

Efficient Methods for Exploring Chemical Space in Computational Drug Discovery



Alexander David Wade

Department of Physics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Selwyn College

September 2020

Declaration

The work in this thesis was completed between October 2017 and October 2020 in the Theory of Condensed Matter (TCM) group at the Cavendish Laboratory, Cambridge. This work was supervised by Dr D.J. Huggins. Chapter 1 outlines the scope of this thesis and introduces the general theory used. The subsequent chapters contain original work which is submitted to peer review or has been published as follows:

- Chapter 2: Wade, Alexander D., Andrea Rizzi, Yuanqing Wang, and David J. Huggins. "Computational Fluorine Scanning Using Free-Energy Perturbation." *Journal of chemical information and modeling* 59, no. 6 (2019): 2776-2784.
- Chapter 3: Wade, Alexander D., and David J. Huggins. "Optimization of Protein–Ligand Electrostatic Interactions Using an Alchemical Free-Energy Method." *Journal of chemical theory and computation* 15, no. 11 (2019): 6504-6512.
- Chapter 4: Wade, Alexander D., and David J. Huggins. "Identification of Optimal Ligand Growth Vectors Using an Alchemical Free-Energy Method." *Journal of chemical information and modeling*, <https://doi.org/10.1021/acs.jcim.0c00610>, (just accepted)
- Chapter 5: Irwin, Benedict W. J., Wade, Alexander D. and Segall, Matt D. "Guiding Drug Optimization Using Deep Learning Imputation and Compound Generation" *International Pharmaceutical Industry, Drug Discovery, Development & Delivery* 12 no. 2 (2020): 28-31

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Alexander David Wade
September 2020

Abstract

Efficient Methods for Exploring Chemical Space in Computational Drug Discovery

Alexander David Wade

In this work novel computational methods will be developed to efficiently explore chemical space in the search for compounds with desirable properties. To improve the efficiency of exploration two methods will be used: reducing the cost of evaluating a point in chemical space, or reducing the number of points which require evaluating to find the desired compound. The first chapter of this work will introduce the topics relevant to this work, place them in the wider context of drug design and outline the theory used to generate the results presented in subsequent chapters.

The first result of this thesis, discussed in chapter 2, is for the application of free energy methods to the problem of computational fluorine scanning. The application made in this work will allow for all fluorinated analogues of a compound to be tested five times faster than existing computational methods and with comparable predictive accuracy.

In chapters 3 and 4 we will consider the application of numerical methods to ligand-protein binding problems in order to optimize the charge/steric parameters of the ligand and maximize binding affinity of these ligands to a given protein target. In these two optimization-based chapters we will use free energy methods to calculate gradients of the binding free energy with respect to the parameters which describe the ligand, thus allowing optimal sets of parameters to be found efficiently. In chapter 3 we search for optimized sets of charge parameters from which design ideas can be generated and tested; 73% of the design ideas were found to beneficially improve binding affinity. In chapter 4 we find optimized sets of steric parameters from which beneficial growth vectors for methyl groups can be predicted. These predictions correlate with existing free energy methods with a Spearman's rank order correlation of 0.59. The advantage of the optimization methods presented in these chapters are: 1) the methods can generate ideas for mutations which improve ligand binding free energy and 2) these methods require less computational time to explore the same volume of chemical space than existing free energy methods.

Finally, chapter 5 will discuss a collaborative open source work to find new malaria therapeutics. Ligand based machine learning methods will be applied to generate and evaluate the potency of hundreds of thousands of compounds in a manner far faster than is possible with free energy methods. Based on the computational predictions, compounds are selected and evaluated experimentally with one compound tested and verified to be active with a pIC₅₀ of 6.2 in good agreement with the computational prediction of 6.42 ± 0.75 .

Acknowledgements

I would like to gratefully acknowledge my supervisor Dr. David Huggins for his direction and supervision throughout the PhD. Also, Dr. Ben Irwin, a former member of the Huggins lab, for his valuable guidance with this work. The following specific contributions have been made to this work:

- Andrea Rizzi and Yuanqing Wang, members of John Chodera's lab, contributed to the Fluorify code used in chapter 2. Adding functionality for the automatic preparation of ligand protein systems.
- Dr. David Huggins wrote the code to perform the free energy perturbation calculations involving methylations used in chapter 4.
- Dr. Ben Irwin and Dr. Mario Öeren built the *Plasmodium falciparum* activity model used in chapter 5.
- Dr. Peter Hunt assessed ideated compounds for reactivity and synthesizability in chapter 5.
- Dr. Edwin Tse synthesised and sent for assay the compounds generated in chapter 5.

I would like to deeply thank my peers Michael Hutcheon, Ben Irwin, Juraj Mavračić and Nick Woods for making this work significantly more enjoyable, for valuable discussions about this work and for help with proofreading this thesis.

Table of contents

List of figures	x
List of tables	xiii
Nomenclature	xv
1 Theory	1
1.1 Introduction	1
1.2 Drug Design	2
1.2.1 Target-based vs. Phenotypic	2
1.2.2 Assays	2
1.2.3 Hit Identification	3
1.2.4 Hit to Lead	4
1.2.5 Lead Optimization and Late Stage	6
1.2.6 Chemical Space	6
1.3 Solving Newton's Equations of Motion on a Computer	8
1.3.1 Molecular Dynamics	9
1.3.2 Force Fields	13
1.4 Free Energy	16
1.4.1 Free Energy Methods	19
1.5 Ligand Binding	25
1.6 Optimization	26
1.6.1 Techniques	28
1.7 Machine Learning	30
1.8 Parallelization	34
2 Computational Fluorine Scanning	36
2.1 Introduction	36
2.2 Methods	38

2.2.1	System Setup	41
2.2.2	Molecular Dynamics	41
2.2.3	Perturbative Fluorine Scanning	41
2.2.4	FEP Calculations	42
2.2.5	Summary of Methods	43
2.3	Results	43
2.4	Conclusion	48
3	Charge Optimization	50
3.1	Introduction	50
3.2	Methods	51
3.2.1	System Setup and Molecular Dynamics	52
3.2.2	Workflow	52
3.2.3	Optimization	53
3.2.4	SSP Objective	56
3.2.5	SSP Gradient	56
3.2.6	Optimization Validation	58
3.2.7	FEP Calculations	60
3.2.8	Summary of Methods	62
3.3	Results	62
3.4	Conclusion	68
4	Steric Optimization	70
4.1	Introduction	70
4.2	Methods	72
4.2.1	System Setup and Molecular Dynamics	73
4.2.2	Workflow	74
4.2.3	Optimization	75
4.2.4	MBAR Objective	76
4.2.5	SSP Gradient	77
4.2.6	Optimization Validation	77
4.2.7	FEP Scans	80
4.2.8	Summary of Methods	81
4.3	Results	83
4.4	Conclusion	99

5	Machine Learnt Compound Generation	102
5.1	Introduction	102
5.2	Methods	103
5.2.1	Bottom-Up Generation	104
5.2.2	Top-Down Generation	105
5.2.3	Other OSM methods	109
5.2.4	Summary of Methods	109
5.3	Results	110
5.4	Discussion	116
5.5	Conclusion	117
6	Final Discussion of Results	118
	References	120
	Appendix A Computational Fluorine Scanning	138
	Appendix B Charge optimization	146
	Appendix C Sterics optimization	154

List of figures

1.1	Alchemical topologies	23
1.2	Alchemical thermodynamic cycle	24
1.3	Recurrent neural network training	32
2.1	FXa inhibitors	37
2.2	2D ligands for fluorine scans	39
2.3	Hybrid topology for fluorine scanning	43
2.4	Fluorine scanning convergence	44
2.5	Fluorine scanning correlation	47
3.1	Charge optimization work flow	53
3.2	Charge optimization objective convergence	55
3.3	Androgen receptor ligand with example optimized hydrogens	57
3.4	Dot product of the normalized optimized charges with the normalized original charges	59
3.5	Convergence plots for charge optimization	61
3.6	Optimized ligands colored by change in charge	63
4.1	Workflow for sterics optimization	75
4.2	Sterics optimization line search	77
4.3	Sterics objective sampling convergence	78
4.4	Sterics gradient sampling convergence	79
4.5	Sterics convergence over optimization iteration	80
4.6	$\Delta\Delta G_{scan}$ convergence of methylations	82
4.7	Androgen receptor ligand with optimized sterics	84
4.8	SARS PLPro ligand with optimized sterics	86
4.9	Renin ligand with optimized sterics	88
4.10	Menin A ligand with optimized sterics	89
4.11	Menin B ligand with optimized sterics	91

4.12	Thrombin A ligand with optimized sterics	92
4.13	Thrombin B ligand with optimized charges	94
4.14	Thrombin C ligand with optimized sterics	95
4.15	Thrombin D ligand with optimized sterics	98
4.16	Final optimized thrombin ligand	100
5.1	Examples of soluble and insoluble molecules produced by the RNN.	107
5.2	An example compound generated using an optimized vector as input to the RNN.	108
5.3	An example compound generated using a known active vector as input to the RNN.	109
A.1	Renin complex <i>RMSD</i> over trajectory length	141
A.2	DPP4 complex <i>RMSD</i> over trajectory length	142
A.3	Menin complex <i>RMSD</i> over trajectory length	142
A.4	P38 complex <i>RMSD</i> over trajectory length	143
A.5	FXa complex <i>RMSD</i> over trajectory length	143
A.6	CDK2 complex <i>RMSD</i> over trajectory length	144
A.7	AKT complex <i>RMSD</i> over trajectory length	144
A.8	JAK complex <i>RMSD</i> over trajectory length	145
A.9	Androgen receptor complex <i>RMSD</i> over trajectory length	145
B.1	Optimized Fxa ligand using 0.05 q_e rmsd limit	146
B.2	Optimized Fxa ligand using 0.03 q_e rmsd limit	147
B.3	Optimized Fxa ligand using 0.01 q_e rmsd limit	147
B.4	Optimized P38 ligand using 0.05 q_e rmsd limit	148
B.5	Optimized P38 ligand using 0.03 q_e rmsd limit	148
B.6	Optimized P38 ligand using 0.01 q_e rmsd limit	149
B.7	Optimized AR ligand using 0.05 q_e rmsd limit	149
B.8	Optimized AR ligand using 0.03 q_e rmsd limit	150
B.9	Optimized AR ligand using 0.01 q_e rmsd limit	150
B.10	Investigation for finite size effects of unbalanced charge in charge optimization calculations	152
B.11	Investigation for sampling needed to converge MBAR calculation of charge gradient	153
C.1	First set of steric optimization verification calculations	157
C.2	Second set of steric optimization verification calculations	158

C.3	Third set of steric optimization verification calculations	159
C.4	Convergence plots for $\Delta\Delta G_{calc}$ in SARS system	160
C.5	Convergence plots for $\Delta\Delta G_{calc}$ in renin system	161
C.6	Convergence plots for $\Delta\Delta G_{calc}$ in menin A system	161
C.7	Convergence plots for $\Delta\Delta G_{calc}$ in menin B system	162
C.8	Convergence plots for $\Delta\Delta G_{calc}$ in thrombin A system	162
C.9	Convergence plots for $\Delta\Delta G_{calc}$ in thrombin C system	163
C.10	Convergence plots for $\Delta\Delta G_{calc}$ in thrombin D system	163
C.11	Example of instability causing close contact in androgen receptor system . .	164

List of tables

2.1	Experimental binding free energy for fluorine scanning case studies	40
2.2	Binding free energy calculation results from fluorine scans	46
2.3	Experimental and computational correlation for fluorine scan results	48
3.1	Ligand structures for charge optimization	52
3.2	Dot products of the normalized vector of optimal charges	60
3.3	Summary of methods for charge optimization	62
3.4	Summary of results for ideated compounds from charge optimization	64
4.1	Ligand structures for sterics optimization	72
4.2	Steric test case PDB IDs	74
4.3	Reproducibility of steric optimizations	78
4.4	Summary of methods for steric optimization	83
4.5	Data for optimized sterics of androgen receptor ligand	84
4.6	Data for optimized sterics of SARS PLPro ligand	87
4.7	Experimental data comparison for predicted SARS PLPro ligand growths .	87
4.8	Data for optimized sterics of renin ligand	89
4.9	Data for optimized sterics of menin A ligand	90
4.10	Data for optimized sterics of Thrombin A ligand	93
4.11	Data for optimized sterics of Thrombin C ligand	96
4.12	Data for optimized sterics of Thrombin D ligand	97
4.13	Experimental data comparison for predicted thrombin ligand growths . . .	99
5.1	Generated and optimized descriptors in toy solubility example	107
5.2	Compounds generated by an RNN using an optimized Alchemite vector as input	111
5.3	Predicted activity for alch opt compounds	112
5.4	Compounds generated by an RNN using an active vector as input	113
5.5	Predicted activity for active vector compounds	113

5.6	Compounds generated using a medicinal chemistry expansion	114
5.7	Predicted activity for med chem compounds	114
5.8	Experimental pIC50 for OSM compounds	115
A.1	All un-averaged data from fluorine scan calculations	138
B.1	Verification calculation for charge optimized free energies	151
C.1	Full presentation of ligands used in steric optimization	154
C.2	Data for all steric optimization verification calculation	157
C.3	Optimized steric parameters for menin B test case	165
C.4	Optimized charge parameters for thrombin B test case	166

Nomenclature

Acronyms / Abbreviations

ADME Absorption, distribution, metabolism, and excretion

BAR Bennet acceptance ratio

DNN Deep neural network

EXP Exponential averaging

FEP Free energy perturbation

FF Force fields

GD Gradient descent

CPU Central processing unit

GPU Graphics processing unit

GRU Gated recurrent unit

HPC High performance computing

HREX Hamiltonian replica exchange

HTS High throughput screening

LSTM Long short-term memory

MBAR Multi-state Bennet acceptance ratio

MCS Maximum common substructure

MD Molecular dynamics

ML Machine learning

MMPBSA Molecular mechanics-Poisson-Boltzmann/surface area

MOA Mechanism of action

OSM Open source malaria

pfal Plasmodium falciparum

PFS Perturbative fluorine scanning

PME Particle mesh-Ewald

QSAR Quantitative structure–activity relationship

RNN Recurrent neural network

SAR Structure activity relationship

SSP Single step perturbation

TI Thermodynamic integration

Chapter 1

Theory

1.1 Introduction

In this work we will be studying complex biomolecular systems, in particular, we will primarily be addressing the protein-ligand binding problem. Proteins are prevalent structures in nature found in biological settings. The relative simplicity of a protein's building blocks, amino acids, betrays the complexity of their structure and function. In eukaryotes proteins are built from strings of only 21 types of amino acids, these are 20 standard residues plus selenocysteine. the combinatorics of how these amino acids can be arranged allows for a vast number of unique strings to exist. This is not to mention the enormous complexity of how these strings fold into 3D structures. A central dogma of structural biology is that sequence informs structure and that structure informs function. Whilst this dogma has been challenged by the developing realization that a significant number of proteins are in fact disordered, this idea remains relevant to highlight that, with such a vast number of potential protein structures there also exists an enormous number of potential protein functions.

To tackle disease an attack vector of modern medicine is to modulate the function of proteins in the human body using small molecules/drugs which bind non-covalently to the protein. In the case of inhibitors the binding free energy, between drug and protein, is a good predictor for the effectiveness of the drug. However, not every drug is capable binding strongly to every protein and the field of drug design is concerned with the development of small molecules which can favourably interact with the specific proteins relevant for a given human disease. Before a drug design effort can begin one must have a good understanding of the clinical or molecular pathology of a disease in order to know what observations should be made to determine the efficacy of a drug. The determination of this pathology is an extremely difficult task, largely outside the scope of this work. What is relevant to this work is generally how this determination is made, as this will inform what information is available to guide a

drug design effort. The next sections will discuss how determining this pathology influences the computational methods of a drug discovery effort, outline the methods used in drug discovery in general and highlight some difficulties within the pipeline of drug discovery.

1.2 Drug Design

1.2.1 Target-based vs. Phenotypic

In modern drug discovery, methodologies fall into one of two categories, target based and phenotypic. Target based methods leverage the advances in genetics and molecular biology to specify a protein/molecular target for the drug discovery effort [1]. Whereas phenotypic methods look at disease related responses to a drug *in vivo* or *ex vivo* [2]. The advantage of a phenotypic method is clear in the case a molecular target for a disease is unknown, poorly understood or difficult to obtain a crystal structure for [3, 4], where the determination of crystal structures is one of the major bottlenecks to target based drug design [5]. In spite of this potential advantage, phenotypic studies may still not perform well if the molecular targets are not well understood. Generally, a poor understanding of a disease will complicate target deconvolution and the determination for the drugs mechanism of action (MOA) later in the drug development process which may be an obstacle to clinical trials [2]. In the example of Alzheimer's disease, where the precise mechanism of amyloid- β oligomers in the progression of Alzheimer's disease is not understood, phenotypic methods have had success, suggesting new possible treatments [6], but have also presented significant difficulty in predicting clinical efficacy of drugs [7].

Target based methods, in contrast to phenotypic, need to have validated a target *a priori*, and as such determining the MOA can be much easier. There are numerous other advantages to a target based approach and they can be cheaper and faster avoiding the more costly *in vivo* and *ex vivo* assays in the early stages of a drugs' development. Some disadvantages to target based methods stem from the abstraction of the problem away from the complexity of full human clinical studies. As such the transferability of target based methods may in general be lower than phenotypic methods [1].

1.2.2 Assays

As compounds are developed throughout a drug discovery campaign there should exist some metric by which progress is measured as these compounds move through the pipeline to become safe and effective drugs. Whilst this metric is a complex multi-valued objective which changes as the campaign advances, we can discuss it usefully in terms of individual

assays for properties of the compounds. In this work, we will divide these assays into two categories those pertaining to the binding affinity of a drug to a protein, and those pertaining to absorption, distribution, metabolism, and excretion (ADME) properties of the drug. The former set of assays are most important here as it is the computation of ligand-protein affinities which is the aim of this work and, as such, these binding assays will be the focus of the discussion here.

The methods which can be used to assay compounds are dependent on what information is available to the study and whether one is using target based or phenotypic methods. In a phenotypic regime, it is common to use cell based methods. As the name suggests, cell based assays use observations of the cell as end points. Examples of some observations which could be made are cell proliferation, motility and morphology amongst other more complex end points [8]. One assay method common in studies where a target is available is to use recombinant methods to create mammalian cells, which over-express the protein target [9]. This protein can then be purified [10] and an affinity measured between a compound and the protein. This affinity can be measured using many methods [11, 12] including: ligand labeled methods, such as fluorescent ligand binding assays [13], ligand unlabeled methods, such as surface plasmon resonance [14] or structure based methods, such as nuclear magnetic resonance, [15] or X-ray crystallography [16]. Predictive computational methods can also be considered as assays for molecules. The variety of methods, which can be used to model protein-ligand interactions, is large and attention will be given to specific methods later in this work. Here we will comment that generally computational methods will estimate the binding free energy of a ligand to a protein or use some approximation to this value in order to assay molecules for affinity.

1.2.3 Hit Identification

In the early stages of a drug discovery effort the focus is to find so-called hit molecules. A hit molecule is defined here as a molecule which has a desired response when tested, where the desired response and nature of the test depend on the assay used. Two examples of hit identification methods commonly used are high throughput screening (HTS), where a large curated library of small molecules are assayed, and focused screening, which is very similar to HTS, but as the name suggests the library of small molecules is focused using knowledge of the target [17]. Fragment screening, although a similar premise to high throughput and focused screening, has a defining difference. This difference is that for fragment screening only the libraries contain weakly binding and light molecules < 300 daltons [18] are used, the goal is to find many hit fragments which interact with the protein and to join or grow these fragments to develop the hit molecule. Whilst HTS can be applied in phenotypic studies

[19, 20] focused and fragment screens fall into the domain of structure based methods as they rely on knowledge of the molecular target. The assay or test used in hit identification can also be computational and this would be referred to as a virtual screen. For example docking models can be used to score ligands in the active site of a protein. These scores are generally based on steric and chemical complementarity between ligand and target [21]. There are numerous variants of these docking algorithms, which range from explicit all-atom or grid-based methods, and can include ligand or receptor flexibility [22].

1.2.4 Hit to Lead

Once a series of hit molecules have been found, effort must be made to develop these into effective therapeutics. A key objective in this development is to increase the efficacy of the drug which is closely linked to the binding affinity of ligand to protein, this is called the hit to lead stage. In this stage the assays we have discussed can be used to test many ideated compounds for activity. From these data a structure activity relationship (SAR) can be built to inform the design of the drug. During this stage, attention is given to the ADME properties of compounds with the intent to avoid later problems with these properties during final drug candidate selection. Again, we do not consider optimization of ADME properties in detail in this work and so it suffices to say, that during the hit to lead stage, assays which measure properties of the compound such as solubility [23] or toxicity [24] are carried out, and the data from these assays contributes alongside activity data to compound development.

The ideas for which compounds should be synthesized and tested can come from many sources. Commonly the intuition of a medicinal chemistry team can be combined with existing data for a project and also computational modeling to suggest molecules to test; a particular interest here is in these computational models. The methods to model the activity of molecules depends, again, on whether this is a target based or phenotypic effort, but now also on the maturity of the campaign and target. That is to say, if there is significant data for the affinity of a compound series against a target then it is more likely that accurate ligand based machine learning (ML) models can be built. There are significant contemporary efforts to develop ligand based methods [25–27]. These ligand based methods remove consideration of the protein binding pocket instead focusing only on the properties of the ligand to make predictions about activity. In the absence of abundant SAR data from a mature campaign, physics based methods can be used instead. In this work, we are particularly interested in atomistic models and free energy methods which allow for the computation of ligand-protein binding affinities. These methods have been widely applied within the drug discovery pipeline with success at predicting potency improving mutations to compounds [28, 29].

There exists a large class of methods called computational *de novo* design programs which aim to grow molecules, one atom or group of atoms at a time to improve their binding affinity. Whilst the scope of these methods can extend outside of the hit to lead process they are worth discussing here as they have the same objective: improve ligand-receptor affinity. Some computational *de novo* design programs emerged as early as 1989, namely HSITE/2D Skeletons [30]. This receptor based algorithm took as input an experimental 3D crystal structure from the Brookhaven Protein Data Bank and would compute a map of potential hydrogen bonding regions. Interestingly, this very early algorithm only used five minutes of processor time on an IBM 3084q mainframe computer operated by the Cambridge University Computing Service from 1982 to 1995 [31]. The idea was that from a map of potential hydrogen bonding regions a ligand with complimentary interactions could then be designed, and so the binding affinity improved.

HSITE/2D Skeletons is a structure-based method, using information about the protein target, in this space of methods there are a number of diverse strategies which have been explored. The commonality in receptor-based methods regards the use of information about the 3D structure of receptors in any calculation for ligand protein affinity. Typically this metric of affinity is described as ‘complementarity’ between the ligand and receptor. This complementarity can be measured in a number of ways for example the LUDI [32, 33] series of programs uses a weighted sum of enthalpic and entropic terms, stemming from ideas such as lipophilic contact, rotatable bond in the ligand etc. The SMOG [34, 35] set of programs use the observed frequencies of atom-atom contacts in the ligand-protein complex. Generally these receptor-based methods will a) use either heuristic [36] or knowledge-based scores [33, 35], or b) use an explicit force field to calculate the potential of the complex and score the ligand [37]. As commented by Schneider *et al.* [38], in their comprehensive review of computational *de novo* design methods, these receptor-based methods aim to approximate the full binding free energy in their scores for affinity. Using approximations to the binding free energy allow these methods to be quite rapid and explore a large area of chemical space, but this can come at the cost of accuracy for any predictions. Several of these computational *de novo* design programs incorporate dynamic structures [39, 40, 37] with the aim of reducing inaccuracy stemming from considering only a static protein structure. Which is to say if only a protein static structure is considered then all physical information about its dynamic structure is not included in the physical model and this will result in a poor model of the protein ligand system for systems where allosteric effects are important [41] or the binding site is a cryptic pocket [42] as examples.

1.2.5 Lead Optimization and Late Stage

The stage following the hit to lead stage is the lead optimization stage. In this stage the aim is to improve the compounds output from the hit to lead stage, particularly any problems with ADME properties must be resolved without sacrificing the affinity which has been developed for the target protein during the hit to lead stage.

The final stages of a drug discovery effort involve rounds of clinical trials. These trials are beyond the scope of this work however it suffices to say that these clinical trials carry with them significant ethical implications [43] in addition to being primary contributors to the cost of drug development [44]. It is therefore imperative that we are confident in the safety and efficacy of any prediction generated in previous stages, which make it as far as a clinical setting.

1.2.6 Chemical Space

The experimental and computational methods described in the Drug Design section can all be applied in aid of searching the chemical space for molecules with some set of desired properties. One of the principal difficulties of operating in chemical space is the cost involved for testing a point in this space. However, arguably a more vexing obstacle to finding useful drugs in chemical space is its large unsearchable size.

By some estimates the number of molecules with 30 atoms containing only C, N, O and S is 10^{60} [45], and this scales combinatorially with the number of atoms. Taking these rules for the number of atoms and type as a rough proxy for drug-likeness we can see that this space of molecules is impractically large to consider searching with brute force methods. The size of this space can be somewhat reduced by considering Lipinski's rule of five [46], which refines the estimate of 10^{60} molecules to look at molecules with drug-likeness and bioavailability. Lipinski et al. developed these rules by looking at a library of several thousand drugs which had reached phase 2 clinical trials, assuming that reaching phase 2 assured that the drug had favorable physico-chemical properties. The observations made about this library were that the majority of molecules had a molecular mass < 500 daltons, calculated partition coefficient ($\log_{10} P$) < 5 , and the sum no greater than 10 nitrogen and oxygen atoms, as such these are the rules by which one can refine chemical space to smaller set, which should be both easier to explore and result in drugs more likely to succeed in clinical trials. There are complications to these rules and they can't be applied generally, with many studies highlighting drugs beyond the rule of five [47], suggesting revised rules for fragment identification [48] and highlighting additional important properties such as number of rotatable bonds [49] or polar

surface area [50]. Even with consideration of these complications the space of drug-like molecules remains large and enumerating it fully is an intractable problem.

Given the size and cost of exploring this space, one could ask: should any hope of rationally designing drugs be abandoned? Despite the vastness of drug-like space, rational drug design has been relatively successful at finding new drugs, aided in part by computational drug design. In order to explore drug-like space with some utility, two core ideas can be used. These are, 1) to find methods which reduce the number of points which need testing and 2) reduce the cost of testing each point in the space.

With regards to point 1), to reduce the number of points which need testing, numerous methods can be used. Some methods use advanced data collection techniques, database building, to collate data from many sources and make it available across groups. This has become a more prevalent technique with the rise of machine learning and natural language processing. However, these methods will not be discussed in detail here and instead predictive methods will be the focus of this work. The predictive methods we have discussed so far in this work all construct a model of reality. These models can range from a medicinal chemist's intuition, to a deep neural network, to classical atomistic simulation. Modelling allows many ideas to be tested and predict if they are beneficial without incurring the full cost of synthesizing and testing the molecules. Most of these modelling methods, however, can only be applied using trial and error. In light of the number of possible drug like chemicals which could be tested, trial and error is not an efficient method with which to explore chemical space. In this work we consider if we can move away from trial and error and reduce the number of points we test in chemical space using numerical optimization methods. The theory relevant to this idea will be discussed in Optimization section 1.6 and applied in chapters 3/4.

Considering point 2), currently the full pipeline of drug discovery is enormously expensive, taking more than a decade and costing more than a billion dollars to produce one drug [51]. In a research setting testing one molecule at the hit to lead stage is estimated to take roughly a week and cost around \$2000. Comparatively using the computational method explored in this work to assess a relatively small mutation such as a methylation would cost 15 hours of computer time equating to roughly \$4. It's clear from these costs that in general computational methods are far cheaper tools with which to explore chemical space. In this work we will aim to develop novel computational methods for drug discovery which improve on the efficiencies of current methods to assess protein-ligand binding affinities and this idea will be explored in detail in all the results chapters.

To address the problem of protein-ligand binding there exist numerous computational methods which populate varying levels of theory and computational cost. In this work we

primarily use free energy methods with molecular dynamics to assess protein-ligand affinity but will also consider some ligand based machine learning methods. With the current level of computational resources available free energy methods and molecular dynamics allow for 10s-100s of ligand bindings to be evaluated on the time scale of days-weeks, and provide good predictive capability for drug design [52]. The motivation to study ligand-protein binding computationally arises from not only the cost savings but also the complexity of these systems. The affinity of ligand to protein is a complex balance of enthalpic and entropic terms in both the bound and unbound states of the ligand, and this balance is very difficult to assess qualitatively with human intuition and predictions for what changes to a ligand will improve binding affinity can often be misleading [53]. Computational free energy methods allow us to assess this balance rigorously providing quantitative predictions for binding affinity. These free energy methods are built on several pillars of theory, and the following sections will address these pillars in detail. To study proteins-ligand binding we first need tools to interrogate the dynamics of these systems one candidate method for such a study would be molecular dynamics.

1.3 Solving Newton's Equations of Motion on a Computer

Molecular dynamics (MD) is at the core of the free energy calculations we will perform in this work. MD relies on solving Newton's equations of motion, for which it is known that analytic solutions do not exist in general, this holds especially true in the case of the complex biological systems considered here. In order to solve Newton's equations numerically, we can first define a three-dimensional space and call this the simulation cell. Typically, the simulation cell is bounded in real space, and therefore a choice is required for the type of spatial boundary conditions; for biological simulations most commonly periodic boundary conditions are chosen. Within this region, we define a system which is composed of many individual objects, indexed by i . These objects are assigned, at least, a position at a point in the region, \mathbf{r}_i , and a mass, m_i . The system as we have defined here would simply be a box of static objects. We can make this system more interesting, however, by adding initial velocities for the objects and interaction potentials between them. At a given time t , the forces on our object can be calculated by taking the gradient of the potential, $U(\mathbf{r})$, in the three dimensional space,

$$m_i \frac{\partial^2 \mathbf{r}_i}{\partial t^2} = \mathbf{F}_i = -\nabla U(\mathbf{r}_i). \quad (1.1)$$

This equation uniquely determines the force on each object, and thus the system can be propagated in time by updating the positions of particles from their velocities, and updating their velocities from their accelerations. It is not practical to solve these equations by hand for any systems of interest in the present context, so instead these equations are solved numerically by a computer.

We are free to use any computational resource, however, the architecture of the computer will inform what system size and simulation length can practically be attained, and will also inform the precise algorithmic steps used to solve Newton's equation. As an example imagine a computer designed to use ternary instead of binary, in these cases the algorithmic steps for addition and multiplication of floating point numbers are not the same and therefore the precise steps to solve Newton's equation on these computers are also not the same. However, we will ignore any considerations for the algorithms' implementation stemming from low level ideas such as variations in computer silicon design and programming languages. Instead we look at the level of a single modern desktop computer where there exists two common resources: the central processing unit (CPU) and graphics processing unit (GPU), both these resources have their own streams of code designed to take advantage of their respective architectures [54–56]. A level above the desktop computer is high performance computing (HPC) centers which, for our purposes, can be conceptualized as hundreds of interconnected desktop machines. Again, at this level there are additional considerations for the computer architecture which must be made to inform algorithm design; here we can examine this idea specifically with reference to parallelization. As an example, the number of instructions a CPU can perform whilst waiting to communicate with another CPU in a HPC is much lower than the number of instructions a GPU could perform whilst waiting to communicate with another GPU. It would therefore make sense to spatially decompose a simulation over CPU based machines in a HPC but not over GPU based machines and we see this borne out in modern GPU and CPU molecular dynamics codes. We can state this more generally by saying that communication between the GPU and any other part of the computer is relatively slow. This idea is important to mention now as it will inform our choices for software design later in the parallelization section 1.8 of this thesis.

1.3.1 Molecular Dynamics

Molecular dynamics is a mature class of computer simulation techniques. In contemporary materials science it leverages the idea of computationally solving Newton's equations to study a wide range of physical systems. Originating in the 1960s the ubiquity of the application of MD has grown over time [57] and MD is now routinely used in the fields of material science [58], biochemistry [59] and biophysics [60]. MD has applications in many types

of simulations, for example coarse grained simulations, but in this work we examine the application of MD, with the use of classical all atom force fields discussed in section 1.3.2, to study the dynamics of a system at an atomistic level. Particularly relevant here is the application of MD to the simulation of proteins, and one of the first simulations in this domain was performed by McCammon *et al.* in 1977 [61] which examined the dynamics of protein folding.

At its core MD uses the ideas we have already discussed about numerically solving Newton's equations. To obtain this numerical solution for a bio-molecular system one would like to integrate forward in time the positions of the atoms. The forward propagation in time of a physical model of a system will be referred to as simulating the system in this work. To perform this simulation we consider atoms as objects in a three dimensional space, with the interaction potentials between the atoms chosen to reproduce the physical properties of the atoms or system we are interested in studying. The choice of potential and its parameterization will be discussed in more detail in the force fields section 1.3.2. To perform the forward propagation of the atom positions in time an integrator, such as the Verlet algorithm, can be used. To reach the equations used in the Verlet algorithm we can consider time series expansions for the position of a particle \mathbf{r}_i at time $t + \Delta t$ and $t - \Delta t$ which gives,

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \frac{\delta \mathbf{r}_i}{\delta t} \Delta t + \frac{1}{2} \frac{\delta^2 \mathbf{r}_i}{\delta t^2} (\Delta t)^2 + \frac{1}{3} \frac{\delta^3 \mathbf{r}_i}{\delta t^3} (\Delta t)^3 + \mathcal{O}(\Delta t)^4 \quad (1.2)$$

$$\mathbf{r}_i(t - \Delta t) = \mathbf{r}_i(t) - \frac{\delta \mathbf{r}_i}{\delta t} \Delta t + \frac{1}{2} \frac{\delta^2 \mathbf{r}_i}{\delta t^2} (\Delta t)^2 - \frac{1}{3} \frac{\delta^3 \mathbf{r}_i}{\delta t^3} (\Delta t)^3 + \mathcal{O}(\Delta t)^4 \quad (1.3)$$

and combining these equations yields the update formula,

$$\mathbf{r}_i(t + \Delta t) = -\mathbf{r}_i(t - \Delta t) + 2\mathbf{r}_i(t) + \frac{\delta^2 \mathbf{r}_i}{\delta t^2} (\Delta t)^2 + \mathcal{O}(\Delta t)^4 \quad (1.4)$$

It can be seen here that when using the Verlet algorithm the positions are updated with an accuracy of $\mathcal{O}(\Delta t)^4$. There exist numerous algorithms with improved accuracy when compared to the Verlet algorithm [62]. An important point to consider is that typical binding free energy calculations considered in this work are dominated by statistical error [62, 63], dependant on the amount of sampling performed, and so errors associated with the integration method are less important. In this work the integrator used is a variant of the Verlet algorithm

called the leapfrog scheme. These integrators are highly similar and share the favourable properties such as being symplectic and time reversible which practically translates into accuracy and stability of the integration after longer times [64, 65].

In this work the simulations are performed using Langevin dynamics. Langevin dynamics add a friction and random force to the systematic forces present in molecular dynamics. The origin of these forces is phenomenological but their inclusion is physically motivated in order to couple the simulated system to a heat bath [66]. Heat from the system can be exchanged with this bath allowing for the temperature of the system to be controlled. The reason one might want to perform MD simulations at a constant temperature is to perform a simulation in a specific thermodynamic ensemble. For example this might be the canonical (NVT) ensemble, in this ensemble the number of particles N , the volume of the system V and the temperature T are all held constant. Another ensemble that could be used is the isothermal-isobaric ensemble (NPT) ensemble, where there is a constant number of particles, N constant pressure P and constant temperature T . The ensemble used in a simulation depends on the ensemble of the system the user wishes to study, in this work the NVT or NPT ensembles will be used.

The idea of different thermodynamic ensembles in which T or P could be held constant has been discussed, to perform a simulation in these ensembles the idea of thermostats and barostats must be introduced. A simple example of a stochastic thermostat [67] would be the Andersen thermostat and the basic idea here is that the velocity of random particles will be reassigned a new velocity from the Maxwell distribution at temperature T . A note here should be that stochastic thermostats should not be used to interrogate the dynamic properties of a system, such as diffusion, and this is because the velocity re-scaling, for example, influences the dynamics of the system. In this work we use a stochastic thermostat which is included as part of Langevin dynamics.

In the domain of barostats there are many methods that could be used [68, 69]. The aim of any barostat is to control the pressure of the simulation and this can be achieved typically by adjusting the size of the simulation cell and re-scaling the atomic positions in this cell. Just because a barostat controls the pressure it does not guarantee that one is sampling in the correct ensemble and care should be taken in choosing a barostat. For example the Berendsen [69] barostat is known to not sample correctly the expected volume fluctuations in the isothermal-isobaric ensemble [70] with evidence this effects the results of free energy calculations [71]. In this work we use the Monte Carlo barostat which resizes the simulation and accepts or rejects this volume change based on the Metropolis algorithm [68].

Applying MD as described is referred to as vanilla MD and will allow us to simulate a biomolecular system. From this simulation we can calculate time average macrostate

properties such as temperature, or most relevant here, the free energy. This time averaged property is calculated from a time series of snapshots for the microstate of the system. To do this calculation we use the ergodic hypothesis which states that, a property averaged over many repeats of some experiment will be equal to the time average of the same property, in the limit that the number of repeats and the length of time are both large. This hypothesis relies on the assumption that if the system is propagated forward indefinitely every state of the system will be visited. Precisely how the free energy can be calculated will be considered in detail in the free energy section 1.4.

One drawback of vanilla MD is that we have no control over what states are sampled. Therefore, if a study is interested in an event which occurs with a low probability (with the probability of a state given by the Boltzmann distribution, see Section 1.4) then a large amount of sampling may be required in order to capture the event. If the potential energy landscape of a biomolecular system is imagined as many low energy configurations separated by high energy barriers, meaning that crossing the high energy barriers is a rare event, then, to collect comprehensive sampling for all the configurations of the systems with brute force MD will be unfeasible. As an aside these minimums in the free energy landscape of a protein for example, exist to both guide the folding [72] and insure the kinetic stability of the protein's folded state, possibly allowing it to maintain its native state and physiological function for longer [73]. Adequate sampling of the configurational space of a system is key to our assumption of ergodicity and will be crucial to calculating accurate free energy differences in later chapters. As such it would be beneficial if the inefficiency in MD could be addressed to allow these low energy configurations to be sampled effectively.

The class of methods that can accelerate MD exploration of configurational space is referred to as enhanced sampling. There are several methods which can be used to accelerate the sampling of configurational space and generally these methods might look like: adding a biasing potential to the potential energy landscape to reduce the height of the high energy barriers, making them easier to cross (umbrella sampling) [74], or using the temperature of the system to accelerate jumps over the barriers (simulated tempering) [75, 76] by proposing temperature changes to the system which can be accepted/rejected based on a Metropolis criteria.

One extension to the idea of using temperature to accelerate sampling is parallel tempering also known as replica exchange [77, 78]. These methods involve running many replicas of a simulation in parallel at different temperatures. One cold replica is run at the studied temperature and so unbiased sampling can be collected from this replica, whilst the other higher temperature replicas exist to accelerate sampling. The configurations of the hot replicas are exchanged with the configurations of the cold replica based on a, Monte Carlo

acceptance criteria [79]. It has been demonstrated in the literature that, using the same set thermodynamics states, simulated tempering can achieve higher exchange rates than replica exchange [80]. One advantage of replica exchange is that it is relatively trivial to parallelize the many replicas over a HPC. A limitation of replica exchange methods is that a large number of replicas are needed to efficiently sample a system with a large number of degrees of freedom [81]. With too few replicas the phase space overlap of adjacent replicas will be small resulting in a low acceptance rate for configuration exchanging moves, and as such the exploration of phase space have less efficiency compared to a high exchange rate. To address this problem the methodology of replica exchange can be further extended to allow for the Hamiltonian of the system to be modified. These methods which modify the Hamiltonian are called Hamiltonian replica exchange (HREX) [82] methods and they give far more freedom for what aspects of the system can be modified, beyond changing the temperature which we have just seen for the tempering methods. As an example, replica exchange with solute tempering, also known as REST [81], effectively allows for 'heating' of only the solute molecule in a large protein-ligand system and this avoids addressing a large number of degrees of freedom, thus reducing the number of replicas required. Later in this work, chapter 4, steric optimization, the OpenMMTools [83] implementation of HREX will be used to perform a set of relative free energy calculations.

1.3.2 Force Fields

In the discussion of MD we considered the idea of modeling the interactions of atoms using potential energy functions. In this section the theory of these potential energy functions will be explored in the context of all atom force fields (FFs). In an all atom FF one atom is represented by one object in the simulation. Alternative methods, not discussed here, exist which involve mapping many atoms onto one simulation object; these are referred to as coarse grained [84].

The choice for a functional form to represent the interactions of an atom is free. However, it is typical to decompose a system into interaction types and to choose one type of potential to represent each type of interaction in the system. For example, a harmonic well can be used to describe the bond interactions between atoms. To capture the variation in the character of interactions between different atom types, the function is chosen to include parameters which take different values to describe different atoms. The choice for these parameters can be made by a fitting procedure with the objective of minimizing an error function. An example of such an error function would be the difference between some computed and experimental property of the system. A FF is then a collection of the potential energy functions and the values which parameterize all relevant atom types. The potentials chosen are all implicit or

explicit functions of the atomic positions and therefore the force field can take as input the position and type of atoms in a system and output a potential energy. The partial derivative of the potentials in the FF with respect to the position of the atoms can be used in our description of Newton's equation to perform an MD simulation.

For biomolecular simulation there are a set of common interaction types into which the system is decomposed, and these interactions are represented with a common set of functions. The potentials normally considered are divided into non-bonded and bonded potentials. The bonded potentials can be subdivided into bond, angle and torsion potentials. The bonded potentials are relatively cheap to compute and an example for the functional forms of these potential is given as follows,

$$U_{bonded} = \sum_{bonds} k_b(r - r_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{torsions} k_\phi(1 + \cos(p\phi - \phi_0)), \quad (1.5)$$

here r is the separation of two atoms in a bond, θ is the angle between three atoms, and ϕ is the dihedral angle between four atoms. r_0 , θ_0 and ϕ_0 describe the equilibrium position of the bond, angle and torsion respectively. k_b , k_θ and k_ϕ are spring constants which describe the stiffness of the bond, angle and torsion respectively. The parameter p defines the periodicity of the torsion. In general more than one torsion term will contribute to the bonded potential. To parameterize the bonded potentials, as outlined here, a choice must be made for seven parameters.

The non-bonded interactions are much more expensive, compared to the bonded, to compute, and in an absolute worst case scenario they would need to be computed in a pair-wise fashion between all N atoms in the system. As such the computational cost of computing the non-bonded terms scales like N^2 whereas the bonded terms will scale like only N . The non-bonded interactions can be decomposed further into two interaction types. These are the Lennard-Jones and electrostatic interactions. The steric component and attractive components of the non-bonded is typically represented with a Lennard-Jones (LJ) potential, the electrostatics are commonly represented using a Coulomb potential.

$$U_{nonbonded} = \sum_{j=1}^{N-1} \sum_{i=j+1}^N \left\{ \epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\} \quad (1.6)$$

Here ϵ_{ij} is a parameter which defines the well depth of the LJ between atoms i and j , σ_{ij} defines the position for the minimum of the well in the LJ potential between atoms i and j , q_i and q_j define the charges of atoms i and j . ϵ_{ij} and σ_{ij} are computed by combining the ϵ or σ for atoms i and j through a combination rule which varies between FFs; an

example of a combination rule would be the Lorentz-Berthelot rule [85] which takes the arithmetic and geometric means of sigma and epsilon parameters respectively. Typically in biomolecular simulations it is best practice to compute the electrostatics using particle mesh-Ewald (PME) [86]. PME is an improvement to the Ewald summation [87] method for computing electrostatics and these two methods allow for the long range component of the electrostatic potential to be included in a calculation without the expense of computing all pairwise electrostatic interactions. Ewald methods allow for the faster computation of the long range component to the electrostatics by splitting the interactions into a short-range contribution, and a long-range contribution. This short range contribution can be evaluated in real space and the long-range in Fourier space. Both these summations converge quickly in their respective spaces (real and Fourier) and therefore a truncation may be made with little loss of accuracy but a computational time saving.

The interactions detailed here are approximations to the true interaction potential between atoms. A significant approximation pertains to the polarizability of atoms. Here we have assumed that each atom has a fixed partial charge. A more accurate description of the partial charge is that it is a function of the atomic environment of an atom. There exist classes of polarizable FFs for MD [88, 89], however, due to their increased computational expense these force fields are rarely used for simulation of proteins. Recent studies have shown that polarizable FFs may provide significant utility for describing the internal electric fields of enzymes [90] but these validation studies are rare due to the aforementioned computational cost of simulating with polarizable FFs.

Key to the study of protein-ligand interactions is the accuracy of the FFs used. Improving the accuracy of these FFs has been a major focus of work from many groups. In the space of biomolecular simulations there are a number of FFs which have become ubiquitous and are periodically updated with the aim of improving their accuracy [91–96]. The most widely used of these FFs for studies of proteins are the AMBER [91], CHARMM [94], GROMOS [97] and OPLS [98] FFs. The variation in these FFs stems from the precise inter and intra molecular potentials used and the parameterization of those potentials. Some notable differences between FFs are for the empirical scaling of 1-4 interactions, the treatment of improper dihedrals, the inclusion of additional potentials such as the Urey-Bradley angle terms and the previously mentioned combination rules.

One difficulty in modelling protein-ligand systems is for the parameterization of the small molecules that bind to the protein. The problem is that the more generic atom types seen in proteins may not accurately describe the more unique chemistry and atoms that can be generated in small molecules. The AMBER GAFF [99] FF provides a solution to this problem and allows for the generation of bespoke parameters for small molecules.

Another important consideration for protein simulations, not yet discussed, is water models. Water molecules make up the majority of the atoms in a protein ligand simulation and the water model used will significantly impact the structures of the protein that will be sampled [100, 101]. For biomolecular simulation the most common all-atom water models used are three point water models. These treat each atom in the water molecule with one interaction site. Commonly used models are TIP3P [102], SPC [103], TIP3P-Ew [104], SPC-E [105]. There exists more accurate models which can better capture hydrogen bond angles and the thermodynamic properties of water such as TIP4P/2005 [106], TIP4P-Ew [107] and TIP5P [108]. These models include additional sites designed to capture the physics of the oxygen lone pair. Whilst these models are more accurate, they are used less commonly in favour of the faster three-point models. In order to achieve good simulation accuracy, some meta considerations for the choice of water model would be, 1) choose a water model that was parameterized with the protein FF one has selected, and 2) use a FF which was parameterized with broadly the same settings and thermodynamic conditions one intends to use for the simulation.

The previous two sections have outlined the theory for the application of molecular dynamics and biological force fields to simulate protein ligand systems. We have seen in the Molecular Dynamics section that from these simulations we can collect a time series of snapshots for the microstate of the system and use this to calculate time average macrostate properties. An important property for this work is the free energy and the following sections will discuss how this free energy can be calculated.

1.4 Free Energy

In this section the theory for the Zwanzig equation [109] will be outlined. This equation is central to the calculations performed in this work for free energy differences. When describing a thermodynamic system, one can use the idea of a thermodynamic ensemble and we have previously discussed the NVT and NPT ensembles. Within an ensemble we can define a thermodynamic potential that will be minimized when the system is at equilibrium. This minimization comes from the principle of minimum energy and more fundamentally the principle of maximum entropy [110] which states 'the equilibrium value of any unconstrained internal parameter is such as to maximize the entropy for the given value of the total internal energy'. For the canonical ensemble this potential is the Helmholtz free energy,

$$A = U - TS, \tag{1.7}$$

where U is the internal energy, T the temperature and S the entropy.

One may wish to perform experiments in an environment where pressure and temperature are held constant. Such an environment is called the isobaric-isothermal or NPT ensemble. The thermodynamic potential associated with this ensemble is the Gibbs free energy defined as,

$$G = U + PV - TS, \quad (1.8)$$

where P is the pressure and V the volume. For a system held at constant pressure and temperature G will be minimized at equilibrium. It is often easier to deal with changes or differentials in these potentials thus, to exploit this, equation 1.8 can be rewritten as,

$$dG = dU + PdV - TdS. \quad (1.9)$$

We can see from equations 1.8 and 1.9 that for a calculation of G or dG we would need to know P , V and T and calculate U and S . To calculate U and S we would need a lot of information about the partition function of a system. If we know the partition function in full and how it varies with temperature, then we can calculate U and S and use these to calculate a value for G using equation 1.8. Problems arise with trying to calculate G like this, and as enumerating all states and their energies to get the partition function for complex systems, such as protein-ligand systems is not possible analytically, a different approach is needed. We can explore this different approach if first four definitions are made. Operating in the canonical ensemble for simplicity, the first definition we make is for the continuous Gibbs entropy as follows,

$$S = -k_b \int p(\vec{q}) \log p(\vec{q}) d\vec{q}, \quad (1.10)$$

where k_b is the Boltzmann constant and $p(\vec{q})$ is the probability of a system being in some state with a configuration and momentum in phase space \vec{q} . The second thing we need to define is the Boltzmann distribution which tells us the probability of being in the state defined by \vec{q} if we know the energy of that state $U(\vec{q})$,

$$p(\vec{q}) = \frac{1}{Z} e^{-\beta U(\vec{q})}. \quad (1.11)$$

Here, Z is the configuration integral and we are defining $\beta = 1/k_b T$. A constraint on any problem is that all probabilities $p(\vec{q})$ sum to unity and, therefore, equation 1.11 also allows

us to define the configuration integral as,

$$Z = \int e^{-\beta U(\vec{q})} d\vec{q}. \quad (1.12)$$

We are free to use the configuration integral here, in place of the full partition function, because the equations we are deriving are for a free energy differences, and the difference in free energies we are calculating will be between states with equal mass and thus the kinetic contribution to the partition function will cancel [111, 112]. The final formula we need to define is for the expectation of an observable as,

$$\langle O \rangle = \int O P(\vec{q}) d\vec{q}, \quad (1.13)$$

the angled brackets here denote an expectation value or an ensemble average. Substituting equation 1.11 into 1.10 and rearranging using equation 1.13 gives a formula for the entropy in terms of the configuration integral,

$$S = -k_b \frac{1}{Z} \int e^{-\beta U(\vec{q})} \log\left(\frac{1}{Z} e^{-\beta U(\vec{q})}\right) d\vec{q}, \quad (1.14)$$

$$= -k_b \frac{1}{Z} \int e^{-\beta U(\vec{q})} (-\log Z - \beta U(\vec{q})) d\vec{q}, \quad (1.15)$$

$$= k_b \log Z + \frac{U}{T}. \quad (1.16)$$

Thus, with the entropy written like this, we can relate the Helmholtz free energy and configuration integral as,

$$A = U - TS = -k_b T \log Z. \quad (1.17)$$

With a free energy written in this form we are still not any closer to being able to calculate A for complex systems as we still rely on the configuration integral. However, writing A like this makes it easier to follow the steps which will provide the free energy in a more easily computable form. As mentioned above we are mostly calculating differences in free energies and formulating A as ΔA here will offer a path forwards,

$$\Delta A = A_1 - A_0 = -k_b T \log Z_1 / Z_0; \quad (1.18)$$

here Z_1 and Z_0 are the configuration integrals for some systems 0 and 1 and the free energy difference between them is ΔA . We can then make this more explicit using the definition for

the configuration integral,

$$\Delta A = -k_b T \log \frac{\int e^{-\beta U_1(\vec{q})} d\vec{q}}{\int e^{-\beta U_0(\vec{q})} d\vec{q}}, \quad (1.19)$$

we can then apply a numerical trick and multiply the numerator of the logarithm by 1 rewritten as $e^{-\beta U_0(\vec{q})} e^{+\beta U_0(\vec{q})}$,

$$\Delta A = -k_b T \log \frac{\int e^{-\beta U_1(\vec{q})} e^{-\beta U_0(\vec{q})} e^{+\beta U_0(\vec{q})} d\vec{q}}{\int e^{-\beta U_0(\vec{q})} d\vec{q}}. \quad (1.20)$$

Writing the free energy like this allows us to use equation 1.13 to identify the argument of the log as the ensemble average of $e^{-\beta(U_1-U_0)}$ which allows us to reach the Zwanzig equation,

$$\Delta A = -k_b T \log \langle e^{-\beta(U_1-U_0)} \rangle_0. \quad (1.21)$$

The 0 subscript on the ensemble average denotes that this is an ensemble average in the 0 state. This formulation of the free energy removes the need for the configuration integral of a system to be computable and using the Zwanzig equation it is possible to calculate free energy changes using the sampling that can be collected from MD simulation.

1.4.1 Free Energy Methods

With the theory for the Zwanzig equation outlined in the previous section, we will now discuss the practical application of this equation – that is, to compute free energies – and the surrounding bodies of theory. In the space of methods which allow for protein-ligand binding affinities to be computed free energy methods have gained significant popularity. The relevant mathematics for free energy methods, namely the Kirkwood and Zwanzig equations can be credited to Lev Landau, John Kirkwood and Robert Zwanzig and originated in the middle of the last century. Methods based on the Kirkwood and Zwanzig equations are thermodynamic integration (TI) and exponential averaging (EXP) respectively.

Equation 1.21 is the Zwanzig equation and is the central equation in the EXP method. From this equation we can extract the core ideas of the EXP methods as well as the core ideas of the broader set of all free energy methods. Equation 1.21 shows us that sampling should be collected for some system 0, and using this sampling the potential should be calculated in systems 0 and 1. As an example, to calculate the potential U_1 in equation 1.21 a time series of configurations can be collected with MD using the Hamiltonian of system 0, this time series is then post processed with the potential being calculated using the Hamiltonian of system 1, and this gives U_1 . The different Hamiltonians which must be adopted to perform

the free energy calculation are generally all built as one Hamiltonian with a controlling parameter λ which turns on and off the relevant interactions to switch between Hamiltonians. If λ took the value of 0 for example the Hamiltonian would be that of system 0. Typically, in these simulations there are more than just the end states, 0 and 1, and so λ takes many values between 0 and 1. The different values λ takes between 0 and 1 stratify this simulation and by convention the different λ values and associated strata are called states or the λ windows of the simulation. To answer the question of why these λ windows are introduced we should consider that the EXP method is a perturbative one, formally exact only in the limit of infinite sampling. For finite sampling if the difference between states 0 and 1 is large then the calculated free energy will take a long time to converge and this is where the idea of intermediate λ windows emerges, designed to reduce the difference between adjacent states [113, 114].

TI in contrast to EXP is not perturbative, TI methods exist parallel to methods based on the Zwanzig equations. Whilst we do not use any TI based methods in this work it is an equally influential set of methods in the space of free energy calculations. We define the Kirkwood equation, used in the TI methods, here as,

$$\Delta A = \int \left\langle \frac{\delta U(\lambda)}{\delta \lambda} \right\rangle_{\lambda} d\lambda. \quad (1.22)$$

We will mention here that TI based methods have several advantages, with the literature demonstrating their accuracy to be equivalent to the best perturbative methods [71], but with much less computation required to perform the analysis [115]. Some ideas to consider with TI methods is that they rely on the calculation of $\frac{\delta U}{\delta \lambda}$. If an MD code does not calculate this quantity internally this can add a layer of difficulty to using TI instead of EXP, because the calculation of $\frac{\delta U}{\delta \lambda}$ is not as straightforward as the calculation of U . Additionally TI methods must still make use of λ windows, this is to ensure an accurate calculation for the numerical integration of $\frac{\delta U}{\delta \lambda}$ over λ .

Returning to the ideas of perturbative methods. One drawback here is that for a free energy calculation, lengthy MD simulations must be performed for all end and intermediate λ windows. If, however, the perturbation between end states remains small enough, such that the phase space overlap between end states is large, the Zwanzig equation and the EXP method is sufficient without any intermediate states. Assessing of the overlap between adjacent λ windows is a difficult task and normally relies on checking several metrics these would be: 1), the convergence of the computed free energy with respect to sampling, with respect to the number of lambda windows and with respect to repeats of the simulation, 2) computing an overlap matrix [116] which details the probability of observing a sample from

λ window adjacent in λ windows and gives a measure of the phase space overlap of these windows, and 3) using several methods in cooperation with each other. For point 3) here one could perform the free energy calculation using EXP and TI if the answers from both methods are well agreed this is a good indication the free energy calculation is converged. Applying the Zwanzig with no intermediate states is referred to as single step perturbation (SSP) and this is a free energy method used multiple times throughout this thesis to achieve rapid free energy estimates. Numerous studies have used SSP [117–120], demonstrating that it is applicable to relative free energy calculations [121, 122] and can be significantly faster than standard FEP [123].

In addition to adding intermediate states the other way in which perturbative methods can be modified to evaluate larger perturbations efficiently is to move away from older methods such as EXP to more contemporary methods. Building on the EXP method there are two estimators: the Bennet acceptance ratio (BAR) [124] and the multistate Bennet acceptance ratio (MBAR) [114]. BAR is a set of self consistent equations which takes sampling from the two end states of a perturbation to calculate the free energy change between these states. It has been demonstrated in the literature that BAR is a more efficient estimator than EXP when applied to realistic atomistic simulations [113]. MBAR is a generalization of the BAR method which can use sampling from any number of states. We will apply MBAR multiple times in this work to different relative free energy calculations. In order to perform these calculations the end states of the perturbation must be specified. The intermediate states will be built by linearly interpolating any parameters of the system which are changing between the two end states¹. Sampling can then be collected from all intermediate states and combined using the MBAR estimator [114].

The choice for the type of perturbation we apply gives rise to two streams of methods; these are alchemical or geometrical methods. Geometrical methods exist parallel to alchemical methods and can allow one to look at kinetics of processes, but these methods are not used or discussed in this work. Jorgensen *et al.* were the first to apply alchemical free energy perturbation (FEP) methods to a chemical transformation in 1985 [125]. The application was made to calculate the hydration free energy difference between methanol and ethane. This involves creating a perturbation to the system by ‘turning off’ the atoms in a hydroxyl group while ‘turning on’ the atoms in a methyl group. The idea of ‘turning on/off’ atoms refers to scaling the potential energy function associated with these atoms with the controlling λ parameter. This concept is often referred to as ‘exploiting the malleability of the potential energy function’.

¹For absolute free energy calculations it is necessary to change charge and LJ parameters separately to avoid instability in the simulation arising from unshielded charges.

Within alchemical methods there are two types of calculation which can be performed; these are absolute and relative calculation. Absolute refers to turning off an atom or molecule in its entirety, this is generally a large perturbation which requires many λ windows and significant sampling to converge. For protein ligand binding it is more common to use a relative approach to transform one ligand into another and this often reduces to a transformation of one chemical group into another, as was seen in the early example of Jorgensen's work with the methanol to ethane transformation. Relative methods reduce the size of the perturbation and eliminate problems such as the 'floating ligand' problem² thereby reducing the complexity of the thermodynamic cycle required. The limitation for relative methods is apparent if the maximum common substructure (MCS) of the two ligands is small or zero, which would be the case if one wanted to hop from one drug series to another.

For relative calculations numerous topology schemes can be used. In the context of this work this refers to how we transform one ligand into another, and, for clarity, we define the common methods explicitly. There are three methods commonly used, single, dual and hybrid topology. Here we will define the single topology approach as one where MCS of two ligands does not respect atom element type. Therefore during the alchemical transformation atoms are free to change their element. Any uncommon atoms are treated with dummy atoms. A dual topology approach is defined as one in which the MCS of two ligands does respect atom element type. Therefore atoms cannot change their element during the alchemical transformation and all atoms involved in the transformation will be uncommon and must be added as dummies. A hybrid topology is a special case of a dual topology where whilst atoms can't change elements they are free to change type, so the carbon in C-F can change its parameters to become a C-H without the use of a dummy. Figure 1.1 shows the transformation between a chlorine atom and OH group using the a) single, b) dual and c) hybrid topologies diagrammatically.

We have seen the various methods by which one can calculate free energy change including absolute and relative free energy changes. The question remains, how can these methods be applied to calculate ligand binding affinities. In this work we mainly investigate relative binding free energies between two highly related ligands named here as ligands A and B. This relative binding free energy, $\Delta\Delta G_{binding}$, can be constructed in two ways, 1) as the difference between $\Delta G_{binding}$ of ligand A and B, where $\Delta G_{binding}$ is defined as the free energy change associated with the binding event of a ligand to the protein or 2) as the difference between $\Delta G_{mutation}$ in the bound and unbound states, where $\Delta G_{mutation}$ is defined as the free energy difference of mutating one ligand into the other. These two constructions are

²The floating ligand problem occurs in absolute binding free energy calculations when a weakly or non interacting ligand is free to 'float around' and explore the entire volume of a simulation space, resulting in poor convergence of the free energy calculation.

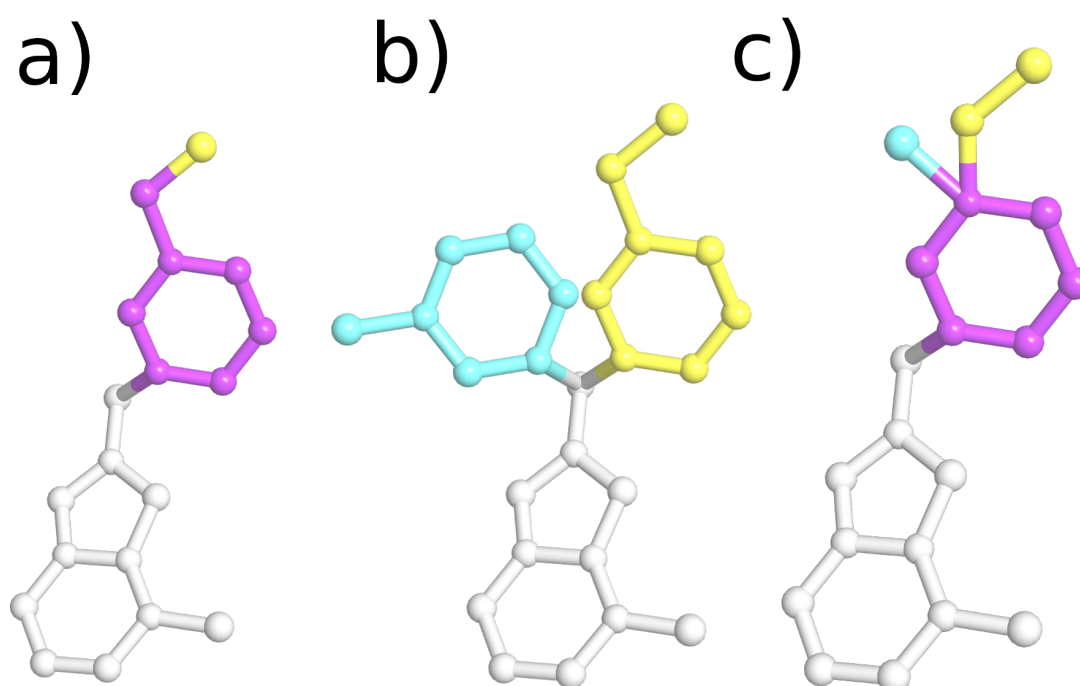


Fig. 1.1 Graphical depiction of topology schemes commonly used in alchemical free energy methods in an example transformation of a chlorine atom to an OH group. a) single, b) dual and c) hybrid topologies. Hydrogen atoms on carbons are omitted for clarity. Parameters of cyan and yellow atoms are fixed, magenta atoms are free to change their parameters.

theoretically equivalent but the use of $\Delta G_{\text{mutation}}$ has better convergence and is standard best practice in relative binding free-energy calculations [126] and as such we use the difference $\Delta G_{\text{mutation}}$ to calculate $\Delta\Delta G_{\text{binding}}$ in this work. See figure 1.2 for a graphical depiction of the thermodynamic cycle used. Whilst the legs of this cycle are combined in the figure, in practise each leg is a separate simulation.

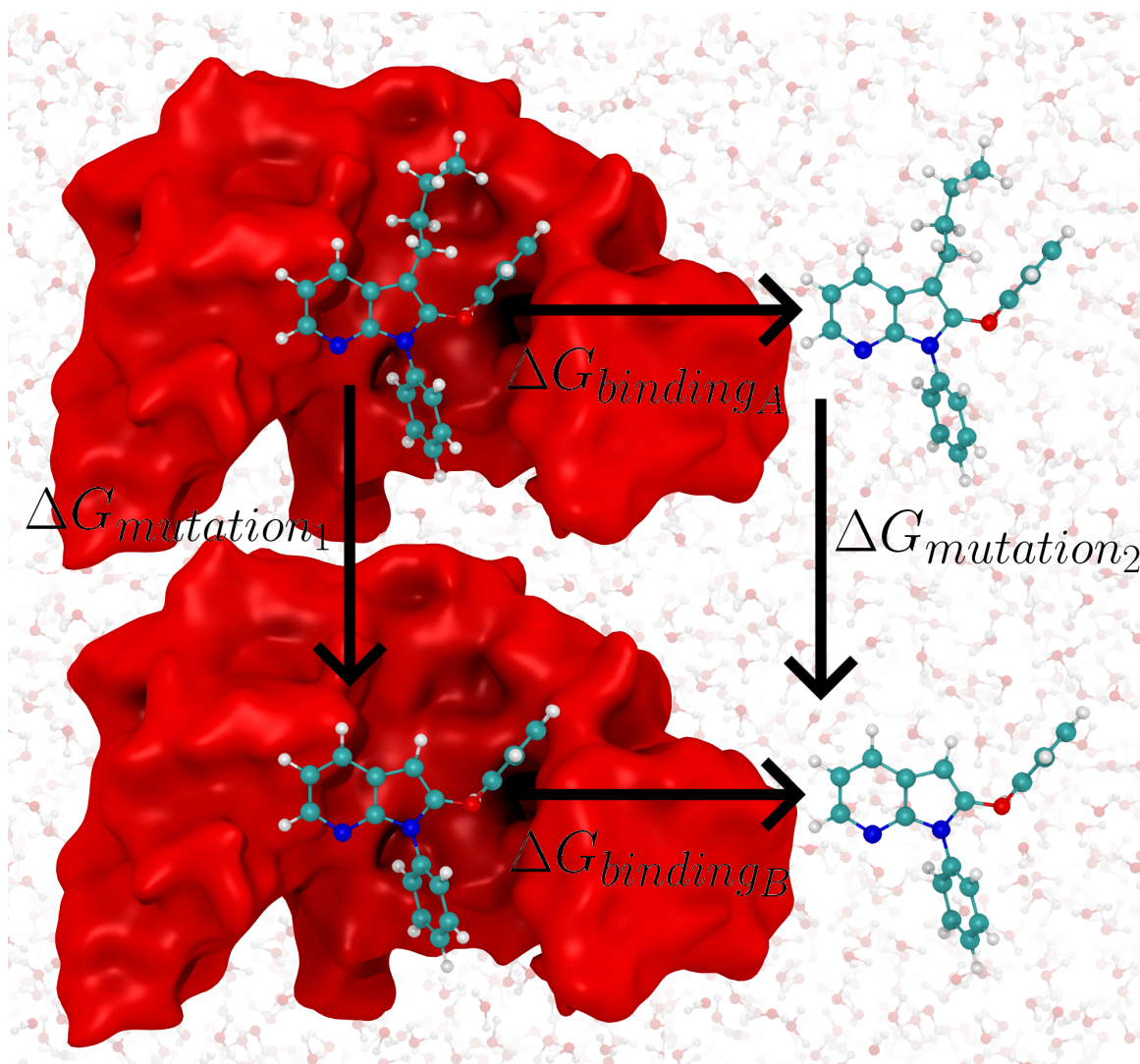


Fig. 1.2 Graphical depiction of the thermodynamic cycle used throughout this work to calculate relative binding free energies between ligands. $\Delta G_{\text{binding}A/B}$ are the binding free energies associated with the binding event of ligand A/B respectively. $\Delta G_{\text{mutation}1}$ is the binding free energy associated with the mutation from ligand A to B in the bound state. $\Delta G_{\text{mutation}2}$ is the same mutation in the unbound state. Red surface depicts the surface of a protein that the ligand is bound to.

1.5 Ligand Binding

We have defined one important metric for the binding affinity of a ligand to a protein and this is the binding free energy. In the literature however there exist several other metrics by which this affinity can be assessed. These are binding constants and IC50s. An IC50 is an experimental measure and proxy for binding affinity, it is used commonly in drug design setting to cheaply (relative to isothermal titration calorimetry) to assess binding affinities of small molecules to proteins, more detail on IC50s is provided later in this section. In the case of a ligand, L , binding to a protein, P , to form a complex, PL , we can consider the following reversible reaction.



The forward reaction described in this relation is the ligand binding to the protein and the backwards reaction is the unbinding event. The forward and backwards reactions have associated rates, these are $k_{forwards}$ and $k_{backwards}$ and their ratio give a binding constant called the association constant K_a ,

$$K_a = \frac{k_{forwards}}{k_{backwards}} = \frac{[PL]}{[L][P]}; \quad (1.24)$$

the square brackets here, $[]$, represent a concentration. This K_a is easily related to the binding free energy ΔG as follows,

$$\Delta G = -RT \log K_a, \quad (1.25)$$

here R is the gas constant. If we have access to either the binding free energy or association constant then affinity between the protein and ligand is well quantified.

In this work the binding events we will consider are the binding of an inhibitor to an enzyme. An enzyme is a particular class of protein which catalyses some reaction on a substrate to produce a product. Both the substrate and inhibitor may bind to the protein and so this complicates our assessment of the affinity. An inhibitor can bind to a protein either competitively, uncompetitively or noncompetitively. Competitive means that the inhibitor and substrate bind to the same location on the enzyme, uncompetitive means that the inhibitor only binds to the complex of substrate and enzyme and noncompetitive means that the inhibitor binds to a different site on the enzyme than the substrate.

To measure the affinity of an inhibitor to an enzyme experimentally typically an IC50 or K_i value is used and these two quantities are related. The experimental techniques used to

measure these quantities vary both in methodology and expense with IC50 being the cheaper [127] and so the most common measurement seen in the literature. An IC50 is the inhibitor concentration at which the remaining activity of the enzyme is half. The K_i is a binding constant called the inhibition constant. K_i represents the ratio of the binding and unbinding rate for inhibitor and enzyme. K_i can be converted to a binding free energy similar to K_a in equation 1.25.

The precise mechanism for these measurements of IC50 or K_i is not important here, however, we are interested in the relationship between these quantities. To calculate a K_i one can measure the rate of the enzyme-catalyzed reaction whilst independently varying the concentration of the substrate [S] and the concentration of the inhibitor. To calculate an IC50 one would do the same but at a constant [S]. From this, we can see why calculating the IC50 is cheaper since only one scan over the inhibitors concentration is needed to calculate the IC50 value. To convert IC50 to K_i the [S] and the Michaelis constant K_m are needed for competitive and uncompetitive binding. Where K_m is the concentration of the substrate for which the rate of reaction is half its potential maximum. The desire to convert these quantities stems from a drive to compare any computational binding affinities to experiment. However, a problem can arise here when converting IC50 to K_i as [S] and K_m are rarely reported in the literature. To avoid this problem, K_i can be approximated as the IC50 if [S] is close to zero for competitive inhibitors and K_i can be approximated as IC50 if [S] is very large for uncompetitive inhibitors. For a noncompetitive inhibitor it is true that the IC50 and K_i are equal without approximation. Throughout this work many values of experimental binding free energy will be presented, roughly half are derived from K_i values the remaining half come from IC50s. None of the IC50 conversions are corrected for [S] for the reasons outlined here. Two more measures of affinity are commonly seen in the literature. These are pK or pIC50 where the p here means that either the binding constant or IC50 value is the argument of $-\log_{10}$.

1.6 Optimization

Using binding free energies and IC50 values a quantitative assessment can be made for affinity of a protein to a ligand. A central goal of drug discovery is to be able to select a drug from the set of all drugs which is assessed to have a favourable binding affinity to a given protein target. We have discussed, however, there is a significant problem with this selection process stemming from the huge number of potential drugs which exist. Many problems in science involve finding a particular element of a large space impracticable to search exhaustively. For these problems it is rare to restrict the exploration of this space

to the brute force testing of points and application of numerical optimization methods can yield solutions much more efficiently. For any continuous optimization problem an objective function $f(x)$ and the domain x of this objective ideally would be specified as follows.

The objective function is ideally:

1. A locally convex function, i.e. a (locally) unique minimum exists, with a corresponding unique minimiser.
2. A continuous function of its arguments, i.e a small change in x gives a small change in $f(x)$ for all x .
3. Twice differentiable so the Jacobian and Hessian can be calculated.

The domain x would ideally:

4. Consist of real numbers.
5. Failing this x should be part of a convex set, i.e. any linear combination for allowed values of x are also allowed .
6. If not convex the values of x which are allowed should be easily checkable.

Additionally we might want: a) to limit the dimensionality of the domain to be small, b) to have a good starting position in the domain and c) to have a reasonable idea that the objective is not full of local minima.

Building the objective like this allows numerical optimization techniques to be applied more effectively, the problem of rational drug design as a whole is extremely difficult to map onto an optimization problem which adheres to the above. This is partly because the full pipeline of drug discovery has a very complex multi-valued objective. However, even if this full pipeline is cut down into more tangible pieces, difficulties can still arise. Principally, one of the problems pertains to how the ligand binding problem can be modelled in a way that is amenable to numerical optimization. For example, molecules could be featurized into a set of continuous real numbers such as molecular fingerprints. This might be ideal for optimization and machine learning methods have somewhat helped to solve the problem of how a molecular figure print might be used as input to algorithms which predict binding free energy. However, converting the optimized solution, a vector of real numbers, back into a molecule is in general a difficult problem [128]. This conversion problem could be solved if instead of featurizing the molecule as a vector of real numbers the molecule is featurized into chemical groups. If chemical groups are used as the optimization domain with groups added

and subtracted to try and improve binding affinity [35, 36] the problem of converting a vector of real numbers into a molecule is sidestepped. This domain, however, does not adhere to point 4), 5) or 6) in our list of what makes a good optimization domain. This idea will be explored further in chapters 3/4 and we will show that the real numbers which parameterize atomistic FFs are a good domain for the optimization of binding affinity.

1.6.1 Techniques

The methods which can be used to explore a space in a directed way are numerous but have a great deal of commonality. In the space of gradient based methods, one of the simplest methods for numerical optimization is gradient descent (GD) which can be described as follows. Given some starting point in a space x and a step size α , a local minimum can be found in a function $f(x)$ using two algorithmic steps:

1. Calculate a gradient of $f(x)$ with respect to x , this is $f'(x)$
2. Move down hill $x_{n+1} = x_n - \alpha f'(x_n)$

We refer to point 2. as the update formula. In the update formula, x_n denotes the value of x for the n th iteration of the optimizer. These steps are applied iteratively until no step can be found to reduce $f(x)$. The choice for the value of α is commonly made using a line search because, whilst there theoretically exists some exact α value which minimizes $f(x)$ in the search direction, for practical reasons it can be easier to take finite steps in the search direction until an approximate minimum is found. Other methods to choose the step size exist such as trust region methods [129]. For gradient descent methods, the choice for the search direction as $-f'(x)$ comes from the knowledge that moving against the gradient is the direction along which $f(x)$ will decrease the fastest for small steps. One of the strengths of GD is that it should allow for a local minimum in $f(x)$ to be found whilst only computing the gradient and avoiding calculation of the more costly Hessian $f''(x)$. However, GD can be extremely slow for some problems, a qualitative example would be a function which has a minimum at the end of a long valley. For such a function, GD will exhibit a characteristic zigzag search through the valley to reach the minimum, where this zigzag results in many more evaluations of $f(x)$ than may be necessary. As mentioned, however, GD is a relatively simple numerical optimization algorithm and there exists many alternative algorithms which can address GDs weaknesses.

Newton's method is another popular algorithm which can address some of the shortcomings of GD. Newton's method for optimization can be reached by considering a second order

Taylor series expansion of $f(x_n + p)$ around x_n ,

$$f(x_n + p) = f(x_n) + f'(x_n)p + 1/2f''(x_n)p^2, \quad (1.26)$$

the step p which minimizes $f(x_n + p)$ can be found by taking the gradient of equation 1.26 with respect to p and setting the result to zero, which yields a search direction $-f'(x)/f''(x)$ and so the Newton update formula is as follow,

$$x_{n+1} = x_n - \alpha f'(x)/f''(x). \quad (1.27)$$

The parameter α controls the size of the step in the Newton direction; the derivation suggests $\alpha = 1$, but we are at liberty to adjust α for the sake of stability. The update formula looks similar to the GD update formula but the search direction is now modified by a factor $1/f''(x)$, which is including information for the curvature of the function. Having information about this curvature, moves the step direction away from the steepest descent and towards the local minimum, assuming the function is quadratic. Calculating the Hessian can be expensive and inverting it can cause numerical problems [129, 130]. We can avoid calculating the full Hessian by using quasi-Newton methods. Quasi-Newton methods forego the inclusion of the numerically exact Hessian in the search direction, opting instead to use an approximation for the Hessian which is updated with information from previous iterations of the optimizer. An example of such a quasi-Newton method would be the Broyden–Fletcher–Goldfarb–Shanno algorithm [131].

In this work the optimizers we will use are GD and sequential quadratic programming (SQP). SQP is an iterative optimizer that can generate steps for a constrained nonlinear objective [129]. Both the GD and Newton's methods we have described here are unconstrained optimizations. The application of constraints is relatively self explanatory in that they constrain the minimizer to a subset of possible values; their effect on the difficulty of the optimization problem is less trivial. Constraints can reduce the number of feasible local minimums in a function, which may help or hinder a user to find a global or local minimum [129]. To apply SQP to the optimization problems in this work we use the SciPy implementation which is a wrapper around an older, 1988, implementation by Dieter Kraft [132]. The advantage of the SciPy implementation of SQP is that it allows for the use of any arbitrary combination of constraints and the application of these constraints can be seen in chapter 3: charge optimization.

1.7 Machine Learning

Machine learning (ML) is a field that relies heavily on optimization methods, and as such has advanced the theory and use of optimization methods. At their core most ML methods are optimization problems where parameters must be found to extremize an objective. Some of the most fundamental methods included under the umbrella of ML are well established algorithms such as linear regression or k -nearest neighbor. However, the rapid development of the ML field has seen newer methods such as artificial neural networks (NNs) applied across a myriad of fields, ranging from image and speech recognition [133] to chemical compound generation [134] and property prediction [135]. To examine these more contemporary ML methods we will now consider examples of some classes of NNs in more detail.

Deep neural networks (DNNs) are a commonly used architecture in ML methods and can be applied to property prediction problems [136, 137]. DNNs can be imagined like a fitting problem where we have a set of inputs, x , containing d points. These inputs have a corresponding output, $y^{observed}$, which also contains d points, both inputs and outputs are indexed by k . We would like to approximate these data points with a function. Choosing a toy example for the function as a straight line, $y^{predicted} = Ax + b$, there are two choices for parameters which must be made: these are the gradient of the line A and the intercept b . The objective is then to minimize the error between the straight line and the data points $y^{observed}$. The choice for A and b could be made analytically and this would then be an application of the linear regression method, equally A and b could be found by applying the optimization techniques we have already discussed in the previous section. Numerically solving for A and b allows for the introduction of more complicated functions which take a vector of x values as input and operate on the input with more sophisticated functions.

For a more realistic application of a DNN the function we use to predict a single value $y_k^{observed}$ would take a vector of x_k values, \mathbf{x}_k , as input and the function for $y_k^{predicted}$ is therefore,

$$y_k^{predicted} = \mathbf{w} \cdot \mathbf{x}_k + b_k, \quad (1.28)$$

many A values are now required one for each element of the input vector, \mathbf{x}_k , and these have been combined into a vector \mathbf{w} . When the A values are combined like this they are collectively referred to as the weights of the network. The b_k values can be grouped over all inputs into a vector of values generally referred to as the biases of the network, \mathbf{b} . Equation 1.28 corresponds to a network consisting of one layer. More layers can be added by passing the output of equation 1.28 to the input of another layer of similar linear equations. To stack these linear layers together in a useful way an activation function [138] such as hyperbolic

tan or the sigmoid function must be added to each layer. The objective function for a DNN as described here could be,

$$\min_{\mathbf{w}, \mathbf{b}} DNN(\mathbf{x}, \mathbf{w}, \mathbf{b}) = \frac{1}{d} \sum_{k=1}^d |y_k^{predicted}(\mathbf{x}_k, \mathbf{w}, b_k) - y_k^{observed}|. \quad (1.29)$$

In the terminology of ML methods, this objective is referred to as an L1 loss function. Here d denotes the number of predicted data points, $y_k^{predicted}$. The details of the function we are using to approximate the true function of $y^{observed}$ are contained within $y^{predicted}(\mathbf{x}_k, \mathbf{w}, b_k)$. In order to find the values of \mathbf{w} , \mathbf{b} , which minimize the objective, the gradient of the objective with respect to the weights and biases of the network can be computed and the optimization techniques described in the optimization section could be applied to find a minimum. More commonly in ML, optimizers such as stochastic gradient descent [139] or ADAM [140] are used but we do not discuss the detail of these optimizers here. More advanced applications of the artificial neural networks described here have been made in many fields. In the domain of material and compound property prediction these networks have successfully been applied to a wide range of systems ranging from alkanes [136] to metal alloys [137] and most relevant here the properties of drugs such as activity against a target [141] or toxicity in the human body [142].

The major choices in network architecture are a) the functions used in each layer of the network b) how information is passed between layers of a network and c) the objective function. These three choices inform what problems the NN can be applied to solve. We will see a large variation in these three areas now when examining another class of NN called recurrent neural networks (RNN). RNNs are a type of NN specifically designed to operate on sequential data. An ideal application of an RNN would be text generation for example. A trained RNN could take as input the first word of a sentence and then generate sequential words in that sentence.

The idea of acting on sequential data in this recurring fashion is built into the architecture of the RNN, to describe this we introduce the notion of a cell. At training time, each cell in the network takes as input the previous word in the sentence (the first cell is initialized with an arbitrary start token). Each cell then outputs a probability distribution over all possible words, figure 1.3 shows this RNN architecture diagrammatically. The objective function used to learn the correct probability distributions is to maximize the likelihood that the correct word will be selected, where correct is defined relative to a training sentence and this objective can

be written as,

$$\min_{\mathbf{w}, \mathbf{b}} RNN(\mathbf{x}, \mathbf{w}, \mathbf{b}) = - \sum_{t=1}^T Q_t \log P(g^t(\mathbf{x}, \mathbf{w}, \mathbf{b}) | g^{t-1}, \dots, g^1), \quad (1.30)$$

here t is the index of the word in a sentence of length T , g^t is a word in a sentence at index t and $P(g^t | g^{t-1}, \dots, g^1)$ is the probability distribution for generating g^t given that g^{t-1}, \dots, g^1 was previously generated. \mathbf{x} is the start token which is constant here but could vary to bias the output of the network and Q_t is the target probability distribution which represents the correct word in the training sentence at position t . Generally a RNN would be shown many sentences at train time with weights and biases found to minimize equation 1.30 summed over all training sentences.

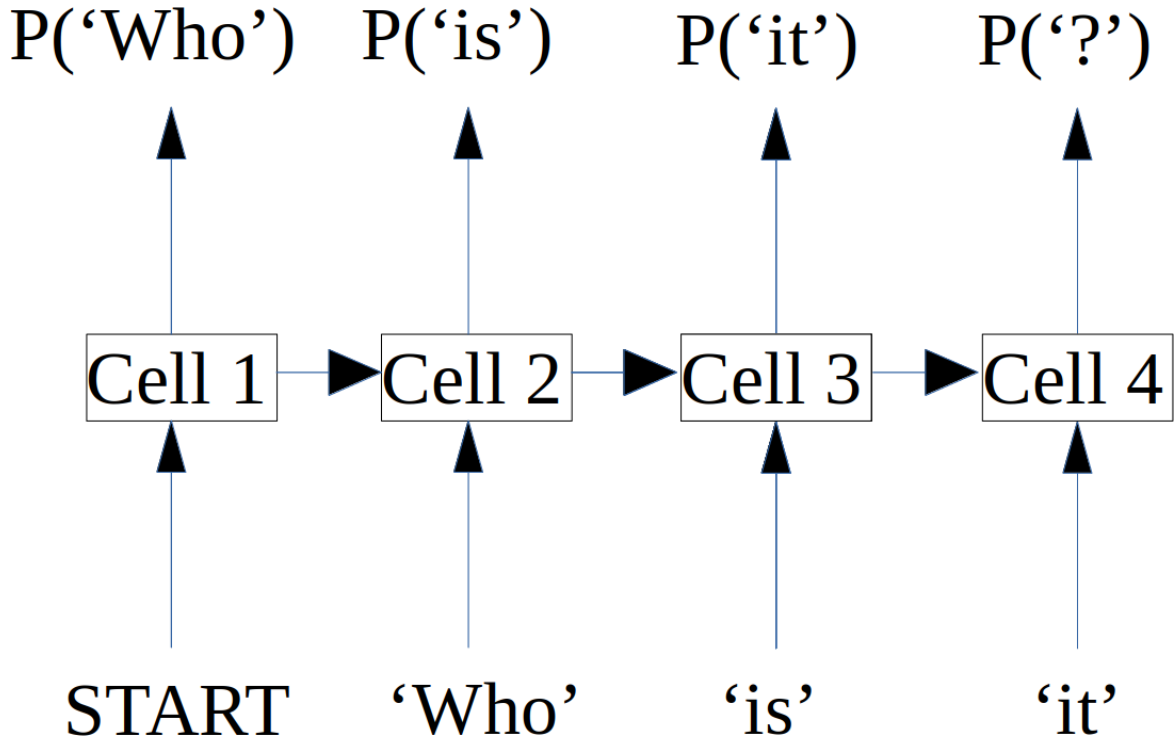


Fig. 1.3 Schematic for how information is passed between RNN cells at training time. In this instance the RNN should generate a probability distribution which maximises the probability of generating the sentence ‘Who is it?’

An important detail we have omitted is the architecture inside each cell which determines how $g^t(\mathbf{x}, \mathbf{w}, \mathbf{b})$ is a function of the weights and biases of the network, whilst we will not explore this in detail here it is suffice to say that there are two cell architectures which are commonly used for contemporary RNNs; these are long short-term memory (LSTM) and gated recurrent units (GRU). These networks are specifically designed to lessen the impact

of the vanishing gradient problem which was detrimental to early RNNs [143]. GRUs and LSTM have been successfully applied to generate human readable text [144] and interpret speech [133]. Most relevant here is the application of RNNs to *de novo* drug design by Olivecrona *et al.* [128, 134] in these works an application of a GRU based RNN was made to learn a string based syntax for compounds known as SMILES. This was achieved by training the RNN with data consisting of small (10-50 heavy atoms) molecules selected from the ChemBL data base containing only H, B, C, N, O, F, Si, P, S, Cl, Br, I elements. This training data contained 1.5 million compounds in total. This RNN was capable of generating valid and novel SMILES strings; several other groups have applied these RNN architectures to similar problems [145, 146].

In the form we have described RNNs here they could already be applied to *de novo* drug design. Simplified, this process would consist of two stages: 1) generation of many molecules by the RNN (virtual library) 2) prediction of activity for generated compounds with a discriminative model. This method resembles very closely established compound library screens, a more innovative application of RNNs is to solve the inverse quantitative structure–activity relationships (QSAR) problem. We saw in the discussion of DNN that a model of this type requires an input vector \mathbf{x} . In the context of compound property prediction \mathbf{x} can be a set of continuous values describing features of the molecules. These features can be picked in a supervised or unsupervised fashion but regardless of how they are chosen the features define a vector space. Within this space each point can be assigned a predicted value for a property, for example protein binding affinity, by the DNN. If a point is found to have predicted properties which are beneficial we would like to know what compound corresponds to this point in the vector space. How to decode this vector into a compound is the inverse QSAR problem.

Olivecrona and Blaschke *et al.* deal with this inverse QSAR problem in two ways, first using a reward based method [128]. A reward layer can be considered as an additional objective which when included in an RNN network, via an additional round of training, will add a bias to generated compounds to satisfy the reward layer objective. Reward layers have been used widely in ML applications across many fields [147, 148]. The reward layer here acts as an additional *post hoc* tool to tune the types of molecules that are generated, beyond the primary requirement of validity. In the context of the inverse QSAR problem Olivecrona *et al.* used this method to generate molecules judged active by a DRD2 activity model.

An alternative method to solve the inverse QSAR problem explored by Blaschke *et al.* allows the vector that the user wishes to decode to be an input to the RNN [149]. To achieve this all compounds in the RNN training set are converted into vectors of molecular descriptors, using an unsupervised method [150]. During the training of the network the RNN is passed,

as input, the vector associated with a particular compound, whilst learning to generate that compound. So the 'arbitrary' start token we described for the general application of a RNN has now been swapped for a specific vector which biases the RNN to generate a specific SMILES string. This RNN was combined with the DRD2 activity model to search for molecules judged to be active. The DRD2 model also used a vector of descriptors as input, to search this vector space a Bayesian optimization method was used. When an vector of descriptors was found, judged to be highly active by the DRD2 model, it could be passed to the RNN which would decode it into a molecule. In chapter 5 of this thesis a similar method will be used to solve the inverse QSAR problem for an existing *plasmodium falciparum* (the parasite which causes the disease malaria in humans) activity model with the aim of helping to generate effective therapeutics against malaria.

1.8 Parallelization

In this work the term efficiency, with reference to exploring chemical space, will mostly be used to mean increasing the amount of information we can glean per clock cycle of a computational resource. It is worth briefly considering efficiency in a different context, which is in terms of computer wall time. Wall time refers to the real time passing on a clock on the wall as opposed to the total CPU time which is a function of the number of processes we are running in parallel on a CPU. Wall time is a critical component to time sensitive research, such as the search for therapeutics to a newly emerged disease and reducing the wall time of these calculations can be of great benefit.

In this work the only MD engine used was OpenMM which is already optimized to run in parallel on the many compute units inside one GPU. There will be multiple opportunities throughout this work where we can exploit the embarrassingly parallel nature of a task to parallelize it over many GPUs and reduce the wall time of our calculations. An embarrassingly parallel problem is one which can be easily decomposed over a domain such that little or no communications are needed between the decomposed elements. On a scale of difficulty for domains over which MD can be decomposed from impossible to embarrassing we would have: a decomposition over sequential time is impossible, over simulation space is difficult, and over repeats or replicas is embarrassing. The relative speed of communications and computation for GPU hardware was previously referenced in section 1.3 and we commented that computation was much faster than communications. Therefore for any calculations we do in this work using GPUs we will avoid any difficult parallelization to minimize communications.

The codes presented in this work (*Fluorify*, *LigandOptimiser*) make an effort to parallelize the work of calculating the dynamics and FEP calculations wherever possible. The decomposition of these tasks is different. Dynamics is decomposed by trajectory length i.e. requesting 50 ns over 4 GPUs will compute 4 independent trajectories of 12.5 ns per GPU. FEP is decomposed by λ windows i.e requesting 16 windows over 4 GPU will compute 4 FEP windows per GPU. There are additional decompositions for other calculations we will perform that we have not discussed here but the idea is always the same: a decomposition of the embarrassingly parallel part of a given task over GPUs.

Chapter 2

Computational Fluorine Scanning

We will now present the results of this thesis, each chapter in these results represents a distinct piece of work, as such each chapter contains its own introduction, methods, results and conclusion sub chapters. These chapters will include discussion for how each piece constitutes part of this whole thesis. This discussion will reference how the theory used and method developed allows for points in chemical space to be tested faster and or allow for a reduced number of points to be tested in a search for compounds which satisfy a given objective.

2.1 Introduction

In this chapter we present perturbative fluorine scanning, a computational fluorine scanning method using free-energy perturbation calculations. This method can be applied to molecular dynamics simulations of a single compound and make predictions for the best binders out of numerous fluorinated analogues. We tested the method on nine test systems: renin, DPP4, menin, P38, factor Xa, CDK2, AKT, JAK2, and androgen receptor. The predictions from the presented fluorine scanning method are in excellent agreement with more rigorous alchemical free-energy calculations and in good agreement with experimental data for most of the test systems. However, the agreement with experiment was very poor in some of the test systems and this highlights the need for improved force fields in addition to accurate treatment of tautomeric and protonation states. The method is of particular interest due to the wide use of fluorine in medicinal chemistry to improve binding affinity and ADME properties.

Fluorine scanning is a common technique in medicinal chemistry and involves systematic replacement of hydrogen with fluorine [151–156]. It can improve binding affinity as well as ADME properties [157, 158]. Fluorinations are able to influence the conformation of a molecule as well as its ability to pass through a cell membrane or, relevant here the molecules

potency against a target [157], these changes to the properties are mainly related to the electronegativity of fluorine which is uniquely high. A striking example for the potential of fluorine scanning can be seen in its application to factor Xa [159]. In the work of Lee *et al.* the authors found that a modification from a hydrogen in compound **1** to a fluorine in compound **2** improved the binding affinity by approximately 55-fold, see figure 2.1.

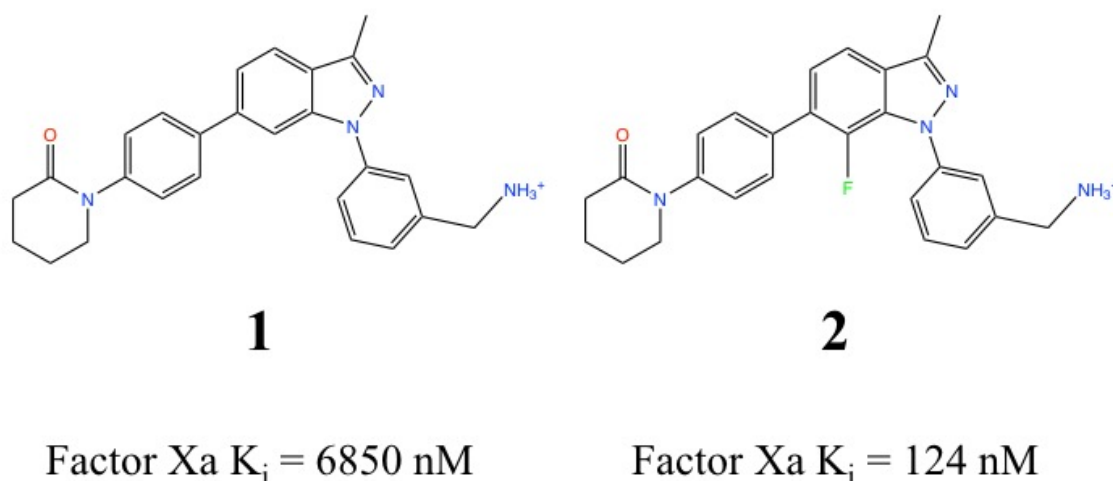


Fig. 2.1 Two factor Xa inhibitors and their K_i values

One of the drawbacks of fluorine scanning is the requirement to test each hydrogen-to-fluorine mutation individually. For example, testing all the hydrogen-to-fluorine mutations in compound **1** requires synthesizing and assaying 16 compounds. Testing combinations of two hydrogen-to-fluorine mutations is an order of magnitude more challenging.

Computational fluorine scanning using a molecular mechanics-Poisson-Boltzmann/ surface area (MMPBSA) method has been suggested in the past as a way to design molecules with improved binding affinity *in silico* [160]. However, MMPBSA calculations use a simplified implicit solvent model. Implicit water models fail to model key elements of protein ligand systems, for example implicit water models are unable to deal effectively with interfacial water molecules. These occur commonly and are very difficult to treat effectively with implicit solvent approaches. Additionally the accuracy of implicit water models is lacking for many types of free energy calculations [161–163], particularly relevant here are implicit model's use in the calculation of binding free energies [164]. As such alchemical free-energy methods with explicit solvent are increasingly used in place of MMPBSA, calculations and

we therefore propose the application of FEP calculation here for the purpose of computational fluorine scanning.

To perform full FEP calculations for every possible fluorination, however, may not be the most efficient way in which to perform a computational fluorine scan. Instead the common reference state for all possible fluorination can be exploited. In order to leverage this common reference state the theory of single step perturbation can be used as discussed in the free energy section 1.4. SSP methods allow for the free energy changes to all perturbed fluorinated states to be calculated from a single simulation in the reference un-fluorinated state. We refer to this single step fluorine scan as perturbative fluorine scanning (PFS)

To carry out these PFS calculations we have developed a tool, *Fluorify*¹, for executing the pipeline for these calculations automatically. *Fluorify* uses OpenMM [165] as both a molecular dynamics engine and library to create the modified alchemical systems. *Fluorify* will generate all of the required mutant ligands from an input wild type ligand; these mutants are automatically parameterized, built into complex systems, simulated and analysed.

2.2 Methods

We consider the effect of hydrogen to fluorine mutations for the hydrogens attached to aromatic carbons of nine different protein ligand binding systems, detailed in table 2.1. The chemical structures of these ligands are shown in figures 2.2, all mutated hydrogens are named explicitly.

Examining all of the systems we are considering here, table 2.1 shows the experimental data that is available for hydrogen to fluorine mutations taken from the respective papers where these systems were investigated [166, 159, 167, 169, 170, 168, 171–173]. Experimental $\Delta\Delta G$ values in table 2.1 are calculated from K_i or IC_{50} values found in their respective references, see references for experimental methodologies. Errors for these experimental values are reported here when provided in the original work. It should be noted that manual preparation of the ligands was required. This manual preparation involved changing the ligand structure from that provided in the Protein Data Bank [174] (PDB) to the highly related structure for the start point of the experimental fluorine mutation examined. These changes are reflected in the chemical structure shown in figure 2.2. In addition, DPP4 was modelled as a monomer rather than the dimer in the crystal structure. This is done for the significant computational saving of reducing the number of atoms by a factor of two, under the assumption that this ligands distance, 20 Å, from the dimer interface is sufficient that modeling as a monomer as little effect on the results.

¹*Fluorify* available at <https://github.com/adw62/Fluorify>.

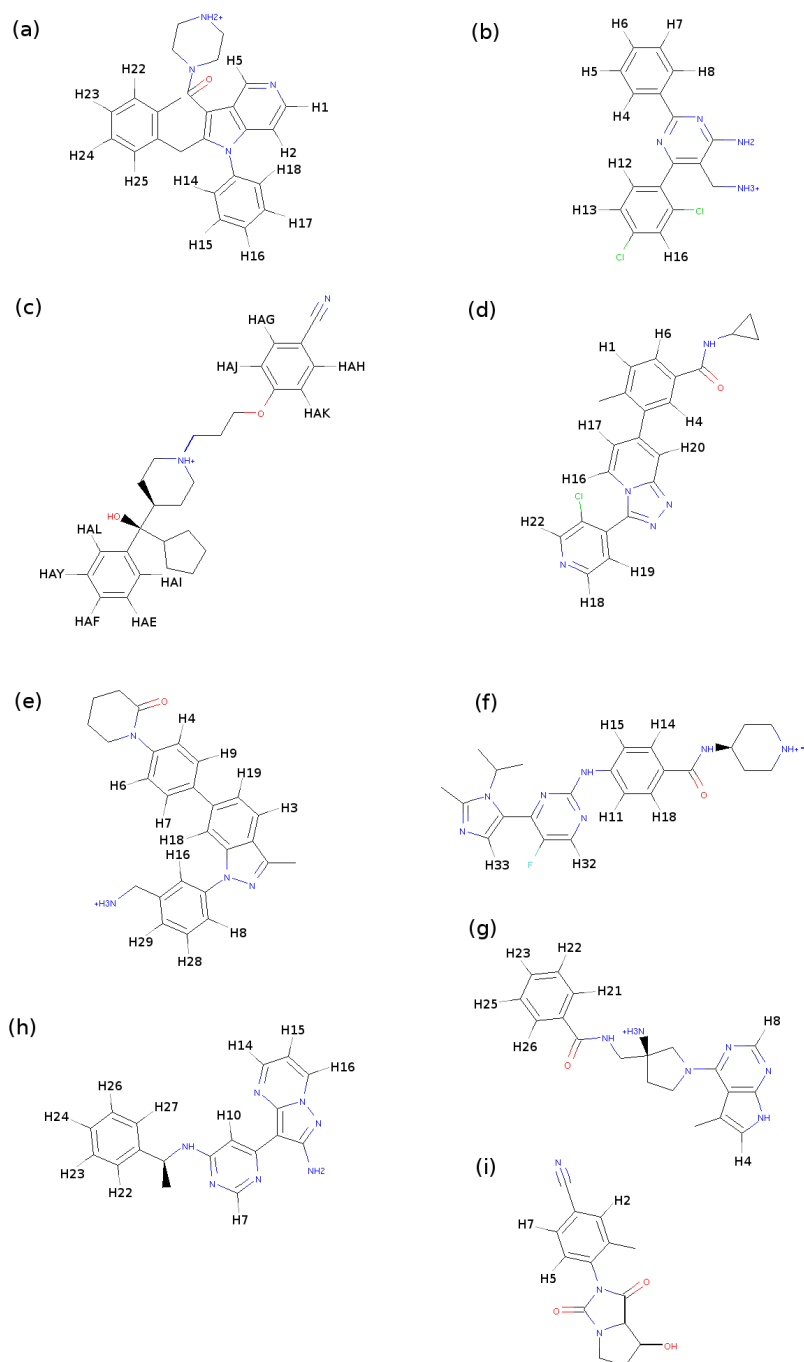


Fig. 2.2 Chemical structures depicting the wild type ligands in each system. (a) Renin [166], (b) DPP4 [167], (c) Menin [168], (d) P38 [169], (e) FXa [159], (f) CDK2 [170], (g) AKT [171], (h) JAK2 [172], (i) Androgen receptor [173].

Table 2.1 $\Delta\Delta G$ for hydrogen to fluorine mutations obtained experimentally. PDB and PDBIDs taken from the Protein Data Bank. The second column denotes the system as specified by figure 2.2. Third column shows hydrogen being mutated as specified by figure 2.2. Values in parentheses are the errors as reported in experimental work.

PDBID	System	Hydrogen	Experimental $\Delta\Delta G$ (kcal/mol)
3OOT[166]	Renin (a)	H22	-2.47
1RWQ[167]	DPP4 (b)	H5	-2.31
	DPP4 (b)	H6	-2.31
	DPP4 (b)	H7	-2.31
	DPP4 (b)	H4	0.91
	DPP4 (b)	H8	0.91
4OG6[168]	Menin (c)	HAY	-0.40(-0.51, -0.30)
3S3I[169]	P38 (d)	H1	-2.26(-2.67, -1.86)
2RA0[159]	FXa (e)	H18	-2.37
2W17[170]	CDK2 (f)	H18	-2.12
3MVH[171]	AKT (g)	H23	-1.26
	AKT (g)	H22	-0.29
	AKT (g)	H25	-0.29
	AKT (g)	H26	-0.20
3IOK[172]	JAK2 (h)	H24	-1.15
2NW4[173]	Androgen receptor (i)	H2	-1.11

The proteins studied in this chapter are renin, menin, DPP4, FXa, CDK2, AKT, p38 and JAK2. Renin is relevant in the pathology of diseases related to blood pressure such as cardiovascular disease [175]. Menin interacts with a protein called mixed lineage leukemia 1 and this protein-protein interaction is critical in leukemias [168]. DPP4 is responsible in the human body for cleaving a protein which stimulates the secretion of insulin and its inhibition is relevant to treating type 2 diabetes [167]. FXa plays a role in coagulation, its inhibition is relevant in anticoagulation therapies, such the prevention of recurrent deep vein thrombosis [159]. CDK2 is important to the cell progression of eukaryotic cells and its inhibition can be used in the treatment of cancer [176]. AKT inhibition can also play a role in the treatment of cancers, with AKT also relevant in cell proliferation, and evidence suggesting AKT can be deregulated in malignant cells [172]. The inhibition of P38 is sought in the pursuit of anti-inflammatory therapeutics which can be used to treat inflammatory diseases such as rheumatoid arthritis [169]. Mutations in the JAK2 protein have been implicated in the pathology of myeloproliferative disorders which include several leukemias, inhibitors of JAK2 are sought as therapeutics to these disorders [172].

2.2.1 System Setup

All FEP and PFS calculations performed in this work were made with *Fluorify* the details for each stage of these calculations are as follows. The co-crystal structure for the nine systems examined here are taken from the PDB with PDBIDs shown in table 2.1. To prepare these systems, non-standard residues were converted to their standard equivalents with *pdbfixer* [177]. Selenomethionines were changed to methionines, sidechains were added using Schrödinger's Preparation Wizard [178], which was also used to assign protonation state of all ionizable residues. All buffer solvents and ions were removed. The hydrogen atom positions were then built using *tleap* and forcefield parameters and partial charges were assigned from the AMBER ff14SB force field [93]. Parameters for the inhibitors were generated using *Antechamber* [179] with AMBER GAFF 2 [99] and AM1-BCC [180]. These structures and parameters were then passed to YANK's [181] 0.23.7 automatic setup pipeline to build solvated ligand-protein and ligand systems. For solvation, TIP4P-EW [107] was used; at this stage a salt concentration of 150mM and any required counter-ions were added. In every case, the edge of the solvation box was 15 Å from any atom of the receptor and ligand.

2.2.2 Molecular Dynamics

All simulations were performed with OpenMM 7.3.0 [54] as follows. First OpenMM's default minimizer was used to minimize all structures. Then equilibration was performed in the NPT ensemble for 500 ps at 300k and 1 atm using a Langevin integrator and Monte Carlo barostat. MD simulations were performed in the NPT ensemble using a time step of 2 fs. Van der Waals interactions were truncated at 11.0 Å with switching at 9.0 Å. Electrostatics were modeled using particle mesh Ewald method with a cutoff of 11 Å. All other simulation parameters were left as default. We ran triplicate simulations of the non-fluorinated compound with the ligand in complex and in solution, for 50 ns. Snapshots were collected every 5 ps. The stability of complex structures was assessed with a calculation of protein RMSD over trajectory lengths, see appendix figures A.1-A.9 for plots.

2.2.3 Perturbative Fluorine Scanning

During perturbative fluorine scanning the binding free energy $\Delta G_{\text{mutation}}$ between the original and fluorinated ligands was calculated with an SSP methodology which applied the Zwanzig

equation as follows,

$$\Delta G_{mutation} = -k_b T \log \langle e^{-\beta(U_{fluorinated} - U_{original})} \rangle_{original}, \quad (2.1)$$

here $U_{fluorinated}$ is the potential energy of the system calculated in the fluorinated Hamiltonian and $U_{original}$ the potential from the original un-fluorinated Hamiltonian. The angled brackets $\langle \rangle_{original}$ denote that the sampling is taken from the system using the original Hamiltonian. To calculate a relative binding free energy $\Delta G_{mutation}$ was combined from the bound and unbound states as shown in figure 1.2.

To switch between Hamiltonians, van der Waals, charge, bond, angle and torsion parameters were all assumed to change. Since this was a post analysis and the dynamics were collected from the system with the un-fluorinated bond parameters, the change in bonded parameters had no effect on the dynamics of the molecule's geometry. Whilst this should be negligible when considering the change in geometry of non-perturbed atoms this may not be true for the atoms perturbed from hydrogen to fluorine where the C-F bond should be longer than C-H. To include this in the model, we used a hybrid topology approach where massless interaction sites at the position of all possible fluorine mutations were added. The position of these fluorines was defined relative to the position of their parent hydrogen such that the C-F distance was always 1.24 times the C-H distance [182]. During the simulation the LJ, charge, bond, angle and torsions parameters of these additional fluorine sites are turned off. When mutating to a fluorinated system, the relevant hydrogen was turned off and fluorine LJ and charge parameters were applied to the additional site this is demonstrated in figure 2.3. The torsion and angle parameters were mutated from the hydrogenated system to fluorinated system, but the torsions and angles remain on the parent hydrogen and were not transferred to the virtual fluorine. This has no effect on the energy as the angles remain the same. When simulating these systems, all hydrogen bonds were constrained, since the position of the fluorine was defined relative to the position of its parent hydrogen it was also implicitly constrained. We therefore make the assumption that the C-F bond oscillations were negligible. To prevent the hybrid topologies from interacting, the additional fluorine was excluded from interacting with their parent hydrogens.

2.2.4 FEP Calculations

To validate the PFS result, we compare it against standard alchemical relative binding free energy calculations using the MBAR [114] estimator, see the free energy theory section for more details on this. These FEP MBAR calculations use the same fluorinated/un-fluorinated end states and hybrid topology described previously for PFS. We used a total of 12 equally

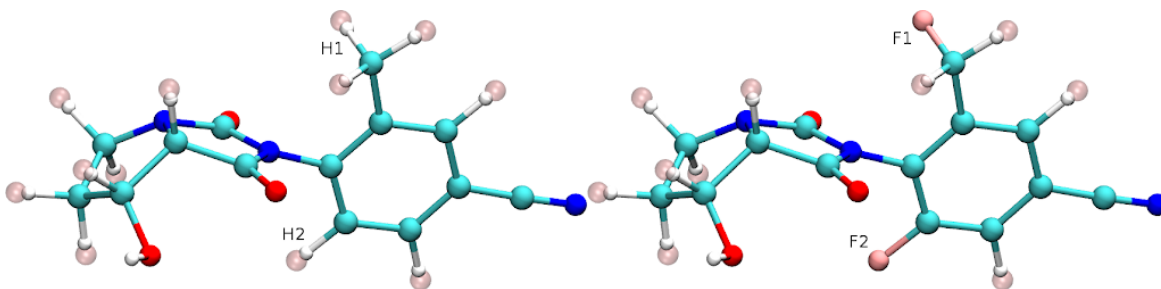


Fig. 2.3 Left panel Androgen inhibitor with all fluorines turned off. Right panel Androgen inhibitor with H1 and H2 transformed to F1 and F2.

spaced lambda windows in which LJ, charge, angle and torsion parameters were interpolated simultaneously and lineally from the wild-type to the mutated state as described for the PFS calculations. All windows were sampled independently with 2 ns of Langevin dynamics. Giving a total of 24 ns of sampling per mutant half of the 50 ns used for all mutants of a ligand in PFS. All simulation conditions were identical to the PFS molecular dynamics calculation described above. The samples were processed with MBAR analysis using the PyMBAR 3.0.1 [114] python library. This FEP protocol is run automatically as part of the Fluorify package to check the $\Delta\Delta G$ for the top-ranked mutants as determined by PFS.

2.2.5 Summary of Methods

The FEP calculations performed in this work can be summarized as follows. Perturbative fluorine scanning calculations were performed using SSP theory and any $\Delta\Delta G$ values from these calculations are denoted as PFS $\Delta\Delta G$. These SSP calculations were then verified with full FEP calculations analysed with MBAR, $\Delta\Delta G$ values from these verification calculations are denoted as FEP $\Delta\Delta G$. All experimental $\Delta\Delta G$ will be denoted as EXP $\Delta\Delta G$.

2.3 Results

We first analysed the convergence of PFS predictions as the simulation time was increased. Figure 2.4 shows the PFS $\Delta\Delta G$ predictions for the factor Xa test case. From figure 2.4 it can be seen that the $\Delta\Delta G$ calculations are well converged within 50 ns of sampling.

The results for each system are $\Delta\Delta G$ s calculated by PFS and FEP and table 2.2 shows the results of these $\Delta\Delta G$ calculations. $\Delta\Delta G$ values are presented for some number of the best binding ligands where the number is taken to be either three or the rank (as determined by PFS) of the best experimental mutant. Such that, PFS and FEP results are always presented for the experimental mutant measured to be the tightest binder. All computational values in

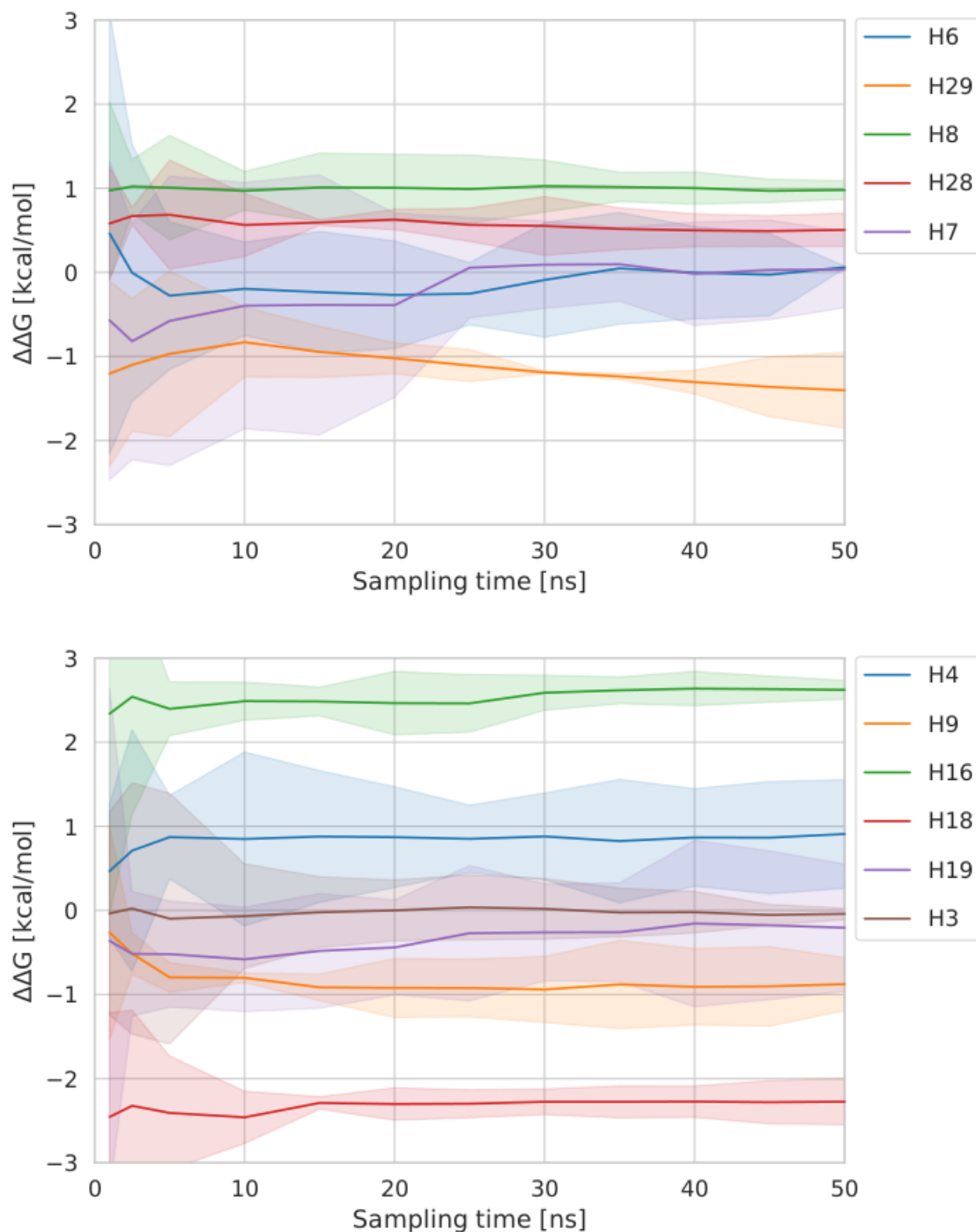


Fig. 2.4 Convergence of the PFS $\Delta\Delta G$ predictions for the hydrogen to fluorine mutations in the factor Xa test case as the simulation time is increased, H labels shown in figure 2.2. Calculations were performed at 1.0 ns, 2.5 ns and then from 5.0 ns to 50.0 ns in 5.0 ns increments. $\Delta\Delta G$ reported as mean of three replicates with shaded area showing 95% confidence interval computed as $\text{mean} \pm t_2 \cdot \text{SEM}$, where t_2 is the t-distribution statistic with two degrees of freedom, and SEM is the standard error of the mean computed from the sample standard deviation of the three independent replicate predictions.

table 2.2 are the average of three calculations, unaveraged values for PFS are reported in table A.1 along with $\Delta\Delta G$ values from PFS calculations for all possible aromatic hydrogen to fluorine mutations.

Overall PFS demonstrated excellent agreement with FEP with a mean unsigned difference of 0.3 kcal/mol. PFS also shows good agreement with experiment in most systems with a mean unsigned error of 1.2 kcal/mol and this is comparable to the mean unsigned error of FEP and experiment, 1.1 kcal/mol. To examine the agreement of these methods more closely the correlation of their results should be inspected. The correlation between PFS and FEP is good as seen in figure 2.5 where fitting a trend line yields a R^2 value of 0.9. This correlation is favourably impacted by the uppermost top right data point and excluding this data point R^2 would be 0.8.

In terms of agreement with experimental data, errors are greater than 2.0 kcal/mol in only three cases (all for DPP4). Overall correlation is poor, however, because the experimental range for $\Delta\Delta G$, 0.40 - 2.47 kcal/mol, is small. Table 2.3 shows the correlation R^2 , the mean unsigned difference and *RMSD* between the PFS, FEP and experimental data. Despite poor experimental correlations, both PFS and FEP have a reasonable accuracy in terms of mean unsigned difference.

Looking at each test case individually, we see that PFS is a reasonably good predictor of the mutant highlighted by experimental work. For systems renin, menin, P38, FXa, JAK2 and Androgen receptor PFS correctly predicts the mutant highlighted by experimental work. System DPP4 was more challenging, the top mutants, H5, H6 and H7 all have the same free energy. PFS ranks one of the best mutants, H7, as 3rd and incorrectly calculated the best mutant as H13 and second best as H16. FEP does better, again incorrectly ranking the best mutant as H13 but ranking two of the best experimental mutations H5 and H7 as second and third respectively. Whilst PFS and FEP are well agreed (within 1 kcal/mol) for this test case neither of these methods predict the best experimental mutant correctly. This may be due to the system preparation, modelling the DPP4 monomer rather than the dimer. Additionally it can be seen from figure 2.2b that H13 and H16 are on a phenyl already selected as favourable for chlorination and this may explain why PFS indicates these positions over the best position determined by experiment. For system CDK2, PFS fails to predict the top mutant however this failure is mirrored in FEP. The predictions made by PFS and FEP for the $\Delta\Delta G$ of the top experimental mutant agree within 1 kcal/mol. However neither are within 1 kcal/mol of the experimental $\Delta\Delta G$. PFS and FEP select H33 as the best position for fluorination. It can be seen in figure 2.2f that H33 is close to a position already selected as favourable for fluorination and this might explain why it selected over the position highlighted in the experimental work.

Table 2.2 PFS, FEP and experimental $\Delta\Delta G$ s for all test cases. Results are reported as the mean $\Delta\Delta G$ s of three replicates with 95% confidence interval reported between square brackets computed as $\text{mean} \pm t_2 \cdot \text{SEM}$, where t_2 is the t-distribution statistic with two degrees of freedom, and SEM is the standard error of the mean computed from the sample standard deviation of the three independent replicate predictions. Symmetry related positions are indicated by †.

System/Hydrogen	PFS $\Delta\Delta G$ [kcal/mol]	FEP $\Delta\Delta G$ [kcal/mol]	EXP $\Delta\Delta G$ [kcal/mol]
(a) Renin			
H22	-1.8 [-2.4, -1.2]	-1.4 [-1.8, -0.9]	-2.47
H25	-1.6 [-2.1, -1.0]	-1.5 [-2.3, -0.8]	
H15	-0.6 [-1.0, -0.1]	-0.4 [-1.1, 0.4]	
(b) DPP4			
H13	-1.2 [-3.6, 1.3]	-0.8 [-2.1, 0.6]	-2.31† -2.31† 0.91† -2.31 0.91†
H16	-0.5 [-1.7, 0.8]	-0.2 [-1.7, 1.2]	
H7	-0.3 [-1.0, 0.4]	-0.2 [-0.8, 0.4]	
H5	-0.2 [-0.6, 0.1]	-0.4 [-0.7, 0.0]	
H4	-0.2 [-0.6, 0.2]	0.2 [-0.6, 1.1]	
H6	-0.2 [-0.3, 0.0]	-0.2 [-0.3, -0.1]	
H8	-0.1 [-1.5, 1.3]	0.0 [-0.7, 0.6]	
(c) Menin			
HAY	-1.5 [-2.0, -1.0]	-1.4 [-2.3, -0.5]	-0.40
HAI	-1.3 [-1.6, -0.9]	-0.7 [-1.2, -0.3]	
HAL	-0.9 [-1.1, -0.6]	-0.8 [-1.0, -0.5]	
(d) P38			
H1	-2.2 [-2.8, -1.6]	-2.2 [-2.7, -1.6]	-2.26
H19	-1.9 [-2.1, -1.6]	-1.6 [-1.7, -1.4]	
H16	-0.6 [-0.9, -0.3]	-0.3 [-0.5, -0.1]	
(e) Fxa			
H18	-2.3 [-2.5, -2.0]	-2.2 [-2.3, -2.1]	-2.37
H29	-1.4 [-1.9, -1.0]	-0.6 [-1.2, -0.1]	
H9	-0.9 [-1.2, -0.6]	-0.8 [-1.3, -0.4]	
(f) CDK2			
H33	-1.0 [-1.4, -0.5]	-0.6 [-1.4, 0.1]	-2.12
H14	-0.3 [-1.2, 0.6]	-0.4 [-1.6, 0.8]	
H18	-0.2 [-0.9, 0.5]	0.1 [-0.5, 0.7]	
(g) AKT			
H22	-2.2 [-2.6, -1.8]	-0.9 [-2.3, 0.4]	-0.29† -0.29† -0.20 -1.26
H25	-1.3 [-2.1, -0.5]	-0.8 [-1.0, -0.6]	
H26	-1.2 [-2.2, -0.3]	-1.9 [-2.3, -1.6]	
H23	-0.7 [-1.1, -0.3]	-0.5 [-0.8, -0.1]	
(h) JAK2			
H24	-2.0 [-2.4, -1.6]	-2.1 [-2.7, -1.5]	-1.15
H27	-1.4 [-2.2, -0.6]	-1.2 [-1.7, -0.6]	
H14	-1.0 [-1.3, -0.7]	-0.8 [-1.4, -0.2]	
(i) Androgen receptor			
H2	-2.5 [-3.7, -1.3]	-2.5 [-2.8, -2.1]	-1.11
H7	-0.3 [-0.4, -0.2]	-0.3 [-0.3, -0.2]	
H5	3.5 [2.9, 4.0]	3.5 [2.9, 4.1]	

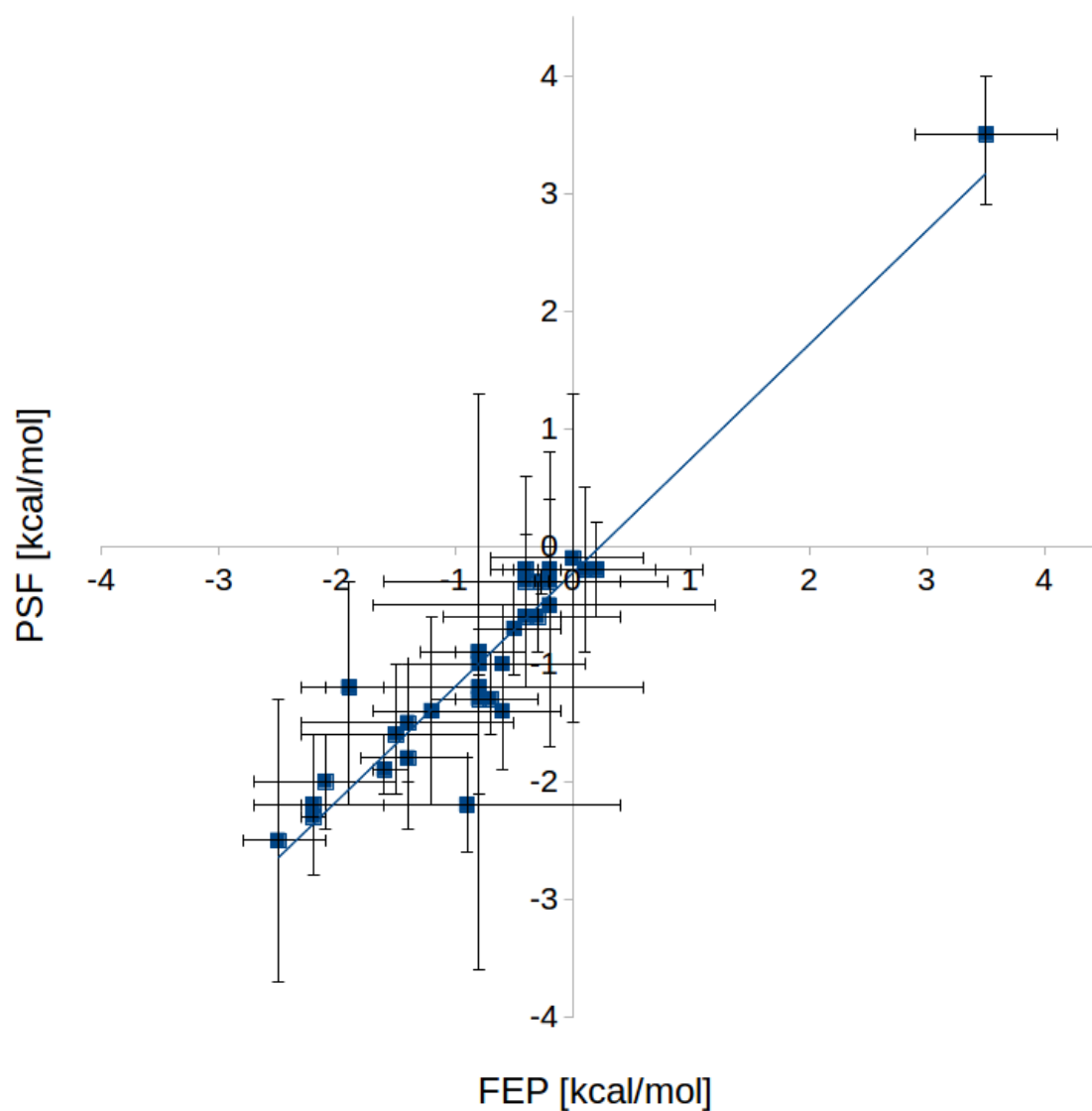


Fig. 2.5 Plot for $\Delta\Delta G$ for fluorine mutants for calculations from perturbative fluorine scanning (PFS) and FEP for all data points in table 2.2. $\Delta\Delta G$ s are reported as the means with error bars as 95% t-based confidence interval computed from the three independent replicate calculations.

Table 2.3 Correlation R^2 , mean unsigned difference, and $RMSE$ between the PFS, FEP and experimental data for all data points in table 2.2. 95% confidence intervals were estimated with the bias-corrected and accelerated bootstrap method and are reported between square brackets.

Comparison	R^2	Mean Unsigned Difference [kcal/mol]	$RMSE$ [kcal/mol]
PFS vs FEP	0.9 [0.6,1.0]	0.3 [0.2,0.4]	0.4 [0.2,0.5]
FEP vs EXP	0.0 [0.0,0.3]	1.1 [0.8,1.5]	1.3 [1.0,1.6]
PFS vs EXP	0.0 [0.0,0.2]	1.2 [0.9,1.5]	1.3 [1.0,1.6]

2.4 Conclusion

We have developed a new methodology for calculating relative binding affinities, which we term perturbative fluorine scanning. For a typical small molecule inhibitor, PFS applied to molecular dynamics simulations of a single molecule has the potential to combinatorially assess all possible fluorination sites yielding millions of predictions. These predictions can then be further assessed using more rigorous approaches and would be particularly useful in medicinal chemistry, providing insight for which analogs to synthesize. The PFS method is simple and could easily be improved by enhanced sampling techniques such as replica exchange or solute tempering.

The change in binding affinity for a wide range of hydrogen to fluorine mutations has been investigated computationally. Two computational methods were applied: FEP and PFS. It was demonstrated that FEP and PFS are in excellent agreement. However, the correlation between the computational methods and experiment for the $\Delta\Delta G$ calculations was not good. This poor correlation could come from many potential sources, such as systematic errors in the force fields or differences between computational and experimental systems. For example DPP4 has the worst accuracy, compared to experiment, of any system investigated in this work, this may stem from simulating it as a monomer compared to its dimer biological unit. Another potential source of error could be the fluorination making a significant change to the protonation, tautomeric, or conformational states of the ligand, an effect which we do not account for here. The poor correlation with experiment does not raise a major concern regarding the PFS method since it is similar for both FEP and PFS. Additionally, the mean unsigned error for both methods remains low at 1.1 kcal/mol and 1.2 kcal/mol for FEP and PFS respectively and this is very close to 1.0 kcal/mol which has been suggested as the acceptable error for free energy calculations. Where PFS performs well is in reproducing FEP results (both $\Delta\Delta G$ values and rankings) with good correlation, $R^2 = 0.9$ and high accuracy, mean unsigned difference = 0.3 kcal/mol at a fraction of the computational cost.

The scope of this method could also be expanded significantly by considering additional mutations such as chlorination, aromatic C to N and methylation. These changes, either alone or in combination, can be considered from single simulations of the bound and unbound states. In this work we have focused on hydrogen to fluorine changes which are particularly attractive in a medicinal chemistry context due to the potential for fluorine to act as a metabolic block in addition to a source of increased binding affinity [183]. We envisage the use of PFS to identify hydrogen to fluorine changes predicted to increase binding affinity in addition to hydrogen to fluorine changes predicted to improve ADME properties whilst maintaining binding affinity.

The results in figure 2.4 suggest that the molecular dynamics simulations need to be run for at least 30.0 ns and with multiple replicates to reach converged predictions. PFS consumes far less computational resources compared to traditional FEP approaches. For example, the FEP calculations in this work used 24.0 ns of sampling for a single mutant whereas PFS used 50.0 ns of sampling for all possible mutants (For the FXa test case this is 11 single hydrogens, 55 pairs of hydrogens and increasingly more for additional mutations). As a point of reference the FXa case has 99,000 and 13,000 atoms in the complex and solvent systems respectively. Run in parallel across 4 NVIDIA P100 GPUs using OpenMM 7.3.0 [54] and CUDA 8.0 it takes approximately 8.5 hours to collect 50 ns for both the complex and solvent systems. Using the methods outlined above PFS analysis then takes 1 hour to calculate $\Delta\Delta G$ for all 11 mutant ligands. Comparatively with the same hardware and software, full FEP takes 4 hours to compute $\Delta\Delta G$ for one mutant ligand. As such we have significantly reduced the constant of proportionality between number of mutants assessed and computation time by around a factor of 50 times. This allows significantly more fluorinated analogues to be tested using the same computational time and demonstrates a clear gain in efficiency with which to explore chemical space.

Chapter 3

Charge Optimization

3.1 Introduction

In this chapter we will develop methods to use numerical optimization techniques with alchemical free energy methods, allowing for more directed and efficient explorations of chemical space. In order to make this application we will now return to the idea, referenced in the free energy methods section 1.4, which was that free energy methods ‘exploit the malleability of the potential energy function’. This idea is relevant because just as we are free to play with the potential energy function to evaluate the Zwanzig equation in an alchemical method, we are also free to play with the potential energy function to construct an optimization problem in such a way that it can be solved as efficiently as possible. A good starting place to make the optimization efficient is to use the set of real numbers which parameterize our potential energy function as the domain of the optimization. Using a domain consisting of real numbers offers significant advantages for optimization problems and these advantages are discussed in detail in the optimization section 1.7. Additionally this domain incurs a limited inverse QSAR problem and it will be shown that the results of any optimization in this domain can be analysed by eye to generate design ideas for ligand mutations beneficial to the binding affinity. The disadvantage to this domain is the high dimensionality of this space, for example optimizing the charges of a molecule with 60 atoms gives a 60-dimensional space. These advantages and disadvantages will be discussed in more detail below as they present themselves and the following piece of work will address the computational cost of operating in this high dimensional space.

In this chapter, we present an explicit solvent alchemical free-energy method for optimizing the partial charges of a ligand to maximize the binding affinity with a receptor. This methodology can be applied to known ligand-protein complexes to determine an optimized set of ligand partial atomic charges. Three protein-ligand complexes have been optimized in

this work: FXa, P38 and the androgen receptor. The sets of optimized charges can be used to identify design principles for chemical changes to the ligands which improve the binding affinity for all three systems. In this work, beneficial chemical mutations are generated from these principles and the resulting molecules tested using free-energy perturbation calculations. We show that three quarters of our chemical changes are predicted to improve the binding affinity, with an average improvement for the beneficial mutations of approximately 1 kcal/mol. In the cases where experimental data is available, the agreement between prediction and experiment is also good. The results demonstrate that charge optimization in explicit solvent is a useful tool for predicting beneficial chemical changes such as pyridinations, fluorinations, and oxygen to sulphur mutations.

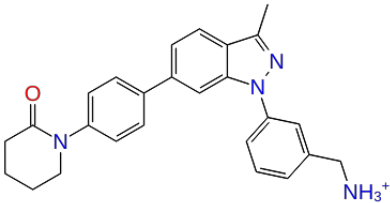
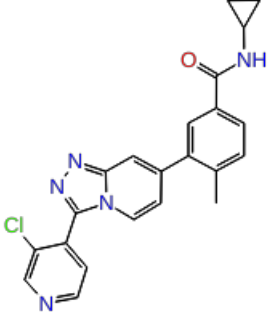
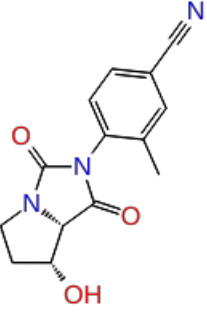
Charge optimization methods have been previously developed by Tidor and co-workers using an implicit water treatment of electrostatics [184, 185]. Poisson-Boltzmann calculations are performed on the bound and unbound states in order to find the optimal partial charges of a given molecule [186–189]. This approach has since been used by other academic groups [190, 191], employed in industry [192], and been extended to consider induced fit effects [193]. However, the approach suffers from the same deficiencies of all implicit water approaches discussed in the previous chapter. Additionally in this previous work the receptor and complex were assumed to be rigid. It is known that this may play a significant role in binding free energies [193]. Due to advances in available computing power, explicit water approaches to charge optimization are now possible. We propose to exploit these computational advances by applying SSP to the bound and unbound states of small molecule inhibitors to develop a method for electrostatic charge optimization in explicit solvent.

Combining SSP with explicit water MD calculations and flexible receptors and complexes has the potential to develop a more accurate charge optimizer. To carry out these charge optimizations, we developed a tool to automate their execution. This tool is freely available at https://github.com/adw62/Ligand_Charge_Optimiser. Our ligand charge optimizer uses OpenMM [165] as both an MD engine and a tool to create the modified alchemical systems. The software will generate all of the required mutant ligands from an input wild type ligand. These mutants are automatically parameterized, built into the complex systems, simulated, and optimized.

3.2 Methods

We optimize the ligand partial charges for three protein test cases: FXa, P38 kinase, and the androgen receptor. The chemical structures of the ligands studied are shown in table 3.1. The ligands were built from highly related molecules in the Protein Databank [174]: 2RA0 [159]

Table 3.1 Name of the target for each system with ligands 2D chemical structure and the PDBIDs of the target ligand complex.

Target	Ligand	PDBID
Factor Xa		2RA0[159]
P38		3S3I[169]
Androgen Receptor		2NW4 [173]

(FXa), 3S3I [169] (P38), and 2NW4 [173] (androgen receptor). These small modifications are made from the PDB to allow comparison with experimental data [159, 169, 173].

3.2.1 System Setup and Molecular Dynamics

Complex structure preparation and molecular dynamics protocol are unchanged from the computational fluorine scanning chapter 2.

3.2.2 Workflow

In this work, several free energy methods were used in different contexts. For clarity we will now outline these methods for future reference. After models for the protein ligand system were built, we performed an optimization of the charge parameters of an inhibitor. The

algorithm which was used to find a local minimum in the objective function was the SciPy 1.1.0 [194] implementation of the Sequential Least Squares Programming algorithm [132]. In order to perform this optimization two calculations are needed: one for the objective and one for the gradient; both of these calculations are performed using SSP theory. The details of how these calculations are performed can be found in the Optimization 3.2.3, SSP Objective 3.2.4 and SSP Gradient 3.2.5 sections below. Once this optimization is assessed to be converged, the output is analysed to generate design ideas for ligand mutations which should be beneficial. We then calculate the relative binding free energy for these designed mutations. These calculations, to test the design, ideas are performed using MBAR. We name these calculations full FEP with more details in the FEP Calculations section below. Figure 3.1 illustrates how each of these calculations are combined in the full workflow.

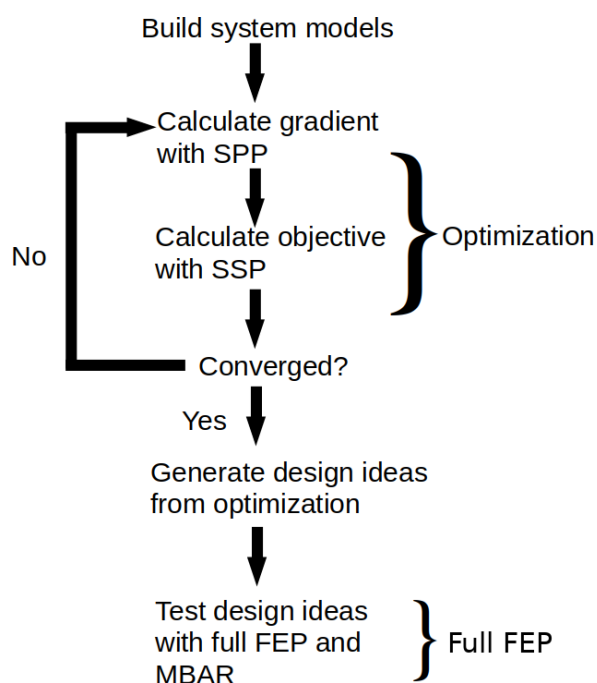


Fig. 3.1 Diagrammatic workflow for calculations performed in this work. SSP objective and SSP gradient method uses an exponential averaging method to calculate free energies. Full FEP uses the MBAR estimator.

3.2.3 Optimization

The objective function and the constraints of the optimization performed in this chapter are shown in equation 3.1-3.4,

$$\min_{q_i} = \Delta G_{binding}(q_i) - \Delta G_{original}, \quad (3.1)$$

$$s.t. \quad \sum_{i=1}^M = net\ charge, \quad (3.2)$$

$$\sqrt{\frac{\sum_{i=1}^M (q_{i,0} - q_{i,n+1})^2}{M}} \leq rmsd\ limit, \quad (3.3)$$

$$|(q_{i,n} - q_{i,n+1})| \leq 0.01e. \quad (3.4)$$

In the objective $\Delta G_{binding}(q_i)$ was defined as the difference in free energy between the bound and unbound states of the ligand with charges q_i . $\Delta G_{original}$ was defined as the difference in free energy between the bound and unbound states of the ligand with the charges of the original unoptimized ligand, $\Delta G_{original}$ is thus a constant. This objective function is a relative binding free energy which was calculated, as discussed in section 1.4.1, as the difference of $\Delta G_{mutation}$ in the bound and unbound states. For brevity we will now refer to the relative binding free energy calculated in the objective as $\Delta \Delta G_{opt}$. In the objective and constraints q_i are the charges of the ligand and $q_{i,n}$ the charges for iteration n of the optimization, M is the number of atoms in the ligand and *net charge* the total *net charge* of the ligand. Equation 3.2 constrains the *net charge* of the ligand to be constant. Equation 3.3 constrains the root mean squared difference between the ligands original charges, $q_{i,0}$, and the ligand's charges in the next optimization step, $q_{i,n+1}$, to be less than some value *rmsd limit*. These *rmsd limits* were chosen to limit the change in binding free energy to a sensible range < 10 kcal/mol. Without this limit, the optimization continued to very large unphysical values of binding free energy because the atomic partial charges can reach unphysical values. Equation 3.4 bounds the perturbation to each atom to be less than $0.01e$ per optimization iteration, where e is the elementary charge. This constraint ensured that the optimizer never took a large step for which a converged value of the objective, $\Delta \Delta G_{opt}$, could not be calculated using SSP.

With a hand picked limit of $0.01 e$ in equation 3.4 a determination of how much sampling was required to give converged calculation of $\Delta \Delta G_{opt}$ with a perturbation to each atom of $0.01 e$ was made in figure 3.2. The amount of sampling needed was determined to be 2.5 ns, this was then the amount of sampling used in this work to calculate the objective and gradient for each optimization step. We will now discuss the details of how this SSP calculation can be used to efficiently calculate the objective and gradient for this optimization.

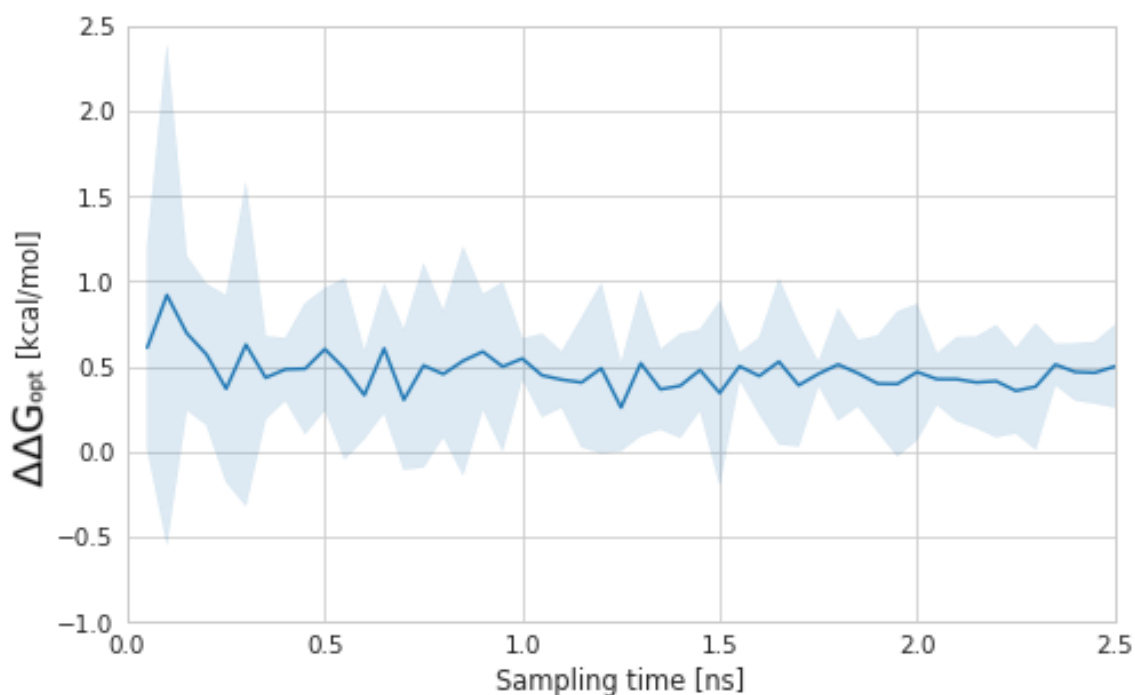


Fig. 3.2 Convergence of the $\Delta\Delta G_{opt}$ predictions in the Factor Xa test case for a perturbation of +0.01 e to half the atoms and -0.01 e to the other half (maintaining the net charge) as the simulation time is increased, calculations were performed at 0.01 ns and then from 0.05 ns to 2.5 ns in 0.05 ns increments. The values of $\Delta\Delta G_{opt}$ are reported as mean of three replicates with the shaded area showing the 95% confidence interval computed as $\text{mean} \pm t_2 \cdot \text{SEM}$, where t_2 is the t-distribution statistic with two degrees of freedom, and SEM is the standard error of the mean computed from the sample standard deviation of the three independent replicate predictions.

3.2.4 SSP Objective

The objective was calculated with a straightforward application of the Zwanzig equation to calculate $\Delta G_{\text{mutation}}$ in the bound and unbound states as follows,

$$\Delta G_{\text{mutation}} = -k_b T \log \langle e^{-\beta(U_{\text{perturbed}} - U_{\text{unperturbed}})} \rangle_{\text{unperturbed}}, \quad (3.5)$$

here $U_{\text{perturbed}}$ and $U_{\text{unperturbed}}$ are the potential energies of the system calculated using the Hamiltonian of the perturbed system and the unperturbed system, respectively. To change Hamiltonian, the charges were switched from unperturbed, $q_{i,n}$, to perturbed values, $q_{i,n+1}$. However, Lennard-Jones, bonded, angle, and torsion parameters did not change. We combined $\Delta G_{\text{mutation}}$ in the bound and unbound states as seen in figure 1.2 and this gave a value for $\Delta\Delta G_{\text{opt}}$.

Since an SSP method is being used, efforts were made to avoid poor overlap between the end states of any perturbations. To achieve this two ideas were used 1) constraints were applied to the optimization and the relevant constraint has already been discussed, see equation 3.4. 2) the system was re-sampled after every optimization step with 2.5 ns. If this re-sampling was not done then the difference between the perturbed and unperturbed systems would grow over the course of the optimization, reducing the overlap in phase space and so reducing the applicability of SSP. Resampling also had one advantage, as it allowed for a calculation of a reverse alchemical step. Therefore, $\Delta\Delta G_{\text{opt}}$ for both the forwards and backwards alchemical transformation were calculated for every step and the $\Delta\Delta G_{\text{opt}}$ in the results are reported as an average of the forwards and backwards transformations.

This need to re-sample is not ideal and reduces the efficiency of applying SSP. For the calculation of the objective, SSP may not be the best choice of method as the size of any optimization step is limited by the need to maintain good overlap between the end states of the calculation of the objective. In chapter 4, we will explore the use of MBAR to calculate the objective and this will eliminate the problems highlighted with SSP here. Despite some of the drawbacks of SSP mentioned here, one area in which SSP is extremely useful is for calculating the gradient of the objective and we will explore this idea in the next section.

3.2.5 SSP Gradient

The gradient of the objective with respect to the force field parameters, $\Delta\Delta G_{\text{grad}}$, was calculated as shown in equation 3.6. For a force field parameter, here we use the partial charge, q , note that this could be any parameter which is an argument of the potential.

$$\nabla(\Delta G_{binding}(q_i) - \Delta G_{original}) = \frac{\Delta G_{binding}(q_i + h) - \Delta G_{binding}(q_i)}{h}. \quad (3.6)$$

Here h is a finite difference of $0.00015 e$. This is chosen by hand by observing the variation in the gradient as this value is changed, a value of $0.00015 e$ was chosen as the resulting calculations of the gradient were observed not to depend on small changes to this value. The numerator of the RHS of equation 3.6 is a $\Delta\Delta G_{binding}$ and can be calculated using an SSP approach, using sampling only from a central state containing the ligand using values q_i as detailed in equation 3.5. This calculation of the gradient shows the advantage of SSP as numerous (10s-100s) evaluations of the RHS of equation 3.6 are required, one for each dimension in q space, and they can all be calculated from the sampling of one central state. The number of dimensions in q space comes from the number of optimized atoms in the molecule. Figure 3.3 shows the ligand from the androgen receptor test case where as an example there are 10 named hydrogens which could be optimized and therefore there would be a 10 dimensional q space.

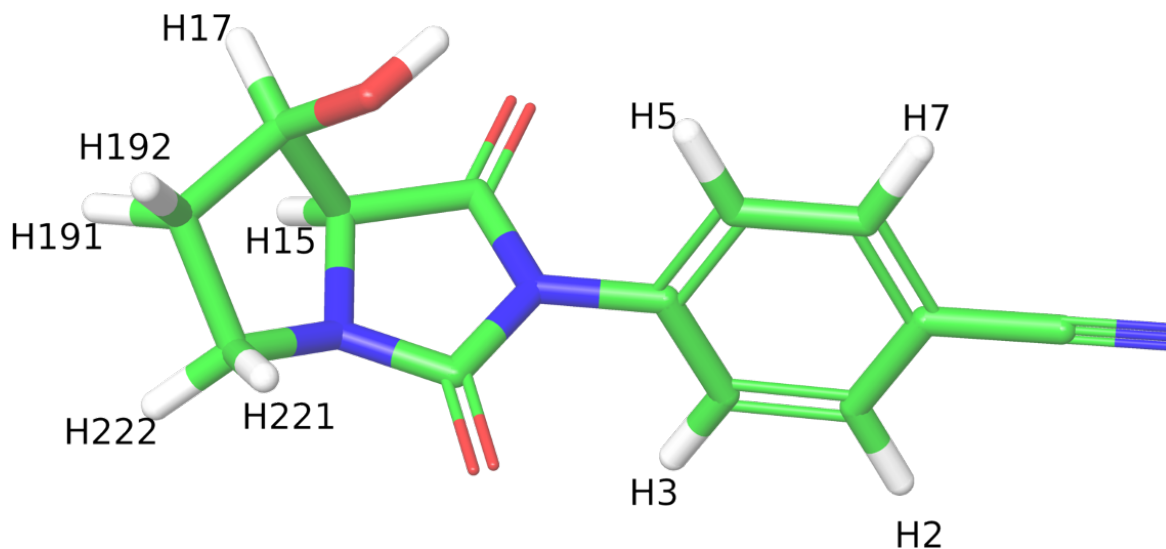


Fig. 3.3 3D structure for the androgen receptor inhibitor with example optimized hydrogens explicitly labeled.

The finite difference in equation 3.6 is between molecules that are extremely similar, differing only by $0.00015 e$ in one atom's charge parameter. There is therefore likely to be a large sampling overlap between these states allowing SSP to be applied. Of note is that for each finite difference calculation the charge of the simulation box has been changed by $0.00015 e$. The potential for finite size effects [195] caused by this change were investigated

and the padding of the simulation with solvent chosen to negate these effects, see appendix figure B.10.

3.2.6 Optimization Validation

For each optimization, we would like to inspect the convergence of the result over the number of optimization steps. In the methodology, mention was made to limiting the RMSD between the original and optimized charges to some value *rmsd limit*. The values which are chosen for the *rmsd limit* are 0.01, 0.03, and 0.05 *e*. The optimization is therefore repeated with the RMSD bound to these three values. With an *rmsd limit* of 0.01 *e*, the optimizer is limited to seven steps, as adequate convergence is seen at this point. For the larger values of *rmsd limit*, the convergence is slower and, therefore, for optimizations with an *rmsd limit* of 0.03 and 0.05 *e* the optimizer is limited to 20 steps. A good metric to analyse the results of the optimization is the set of optimized charges taken as a vector to assess convergence across simulation steps, we take the dot product of the normalized vector of new charges with the normalized vector of original charges for each step of the optimization and present the result in figure 3.4.

Figure 3.4 shows that the direction of the charge vectors over all systems and values of *rmsd limit* are well converged. The direction of these charge vectors represents where the charge is being applied on the molecule and this is the information that will be used in the results section to make chemical mutations to improve the binding affinity of these ligands. It can also be seen that the dot product between the original and optimized charges is different for different values of *rmsd limit*. To quantify this difference, the dot product between the set of optimal charges obtained for *rmsd limit* values of 0.01 with 0.03 and 0.05 *e* can be taken and, the results of these projections can be seen in table 3.2.

In table 3.2 we can see the dot product of the optimized charges from the optimization with an *rmsd limit* of 0.01 *e* with themselves returns 1.00 as expected. The dot product of the vector of charges with *rmsd limit* = 0.01 *e* with *rmsd limit* = 0.03 *e* also returns 1.00 as these vectors are extremely similar in direction. The dot product of the vector of charges with *rmsd limit* = 0.01 *e* with *rmsd limit* = 0.05 *e* returns approximately 1.00 as these vectors are extremely similar in direction but not as close as 0.01 *e* with 0.03 *e*. Therefore we conclude from the results in table 3.2 that sets of charges for the same system are pointing in the same direction. Thus only the value of the charge changes are dependent on the *rmsd limit* value, whilst the direction and relative magnitude of the charge changes are completely consistent. This is an important result because it shows that the design principles identified by the approach will not depend on the arbitrary choice of the *rmsd limit*. The invariance in where the charge is being applied can also be seen by eye if the atoms are colored by change

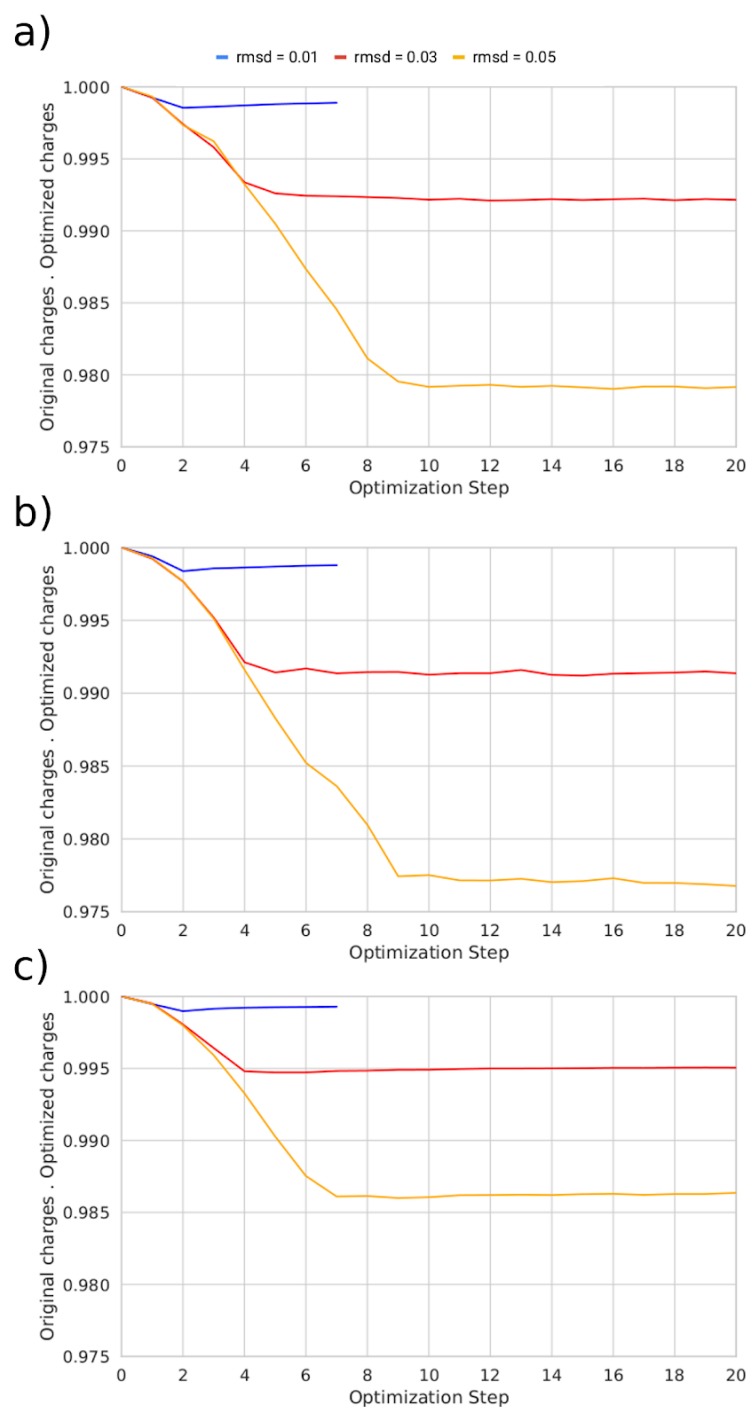


Fig. 3.4 Dot product of the normalized optimized charges with the normalized original charges showing variation of charge vector direction with step. Results are shown for RMSD limits 0.01, 0.03 and 0.05 e . With the FXa, P38 and androgen receptor systems labeled a), b) and c) respectively.

Table 3.2 Dot products of the normalized vector of optimal charges using an *rmsd limit* of 0.01 e with the normalized vector of optimal charges using different *rmsd limit* values.

RMSD [e]	FXa	P38	Androgen Recp.
0.01	1.00	1.00	1.00
0.03	1.00	1.00	1.00
0.05	0.99	0.98	0.99

in charge. Figures illustrating this are presented in the appendix figures B.1-B.9 for all sets of optimized charges.

To inspect the convergence of the optimization we can also look at the value $\Delta\Delta G_{opt}$ over the course of the optimization, and this can be seen in figure 3.5. With an *rmsd limit* of 0.01 e figure 3.5 demonstrates that $\Delta\Delta G_{opt}$ is well converged for all systems. For *rmsd limits* of 0.03 or 0.05 e, the results are only well-converged for the androgen receptor system, figure 3.5c. This suggests that $\Delta\Delta G_{opt}$ for the optimized set of charges is dependent on the value of *rmsd limit*, and that $\Delta\Delta G_{opt}$ is slow to converge for larger ligands such as those in the p38 and FXa test cases.

Additional calculations were made to verify the cumulative sum of $\Delta\Delta G_{opt}$ over all optimization steps. This involved calculating the relative binding free energy between the original and optimized parameters of the inhibitor with MBAR. The mean absolute error between the values of $\Delta\Delta G_{opt}$ calculated by SSP and the verification calculation calculated with MBAR was 1.04 kcal/mol. Details and results of this calculation can be seen in the appendix table B.1.

3.2.7 FEP Calculations

The optimized charges are used to generate design ideas for chemical mutations. These chemical mutations are applied to the ligands with the aim of improving their binding free energy. The relative binding free energy change of the ideated compounds was tested using full FEP calculations and analysed with MBAR. We term the calculated binding free energy for these tests FEP $\Delta\Delta G$. These full FEP calculations share their name with the full FEP calculations discussed in chapter 2 because these calculations use the same methodology and code.

To perform these calculations the charges, van der Waals and bonded terms were all interpolated simultaneously from the original to the designed state. All windows were sampled independently with 2 ns of Langevin dynamics, giving a total of 24 ns of sampling per mutant, unless specified otherwise. In the case of hydrogen to fluorine mutations

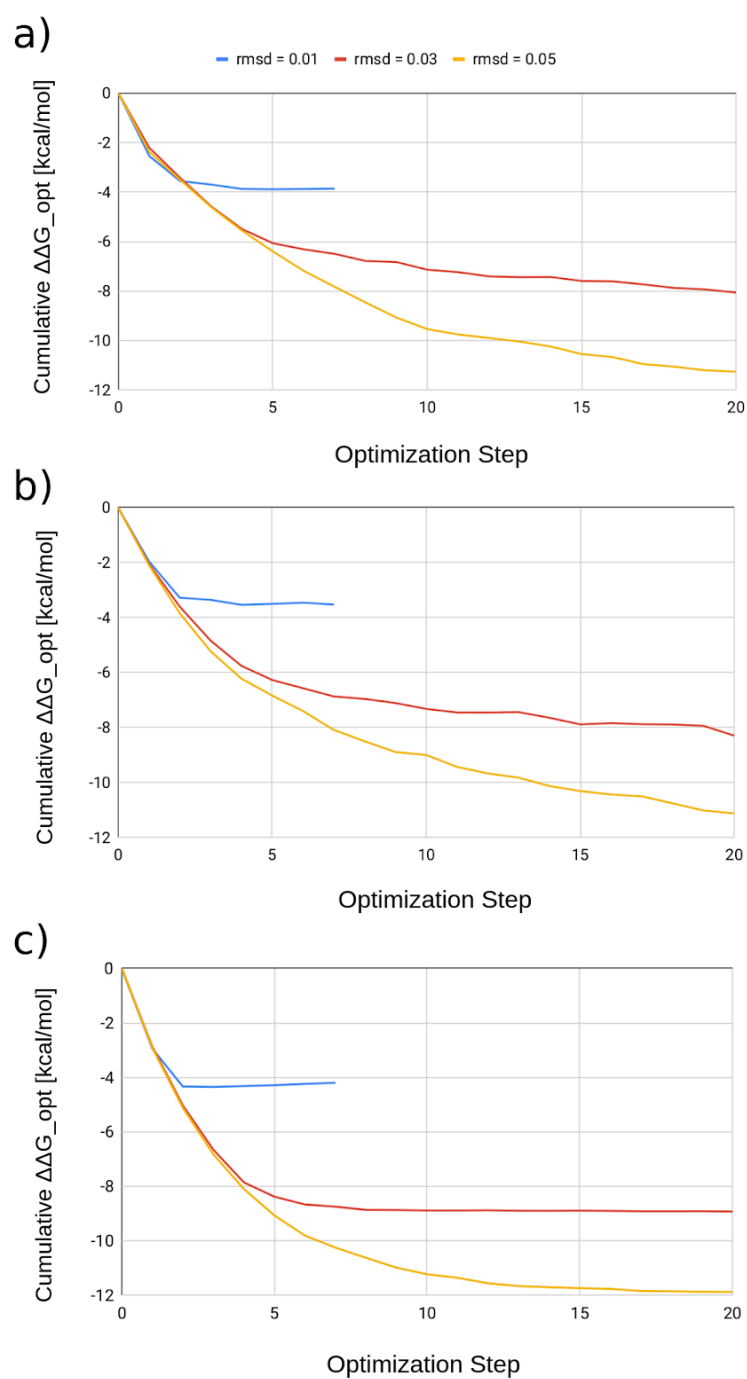


Fig. 3.5 Cumulative sum of $\Delta\Delta G_{opt}$ averaged over three replicates for each step of the optimizer. Three optimizations are shown with RMSD bound to 0.01, 0.03 and 0.05 e . With the FXa, P38 and androgen receptor systems labeled a), b) and c) respectively.

Table 3.3 Abbreviated terms used to reference various relative binding free energies calculated in this work with a brief description of the calculation for the reader’s reference

Free energy value	Description
$\Delta\Delta G_{opt}$	The value for the objective. Calculated between the original and optimized charges with SSP.
$\Delta\Delta G_{grad}$	The value for the gradient. Calculated between many highly related ligands using SSP.
FEP $\Delta\Delta G$	The value for mutations calculated with MBAR.
$\Delta\Delta G_{exp}$	The value for mutations determined experimentally.

the original hydrogen was constrained, therefore its associated C-H bond could not be interpolated to a C-F bond. When neglecting the interpolation of this bond the fluorine appeared at the position of the hydrogen, instead of the true physical position of the fluorine. To avoid this issue, for mutations involving constrained hydrogen, we used the same hybrid topology approach discussed in chapter 2.

3.2.8 Summary of Methods

To summarize, in this chapter an optimization of the charge parameters of several ligands was performed. The objective of this optimization was to find a set of charges to minimize the relative binding free energy of the ligand to a protein. This objective, $\Delta\Delta G_{opt}$, and gradient, $\Delta\Delta G_{grad}$, were both computed using SSP calculations. The results of the optimizer are used to predict beneficial areas on the ligand for chemical mutations. To validate these predictions the relative binding free energy for these mutations were computed using FEP and the MBAR estimator and we call these calculations full FEP calculations and denote the values calculated as FEP $\Delta\Delta G$. For the reader’s reference table 3.3 contains all the abbreviated terms used and a brief description of the calculation.

3.3 Results

To visualize the results of the optimization, the sets of optimal charges are used to color the atoms of the ligands and these visualizations are shown in figure 3.6. We developed specific design ideas to improve $\Delta\Delta G_{binding}$ based on the changes in charge. First analyzing the Fxa ligand, three options are selected:

- Replacing the hydrogen with a fluorine at position 1a.

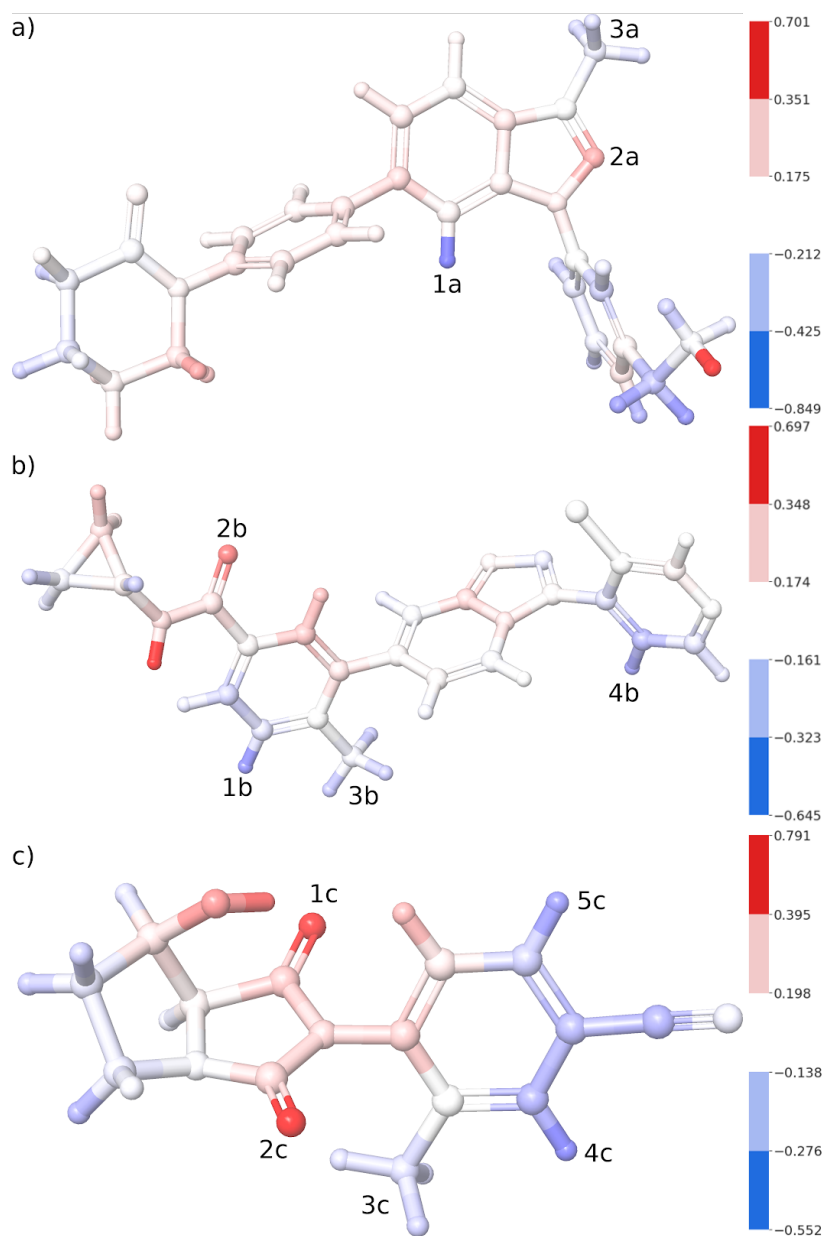


Fig. 3.6 Panels a), b) and c) show the FXa, P38 and androgen receptor ligands with atoms colored by change in charge relative to the original partial charges. The optimised charge is taken from the optimisation with RMSD bound to $0.03 e$. Blue represents atoms which are more negative and red represents atoms which are more positive. Selected sites for chemical modification are numbered.

- Replacing the nitrogen with a carbon at position 2a.
- Replacing one or more of the hydrogens with a fluorine atom on the methyl group at position 3a.

Analyzing the P38 ligand, four options are selected:

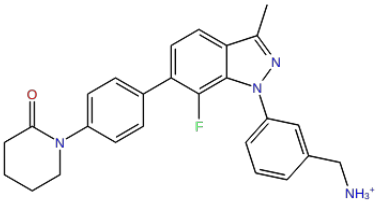
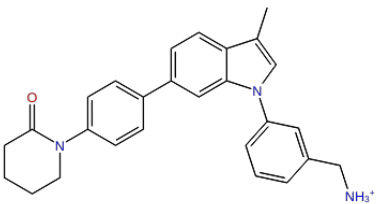
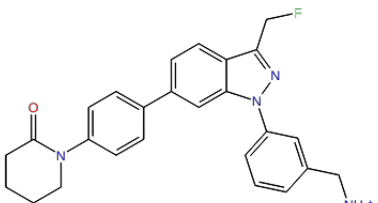
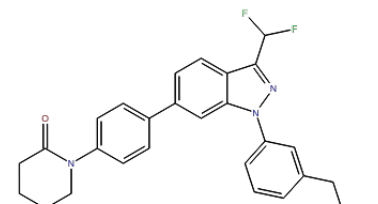
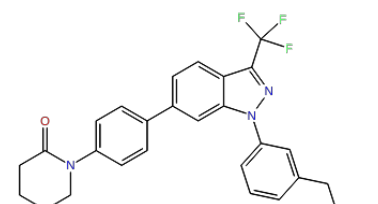
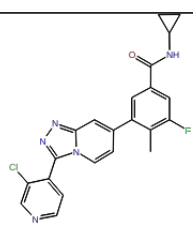
- Replacing the hydrogen with a fluorine at position 1b or 4b.
- Replacing the carbon with a nitrogen at position 1b or 4b.
- Replacing the oxygen with a sulphur at position 2b .
- Replacing one or more of the hydrogens with a fluorine atom on the methyl group at position 3b.

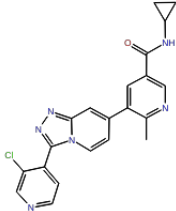
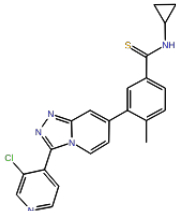
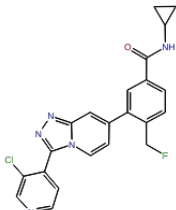
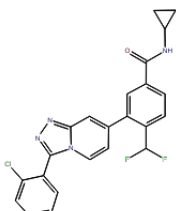
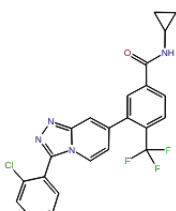
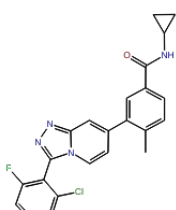
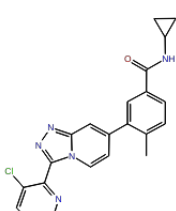
The final set of changes apply to the ligand of the androgen receptor with three options selected:

- Replacing the oxygen with a sulphur at position 1c and 2c
- Replacing the hydrogen with a fluorine at position 3c, 4c or 5c.
- Replacing the bonded carbon with a nitrogen at positions 4c or 5c.

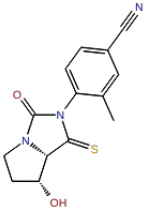
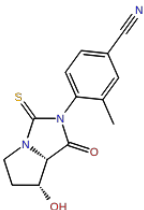
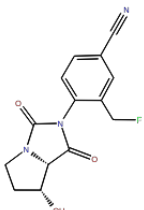
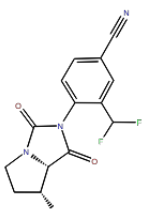
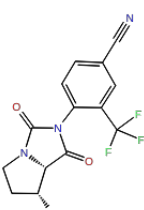
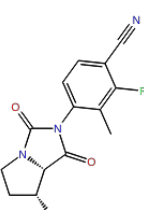
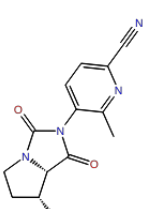
FEP $\Delta\Delta G$ for all of these changes was calculated using the FEP protocol discussed in the methods section. Each FEP calculation was performed in triplicate and the averaged results of these calculations can be seen in table 3.4.

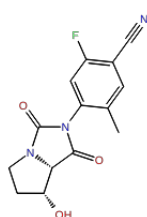
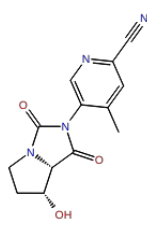
Table 3.4 FEP $\Delta\Delta G$ for proposed chemical mutations to the FXa, P38 and androgen receptor ligands calculated with FEP. The positions denoted numerically correspond to numerical positions in figure 3.6. FEP predictions are reported as the mean value of three replicates with 95% confidence interval reported between square brackets computed as mean \pm t2· SEM, where t2 is the t-distribution statistic with two degrees of freedom, and SEM is the standard error of the mean computed from the sample standard deviation of the three independent replicate predictions. The asterisk label * indicates single or double fluorinations of a methyl. These are averaged over every hydrogen or pair of hydrogen in the methyl and as such this data represents the averaging of nine replicates with the confidence interval reported such that t2 is now the t-distribution statistic with eight degrees of freedom. The obelisk label † denotes calculations that were slow to converge and run with 24 lambda windows of 2 ns. The diesis ‡ label denotes data taken from the previous chapter.

Position and mutation	Mutant	FEP $\Delta\Delta G$ [kcal/mol]
(1a) Hydrogen to fluorine		-2.2 [-2.3, -2.1]‡
(2a) Nitrogen to carbon		3.0 [1.9, 4.1]
(3a) Hydrogen to fluorine		-0.1 [-0.2, 0.0]*
(3a) Double hydrogen to fluorine		-0.6 [-0.7, -0.5]*
(3a) Triple hydrogen to fluorine		-0.8 [-1.2, -0.5]
P38		
(1b) Hydrogen to fluorine		-2.2 [-2.7, -1.6]‡

(1b) Carbon to nitrogen		2.0 [1.2, 2.7]
(2b) Oxygen to sulphur		0.0 [-2.7 , 2.7]†
(3b) Hydrogen to fluorine		0.3 [0.1, 0.4]*
(3b) Double hydrogen to fluorine		-0.5[-0.9, -0.1]*
(3b) Triple hydrogen to fluorine		1.0 [-0.1, 2.1]
(4b) Hydrogen to fluorine		-1.6 [-1.7, -1.4]‡
(4b) Carbon to nitrogen		-0.4 [-1.7, 1.0]

Androgen Receptor

(1c) Oxygen to sulphur		-2.1 [-3.3 , -0.9]
(2c) Oxygen to sulphur		-2.0 [-3.2 , -0.8]
(3c) Hydrogen to fluorine		-0.6 [-0.8, -0.4]*
(3c) Double hydrogen to fluorine		-1.6[-1.8 , -1.4]*
(3c) Triple hydrogen to fluorine		-0.1 [-0.9, 1.0]
(4c) Hydrogen to fluorine		-2.5[-2.8, -2.1]‡
(4c) Carbon to nitrogen		-0.9 [-1.4, -0.5]

(5c) Hydrogen to fluorine		-0.3 [-0.3, -0.2]‡
(5c) Carbon to nitrogen		1.1 [0.0, 2.2]

The atoms indicated by the optimization to beneficially be more negative in figure 3.6 line up with experimental work on these test cases [159, 169, 173]. Mutants 1, 6, and 19 are predicted by full FEP to be beneficial (-2.2 kcal/mol, -2.2 kcal/mol, and -2.5 kcal/mol respectively) and this is in good agreement with experimental data (-2.1 kcal/mol, -2.3 kcal/mol, and -1.1 kcal/mol respectively). Experimental data does not exist for the remaining proposed mutations. However, 73% of the mutations in table 3.4 are predicted to be favourable by FEP. Both the FXa and androgen systems have a higher success rate with 80% and 89% of ideas from charge optimization being beneficial as assessed by FEP respectively. P38 has a lower (though still promising) success rate with 50% of mutations being beneficial as assessed by FEP.

3.4 Conclusion

We have demonstrated ligand charge optimization in explicit solvent to be a useful tool to rationally design ligands with improved binding affinities. The electrostatics of three ligand receptor systems were systematically optimized using the alchemical SSP method, yielding sets of optimal ligand charges. These sets of optimal charges were used to generate design principles for chemical mutations to the ligands that would yield improved receptor binding affinity. These chemical mutations were assessed with a more rigorous FEP method. Using FEP, 73% of the predicted chemical mutations were found to be beneficial. The average improvement of the beneficial mutations was approximately 1 kcal/mol. In three of these cases, experimental data exists and is in excellent agreement with calculations, with mutants 1, 6, and 19 in table 3.4 predicted by FEP to be beneficial (-2.2 kcal/mol, -2.2 kcal/mol, and -2.5 kcal/mol respectively) compared to the experimental data (-2.1 kcal/mol, -2.3 kcal/mol, and -1.1 kcal/mol respectively).

The major advantage of SSP shown in this work is in the calculation of the gradient. SSP allows for many highly related mutations to be assessed quickly, as is required to calculate the gradient via a finite difference method. For comparison, to collect 2.5 ns of sampling for the FXa system with 99,000 and 13,000 atoms in the complex and solvent systems respectively takes 29 minutes. To calculate a gradient from this sampling takes 15 minutes and so, including sampling, this totals to 44 minutes per gradient. To calculate the free energy change, for a perturbation of 0.00015 e to one ligand atom, with full FEP (assuming 1.0 ns of sampling converges the $\Delta\Delta G_{binding}$ calculation, see appendix figure B.11 for convergence plot) takes 14 minutes. The FXa ligand contains 58 atoms with partial charges these must all be perturbed in the complex and solvent systems, so in total each gradient calculated with FEP would take approximately 27 hours. This advantage would only be compounded if a more complex optimization scheme, which required a calculation of the Hessian, was used. Both these example FEP and SSP calculations of the gradient are run in parallel, see the parallelization chapter for detail, across 4 NVIDIA P100 GPUs using OpenMM 7.3.0 and CUDA 10.0.

Beyond the calculation of the gradient the overall saving in computer time can be seen by considering that to test one fluorinated analogue of the FXa ligand with full FEP, takes roughly 4 hours, as calculated in chapter 2. Therefore to test all fluorinated analogues would take roughly three days of computer time. Using the optimization method presented here we can assess qualitatively the benefit of adding, not only a fluorine at all possible positions but also nitrogens, sulphurs and potentially any charged atoms or groups. Assuming the optimization needs 48 steps, including calculations for all values of *rmsd limit*, with each step for FXa taking less than 1 hour, this equates to roughly two days of computer time. It can therefore be seen that the method presented here allows for design ideas to be generated quickly, assessing the relative benefit of mutating atoms on a ligand far quicker than traditional FEP. This increased speed represents a significant gain in efficiency for the exploration of chemical space. Additionally the ideas generated are easy to interpret when compared to faster ML methods. This is because the inverse QSAR problem we have created here (mapping atom type onto the real numbers) was easily solvable by eye. Problems may arise if the optimizer were to pick very unphysical charges however this has been addressed here with some simple constraints.

Chapter 4

Steric Optimization

4.1 Introduction

In this chapter, a novel method to rationally design inhibitors with improved steric contacts and enhanced binding free energies is presented. This new method uses alchemical single step perturbation calculations to rapidly optimize the van der Waals interactions of a small molecule in a protein-ligand complex in order to maximize its binding affinity. The results of the optimizer are used to predict beneficial growth vectors on the ligand and good agreement is found between the predictions from the optimizer and a more rigorous free energy calculation, with a Spearman's rank order correlation of 0.59. We use the Spearman's rank order correlation under the assumption that the rank order of compounds would be the most useful information to the drug discovery campaigns where we envisage these methods could be applied. The advantage of the method presented here is the significant 10x speed up over more rigorous free energy calculation and sublinear scaling with the number of growth vectors assessed. Where experimental data was available, mutations from hydrogen to a methyl group at sites highlighted by the optimizer were calculated with MBAR and the mean unsigned error between experimental and calculated values of the binding free energy was 0.83 kcal/mol.

This work has significant overlap with the charge optimization method presented in the previous chapter. Whilst the objective function remains the same here the primary difference is that we now consider this objective as a function of the steric radius of the ligand atoms, σ . Optimizing the binding free energy with respect to σ introduces some new challenges. These challenges stemmed mostly from the larger perturbations to phase space introduced by perturbing the steric parameters compared to charge parameters, in addition to the larger size of the groups any design ideas might suggest. For example previously we have considered adding mostly fluorine, which we know to be a small perturbation treatable with SSP, however

in this chapter the design ideas will point to the addition of larger groups such as methyls. To address these challenges, several changes are made to the previous charge optimization methodology and these changes are detailed in the methods section 4.2.

Given the overlap with the charge optimization chapter similar comparisons can be drawn between this work and the charge optimization work of Tidor *et al.* This work builds on the work of Tidor *et al.* by adopting more contemporary molecular dynamic methodologies such as flexible ligand/receptor, explicit water and PME electrostatic. Other relevant comparisons can be drawn between this work and any other computational methods which grow ligands with the aim of improving binding affinity. Whilst this work focuses specifically on the hit-to-lead subset of the drug discovery pipeline, we will draw some comparison here to the broader set of *de novo* design programs discussed in section 1.2.4. In section 1.2.4, we discussed how *de novo* methods score ligands based on the complementarity with the receptor and mentioned a review of these methods by Schneider *et al.* [38], which commented that, receptor based *de novo* methods generally aim to approximate the full binding free energy in their scores for affinity. The work presented here does not make this approximation and instead the binding free energy is calculated explicitly using alchemical free energy methods.

One additional variable in the design of *de novo* algorithms not discussed in section 1.2.4 was how affinity scores are used to build an optimal ligand. Whilst there exists several methods which are used to build ligands such as fragment linking [36], for this work we are most interested in ligand growing methods. Ligand growing, as the name suggests, involves adding chemical entities to the ligand sequentially to improve the complementarity score. The chemical entities can range from individual atoms [196] to whole chemical groups [197]. In the former case some efforts are typically made to restrict the exploration of chemical space and limit combinatorial explosions in the number of possible ligands that could be generated [196]. In this chapter, these combinatorics were dealt with by only considering the addition of methyl groups.

The method used to build tight binding ligands in this work is similar to the growth methods used in previous work, however, only growths motivated by the optimization of the binding free energy with respect to the steric parameters of the ligand will be explored. Using information from the gradient and optimization allows for a more directed exploration of chemical space and should reduce the number of points which must be tested to find a beneficial ligand mutation.

4.2 Methods

The goal of this work is to optimize the steric parameters of a ligand in a protein-ligand complex and use the results of this optimization to predict beneficial growth vectors on the ligand more efficiently than existing FEP methods. To this end experimental data has been collected to retrospectively validate this method from the following systems: androgen receptor (AR)[198, 199], renin[166], menin[168], thrombin[200], and SARS PL protease[201, 202]. These systems were chosen as they contain data for free energy differences for changes from hydrogens to methyl groups. Two calculations were performed with menin using two different ligands, these are termed menin A and B. Four calculations were performed for the thrombin systems using four different ligands, these are termed thrombin A, B, C and D. All ligands used in this work are shown in table 4.1.

Table 4.1 Ligands used in androgen receptor, SARS PL protease, renin, menin and thrombin systems.

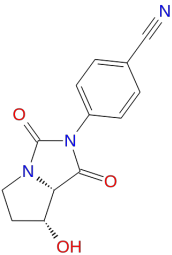
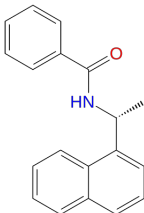
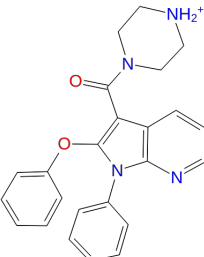
System	Ligand ref.	2D structure
Androgen receptor	A	
SARS PL Protease	A	
Renin	A	

Table 4.2 PDBIDs for each system used in this work.

System	PDBID
Androgen receptor	2NW4 [199]
SARS PL Pro	3E9S [202]
Renin	3OOT [166]
Menin	4OG6 [168]
Thrombin	2ZNK [200]

this is now replaced with a custom system builder developed for this work. This was done to ensure no variation in the preparation of systems for $\Delta\Delta G_{opt}$ and MBAR FEP calculations. Additionally water is now modelled with the SPC/E water model and the edge of the solvation box was set to be 10.0 Å from any atom of the receptor or ligand. Van der Waals interactions were truncated at 10.0 Å. Electrostatics were modelled using the particle mesh Ewald method with a cutoff of 10.0 Å. The PDBIDs for the crystal structures of the systems used as input to this work are shown in table 4.2.

4.2.2 Workflow

In this work, several free energy methods were used in different contexts. For clarity we will now outline these methods for future reference. After models for the protein ligand system were built we performed an optimization of the van der Waals parameters of an inhibitor in the protein ligand complex. This is done using a gradient descent (GD) algorithm with a line search. In order to perform this optimization two calculations are needed: one for the objective and one for the gradient. The objective was calculated with MBAR we therefore name this the MBAR objective and the gradient was calculated with SSP we therefore name this the SSP gradient. The details of how these calculations are performed can be found in the optimization, SSP gradient and MBAR objective sections below. Once this optimization is converged, the output is analysed to determine growth vectors which should be beneficial. We then calculate the relative binding free energy of adding methyl groups at the locations specified by the optimizer. This calculation is performed using MBAR with Hamiltonian replica exchange. We name these calculations FEP scans with more details in the FEP scans section below. Figure 4.1 illustrates how each of these calculations are combined in the full workflow.

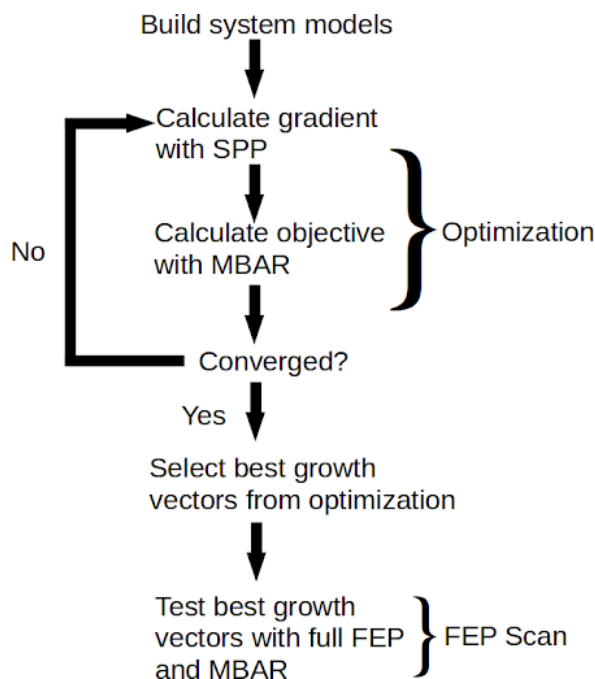


Fig. 4.1 Diagrammatic workflow for calculations performed in this work. SSP gradient method uses an exponential averaging method to calculate free energies. MBAR objective and FEP scan both use the MBAR estimator.

4.2.3 Optimization

Optimizations were performed using the Ligand-Optimiser code which is available on Github at <https://github.com/adw62/Ligand-Optimiser>. The optimizations were applied to the σ parameter of the LJ potential shown in the equation 1.6. Note that this method is agnostic to the exact form of the LJ potential used and the exact meaning of σ in that form. The objective function for the optimization performed in this work is defined in equation 4.1,

$$\min_{\sigma_i} \Delta G_{binding}(\sigma_i) - \Delta G_{original}, \quad (4.1)$$

where $\Delta G_{binding}(\sigma_i)$ is the binding free energy difference between the bound and unbound states of the ligand and receptor for a ligand with M atoms using σ_i parameters. $\Delta G_{original}$ is the binding free energy for the ligand using its original parameters and therefore is a constant for each given system.

To perform this optimization an implementation of the gradient descent algorithm with a line search was used. The line search used a step size of 0.6 nm and a convergence tolerance of 0.15 nm in σ . The objective function in equation 4.1 can be considered as the difference in binding free energies between two ligands with binding free energy $\Delta G_{binding}(\sigma_i)$ and

$\Delta G_{original}$. From this point forward we define this relative binding free energy as $\Delta\Delta G_{opt}$. As previously seen in this work $\Delta\Delta G_{opt}$ was calculated as the difference between $\Delta G_{mutation}$ in the bound and unbound states. In this chapter we calculate the objective using MBAR and the gradient using SSP and the following section discusses these calculations in more detail.

4.2.4 MBAR Objective

In chapter 3 the objective, $\Delta\Delta G_{opt}$, was calculated with SSP we now switch to MBAR. There are two reasons for this: 1) we are generally dealing with larger perturbations in the sterics, which are harder to converge, 2) using MBAR should allow for the optimizer to take large, potentially more efficient, steps per iteration without the poor convergence that SSP suffers from.

The objective was calculated by simulating 24 alchemical windows, the two end states were a) the system with the Hamiltonian using the sigmas for the current optimization step n and b) the system with the Hamiltonian using the sigmas from optimization step $n + 1$. The intermediate states are created by linearly interpolating the sigma parameters only. Charge, bond, angle and torsion parameters do not change. The sigmas for the optimization step $n + 1$ were determined with the standard GD algorithm using a maximum step size, α of 0.6 nm. Sampling from all windows was collected and the potentials evaluated from this sampling were combined using the MBAR estimator giving a value for $\Delta G_{mutation}$. Combining $\Delta G_{mutation}$ from the bound and unbound states gave $\Delta\Delta G_{opt}$ values across all lambda windows. The lambda windows of the free energy calculation were treated as a line search in sigma space and therefore the lambda window with the minimum in $\Delta\Delta G_{opt}$ was selected. The sigma parameters corresponding to that lambda window then became the accepted parameters for the next step of optimization. If the minimum was found to be at the last lambda value then the line search was extended by another 0.6 nm without recalculating the gradient. An example of this line search is shown in figure 4.2, where the σ s corresponding to the seventh window in a line search using 12 MBAR windows would be selected as, $\sigma_{n+1} = \sigma_n - \frac{7}{12} * \alpha * gradient$.

Each evaluation of the objective used 6 ns of sampling in AR, SARS PLPro, and menin A systems. Renin, menin B and the four thrombin systems used 12 ns as these optimizations were observed taking steps which correspond to larger values of $\Delta\Delta G_{opt}$. Figure 4.3 shows an example for how a calculation for $\Delta\Delta G_{opt}$ varies with sampling.

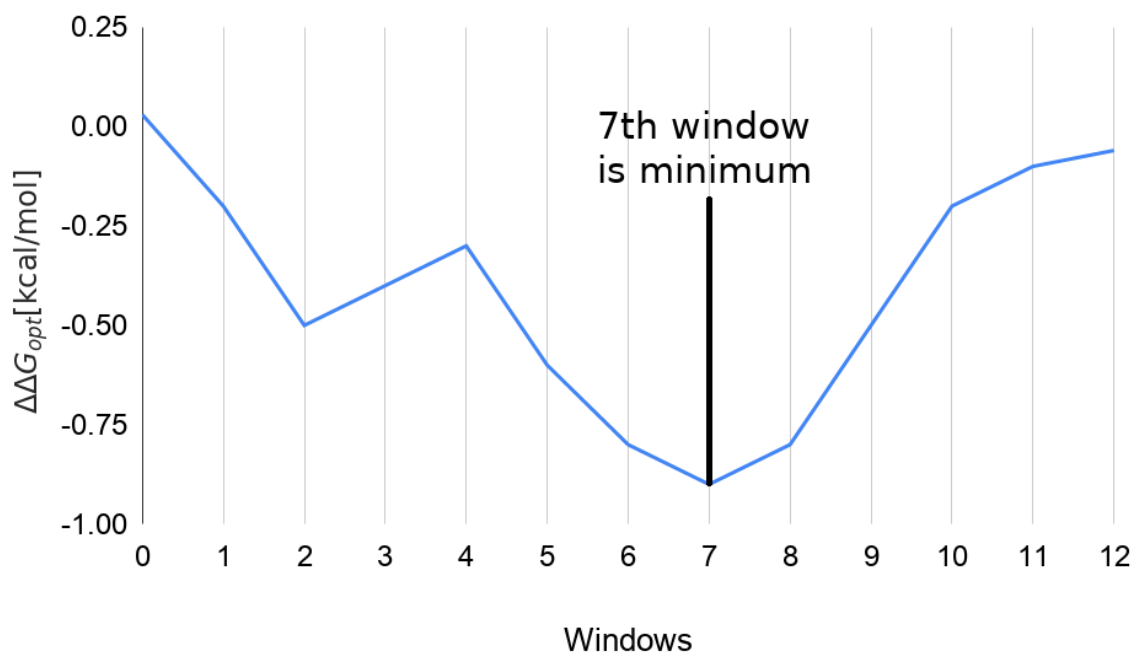


Fig. 4.2 Example line search in GD algorithm. Here the σ s corresponding to the seventh window would be selected for the next step of the optimizer.

4.2.5 SSP Gradient

The calculation of the SSP gradient is the same as in the Charge Optimization chapter 3. However the gradient is now taken with respect the σ parameters of the ligand using a finite difference of 0.00015 nm. Figure 4.4 shows an example for how this gradient varies with sampling for the androgen receptor ligand.

4.2.6 Optimization Validation

To test the reproducibility of the optimization we examined how the optimized values for the sigma of each hydrogen varied across repeats. Three repeats of an optimization for the androgen receptor were made and the averaged set of optimized σ are presented in table 4.3. We can see from table 4.3 that the variance in the optimized σ is small relative to the difference between original and optimized σ values. We are therefore confident that the solution reached by the optimizer is consistent.

To ensure each optimization is converged over optimization iteration the cumulative $\Delta\Delta G_{opt}$ is plotted over optimization steps and presented in figure 4.5. We can see from 4.5 that for all systems the optimization is well converged within ten iterations.

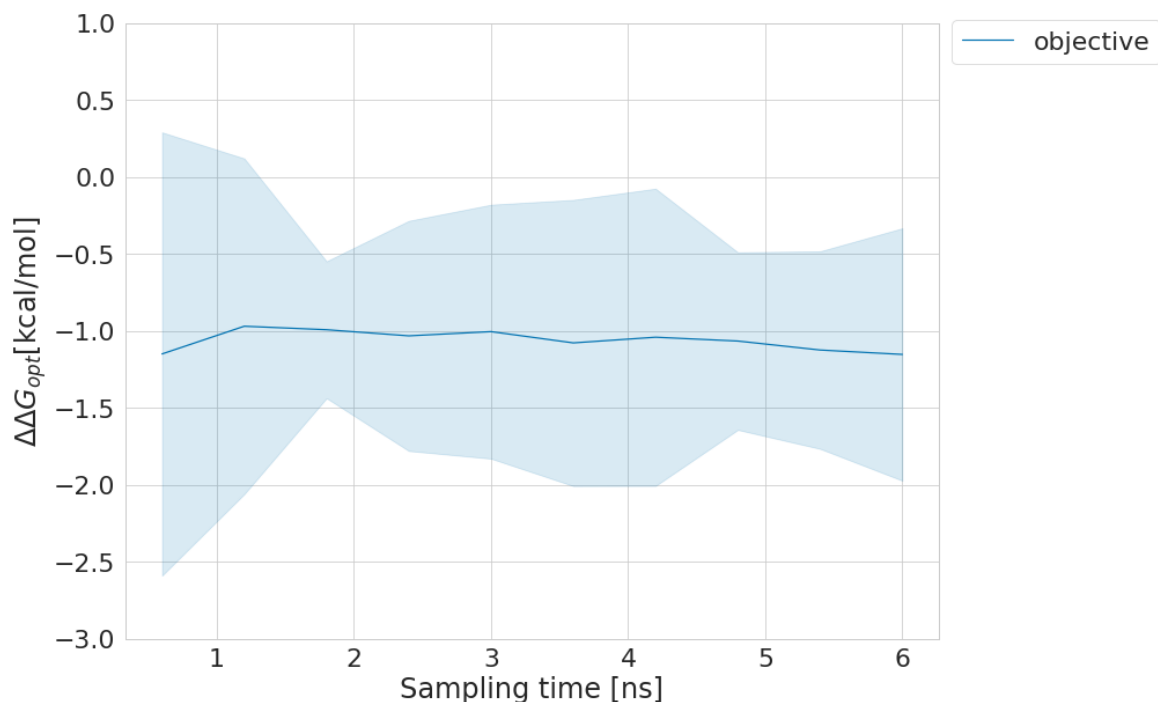


Fig. 4.3 Convergence for a calculation of the objective in the androgen receptor test case. $\Delta\Delta G_{opt}$ is reported as mean of three replicates with shaded area showing 95% confidence interval computed as $\text{mean} \pm t2 \cdot \text{SEM}$, where $t2$ is the t-distribution statistic with two degrees of freedom, and SEM is the standard error of the mean computed from the sample standard deviation of the three independent replicate predictions.

Table 4.3 Original and optimized sigmas for every atom in the androgen receptor ligand alongside the variance. Averages and variances are taken from three repeats. All hydrogen names are given in figure 3.3

Hydrogen	Original σ [nm]	Average optimized σ [nm]	Variance optimized σ [nm ²]
H5	0.263	0.154	0.006
H7	0.263	0.301	0.000
H15	0.242	0.032	0.001
H17	0.242	0.245	0.000
H191	0.260	0.285	0.001
H192	0.260	0.427	0.002
H221	0.242	0.233	0.000
H222	0.242	0.296	0.000
H3	0.263	0.381	0.002
H2	0.263	0.264	0.001

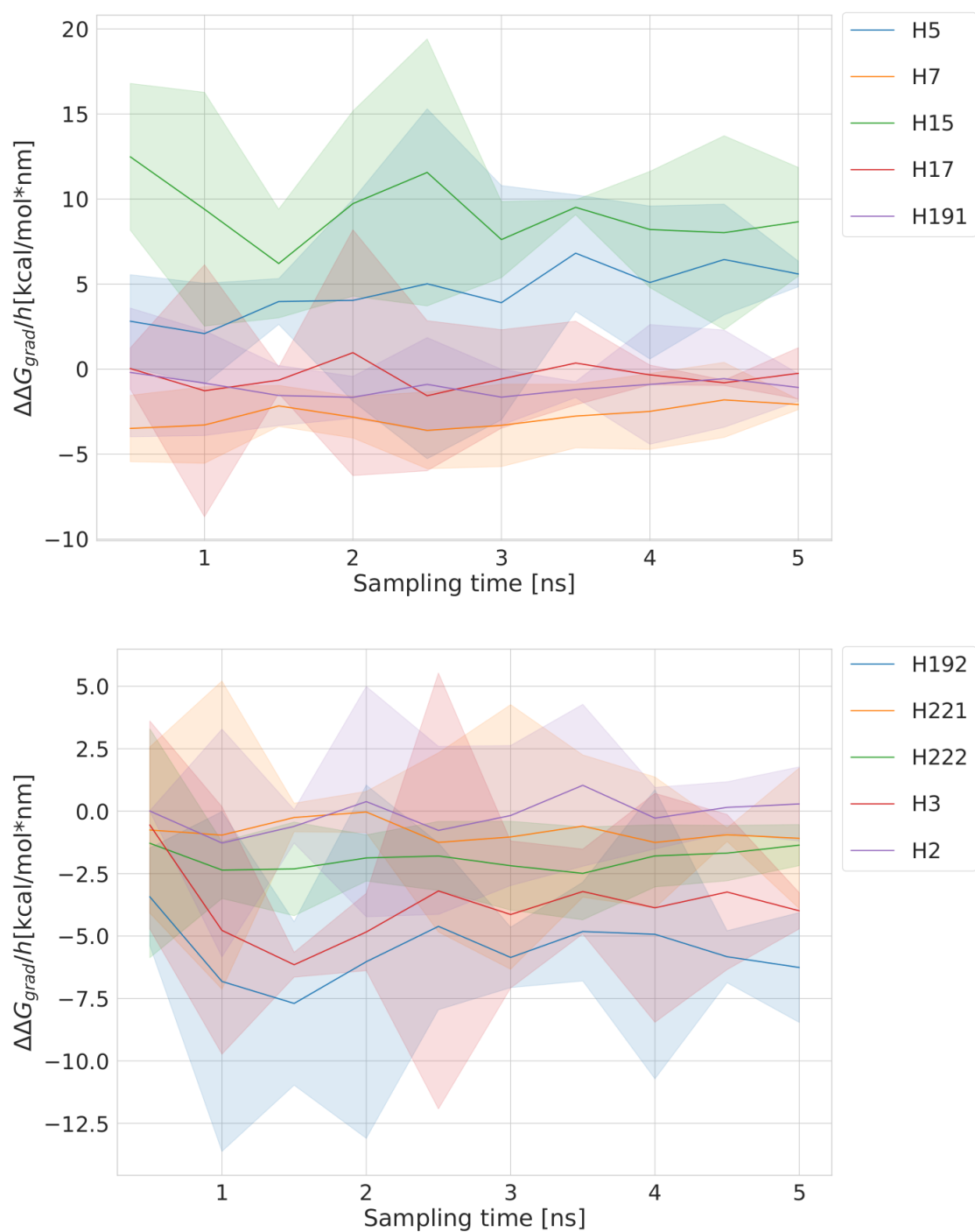


Fig. 4.4 Convergence for a calculation of the gradient in the androgen receptor test case. ΔG_{grad} is reported as the mean of three replicates with the shaded area showing 95% confidence interval computed as mean \pm t2*SEM, where t2 is the t-distribution statistic with two degrees of freedom. Atom names in the legend can be seen in figure 3.3

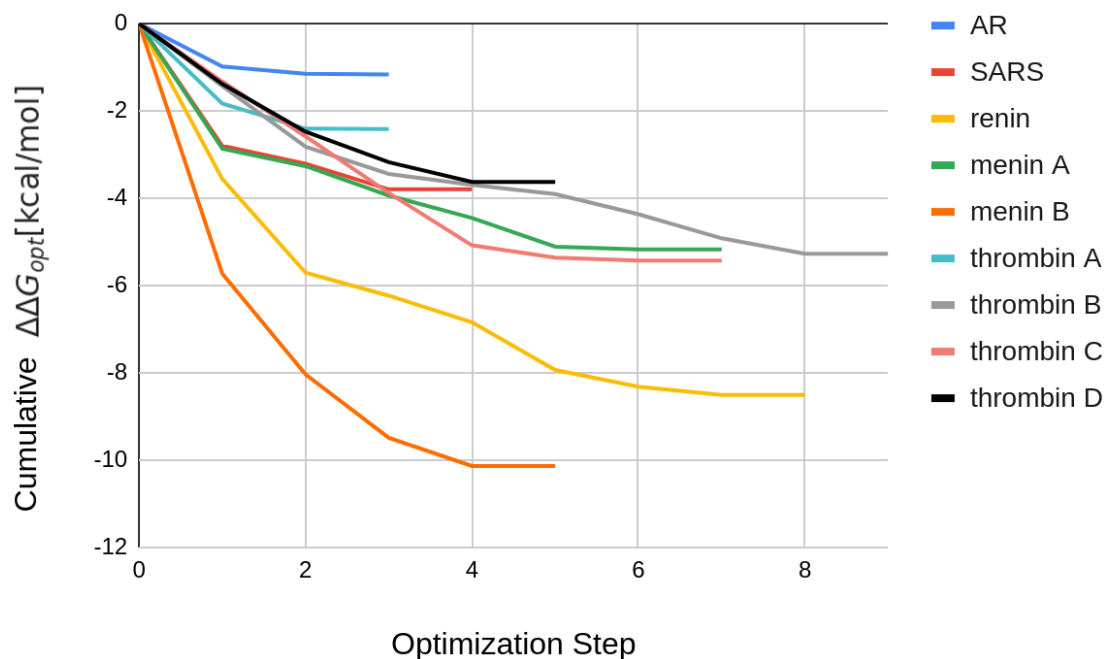


Fig. 4.5 Cumulative $\Delta\Delta G_{opt}$ calculated by summing the MBAR objective for each step of the optimization plotted against optimization step for every system.

Additional calculations were made to verify the value of the cumulative sum $\Delta\Delta G_{opt}$. This involved calculating the relative binding free energy between the original and optimized parameters of the inhibitor with MBAR. The mean absolute error between the values of $\Delta\Delta G_{opt}$ and the verification calculation was 1.57 kcal/mol. Details and results of this calculation can be seen in the appendix (figure C.1-C.3 and table C.2).

4.2.7 FEP Scans

The results of the optimization were analyzed and the best growth vectors on the ligand were determined. The best growth vectors were tested using the following protocol named FEP scans which used the MBAR method to calculate free energy changes. The binding free energy associated with these FEP scans will be termed $\Delta\Delta G_{scan}$. $\Delta\Delta G_{scan}$ values were calculated using an unpublished code written by David Huggins. This code uses OpenMMTools [83] to collect sampling using HREX. In previous chapters HREX was not used, for the full FEP calculations, we use HREX here to aid the convergence of the larger perturbations which will be calculated in this chapter. Each Hamiltonian in the replica exchange was an intermediate state in the alchemical transformation between two ligands using a hybrid topology. The end states of this transformation where a) the un-methylated

ligand and b) the methylated ligand. For these calculations the charge, van der Waals, bond, angle and torsion parameters are all interpolated between the end states. A softcore potential is used to interpolate the LJ potential for these FEP scan calculations. A softcore potential was omitted in previous chapters, but is included here to aid convergence of the calculations.

In this work we focus on transforming only hydrogens to methyls. This makes the exploration of the chemical space tractable. Transformations to methyl at all possible sites allows for a thorough validation of the optimizations because a more complete comparison can be made between the ranking of growth vectors by the optimization and the FEP scans. Exhaustive testing of all methyls is done here for the purpose of validation. If applying this method in a drug discovery effort, only the top growth vectors ranked highly by the optimizer would need to be tested using FEP scans.

To disappear a hydrogen and appear a methyl each FEP scan used 33 ns of sampling split across 22 alchemical windows, except renin which used 55 ns as this was observed to require more sampling to converge. Hamiltonian swapping was performed every 5 ps. All calculations were performed in triplicate. In figure 4.6 we show the convergence for making all possible hydrogen to methyl mutations on the androgen receptor ligand. All remaining convergence graphs for the FEP scan calculations can be seen in figures C.4-C.10.

When trying to calculate all hydrogen to methyl mutations we encountered some methyls which introduced numerical instability into the simulation. The causes of these instabilities in the SARS PLPro, menin and thrombin test cases were very close contact between an existing part of the ligand and the added methyl. In the androgen receptor test case the cause was close contact between the protein and the added methyl. We show the androgen receptor case in figure C.11. Any $\Delta\Delta G_{scan}$ which could not be calculated due to numerical instability is given an NA value in the results.

4.2.8 Summary of Methods

To summarize, in this work an optimization of the sigma parameters of several ligands was performed. The objective of this optimization was to find a set of sigmas to minimize the relative binding free energy of the ligand to a protein. This objective, $\Delta\Delta G_{opt}$, was evaluated using FEP calculations and the MBAR estimator. To calculate the gradient, $\Delta\Delta G_{grad}$, SSP calculations were used. The results of the optimizer are used to predict beneficial growth vectors on the ligand. To validate these predicted growths, hydrogen to methyl mutations were computed, $\Delta\Delta G_{scan}$, using FEP and the MBAR estimator with HREX. For the reader's reference table 4.4 contains all the abbreviated terms used and a brief description of the calculation.

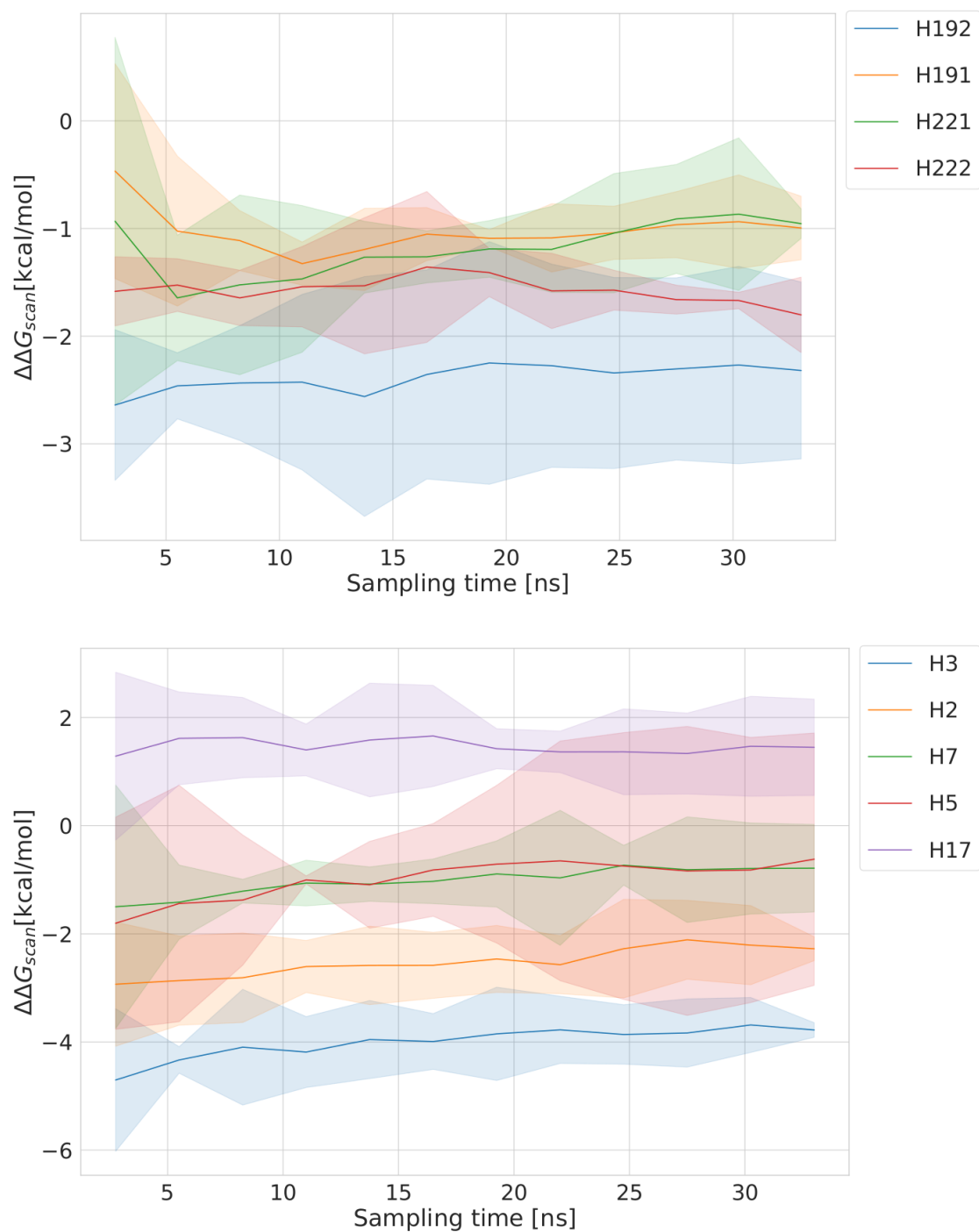


Fig. 4.6 $\Delta\Delta G_{scan}$ for each methylation in the androgen receptor system as the amount of sampling is increased. $\Delta\Delta G_{scan}$ are reported as mean of three replicates with shaded area showing 95% confidence interval computed as $\text{mean} \pm t_2 \cdot \text{SEM}$, where t_2 is the t-distribution statistic with two degrees of freedom, and SEM is the standard error of the mean computed from the sample standard deviation of the three independent replicate predictions. All hydrogen names are taken from figure 3.3.

Table 4.4 Abbreviated terms used to reference various relative binding free energies calculated in this work with a brief description of the calculation for the reader's reference

Free energy value	Description
$\Delta\Delta G_{opt}$	The value for the objective. Calculated between the original and optimized sigmas with MBAR.
$\Delta\Delta G_{grad}$	The value for the gradient. Calculated between many highly related ligands using SSP.
$\Delta\Delta G_{scan}$	The value for one or many hydrogen to methyl mutations calculated with MBAR and HREX.
$\Delta\Delta G_{exp}$	The value for mutations determined experimentally.

4.3 Results

The results from the optimizations performed in this work will be considered one system at a time. First looking at the androgen receptor. Each optimization produces a set of optimized σ 's which minimizes the binding free energy. We show this result with figures, made such that any optimized hydrogens are sized in proportion to their calculated $\Delta\sigma$, where $\Delta\sigma$ is the difference between the atoms optimized and original σ . One reason we create these figures is to allow the continuous values of σ in the optimization to be converted by eye into the discrete values of σ associated with adding any atom(s). In the following calculations we denote symmetry related positions with an obelisk symbol (\dagger). For optimization calculations no atoms are considered symmetric. This is because the optimizer assigns different sigmas to symmetric atoms, and this breaks any symmetry. In the context of these simulations, symmetric methyl hydrogens rapidly interconvert due to rotation, but symmetric hydrogens on aromatic rings do not. Thus, hydrogens on methyls are considered symmetric for the $\Delta\Delta G_{scan}$ calculations, but hydrogens on aromatic rings are not.

In this results section all $\Delta\Delta G_{scan}$ values are reported as the mean of three replicates with bracketed values showing the 95% confidence intervals computed as $\text{mean} \pm t_2 \cdot \text{SEM}$, where t_2 is the t-distribution statistic with two degrees of freedom, and SEM is the standard error of the mean computed from the sample standard deviation of the three independent replicate predictions.

Figure 4.7 shows the relative size of the optimized hydrogens in the androgen receptor test case. The largest hydrogens are H192 and H3 with H3 as the position chosen in previous experimental work to be methylated. The values of $\Delta\sigma$ from the optimization are tabulated along side calculated values for $\Delta\Delta G_{scan}$ and available experimental free energies are reported as $\Delta\Delta G_{exp}$.

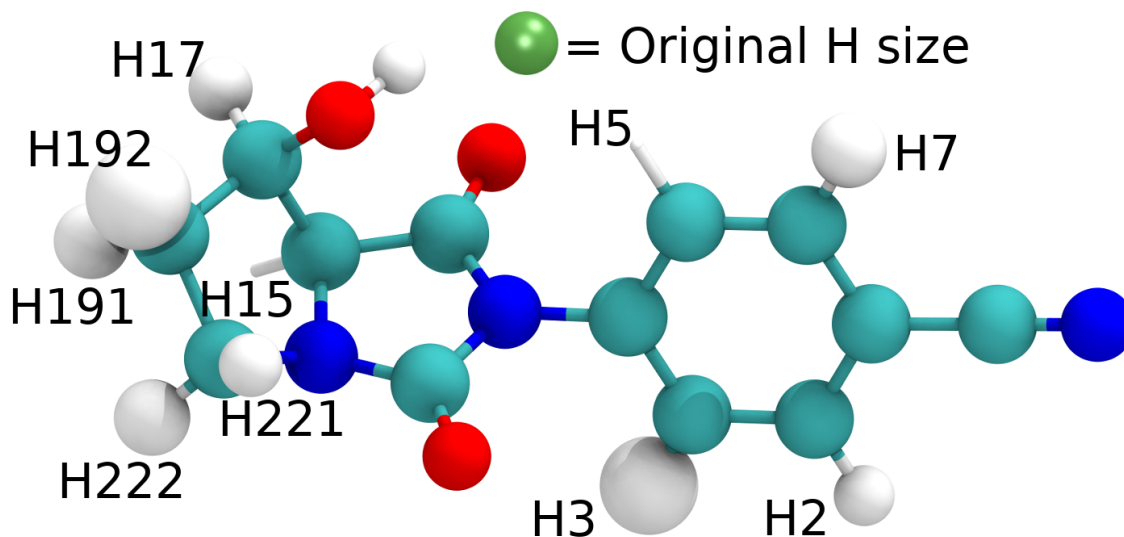


Fig. 4.7 Androgen receptor ligand with all optimized hydrogens sized in proportion to their calculated $\Delta\sigma$

Table 4.5 Comparison of $\Delta\sigma$, $\Delta\Delta G_{scan}$ for mutating a hydrogen to methyl for the androgen receptor test case alongside experimental values $\Delta\Delta G_{exp}$ [173].

Hydrogen	$\Delta\sigma$ [nm]	$\Delta\Delta G_{scan}$ [kcal/mol]	$\Delta\Delta G_{exp}$ [kcal/mol]
H192	0.148	-2.32 [-3.14, -1.50]	-0.83
H3	0.122	-3.78 [-3.91, -3.64]	
H191	0.041	-1.00 [-1.29, -0.70]	
H222	0.041	-1.80 [-2.15, -1.45]	
H7	0.038	-0.79 [-1.60, 0.02]	-0.83
H17	-0.004	1.45 [0.56, 2.34]	
H221	-0.004	-0.95 [-1.09, -0.82]	
H2	-0.015	-2.28 [-2.50, -2.05]	
H5	-0.187	-0.62 [-2.95, 1.71]	-0.83
H15	-0.237	NA	

The results in table 4.5 show that for the AR test case the optimization and FEP scan rank the experimentally verified methylation, H3, second and first respectively. The overall agreement in the ranking between computational methods can be calculated with the Spearman's Rank-Order Correlation, ρ . Here we calculated ρ between the ranking from the optimization and FEP scan (any hydrogens without calculated value for $\Delta\Delta G_{scan}$ are ranked last) and for the AR test case ρ was calculated as 0.7. The agreement between the experimental $\Delta\Delta G_{exp}$ and computational $\Delta\Delta G_{scan}$ is not good in this case, differing by more than 1 kcal/mol. Looking at the outlying data in table 4.5 it can be seen that whilst the optimization ranks H2 as not beneficial the FEP scan calculated it to be a beneficial position for a methyl. We speculate that this is because during the optimization all σ values are changed simultaneously in the same system. This means that if both H2 and H3 grow simultaneously they will see each other with increased radius in the simulation. It may be possible that some growths which are close in proximity can interfere with each other such that only one position grows to maximize the binding affinity. This effect would not be seen for the FEP scan calculation as each methylation is a separate calculation and as such the methylation at H2 and H3 do not need to be accommodated simultaneously.

Figure 4.8 provides the result from the optimizer for the SARS PLPro system and shows that three growth vectors have been highlighted. Two of these growths, H4 and H9, are adjacent to each other pointing in approximately the same direction, the other, H10, is separately located on the ortho position of a phenyl; it is this ortho position that was the experimentally chosen position in previous work.

Looking more closely at the SARS PLPro result in table 4.6 it can be seen that the experimentally chosen hydrogen, H10, is ranked second by the optimizer and first by the FEP scan. H4, the position most favoured by the optimizer, is confirmed to be a beneficial position for methylation by the FEP scan with a $\Delta\Delta G_{scan}$ of -0.45 [-1.09, 0.19]. For the SARS PLPro test case the correlation in the ranking of growth vectors by the optimizer and FEP scan is good with ρ calculated to be 0.8. The experimental data which exists for the SARS PLPro systems is for ligands with methyls at positions H10/H13, H17/H19 and H21 but no reference data exists for the ligand with a hydrogen at all of these sites. To compare to experiment we therefore use the $\Delta\Delta G_{scan}$ values in table 4.6 to calculate the relative binding free energy change for permuting the methyl on sites H10/H13, H17/H19 and H21 and present this result in table 4.7.

Table 4.7 shows that the $\Delta\Delta G_{scan}$ for moving the methyl from H19 to H21 is well agreed with experiment, within 1 kcal/mol. For the H10 to H19 and H10 to H21 transformations disagreement with experiment is greater than 1 kcal/mol. However, the ranking in table 4.6

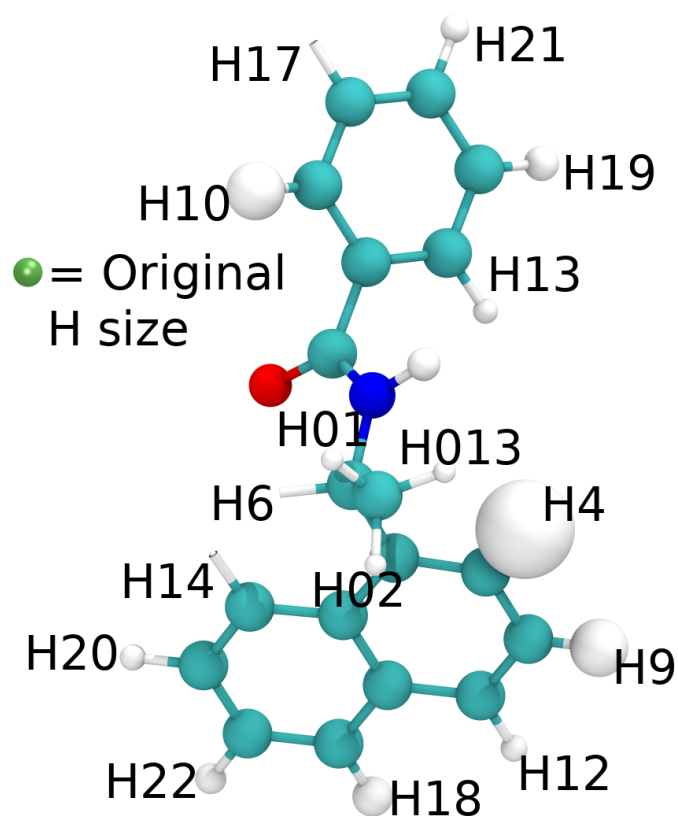


Fig. 4.8 SARS PLPro ligand with all optimized hydrogens sized in proportion to their calculated $\Delta\sigma$

Table 4.6 Comparison of $\Delta\sigma$, $\Delta\Delta G_{scan}$ for mutating a hydrogen to methyl for the SARS PLPro test case alongside the available data for an experimental ranking for the benefit of a methyl mutation [201].

Hydrogen	$\Delta\sigma$ [nm]	$\Delta\Delta G_{scan}$ [kcal/mol]	Experimental Rank
H4	0.270	-0.45 [-1.09, 0.19]	1st
H10	0.115	-1.70 [-2.35, -1.05]	
H9	0.115	-0.73 [-0.94, -0.53]	
H18	0.047	-0.16 [-0.46, 0.13]	2nd
H19	0.028	-0.26 [-0.65, 0.14]	
H22	0.013	-0.34 [-0.75, 0.07]	
H21	0.002	-0.10 [-0.27, 0.08]	3rd
H12	-0.004	-0.18 [-0.67, 0.31]	
H20	-0.007	-0.58 [-0.79, -0.36]	
H13	-0.011	-1.14 [-1.17, -1.11]	
H01	-0.02	0.34 [-0.02, 0.70]†	
H013	-0.021	0.34 [-0.02, 0.70]†	
H012	-0.023	0.34 [-0.02, 0.70]†	
H14	-0.080	NA	
H6	-0.103	NA	
H17	-0.106	1.05 [0.89, 1.21]	

Table 4.7 $\Delta\Delta G_{scan}$ for mutating a methyl from one position to another on the ligand compared to experimental values, $\Delta\Delta G_{exp}$, in the SARS PLPro test case. Values in parentheses are experimental free energy \pm reported experimental uncertainty [201].

Methyl Mutation	$\Delta\Delta G_{scan}$ [kcal/mol]	$\Delta\Delta G_{exp}$ [kcal/mol]
H10 to H19	1.44 [0.68, 2.20]	0.32(0.12, 0.52)
H10 to H21	1.60 [0.93, 2.27]	0.72(0.63, 0.81)
H19 to H21	0.16 [-0.27, 0.59]	0.40(0.19, 0.62)

is much more accurate with the experimentally preferred methylation ranked first and second by the FEP scan and optimization respectively.

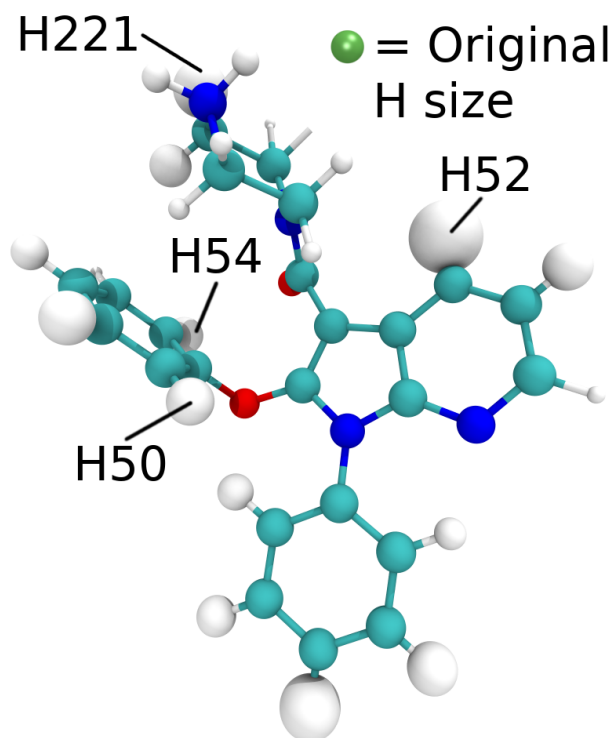


Fig. 4.9 Renin ligand with all optimized hydrogens sized in proportion to their calculated $\Delta\sigma$. See all named hydrogens in table C.1.

The data for the renin system are detailed in figure 4.9 and table 4.8, from these data it can be seen that there is one clear growth vector highlighted by the optimizer H52, this is not in the direction chosen experimentally (H50/H54). There is, however, agreement between the optimizer and the FEP scan which also ranked H52 first with a $\Delta\Delta G_{scan}$ of -1.56 [-1.85, -1.27]. This is an example where the ranking between the computational and experimental methods are not agreed but the rankings between computational methods are agreed and this may result from inaccuracy in the simulation methodology or the force field.

In this renin test case the result for H221 in table 4.8 is outlying and is not well agreed between the optimizer and the FEP scan, we speculate that this disagreement is caused by the disruption of favourable charged interaction between the protein and the amine group adjacent to H221. These interactions may be disrupted by the methyl group added in the FEP scan but not by the comparatively smaller change in sigma seen during the optimization. Overall the correlation in the ranking between the computational methods in the renin system is high with a calculated ρ of 0.7. The difference between the calculated and experimental binding affinity is between 1 and 2 kcal/mol.

Table 4.8 Comparison of $\Delta\sigma$, $\Delta\Delta G_{scan}$ for mutating a hydrogen to methyl for the renin test case alongside experimental data, $\Delta\Delta G_{exp}$ [166].

Hydrogen	$\Delta\sigma$ [nm]	$\Delta\Delta G_{scan}$ [kcal/mol]	$\Delta\Delta G_{exp}$
H52	0.256	-1.56 [-1.85, -1.27]	-1.75
H38	0.158	-1.05 [-1.28, -0.82]	
H221	0.158	1.69 [1.01, 2.36]	
H39	0.088	-0.79 [-1.46, -0.13]	
H31	0.084	-0.36 [-0.76, 0.04]	
H6	0.080	-0.31 [-0.54, -0.08]	
H54	0.080	-0.28 [-0.49, -0.07]	
H222	0.077	0.67 [-0.56, 1.91]	
H50	0.069	0.68 [0.26, 1.11]	
H37	0.052	-0.94 [-1.14, -0.74]	
H36	0.037	0.46 [0.07, 0.86]	
H99	0.017	0.17 [-0.54, 0.87]	
H40	-0.006	0.34 [-0.92, 1.59]	
H191	-0.060	6.38 [3.89, 8.88]	
H202	-0.065	1.65 [-0.33, 3.63]	
H192	-0.067	5.80 [5.10, 6.51]	
H201	-0.070	9.14 [8.35, 9.93]	
H1	-0.089	1.42 [1.15, 1.70]	
H53	-0.099	0.46 [0.08, 0.85]	
H231	-0.102	1.78 [-3.86, 7.42]	
H232	-0.176	0.70 [-2.30, 3.70]	

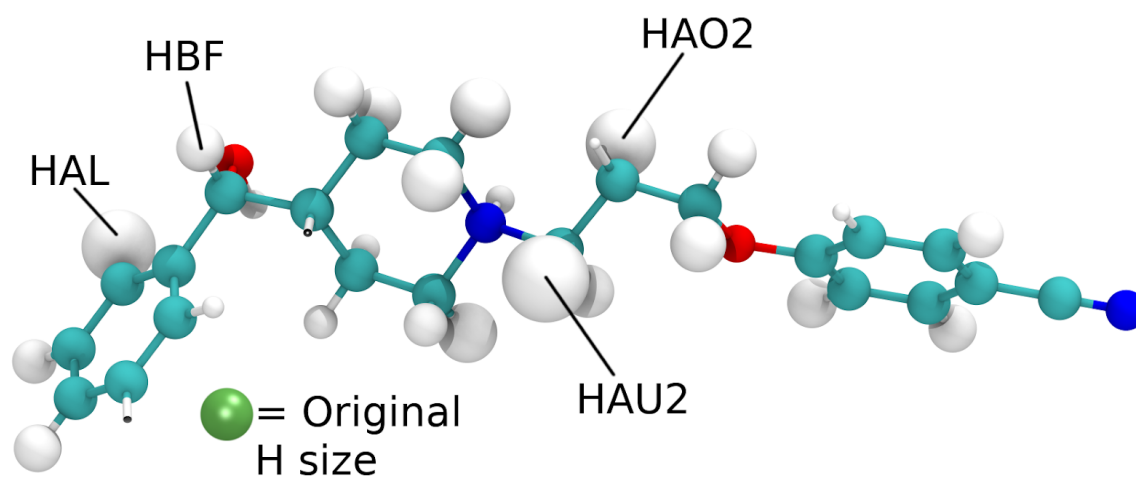


Fig. 4.10 Menin A ligand with all optimized hydrogens sized in proportion to their calculated $\Delta\sigma$. See all named hydrogens in table C.1.

Table 4.9 Comparison of $\Delta\sigma$, $\Delta\Delta G_{scan}$ for mutating a hydrogen to methyl for the menin A test case alongside experimental data, $\Delta\Delta G_{exp}$ [168]

Hydrogen	$\Delta\sigma$ [nm]	$\Delta\Delta G_{scan}$ [kcal/mol]	$\Delta\Delta G_{exp}$
HAU2	0.167	0.46 [-0.27, 1.20]	0.14
HAL	0.111	-1.57 [-2.24, -0.90]	
HAO2	0.094	-0.49 [-1.19, 0.20]	
HAV1	0.059	0.76 [-0.40, 1.92]	
HAV2	0.051	0.78 [-0.66, 2.23]	
HAW1	0.051	0.35 [-0.96, 1.65]	
HAP1	0.033	1.62 [0.69, 2.54]	
HAU1	0.021	1.95 [1.07, 2.84]	
HAS1	0.009	-0.05 [-0.61, 0.52]	
HAI	0.006	0.72 [-0.58, 2.01]	
HAP2	0.003	0.75 [0.30, 1.20]	
HAS2	0.000	0.28 [-0.58, 1.14]	
HBF	-0.001	-0.22 [-0.78, 0.34]	
HAG	-0.001	-0.09 [-0.22, 0.05]	
HAF	-0.003	0.16 [-0.55, 0.86]	
HAH	-0.010	1.42 [1.00, 1.84]	
H3	-0.017	0.83 [0.28, 1.38]	
HAW2	-0.02	0.32 [-0.22, 0.85]	
HAT2	-0.062	0.03 [-0.35, 0.41]	
HAT1	-0.083	0.29 [0.23, 0.35]	
HAI	-0.112	NA	
HAK	-0.1300	0.99 [-0.19, 2.17]	
HAO1	-0.158	1.92 [0.72, 3.12]	
HBD	-0.196	NA	
HAE	-0.207	0.29 [-0.78, 1.36]	

The result for the menin A system can be seen in figure 4.10 and table 4.9. These results show that three growth vectors are highlighted: HAL, HAU2 and HAO2. Our method agrees with the experimental result that HBF is not a position which can beneficially accommodate a methyl group this can be seen as the $\Delta\sigma$ for HBF is not a large positive.

It is worth noting that the HBF site is beneficial for groups larger than methyl such as isopropyl or cyclohexyl with a $\Delta\Delta G_{exp}$ of -2.41 or -2.94 kcal/mol respectively. From this test case we would conclude that in its current form the optimization cannot provide information that this is a beneficial spot for larger mutations. The highlighted position, HAL and HAO2, are well agreed between the optimization and FEP scan, ranked second and third by the optimizer and first and second by the FEP scan. HAU2 is not agreed between the two computational methods. The overall rank correlation, ρ , for the menin A system was lower than seen in the other test cases with a value of 0.3. The experimental value for adding a methyl at HBF was well agreed with experiment, 0.14 kcal/mol compared to the $\Delta\Delta G_{scan}$ value of -0.22 [-0.78, 0.34].

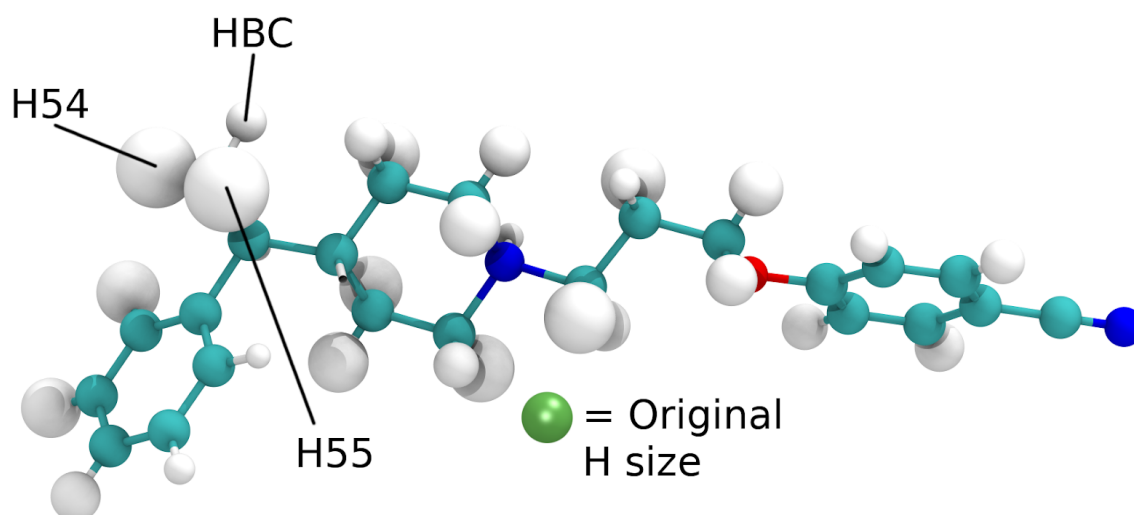


Fig. 4.11 Menin B ligand with all optimized hydrogens sized in proportion to their calculated $\Delta\sigma$. See all named hydrogens in table C.1.

As mentioned, the site at HBF in the test case menin A will beneficially accommodate mutations larger than a methyl and this will now be investigated with system menin B. The menin B ligand is the same as a menin A with a methyl added at HBF. When ligand menin B is optimized the newly added hydrogens H55 and H54 are now found to be the most beneficial by the optimizer. This can be seen in figure 4.11 where H55 and H44 are ranked first and second, see table A5 for all values for optimized σ . The experimental data for growths of ligand menin B correspond to adding methyls to H54 and H55 simultaneously.

Therefore the $\Delta\Delta G_{scan}$ values calculated are now also for adding two methyls, one at H54 and one at H55. These $\Delta\Delta G_{scan}$ calculations for two methyls give a value of -3.11 [-4.02, -2.2] kcal/mol which agrees well with the experimental value -2.55 kcal/mol.

The final test case to consider is thrombin. In this test case several mutations are strung together and the optimization method applied iteratively to build a more tightly binding inhibitor, we also include a round of charge optimization as discussed in chapter 2. We perform this test case as we imagine this method might be applied in a drug discovery setting. This means that not all methylations will be tested by computing $\Delta\Delta G_{scan}$, in every round of optimization, only those which would contribute new information to the study. The goal here was to use information from the optimizer to move from thrombin A shown in table 4.1 which has an experimental binding free energy of -7.1 kcal/mol to a ligand with improved binding free energy.

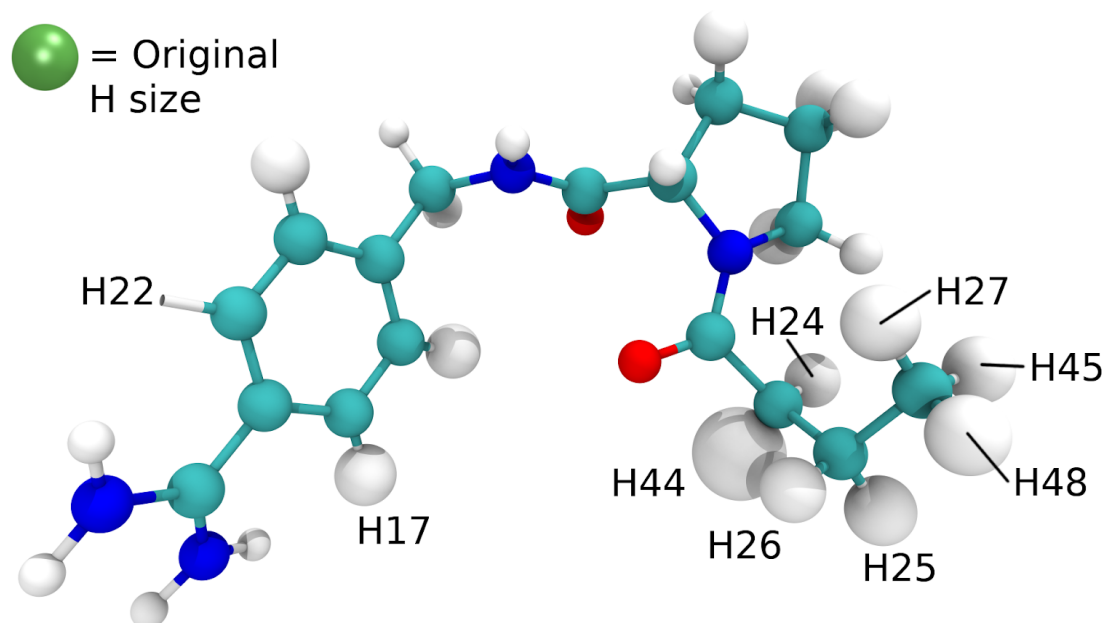


Fig. 4.12 Thrombin A ligand with all optimized hydrogens sized in proportion to their calculated $\Delta\sigma$. See all named hydrogens in table C.1.

Considering the result of the thrombin A optimization in figure 4.12 and table 4.10 it can be seen that H44 is ranked first by the optimization and H27/H45/H48 are ranked first by the FEP scan. The ranking of H17 is not well agreed between the optimizer and the FEP scan and this may be a similar effect to that seen for the outlier in the renin test case. Where perhaps in the FEP scan the added methyl is disrupting charged interaction between the protein and the amine groups on the ligand nearby to H17. Another possible cause for the disagreement might be explained by looking at H22. In the original thrombin A ligand

Table 4.10 Comparison of $\Delta\sigma$, $\Delta\Delta G_{scan}$ for mutating a hydrogen to methyl for the Thrombin A test case alongside experimental data, $\Delta\Delta G_{exp}$. Values in parentheses are experimental free energy \pm reported experimental uncertainty [203].

Hydrogen	$\Delta\sigma$ [nm]	$\Delta\Delta G_{scan}$ [kcal/mol]	$\Delta\Delta G_{exp}$
H44	0.232	-0.57 [-0.89, -0.24]	
H32	0.056	-0.07 [-0.36, 0.21]	
H27	0.046	-1.75 [-2.66, -0.83] [†]	
H25	0.045	-0.24 [-0.49, 0.02]	-0.24(-0.39, -0.09)
H48	0.040	-1.75 [-2.66, -0.83] [†]	
H33	0.033	0.34 [-0.65, 1.33]	
H17	0.028	5.11 [3.78, 6.44]	
H45	0.018	-1.75 [-2.66, -0.83] [†]	
H16	0.012	-0.80 [-1.05, -0.55]	
H24	0.003	NA	
H26	-0.001	-0.76 [-1.21, -0.31]	-0.24(-0.39, -0.09)
H31	-0.010	1.08 [0.79, 1.37]	
H29	-0.032	2.70 [1.79, 3.60]	
H23	-0.035	1.10 [0.86, 1.34]	
H36	-0.075	5.45 [5.03, 5.87]	
H34	-0.086	1.56 [1.47, 1.66]	
H28	-0.136	2.94 [1.82, 4.06]	
H30	-0.144	2.46 [1.49, 3.44]	
H37	-0.153	3.56 [-0.66, 7.78]	
H22	-0.333	6.09 [3.53, 8.65]	

H17 is symmetric to H22 and the optimizer and FEP scan are in agreement that H22 is very unfavourable. Therefore it may also be possible that in the FEP scan simulation H17 and H22 are adequately interconverting to converge to the solution that H17/H22 is an unfavourable position but this interconversion is not occurring sufficiently for the optimization simulations, which are shorter. Overall the ρ calculated between the ranking from the optimization and FEP scan for the thrombin A system was 0.7.

Based on the results of figure 4.12 and table 4.10 either H44 or H27/H45/H48 could be chosen for methylation and here we show the result for methylating H44. The thrombin A ligand (table 4.1) with a methyl added at H44 is the ligand thrombin B in (table 4.1) and so ligand thrombin B was then considered for an additional round of optimization. With the retrospective knowledge that the H44 is a beneficial location for an amine group the next round of optimization is performed for the partial charges of ligand thrombin B. The charge optimization used was identical to the steric optimization with the exception that the parameters of the optimization were the partial charges of the atoms instead of the atoms σ . The methodology to perform this charge optimization has been explored in chapter 2.

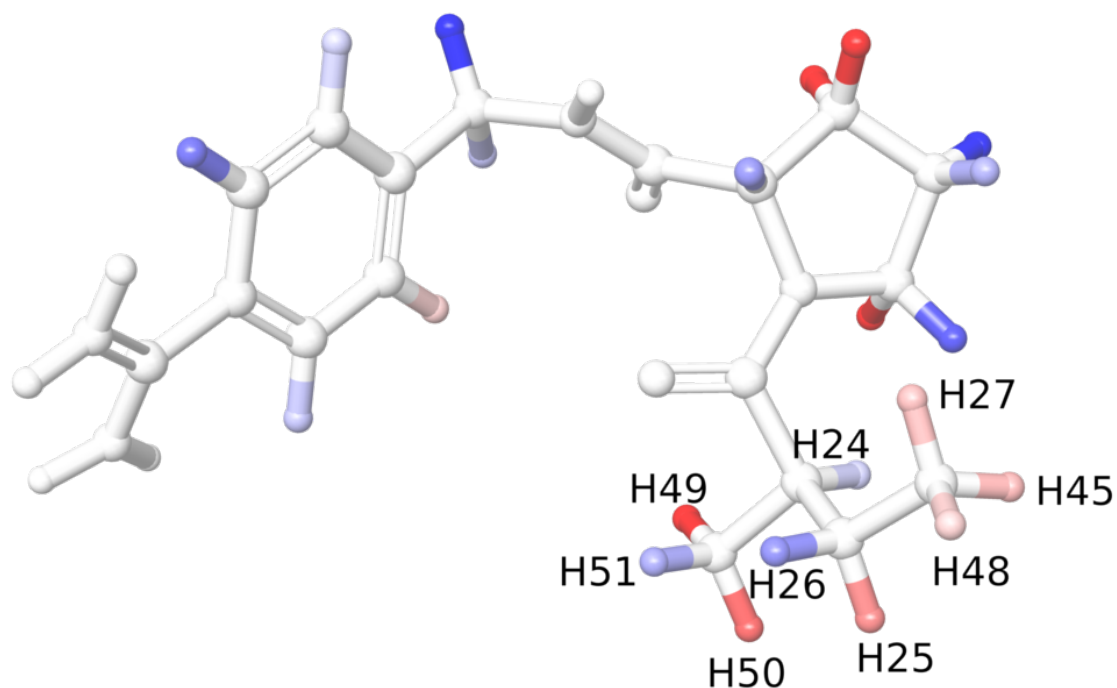


Fig. 4.13 Thrombin B ligand with all optimized hydrogens colored in proportion to their calculated Δq . Red is more positive and blue more negative. See all named hydrogens in table C.1.

Figure 4.13 presents the result of the charge optimization and shows that hydrogens H49 and H50 can be made more positive to improve the binding free energy. The total change in charge for H49, H50 and H51 made by the optimizer was $+0.3 e$, with all changes in partial charge shown in table C.4. Adding a charged amine group at this position is known experimentally to be beneficial, giving a change in binding affinity of -2.06 kcal/mol. We therefore chose to add an amide group to ligand thrombin B (table 4.1) which gave ligand thrombin C (table 4.1). We then continued to optimize with another round for the thrombin C ligand. We are now once again optimizing the sigma parameters of ligand thrombin C and the result of this optimization are presented in figure 4.14 and table 4.11.

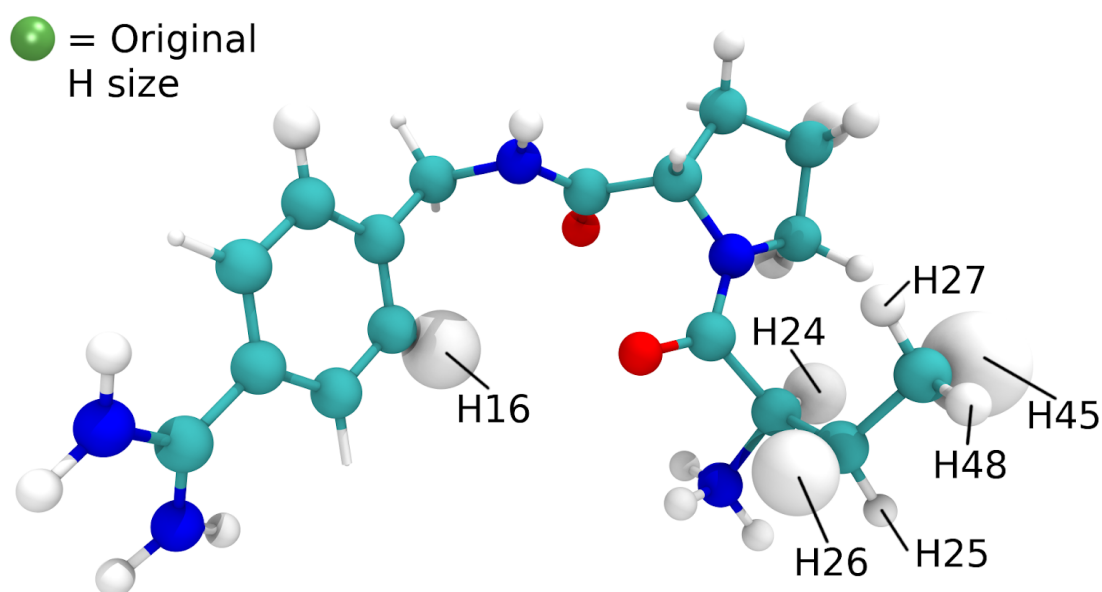


Fig. 4.14 Thrombin C ligand with all optimized hydrogens sized in proportion to their calculated $\Delta\sigma$. See all named hydrogens in table C.1.

In table 4.11 only a subset of all possible $\Delta\Delta G_{scan}$ values are calculated, the calculated values of $\Delta\Delta G_{scan}$ are chosen in the areas where growth were made to the ligand experimentally. Looking at figure 4.14 and table 4.11 three growth vectors are highlighted: H45, H16 and H26 which are all confirmed to be beneficial by the FEP scan. The ρ calculated between the ranking from the optimization and FEP scan for the thrombin C system was 0.6. The growth vector ranked first by the optimizer and the methylation scan was H45 with a $\Delta\Delta G_{scan}$ of -1.68 $[-2.27, -1.09]$ kcal/mol. Since the optimizer and FEP scan both rank H45 first it was selected to be mutated to a methyl. Adding a methyl at H45 of ligand thrombin C (table 4.1) gives ligand thrombin D (table 4.1), we therefore performed a final round of optimization on ligand thrombin D.

Table 4.11 Comparison of $\Delta\sigma$, $\Delta\Delta G_{scan}$ for mutating a hydrogen to methyl for the Thrombin C test case alongside experimental data, $\Delta\Delta G_{exp}$. Values in parentheses are experimental free energy \pm reported experimental uncertainty [203].

Hydrogen	$\Delta\sigma$ [nm]	$\Delta\Delta G_{scan}$ [kcal/mol]	$\Delta\Delta G_{exp}$
H45	0.305	-1.68 [-2.27, -1.09]†	-0.60(-0.83, -0.29)
H16	0.228	-0.54 [-0.87, -0.22]	
H26	0.196	-0.52 [-1.11, 0.07]	
H24	0.138		
H32	0.079		
H33	0.038		
H23	0.028		
H27	0.007	-1.68 [-2.27, -1.09]†	
H48	-0.001	-1.68 [-2.27, -1.09]†	
H31	-0.003		
H29	-0.027		
H25	-0.029	-0.29 [-0.93, 0.34]	-0.60(-0.83, -0.29)
H34	-0.040		
H37	-0.098		
H28	-0.099		
H22	-0.109		
H17	-0.153		
H30	-0.159		
H36	-0.180		

Table 4.12 Comparison of $\Delta\sigma$, $\Delta\Delta G_{scan}$ for mutating a hydrogen to methyl for the Thrombin D test case.

Hydrogen	$\Delta\sigma$ [nm]	$\Delta\Delta G_{scan}$ [kcal/mol]
H16	0.177	
H26	0.145	-1.35 [-1.76, -0.95]
H24	0.098	1.57 [0.26, 2.87]
H53	0.089	-1.02 [-1.32, -0.72]†
H52	0.075	-1.02 [-1.32, -0.72]†
H23	0.019	
H48	0.015	-0.64 [-1.13, -0.14]
H31	0.012	
H54	0.008	-1.02 [-1.32, -0.72]†
H27	0.008	-0.67 [-0.74, -0.60]
H33	0.005	
H25	0.000	-0.27 [-1.13, 0.60]
H32	-0.004	
H29	-0.038	
H34	-0.065	
H22	-0.079	
H28	-0.126	
H17	-0.133	
H37	-0.156	
H30	-0.156	
H36	-0.164	

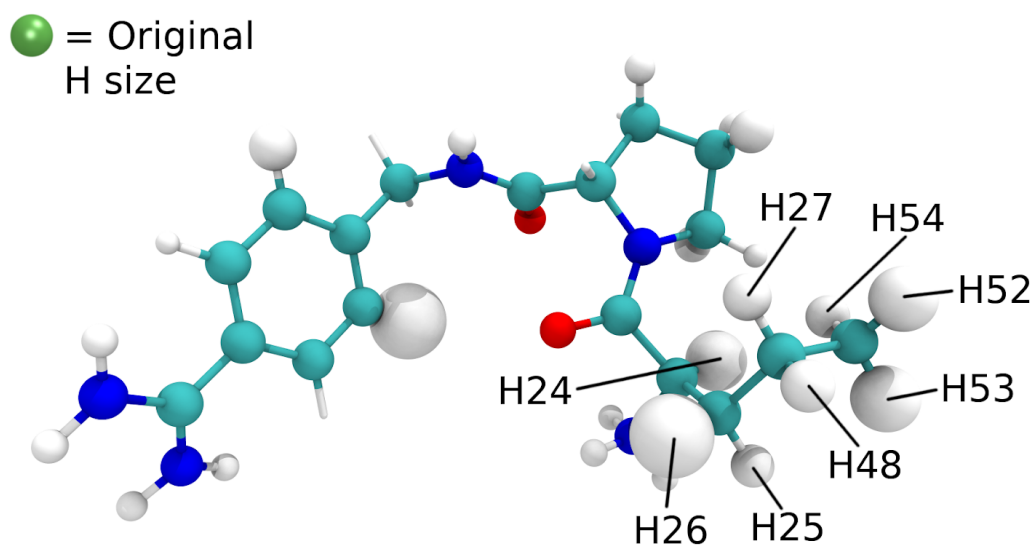


Fig. 4.15 Thrombin D ligand with all optimized hydrogens sized in proportion to their calculated $\Delta\sigma$. See all named hydrogens in table C.1.

Again only a subset of all possible methylations are calculated in table 4.12. With the calculated values of $\Delta\Delta G_{scan}$ chosen in the area growths were made to the ligand experimentally. For the thrombin D test case figure 4.15 and table 4.12 show that the optimizer ranks H16 and H26 as the first and second best growth vectors. The FEP scan ranks H26 as the best growth vector. The ρ calculated between the ranking from the optimization and FEP scan for the thrombin C system was 0.4. Based on the information from the optimizer and FEP scan H26 was selected to be methylated. To compare to experiment the $\Delta\Delta G_{scan}$ from thrombin C and thrombin D were combined. The $\Delta\Delta G_{scan}$ for H45 from thrombin C was combined with the $\Delta\Delta G_{scan}$ for a methylation in thrombin D, which gave the free energy change for adding a methyl at both sites simultaneously, these calculations are presented in table 4.13.

H27 or H48 are the positions methylated in the experimental work and it can be seen in table 4.13 that computationally these positions are also beneficial to add a methyl. However H27 and H48 are not ranked highly, instead H26 was ranked first. Comparing the $\Delta\Delta G_{scan}$ for H27 or H48 to the experimental values gives reasonable agreement with experiment, within 1 kcal/mol for both H27 and H48.

Finally combining all the mutations selected in the thrombin optimization iterations the final computationally ligand is presented alongside an experimentally optimized ligand [200] in figure 4.16. Both our computationally optimized and the experimentally optimized ligand originate from the ligand thrombin A (table 4.1). It can be seen in figure 4.16 that our

Table 4.13 Comparison for $\Delta\sigma$, $\Delta\Delta G_{scan}$ for mutating thrombin C H45 to a methyl and one other named hydrogen in thrombin D to a methyl and any experimental values $\Delta\Delta G_{exp}$ for the thrombin D test case. Values in parentheses are experimental free energy \pm reported experimental uncertainty [200].

Hydrogen	$\Delta\Delta G_{scan}$ [kcal/mol]	$\Delta\Delta G_{exp}$ [kcal/mol]
H26	-3.03 [-3.74, -2.32]	
H52	-2.70 [-3.36, -2.04]†	
H53	-2.70 [-3.36, -2.04]†	
H54	-2.70 [-3.36, -2.04]†	
H27	-2.35 [-2.94, -1.76]	-1.20(-1.62, -1.00)
H48	-2.32 [-3.09, -1.55]	-1.20(-1.62, -1.00)
H25	-1.95 [-3.00, -0.90]	
H24	-0.11 [-1.54, 1.32]	

optimized ligand is very similar to a ligand determined experimentally, differing only in the placement of one methyl group. The binding free energy of the experimental ligand is -10.4 kcal/mol [200], we can use this value in combination with our calculation for removing a methyl at H27 and adding a methyl at H26 (table 4.12) to estimate the binding free energy of our computationally optimized ligand. The binding free energy of our optimized ligand is calculated to be -11.1 [-10.7, -11.5] significantly better than the starting ligand (thrombin A) which had an experimental binding free energy of -7.1 kcal/mol.

4.4 Conclusion

In this work a novel method to optimize the van der Waals interactions of small molecule inhibitors to maximize the binding affinity to a receptor has been developed. This method combines rapid single step perturbation calculation with MBAR calculations to calculate the gradient and objective respectively, allowing optimized inhibitors to be found quickly. This new method was applied to nine inhibitors across five diverse test systems: androgen receptor, SARS PL protease, renin, menin and thrombin. Good agreement was found between the beneficial growth vectors identified by the optimization and the results of full FEP calculations to exhaustively calculate methylations, with a Spearman's rank order correlation of 0.59. One advantage of this method is that it allows for all growth vectors on a ligand to be systematically assessed for their potential benefit to the binding free energy. Which growth vectors are best can be challenging to assess by eye, as an example in the androgen receptor test case the growth vector of H192 is found to beneficially accommodate a hydrogen. An adjacent hydrogen to H192, H17, is found by FEP to be a very unbeneficial position for

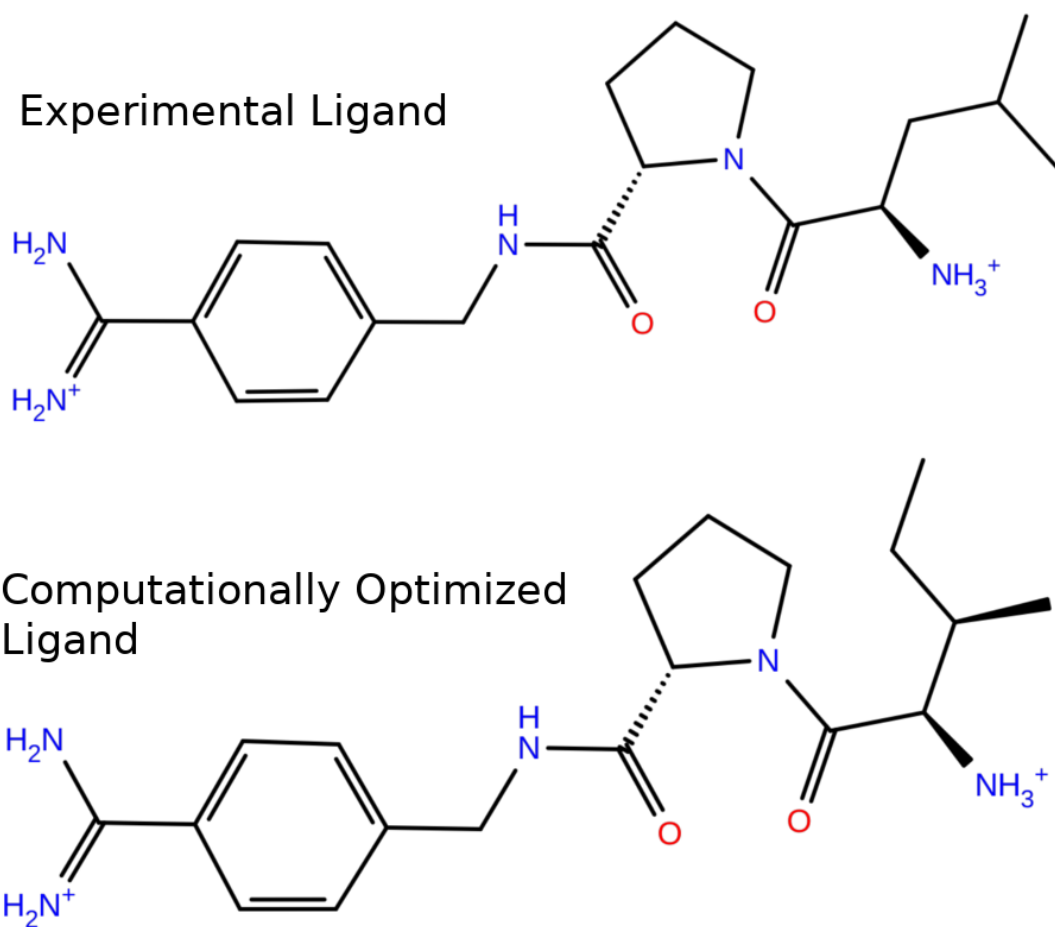


Fig. 4.16 Ligand with improved binding free energy found experimentally and final computationally optimized ligand built by choosing perturbations highlighted by our optimization methodology.

a growth this is despite the growths at H192 and H17 having their closest contacts in the protein be 1.91 and 1.93 angstrom respectively. Whilst this argumentation is not rigorous it does serve to highlight the point that in general determining the best growth vector from the crystal structure of the protein ligand complex can be a difficult task. The method developed in this work is found to be approximately 10 times faster than testing all possible growth vectors with FEP. Using a Nvidia P100 GPU, an optimization for the androgen receptor test case (containing around 4000 atoms in the ligand-protein simulation) takes approximately 13 hours wall time. This is compared to 15 hours per growth vector totaling 150 hours to test all growth vectors with full FEP. Additionally the scaling of the optimization compute time with the number of growth vectors is sublinear compared to linear scaling in the full FEP scan case. The sublinear scaling is a result of using single step perturbation theory in the optimization calculations that allowed any number of growth vectors to be assessed with one molecular dynamics simulation. Where experimental data was available mutations highlighted by the optimizer were tested with full FEP and the mean unsigned error between experimental and calculated values of the binding free energy was 0.83 kcal/mol. We suggest that optimization methods such as this will be useful in a drug discovery setting to identify beneficial growth vectors during the hit-to-lead process, reducing the need for costly trial and error in both computational and experimental campaigns.

This work could be extended by investigating the information that could be extracted by optimizing any other parameters of the system. A natural example might be the epsilon parameter in the LJ potential which may provide more information as to what groups might be beneficially accommodated beyond methyl groups. In theory any parameter of the system could be optimized and looking at the bond, angle or torsion parameters using the methods developed here may also be interesting.

Chapter 5

Machine Learnt Compound Generation

5.1 Introduction

In the previous chapter we applied physics-based models to explore chemical space and exploited SSP methods to perform our exploration using reduced computational resources. These physics-based models rely on particular *a priori* knowledge and input. Namely, the input is a biological target and a crystal structure of this target, but this information is not always available. We now explore the use of ligand-based methods for the generation of compound ideas when no molecular target exists.

In this chapter we will focus on the application of ligand based methods to aid in the development of new antimalarial therapeutics. Multiple ligand based methods will be applied to explore chemical space, where the points in this space are tested using an existing ML model based on the Alchemite [135–137] software. It will be shown that the ML models can predict compound activity quickly relative to FEP methods and that this increased speed, combined with a variety of methods for compound generation, allows for hundreds of thousands of compounds to be tested.

The context for the application of ligand based methods in this chapter is the Open Source Malaria (OSM) consortium [204]. OSM is an initiative which has been evaluating and predicting the activity of compounds against *Plasmodium falciparum* (*pfal*), the parasite which causes the disease malaria in humans. The OSM project has seen multiple rounds of model building. Most recently the focus of these rounds has been a series of compounds based on a triazolopyrazine core, referred to internally as ‘Series 4’ compounds. Whilst the mechanism of action for these Series 4 compounds is not known it is suggested in other work [205] that they kill the parasite by inhibiting an interacellular ion regulatory pump PfATPA and may belong to a larger class of compounds with diverse structures which may target this pump [206]. No crystal structure for this pump exists and as such it is not possible to use

structure based methods in this campaign. Whilst many anti-malaria therapeutics already exist, novel mechanisms of action are particularly important for malaria drugs due to the developing risk of resistance to currently available treatments [207].

The early rounds of the OSM consortium focused on model building for the prediction of *pfal* inhibition; the techniques used in these rounds ranged from pharmacophore [208] to QSAR [209] and DNN models. These rounds ran from 2015-2017 and whilst they were successful in demonstrating the diversity of modern ML methods and the potential of open source initiatives for crowd sourcing expertise, none of the models in these rounds were found to be highly predictive. Given the continued rapid development of machine learning methods in recent years and the increasing application of these methods to aid drug design [27], the OSM consortium decided to run another predictive round in July 2019. This most recent round of the OSM project, round two, involved both model building and now also prospective suggestions for potent compounds to be tested experimentally. Round two saw the participation of eleven independent groups from both academic and industry backgrounds. The groups with the four most predictive models were asked to generate potent compounds which were then synthesized and tested for potency. It is the generation of these compounds that will be a major focus of this chapter.

Some general applications of ML to property prediction have been explored in the machine learning section 1.7 of this thesis and, whilst the specifics of these methods are not the focus here, the overall relative cost of these methods compared to FEP is of interest. The ML models which are used to predict activity can, in general, test points in chemical space extremely cheaply. This allows for predictions for millions of compounds to be computed in the time it would take to compute one FEP prediction. This huge gain in speed can come at the cost of physical accuracy, but this drawback can be mitigated in a few ways. For example, the accuracy can be improved by both remaining close to the training set of a predictive model, and by using robust uncertainty quantification for predictions. In the section 1.7 we also discussed the work of Olivecrona and Blaschke *et al.* and their development of RNNs for the generation of *de novo* compounds. This RNN architecture, and the surrounding theory, will be used in the following methods section in a top down approach to generate compounds to inhibit *pfal*.

5.2 Methods

The work presented here began with the author of this thesis joining one of the teams (Optibrium/Intellegens) already participating in the OSM consortium. At the time of joining, Optibrium/Intellegens had already built a predictive *pfal* model assessed to be in the top four

most predictive models by the OSM consortium. This model was created using Alchemite [135], a software which has shown itself to be useful when applied to the noisy and sparse data sets common in drug discovery projects [210, 211]. Alchemite is based on a similar DNN architecture to that discussed in section 1.7, to recap: 1) the DNN takes as input a vector of molecular descriptors 2) then operates on this vector with a set of linear equations which were parameterized to minimize the error between experimental and predicted data points 3) proceeds to combine the outputs of the linear equations into a value for one or more properties to be predicted. More specifically the Alchemite model is built using 3 interconnected linear layers using a hyperbolic tangent activation function, see the original work of Conduit *et al.* [137] for full details of this method. In the work presented here the input to the Alchemite model is a set of 330 molecular descriptors combined with available values for experimental assays (endpoints) and the output is predictions for the values of endpoints not provided in the input. These molecular descriptors are composed of whole molecular properties such as molecular weight or topological polar surface area, in addition to counts of atoms and chemical groups. The output of this model were predictions for *pfal* pIC50 and ion regulation activities, amongst other properties. The novel aspects of Alchemite that deviate from a general application of a DNN are as follows. The input to the network consists of both molecular descriptors and available assay data. This allows for the correlation between descriptors and endpoints and the correlation between different endpoints to both be used in the prediction of unknown endpoints [135, 137]. Uncertainty estimation is also included in the Alchemite method; this is implemented by training an ensemble of networks with different weights. The variation in the predictions from this ensemble of networks is taken to represent the uncertainty of the prediction [212, 213]. With this Alchemite model in hand to test compounds for *pfal* inhibition, we will now discuss the various methods used in this work for compound generation.

5.2.1 Bottom-Up Generation

One method for compound generation used in this work was referred to as a ‘bottom up’ approach. This involved starting from the triazolopyrazine core and, from there, applying three generations of medicinal chemistry transformations using the Nova module of the StarDrop program [214]. These transformations include roughly twenty thousand molecular modifications such as removing atoms, and adding rings or functional groups. The 50% most diverse compounds in a generation continue to the next round of modification. Filters are applied at generation time to remove any duplicates or compounds with undesirable groups via a substructure match against Stardrops’ curated libraries of undesirable fragments. The result of this generation was hundreds of thousands of compounds which were scored based

on the Alchemite model for *pfal* activity. A multiparameter scoring function was then applied to select molecules with the highest activities balanced across all assays and low uncertainty in the assay predictions.

5.2.2 Top-Down Generation

In the machine learning section 1.7 we saw the application of RNNs to produce valid SMILES strings. The idea of solving for active molecules in a vector space and converting these vectors to molecules with an RNN was also explored. The translation of the vector to a compound with an RNN helps to solve the inverse QSAR problem which arises from representing molecules as a vector of molecular descriptors. Here we apply these methods for the purpose of generating compounds potent against *pfal*; we refer to these methods as a ‘top down’ approach.

As discussed in section 1.7 the previous works in the domain of compound generation used an unsupervised method [149] to generate the vector space. Points in this vector space could then be tested for activity in a discriminative model using the vector as input. If that point vector space was found to be active then the vector was decoded into a molecule using a RNN. For the work presented here the Alchemite model can be used as the discriminative model. At the beginning of this project the Alchemite *pfal* activity model already existed and therefore was fixed, as such its input was also fixed as a vector of 330 molecular descriptors, which we will now call the input vector. Given that the input to the model could not be changed, we modified an RNN from previous work to take this same input vector and produce SMILES strings, an open source version of this modified code is available here: <https://github.com/adw62/ODO>.

Mirroring the methods discussed in section 1.7, the RNN is passed the input vector which is then passed internally by the RNN between a series of cells. As an output each cell generated a probability distribution over all possible characters which can appear in a SMILES string. At training time the weights of the network were optimized to maximize the probability which is assigned to the ‘correct’ character. Where correct is defined by the next character in a training SMILES string. At run time the character with the highest probability was selected for each cell and these characters were arranged sequentially to give the generated compound. To train this RNN we used the same data set used by Olivecrona *et al.* [128], however, in this work we also added all compounds already assayed in the OSM data set [215]. As a reminder this data set consisted of small (10-50 heavy atoms) molecules selected from the ChemBL database. These molecules were filtered to contain only H, B, C, N, O, F, Si, P, S, Cl, Br, I elements. This training data contained 1.5 million compounds in total.

The first tests of this method were made using a simple toy model for the solvation of small molecules. This model was also created using Alchemite, but the input vector was reduced from 330 molecular descriptors, in the case of the *pfal* inhibition model, to just four. These descriptors were: molecular weight (MW), hydrogen bond donors (HBD), hydrogen bond acceptors (HBA) and topological polar surface area (TPSA), where we call these descriptors the ‘solubility descriptors’. The idea is that if a vector of these solubility descriptors is input to the RNN, then the RNN will generate as an output molecules which have similar values for the solubility descriptors which were input. The question, then, is how should the input vector be chosen to obtain a desired solubility?

The input to the Alchemite solubility model comes from a vector space. As has been shown in previous work [149], it is possible to perform an optimization in this vector space to find a vector which minimizes the difference between a target and predicted property, in this case, solubility. The Alchemite program can perform this optimization, and we call the output of this optimization the ‘optimized vector’. If we pass the optimized vector to our RNN we can generate compounds that should have the target solubility. By converting a vector into a molecule, the RNN helps to solve our inverse QSAR problem.

Examples were made for this procedure, one targeting a high solubility with $\log_{10} S = 6.50$, and one targeting molecules with a lower solubility $\log_{10} S = -3.00$. Solving the solubility model for descriptors which match these targeted solubilities gave the vectors shown in table 5.1, named Optimized 1 and Optimized 2. We can see in table 5.1 that the target and optimized solubility are in good agreement for the high solubility example, $\log_{10} S = 6.50$ vs. $\log_{10} S = 6.47$, but for low solubility there is a larger discrepancy between target and optimized solubility, $\log_{10} S = -3.00$ vs. $\log_{10} S = -2.24$. This discrepancy is a consequence of the properties of the function Alchemite has parameterized for predicting solubility, and whether or not a point in that function can be found by the optimizer which gives the targeted solubility.

If the vectors Optimized 1 and Optimized 2 in table 5.1 are passed to the RNN and the compounds generated analysed then the average descriptors of the generated compounds are given by the vectors named in table 5.1 as Generated 1 and Generated 2 respectively. It can be seen that the generated descriptors match the optimized descriptors well, with the exception of molecular weight for the Optimized 1 vector, and HBD for the Optimized 2 vector. This may be caused by the extreme values for the descriptors found by the optimizer. If the optimized vector contains extreme values for the descriptors, such as MW 907, outside of the compounds that the RNN was shown at training time (small molecules from ChEMBL) then generated compounds would be expected to struggle to match the optimized descriptors.

Table 5.1 Solved descriptors for high and low solubility from Alchemite solubility model are given in vectors Optimized 1 and Optimized 2. Average descriptors of generated molecules using Optimized 1 and Optimized 2 as input to the RNN are shown by Generated 1 and Generated 2 respectively. Molecular weight is (MW), hydrogen bond donors (HBD), hydrogen bond acceptors (HBA), topological polar surface area (TPSA). $\log_{10} S$ is the solubility predicted for these vectors by the model.

	MW [daltons]	HBD	HBA	TPSA [\AA^2]	$\log_{10} S$
Optimized 1	907	0.4	1.3	47.3	-2.24
Generated 1	755	0.3	2.6	47.4	-0.74
Optimized 2	211	13.8	5.6	103.5	6.47
Generated 2	224	6.9	6.9	101.5	5.74

The average solubility for the compounds generated was assessed by the Alchemite model and the Optimized 1 vector generated molecules with an average solubility $\log_{10} S = -0.74$ with a standard deviation of 1.42, and the Optimized 2 vector generated molecules with an average $\log_{10} S = 5.7$ with a standard deviation of 0.25. The difference in the optimized and generated solubility most likely stems from the difference in Optimized and Generated vectors in table 5.1. One additional consideration is that the RNN may fail to grow large heavy molecules because of the increasing probability of the SMILES terminating as they increase in length. Figure 5.1 shows examples of high and low solubility molecules that have been generated.

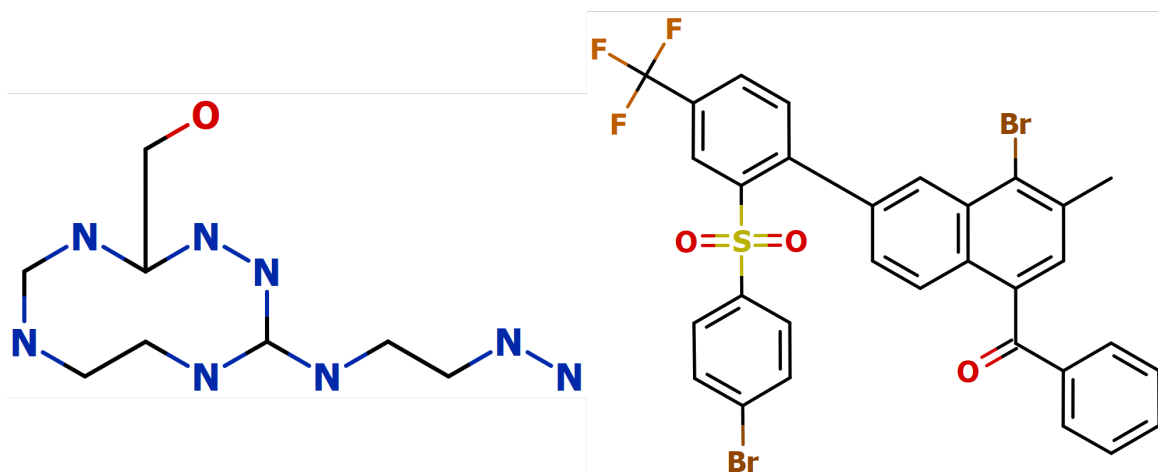


Fig. 5.1 Examples of soluble and insoluble molecules produced by the RNN. Soluble and insoluble molecules depicted in left and right panels respectively.

The compounds in figure 5.1 do not appear in the training set of the RNN meaning we have some confidence that this method allows us to generate novel compounds with a desired target property in the case of this toy solubility example. One of the clear problems with these generated compounds, however, is their impossible chemistry. This is a problem that we will see repeated when ‘optimized’ vectors are used to generate compounds for *pfal* activity.

To generate compounds for *pfal* activity the same method tested with the toy solubility model can be applied to the *pfal* inhibition model. An example molecule generated with this method is shown in figure 5.2. Whilst the molecule in figure 5.2 is calculated to be active by the Alchemite model, and is technically a valid SMILES, it can be seen by eye that there are significant problems with the chemistry of this molecule. It may be possible to address these problems with additional filters for synthesizability or introducing a drug likeness score in the RNN at training time, but we do not examine these ideas here and instead our application of this generation method will use a filter of human intuition at the end to ‘correct’ any high scoring compounds to be more synthesizable and less reactive.

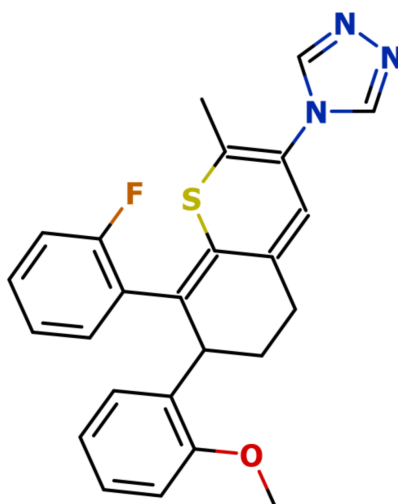


Fig. 5.2 An example compound generated using an optimized vector as input to the RNN.

To generate the compound in figure 5.2, an optimized vector was passed to the RNN. It will be shown in the results that compounds generated with this method can be scored active by the *pfal* model. However, clearly the output compounds generated by the RNN in this case have problems with their chemistry. One solution to this was instead of passing an optimized vector to the RNN we passed a vector for a known experimentally active Series 4 compound. If a known active vector was passed to the RNN, the compound generated by the RNN resemble very closely the input molecule/Series 4, and as such was far more reasonable in terms of both their chemistry and synthesizability, this can be seen in figure 5.3.

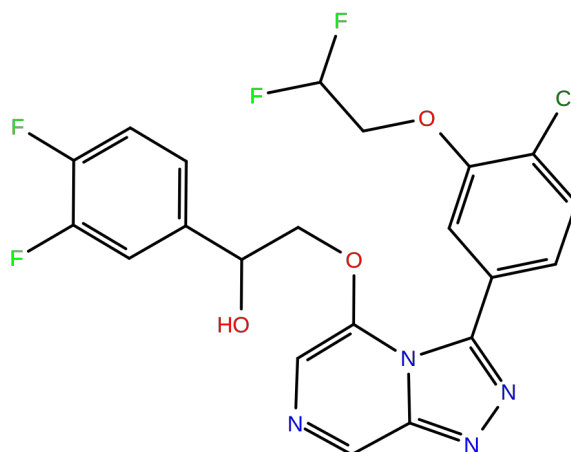


Fig. 5.3 An example compound generated using a known active vector as input to the RNN.

5.2.3 Other OSM methods

As discussed in the introduction to this chapter, round two of the OSM saw the participation of eleven independent groups from both academic and industry backgrounds which all used different methods to build their predictive models and generate compound suggestions. We will discuss here briefly the methods used by the other collaborating groups for compound generation. Molomics are an industry entry and used a collective intelligence approach to generate compounds. This involved using a ‘live design’ application where many users can draft molecules collaboratively and have their idea scored in real time by the Molomics *pfal* inhibition model. Molomics submissions for synthesis were selected by consensus decision with a focus on synthesizability. David Guan was an academic entry and to generate compounds they used a population-based approach named ChemGE based on the methods of grammatical evolution [216]. This method was biased towards generating useful molecules similar to the triazolopyrazine core by using a similarity score as well as scores for synthetic accessibility, lipophilicity, and an aromatic ring penalty.

5.2.4 Summary of Methods

To summarize, in this work we will use three methods for molecule generation, these are: a) a bottom-up approach where medicinal chemistry transforms are applied to a compound series core, we will refer to these compounds as ‘med chem’ compounds. b) a top-down approach where compounds are generated from an optimized vector of descriptors found by ‘solving’ an Alchemite model, we will call compounds generated with this method ‘alch opt’ compounds. Finally c) another top-down approach, where compounds are generated using a

vector of descriptors describing a known active molecule as input to the RNN, we will call compounds generated with this method ‘active vector’ compounds.

5.3 Results

Based on the models and compound generation techniques discussed above, we generated compounds to be submitted to the OSM consortium. The specifications given by the OSM consortium were the following: of the four compound submissions made by each group, two molecules should be structurally distinct from Series 4, and two should be based on Series 4. Our alch opt compounds were the most structurally distinct from Series 4, but as mentioned human intervention was required to address reactivity and synthesizability issues with these compounds. Table 5.2 shows three raw alch opt molecules generated by the RNN alongside the human corrected molecules. Table 5.3 shows activities predicted by the Alchemite’s *pfal* inhibition model for all of these alch opt compounds.

For all the generation methods used here the presented molecules were selected from a much larger set of generated molecules. They were selected using a multiparameter scoring function tuned for a good balance of activities across assays and low uncertainty in the assay predictions. For our discussion of the results the criteria for active potency used by the OSM group will be used here and this a $pIC_{50} > 6$. One point to note is that the Alchemite model generates output predictions for multiple assays including multiple pIC_{50} assays from different labs (Avery, Dundee and GSK). Provided that measurements of pIC_{50} are performed accurately there should only be small variations between the pIC_{50} value of different labs. The Alchemite model, however, has no knowledge of this and is free to predict different values of pIC_{50} for different labs. Anecdotally the model is more likely to assign different pIC_{50} s across labs the further the input molecule was from the training set of molecules use to train the Alchemite model.

Table 5.2 contains three molecules which are raw output from the RNN (1a)-3a), and three molecules which are corrected with human intuition (1b)-3b). In table 5.3 it was predicted that all molecules would be considered active in the Avery and GSK pIC_{50} assays; none of these molecules were predicted to be active in the Dundee assay. In the experimental work ion regulation activity is assessed using a NA^+ reactive florescent dye, sodium-binding benzofuran isophthalate [217]. When recorded in the OSM data set the raw data from the observation of this experiment are converted to binary where 1 denotes ion regulation and 0 denotes no ion regulation. It can therefore be seen in table 5.3 that the predicted ion regulation for compounds 1a-3b is low or none.

Table 5.2 a) compounds are alch opt compounds, generated by the RNN using an optimized Alchemite vector as input. b) compounds are the compounds which have been ‘corrected’ with human intuition.

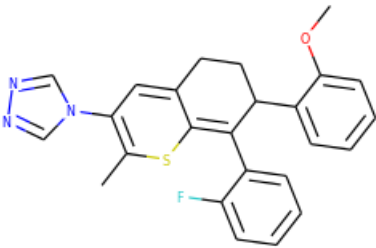
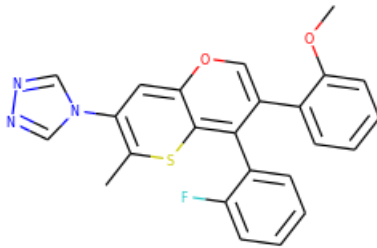
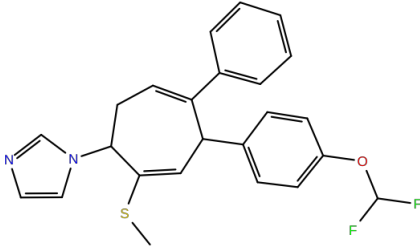
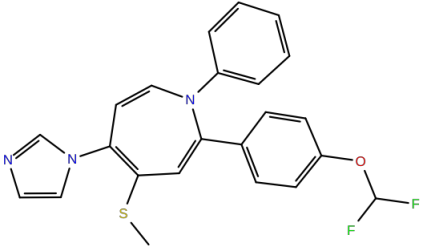
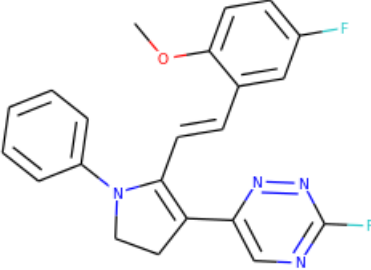
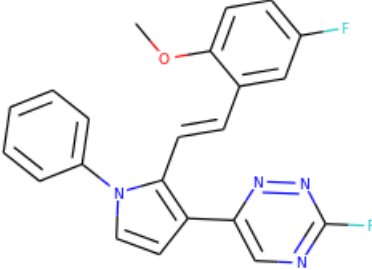
ID	(a) Compounds	(b) Compounds
		
1a/b		
2a/b		
3a/b		

Table 5.3 Predicted activity for alch opt compounds. pIC50 is the *pfal* pIC50. Square brackets show the prediction \pm the uncertainty predicted by Alchemite's ensemble based method [212, 213]. Ion regulation activity varies between 1.0 and 0.0 where 1.0 indicates ion regulation was observed and 0.0 indicates ion regulation was not observed.

ID	ion regulation activity	pIC50 (Avery)	pIC50 (Dundee)	pIC50 (GSK)
1a	0.25 [-0.13, 0.64]	7.56 [6.54, 8.58]	5.37 [4.73, 6.02]	7.23 [6.76, 7.70]
1b	0.09 [-0.15, 0.32]	6.88 [5.74, 8.03]	5.18 [4.54, 5.82]	7.22 [6.74, 7.70]
2a	0.35 [-0.11, 0.81]	7.32 [6.17, 8.46]	5.21 [4.69, 5.74]	7.02 [6.52, 7.52]
2b	0.18 [-0.18, 0.54]	7.24 [6.30, 8.19]	5.34 [4.79, 5.89]	7.26 [6.83, 7.70]
3a	0.15 [-0.20, 0.49]	6.99 [5.61, 8.38]	5.26 [4.70, 5.82]	6.92 [6.28, 7.55]
3b	0.27 [-0.16, 0.70]	6.19 [4.78, 7.60]	5.45 [4.86, 6.04]	6.22 [5.44, 7.00]

The next round of generation pertains to the active vector compounds discussed in the methods section. The compounds generated at this stage were filtered for duplicates in the OSM data set and many compounds were found to be unique. Table 5.4 shows high activity compounds selected from this round of generation, and table 5.5 shows the predicted activity for these molecules. The pIC50 activities predicted in table 5.5 were lower compared to the alch opt compounds in table 5.3, but the predicted pIC50s were still promising and both molecules 4 and 5 were predicted to be active in the Dundee and GSK assays. The predicted ion regulation, however, was significantly higher than for the alch opt compounds.

The final round of generation then considered the med chem compounds. Using the bottom-up method discussed prior, hundreds of thousands of compounds were generated using medicinal chemistry transformations applied to the series core, and the best compounds that were selected from this generation are presented in table 5.6. The predicted activities of compounds in table 5.6 are shown in table 5.7. Overall the predictions were good; all assays were predicted to be active with the exception of the Avery assay for compound 7. Again the ion regulation activity for the compounds in table 5.7 was much higher than the alch opt compounds in table 5.3.

Based on the brief given by the OSM consortium and considerations for reactivity and synthesizability the 4 molecules we chose for submission were 1b, 2b, 6 and 7. All these compounds were then considered for synthetic routes by the OSM consortium. It can be seen in table 5.7 that compound 6 has reactivity issues with the difluoromethyl. After submission this was flagged by eye as having potential to eliminate to HF, a neurotoxin, and therefore compound 6 was not considered for synthesis. Compounds 1b and 2b were found by the OSM consortium not to have reasonable routes for synthesis. With this in mind only compound 7 was synthesized and tested at the Dundee lab and found to have a pIC50 of 6.2 very close to the predicted value of 6.4.

Table 5.4 Compounds generated using a known active vector as input to the RNN.

ID	Compound
4	
5	

Table 5.5 Predicted activity for active vector compounds. pIC50 is the *pfal* pIC50. Square brackets show the prediction \pm the uncertainty predicted by Alchemite's ensemble based method [212, 213]. Ion regulation activity varies between 1.0 and 0.0 where 1.0 indicates ion regulation was observed and 0.0 indicates ion regulation was not observed.

ID	ion regulation activity	pIC50 (Avery)	pIC50 (Dundee)	pIC50 (GSK)
4	0.90 [0.61, 1.18]	5.19 [4.07, 6.31]	6.38 [5.65, 7.11]	6.46 [5.96, 6.95]
5	0.90 [0.62, 1.18]	5.69 [4.23, 7.15]	6.33 [5.54, 7.11]	6.57 [6.09, 7.06]

Table 5.6 Compounds generated by expanding series 4 triazolopyrazine core with medicinal chemistry transformations.

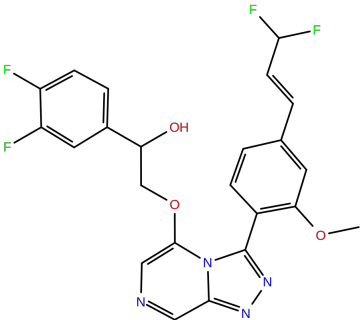
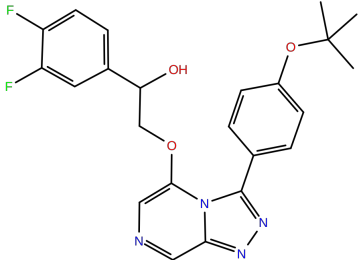
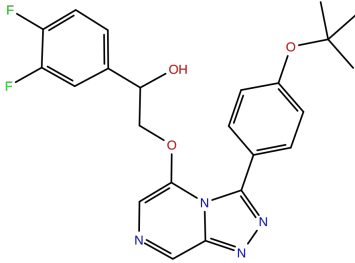
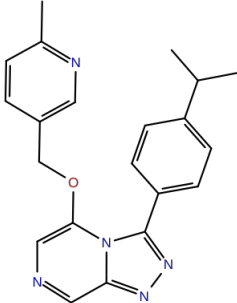
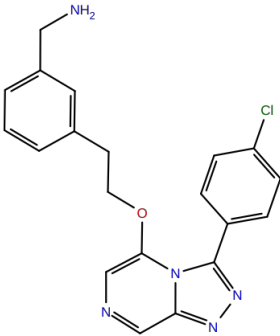
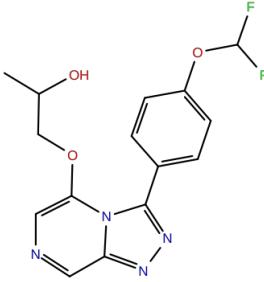
ID	Compound
6	
7	

Table 5.7 Predicted activity for med chem compounds. pIC50 is the *pfal* pIC50. Square brackets show the prediction \pm the uncertainty predicted by Alchemite's ensemble based method [212, 213]. Ion regulation activity varies between 1.0 and 0.0 where 1.0 indicates ion regulation was observed and 0.0 indicates ion regulation was not observed.

ID	ion regulation activity	pIC50 (Avery)	pIC50 (Dundee)	pIC50 (GSK)
6	0.92 [0.67, 1.17]	6.32 [4.8, 7.83]	6.61 [5.94, 7.27]	6.54 [5.95, 7.12]
7	0.91 [0.63, 1.19]	5.31 [4.17, 6.46]	6.42 [5.67, 7.17]	6.47 [6.0, 6.95]

Table 5.8 Experimental pIC50 for OSM compounds. The Optibrium/Intellegens entrant is compound 7 designed in this work. Davy Guan, Exscientia and Molomics were collaborating groups who also submitted molecules to the OSM consortium.

Entrant	Compound	pIC50 (Dundee)
Optibrium/Intellegens		6.2
Davy Guan		<4.6
Exscientia		5.0
Molomics		<4.6

Looking at all synthesized compounds submitted across all collaborating groups (Davy Guan, Exscientia and Molomics) in table 5.8 it can be seen that overall this work is a success and that 25% of the synthesised molecules were found to be active. This is a small sample size and so caution should be taken in drawing any conclusions, but it is interesting to note that the active rate achieved in these computational works is comparable to the active rate for the drug discovery effort to date, which has so far been driven by the intuition of medicinal chemists. For comparison, there are around 232 existing compounds evaluated by the Dundee lab in the OSM series 4 data set [215] and 55 of these are considered active which gives roughly a 25% active rate.

5.4 Discussion

The bulk of this work was performed within strict time constraints dictated by the timeline on which the OSM consortium was operating. As such many ideas and potential improvements to methods used in this work could not be explored. A discussion on some of the shortcomings of this work and potential avenues for improvement in future work is now given.

One issue with the generation of compounds that repeatedly occurred in this work was the generation of invalid chemistry. This became a problem increasingly as the input to the RNN became more distinct from the training data of the RNN [149]. This can be seen clearly when comparing the molecules generated from the optimized vector against molecules generated from the known active vector. Where the compounds generated from the known active vector, which is safely within the training set, show far fewer problems with their chemistry than compounds generated from the optimized vector. The solution here would be to restrict the optimization to search a vector space which is well overlapped with the space of vectors with which the RNN was trained. One caveat here though is that any restriction of the search space has the potential to increase the difficulty of the optimization. This may result in less potent solutions being found but the value of these solutions will be much greater as they can be decoded by the RNN more robustly.

Another similar issue pertains to the overlap of *pfal* model vector input with the *pfal* model training data. Again, if the vector input to the model is far from the training data, the model's predictions will be inaccurate. Since this model was trained on Series 4 compounds the prediction for compounds structurally distinct from the training set, the predicted activities of compounds 1a-3b for example, are highly likely inaccurate. To resolve this any predictions made by the models should be first validated by checking the input is close to the training data. This solution is at odds, however, with the design brief of the OSM consortium which asked for structurally diverse molecules. Whilst the OSM consortium was interested in

structurally diverse compounds, in fact, none of the submitted molecules that constituted diverse compounds, from any group, were synthesized because either no submission was made or no route to synthesize the compounds could be found. The difficulty in accurately predicting activity and finding synthetic routes for structurally diverse molecules highlights some major obstacles for ligand based methods in computational drug discovery.

5.5 Conclusion

In this chapter we have seen the application of various machine learning methods in the domains of activity prediction and compound generation. The application of these methods has allowed for hundreds of thousands of compounds to be tested for activity against *Plasmodium falciparum*. Using a combination of generation methods and significant human intervention a compound was selected which was then confirmed experimentally to be active with close agreement between the predicted, 6.4, and experimental values, 6.2, for pIC₅₀. The selected compound would not have been chosen by traditional methods based on the SAR data available to the project but since it was confirmed to be active it represents a valuable data point to the OSM project and demonstrates the value which can be added by ML methods for activity prediction and compound generation. The reduced computational cost of the ligand based methods applied here allowed for far more compounds to be tested than any other method explored in this thesis and demonstrates that ML based modelling can play a useful role in the efficient exploration of chemical space.

Chapter 6

Final Discussion of Results

In this work several novel methods for improving the efficiency of ligand-binding affinity calculations have been presented. Approaches based on free energy methods were presented in chapters 2, 3, and 4. In chapter 2 a computational fluorine scanning method was presented which used single step perturbation theory to quickly assess the relative binding affinity of fluorinated analogues. The predictions from this method were in excellent agreement with more rigorous alchemical free-energy calculations, and in good agreement with experimental data for most of the test systems. However, the agreement with experiment was very poor in some of the test systems, particularly the DPP4 system, and it was hypothesised that this may stem from simulating the system as a monomer compared to its dimer biological unit. In terms of improved computational efficiency, PFS was five times faster to calculate the same number of relative binding free energies than full FEP calculations. The example that was given was that 11 fluorinated FXa compounds could be tested in 9.5 hours with PFS compared to 44 hours with full FEP.

Chapter 3 presented a method to systematically optimize a set of ligand charges to maximize the ligand-protein binding affinity. This optimization method used SSP to calculate both the objective and gradient in the search for a set of optimal charges. From this set of optimal charges design ideas for beneficial chemical mutations could be extracted. When tested with more rigorous free energy methods it was shown that 73% of these design ideas were beneficial. It was demonstrated that charge optimization in an explicit solvent was a useful tool for predicting beneficial chemical changes such as pyridinations, fluorinations, and oxygen to sulphur mutations. We estimated in this chapter that to test all of the fluorinated analogues of the FXa ligand would take roughly three days of computer time with full FEP compared to the two days required for the optimization to qualitatively assess all allowed fluorination, in addition to pyridination or oxygen to sulphur mutations.

In chapter 4 the charge optimization method was extended to consider the steric parameters of a ligand. Modifications were made to the optimization method in this chapter such that the MBAR method was used to calculate the objective and single step perturbation used to calculate the gradient. Again, from the optimized set of parameters design ideas for affinity improving mutations could be extracted. In this chapter the design ideas pertain exclusively to beneficial growth vectors for methylation. The predictions from the optimizer for the ranking of growth vectors correlated with existing free energy methods with a Spearman's rank order correlation of 0.59. Using this optimization method allowed for approximately a ten times speed up in the testing of all growth vectors. With an example given for the androgen receptor test case that the optimization took approximately 13 hours of wall time to rank all growth vectors compared to 15 hours per growth vector totaling 150 hours to test all growth vectors with full FEP.

Finally, chapter 5 discussed a collaborative effort performed with the OSM consortium. We presented an application of machine learned property prediction and compound generation to find compounds potent against *Plasmodium falciparum*. This work involved the leveraging of machine learning methods for the testing of hundreds of thousands of compounds, far more than would have been possible with free energy methods. The result of this chapter was a selection of compounds predicted to be active from which one was evaluated experimentally by the OSM and verified to be active with a pIC₅₀ of 6.2 in good agreement with the computational prediction of 6.42 ± 0.75 .

References

- [1] Glenn E Croston. The utility of target-based discovery. *Expert Opin. Drug Discov.*, 12(5):427–429, May 2017.
- [2] John G Moffat, Fabien Vincent, Jonathan A Lee, Jörg Eder, and Marco Prunotto. Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nat. Rev. Drug Discov.*, 16(8):531–543, August 2017.
- [3] David Hardy, Roslyn M Bill, Anass Jawhari, and Alice J Rothnie. Overcoming bottlenecks in the membrane protein structural biology pipeline. *Biochem. Soc. Trans.*, 44(3):838–844, June 2016.
- [4] Alexander McPherson and Jose A Gavira. Introduction to protein crystallization. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.*, 70(Pt 1):2–20, January 2014.
- [5] Jessica Lynn Grey and David H Thompson. Challenges and opportunities for new protein crystallization strategies in structure-based drug design. *Expert Opin. Drug Discov.*, 5(11):1039–1045, November 2010.
- [6] Takayuki Kondo, Masashi Asai, Kayoko Tsukita, Yumiko Kutoku, Yutaka Ohsawa, Yoshihide Sunada, Keiko Imamura, Naohiro Egawa, Naoki Yahata, Keisuke Okita, Kazutoshi Takahashi, Isao Asaka, Takashi Aoi, Akira Watanabe, Kaori Watanabe, Chie Kadoya, Rie Nakano, Dai Watanabe, Kei Maruyama, Osamu Hori, Satoshi Hibino, Tominari Choshi, Tatsutoshi Nakahata, Hiroyuki Hioki, Takeshi Kaneko, Motoko Naitoh, Katsuhiko Yoshikawa, Satoko Yamawaki, Shigehiko Suzuki, Ryuji Hata, Shu-ichi Ueno, Tsuneyoshi Seki, Kazuhiro Kobayashi, Tatsushi Toda, Kazuma Murakami, Kazuhiro Irie, William L Klein, Hiroshi Mori, Takashi Asada, Ryosuke Takahashi, Nobuhisa Iwata, Shinya Yamanaka, and Haruhisa Inoue. Modeling alzheimer’s disease with iPSCs reveals stress phenotypes associated with intracellular A β and differential drug responsiveness. *Cell Stem Cell*, 12(4):487–496, April 2013.
- [7] C Dirk Keene, Martin Darvas, Brian Kraemer, Denny Liggitt, Christina Sigurdson, and Warren Ladiges. Neuropathological assessment and validation of mouse models for alzheimer’s disease: applying nia-aa guidelines. *Pathobiology of Aging & Age-related Diseases*, 6, 2016.
- [8] Elisa Michelini, Luca Cevenini, Laura Mezzanotte, Andrea Coppa, and Aldo Roda. Cell-based assays: fuelling drug discovery. *Anal. Bioanal. Chem.*, 398(1):227–238, September 2010.
- [9] Molly Hunter, Ping Yuan, Divya Vavilala, and Mark Fox. Optimization of protein expression in mammalian cells. *Curr. Protoc. Protein Sci.*, 95(1):e77, February 2019.

- [10] Paul T Wingfield. Overview of the purification of recombinant proteins. *Curr. Protoc. Protein Sci.*, 80:6.1.1–6.1.35, April 2015.
- [11] Lutea A A de Jong, Donald R A Uges, Jan Piet Franke, and Rainer Bischoff. Receptor-ligand binding assays: technologies and applications. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.*, 829(1-2):1–25, December 2005.
- [12] Thomas D Pollard. A guide to simple and informative binding assays. *Mol. Biol. Cell*, 21(23):4061–4067, December 2010.
- [13] Leigh A Stoddart, Carl W White, Kim Nguyen, Stephen J Hill, and Kevin D G Pfleger. Fluorescence- and bioluminescence-based approaches to study GPCR ligand binding. *Br. J. Pharmacol.*, 173(20):3028–3037, 2016.
- [14] Davide Capelli, Chiara Parravicini, Giorgio Pochetti, Roberta Montanari, Caterina Temporini, Marco Rabuffetti, Maria Letizia Trincavelli, Simona Daniele, Marta Fumagalli, Simona Saporiti, Elisabetta Bonfanti, Maria P Abbraccio, Ivano Eberini, Stefania Ceruti, Enrica Calleri, and Stefano Capaldi. Surface plasmon resonance as a tool for ligand binding investigation of engineered GPR17 receptor, a G protein coupled receptor involved in myelination. *Front Chem*, 7:910, 2019.
- [15] Christoph Nitsche and Gottfried Otting. NMR studies of ligand binding. *Curr. Opin. Struct. Biol.*, 48:16–22, February 2018.
- [16] Laurent Maveyraud and Lionel Mourey. Protein x-ray crystallography and drug discovery. *Molecules*, 25(5), February 2020.
- [17] C John Harris, Richard D Hill, David W Sheppard, Martin J Slater, and Pieter F W Stouten. The design and application of target-focused compound libraries. *Comb. Chem. High Throughput Screen.*, 14(6):521–531, July 2011.
- [18] Duncan E Scott, Anthony G Coyne, Sean A Hudson, and Chris Abell. Fragment-based approaches in drug discovery and chemical biology. *Biochemistry*, 51(25):4990–5003, June 2012.
- [19] Hugo Brito, Vanda Marques, Marta B Afonso, Dean G Brown, Ulf Börjesson, Nidhal Selmi, David M Smith, Ieuan O Roberts, Martina Fitzek, Natália Aniceto, et al. Phenotypic high-throughput screening platform identifies novel chemotypes for necroptosis inhibition. *Cell Death Discov.*, 6(1):1–13, 2020.
- [20] Yoram Etzion and Anthony J Muslin. The application of phenotypic high-throughput screening techniques to cardiovascular research. *Trends Cardiovasc. Med.*, 19(6):207–212, August 2009.
- [21] Natasja Brooijmans and Irwin D Kuntz. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.*, 32:335–373, January 2003.
- [22] Pradeep Anand Ravindranath, Stefano Forli, David S Goodsell, Arthur J Olson, and Michel F Sanner. AutoDockFR: Advances in Protein-Ligand docking with explicitly specified binding site flexibility. *PLoS Comput. Biol.*, 11(12):e1004586, December 2015.

- [23] J Huuskonen. Estimation of aqueous solubility in drug design. *Comb. Chem. High Throughput Screen.*, 4(3):311–316, May 2001.
- [24] Igor V Tetko and Alexander Tropsha. Joint virtual special issue on computational toxicology. *J. Chem. Inf. Model.*, 60(3):1069–1071, March 2020.
- [25] Lars Ruddigkeit, Lorenz C Blum, and Jean-Louis Reymond. Visualization and virtual screening of the chemical universe database GDB-17. *J. Chem. Inf. Model.*, 53(1):56–65, January 2013.
- [26] Markus Hartenfeller, Heiko Zettl, Miriam Walter, Matthias Rupp, Felix Reisen, Ewgenij Proschak, Sascha Weggen, Holger Stark, and Gisbert Schneider. DOGS: reaction-driven de novo design of bioactive compounds. *PLoS Comput. Biol.*, 8(2), 2012.
- [27] W Patrick Walters and Mark Murcko. Assessing the impact of generative AI on medicinal chemistry. *Nat. Biotechnol.*, 38(2):143–145, February 2020.
- [28] Paul R Leger, Dennis X Hu, Berenger Biannic, Minna Bui, Xiping Han, Emily Karbarz, Jack Maung, Akinori Okano, Maksim Osipov, Grant M Shibuya, Kyle Young, Christopher Higgs, Betty Abraham, Delia Bradford, Cynthia Cho, Christophe Colas, Scott Jacobson, Yamini M Ohol, Deepa Pookot, Payal Rana, Jerick Sanchez, Niket Shah, Michael Sun, Steve Wong, Dirk G Brockstedt, Paul D Kassner, Jacob B Schwarz, and David J Wustrow. Discovery of potent, selective, and orally bioavailable inhibitors of USP7 with in vivo antitumor activity. *J. Med. Chem.*, 63(10):5398–5420, May 2020.
- [29] Robert Abel, Lingle Wang, Edward D Harder, B J Berne, and Richard A Friesner. Advancing drug discovery through enhanced free energy calculations. *Acc. Chem. Res.*, 50(7):1625–1632, July 2017.
- [30] D J Danziger and P M Dean. Automated site-directed drug design: a general algorithm for knowledge acquisition about hydrogen-bonding regions at protein surfaces. *Proc. R. Soc. Lond. B Biol. Sci.*, 236(1283):101–113, March 1989.
- [31] Karen Sparck Jones. *A brief informal history of the Computer Laboratory*, 2001 (accessed 2020-08-24). <https://www.cl.cam.ac.uk/events/EDSAC99/history.html>.
- [32] H J Böhm. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput. Aided Mol. Des.*, 8(3):243–256, June 1994.
- [33] H J Böhm. Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J. Comput. Aided Mol. Des.*, 12(4):309–323, July 1998.
- [34] Robert S DeWitte and Eugene I Shakhnovich. SMOG: de novo design method based on simple, fast, and accurate free energy estimates. 1. methodology and supporting evidence. *J. Am. Chem. Soc.*, 118(47):11733–11744, January 1996.
- [35] A V Ishchenko and E I Shakhnovich. Small molecule growth 2001 (SMoG2001): An improved knowledge-based scoring function for protein ligand interactions. *J. Med. Chem.*, 2002.

- [36] Renxiao Wang, Ying Gao, and Luhua Lai. LigBuilder: A Multi-Purpose program for Structure-Based drug design. *Molecular modeling annual*, 6(7):498–516, August 2000.
- [37] J Zhu, H Fan, H Liu, and Y Shi. Structure-based ligand design for flexible proteins: application of new F-DycoBlock. *J. Comput. Aided Mol. Des.*, 15(11):979–996, November 2001.
- [38] Gisbert Schneider and Uli Fechner. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discov.*, 4(8):649–663, August 2005.
- [39] David A Pearlman and Mark A Murcko. CONCEPTS: New dynamic algorithm for de novo drug suggestion. *J. Comput. Chem.*, 14(10):1184–1193, October 1993.
- [40] D A Pearlman and M A Murcko. CONCERTS: dynamic connection of fragments as an approach to de novo ligand design. *J. Med. Chem.*, 39(8):1651–1663, April 1996.
- [41] Michele Perrella and Rosaria Russo. Allosteric proteins: Lessons to be learned from the hemoglobin intermediates. *Physiology*, 18(6):232–236, 2003.
- [42] Sandor Vajda, Dmitri Beglov, Amanda E Wakefield, Megan Egbert, and Adrian Whitty. Cryptic binding sites on proteins: definition, detection, and druggability. *Curr Opin Chem Biol*, 44:1–8, 2018.
- [43] Cecilia Nardini. The ethics of clinical trials. *Ecancermedicalscience*, 8:387, January 2014.
- [44] Linda Martin, Melissa Hutchens, Conrad Hawkins, and Alaina Radnov. How much do clinical trials cost? *Nat. Rev. Drug Discov.*, 16(6):381–382, June 2017.
- [45] R S Bohacek, C McMartin, and W C Guida. The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.*, 16(1):3–50, January 1996.
- [46] Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, 23(1):3–25, January 1997.
- [47] Ming-Qiang Zhang and Barrie Wilkinson. Drug discovery beyond the 'rule-of-five'. *Curr. Opin. Biotechnol.*, 18(6):478–488, December 2007.
- [48] Miles Congreve, Robin Carr, Chris Murray, and Harren Jhoti. A 'rule of three' for fragment-based lead discovery? *Drug Discov. Today*, 19(8):876–877, 2003.
- [49] Daniel F Veber, Stephen R Johnson, Hung-Yuan Cheng, Brian R Smith, Keith W Ward, and Kenneth D Kopple. Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.*, 45(12):2615–2623, June 2002.
- [50] D E Clark and S D Pickett. Computational methods for the prediction of 'drug-likeness'. *Drug Discov. Today*, 5(2):49–58, February 2000.

- [51] Joseph A DiMasi, Henry G Grabowski, and Ronald W Hansen. Innovation in the pharmaceutical industry: New estimates of R&D costs. *J. Health Econ.*, 47:20–33, May 2016.
- [52] Lingle Wang, Yujie Wu, Yuqing Deng, Byungchan Kim, Levi Pierce, Goran Krilov, Dmitry Lupyan, Shaughnessy Robinson, Markus K Dahlgren, Jeremy Greenwood, et al. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.*, 137(7):2695–2703, 2015.
- [53] Caterina Bissantz, Bernd Kuhn, and Martin Stahl. A medicinal chemist’s guide to molecular interactions. *J. Med. Chem.*, 53(14):5061–5084, July 2010.
- [54] P Eastman, J Swails, J D Chodera, and others. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.*, 2017.
- [55] Mark T Nelson, William Humphrey, Attila Gursoy, Andrew Dalke, Laxmikant V Kalé, Robert D Skeel, and Klaus Schulten. NAMD: a parallel, Object-Oriented molecular dynamics program. *The International Journal of Supercomputer Applications and High Performance Computing*, 10(4):251–268, December 1996.
- [56] David A Case, T A Darden, T E Cheatham, Carlos L Simmerling, Junmei Wang, Robert E Duke, Ray Luo, Mrcw Crowley, Ross C Walker, Wei Zhang, and Others. Amber 10. Technical report, University of California, 2008.
- [57] Tamar Schlick, Rosana Collepardo-Guevara, Leif Arthur Halvorsen, Segun Jung, and Xia Xiao. Biomolecular modeling and simulation: a field coming of age. *Q. Rev. Biophys.*, 44(2):191, 2011.
- [58] D Wolf, V Yamakov, S R Phillpot, A Mukherjee, and H Gleiter. Deformation of nanocrystalline materials by molecular-dynamics simulation: relationship to experiments? *Acta Mater.*, 53(1):1–40, January 2005.
- [59] Kimberly Chenoweth, Adri C T van Duin, and William A Goddard, 3rd. ReaxFF reactive force field for molecular dynamics simulations of hydrocarbon oxidation. *J. Phys. Chem. A*, 112(5):1040–1053, February 2008.
- [60] Martin Karplus and J Andrew McCammon. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.*, 9(9):646–652, September 2002.
- [61] J A McCammon, B R Gelin, and M Karplus. Dynamics of folded proteins. *Nature*, 267(5612):585–590, June 1977.
- [62] Benedict Leimkuhler and Charles Matthews. Rational construction of stochastic numerical methods for molecular sampling. *Applied Mathematics Research eXpress*, 2013(1):34–56, 2013.
- [63] Agastya P Bhati, Shunzhou Wan, David W Wright, and Peter V Coveney. Rapid, accurate, precise, and reliable relative free energy prediction using ensemble based thermodynamic integration. *J. Chem. Theory Comput.*, 13(1):210–222, 2017.

- [64] Robert D Skeel, Guihua Zhang, and Tamar Schlick. A family of symplectic integrators: stability, accuracy, and molecular dynamics applications. *J. Sci. Comput.*, 18(1):203–222, 1997.
- [65] MBBJM Tuckerman, Bruce J Berne, and Glenn J Martyna. Reversible multiple time scale molecular dynamics. *J. Chem. Phys.*, (3):1990–2001, 1992.
- [66] Tamar Schlick. *Molecular modeling and simulation: an interdisciplinary guide: an interdisciplinary guide*, volume 21. Springer Science & Business Media, 2010.
- [67] Hans C Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.*, 72(4):2384–2393, 1980.
- [68] Kim-Hung Chow and David M Ferguson. Isothermal-isobaric molecular dynamics simulations with monte carlo volume sampling. *Comput. Phys. Commun.*, 91(1-3):283–289, 1995.
- [69] Herman JC Berendsen, JPM van Postma, Wilfred F van Gunsteren, ARHJ DiNola, and Jan R Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81(8):3684–3690, 1984.
- [70] Mattia Bernetti and Giovanni Bussi. Pressure control using stochastic cell rescaling. *J. Chem. Phys.*, 153(11):114107, 2020.
- [71] Andrea Rizzi, Travis Jensen, David R Slochower, Matteo Aldeghi, Vytautas Gapsys, Dimitris Ntekoumes, Stefano Bosisio, Michail Papadourakis, Niel M Henriksen, Bert L De Groot, et al. The sampl6 sampling challenge: assessing the reliability and efficiency of binding free energy calculations. *J. Comput. Aided Mol. Des.*, 34(5):601–633, 2020.
- [72] Jerelle A Joseph, Konstantin Röder, Debayan Chakraborty, Rosemary G Mantell, and David J Wales. Exploring biomolecular energy landscapes. *Chem. Comm.*, 53(52):6974–6988, 2017.
- [73] Jose M Sanchez-Ruiz. Protein kinetic stability. *Biophys. Chem.*, 148(1-3):1–15, 2010.
- [74] Johannes Kästner. Umbrella sampling. *WIREs Comput Mol Sci*, 1(6):932–942, November 2011.
- [75] E Marinari and G Parisi. Simulated tempering: A new monte carlo scheme. *EPL*, 19(6):451, July 1992.
- [76] A P Lyubartsev, A A Martsinovski, S V Shevkunov, and P N Vorontsov-Velyaminov. New approach to monte carlo calculation of the free energy: Method of expanded ensembles. *J. Chem. Phys.*, 96(3):1776–1783, 1992.
- [77] Charles J Geyer and Elizabeth A Thompson. Annealing markov chain monte carlo with applications to ancestral inference. *J. Am. Stat. Assoc.*, 90(431):909–920, September 1995.
- [78] Christopher J Woods, Jonathan W Essex, and Michael A King. The development of Replica-Exchange-Based Free-Energy methods. *J. Phys. Chem. B*, 107(49):13703–13710, December 2003.

- [79] John D Chodera and Michael R Shirts. Replica exchange and expanded ensemble simulations as gibbs sampling: simple improvements for enhanced mixing. *J. Chem. Phys.*, 135(19):194110, November 2011.
- [80] Sanghyun Park. Comparison of the serial and parallel algorithms of generalized ensemble simulations: An analytical approach. *Phys. Rev. E*, 77(1):016709, 2008.
- [81] Lingle Wang, Richard A Friesner, and B J Berne. Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (REST2). *J. Phys. Chem. B*, 115(30):9431–9438, August 2011.
- [82] Giovanni Bussi. Hamiltonian replica exchange in gromacs: a flexible implementation. *Mol. Phys.*, 112(3-4):379–384, 2014.
- [83] J Chodera, A Rizzi, L Naden, K Beauchamp, P Grinaway, J Fass, B Rustenburg, G A Ross, A Simmonett, and D W H Swenson. choderalab/openmmtools: 0.14. 0-exact treatment of alchemical PME electrostatics, water cluster test system, optimizations, 2018 (accessed 2020-08-24). <https://doi.org/10.5281/zenodo.1161149>.
- [84] Siewert J Marrink, H Jelger Risselada, Serge Yefimov, D Peter Tieleman, and Alex H de Vries. The MARTINI force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B*, 111(27):7812–7824, July 2007.
- [85] Jérôme Delhommelle and Philippe Millié. Inadequacy of the Lorentz-Berthelot combining rules for accurate predictions of equilibrium properties by molecular simulation. *Mol. Phys.*, 99(8):619–625, April 2001.
- [86] Ulrich Essmann, Lalith Perera, Max L Berkowitz, Tom Darden, Hsing Lee, and Lee G Pedersen. A smooth particle mesh ewald method. *J. Chem. Phys.*, 103(19):8577–8593, November 1995.
- [87] Abdulnour Y Toukmaji and Board, John A. Ewald summation techniques in perspective: a survey. *Comput. Phys. Commun.*, 95(2):73–92, June 1996.
- [88] Christopher M Baker, Victor M Anisimov, and Alexander D MacKerell, Jr. Development of CHARMM polarizable force field for nucleic acid bases based on the classical drude oscillator model. *J. Phys. Chem. B*, 115(3):580–596, January 2011.
- [89] Yue Shi, Zhen Xia, Jiajing Zhang, Robert Best, Chuanjie Wu, Jay W Ponder, and Pengyu Ren. Polarizable atomic Multipole-Based AMOEBA force field for proteins. *J. Chem. Theory Comput.*, 9(9):4046–4063, September 2013.
- [90] Richard T Bradshaw, Jacek Dziedzic, Chris-Kriton Skylaris, and Jonathan W Essex. The role of electrostatics in enzymes: do biomolecular force fields reflect protein electric fields? *J. Chem. Inf. Model.*, 2020.
- [91] Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, Paul Maragakis, John L Klepeis, Ron O Dror, and David E Shaw. Improved side-chain torsion potentials for the amber ff99SB protein force field. *Proteins: Struct. Funct. Bioinf.*, 78(8):1950–1958, 2010.

- [92] Lee-Ping Wang, Keri A McKiernan, Joseph Gomes, Kyle A Beauchamp, Teresa Head-Gordon, Julia E Rice, William C Swope, Todd J Martínez, and Vijay S Pande. Building a more predictive protein force field: A systematic and reproducible route to AMBER-FB15. *J. Phys. Chem. B*, 121(16):4023–4039, April 2017.
- [93] James A Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E Hauser, and Carlos Simmerling. ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.*, 11(8):3696–3713, August 2015.
- [94] Alex D MacKerell, Jr, Donald Bashford, Mldr Bellott, Roland Leslie Dunbrack, Jr, Jeffrey D Evanseck, Martin J Field, Stefan Fischer, Jiali Gao, H Guo, Sookhee Ha, and Others. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102(18):3586–3616, 1998.
- [95] A D MacKerell, Jr, N Banavali, and N Foloppe. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers*, 56(4):257–265, 2000.
- [96] Jing Huang and Alexander D MacKerell, Jr. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *J. Comput. Chem.*, 34(25):2135–2145, 2013.
- [97] Wilfred F van Gunsteren, S R Billeter, A A Eising, P H Hünenberger, Pkhc Krüger, A E Mark, W R P Scott, and I G Tironi. Biomolecular simulation: the GROMOS96 manual and user guide. *Vdf Hochschulverlag AG an der ETH Zürich, Zürich*, 86, 1996.
- [98] William L Jorgensen, David S Maxwell, and Julian Tirado-Rives. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, 118(45):11225–11236, 1996.
- [99] Junmei Wang, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. Development and testing of a general amber force field. *J. Comput. Chem.*, 25(9):1157–1174, July 2004.
- [100] Stefano Piana, Alexander G Donchev, Paul Robustelli, and David E Shaw. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J. Phys. Chem. B*, 119(16):5113–5123, April 2015.
- [101] Alexey V Onufriev and Saeed Izadi. Water models for biomolecular simulations. *WIREs Comput Mol Sci*, 8(2):e1347, March 2018.
- [102] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, July 1983.
- [103] HJC Berendsen, JPM Postma, WF Van Gunsteren, and a J Hermans. Intermolecular forces, 1981.
- [104] Daniel J Price and Charles L Brooks, 3rd. A modified TIP3P water potential for simulation with ewald summation. *J. Chem. Phys.*, 121(20):10096–10103, November 2004.

- [105] Pekka Mark and Lennart Nilsson. Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K. *J. Phys. Chem. A*, 105(43):9954–9960, November 2001.
- [106] J L F Abascal and C Vega. A general purpose model for the condensed phases of water: TIP4P/2005. *J. Chem. Phys.*, 123(23):234505, December 2005.
- [107] Hans W Horn, William C Swope, Jed W Pitera, Jeffry D Madura, Thomas J Dick, Greg L Hura, and Teresa Head-Gordon. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J. Chem. Phys.*, 120(20):9665–9678, May 2004.
- [108] Michael W Mahoney and William L Jorgensen. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J. Chem. Phys.*, 112(20):8910–8922, May 2000.
- [109] Robert W Zwanzig. Erratum: High-temperature equation of state by a perturbation method. i. nonpolar gases. *J. Chem. Phys.*, 22(12):2099–2099, 1954.
- [110] Herbert B Callen. *Thermodynamics and an Introduction to Thermostatistics*. John Wiley & Sons, 1985.
- [111] Michael K Gilson, James A Given, Bruce L Bush, and J Andrew McCammon. The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophys. J.*, 72(3):1047–1069, 1997.
- [112] Jessica MJ Swanson, Richard H Henchman, and J Andrew McCammon. Revisiting free energy calculations: a theoretical connection to mm/pbsa and direct calculation of the association free energy. *Biophys. J.*, 86(1):67–74, 2004.
- [113] Michael R Shirts and Vijay S Pande. Comparison of efficiency and bias of free energies computed by exponential averaging, the bennett acceptance ratio, and thermodynamic integration. *J. Chem. Phys.*, 122(14):144107, April 2005.
- [114] Michael R Shirts and John D Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.*, 129(12):124105, September 2008.
- [115] Himanshu Paliwal and Michael R Shirts. A benchmark test set for alchemical free energy transformations and its use to quantify error in common free energy methods. *J. Chem. Theory Comput.*, 7(12):4115–4134, 2011.
- [116] Antonia SJS Mey, Bryce Allen, Hannah E Bruce Macdonald, John D Chodera, Maximilian Kuhn, Julien Michel, David L Mobley, Levi N Naden, Samarjeet Prasad, Andrea Rizzi, et al. Best practices for alchemical free energy calculations. *arXiv preprint arXiv:2008.03067*, 2020.
- [117] Haiyan Liu, Alan E Mark, and Wilfred F van Gunsteren. Estimating the relative free energy of different molecular states with respect to a single reference state. *J. Phys. Chem.*, 100(22):9485–9494, January 1996.

- [118] Tiziana Z Mordasini and J Andrew McCammon. Calculations of relative hydration free energies: a comparative study using thermodynamic integration and an extrapolation method based on a single reference state. *J. Phys. Chem. B*, 104(2):360–367, 2000.
- [119] E Prabhu Raman, Kenno Vanommeslaeghe, and Alexander D MacKerell Jr. Site-specific fragment identification guided by single-step free energy perturbation calculations. *J. Chem. Theory Comput.*, 8(10):3513–3525, 2012.
- [120] Martin Stroet, Katarzyna B Koziara, Alpeshkumar K Malde, and Alan E Mark. Optimization of empirical force fields by parameter space mapping: A single-step perturbation approach. *J. Chem. Theory Comput.*, 13(12):6201–6212, 2017.
- [121] Chris Oostenbrink and Wilfred F Van Gunsteren. Single-step perturbations to calculate free energy differences from unphysical reference states: Limits on size, flexibility, and character. *J. Comput. Chem.*, 24(14):1730–1739, 2003.
- [122] Chris Oostenbrink and Wilfred F van Gunsteren. Free energies of ligand binding for structurally diverse compounds. *Proc. Natl. Acad. Sci. U. S. A.*, 102(19):6750–6754, May 2005.
- [123] E Prabhu Raman, Sirish Kaushik Lakkaraju, Rajiah Aldrin Denny, and Alexander D MacKerell Jr. Estimation of relative free energies of binding using pre-computed ensembles based on the single-step free energy perturbation and the site-identification by ligand competitive saturation approaches. *J. Comput. Chem.*, 38(15):1238–1251, 2017.
- [124] Charles H Bennett. Efficient estimation of free energy differences from monte carlo data. *J. Comput. Phys.*, 22(2):245–268, October 1976.
- [125] W L Jorgensen and C Ravimohan. Monte carlo simulation of differences in free energies of hydration. *J. Chem. Phys.*, 1985.
- [126] Zoe Cournia, Bryce Allen, and Woody Sherman. Relative binding free energy calculations in drug discovery: Recent advances and practical considerations. *J. Chem. Inf. Model.*, 57(12):2911–2937, December 2017.
- [127] Benjamin T Burlingham and Theodore S Widlanski. An intuitive look at the relationship of k_i and IC_{50} : A more general use for the dixon plot. *J. Chem. Educ.*, 80(2):214, February 2003.
- [128] Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *J. Cheminform.*, 9(1):48, September 2017.
- [129] Jorge Nocedal and Stephen Wright. *Numerical Optimization*. Springer Science & Business Media, December 2006.
- [130] Nicholas J Higham. *Accuracy and stability of numerical algorithms*. SIAM, 2002.
- [131] Dong C Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Math. Program.*, 45(1-3):503–528, 1989.

- [132] Dieter Kraft. A software package for sequential quadratic programming. *Forschungsbericht- Deutsche Forschungs- und Versuchsanstalt für Luft- und Raumfahrt*, 1988.
- [133] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013.
- [134] Thomas Blaschke, Josep Arús-Pous, Hongming Chen, Christian Margreitter, Christian Tyrchan, Ola Engkvist, Kostas Papadopoulos, and Atanas Patronov. Reinvent 2.0: An ai tool for de novo drug design. *J. Chem. Inf. Model.*, 2020.
- [135] T M Whitehead, B W J Irwin, P Hunt, M D Segall, and G J Conduit. Imputation of assay bioactivity data using deep learning. *J. Chem. Inf. Model.*, 59(3):1197–1204, March 2019.
- [136] Pavao Santak and Gareth Conduit. Predicting physical properties of alkanes with neural networks. *Fluid Phase Equilib.*, 501:112259, December 2019.
- [137] B D Conduit, N G Jones, H J Stone, and G J Conduit. Design of a nickel-base superalloy using a neural network. *Mater. Des.*, 131:358–365, October 2017.
- [138] Bekir Karlik and A Vehbi Olgac. Performance analysis of various activation functions in generalized MLP architectures of neural networks. *International Journal of Artificial Intelligence and Expert Systems*, 1(4):111–122, 2011.
- [139] Léon Bottou. Stochastic gradient descent tricks. In Grégoire Montavon, Geneviève B Orr, and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade: Second Edition*, pages 421–436. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.
- [140] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [141] Wen Zhang, Weiran Lin, Ding Zhang, Siman Wang, Jingwen Shi, and Yanqing Niu. Recent advances in the machine Learning-Based Drug-Target interaction prediction. *Curr. Drug Metab.*, 20(3):194–202, 2019.
- [142] Chuipu Cai, Pengfei Guo, Yadi Zhou, Jingwei Zhou, Qi Wang, Fengxue Zhang, Jiansong Fang, and Feixiong Cheng. Deep Learning-Based prediction of Drug-Induced cardiotoxicity. *J. Chem. Inf. Model.*, 59(3):1073–1084, March 2019.
- [143] Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *Int. J. Uncertainty Fuzziness Knowledge Based Syst.*, 06(02):107–116, April 1998.
- [144] Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.
- [145] Marwin H S Segler, Thierry Kogej, Christian Tyrchan, and Mark P Waller. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent Sci*, 4(1):120–131, January 2018.

- [146] Josep Arús-Pous, Atanas Patronov, Esben Jannik Bjerrum, Christian Tyrchan, Jean-Louis Reymond, Hongming Chen, and Ola Engkvist. SMILES-based deep generative scaffold decorator for de-novo drug design. *J. Cheminform.*, 12(1):38, May 2020.
- [147] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.
- [148] Richard S Sutton and Andrew G Barto. *Reinforcement Learning, second edition: An Introduction*. MIT Press, November 2018.
- [149] Thomas Blaschke, Marcus Olivecrona, Ola Engkvist, Jürgen Bajorath, and Hongming Chen. Application of generative autoencoder in de novo molecular design. *Mol. Inform.*, 37(1-2), January 2018.
- [150] G E Hinton and R R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, July 2006.
- [151] Ikumi Hyohdoh, Noriyuki Furuichi, Toshihiro Aoki, Yoshiko Itezono, Haruyoshi Shirai, Sawako Ozawa, Fumio Watanabe, Masayuki Matsushita, Masahiro Sakaitani, Pil-Su Ho, et al. Fluorine scanning by nonselective fluorination: enhancing raf/mek inhibition while keeping physicochemical properties. *ACS Med. Chem. Lett.*, 4(11):1059–1063, 2013.
- [152] Yan Lou, Zachary K Sweeney, Andreas Kuglstatter, Dana Davis, David M Goldstein, Xiaochun Han, Junbae Hong, Buelent Kocer, Rama K Kondru, Renee Litman, et al. Finding the perfect spot for fluorine: Improving potency up to 40-fold during a rational fluorine scan of a bruton’s tyrosine kinase (btk) inhibitor scaffold. *Bioorg. Med. Chem. Lett.*, 25(2):367–371, 2015.
- [153] Martin Morgenthaler, Johannes D Aepli, Fiona Grüniger, Daniel Monn, Björn Wagner, Manfred Kansy, and François Diederich. A fluorine scan of non-peptidic inhibitors of neprilysin: Fluorophobic and fluorophilic regions in an enzyme active site. *J. Fluor. Chem.*, 129(9):852–865, September 2008.
- [154] Timothée Naret, Jérôme Bignon, Guillaume Bernadat, Mohamed Bencheikroun, Helene Levaique, Christine Lenoir, Joelle Dubois, Alain Pruvost, François Saller, Delphine Borgel, Boris Manoury, Veronique Leblais, Romain Darrigrand, Sébastien Apcher, Jean-Daniel Brion, Etienne Schmitt, Frédéric R Leroux, Mouad Alami, and Abdallah Hamze. A fluorine scan of a tubulin polymerization inhibitor isocombretastatin a-4: Design, synthesis, molecular modelling, and biological evaluation. *Eur. J. Med. Chem.*, 143:473–490, January 2018.
- [155] Jacob A Olsen, David W Banner, Paul Seiler, Ulrike Obst Sander, Allan D’Arcy, Martine Stihle, Klaus Müller, and François Diederich. A fluorine scan of thrombin inhibitors to map the fluorophilicity/fluorophobicity of an enzyme active site: Evidence for c f c o interactions. *Angewandte Chemie International Edition*, 42(22):2507–2511, 2003.

- [156] Eliane Schweizer, Anja Hoffmann-Röder, Kaspar Schärer, Jacob A Olsen, Christoph Fäh, Paul Seiler, Ulrike Obst-Sander, Björn Wagner, Manfred Kansy, and François Diederich. A fluorine scan at the catalytic center of thrombin: C-f, c-h, and c-ome bioisosterism and fluorine effects on pka and log d values. *ChemMedChem: Chemistry Enabling Drug Discovery*, 1(6):611–621, 2006.
- [157] Eric P Gillis, Kyle J Eastman, Matthew D Hill, David J Donnelly, and Nicholas A Meanwell. Applications of fluorine in medicinal chemistry. *J. Med. Chem.*, 58(21):8315–8359, 2015.
- [158] Nicholas A Meanwell. Fluorine and fluorinated motifs in the design and application of bioisosteres for drug design. *J. Med. Chem.*, February 2018.
- [159] Yu-Kai Lee, Daniel J Parks, Tianbao Lu, Tho V Thieu, Thomas Markotan, Wenxi Pan, David F McComsey, Karen L Milkiewicz, Carl S Crysler, Nisha Ninan, et al. 7-fluoroindazoles as potent and selective inhibitors of factor xa. *J. Med. Chem.*, 51(2):282–297, 2008.
- [160] Bernd Kuhn and Peter A Kollman. A ligand that is predicted to bind better to avidin than biotin: Insights from computational fluorine scanning. *J. Am. Chem. Soc.*, 122(16):3909–3916, April 2000.
- [161] Jin Zhang, Haiyang Zhang, Tao Wu, Qi Wang, and David van der Spoel. Comparison of implicit and explicit solvent models for the calculation of solvation free energy in organic solvents. *J. Chem. Theory Comput.*, 13(3):1034–1043, March 2017.
- [162] Robert C Harris and B Montgomery Pettitt. Examining the assumptions underlying continuum-solvent models. *J. Chem. Theory Comput.*, 11(10):4593–4600, October 2015.
- [163] Haiyang Zhang, Tianwei Tan, and David van der Spoel. Generalized born and explicit solvent models for free energy calculations in organic solvents: Cyclodextrin dimerization. *J. Chem. Theory Comput.*, 11(11):5103–5113, November 2015.
- [164] Haiyang Zhang, Chunhua Yin, Hai Yan, and David van der Spoel. Evaluation of generalized born models for large scale affinity prediction of cyclodextrin Host–Guest complexes. *J. Chem. Inf. Model.*, 56(10):2080–2092, October 2016.
- [165] Peter Eastman, Mark S Friedrichs, John D Chodera, Randall J Radmer, Christopher M Bruns, Joy P Ku, Kyle A Beauchamp, Thomas J Lane, Lee-Ping Wang, Diwakar Shukla, Tony Tye, Mike Houston, Timo Stich, Christoph Klein, Michael R Shirts, and Vijay S Pande. OpenMM 4: A reusable, extensible, hardware independent library for high performance molecular simulation. *J. Chem. Theory Comput.*, 9(1):461–469, January 2013.
- [166] Hans Matter, Bodo Scheiper, Henning Steinhagen, Zsolt Böcskei, Valérie Fleury, and Gary McCort. Structure-based design and optimization of potent renin inhibitors on 5- or 7-azaindole-scaffolds. *Bioorg. Med. Chem. Lett.*, 21(18):5487–5492, September 2011.

- [167] Jens-Uwe Peters, Silja Weber, Stéphane Kritter, Peter Weiss, Angelina Wallier, Markus Boehringer, Michael Hennig, Bernd Kuhn, and Bernd-Michael Loeffler. Aminomethylpyrimidines as novel DPP-IV inhibitors: a 10(5)-fold activity increase by optimization of aromatic substituents. *Bioorg. Med. Chem. Lett.*, 14(6):1491–1493, March 2004.
- [168] Shihan He, Timothy J Senter, Jonathan Pollock, Changho Han, Sunil Kumar Upadhyay, Trupta Purohit, Rocco D Gogliotti, Craig W Lindsley, Tomasz Cierpicki, Shaun R Stauffer, and Jolanta Grembecka. High-Affinity Small-Molecule inhibitors of the Menin-Mixed lineage leukemia (MLL) interaction closely mimic a natural Protein–Protein interaction. *J. Med. Chem.*, 57(4):1543–1556, February 2014.
- [169] Josep Aiguadé, Cristina Balagué, Inés Carranco, Francisco Caturla, María Domínguez, Paul Eastwood, Cristina Esteve, Jacob González, Wenceslao Lumeras, Adelina Orellana, Sara Preciado, Ramón Roca, Laura Vidal, and Bernat Vidal. Novel triazolopyridylbenzamides as potent and selective p38 α inhibitors. *Bioorg. Med. Chem. Lett.*, 22(10):3431–3436, May 2012.
- [170] Clifford D Jones, David M Andrews, Andrew J Barker, Kevin Blades, Paula Daunt, Simon East, Catherine Geh, Mark A Graham, Keith M Johnson, Sarah A Loddick, Heather M McFarland, Alexandra McGregor, Louise Moss, David A Rudge, Peter B Simpson, Michael L Swain, Kin Y Tam, Julie A Tucker, and Mike Walker. The discovery of AZD5597, a potent imidazole pyrimidine amide CDK inhibitor suitable for intravenous dosing. *Bioorg. Med. Chem. Lett.*, 18(24):6369–6373, December 2008.
- [171] Kevin D Freeman-Cook, Christopher Autry, Gary Borzillo, Deborah Gordon, Elsa Barbacci-Tobin, Vincent Bernardo, David Briere, Tracey Clark, Matthew Corbett, John Jakubczak, Shefali Kakar, Elizabeth Knauth, Blaise Lippa, Michael J Luzzio, Mahmoud Mansour, Gary Martinelli, Matthew Marx, Kendra Nelson, Jayvardhan Pandit, Francis Rajamohan, Shaughnessy Robinson, Chakrapani Subramanyam, Liuqing Wei, Martin Wythes, and Joel Morris. Design of selective, ATP-competitive inhibitors of akt. *J. Med. Chem.*, 53(12):4615–4622, June 2010.
- [172] Mark W Ledeboer, Albert C Pierce, John P Duffy, Huai Gao, David Messersmith, Francesco G Salituro, Suganthini Nanthakumar, Jon Come, Harmon J Zuccola, Lora Swenson, Dina Shlyakter, Sudipta Mahajan, Thomas Hooock, Bin Fan, Wan-Jung Tsai, Elaine Kolaczowski, Scott Carrier, James K Hogan, Richard Zessis, S Pazhanisamy, and Youssef L Bennani. 2-aminopyrazolo[1,5-a]pyrimidines as potent and selective inhibitors of JAK2. *Bioorg. Med. Chem. Lett.*, 19(23):6529–6533, December 2009.
- [173] Lawrence G Hamann, Mark C Manfredi, Chongqing Sun, Stanley R Krystek Jr, Yanting Huang, Yingzhi Bi, David J Augeri, Tammy Wang, Yan Zou, David A Betebenner, et al. Tandem optimization of target activity and elimination of mutagenic potential in a potent series of n-aryl bicyclic hydantoin-based selective androgen receptor modulators. *Bioorg. Med. Chem. Lett.*, 17(7):1860–1864, 2007.
- [174] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

- [175] Matthew R Weir. Effects of renin-angiotensin system inhibition end-organ protection: can we do better? *Clinical therapeutics*, 29(9):1803–1824, 2007.
- [176] Ian R Hardcastle, Christine E Arris, Johanne Bentley, F Thomas Boyle, Yuzhu Chen, Nicola J Curtin, Jane A Endicott, Ashleigh E Gibson, Bernard T Golding, Roger J Griffin, et al. N2-substituted o 6-cyclohexylmethylguanine derivatives: potent inhibitors of cyclin-dependent kinases 1 and 2. *J. Med. Chem*, 47(15):3710–3722, 2004.
- [177] P Eastman. *Pdbfixer*, 2015 (accessed 2020-08-24). <https://github.com/pandegroup/pdbfixer>.
- [178] G Madhavi Sastry, Matvey Adzhigirey, Tyler Day, Ramakrishna Annabhimoju, and Woody Sherman. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided Mol. Des.*, 27(3):221–234, March 2013.
- [179] Junmei Wang, Wei Wang, Peter A Kollman, and David A Case. Antechamber: an accessory software package for molecular mechanical calculations. *J. Am. Chem. Soc.*, 222:U403, 2001.
- [180] Araz Jakalian, David B Jack, and Christopher I Bayly. Fast, efficient generation of high-quality atomic charges. am1-bcc model: Ii. parameterization and validation. *J. Comput. Chem*, 23(16):1623–1641, 2002.
- [181] A Rizzi, P B Grinaway, D L Parton, M R Shirts, K Wang, P Eastman, M Friedrichs, V S Pande, K Branson, and D L Mobley. YANK: A GPU-accelerated platform for alchemical free energy calculations [internet], 2018 (accessed Aug 23, 2018). <http://getyank.org>.
- [182] Frank H Allen, Olga Kennard, David G Watson, Lee Brammer, A Guy Orpen, and Robin Taylor. Tables of bond lengths determined by x-ray and neutron diffraction. part 1. bond lengths in organic compounds. *J. Chem. Soc. Perkin Trans. 2*, (12):S1–S19, 1987.
- [183] Poonam Shah and Andrew D Westwell. The role of fluorine in medicinal chemistry. *J. Enzyme Inhib. Med. Chem.*, 22(5):527–540, October 2007.
- [184] Lee-Peng Lee and Bruce Tidor. Optimization of electrostatic binding free energy. *J. Chem. Phys.*, 106(21):8681–8690, June 1997.
- [185] E Kangas and B Tidor. Charge optimization leads to favorable electrostatic binding free energy. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics*, 59(5 Pt B):5958–5961, May 1999.
- [186] Erik Kangas and Bruce Tidor. Electrostatic specificity in molecular ligand design. *J. Chem. Phys.*, 112(20):9120–9131, May 2000.
- [187] L P Lee and B Tidor. Optimization of binding electrostatics: Charge complementarity in the barnase-barstar protein complex. *Protein Sci.*, 2001.

- [188] J P Bardhan, J H Lee, S S Kuo, M D Altman, B Tidor, and J K White. Fast methods for biomolecule charge optimization. *Modeling and Simulation of Microsystems (Nanotech)*, 2003.
- [189] Erik Kangas and Bruce Tidor. Electrostatic complementarity at ligand binding sites: Application to chorismate mutase. *J. Phys. Chem. B*, 105(4):880–888, February 2001.
- [190] Peter A Sims, Chung F Wong, and J Andrew McCammon. Charge optimization of the interface between protein kinases and their ligands. *J. Comput. Chem.*, 25(11):1416–1429, August 2004.
- [191] Traian Sulea and Enrico O Purisima. Optimizing ligand charges for maximum binding affinity. a solvated interaction energy approach. *J. Phys. Chem. B*, 105(4):889–899, February 2001.
- [192] Kathryn A Armstrong, Bruce Tidor, and Alan C Cheng. Optimal charges in lead progression: a structure-based neuraminidase case study. *J. Med. Chem.*, 49(8):2470–2477, April 2006.
- [193] Yang Shen, Michael K Gilson, and Bruce Tidor. Charge optimization theory for Induced-Fit ligands. *J. Chem. Theory Comput.*, 8(11):4580–4592, November 2012.
- [194] Travis E Oliphant. SciPy: Open source scientific tools for python. *Comp.Sci. Eng.*, 9(1):10–20, 2007.
- [195] Maria M Reif and Philippe H Hünenberger. Computation of methodology-independent single-ion solvation properties from molecular simulations. III. correction terms for the solvation free energies, enthalpies, entropies, heat capacities, volumes, compressibilities, and expansivities of solvated ion. *J. Chem. Phys.*, 134(14):144103, April 2011.
- [196] Yoshihiko Nishibata and Akiko Itai. Automatic creation of drug candidate structures based on receptor structure. starting point for artificial lead generation. *Tetrahedron*, 47(43):8985–8990, November 1991.
- [197] V J Gillet, W Newell, P Mata, G Myatt, S Sike, Z Zsoldos, and A P Johnson. SPROUT: recent developments in the de novo design of molecules. *J. Chem. Inf. Comput. Sci.*, 34(1):207–217, January 1994.
- [198] Lawrence G Hamann, Mark C Manfredi, Chongqing Sun, Stanley R Krystek Jr, Yanting Huang, Yingzhi Bi, David J Augeri, Tammy Wang, Yan Zou, David A Betebenner, et al. Tandem optimization of target activity and elimination of mutagenic potential in a potent series of n-aryl bicyclic hydantoin-based selective androgen receptor modulators. *Bioorg. Med. Chem. Lett.*, 17(7):1860–1864, 2007.
- [199] Jacek Ostrowski, Joyce E Kuhns, John A Lupisella, Mark C Manfredi, Blake C Beehler, Stanley R Krystek, Jr, Yingzhi Bi, Chongqing Sun, Ramakrishna Seethala, Rajasree Golla, Paul G Sleph, Abera Fura, Yongmi An, Kevin F Kish, John S Sack, Kasim A Mookhtiar, Gary J Grover, and Lawrence G Hamann. Pharmacological and x-ray structural characterization of a novel selective androgen receptor modulator: potent hyperanabolic stimulation of skeletal muscle with hypostimulation of prostate in rats. *Endocrinology*, 148(1):4–12, January 2007.

- [200] Bernhard Baum, Laveena Muley, Michael Smolinski, Andreas Heine, David Hangauer, and Gerhard Klebe. Non-additivity of functional group contributions in Protein–Ligand binding: A comprehensive study by crystallography and isothermal titration calorimetry. *J. Mol. Biol.*, 397(4):1042–1054, April 2010.
- [201] Arun K Ghosh, Jun Takayama, Yoann Aubin, Kiira Ratia, Rima Chaudhuri, Yahira Baez, Katrina Sleeman, Melissa Coughlin, Daniel B Nichols, Debbie C Mulhearn, and Others. Structure-Based design, synthesis, and biological evaluation of a series of novel and reversible inhibitors for the severe acute respiratory syndrome- coronavirus Papain-Like protease. *J. Med. Chem.*, 52(16):5228–5240, 2009.
- [202] Kiira Ratia, Scott Pegan, Jun Takayama, Katrina Sleeman, Melissa Coughlin, Surendranath Baliji, Rima Chaudhuri, Wentao Fu, Bellur S Prabhakar, Michael E Johnson, Susan C Baker, Arun K Ghosh, and Andrew D Mesecar. A noncovalent class of papain-like protease/deubiquitinase inhibitors blocks SARS virus replication. *Proc. Natl. Acad. Sci. U. S. A.*, 105(42):16119–16124, October 2008.
- [203] Laveena Muley, Bernhard Baum, Michael Smolinski, Marek Freindorf, Andreas Heine, Gerhard Klebe, and David G Hangauer. Enhancement of hydrophobic interactions and hydrogen bond strength by cooperativity: synthesis, modeling, and molecular dynamics simulations of a congeneric series of thrombin inhibitors. *J. Med. Chem.*, 53(5):2126–2135, 2010.
- [204] Matthew H Todd. *OSM Website Landing Page*, 2020 (accessed 2020-08-24). <http://opencoursemalaria.org/>.
- [205] Melanie Ridgeway Kieran Kirk, Adele Lehane. *Evaluation of Series 4 Compounds vs ATP4-Resistant Mutants*, 2015 (accessed 2020-08-24). http://malaria.ourexperiment.org/biological_data/11448/post.html.
- [206] Kieran Kirk. Ion regulation in the malaria parasite. *Annu. Rev. Microbiol.*, 69:341–359, 2015.
- [207] Edwin G Tse, Marat Korsik, and Matthew H Todd. The past, present and future of anti-malarial medicines. *Malar. J.*, 18(1):93, March 2019.
- [208] Sheng-Yong Yang. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discov.*, 15(11-12):444–450, 2010.
- [209] Markus A Lill. Multi-dimensional qsar in drug discovery. *Drug Discov.*, 12(23-24):1013–1017, 2007.
- [210] Benedict W J Irwin, Julian R Levell, Thomas M Whitehead, Matthew D Segall, and Gareth J Conduit. Practical applications of deep learning to impute heterogeneous drug discovery data. *J. Chem. Inf. Model.*, June 2020.
- [211] Benedict W J Irwin, Samar Mahmoud, Thomas M Whitehead, Gareth J Conduit, and Matthew D Segall. Imputation versus prediction: applications in machine learning for drug discovery. *Future Drug Discov.*, 2(2):FDD38, April 2020.
- [212] P C Verpoort, P MacDonald, and G J Conduit. Materials data validation and imputation with an artificial neural network. *Comput. Mater. Sci.*, 147:176–185, May 2018.

-
- [213] G Papadopoulos, P J Edwards, and A F Murray. Confidence estimation methods for neural networks: a practical comparison. *IEEE Trans. Neural Netw.*, 12(6):1278–1287, 2001.
- [214] Matthew Segall, Ed Champness, Chris Leeding, Ryan Lilien, Ramgopal Mettu, and Brian Stevens. Applying medicinal chemistry transformations and multiparameter optimization to guide the search for high-quality leads and candidates. *J. Chem. Inf. Model.*, 51(11):2967–2976, November 2011.
- [215] Edwin Tse and Matthew H Todd. *OSM Sources of Data*, 2020 (accessed 2020-08-24). <https://github.com/OpenSourceMalaria/Series4/wiki/Sources-of-Data>.
- [216] Naruki Yoshikawa, Kei Terayama, Masato Sumita, Teruki Homma, Kenta Oono, and Koji Tsuda. Population-based de novo molecule generation, using grammatical evolution. *Chem. Lett.*, 47(11):1431–1434, November 2018.
- [217] Silke D Meier, Yury Kovalchuk, and Christine R Rose. Properties of the new fluorescent na⁺ indicator corona green: comparison with sbfi and confocal na⁺ imaging. *J. Neurosci. Methods*, 155(2):251–259, 2006.

Appendix A

Computational Fluorine Scanning

Table A.1 presents the $\Delta\Delta G$ for all possible single aromatic hydrogen to fluorine mutations on all nine test system presented in chapter 2 calculated with perturbative fluorine scanning (PFS)

Table A.1 Peterbative fluorine scanning (PFS) $\Delta\Delta G$ s for all aromatic hydrogen to fluorine substitutions for all test cases, in kcal/mol. See figure 2.2 for hydrogen labels. R1-3 are $\Delta\Delta G$ collected from repeat calculations, AVG is mean of all repeats.

System	R1 [kcal/mol]	R2 [kcal/mol]	R3 [kcal/mol]	AVG [kcal/mol]
Renin[166]				
H22	-1.69	-2.11	-1.69	-1.83
H25	-1.32	-1.58	-1.76	-1.56
H15	-0.49	-0.78	-0.4	-0.55
H1	-1.05	-0.05	-0.47	-0.52
H16	-0.25	-0.56	-0.26	-0.35
H14	-0.38	0.1	-0.17	-0.15
H17	0.08	-0.2	-0.27	-0.13
H23	-0.47	0.98	-0.78	-0.09
H2	-0.54	0.65	-0.04	0.02
H18	0.22	0.38	0.03	0.21
H5	0.2	0	1.35	0.52
H24	0.56	0.51	0.53	0.53
DPP4[167]				
H13	-1.42	-0.08	-2	-1.17
H16	-0.06	-0.33	-1.02	-0.47
H7	-0.53	0.01	-0.38	-0.3

H5	-0.27	-0.08	-0.33	-0.23
H4	-0.07	-0.09	-0.36	-0.17
H6	-0.23	-0.16	-0.07	-0.15
H8	0.17	-0.77	0.22	-0.12
H12	-0.04	-0.09	0.23	0.03
Menin[168]				
HAY	-1.32	-1.72	-1.47	-1.5
HAI	-1.37	-1.1	-1.33	-1.27
HAL	-0.74	-0.94	-0.91	-0.86
HAH	0.13	-0.38	0.02	-0.08
HAE	-0.2	0.24	-0.12	-0.03
HAF	0.41	0.32	0.23	0.32
HAG	0.35	0.52	0.46	0.44
HAI	0.36	0.63	0.61	0.54
HAK	0.99	1.12	1.14	1.08
P38[169]				
H1	-2.4	-2.32	-1.96	-2.23
H19	-2	-1.79	-1.85	-1.88
H16	-0.62	-0.74	-0.48	-0.61
H17	-0.47	-0.24	-0.7	-0.47
H18	0.3	0.13	0.11	0.18
H22	0.41	0.51	0.65	0.52
H20	1.3	0.95	1.24	1.16
H6	1.04	1.03	1.68	1.25
H4	1.79	2.8	2.66	2.42
FXa[159]				
H18	-2.4	-2.21	-2.22	-2.27
H29	-1.31	-1.61	-1.28	-1.4
H9	-0.74	-0.89	-0.99	-0.88
H19	-0.24	0.12	-0.49	-0.21
H3	-0.07	-0.02	-0.03	-0.04
H7	0.21	0.05	-0.15	0.03
H6	0.05	0.06	0.06	0.06
H28	0.54	0.41	0.56	0.51
H4	0.84	1.19	0.69	0.91
H8	0.99	1.01	0.93	0.98

H16	2.57	2.65	2.64	2.62
CDK2[170]				
H33	-1.15	-1	-0.8	-0.99
H14	-0.24	-0.71	0	-0.32
H18	-0.52	-0.1	-0.02	-0.21
H15	0.67	0.4	1.66	0.91
H11	2.51	2.66	1.16	2.11
H32	2.98	1.82	1.68	2.16
AKT[171]				
H22	-1.99	-2.32	-2.24	-2.18
H25	-0.89	-1.45	-1.48	-1.27
H26	-0.92	-1.69	-1.13	-1.25
H23	-0.81	-0.5	-0.71	-0.67
H21	-0.5	-0.75	-0.76	-0.67
H8	0.03	-0.15	-0.19	-0.1
H4	-0.05	0.06	0.55	0.19
JAK2[172]				
H24	-1.99	-2.17	-1.83	-2
H27	-1.77	-1.39	-1.1	-1.42
H14	-1.16	-0.95	-0.9	-1
H26	-1.19	-1.04	-0.64	-0.96
H23	-0.79	-1.01	-0.92	-0.91
H15	-0.32	-0.48	-0.97	-0.59
H7	-0.31	0.19	0.12	0
H22	0.61	0.73	0.79	0.71
H16	2.49	2.51	2.39	2.46
H10	3.62	4.26	4.21	4.03
Androgen Recp.[173]				
H2	-2.84	-2.66	-1.93	-2.47
H7	-0.3	-0.25	-0.36	-0.3
H5	3.52	3.68	3.23	3.48

Complex RMSDs

To assert that the complexes of the test cases are maintained during the simulations in chapter 2 a plot of root mean squared difference is made for the protein in each system figures

A.1-A.9 show that the RMSD for all complex systems is well converged within 2500 frames (12.5ns).

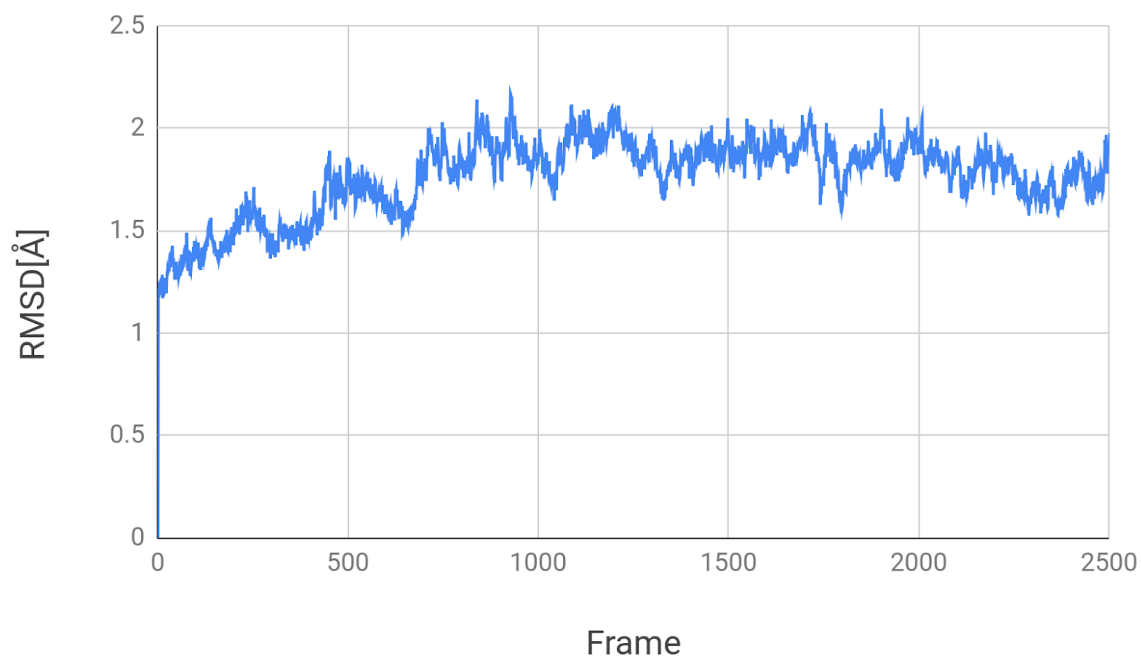


Fig. A.1 Plot of root mean squared difference for renin complex across 2500 frames.

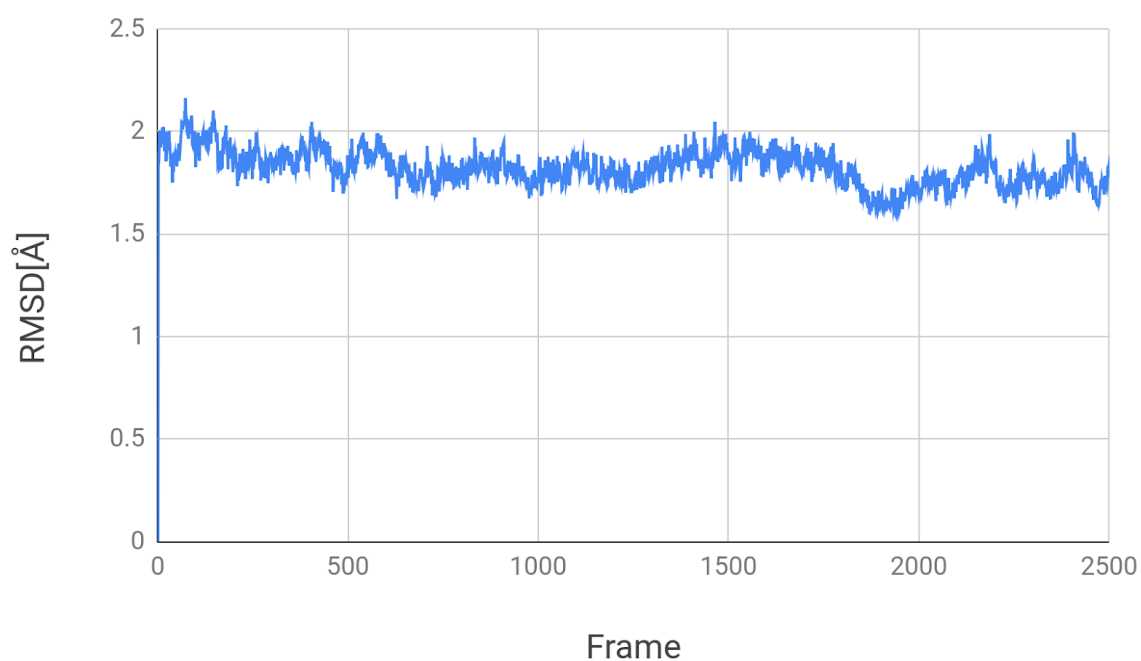


Fig. A.2 Plot of root mean squared difference for DPP4 complex across 2500 frames.

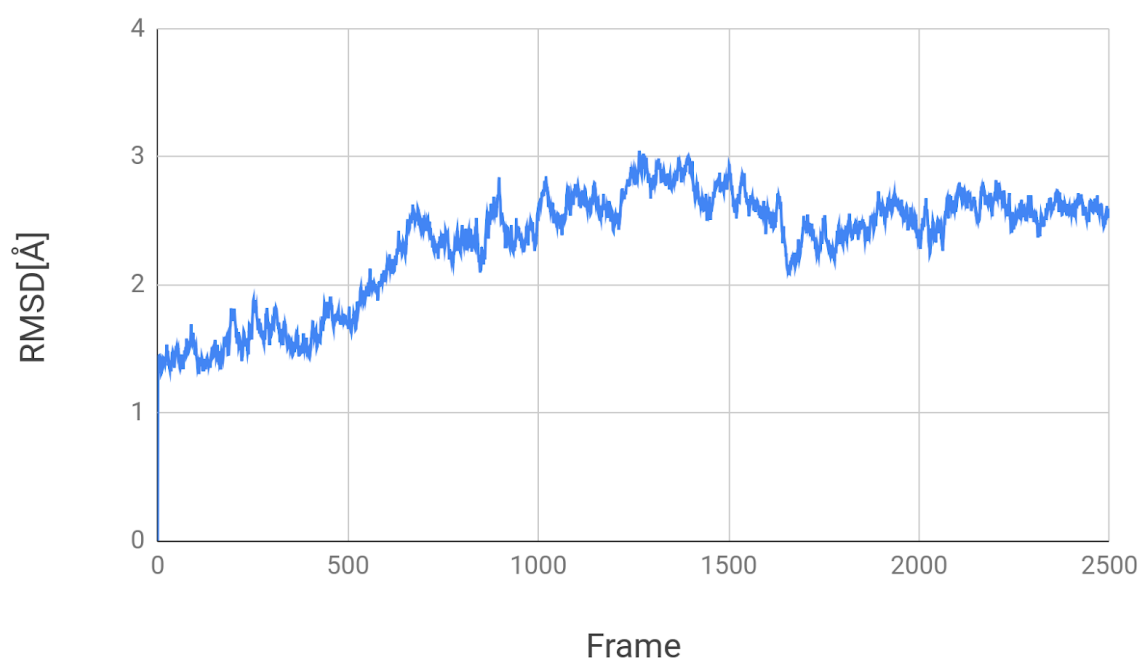


Fig. A.3 Plot of root mean squared difference for menin complex across 2500 frames.

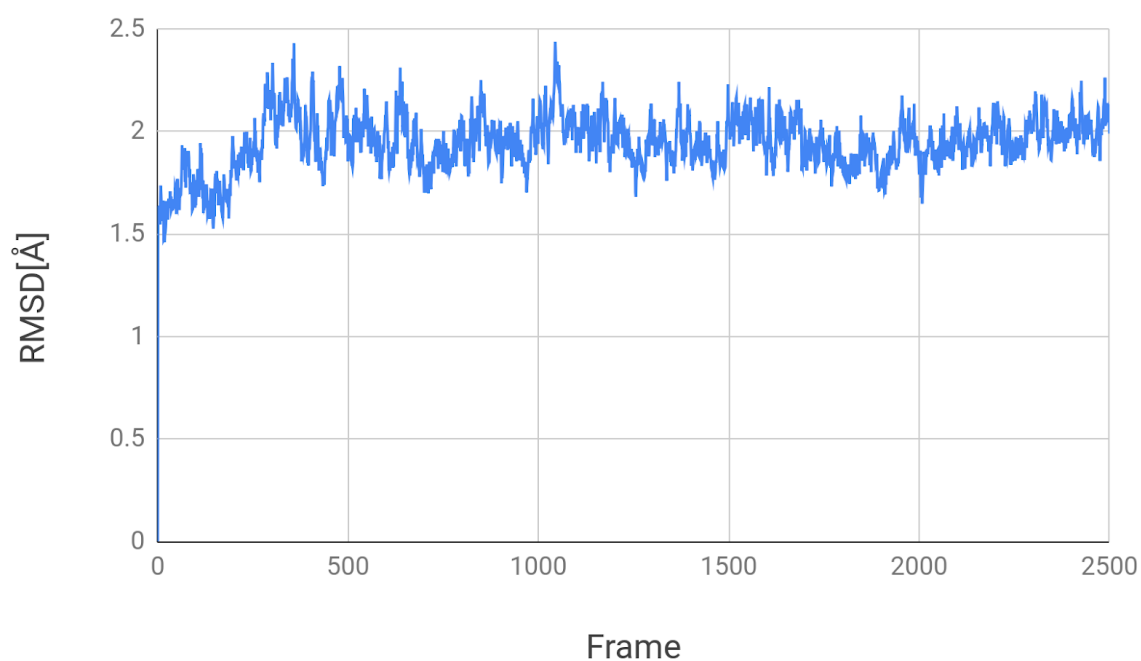


Fig. A.4 Plot of root mean squared difference for P38 complex across 2500 frames.

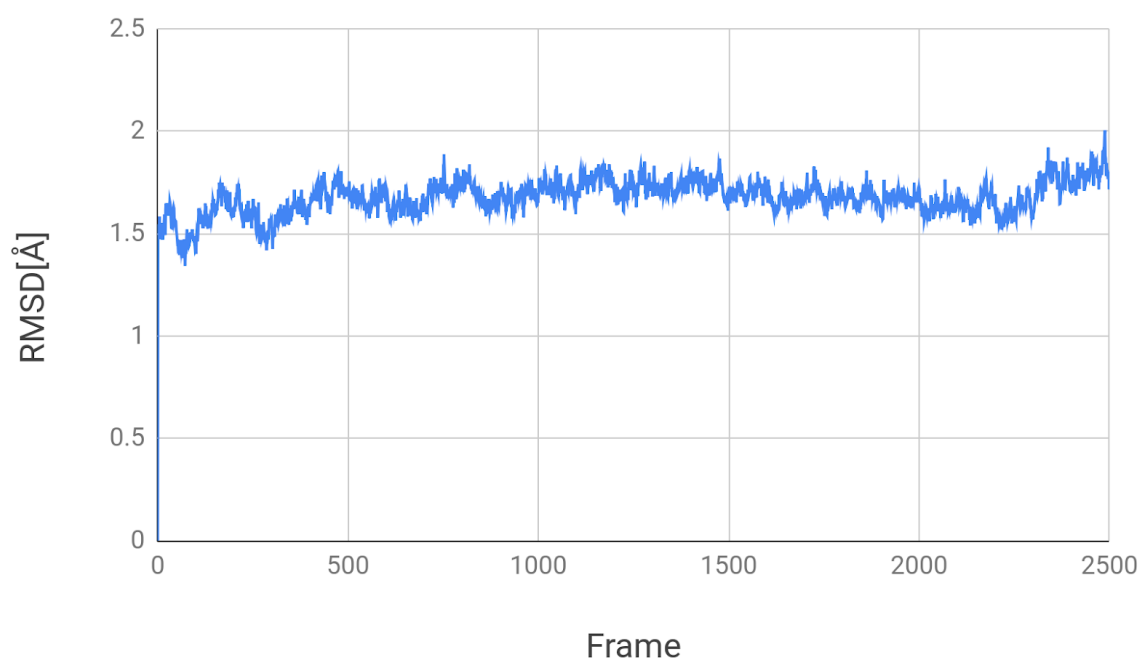


Fig. A.5 Plot of root mean squared difference for FXa complex across 2500 frames.

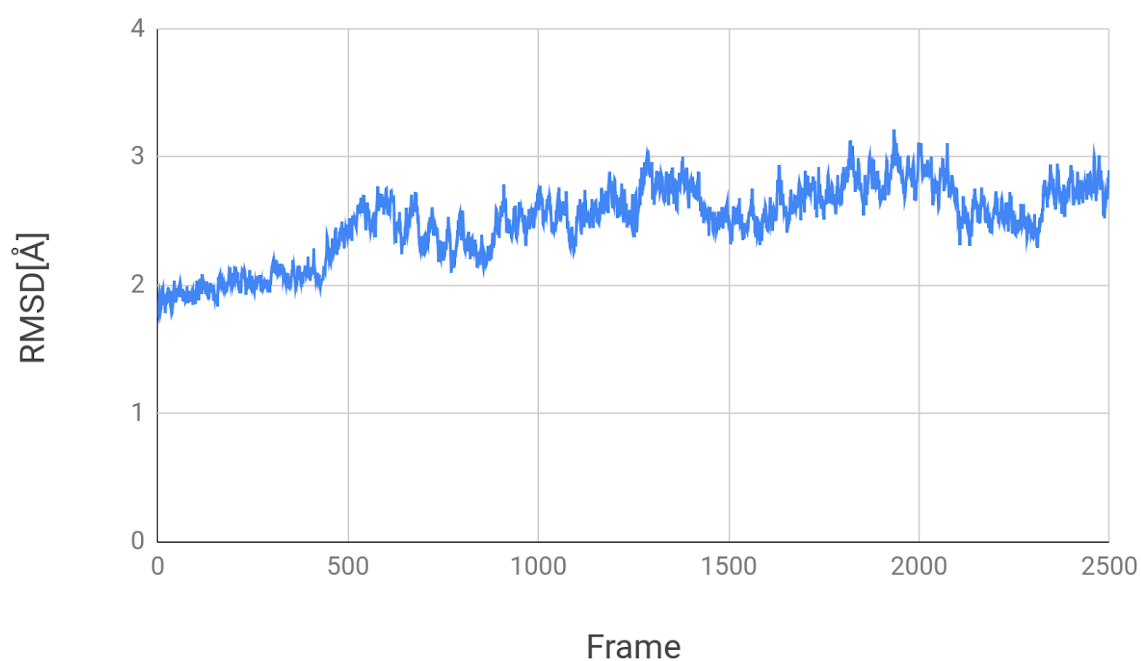


Fig. A.6 Plot of root mean squared difference for CDK2 complex across 2500 frames.

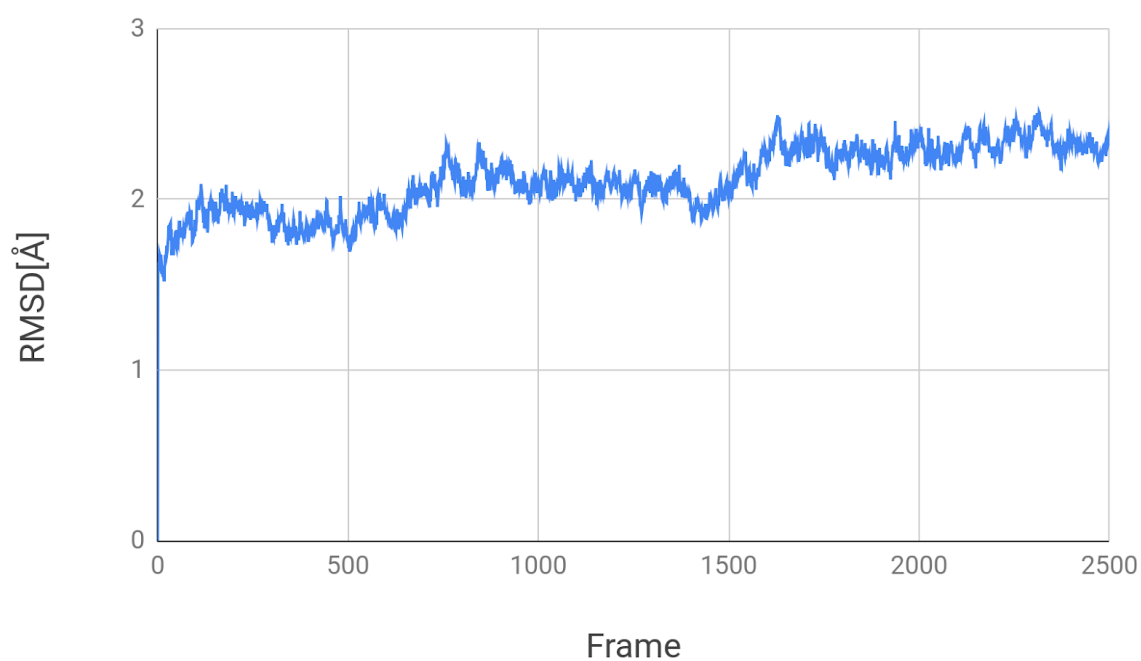


Fig. A.7 Plot of root mean squared difference for AKT complex across 2500 frames.

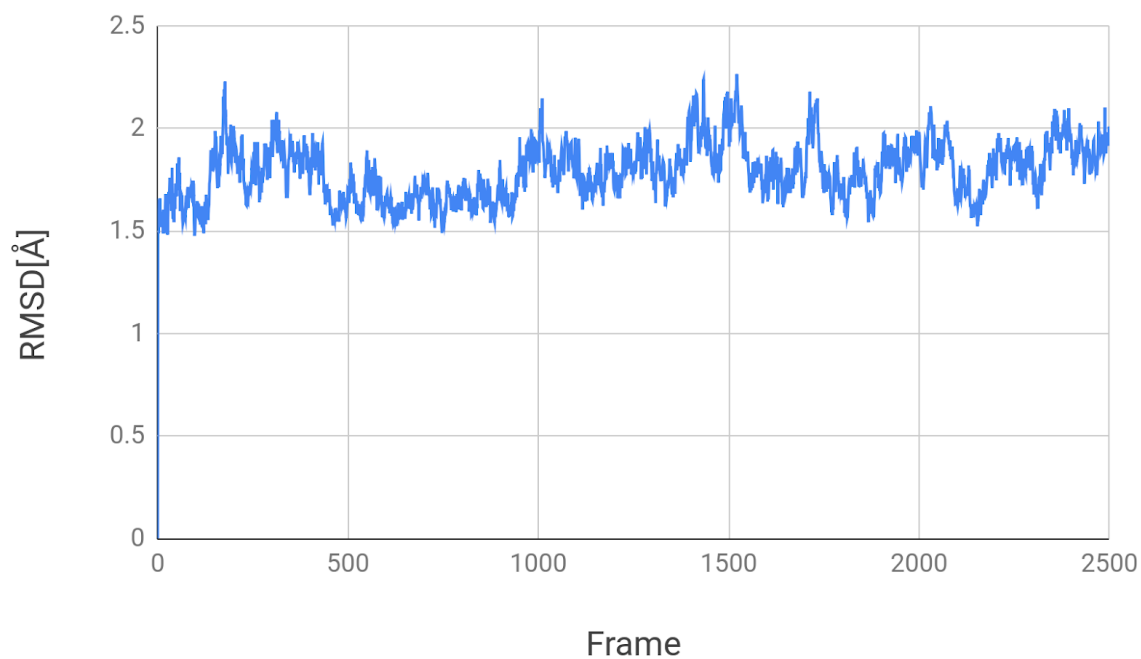


Fig. A.8 Plot of root mean squared difference for JAK complex across 2500 frames.

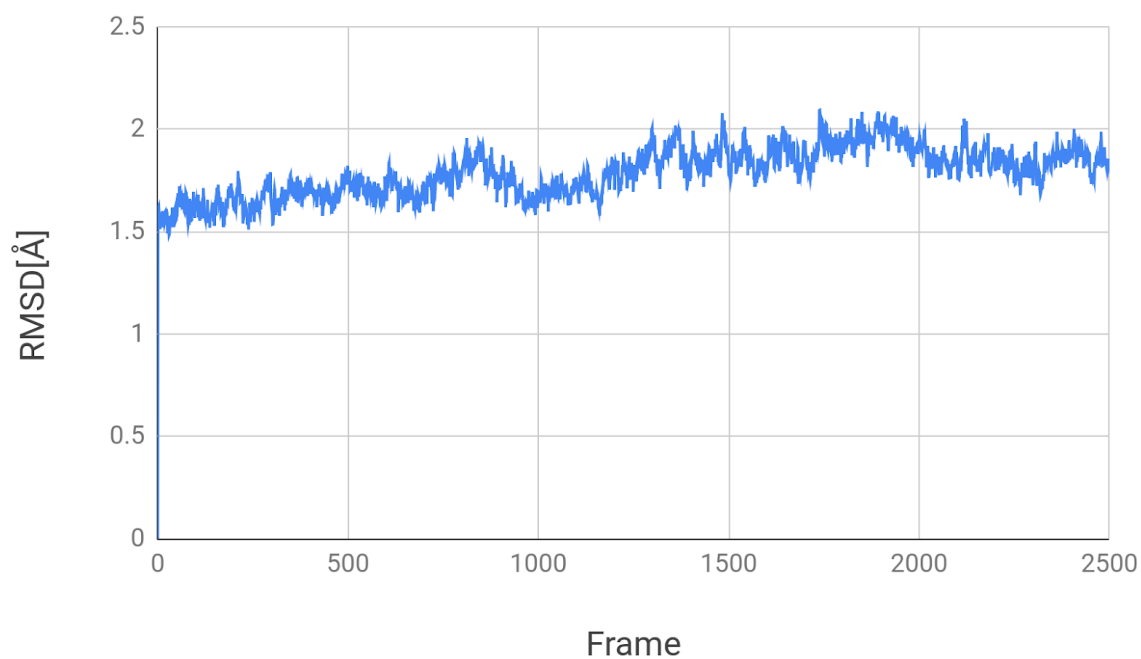


Fig. A.9 Plot of root mean squared difference for androgen receptor complex across 2500 frames.

Appendix B

Charge optimization

To visually assert that the charges optimized in chapter 3 are being placed in the same places independent of the RMSD limit chosen plots are made for all optimized charges. Figures B.1-B.9 show all inhibitors colored by change in charge across for all RMSD limits. Figures B.1-B.9 show that the information about which atoms could be beneficially changed to be more positive or negative is largely invariant as the RMSD limit is changed.

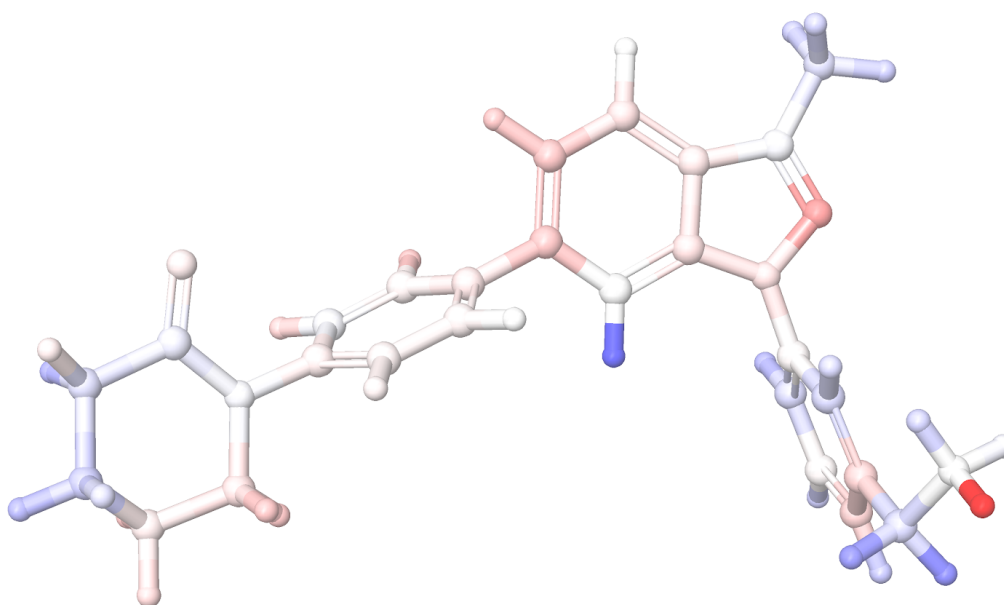


Fig. B.1 Fxa ligand with atoms colored by charge such that more negative atoms are blue and more positive atoms are red. Optimal set of charges computed with and RMSD limit of $0.05 q_e$.

In chapter 3 the cumulative sum of the $\Delta\Delta G_{opt}$, calculated with SSP theory, was plotted for all optimization steps, see figure 3.5. We term the total cumulative sum $\Delta\Delta G_{opt}$ values

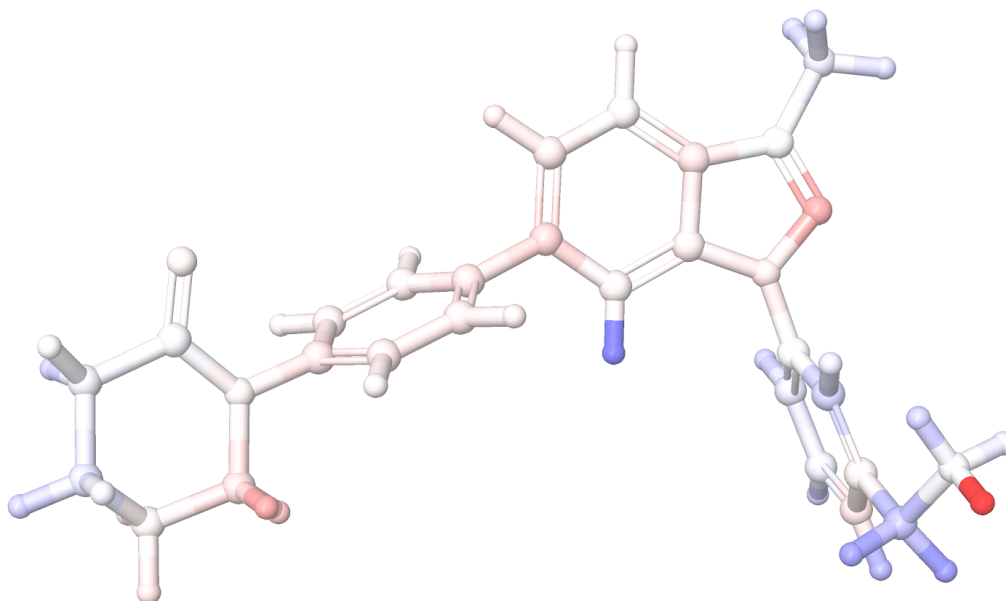


Fig. B.2 Fxa ligand with atoms colored by charge such that more negative atoms are blue and more positive atoms are red. Optimal set of charges computed with and RMSD limit of $0.03 q_e$.

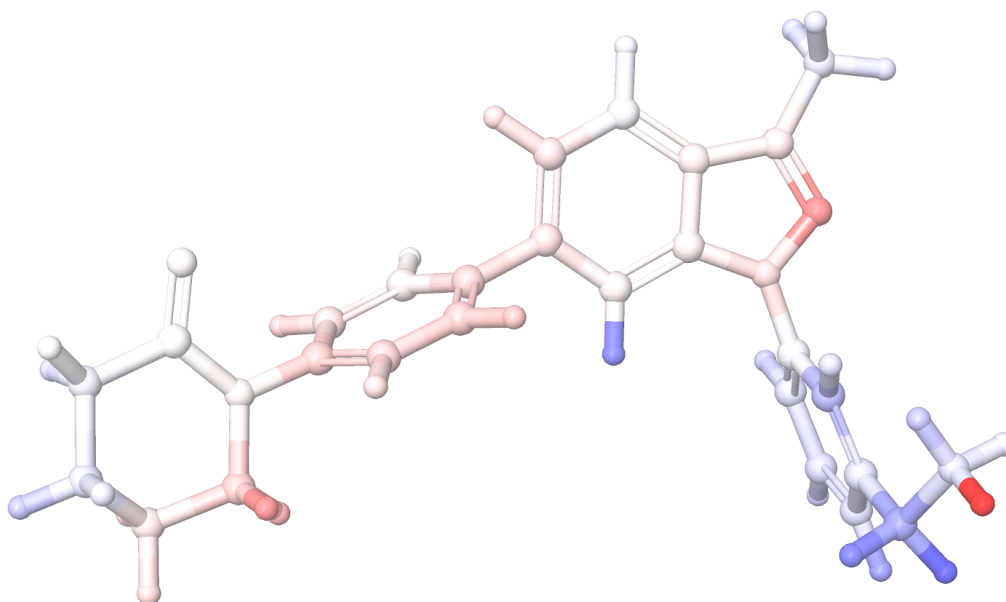


Fig. B.3 Fxa ligand with atoms colored by charge such that more negative atoms are blue and more positive atoms are red. Optimal set of charges computed with and RMSD limit of $0.01 q_e$.

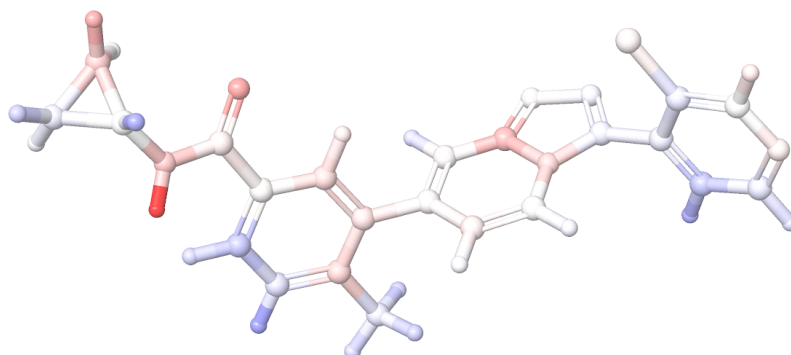


Fig. B.4 P38 ligand with atoms colored by charge such that more negative atoms are blue and more positive atoms are red. Optimal set of charges computed with and RMSD of $0.05 q_e$.

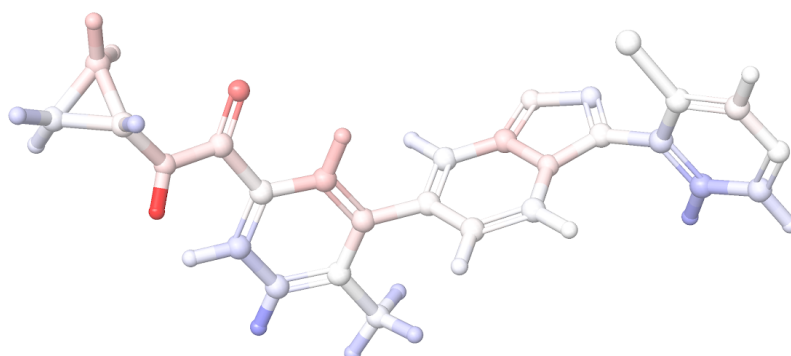


Fig. B.5 P38 ligand with atoms colored by charge such that more negative atoms are blue and more positive atoms are red. Optimal set of charges computed with and RMSD limit of $0.03 q_e$.

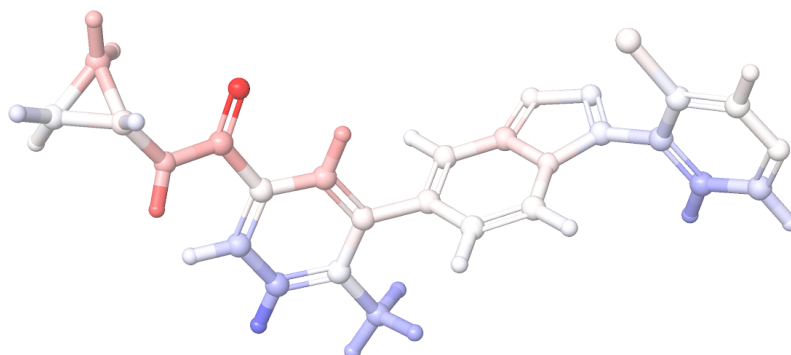


Fig. B.6 P38 ligand with atoms colored by charge such that more negative atoms are blue and more positive atoms are red. Optimal set of charges computed with and RMSD limit of $0.01 q_e$.

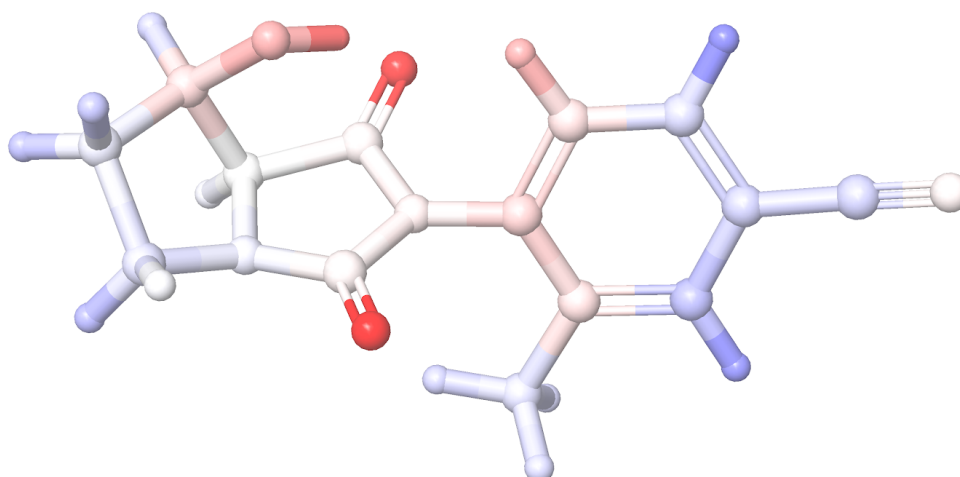


Fig. B.7 AR ligand with atoms colored by charge such that more negative atoms are blue and more positive atoms are red. Optimal set of charges computed with and RMSD limit of $0.05 q_e$.

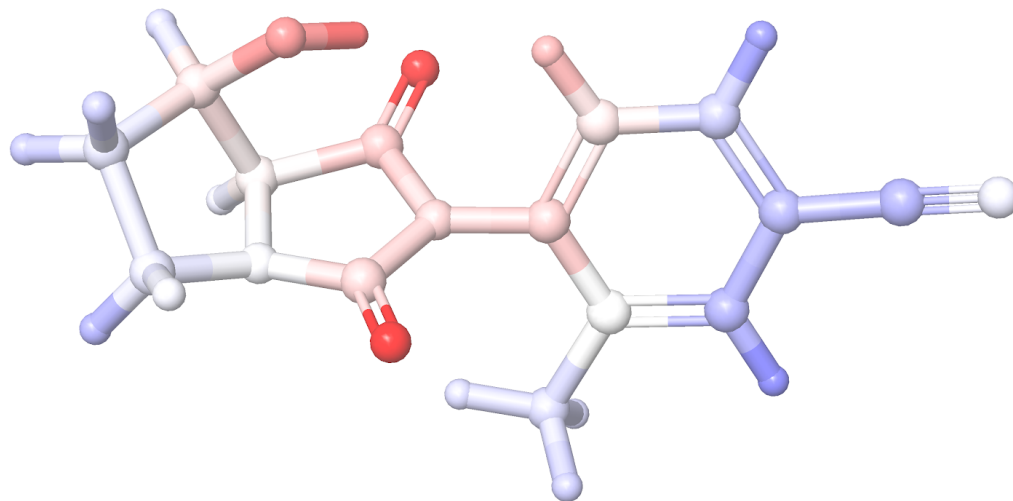


Fig. B.8 AR ligand with atoms colored by charge such that more negative atoms are blue and more positive atoms are red. Optimal set of charges computed with and RMSD limit of 0.03 q_e .

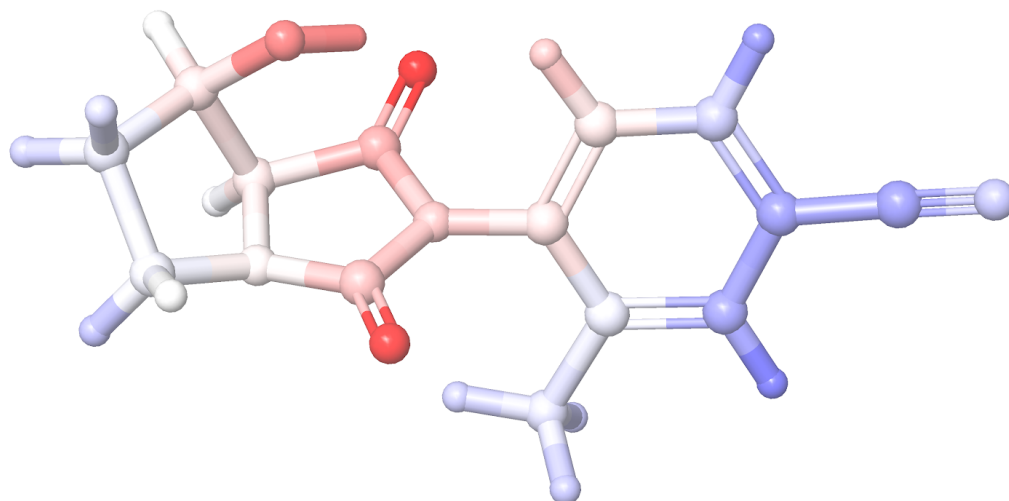


Fig. B.9 AR ligand with atoms colored by charge such that more negative atoms are blue and more positive atoms are red. Optimal set of charges computed with and RMSD limit of 0.01 q_e .

Table B.1 Calculated $\Delta\Delta G_{total}$ for the set of optimal charges. $\Delta\Delta G_{total}^{SSP}$ values are calculated by summing the average of forward and backwards SSP calculations made for each step of the optimizer. $\Delta\Delta G_{total}^{FEP}$ values are calculated from an alchemical transformation from the original charges to the optimal charges using MBAR. SSP and FEP predictions are reported as the mean of three replicates with 95% confidence interval reported between square brackets computed as mean \pm t2·SEM, where t2 is the t-distribution statistic with two degrees of freedom, and SEM is the standard error of the mean computed from the sample standard deviation of the three independent replicate predictions.

FXa			
RMSD (<i>e</i>)	0.05	0.03	0.01
$\Delta\Delta G_{total}^{FEP}$ [kcal/mol]	-8.7 [-9.2, -8.2]	-6.3 [-6.5, -6.1]	-3.1 [-3.4, -2.9]
$\Delta\Delta G_{total}^{SSP}$ [kcal/mol]	-11.3 [-12.4, -10.1]	-8.1 [-9.1, -7.0]	-3.9 [-4.4, -3.3]
P38			
RMSD (<i>e</i>)	0.05	0.03	0.01
$\Delta\Delta G_{total}^{FEP}$ [kcal/mol]	-9.4 [-10.8, -8.0]	-6.6[-7.2, -6.0]	-3.2[-3.4, -3.1]
$\Delta\Delta G_{total}^{SSP}$ [kcal/mol]	-11.1[-11.5, -10.8]	-8.3[-8.7, -7.9]	-3.5[-4.0, -3.1]
Androgen Receptor			
RMSD (<i>e</i>)	0.05	0.03	0.01
$\Delta\Delta G_{total}^{FEP}$ [kcal/mol]	-11.5[-12.0, -10.9]	-8.8[-8.8, -8.8]	-4.2[-4.2, -4.1]
$\Delta\Delta G_{total}^{SSP}$ [kcal/mol]	-11.9 [-12.3, -11.5]	-8.9[-9.2, -8.7]	-4.2[-4.4, -4.0]

as $\Delta\Delta G_{total}^{SSP}$. We compare this $\Delta\Delta G_{total}^{SSP}$ for each set of optimized charges with full MBAR FEP calculations, see table B.1. These full FEP calculations use the original and optimized charges as the two end states and calculate a relative binding free energy we term $\Delta\Delta G_{total}^{FEP}$.

Table B.1 shows that the SSP and FEP calculations are well agreed with an RMSD limit of 0.01 *e* (differing by less than 1.0 kcal/mol in all cases). For an RMSD limit of 0.03 and 0.05 *e*, SSP and FEP are less well agreed (differing by more than 1.0 kcal/mol in some cases). Table B.1 also shows clearly that changing the RMSD limit changes the calculated $\Delta\Delta G_{total}$. The relation here is that increasing the RMSD limit bound increases how much the charges can be changed and so increases the change in $\Delta\Delta G_{total}$. However, as discussed in chapter 3, the convergence of $\Delta\Delta G_{total}$ is an unnecessary condition, providing no additional information. It is only critical that the direction of the charge vectors are well converged and consistent for all RMSD limit values for all test cases.

It was discussed in the methods, of chapter 3, that the charge perturbation of 0.00015 *e* used to calculate the gradient is unbalanced, in that no counter charge is added to keep the total charge of the system neutral. This may lead to finite size effects if the periodic images of these charges interact with each other. To investigate this we calculate the $\Delta G_{binding}$ values

resulting from a perturbation of $0.00015\ e$ whilst varying the size of the simulation box. Where $\Delta G_{binding}$ is a ΔG between end states with unperturbed and perturbed charges in the bound (complex) or unbound (solvent) systems. Figure B.10 shows that the calculation of ΔG bound and unbound are not dependent on the size of the simulation box (for the size of the simulation box considered in this work).

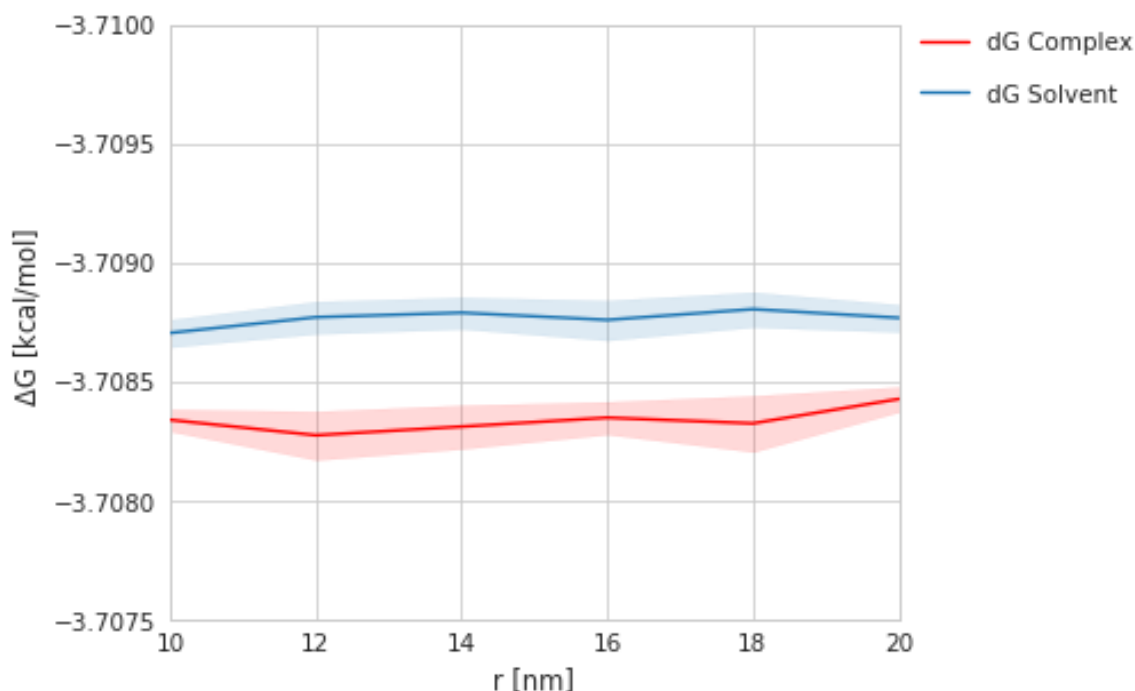


Fig. B.10 ΔG values for a $0.00015\ e$ perturbation to one atom against r (where r is the minimum padding of solvent added between the protein and edge of the box). ΔG values are calculated using SSP and 2.5 ns of sampling. ΔG s are reported as the mean of six replicates with shaded area showing 95% confidence interval computed as $\text{mean} \pm t_2 \cdot \text{SEM}$, where t_2 is the t-distribution statistic with five degrees of freedom, and SEM is the standard error of the mean computed from the sample standard deviation of the six independent replicate predictions.

To compare the speed of SSP and full MBAR FEP for computing the gradient of $\Delta\Delta G_{binding}$ w.r.t all charges, the convergence of $\Delta\Delta G_{binding}$ with sampling time needed to be investigated when $\Delta\Delta G_{binding}$ was calculated with full MBAR FEP. Figure B.11 presents the results of this investigation and shows that for a perturbation of $0.00015\ e$ to one charge of the ligand 1 ns is sufficient to calculate converged $\Delta\Delta G_{binding}$.

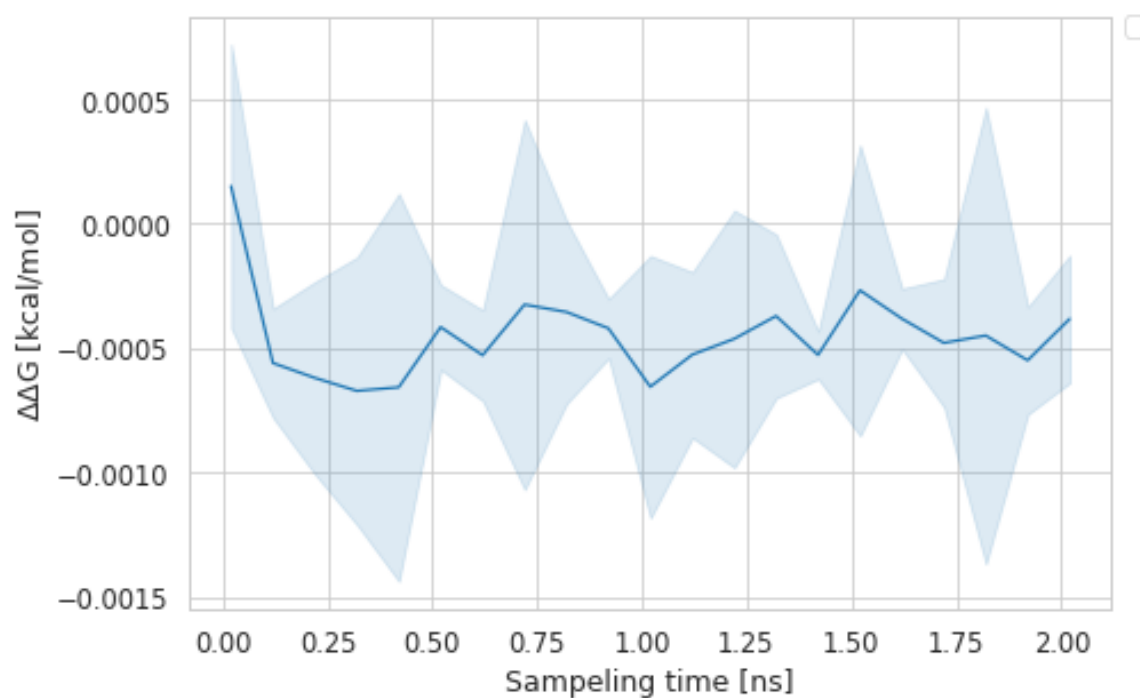


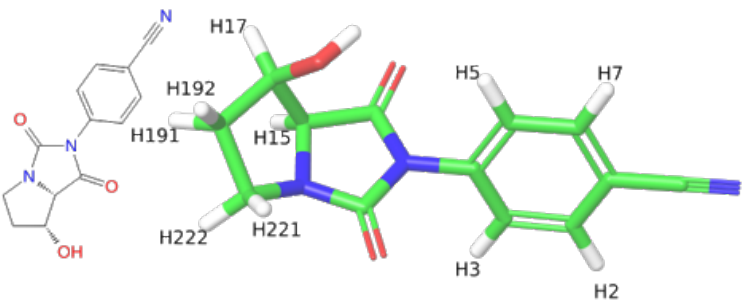
Fig. B.11 Convergence of the $\Delta\Delta G_{binding}$ predictions in the Factor Xa test case for a perturbation of $0.00015 e$ to one charge as the simulation time is increased, calculations were performed from 0.02 ns to 2.1 ns in 0.02 ns increments. The values of $\Delta\Delta G_{binding}$ are reported as mean of three replicates with the shaded area showing the 95% confidence interval computed as $\text{mean} \pm t_2 \cdot \text{SEM}$, where t_2 is the t-distribution statistic with two degrees of freedom, and SEM is the standard error of the mean computed from the sample standard deviation of the three independent replicate predictions.

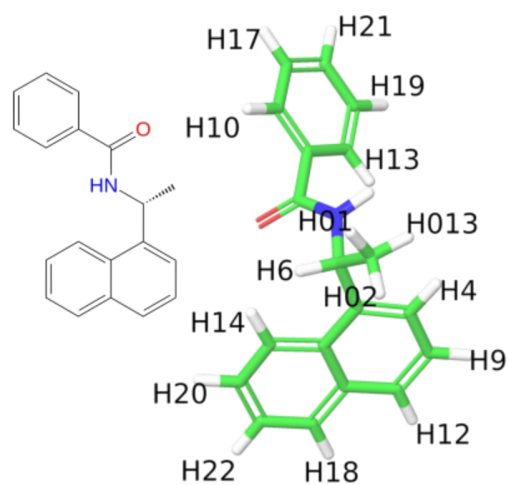
Appendix C

Sterics optimization

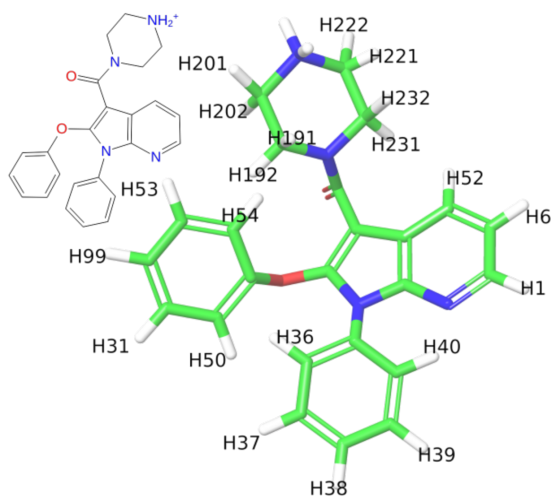
Table C.1 shows 2D and 3D structures for all ligands used in chapter 3 with all optimized hydrogens explicitly named.

Table C.1 2D and 3D structures for ligands used in androgen receptor, SARS PL protease, renin, menin (A/B), thrombin (A/B/C/D). The 3D poses of these ligands are manipulated to make explicit hydrogen labels clear.

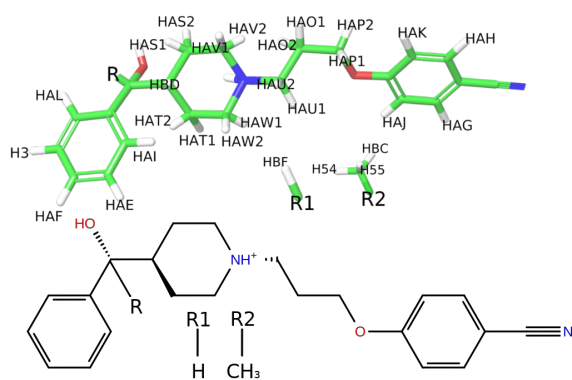
System	2D/3D Structures
Androgen receptor	



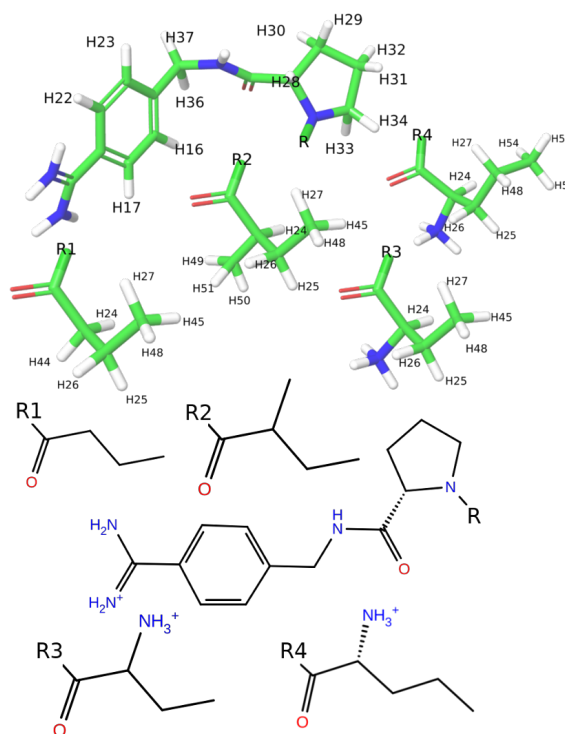
SARS PLPro



Renin



Menin A/B



Thrombin A/B/C/D

In section 4.2.6 reference is made to Additional calculations were made to verify the value of the total cumulative sum of $\Delta\Delta G_{opt}$. We now perform these calculations and this involves calculating the relative binding free energy between the original and optimized steric parameters of the inhibitor with MBAR. We name this calculation the optimization validation.

To verify these total cumulative sums of $\Delta\Delta G_{opt}$ values, optimization validation calculations are performed using the original and optimized sigmas as end states; we term this relative free energy change $\Delta\Delta G_{optval}$. The start and end state of $\Delta\Delta G_{optval}$ calculations were the original ligand sigmas and the optimized ligand sigmas respectively and these end states were linearly interpolated between using the lambda schedule. This will most likely be a different thermodynamic path than the optimizer took between these end states and is therefore a good verification of $\Delta\Delta G_{opt}$.

Optimization validation calculations were performed using 12 alchemical windows and a total of 21 ns of sampling for the AR, SARS and menin systems. 24 windows and 108 ns total sampling was used for the renin and thrombin systems as these were observed to be harder to converge using less sampling. The convergence for these optimization validation calculations can be seen in Figure C.1-C.3.

The calculated values for $\Delta\Delta G_{opt}$ and $\Delta\Delta G_{optval}$ taken from Figures C.1-C.3 using all of the sampling are presented in Table C.2.

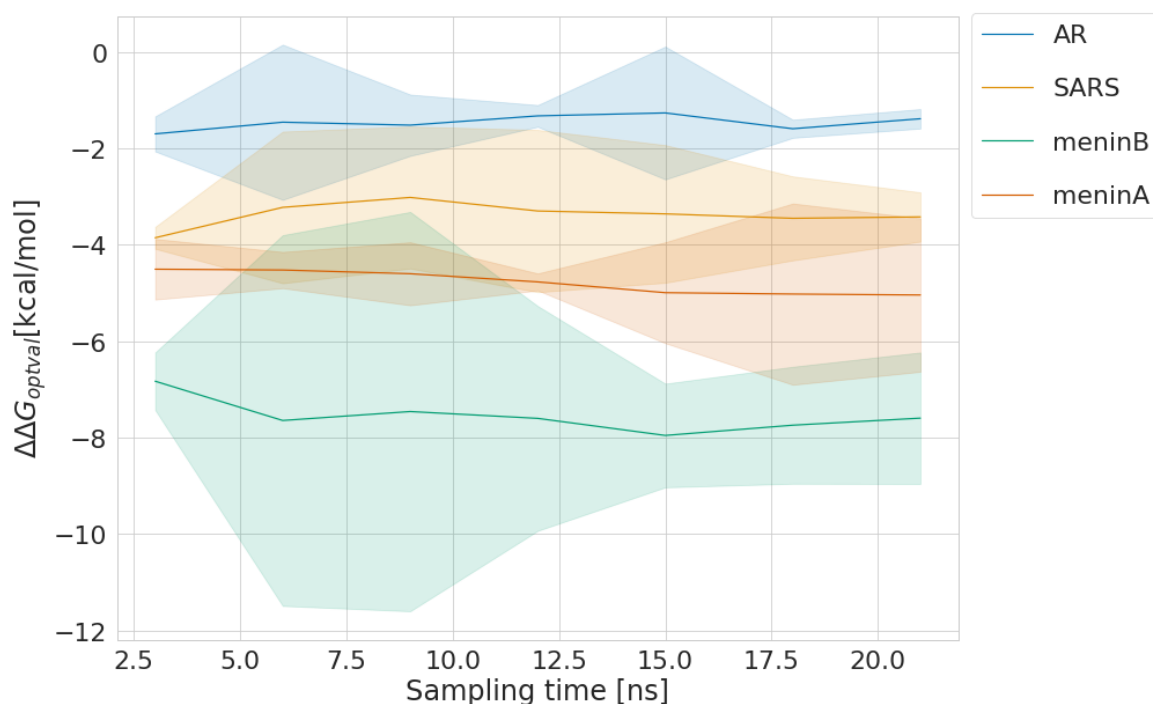


Fig. C.1 Convergence of calculation of $\Delta\Delta G_{optval}$ for systems using 21 ns of sampling. $\Delta\Delta G_{optval}$ is reported as the mean of three replicates with the shaded area showing 95% confidence interval computed as $\text{mean} \pm t_2 \cdot \text{SEM}$, where t_2 is the t-distribution statistic with two degrees.

Table C.2 Calculated $\Delta\Delta G_{opt}$ and their verification $\Delta\Delta G_{optval}$ values for all AR, SARS, menin, renin and thrombin systems.

System	$\Delta\Delta G_{opt}$ [kcal/mol]	$\Delta\Delta G_{optval}$ [kcal/mol]
AR	-1.16	-1.38 [-1.59, -1.18]
SARS	-3.80	-3.42 [-3.93, -2.91]
menin A	-5.17	-5.04 [-6.63, -3.45]
menin B	-10.13	-7.59 [-8.96, -6.23]
renin	-8.51	-1.31 [-2.36, -0.27]
thrombin A	-2.42	-1.59 [-1.86, -1.31]
thrombin B	-5.27	-4.16 [-4.95, -3.37]
thrombin C	-5.43	-4.30 [-5.84, -2.77]
thrombin D	-3.63	-4.18 [-4.60, -3.76]

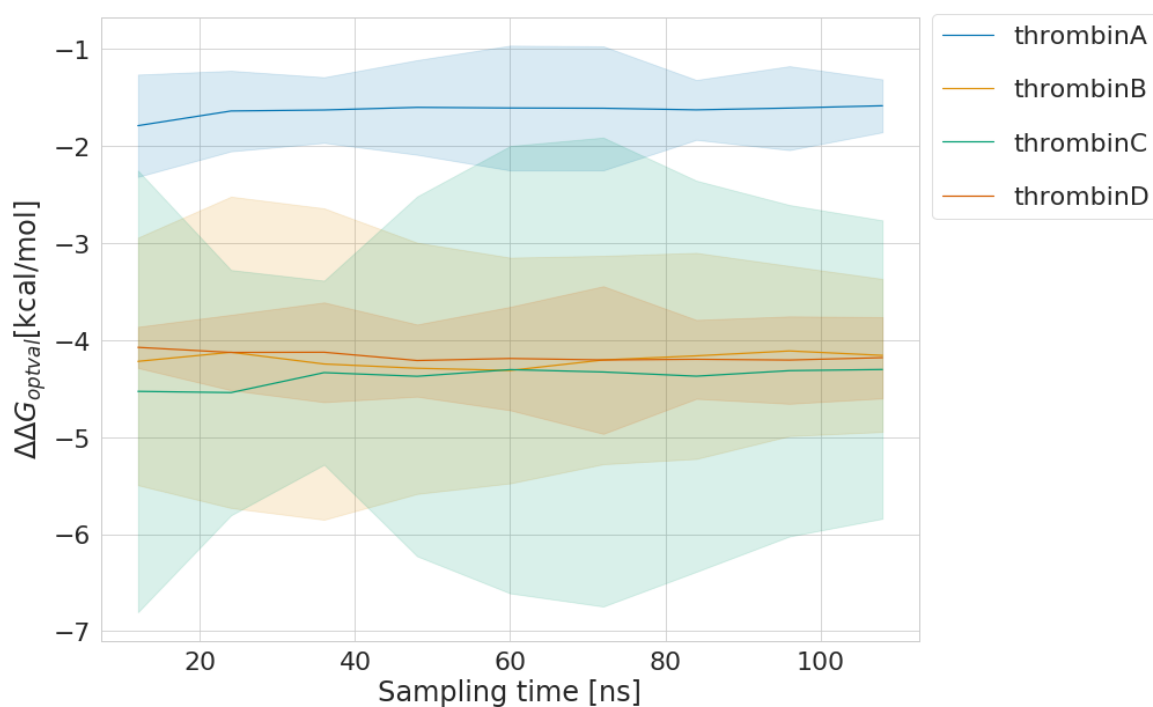


Fig. C.2 Convergence of calculation of $\Delta\Delta G_{optval}$ for systems using 108 ns of sampling. $\Delta\Delta G_{optval}$ is reported as the mean of three replicates with the shaded area showing 95% confidence interval computed as $\text{mean} \pm t_2 \cdot \text{SEM}$, where t_2 is the t-distribution statistic with two degrees.

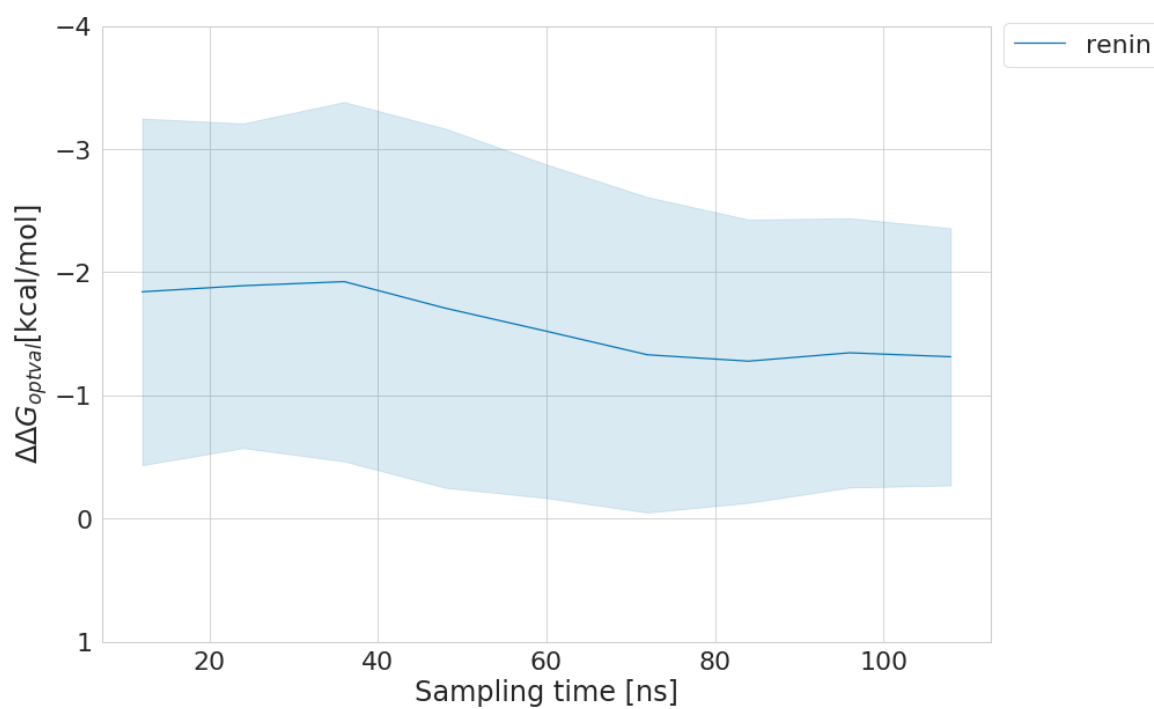


Fig. C.3 Convergence of calculation of $\Delta\Delta G_{optval}$ for systems using 108 ns of sampling. $\Delta\Delta G_{optval}$ is reported as the mean of three replicates with the shaded area showing 95% confidence interval computed as $\text{mean} \pm t2 \cdot \text{SEM}$, where $t2$ is the t-distribution statistic with two degrees.

In the chapter 3 many calculations are made for the mutation of the hydrogens on a ligand to a methyl group using MBAR and HREX. The convergence graphs for these systems are presented here. Where $\Delta\Delta G_{calc}$ in figures C.4-C.10 correspondence to $\Delta\Delta G_{scan}$ in the main text. In figures C.4-C.10 $\Delta\Delta G_{calc}$ are reported as mean of three replicates with shaded area showing 95% confidence interval computed as $\text{mean} \pm t_2 \cdot \text{SEM}$, where t_2 is the t-distribution statistic with two degrees of freedom, and SEM is the standard error of the mean computed from the sample standard deviation of the three independent replicate predictions.

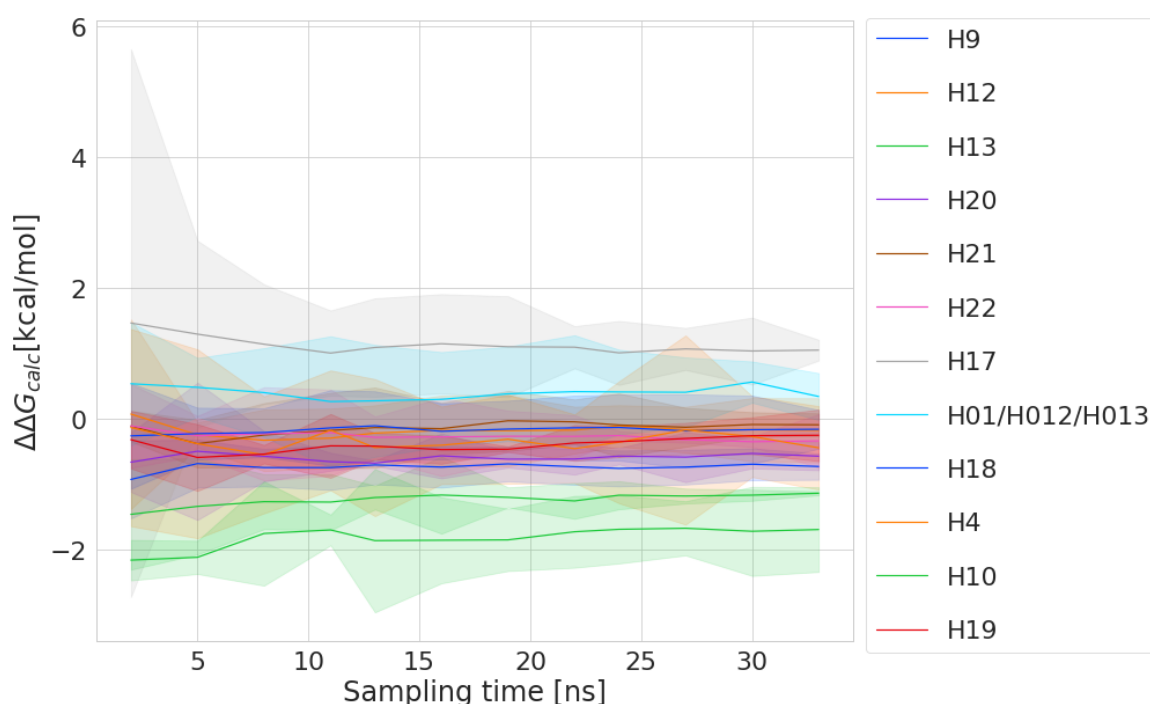


Fig. C.4 $\Delta\Delta G_{calc}$ for each methylation in the SARS system as the amount of sampling is increased.

In the main work reference is made to instability in the molecular dynamics calculation arising when mutated methyl have close contact with the protein structure. Figure C.11 shows an example of this for the androgen receptor system. Where the labeled distances measure the close contact between the added meth and atoms in the protein.

For brevity the optimized sigmas for the menin B system and the optimized charges of thrombin B are omitted from the main text here they are presented in full.

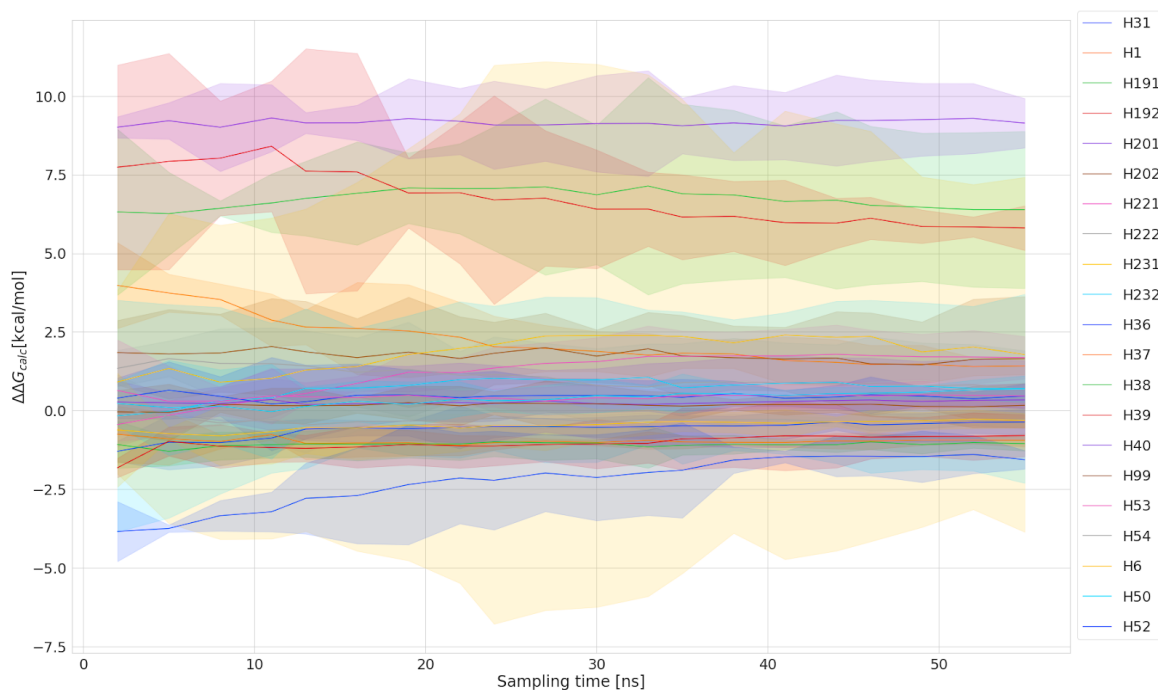


Fig. C.5 $\Delta\Delta G_{calc}$ for each methylation in the renin system as the amount of sampling is increased.

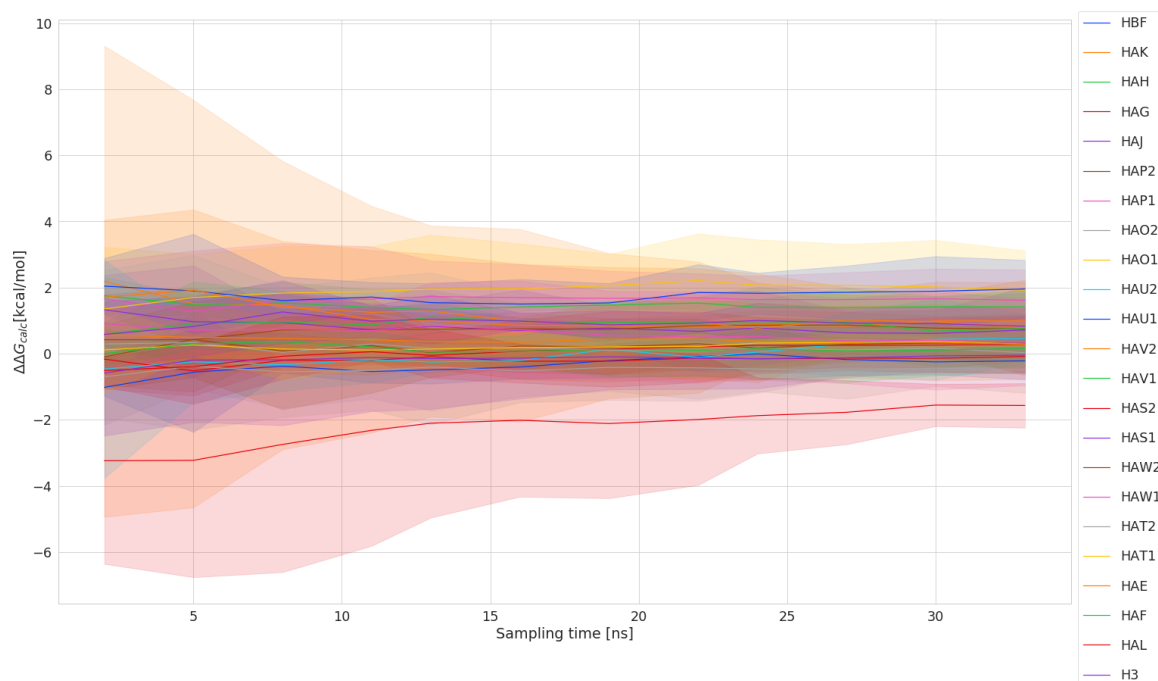


Fig. C.6 $\Delta\Delta G_{calc}$ for each methylation in the menin A system as the amount of sampling is increased.

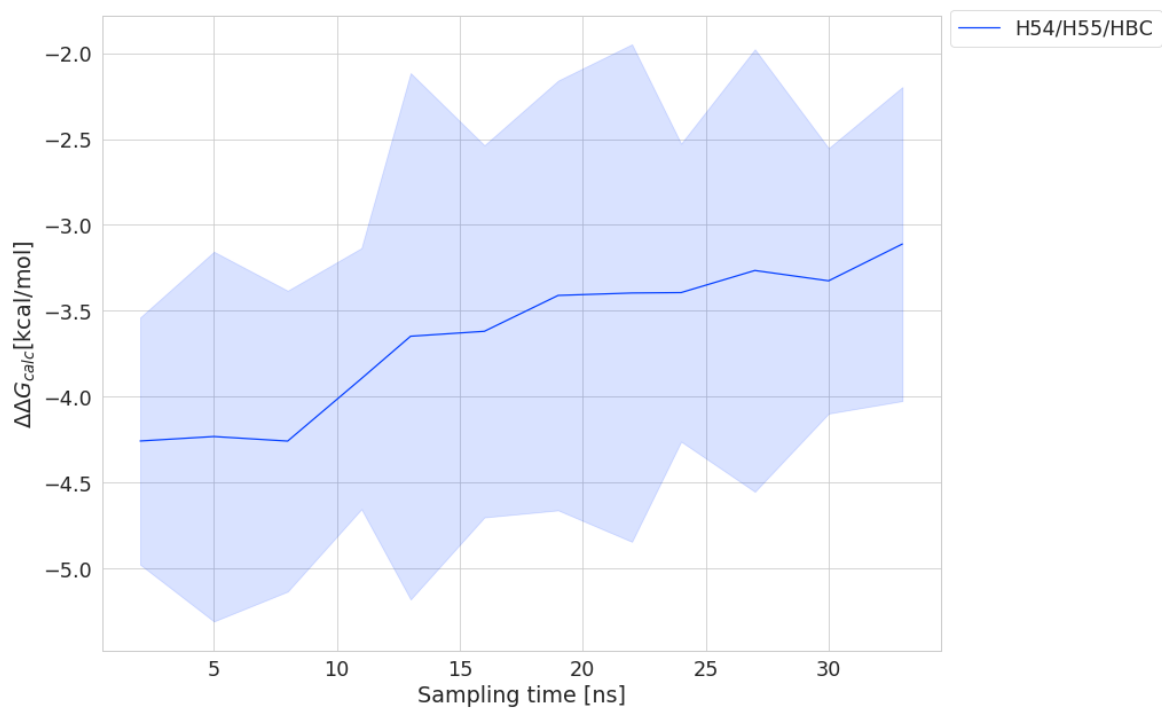


Fig. C.7 $\Delta\Delta G_{calc}$ for each methylation in the menin B system as the amount of sampling is increased.

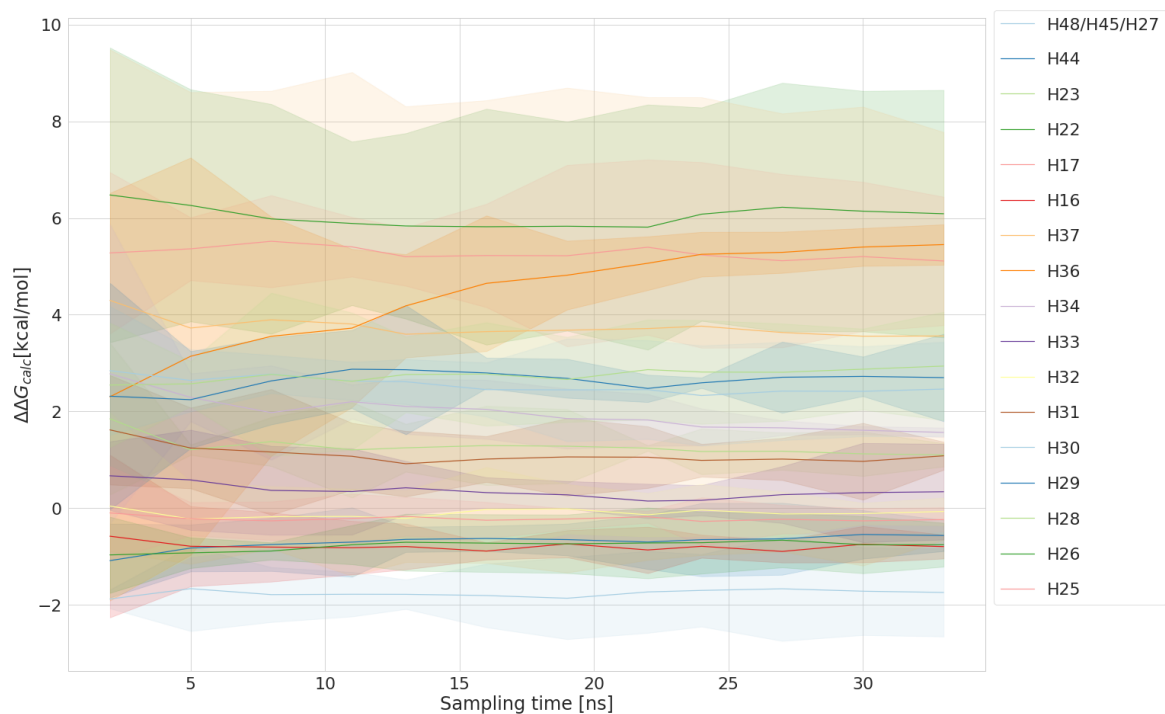


Fig. C.8 $\Delta\Delta G_{calc}$ for each methylation in the thrombin A system as the amount of sampling is increased.

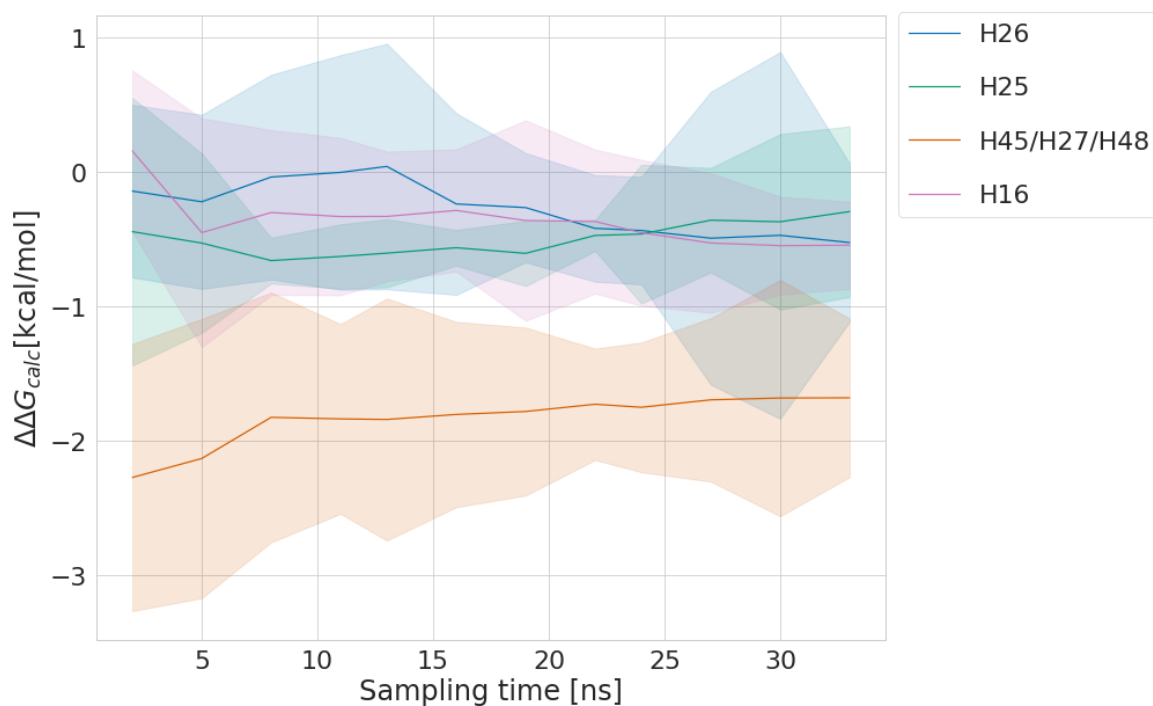


Fig. C.9 $\Delta\Delta G_{calc}$ for each methylation in the thrombin C system as the amount of sampling is increased.

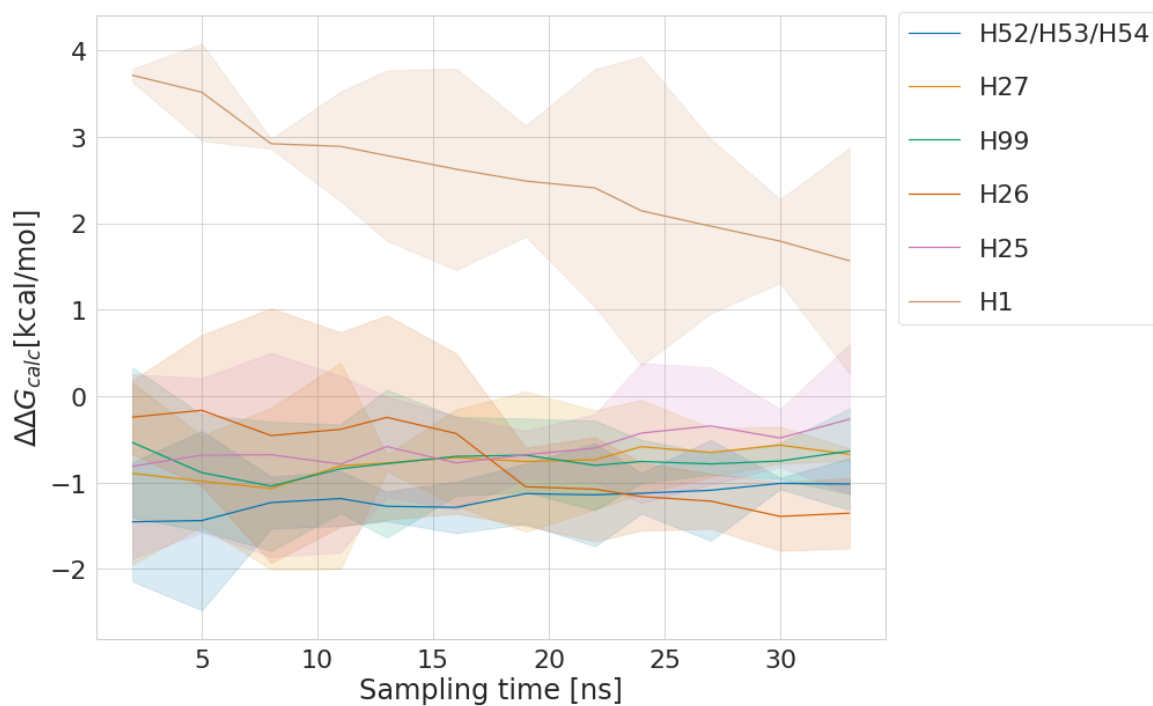


Fig. C.10 $\Delta\Delta G_{calc}$ for each methylation in the thrombin D system as the amount of sampling is increased.

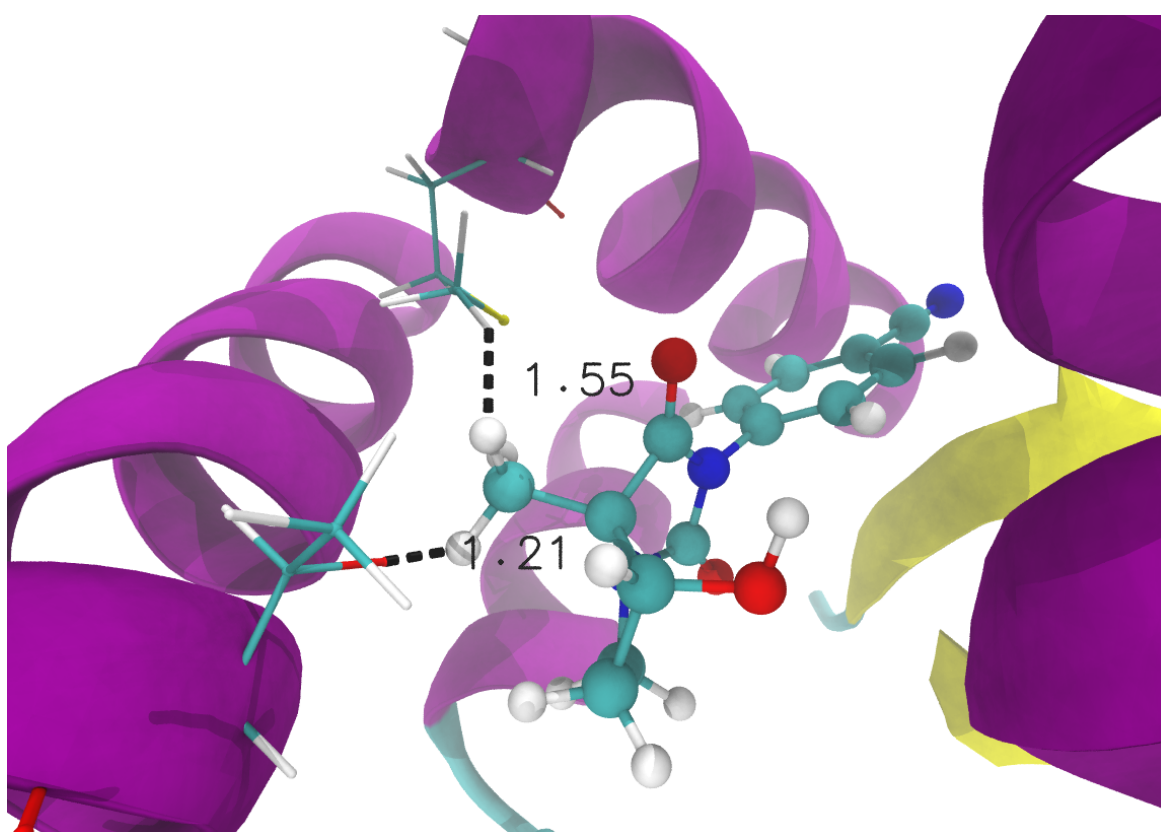


Fig. C.11 androgen receptors system with a methylation at H15 (Table C.1). Closest distance to atoms in protein side chains are labeled in angstroms.

Table C.3 $\Delta\sigma$ for each atom named in Table C.1 for the menin B test case.

Hydrogen	$\Delta\sigma$ [nm]
H55	0.296
H54	0.257
HAU2	0.180
HAW1	0.146
HAO2	0.119
HAL	0.111
HAT1	0.108
HAS2	0.103
HAV1	0.090
HAT2	0.086
H3	0.0810
HAP2	0.035
HAP1	0.011
HAU1	0.009
HAV2	0.007
HAF	-0.003
HAG	-0.005
HAI	-0.035
HAW2	-0.041
HAH	-0.050
HAS1	-0.052
HBC	-0.080
HAK	-0.106
HAO1	-0.159
HAJ	-0.160
HAI	-0.213
HBD	-0.424

Table C.4 Difference between original and optimized charges for thrombin B test case. All hydrogen names given in Table C.1

Hydrogen	$\Delta q e$
H49	0.276
H30	0.197
H33	0.194
H29	0.180
H50	0.124
H25	0.097
H45	0.060
H27	0.055
H16	0.041
H48	0.041
H23	-0.038
H24	-0.039
H17	-0.046
H36	-0.051
H31	-0.066
H51	-0.067
H28	-0.086
H26	-0.090
H22	-0.112
H34	-0.134
H32	-0.194
H37	-0.343