogen derivative to the OCF_3 derivative, with the ligand bound to the receptor. The forward (black line) and the reverse (green line) simulation time series are shown. The horizontal grey strip indicates the equilibrated region.

of data remain within error of the final value and so the calculation is considered converged. Furthermore, the relative free energies of the forwards and backwards transformations were compared, as the values are expected to be identical but of opposite sign. The H → OCF_3 perturbation of the free ligand gave a relative free energy change of -8.9 kcal mol^{-1}, while the OCF_3 → H perturbation resulted in a relative free energy change of 8.5 kcal mol^{-1}. The same perturbations gave relative free energy changes of -10.4 kcal mol^{-1} and 9.8 kcal mol^{-1} when the ligand was bound. The relative free energies reflect that, due to its stronger non-covalent interactions, the OCF_3 derivative is favoured over the hydrogen substituent in both solvent and in complex with the receptor. The similarity for the forwards and backwards transformation in both states suggests that conditions are met for the convergence of the alchemical simulations.

Calculated relative free energies of binding are shown in Table 4.3. The ΔG values for the transformations of the free ligand and in complex with the receptor were evaluated with the BAR method,[98] using the ParseFEP tool in VMD.[268] $\Delta\Delta G$ values were obtained by taking the difference between each of these transformations. These theoretical values are compared with $\Delta\Delta G$ values obtained from the experimental pIC_{50} values. For all ten alchemical transformations, the sign of the relative free energy difference is correctly predicted, although the magnitude is almost always over-estimated but in a non-systematic manner. As all derivatives, with the exception of CF_3, were measured as racemic mixtures, there is a probable ±0.3 log error on each experimental binding free energy, which should also be taken into account. The calculated values are within 1.5 kcal mol^{-1} of experiment, with the exception of the $H \rightarrow CH_3$, $CH_3 \rightarrow OCH_3$ and $OCH_3 \rightarrow OCF_3$ aryl transformations. In order to compare the results of the FEP simulations with simpler methods, we investigate the relationship between docking scores and experimental activity (Table 4.1). Upon docking of the ligands into the equilibrated receptor, a score is obtained which reflects the quality of the docking of the RGD mimetic to the αvβ6 binding site. Although the values are somewhat empirical and cannot be directly related to binding affinity, docking scores can be useful for ranking compounds or distinguishing between active and non-active compounds.[269] The docking scores shown in Table 4.1 do not correlate with the pIC_{50} values. Despite this, the scores do consistently indicate that the (S)-enantiomers will be more active than the (R)-enantiomers, as expected. As the relative free energies obtained for the ligands agree more

Table 4.3: Relative free binding energy of ligand derivatives. Experimental $\Delta\Delta G$ values have an error of 0.6 kcal mol^{-1}.

	$\Delta\Delta G$ (kcal mol^{-1})		
Transformation	Experimental	Calculated	Absolute Difference
CH_3 → OCH_3	-0.1	-2.6 ± 0.3	2.5
OCH_3 → OCF_3	-0.3	-2.3 ± 0.3	2.0
H → F	-0.5	-1.9 ± 0.2	1.4
OCF_3 → CF_3	-0.5	-0.5 ± 0.4	0.0
H → CH_3	-1.0	-2.9 ± 0.3	1.9
CH_3 → CF_3	-1.0	-1.4 ± 0.3	0.4
H → OCH_3	-1.1	-1.8 ± 0.3	0.7
H → OCF_3	-1.4	-1.9 ± 0.4	0.5
F → CF_3	-1.4	-1.3 ± 0.3	0.1
H → CF_3	-1.9	-2.3 ± 0.3	0.4

closely with experiment, compared to the docking scores, we suggest that a more physically realistic model has resulted from FEP simulations for this system, albeit at higher computational cost.

There is a lack of correlation in the rank order, with respect to the hydrogen derivative, between the calculated and measured affinities. This is not atypical of many such examples in the field[270] where there is a narrow range of activity exhibited by the ligands. However, it is also important to note that when forming a perturbation map with the hydrogen, fluorine, CF_3 and OCF_3 substituents, to make a closed thermodynamic cycle, the summation of the estimated free energy along each edge is -0.9 kcal mol^{-1}. This is the hysteresis of cycle closure and is substantially higher than the triplicate error estimate of 0.2-0.4 kcal mol^{-1}. This analysis prompts caution in using this FEP methodology for ranking this ligand series. Despite this, the sign of the free energy change is correctly predicted in each case, suggesting it is a useful tool for pairwise comparisons of activity.

The close match between the values for each of the derivatives shows that FEP calculations do still model the system well, giving a solid foundation for the associated analysis. While the equilibrium MD simulations provide information on the dynamic behaviour of interacting residues (as detailed in Table 4.1), FEP simulations provide a rigorous way to quantify various physical factors, such as changes in hydrophobic and hydrogen bonding interactions. Taking the transformation from the hydrogen derivative to the OCF_3 derivative as an example, the more lipophilic OCF_3 substituent is expected to have more favorable interactions with the hydrophobic binding pocket in the β6 subunit. This is reflected by the experimental free energy of binding, which shows a difference of -1.4 kcal mol^{-1} between the two derivatives. Calculated values show a similar relative binding free energy of -1.9 kcal mol^{-1}.

4.4 Conclusions

In this work, MD and FEP simulations aid the investigation of potential antagonists of αvβ6. An MD simulation of αvβ6 bound with the the pro domain of TGF-β1, starting from an X-ray crystal structure,[19] was performed in order to understand the binding site interactions of the natural ligand. MD simulations on a series of compounds, based on an RGD mimetic, were also performed starting from docked structures. From these studies the importance of the canonical interactions, the bidentate interaction of the 1,8-

naphthyridine with (αv)-Asp218 and the acid binding to the MIDAS site, are clear. These interactions are supported by unpublished SAR which (i) demonstrates the distance between the base and the acid is critical and (ii) shows that maintaining αvβ6 affinity whilst structurally modifying the acid and 1,8-naphthyridine is particularly challenging. The calculations suggest the substituted aryl ring is essentially solvent exposed. Other work[8,234,248] has found that about two log units of potency can be gained with the optimal substituent both with this chemotype[8,234] and others,[248] but this involves larger substituents which can form additional interactions with the receptor.

FEP simulations have enabled us to estimate the relative free energies of binding between pairs of RGD mimetics. As the range in pIC_{50} of the subset of compounds studied was 5.2 to 7.1, which is relatively narrow, future work should expand the substituents studied to include more potent compounds. Furthermore, in order to rank this series of β6 antagonists, all perturbations should be linked in some way so that the energies can be compared. Therefore, an alternative perturbation map should be considered. As suggested by Cournia et al.,[196] instead of performing perturbations to each ligand from a single reference compound, substituents should be connected in a single graph, arranged by similarity. By connecting each substituent to at least two other derivatives in a closed cycle, it is also possible to compute sampling errors as the total free energy change in a closed thermodynamic cycle should equal zero. Nevertheless, by comparing the binding free energies calculated in this study with the difference in pIC_{50} value for each pair, we have shown that this integrin system, along with this series of ligands is amenable to study by FEP, with a good level of accuracy.

From a drug design perspective, lead optimisation of αv integrin inhibitors (αvβ6 inhibitors for example) has been driven empirically[8] and a priori selection of optimal substituents on the carbo-aromatic is difficult. This is because the current understanding in how the substituted aryl part of the molecule affects potency and selectivity through interactions with the specificity determining loop in the binding site is poor.[7,8] As a result, lead optimisation requires a large team of synthetic chemists and substantial budget and can mean progress towards a clinical candidate becomes slow. There is therefore great value in any reliable method which transitions molecular design from empiricism to theory.

Based on our previous work,[112] large numbers of potential inhibitors can be generated computationally featuring multi-substituted aryl motifs. Whilst this has advantages, multi-substituted aromatics can be difficult to synthesise which exacerbates what are already often challenging and long syntheses of the inhibitor. So some additional method to select the best compounds to make is needed and this illustrates the value of this work as it paves the way towards a more robust computational prediction of affinity, which should be valuable in prioritising compounds for synthesis.

Chapter 5

Structural Variation of Protein–Ligand Complexes of the First Bromodomain of BRD4

5.1 Introduction

BRD4, a member of the bromodomain and extra-terminal domain (BET) family, plays a key role in several diseases, especially cancers. Increased interest in BRD4 as a therapeutic target has resulted in a wealth of publicly available X-ray crystal structures of the protein in complex with small molecule inhibitors over the recent decade. These structures provide valuable atomistic insight into its binding site interactions. However, this also means that it is increasingly difficult to choose which crystal structure is preferred as the starting point of any *in silico* study. Within this chapter, we use the structural information available to help understand the flexibility of

BRD4 and examine the effects of small molecule inhibitors, with a focus on the BD1 binding pocket. This analysis provides guidance in selecting crystal structures and other features, such as crystallographic water molecules.

5.1.1 Structure and binding of BRD4

The BET family consists of the BRD2, BRD3, BRD4 and bromodomain testis-specific (BRDT) proteins. Like its family members, BRD4 consists of two N-terminal BDs (BD1 and BD2) and an extra C-terminal domain (ET).[271] Each BD is composed of four helices (αZ, αA, αB and αC), which are connected by the ZA loop and BC loop, creating a binding site (Figure 5.1). BRD4 recognizes the acetylated N-terminal tails of histones through interactions of its BDs. BRD4-BD1 recognises histone H4, which is anchored by a hydrogen bonding interaction between the carbonyl oxygen atom on the acetylated lysine and an asparagine residue Asn140 on the BC loop of the receptor.[272] A second interaction is formed through a water mediated hydrogen bond between the acetyl lysine and tyrosine residue Tyr97 on the ZA loop. Additional binding site residues create a deep hydrophobic cavity, with Trp81 and Met149 also considered key residues in H4 and small molecule inhibitor binding.[273]

At the base of the BRD4-BD1 binding pocket, there is a network of highly conserved water molecules, which is important in ligand binding and stabilising the protein structure.[273–275] A study investigating the structural and thermodynamic properties of the crystallographic water molecules found that it is energetically unfavourable to displace the water molecules with a small drug-like compound, as this would require a large amount of energy to

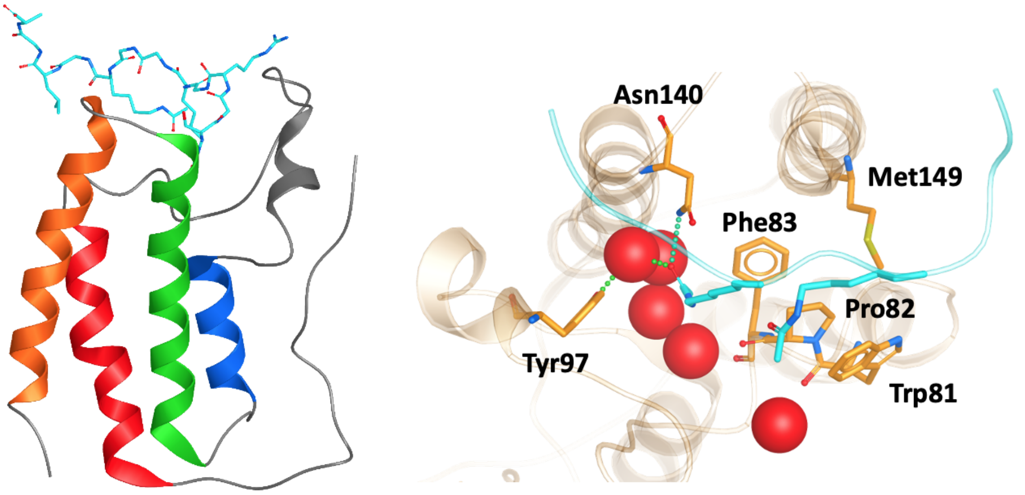


Figure 5.1: (Left) Structure of BRD4-BD1 (PDB 3UVW) with histone H4 (light blue) bound. Secondary structures of α-helices αZ, αA, αB and αC are coloured in red, blue, green and orange, respectively. (Right) Key binding site residues are highlighted as sticks. Histone H4 is shown in light blue, with acetylated lysine residues shown as sticks. Water molecules at the base of the binding site are shown as red spheres.

compensate breaking the hydrogen bonding network.[276] Furthermore, MD simulations demonstrated a high occupancy for several of the water sites. The authors used these findings to develop a docking based VS protocol to identify novel inhibitors towards BRD4, highlighting the importance of the water network in SBDD.

5.1.2 Therapeutic interest

Histones are proteins that provide structural support to the packaging of DNA into chromosomes. Histones H2A, H2B, H3 and H4 form a core that are surrounded by segments of DNA, as shown in Figure 5.2. Through binding to the acetylated tail of histones, BET proteins play a crucial role in regulating gene expression.[277] Furthermore, as histone acetylation readers, they contribute to tumorigenesis (the production or formation of tumours), mak-

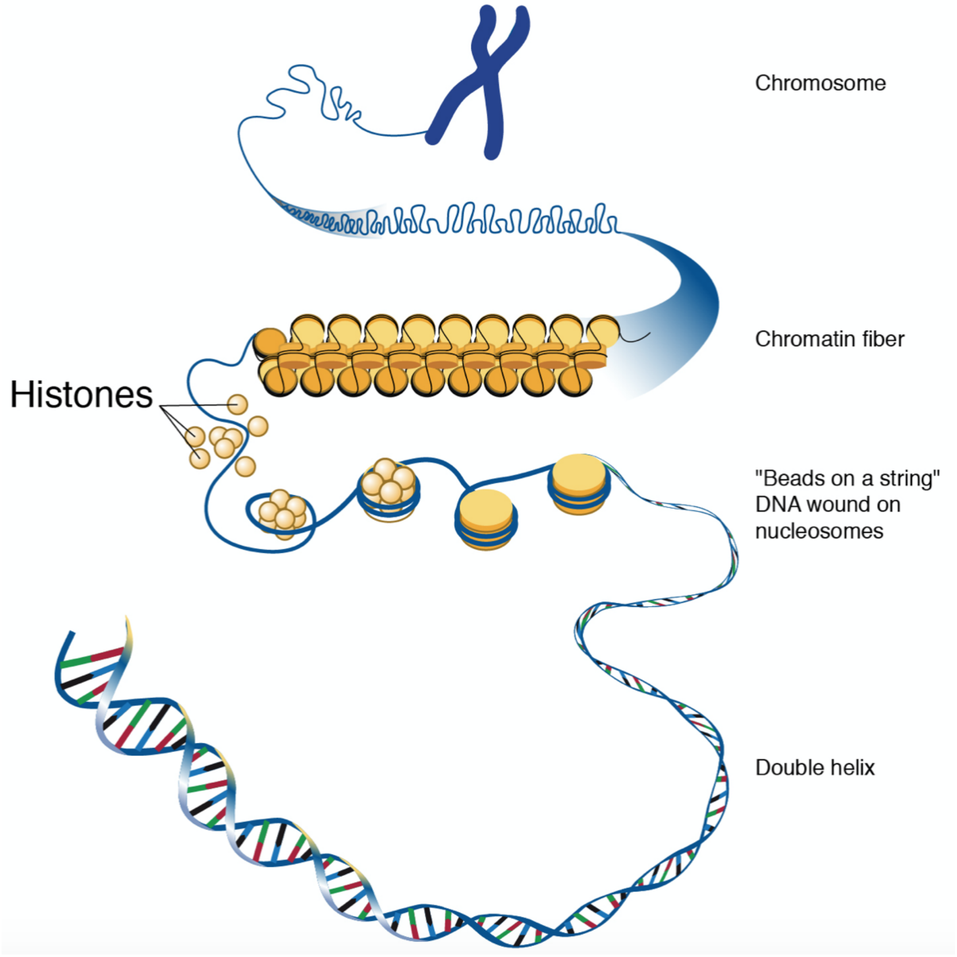


Figure 5.2: Representation of the packing of DNA into chromosomes. [Figure reproduced with permission from the National Human Genome Research Institute (https://www.genome.gov July 2021).]

ing them important targets for the development of small molecule drugs to inhibit these epigenetic interactions. The most extensively studied member of the BET family is BRD4, due to its promise as a therapeutic target for diseases such as cancer, neurodegenerative disorders, inflammation and obesity.[278–282]

Small molecule inhibitors

Over the last decade, there have been many small molecule inhibitors published, with some of them reaching human clinical trials.[283–288] Contained within these inhibitors is a wide chemical diversity with core motifs encompassing azepines, 3,5-dimethyl isoxazoles, pyridones, triazolopyridines, tetrahydroquinolines, 4-acyl pyrroles and 2-thiazolidinones.[289,290] Each of these structural classes contains a unique warhead, which competes with H4 to replicate the interactions with Asn140 and Tyr97. An additional common feature of small drug-like compounds bound to BRD4-BD1 is a lipophilic group, which can extend into the binding pocket and interact with the hydrophobic WPF shelf (Trp81-Pro82-Phe83).

The pool of BRD4 inhibitors continues to grow, with most identified through fragment or SBDD based on properties of known BRD4 inhibitors.[291–295] Additionally, a recent review[296] identified three novel strategies in targeting BRD4, including bivalent BRD4 inhibitors, proteolytic targeting chimeric molecules and re-purposing of kinase inhibitors. The selectivity of inhibitors is also being targeted.[297–299] Initially many compounds were developed as BRD4 inhibitors. However, due to the structural similarity across the BDs of BET proteins, many of these were pan-inhibitors which could cause adverse side effects such as dizziness and nausea.[300,301] Gilan et al.[297] used structure-based design to discover compounds that interact specifically with either the BD1 or BD2 of BET proteins, providing new insights into improving therapeutic strategies with fewer toxic side effect. Speck-Planche et al.[299] have also developed the first multi-target quantitative structure activity relationship (mt-QSAR) model, which can predict BET inhibitor potency against BRD2, BRD3 and BRD4.

5.1.3 Computational approaches for BRD4 inhibitor design

As in many drug discovery campaigns, computational methods remain an important tool in finding BRD4 inhibitors.[291,293,302,303] Structure based virtual screening methods, such as docking and 3D-QSAR, can facilitate high-throughput approximations of binding affinities. MD simulations expand on static representations of protein-ligand complexes by providing a more dynamic view and develop our understanding of structural patterns and interactions, which lead to high potency and selectivity.[304,305] As a crucial goal of drug development is designing a compound that binds competitively and strongly, alchemical free energy calculations are becoming increasingly important as, when done well, they can provide accurate estimations of binding free energies and present a way to minimize the number of compounds that are made in the laboratory. Pan et al.[291] discovered a BRD4-histone deactylase (HDAC) inhibitor, a promising therapeutic strategy for colorectal carcinoma, through high-throughput rigid molecular docking of fragments of known BRD4 inhibitors, embedded into the fragment-like library of ZINC.[306] A flexible docking method was then implemented on the top 200 scoring fragments. Following this docking, 24 compounds were synthesised and tested based on the 10 top scoring fragments, resulting in a promising lead compound. The extensive research already conducted on BRD4 and its inhibitors also makes it an excellent test case for the development of novel computational workflows. For example, Fusani et al.[307] merged active learning with the comparative binding energy (COMBINE) method and demonstrated its performance using a BRD4 dataset. Active learning was used to introduce an uncertainty estimation component to the COMBINE method, which is a powerful tool in studying the structural information of protein-ligand complexes and deriving QSAR for structurally similar series of compounds.

A fundamental component of structure based *in silico* methods is the use of X-ray crystal structures. It is therefore important to be able to rely on the starting conformations of proteins in order to make accurate predictions. For example, different active site conformations may lead to different binding poses, which can severely impact predicted binding affinities.[212,308] While MD simulations, combined with enhanced sampling methods, can be used to find the most stable binding pose, these methods come with a computational expense and it is still sensible to choose the starting structure with care. In a study on the T4 lysozyme L99A, Lim et al.[309] demonstrated that predicted relative free energy values are sensitive to initial protein conformation, even when using enhanced sampling.

In this chapter, X-ray crystal structures of BRD4-BD1 complexes in the PDB are examined and structural clustering is performed, to identify the variation of conformations and the best static representative structures of the receptor. To compare the binding site of multiple complexes and identify ligands which cause structural variation, we use WONKA.[310,311] WONKA is a tool for ligand-based, residue-based and water-based analyses of protein-ligand structural ensembles. The advantage of using WONKA over other visualisation and analysis tools, such as PyMOL,[312] is its ability to identify trends within a data set. It can identify patterns between structure and individual ligand complexes, and these observations are displayed on a web based graphical user interface. WONKA also analyses water displacements and relates which ligands displace conserved water molecules. Therefore, we are able to explore the extent of the conservation of crystallographic water molecules in the BRD4-BD1 binding pocket and highlight functional groups present in the ligands, which displace the usually highly conserved water network. Molecular docking and absolute FEP calculations are also used to assess the accuracy of predicted binding poses, with and without the water network present.

5.2 Materials and Methods

5.2.1 Structural clustering

A survey of the PDB reveals, at the time of this study, 323 X-ray crystal structures of BRD4 in complex with a variety of ligands. To identify the common sequence, multiple sequence alignment was performed using the Clustal Omega[313] alignment tool in Chimera.[314] Sequence alignment shows that 26 of the ligands are co-crystallised with BD2 and are therefore discounted. In total, 297 BRD4-BD1 complexes are taken forward for further analysis.

To prepare the receptor structures for clustering, the ligands were first removed, multiple sequence alignment was performed and the common sequence across all structures was retained, with the remaining tails removed. Protein structural clustering was performed using ClusCo,[315] a software tool for the clustering and comparison of protein models. This utilizes an open source K-means[316] code for Hierarchical Agglomerative Clustering. To identify a sensible number of groups to cluster the structures into, all-vs-all pairwise RMSD values were calculated. RMSD was based on Cα atoms. The centroid of the whole ensemble was established by clustering with the cluster number set to one.

Although there is no specified upper limit to the number of crystal structures that WONKA can analyse, a cutoff of 100 structures was found to be preferable for the software to perform smoothly. To ensure that effects of a wide range of ligand activity and structural diversity were captured, the crystal structures studied using WONKA were chosen based on structural clustering of the co-crystallised compounds. Out of the 297 crystal structures of BRD4-BD1, there are 266 unique ligands in complex with the receptor, which have accessible experimental data. Compounds that show little or no activity ($pIC_{50} \leq 5$) were discounted, leaving 175 compounds to be clustered. Ligand structural clustering was performed in DataWarrior,[317] using FragFp descriptors and Tanimoto similarity (T). FragFp is the default descriptor in DataWarrior and is a substructure fragment library based on 512 predefined structure fragments, which occur frequently in typical organic molecule structures. Tanimoto similarity is calculated by dividing the number of common features between two compounds by the total number of features available. In this work, two structures are considered to be similar if $T \geq 0.8$ and therefore are grouped into the same cluster. Representative compounds from each cluster, and their respective crystal structures, were chosen for analysis using WONKA.

5.2.2 Molecular docking

The representative compounds from ligand-based clustering were selected for molecular docking. Ligand coordinates were extracted from their original crystal structures and docked against the centroid crystal structure of BRD4-BD1 (PDB 4BJX), with and without the water network included as part of the receptor. All other crystallographic water molecules were removed in both cases. To prepare the crystal structure for docking, the receptor generation software as part of the OpenEye docking toolkit[37,253] was used. The co-crystallised compound with ID 73B was assigned as the ligand and therefore did not interact with docked molecules. A box centred around the original ligand with sides of length 15.7 Å × 20.7 Å × 19.0 Å was situated to cover the BRD4-BD1 binding cavity, giving a total receptor volume of 6151 Å^3. Constraints were applied to ensure a heavy atom contact with Asn140. Compounds to be docked were protonated according to physiological pH and prepared using OpenEye OMEGA.[253] Conformers were generated using a truncated form of the MMFF94s force field[254] with a maximum energy difference of 20 kcal mol^{-1} set from the lowest energy conformer. A maximum of 1000 conformers was allowed and those within 0.5 Å of any others were considered as duplicates and removed. Docking was performed using OpenEye FRED.[37] Compounds were docked using the high resolution setting with rotational and translational step sizes of 1 Å. To measure the accuracy of the docking with and without the water network present, heavy atom RMSD was calculated between the docked poses and the crystallographic poses of each of the ligands.

5.2.3 Absolute free energy calculations

Alchemical free energy calculations are a robust and increasingly common way to calculate ligand binding affinities. Absolute FEP simulations estimate binding free energies by calculating the free energy difference between the bound state of a ligand and the ligand free in solvent. To illustrate the impact of starting structure on these predictions, two calculations were performed, each starting from a different binding pose of the same ligand. Initial coordinates were obtained from the molecular docking study, where the ligand (PDB 83T) was docked to the centroid structure of BRD4-BD1 (PDB 4BJX), with and without active site crystallographic water molecules. This binding site water network was retained for the FEP calculation in the case where it had been involved in the docking. System preparation and the simulations were performed using GROMACS 2020.3[175] and the CHARMM force field.[318] Ligand parameters were generated using CGenFF.[240] The complexes were solvated in a dodecahedral box with an edge distance of 3 nm, to construct an explicitly modelled solvent consisting of around 32,000 TIP3P water molecules[182] and two Cl^- ions to give a net neutral charge. In absolute FEP simulations, ligand binding affinity is estimated by calculating the difference between the free energy change of decoupling the ligand from solution and decoupling from the receptor. Therefore, a system was also prepared with the ligand free in solution. Upon setup, systems were minimized for 200 ps using a steepest descent algorithm. Equilibration was performed for 200 ps in the NVT ensemble with harmonic position restraints applied to the heavy atoms with a force constant of 1000 kJ mol^{-1} nm^{-2}. Temperature coupling was achieved by using velocity re-scaling with a stochastic term and a reference temperature of 298 K. A further equilibration was performed for 4 ns in the NPT ensemble with a Berendsen pressure and temperature coupling scheme. The decoupling of the ligands from the receptor was split into 30 lambda windows, where ligand restraints were applied and vdW and Coulomb interactions were gradually turned off. The relative positions of the ligands with respect to the receptor were restrained by one bond, two angles and three dihedral harmonic potentials. To account for these restraints, a correction is applied to the free energy of binding (Equation 2.20). For the free ligand in solvent, vdW and Coulomb interactions were decoupled over 20 lambda windows. Each lambda window consisted of 1 ns of equilibration in the NPT ensemble, followed by 2 ns of data collection. Free energy changes were evaluated with the BAR method[98] as implemented in GROMACS.

5.3 Results and Discussion

5.3.1 Receptor based structural clustering

Figure 5.3 shows the distribution of the resolution of the crystal structures identified in the PDB; they have an average resolution of 1.60 Å. Overall, we can consider the structures to be high-resolution and have confidence in the quality of the crystal structure data.

Structural differences within the ensemble of crystal structures were measured using pairwise RMSD. The median and mean pairwise RMSDs are 0.56 Å and 0.58 ± 0.22 Å respectively. These RMSD values are small and

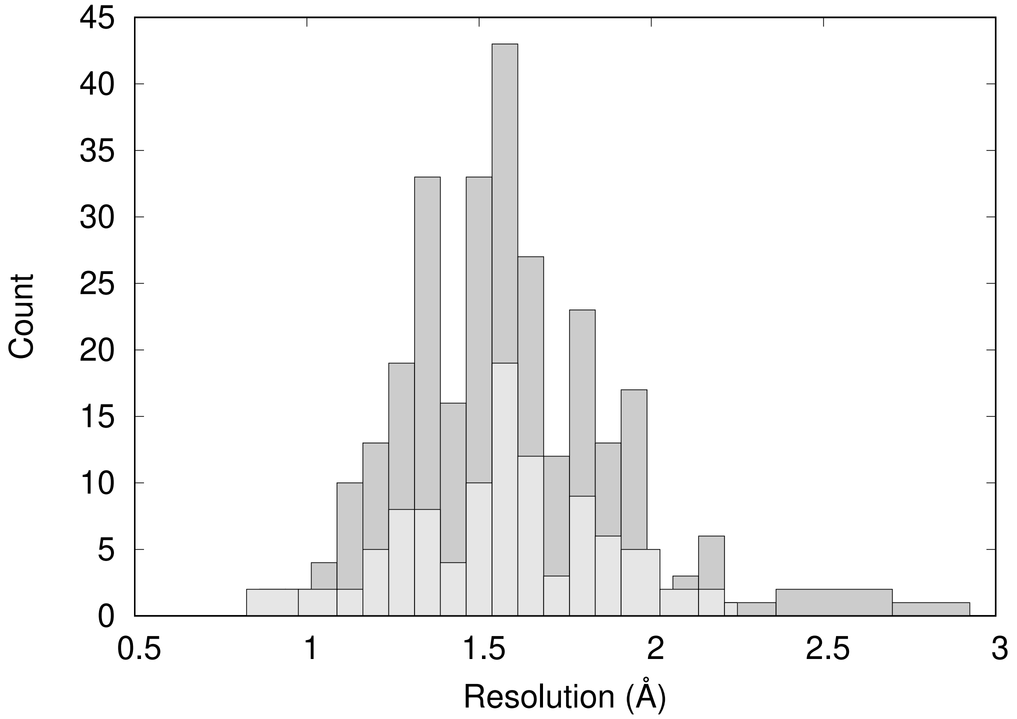


Figure 5.3: Distribution of the resolution of 297 X-ray crystal structures of BRD4-BD1 complexes (dark grey) and the 101 crystal structures analysed using WONKA (light grey).

suggest high similarity between most protein structures within the ensemble. The structural similarity between the superimposed structures can be observed in Figure 5.4. The maximum pairwise RMSD is 1.70 Å between PDB entries 5KU3 and 6V1L. The overlay of these structures (Figure 5.4) suggests the largest structural variance occurs in the tail leading to the N-terminus. There is a small amount of deviation in the ZA loop. Given the narrow distribution of RMSD values, it is sensible to group the structures into five clusters. Figure 5.5 shows an overlay of the representation crystal structures of the five structural clusters. The position of the binding site residues further demonstrates the similarity between different structures of the receptor. The centroid of the whole ensemble is the crystal structure with PDB code 4BJX, which has a resolution of 1.59 Å. With such a high number

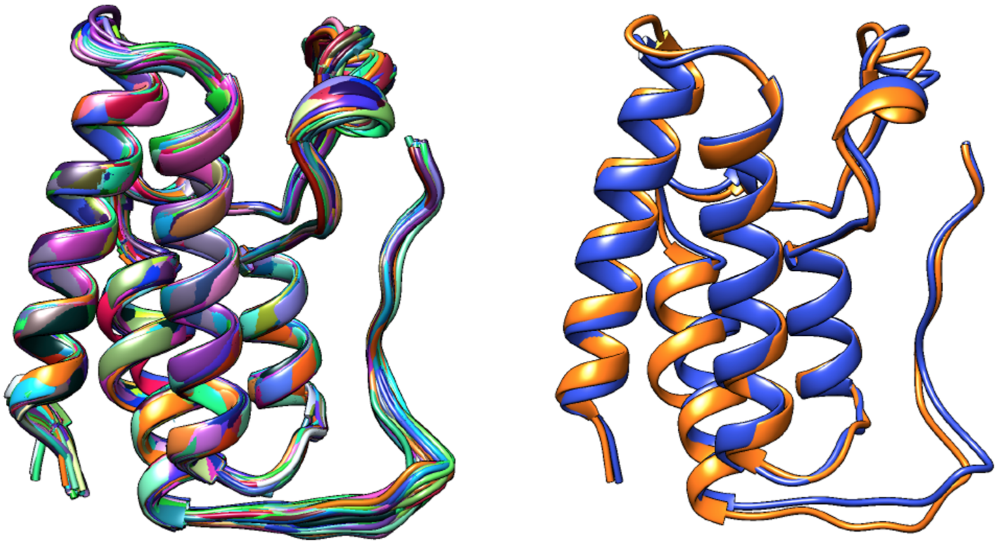


Figure 5.4: (Left) Superimposed structures of 297 BRD4-BD1 X-ray crystal structures available in the PDB. (Right) Comparison of structures PDB 5KU3 (blue) and 6V1L (orange), which have the highest pairwise RMSD.

of crystal structures available for BRD4, these results can aid the selection of the most representative structures to use for the computational study of BRD4-BD1.

5.3.2 Ligand based structural clustering

Clustering the co-crystallised ligands, which have experimental pIC_{50} values of ≥ 5, based on structural similarity resulted in 101 groups. The distribution of ligand activity for the whole data set and the 101 representative compounds from each cluster is shown in Figure 5.6. The representative compounds have a pIC_{50} range of 5.0 to 8.8 and cover a relatively wide span of activity. Therefore, no further filtering of the compounds was performed and the complexes containing these ligands were taken forward for analysis using WONKA.

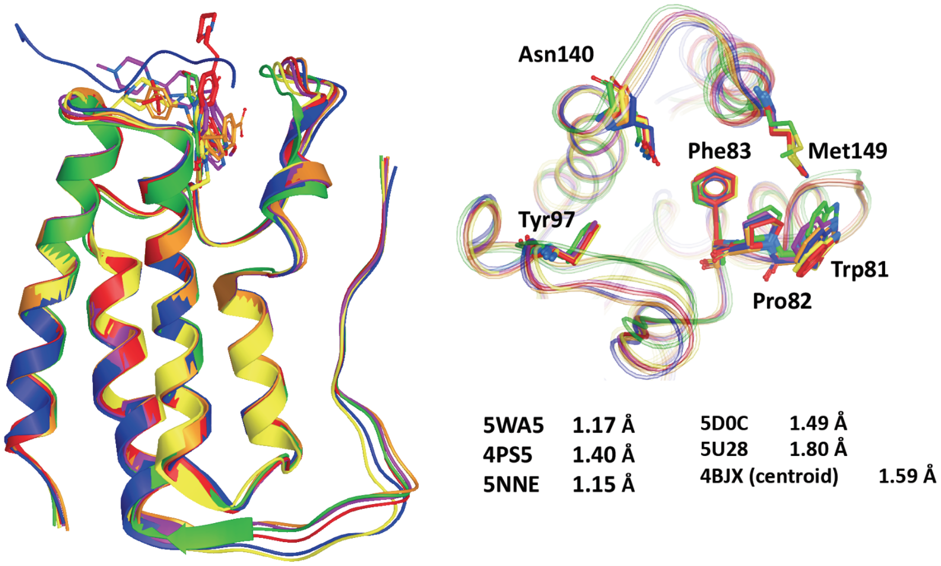


Figure 5.5: Representative X-ray crystal structures after grouping 297 structures of BRD4-BD1 into five clusters. PDB codes are 5WA5 (purple), 4PS5 (red), 5NNE (blue), 5D0C (yellow), 5U28 (lime) and 4BJX (orange). (Left) Secondary structures of the five representative structures. (Top right) Active sites with key binding residues highlighted as sticks. (Bottom right) X-ray crystal structure resolutions.

5.3.3 Structural diversity of the binding site

WONKA enables the identification of trends in the position of active site residues for multiple crystal structures of the same protein. Figure 5.7 shows the superimposition of the key binding site residues in BRD4-BD1 for each of the co-crystal structures, identified by ligand based structural clustering. WONKA clusters a particular residue's conformations into different groups based on an all-vs-all heavy atom RMSD of 2.5 Å between like-residues in a structural ensemble. Asn140, Tyr97, Met149, Trp81, Pro82 and Phe83 show 5, 3, 6, 9, 4 and 2 clusters respectively.

Visual inspection reveals that all key binding site residues, with the exception of Trp81, have a high level of conformational similarity across all

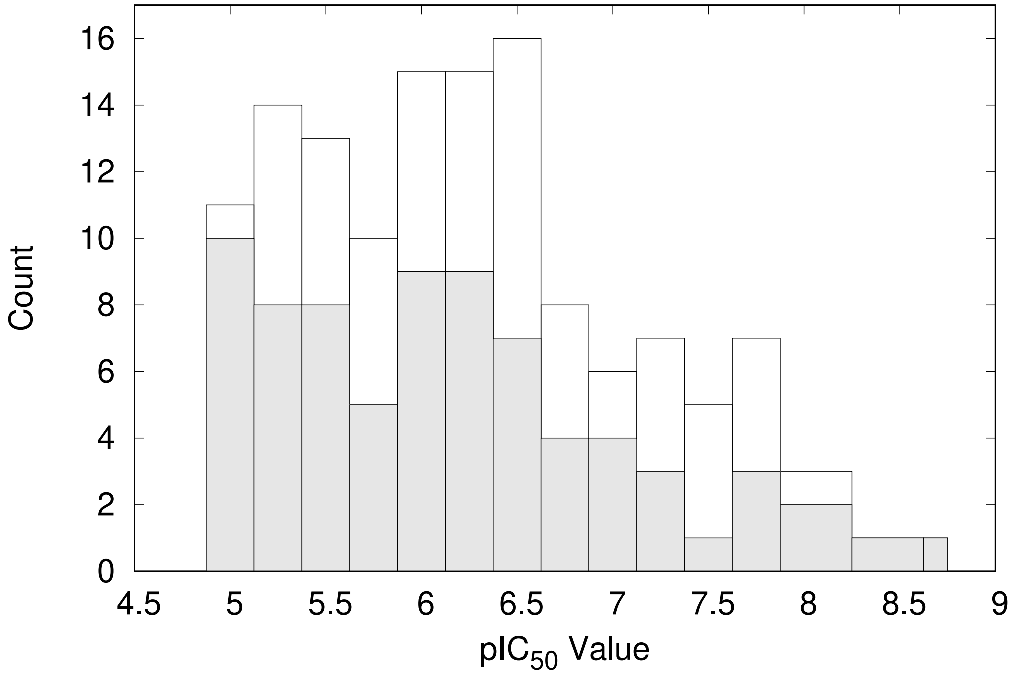


Figure 5.6: Distribution of pIC_{50} values for the representative 101 compounds (shaded) compared to the total data set (open).

of the crystal structures, regardless of the structure or activity of the bound ligand. There are three structures of Trp81 that show dissimilar conformations, highlighted in blue, red and green in Figure 5.7. There are multiple factors that could play a role in the observed disorder of Trp81. It is important to recognise that these crystal structures provide information on only one static conformation. A protein is flexible in solution, and it is possible that the positions of these Trp81 residues are more ordered when sampling a different conformation. Interactions with other amino acids within the protein and with the bound ligand can also influence the position of binding site residues. The bound ligands that correlate with these deviations in Trp81 position are shown in Figure 5.8. The compounds have experimental pIC_{50} values of 5.26, 5.89 and 5.20, which are towards the lower end of the range.

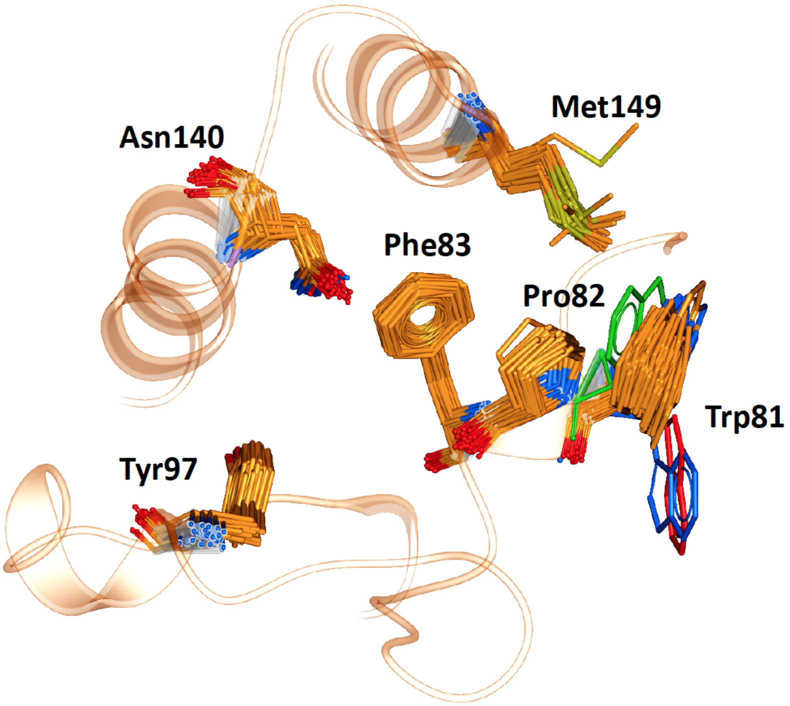


Figure 5.7: The position of key binding site residues in BRD4-BD1 over 101 X-ray crystal structures. Conformations of Trp81, which show the largest deviation, are shown in blue, red and green.

As these values were measured using different biological assays, we should be cautious about directly comparing them with each other and the remaining dataset.[319] However, for a potent compound, we would expect a pIC_{50} value upwards of 7.5, regardless of the assay conditions. A possible reason for these Trp81 positions could be that any hydrophobic interactions formed by the ligand with Trp81 do not outweigh the stability gained by polar ligand atoms forming solvent interactions. No additional binding site residue interactions are observed, in place of an interaction with Trp81. Therefore, compounds that result in this disorder of Trp81 are not desirable and do not correlate with higher binding affinity. Furthermore, it would be sensible to use structures with a 'regular' Trp81 position for the basis of computational studies.

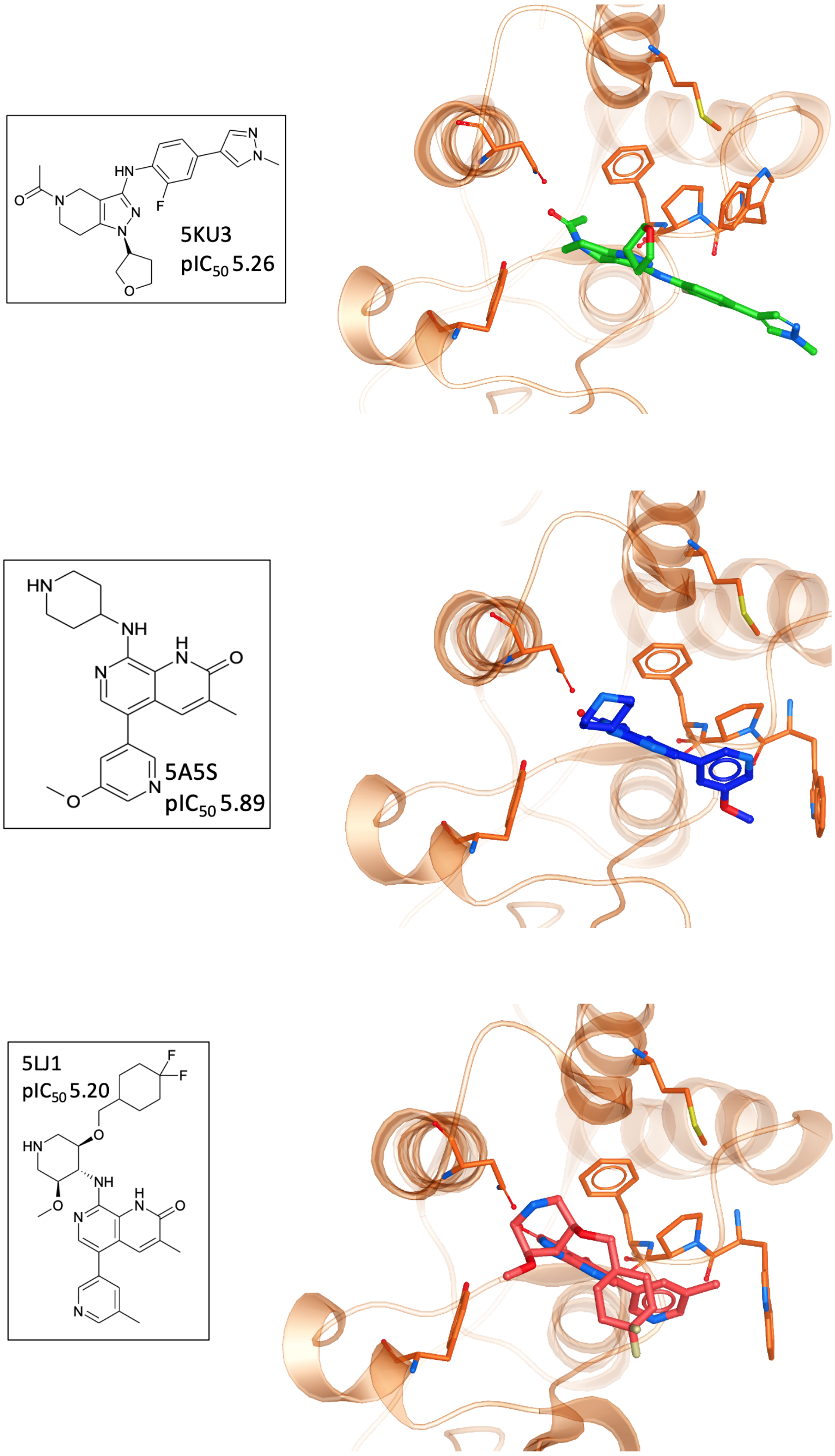


Figure 5.8: The three crystal structures and corresponding ligands, which contain Trp81 conformations most dissimilar from the whole PDB ensemble.

5.3.4 Active site water molecules

It is well documented that a network of conserved water molecules plays an important role in BRD4-BD1 ligand binding.[274–276,297,320,321] Including crystallographic water molecules is an important consideration for computational studies, and many tools have been developed to locate water molecules in protein binding sites.[322–325] Crystallographic water molecules can drastically change the binding mode in protein-ligand docking[276] and also provide stability to the system in more advanced methods such as binding free energy calculations.[196] As part of our exploration of X-ray crystal structures, we used WONKA to analyse the occupancy of the water network, which lines the BRD4-BD1 binding pocket, as shown in Figure 5.9. From the positions of the water molecules, we can identify that the water molecule at site 2 mediates a hydrogen bond between Tyr97 and bound ligands. The size of the red spheres, which represent crystallographic water molecules, reflect how conserved they are across the ensemble of crystal structures. For example, the sphere at site 1 is the largest as all crystal structures, except from one, contains a water molecule at this position. All water clusters within 8 Å of the ligand are displayed and there is a maximum distance of 1.5 Å between a point in a cluster and the cluster centre.

Using WONKA, we can easily identify the ligands which displace the water molecules in the sites where they are not present. The only crystal structure with no water molecules at site 1 or 2 is PDB 6MH1 (Figure 5.10). Water molecules at sites 3 and 4 are also displaced. Divakaran et al.[326] acknowledge the reorganisation of the usually conserved water network and

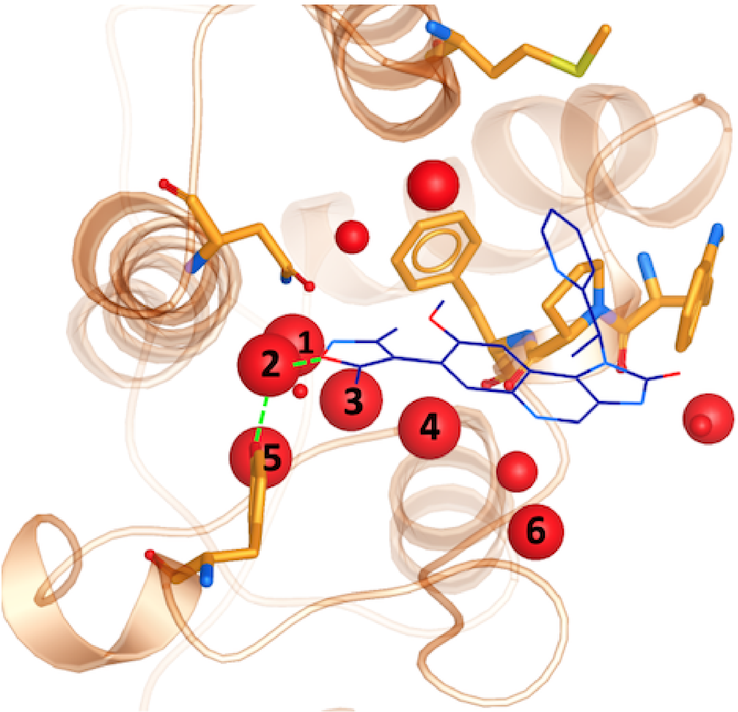


Figure 5.9: Crystallographic water molecules (red) in the binding pocket of BRD4-BD1 (orange). The size of the spheres indicate the extent of their conservation across 101 crystal structures. For perspective, a small molecule binder (PDB 3ZYU) is shown in blue.

attribute it to the fluorophenyl group of the ligand. A moderate pIC_{50} value of 5.77 was measured for this compound. However, increased selectivity over other BET receptors was observed and it was hypothesised that this was in part due to the displacement of the water molecules. In our ensemble of crystal structures, which were analysed using WONKA, there are no other fluorine containing groups which occupy the same region of the binding site, which perhaps explains why sites 1, 2, 3 and 4 are occupied by water molecules for the majority of remaining complexes.

Beside the structure previously discussed, there is one other crystal structure, PDB 5I88, which does not show water molecules at sites 3 and 4. The butenyl group on the ligand displaces the water network and induces a rear-

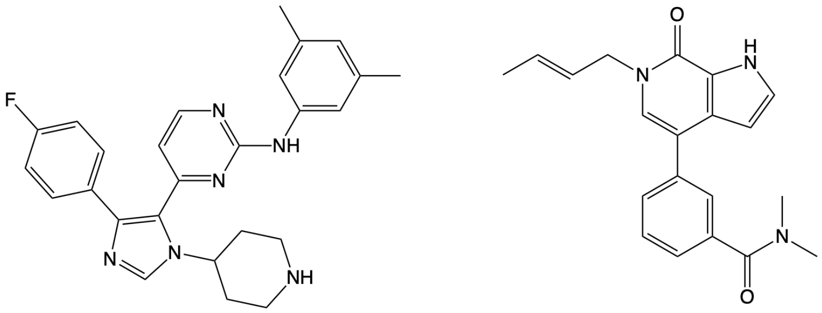


Figure 5.10: The structure of the compound that displaces crystallographic water molecules at sites 1-4 is shown on the left. The structure on the right displaces water molecules at sites 3 and 4.

rangement. The addition of a butenyl group to the compound corresponds with a reduction in activity as the butenyl containing compound has a pIC_{50} of 6.43, while its equivalent without the butenyl group has a pIC_{50} of 7.04. Crawford et al.[275] suggest that, while there may be multiple parameters that contribute to the decreased potency, the position of the active site water molecules may play a role.

Site 5 is occupied by a water molecule in all crystal structures, with the exception of PDB 4GPJ and 5DLZ. The displacement of this water molecule does not correlate with high activity ligands. Furthermore, there are ten crystal structures in our data set, which do not contain a water molecule at site 6. The corresponding ligands for these crystal structures have a pIC_{50} range of 5.30 to 7.30. The remaining water molecules depicted in Figure 5.9 are present in $\leq$ 80% of the crystal structures. While we have not found a correlation between the displacement of specific water molecules in the network and the activity of the co-crystallised compounds, we have demonstrated the extent of their conservation. We expect this analysis to be a useful tool in selecting the best crystal structure and number of crystallographic water molecules to retain in computational studies of BRD4-BD1.

5.3.5 Molecular docking

To demonstrate the importance of the conserved binding site water network in modelling an experimentally accurate system, molecular docking was performed with and without the presence of the water network. Ligands from 101 crystal structures of BRD4-BD1 complexes were docked against structure PDB 4BJX. To compare the two setups, the RMSD values between the docked poses and the crystallographic poses were calculated. On average, the improvement in RMSD when including the water molecules was 1.52 Å. The distribution of RMSD values for each data set is shown in Figure 5.11. Furthermore, 82% of the ligand poses were better predicted when including water molecules as part of the receptor. For example, Figure 5.12 shows a large difference in bound conformation of one of the compounds. The docked pose has a RMSD of 0.21 Å when water molecules are included and 1.54 Å when docked without water molecules. Different functional groups occupying different regions of the active site, such as in this example, can lead to large differences in predicted activity when using more involved, but more accurate, methods such as free energy calculations. The accuracy of the binding poses when docking with the conserved water molecules indicates that these water molecules should be retained in computational studies of BRD4-BD1. Furthermore, this aids the design of new inhibitors. Compounds should be designed with these solvent interactions in mind, while all other crystallographic water molecules are likely to be able to be displaced.

The receptor used in this study was the centroid of the ensemble of 297

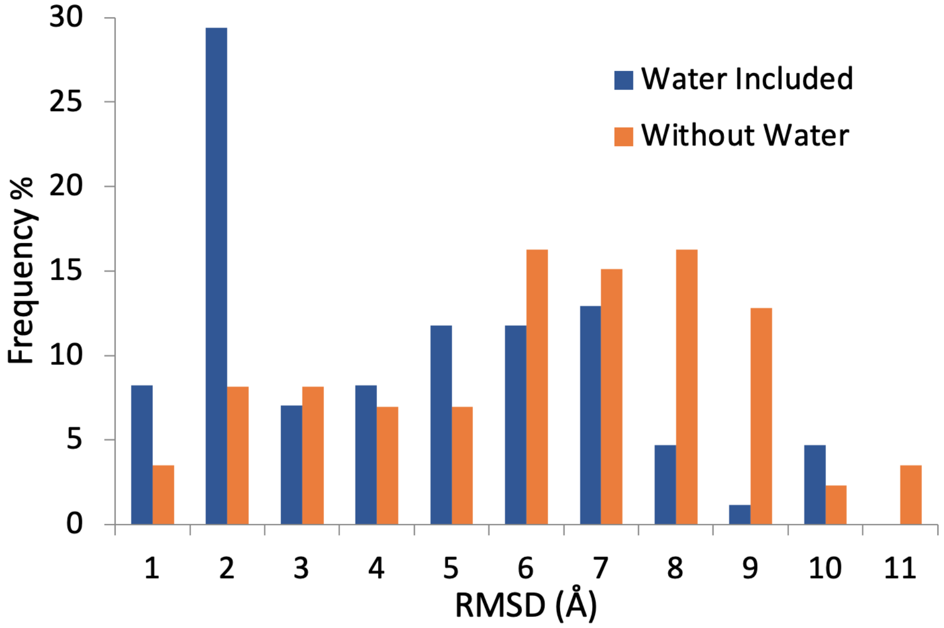


Figure 5.11: Distribution of RMSD values between docked poses and crystallographic poses. Ligands from 101 crystal structures of BRD4-BD1 complexes were docked with and without the retention of crystallographic water molecules.

PDB crystal structures of BRD4-BD1. Regardless of the inclusion of crystallographic water molecules as part of the receptor, all docked compounds show a good similarity to their original crystallographic conformations. This indicates that crystal structure 4BJX is a suitable starting conformation for the in silico study of BRD4-BD1.

5.3.6 Binding free energies

To further illustrate the impact of binding site water molecules, the free energy of binding was calculated for the two docked poses shown in Figure 5.12. The orange structure is the co-crystallised binding conformation of the compound. The structure in light blue is the docked pose when including the active site water network and the green pose is the docked pose obtained

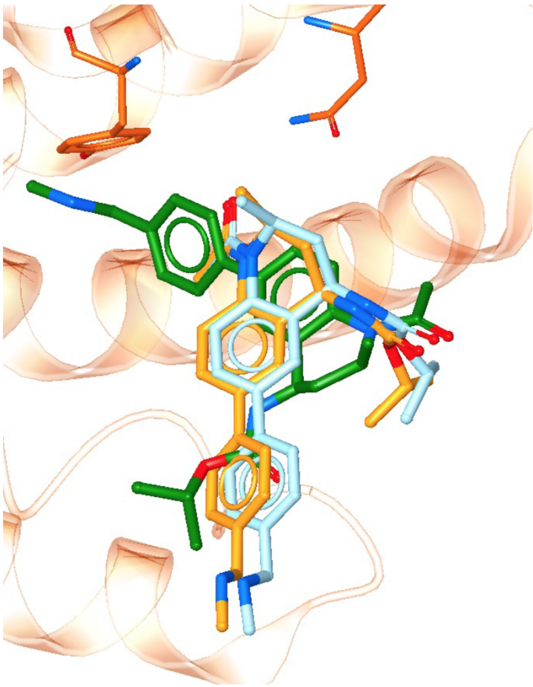


Figure 5.12: Docked poses with (light blue) and without (green) including crystallographic water molecules as part of the receptor. The crystallographic pose (orange) is also shown for comparison.

without including the water molecules. The experimental binding free energy for the compound is -8.9 kcal mol^{-1}.[327] Absolute FEP resulted in a predicted binding free energy of -11.2 ± 0.6 kcal mol^{-1} for the light blue binding pose and a value of -1.5 ± 0.5 kcal mol^{-1} for the green pose. A full breakdown of the free energies for each leg of the thermodynamic cycle (Figure 2.10) can be found in Table 5.1. There are a number of factors that can affect the accuracy of FEP calculations, such as the quality of the small molecule force field parameters and the number of lambda windows used. However, the only difference in procedure between our two calculations was the starting pose of the ligand. The large difference between our predicted free energy values demonstrates the importance of starting structure in these types of calculations. Furthermore, the inclusion of water molecules resulted in a free energy of binding which is closer to experiment. This supports our

Table 5.1: Calculated free energy values for the thermodynamic cycle (Figure 2.10) of an absolute FEP calculation. $\Delta G^{prot}_{elec+vdW+restr}$ corresponds to the free energy change when decoupling the ligand from complex, with position restraints on the ligand. $\Delta G^{solv}_{elec+vdW}$ is the free energy change when decoupling the ligand from solvent and $\Delta G^{solv}_{restr\_on}$ is the free energy correction to account for the ligand restraints. Ligand binding free energy, ΔG^{o}_{b}, is the sum of these three components. Values are shown for an absolute FEP calculation of a BRD4 inhibitor with and without the inclusion of binding site crystallographic water molecules. Free energy values are shown in kcal mol^{-1}.

	Receptor Setup	
	Docking with Water	Docking without Water
$\Delta G^{prot}_{elec+vdW+restr}$	-75.8 ± 0.6	-67.9 ± 0.5
$\Delta G^{solv}_{elec+vdW}$	61.7 ± 0.2	61.7 ± 0.2
$\Delta G^{solv}_{restr\_on}$	2.9 ± 0.0	4.7 ± 0.0
ΔG^{o}_{b}	-11.2 ± 0.6	-1.5 ± 0.5
$\lvert\Delta G^{o}_{b} - \Delta G_{exp}\rvert$	2.3 ± 0.6	7.4 ± 0.5

conclusion that these water molecules are crucial for accurately modelling a BRD4 system.

5.4 Conclusions

There has been increasing interest in BRD4 as a therapeutic target, resulting in a large number of X-ray crystal structures of the receptor in complex with small molecule ligands. In this chapter, we examined an ensemble of structures of BRD4-BD1 complexes in order to compare different conformations of the protein, without the need, for example, to carry out MD simulations. By superimposing 297 crystal structures of BRD4-BD1 and calculating pairwise RMSD values, we have found a high level of similarity between the conformations, regardless of the bound ligand. Clustering algorithms identify

PDB 4BJX as the centroid of the ensemble and clustering into five groups gave structures 5WA5, 4PS5, 5NNE, 5D0C and 5U28 as the representative structures of each cluster.

To achieve a more detailed view of the binding site, we used WONKA to compare the conformations of individual residues that are important for histone and small molecule binding. In this analysis the positions of Asn140, Tyr97, Met149, Trp81, Pro82 and Phe83 in 101 X-ray crystal structures were compared. With the exception of a handful of Trp81 conformations, the positions of these residues were extremely similar. This shows the size and shape of the BD1 cavity remains unchanged with different ligands bound, highlighting the importance of the chemical features needed in a potential inhibitor. A polar group at the head of the ligand is necessary to form both the interaction with Asn140 and a water mediated interaction with Tyr97. Simultaneously, a lipophilic group is needed to extend into the hydrophobic cavity of BD1 and strengthen ligand binding.

Water molecules also play an important role in BRD4 ligand binding. Therefore, we examined the conservation of crystallographic water molecules in the binding site. Analysis in WONKA showed that the majority of crystal structures contain the four or five water molecules generally considered important for ligand binding. In total, there are up to 11 water molecules within 8 Å of the bound ligands, which are largely conserved across the ensemble. While there have been previous studies on this highly conserved water network,[275,276] ours is the first to consider such a high number of experimental structures. Our work demonstrates the extent of the conservation and, through molecular docking and absolute FEP calculations, highlights the importance of retaining binding site water molecules in computational studies. Through this examination of BRD4-BD1 crystal structures, we have provided a quantitative basis to facilitate the selection of structures in future computational studies.

Chapter 6

Alchemical Free Energy Methods Applied to BRD4-Ligand Complexes

6.1 Introduction

Alchemical free energy calculations are increasingly gaining importance due to their application in drug design and development.[101] The accurate and reliable prediction of ligand binding free energies provides guidance and confidence in the synthesis of molecules with the potential to be lead compounds. A common use of alchemical methods, such as FEP, is in post-docking refinement, where more accurate predictions of binding affinity, compared to docking scores, are desired.[328,329] This often involves small modifications made to a hit compound to increase its potency, or improve physicochemical properties without compromising potency. An improvement in the computational expense of alchemical methods would mean that they hold a lot of promise for the high throughput estimation of binding free energies in drug discovery projects, in both an industrial and academic setting. In this chapter, the application of relative FEP and MSλD simulations to a set of BRD4-BD1 inhibitors is discussed. In particular, we compare the accuracy, manual intervention required and computational expense of each of the methods.

As discussed in the previous chapter, BRD4 is a member of the BET family and the development of a drug for its inhibition is of interest for the treatment of several diseases, most notably cancers.[278–282] As a result, BRD4-BD1 is the target of an ongoing collaboration between UoN, GSK and the University of Strathclyde. Computational methods are employed to design and assess compounds, which then actively guides the synthesis of novel potential inhibitors of BRD4-BD1. The aim is to incorporate alchemical free energy calculations into the evaluation of the compounds, prior to synthesis. For the assessment of relative FEP and MSλD approaches to this system, we use a set of GSK inhibitors that has been previously studied *in silico* by Coveney and coworkers.[330]

6.1.1 Tetrahydroquinoline series of BRD4-BD1 inhibitors

The compounds studied are based on a tetrahydroquinoline (THQ) scaffold and represent a good range of chemical functionality and binding affinities. The scaffold and its substituents are shown in Figure 6.1. There are four points of substitution, which we refer to as sites 1 to 4 and correspond to the labelled R groups on Figure 6.1. All derivatives of the scaffold have a net neu-

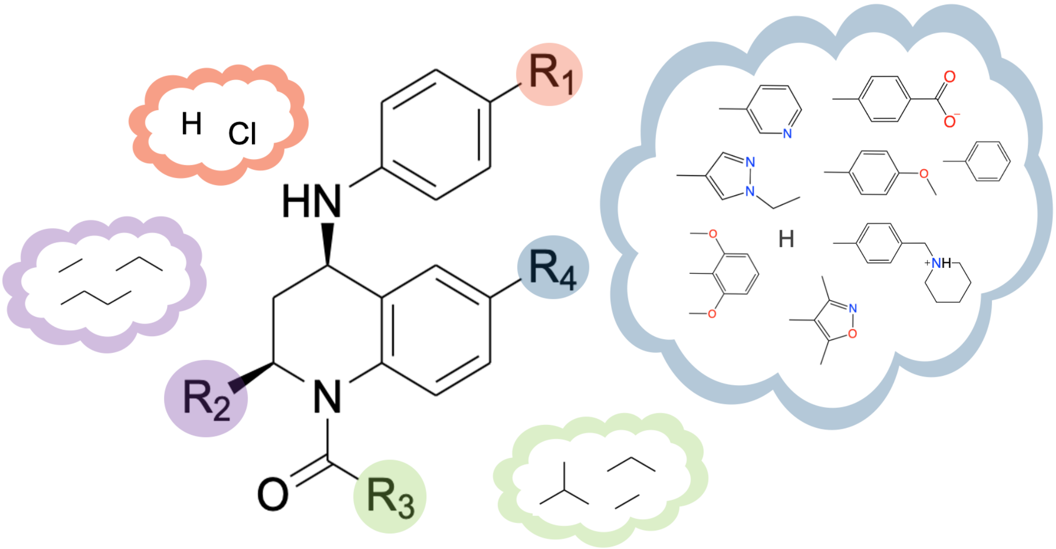


Figure 6.1: THQ scaffold of a series of BRD4-BD1 inhibitors. Substituents at each site that have been previously synthesised and characterised are shown.

tral charge except for those with the benzoic and piperidine substituents at site 4. These groups are charged under physiological conditions and present an opportunity for the refinement of RBFE calculations that involve a change in charge.

Experimental data are available for 15 compounds, based on different combinations of the substituents on the scaffold. These have a pIC_{50} range of ≤ 4.3 to 7.9, which corresponds to a binding free energy range of ~5 kcal mol^{-1}.[330] This range in activity, coupled with the relatively small modifications on each of the sites, makes this series of compounds a good test case for RBFE calculations. A previous study by Wan et al.[330] described binding free energy calculations on this series using two free energy protocols. The first approach is termed "enhanced sampling of molecular dynamics with approximation of continuum solvent" (ESMACS)[331] and is based on MM-PBSA, where solvent is treated as a continuous approximation. The second approach involved TI with enhanced sampling (TIES).[332] ESMACS was used for the full set of compounds, while the TIES calculations were split into three subsets of compounds, so that perturbations involved derivatives with the same net charge. A good correlation with experimental data was found, with a Spearman rank correlation coefficient, r_s, of 0.78 for the EMACS 3-trajectory calculations and 0.92 for TIES. Furthermore, the ESMACS protocol showed good reproducibility, with a Spearman correlation of 0.98 ± 0.02 between two independent studies performed on different supercomputers.

In this study, we investigate how the calculation of RBFE compares when using relative FEP[96,97] and MSλD[108] protocols. Relative FEP simulations involve splitting an alchemical perturbation into a series of λ windows, where a substituent is transformed into another with the progression of a λ variable. In contrast, MSλD calculations utilise λ as a dynamic variable that propagates throughout a simulation, along with the coordinates..[106,107] The introduction of additional λ coordinates means that λ-dynamics can be performed on more than one substituent in a single calculation. Therefore, RBFEs can be obtained in far fewer calculations than FEP approaches and on much quicker timescales. This concept is demonstrated herein, where the computational expense and accuracy, compared to experiment, of these methods is investigated.

6.2 Materials and Methods

6.2.1 Molecular docking

Receptor coordinates were taken from the X-ray crystal structure (PDB: 4BJX) of BRD4-BD1 in complex with small molecule inhibitor, I-BET726.[330] To prepare for docking, the protein structure was minimised for 20,000 steps using a conjugate gradient and line search algorithm and equilibrated for 1.5 ns in the NVT ensemble and 18.5 ns in the NPT ensemble. The co-crystallised ligand was retained for the equilibration period. Solvation and periodic image set up for the equilibration period is outlined in the relative FEP methodology section below. Once the protein structure was equilibrated, all water molecules were removed, with the exception of the high conserved network of five water molecules, which line the binding pocket of BRD4-BD1. Using receptor generation software as part of the OpenEye docking toolkit,[37,253] I-BET726 was assigned as the ligand and is treated as non-interacting during the molecular docking. A box centered around the original ligand with sides of length 17.7 × 19.7 × 17.0 Å was situated to cover the BRD4-BD1 binding cavity fully, giving a total receptor volume of 5906 Å^3. The 15 THQ compounds were protonated according to physiological pH and prepared using OpenEye OMEGA.[253] Conformers were generated using a truncated form of the MMFF94s force field[254] with a maximum energy difference of 20 kcal mol^{-1} set from the lowest energy conformer. A maximum of 1000 conformers was allowed and those within 0.5 Å of any others were considered duplicates and removed. Docking was performed using OpenEye FRED[37] using the high resolution setting with rotational and translational step sizes of 1 Å. Once docked, OpenEye FRED provides ten sets of ligand coordinates that display the best docking scores. With the exception of compound **7**, all compounds exhibited one conserved binding mode, with little variation between each set of the ten best coordinates. Within the small movements of this binding mode, the pose taken forward for each compound was chosen to optimise the overlap between the common core of the THQ scaffold. Compound **7** displayed two binding modes, with the common binding pose also taken forward for free energy of binding evaluation.

6.2.2 Multi-site λ-dynamics simulations

Atoms belonging to all derivatives of the THQ scaffold were identified using a MCS search. The common core used for the neutral set of substituents is shown in red in Figure 6.2 and the core used for the charged substituents is shown in blue. All remaining atoms were fragments or anchor atoms, which are coupled and decoupled from the system as their corresponding λ variables propagate through the simulation. Fragments correspond to the parts of the compound that are treated as substituents. Anchor atoms are the attachment points between the common core and the fragments and become part of the substituents once the simulation is initiated. Once an initial common core was identified, the core, fragments and anchor atoms were manually altered so that additional atoms became part of the fragments. Although all atom types on the amide and THQ groups of the ligand scaffold are consistent, regardless of the substituents, not all atoms are chosen to belong to the common core. This is to allow a change in the partial charges assigned

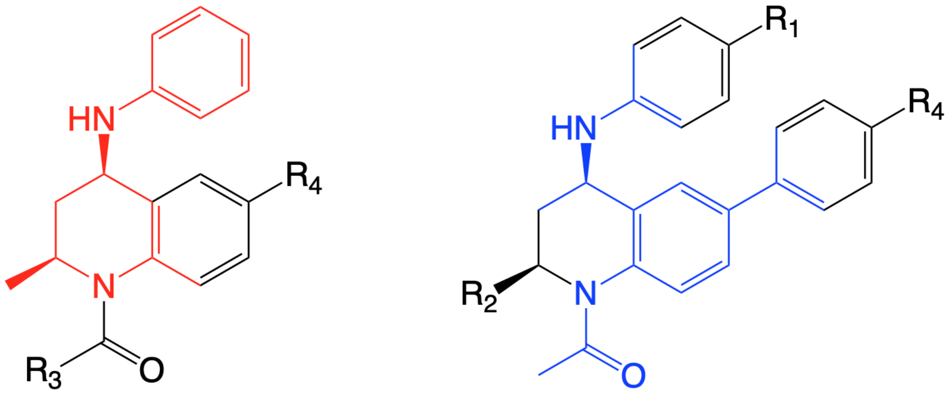


Figure 6.2: Common cores used in MSλD calculations. Atoms in red were used as the core for calculations involving neutral substituents. Atoms in blue were used as the core for charged substituents. All other atoms on the THQ scaffold, are treated as substituents or anchor atoms, which are perturbed during MSλD.

to each of the atoms, which are affected by the substituent attached, thereby enabling a better representation of the electrostatics of the ligand.

During λ dynamics, the charge of the compound must sum to an integer net charge, regardless of the combination of substituents at each site. Therefore, the partial charges of substituents at one particular site are normalised so that each substituent has the same total net partial charge. An exception is when a charge perturbation is performed, with the addition of a protonated or deprotonated substituent. For example, the net charge of compound **15** is -1 (compound indexing shown in Table 6.2). Therefore, the compound changes from neutral to charged during the λ dynamics simulation. The alteration of partial charges for the preparation of λ dynamics is termed charge re-normalisation and is performed using an algorithm, developed by the Brooks group (to be published). Initial partial charges were obtained from atom type matching with existing parameters in the CHARMM force field, using CGenFF.[123] There was an average RMSD of 0.015 e between the original CGenFF charges and the adjusted charges. All other parameters, attributed to bond lengths, angles and dihedral angles, remained unchanged from the CGenFF initial guess parameters.

Two systems were built, one composed of the ligand in solution and the second with the ligand in complex with BRD4-BD1 (PDB: 4BJX[330]). The ligand, receptor and solvent coordinates for the complex site were obtained from the equilibrated structure and molecular docking, as detailed above. Ligand topologies were constructed using a multiple topology approach.[106,208] This is a similar method to the dual topology approach in FEP, where all substituents explicitly exist in the topology, attached to the same common core. For the ligand in solvent system, the ligand was solvated in a cubic periodic boundary cell with 1755 TIP3P water molecules.[182] All following simulations were performed using the CHARMM molecular simulation package with the domain decomposition (DOMDEC) computational kernels on GPU.[69,333,334] MD simulations were run in the NPT ensemble at 298 K and 1 atm using a Nosé-Hoover thermostat[335] and Langevin pressure piston with a friction coefficient of 20 ps^{-1}.[257] A timestep of 2 fs was used, with hydrogen-heavy atom bond lengths constrained with the SHAKE algorithm.[168] A cutoff distance of 12 Å was used for all long-range interactions, with a switching function at a distance of 10 Å.

The THQ compounds were split into three sets: those with a net neutral charge (compounds **1** to **9**), those with a net charge of +1 (compounds **10** to **12**) and finally those with a net charge of -1 (compounds **13** to **15**). Considering only compounds with a net neutral charge, there is one substituent at site 1, one at site 2, three at site 3 and seven at site 4. Similar to FEP, the accuracy of MSλD is impacted by the sizes of the perturbations. Although there are no hard rules about the number of substituents or sites that can be handled, generally, the smaller the perturbation between each substituent, the more substituents or sites that can be used. In our dataset, seven quite varied substituents on one site, along with other sites of substitution, means that it is sensible to split it into two sets of calculations. Substituents on site 4 were split into two groups based on their similarity, with the phenyl, methoxyphenyl, isoxazole and ethylpyrazole substituents in one group and the phenyl, hydrogen, pyridyl and dimethoxyphenyl substituents in the second. The phenyl substituent was included in both sets as the reference compound. For comparison, a single MSλD calculation with all neutral substituents was performed.

For all MSλD calculations, the ALF algorithm was used to identify appropriate biasing potentials to flatten the potential energy landscape between substituents.[209] A soft-core potential was used to scale all nonbonded interactions by λ and to prevent end-point singularities.[200,201,208] To identify initial biases for the complex system, 50 simulations of 100 ps each were performed, followed by 30 simulations of 1 ns to refine the biases. ALF was performed for the ligand in solution system for 50 simulations of 100 ps, followed by 20 simulations of 1 ns. Production simulations were run for 20 and 50 ns for the solution and complex systems respectively, with the first 5 ns of each discarded as equilibration. Five replicas of each production run were performed using a different random seed. End-state populations were binned using a $\lambda \geq 0.99$ cutoff criterion and the final relative free energy of binding values were calculated by Boltzmann re-weighting end-state populations to the original biases and then using Equation 6.1.[106,207] Uncertainties were calculated as the standard deviation of the mean value over the five independent runs.

$$\Delta\Delta G_{i\rightarrow j} = -k_B \ln \frac{P_j}{P_i} \tag{6.1}$$

Equation 6.1 shows how relative binding free energies are calculated as the ratio of the amount of time one ligand is sampled compared to a reference ligand. In our calculations, compound **3** was chosen as the reference ligand, because the hydrogen group at site 1 and methyl groups at sites 2 and 3 are the most common substituents at these sites across all of the compounds. Furthermore, the phenyl group at site 4 is most similar to all other substituents at this position and therefore involves the smallest perturbation between substituents.

As changing the net charge of the compound adds a layer of complexity, MSλD calculations involving charged substituents were constructed in a different way to the neutral substituent calculations. Separate simulations were performed with the neutral form of each charged substituent as the reference compound. For example, for the negatively charged compounds, benzoic acid was used as the reference substituent on site 4. The deprotonated form, benzoate, was included as a substituent for MSλD. Substituents attributed to compound **3** were also included in the MSλD calculation, so that the relative binding free energy with respect to compound **3** could be calculated, for consistency. Using benzoic acid on site 4 as the reference, compared to a phenyl group, meant there was a smaller perturbation and the change in net charge could be accounted for more effectively. The same approach was used for compounds with a piperidine substituent, which is protonated at physiological pH.

6.2.3 Relative free energy perturbation simulations

Dual topologies were constructed, with compound **3** as the reference compound, for each alchemical transformation. For example, Figure 6.3 shows the ligand topology for the transformation of compound **3** to compound **1**. When λ=0, the phenyl group is interacting with the system and when λ=1, the methoxybenzene is interacting. Using input generated by CHARMM-GUI,[255] all complex systems were solvated in a cubic periodic boundary cell with edge distances of 18 Å to construct an explicitly modelled solvent consisting of around 22,000 TIP3P water molecules.[182] Depending on the net charge of the ligand, Na^+ or Cl^- ions were added, to neutralise the system. To optimise the solvent positions, all heavy atoms were fixed, except for water molecules, during 50 steps of steepest descent and 50 steps of Adopted Basis Newton-Raphson minimisation. Potential energy evaluations were performed with the CHARMM force field.[119] To ensure a fair comparison of binding free energies obtained from FEP and MSλD calculations, the charge renormalised ligand parameters, adapted from CGenFF,[240] were used. Systems containing the ligand in solution, without the receptor, were also set up using input from CHARMM-GUI.[255] Ligands were solvated in a cubic periodic boundary cell with around 2,300 TIP3P water molecules. Minimisation and equilibration were performed using the same protocol as for the protein-ligand complexes.

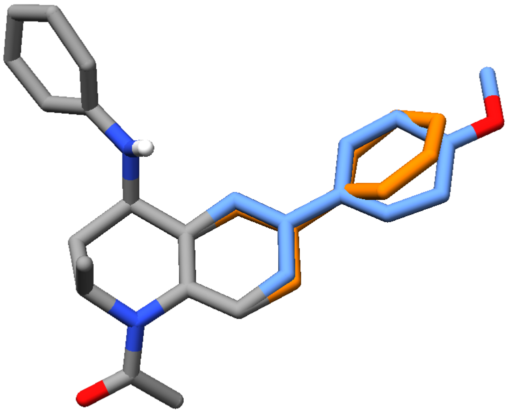
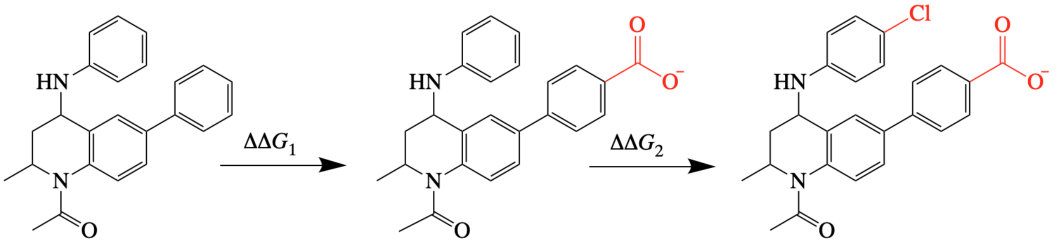


Figure 6.3: Dual topology constructed for the alchemical transformation of a phenyl group (orange) to a methoxybenzene group (blue), attached to a THQ scaffold (grey).

Once set up, all systems were minimised for 20 ps using a conjugate gradient and line search algorithm using the NAMD simulation software.[173] Protein backbone and side chain restraints were applied using harmonic constraints with force constants of 10 kcal mol^{-1} Å^{-2} and 5 kcal mol^{-1} Å^{-2} during a heating period of 50 ps. Systems were heated to 298 K in increments of 10 K. Restraints were removed for 0.1 ns of equilibration in the NVT ensemble and 4.9 ns in the NPT ensemble, with a 2 fs timestep. The temperature was controlled using Langevin dynamics parameters, with a friction coefficient of 5 ps^{-1} for all equilibration and FEP simulations. Constant pressure was maintained using the Langevin piston Nosé-Hoover method[257] with a target pressure of 1 atm. During equilibration, a cutoff distance of 12 Å was used for vdWs pairs, with a switching function at a distance of 10 Å. Long range electrostatic interactions were computed using the PME method.[258] The SHAKE algorithm[168] was used to fix all bond lengths involving hydrogen atoms.

To develop an efficient protocol for FEP calculations on this series of BRD4 inhibitors, a series of benchmark calculations was performed. The



3

15

Figure 6.4: To calculate the binding free energy of compound **15**, relative to compound **3**, an intermediate step is required. The relative binding free energy is the sum of $\Delta\Delta G_1$ and $\Delta\Delta G_2$. Substituents being added or transformed are shown in red.

relative free energy of binding of compound **1**, with respect to compound **3**, was calculated using 8, 10, 16, 20 and 25 λ windows. For each λ window, 2 ns of equilibration was performed, followed by 1 ns of data collection. Electrostatic interactions of outgoing atoms were decoupled from the system from λ=0 to λ=0.5, while the electrostatics for incoming atoms were coupled to the system from λ=0.5 to λ=1. For all simulations, a soft-core potential was used to avoid "end-point catastrophes". The effect of reducing the length of the data collection period for each λ window was then tested by performing the perturbation with 20 λ windows, 2 ns of equilibration and 0.5 ns of data collection. Finally, equilibration of lengths 1 ns and 0.5 ns were tested, using 20 λ windows and 1 ns of data collection. The average value over three replicas was calculated for each combination of FEP parameters, with free energy values evaluated using the BAR method[98] as implemented in the ParseFEP tool in VMD.[268]

Once the optimal number of λ windows, equilibration length and data collection length were established, the relative free energies of binding were calculated for the remaining compounds. As substituents on two sites of the common scaffold are modified, compared to compound **3**, for compounds **8**, **9** and **11** to **15**, an intermediate FEP step was required. For example, to calculate the relative free energy of binding of compound **15**, FEP calculations were performed for the changes shown in Figure 6.4. First, site 4 was perturbed from a phenyl group to a benzoic acid substituent. In a separate simulation, the hydrogen atom on site 1 was then transformed to a chlorine substituent. The sum of the free energy changes for these transformations resulted in the total relative free energy of binding of compound **15**, with respect to compound **3**. Compound **1** served as the reference for transformations to compounds **8** and **9**, and compound **10** was the reference for transformations to compounds **11** and **12**. Therefore, including replicas, reverse transformations and ligand in solution simulations, to obtain the full RBFE data set for the 14 compounds, with respect to compound **3**, a total of 168 FEP simulations were required.

6.3 Results and Discussion

Relative FEP parameters such as number of λ windows, equilibration and data collection length is often a balance between obtaining sufficient sampling of each λ state, while keeping the calculation to a reasonable timescale. Therefore, we firstly present our findings for the most effective parameters to use for our system of interest. Next, we discuss the calculation of the biasing potentials for the MSλD calculations. On demonstration of the reliability of our procedures, we compare the accuracy of relative FEP and MSλD with respect to experimental binding affinities. Lastly, an assessment of the investment required for each method, in terms of both computational and human time, is presented.

6.3.1 Relative FEP Benchmarking

To establish the best number of λ windows to use for relative FEP calculations on this series of BRD4 inhibitors, perturbations from compound **3** to compound **1** were performed with 8, 10, 16, 20 and 25 windows. This alchemical perturbation involved the transformation of a phenyl substituent on site 4 of the THQ compound to a methoxybenzene substituent. To assess the performance of the calculations, three criteria were taken into account. First, a comparison between the predicted relative free energy of binding and the experimental value was made. Second, the standard deviation of the mean value of the BAR error over three independent replica runs was calculated. Third, the convergence was measured by plotting the relative binding free energy calculated using an increasing fraction of the simulation data. The free energies using the reverse proportion of the data were also plotted. Convergence plots are important for ensuring that the free energy is being measured for an equilibrated system. This graphical method of assessing convergence, outlined by Klimovich et al.,[206] helps identify any non-equilibrated regions throughout the simulation.

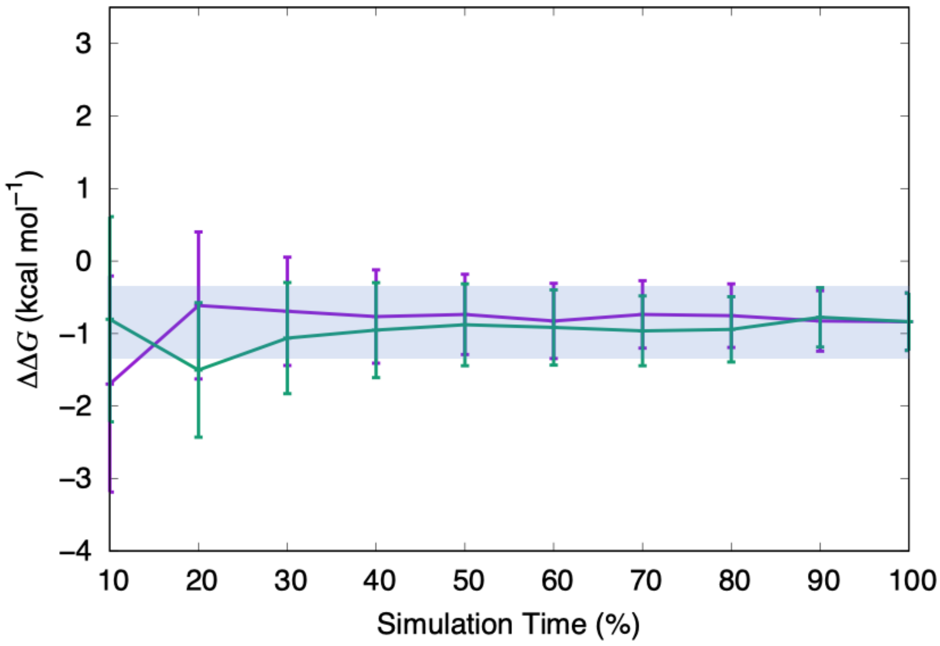
Table 6.1 shows the mean predicted relative binding free energies over three replicas, their errors and the absolute difference with experimental values. All predicted values are within chemical accuracy of the experimental values, which is generally considered to be 1 kcal mol^{-1}. However, there is an increase in their absolute differences with a decreasing number

Table 6.1: Benchmarking of relative FEP protocols. Varying numbers of λ windows, equilibration time and data collection time were tested. RBFE predictions are compared to experiment.

λ Windows	Equilibration (ns)	Data Collection (ns)	$\Delta\Delta G_{calc}$ (kcal mol^{-1})	Error (kcal mol^{-1})	Absolute Difference (kcal mol^{-1})
25	2	1	-0.5	0.3	0.2
20	2	1	-0.8	0.4	0.5
16	2	1	-0.6	0.5	0.3
10	2	1	-1.2	0.6	0.9
8	2	1	0.3	0.6	0.8
20	2	0.5	-0.8	0.6	0.5
20	1	1	-1.1	0.4	0.8
20	0.5	1	-0.5	0.4	0.2

of λ windows. Furthermore, the error also increases. This is to be expected, as decreasing the number of intermediate steps between the transformation means that there will be a poorer overlap of phase space between each window. For reliable estimations, an error of no more than 0.5 kcal mol^{-1} is desirable. This corresponds to a variation in a pIC$_{50}$ value of approximately 0.4. With this in mind, FEP with 20 or 16 λ windows appears to be the best approach. Figure 6.5 shows the convergence plots for these perturbations. Convergence plots for all benchmark FEP calculations can be found in Appendix B. An agreement, within error, between the forward and reverse free energies is a sign of an equilibrated system. The shaded bar on the plots indicates an error range of 0.5 kcal mol^{-1}, centred on the final relative free energy value. These plots show that FEP with 20 λ windows results in free energies that are better converged. Therefore, relative binding free energies in this study are predicted using 20 intermediate steps between the initial and final states.

In an attempt to gain computational speed, perturbations with data col-



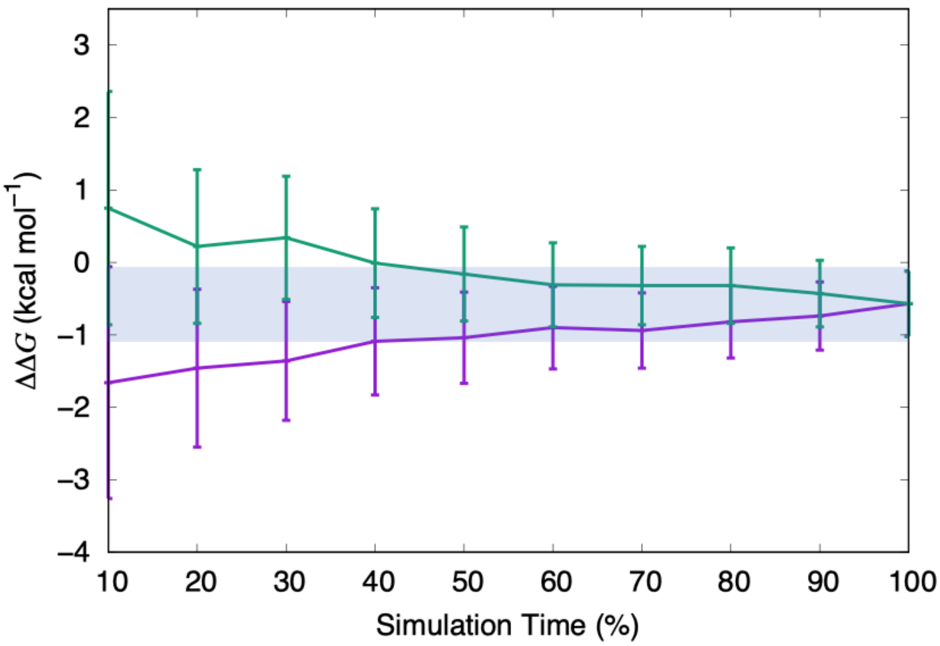


Figure 6.5: Convergence assessment of the transformation of a phenyl substituent at site 4 to a methoxybenzene substituent. (Top) Using 20 λ windows with 2 ns of equilibration and 1 ns of data collection. (Bottom) Using 16 λ windows with 2 ns of equilibration and 1 ns of data collection. The forward (purple line) and the reverse (green line) simulation time series are shown. The horizontal shaded bar indicates the equilibrated region.

lection periods of 0.5 ns for each λ window were tested. This resulted in an error of 0.6 kcal mol^{-1} (Table 6.1). Furthermore, poor convergence (Figure B.4) was observed. Therefore, 1 ns of data collection for each λ window was performed for all FEP calculations. Equilibration periods of 1 ns and 0.5 ns were also tested for each λ window. Reducing the equilibration of the windows to 0.5 ns did not affect the error or convergence of the predicted relative binding free energies. Therefore, we conclude that a protocol of using 20 λ windows with 0.5 ns of equilibration and 1 ns of data collection results in a good compromise between accuracy and computational efficiency.

6.3.2 Adaptive landscape flattening

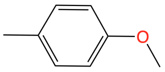
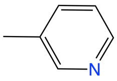
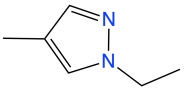
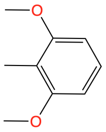
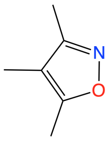
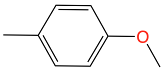
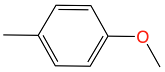
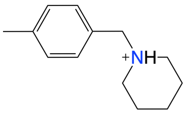
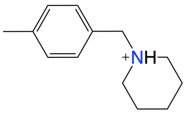
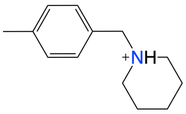
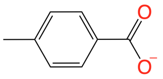
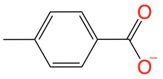
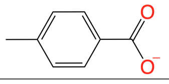
ALF is the process of calculating the biases to flatten the alchemical potential energy landscape between substituents on a given site, to ensure sufficient sampling of all substituents.[208,209] To assess the fixed biases that were used for MSλD, their convergence along the serial ALF simulations was investigated. Figure 6.6 shows that at the end of each ALF process, the biases were stable and therefore suitable to be used for data collection.

6.3.3 Relative binding free energies

Accuracy and reliability

Relative binding free energies are shown in Table 6.2. Results shown for the neutral compounds using MSλD are RBFEs calculated from splitting the compounds into two separate calculations, as this improved the accuracy. RBFE predictions when including all substituents in one calculation can be

Table 6.2: Predictions of binding affinity for a series of BRD4-BD1 inhibitors based on a THQ scaffold (Figure 6.1). Predictions calculated using MSλD and relative FEP are compared to experiment. All relative free energy values are shown in kcal mol^{-1}.

ID	R1	R2	R3	R4	$\Delta\Delta G_{exp}$	$\Delta\Delta G_{MS\lambda D}$	$\Delta\Delta G_{FEP}$	$\lvert\Delta\Delta G_{MS\lambda D}$ $-\Delta\Delta G_{exp}\rvert$	$\lvert\Delta\Delta G_{FEP}$ $-\Delta\Delta G_{exp}\rvert$
1	H	Me	Me		-0.3 ± 0.1	-0.4 ± 0.1	-0.5 ± 0.4	0.1	0.2
2	H	Me	Me	H	1.6 ± 0.1	2.3 ± 0.1	2.1 ± 0.4	0.7	0.5
4	H	Me	Me		0.0 ± 0.1	0.6 ± 0.4	0.4 ± 0.1	0.6	0.4
5	H	Me	Me		-1.5 ± 0.1	1.0 ± 0.1	0.2 ± 0.4	2.5	1.7
6	H	Me	Me		1.6 ± 0.1	1.4 ± 0.1	0.4 ± 0.6	0.2	1.2
7	H	Me	Me		1.3 ± 0.1	1.4 ± 0.1	1.3 ± 0.4	0.1	0.0
8	H	Me	Et		0.4 ± 0.1	1.0 ± 0.1	-1.2 ± 0.6	0.6	1.5
9	H	Me	i-Pr		≥ 3.4	1.7 ± 0.8	-1.8 ± 0.6	≥ 1.7	≥ 5.2
10	H	Me	Me		-1.1 ± 0.2	-0.2 ± 0.2	-0.1 ± 0.5	0.9	1.0
11	H	Et	Me		0.0 ± 0.4	0.0 ± 0.2	0.6 ± 0.6	0.0	0.6
12	H	Pr	Me		1.8 ± 0.1	2.0 ± 0.2	1.7 ± 0.6	0.2	0.1
13	H	Pr	Me		1.9 ± 0.1	1.8 ± 0.1	1.4 ± 0.6	0.1	0.5
14	H	Et	Me		0.1 ± 0.3	-0.2 ± 0.1	-0.4 ± 0.6	0.3	0.5
15	Cl	Me	Me		-1.4 ± 0.2	-0.6 ± 0.1	-1.6 ± 0.6	0.8	0.2

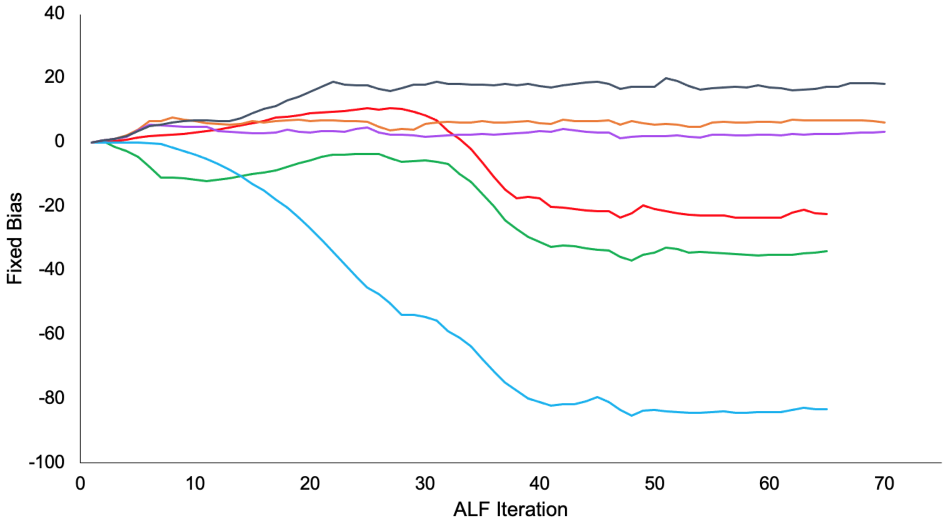


Figure 6.6: Convergence of the fixed bias for each substituent at site 4 as the ALF simulations progress. Substituents at site 4 include methoxyphenyl (red), ethylpyrazole (green), isoxazole (light blue), hydrogen (orange), pyridyl (purple) and dimethoxyphenyl (dark blue).

found in Appendix C. Overall, the two methods have similar levels of accuracy compared to experiment. MSλD calculations resulted in an average difference of 0.6 ± 0.7 kcal mol^{-1} to experiment and for relative FEP predictions this was 1.0 ± 1.3 kcal mol^{-1}. Furthermore, when discounting the large deviation from experiment found for compound **9**, the average differences for the MSλD and relative FEP calculations become 0.6 ± 0.7 kcal mol $^{-1}$ and 0.7 ± 0.5 kcal mol^{-1}, respectively, showing there is little difference in accuracy between the two methods. The Spearman correlation (r_s) between the rank order of the predicted and experimental RBFEs have also been calculated, which shows that both methods have a good, and comparable, correlation with experiment. RBFE predictions calculated using MSλD have a r_s of 0.80, while relative FEP predictions have a r_s of 0.70. With this small dataset, these differences in r_s are not statistically significant. These results show that MSλD and relative FEP (using the λ window parameters selected from benchmarking) are accurate methods for the prediction of RBFEs and identifying highly active compounds out of a set of congeneric compounds. Whilst the comparison to experiment is similar to the EMACS (r_s 0.78) and TIES (r_s of 0.92) method presented by Wan et al.,[330] MSλD predicts $\Delta\Delta G$ values for the combinatorial set of substituents at each site and so a larger space of 28 compounds is explored using the four MSλD simulations presented in this work. This is discussed in more detail in the computational expense section.

As discussed previously, all neutral substituents on site 4 were initially included as part of one MSλD calculation. For comparison, the substituents were also split into two calculations. The average RBFE compared to experiment was 1.4 ± 1.4 kcal mol^{-1} when the substituents were included in one simulation, while the difference was 0.8 ± 0.8 kcal mol^{-1} when splitting them into two sets of calculations. Furthermore, RBFE predictions obtained from one calculation have a r_s of 0.30, compared to a r_s of 0.84 for the two sets. The increased accuracy when splitting the substituents into two calculations is not surprising. When including all site 4 substituents with a net neutral charge, there are seven possible substituents, which means that all combinations of physically meaningful end points are sampled less during the simulation and less likely to achieve converged results. This is also reflected by the larger uncertainties of the single MSλD simulation, which have an average of 0.4 ± 0.2 kcal mol^{-1} compared to 0.2 ± 0.2 kcal mol^{-1} for the two calculations. Solutions for more accurate predictions in a single simulation could be to use longer simulation times or enhanced sampling methods. A study by Vilseck et al.[109] demonstrated that accuracy within 0.8 kcal mol^{-1}

can be achieved for perturbation sites with seven substituents when using MSλD with biasing potential replica exchange,[336] to enhance end-state sampling.

A common limitation to RBFE methods is their lack of reproducibility.[337] Like all MD-based methods, this arises from the ensemble averaging of macroscopic properties over microscopic states. Therefore, the quality of the predictions relies on how well the microscopic states have been sampled. To address this issue, it is common practice to run multiple independent calculations with different initial velocities and average the results across the replicas. Uncertainties can be estimated by calculating the standard deviation around the averaged free energies. In our calculations, five replicas were performed for the MSλD calculations and three replicas were performed for the relative FEP calculations. Three replicas were chosen for relative FEP due to the significantly higher computational cost associated with this method (discussed in the next section). The uncertainties associated with the predictions were lower for the MSλD calculations, with an average of 0.2 ± 0.2 kcal mol^{-1}, compared to an average of 0.5 ± 0.1 kcal mol^{-1} for the relative FEP calculations. Therefore, more reliable estimations of binding affinity are achieved using MSλD, especially when there are more than two sites of perturbation. In these cases, to obtain RBFE values using relative FEP, intermediate transformations are necessary and the uncertainty accumulates over the two simulations (free energies and their associated uncertainties for the intermediate calculations can be found in Appendix D). Using MSλD, only one calculation is required, with an uncertainty that is comparable to when there is only one site of perturbation.

Outliers

Compound **9** has a pIC_{50} of ≤ 4.3 and an experimental RBFE of ≥ 3.4 kcal mol^{-1} with respect to compound **3**, indicating that it has no activity towards BRD4-BD1. The difference in substituents between compound **9** and compound **1**, which has a pIC_{50} of 7.0 ± 0.1, is an isopropyl group at site 2, compared to a methyl group. As noted by Wan et al.,[330] the position of site 2 occupies a small lipophilic site in the BRD4-BD1 binding pocket, which offers little room for large substituents without structural reorganisation. Therefore, we infer that the isopropyl group is too large for this part of the binding pocket. A representative compound in the binding site of BRD4-BD1 is shown in Figure 6.7. The large discrepancy between the experimental and predicted RBFEs for compound **9** suggests that MSλD and relative FEP methods are less accurate when predicting non-binders. Additionally, isopropyl is not well represented in the CGenFF force field,[123] particularly the dihedral angle parameters when attached to an amide, which may also contribute to the deviation from experiment. A difference larger than 1.5 kcal mol^{-1} from experiment was also found for compound **5** for both RBFE methods. Investigation into the force field parameters and interactions made by the pyrazole derivative at site 4 of compound **5** is ongoing to try and identify a reason for this difference.

Charge perturbations

Perturbations that involve a change in net charge of the ligand are difficult and should generally be avoided. Cournia et al.[328] explain that this is due to

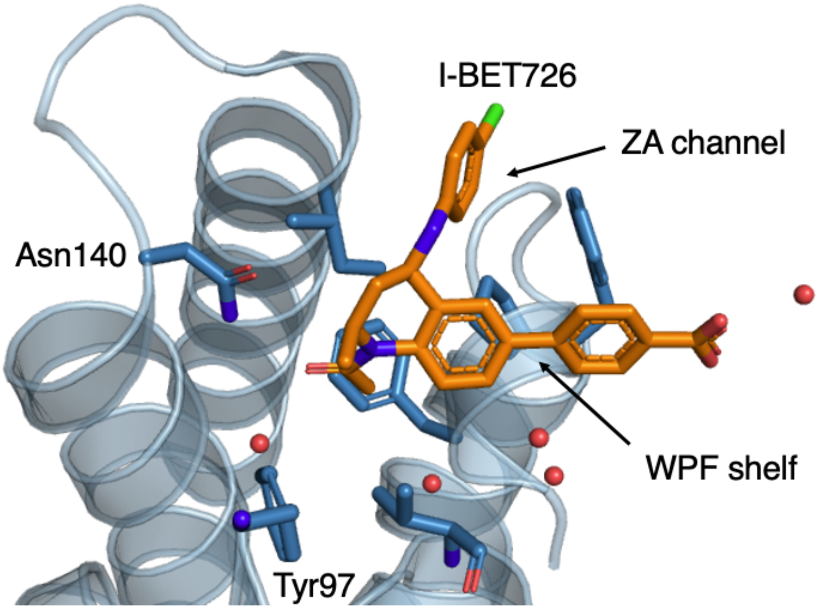
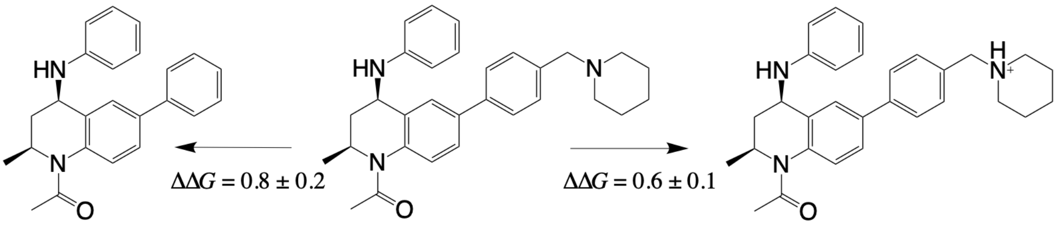


Figure 6.7: Binding site of BRD4-BD1 with inhibitor I-BET726 bound (PDB 4BJX[31]). I-BET726 is compound **15** in the compound series of interest in this work. I-BET726 is represented as stick in orange, the protein is shown as blue cartoon and sticks and water molecules are shown as red spheres.

the PME treatment of long-range interactions, which is likely to introduce an error when changing the net charge of the system. Additionally, care must be taken to ensure that enough time is allowed for the rearrangement of solvent molecules around the ligand when there is a change in charge. Cournia et al. advise that changes in charge should be made to the ligand experimentally, with the results forming the basis for a new series of compounds, with a consistent net charge. Despite this, we believe there was value in investigating how MSλD handles changes in the charge of a ligand, with relative FEP as a comparison, especially as there are few examples in the literature.

As described in the methods section, the setup of MSλD calculations was slightly modified for the positively charged piperidine and the negatively charged benzoic acid substituents. A separate MSλD calculation was performed for the positively and negatively charged set of compounds, where



3

10

Figure 6.8: Setup for charge perturbations using MSλD. In this example, the neutral form of compound **10** is used as the reference to calculate the RBFE compared to compound **3** and the protonated form of compound **10**.

the neutral form of the substituent at site 4 was used for each. A phenyl group at site 4 was included in the multiple topology setup so that the RBFE with respect to compound **3** could still be calculated. Figure 6.8 illustrates the changes in binding free energy calculated for the MSλD perturbation of compound **3** to compound **10**. It should be noted that ethyl and propyl substituents at site 2 were also included so that values of RBFE were obtained for compounds **11** and **12** in the same simulation. Using this approach, the average difference from experiment for the charged compounds was 0.4 ± 0.4 kcal mol^{-1}. In comparison, MSλD calculations for the charged substituents without using the neutral reference compound showed an average difference of 0.9 ± 0.3 kcal mol^{-1} from experiment. Therefore, the impressive agreement with experiment shown by our protocol demonstrates that there is a benefit to using a neutral intermediate compound.

Relative FEP predictions for charge perturbations at site 4 also show good agreement with experiment, with an average difference of 0.6 ± 0.3 kcal mol^{-1}. The position of site 4 on the THQ scaffold fills the narrow ZA channel in the binding site of BRD4-BD1 and points towards the solvent exposed region (Figure 6.7). It appears that both MSλD and relative FEP methods accurately predict RBFEs that involve a charge perturbation at this region of the binding pocket.

6.3.4 Computational expense

To estimate the computational expense of relative FEP and MSλD calculations applied to this compound series, the simulation time required for each method is calculated. Over four MSλD calculations, 119 ns of ALF and 210 ns of data collection is required. This means the full set of RBFE predictions using MSλD can be calculated with 329 ns of simulation time. This is for predictions where the neutral substituents at site 4 have been split into two calculations, with five replicas performed for each. In contrast, 240 ns of simulation time is required for the RBFE prediction of one pairwise set of compounds using relative FEP, totalling 3360 ns of simulation time for the full set of 14 predictions. Therefore, the MSλD calculations require less simulation time by a factor of ~10, compared to relative FEP, when considering these 14 compounds. However, as MSλD calculates RBFE for all combinations of substituents at each site, there is a simulation time saving of a factor of 18, when considering the total molecule space explored. Taking this into account, MSλD provided values for an additional 14 compounds, beyond the 14 presented in Table 6.2. The compound predicted as having the best binding affinity, relative to compound **3** out the compounds with experimental data was compound **15**. This matches experiment, with it having the highest pIC_{50}.[330] From the additional perturbations that MSλD provided, we found that a methyl to ethyl perturbation on site two of compound 15 results in an equivalent binding affinity. It is also possible for further substituents to be considered at each site with limited additional cost, which would substantially extend the number of compounds evaluated overall.

A nontrivial aspect of relative free energy calculations is the manual time it takes to setup a simulation. These setups are often complicated and prone to human error and although tools for their automation are being developed, most are in their early stages or are limited to specific simulation programs.[338–340] Therefore, even with advancements in computational resources and GPU acceleration,[74] the "human time" required for these calculations often becomes a limitation for the rapid estimation of RBFE for large compound data sets, especially in an academic setting. We have found that for an experienced user and once the initial input scripts have been written, the setup of one MSλD calculation is comparable to the setup of one relative FEP calculation. The difference occurs when considering that one MSλD calculation can provide a large number of binding affinity predictions, whereas a separate simulation is required for every pairwise set of compounds when using relative FEP. Therefore, MSλD shows potential for the high-throughput prediction of accurate binding affinities.

6.4 Conclusions

In this chapter, we have presented an investigation into the applicability of MSλD and relative FEP calculations to a series of inhibitors of BRD4-BD1, a prominent therapeutic target. First, benchmarking of relative FEP protocols was performed. Varying numbers of λ windows, equilibration and data collection periods were used, with the accuracy, uncertainty and convergence tested for each combination. We found that using 20 λ windows with 0.5 ns of equilibration and 1 ns of data collection was optimal and presented a good compromise between accuracy and efficiency. When applied to the full set of 14 compounds, relative FEP resulted in RBFE predictions with an average accuracy of 0.6 ± 0.6 kcal mol^{-1}, when discounting one outlier.

The THQ scaffold has four sites of perturbation, with two substituents at site 1, three at site 2, three at site 3 and nine at site 4. Two of the substituents at site 4 have a charge under physiological conditions and were investigated using separate simulations. To test how well MSλD handles the remaining combinations, all $2 \times 3 \times 3 \times 7$ perturbations were considered simultaneously within a single calculation. This resulted in an average accuracy of 1.4 ± 1.4 kcal mol^{-1} and limited correlation between the computed and experimental rank order ($r_s = 0.30$). MSλD achieved more accurate results when splitting the neutral set of substituents into two independent simulations, with an average accuracy of 0.6 ± 0.7 kcal mol^{-1} for the 14 compounds with experimental values available.

MSλD and relative FEP simulations achieved comparable levels of accuracy for this dataset. However, the difference lies in the computational cost of the methods. Comparing the amount of simulation time required for each, MSλD required a factor of ~10 less than relative FEP simulations when considering only those compounds with known free energies, but is a factor of ~18 quicker when the entire molecule space is considered. Furthermore, a much larger number of compounds can be evaluated using a single MSλD calculation, compared to relative FEP, which also saves on manual setup time. As one of the critical limitations of relative FEP is its computational cost, MSλD is a promising alternative for the accurate prediction of ligand binding affinity. The next step in our wider BRD4 study is to apply MSλD to a novel set of compounds, for prospective predictions, the guidance of synthetic decisions and further validation of the method.

Chapter 7

Concluding Remarks

The insights into protein-ligand interactions that computational approaches provide are crucial to the early stages of modern drug discovery and design processes. Molecular docking and MD simulations give understanding of protein structure, active sites and the important interactions to target when designing potential drugs. Furthermore, accurate estimations of binding affinity obviate the need to make every compound in a series. Therefore, investigating potential inhibitors *in silico* expedites the drug discovery pipeline, as computational methods are generally quicker than unfamiliar synthetic routes. There is also the benefit of saving synthetic resources and costs, which is important in the current 2021 climate where sustainable chemistry is paramount. In this thesis the application of CADD to two biological systems of therapeutic interest was explored, with a focus on free energy calculations for the prediction of protein-ligand binding free energies.

The accuracy of MM methods ultimately rely on the quality of the force field in which the potential energy is calculated from. Non-bonded and bonded interactions require parameterisation for quantities such as equilibrium bond lengths, force constants and atomic charges. These parameters are well developed for proteins, mostly from fitting to QM properties of amino acids, with widely used biological force fields including CHARMM,[119] AMBER[118] and OPLS.[121] However, small organic compounds are generally more poorly represented. This can be an issue for modelling protein-ligand complexes, especially when the desired accuracy of binding free energy predictions to guide synthetic decisions is generally assumed to be 0.6-1.0 kcal mol^{-1}.[341] In Chapter 3, we recognise the importance of reliable force field parameters and develop CHARMM force field compatible quantities for a small molecule inhibitor of the protein αvβ6.[27] These parameters are then used for MD and RBFE predictions in Chapter 4. We expect the parameters developed for the 1,8-naphthyridine moiety to aid future computational studies, beyond our own, as naphthyridine is a well utilised group in medicinal chemistry.[233]

Research into more efficient ways to develop parameters and automate these processes is ongoing by multiple groups in the MM community.[341–344] Although, *ad hoc* parameter optimisation is important for reliable results, it can often be too laborious for large compound libraries. One example of work being carried out is the Open Force Field Initiative.[342] This is a collaboration between a network of academic and industry researchers to improve techniques for the parameterisation of small molecule and biomolecular force fields. This initiative recognises that a limitation of current optimisation regimes is the atom-typing of atoms based on their local chemical environments, which adds a level of manual complexity to the process. Instead, the SMIRKS Native Open Force Field (SMIRNOFF) format is used, which assigns parameters based on the full chemical environment of an atom through direct chemical perception.[345] Machine learning is also a promising tool for improving the speed, accuracy and transferability of parameter development.[341,343,346] For example, the program *Parameterize*[343] is an automated force field parameterisation method that uses density functional theory (DFT) and neural network potentials. This method produces small molecule parameters quickly and more accurately than the small molecule force field for AMBER (GAFF2).[122]

In Chapter 4, MD and FEP simulations were utilised for the investigation of a series of small molecules for the inhibition of αvβ6, a protein linked to the initiation and progression of the chronic lung disease IPF.[27] MD simulations highlighted the importance of targeting a set of key binding site interactions, as these were maintained throughout the simulations, regardless of the substituents attached to the core scaffold. More specifically, the MD simulations confirm that a bidentate hydrogen bonding interaction with (αv)-Asp218 and metal chelate interaction with a Mg^{2+} ion in the binding site are critical interactions for inhibitors of αvβ6. FEP simulations provided estimates of ligand binding affinity with an average accuracy of 1.5 kcal mol^{-1}, when compared to experiment. When considering the narrow range of activity shown by this compound series and the probable experimental error, this level of accuracy is sufficient to demonstrate that this integrin system, along with this series of ligands, is amenable to FEP. Substitution on the scaffold involved the addition of a group onto an aryl ring, which significantly increases the complexity and length of the synthesis of this compound. Therefore, accurate predictions of binding affinity are valuable for providing guidance on which compounds should be prioritised for synthesis.

Proteins are highly dynamic and possess inherent flexibility so that they can adapt to form interactions and achieve their function. As X-ray crystal structures represent only one static conformation of a protein, the choice of crystal structure as a starting point for computational modelling can be important. For example, ligand conformation predictions in molecular docking can be influenced by different side chain arrangements of an active site. Furthermore, free energy calculations are even more sensitive to the initial structure of a protein, compared to docking, as the whole protein structure contributes to the results, not just the binding site residues.[196] Therefore, equilibration of any structure is important to ensure that it is close to the correct energy minimum before starting calculations of free energy. To expedite this process, it is logical, if presented with a choice, to choose an initial crystal structure that is close to an ideal conformation.

BRD4 is a target of therapeutic interest and at the time of our study, there were over 300 X-ray crystal structures of BRD4-BD1 publicly available. This presented a challenge for the selection of an initial structure to serve as the basis for computational modelling. In Chapter 5, an analysis of these crystal structures was performed. Structural alignment of BRD4-BD1 complexes showed a high level of similarity between the structures, regardless of the bound ligand. We employed WONKA,[310,311] a tool for detailed analyses of protein binding sites, to compare the active site of over 100 of the crystal structures. The positions of key binding site residues show a high level of conformational similarity, with the exception of Trp81. A focused analysis on the highly conserved water network in the binding site of BRD4-BD1 is performed to identify the positions of these water molecules across the crystal structures. The importance of the water network was illustrated using molecular docking and absolute FEP simulations. 82% of the ligand poses were better predicted when including water molecules as part of the receptor. Our analysis provides guidance for the design of new BRD4-BD1 inhibitors and the selection of the best structure of BRD4-BD1 to use in SBDD, which is important for faster and more cost-efficient lead discovery.

Accurate predictions of binding affinity are valuable in guiding lead design and optimisation. Molecular docking has the functionality to score compounds, based on binding affinity. However, predictions are often compromised by the simplicity of the scoring functions and improvement in their accuracy remains a challenge. Alternatively, alchemical free energy calculations present methods for more rigorous and accurate estimations of binding affinity, although the complexity and computational demand of these methods means they present their own challenges. In Chapter 6, relative FEP and MSλD simulations are employed to calculate the RBFE for a series of BRD4-BD1 inhibitors, based on a THQ scaffold. The two methods achieved comparable levels of accuracy, with an average difference to experiment of 0.6 ± 0.7 kcal mol^{-1} for the relative FEP calculations and 0.6 ± 0.6 kcal mol^{-1} for the MSλD calculations. However, the computational cost of the MSλD calculations was significantly lower, with a factor of ~10 less simulation time required. This study demonstrates the value of MSλD and its potential for the fast and accurate prediction of ligand binding affinity.

Bibliography

(1) F. R. Lichtenberg, *Int Health*, 2019, **11**, 403–416.

(2) M. A. Mohammed, R. J. Moles and T. F. Chen, *Ann. Pharmacother.*, 2016, **50**, 862–881.

(3) S. Talebian, G. G. Wallace, A. Schroeder, F. Stellacci and J. Conde, *Nat. Nanotechnol.*, 2020, **15**, 618–621.

(4) *Medical News Today*, www.medicalnewstoday.com/articles/322409, 2021.

(5) P. Horby, W. Shen Lim, J. R. Emberson, M. Mafham and J. L. Bell, *N. Engl. J. Med.*, 2021, **384**, 693–704.

(6) D. Brown, M. M. Flocco, X. Barril, R. Soliva, B. Davis, J. Hubbard, A. Leach, M. Hann, J. N. Burrows, E. Griffen, J. Liebschutz, S. D. Jones, M. R. Wiley and M. S. Young, *Structure-based drug discovery : an overview*, RSC Publishing, Cambridge, 2006.

(7) R. J. Hatley, S. J. F. Macdonald, R. J. Slack, J. Le, S. B. Ludbrook and P. T. Lukey, *Angew. Chem., Int. Ed.*, 2018, **57**, 3298–3321.

(8) J. Adams, E. C. Anderson, E. E. Blackham, Y. W. R. Chiu, T. Clarke, N. Eccles, L. A. Gill, J. J. Haye, H. T. Haywood, C. R. Hoenig, M. Kausas, J. Le, H. L. Russell, C. Smedley, W. J. Tipping, T. Tongue, C. C. Wood, J. Yeung, J. E. Rowedder, M. J. Fray, T. McInally and S. J. F. Macdonald, *ACS Med. Chem. Lett.*, 2014, **5**, 1207–1212.

(9) C. P. Tinworth and R. J. Young, *J. Med. Chem.*, 2020, **63**, 10091–10108.

(10) C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Delivery Rev.*, 2001, **46**, 3–26.

(11) S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg and A. L. Schacht, *Nat. Rev. Drug Discovery*, 2010, **9**, 203–214.

(12) D. G. Brown and H. J. Wobst, *J. Med. Chem.*, 2021, in press.

(13) *The Drug Development Process*, https://www.fda.gov/patients/learn-about-drug-and-device-approvals/drug-development-process, 2018.

(14) I. M. Kapetanovic, *Chem Biol Interact.*, 2008, **171**, 165–176.

(15) I. Kola and J. Landis, *Nat. Rev. Drug Discovery*, 2004, **3**, 711–716.

(16) S. Mandal, M. Moudgil and S. K. Mandal, *Eur. J. Pharmacol.*, 2009, **625**, 90–100.

(17) U. Gore, Mohini and Jagtap, *Computational Drug Discovery and Design*, ed. R. Baron, Springer New York, New York, 2012, p. 628.

(18) H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. D. Westbrook and C. Zardecki, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2002, **58**, 899–907.

(19) X. Dong, N. E. Hudson, C. Lu and T. A. Springer, *Nat. Struct. Mol. Biol.*, 2014, **21**, 1091–1096.

(20) O. Mirguet, Y. Lamotte, C. W. Chung, P. Bamborough, D. Delannée, A. Bouillot, F. Gellibert, G. Krysa, A. Lewis, J. Witherington, P. Huet, Y. Dudit, L. Trottet and E. Nicodeme, *ChemMedChem*, 2014, **9**, 580–589.

(21) M. M. Bluhm, G. Bodo, H. M. Dintzis and J. C. Kendrew, *Proc. R. Soc. A*, 1958, **246**, 369–389.

(22) M. F. Perutz and L. Mazzarella, *Nature*, 1947, **159**, 671.

(23) C. Blake, D. Koenig, G. Mair, A. North, D. Phillips and V. Sarma, *Nature*, 1965, **206**, 757–761.

(24) D. Matthews, R. Alden, J. Bolin, S. Freer, R. Hamlin, N. Xuong, J. Kraut, M. Poe, M. Williams and K. Hoogsteen, *Science*, 1977, **197**, 452–455.

(25) L. F. Kuyper, B. Roth, D. P. Baccanari, R. Ferone, C. R. Beddell, J. N. Champness, D. K. Stammers, J. G. Dann and F. E. A. Norrington, *J. Med. Chem.*, 1985, **28**, 303–311.

(26) K. A. Armacost, S. Riniker and Z. Cournia, *J. Chem. Inf. Model.*, 2020, **60**, 5283–5286.

(27) E. E. Guest, S. A. Oatley, S. J. F. Macdonald and J. D. Hirst, *J. Chem. Inf. Model.*, 2020, **60**, 5487–5498.

(28) C. John Harris, R. D. Hill, D. W. Sheppard, M. J. Slater and P. F.W. Stouten, *Comb. Chem. High Throughput Screening*, 2011, **14**, 521–531.

(29) M. E. Welsch, S. A. Snyder and B. R. Stockwell, *Curr. Opin. Chem. Biol.*, 2010, **14**, 347–361.

(30) D. B. Kitchen, H. Decornez, J. R. Furr and J. Bajorath, *Nat. Rev. Drug Discovery*, 2004, **3**, 935–949.

(31) A. Wyce, G. Ganji, K. N. Smitheman, C. Chung, S. Korenchuk, Y. Bai, O. Barbash, B. C. Le, P. D. Craggs, M. T. McCabe, K. M. Kennedy-Wilson, L. V. Sanchez, R. L. Gosmini, N. Parr, C. F. McHugh, D. Dhanak, R. K. Prinjha, K. R. Auger and P. J. Tummino, *PLoS ONE*, 2013, **8**, 1–16.

(32) D. G. Alberg and S. L. Schreiber, *Science*, 1993, **262**, 248–250.

(33) N. S. Pagadala, K. Syed and J. Tuszynski, *Biophys. Rev.*, 2017, **9**, 91–102.

(34) D. E. Koshland, *Science*, 1963, **142**, 1533–1542.

(35) G. G. Hammes, *Biochemistry*, 2002, **41**, 8221–8228.

(36) M. Rarey, B. Kramer, T. Lengauer and G. Klebe, *J. Mol. Biol.*, 1996, **261**, 470–489.

(37) M. McGann, *J. Chem. Inf. Model.*, 2011, **51**, 578–596.

(38) W. J. Allen, T. E. Balius, S. Mukherjee, S. R. Brozell, D. T. Moustakas, P. T. Lang, D. A. Case, I. D. Kuntz and R. C. Rizzo, *J. Comput. Chem.*, 2015, **36**, 1132–1156.

(39) R. Dayam and N. Neamati, *Bioorg. Med. Chem.*, 2004, **12**, 6371–6381.

(40) G. Jones, P. Willett, R. C. Glen, A. R. Leach and R. Taylor, *J. Mol. Biol.*, 1997, **267**, 727–748.

(41) T. A. Halgren, R. B. Murphy, R. A. Friesner, H. S. Beard, L. L. Frye, W. T. Pollard and J. L. Banks, *J. Med. Chem.*, 2004, **47**, 1750–1759.

(42) O. Trott and A. J. Olson, *J. Comput. Chem.*, 2009, **31**, 455–461.

(43) J. Wang and N. V. Dokholyan, *J. Chem. Inf. Model.*, 2019, **59**, 2509–2515.

(44) M. A. Phillips, M. A. Stewart, D. L. Woodling and Z. Xie, in *Molecular Docking*, ed. D. Vlachakis, IntechOpen, 1st edn., 2018, vol. 32, ch. 8, pp. 137–144.

(45) L. Pinzi and G. Rastelli, *Int. J. Mol. Sci.*, 2019, **20**, 1–23.

(46) M. Kumari, S. Chandra, N. Tiwari and N. Subbarao, *BMC Struct. Biol.*, 2016, **16**, 1–11.

(47) J. Fan, A. Fu and L. Zhang, *Quant. Biol.*, 2019, **7**, 83–89.

(48) M. Fan, J. Wang, H. Jiang, Y. Feng, M. Mahdavi, K. Madduri, M. T. Kandemir and N. V. Dokholyan, *J. Phys. Chem. B*, 2021, **125**, 1049–1060.

(49) E. Yuriev and P. A. Ramsland, *J. Mol. Recognit.*, 2013, **26**, 215–239.

(50) T. M. Menchaca, C. Juarez-Portilla and R. C. Zepeda, in *Drug Discovery and Development - New Advances*, ed. V. Gaitonde, IntechOpen, 2013, vol. 32, ch. 2, pp. 137–144.

(51) B. Sadjad and Z. Zsoldos, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 2011, **8**, 1120–1133.

(52) P. J. Ballester and J. B. O. Mitchell, *Bioinformatics*, 2012, **26**, 1169–1175.

(53) C. Bissantz, B. Kuhn and M. Stahl, *J. Med. Chem.*, 2010, **53**, 5061–5084.

(54) G. Klebe, *Drug Discovery Today*, 2006, **11**, 580–594.

(55) Q. Lu, L. W. Qi and J. Liu, *J. Theor. Comput. Chem.*, 2019, **18**, 1–12.

(56) C. N. Cavasotto and J. I. Di Filippo, *Arch. Biochem. Biophys.*, 2021, **698**, 108730.

(57) S. L. Kinnings, N. Liu, P. J. Tonge, R. M. Jackson, L. Xie and P. E. Bourne, *J. Chem. Inf. Model.*, 2011, **51**, 1195–1197.

(58) F. Gentile, V. Agrawal, M. Hsing, A. T. Ton, F. Ban, U. Norinder, M. E. Gleave and A. Cherkasov, *ACS Cent. Sci.*, 2020, **6**, 939–949.

(59) S. Pushpakom, F. Iorio, P. A. Eyers, K. J. Escott, S. Hopper, A. Wells, A. Doig, T. Guilliams, J. Latimer, C. McNamee, A. Norris, P. Sanseau, D. Cavalla and M. Pirmohamed, *Nat. Rev. Drug Discovery*, 2018, **18**, 41–58.

(60) R. R. Deshpande, A. P. Tiwari, N. Nyayanit and M. Modak, *Eur. J. Pharmacol.*, 2020, **886**, 173430.

(61) J. Wang, *J. Chem. Inf. Model.*, 2020, **60**, 3277–3286.

(62) A. D. Elmezayen, A. Al-Obaidi, A. T. Şahin and K. Yelekçi, *J. Biomol. Struct. Dyn.*, 2021, **39**, 1–13.

(63) M. Hakmi, E. M. Bouricha, I. Kandoussi, J. E. Harti and A. Ibrahimi, *Bioinformation*, 2020, **16**, 301.

(64) O. M. H. Salo-ahen, I. Alanko, R. Bhadane, A. M. J. J. Bonvin, R. V. Honorato, S. Hossain, A. H. Juffer, A. Kabedev, M. Lahtela-kakkonen, A. S. Larsen, E. Lescrinier and P. Marimuthu, *Processes*, 2021, **9**, 1–60.

(65) R. A. Copeland, *Nat. Rev. Drug Discovery*, 2016, **15**, 87–95.

(66) D. A. Schuetz, W. E. A. de Witte, Y. C. Wong, B. Knasmueller, L. Richter, D. B. Kokh, S. K. Sadiq, R. Bosma, I. Nederpelt, L. H. Heitman, E. Segala, M. Amaral, D. Guo, D. Andres, V. Georgi, L. A. Stoddart, S. Hill, R. M. Cooke, C. De Graaf, R. Leurs, M. Frech, R. C. Wade, E. C. M. de Lange, A. P. IJzerman, A. Müller-Fahrnow and G. F. Ecker, *Drug Discovery Today*, 2017, **22**, 896–911.

(67) M. Karplus, J. A. McCammon and B. R. Gelin, *Nature*, 1977, **267**, 585–590.

(68) M. Karplus, *Biopolymers*, 2003, **68**, 350–358.

(69) B. R. Brooks, C. L. Brooks, A. D. Mackerell Jr., L Nilsson, R. J. Petrella, B Roux, Y Won, G Archontis, C Bartels, S Boresch, A Caflisch, L Caves, Q Cui, A. R. Dinner, M Feig, S Fischer, J Gao, M Hodoscek, W Im, K Kuczera, T Lazaridis, J Ma, V Ovchinnikov, E Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M Schaefer, B Tidor, R. M. Venable, H. L. Woodcock, X Wu, W Yang, D. M. York and M Karplus, *J. Comput. Chem.*, 2009, **30**, 1545–1614.

(70) M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindahl, *SoftwareX*, 2015, **1-2**, 19–25.

(71) D. A. Case, T. E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K. M. Merz Jr., A. V. Onufriev, C. Simmerling, B. Wang and R. J. Woods, *J. Comput. Chem.*, 2013, **236**, 47–56.

(72) J. C. Phillips, K. Schulten, A. Bhatele, C. Mei, Y. Sun, E. J. Bohm and L. V. Kale, *J. Comput. Chem.*, 2016, **26**, 60–76.

(73) N. Kondratyuk, V. Nikolskiy, D. Pavlov and V. Stegailov, *Int. J. High Perform. Comput. Appl.*, 2021, 1–13.

(74) J. C. Phillips, D. J. Hardy, J. D. Maia, J. E. Stone, J. V. Ribeiro, R. C. Bernardi, R. Buch, G. Fiorin, J. Hénin, W. Jiang, R. McGreevy, M. C. Melo, B. K. Radak, R. D. Skeel, A. Singharoy, Y. Wang, B. Roux, A. Aksimentiev, Z. Luthey-Schulten, L. V. Kalé, K. Schulten, C. Chipot and E. Tajkhorshid, *J. Chem. Phys.*, 2020, **153**, 1–33.

(75) T. S. Lee, D. S. Cerutti, D. Mermelstein, C. Lin, S. Legrand, T. J. Giese, A. Roitberg, D. A. Case, R. C. Walker and D. M. York, *J. Chem. Inf. Model.*, 2018, **58**, 2043–2050.

(76) J. Jung, W. Nishima, M. Daniels, G. Bascom, C. Kobayashi, A. Adedoyin, M. Wall, A. Lappala, D. Phillips, W. Fischer, C.-S. Tung, T. Schlick, Y. Sugita and K. Y. Sanbonmatsu, *J. Comput. Chem.*, 2019, **40**, 1919–1930.

(77) C. Zhang, L. J. Feng, Y. Huang, D. Wu, Z. Li, Q. Zhou, Y. Wu and H. B. Luo, *J. Chem. Inf. Model.*, 2017, **57**, 355–364.

(78) J. D. Durrant, S. E. Kochanek, L. Casalino, P. U. Ieong, A. C. Dommer and R. E. Amaro, *ACS Cent. Sci.*, 2020, **6**, 189–196.

(79) Y. Sugita and Y. Okamoto, *Chem. Phys. Lett.*, 1999, **314**, 141–151.

(80) A. Laio and M. Parrinello, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 12562–12566.

(81) C. Tsallis and D. A. Stariolo, *Phys. A (Amsterdam, Neth.)*, 1996, **233**, 395–406.

(82) R. C. Bernardi, M. C. Melo and K. Schulten, *Biochim. Biophys. Acta, Gen. Subj.*, 2015, **1850**, 872–877.

(83) J. Romanowska, D. B. Kokh, J. C. Fuller and R. C. Wade, in *Thermodynamics and Kinetics of Drug Binding*, John Wiley Sons, Ltd, 2015, ch. 11, pp. 211–235.

(84) G. Vauquelin, *MedChemComm*, 2018, **9**, 1426–1438.

(85) D. Gobbo, V. Piretti, R. M. C. Di Martino, S. K. Tripathi, B. Giabbai, P. Storici, N. Demitri, S. Girotto, S. Decherchi and A. Cavalli, *J. Chem. Theory Comput.*, 2019, **15**, 4646–4659.

(86) H. M. Senn and W. Thiel, *Angew. Chem., Int. Ed.*, 2009, **48**, 1198–1229.

(87) A. Romero-Rivera, M. Garcia-Borràs and S. Osuna, *Chem. Commun.*, 2017, **53**, 284–297.

(88) B. Saïd-Salim, B. Mathema and B. N. Kreiswirth, *Infect. Control Hosp. Epidemiol.*, 2003, **24**, 451–455.

(89) V. Dalal, P. Dhankhar, V. Singh, V. Singh, G. Rakhaminov, D. Golemi-Kotra and P. Kumar, *Protein J.*, 2021, **40**, 148–165.

(90) A. De Ruiter and C. Oostenbrink, *Curr. Opin. Chem. Biol.*, 2011, **15**, 547–552.

(91) J. Srinivasan, J. Miller, P. A. Kollman and D. A. Case, *J. Biomol. Struct. Dyn.*, 1998, **16**, 671–682.

(92) P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case and T. E. Cheatham, *Acc. Chem. Res.*, 2000, **33**, 889–897.

(93) E. Wang, H. Sun, J. Wang, Z. Wang, H. Liu, J. Z. Zhang and T. Hou, *Chem. Rev.*, 2019, **119**, 9478–9508.

(94) J. Srinivasan, T. E. Cheatham, P. Cieplak, P. A. Kollman and D. A. Case, *J. Am. Chem. Soc.*, 1998, **120**, 9401–9409.

(95) M. Arba, S. Ihsan, L. O. A. N. Ramadhan and D. H. Tjahjono, *Comput. Biol. Chem.*, 2017, **67**, 9–14.

(96) R. W. Zwanzig, *J. Chem. Phys.*, 1954, **22**, 1420–1426.

(97) P. A. Bash, U. C. Singh, F. K. Brown, R. Langridge and P. A. Kollman, *Science*, 1987, **235**, 574–576.

(98) C. H. Bennett, *J. Comput. Phys.*, 1976, **22**, 245–268.

(99) J. G. Kirkwood, *J. Chem. Phys.*, 1935, **3**, 300–313.

(100) W. L. Jorgensen and C. Ravimohan, *J. Chem. Phys.*, 1985, **83**, 3050–3054.

(101) L. F. Song and K. M. Merz, *J. Chem. Inf. Model.*, 2020, **60**, 5308–5318.

(102) M. Aldeghi, J. P. Bluck and P. C. Biggin, in *Computational Drug Discovery and Design*, 2018, pp. 199–232.

(103) Z. Cournia, B. Allen and W. Sherman, *J. Chem. Inf. Model.*, 2017, **57**, 2911–2937.

(104) F. Deflorian, L. Perez-Benito, E. B. Lenselink, M. Congreve, H. W. Van Vlijmen, J. S. Mason, C. De Graaf and G. Tresadern, *J. Chem. Inf. Model.*, 2020, **60**, 5563–5579.

(105) R. Abel, S. Mondal, C. Masse, J. Greenwood, G. Harriman, M. A. Ashwell, S. Bhat, R. Wester, L. Frye, R. Kapeller and R. A. Friesner, *Curr. Opin. Struct. Biol.*, 2017, **43**, 38–44.

(106) J. L. Knight and C. L. Brooks, *J. Comput. Chem.*, 2009, **30**, 1692–1700.

(107) X. Kong and C. L. Brooks, *J. Chem. Phys.*, 1996, **105**, 2414–2423.

(108) R. L. Hayes, K. A. Armacost, J. Z. Vilseck and C. L. Brooks, *J. Phys. Chem. B*, 2017, **121**, 3626–3635.

(109) J. Z. Vilseck, N. Sohail, R. L. Hayes and C. L. Brooks, *J. Phys. Chem. Lett.*, 2019, **10**, 4875–4880.

(110) M. R. Shirts and D. L. Mobley, *Methods Mol. Biol.*, 2013, **924**, 271–311.

(111) T. McInally and S. J. Macdonald, *J. Med. Chem.*, 2017, **60**, 7958–7964.

(112) D. Oglic, S. A. Oatley, S. J. F. Macdonald, T. McInally, R. Garnett, J. D. Hirst and T. Gärtner, *Mol. Inf.*, 2018, **37**, 1–15.

(113) E. E. Guest, S. D. Pickett and J. D. Hirst, *Org. Biomol. Chem.*, 2021, **19**, 5632–5641.

(114) F. Neese, M. Atanasov, G. Bistoni, D. Maganas and S. Ye, *J. Am. Chem. Soc.*, 2019, **7**, 2814–2824.

(115) J. M. Combes, P Duclos and R Seiler, in *Rigorous Atomic and Molecular Physics*, ed. G Velo and A. S. Wightman, Springer US, Boston, MA, 1981, pp. 185–213.

(116) J. Lennard-Jones, *Proc. Royal Society London A*, 1924, **4**, 463–477.

(117) J. A. Rackers and J. W. Ponder, *J. Chem. Phys.*, 2019, **150**, 1–22.

(118) V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, *Proteins: Struct., Funct., Bioinf.*, 2006, **65**, 712–725.

(119) R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig and A. D. MacKerell, *J. Chem. Theory Comput.*, 2012, **8**, 3257–3273.

(120) C. Oostenbrink, A. Villa, A. E. Mark and W. F. Van Gunsteren, *J. Comput. Chem.*, 2004, **25**, 1656–1676.

(121) G. A. Kaminski, R. A. Friesner, J. Tirado-Rives and W. L. Jorgensen, *J. Phys. Chem. B*, 2001, **105**, 6474–6487.

(122) J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *J. Comput. Chem.*, 2004, **25**, 1157–1174.

(123) K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov and A. D. MacKerell, *J. Comput. Chem.*, 2010, **31**, 671–690.

(124) T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 490–519.

(125) T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 520–552.

(126) T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 553–586.

(127) T. A. Halgren and R. B. Nachbar, *J. Comput. Chem.*, 1996, **615**, 587–615.

(128) T. A. Halgren, *J. Comput. Chem.*, 1996, **641**, 616–641.

(129) E. Harder, W. Damm, J. Maple, C. Wu, M. Reboul, J. Y. Xiang, L. Wang, D. Lupyan, M. K. Dahlgren, J. L. Knight, J. W. Kaus, D. S. Cerutti, G. Krilov, W. L. Jorgensen, R. Abel and R. A. Friesner, *J. Chem. Theory Comput.*, 2016, **12**, 281–296.

(130) X. Daura, A. E. Mark and W. F. Van Gunsteren, *J. Comput. Chem.*, 1998, **19**, 535–547.

(131) L. D. Schuler, X. Daura and W. F. van Gunsteren, *J. Comput. Chem.*, 2001, **22**, 1205–1218.

(132) B. A. Horta, P. T. Merz, P. F. Fuchs, J. Dolenc, S. Riniker and P. H. Hünenberger, *J. Chem. Theory Comput.*, 2016, **12**, 3825–3850.

(133) P. J. Goodford, *J. Med. Chem.*, 1985, **28**, 849–857.

(134) D. G. Levitt and L. J. Banaszak, *J. Mol. Graphics*, 1992, **10**, 229–234.

(135) R. A. Laskowski, *J. Mol. Graphics*, 1995, **13**, 323–330.

(136) G. P. Brady and P. F. Stouten, *J. Comput.-Aided Mol. Des.*, 2000, **14**, 383–401.

(137) M. Mezei, *J. Mol. Graphics Modell.*, 2003, **21**, 463–472.

(138) L. G. Ferreira, R. N. Dos Santos, G. Oliva and A. D. Andricopulo, *Molecules*, 2015, **20**, 13384–13421.

(139) G. M. Morris and M. Lim-Wilby, in *Molecular Modeling of Proteins*, 2008, vol. 443, pp. 365–382.

(140) Z. Zsoldos, D. Reid, A. Simon, S. B. Sadjad and A. P. Johnson, *J. Mol. Graphics Modell.*, 2007, **26**, 198–212.

(141) A. T. Brint and P. Willett, *J. Chem. Inf. Comput. Sci*, 1987, **27**, 152–158.

(142) D. Fischer, R. Norel and H. Wolfson, *Proteins: Struct., Funct., Genet.*, 1993, 16278–292.

(143) R. Norel, D. Fischer, H. J. Wolfson and R. Nussinov, *Protein Eng.*, 1994, **7**, 39–46.

(144) M. Rarey, B. Kramer, T. Lengauer and G. Klebe, *J. Mol. Biol*, 1996, **261**, 470–489.

(145) R. L. DesJarlais, R. P. Sheridan, S. J. Dixon, I. D. Kuntz and R. Venkataraghavan, *J. Med. Chem.*, 1986, **29**, 2149–2153.

(146) A. R. Leach and I. D. Kuntz, *J. Comput. Chem.*, 1992, **13**, 730–748.

(147) D. S. Goodsell, H. Lauble, C. D. Stout and A. J. Olson, *Proteins: Struct., Funct., Bioinf.*, 1993, **17**, 1–10.

(148) T. N. Hart and R. J. Read, *Proteins: Struct., Funct., Bioinf.*, 1992, **13**, 206–222.

(149) G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew and A. J. Olson, *J. Comput. Chem.*, 1998, **19**, 1639–1662.

(150) X. Y. Meng, H. X. Zhang, M. Mezei and M. Cui, *Curr. Comput.-Aided Drug Des.*, 2011, **7**, 146–157.

(151) P. C. Hawkins, A. G. Skillman, G. L. Warren, B. A. Ellingson and M. T. Stahl, *J. Chem. Inf. Model.*, 2010, **50**, 572–584.

(152) S. Y. Huang, S. Z. Grinter and X. Zou, *Phys. Chem. Chem. Phys.*, 2010, **12**, 12899–12908.

(153) H.-J. Böhm, *J. Comput.-Aided Mol. Des.*, 1994, **8**, 243–256.

(154) C. A. Sotriffer, P. Sanschagrin, H. Matter and G. Klebe, *Proteins: Struct., Funct., Genet.*, 2008, **73**, 395–419.

(155) A. Krammer, P. D. Kirchhoff, X. Jiang, C. M. Venkatachalam and M. Waldman, *J. Mol. Graphics Modell.*, 2005, **23**, 395–407.

(156) M. M. Mysinger and B. K. Shoichet, *J. Chem. Inf. Model.*, 2010, **50**, 1561–1573.

(157) A. M. Ruvinsky, *J. Comput. Chem.*, 2007, **28**, 1364–1372.

(158) R. Wang, Y. Lu and S. Wang, *J. Med. Chem.*, 2003, **46**, 2287–2303.

(159) S. Tietze and J. Apostolakis, *J. Chem. Inf. Model.*, 2007, **47**, 1657–1672.

(160) S. Yin, L. Biedermannova, J. Vondrasek and N. V. Dokholyan, *J. Chem. Inf. Model.*, 2008, **48**, 1656–1662.

(161) A. N. Jain, *J. Med. Chem.*, 2003, **46**, 499–511.

(162) Z. R. Xie and M. J. Hwang, *BMC Bioinformatics*, 2010, **11**, 1–16.

(163) H. Fan, D. Schneidman-Duhovny, J. J. Irwin, G. Dong, B. K. Shoichet and A. Sali, *J. Chem. Inf. Model.*, 2011, **51**, 3078–3092.

(164) H. M. Kumalo, S. Bhakat and M. E. Soliman, *Molecules*, 2015, **20**, 1984–2000.

(165) M. P. Allen and D. J. Tildesley, *Computer simulation of liquids*, Oxford university press, 2017.

(166) R. Hockney and J. Eastwood, *Computer simulation using particles*, 1988.

(167) W. C. Swope, H. C. Andersen, P. H. Berens and K. R. Wilson, *The Journal of Chemical Physics*, 1982, **76**, 637–649.

(168) J.-P. Ryckaert, G. Ciccotti and H. J. C. Berendsen, *J. Comput. Phys.*, 1977, **23**, 327–341.

(169) T. A. Collier, T. J. Piggot and J. R. Allison, in *Protein Nanotechnology: Protocols, Instrumentation and Applications*, ed. J. A. Gerrard and L. J. Domigan, 3rd edn., 2020, vol. 2073, ch. 17, pp. 311–327.

(170) D. A. S. K. Humphrey, W., *VMD-VisualMolecularDynamics*, 1996.

(171) Schrödinger LLC, *The PyMOL Molecular Graphics System, Version 2.1.1*, 2015.

(172) E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng and T. E. Ferrin, *J. Comput. Chem.*, 2004, **25**, 1605–1612.

(173) J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé and K. Schulten, *J. Comput. Chem.*, 2005, **26**, 1781–802.

(174) D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang and R. J. Woods, *J. Comput. Chem.*, 2005, **26**, 1668–1688.

(175) H. J. C. Berendsen, D. van der Spoel and R. van Drunen, *Comput. Phys. Commun.*, 1995, **91**, 43–56.

(176) W. Kauzmann, in *Advances in Protein Chemistry*, ed. C. B. Anfinsen, M. L. Anson, K. Bailey and J. T. Edsall, Academic Press, 1959, vol. 14, pp. 1–63.

(177) W. H. Orttung, *Ann. N. Y. Acad. Sci.*, 1977, **303**, 22–37.

(178) D. Bashford and D. A. Case, *Annu. Rev. Phys. Chem.*, 2000, **51**, 129–152.

(179) S. J. Bachmann and W. F. Van Gunsteren, *J. Chem. Phys.*, 2014, **141**, 22D515.

(180) W. L. Jorgensen, *J. Am. Chem. Soc.*, 1981, **103**, 335–340.

(181) H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren and J Hermans, in *Intermolecular Forces: Proceedings of the Fourteenth Jerusalem Symposium*, ed. B. Pullman, Springer Netherlands, Dordrecht, 1981, pp. 331–342.

(182) W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J. Chem. Phys.*, 1983, **79**, 926–935.

(183) A. Hussain, Ph.D. Thesis, University of Nottingham, 2012.

(184) P. P. Ewald, *Annalen der Physik*, 1921, **369**, 253–287.

(185) T. Darden, D. York and L. Pedersen, *J. Chem. Phys.*, 1993, **98**, 10089–10092.

(186) B. Gautam, in *Homology Molecular Modeling - Perspectives and Applications*, ed. R. T. Maia, IntechOpen, 2021, ch. 7, pp. 265–280.

(187) J. C. Meza, *Wiley Interdiscip. Rev. Comput. Stat.*, 2010, **2**, 719–722.

(188) R. Fletcher and C. M. Reeves, *J. Comput.*, 1964, **7**, 149–154.

(189) T. J. Ypma, *SIAM Rev.*, 1995, **37**, 531–551.

(190) J. W. Carter, A. S. Tascini, J. M. Seddon and F. Bresme, in *Computational Tools for Chemical Biology*, ed. S. Martin-Santamaria, Royal Society of Chemistry, 2017, ch. 2, pp. 39–68.

(191) Z. Lin, J. Zou, S. Liu, C. Peng, Z. Li, X. Wan, D. Fang, J. Yin, G. Gobbo, Y. Chen, J. Ma, S. Wen, P. Zhang and M. Yang, *J. Chem. Inf. Model.*, 2021, **61**, 2720–2732.

(192) R. W. Zwanzig, *J. Chem. Phys.*, 1955, **23**, 1915–1922.

(193) M. Aldeghi, A. Heifetz, M. J. Bodkin, S. Knapp and P. C. Biggin, *Chem. Sci.*, 2016, **7**, 207–218.

(194) S. Boresch, F. Tettinger, M. Leitgeb and M. Karplus, *J. Chem. Phys. B*, 2003, **107**, 9535–9551.

(195) S. E. Boyce, D. L. Mobley, G. J. Rocklin, A. P. Graves, K. A. Dill and B. K. Shoichet, *J. Mol. Biol.*, 2009, **394**, 747–763.

(196) Z. Cournia, B. Allen and W. Sherman, *J. Chem. Inf. Model.*, 2017, **57**, 2911–2937.

(197) J. Gao, K. Kuczera, B. Tidor and M. Karplus, *Science*, 1989, **244**, 1069–1072.

(198) D. A. Pearlman, *J. Phys. Chem.*, 1994, **98**, 1487–1493.

(199) M. R. Shirts and D. L. Mobley, *Methods Mol. Biol.*, 2013, **924**, 271–311.

(200) M. Zacharias, T. P. Straatsma and J. A. McCammon, *J. Chem. Phys.*, 1994, **100**, 9025–9031.

(201) T. C. Beutler, A. E. Mark, R. C. van Schaik, P. R. Gerber and W. F. van Gunsteren, *Chem. Phys. Lett.*, 1994, **222**, 529–539.

(202) M. R. Shirts and V. S. Pande, *J. Chem. Phys.*, 2005, **122**, 144107.

(203) N. Lu, J. K. Singh and D. A. Kofke, *J. Chem. Phys.*, 2003, **118**, 2977–2984.

(204) M. R. Shirts and J. D. Chodera, *J. Chem. Phys.*, 2008, **129**, 124105.

(205) M. R. Shirts, E. Bair, G. Hooker and V. S. Pande, *Phys. Rev. Lett.*, 2003, **91**, 1–4.

(206) P. V. Klimovich, M. R. Shirts and D. L. Mobley, *J. Comput.-Aided Mol. Des.*, 2015, **29**, 397–411.

(207) X. Kong and C. L. Brooks, *J. Chem. Phys.*, 1996, **105**, 2414–2423.

(208) J. L. Knight and C. L. Brooks, *J. Chem. Theory Comput.*, 2011, **7**, 2728–2739.

(209) R. L. Hayes, K. A. Armacost, J. Z. Vilseck and C. L. Brooks, *J. Chem. Phys. B*, 2017, **121**, 3626–3635.

(210) J. L. Knight and C. L. Brooks, *J. Comput. Chem.*, 2011, **32**, 3423–3432.

(211) Z. Guo, C. L. Brooks and X. Kong, *J. Phys. Chem. B*, 1998, **102**, 2032–2036.

(212) M. Suruzhon, M. S. Bodnarchuk, A. Ciancetta, R. Viner, I. D. Wall and J. W. Essex, *J. Chem. Theory Comput.*, 2021, **17**, 1806–1821.

(213) E. King, R. Qi, H. Li, R. Luo and E. Aitchison, *J. Chem. Theory Comput.*, 2021, **17**, 2541–2555.

(214) L. Pérez-Benito, N. Casajuana-Martin, M. Jiménez-Rosés, H. Van Vlijmen and G. Tresadern, *J. Chem. Theory Comput.*, 2019, **15**, 1884–1895.

(215) SC '06: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing, Association for Computing Machinery, New York, NY, USA, 2006.

(216) K. Roos, C. Wu, W. Damm, M. Reboul, J. M. Stevenson, C. Lu, M. K. Dahlgren, S. Mondal, W. Chen, L. Wang, R. Abel, R. A. Friesner and E. D. Harder, *J. Chem. Theory Comput.*, 2019, **15**, 1863–1874.

(217) L. Richeldi, H. R. Collard and M. G. Jones, *Lancet*, 2017, **389**, 1941–1952.

(218) V. Navaratnam, K. M. Fleming, J. West, C. J. P. Smith, R. G. Jenkins, A. Fogarty and R. B. Hubbard, *Thorax*, 2011, **66**, 462–467.

(219) D Forman, F Bray, D. H. Brewster, C. G. Mbalawa, B Kohler, M Piñeros, E Steliarova-Foucher, R Swaminathan and J Ferlay, *Cancer Incidence in Five Continents*, IARC Scientific Publications, 2014, vol. X.

(220) J. Hutchinson, A. Fogarty, R. Hubbard and T. McKeever, *Eur. Respir. J.*, 2015, **46**, 795–806.

(221) R. Du Bois and T. E. King, *Thorax*, 2007, **62**, 1008–1012.

(222) R. J. Panos, R. L. Mortenson, S. A. Niccoli and T. E. King Jr., *Am. J. Med.*, 1990, **88**, 396–404.

(223) A. Tzouvelekis, R. Toonkel and T. Karampitsakos, *Oncotarget*, 2018, **5**, 1–8.

(224) K. R. Flaherty, T. V. Colby, W. D. Travis, G. B. Toews, J. Mumford, S. Murray, V. J. Thannickal, E. A. Kazerooni, B. H. Gross, J. P. Lynch and F. J. Martinez, *Am. J. Respir. Crit. Care Med.*, 2003, **167**, 1410–1415.

(225) A. G. Nicholson, L. G. Fulford, T. V. Colby, R. M. Du Bois, D. M. Hansell and A. U. Wells, *Am. J. Respir. Crit. Care Med.*, 2002, **166**, 173–177.

(226) U. Costabel, R. M. Du Bois and J. J. Egan, *Diffuse Parenchymal Lung Disease*, Karger, 2007, vol. 26, pp. 101–109.

(227) S. Ghatak, V. C. Hascall, R. R. Markwald, C. Feghali-Bostwick, C. M. Artlett, M. Gooz, G. S. Bogatkevich, I. Atanelishvili, R. M. Silver, J. Wood, V. J. Thannickal and S. Misra, *J. Biol. Chem.*, 2017, **292**, 10490–10519.

(228) D. Sheppard, *Proc. Am. Thorac. Soc.*, 2006, **3**, 413–417.

(229) R. Hatley, S. Macdonald, R. Slack, J. Le, S. Ludbrook and P. Lukey, *Angew. Chem., Int. Ed.*, 2017, **57**, 3298–3321.

(230) S. Raab-Westphal, J. F. Marshall and S. L. Goodman, *Cancers*, 2017, **9**, 1–28.

(231) A. Becker, O. Von Richter, A. Kovar, H. Scheible, J. J. Van Lier and A. Johne, *J. Clin. Pharmacol.*, 2015, **55**, 815–824.

(232) O. L. Chinot, *Lancet Oncol.*, 2014, **15**, 1044–1045.

(233) A. Madaan, R. Verma, V. Kumar, A. T. Singh, S. K. Jain and M. Jaggi, *Arch. Pharm. Chem. Life Sci.*, 2015, **348**, 837–860.

(234) H. Robinson, S. A. Oatley, J. E. Rowedder, P. Slade, S. J. Macdonald, S. P. Argent, J. D. Hirst, T. McInally and C. J. Moody, *Chem. - Eur. J.*, 2020, **26**, 7678–7684.

(235) A. D. MacKerell, M. Feig and C. L. Brooks, *J. Am. Chem. Soc.*, 2004, **126**, 698–699.

(236) K. Vanommeslaeghe and A. D. Mackerell Jr., *J. Chem. Inf. Model.*, 2012, **52**, 3144–3154.

(237) K. Vanommeslaeghe, E. P. Raman and A. D. Mackerell Jr., *J. Chem. Inf. Model.*, 2012, **52**, 3155–3168.

(238) Y. Shao, Z. Gan, E. Epifanovsky, A. T. Gilbert, M. Wormit, J. Kussmann, A. W. Lange, A. Behn, J. Deng, X. Feng, D. Ghosh, M. Goldey, P. R. Horn, L. D. Jacobson, I. Kaliman, R. Z. Khaliullin, T. Kus, A. Landau, J. Liu, E. I. Proynov, Y. M. Rhee, R. M. Richard, M. A. Rohrdanz, R. P. Steele, E. J. Sundstrom, H. L. Woodcock, P. M. Zimmerman, D. Zuev, B. Albrecht, E. Alguire, B. Austin, G. J. Beran, Y. A. Bernard, E. Berquist, K. Brandhorst, K. B. Bravaya, S. T. Brown, D. Casanova, C. M. Chang, Y. Chen, S. H. Chien, K. D. Closser, D. L. Crittenden, M. Diedenhofen, R. A. Distasio, H. Do, A. D. Dutoi, R. G. Edgar, S.

Fatehi, L. Fusti-Molnar, A. Ghysels, A. Golubeva-Zadorozhnaya, J. Gomes, M. W. Hanson-Heine, P. H. Harbach, A. W. Hauser, E. G. Hohenstein, Z. C. Holden, T. C. Jagau, H. Ji, B. Kaduk, K. Khistyaev, J. Kim, J. Kim, R. A. King, P. Klunzinger, D. Kosenkov, T. Kowalczyk, C. M. Krauter, K. U. Lao, A. D. Laurent, K. V. Lawler, S. V. Levchenko, C. Y. Lin, F. Liu, E. Livshits, R. C. Lochan, A. Luenser, P. Manohar, S. F. Manzer, S. P. Mao, N. Mardirossian, A. V. Marenich, S. A. Maurer, N. J. Mayhall, E. Neuscamman, C. M. Oana, R. Olivares-Amaya, D. P. Oneill, J. A. Parkhill, T. M. Perrine, R. Peverati, A. Prociuk, D. R. Rehn, E. Rosta, N. J. Russ, S. M. Sharada, S. Sharma, D. W. Small, A. Sodt, T. Stein, D. Stück, Y. C. Su, A. J. Thom, T. Tsuchimochi, V. Vanovschi, L. Vogt, O. Vydrov, T. Wang, M. A. Watson, J. Wenzel, A. White, C. F. Williams, J. Yang, S. Yeganeh, S. R. Yost, Z. Q. You, I. Y. Zhang, X. Zhang, Y. Zhao, B. R. Brooks, G. K. Chan, D. M. Chipman, C. J. Cramer, W. A. Goddard, M. S. Gordon, W. J. Hehre, A. Klamt, H. F. Schaefer, M. W. Schmidt, C. D. Sherrill, D. G. Truhlar, A. Warshel, X. Xu, A. Aspuru-Guzik, R. Baer, A. T. Bell, N. A. Besley, J. D. Chai, A. Dreuw, B. D. Dunietz, T. R. Furlani, S. R. Gwaltney, C. P. Hsu, Y. Jung, J. Kong, D. S. Lambrecht, W. Liang, C. Ochsenfeld, V. A. Rassolov, L. V. Slipchenko, J. E. Subotnik, T. Van Voorhis, J. M. Herbert, A. I. Krylov, P. M. Gill and M. Head-Gordon, *Mol. Phys.*, 2015, **113**, 184–215.

(239) O. Guvench and A. D. MacKerell, *J. Mol. Model.*, 2008, **14**, 667–679.

(240) K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov and A. D. MacKerell, *J. Comput. Chem.*, 2010, **31**, 671–690.

(241) C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.*, 2016, **72**, 171–179.

(242) Y.-H. Kiang, W. Xu, P. W. Stephens, R. G. Ball and N. Yasuda, *Cryst. Growth Des.*, 2009, **9**, 1833–1843.

(243) W. A. Kinney, D. K. Luci, R. J. Santulli, D. A. Gauthier, B. A. Tounge, S. Ghosh, J. C. Proost, B. De Corte, R. A. Galemmo, J. M. Lewis, W. E. Dorsch, M. W. Wagaman, B. P. Damiano and B. E. Maryanoff, *Heterocycles*, 2004, **62**, 543.

(244) W. Ma, F. Chen, Y. Liu, Y.-M. He and Q.-H. Fan, *Org. Lett.*, 2016, **18**, 2730–2733.

(245) X. Chen, H. Zhao, C. Chen, H. Jiang and M. Zhang, *Chem. Commun.*, 2018, **54**, 9087–9090.

(246) A. Kotecha, Q. Wang, X. Dong, S. L. Ilca, M. Ondiviela, R. Zihe, J. Seago, B. Charleston, E. E. Fry, N. G. A. Abrescia, T. A. Springer, J. T. Huiskonen and D. I. Stuart, *Nat. Commun.*, 2017, **8**, 15408.

(247) X. Dong, B. Zhao, R. E. Iacob, J. Zhu, A. C. Koksal, C. Lu, J. R. Engen and T. A. Springer, *Nature*, 2017, **542**, 55–59.

(248) P. A. Procopiou, N. A. Anderson, J. Barrett, T. N. Barrett, M. H. Crawford, B. J. Fallon, A. P. Hancock, J. Le, S. Lemma, R. P. Marshall, J. Morrell, J. M. Pritchard, J. E. Rowedder, P. Saklatvala, R. J. Slack, S. L. Sollis, C. J. Suckling, L. R. Thorp, G. Vitulli and S. J. Macdonald, *J. Med. Chem.*, 2018, **61**, 8417–8443.

(249) F. S. Di Leva, S. Tomassi, S. Di Maro, F. Reichart, J. Notni, A. Dangi, U. K. Marelli, D. Brancaccio, F. Merlino, H. J. Wester, E. Novellino, H. Kessler and L. Marinelli, *Angew. Chem., Int. Ed.*, 2018, **57**, 14645–14649.

(250) T. G. Kapp, F. Saverio Di Leva, J. Notni, A. F. Räder, M. Fottner, F. Reichart, D. Reich, A. Wurzer, K. Steiger, E. Novellino, U. Kiran Marelli, H.-J. Wester, L. Marinelli and H. Kessler, *J. Med. Chem.*, 2018, **61**, 2490–2499.

(251) M. Ghitti, A. Spitaleri, B. Valentinis, S. Mari, C. Asperti, C. Traversari, G. P. Rizzardi and G. Musco, *Angew. Chem., Int. Ed.*, 2012, **51**, 7702–7705.

(252) T. N. Barrett, J. A. Taylor, D. Barker, P. A. Procopiou, J. D. F. Thompson, J. Barrett, J. Le, S. M. Lynn, P. Pogany, C. Pratley, J. M. Pritchard, J. A. Roper, J. E. Rowedder, R. J. Slack, G. Vitulli, S. J. F. Macdonald and W. J. Kerr, *J. Med. Chem.*, 2019, **62**, 7543–7556.

(253) P. C. Hawkins, A. G. Skillman, G. L. Warren, B. A. Ellingson and M. T. Stahl, *J. Chem. Inf. Model.*, 2010, **50**, 572–584.

(254) T. A. Halgren, *J. Comput. Chem.*, 1999, **20**, 720–729.

(255) S. Jo, T. Kim, V. G. Iyer and W. Im, *J. Comput. Chem.*, 2008, **29**, 1859–1865.

(256) D. Beglov and B. Roux, *J. Chem. Phys.*, 1994, **100**, 9050–9063.

(257) S. E. Feller, Y. Zhang, R. W. Pastor and B. R. Brooks, *J. Comput. Phys.*, 1995, **103**, 4613–4621.

(258) S. E. Feller, R. W. Pastor, A. Rojnuckarin, S. Bogusz and B. R. Brooks, *J. Phys. Chem.*, 1996, **100**, 17011–17020.

(259) P. A. Bash, U. Chandra Singh, F. K. Brown, R. Langridge and P. A. Kollman, *Science*, 1987, **235**, 574–576.

(260) J. P. Xiong, T. Stehle, R. Zhang, A. Joachimiak, M. Frech, S. L. Goodman and M. A. Arnaout, *Science*, 2002, **296**, 151–155.

(261) W. Xia and T. A. Springer, *Proc. Natl. Acad. Sci. U. S. A.*, 2014, **111**, 17863–17868.

(262) O. V. Maltsev, U. K. Marelli, T. G. Kapp, F. S. Di Leva, S. Di Maro, M. Nieberler, U. Reuning, M. Schwaiger, E. Novellino, L. Marinelli and H. Kessler, *Angew. Chem., Int. Ed.*, 2016, **55**, 1535–1539.

(263) M. Civera, D. Arosio, F. Bonato, L. Manzoni, L. Pignataro, S. Zanella, C. Gennari, U. Piarulli and L. Belvisi, *Cancers*, 2017, **9**, 1–13.

(264) J. P. Gallivan and D. A. Dougherty, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**, 9459–9464.

(265) K. Kumar, S. M. Woo, T. Siu, W. A. Cortopassi, F. Duarte and R. S. Paton, *Chem. Sci.*, 2018, **9**, 2655–2665.

(266) E. M. Duffy, W. L. Jorgensen and P. J. Kowalczyk, *J. Am. Chem. Soc.*, 1993, **115**, 9271–9275.

(267) S. Tsuzuki, M. Mikami and S. Yamada, *J. Am. Chem. Soc.*, 2007, **129**, 8656–8662.

(268) P. Liu, F. Dehez, W. Cai and C. Chipot, *J. Chem. Theory Comput.*, 2012, **8**, 2606–2616.

(269) S. F. Sousa, P. A. Fernandes and M. J. Ramos, *Proteins: Struct., Funct., Genet.*, 2006, **65**, 15–26.

(270) L. Wang, Y. Wu, Y. Deng, B. Kim, L. Pierce, G. Krilov, D. Lupyan, S. Robinson, M. K. Dahlgren, J. Greenwood, D. L. Romero, C. Masse, J. L. Knight, T. Steinbrecher, T. Beuming, W. Damm, E. Harder, W. Sherman, M. Brewer, R. Wester, M. Murcko, L. Frye, R. Farid, T. Lin, D. L. Mobley, W. L. Jorgensen, B. J. Berne, R. A. Friesner and R. Abel, *J. Am. Chem. Soc.*, 2015, **137**, 2695–2703.

(271) S. Mujtaba, L. Zeng and M. M. Zhou, *Oncogene*, 2007, **26**, 5521–5527.

(272) P. Filippakopoulos, S. Picaud, M. Mangos, T. Keates, J. P. Lambert, D. Barsyte-Lovejoy, I. Felletar, R. Volkmer, S. Müller, T. Pawson, A. C. Gingras, C. H. Arrowsmith and S. Knapp, *Cell*, 2012, **149**, 214–231.

(273) M. Jung, M. Philpott, S. Müller, J. Schulze, V. Badock, U. Eberspächer, D. Moosmayer, B. Bader, N. Schmees, A. Fernández-Montalván and B. Haendler, *J. Biol. Chem.*, 2014, **289**, 9304–9319.

(274) P. Filippakopoulos and S. Knapp, *Nat. Rev. Drug Discovery*, 2014, **13**, 337–356.

(275) T. D. Crawford, V. Tsui, E. M. Flynn, S. Wang, A. M. Taylor, A. Côté, J. E. Audia, M. H. Beresini, D. J. Burdick, R. Cummings, L. A. Dakin, M. Duplessis, A. C. Good, M. C. Hewitt, H. R. Huang, H. Jayaram, J. R. Kiefer, Y. Jiang, J. Murray, C. G. Nasveschuk, E. Pardo, F. Poy, F. A. Romero, Y. Tang, J. Wang, Z. Xu, L. E. Zawadzke, X. Zhu, B. K. Albrecht, S. R. Magnuson, S. Bellon and A. G. Cochran, *J. Med. Chem.*, 2016, **59**, 5391–5402.

(276) H. Zhong, Z. Wang, X. Wang, H. Liu, D. Li, H. Liu, X. Yao and T. Hou, *Phys. Chem. Chem. Phys.*, 2019, **21**, 25276–25289.

(277) Y. Cheng, C. He, M. Wang, X. Ma, F. Mo, S. Yang, J. Han and X. Wei, *Signal Transduct Target Ther.*, 2019, **4**, 1–39.

(278) B. Padmanabhan, S. Mathur, R. Manjula and S. Tripathi, *J. Biosci*, 2016, **41**, 295–311.

(279) E. Korb, M. Herre, I. Zucker-Scharff, R. B. Darnell and C. D. Allis, *Nat. Neurosci.*, 2015, **18**, 1464–1473.

(280) A. C. Belkina and G. V. Denis, *Nat. Rev. Cancer*, 2012, **12**, 465–477.

(281) P. Filippakopoulos, J. Qi, S. Picaud, Y. Shen, W. B. Smith, O. Fedorov, E. M. Morse, T. Keates, T. T. Hickman, I. Felletar, M. Philpott, S. Munro, M. R. McKeown, Y. Wang, A. L. Christie, N. West, M. J. Cameron, B. Schwartz, T. D. Heightman, N. La Thangue, C. A. French, O. Wiest, A. L. Kung, S. Knapp and J. E. Bradner, *Nature*, 2010, **468**, 1067–1073.

(282) L. F. Chen, Y. Mu and W. C. Greene, *EMBO J.*, 2002, **21**, 6539–6548.

(283) S. Picaud, C. Wells, I. Felletar, D. Brotherton, S. Martin, P. Savitsky, B. Diez-Dacal, M. Philpott, C. Bountra, H. Lingard, O. Fedorov, S. Müller, P. E. Brennan, S. Knapp and P. Filippakopoulos, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 19754–19759.

(284) O. Mirguet, R. Gosmini, J. Toum, C. A. Clément, M. Barnathan, J. M. Brusq, J. E. Mordaunt, R. M. Grimes, M. Crowe, O. Pineau, M. Ajakane, A. Daugan, P. Jeffrey, L. Cutler, A. C. Haynes, N. N. Smithers, C. W. Chung, P. Bamborough, I. J. Uings, A. Lewis, J. Witherington, N. Parr, R. K. Prinjha and E. Nicodème, *J. Med. Chem.*, 2013, **56**, 7501–7515.

(285) M. M. Coudé, T. Braun, J. Berrou, M. Dupont, S. Bertrand, A. Masse, E. Raffoux, R. Itzykson, M. Delord, M. E. Riveiro, P. Herait, A. Baruchel, H. Dombret and C. Gardin, *Oncotarget*, 2015, **6**, 17698–17712.

(286) K. T. Siu, H. Eda, L. Santo, J. Ramachandran, M. Koulnis, J. Mertz, R. J. Sims, M. Cooper and N. S. Raje, *Blood*, 2015, **126**, 4255.

(287) G. Shapiro, A. Dowlati, P. LoRusso, J. Eder, A. Anderson, K. Đ, M. Kagey, C. Sirard, J. Bradner and S. Landau, *Mol. Cancer Ther.*, 2015, **14**, A49–A49.

(288) A. Sarthy, L. Li, D. H. Albert, X. Lin, W. Scott, E. Faivre, M. H. Bui, X. Huang, D. M. Wilcox, T. Magoc, F. G. Buchanan, P. Tapang, G. S. Sheppard, L. Wang, S. D. Fidanze, J. Pratt, D. Liu, L. Hasvold, P. Hessler, T. Uziel, L. Lam, G. Rajaraman, G. Fang, S. W. Elmore, S. H. Rosenberg, K. McDaniel, W. Kati and Y. Shen, *Cancer Res.*, 2016, **76**, 4718 LP –4718.

(289) T. Lu, W. Lu and C. Luo, *Expert Opin. Ther. Pat.*, 2020, **30**, 57–81.

(290) Z. Liu, P. Wang, H. Chen, E. A. Wold, B. Tian, A. R. Brasier and J. Zhou, *J. Med. Chem.*, 2017, **60**, 4533–4558.

(291) Z. Pan, X. Li, Y. Wang, Q. Jiang, L. Jiang, M. Zhang, N. Zhang, F. Wu, B. Liu and G. He, *J. Med. Chem.*, 2020, **63**, 3678–3700.

(292) J. P. Hilton-Proctor, O. Ilyichova, Z. Zheng, I. G. Jennings, R. W. Johnstone, J. Shortt, S. J. Mountford, M. J. Scanlon and P. E. Thompson, *Eur. J. Med. Chem.*, 2020, **191**, 112120.

(293) Z. Li, S. Xiao, Y. Yang, C. Chen, T. Lu, Z. Chen, H. Jiang, S. Chen, C. Luo and B. Zhou, *J. Med. Chem.*, 2020, **63**, 3956–3975.

(294) F. Jiang, Q. Hu, Z. Zhang, H. Li, H. Li, D. Zhang, H. Li, Y. Ma, J. Xu, H. Chen, Y. Cui, Y. Zhi, Y. Zhang, J. Xu, J. Zhu, T. Lu and Y. Chen, *J. Med. Chem.*, 2019, **62**, 11080–11107.

(295) C. R. Wellaway, D. Amans, P. Bamborough, H. Barnett, R. A. Bit, J. A. Brown, N. R. Carlson, C. W. Chung, A. W. Cooper, P. D. Craggs, R. P. Davis, T. W. Dean, J. P. Evans, L. Gordon, I. L. Harada, D. J. Hirst, P. G. Humphreys, K. L. Jones, A. J. Lewis, M. J. Lindon, D. Lugo, M. Mahmood, S. McCleary, P. Medeiros, D. J. Mitchell, M. O'Sullivan, A. Le Gall, V. K. Patel, C. Patten, D. L. Poole, R. R. Shah, J. E. Smith, K. A. Stafford, P. J. Thomas, M. Vimal, I. D. Wall, R. J. Watson, N. Wellaway, G. Yao and R. K. Prinjha, *J. Med. Chem.*, 2020, **63**, 714–746.

(296) D. Liang, Y. Yu and Z. Ma, *Eur. J. Med. Chem.*, 2020, **200**, 112426.

(297) O. Gilan, I. Rioja, K. Knezevic, M. J. Bell, M. M. Yeung, N. R. Harker, E. Y. Lam, C. Chung, P. Bamborough, M. Petretich, M. Urh, S. J. Atkinson, A. K. Bassil, E. J. Roberts, D. Vassiliadis, M. L. Burr, A. G. Preston, C. Wellaway, T. Werner, J. R. Gray, A. M. Michon, T. Gobbetti, V. Kumar, P. E. Soden, A. Haynes, J. Vappiani, D. F. Tough, S. Taylor, S. J. Dawson, M. Bantscheff, M. Lindon, G. Drewes, E. H. Demont, D. L. Daniels, P. Grandi, R. K. Prinjha and M. A. Dawson, *Science*, 2020, **368**, 387–394.

(298) R. J. Watson, P. Bamborough, H. Barnett, C. W. Chung, R. Davis, L. Gordon, P. Grandi, M. Petretich, A. Phillipou, R. K. Prinjha, I. Rioja, P. Soden, T. Werner and E. H. Demont, *J. Med. Chem.*, 2020, **63**, 9045–9069.

(299) A. Speck-Planche and M. T. Scotti, *Mol. Diversity*, 2019, **23**, 555–572.

(300) S. Postel-Vinay, K. Herbschleb, C. Massard, V. Woodcock, J.-C. Soria, A. O. Walter, F. Ewerton, M. Poelman, N. Benson, M. Ocker, G. Wilkinson and M. Middleton, *Eur. J. Cancer*, 2019, **109**, 103–110.

(301) M. Pervaiz, P. Mishra and S. Günther, *Chem. Rec.*, 2018, **18**, 1808–1817.

(302) B. K. Allen, S. Mehta, S. W. Ember, J. Y. Zhu, E. Schönbrunn, N. G. Ayad and S. C. Schürer, *ACS Omega*, 2017, **2**, 4760–4771.

(303) A. Tahir, R. D. Alharthy, S. Naseem, N. Mahmood, M. Ahmed, K. Shahzad, M. N. Akhtar, A. Hameed, I. Sadiq, H. Nawaz and M. Muddassar, *Molecules*, 2018, **23**, 1–16.

(304) J. Su, X. Liu, S. Zhang, F. Yan, Q. Zhang and J. Chen, *Chem. Biol. Drug Des.*, 2018, **91**, 828–840.

(305) J. Su, X. Liu, S. Zhang, F. Yan, Q. Zhang and J. Chen, *Chem. Biol. Drug Des.*, 2019, **93**, 163–176.

(306) J. J. Irwin and B. K. Shoichet, *J. Chem. Inf. Model.*, 2005, **45**, 177–182.

(307) L. Fusani and A. C. Cabrera, *J. Comput.-Aided Mol. Des.*, 2019, **33**, 287–294.

(308) S. E. Nichols, R. Baron and J. A. McCammon, in *Computational Drug Discovery and Design*, ed. R. Baron, Springer New York, New York, NY, 2012, pp. 93–103.

(309) N. M. Lim, L. Wang, R. Abel and D. L. Mobley, *J. Chem. Theory Comput.*, 2016, **12**, 4620–4631.

(310) A. R. Bradley, I. D. Wall, F. Von Delft, D. V. Green, C. M. Deane and B. D. Marsden, *J. Comput.-Aided Mol. Des.*, 2015, **29**, 963–973.

(311) C. M. Deane, I. D. Wall, D. V. Green, B. D. Marsden and A. R. Bradley, *Acta Crystallogr., Sect. D: Struct. Biol.*, 2017, **73**, 279–285.

(312) Schrödinger, LLC, "The PyMOL Molecular Graphics System, Version~1.8", 2015.

(313) F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson and D. G. Higgins, *Mol. Syst. Biol.*, 2011, **7**, 1–6.

(314) E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng and T. E. Ferrin, *J. Comput. Chem.*, 2004, **25**, 1605–1612.

(315) M. Jamroz and A. Kolinski, *BMC Bioinf.*, 2013, **14**, 62.

(316) J. MacQueen, *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob.*, 1967, **1**, 281–297.

(317) T. Sander, J. Freyss, M. Von Korff and C. Rufener, *J. Chem. Inf. Model.*, 2015, **55**, 460–473.

(318) J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. De Groot, H. Grubmüller and A. D. MacKerell, *Nat. Methods*, 2016, **14**, 71–73.

(319) T. Kalliokoski, C. Kramer, A. Vulpetti and P. Gedeck, *PLoS ONE*, 2013, **8**.

(320) J. P. Hilton-Proctor, O. Ilyichova, Z. Zheng, I. G. Jennings, R. W. Johnstone, J. Shortt, S. J. Mountford, M. J. Scanlon and P. E. Thompson, *Bioorg. Med. Chem.*, 2019, **27**, 115157.

(321) E. Watts, D. Heidenreich, E. Tucker, M. Raab, K. Strebhardt, L. Chesler, S. Knapp, B. Bellenie and S. Hoelder, *J. Med. Chem.*, 2019, **62**, 2618–2637.

(322) L. Fusani, I. Wall, D. Palmer and A. Cortes, *Bioinformatics*, 2018, **34**, 1947–1948.

(323) R. Abel, T. Young, R. Farid, B. J. Berne and R. A. Friesner, *J. Am. Chem. Soc.*, 2008, **130**, 2817–2831.

(324) C. N. Nguyen, T. Kurtzman Young and M. K. Gilson, *J. Chem. Phys.*, 2012, **137**, 973–980.

(325) D. Beglov and B. Roux, *J. Phys. Chem. B*, 1997, **101**, 7821–7826.

(326) A. Divakaran, S. K. Talluri, A. M. Ayoub, N. K. Mishra, H. Cui, J. C. Widen, N. Berndt, J. Y. Zhu, A. S. Carlson, J. J. Topczewski, E. K. Schonbrunn, D. A. Harki and W. C. Pomerantz, *J. Med. Chem.*, 2018, **61**, 9316–9334.

(327) S. J. Atkinson, P. E. Soden, D. C. Angell, M. Bantscheff, C. W. Chung, K. A. Giblin, N. Smithers, R. C. Furze, L. Gordon, G. Drewes, I. Rioja, J. Witherington, N. J. Parr and R. K. Prinjha, *MedChemComm*, 2014, **5**, 342–351.

(328) Z. Cournia, B. K. Allen, T. Beuming, D. A. Pearlman, B. K. Radak and W. Sherman, *J. Chem. Inf. Model.*, 2020, **60**, 4153–4169.

(329) M. Kuhn, S. Firth-Clark, P. Tosco, A. S. Mey, M. MacKey and J. Michel, *J. Chem. Inf. Model.*, 2020, **60**, 3120–3130.

(330) S. Wan, A. P. Bhati, S. J. Zasada, I. Wall, D. Green, P. Bamborough and P. V. Coveney, *J. Chem. Theory Comput.*, 2017, **13**, 784–795.

(331) S. Wan, B. Knapp, D. W. Wright, C. M. Deane and P. V. Coveney, *J. Chem. Theory Comput.*, 2015, **11**, 3346–3356.

(332) A. P. Bhati, S. Wan, D. W. Wright and P. V. Coveney, *J. Chem. Theory Comput.*, 2017, **13**, 210–222.

(333) B. Brooks, R. Bruccoleri, B. Olafson, D. States, S. Swaminathan and M. Karplus, *J. Comput. Chem.*, 1983, **4**, 187–217.

(334) A. P. Hynninen and M. F. Crowley, *J. Comput. Chem.*, 2014, **35**, 406–413.

(335) W. G. Hoover, *Phys. Rev. A*, 1985, **31**, 1695–1697.

(336) K. A. Armacost, G. B. Goh and C. L. Brooks, *J. Chem. Theory Comput.*, 2015, **11**, 1267–1277.

(337) A. P. Bhati, S. Wan, Y. Hu, B. Sherborne and P. V. Coveney, *J. Chem. Theory Comput.*, 2018, **14**, 2867–2880.

(338) H. H. Loeffler, J. Michel and C. Woods, *J. Chem. Inf. Model.*, 2015, **55**, 2485–2490.

(339) W. Jespers, M. Esguerra, J. Åqvist and H. Gutiérrez-De-Terán, *J. Cheminf.*, 2019, **11**, 1–16.

(340) L. Carvalho Martins, E. A. Cino and R. S. Ferreira, *J. Chem. Theory Comput.*, 2021, **17**, 4262–4273.

(341) D. J. Cole, J. T. Horton, L. Nelson and V. Kurdekar, *Future Med. Chem.*, 2019, **11**, 2359–2363.

(342) *Open Force Field Initiative*, https://openforcefield.org, 2021.

(343) R. Galvelis, S. Doerr, J. M. Damas, M. J. Harvey and G. De Fabritiis, *J. Chem. Inf. Model.*, 2019, **59**, 3485–3493.

(344) J. T. Horton, A. E. Allen, L. S. Dodda and D. J. Cole, *J. Chem. Inf. Model.*, 2019, **59**, 1366–1381.

(345) D. L. Mobley, C. C. Bannan, A. Rizzi, C. I. Bayly, J. D. Chodera, V. T. Lim, N. M. Lim, K. A. Beauchamp, D. R. Slochower, M. R. Shirts, M. K. Gilson and P. K. Eastman, *J. Chem. Theory Comput.*, 2018, **14**, 6076–6092.

(346) J. S. Smith, O. Isayev and A. E. Roitberg, *Chemical Science*, 2017, **8**, 3192–3203.

Appendix A

Protein-ligand interactions of integrin inhibitors during MD simulations

In Chapter 4, MD simulations of protein-ligand complexes of integrin inhibitors are described. The interaction frequencies between certain binding site residues and the ligands are calculated and discussed. Interaction frequency is calculated by the proportion of frames with the interaction of interest present with respect to the total number of frames, averaged over five 10 ns simulations. Table A.1 shows a full description of the interaction frequency of the canonical interactions. Interactions with binding site residues, additional to those shown in the main body of Chapter 4 are also shown. Interactions H1, H2, M1 and M2 refer to those labelled in Figure A.1.

Table A.1: Frequency in % of simulation time that interactions between atoms involving , (αv)-Asp^{218}, and Mg^{2+} are present. Interaction H3 refers to the interaction between the H1 polar hydrogen and H2 (αv)-Asp^{218} oxygen atom, i.e. the diagonal interaction. Interaction H4 refers to the other diagonal hydrogen bond between the polar H2 hydrogen atom and the H1 (αv)-Asp^{218} oxygen atom. The proportion of time each derivative of the ligand interacts with residues (β6)-Ala^{126} and (β6)-Asn^{218} is also shown.

				Interaction Frequency (% Time)							
Substituent	Conformation	pIC$_{50}$	Docking Score	H1	H2	H3	H4	M1	M2	(β6)-Ala^{126}	(β6)-Asn^{218}
H	*S*	5.7	-9.99	89	89	77	0	100	0	67	47
	R		-9.19	99	96	43	7	100	100	0	0
F	*S*	6.1	-9.64	99	96	37	3	100	46	45	27
	R		-9.29	29	30	30	1	89	80	0	20
CH_3	*S*	6.4	-7.96	100	100	34	4	100	73	10	6
	R		-4.64	100	100	54	5	100	100	0	0
OCH_3	*S*	6.5	-9.38	65	66	49	2	100	100	0	3
	R		-6.57	100	100	58	2	100	100	0	0
OCF_3	*S*	6.7	-8.49	100	98	36	6	100	40	60	28
	R		-3.59	99	99	19	18	100	0	0	0
CF_3	*S*	7.1	-9.33	75	74	52	2	100	100	0	33
	R	5.2	-5.70	26	19	14	4	0	100	0	0

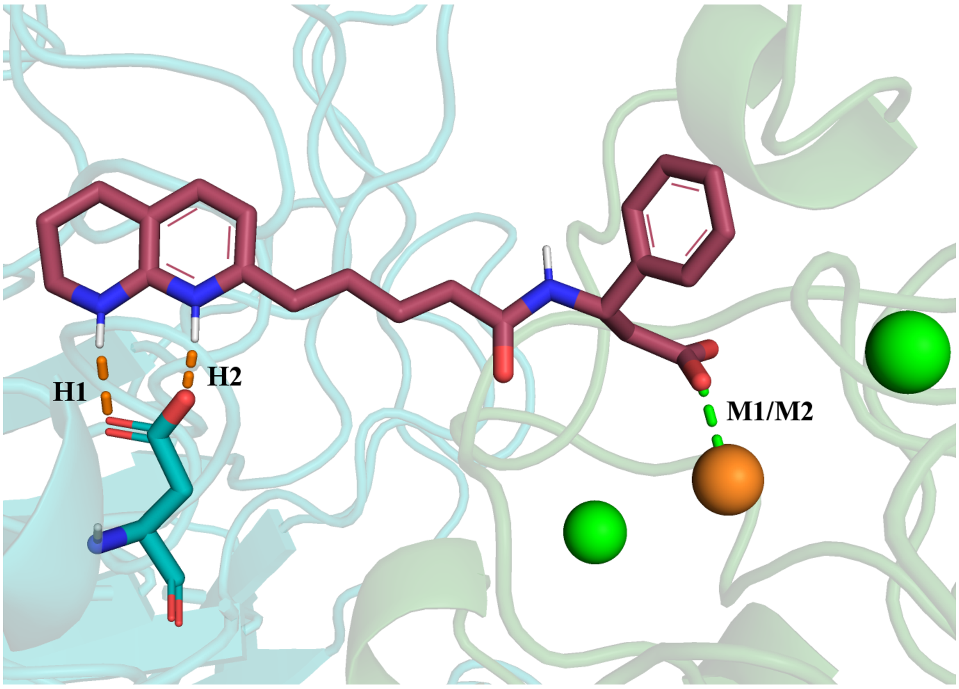


Figure A.1: The RGD mimetic bound to αvβ6. Canonical interactions are labelled. Whether the interaction is M1 or M2 depends on which carboxyl oxygen atom on the RGD mimetic is interacting with the receptor.

Appendix B

Convergence assessment of relative FEP simulations for the calculation of binding free energies for BRD4-ligand complexes

Chapter 6 describes the benchmarking of relative FEP simulations with varying numbers of λ windows, equilibration length and data collection length. The following plots show the convergence of the binding free energy with each combination of parameters tested. The forward (purple lines) and the reverse (green lines) simulation time series are shown and the horizontal shaded bar indicates ±0.5 kcal mol^{-1} of the final value.

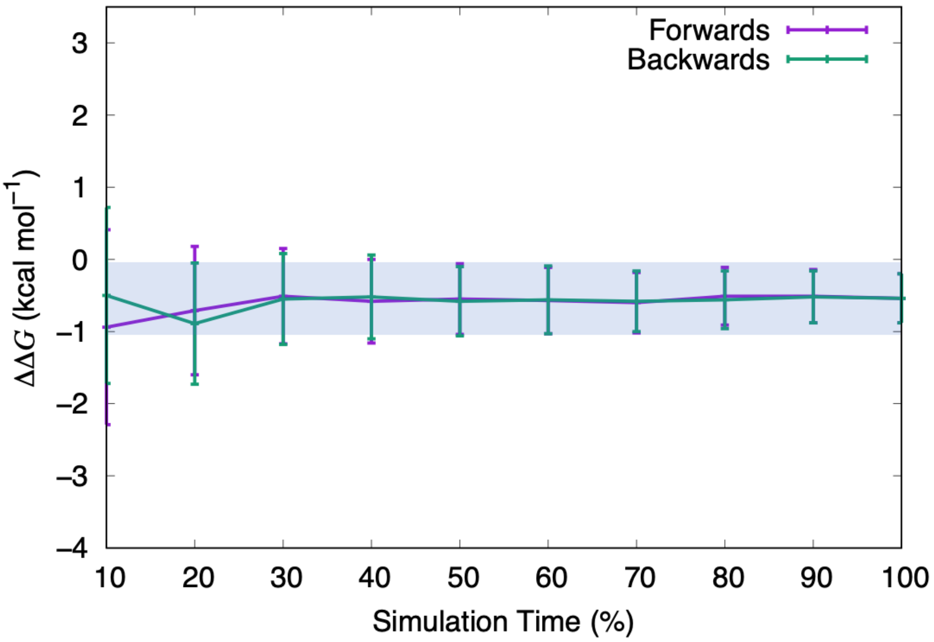


Figure B.1: Convergence assessment of a relative FEP simulation using 25 λ windows with 2 ns of equilibration and 1 ns of data collection.

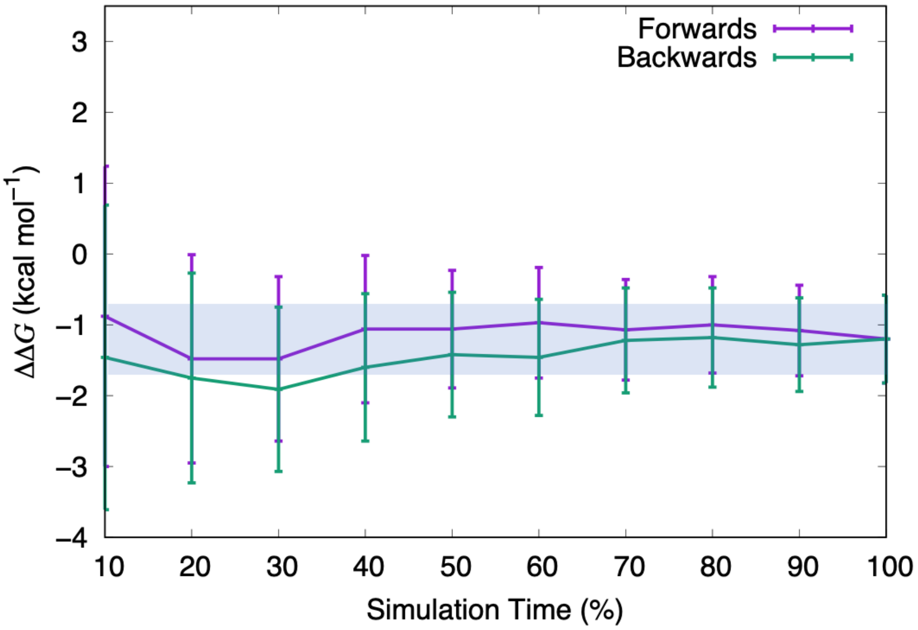


Figure B.2: Convergence assessment of a relative FEP simulation using 10 λ windows with 2 ns of equilibration and 1 ns of data collection.

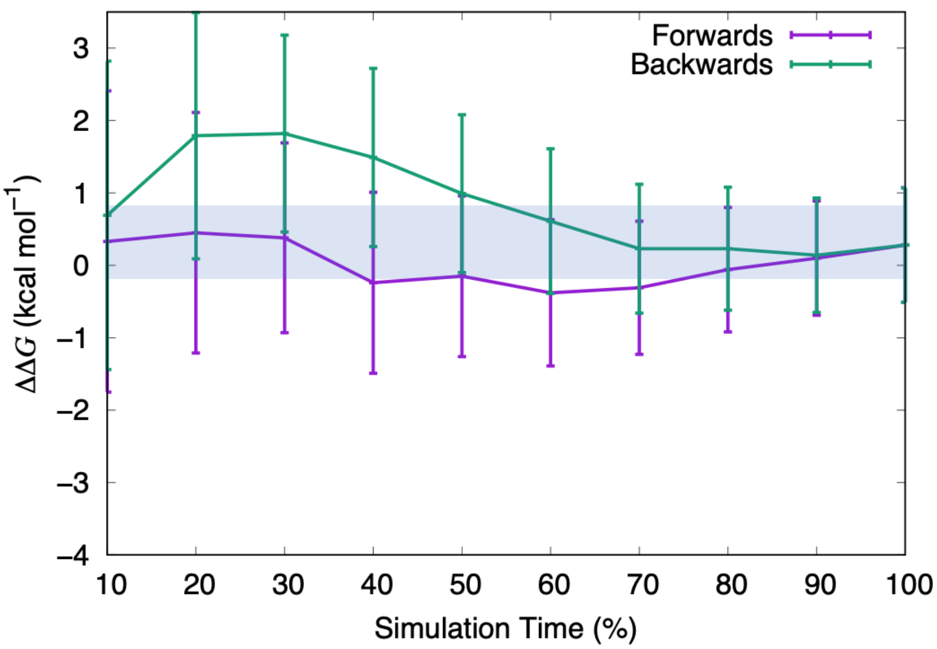


Figure B.3: Convergence assessment of a relative FEP simulation using 8 λ windows with 2 ns of equilibration and 1 ns of data collection.

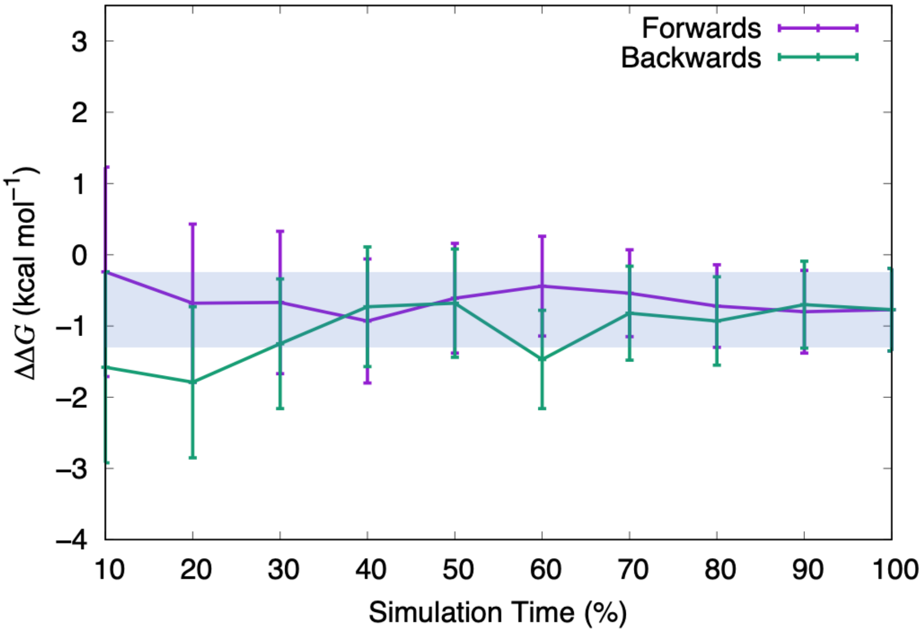


Figure B.4: Convergence assessment of a relative FEP simulation using 20 λ windows with 2 ns of equilibration and 0.5 ns of data collection.

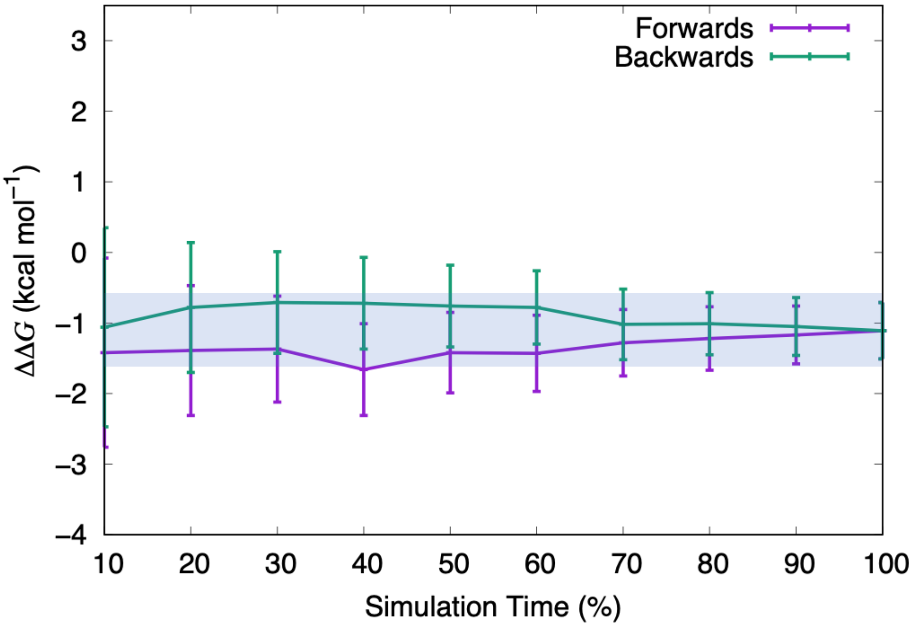


Figure B.5: Convergence assessment of a relative FEP simulation using 20 λ windows with 1 ns of equilibration and 1 ns of data collection.

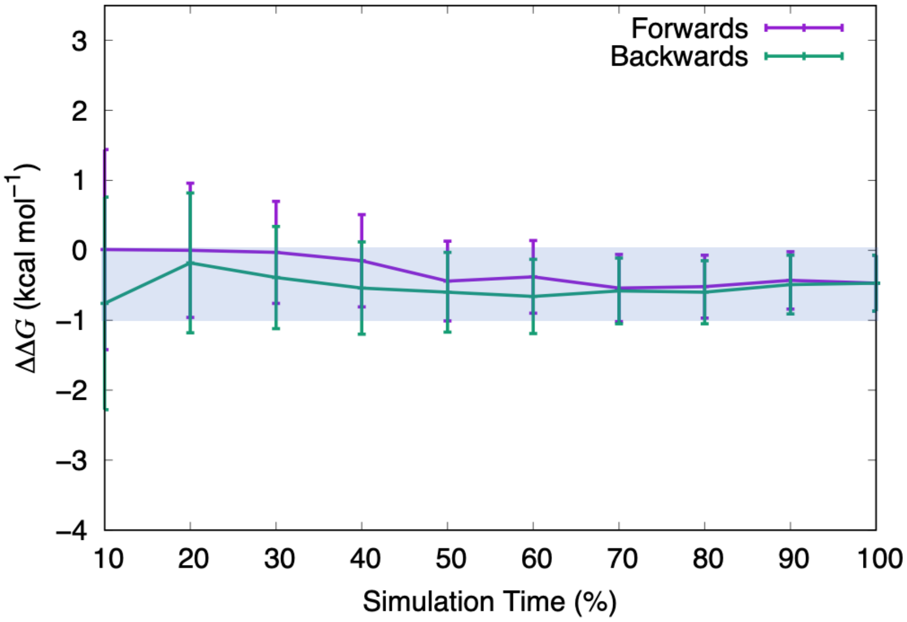


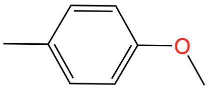
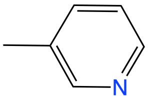
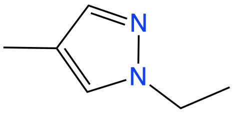
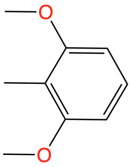
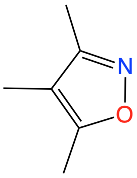
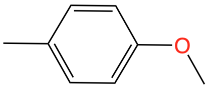
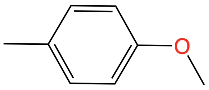
Figure B.6: Convergence assessment of a relative FEP simulation using 20 λ windows with 0.5 ns of equilibration and 1 ns of data collection.

Appendix C

RBFE predictions of BRD4-BD1 inhibitors with a net neutral charge calculated using a single MSλD simulation

Chapter 6 describes the calculation of RBFEs for a series of BRD4-BD1 inhibitors, based on a THQ scaffold. The predicted relative binding affinities for the compounds with a net neutral charge, calculated using a single MSλD simulation are shown below. In the main body of the text, results are presented for the compounds when they were split into two sets of MSλD simulations, which achieved more accurate results compared to the ones presented below.

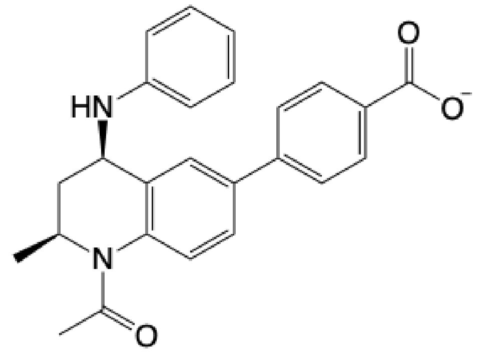
Table C.1: RBFE predictions for a series of BRD4-BD1 inhibitors calculated using a single MSλD simulation. Relative free energy changes are shown in kcal mol^{-1}. R positions correspond to Figure 6.2 in the main text.

ID	R1	R2	R3	R4	$\Delta\Delta G_{exp}$	$\Delta\Delta G_{MS\lambda D}$	$\lvert\Delta\Delta G_{MS\lambda D}$ $-\Delta\Delta G_{exp}\rvert$
1	H	Me	Me		-0.3 ± 0.1	0.3 ± 0.2	0.6
2	H	Me	Me	H	1.6 ± 0.1	1.1 ± 0.7	0.5
4	H	Me	Me		0.0 ± 0.1	0.7 ± 0.4	0.7
5	H	Me	Me		-1.5 ± 0.1	2.1 ± 0.4	3.6
6	H	Me	Me		1.6 ± 0.1	0.2 ± 0.4	1.4
7	H	Me	Me		1.3 ± 0.1	1.2 ± 0.2	0.1
8	H	Me	Et		0.4 ± 0.1	-0.2 ± 0.4	0.6
9	H	Me	i-Pr		≥ 3.4	-0.1 ± 0.3	≥ 3.5

Appendix D

RBFE predictions for intermediate compounds using relative FEP

In Chapter 6, relative FEP simulations are used to calculate relative binding affinities for a series of BRD4-BD1 inhibitors. A few of these perturbations involved changing substituents on more than one attachment point of the scaffold. Therefore, intermediate transformations were required. The results of which are shown below. The structure of compound **15h** is shown in Figure D.1.



15h

Figure D.1: The structure of compound **15h**, which is used as an intermediate FEP compound for the calculation of RBFE for compounds **13**, **14** and **15**, with respect to compound **3**.

Table D.1: RBFE predictions for intermediate compounds using relative FEP. Compound numbers correspond to those in Table 6.2.

Transformation	$\Delta\Delta G_{calc}$ (kcal mol^{-1})	Uncertainty (kcal mol^{-1})
3 → 10	-0.1	0.5
10 → 11	0.7	0.2
10 → 12	1.9	0.2
3 → 15h	-0.7	0.6
15h → 13	2.1	0.3
15h → 14	0.3	0.2
15h → 15	-0.9	0.2