Free Energy Calculations in Drug Design: Application to Bromodomains



Matteo Aldeghi

Lincoln College

University of Oxford

A thesis submitted in partial fulfilment of
the requirements for the degree of

Doctor of Philosophy

Michaelmas 2016

To my parents
for their love, support, and encouragement

Acknowledgements

Firstly, I would like to express my gratitude to my supervisor Prof. Phil Biggin, whose door was always open for advice, support, and discussion on research or any other matter. He has managed to provide valuable guidance while leaving me much independence on the direction and details on how to conduct the project, which allowed me to grow as a scientist and contributed toward my enjoyment of these years of doctoral research. I also thank my second supervisor Prof. Stefan Knapp, as he too has always been available for advice and support whenever needed. His group provided a unique contribution to the project, and without him large part of the work here presented would not have been possible.

I would also like to extend my gratitude to my industrial supervisors and collaborators from Evotec: Dr Mike Bodkin, Dr Alex Heifetz, and Dr Richard Law. Mike co-supervised the doctoral research with Phil and Stefan and provided valuable feedback, suggestions, and an industrial perspective that added much to the quality of the work. Alex has always been very supportive and helpful since my first visit to Evotec, and it was a pleasure to work with him on a number of additional projects that are unfortunately outside the scope of this thesis. Richard was pivotal in setting up the project and has always been encouraging and ready to help in my research and professional development.

My wonderful experience in Oxford has also much to do with the people I have met in the Structural Bioinformatics and Computational Biochemistry group. An extensive list of all present and past members would be too long, so I limit myself to acknowledge some of my fellow D.Phil. students who are due to complete their degree with me this year: Nat Willems, Ben Morris, Laura Domicevica, Erin Cutts, and Vishal Maingi. I would also like to thank George Hedger, Jan Domański, David Shorthouse, Jo Lee, Jemma Trick, Amanda Buyan, Nick Michelarakis, Tom Newport, and Fiona Naughton. In addition, I wish to acknowledge Dr Maria Musgaard, Dr Heidi Koldsø, Dr Teresa Paramo, and Dr Anna Duncan for having always been ready to help and for the great time in and outside the lab. I would further like to thank Dr Greg Ross, who helped me in the first few weeks of the D.Phil. and provided key guidance and advice on the work presented in Chapter 6.

Outside the lab I have been fortunate to have made many friends, especially at Lincoln College and the Doctoral Training Centre. In particular, I am thankful

for the time spent with Lukas Hutter, Oakley Cox, Phil Gerken, Finn Wolfreys, Bobojon Nazarov, Theo Kapellos, Ilya Chevyrev, James Mitchell, Sam Demharter, Marcin Paczkowski, Jonny Brooks, Mike Barber, Malte Lücken, Dhruva Raman, and Pedro Aparicio. Unfortunately, this acknowledgment section is too brief to give credits to all the people who contributed towards my experience in Oxford and the UK, and towards who I am today. I extend my gratitude to all of them.

I thank my family, and especially my parents and grandparents, for their love and for always having been supportive of me. Finally, I thank you Alycia, for the time spent together in Oxford, Grenoble, and elsewhere, for your patience, and for always having been present despite the distance.

Abstract

Computer simulations of biomolecules have been improving at a pace that is faster than Moore's law for microprocessors in the last few decades. Thanks to advances in theory, hardware, and algorithms it is increasingly possible to study biological processes at relevant spatial and temporal resolutions, and to exploit simulation for quantitative predictions. One area that can potentially benefit greatly from such computational predictions is that of drug discovery. Since the inception of the concept of rational drug design, the prediction of how tightly an organic molecule binds to a macromolecular partner has been one of the chief objectives of computational chemistry. Computers already play a fundamental support role during the drug discovery process, and today many novel approaches that aim at studying the details of drug binding and predicting binding affinity are being actively investigated. In this thesis, I report a series of studies that aim to evaluate the potential utility of free energy calculations based on molecular simulations for drug design. In particular, I focus on the prediction of small-molecule binding affinities to the epigenetic target of bromodomains. Bromodomains are small protein modules that have been found in 46 human proteins involved in gene regulation. Given their role in various diseases, in particular cancer and inflammation, a number of bromodomain inhibitors are currently being investigated both in the laboratory and the clinic. Here, it is shown how thorough calculations based on explicit-solvent simulations and all-atom force fields can accurately reproduce binding free energies for this protein family. Rigorous free energy calculations are also compared to more approximate methods based on the post-processing of the simulation trajectories in implicit solvent. Finally, a recently proposed method for the estimation of water binding free energy is employed to study water displaceability from bromodomain binding pockets.

Contents

List of Figures **xiii**

List of Tables **xv**

List of Abbreviations **xvii**

1 Introduction **1**

1.1 Structure-based drug design 2

1.1.1 The role of simulations 4

1.1.2 Free energy calculations 6

1.2 Bromodomains 11

1.2.1 Structure and function of the bromodomain 11

1.2.2 Bromodomains as a therapeutic target 15

1.2.3 Current state of bromodomain inhibitors 18

1.3 Ligand binding assays 27

1.3.1 Isothermal titration calorimetry 28

1.4 Aims and motivation 33

2 Theory and Methods **35**

2.1 Molecular mechanics 36

2.1.1 Empirical force fields 36

2.1.2 Water models 40

2.2 Energy minimisation 41

2.3 Molecular dynamics simulations 42

2.3.1 Finite difference methods 44

2.3.2 Constraints 46

2.3.3 Ensembles and temperature/pressure coupling 46

2.3.4 Periodic boundary conditions 50

2.3.5 Treatment of long-range interactions 51

2.4 Monte Carlo simulations 54

2.4.1 Grand Canonical Monte Carlo simulations 56

2.5 Replica exchange 57

2.6 Free energy calculations 59

2.6.1 Alchemical methods . . . 61
2.6.2 Implicit solvent end-point methods . . . 66
2.6.3 Grand canonical integration . . . 69

3 Predicting ligand affinities to a single target 71
3.1 Introduction . . . 72
3.2 Methods . . . 74
3.2.1 System Setup . . . 74
3.2.2 Free Energy Calculations . . . 75
3.2.3 Data Analysis . . . 77
3.2.4 Docking . . . 78
3.3 Performance when starting from X-ray structures . . . 79
3.4 Performance when starting from docked poses . . . 83
3.4.1 Ligand placement and identification of crystallographic poses 84
3.4.2 Ranking the ligands by affinity . . . 87
3.5 Discussion . . . 90
3.6 Conclusions . . . 93

4 Predicting ligand selectivity 95
4.1 Introduction . . . 96
4.2 Methods . . . 98
4.2.1 System setup . . . 98
4.2.2 Derivation of torsional parameters . . . 99
4.2.3 Docking . . . 100
4.2.4 Preliminary Molecular Dynamics Simulations . . . 101
4.2.5 Free Energy Calculations . . . 101
4.2.6 Data Analysis . . . 103
4.3 Small molecule parameters: torsion angles . . . 104
4.4 RVX-208 and RVX-OH . . . 106
4.4.1 Binding pose prediction . . . 107
4.4.2 Affinity predictions . . . 109
4.5 Bromosporine . . . 110
4.5.1 Binding pose prediction . . . 111
4.5.2 Affinity predictions . . . 112
4.5.3 Effect of partial charges . . . 115
4.5.4 Effect of torsional parameters . . . 117
4.6 Discussion . . . 121
4.7 Conclusions . . . 124

5 Comparison to MMPBSA **127**
5.1 Introduction . . . 128
5.2 Methods . . . 131
5.2.1 MMPBSA calculations . . . 131
5.2.2 Entropy calculations . . . 132
5.2.3 Data analysis . . . 132
5.3 Case 1: predicting ligand binding to BRD4(1) from X-ray structures 133
5.4 Case 2: predicting ligand binding to BRD4(1) from docking poses . 137
5.4.1 Ranking of docking poses . . . 137
5.4.2 Affinity predictions . . . 138
5.5 Case 3: predicting RVX-OH vs RVX-208 binding and selectivity . . 141
5.5.1 Identification of correct RVX-OH pose . . . 141
5.5.2 Affinity predictions . . . 142
5.6 Case 4: predicting the selectivity profile of bromosporine . . . 146
5.7 Discussion . . . 147
5.8 Conclusion . . . 152

6 Water stability in bromodomain binding pockets **155**
6.1 Introduction . . . 156
6.2 Methods . . . 158
6.2.1 Systems preparation . . . 158
6.2.2 Water hydration free energy . . . 159
6.2.3 Grand Canonical simulations . . . 160
6.3 Water placement . . . 163
6.4 Stability of the water network . . . 166
6.5 Stability of individual waters . . . 170
6.6 Discussion . . . 176
6.7 Conclusion . . . 178

7 Summary **181**

Appendices

A Supplementary material for Chapter 3 **189**

B Supplementary material for Chapter 4 **193**

C Supplementary material for Chapter 5 **199**

D Supplementary material for Chapter 6 **207**

References **213**

List of Figures

1.1 Overview of the drug discovery process 3
1.2 Phylogenetic tree of the human BRD family 12
1.3 Domain organisation of example proteins that contain BRDs 13
1.4 Bromodomain fold and acetyl-lysine binding pocket 14
1.5 Structure-guided sequence alignment of the human bromodomain family 16
1.6 Example bromodomain inhibitors by target family 19
1.7 Schematic representation of ITC experimental setup and its data . . 29

2.1 Summary of energy terms in a typical biomolecular force field . . . 37
2.2 Examples of common fixed-charge water models 40
2.3 Periodic boundary conditions and examples of periodic cells 51
2.4 Replica exchange scheme 58
2.5 Thermodynamic cycle for absolute binding free energy calculations . 62
2.6 Thermodynamic cycle for MMPBSA binding free energy calculations 67
2.7 Basic principles of free energy calculations based on GCMC 70

3.1 Set of eleven ligands considered in the study 73
3.2 Crystal structures of the ligands in complex with BRD4(1) 80
3.3 Docking poses 84
3.4 Multiple potential binding conformations of ligand 3 87
3.5 Scatter plot of experimental versus calculated affinities and relative errors 88

4.1 Ligands and proteins considered in the study 97
4.2 Torsion scans for the biaryl dihedral angles for RVX-OH/RVX-208 and bromosporine 105
4.3 X-ray and docking poses of RVX-OH and RVX-208 107
4.4 Calculated versus experimental affinities for RVX-OH and RVX-208 109
4.5 Bromosporine docking poses 111
4.6 Calculated versus experimental affinities for bromosporine 113
4.7 Scatter plot of experimental versus calculated affinities for bromosporine with optimised parameters 118

5.1 Pearson and Spearman correlations for ABFE and MMPBSA for Case 1 . . . 134
5.2 Correlation difference between MMPBSA and ABFE calculations for Case 1 . . . 135
5.3 Number of correctly ranked poses for the ligands docked in Case 2 . 138
5.4 Pearson and Spearman correlations for ABFE and MMPBSA for Case 2 . . . 139
5.5 Correlation difference between MMPBSA and ABFE calculations for Case 2 . . . 140
5.6 Number of correctly ranked poses for the ligands docked in Case 3 . 143
5.7 Pearson and Spearman correlations for ABFE and MMPBSA for Case 3 . . . 144
5.8 Correlation difference between MMPBSA and ABFE calculations for Case 3 . . . 145
5.9 Pearson and Spearman correlations for ABFE and MMPBSA for Case 4 . . . 147
5.10 Correlation difference between MMPBSA and ABFE calculations for Case 4 . . . 148

6.1 Water network structure and proteins considered in the study . . . 157
6.2 Water networks as predicted by the simulations . . . 163
6.3 Water placement for the three water models . . . 164
6.4 Stability of the water network . . . 167
6.5 Stability of individual water sites for all bromodomains considered . 171
6.6 Swarm plots of individual water stabilities . . . 174

B.1 Quality of fit between MM and QM energies for the optimised bromosporine groups . . . 198

C.1 Scatter plot of experimental versus calculated affinities for Case 1 . 200
C.2 Scatter plot of experimental versus calculated affinities for Case 2 . 201
C.3 Scatter plot of experimental versus calculated affinities for Case 3 . 202
C.4 Scatter plot of experimental versus calculated affinities for Case 4 . 203

D.1 Distributions of placement errors and free energies for the three water models . . . 212

List of Tables

1.1 Bromodomain inhibitors currently in clinical trial 26

2.1 Comparison of parameters for a few water models 41

3.1 Summary of free energy calculation results based on crystal structures 81
3.2 Summary of docking results 86
3.3 Summary of the free energy calculation results based on docking . . 89

4.1 Torsional parameters for the biaryl torsions in bromosporine and RVX-OH/RVX-208 100
4.2 Summary of results for RVX-OH and RVX-208 108
4.3 Bromosporine calculated binding free energies 114
4.4 Bromosporine calculated binding free energies with AM1-BCC charges 116
4.5 Calculated binding free energies for bromosporine with optimised torsional parameters for the sulfonamide and the carbamate groups and the biaryl torsion 120

5.1 Pearson and Spearman correlations for all MMPBSA calculations. . 150

A.1 Physicochemical properties of the ligands considered 190
A.2 Full breakdown of free energy results starting from the crystal structures 191
A.3 Full breakdown of free energy results starting from the docking poses 192

B.1 Information on the protein models used in the calculations 194
B.2 Full breakdown of free energy components for the calculations of RVX-OH and RVX-208 195
B.3 Full breakdown of free energy components for the calculations of bromosporine 196
B.4 Optimised torsional parameters for bromosporine 197

C.1 All free energy results for Case 1 204
C.2 All free energy results for Case 2 204
C.3 All free energy results for Case 3 205
C.4 All free energy results for Case 4 205

D.1 Structures used for the GCMC calculation and overall results 208
D.2 Results obtained with the TIP3P water model 209
D.3 Results obtained with the TIP4P water model 210
D.4 Results obtained with the SPC water model 211

List of Abbreviations

ABFE . . . . . Absolute binding free energy.

BET . . . . . Bromodomain and extra terminal.

BRD . . . . . Bromodomain.

FEP . . . . . Free energy perturbation.

FRET . . . . . Fluorescence resonance energy transfer.

GCMC . . . . Grand canonical Monte Carlo.

GCI . . . . . Grand canonical integration.

HAT . . . . . Histone acetyltransferase.

HDAC . . . . . Histone deacetylase.

ITC . . . . . Isothermal titration calorimetry.

Kac . . . . . Acetylated lysine.

MBAR . . . . Multiple Bennett's acceptance ratio.

MC . . . . . Monte Carlo.

MD . . . . . Molecular dynamics.

MMPBSA . . Molecular mechanics Poisson-Boltzmann surface area.

MUE . . . . . Mean unsigned error.

PDB . . . . . Protein Data Bank.

SPR . . . . . Surface plasmon resonance.

RBFE . . . . . Relative binding free energy.

RMSD . . . . . Root mean squared deviation.

RMSE . . . . . Root mean squared error.

vdW . . . . . van der Waals.

If we were to name the most powerful assumption of all, which leads one on and on in an attempt to understand life, it is that all things are made of atoms, and that everything that living things do can be understood in terms of the jigglings and wigglings of atoms.

— Richard Feynman

1 Introduction

Contents

1.1 Structure-based drug design 2
1.1.1 The role of simulations 4
1.1.2 Free energy calculations 6
1.2 Bromodomains 11
1.2.1 Structure and function of the bromodomain 11
1.2.2 Bromodomains as a therapeutic target 15
1.2.3 Current state of bromodomain inhibitors 18
1.3 Ligand binding assays 27
1.3.1 Isothermal titration calorimetry 28
1.4 Aims and motivation 33

In this chapter I will introduce the topic of simulation-based free energy calculations and that of bromodomains as a therapeutic target. In doing so, I will briefly describe the role of computer simulations in drug discovery, and then the state of the art in binding affinity predictions. Furthermore, I will provide an overview of the structure and function of bromodomains, as well as their role in disease, potential as therapeutic target, and the current state of bromodomain inhibitor development. Finally, I will concisely describe the biophysical technique of isothermal titration calorimetry, which is used for the accurate experimental measurement of binding free energies. This chapter will serve the purpose of providing the necessary background

information in order to contextualise the results of this thesis, and to provide support to the motivation and aims of the work presented.

Small portions of this chapter have been based on text written for the following publications:

P.C. Biggin, M. Aldeghi, M.J. Bodkin, A. Heiftez. "Beyond Membrane Protein Structure: Drug Discovery, Dynamics and Difficulties". In: *The Next Generation in Membrane Protein Structure Determination*. Ed. by Isabel Moraes. Advances in Experimental Medicine and Biology. Springer, 2016. Chap. 12, pp. 161-181.

M. Aldeghi and P.C. Biggin. "Advances in Molecular Simulations". In: *Comprehensive Medicinal Chemistry III*. Elsevier, 2017. *In press*.

1.1 Structure-based drug design

The goal of drug discovery is to find a therapy for a given medical need. At the molecular level, this is achieved by the modulation of a biological target molecule, typically exploiting non-covalent interactions with small organic molecules or larger biomolecules. Assuming the relevance and potential of the biological target for the disease of interest is established, the challenge is to find that molecule that can bind the target of interest and elicit the desired response, while also being non-toxic and bioavailable (the typical stages of the drug discovery and development process are summarised in Figure 1.1). Considering the size of the drug-like chemical space has been estimated to be between 10^{23} and 10^{60} [1], the task corresponds to the ordeal of finding a needle in a haystack. Therefore, during the past century there has been a transition from the trial-and-error approach of classical pharmacology to a so-called structure-based drug design approach (although phenotypic screening is still valuable) [2], which tries to engineer a drug or a ligand, based on the knowledge of the three-dimensional structure of the target, or on the physicochemical patterns shared by known binders [3]. A recent study places the development cost of a new drug at around $2.6 billions [4], when including the price of failures and the expected returns forgone by investors while a drug is in development. It is therefore important to be able to quickly identify the most promising drug candidates in

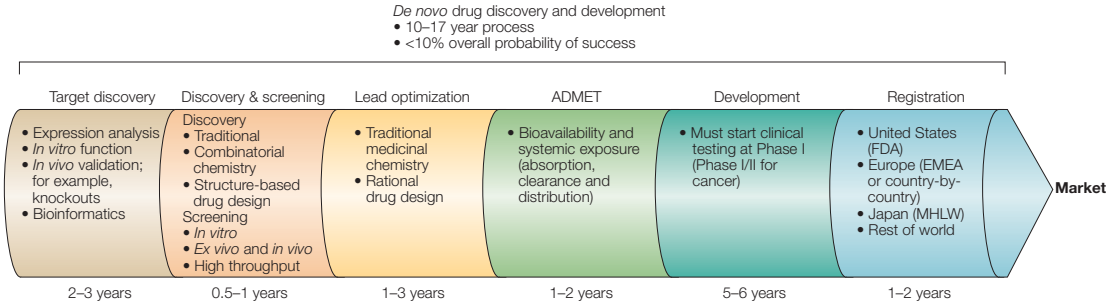


Figure 1.1: Overview of the stages involved in a typical drug discovery campaign from target selection to market. Reproduced from Ashburn & Thor [5] with permission granted by Nature Publishing Group.

order to reduce the number of failures and to bring a product to market as quickly as possible. Being able to discover new drugs quickly and cheaply is also important in light of the constant and fast evolution of pathogens, leading to new human diseases and the rise of resistance. Faster and less expensive development cycles are indeed among the goals of structure-based drug design.

In this approach, computers play a central role in aiding the scientist decision-making process [6]. Structural information about the target is valuable as it can simplify the hit identification and lead optimisation stages. In fact, in structure-based drug design (SBDD), a structure of the biological target is typically obtained via X-ray crystallography or nuclear magnetic resonance (NMR) and it is used as a template for the design of small molecule ligands. An initial hit can be identified experimentally by high-throughput or fragment screening, but also computationally using virtual screening (VS). VS can also be used for the design of the screening library in order to reduce the number of compounds to test. Docking in particular is a popular computational approach used for VS where a search algorithm is used to propose ligand binding poses based on the structure of the binding pocket, and then a scoring function that is fast to compute is employed to rank a large number of ligands [7–11]. Most often the initial hit lacks the properties of a lead compound, such as potency or bioavailability. The information provided from an apo structure of the target protein, or even better of the protein-ligand complex, can then be used to guide the medicinal chemistry efforts aimed at achieving the

desired properties by chemically modifying the initial hit. Again, computation is typically employed, not only for the visualisation of structures, but also to provide further information that is hard to gather from visual inspection, such as the energetics of specific protein-ligand interactions. Many approaches can be employed at this stage, from the use of docking and scoring, to molecular simulations and quantum mechanics (QM) calculations [6, 12].

While here we focus on structure-based drug design, it is important to remember that in many drug discovery campaigns a structure of the target is not available. One approach is to try to predict the structure of the protein of interest, most often via homology modelling. However, ligand-based approaches are widely adopted too. Various algorithms and pipelines can be used for library design, trying to maximise the chances of finding a hit during a screening campaign, by for instance increasing the structural diversity of the ligand set, or by building a focussed library based on the target protein family. At the hit-to-lead or lead optimisation stages, quantitative structural-activity or -property relationships (QSAR/QSPR) can instead be built based on the structure and physicochemical properties of a set of different compounds for which experimental data is available or has been collected [13].

1.1.1 The role of simulations

As in most sciences or businesses nowadays, computers and computation have an increasingly pervasive role and are used to perform a wide range of tasks, from data storage and database management, to automated data analysis, visualisation, and complex calculations. However, what is the role of computer physics-based simulations in drug discovery?

Structural studies have demonstrated the importance of protein and ligand flexibility for binding. In fact, while static models provide valuable insights into macromolecular structure, molecular recognition and drug binding are dynamic events. However, obtaining multiple protein conformations through NMR or crystallography is very expensive and labour intensive, so that computational techniques have been sought in order to predict protein motions [14]. A quantum-mechanical

description of these motions for large molecules like proteins is computationally too intensive for even the current best supercomputers, and the situation is not expected to change substantially in the near future either. Molecular dynamics (MD) simulations for proteins, developed in the late 1970s [15], reduce the computational complexity by using simple approximations based on Newtonian physics to simulate atomic motions [14]. The computational cost of MD simulations is still not negligible, in particular when the relevant biological timescales are longer than microseconds. Nonetheless, it is accessible to current levels of computer power often available in-house or in regional facilities. Therefore, many studies based on MD simulations have been performed in order to elucidate the binding mechanism and the role of different protein conformations for a number of drug molecules [14].

Typical applications of simulations for drug discovery have been the identification of cryptic and allosteric binding sites, the improvement of screening campaigns by the use of multiple protein conformations, the study of binding pathways, and the prediction of binding affinities [14]. For instance, Wassman *et al.* [16] identified a transiently open binding pocket in the tumour suppressor p53 core using MD simulations. Although different compounds are known to restore wild-type function in mutant p53, their binding sites and mechanisms of action are elusive. Mutational studies confirmed the cryptic pocket as the potential binding site of one of these reactivations compounds. The authors then performed ensemble-based virtual screening against the newly discovered pocket and managed to identify stictic acid as a possible p53 reactivation compound: the activity of the compound was demonstrated in human osteosarcoma cells. Amaro *et al.* [17] discovered instead drug-like inhibitors of a key enzyme in the editing machinery of *Trypanosoma brucei*, the parasite causing sleeping sickness. These inhibitors were identified using an approach that employed MD to account for protein flexibility, with a number of simulation snapshots used to refine the initial docking results. The ranking of the ligands resulted in an initial recommended set of eight ligands, two of which exhibited μM activity against the intended target. A subsequent similarity search

based on the most active compound, followed by rescoring, again using ensemble-based VS, resulted in the identification of six additional ligands, two of which were confirmed to bind to the target with with IC_{50} values of $\sim$1 μM.

Simulations have also been employed to elucidate the binding mechanism of different drugs. Limongelli *et al.* [18] have studied the full dissociation process of SC-558, a highly selective cyclooxygenase-2 (COX-2) inhibitor, through the use of advanced MD simulations. They discovered an alternative binding mode to the one observed experimentally, and suggested this as the reason behind the selectivity and the long residence time of the inhibitor for this isoform. Grazioso *et al.* [19] used a similar method to study glutamate uptake from the synaptic cleft and its release into the intracellular medium. The authors managed to describe the opening mechanism, which involved a large-scale hairpin motion, and the associated free energy profile. Herbert *et al.* [20] reported the discovery of the first extracellular inhibitor of the fibroblast growth factor receptor (FGFR), a tyrosine kinase receptor involved in cell growth, proliferation, and survival. In this study, simulations were used in order to study the ligand's mode of action. Experimental methods had been unsuccessful in identifying the inhibitor's binding site, however, the simulations correctly predicted the dissociation constant of the inhibitor, as well as for other lower potency analogues, validating the *in silico* analysis.

1.1.2 Free energy calculations

One of the "holy grails" of computational drug design is the accurate prediction of the affinity of a drug for its target protein. Despite the development of pharmacologically active molecules being a multifactorial optimisation problem, where other considerations too, such as bioavailability and toxicity, play an important role, high affinity of a compound for its intended biological target is a necessary requirement for achieving a potent, selective and ultimately efficacious drug. Unfortunately, even when structural information is available, solvent effects, conformational changes of the protein and/or the ligand and entropy-enthalpy compensation make the rationalisation of the ligand-macromolecule association process a very complex

task [21, 22]. However, thanks to important advances in theory and computing, particularly in the last decade, the prediction of binding affinities using physics-based computer simulations holds the promise [23, 24] to achieve reliable binding energy estimates by naturally taking into account complicating effects due to the discrete nature of solvent and entropy changes upon binding.

Alchemical free energy calculations and steered methods based on all-atom MD simulations in explicit solvent are two of the approaches that operate at the highest level of theoretical rigour and that are also accessible to current typical levels of computational power. Alchemical methods are based on a non-physical thermodynamic cycle, where the binding free energy is computed as the sum of multiple steps during which the ligand is "inserted" or "removed" from different environments, such as a bound and unbound state [25]. Steered or pulling method approaches follow instead a physical pathway, by applying a force that pulls the ligand away from the protein. This is typically achieved either with non-equilibrium simulations using the Jarzynski relationship [26–28], or by harmonically restraining the ligand at different distances from the binding pocket and then computing a potential of mean force [25, 29, 30]. Alternative popular approaches include endpoint methods that involve implicit solvent post-processing of explicit-solvent simulations, such as molecular mechanics with Poisson-Boltzmann or generalised Born and surface area (MMPBSA and MMGBSA) methods [31–35]. Another promising approach is metadynamics [36–38], where biasing energies are added in order to sample multiple binding events, using a funnel-shaped restraining potential [39, 40] or path collective variables [41, 42]. The advantage of simulation-based approaches over scoring functions is that the full flexibility of protein and ligand is taken into account, which is fundamental as it is the ensemble of conformations that determines the free energy rather than any single one.

While the estimation of relative binding free energies (RBFE), i.e. the difference in binding affinities between two ligands, appears to be ready to be applied to a range of biological targets (albeit with still many exclusions), the prediction of absolute binding affinities still faces many challenges [43, 44]. Recent studies have

shown how the prediction of relative affinities can guide medicinal chemistry efforts in lead optimisation [24]. The Jorgensen group has been combining computation with medicinal chemistry for the development of potent HIV reverse transcriptase inhibitors [45, 46]. In one of these studies, Bollini *et al.* [45] used relative free energy calculations to guide the modification of an initial 5 μM hit into a sub-nanomolar ligand; the use of the computational method was crucial for the identification of optimal substitution patterns. Similarly, Jorgensen *et al.* [46] reported the optimisation of three low affinity hits into low nanomolar inhibitors of both the wild type and the Y181C variant of HIV-1 reverse transcriptase. Thanks to the use of affinity predictions, these potent and dual-target inhibitors were achieved after the synthesis of only $\sim$30 compounds [46]. Recently, Wang *et al.* [47] reported a large retrospective study with about 200 ligands and 10 protein targets, using an improved small molecule force field. Their predictions correlated extremely well (weighted average R-value of 0.75) and were mostly within 1.0 kcal/mol of the experimental values.

The calculation of absolute binding free energies (ABFE) is on the other hand still a challenge as shown by a recent SAMPL exercise, where participants were asked to predict affinities of HIV integrase inhibitors in the catalytic core domain [48]. While in RBFE calculations sampling issues might have small effects [44], and indeed are still not to be underestimated even in apparently simple cases [49], these are likely to have large repercussions when calculating absolute binding free energies. Mobley *et al.* [50] showed how missing to sample even a valine side-chain reorientation upon binding can lead to an error of several kcal/mol in the binding affinity prediction. For these reasons, prospective applications of ABFE calculations are still rare and most efforts have focused on retrospective validation of the methodology, with one of the most studied macromolecular systems having been the binding pocket of engineered T4 lysozymes.

The advantage of absolute over relative binding free energy calculations is in the fact that the former are more general. That is, in theory they can be applied to any protein-ligand pair, so that one can compare the binding of two very dissimilar

ligands, or the selectivity of one ligand for two different proteins, or multiple binding poses for the same ligand. On the other hand, RBFE calculations tend to be restricted to series of similar compounds. This is very useful at the lead optimisation stage; however, it limits the scope for other applications during the early stages of the drug discovery pipeline. Both types of calculations are, however, actively being studied and improved, so that the field of applicability of both approaches is expected to grow the more efforts are put into their development and validation. For instance, it has recently been shown how it could be possible to combine Hamiltonian replica exchange with Monte Carlo ligand translation/rotation moves to simultaneously identify ligand binding sites and orientations and estimate their binding free energies [51]. A current challenge in RBFE is instead the implementation of alchemical transformations that allows the insertion of new ring structures in a compound [52]. While this is not currently reliably achievable, it might become routine in a few years, expanding the scope of RBFE calculations.

While even a level of accuracy of about 2 kcal/mol as root mean square error (RMSE) would already be useful in a drug design context when screening large libraries [44, 53], the performance of ABFE calculations still needs to be validated against more relevant targets and small molecules. Applications to complex proteins such as ion channels and GPCRs where conformational changes play an important role in the function of the receptor [54–56] would largely amplify the current limitations that need to be overcome in order to obtain reliable predictions. Similarly to the sampling issue of a side-chain rotation, slow degrees of freedoms like conformational changes upon binding cause significant sampling issues for the timescales currently accessible computationally. Overall, ABFE calculations are still computationally challenging and, when considering calculations based on alchemical cycles, their performance has been limitedly tested against experimental data, with a focus on a few model systems [57–59]. Careful retrospective validation is therefore still needed before the methodology can be confidently applied in a prospective drug discovery scenario.

Free energy calculations for water molecules

From a theoretical perspective, the binding free energy of water molecules to a protein binding site can be calculated as for any other molecule. Nonetheless, given the small size of a water molecule and the fact it is also the solvent, there are a number of technical challenges for the straightforward application of these approaches. For instance, water molecules often do not have a well-defined binding pose like larger ligands, and they tend to form networks that are very dynamic and respond quickly to small protein conformational changes. In addition, being the solvent, the definition of its "unbound" end state tend not to be straightforward and somewhat artificial.

Thus, a set of alternative methodologies for the calculation of water binding free energy, which provides an estimate of water stability and is the negative of the free energy penalty associated with water displacement, have been sought by a few research groups. Currently, there are a number of approaches available for the prediction of water site location and stability in protein cavities. Among the faster approaches are methods based on docking and scoring, such as WaterDock [60], and on integral equation theory of liquids, such as the 3D reference interaction site model (3D-RISM) [61, 62]. These methods allow for fast water placement, yet do not provide an accurate estimate of water stability, and do not account for cooperative effects present in water networks. Approaches based on inhomogeneous fluid solvation theory (IFST) [63, 64], such as WaterMap [65] and grid inhomogeneous solvation theory (GIST) [66], or on grid cell theory (GCT) [67], rely on the post-processing of an MD or MC simulation to estimate entropic and enthalpic contributions to water binding by analysing the solvent energetic interactions with the system as well as its translational and orientational degrees of freedom. These approaches provide the lowest free energy distribution of solvent molecules for the chosen force field, and as such are rigorous water placement methods. On the other hand, the estimation of water binding free energies is still approximate, with the calculation of entropic contributions being particularly challenging. Alchemical free energy methods such as double-decoupling can return the exact binding free energy for a water molecule to a specific site, but it relies on the knowledge of the position

of the water site in advance. In addition, the calculation of the hydration free energy of a water network containing multiple molecules with alchemical methods is very demanding, full of potential pitfalls, and still not free of assumptions and approximations. An approach recently proposed by Ross *et al.* [68] is based instead on Monte Carlo simulation in the grand canonical ensemble (GCMC) and tries to tackle the limitations of the above mentioned techniques. The method has the advantage of being able to identify preferred water molecules positions and at the same time rigorously calculate hydration free energies in a single set of simulations. In addition, the approach naturally takes into account cooperative effects in water networks. By post-processing the trajectories, it is also possible to extract approximate binding free energies for any water site or water network present within the grand canonical region of the simulation.

1.2 Bromodomains

The work presented in this thesis is focussed on a specific protein family, that of bromodomains (BRDs). These domains have been found in many human proteins and are an epigenetic drug target currently under investigation; as such the number of BRD inhibitors and associated high-quality experimental data available are increasing rapidly. Bromodomains bind to acetylated lysine (Kac), without catalysing any reaction or resulting in large conformational changes, and their structure is fairly small (about 120 residues in length) and rigid. For these reasons, BRDs are interesting model systems to be used for the validation of quantitative computational approaches. In this section, I will briefly provide an overview of bromodomains function, structure, and therapeutic potential, as well as the current landscape of BRD inhibitors.

1.2.1 Structure and function of the bromodomain

BRDs are epigenetic mark readers that specifically recognise ϵ-N-lysine acetylation motifs [69, 70]. Acetylation neutralises the positive charge of lysine residues, thereby reducing the ability of histones to neutralise the DNA negative charge, resulting in an

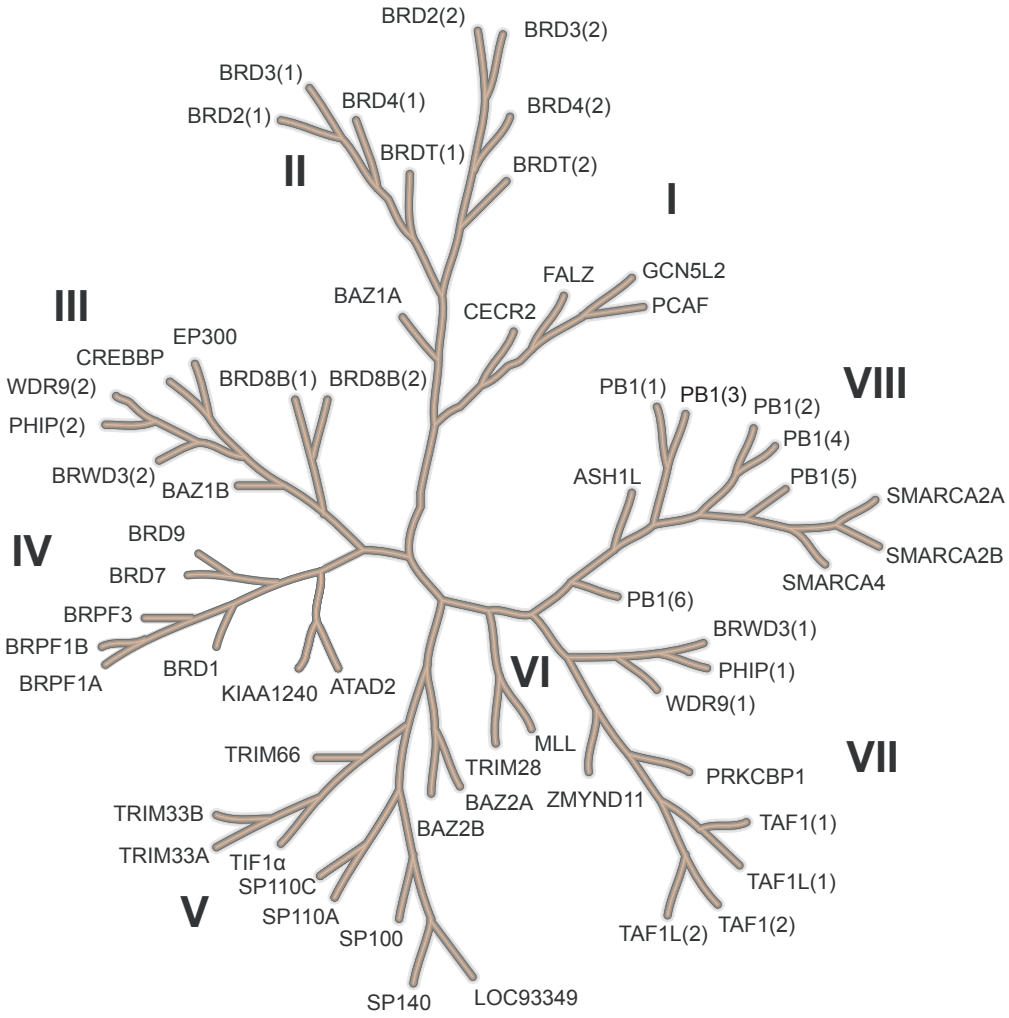


Figure 1.2: Phylogenetic tree of the human BRD family. The different subfamilies are identified by roman numbers (I to VIII).

open chromatin structure. This open state allows polymerases, transcription factors, and other transcriptional complexes access to DNA, leading to gene transcription [70, 71]. Histone acetyltransferases (HATs; the "writers") and histone deacetylases (HDACs, the "erasers") regulate acetylation levels by, respectively, adding and removing the acetyl group from lysine residues on histones. The role of BRDs is then to "read" this epigenetic information and to allow the recruitment of chromatin- and transcription-associated proteins to specific sites on chromatin [72].

The name *bromodomain* derives from the gene that was first observed to contain this evolutionary conserved motif: the *brahma* gene of *Drosophila* [73]. The discovery was made in 1992, yet, while it was known from genetic studies that this was a

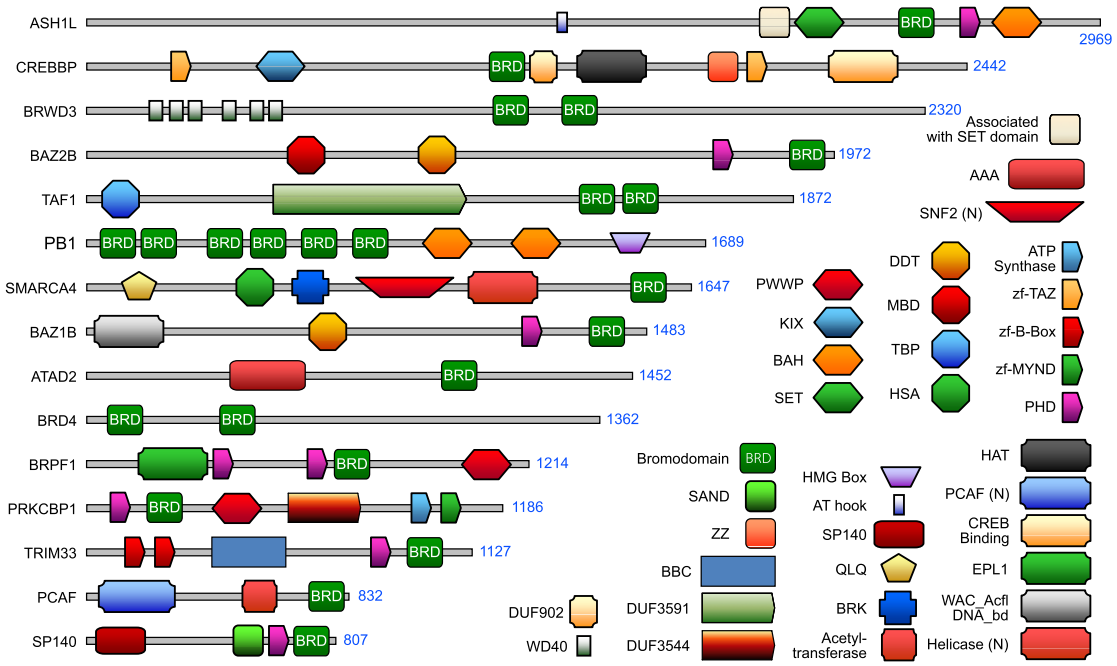


Figure 1.3: Domain organisation of example proteins that contain BRDs. The name and length of the proteins are shown on the bar chart in the eft panel, and the positions the different domains are highlighted as shown in the legend on the right. Reproduced from Filippakopoulos *et al.*[70].

conserved domain in multiple proteins and species [74], its structure and functional significance remained unknown until the end of the decade. In fact, the structure of the BRD of the human transcriptional activator PCAF (p300/CBP associated factor) was solved by NMR in 1999 [75], and shortly after it was shown by crystallography how BRDs could selectively bind to acetylated H4 peptides [76, 77].

BRDs and lysine mediated interactions are important for many cellular processes. In the human genome, 61 BRDs have been found so far, and they are present in 46 nuclear and cytoplasmic proteins. The members of the BRD family are typically clustered into eight groups according to sequence and structural similarity, as shown in Figure 1.2 [70]. BRDs work with numerous transcription-associated proteins, such as chromatin remodellers, histone acetyl- and methyl-transferases, to regulate gene transcription. Most BRD-containing proteins are comprised of multiple domains (Figure 1.3) that work alone or together with one another in order to carry out their complex function. For instance, bromodomain and extra terminal domain (BET)

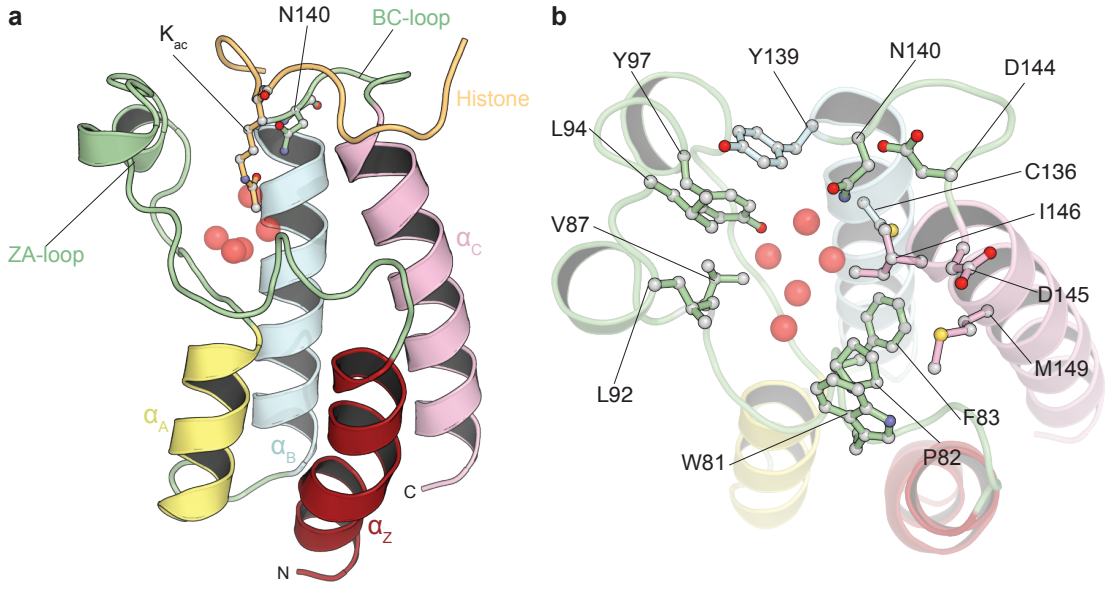


Figure 1.4: Bromodomain fold and acetyl-lysine binding pocket. (a) Cartoon representation of the structure of BRD4(1) bromodomain in complex with an acetylated peptide. Crystallographically observed water molecules are represented as red spheres. (b) BRD4(1) acetyl-lysine binding site with key interacting residues labeled.

proteins contain two conserved N-terminal BRDs, an extra terminal domain (ET) and a C-terminal recruitment domain (CT motif). BET proteins bind to the positive elongation factor P-TEFb via their CT motif, securing the complex to acetylated histone tails via their two BRDs, and resulting in the assembly of the transcriptional machinery [78]. In transcriptional co-activators like the HATs PCAF or GCN5, the BRD module anchors these proteins to an acetylated residue on chromatin so that they can drive further acetylation [72, 79]. The proteins SMARCA2 and SMARCA4 are instead involved in the chromatin remodelling process [70].

All BRDs share a conserved fold formed by a left-handed bundle of four alpha helices (α_Z, α_A, α_B, α_C), as shown in Figure 1.4. The helices are linked by loop regions of variable length: the ZA and BC loops. The BRD binding pocket where Kac binds is sandwiched between these two loops. As represented in Figure 1.4, Kac binds the pocket by hydrogen bonding a conserved asparagine residue at the end of the α_B helix and a conserved water molecule through its carbonyl group, while the methyl moiety of the acetyl group is buried deep in the pocket and sits

just above a conserved network of water molecules.[70] This network of waters is in fact maintaned across many BRDs and it has proved very stable, with the first instance of full water-network displacement being reported only recently [80]. It has also been demonstrated how some BRDs can bind to two acetylated lysine marks, which are recognised by the same BRD module [81], a property shared by all BET BRDs [70]. As can be seen in Figure 1.4b and Figure 1.5, BET BRDs have a conserved WPF motif on the ZA loop. This group of residues forms a bulge that creates two sub-pockets often exploited by small molecules inhibitors; in particular, the cavity between the WPF motif and the rest of the ZA loop is referred to as the "ZA channel", and the one between this motif and the BC loop as the "WPF shelf".

The conserved asparagine residue mentioned above is present in 48 of the 61 human BRDs. However, in eight BRDs this residue is replaced by tyrosine, in four by threonine, and in one by aspartate (Figure 1.5). Growing evidence indicate that some of these atypical BRDs might have functions other than Kac binding alone. For instance, TRIM28 (also called TIF1β) appears to function in tandem with a PHD finger domains as a small ubiquitin-like modifier (SUMO) E3 ligase for transcriptional silencing [82, 83]. The exact role of the atypical BRD module in other proteins, like the methyltransferases ASH1L and MLL, is still unclear [72].

1.2.2 Bromodomains as a therapeutic target

Histone acetylation is a hallmark of open chromatin structure and transcriptional activation. Altered acetylation levels have been linked to a number of diseases, and particularly cancer, where this state promotes the expression of oncogenes resulting in proliferation and tumorigenesis [69, 70, 84]. The disease is often not caused by defects in the BRD module itself. However, a number of oncogenic rearrangements and overexpression of BRDs have been linked to the development of malignancies.

Among oncogenic rearrangements involving BRDs is the fusion of BRD4 or BRD3 with the nuclear protein in testis (NUT). This translocation results in an aggressive form of poorly differentiated carcinoma mainly located in the midline, such as the head or neck [85], thus referred to as NUT midline carcinoma. BRD-NUT

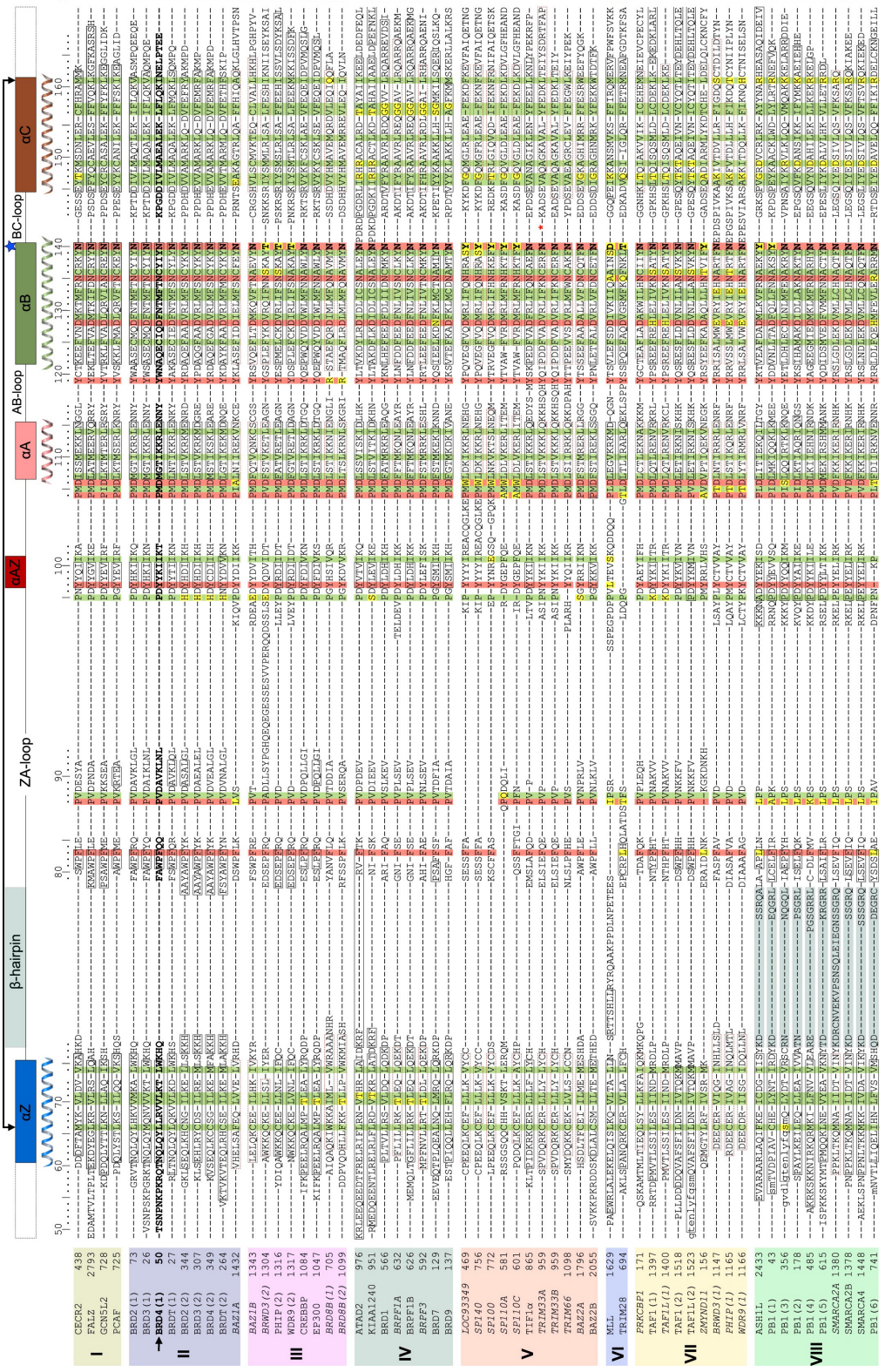


Figure 1.5: Structure-guided sequence alignment of the human bromodomain family. The sequence region used for the generation of the phylogenetic tree is indicated by a black arrow. BRD4(1), in bold and highlighted by an arrow, has been chosen as a reference sequence and the corresponding numbering is shown on top of the alignment. A red asterisk indicates a unique 17 residue insertion between helices B and C in the case of TRIM33A. Reproduced from Filippakopoulos *et al.* [70].

impairs cellular differentiation, and the specific recruitment of CBP/p300 stimulates HAT activity, resulting in hyperacetylated chromatin domains, formation of nuclear foci, and p53 inactivation [86]. The thienotriazolodiazepine compound JQ1 was shown to selectively inhibit the BET BRDs, including BRD4, resulting in epithelial differentiation, tumour shrinkage and survival in BRD4-NUT xenograft mice [87]. Translocations, mutations and overexpression of the CREBBP bromodomain have instead been linked to leukaemia [88, 89], B-cell non-Hodgkin's lymphoma [90], and advanced prostate cancer [91].

Bromodomain overexpression has often been reported in different cancers too. ATAD2 is known to be overexpressed in over 70% of breast tumours, with higher expression levels being correlated with poor survival and recurrence [92]. Its ability to promote cell proliferation was demonstrated to be disabled by mutation of its BRD module [93]. ATAD2B is instead highly expressed in glioblastoma, oligodendroglioma, and breast carcinoma [94]. The BRDT bromodomain has been found to be frequently overexpressed in non-small-cell lung cancer [95] and several other cancers [96]. BRD4 has been one of the most heavily investigated bromodomains thus far, so that its role in cancer is better understood. This BRD is a key regulator of cell cycle control and transcriptional elongation of growth-promoting genes: its role in the recruitment of P-TEFb to transcriptional start sites provides a strategy to targeting cyclin-dependent kinase 9 (CDK9) as a therapy for chronic lymphocytic leukemia [97]. Inhibition of BET BRDs also represses the MYC oncogene, which is often overexpressed in multiple myeloma and medulloblastoma [98, 99].

BRDs have also been investigated as targets for the treatment of inflammation, due to their role in the transcriptional control of proinflammatory cytokines [85]. The BET inhibitor I-BET was shown to disrupt chromatin complexes responsible for the expression of inflammatory genes, resulting in the protection against endotoxic shock and bacteria-induced sepsis in mice [100]. The nuclear factor κB (NF-κB) is a master regulator of the cellular immune response, and activation of proinflammatory genes is linked to activation of NF-κB [101]. Lysine acetylation leads to the

recruitment of the NF-κB co-activators BRD4 and P-TEFb, which results in the transcription of NF-κB target genes. BRD2 mutations have been linked instead to rheumatoid arthritis, with hypomorphic mice being obese and having reduced inflammation in adipose tissue [102].

BET BRDs have also been implicated in viral infection, and specifically in the replication of viral genome and transcriptional regulation of viral proteins [69]. In fact, BET inhibitors have been shown to be able to cause awakening of latent HIV [103–105]. This is important as it suggests a potential strategy for the eradication of the virus, which would stop this infection from being a chronic illness. Pausing of initiated RNA Polymerase II (Pol II) on integrated proviral DNA is a major rate-limiting step in viral gene expression [106, 107]. To overcome this restriction, the viral protein Tat recruits P-TEFb (composed of CDK9 and Cyclin T1 domains) to the HIV long terminal repeats (LTRs). Enzymatically active CDK9 then phosphorylates the C-terminal domain of Pol II and the negative elongation factors, thereby increasing the rate of HIV-1 transcription [108]. The competition of BRD4 for P-TEFb binding might therefore be at the basis of impaired reactivation of HIV latency, which is then removed upon BRD4 inhibition [109]. BRD2 as well appears to be implicated in the regulation of HIV latency [110]. However, it does not do so through interaction with P-TEFb, but via association with the E2F-1 transcription factor, which binds to the HIV enhancer with NF-κB to repress transcription. BRD2 inhibition thus causes the dissociation of the repressor complex from the HIV LTRs [111]. BET BRD inhibition has therefore a dual mechanism of HIV reactivation. BRD2 and BRD4 are also known to bind viral proteins of herpesvirus, human papilloma virus (HPV), and Merkel cell polyoma virus (MCV) [69, 112]. For instance, BRD4 was found to be essential for HPV transcription and replication [113].

1.2.3 Current state of bromodomain inhibitors

Novel BRD inhibitors are finding broad application in medicine and basic biological research, with a few compounds currently undergoing clinical trials [69]. Since the

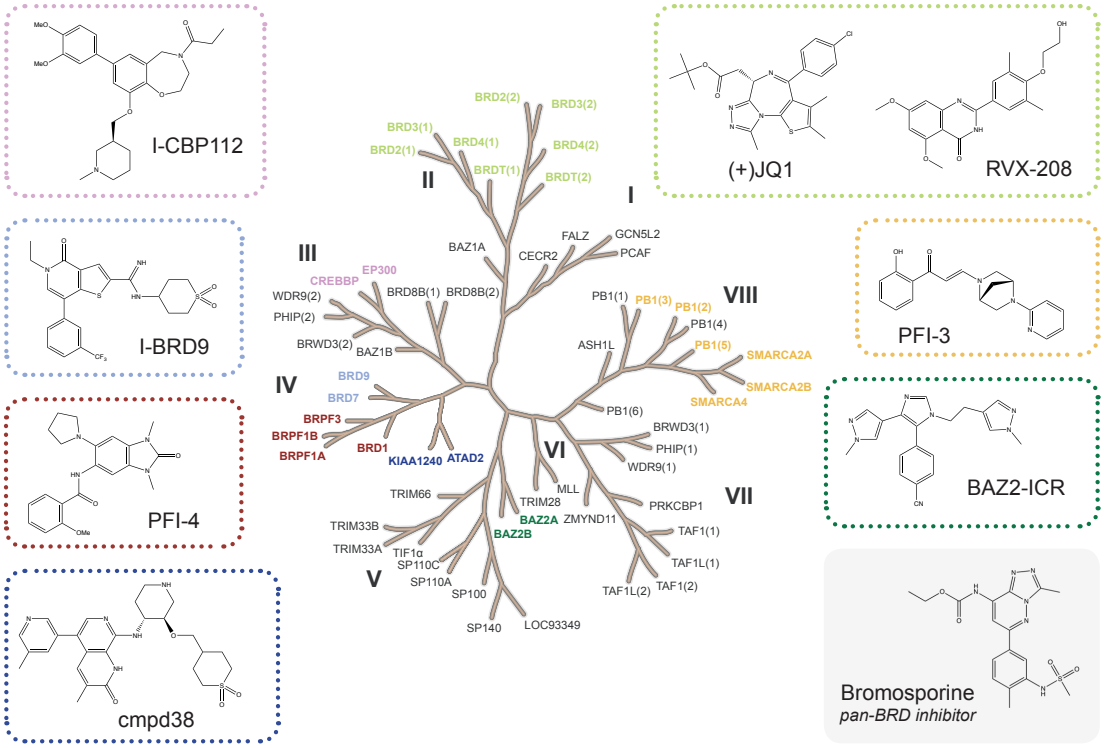


Figure 1.6: Example bromodomain inhibitors by target family.

development of the first BRD inhibitor for PCAF and CREBBP by the Zhou group [114, 115], but in particular the publication of two potent pan-BET inhibitors in 2010 [87, 100], BRDs attracted large interest as a promising therapeutic target and many new ligands have been discovered in the last five years. Here, I provide a brief overview of the most notable chemical scaffolds available, as well as the current state of BRD inhibitors undergoing clinical trials.

BET bromodomain inhibitors

Thanks to a clear link to cancer pathogenesis and demonstrated druggability [87, 100], the BET family of BRDs has been the most intensively studied in the last few years [116]. BET proteins comprise BRD2, BRD3, BRD4 and the testes specific BRDT, and are characterised by the presence of two tandem BRDs, an N-terminal and a C-terminal (or first and second) BRDs. These first and second BRDs are more similar among themselves than between the two found in a single BET protein; e.g. BRD2(1) is closer to BRD4(1) than to BRD2(2) [70]. As mentioned previously,

BRD4 in particular is known to associate with P-TEFb [78], which results in transcriptional elongation, and inhibition of BRD4 binding to histones appears to have a therapeutic effect in cancer, inflammation and HIV-1 infection [69].

Among the most successful BET inhibitors are diazepine-based ligands. I-BET762 (also known simply as I-BET, or GSK525762) was discovered through phenotypic screening during a campaign aiming at the identification of small molecule up-regulators of apolipoprotein A1 (ApoA1), a constituent of high-density lipoproteins involved in the regulation of cholesterol levels and an arteriosclerosis target [117]. Enhancers of ApoA1 expression were initially identified with a luciferase reporter-based screen, and then a chemoproteomic approach was used in order to determine the targets of the ligand, which were found to be BRD2, BRD3 and BRD4. I-BET762 binds to BRD4 with an affinity of 55 nM as determined by ITC, and was shown to displace a tetra-acetylated histone H4 peptide through fluorescence resonance energy transfer (FRET). The ligand (+)-JQ1 (Figure 1.6) was developed by Filippakopoulos *et al.* [87] and was based on the thienotriazolodiazepine core firstly described in a patent by Yoshitomi Pharma (later Mitsubishi Pharma) in 1996, when compounds based on this scaffold were proposed for the treatment of autoimmune diseases like ulcerative colitis and Crohn's disease [118, 119]. However, it was only in 2009 that BET BRDs were identified as the targets of these ligands [120]. (+)-JQ1 was shown to be selective for BET BRDs and its affinity for BRD4(1) and BRD4(2) was determined by ITC to be 50 nM and 90 nM, respectively [87]. Another thienotriazolodiazepine ligand very similar to (+)-JQ1 is OTX015, which was too developed by Mitsubishi Pharma and later licensed to OncoEthix, and is currently in clinical trials for different tumours [121].

Another privileged moiety used for mimicking Kac in BET BRDs is 3,5-dimethyloxazole, which has been discovered independently by different groups [122–127], and has been adopted for targeting the CREBBP bromodomain too [122, 128, 129]. The Structural Genomics Consortium (SGC) discovered this scaffold through fragment-screening, which was then turned into a more potent set of ligands by fragment growth into the ZA channel and WPF shelf [122]. Some of the most

effective compounds also caused reduced viability of lung adenocarcinoma cell lines [123]. A fragment approach was also used by Constellation Pharmaceuticals for the discovery of an amino-isoxazole fragment that was later turned into a ligand with IC_{50} of ∼260 nM, which is an analogues of (+)-JQ1 where the methyl-triazole ring is replaced by a methyl-isoxazole [130]. Through a structure activity relationship (SAR) study, the same group modified further the ligand in order to obtain the drug-like compound CPI-0610, which was shown to reduce MYC expression and tumor growth in a mouse xenograft model, as well as acceptable toxicity in rat and dog [131]. Scientists at GlaxoSmithKline (GSK) used instead a phenotypic screen for the identification of I-BET151, which was developed in parallel to I-BET762 [124]. I-BET151 binds to BRD4 (both first and second bromodomains; residues 1-477) with a K_d of 100 nM as measured by surface plasmon resonance (SPR), and showed efficacy against human and murine MLL-fusion leukaemic cell lines [124], in addition to an improved pharmacokinetic profile as compared to the triazolodiazepine scaffold [126].

Among other interesting BET inhibitors is RVX-208 (Figure 1.6), a ligand with a quinazolinone core originally discovered by Resverlogix Corp. as an ApoA1 upregulator [132, 133]. Similarly to I-BET762, it was only later discovered that the compound acts via binding to BET BRDs, and Picaud *et al.* [134] also showed by ITC that RVX-208 is slightly selective for the second BET BRDs. For instance, the ligands binds to BRD4(1) with a K_d of 1.1 μM and to BRD4(2) with with a K_d of 140 nM [134]. RVX-208 is currently in clinical trials for the treatment of atherosclerosis, diabetes, and cardiovascular diseases [133, 135]. Another quinazolinone based compound is PFI-1, which was identified by Fish *et al.* [136] via a fragment screen. The inhibitor showed antiproliferative effect in leukemic cell lines, as well as G1 cell-cycle arrest, downregulation of MYC expression, induction of apoptosis and differentiation in primary leukemic blasts. Interestingly, cells exposed to PFI-1 also showed downregulation of Aurora B kinase, a well-established oncology target [137]. Finally, particularly intriguing are also the recently discovered dual BRD-kinase inhibitors, which open the possibility of a poly-pharmacology approach for the

treatment of cancers that are affected by multiple pathways. Ciceri *et al.* [138] screened a library of about 600 kinase inhibitors against BRD4(1) and found that nine inhibitors strongly inhibited binding of a polyacetylated histone H4 peptide. Most notably, the compounds BI-2536 and TG-101348, which are clinical PLK1 and JAK2/FLT3 kinase inhibitors, were found to have nanomolar activity against BRD4(1), with K_d of 37 nM and 164 nM (ITC), respectively. The structures of the complexes were solved by X-ray crystallography, showing how the ligands act indeed as Kac mimetics. In addition, disruption of BRD4 chromatin binding was demonstrated by a fluorescence recovery after photobleacing (FRAP) assay. In a separate and concurrent study, Ember *et al.* [139] used a kinase library of similar size (∼600 ligands) and a co-crystallization screen to discover 14 kinase inhibitors that are able to bind BRD4. Encouragingly, the most potent compounds identified corresponded to the ones also found by Ciceri *et al.* [138]. The analysis of the 14 crystal structures obtained revealed that the kinase inhibitors hinge-binding motifs can adopt different binding modes in the BRD binding pocket, suggesting BRDs as potential off-targets to many kinase inhibitors [139].

Other bromodomain inhibitors

A number of BRDs outside the BET family are now targeted by different ligands, despite not having received the same extent of attention as the BET BRD inhibitors. In particular, a number of ligands have been developed for CREBBP. Early work on these inhibitors was carried out by the Zhou group, which identified 14 N-acetylated fragments binding to CREBBP, among which two were shown to block the interaction between the BRD and p53 in a concentration dependent manner[140]. The same laboratory also identified azobenzene binders through an NMR screen, which were then optimised and the most active compound achieved an IC_{50} of 19 μM in a cellular assay [141]. More recently, Rooney *et al.* [142] developed potent submicromolar Kac mimics targeting CREBBP based on dihydroquinoxalinone fragments. The side chain of this series was shown to bind in an induced-fit pocket forming a cation-π interaction with R1173. In addition, the highest affinity

compounds showed good activity in FRAP assays in U2OS cells. As mentioned previously, 3,5-dimethyloxazole ligands were also found to bind CREBBP [122, 130]. In fact, this moiety was used to obtain the compound SGC-CBP30, which is a low nanomolar inhibitor of CREBBP and the closely related EP300 bromodomain with 40-fold selectivity over BRD4(1) [129]. The inhibitor I-CBP112, developed by the SGC and GSK, is instead based on benzoazepine scaffold, where the Kac mimicking moiety is a carbonyl group branching off the oxazepine ring [143]. The compound selectively targets the CREBBP and EP300 bromodomains, and its use on mouse leukemic cell lines resulted in impaired colony formation and induced cellular differentiation without significant cytotoxicity [143].

A number of ligands are now available also for the BRPF1 bromodomain. Demont *et al.* [144] reported the discovery of a selective series of BRPF1 inhibitors based on benzimidazolone scaffold, with selectivity over BRPF2 (i.e. BRD1) and BRPF3 and BET BRDs. Palmer *et al.* [145] recently reported the discovery of IACS-9571 instead, a potent dual TRIM24-BRPF1 bromodomain inhibitor with low nanomolar potency *in vitro* and cellular activity too (EC_{50} of 50 nM). The ligand is selective against a panel of other BRDs, but it is not particularly selective within the BRPF family. Zhu and Caflisch [146] performed a fragment screen by high-throughput docking that resulted in the identification of six chemotypes binding to BRPF1. The crystal structures of BRPF1 in complex with 19 compounds were also resolved, showing conservation of the overall structure of the acetyl lysine binding site. Finally, the SGC has reported on their website (`http://www.thesgc.org/chemical-probes/PFI-4`) the discovery of PFI-4 (Figure 1.6), a compound binding to BRPF1 with K_d of 13 nM (ITC), in collaboration with Pfizer.

Due to being associated to a wide variety of cancers, the BRD of ATAD2 has been recently targeted by fragment screening, despite being considered a challenging target. Harner *et al.* [147] used an NMR-based screen and identified a number of low affinity chemotypes, which could be optimised into more potent ligands. Demont *et al.* [148] have instead discovered two micromolar binders of ATAD2 through a fragment screen using a focussed library. Bamborough *et al.* [149] have reported the

structure-based optimisation of a series of naphthyridones from micromolar leads to potent nanomolar inhibitors of ATAD2 with 100-fold selectivity over BET BRDs.

The BRDs of proteins BAZ2A and BAZ2B have also been successfully targeted recently. The Institute of Cancer Research (ICR), in collaboration with the SGC, described in 2015 the discovery of the potent and selective chemical probe BAZ2-ICR (Figure 1.6) to target BAZ2A and BAZ2B bromodomains (K_d of 109 nM and 170 nM for BAZ2A and BAZ2B respectively, as measured by ITC), which mimics acetylated lysine through a methyl-pyrazole group [150]. In the same year, Chen *et al.* [151] reported the development of GSK2801, a potent, selective, as well as cell active inhibitor of BAZ2A/B bromodomains. The compound was shown to bind with K_d of 257 nM and 136 nM for BAZ2A and BAZ2B, respectively, and crystal structures demonstrated the canonical Kac competitive binding mode. GSK2801 also showed cellular activity in a FRAP assay where BAZ2A was displaced from acetylated chromatin. In addition, this ligand displayed a reasonable pharmacokinetic profile in mice.

BRD7 and BRD9, from the BRD family IV like BRPF1 and ATAD2, are less well studied as compared to the BET BRDs. However, there have been multiple inhibitors developed in the last two years that will prove useful to probe their function. Two of these ligands, one a quinolone-fused lactam (LP99) [152], and the other one an imidazopyridine [153], bind to BRD9 with low nanomolar affinities (99 nM and 68 nM, respectively; ITC) as well as to BRD7, while showing little to no activity against other BRDs. Also a series of ligands based upon 9H-purine scaffolds have been recently reported to bind BRD9, though with a more limited selectivity over BET BRDs [154]. Finally, Theodoulou *et al.* [155] reported the discovery of the thienopyridine I-BRD9 (Figure 1.6), which has low nanomolar affinity for BRD9 and >200-fold selectivity for BRD7, as well as >700-fold selectivity over BET BRDs. I-BRD9 was used to identify genes regulated by BRD9 in a cell-based assay and it revealed a number of genes involved in oncology and immune response pathways [155].

A number of potent ligands binding to the PB1 and SMARCA bromodomains with a distinctive binding mode that displaces the conserved network of waters have recently been reported. Fedorov *et al.* [80] discovered that salicylic acid could bind to the BRDs of family VIII and displace the four conserved waters at the bottom of the binding pocket. The fragment was then expanded into the potent sub 100 nM inhibitor PFI-3 (Figure 1.6), which is selective for the PB1 and SMARCA bromodomains. Soon after, Gerstenberger *et al.* [156] reported the development of another inhibitor based on the same salicylic acid fragment hit. Then, Sutherell *et al.* [157] discovered a ligand based on a chlorinated quinazolinone scaffold that bind to the PB1 and SMARCA bromodomains with selectivity over other BRD families.

Atypical BRDs in which the conserved asparagine residue is replaced by a tyrosine or threonine have been elusive targets. Nonetheless, Cox *et al.* [158] have recently employed a fragment screening strategy where a chemically-poised library was screened using X-ray crystallography as the primary technique in order to identify PHIP(2) bromodomain hits. 406 fragments were tested and four were found to bind the BRD. As the library was design for rapid synthetic expansion by parallel chemistry, three distinct series of fragments (based on thioureas, N-benzylamides, and oxazoles scaffolds) were quickly derived from the initial hits. The resulting molecules were in the range of 100 μM to 300 μM, but had good ligand efficiencies and will provide a good starting point for fragment growth and hit optimisation.

Finally, the SGC discovered a broad-spectrum BRD inhibitor, which was named *bromosporine* (Figure 1.6) in reference to the pan-kinase inhibitor *staurosporine*. Bromosporine binds across all BRD families with affinities that range from μM to single digit nM [159]. The compound will be useful as a tool for probing the function of BRDs for which no other ligands are known.

Clinical applications

While the design and development of BRD inhibitors is interesting from a SBDD perspective, their therapeutic value will only be confirmed through clinical trial. With a number of studies showing promising effects of BET BRD inhibition, it

Table 1.1: Bromodomain inhibitors currently in clinical trial. Data initially taken from reference [72] and updated as of August 2016 based on information available at `https://clinicaltrials.gov`.

Compound	Company	Indication	Stage (status)	ClinicalTrials.gov ID
I-BET762	GlaxoSmithKline	NUT midline carcinoma and other cancers	Phase I (recruiting)	NCT01587703
		Relapsed, refractory Haematologic malignancies	Phase I (recruiting)	NCT01943851
OTX015	OncoEthix	Acute myeloid leukaemia and diffuse large B-cell lymphoma	Phase I (ongoing, not recruiting)	NCT01713582
		Advanced solid tumours	Phase IB (ongoing, not recruiting)	NCT02259114
		Acute myeloid leukaemia	Phase IB/II (withdrawn)	NCT02303782
		Glioblastoma multiforme	Phase IIA (terminated)	NCT02296476
	Merck & Co.	Advanced solid tumours	Phase I (recruiting)	NCT02698176
		Haematologic malignancies	Phase I (recruiting)	NCT02698189
CPI-0610	Constellation Pharma	Progressive lymphoma	Phase I (recruiting)	NCT01949883
		Multiple myeloma	Phase I (recruiting)	NCT02157636
		Acute leukemia, myelodysplastic syndrome, myelodysplastic/myeloproliferative neoplasms	Phase I (recruiting)	NCT02158858
TEN-010	Hoffmann-La Roche	Acute myeloid leukaemia and myelodysplastic therapeutics syndrome	Phase I (recruiting)	NCT02308761
		Solid tumours	Phase I (recruiting)	NCT01987362
RVX-208	Resvelogix Corp.	Dyslipidaemia, atherosclerosis, acute coronary syndrome, cardiovascular disease	Phase II (completed)	NCT00768274
		Atherosclerosis and coronary artery disease	Phase II (completed)	NCT01058018
		Coronary artery disease	Phase II (completed)	NCT01067820
		Coronary artery disease and dyslipidaemia	Phase II (completed)	NCT01423188
		Diabetes	Phase II (completed)	NCT01728467
		Dyslipidaemia and coronary artery disease	Phase II (terminated)	NCT01863225
		High-risk cardiovascular disease	Phase III (recruiting)	NCT02586155

is not surprising that pharmaceutical companies have invested efforts into the progression of these inhibitors into the clinic.

Table 1.1 summarises all ongoing clinical trials involving BRD inhibitors as of August 2016. I-BET762 is undergoing two Phase I trials, one for NUT midline carcinoma and another one for refractory haematologic malignancies. OTX015 is undergoing a number of Phase I trials for the treatment of leukaemia, solid tumours, and haematologic malignancies, while its Phase II trial for the treatment of glioblastoma multiforme was terminated. The compound is now being trialled by Merck & Co., which acquired OncoEthix in late 2014. TEN-010 is another BET BRD inhibitor currently in Phase I for the treatment of leukaemia and solid tumours. The compound was initially discovered by Tensha Therapeutics, which was then acquired by Hoffmann-La Roche in January 2016, who is now the official sponsor. RVX-208 is, at the moment, the compound that is furthest along the clinical development pipeline, with Phase II studies having been completed for a number of diseases related atherosclerosis, coronary artery disease, dyslipidaemia, and diabetes mellitus. Moreover, Phase III trials were begun in late 2015 for the

treatment of high-risk cardiovascular disease patients with type 2 diabetes mellitus.

1.3 Ligand binding assays

A variety of biochemical binding assays are available to determine the affinity of a ligand for a protein of interest. More specifically for BRDs, a few assays have been proved particularly useful. In high-throughput screens, a thermoshift and a bead-based proximity assay have been widely used. The thermal shift assay, also known as *thermofluor* or *differential scanning fluorometry*, is based on the principle of higher thermal stability for proteins in complex with a ligand than in apo form. A fluorescent dye such as SYPRO Orange is added to the protein solution, and the melting temperature (T_M) of the macromolecule is determined by measuring fluorescence while raising the temperature. As the dye binds nonspecifically to hydrophobic surfaces, and water strongly quenches its fluorescence, the recorded signal is proportional to protein unfolding. When a strong small molecule binder associates with the target protein, it increases the protein T_M. This shift in melting temperature can then be used for the identification of ligands binding the protein of interest [160]. Another popular approach used for screening large compounds libraries against BRDs is a bead-based proximity assay, known by the commercial name of Amplified Luminescence Proximity Homogeneous Assay (AlphaScreen^{TM}) [161]. This assay works by detecting the proximity of a donor and an acceptor bead, conjugated to the binding partners of interest. The acceptor bead emits light when it is stimulated, and stimulation occurs when it is in proximity of the donor bead. The light emitted can then be detected using a photometer. In their assays, the SGC has typically used donor beads coated with streptavidin, allowing conjugation to biotinylated peptides containing Kac, and acceptor beads coated with a nickel chelator for the conjugation to His-tagged BRDs [162]. Laser excitation at 680 nm of a photosensitizer within the donor bead converts ambient oxygen to singlet oxygen. If the acceptor bead is within approximately 200 nm of the donor bead, the singlet oxygen initiates a chemiluminescent reaction in the acceptor bead, which emits light in the 520 to 620 nm range. BRD inhibitors

disrupt the interaction between the target BRD and the peptide conjugated to the donor bead, thus causing the singlet oxygen to return to its ground state without a signal being emitted by the acceptor bead [161, 162]. Both these screens are suitable to high-throughput campaigns, yet not for the accurate determination of binding free energies, despite the strength of the signal often correlates with K_d. AlphaScreen can accurately measure IC_{50} values, but these depend on the affinity of the displaced peptide. For the more accurate determination of binding free energies, biophysical methods are considered to be more reliable. In particular, isothermal titration calorimetry (ITC) tends to be considered the gold standard in protein-ligand binding affinity determination and it has been used to measure the binding constant of many lead compounds and optimised chemical probes [163–167]. In this thesis, the vast majority of the computationally derived binding free energies are assessed in comparison to high-quality calorimetry data. For this reason, I will provide a brief introduction of the concepts behind ITC.

1.3.1 Isothermal titration calorimetry

The use of calorimetry for the study of biological processes dates back to the late 18^{th} century when Lavoisier and Laplace used an ice calorimeter to measure the metabolic heat produced by guinea pigs [168]. More modern titration calorimetry was first introduced in the 1960s by Christensen and Izatt, who described it as method for the simultaneous determination of association constants and enthalpies in acid-base equilibria and metal ions complexation reactions [169, 170]. In 1978, Beaudette and Langerman reported one of the first calorimetric binding studies of a biological system, in which they measured the ΔH and ΔG for the binding of adenosine diphosphate (ADP) to bovine liver glutamate dehydrogenase [171]. The year after, Langerman and Biltonen published an extensive review on microcalorimeters for biological chemistry that included a discussion on available instrumentation, experimental design, data analysis and interpretation [172]. Despite this, commercial titration calorimeters for the study of biological systems was not available for another decade [166]. ITC is nowadays widely used for the study of molecular association

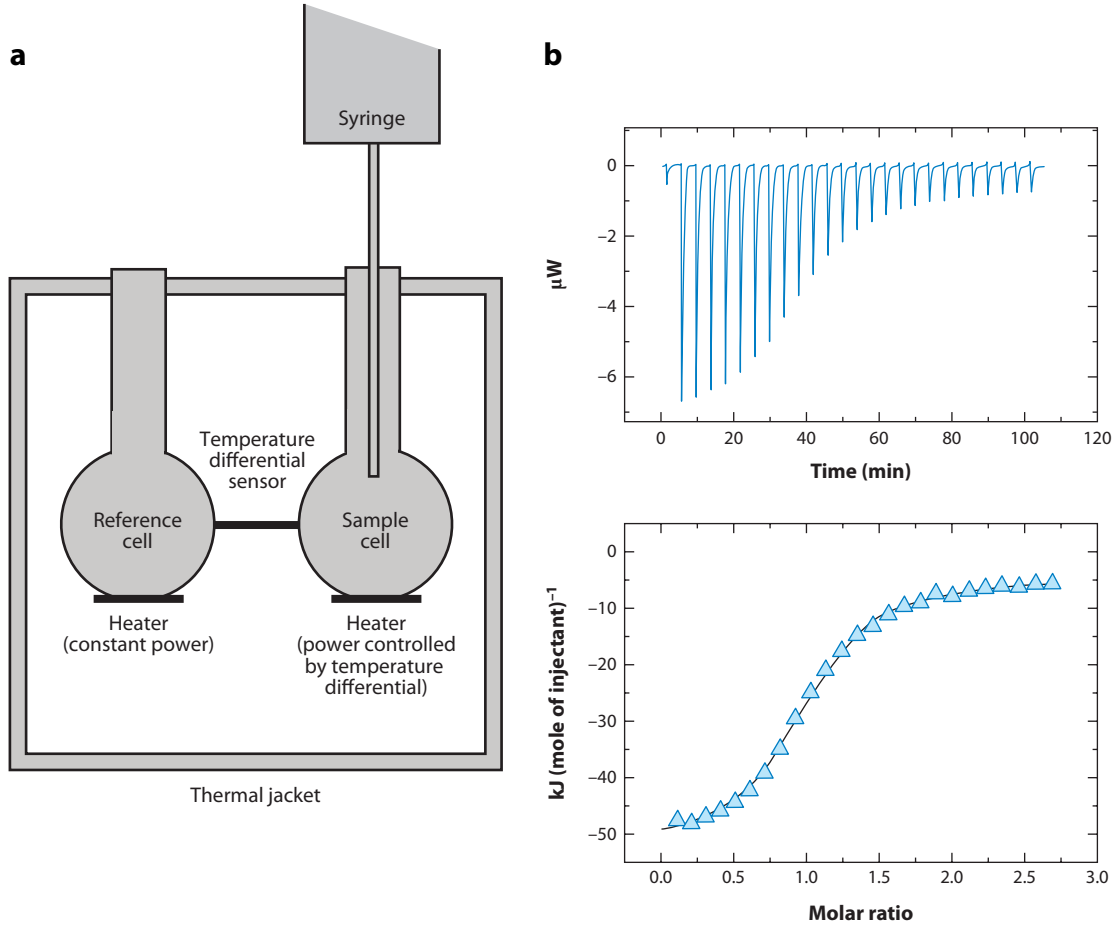


Figure 1.7: Schematic representation of ITC and its data. a) Typical design of a power-compensating ITC. b) Data obtained from an ITC experiment; at the top the raw data (applied power as a function of time), and at the bottom the integrated heats of injection with the non-linear curve fit. Reproduced from Chodera & Mobley [22] with permission granted by Annual Reviews.

by many laboratories using robust commercial instrumentation that makes the technique available outside the small number of groups that focus on the development of the method [173]. Currently available calorimeters can measure heats as small as 0.1 μcal and dissociation constants as little as 10^{-9} M [166]. The upper and lower K_d detection limits of ITC are ultimately dependant on the sensitivity of the instrument to the heat signal, which is proportional to the molar enthalpy of the reaction and the amount of protein-ligand complex formed.

The use of ITC for the determination of binding constants has a number of advantages over other binding assays. In fact, ITC can be used with spectroscopically

silent molecules, so that a chromophores or fluorophores are not needed; opaque, turbid, or heterogeneous solutions, such as a cell suspension, can be used; and a range of biologically relevant conditions, with varied pH, temperature, and salt concentration, can be tested [166]. Overall, if carried out carefully, the ITC experiment is able to return accurate binding affinities that are not affected by the presence of large molecular tags, as in the AlphaScreen^{TM} assays, or by extreme conditions and an indirect measure of binding as in thermoshift assays. In addition, with a single experiment it is possible to obtain ΔH, ΔS, and ΔG.

Figure 1.7 schematises the basic principle behind an ITC experiment. Modern calorimeters are based on a power compensation design for the measurement of reaction heat [22, 166, 173]. As shown in Figure 1.7a, the calorimeter cell is maintained at constant temperature through the use of a thermal jacket (hence the prefix *isothermal*). A reference cell and the sample cell are then in equilibrium and have the same temperature. When the syringe injects an aliquot of ligand solution into the protein solution, or vice-versa, the chemical reaction takes place and any heat generated is detected as a temperature difference between the reference and sample cells. Both cells are connected to heaters; however, while the reference cell is subjected to constant power, the power applied to the sample cell quickly adapts to the temperature difference detected so that the temperature is brought back to the value that was set to be constant. This power compensation (Figure 1.7b, top) is the raw signal detected by the instrument. Then, integrating the signal over time, one can recover the heat associated with each injection (Figure 1.7b, bottom), which is typically plotted against the molar ratio of the two reactants [22, 166, 173].

The calorimetry experiment is designed with excess of reactant "A" in the cell (e.g. protein), so that during the first injections all reactant "B" present in the aliquots injected by the syringe (e.g. ligand) is consumed. As the experiment continues, the ratio of free protein and ligand versus complex ("A" and "B" versus "AB") is determined by the equilibrium constant of the reaction. At the end of the experiment, there is an excess of "B" so that the injection of additional aliquots from the syringe does not result in any heat, given that all binding sites are occupied.

Before the start of the experiment, the concentrations of both protein and ligand are chosen so to be able to obtain heats large enough to be measured, while at the same time avoiding a step-function shaped thermogram, which is often also the result of particularly strong binders ($K_d < 1$ nM) [22, 166, 173]. From the ITC experiment it is possible to directly obtain binding enthalpy (ΔH), dissociation constant (K_d), and stoichometry (n) of the reaction, based on a non-linear fit of the titration data. ΔH is obtained as the amount of heat released per mole of ligand bound from the difference in the heat signal between the initial and final peaks. The molar ratio at the centre of the binding isotherm returns the reaction stoichiometry; n might deviate from the expected stoichiometry as the result of concentration errors of the titrand being absorbed into this term. K_d is instead given by the slope of the curve at the inflection point. ΔG is then obtained from the dissociation constant as $\Delta G = RTln(K_d/c^{\ominus})$, where R is the gas constant, T the absolute temperature, and $c^{\ominus} = 1$ mol/L, while the binding entropy is obtained as $T\Delta S = \Delta H - \Delta G$ [22, 166, 173].

Sources of error in ITC

A number of issues might arise during the ITC experiment. A typical problem is the presence of heat of mixing caused by a mismatch between the buffer composition in the the syringe and sample cell. This problem can be avoided by dialysis of the protein solution, and by then using the dialysate as the solvent for the ligand. However, heat caused by the dilution of the small molecule itself can still be problematic. In this circumstance it is possible to observe residual heat at the end of the experiment, in the final peaks, when all binding sites are expected to be saturated. In this case a ligand blank titration is usually carried out, so to be able to subtract the dilution heat from the heat due to the (usually) exothermic association reaction. In theory, the same problem exists for the macromolecule, however the heat of dilution of the protein is usually negligible. An instrumental blank experiment where no ligand or protein is present (i.e. titrating buffer into

buffer) is also sometimes performed before carrying out the actual ITC measurement, in order to test the correct functioning of the instrument [166].

Apart from the above mentioned challenges, it has been shown how concentration errors can largely affect the results obtained from ITC [174]. In fact, the accurate measurement of ligand and protein concentrations can be challenging, considering issues such as organic molecule solubility and degradation and protein stability, as well as the limitations of approaches such as UV spectrophotometry. For instance, if the ligand aggregates it is not available anymore for binding, resulting in a concentration that is effectively lower than the one measured. Furthermore, non-specific binding can be a source of error, which would too reduce the fraction of available ligand in solution. The non-specific binding of a ligand to a protein might also manifest itself in the measurement of an unexpected non-integer stoichiometry n. On the other hand, the binding to N specific binding sites should result in an integer $n = N$. In the inter-laboratory assessment by the Molecular Interaction Research Group of the Association of Biomolecular Resource Facilities (ABRF-MIRG'02), identical samples for the 1:1 binding of 4-carboxy-benzensulfonamide (CBS) to bovine carbonic anhydrase II (CA II) were distributed to 14 different laboratories with ITC expertise. The groups were then asked to report the measured association constants and binding enthalpies. The reported ΔH and K_d showed large standard deviations over independent measurements: the mean and standard deviation of the binding enthalpy was -10.4 ± 2.5 kcal/mol, and $(1.00 \pm 0.22) \times 10^6$ M^{-1} for the association constant [174]. Myszka and co-workers found that the main source of error in the study was the failure to accurately determine the titrant (the CBS ligand in this case) concentration. In fact, even though concentration errors in the titrand (here the protein CA II) are absorbed into the stoichiometry parameter, n, in typical fitting procedures, the concentration of the titrant is assumed to be known exactly [22]. This assumption results in a direct relationship between the concentration error and the error in K_a (or K_d) and ΔH, so that a 10% error in the titrant concentration will result in a 10% error in the measured association constant and binding enthalpy. It is important to note that, thanks

to the logarithmic relationship between K_d and ΔG, the estimate of binding free energy tends to be quite robust despite the above discussed challenges. In fact, at 298.15 K, an error of 20% in K_d translates into the relatively small error of 0.11 kcal/mol in ΔG; if we assume K_{d_1} is the correct binding affinity, and $K_{d_2} = 1.2 \cdot K_{d_1}$ (that is, 20% larger), then the difference between the two binding free energies is: $\Delta\Delta G = RTln(K_{d_2}/K_{d_1}) = RTln(1.2) = 0.11$ kcal/mol, where $T = 298.15$ K and $R = 1.987 \cdot 10^{-3}$ kcal mol^{-1} K^{-1}. Indeed, in the ABRF-MIRG'02 study, while the standard deviation in ΔH was of 2.5 kcal/mol ($\sim$23% in relative terms), the standard deviation in ΔG was of 0.13 kcal/mol, a relative error of only $\sim$1.6% compared to the mean measured ΔG of -8.18 kcal/mol [174]. Note that this inter-laboratory error is still larger than what typically reported when only a single ITC experiment is performed (which tends to be about $0.02 - 0.05$ kcal/mol), as the error shown is related to the curve fitting procedure only. For this reason, in this thesis I will use 0.13 kcal/mol as the uncertainty of the binding free energy derived by ITC. Since $T\Delta S$ is then derived as $\Delta H - \Delta G$, the robustness of the binding free energy, coupled with the uncertainty in the binding enthalpy, can result in apparent entropy-enthalpy compensation that is caused simply by large and opposite errors in the enthalpic and entropic contributions [22].

1.4 Aims and motivation

This thesis focuses on the use of computer simulation for the study of protein-ligand binding. More specifically, the aim of the work presented here is to validate novel free energy methods for their application to drug design. As discussed in this introductory chapter, free energy calculations based on statistical mechanics principles are becoming increasingly accessible thanks to increasing levels of computer power, with relative binding free energy (RBFE) calculations that have already started being commercialised by software companies and adopted by the pharmaceutical industry in lead development projects. On the other hand, absolute binding free energy (ABFE) calculations based on alchemical cycles have been so far confined to testing on engineered model systems with limited pharmacological interest. Despite

the computational cost of these calculations remaining high, such that it hinders a large scale analysis of performance, it is important to start evaluating the expected prospective value of these approaches by careful retrospective testing. This will provide not only an estimate of their accuracy and precision, but also allow the identification of sources of errors, which might indicate the challenges that need to be overcome in order to advance these techniques. For these reasons, BRDs were chosen as a computationally approachable yet pharmacologically highly relevant target for the testing of ABFE calculations. In Chapters 3 and 4, I will test the performance of an ABFE protocol based on a standard implementation of alchemical transformations for the prediction of the binding affinities of different drug-like molecules binding to multiple BRDs. In Chapter 5 I will then compare the performance of the ABFE approach, in terms of correlation with experiment, against a well-established and widely used implicit solvent end-point method. Finally, in Chapter 6, I will apply a free energy method based on Monte Carlo simulations in the grand canonical ensemble for the estimation of water stability in BRD cavities, in order to provide guidance to medicinal chemists on which waters and water networks might be more easily displaced.

Thermodynamics is a funny subject. The first time you go through it, you don't understand it at all. The second time you go through it, you think you understand it, except for one or two small points. The third time you go through it, you know you don't understand it, but by that time you are so used to it, it doesn't bother you any more.

— Arnold Sommerfeld

2 Theory and Methods

Contents

2.1 Molecular mechanics 36
2.1.1 Empirical force fields 36
2.1.2 Water models 40
2.2 Energy minimisation 41
2.3 Molecular dynamics simulations 42
2.3.1 Finite difference methods 44
2.3.2 Constraints 46
2.3.3 Ensembles and temperature/pressure coupling 46
2.3.4 Periodic boundary conditions 50
2.3.5 Treatment of long-range interactions 51
2.4 Monte Carlo simulations 54
2.4.1 Grand Canonical Monte Carlo simulations 56
2.5 Replica exchange 57
2.6 Free energy calculations 59
2.6.1 Alchemical methods 61
2.6.2 Implicit solvent end-point methods 66
2.6.3 Grand canonical integration 69

The work presented in this thesis is based on the use of a number of computer simulation techniques. In this section the fundamental concepts behind these approaches are briefly reviewed. An in-depth description of all methods is out of the scope of this chapter, as it would require more than the space allowed for the doctoral thesis, and excellent textbooks have been written on these subjects by Leach [175],

Tuckerman [176], and Pohorille & Chipot [177]. Hence, I will focus on a high-level description of the core techniques that have been used, as well as some of the most fundamental aspects of the theory and algorithms behind their implementation.

2.1 Molecular mechanics

The behaviour of any molecule can be described by the interactions between its atoms, including both nuclei and electrons. The exact description of the atomic electronic structure requires a quantum mechanical treatment (QM) of the system through the Schrödinger equation. Thanks to the Born-Oppenheimer approximation, computation of the energy and wave function of a molecule is simplified, yet still very expensive computationally. Even if some electrons are ignored, as in semi-empirical approaches, problems relating the dynamics of large biological molecules are still out of reach for current levels of computer power. Molecular mechanics (MM) offers a computationally cheap treatment of atomic particles that is based on classical physics and the validity of a number of assumptions. The first of these is again the Born-Oppenheimer approximation, as MM ignores the motions of electrons and derives the energy of the system only as a function of nuclear positions. It follows that from these models it is not possible to derive any property that depends upon the electronic distribution in a molecule. A second assumption is that simple functions are sufficient to describe interactions within a system to an acceptable degree of accuracy. Finally, parameter transferability is a key attribute of MM models, as it enables us to derive parameters for a limited number of test cases, yet apply them to a much wider number of problems [175].

2.1.1 Empirical force fields

Molecular mechanics models are called "force fields", and most of the currently popular ones are characterised by the use of a set of bonded and non-bonded terms that define the potential energy (U) of the system as a function of atomic coordinates:

$$U(\mathbf{r}^N) = \sum U_{bonded} + \sum U_{non-bonded} \tag{2.1}$$

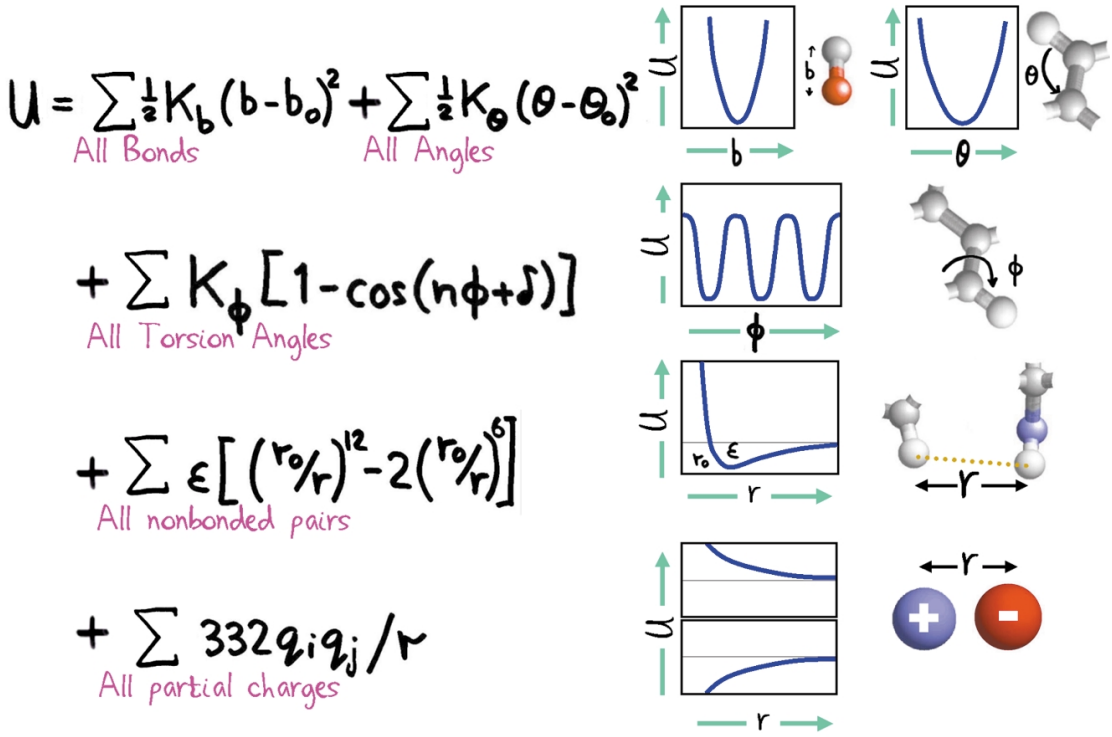


Figure 2.1: Summary of energy terms in a typical biomolecular force field. Reproduced from Levitt [178] with permission granted by Nature Publishing Group.

In the next sections I will describe all of the above terms in more detail, which are also summarised in Figure 2.1. The force constants and equilibrium values used by these models are derived by fitting to experimental measurements such as electron diffraction, crystallography and spectroscopy, as well as to QM calculations; hence the designation of *empirical*. The exact parametrisation procedure and data used is different between force field models, and as such it is inappropriate to mix and match parameters from different force fields [175].

Bonded interactions

Bonded interactions are intramolecular in nature and model forces arising from bonds stretching and angle bending, as well as torsional terms that are introduced to describe rotational barriers around a bond.

$$U(\mathbf{r}^N) = \sum U_{bonds} + \sum U_{angles} + \sum U_{dihedrals} \tag{2.2}$$

The potential energy for bond stretching and angle bending is generally described via Hooke's law, in which the energy varies with the square of the displacement from the reference bond length (l_0) or angle (θ_0):

$$U_{bonds} = \sum_{bonds} \frac{k_l}{2}(l - l_0)^2 \tag{2.3}$$

$$U_{angles} = \sum_{angles} \frac{k_\theta}{2}(\theta - \theta_0)^2 \tag{2.4}$$

The force constants k determines the energetic penalty of deviating from the reference values. Substantial energies are typically required to move these terms away from their equilibrium values, therefore most of the conformational changes encountered during a simulation are a result of non-bonded and torsional effects.

The energy of dihedral (or torsion) angles is described by a periodic function instead, typically as a cosine series expansion. For instance, the Amber force field adopts the following functional form:

$$U_{dihedrals} = \sum_{dihedrals} k_\phi(1 + cos(n\phi - \phi_0)) \tag{2.5}$$

where ϕ is the torsion angle, ϕ_0 is the phase, and n is the periodicity.

Improper torsions are often introduced in order to enforce the planarity of certain groups, such as benzene, or to keep the stereochemistry of a molecule constant during the simulation. These can have the same functional form as proper dihedrals, or can have an harmonic form, similarly to the bond and angle terms [175].

Non-bonded interactions

Non-bonded terms model the interatomic interactions between particles. However, it is important to note that they also act intramolecularly, so that, for instance, the energy of the system around a torsion angle will depend also on the non-bonded interactions between atoms in the molecule. Two types of non-bonded interactions are present in classical force fields: electrostatic interactions, described by Coulomb's law, and van der Waals (vdW) interactions, described by the Lennard-Jones potential [175].

$$U_{non-bonded} = \sum U_{vdW} + \sum U_{elec} \tag{2.6}$$

In classical fixed-charged force fields, partial charges are assigned to each particle. Thus, electrostatic interactions arise between any pair of particles carrying a charge. The potential energy associated with these interactions is modelled using Coulomb's law:

$$U_{elec} = \sum_{pairs(i,j)} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{2.7}$$

where q_i and q_j are the partial charges of particle i and j, ϵ_0 is the permittivity of free space, and r_{ij} is the interatomic distance.

The van der Waals interactions are instead typically modelled using a 6-12 Lennard-Jones potential, where the r^{-12} term describes the Pauli repulsion at short ranges and the r^{-6} term describes attraction at long ranges due to dispersion forces:

$$U_{vdW} = 4\epsilon_{ij}\left(\left[\frac{\sigma_{ij}}{r_{ij}}\right]^{12} - \left[\frac{\sigma_{ij}}{r_{ij}}\right]^{6}\right) \tag{2.8}$$

where σ_{ij} is the collision diameter (the distance at which the energy is zero), ϵ_{ij} is the well depth, and r_{ij} is the interatomic distance (Figure 2.1). σ_{ij} and ϵ_{ij} are usually obtained from the σ and ϵ values of the two interacting atom types via the Lorentz-Berthelot combination rules:

$$\sigma_{ij} = \frac{1}{2}(\sigma_i + \sigma_j) \tag{2.9}$$

$$\epsilon_{ij} = \sqrt{\epsilon_i + \epsilon_j} \tag{2.10}$$

van der Waals parameters are more difficult to derive than bonded parameters or partial charges, and often they result from iterative optimisation procedures aimed at matching experimental data such as liquid densities and heats of vaporisation [175].

2.1.2 Water models

Molecular mechanics models of liquid water are of utmost importance for computer simulations of biological material. Water makes up a large portion of biological tissue, and it is the solvent in which (soluble) proteins are stable. Thus, simulations of proteins are most often performed in aqueous environment, just like biochemical assays. Considering that there may be a large number of water molecules in a simulation, the computational efficiency of the model is often an important factor. Simple fixed-charge water models tend to be rigid and have between three and five interaction sites, so that they do not require a large amount of computation and thousands of water molecules can be included in a simulation box [175]. Table 2.1 reports the parameters of a few popular water models, while Figure 2.2 schematises the their geometry. The TIP3P and SPC water models have three electrostatic interaction sites, and a single vdW site located on the oxygen atom [179, 180]. The TIP4P model uses four interaction sites, with the negative partial charge placed on an additional site on the bisector of the HOH angle towards the hydrogen atoms [181]. In the five-sites TIP5P model, the negative partial charge is placed instead on two additional dummy atoms that represent the two oxygen lone pairs [182]. In the work presented in this thesis, I adopted the TIP3P water model, as previous studies suggested a better performance for hydration free energy calculations compared to other models [183].

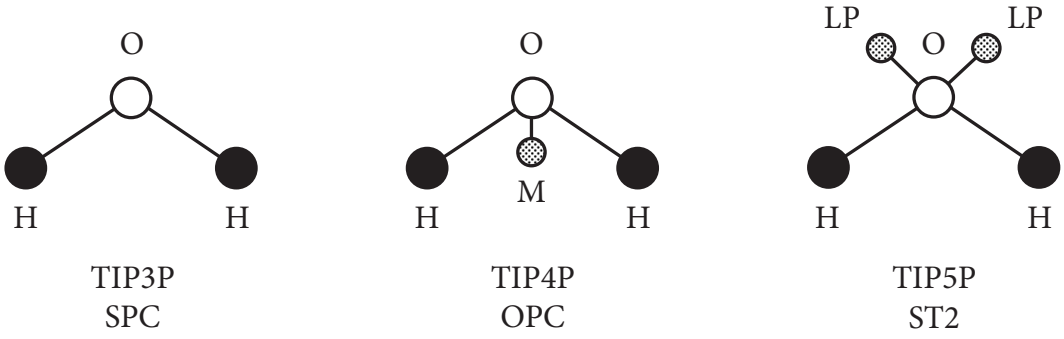


Figure 2.2: Examples of common fixed-charge water models (Table 2.1). "O" is oxygen; "H" is hydrogen; "LP" is lone pair; "M" is a massless particle used in 4-site models.

Table 2.1: A comparison of a few three-, four-, and five-point water models. A missing number means the parameter does not apply to the water model. "O" is oxygen; "H" is hydrogen; "LP" is lone pair; "M" is the massless particle used in 4-site models.

	TIP3P[179]	SPC[180]	TIP4P[181]	OPC[184]	TIP5P[182]	ST2[185]
r_{OH} (Å)	0.9572	1.0	0.9572	0.8724	0.9572	1.0
r_{OM} (Å)			0.15	0.1594		
r_{OLP} (Å)					0.70	0.80
$\angle_{H-O-H}$ (deg)	104.52	109.47	104.52	103.6	104.52	109.47
$\angle_{LP-O-LP}$ (deg)					109.47	109.47
q_O	−0.834	−0.82				
q_H	0.417	0.41	0.52	0.6791	0.241	0.243 57
q_M			−1.04	−1.3582		
q_{LP}					−0.241	−0.243 57
ϵ_O (kJ/mol)	0.6364	0.650	0.6480	0.8903	0.6694	0.316 94
σ_O (Å)	3.150 61	3.166	3.153 65	3.1666	3.12	3.1

2.2 Energy minimisation

Physics-based computer simulations are initialised by a set of coordinates, usually obtained from an experimentally determined protein structure, or a homology model. Protein-ligand complexes can also be obtained by docking. Incorrect geometries and unfavourable contacts are therefore common, and may lead to instabilities in the simulation due to large forces. Energy minimisation, or geometry optimisation, is a process that tries resolve the issue by searching for a stationary point, i.e. a local minimum, on the complex potential energy surface (PES) of the system [175]. Mathematically, the problem corresponds to finding the set of coordinates for which the first derivative of the potential energy function is zero, and its second derivative is positive:

$$\frac{\partial U}{\partial \mathbf{r}} = 0 \quad and \quad \frac{\partial^2 U}{\partial \mathbf{r}^2} > 0 \tag{2.11}$$

Due to the complexity of the potential energy function it is not possible to find its minimum analytically, thus numerical methods are required. Minimisation algorithms can be classified in two groups, depending on whether they use the derivative of the energy with respect to the coordinates or not. Among non-derivative methods are the *simplex* and *sequential univariate* methods. Derivative

methods are however more popular in molecular modelling, as the first and second derivatives of the energy provide useful information for the minimisation problem: the direction of the first derivative (the gradient) indicates the location of the minimum, and its magnitude the steepness of the local slope. The most common derivative methods are first-order algorithms like the *steepest descent* and *conjugate gradient*, and second-order ones like the *Newton-Raphson*. The *steepest descent* method has been used in this work and it can rapidly guide the system toward a minimum, despite the oscillatory behaviour once close to it. The algorithm iteratively changes the coordinates of the atoms by moving them in the negative direction of the gradient of the potential energy [175].

2.3 Molecular dynamics simulations

In molecular dynamics (MD) simulations, given a set of atomic coordinates, the atom positions are propagated through time using Netwon's second law of motion. The system is thus allowed to evolve in time, based on the potential energy function defined by the force field. Using the overdot notation for differentiation with respect to time, for a particle i:

$$\mathbf{F}_i = m\mathbf{a}_i \tag{2.12}$$

$$-\nabla_{\mathbf{r}_i} U(\mathbf{r}_i) = m\ddot{\mathbf{r}}_i(t) \tag{2.13}$$

$$\int\int -\nabla_{\mathbf{r}_i} U(\mathbf{r}_i) dtdt = m\mathbf{r}_i(t) \tag{2.14}$$

$$\mathbf{r}_i(t) = \frac{1}{m}\int\int -\nabla_{\mathbf{r}_i} U(\mathbf{r}_i) dtdt \tag{2.15}$$

The equations of motion can thus yield a time-resolved trajectory of the system of interest. However, due to the many-body nature of the problem, the above second-order differential equation (eq. 2.15) needs to be discretised and solved numerically, and the instantaneous forces acting on each atom and their initial positions and velocities need to be specified [175]. As already mentioned, the initial coordinates of the system are typically obtained from a crystal structure. Initial

velocities are instead randomly assigned from the Maxwell-Boltzmann distribution at a given temperature:

$$p(\mathbf{v}_i) = \Big(\frac{m_i}{2\pi k_B T}\Big)^{1/2} \cdot exp\Big(-\frac{1}{2}\frac{m_i \mathbf{v}_i}{k_B T}\Big) \tag{2.16}$$

where $p(\mathbf{v})$ is the probability distribution of the velocity $\mathbf{v}$ [175].

The trajectories resulting from the simulation describe the time evolution of the system in phase space and are defined by the position and velocity vectors of all particles. Thanks to the *ergodic hypothesis*, any ensemble property A can be calculated from an MD trajectory as a time average:

$$\langle A \rangle = \frac{1}{M}\sum_{i=1}^{M} A(\mathbf{p}^N, \mathbf{r}^N) \tag{2.17}$$

where $\mathbf{p}^N$ are the particles' momenta [175]. The total energy E of the system being simulated is the sum of the kinetic and potential energy, and it can be shown that the equations of motion satisfy the conservation of energy. This can be verified by differentiating the energy with respect to time:

$$E = \sum_{i=1}^{N} \frac{1}{2} m_i \dot{\mathbf{r}}_i^2 + U(\mathbf{r}_1, \ldots, \mathbf{r}_N) \tag{2.18}$$

$$\frac{dE}{dt} = \sum_{i=1}^{N} m_i \dot{\mathbf{r}}_i \cdot \ddot{\mathbf{r}}_i + \sum_{i=1}^{N} \frac{\partial U}{\partial \mathbf{r}_i} \cdot \dot{\mathbf{r}}_i \tag{2.19}$$

$$= \sum_{i=1}^{N} \dot{\mathbf{r}}_i \cdot \left[m_i \ddot{\mathbf{r}}_i + \frac{\partial U}{\partial \mathbf{r}_i} \right] \tag{2.20}$$

$$= \sum_{i=1}^{N} \dot{\mathbf{r}}_i \cdot [m_i \ddot{\mathbf{r}}_i - \mathbf{F}_i] \tag{2.21}$$

$$= 0$$

where the last line follows from eq. 2.12 [176].

In the following sections, I will briefly describe some of the fundamental aspects of MD theory and its algorithmic implementation.

2.3.1 Finite difference methods

As explained above, the equations of motion cannot be solved analytically, thus they are integrated using a *finite difference method*. The integration is broken down into many steps, each separated in time by a fixed time Δt, during which the force is assumed to be constant [175].

The first step in a finite difference method is to calculate the force acting on each particle at time t, as a vector sum of its interactions with the other particles. The force is then used to calculate the acceleration, velocity, and position of each particle at time $t + \Delta t$ using a finite difference algorithm. The derivatives of the positions $\mathbf{r}$, defining the velocities $\mathbf{v}$ and accelerations $\mathbf{a}$, are approximated as a Taylor expansion series:

$$\mathbf{r}(t+\Delta t) = \mathbf{r}(t) + \mathbf{v}(t)\Delta t + \frac{1}{2}\mathbf{a}(t)\Delta t^2 + \frac{1}{6}\mathbf{b}(t)\Delta t^3 + \frac{1}{24}\mathbf{c}(t)\Delta t^4 + \dots \tag{2.22}$$

$$\mathbf{v}(t+\Delta t) = \mathbf{v}(t) + \mathbf{a}(t)\Delta t + \frac{1}{2}\mathbf{b}(t)\Delta t^2 + \frac{1}{6}\mathbf{c}(t)\Delta t^3 + \dots \tag{2.23}$$

$$\mathbf{a}(t+\Delta t) = \mathbf{a}(t) + \mathbf{b}(t)\Delta t + \frac{1}{2}\mathbf{c}(t)\Delta t^2 + \dots \tag{2.24}$$

$$\mathbf{b}(t+\Delta t) = \mathbf{b}(t) + \mathbf{c}(t)\Delta t + \dots \tag{2.25}$$

$$\vdots$$

One of the most popular finite difference integrators is the *Verlet algorithm* [186]. This method uses the positions from $t - \Delta t$ and the positions and accelerations at t in order to calculate the positions at time $t + \Delta t$:

$$\mathbf{r}(t+\Delta t) = \mathbf{r}(t) + \mathbf{v}(t)\Delta t + \frac{1}{2}\mathbf{a}(t)\Delta t^2 + \dots \tag{2.26}$$

$$\mathbf{r}(t-\Delta t) = \mathbf{r}(t) - \mathbf{v}(t)\Delta t + \frac{1}{2}\mathbf{a}(t)\Delta t^2 - \dots \tag{2.27}$$

Adding these two equations, and ignoring higher order terms:

$$\mathbf{r}(t+\Delta t) = 2\mathbf{r}(t) - \mathbf{r}(t-\Delta t) + \mathbf{a}(t)\Delta t^2 \tag{2.28}$$

The velocities are not returned by the Verlet algorithm, however, they can be obtained in a variety of ways, including:

$$\mathbf{v}(t+\Delta t) = \frac{\mathbf{r}(t+\Delta t) - \mathbf{r}(t-\Delta t)}{2\Delta t} \tag{2.29}$$

Multiple variations of the Verlet algorithm have been developed. An improved version used in Gromacs [187] is the *leap-frog algorithm* [188, 189], which uses velocities at $t - \Delta t$ and accelerations at t in order to derive velocities at $t + \Delta t$, which are then used to obtain the positions at $t + \Delta t$; velocities at t are then calculated by interpolation of the velocities at $t - \Delta t$ and $t + \Delta t$:

$$\mathbf{v}\left(t + \frac{\Delta t}{2}\right) = \mathbf{v}\left(t - \frac{\Delta t}{2}\right) + \mathbf{a}(t)\Delta t \tag{2.30}$$

$$\mathbf{r}(t+\Delta t) = \mathbf{r}(t) + \mathbf{v}\left(t + \frac{\Delta t}{2}\right)\Delta t \tag{2.31}$$

$$\mathbf{v}(t) = \frac{1}{2}\left[\mathbf{v}\left(t + \frac{\Delta t}{2}\right) + \mathbf{v}\left(t - \frac{\Delta t}{2}\right)\right] \tag{2.32}$$

One of the advantages of the leap-frog algorithm over the Verlet algorithm is that it explicitly includes the velocities, which are used to calculate the kinetic energy. Other integrators related to Verlet are the *velocity Verlet* [190] and the *Beeman's* [191] algorithms.

Stochastic dynamics

Stochastic or Langevin dynamics adds a friction and a noise term to Newton's equations of motion [192], so that for a particle i:

$$m\ddot{\mathbf{r}}_i = \mathbf{F}_i(\mathbf{r}) - m_i\gamma\dot{\mathbf{r}} + \mathring{\mathbf{r}}_i \tag{2.33}$$

where γ is a friction coefficient and $\mathring{\mathbf{r}}_i(t)$ is a noise term with Gaussian distribution and zero mean. It is possible to see stochastic dynamics as molecular dynamics with stochastic temperature-coupling, as the dispersion of $\mathring{\mathbf{r}}_i(t)$ depends on the reference temperature. It can be shown that the resulting velocity distribution converges to a Maxwell-Boltzmann distribution at the reference temperature [193].

The advantage compared to other temperature-coupling methods described later is that it guarantees the generation of a Boltzmann ensemble. In the work presented in this thesis, stochastic dynamics as implemented in Gromacs has been used for integration of the equations of motion and for thermostatting [192, 193].

2.3.2 Constraints

As a consequence of the discretisation needed for the numerical integration of the equations of motion, which leads to the assumption of constant force between each time step Δt, the size of the time step is limited by the highest frequency motion being simulated due to stability considerations. A time step larger than the rapid vibrational modes would in fact result in an unstable simulation, as pairs of atoms forming the bond would start deviating from the equilibrium distance, resulting in larger and larger forces accumulating on them. In particular, bond stretching between heavy atoms and hydrogen atoms is very fast, despite not being of interest in most MD studies. Therefore, to be able to increase the size of the integration time step, all bonds, or bonds involving hydrogen atoms, are typically constrained to their equilibrium values. A number of constraining algorithms have been developed, in which the basic principle is to correct for deviations from equilibrium bond lengths after the atoms' coordinates have been updated by the integrator [175]. Water models in Gromacs are constrained via the SETTLE algorithm [194], while the P-LINCS algorithm has become the standard for other biomolecules [195, 196]. Through the use of constraints, it is possible to employ an integration time step of 2 fs.

2.3.3 Ensembles and temperature/pressure coupling

Molecular dynamics simulations as presented so far would be able to reproduce the microcanonical ensemble, where the number of particles (N), volume (V), and total energy (E), are constant. However, in order to simulate ensembles that better represent physiological or at least experimental conditions, the canonical (NVT) or isothermal-isobaric (NPT) ensembles are most commonly adopted in biomolecular simulations. The grand-canonical ensemble (μVT) is also sometimes used when one

wants the number of particles in a system to fluctuate. These ensembles require the use of algorithms to keep the temperature and pressure of the system constant.

Thermostats

The temperature is related to the kinetic energy K of the particles:

$$K = \frac{1}{2}\sum_{i=1}^{N} m_i \mathbf{v}_i^2 = \frac{3}{2} N k_B T \tag{2.34}$$

$$T = \frac{2}{3}\frac{K}{N k_B} = \frac{1}{3 N k_B}\sum_{i=1}^{N} m_i \mathbf{v}_i^2 \tag{2.35}$$

The simplest way to control the temperature is to scale the velocities [197]. The velocities at each time step can be scaled by a factor $\lambda = \sqrt{T_{ref}/T}$, where T_{ref} is the desired reference temperature and T is the instantaneous temperature as calculated from the kinetic energy:

$$\Delta T = \frac{1}{3}\sum_{i=1}^{N} \frac{m_i(\lambda \mathbf{v}_i)^2}{N k_B} - \frac{1}{3}\sum_{i=1}^{N} \frac{m_i \mathbf{v}_i^2}{N k_B} \tag{2.36}$$

$$= \frac{1}{3 N k_B}\left(\sum_{i=1}^{N} m_i(\lambda \mathbf{v}_i)^2 - \sum_{i=1}^{N} m_i \mathbf{v}_i^2\right) \tag{2.37}$$

$$= \frac{1}{3 N k_B}\sum_{i=1}^{N} m_i \mathbf{v}_i^2(\lambda^2 - 1) \tag{2.38}$$

$$= (\lambda^2 - 1)T \tag{2.39}$$

$$\lambda = \sqrt{T_{ref}/T} \tag{2.40}$$

However, this approach does not generate the correct thermodynamic properties of the canonical ensemble. Considering the limit in which velocities are rescaled at every time step, the kinetic energy will remain constant in time, with zero fluctuations. This is in contrast to the kinetic energy fluctuations expected from statistical mechanics [176]:

$$\sigma_K^2 = \frac{3}{2} N k_B^2 T^2 \tag{2.41}$$

The *Berendsen thermostat* [198] couples the system to an external heat bath with target temperature T_{ref}. This thermostat assigns a time scale for updating the velocities, rather than scaling them at each time step, allowing for fluctuations around the desired temperature. The coupling parameter τ determines how tightly the system and the bath are coupled together:

$$\frac{dT(t)}{dt} = \frac{1}{\tau}(T_{ref} - T(t)) \tag{2.42}$$

The change in temperature between successive time steps, and the scaling factor for the velocities are in this case the following:

$$\Delta T = \frac{\Delta t}{\tau}(T_{ref} - T(t)) \tag{2.43}$$

$$\lambda = \sqrt{1 + \frac{\Delta t}{\tau}\left(\frac{T_{ref}}{T(t)} - 1\right)} \tag{2.44}$$

This thermostat, however, do not generate rigorous canonical averages either. Methods that generate the correct ensembles are the *Andersen* stochastic thermostat[199] and *extended system* methods such as the *Nosé-Hoover* scheme [200]. As already mentioned, another way of controlling the temperature of the system is via Langevin dynamics [192, 193].

The stochastic *Langevin thermostat* applies friction and random force to momenta [193], where at each time step the change in velocity is defined by the following stochastic differential equation:

$$\Delta \mathbf{v}_i = \mathbf{a}_i \Delta t - \gamma_i \mathbf{v}_i \Delta t + \xi \sqrt{2 k_B T_{ref} \gamma_i \Delta t / m_i} \tag{2.45}$$

The friction coefficient γ_i damps the momenta, and ξ is a normally distributed random number with unit variance. As compared to global thermostats such as Berendsen [198], the stochastic Langevin thermostat acts on each degree of freedom locally and returns the correct distribution of kinetic energies for decoupled states [201]. This is an important consideration when running alchemical free energy calculations, since the ligand can be partially or completely decoupled from the rest of the system in a number of simulations [193].

Barostats

In computer simulations, the pressure is typically calculated via the virial theorem of Clausius [175], where the *virial* (W) for a three-dimensional box is defined as follows:

$$W = \sum_{i=1}^{N} \mathbf{r}_i \cdot \mathbf{F}_i^{TOT} = -3Nk_BT \tag{2.46}$$

The total forces acting on a particle ($\mathbf{F}_i^{TOT}$) can be thought of as being composed of two contributions: one arising from the container's wall ($\mathbf{F}_i^{EXT}$), and another one arising from interatomic interactions ($\mathbf{F}_i^{INT}$):

$$W = \sum_{i=1}^{N} \mathbf{r}_i \cdot \mathbf{F}_i^{EXT} + \sum_{i=1}^{N} \mathbf{r}_i \cdot \mathbf{F}_i^{INT} \tag{2.47}$$

For an ideal gas, the only forces present are the ones due to the container, such that:

$$W = \sum_{i=1}^{N} \mathbf{r}_i \cdot \mathbf{F}_i^{EXT} = -3PV \tag{2.48}$$

When considering a real gas or a liquid, also the forces arising from the particle interactions must be taken into account:

$$W = -3PV + \sum_{i=1}^{N} \sum_{j>i}^{N} \mathbf{r}_{ij} \cdot \mathbf{F}_{ij} \tag{2.49}$$

Combining equations 2.46 and 2.49:

$$-3Nk_BT = -3PV + \sum_{i=1}^{N} \sum_{j>i}^{N} \mathbf{r}_{ij} \cdot \mathbf{F}_{ij} \tag{2.50}$$

$$P = \frac{1}{V}\left[Nk_BT - \frac{1}{3}\sum_{i=1}^{N} \sum_{j>i}^{N} \mathbf{r}_{ij} \cdot \mathbf{F}_{ij}\right] \tag{2.51}$$

As interatomic forces are already calculated as part of the simulation, the computation of the virial comes at virtually no additional cost. It is also possible to see how the pressure can be kept constant by scaling the volume of the simulation box, which is equivalent to scaling the atomic coordinates.

The *Berendsen* barostat follows a similar approach to the Berendsen thermostat [198], by coupling the system to an external pressure "bath":

$$\frac{dP(t)}{dt} = \frac{1}{\tau}(P_{ref} - P(t)) \tag{2.52}$$

$$\lambda = 1 - \kappa\frac{\Delta t}{\tau}(P - P_{ref}) \tag{2.53}$$

$$\mathbf{r}_i' = \lambda^{1/3} \cdot \mathbf{r}_i \tag{2.54}$$

where κ is the isothermal compressibility.

Although the Berendsen algorithm yields a simulation with the correct average pressure, it does not yield the exact NPT ensemble. Therefore, this barostat has been used mainly for equilibration purposes, but not for data collection and analysis. For production simulations, the *Parrinello-Rahman* barostat has been employed [202]. This method is an extended system method that is similar to the Nosé-Hoover approach to temperature control. In these methods, extra degrees of freedom are added and the Hamiltonian. In the Andersen barostat [199], the extra degree of freedom is the volume of the box, which has associated kinetic and potential energy terms and follows modified equations of motion. The Parrinello-Rahman method is an extension of Andersen, in which the box sides are made independent of each other, and the box side vectors are allowed to be non-orthogonal, which means that anisotropic pressure coupling is possible.

2.3.4 Periodic boundary conditions

In MD we are simulating a very small volume, focussing on the time evolution of a small number of atoms. This leads to the question of how to treat the system at the boundaries of the simulation box. Periodic boundary conditions allow to simulate a small number of particles in such a way that these will experience forces as if they were in bulk. The simulation box is tiled in all dimensions in order to produce a periodic array, so that there are no boundaries of the system. While a cubic box is the simplest periodic system to visualise and program, it is not the most computationally efficient [175]. Usually we are most interested in the behaviour of

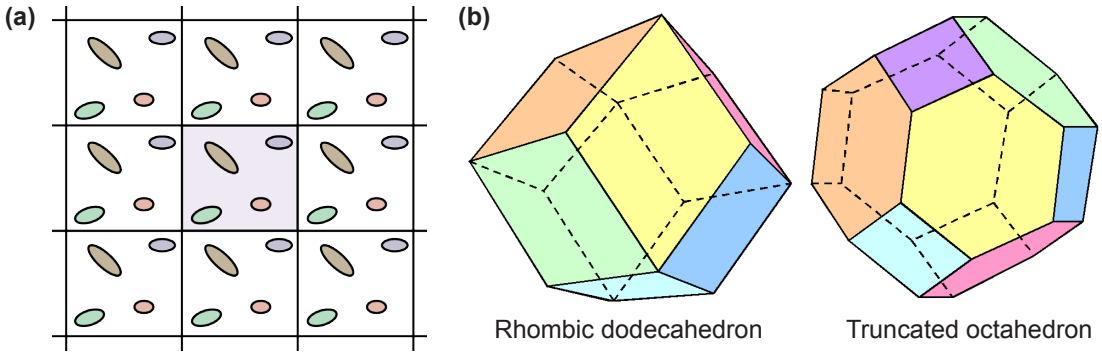


Figure 2.3: (a) Two-dimensional representation of periodic boundary conditions; image by Cnrowley, distributed under CC BY-SA 4.0 via Wikimedia Commons. (b) Shape of rhombic dodecahedral and truncated octahedral periodic cells; images by Svdmolen and available in the public domain.

the solute, and a large number of solvent molecules is therefore undesirable as it contributes to the computational burden without providing additional information. Box shapes such as the *truncated octahedron* and the *rhombic dodecahedron* provide periodic cells that reduce the volume required for the simulation to 0.770 and 0.707 of the cubic volume, respectively. In the work presented here, simulations have been performed in a rhombic dodecahedron box unless otherwise stated.

2.3.5 Treatment of long-range interactions

The most time-consuming part of an MD code is the calculation of the non-bonded energies, since each particle interacts with all other particles in the system. The number of pairwise interactions to be calculate increases as the square of the number of particles, thus being of order N^2.

Long-range Lennard-Jones interactions

The Lenard-Jones (LJ) energy, however, decays fairly rapidly with the distance, with r^{-6} dependence. Hence, it is common to adopt a cut-off so that the LJ interactions will be calculated only for particles within this distance; outside the cut-off the LJ energy is set to zero. In order to know which particles are within this cut-off, a neighbour list is used and updated at regular intervals. For free energy

calculations, even this approximation can affect the accuracy of the results, and methods to address this issue have been developed.

In Gromacs,[187, 203, 204] there is an analytical correction for energy and pressure that tries to remove artefacts due to this cut-off. However, it assumes an isotropic distribution of solvent outside the LJ cut-off, which is not the case when simulating large biomolecules, at least when the cut-off is smaller than the maximum dimension of the molecule. Shirts *et al.* [205] proposed a numerical correction for non-isotropic systems and named it "EXP-LR". The idea is to estimate the free energy difference between the reference state that uses the LJ cut-off, and a state that does not have any LJ cut-off. This is done by building an additional thermodynamic cycle on top of the one already used for the ligand-binding free energy calculation (see § 2.6.1), and re-evaluating the energies of the system without the cut-off by re-running the trajectory already obtained. The method is relatively cheap computationally, as only the potential energy of the system needs to be calculated. In the work presented here, the EXP-LR approach has been used to remove the dependence of the free energy results on the cut-off distance used. More recently, a lattice summation technique called Lennard-Jones particle mesh Ewald (LJ-PME) has been developed and implemented in the Gromacs package [206]. This method efficiently computes LJ energies between particles at a distance larger than the cut-off with only a small approximation error, thus eliminating the need of a post-hoc correction.

Long-range Coulombic interactions

Electrostatic interactions decay with a r^{-1} distance dependence, therefore the error introduced by the use of a cut-off can be large [207]. A number of techniques have been developed to handle long-range electrostatic contributions in periodic systems. Among these, Ewalds methods are the most popular thanks to their accuracy and convergence properties [208–210]. The fundamental idea is to convert a single slowly-converging sum into two quickly-converging terms. The following observation by Ewald [208] is thus the mathematical foundation for all these techniques:

$$\frac{1}{r} = \frac{f(r)}{r} + \frac{1 - f(r)}{r} \tag{2.55}$$

The electrostatic problem is then solved in two parts, where a cut-off defines short- and long-range contributions. The short-range contribution is evaluated in real space, while the long-range one converges rapidly in Fourier space. Further improvements are obtained by mapping the charges onto grid points and using fast Fourier transforms [209]. These algorithms reduce the scaling with respect to the number of particles from N^2 to $NlnN$. A number of variants based on this theme have been proposed, with the particle mesh Ewald (PME) [209, 210] and particle-particle-particle-mesh (PPPM) [211] approaches being among the most computationally efficient [212]. PME has been used in all MD simulations of this thesis.

When using periodic boundary conditions and PME, the average of the electric potential over the computational box is set to zero, contrary to the ideal non-periodic bulk situation where the electric potential would be zero at infinity. The PME definition of zero electrostatic potential can be a problem when calculating the absolute free energy of binding of charged molecules. In fact, as the net charge of ligand is scaled to zero, the simulation box is not neutral anymore, with the consequence of a uniform neutralising background charge being added to the system in order to compensate its net charge [213]. This neutralising charge, along with other artefacts dependent on the box size and caused by periodicity, can affect the results of a charging free energy calculation by introducing spurious interaction between the protein-ligand complex and its periodic replicas. A scheme was therefore recently suggested by Rocklin *et al.* [213] for the correction of these effects. In this work, the semi-analytical correction they proposed has been used:

$$\Delta\Delta G_{correction}(L) = \Delta\Delta G_{NET}(L) + \Delta\Delta G_{USV}(L) + \Delta\Delta G_{RIP}(L) + \Delta\Delta G_{EMP}(L) + \Delta\Delta G_{DSC}(L) \quad (2.56)$$

$$\Delta\Delta G_{NET}(L) = -\frac{\xi_{LS}}{8\pi\epsilon_0}[(Q_P + Q_L)^2 - Q_P^2]\frac{1}{L} \quad (2.57)$$

$$\Delta\Delta G_{USV}(L) = \frac{\xi_{LS}}{8\pi\epsilon_0}\left(1 - \frac{1}{\epsilon_S}\right)[(Q_P + Q_L)^2 - Q_P^2]\frac{1}{L} \quad (2.58)$$

$$\Delta\Delta G_{RIP}(L) = [(I_P + I_L)(Q_P + Q_L) - I_P Q_P]\frac{1}{L^3} \quad (2.59)$$

$$\Delta\Delta G_{EMP}(L) = -\frac{1}{8\pi\epsilon_0}\frac{16\pi^2}{45}\left(1 - \frac{1}{\epsilon_S}\right) \times \left\{\left[\frac{1}{8\pi\epsilon_0}\frac{4\pi}{3}\left(1 - \frac{1}{\epsilon_S}\right)Q_L\right]^{-1} I_{L,SVL}\right\}\frac{1}{L^6} \quad (2.60)$$

$$\Delta\Delta G_{DSC}(N_S, L) = -\frac{\gamma_S Q_L}{6\epsilon_0}\frac{N_S}{L^3} \quad (2.61)$$

where NET stands for periodicity-induced net-charge interactions; USV stands for periodicity-induced undersolvation; RIP stands for residual integrated potential; EMP refers to an empirical term; DSC stands for discrete solvent effects; Q_P and Q_L are the protein and ligand net charges, with $Q_P = 0$ if counterions have been used; L is the length of the edge of the cubic box; I_P and I_L are the residual integrated potentials of the protein and ligand, which need to be obtained via a Poisson-Boltzmann calculation; $I_{L,SLV}$ is the solvation contribution to the ligand RIP; N_S is the number of water molecules in the system; ϵ_0 and ϵ_S are the vacuum and solvent permittivities; ξ_{LS} is the Wigner integration constant, which for a cubic cell is approximately -2.84; and γ_S is the quadrupole-moment trace of the solvent model relative to its single van der Waals interaction site, defined as $\gamma_S = \sum_{i=1}^{n} q_i r_i^2$, where q_i is the partial charge on the interaction site i at distance r_i from the single vdW site of the water model.

2.4 Monte Carlo simulations

Monte Carlo (MC) methods are a class of computational algorithms that rely on repeated random sampling to compute their results [214]. More specifically, the *Metropolis Monte Carlo* method [215] is used for the sampling of physical systems.

As opposed to MD simulations, Monte Carlo simulations have no momenta or time dependence.

MC reproduces the correct statistical ensemble by using acceptance criteria that ensure that the probability of obtaining a certain configuration is equal to its Boltzmann factor. States with higher energy will therefore be generated with lower probability than states with lower energy [175]. From all accepted configurations of the system the properties of interest can thus be calculated by simply averaging over all configurations:

$$\langle A \rangle = \frac{1}{M} \sum_{i=1}^{M} A(\mathbf{r}^N) \tag{2.62}$$

where A is the thermodynamic property of interest, M is the number of accepted configurations, and $\mathbf{r}^N$ are the coordinates of all particles.

In MC, each new configuration is generated by randomly translating and rotating atoms and molecules. The system's energy is defined by a molecular mechanics potential energy function. The newly generated configuration is then accepted if its energy is lower than the energy of its previous configuration [175]. If the energy is higher, the new configuration is accepted according to a Metropolis criterion:

$$rand(0,1) \leq e^{-\Delta E/k_B T} \tag{2.63}$$

The new state with higher energy is therefore accepted only if its Boltzmann factor is larger than a randomly generated number between zero and one. This procedure is then performed a large number of times and the system configuration is written to file at the desired frequency [175]. Each iteration is independent of one another, and the probability of reverting form a state B to a previous state A is as probable as moving from state A to B, satisfying detailed balance:

$$\pi_i p_{ij} = \pi_j p_{ji} \tag{2.64}$$

where p_{ij} is the Markov transition probability from state i to state j, and π is the equilibrium probability of being in states i and j [175].

The MC method for biomolecular sampling has been implemented in a number of different codes. In this thesis, and specifically in chapter 6, the code developed in Prof. Jon Essex's group at the University of Southampton is employed [216].

2.4.1 Grand Canonical Monte Carlo simulations

The grand canonical ensemble (μVT) is a statistical ensemble where the conserved properties are the chemical potential μ, the volume, and the temperature. The key feature of this ensemble is that the number of particles in the simulation are allowed to vary [175]. The molecular species to be inserted/removed are thought to be present in ideal gas state in a reservoir coupled to the condensed phase simulation box, so that the chemical potential of a system where only one species can vary is defined as follows:

$$\mu = \left(\frac{\partial F(N,V,T)}{\partial N}\right)_{V,T} \tag{2.65}$$

The chemical potential controls the ratios of the number or particles in the system and reservoir, with equilibrium being reached when the chemical potentials of the two are equal [217, 218]. In a MC simulation, molecules can be allowed to be inserted or removed from the system by allowing these moves, where the insertion and deletion probabilities can be defined as:

$$p(N_{sys} \rightarrow N_{sys}+1) = min\left[1, \frac{1}{N_{sys}+1} \cdot e^{B} e^{-\Delta E/k_B T}\right] \tag{2.66}$$

$$p(N_{sys} \rightarrow N_{sys}-1) = min\left[1, N_{sys} \cdot e^{B} e^{-\Delta E/k_B T}\right] \tag{2.67}$$

where N_{sys} is the number of molecules of interest present in the system, and B is a parameter known as Adams parameter, which can be defined as [68]:

$$B = \frac{\mu}{k_B T} + ln\left(\frac{V}{\Lambda^3}\right) \tag{2.68}$$

where Λ is the thermal de Broglie wavelength of the molecule, and V is the volume over which the insertion/deletion moves can happen. The Adams parameter

needs to be set before the start of the MC simulation and it will affect the acceptance probability of grand canonical Monte Carlo (GCMC) moves, as well as the equilibrium number of grand canonical molecules in the system [68, 217, 218]. Under these conditions, the chemical potential of the system and the ideal gas reservoir are:

$$\mu_{sys} = \mu'_{sys} + k_B T ln\left(\frac{N_{sys,equil}\Lambda^3}{V}\right) \tag{2.69}$$

$$\mu_{res} = k_B T ln\left(\frac{N_{res,equil}\Lambda^3}{V}\right) \tag{2.70}$$

where μ' is the excess chemical potential. Combining equations 2.68, 2.69, and 2.70 the following definitions of the Adams parameter can be derived:

$$B = \frac{\mu'_{sys}}{k_B T} + ln(N_{sys,equil}) \tag{2.71}$$

$$B = ln(N_{res,equil}) \tag{2.72}$$

It is thus possible to see how the value of B determines the size of the reservoir at equilibrium as well as the average number of grand canonical particles found in the simulation [68].

In this thesis, MC simulations in the grand canonical ensemble have been used in order to estimate the binding free energy of water molecules for different bromodomain binding pockets. In these calculations, the grand canonical solute is water, which is allowed to be inserted/removed from the proteins' cavities [68].

2.5 Replica exchange

Replica exchange methods enhance sampling by running multiple simulations of the same system in parallel at different thermodynamic states. In the first use of the method, Swendsen and Wang used multiple Monte Carlo replicas at different temperatures to speed-up sampling [219]. A set of simulations with increasing temperatures $T_0 < T_1 < T_2 < T_M$ is prepared, where T_0 is the thermodynamic state we are interested in. All the simulations are run in parallel, and the system

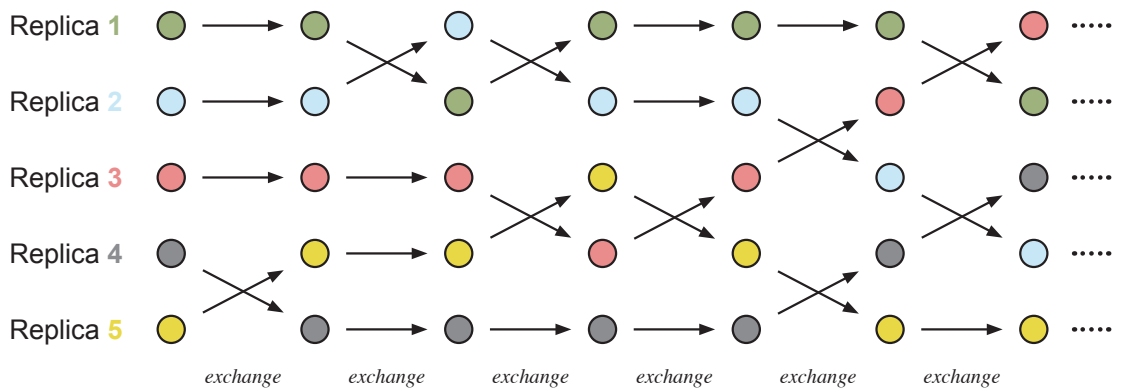


Figure 2.4: Simple scheme showing the principle behind the replica exchange method.

configurations are allowed to swap between replicas (Figure 2.4) according to a Metropolis criterion that depends on the potential energy difference between replicas [220–222]. The exchange probability between replicas is defined as:

$$p(1 \leftrightarrow 2) = min\left(1, exp\left[\left(\frac{1}{k_B T_1} - \frac{1}{k_B T_2}\right)(U_1 - U_2)\right]\right) \tag{2.73}$$

In such a way, energetic barriers are more easily overcome thanks to the larger kinetic energy at higher temperatures, while providing a correct ensemble of conformations at T_0. The acceptance criterion implies that swaps between replicas occur frequently only if the energetic overlap between replicas is sufficient. The computational overhead due to the swaps and the Monte Carlo move is negligible, considering that the potential energies of the systems are evaluated anyway for the calculation of the forces [222].

Temperature is not the only alteration between replicas that can be used. *Hamiltonian replica exchange* is a more general form of replica exchange where it is possible to enhance sampling by using parameters other than temperature [223, 224]. In alchemical free energy calculations, interaction energies are scaled in order to decouple a ligand from its environment. This can be done via a number of equilibrium simulations where the parameter λ determines the scaling of these interactions. In the free energy perturbation formalism, phase space overlap between these equilibrium simulations is important for the estimation of the final free energy. Hence, replica exchange can be used to enhance mixing

among the different thermodynamic states leading to improved overlap [225]. In this circumstance, the exchange probability is:

$$p(1 \leftrightarrow 2) = min\left(1, exp\left[\left(\frac{1}{k_B T}\right)\left((U_1(\mathbf{r}_2) - U_1(\mathbf{r}_1)) + (U_2(\mathbf{r}_1) - U_2(\mathbf{r}_2)\right)\right]\right) \tag{2.74}$$

Typically, swaps are allowed only between neighbouring replicas. However, it is possible to swap configurations across all replicas in a scheme called *Gibbs sampling replica exchange*, leading to improved mixing between states [225]. This approach has been used for the free energy calculations presented in this thesis.

2.6 Free energy calculations

Free energy can arguably be considered the most important property in thermodynamics, as the difference in free energy between two states is what determines the probability of a system being found in one or the other state. This also determines whether a process is energetically favourable so that work is obtainable from it, or unfavourable so that work needs to be done in order for the process to take place [25]. The free energy of a system is directly related to its partition function, and in the canonical ensemble the expression for the Helmholtz free energy can be expressed as follows:

$$A = -k_B T ln Q_{NVT} = U - TS \tag{2.75}$$

where Q_{NVT} is the canonical partition function, U is the internal energy, T is the temperature and S is the entropy of the system [177, 226]. At constant pressure, in the isothermal-isobaric ensemble, the Gibbs free energy is instead defined as:

$$G = A + PV \tag{2.76}$$

$$= U - TS + PV \tag{2.77}$$

$$= H - TS \tag{2.78}$$

where P is the pressure, and $U + PV$ corresponds to the enthalpy H [177]. Properties such as solvation, binding, stability, and phase transitions, just to mention some, can be described as free energy differences. A central goal in structure-based drug design is the prediction of the biological activity of a compound, which is related to its affinity for a biological target. Under isothermal and isobaric conditions, the dissociation constant K_d of a ligand is related to the Gibbs free energy according to the following equation:

$$\Delta G_{bind} = -k_B T ln \frac{K_d}{c^{\ominus}} \tag{2.79}$$

where $c^{\ominus}$ is the standard concentration, 1 molar by convention [53, 226–228]. This is the binding free energy of a single ligand molecule. However, it is common to report the binding free energy per mole of ligand; k_B is thus multiplied by the Avogadro constant N_A. $k_B N_A$ corresponds to the gas constant R. In this thesis we focus on binding free energy calculations in particular, due to their relevance for structure based drug design. Thus, ΔG_{bind} is the quantity we try to predict.

A large number of methods for estimating binding free energies have been developed. Many are still under study, and each has different trade-offs between accuracy and computational efficiency. At the most approximate and fastest end of the spectrum are *scoring* approaches, in which a protein-ligand complex is scored by a function that employs a physical or statistical potential [229]. At the other end are free energy calculations based on atomistic simulations, also called *pathway* methods, which can be broadly divided in *alchemical* and *geometrical* calculations [25]. Alchemical free energy calculations are based on a non-physical thermodynamic cycle, where the binding free energy is computed as the sum of multiple steps during which the ligand is "inserted" or "removed" from different environments [228]. Geometrical calculations follow instead a physical pathway, by applying a force that pulls the ligand away from the protein and calculating the work required to extract the ligand from the pocket [25]. Somewhat in between are so called *end-point* methods, which are typically based on the post-processing of snapshots from MD or MC simulations using implicit solvent models for computational efficiency

[229–231]. It is important to note that while alchemical and geometrical free energy approaches are rigorous from a statistical mechanics viewpoint, i.e. they return the correct free energy difference given a certain potential energy function and ensemble, scoring functions and end-point methods are not, as they try to obtain a quantity that can be a surrogate for the free energy.

In this section, we provide an overview of the theoretical foundations of alchemical binding free energy calculations, as well as end-point methods, with attention to the specific approaches that have been used in the work presented in this thesis. In addition, a recent free energy method based on grand canonical Monte Carlo (GCMC) simulations, which has been employed for the calculation of the binding free energy of water molecules, is briefly reviewed.

2.6.1 Alchemical methods

As already mentioned, in alchemical free energy calculations a chemical species is transformed into another via a non-physical pathway. Many physical processes, including ligand binding, can be expressed as a composition of non-physical transformations. These transformations are accomplished by coupling the Hamiltonian (H) of the system to a parameter λ that interpolates between the two end states of interest. Clearly, this is possible because free energy is a state function; this also implies that the initial and final states must still have physical meaning. The advantage as compared to faster approaches is that the discrete nature of the solvent as well as entropy contributions are automatically taken into account, thanks to the explicit representation of all the atoms and the correct sampling of the statistical ensemble [25, 229]. In this section, I will focus in particular on the thermodynamic cycle behind absolute, rather than relative, binding free energies, in which the two end states of interest are the ligand bound and unbound states. I will also restrict the discussion to equilibrium approaches, in which the calculation is split into a number of equilibrium simulations at different thermodynamic states. Non-equilibrium approaches based on the Jarzinski equality [27, 28] or the Crooks

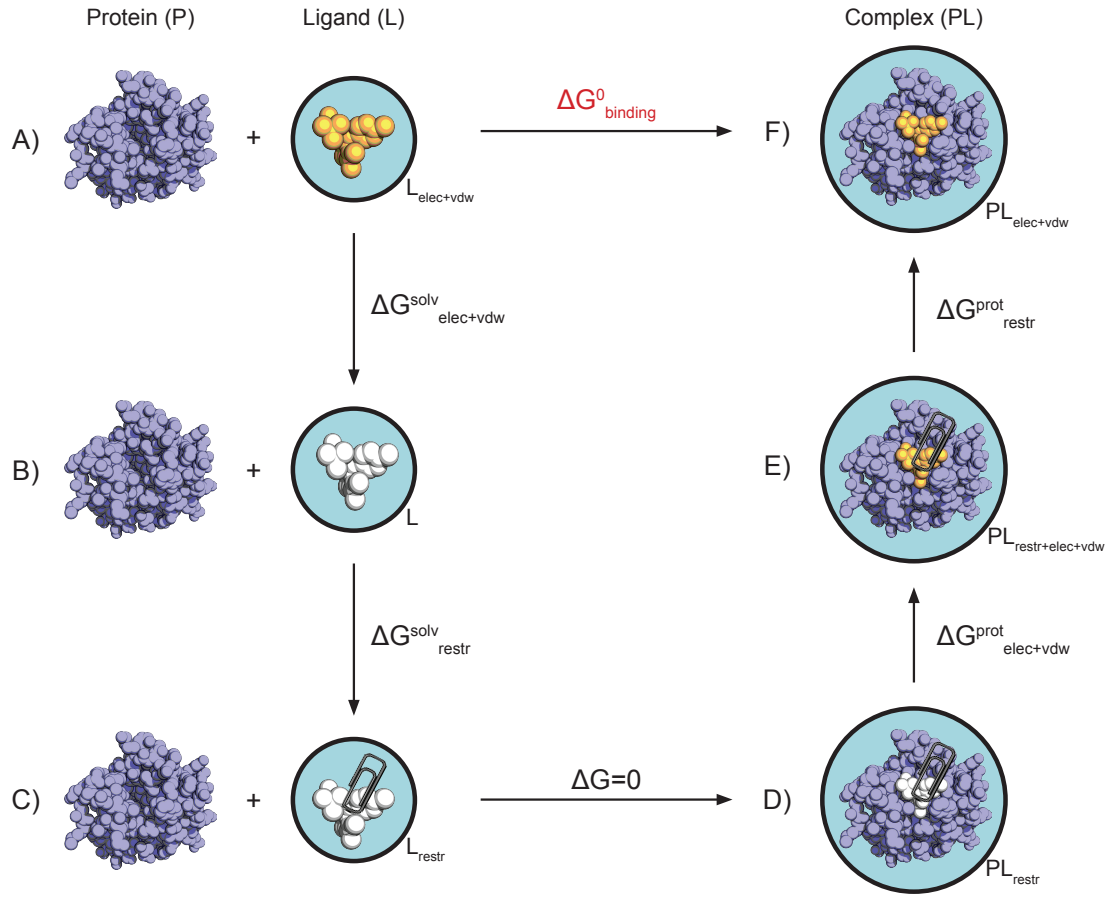


Figure 2.5: Thermodynamic cycle for absolute binding free energy calculations. The fully interacting ligand (orange) in solution at the top left (A) is transformed into a non-interacting solute (B, white) during a series of equilibrium simulations where its electrostatic and van der Waals interactions are scaled to zero, providing the term $\Delta G^{solv}_{elec+vdW}$. The ligand is then restrained while still non-interacting with the environment (C). This step (ΔG^{solv}_{restr}) can be computed analytically when following the protocol proposed by Boresch *et al.* [233]. This state is equivalent to having the non-interacting ligand restrained within the protein cavity (D). The restrained and non-interacting ligand in complex with the protein has its electrostatic and vdW interactions turned back on again (E), giving $\Delta G^{prot}_{elec+vdW}$. The restraints between ligand and protein are then removed (ΔG^{prot}_{restr}), closing the cycle, and the final state is the unrestrained and fully interacting ligand in complex with the protein (F).

fluctuation theorem [232] have not been used in the work presented here and will therefore not be discussed.

The non-physical transformations involved in an alchemical calculation are possible thanks to the malleability of the MM potential energy function, and usually involve the scaling of the Coulombic and LJ interactions between ligand

and environment, although in theory any term could be modified and the free energy difference between initial and final states recovered [25]. The thermodynamic cycle typically employed in absolute binding free energy calculations is presented in Figure 2.5. In this cycle two restraining steps are present. When using the set of restraints proposed by Boresch *et al.* [233], composed of one distance, two angles, and three dihedrals between three atoms of the protein and three of the ligand, the free energy difference associated with one of these two steps can be computed analytically from the following equation:

$$\Delta G^{solv}_{restr} = -k_B T \ln \left[\frac{8\pi^2 V^0}{r_0^2 sin\theta_{A,0} sin\theta_{B,0}} \cdot \frac{(K_r K_{\theta_A} K_{\theta_B} K_{\phi_A} K_{\phi_B} K_{\phi_C})^{\frac{1}{2}}}{(2\pi k_B T)^3} \right] \tag{2.80}$$

where K refers to the force constant of a harmonic potential, r is the equilibrium value of the distance restraint, θ_A and θ_B are the equilibrium values of the two angle restraints, and ϕ_A, ϕ_B, and ϕ_C are the equilibrium values of the three dihedral restraints. Data extracted from all the simulations composing the cycle provide the final free energy difference. In the next paragraphs we review some of the most popular methods for estimating free energies from equilibrium simulations.

Thermodynamic integration

Thermodynamics integration (TI) is a gradient based approach for the estimation of free energy differences from computer simulations. By taking the derivative of the free energy with respect to the parameter λ:

$$\frac{dG}{d\lambda} = \frac{d}{d\lambda} \int e^{-\frac{H(\lambda,\mathbf{x})}{k_B T}} d\mathbf{x} = \left\langle \frac{dH(\lambda,\mathbf{x})}{d\lambda} \right\rangle_\lambda \tag{2.81}$$

$$\Delta G = \int_{\lambda=0}^{\lambda=1} \left\langle \frac{dH(\lambda,\mathbf{x})}{d\lambda} \right\rangle_\lambda \tag{2.82}$$

where the λ is the coupling parameter, H the Hamiltonian of the system, and $\mathbf{x}$ the coordinate and momentum space [53, 227]. In the canonical ensemble, provided that the end-states have the same masses, H reduces to the potential energy U as

the momenta are not dependent on λ. In the NPT ensemble, there is also a pressure-volume contribution to the free energy, however, at constant pressure the change in free energy related to changes in average volume tends to be negligible.[227] $\langle \frac{dH(\lambda,\mathbf{x})}{d\lambda} \rangle_\lambda$ represents an ensemble average of $\frac{dH(\lambda,\mathbf{x})}{d\lambda}$ obtained from a simulation at particular values of λ. Numerical integration is then needed to recover the free energy difference between $\lambda = 0$ and $\lambda = 1$. The trapezoidal rule is usually used for simplicity, but any numerical integration method can be used. TI is quite popular for the estimation of free energy differences, as it is robust and accurate while also easy to use if the simulation code provides the derivatives needed.

Zwanzig relationship

Another approach is based on perturbation theory, and for this reason it is often referred to as *free energy perturbation* (FEP). However, FEP is also at times used to refer to alchemical free energy methods in general, which can make the terminology confusing. Adding to this, other techniques in molecular modelling are referred to as perturbation methods too. Thus, for clarity and as done by some authors[43, 53], I will refer to this approach as *exponential averaging* (EXP). Historically, this has probably been the most well-known method for calculating free energy differences, and it is based on the following formula by Zwanzig [234]:

$$\Delta G = \beta^{-1} ln \Big\langle e^{-\beta[H_1(\mathbf{x}) - H_0(\mathbf{x})]} \Big\rangle_0 = \beta^{-1} ln \Big\langle e^{-\beta \Delta H(\mathbf{x})} \Big\rangle_0 \tag{2.83}$$

where $\beta = (k_B T)^{-1}$. The equation shows that the free energy difference is the logarithm of the ensemble average of the exponential of the Boltzmann weighted energy difference between the Hamiltonians H_0 and H_1. This is why the approach is also called exponential averaging. Note that the energy difference between the two Hamiltonians is evaluated over the same reference system. Despite the equation being exact, it has been shown that EXP convergences only slowly with the amount of data collected, and an average that appears to have converged may only indicate poor overlap between the two states studied [235, 236].

Bennet's acceptance ratio

The free energy obtained via EXP for either the forward or reverse direction converges to the same result with sufficient samples. A simple way to improve EXP is thus to simply perform the calculation in both directions and average the results. However, because of a direct relationship between the distributions of potential energy in the forward and reverse directions, there is a more robust and statistically optimal way to use information from both directions [53]. Bennett's formulation started with the following observation:

$$\Delta G = k_B T ln \frac{\langle a(\mathbf{x}) exp[-\beta(H_0 - H_1)]\rangle_1}{\langle a(\mathbf{x}) exp[-\beta(H_1 - H_0)]\rangle_0} \tag{2.84}$$

which is true for any function $a(\mathbf{x})$ [237]. He then used variational calculus to find a function $a(\mathbf{x})$ that minimises the variance of the free energy, resulting in the following implicit function of ΔG that is solvable numerically:

$$\begin{aligned} \sum_{i=1}^{n_1} \frac{1}{1 + exp[ln(n_1/n_0) + \beta\Delta H_{10} - \beta\Delta G]} \\ - \sum_{j=1}^{n_0} \frac{1}{1 + exp[ln(n_0/n_1) + \beta\Delta H_{01} - \beta\Delta G]} = 0 \end{aligned} \tag{2.85}$$

where n_1 and n_0 are the number of samples from each state.[237] This formula provides a maximum likelihood estimate of the free energy given the samples from the two states [238]. Studies have shown the superiority of the Bennet's acceptance ratio (BAR) over EXP in molecular simulations: significantly less phase space overlap between states is required in order to converge results as compared to EXP [235, 236]. TI, on the other hand, can be as efficient as BAR if the integrand is smooth [236, 239].

Mutistate Bennet's acceptance ratio

A multistate extension of BAR, called the multistate Bennett's Acceptance Ratio (MBAR), has recently been proposed by Shirts *et al.* [240]. In this approach, a series of $K \times K$ weighting functions $a(\mathbf{x})$ are derived to minimise the uncertainties in free energy differences between all K states considered simultaneously. MBAR reduces

to BAR when only two states are considered, and it can also be interpreted as a zero-width weighted histogram analysis method (WHAM) [53, 227, 240]. A python implementation of the MBAR estimator is provided by the authors at `https://github.com/choderalab/pymbar`. In all of the alchemical free energy calculations presented in this thesis, MBAR has been used to estimate free energy differences.

2.6.2 Implicit solvent end-point methods

Computational methods that combine molecular mechanics energy terms with implicit solvent methods have been widely employed for the prediction of ligand binding thanks to the fact they are computationally cheap as compared to exact methods, while being more accurate than scoring functions. The fundamental idea is to process a number of configurations extracted from an MD or MC simulation and to use the MM energy and an implicit solvent model in order to approximate the binding *energy*. An entropic term is sometimes added to the protocol in order to obtain an approximation of the binding *free energy* [230, 241]. In this section I will limit the discussion to the molecular mechanics Poisson-Boltzmann surface area (MMPBSA) approach, as it is the one employed for the calculations in chapter 5. A closely related approach is MMGBSA, which makes use of the Generalised Born model, an approximation of the linearised PB equation, for the implicit treatment of solvent.

The idea behind MMPBSA is based on the cycle shown in Figure 2.6 and starts with the following relation:

$$\Delta G_{bind} = \langle G^{PL} \rangle_{PL} - \langle G^{P} \rangle_{P} - \langle G^{L} \rangle_{L} \tag{2.86}$$

$$\approx \langle G^{PL} - G^{P} - G^{L} \rangle_{PL} \tag{2.87}$$

where PL represent a protein-ligand complex, P the protein, and L the ligand, and where the angle brackets represent an ensemble average. Strictly, the complex, protein and ligand ensembles should be taken from different simulations. However, it is more common to simulate only the complex, and then extract the ligand and protein conformations from that simulation [230, 241]. This is done primarily for

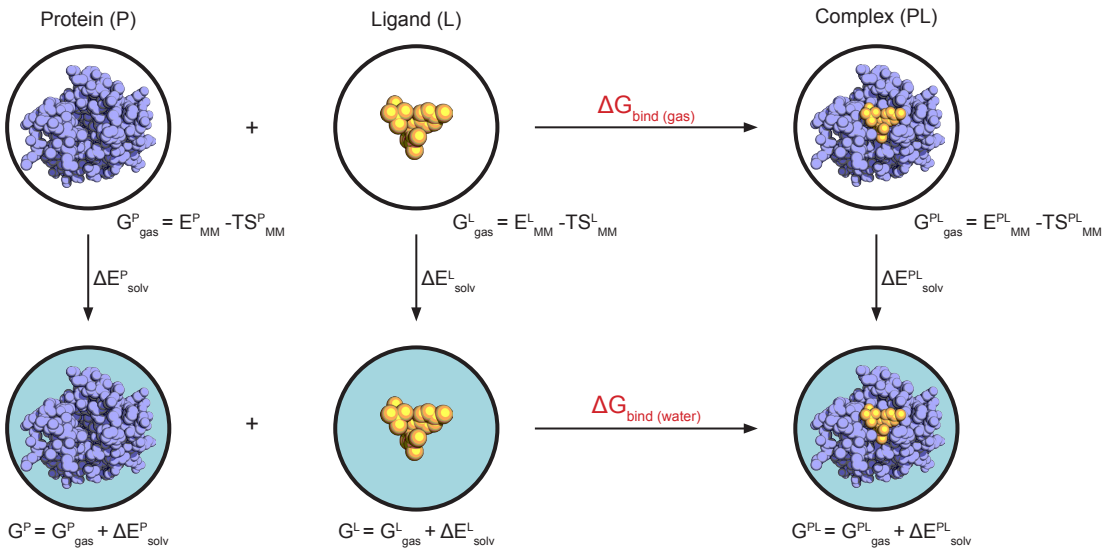


Figure 2.6: Thermodynamic cycle for MMPBSA binding free energy calculations.

computational efficiency, as only one simulation is needed and the calculations are better converged given the same amount of simulated time. In addition, while in theory the single-trajectory approach disregards potentially important ligand and protein structural changes upon binding, it is found to be more accurate than the three-trajectory approach in a system-dependent manner [230, 242–244]. This is a recurrent theme in MMPSBA calculations, where a consensus on an overall best approach or set of parameters to use is lacking, as the outcome is often system-dependent. In practice, this often means that the protocol that is computationally more efficient is chosen.

The free energy for PL, P, and L is approximated as follows:

$$G = E_{MM_{vacuo}} + \Delta E_{solv} - TS_{MM} \tag{2.88}$$

$$E_{MM} = E_{bond} + E_{coul} + E_{LJ} \tag{2.89}$$

$$\Delta E_{solv} = \Delta E_{polar} + \Delta E_{nonpolar} \tag{2.90}$$

E_{MM} is the molecular mechanics energy of the system *in vacuo*, which includes the bonded (E_{bond}) and non-bonded ($E_{coul} + E_{LJ}$) terms. Note that in the single-trajectory approach, there is no E_{bond} contribution towards the final ΔG_{bind}, as

it cancels out between G_{PL}, and $G_P + G_L$. ΔE_{solv} accounts for the solvation contribution to the free energy, and is divided into polar (ΔE_{polar}) and nonpolar ($\Delta E_{nonpolar}$) terms. ΔE_{polar} is associated with the difference in system electrostatic energies between vacuum and solvent, and is calculated using the implicit solvent model chosen. $\Delta E_{nonpolar}$ represents instead the energy associated with the creation of a cavity as well as dispersive interactions in solution and in the protein, and is usually calculated as a linear relation to the solvent accessible surface area (SASA):

$$\Delta E_{nonpolar} = \gamma SASA + b \tag{2.91}$$

where γ is typically between 0.01 - 0.24 kJ mol^{-1} Å^{-2}, and b between 0 - 4.2 kJ/mol [230, 241]. This term tends to be small with minor variation among similar ligands, thereby having little influence on the results [230].

The solute entropic contribution S_{MM} can be estimated through quasi-harmonic or normal mode analysis. However, these methods are time-consuming and dominate the computational cost of the MMPBSA calculations if performed on all snapshots considered. Often their energies carry large uncertainties, thus decreasing the precision of the final ΔG_{bind} estimate [33, 230, 241, 245]. In addition, a number of studies have suggested that including this term can lead to only minor improvements in the predictions [31, 32, 246–248]. It is therefore not clear whether the inclusion of the entropy term is beneficial for the calculations, and consequently, considering also the computational burden, it is often ignored. Duan *et al.* [249] have, however, recently suggested a computationally fast approach for the estimation of the entropy contribution that is based on the fluctuation of the protein-ligand interaction energy:

$$-T\Delta S_{MM} = -(TS^{PL}_{MM} - TS^{P}_{MM} - TS^{L}_{MM}) = k_B T ln\langle e^{\beta \Delta E^{PL}_{int}} \rangle \tag{2.92}$$

where ΔE^{PL}_{int} and its ensemble average can cheaply be calculated over all MD frames:

$$\Delta E_{int}^{PL} = E_{int}^{PL} - \langle E_{int}^{PL} \rangle \tag{2.93}$$

$$\langle E_{int}^{PL} \rangle = \frac{1}{N} \sum_{i=1}^{N} E_{int}^{PL}(t_i) \tag{2.94}$$

Considering the ease of computation, we adopted this approach for the estimation of the entropy term in MMPBSA, and evaluated the effect of its inclusion on the agreement between predictions and experiment.

2.6.3 Grand canonical integration

It has been recently shown how water binding free energies can be obtained from GCMC simulations [68]. By simulating a system at different chemical potentials, it is possible to plot the Adams values (see § 2.4.1) against the average number of water molecules present in the grand canonical region. By integrating the area under the curve, the ΔG associated with the transfer of N water molecules from the gas reservoir to the binding pocket is obtained, hence the name *grand canonical integration* (GCI) [68]. The thermodynamic cycle for this method is shown in Figure 2.7a.

The binding of N water molecules to a protein cavity can be defined as follows:

$$\Delta G_{bind}(N) = \Delta G_{trans}(N) - \Delta G_{hyd}(N) \tag{2.95}$$

where $\Delta G_{trans}(N)$ is the free energy of transferring N water molecules from gas phase to the GCMC region, and $\Delta G_{hyd}(N)$ is the hydration free energy of N water molecules. ΔG_{hyd} for a certain water model can be obtained by double decoupling; these values are also available from the literature. Ross *et al.* [68] have shown that the value of $\Delta G_{trans}(N)$ can be obtained from the following equality:

$$\beta \Delta G_{trans}(N_i \rightarrow N_j) = N_j B_j - N_i B_i + ln\left(\frac{N_i!}{N_j!}\right) - \int_{B_i}^{B_j} N(B) dB \tag{2.96}$$

where N_i is an initial number of water molecules present in the GCMC region, N_j is their final number, B is the Adams parameter, and the integral over $N(B)$ is

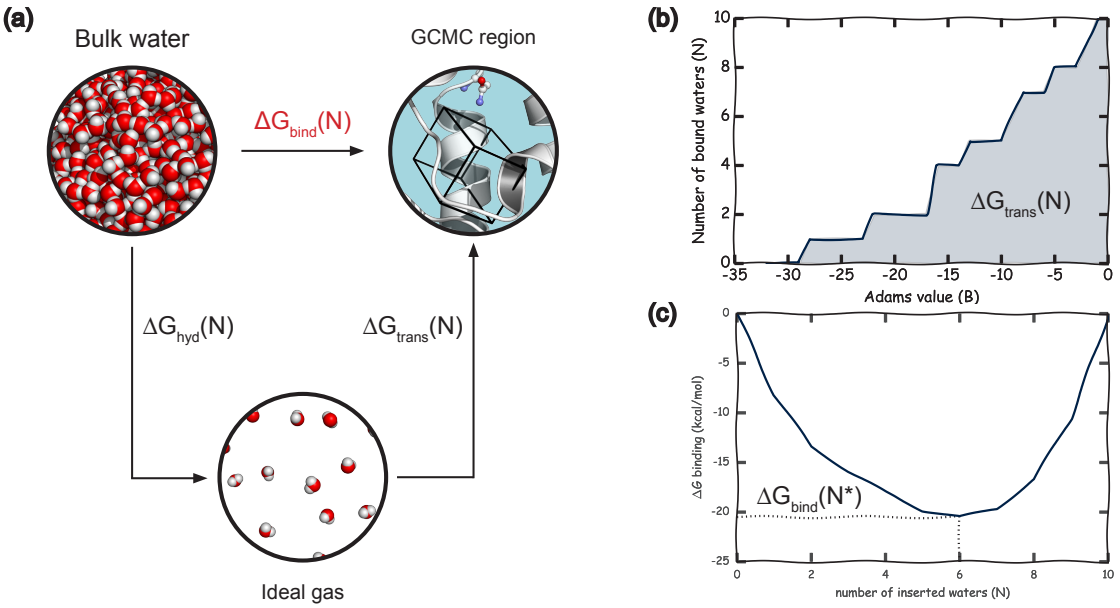


Figure 2.7: Basic principles of GCI. (a) Thermodynamic cycle for water binding free energy calculations using GCI. (b) Chemical potential versus number of bound water molecules, from which $\Delta G_{trans}(N)$ can be obtained via integration. (3) Plot of free energy versus number of waters present in the GCMC region, from which it is possible to identify the equilibrium number of waters and their binding free energy.

the area under the GCMC titration curve as shown in Figure 2.7b. The authors suggest to smooth the titration curve using the following logistic equation:

$$N(B)_{approx} = \sum_{i=1}^{m} \frac{n_i}{1 + exp(w_{0i} - w_i B)} \tag{2.97}$$

where m, n_i, w_{0i}, and w_i are free parameters to be fitted to the titration data. As shown in Figure 2.7c, once $\Delta G_{trans}(N)$ has been obtained, $\Delta G_{bind}(N)$ can too, and by identifying the minimum of this curve it is possible to recover the number of water molecules present in the GCMC region at equilibrium (N^*), as well as the free energy difference associated with their binding to the pocket ($\Delta G_{bind}(N^*)$) [68].

Tools for the setup and analysis of GCMC simulations and grand canonical integration are available as part of the ProtoMS package, and have been used in the work presented in chapter 6 [68, 216].

Costanzia: non chi comincia, ma quel che persevera.

— Leonardo da Vinci

3

Predicting ligand affinities to a single target

Contents

3.1 Introduction 72
3.2 Methods 74
3.2.1 System Setup 74
3.2.2 Free Energy Calculations 75
3.2.3 Data Analysis 77
3.2.4 Docking 78
3.3 Performance when starting from X-ray structures . . . 79
3.4 Performance when starting from docked poses 83
3.4.1 Ligand placement and identification of crystallographic poses 84
3.4.2 Ranking the ligands by affinity 87
3.5 Discussion 90
3.6 Conclusions 93

In this chapter I will report the results of a study where I performed absolute free energy calculations for a set of diverse inhibitors binding to bromodomain-containing protein 4 (BRD4). Good agreement with experiment was obtained in terms of both correlation and absolute error. The performance of the calculations was evaluated retrospectively, starting from the deposited crystal structures of the protein-ligand complexes, and pseudo-prospectively, starting from an apo BRD4 structure and

docking the ligands. Much of this chapter is based upon the following publication: M. Aldeghi, A. Heiftez, M.J. Bodkin, S. Knapp, P.C. Biggin. Accurate calculation of the absolute free energy of binding for drug molecules. *Chemical Science* **2016**, 7, 207-218.

3.1 Introduction

Before this work, absolute binding free energies had been calculated with alchemical methods for a few protein-ligand systems. One of the most studied macromolecular systems has been the engineered binding pocket of T4 lysozyme. Mobley *et al.* [57] studied the binding of thirteen single-ring fragment-like ligands to a L99A hydrophobic T4 lysozyme cavity mutant, obtaining a root mean square (RMS) error compared to isothermal titration calorimetry (ITC) experiments of roughly 1.9 kcal/mol. Boyce *et al.* [58] studied instead the binding of similar fragment-like ligands to a slightly polar model cavity of T4 lysozyme (L99A/M102Q) in a prospective fashion, obtaining a RMS error compared to ITC for the five compounds with measurable affinities of about 1.8 kcal/mol. Another popular test system has been the FK506-binding protein (FKBP12). The series of ligands evaluated with FKBP12 was originally studied experimentally by Holt *et al.* [250] and are drug-like, with multiple rings and several rotatable bonds, although sharing very similar chemical moieties. Shirts first reported a RMS error of about 2.0 kcal/mol for the affinity prediction of nine inhibitors [251], and a following study by Wang *et al.* [252] obtained an error of 2.0-2.5 kcal/mol. For this system the experimental free energies taken as reference were derived from competitive inhibition of FKPB12 activity [250]. Fujitani and coworkers [59] obtained a RMS difference from a linear fit of only 0.4 kcal/mol for eight FKBP-12 inhibitors, however, there was a large offset (-3.2 kcal/mol) relative to experiment. Other calculations have also been reported albeit on smaller numbers of ligands and this makes it harder to establish the actual errors [239, 253–255].

The above mentioned studies were done on a limited number of proteins, and mostly on engineered cavities. In addition, with the fast surge in computer power we

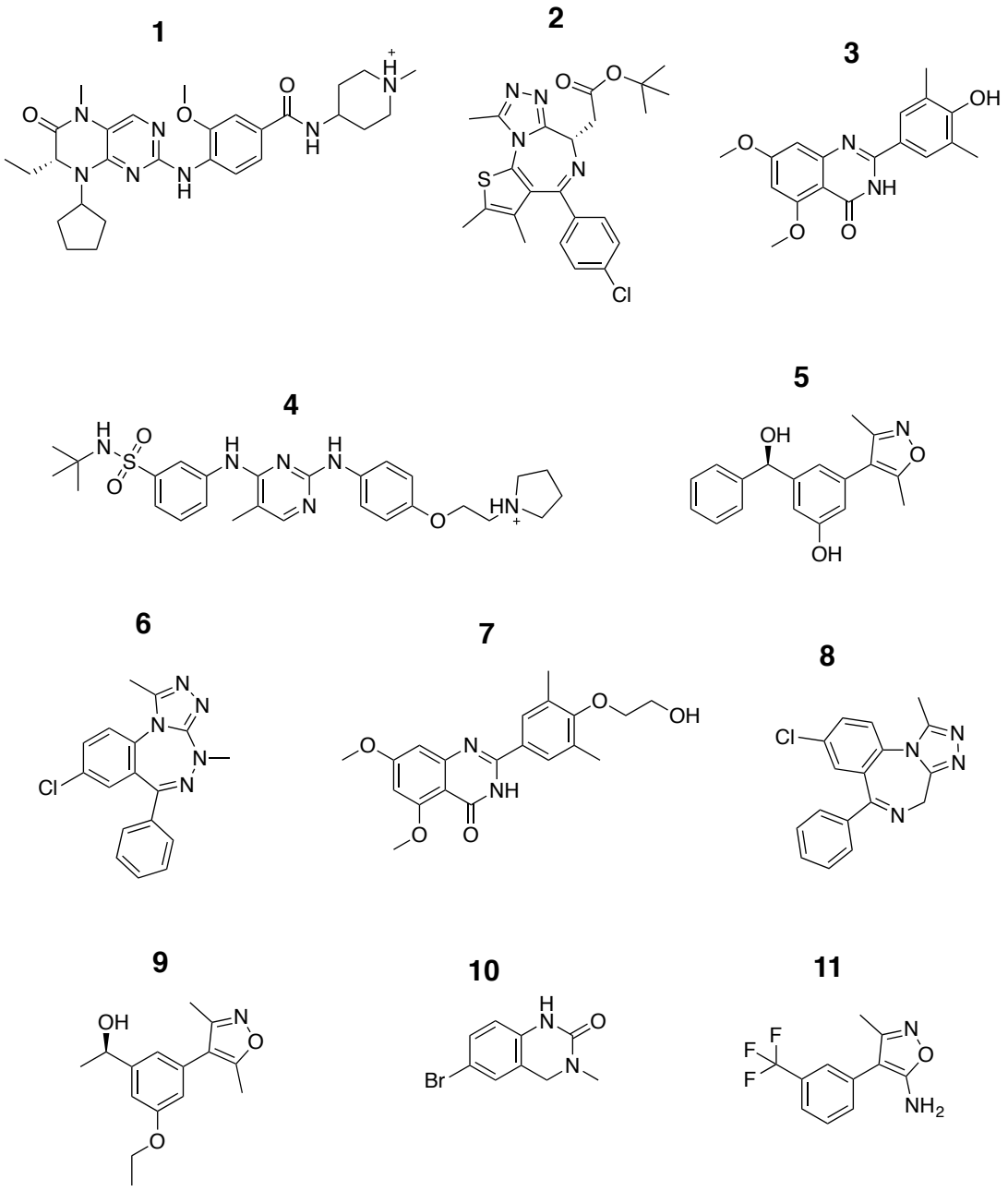


Figure 3.1: Chemical structure of the ligands. The structures of the compounds analysed in this study are shown and are labelled with arabic numerals in descending order of affinity. It is possible to note the presence of a few specific chemotypes. In particular, compounds 1 and 4 are kinase inhibitors carrying the same aminopyrimidine hinge-binding motif; compounds 2 and 8 are benzodiazepines, and compound 6 is a benzotriazepine; compounds 3, 7, and 10 are all quinazolinones; compounds 5, 9, and 11 are based on a methyl-isoxazole scaffold. However, we consider this set of ligands to be diverse in the sense that it does not represent a chemical series where only a few substituents differ between the ligands, but there are a few distinct scaffolds, each with a few representative compounds.

can now access longer timescales and obtain better convergence of the calculations. Therefore, I sought to assess whether absolute free energy calculations based on standard implementations of alchemical transformations are now reaching the point where they can be applied to diverse (i.e. that are not part of the same chemical series), drug-like organic molecules and a pharmacologically relevant target. In order to achieve this goal, I compared predicted binding free energies for 11 diverse small molecule inhibitors (Figure 3.1) that bind to bromodomains (BRDs) with experimental measurements, primarily isothermal titration calorimetry (ITC). I performed a retrospective analysis in order to assess the performance of the calculations in terms of accuracy and precision in a best-case scenario. Subsequently, I carried out a pseudo-prospective study where I repeated the exercise with traditional docking methods to give the initial poses without using any structural information for the protein-ligand complexes. Both studies gave very good agreement with experimental data.

3.2 Methods

3.2.1 System Setup

The initial conformations of the complexes were taken from holo crystal structures available in the Protein Data Bank (PDB) [256] (3U5J, 3U5L, 4OGI, 4OGJ, 3MXF, 4MR3, 4MR4, 3SVG, 4J0R, 4HBV) with the exception of ligand 11, which was modelled based on the structure 3SVG, and from the results of the ligand docking into the apo protein (2OSS). Missing atoms in the crystals were modelled with the WHAT-IF web interface [257, 258] and all organic molecules that were not the ligand of interest were removed from the system, whereas all crystallographic waters were retained. After adding hydrogens with Maestro (v9.5, Schrödinger), ligands were parameterised with the general AMBER force field (GAFF v1.5) [259] and AM1-BCC charges [260, 261] using AmberTools12 [262] provided with the FESetup tool v1.1pre1 [263]. GROMACS topologies and coordinates were generated from the AMBER ones using *acpype* (v.2013-11-28 Rev: 399) [264]. We used the Amber99SB-ILDN force field [265] for the protein and the TIP3P model [179] for water molecules.

The protonation state of the ligands was based on the major species in solution as predicted by pK_a calculations using Marvin Sketch (v6.1.0, ChemAxon). Ligand 11 has a primary amine attached to the oxazole ring. This group may be thought to be protonated in solution due to the basic nature of amines. However, arylamines are less basic than alkylamines, as the lone pair of electrons on the nitrogen is conjugated to the π-electrons of the aromatic ring and are therefore less available for protonation, which disrupts the conjugation [266]. In fact, pK_a calculations predicted the compound to present only in its neutral form at pH 7.0. The tautomers employed were instead chosen manually based on chemical intuition and what was reported in the literature. A better approach would have been to calculate the major tautomer species present in solution using available software, similarly to what was done for the protonation states. Nonetheless, when tautomer calculations were performed (using Marvin Sketch) after the simulations had already run, it was found that the tautomers used were the single major species for all ligands considered. More in general, however, it is important to note that in some cases there might not be a single dominant tautomer or protonation state, but multiple states may be present in similar ratios. In this circumstance, one could run multiple calculations in order to predict the binding free energies of the different species, and then combine these into a single free energy difference. However, in the long term a better approach would be to be able to take the equilibrium between multiple protonation and tautomeric states into account during a single simulation [267–269].

The complexes were then solvated in a dodecahedral box, apart from ligands 1 and 4 that were solvated in a cubic box, with periodic boundary conditions and a minimum distance between the solute and the box of 12 Å. Sodium and chloride ions were added to neutralise the systems.

3.2.2 Free Energy Calculations

Absolute binding free energy calculations were performed beginning from both crystal ligand poses and docked poses as detailed in the Results section, using the non-physical thermodynamic cycle illustrated in Figure 2.5. All simulations

were carried out in GROMACS 4.6.5 [204, 270]. The ligand van der Waals (vdW) interactions were decoupled and the charges annihilated using a linear alchemical pathway with $\Delta\lambda = 0.05$ for the van der Waals and $\Delta\lambda = 0.1$ for the coulombic transformations. For the addition of the ligand restraints instead, 12 non-uniformly distributed λ values were used (0.0, 0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2, 0.3, 0.5, 0.75, 1.0). A total of 42 windows for the complex simulations and 31 windows for the ligand simulations were therefore employed. For each window, 10,000 energy minimisation steps were carried out using a steepest descent algorithm. The system was subsequently simulated for 0.5 ns in the canonical ensemble with harmonic position restraints applied to the solute heavy atoms with a force constant of 1,000 kJ mol^{-1} nm^{-2}. Temperature was coupled using Langevin dynamics [192, 193] with 298.15 K as the reference temperature. A 1 ns position restrained run in the isothermal-isobaric ensemble was then performed using the Berendsen weak coupling algorithm [198]. 10 ns unrestrained production runs were performed for data collection using Hamiltonian-exchange Langevin dynamics with a 2 fs time-step in the NPT ensemble with the Parrinello-Rahman pressure coupling scheme [202]. 3 million swaps between any state pair were attempted every 1000 time steps, following the Gibbs sampling scheme proposed by Chodera and Shirts [225]. This led to acceptance rates between neighbouring states that ranged from 0.1 to 0.3 (mean and standard deviation of 0.2 +/- 0.1), but a probability of jumping to any other state from 0.2 to 0.9 (mean and standard deviation of 0.7 +/- 0.2). This resulted in a total unrestrained simulated time of 43 μs for this study. The relative position and orientation of the bound ligand with respect to the protein were restrained by means of one distance, two angles and three dihedral harmonic potentials with force constant of 10 kcal mol^{-1} $Å^{-2}$ [rad^{-2}]. The contribution of this set of restraints to the free energy can be calculated analytically as described by Boresch *et al.* [233] for the non-interacting ligand in solution (ΔG^{solv}_{restr}), while it has to be evaluated numerically for the interacting ligand in complex with the protein (ΔG^{prot}_{restr}). The equation used to evaluate this contribution also includes a correction for the standard-state dependence of the binding free energy [233]. A soft-core

potential [271] was employed for the van der Waals interactions transformed. For all simulations the particle mesh Ewald (PME) algorithm [210] was used for electrostatic interactions with a real space cut-off of 12 Å, a spline order of 6, a relative tolerance of 10^{-6} and a Fourier spacing of 1.0 Å. A switch function between 9 Å and 10 Å was used for the van der Waals interactions. The P-LINCS constraint algorithm [196] was used only on bonds involving hydrogens (H-bonds). The GROMACS long-range dispersion correction for energy and pressure was used, and an additional long-range dispersion correction (EXP-LR) was applied as described in Shirts *et al.* [205]. For charged compounds (1 and 4), the semi-analytical correction scheme for electrostatic finite size effects proposed by Rocklin *et al.* [213] was employed. The residual integrated potential (RIP) was calculated for 11 frames (every 1 ns, from 1 ns to 10 ns) using APBS 1.3 [272]. Only the mean RIP value was used for the correction; its standard deviation, resulting from the different conformations of the ligand and complex, was always below 0.05 kcal/mol.

3.2.3 Data Analysis

The results were analysed with the implementation of the multiple Bennet acceptance ratio (MBAR) provided with the python package *pymbar* [240] and the alchemical-analysis tool [273]. The first 1 ns of each window was discarded as equilibration period. Prior to the free energy estimate, the data from each lambda state were subsampled in order to include only uncorrelated data-points by calculating the autocorrelation time and statistical inefficiency of the potential energy. 200 boot-strap sets were constructed by random resampling with replacement of the uncorrelated data, where the first set was the original sample [274]. The free energy was estimated with MBAR for all of the 200 sets and the final estimate is the mean of all these free energy estimates. The error of the final estimate is the sample standard deviation of the estimates for all bootstrap samples. For the calculations beginning from the crystal ligand poses, the whole calculations have been repeated three times in order to assess the convergence of the results. The combined free energy estimate and standard deviation for ligand and complex simulations were determined by

taking the mean and sample standard deviation of all the 600 bootstrap samples. Therefore, the final uncertainty is representative of both the statistical uncertainty of the MBAR free energy estimate and the error due to finite sampling. Ligand 11 is an exception, as it was modelled based on ligand 9 and the PDB structure 3SVG, and two equally plausible binding modes were present, where the trifluorotoluene moiety is flipped by 180 degrees. Therefore, two calculation repeats were carried out per binding mode, resulting in a total of four binding free energy calculations; the results of multiple binding modes can be combined in a single binding free energy value as described by Mobley *et al.* [275]. The final binding free energy for each ligand is the difference between the decoupling of the ligand from the water solution and from the solvated complex; the final error in the binding free energies is thus the root sum square of the uncertainties of ligand and complex calculations. 95% confidence intervals (CI) for the performance statistics (mean unsigned error (MUE), root mean square error (RMSE), Pearson correlation (r_p), Spearman correlation (r_s)) were calculated with percentile bootstrap, after building 10^5 bootstrap samples by resampling the experimental and calculated free energies according to their mean and standard deviations and assuming normality. Throughout the text I will report the 95% CI in square brackets immediately after the statistics.

3.2.4 Docking

Rigid docking was performed with MOE v2013.08 using a crystal structure (PDB 2OSS) of the apo protein of BRD4(1). The five highly conserved crystal waters present in all BET bromodomains binding pockets were kept, whereas all other waters and organic molecules were removed. The 2D chemical structures of the ligands were drawn in Marvin Sketch (v6.1.0, ChemAxon) and a stochastic conformational search was performed in order to generate the 3D conformations. The number of conformations was limited to a maximum of 100 per ligand and duplicate conformations (RMSD < 0.25 Å) were removed. The docking protocol employed the pharmacophore placement method and the London ΔG scoring function. Each binding pose was then minimised and rescored with the GBVI/WSA ΔG scoring

function. The pharmacophore query was built based on the properties of the acetyl-lysine found in the PDB structure 3UVW, and consisted of a hydrogen-bond acceptor site, to mimic the acetyl oxygen, and a non-polar site, corresponding to the position of the methyl moiety. The protein was parameterised using Amber ff99SB [276]. The ligand bonded parameters were obtained with 2D Extended Hückel Theory [277], the vdW parameters were derived from GAFF [259] and the charges from Bond Charge Increments [278] according to the AMBER10:EHT force field option in MOE. Duplicate poses were automatically removed based on their hydrogen-bond and hydrophobic patterns, and poses with positive binding free energy as predicted by the GBVI/WSA ΔG scoring function were removed too, as they typically involved clear clashes with the protein atoms. The remaining poses were furthermore clustered by RMSD with a 3 Å cut-off in order to have a coarser landscape of the possible binding poses, considering we were not interested in running the free energy calculations on similar binding orientation that can interconvert within the simulations timescale. Thus, only the best scoring pose within each cluster would be used for free energy calculations. This procedure aimed at reducing the number of calculations to run while maximising the chances of retaining the poses that closely approximate the crystal.

3.3 Performance when starting from X-ray structures

In this part of the study I carried out alchemical binding free energy calculations starting from the crystal structures of BRD4(1) in complex with 11 inhibitors to a common binding site (Figures 3.1 and 3.2). I was first interested in evaluating the performance of the predictions in a favourable scenario, that is, when the binding conformation is known from experiment. The results from this study, therefore, provide a picture of the best accuracy that can be expected given the physical model used. In addition, I was interested in evaluating the precision of the calculations, considering that large and flexible molecules are present in the test set. Large uncertainties in the results when dealing with drug-like molecules would indeed

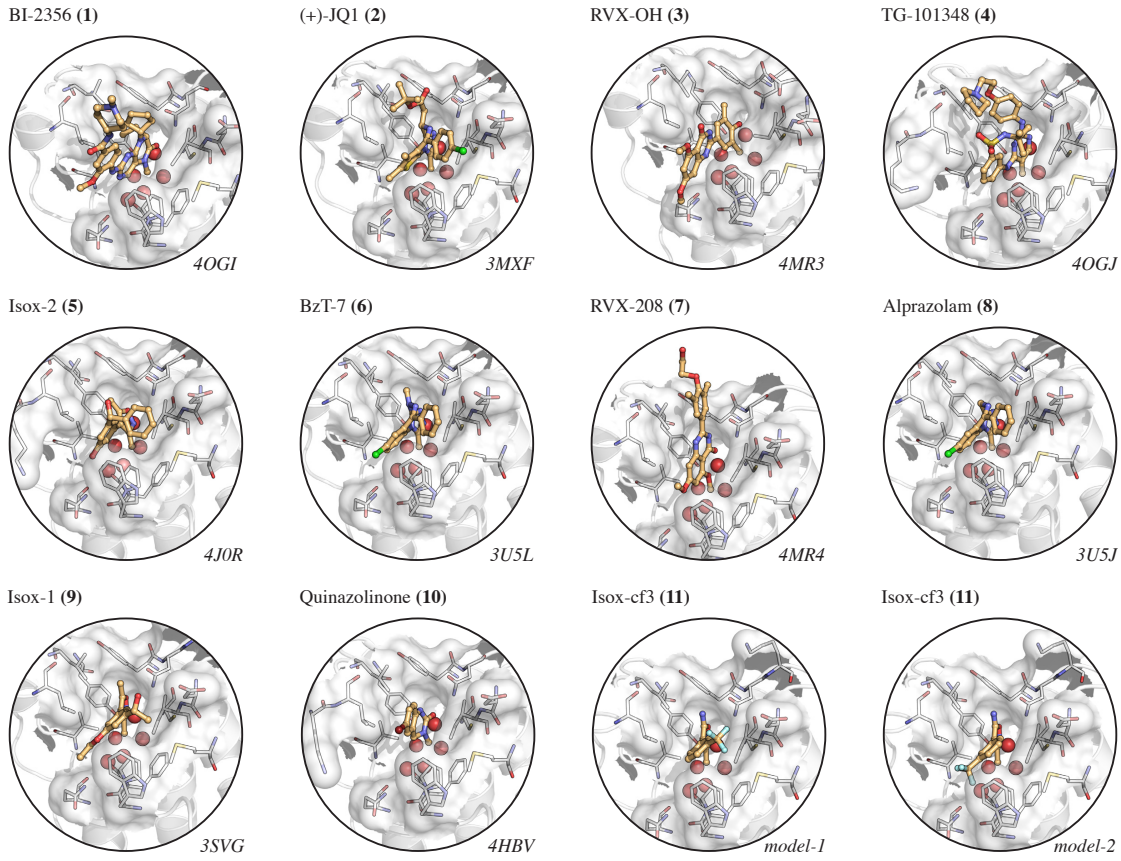


Figure 3.2: Crystal structures used as starting conformations for the free energy calculations. The common or reference names for the ligands used in this work are above the representations of each protein-ligand complex, along with their compound IDs in bold and between brackets. At the bottom-right side of each representation is instead the PDB-ID of each X-ray structure used.

prevent a meaningful assessment of the accuracy of the results. To this end, in addition to the bootstrap analysis to evaluate the statistical uncertainty of the free energy estimator, I repeated the calculations three times, in order to obtain an approximation of the uncertainty due to finite sampling (for ligand 11, four repeats were performed, as explained in § 3.2.1). It was in fact noticed that while bootstrap provided a more realistic uncertainty estimate than the MBAR error estimate alone, it still underestimated the sample standard deviation. Each calculation was the results of 73 10-ns long all-atom molecular dynamics runs for a total simulated time of about 25 μs for this portion of the study.

The set of inhibitors considered comprises mostly drug-like molecules with

Table 3.1: Summary of free energy calculation results based on crystal structures. ΔG_{calc} is the calculated standard binding free energy; ΔG_{exp} is the experimental standard binding free energy (in square brackets, at the end of each line, is the reference for this latter quantity). Reported are also the PDB files used as input and the experimental method used for the affinity measurement. All values are in kcal/mol. All ΔG errors are one standard deviation. $^{\alpha}$ This is an estimate of the typical ITC standard deviation (1σ) based on the variability of the affinity values observed in the ABRF-MIRG002 inter-laboratory assessment [174]; $^{\beta}$ the error represents the standard deviation of two measurements; $^{\gamma}$ error reported as only a single experiment was performed. Values for the difference between ΔG_{calc} and ΔG_{exp} might appear inconsistent due to rounding. In square brackets are reported the 95% confidence intervals of the statistics based on percentile bootstrap.

Cpd	ΔG_{calc}	ΔG_{exp}	$\Delta G_{calc} - \Delta G_{exp}$	PDB	Exp method
1	-10.4 ± 0.6	$-9.8 \pm 0.1^{\alpha}$	-0.6 ± 0.6	4OGI	ITC[138]
2	-9.5 ± 0.4	$-9.6 \pm 0.1^{\alpha}$	0.2 ± 0.4	3MXF	ITC[87]
3	-9.2 ± 0.5	$-9.0 \pm 0.1^{\alpha}$	-0.2 ± 0.5	4MR3	ITC[134]
4	-9.4 ± 0.8	$-8.9 \pm 0.1^{\alpha}$	-0.4 ± 0.8	4OGJ	ITC[138]
5	-8.6 ± 0.3	$-8.8 \pm 0.1^{\beta}$	0.2 ± 0.3	4J0R	SPR[123]
6	-9.9 ± 0.8	$-8.2 \pm 0.1^{\alpha}$	-1.7 ± 0.8	3U5L	ITC[279]
7	-5.9 ± 0.5	$-7.8 \pm 0.1^{\alpha}$	2.0 ± 0.5	4MR4	ITC[134]
8	-7.8 ± 0.3	$-7.4 \pm 0.1^{\alpha}$	-0.4 ± 0.3	3U5J	ITC[279]
9	-7.7 ± 0.4	$-7.3 \pm 0.0^{\beta}$	-0.4 ± 0.4	3SVG	AlphaScreen[123]
10	-5.9 ± 0.2	$-6.3 \pm 0.1^{\beta}$	0.4 ± 0.3	4HBV	AlphaScreen[136]
11	-5.4 ± 0.2	$-5.6 \pm 0.0^{\gamma}$	0.1 ± 0.2	Model	AlphaScreen[280]
Statistics					
Mean Unsigned Error					$0.60[0.47, 0.95]$
Root Mean Square Error					$0.85[0.64, 1.32]$
Pearson r_p					$0.87[0.73, 0.92]$
Spearman r_s					$0.84[0.70, 0.94]$

a diverse range of physicochemical properties: number of atoms from 22 to 77; molecular weight from 241 to 525 Da; number of rotatable bonds from 0 to 11; calculated logP from −0.4 to 5.3 (Table A.1). The range of affinities includes micromolar binders such as ligand 10 (∼80 μM) and 11 (∼23 μM), down to low nanomolar binders such as ligand 1 (∼40 nM) and 2 (∼80 nM). A number of different chemical groups are represented and the dissimilarity of the set provides more confidence that the results obtained are not excessively biased by the limited chemical space considered.

Table 3.1 summarises the results obtained for this retrospective study (full breakdown of the results is available in Table A.2). Most calculations agree extremely

well with the experimentally determined values. Seven out of eleven predictions have errors below 0.5 kcal/mol, and all prediction errors are below 2.0 kcal/mol. This resulted in a MUE of 0.60 [0.47, 0.95] kcal/mol and a RMSE of 0.85 [0.64, 1.32] kcal/mol. The calculated free energies strongly correlate with the experimental ones, as shown in Table 3.1 and Figure 3.5, with a Pearson correlation (r_p) of 0.87 [0.73, 0.92], and manage to rank the ligand affinities effectively with a Spearman correlation (r_s) of 0.84 [0.70, 0.94]). The precision of the calculations is encouraging too, as in only three instances the uncertainty is above 0.5 kcal/mol, and in all case it is below 1.0 kcal/mol.

The predictions for ligands 6 and 7 showed large deviations from the ITC measurements (Table 3.1), with the binding free energy of ligand 6 being overestimate by 1.7 kcal/mol, and the binding free energy of ligand 7 being underestimated by 2.0 kcal/mol. Among the ligand parameters that can affect the accuracy of the physical model are the partial charges used, which affect the inter-and intra-molecular coulombic interactions, with an effect on the energy of different ligand geometries, and the torsional parameters, which define the location of minima and maxima, and the energy barriers around dihedral angles. In fact, it has already been noted in the literature that charges and torsional parameters in particular, can be problematic in small molecule force fields due to their limited transferability [281, 282]. Indeed, when comparing the GAFF energy profile for the biaryl torsion in ligand 7 to QM calculations (HF/6-31G* level of theory), I noticed that the energy barrier to the rotation at 90° and 270° was underestimated. After derivation of improved parameters for this biaryl torsion by fitting to the HF/6-31G* energy profile, the calculated binding free energy shifted toward a more negative value (-6.5 ± 0.4 kcal/mol) and 0.6 kcal/mol closer to the experimental free energy.

Among the additional challenges related instead to ligand 6 is the presence of a chlorine atom that could engage in halogen bonds. Halogen bonds are directional and attractive intermolecular interactions between a halogen atom and a neutral or negatively charged Lewis base. The electropositive region, termed the σ *hole*, is

larger in the heavier and more polarisable halogens, and is enhanced by electron-withdrawing groups covalently attached to the halogen atom [283, 284]. GAFF, like other small molecule force fields, cannot reproduce this type of interaction as the atom-centered partial charge model adopted is inherently symmetrical. Thus, as it has been discussed in the literature [285–287], one way to model the σ hole is through the use of off-center point charges, which break the charge symmetry on halogen atoms. This solution has been recently adopted by the OPLS3 force field, which showed improved performance in hydration and relative binding free energy calculations for small molecules containing halogens [288]. From the X-ray structure of ligand 6 in complex with BRD4(1) (PDB-ID 3U5L) it is not possible to identify a clear halogen bond between the protein and the ligand. Nonetheless, the poor representation of this molecular interaction can also affect the hydration free energy of the compound, as well as its interaction energy with any protein's residue, ultimately affecting the overall binding free energy.

3.4 Performance when starting from docked poses

In order to have an idea of the usefulness of these calculations in a prospective context, I carried out a docking and free energy calculation exercise based upon docking results rather than crystal structures. The main objective of this portion of the study was to evaluate whether the accuracy observed in the retrospective exercise could be achieved with a more prospective docking-based approach. Three-dimensional ligand structures were therefore generated through a conformational search after drawing the ligands in two dimensions. A pharmacophore based on acetyl-lysine was then created and used to aid the docking of the 11 inhibitors to an apo structure of BRD4(1) (PDB-ID 2OSS). Unrestrained molecular dynamics simulations explore binding conformations close to each other, hence the docking poses obtained were clustered by RMSD in order to avoid selecting conformations that interconvert during the simulations time scales. The total number of docked poses obtained was 72, which was reduced to 25 after the clustering procedure and removal of poses with positive binding free energies as predicted by the scoring

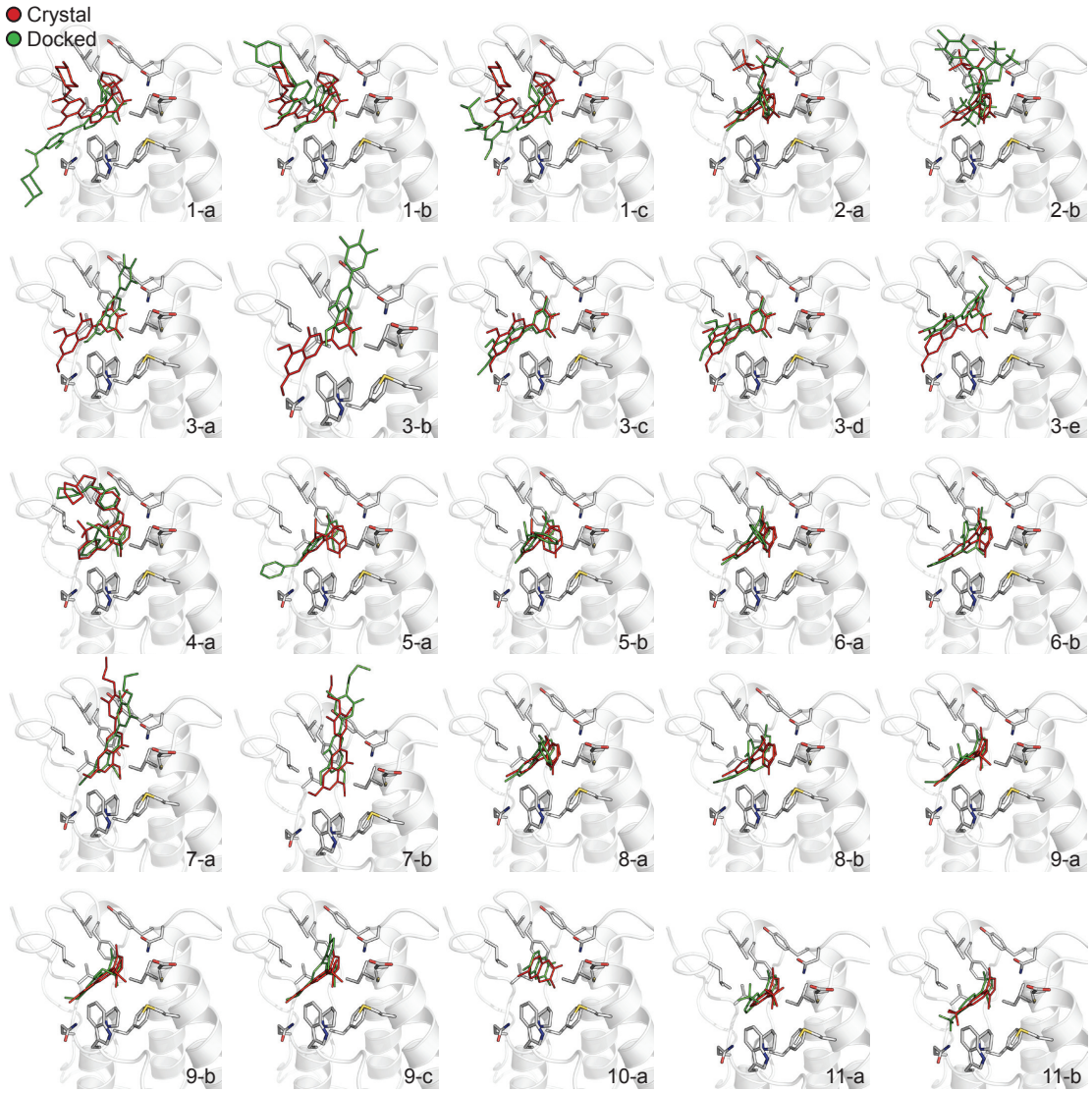


Figure 3.3: Binding poses suggested by docking. In red are the crystallographic structures, and in green are the docked ligands. The ligand number and cluster letter are reported on each pose.

function, as they typically involved clear atom clashes. Figure 3.3 shows these 25 docking poses, and their scores and RMSD to crystal are summarised in Table 3.2.

3.4.1 Ligand placement and identification of crystallographic poses

For all ligands, a binding pose that captures the main features of its interactions with BRD4(1) was among the docking results. The docking software managed to

reproduce the binding mode of the inhibitors well. For eight out of ten ligands the RMSD was below 2.0 Å. In addition, docking correctly identified the pose closest to the crystallographic observed one as being the most favourable for 7 out of 10 ligands (Table 3.2). For ligands 1, 3, and 5, the function used by the docking program scores the experimentally observed pose lower than a different one. On the other hand, free energy calculations identified correctly the most stable binding pose 9 times out of 10. For ligand 6, the MD-based calculations returned a similar binding affinity for two very different poses, probably overestimating the affinity of the pose that is not observed crystallographically.

For ligand 1, two poses actually capture the correct binding mode of the molecules within the BRD4(1) binding pocket (poses 1-a and 1-c), despite the fact that pose 1-a was substantially different from the crystal (RMSD of 8.4 Å). The specific pose observed in the X-ray structure is in fact probably due to crystal contacts between two symmetry-related BRD4(1) molecules. Large part of this inhibitor is solvent exposed and free to explore a number of conformations. While the sampling during the simulations can capture the fact the two poses are similar, despite the large RMSD with respect to the X-ray structure, the docking score cannot and thus assigns different scores to the two poses, suggesting pose 1-a as being substantially more stable than pose 1-c. In the free energy calculations instead, pose 1-a is still assigned a large binding free energy.

The case of ligand 3 is presented in more detail, as it is more complex. Ligands 3 and 7 are closely related; in fact, ligand 3 is the synthetic precursor of ligand 7. Despite their chemical similarity, the two ligands bind the BRD4(1) binding pocket in two very different modes, as shown in Figure 3.4. This substantial change in binding pose is extremely hard to predict by visual inspection or docking alone. Indeed, docking proposed pose 3-a as the most favourable binding pose for ligand 3 with a score of -4.9 kcal/mol, which resembles the pose of ligand 7 (Figure 3.4b), forming two hydrogen bonds with N140 through the dihydroquinazolinone scaffold and burying a methoxy group at the bottom of the pocket. The actual binding mode of the ligand is correctly represented instead by the pose 3-c, which is assigned

Table 3.2: Docking results. Shown are the poses retained after the clustering of the results and ranked by the scoring function used as well as the free energy calculations. An arrow indicates the poses with lowest RMSD, which most closely represent the crystallographic binding conformations.

Ligand	Pose	RMSD to crystal (Å)	Docking Score (kcal/mol)	ΔG_{calc} (kcal/mol)
1	a	8.4	−2.69	-10.3 ± 0.8
	b	4.1	−2.55	-4.6 ± 0.8
	→ c	3.2	−0.79	-10.9 ± 0.8
2	→ a	1.9	−4.67	-10.1 ± 0.4
	b	6.8	−1.73	-6.7 ± 0.4
3	a	6.8	−4.93	-6.2 ± 0.2
	b	7.8	−3.52	-6.5 ± 0.3
	→ c	2.0	−2.57	-10.8 ± 0.2
	d	3.0	−2.46	-10.5 ± 0.2
	e	7.8	−1.57	-7.3 ± 0.2
4	→ a	1.8	−3.40	-9.0 ± 0.8
5	a	5.6	−4.20	-5.8 ± 0.2
	→ b	1.2	−4.18	-8.3 ± 0.2
6	→ a	1.0	−4.58	-10.3 ± 0.3
	b	5.0	−4.50	-10.6 ± 0.3
7	→ a	2.2	−5.18	-6.6 ± 0.3
	b	3.9	−3.80	-5.1 ± 0.1
8	→ a	0.8	−4.15	-10.2 ± 0.2
	b	5.0	−2.36	-6.6 ± 0.1
9	→ a	1.8	−4.19	-7.7 ± 0.1
	b	4.2	−3.94	-7.3 ± 0.2
	c	4.7	−2.11	-6.1 ± 0.1
10	→ a	0.7	−3.19	-6.2 ± 0.1
11	a	n.a.	−4.76	-5.3 ± 0.1
	b	n.a.	−4.44	-5.4 ± 0.1

a worse docking score (−2.6 kcal/mol), and it is characterised by the formation of one hydrogen bond with N140 thanks to the oxygen of the dimethylphenol ring, and the burial of a methyl group in the hydrophobic pocket in an analogous fashion to the binding of the acetyl moiety in Kac. Absolute free energy calculations were carried out starting from all docking poses of ligand 3, as for the other ligands, including pose 3-a and pose 3-c. The binding free energy obtained for the pose that approximates the bound structure in the crystal (pose 3-c), was -10.8 ± 0.2 kcal/mol, whereas the free energy for the pose that binds BRD4(1) similarly to ligand 7 (pose 3-a) was estimated to be -6.2 ± 0.2 kcal/mol. The results therefore unequivocally

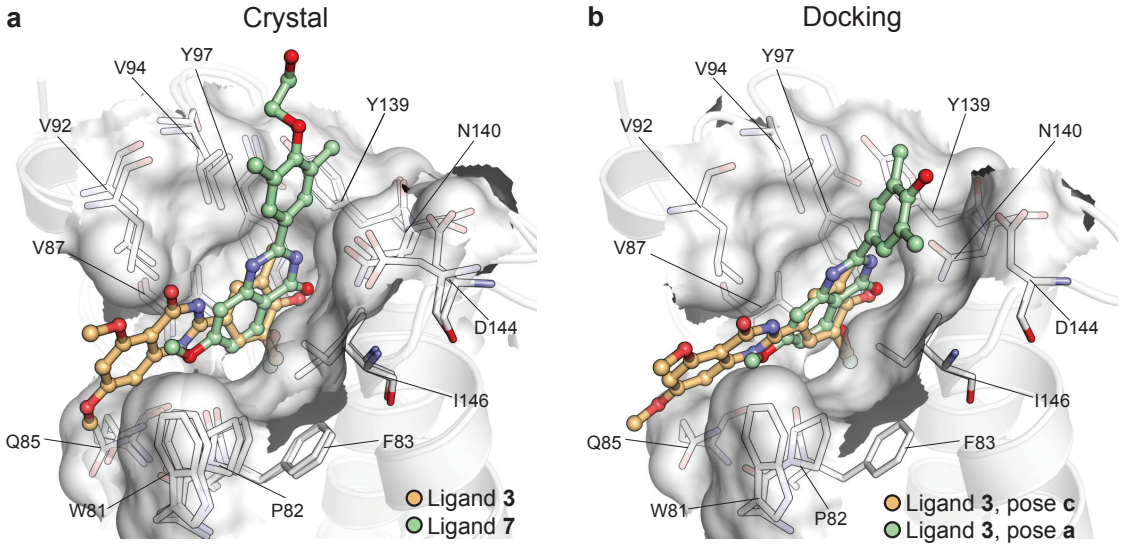


Figure 3.4: Multiple potential binding conformations of ligand 3. (a) Overlay of the crystal structures for ligand 3 and ligand 7 (4MR3 and 4MR4 respectively). (b) Overlay of docking poses 3-a and 3-c.

identified the crystallographic binding pose as being the most favourable one.

For ligand 5, docking scores the two proposed poses similarly, despite pose 5-a being far from the crystallographic structure. This is not the case for the free energy calculations, which identify pose 5-b as being the most stable by about 2.5 kcal/mol.

There is only one case where the free energy calculations were unable to unambiguously identify the most stable binding pose and that is ligand 6. In this case both the scoring function and the MD suggest that the two poses (6-a and 6-b) have similar binding affinity for BRD4(1). Interestingly, ligand 6, when compared to the similar ligand 8, has an additional methyl group on its triazepine ring that can potentially mimic the methyl moiety of the acetylated lysine. Indeed, pose 6-b binds the pocket placing this methyl group similarly to Kac. Pose 6-b might therefore be a legitimate secondary binding pose, even though its binding affinity is probably largely overestimated.

3.4.2 Ranking the ligands by affinity

When using the pharmacophore placement method and the GBVI/WSA ΔG scoring function implemented in MOE 2013.08, docking, despite having provided good

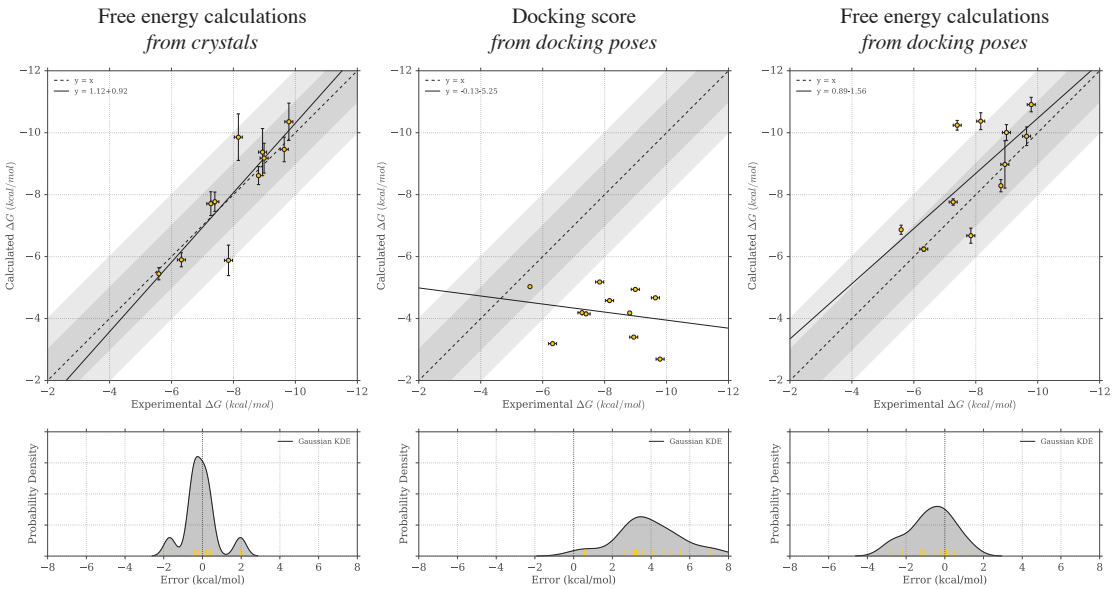


Figure 3.5: Scatter and correlation plots of the results. Correlation plots for the free energy calculations starting from the X-ray structures (left), the docking free energy scores (middle), and the free energy calculations starting from the docked structures (right). Below the scatter plots are rug plots of the errors with a Gaussian kernel density estimate (KDE) of the data.

binding conformations for the ligands, and fairly good relative pose ranking for the same ligand, scored the ligands inaccurately (Figure 3.5(middle)). With RMSE = 4.2 kcal/mol and $r_p = -0.16$, the affinities provided do not help in discriminating between tight and weak binders. Conversely, the free energy calculations based on MD still achieved good agreement with the experimental affinities (Figure 3.5(right)). Table 2 (full breakdown in Table A.3) reports the results of the free energy calculations based on the highest affinity pose for each ligand, along with the RMSD of the poses with respect to the crystal structure, and the affinity predicted by the docking scoring function for the same ligand. The MUE and RMSE for the free energy calculations were respectively 1.01 [0.90, 1.29] kcal/mol and 1.37 [1.25, 1.64] kcal/mol, whereas the correlation to experimental values was of 0.78 [0.68, 0.83] for Pearson's r and of 0.78 [0.55, 0.85] for Spearman's ρ. At the bottom of Figure 3.5 the distribution of the errors in the calculations are shown. The errors when starting from X-ray structures were distributed around zero. When starting from the docking poses, there was a slight tendency for overestimation

Table 3.3: Summary of the free energy calculation results based on docking. Shown are the data for the predicted most stable binding poses for each ligand. ΔG_{calc} is the calculated standard binding free energy; ΔG_{exp} is the experimental standard binding free energy. For comparison, also the affinities predicted with the docking scoring function are reported. All values are in kcal/mol. All errors are one standard deviation. "Pose" indicates whether the lowest energy pose identified corresponds to the crystallographically observed binding mode; also the RMSD of the pose as compared to the crystal is reported. $^{\alpha}$ This is an estimate of the typical ITC standard deviation (1σ) based on the variability of the affinity values observed in the ABRF-MIRG002 inter-laboratory assessment [174]; $^{\beta}$ the error represents the standard deviation of two measurements; $^{\gamma}$ error reported as only a single experiment was performed. Difference values may include rounding effects.

Cpd	ΔG_{calc}	ΔG_{exp}	$\Delta G_{calc} - \Delta G_{exp}$	Pose	RMSD (Å)	Dock ΔG
1	-10.9 ± 0.8	$-9.8 \pm 0.1^{\alpha}$	-1.1 ± 0.8	Yes	3.2	-2.7
2	-10.1 ± 0.4	$-9.6 \pm 0.1^{\alpha}$	-0.5 ± 0.4	Yes	1.9	-4.7
3	-10.8 ± 0.2	$-9.0 \pm 0.1^{\alpha}$	-1.8 ± 0.2	Yes	2.0	-4.9
4	-9.0 ± 0.8	$-8.9 \pm 0.1^{\alpha}$	-0.0 ± 0.8	Yes	1.8	-3.4
5	-8.3 ± 0.2	$-8.8 \pm 0.1^{\beta}$	0.5 ± 0.2	Yes	1.2	-4.2
6	-10.6 ± 0.3	$-8.2 \pm 0.1^{\alpha}$	-2.4 ± 0.3	No	5.0	-4.6
7	-6.6 ± 0.3	$-7.8 \pm 0.1^{\alpha}$	1.2 ± 0.3	Yes	2.2	-5.2
8	-10.2 ± 0.2	$-7.4 \pm 0.1^{\alpha}$	-2.8 ± 0.2	Yes	0.8	-4.2
9	-7.7 ± 0.1	$-7.3 \pm 0.0^{\beta}$	-0.5 ± 0.1	Yes	1.8	-4.2
10	-6.2 ± 0.1	$-6.3 \pm 0.1^{\beta}$	0.1 ± 0.2	Yes	0.7	-3.2
11	-5.4 ± 0.1	$-5.6 \pm 0.0^{\gamma}$	0.2 ± 0.1	n.a.	n.a.	-4.8
Statistics						
Mean Unsigned Error						1.01 [0.90, 1.29]
Root Mean Square Error						1.37 [1.25, 1.64]
Pearson's r						0.78 [0.68, 0.83]
Spearman's ρ						0.78 [0.55, 0.85]

of the affinities; this might however be due to chance, considering that this set of calculations was run once and not in triplicates. On the other hand, the docking scoring function considered here systematically underestimated the ligand affinities.

As mentioned previously in this chapter, MD simulations and free energy calculations can be affected by a ligand's charges and torsional parameters, due to the limited transferability of these across even apparently similar molecules. In particular, torsions between aromatic groups have been identified to be most problematic as they are often poorly parameterised in common small molecule force fields [281, 282]. When starting from the crystal structure of ligand 7, the use of re-fitted torsional parameters resulted in improved agreement with the

experimental binding free energy. Learning from this observation, ligand 7 was later re-parametrised (yet slightly differently) also for the calculations starting from its docking pose, for use in a different study that will be presented in the next chapter. Here, however, I anticipate some of the results as they are relevant for the discussion in this chapter too.

For ligand 3 and 7, restrained electrostatic potential (RESP) partial charges were derived by HF/6-31G* geometry optimisation, followed by a molecular electrostatic potential (MEP) calculation at the HF/6-31G* level of theory. The biaryl torsion present in these molecules was also re-parametrised by fitting to a relaxed torsion scan at the MP2/6-31G* level of theory (Gaussian 09, Rev D.01) [289] performed every 5 degrees. Additional information and discussion on the parametrisation procedure and the resulting parameters are in Chapter 4, where these parameters have been used more extensively for the prediction of the binding free energy of ligands 3 and 7 to multiple BRDs. When considering BRD4(1) as the binding partner, as in this chapter, the new set of parameters returned slightly improved results. In fact, the predicted binding free energy for ligand 7 was of -6.8 ± 0.1 kcal/mol, showing a small improvement of 0.2 kcal/mol. For ligand 3, the predicted binding free energy was of of -9.9 ± 0.1 kcal/mol, showing a larger improvement of 0.9 kcal/mol. Thus, overall, the new set of parameters improved the agreement of the predictions with experiment, by shifting the binding free energies of the two ligands in opposite directions. The improvement for ligand 7 was in this case not as large as what was observed when starting from the X-ray structure. However, the two calculations started from different initial structures, which, combined with finite sampling, can result in the generation of different ensembles. Secondly, the uncertainty in the calculations means that part of the improvements (or lack thereof) observed might be caused by chance.

3.5 Discussion

As discussed by Mobley and Klimovitch [44], reliable binding free energy predictions can have a substantial impact in drug discovery campaigns even with modest levels

of accuracy. In a lead optimisation exercise, screening ∼10-100 molecules per week with 2.0 kcal/mol of noise would reduce the synthetic effort by a factor of 3 when the goal is to achieve a 10-fold improvement in binding affinity (i.e. a 1.4 kcal/mol improvement in binding free energy). Moreover, absolute calculations need only structural information of the target in order to be employed. Despite currently still being computationally expensive, this level of accuracy is encouraging for lead optimisation applications. Assuming steady improvements in hardware and algorithmic performance, in the long term it is possible to foresee applications in lead discovery too as an accurate rescoring method. Furthermore, alchemical calculations were able to resolve ambiguities regarding unexpectedly large differences in binding modes between extremely similar molecules, which suggests a potential application in the modelling of ligand binding poses.

The precision of the calculations was rigorously assessed in order to take into account both the statistical and sampling uncertainties. Even with the largest and most flexible ligands, standard deviations below 1.0 kcal/mol are achievable within the microsecond time-scale. It is important however to remember that the accuracy of these calculations comes at a high computational cost with respect to scoring functions or endpoint methods. Although we have not optimised our protocol for computational efficiency as we were focussing mainly on accuracy, to give an idea of the computational cost, each free energy calculation took more than a day on 42 Intel® Xeon® CPUs (E5-2697 v2, 2.7GHz, 12 cores per CPU) of a Cray® XC30™ supercomputer. On the other hand, the scores from docking are obtained in seconds on a single core. While the use of graphical processing units can substantially accelerate the simulations, the screening of hundreds to thousands of compounds would still be an onerous exercise. Nonetheless, the accurate experimental determination of binding affinities using biophysical methods such as ITC can be very time consuming too. Before the calorimetric experiment can be prepared and executed, protein expression and purification alone can take about a week, and the ligand of interest might need to be synthesised through multiple reaction steps.

A requirement for successful absolute binding free energies is to have a starting structure that captures the main features of the protein-ligand complex. We showed how, thanks to the ability of MD to explore an ensemble of conformations, having a starting ligand pose that considerably deviates from the crystal pose does not affect the results as long as the main features of the binding are maintained. However, the methodology is still heavily dependent on the quality of poses generated by docking. An incorrect pose prediction might lead to a false negative when calculating the ligand affinity if the pose is assigned a very small binding free energy. Nevertheless, in this study, docking was sufficient and succeeded in finding good poses for all of the inhibitors considered, also thanks to the presence of a clear pharmacophore in this case. On the contrary, docking failed in ranking the ligands or estimating their affinities accurately.

Here, I used replica exchange simulation in order to enhance the mixing of the different thermodynamic states along the alchemical pathway. However, with this scheme it is possible to allow exchanges across additional dimensions. For instance, it is possible to add the ligand binding orientation as a second dimension in order to sample and recover the binding affinity of multiple docking poses in a single set of calculations. In addition, one could also imagine sampling multiple potential binding sites, along with different poses for each site. A very similar approach has in fact already been explored by Wang *et al.* [51] using implicit solvent MD simulations and the T4 lysozyme L99A model system.

It is clear that the accuracy of free energy calculations is highly dependent on the quality of the force field parameters [57, 290, 291]. While here I have not explored the effect of different parameter sets for all ligands, the improved results obtained for ligands 3 and 7 exemplify how terms such as charges and torsions can have an effect on the ability of the simulations to reproduce experimental observables. The effect of van der Waals and coulombic non-bonded parameters on binding free energy results has been previously discussed including using QM calculations to handle polarisation [292]. When dealing with drug-like ligands it is apparent that, as already mentioned in this chapter, torsional parameters are poorly transferable

and can thus affect the performance of the calculations [281, 282]. Furthermore, certain types of inter-molecular interactions are poorly represented by current force field models. Halogen bonding in particular is not allowed by the use of a single partial charge placed on the center of halogen atoms. It has, however, been shown how this issue can be tackled by the use of an additional off-site virtual site carrying a positive charge [285–288]. The impact of force field parameters on the accuracy of the calculations will be discussed more thoroughly in the next chapter. However, it is encouraging that current MM parameters, despite their approximations, manage to provide the level of accuracy here presented, which could also be improved by refinement and extension of existing models.

3.6 Conclusions

In this chapter I have shown that for a small and fairly rigid system such as BRD4(1), free energy calculations based on molecular dynamics are able to achieve RMS errors that do not exceed 1.4 kcal/mol when starting from docked structures, and down to 0.8 kcal/mol when using crystal structures and a more expensive protocol. The present results suggest absolute free energy calculations as a potentially accurate affinity prediction method for drug discovery. This is the first study on absolute binding free energy that takes into account a diverse set of drug-like molecules and a biologically relevant target currently investigated for its therapeutic potential. Notably, a similar level of accuracy was recently reported for a large set of molecules in terms of their relative binding free energies [47]. However, caution is still advised as the test still cover a very small portion of biological and chemical space, and performance of the calculations will likely be highly dependent on the problem at hand. Protein that do not present sampling challenges, and ligands that contain moieties that are well represented in the force field training set, are more likely to return good agreement with experiment. This might not be the case for large and flexible proteins such as kinases or GPCRs, or for ligands with less common chemical groups, or simply groups for which experimental data is sparse. It has

been shown, however, how specific issues with force field parameters, such as poor torsion parameters, may be solved with a careful parametrisation procedure.

Essentially, all models are wrong, but some are useful.

— George E.P. Box, in *Empirical Model Building and Response Surfaces*, 1987

4

Predicting ligand selectivity

Contents

4.1 Introduction 96
4.2 Methods 98
4.2.1 System setup 98
4.2.2 Derivation of torsional parameters 99
4.2.3 Docking 100
4.2.4 Preliminary Molecular Dynamics Simulations 101
4.2.5 Free Energy Calculations 101
4.2.6 Data Analysis 103
4.3 Small molecule parameters: torsion angles 104
4.4 RVX-208 and RVX-OH 106
4.4.1 Binding pose prediction 107
4.4.2 Affinity predictions 109
4.5 Bromosporine 110
4.5.1 Binding pose prediction 111
4.5.2 Affinity predictions 112
4.5.3 Effect of partial charges 115
4.5.4 Effect of torsional parameters 117
4.6 Discussion 121
4.7 Conclusions 124

In this chapter I will report the results of a study where I performed absolute free energy calculations for three ligands binding to multiple bromodomains (BRDs). The aim of this work was to assess the performance of the calculations across multiple BRD families, and evaluate their potential application for the prediction

of ligand selectivity. The content of this chapter has been published after the first submission of this thesis in the following journal article:
M. Aldeghi, A. Heiftez, M.J. Bodkin, S. Knapp, P.C. Biggin. Predictions of Ligand Selectivity from Absolute Binding Free Energy Calculations. *Journal of the American Chemical Society* **2017**, 139(2), 946-957.

4.1 Introduction

Binding selectivity is a requirement for the development of a safe drug, and it is a critical property for chemical probes used in preclinical target validation. Large differences in binding affinity between the intended target and similar proteins are often sought in order to avoid unwanted and often unknown side effects [293]. Conversely, in some cases a compound with a specific promiscuous profile might be equally desirable for reasons of efficacy [138, 294]. Yet, the identification and design of selective (or promiscuous) chemical tools is a difficult task, in particular when working on protein families with conserved folds and similar binding pocket residues. A typical example is provided by the protein kinase family, for which different assay panels have been developed in order to determine ligands' selectivity profiles [295, 296]. The development of bromodomain (BRD) chemical probes poses the same selectivity challenge due to their conserved fold.

Engineering selectivity adds considerable complexity to the rational design of new drugs, as it involves the optimisation of multiple binding affinities. Different strategies have been adopted for selectivity design, such as shape and electrostatics complementarity, conformational selection, water displacement, and allosteric binding. Yet, engineering selectivity still remains a challenge for the drug design process [293]. Computationally, the prediction of binding selectivity is a difficult task and general applicable methodologies are still not available to the computational and medicinal chemistry community. Henrich *et al.* [297] evaluated the performance of target-specific scoring functions, developed from sets of known binders, for three protein targets: thrombin, trypsin, and urokinase-type plasminogen activator. The quality of the predictions appeared to be system dependent and the authors

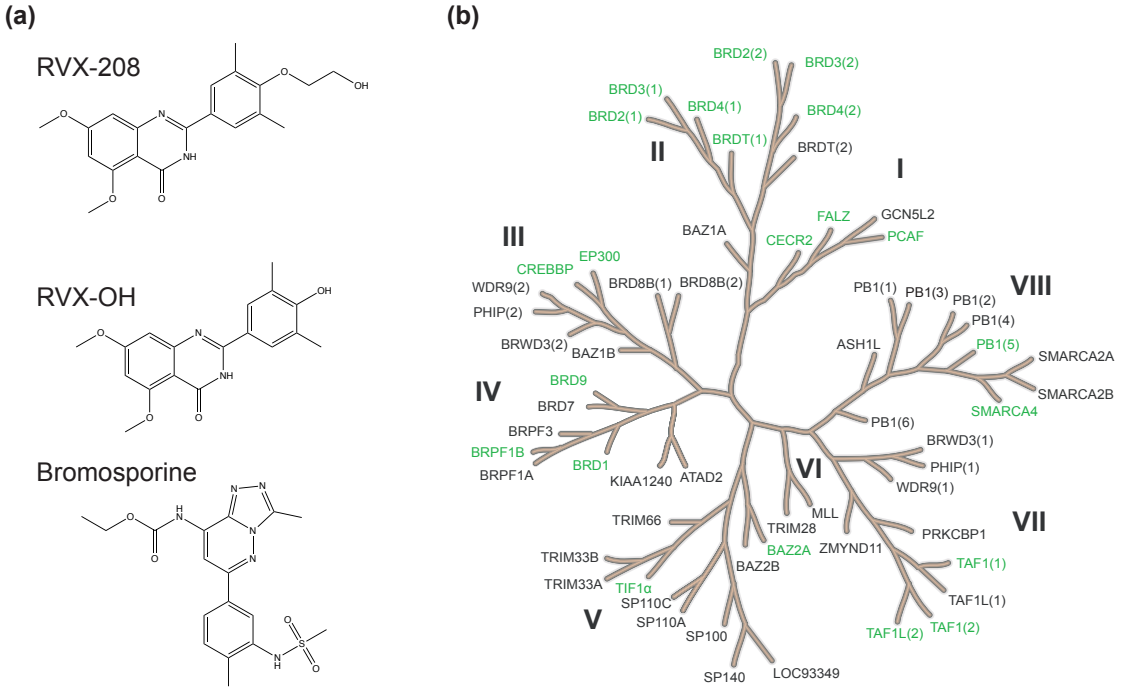


Figure 4.1: Ligands and proteins considered in this study. (a) Chemical structures of the compounds. (b) Phylogenetic tree of the human bromodomain family; in green are all the BRDs included in the study.

discussed how inclusion of protein flexibility and conserved water molecules might result in better docking methods. Absolute binding free energy (ABFE) calculations based on alchemical pathways could offer an alternative and more rigorous approach to the problem, and one that does not require the use of training sets. In fact, these calculations are a more general approach as they do not require knowledge of any affinity value in advance and, in principle, could be used to compare the affinity of any ligand for any protein target. The calculations could thus be employed as a tool to probe ligand selectivity among similar binding pockets, in order to identify the scaffolds providing the desired selectivity profile.

Here, I evaluate the performance of ABFE calculations for the prediction of selectivity by estimating the affinity profile of three bromodomain inhibitors across multiple bromodomain families. The objective was to assess the accuracy of the computational method and its potential utility for the study of ligand selectivity. I calculated the affinities for 35 complexes, involving 3 ligands and 22 bromodomains

(Figure 4.1), without any information on the affinities or structures of the complexes. The results are compared to high-quality isothermal titration calorimetry (ITC) data and the performance of the calculations discussed in relation to their potential application in drug design.

4.2 Methods

4.2.1 System setup

Protein conformations were taken from crystal structures as listed in Table 4.3 (more detailed information on the structures in Table B.1). If multiple chains were present in the asymmetric unit, only monomer A would be used for the calculations. Missing atoms and side chains in the crystal structures were modelled with the WHAT-IF web interface [257, 258] and all organic molecules were removed from the system, whereas all crystallographic water molecules were retained. The binding poses for RVX-208 and RVX-OH were taken from the study described in Chapter 3, where the poses were identified by docking. The preferred binding pose for bromosporine was identified following the same procedure as of Chapter 3, where the ligand was docked into an apo X-ray structure of BRD4(1), PDB-ID 2OSS [70], and then free energy calculations were performed on representative poses after clustering by RMSD. The highest affinity pose for each ligand was then modelled into the pockets of the other bromodomains by structural alignment to BRD4(1) with PyMOL. When placing the ligands into the pockets, crystallographic water molecules within 1.4 Å of the ligand were removed in order to avoid steric clashes. After adding hydrogen atoms with Maestro (v9.5, Schrödinger), the ligands were parameterised with the general AMBER force field [259] (GAFF v1.7) using AmberTools14. Restrained electrostatic potential (RESP) charges [298] (HF/6-31G*//HF/6-31G*) for bromosporine were obtained with the PyRED server [299] using Gaussian 09 (Revision D.01). RESP charges for the ligands RVX-208 and RVX-OH were obtained with a geometry optimisation followed by a molecular electrostatic potential calculation, both at the HF/6-31G* level of theory. The restrained ESP fit was done using antechamber in AmberTools14, while all QM calculations were preformed in Gaussian 09 (Rev.

D.01) [289]. The PyRED server was not employed in the latter cases simply due to its unavailability for a number of months in late 2015. GROMACS topologies and coordinates were generated from the AMBER ones using *acpype* (v.2014-08-27 Rev: 403) [264]. The Amber99SB-ILDN force field was used for the protein and the TIP3P model for water molecules [179, 265]. The complexes were solvated in a dodecahedral box with periodic boundary conditions and a minimum distance between the solute and the box of 12 Å. Sodium and chloride ions were added to neutralise the systems at a concentration of 0.15 M.

4.2.2 Derivation of torsional parameters

The parameters for the two different biaryl torsion angles present in bromosporine and the two RVX ligands and missing from the GAFF force field were derived by fitting to a relaxed torsion scan at the MP2/6-31G* level of theory (Gaussian 09, Rev D.01) [289] performed every 5 degrees (Figure 4.2). Molecular mechanics parameters were searched with a genetic algorithm in which the fitness function was the root mean square deviation between the QM and MM relative energies, using the script provided at `http://www.ub.edu/cbdd/?q=content/small-molecule-dihedrals-parametrization`. For the dihedral scan, only the biaryl fragments were used rather than the whole molecules, in order to avoid potential issues during the QM relaxed potential energy surface scan due to flexible side groups. In addition, this allows the derivation of parameters that could in principle be more transferable to similar torsions in different molecules. The biaryl fragments used for the calculations are highlighted in black on the left-hand side of Figure 4.2. Table 4.1 reports the Amber MM parameters derived for the two biaryl torsions present in bromosporine and RVX-OH and RVX-208.

A number of additional torsions were also optimised specifically for the bromosporine model using the program *paramfit* [300]. The program allows the fitting of multiple dihedrals at the same time. Three moieties were optimised: the benzensulfonamide group, the carbamate group, and again the biaryl torsion. In all three cases, the parameters were derived by fitting the MM energies to single point

Table 4.1: Torsional parameters for the biaryl torsions in bromosporine and RVX-OH/RVX-208. f is the factor by which the torsional barrier is divided; k_ϕ is the force constant in kcal/mol; ϕ_0 is the phase in degrees; n is the periodicity. The atom type "ca" corresponds to an sp^2 aromatic carbon, "cc" and "cd" to a sp^2 carbon in a conjugated ring system, "n" to an amide nitrogen, and "nd" to an sp^2 nitrogen in a conjugated ring system, as per the atom types definitions in GAFF [259].

Dihedral	f	k_ϕ (kcal/mol)	ϕ_0 (°)	n
Bromosporine				
ca-ca-cc-cc	4	3.90	180.0	2
ca-ca-cc-cc	4	0.30	180.0	6
ca-ca-cc-nd	4	3.90	180.0	2
ca-ca-cc-nd	4	0.30	180.0	6
RVXOH and RVX-208				
ca-ca-cc-n	4	3.30	180.0	2
ca-ca-cc-n	4	0.40	0.0	4
ca-ca-cc-nd	4	3.30	180.0	2
ca-ca-cc-nd	4	0.40	0.0	4

energy calculations at the MP2/6-31G* level of theory performed with Gaussian 09 (Rev. A.02) [289]. The parameter search was performed using the hybrid genetic algorithm with a mutation rate of 0.1 and then refined with the simplex algorithm with a convergence limit of 10^{-6}. All other input parameters were used with their default values. For the benzensulfonamide, a total of 514 conformations were used for the fitting of 3 torsions (4 terms per torsion). For the carbamate, a total of 882 conformations were used for the fitting of 7 torsions (2-4 terms per torsion, for a total of 22 terms). For the biaryl, a total of 181 conformations were used for the fitting of 2 torsions (4 terms per torsion). The conformations were derived by systematic variation of the dihedral angles (e.g. for the biaryl torsion, the dihedral was scanned from 0° to 360° every 2°). In all cases, only the force constant and the phase of the torsion were fit, while the periodicity was fixed. The parameters for all these terms are shown in Table B.4.

4.2.3 Docking

The docking protocol employed was the same as the one described in § 3.2.4.

4.2.4 Preliminary Molecular Dynamics Simulations

In order to improve the ligand fit into the binding pockets of the structures used for the free energy calculations, a brief preliminary MD simulation was run for each complex. All simulations were performed in GROMACS 5.0.4 or GROMACS 5.1 [187]. 10,000 energy minimisation steps were carried out using a steepest descent algorithm. The systems were subsequently simulated for 0.5 ns in the canonical ensemble and 1 ns in the isothermal-isobaric ensemble with harmonic position restraints applied to the solute heavy atoms with a force constant of 1,000 kJ mol^{-1} nm^{-2}. Temperature was coupled using Langevin dynamics with 298.15 K as the reference temperature and pressure using the Berendsen weak coupling algorithm [193, 198]. A 15 ns unrestrained run was then performed in the NPT ensemble with the Parrinello-Rahman pressure coupling [202]. The last frame of the run provided the starting system conformation for the free energy calculations. The last 5 ns of the unrestrained simulations were instead used to obtain equilibrium values for the protein-ligand restraints used during the free energy cycle as described in the next section.

4.2.5 Free Energy Calculations

All calculations were performed in GROMACS 5.0.4 [187]. The ligand van der Waals interactions were decoupled and the charges annihilated using a linear alchemical pathway with $\Delta\lambda = 0.05$ for the van der Waals and $\Delta\lambda = 0.1$ for the coulombic transformations. For the addition of the ligand restraints instead, 12 non-uniformly distributed λ values were used (0.0, 0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2, 0.3, 0.5, 0.75, 1.0). A total of 42 windows for the complex simulations and 31 windows for the ligand simulations were therefore employed. For each window, 10,000 energy minimisation steps were carried out using a steepest descent algorithm. The system was subsequently simulated for 0.5 ns in the canonical ensemble with harmonic position restraints applied to the solute heavy atoms with a force constant of 1,000 kJ mol^{-1} nm^{-2}. Temperature was coupled using Langevin dynamics with 298.15 K as the reference temperature [193]. A 1 ns position restrained run in

the isothermal-isobaric ensemble was then performed using the Berendsen weak coupling algorithm [198]. 15 ns unrestrained production runs were performed for data collection using Hamiltonian-exchange Langevin dynamics in the NPT ensemble with the Parrinello-Rahman pressure coupling scheme [202]. 3 million swaps between any state pair were attempted every 1000 time steps, following the Gibbs sampling scheme proposed by Chodera and Shirts [225]. The relative position and orientation of the bound ligand with respect to the protein was restrained by means of one distance, two angles and three dihedral harmonic potentials with force constant of 10 kcal mol^{-1} $Å^{-2}$ [rad^{-2}]. The equilibrium distance and angles for the restraints were taken from the average values of the last 5 ns of the 15 ns preliminary MD simulations. The contribution of this set of restraints to the free energy when the ligand is decoupled can be evaluated analytically as described by Boresch *et al.* [233]. The equation used to evaluate this contribution also includes a correction for the standard-state dependence of the binding free energy [233]. A soft-core potential was employed for the van der Waals interactions transformed [271]. For all simulations the particle mesh Ewald (PME) algorithm [271] was used for electrostatic interactions with a real space cut-off of 12 Å, a spline order of 6, a relative tolerance of 10^{-6} and a Fourier spacing of 1.0 Å. The Verlet cut-off scheme was used with a van der Waals interaction cut-off of 12 Å and a buffer tolerance of 0.005 kJ/mol/ps [301]. The P-LINCS constraint algorithm was used only on H-bonds [196]. The GROMACS long-range dispersion correction for energy and pressure was used, and an additional numerical long-range dispersion correction (EXP-LR) was applied as described in Shirts *et al.* [205] to address the failure of the isotropic assumption of the analytical correction during simulations of protein-ligand complexes. No correction for box-size effects was applied as it has been shown that, contrary to charged ligands, for neutral solutes these effects are negligible [302].

Relative Calculations

Relative free energy calculations were performed in order to estimate the effect of partial charges (AM1-BCC versus RESP model) and torsional parameters on the

calculated binding affinity for bromosporine. In the calculations transforming the partial charges, 11 uniformly distributed λ values were used. For the calculations transforming the torsional parameters, the force constant and phase of the dihedrals were interpolated between the two models using 12 uniformly distributed λ values. The equilibration and production phases were carried out as per the protocol described above for absolute free energy calculations. No restraints have been used in these calculations, and all of the ligand atoms were interacting with the environment during the transformations. For the ligand in solution, 5 repeats were performed in order to determine the standard deviation, while for the protein-ligand complexes simulations a single repeat was performed.

4.2.6 Data Analysis

Free energy estimates were obtained with the implementation of the multiple Bennet acceptance ratio (MBAR) provided with the python package *pymbar* (`https://github.com/choderalab/pymbar`) and using the *alchemical analysis* tool (`https://github.com/MobleyLab/alchemical-analysis`).[240, 273] The data from each lambda state were subsampled in order to include only uncorrelated data-points by calculating the autocorrelation statistical inefficiency of the derivative of the potential energy with respect to lambda ($dU/d\lambda$). Only the $dU/d\lambda$ components that were changing at a particular lambda state were used as the observable for the calculation of the autocorrelation times. The equilibrated region was determined by visualising the forward and reverse convergence plots as suggested by Klimovich *et al.* [273]. Multiple plots were generated by discarding data at increments of 1 ns, from a minimum of 1 ns to a maximum of 12 ns, and identifying the equilibration time that produced a convergence plot with little drift in free energy in any one direction. The uncertainty of each free energy estimate was taken from the MBAR asymptotic variance-derived uncertainty. For the free energy of decoupling the ligand from solution, the mean and sample standard deviation of multiple (five) repeats were used. For bromosporine and RVX-OH, the five initial conformation were determined by the five docked poses; for RVX-208, which only had two poses

returned by the docking and clustering procedure, other three snapshots were extracted from the simulations of the two docking poses in order to provide the additional random starting conformations. The final absolute binding free energy is the difference between the decoupling of the ligand from the water solution and from the solvated complex; the final error in the binding free energies is thus the root sum of squares of the uncertainties of ligand and complex calculations. The binding free energies for the bromosporine models with modified parameters (charges or torsions) were obtained by adding the relative free energy obtained due to the parameters modification to the absolute free energy obtained with the original model. The final uncertainty was derived as the root sum of squares of the standard deviations of the absolute and relative binding free energies. 95% confidence intervals (CIs) for the performance statistics (MUE, RMSE, r_p, r_s) were calculated with percentile bootstrap, after building 10^5 bootstrap samples by resampling the experimental and calculated free energies according to their mean and standard deviations, assuming normality. Throughout the text I will report the 95% CI in square brackets immediately after the statistics.

4.3 Small molecule parameters: torsion angles

Based upon the experience of the calculations described in Chapter 3, and indications from the literature [281, 282], it was noted that torsional parameters in small molecule force fields tend to be problematic due to their limited transferability. In particular, biaryl torsions can have different minima and maxima depending on the nature of the ring substituents, and they often show low energy barriers. This means that in solution many rotameric states are accessible and the proportion of each state depends on the energy around the dihedral. It is thus important to capture the relative energy of the different dihedral angles in order to obtain the correct ensemble of conformations during a simulation. In addition, it was observed that parameters for the biaryl torsions present in the ligands of interest in this study were missing from the GAFF force field. The program *parmchk2*, part of the AmberTools suite, is able to assign parameters based on similarity to other chemical

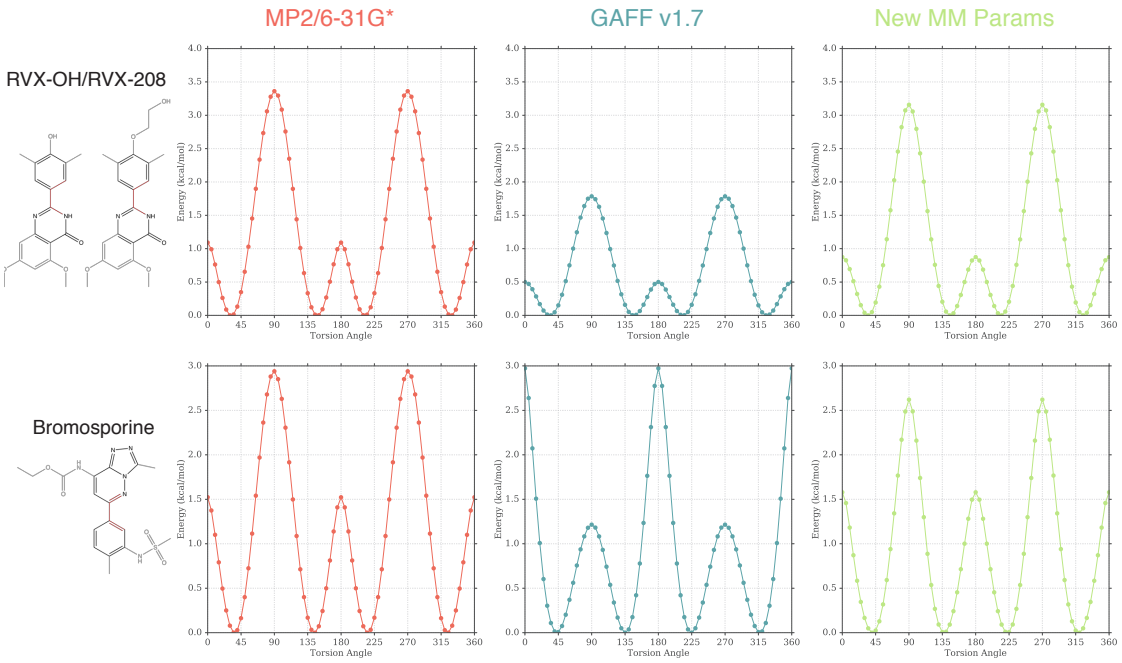


Figure 4.2: QM and MM torsion scans for the biaryl dihedral angles in RVX-OH/RVX-208 and bromosporine.

groups for which parameters are present in the force field. However, the chosen set of parameters are not guaranteed to be accurate for the chemical group present in the molecule, and this prompted us to derive MM parameters for these torsions. In fact, GAFF v1.7 assigns the biaryl torsions in RVX-OH and RVX-208, and bromosporine, the same parameters, which are taken to be the ones for a general torsion around a bond between an sp2 aromatic and an sp2 aliphatic carbon ("X -c2-ca-X" in the Amber notation, where "X" are wildcards representing any atom).

Figure 4.2 shows the resulting energetic barriers around these dihedrals for fragments representing the core of the molecules here considered. It is possible to see that for RVX-OH and RVX-208 the minima and maxima are correctly placed, but the heights of the energy barriers at 0°, 90°, 180° and 270° are about half of that in the QM scan. This results in a high frequency of ring flips during a MD simulation. The newly derived parameters solve much of this issue by increasing the height of the energy barrier to be closer to what was observed in the QM calculations. In the case of the biaryl torsion in bromosporine too, the placement of the minima is correct (approximately at 45°, 135°, 225°, and 315°), but the

height of the energy barriers is not. GAFF places the larger (3 kcal/mol) barriers at 0° and 180°, rather than at 90° and 270°, where there should be instead lower barriers of about 1.5 kcal/mol. The new parameters, on the right of Figure 4.2, better recapitulate the QM relative energies, with ∼1.5 kcal/mol barriers at 0° and 180°, and higher barriers at 90° and 270° of ∼2.6 kcal/mol (although the latter are still slightly underestimated by ∼0.4 kcal/mol).

4.4 RVX-208 and RVX-OH

RVX-208 is an orally available small molecule that is currency being investigated in clinical trials for the treatment of atherosclerosis, diabetes, and cardiovascular diseases, and it has been found to inhibit the bromo and extraterminal (BET) family of bromodomains [132, 134, 135]. RVX-208 targets only this bromodomain family, and it shows a slight selectivity (from about 8 to 21 fold in affinity) for the second (BD2) versus the first bromodomains (BD1) within this group [134]. RVX-OH is the synthetic precursor of RVX-208 and also targets the BET bromodomain family, however, it does not show the same preference for the second bromodomains. The compound in fact rearranges in the binding pockets of the first BET bromodomains and binds with a different pose as compared to RVX-208, where it occupies the WP/ZA shelf (see § 1.2.1, and in particular Figure 1.4). In this way RVX-OH achieves similar affinities for both the first and second BET bromodomains, abrogating the subtle selectivity observed for RVX-208. The crystallographically observed binding poses for RVX-208 and RVX-OH in the binding pockets of a first and second BET bromodomains are shown in Figure 4.3. For ease of description, and consistently with the orientation of the figures in this text, we denote the binding orientation of RVX-OH in BD2 pockets as the “vertical” V-pose, and the binding orientation of RVX-OH in BD1 pockets as the “horizontal” H-pose, despite this naming not having any physical or biological meaning. Note that it is not possible for RVX-208 to bind in the H-pose, due to the presence of an additional hydroxyethyl group that would clash with the protein. These ligands thus represent an interesting case scenario to test the ability of the calculations to identify both

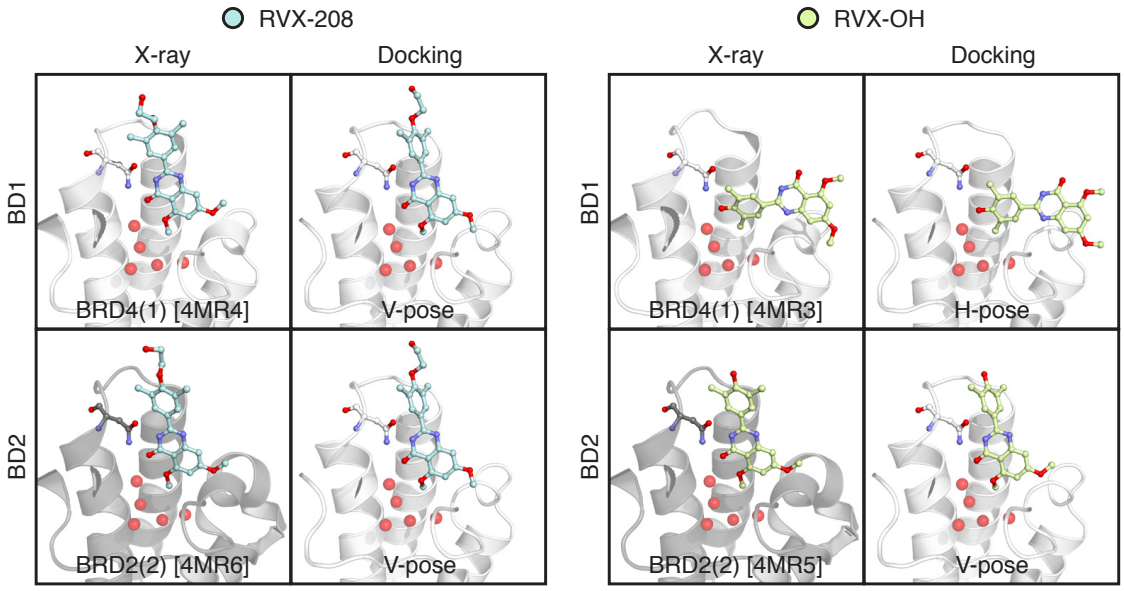


Figure 4.3: Comparison of the X-ray structures of RVX-208 and RVX-OH in first and second BET bromodomains binding pockets, and the corresponding docked structures. It is possible to see how RVX-208 binds with the same pose to both BD1 and BD2, while RVX-OH adopts two distinct binding orientations. The two ligands have been docked into an apo structure of BRD4(1) (PDB-ID 2OSS), and the docking poses best representing the crystal structures are shown. These poses are taken from the work in Chapter 3. BD1s are represented by white cartoons, and BD2s by grey cartoons; RVX-208 is represented in light blue, and RVX-OH in light green; the Asn residue conserved across the whole subfamily II and the conserved network of water molecules are shown as reference points.

the binding modes and slight selectivity of a ligand for multiple proteins. In this study, the 7 BET BRDs (out of 8) with at least one crystal structure available in the PDB and affinity data determined by ITC were considered.

4.4.1 Binding pose prediction

The docking pose for RVX-208 was taken from the work in Chapter 3, in which absolute free energy calculations were shown to be able to identify the crystallographic pose for this ligand in BRD4(1). Similarly, the two docking poses for RVX-OH, corresponding to the V- and H-poses, were taken from the same study. These docking poses were modelled into the structure of other bromodomains by superimposition using PyMOL [303]. Again, in this case too, it was showed how the calculations identified the correct pose for BRD4(1). Here, however, since we are considering both BD1s and BD2s, we will re-score both the V-pose and H-pose

Table 4.2: Summary of the free energy calculation results for RVX-OH and RVX-208, along with information about the experimental X-ray and affinity data. ITC data taken from Picaud *et al.* [134]. All uncertainties are one standard deviation. ITC uncertainties are an estimate based on the standard deviation of the ΔG values observed in the ABRF-MIRG'02 inter-laboratory assessment [174]. Values for the difference $\Delta G_{calc} - \Delta G_{exp}$ might appear inconsistent due to rounding. "Pose" indicates in which one of the two orientations the ligands are expected to bind given the experimental evidence.

Protein	Ligand	Pose	ΔG_{exp} (kcal/mol)	H-pose ΔG_{calc} (kcal/mol)	V-pose ΔG_{calc} (kcal/mol)	$\Delta G_{calc} - \Delta G_{exp}$ (kcal/mol)
BRD2(1)	RVX-OH	H	-8.5 ± 0.1	-9.8 ± 0.1	-6.5 ± 0.1	-1.3 ± 0.2
BRD2(2)	RVX-OH	V	-8.8 ± 0.1	-7.6 ± 0.1	-7.9 ± 0.1	0.9 ± 0.1
BRD3(1)	RVX-OH	H	-8.0 ± 0.1	-6.7 ± 0.1	-5.7 ± 0.1	1.3 ± 0.2
BRD3(2)	RVX-OH	V	-8.8 ± 0.1	-6.6 ± 0.2	-7.4 ± 0.1	1.5 ± 0.2
BRD4(1)	RVX-OH	H	-9.0 ± 0.1	-9.9 ± 0.1	-6.6 ± 0.1	-0.9 ± 0.2
BRD4(2)	RVX-OH	V	-8.8 ± 0.1	-5.0 ± 0.1	-8.5 ± 0.1	0.3 ± 0.2
BRDT(1)	RVX-OH	H	-7.7 ± 0.1	-7.7 ± 0.1	-6.3 ± 0.1	0.0 ± 0.2
BRD2(1)	RVX-208	V	-6.9 ± 0.1		-7.2 ± 0.1	-0.3 ± 0.2
BRD2(2)	RVX-208	V	-8.5 ± 0.1		-7.5 ± 0.1	0.9 ± 0.2
BRD3(1)	RVX-208	V	-7.1 ± 0.1		-5.5 ± 0.1	1.6 ± 0.2
BRD3(2)	RVX-208	V	-8.8 ± 0.1		-8.7 ± 0.1	0.1 ± 0.2
BRD4(1)	RVX-208	V	-7.8 ± 0.1		-6.8 ± 0.1	1.0 ± 0.2
BRD4(2)	RVX-208	V	-9.1 ± 0.1		-9.7 ± 0.1	-0.6 ± 0.2
BRDT(1)	RVX-208	V	-7.0 ± 0.1		-6.5 ± 0.1	0.5 ± 0.2
Statistics						
Mean Unsigned Error						0.81 [0.74, 0.90]
Root Mean Square Error						0.95 [0.87, 1.05]
Pearson's r						0.75 [0.67, 0.80]
Spearman's ρ						0.78 [0.64, 0.85]

for all 7 BRDs considered. The relevant docking poses are shown in Figure 4.3. The two poses for RVX-OH, and the one for RVX-208, were modelled into the other bromodomain structures. In order to allow the binding pockets to relax around the ligands, a short (15 ns) MD simulation of the complexes was run before performing the free energy calculations.

Given the two potential binding poses of RVX-OH for the BET bromodomains, I ran the calculations starting from both orientations for all the seven BET bromodomains here considered (no structure is yet available for BRDT(2)) in order to test whether the computational protocol could reliably identify the correct pose for the first (H-pose) versus second (V-pose) BET bromodomains. Table 4.2 summarises the results for RVX-OH, and shows how the orientation that would be expected to be the more stable based on the crystallographic evidence

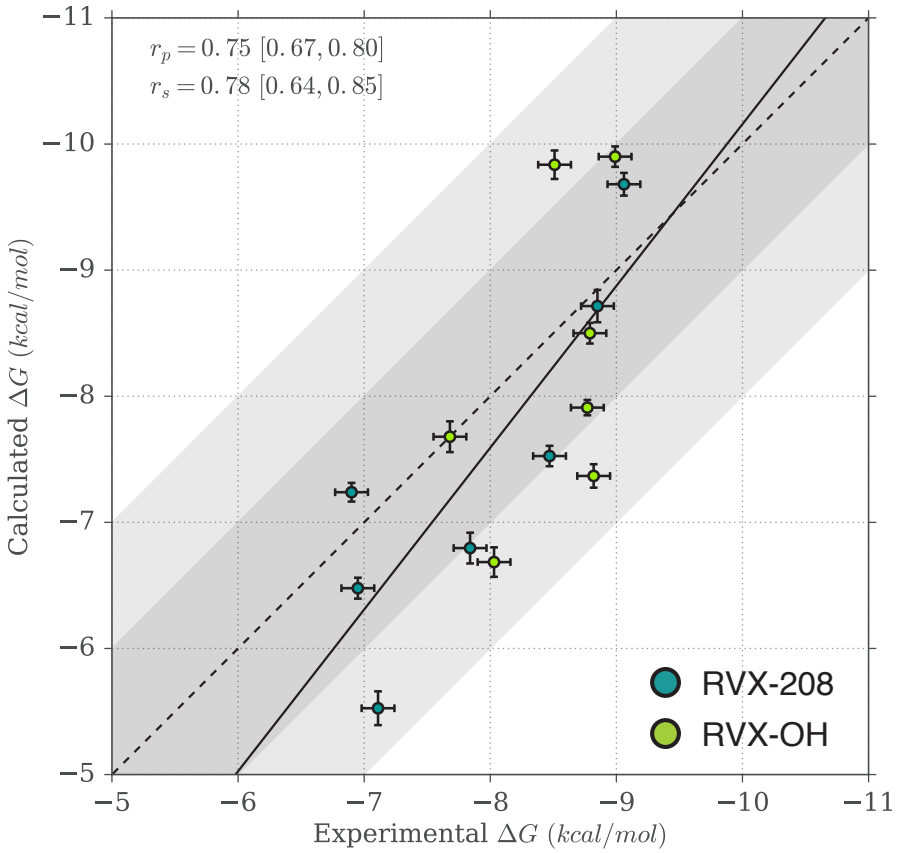


Figure 4.4: Scatter plots of calculated versus experimental affinities for the two RVX compounds. A linear fit of the data is shown as a black line, and the 1 and 2 kcal/mol error boundaries are shown as shaded grey areas.

was always predicted to have a lower binding free energy value, i.e. a higher affinity, than the other competing orientation. For each bromodomain/RVX-OH complex we then took the higher affinity pose as being the most stable orientation and compared the corresponding binding free energy value to the ITC data. As previously mentioned, RVX-208 can bind only in one orientation (V-pose), thus this exercise was not carried out for this ligand.

4.4.2 Affinity predictions

Table 4.2 and Figure 4.4 summarise the correspondence of the computational predictions to the ITC data for the two ligands together (full breakdown of free energy terms in Table B.2). It is possible to see how the predictions reproduced experimental values in absolute terms extremely well in this case. All of the

predicted binding free energies are within 2 kcal/mol of the ITC values, and about two thirds are within 1 kcal/mol. This results in low mean unsigned error (MUE) and root mean squared error (RMSE): 0.81 $[0.74, 0.90]$ and 0.95 $[0.87, 1.05]$ kcal/mol, respectively. The correlation with ITC values is also particularly good for absolute calculations, with Pearson correlation (r_p) of 0.75 $[0.67, 0.80]$ and Spearman rank (r_s) of 0.78 $[0.64, 0.85]$. Considering RVX-208 alone, the calculations achieve $r_p = 0.84$ $[0.76, 0.90]$ and $r_s = 0.75$ $[0.71, 0.93]$. On the other hand, when considering RVX-OH only, lower correlation values with larger uncertainties are returned ($r_p = 0.49$ $[0.27, 0.66]$ and $r_s = 0.46$ $[0.14, 0.71]$) despite the calculations performing well in terms of absolute errors for both ligands; this is expected given the small dynamic range of binding free energies considered for RVX-OH (from -9.0 kcal/mol to -7.7 kcal/mol).

If the predictions were grouped into first [BRD2(1), BRD3(1), BRD4(1), BRDT(1)] and second [BRD2(2), BRD3(2), BRD4(2)] BET bromodomains, it would be possible to anticipate the slight selectivity of RVX-208 for BD2s. In fact, the affinity sample mean and standard deviation of RVX-208 for BD1s was of -6.5 ± 0.7 kcal/mol, while for BD2s was of -8.6 ± 1.1 kcal/mol. A two-tailed t-test returns a p-value of 0.03, which would suggest that the predicted difference in affinity between the two groups is unlikely to be caused by chance. Contrarily, RVX-OH was predicted to have an average affinity of -8.5 ± 1.6 kcal/mol for BD1s and of -7.9 ± 0.6 kcal/mol for BD2s, showing a larger overlap of their distributions and consequently returning a p-value of 0.57 when applying a two-tailed t-test, correctly suggesting the probable lack of any preference for either the first or second BET bromodomains.

4.5 Bromosporine

Bromosporine is a broad-spectrum bromodomain inhibitor that binds across all bromodomain families with a relatively wide range of affinities (from $\sim$25 μM to $\sim$10 nM; from -6.2 kcal/mol to -10.7 kcal/mol in binding free energy terms). As affinities for a number of BRDs have been experimentally determined by isothermal titration calorimetry (Table 4.3) [159], it is an ideal system in order to test the

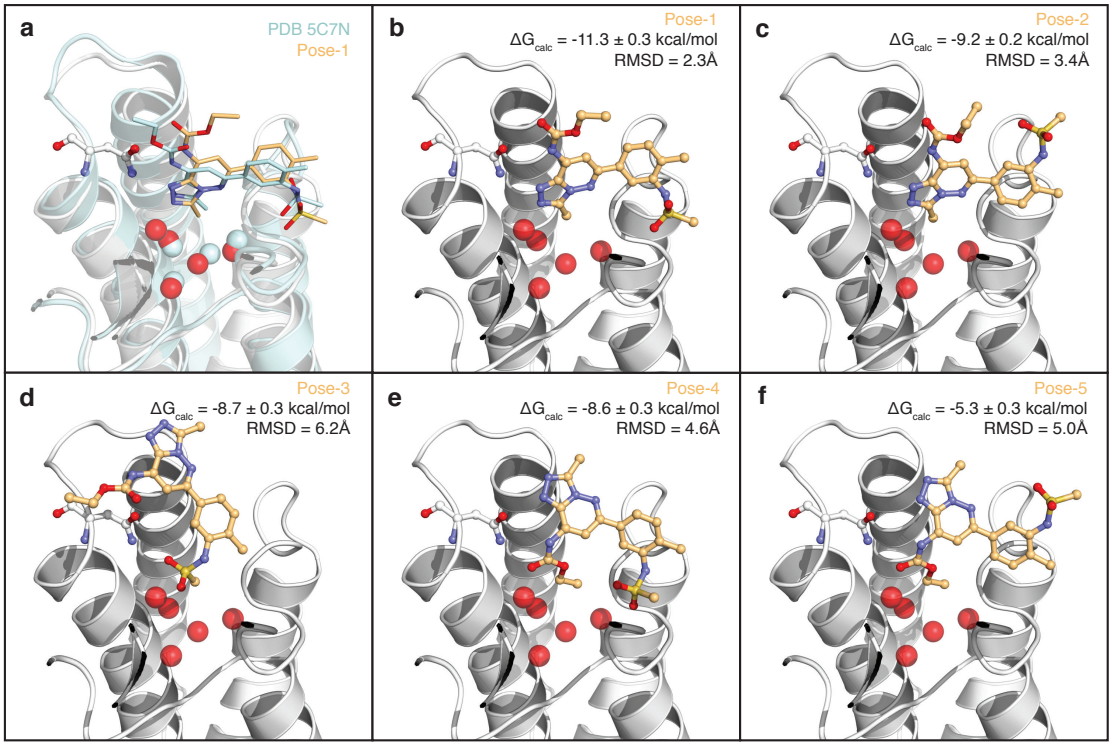


Figure 4.5: Docking poses of bromosporine in the binding pocket of BRD4(1). (a) Comparison of the pose with highest predicted affinity to the X-ray structure of bromosporine in complex with the bromodomains of BRPF1 (PDB-ID 5C7N). (b-e) The five bromosporine poses suggested by docking, with their predicted binding affinity and RMSD with respect to the binding mode in the crystal structure.

ability of free energy calculations to reproduce the affinity profile. As for RVX-OH and RVX-208, I considered all of the bromodomains with at least one crystal structure available in the PDB and affinity data determined by ITC, for a total of 22 bromosporine/BRD pairs.

4.5.1 Binding pose prediction

At the onset of this study, in March 2015, no structure of bromosporine in complex with a bromodomain had been deposited in the PDB. I therefore docked the ligand in the binding pocket of BRD4(1) as done in Chapter 3. In this way, 5 different clusters of poses were identified and free energy calculations were performed on all of them in order to determine the most stable orientation. These were distinct poses that did not interconvert within the simulations timescale. Figure 4.5b-f

provides an overview of these poses, ranked by their predicted affinity. In July 2015 the first bromosporine structure in complex with a human bromodomain (BRPF1B, family IV) was released on the PDB (PDB-ID 5C7N), and we were then able to compare the docked poses to this experimentally resolved structure. Figure 4.5a shows an alignment of the docking pose with highest predicted affinity in BRD4(1) and the X-ray structure of bromosporine in complex with BRPF1B. It is clear that the predicted pose resembles the crystallographic orientation well, with an RMSD of 2.3 Å, mainly due to a rotation of the ethyl carbamate sidechain. The fact that bromosporine was shown to bind to BRPF1B in the same fashion as predicted for BRD4(1) also gave us confidence that the binding mode for this ligand is conserved across bromodomain subfamilies. The second pose (Figure 4.5c) is close to the experimental structure (RMSD of 3.4 Å) as well, but with the sulfonamide group directed toward the solvent rather than making contacts with the protein. The remaining three poses (Figure 4.5d-f) bind the pocket in substantially different orientations as compared to the first two docking poses and the crystallographic pose, as indicated by the larger RMSD of 4.6 Å, 5.0 Å, and 6.2 Å. The binding affinity of bromosporine for BRD4(1) was overestimated by over 1 kcal/mol: the calculated free energy of the most stable pose was predicted to be -11.3 ± 0.3 kcal/mol, while the ITC measure was of -9.7 ± 0.1 kcal/mol. Correspondingly, the docking poses c, d and e, are unexpectedly assigned relatively large affinities while binding very differently when compared to the experimental pose determined by X-ray crystallography.

4.5.2 Affinity predictions

After establishing the most likely binding orientation of bromosporine in BRD pockets, the ligand was modelled into the X-ray structures of 21 other bromodomains, from 7 of the 8 bromodomain families. As for the RVX ligands, a short MD simulation of the complexes was run before performing free energy calculations for all the bromosporine/BRD pairs. The results of the calculations are summarised in Table 4.3 (full breakdown of free energy terms in Table B.3) and Figure 4.6 (left,

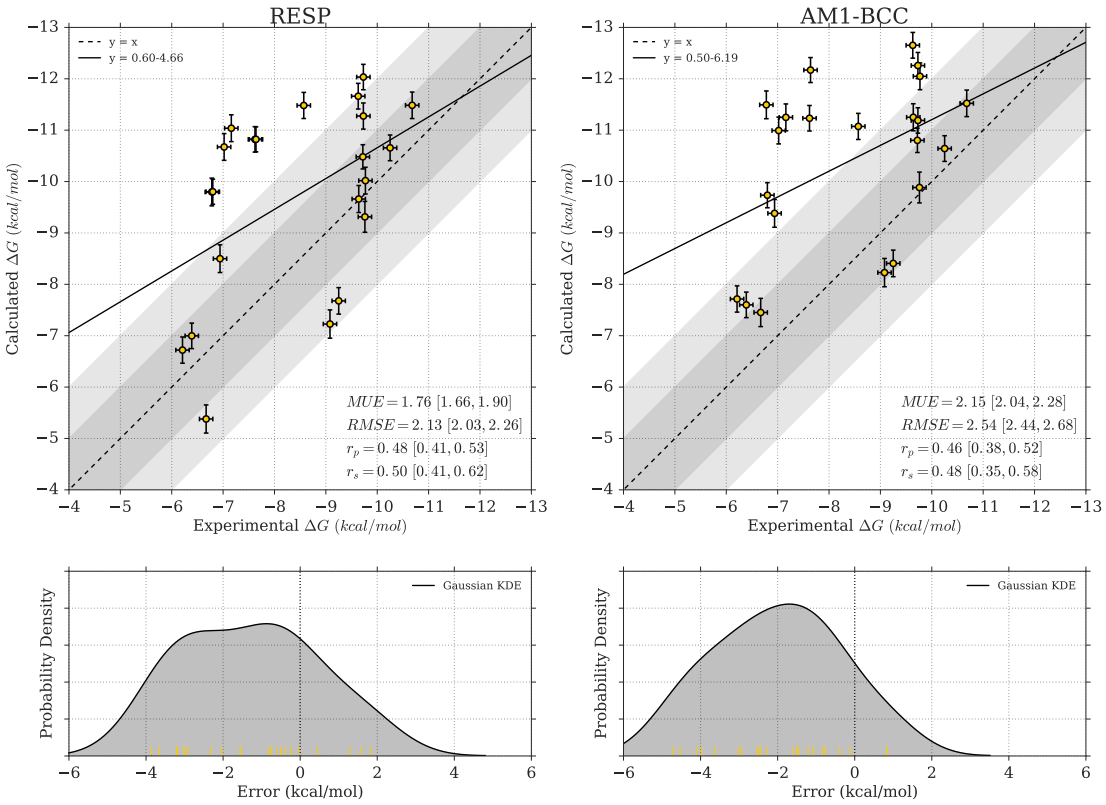


Figure 4.6: Scatter plots of calculated versus experimental affinities for bromosporine, for both the model using RESP and AM1-BCC charges. A linear fit of the data is shown as a black line, and the 1 and 2 kcal/mol error boundaries are shown as shaded grey areas. Below the scatter plots, the error distributions are shown as kernel density estimates.

"RESP"). A reasonable overall agreement with experiment in terms of absolute values was obtained, with a MUE of 1.76 [1.66, 1.90] kcal/mol and a RMSE of 2.13 [2.03, 2.26] kcal/mol. Roughly a third of the results are within 1 kcal/mol of the ITC values, and two thirds within 2 kcal/mol. However, this leaves another third of the results being off by at least 2 kcal/mol, which corresponds to about a 30-fold error in the dissociation constant. The correlation with experiment is reasonable but not as strong as for RVX-OH/RVX-208 or the tests presented in Chapter 3, with $r_p = 0.48$ [0.41, 0.53] and $r_s = 0.50$ [0.41, 0.62]. The distribution of the errors shown at the bottom of Figure 4.6 illustrate how, overall, there is an overestimation issue in this set of predictions, with mean signed error of -1.3 kcal/mol, and most errors being between 0.0 and -4.0 kcal/mol.

Table 4.3: Bromosporine X-ray and affinity data, and absolute free energy calculation results for the bromosporine model with RESP charges. ITC data taken from Picaud *et al.* [159]. All ΔG uncertainties are one standard deviation. ITC uncertainties are an estimate based on the standard deviation of the ΔG values observed in the ABRF-MIRG'02 inter-laboratory assessment [174]. Values for the difference $\Delta G_{calc} - \Delta G_{exp}$ might appear inconsistent due to rounding. In square brackets the 95% confidence intervals of the statistics based on percentile bootstrap are reported.

Protein	Family	PDB-ID	ΔG_{exp} (kcal/mol)	ΔG_{calc} (kcal/mol)	$\Delta G_{calc} - \Delta G_{exp}$ (kcal/mol)
CECR2	I	3NXB	-10.7 ± 0.1	-11.5 ± 0.3	-0.8 ± 0.3
FALZ	I	3UV2	-6.8 ± 0.1	-9.8 ± 0.2	-3.0 ± 0.3
PCAF	I	3GG3	-7.0 ± 0.1	-10.7 ± 0.3	-3.7 ± 0.3
BRD2(1)	II	4ALG	-9.2 ± 0.1	-7.7 ± 0.3	1.6 ± 0.3
BRD2(2)	II	4MR5	-9.6 ± 0.1	-9.7 ± 0.3	-0.0 ± 0.3
BRD3(1)	II	3S91	-9.1 ± 0.1	-7.2 ± 0.3	1.9 ± 0.3
BRD3(2)	II	3S92	-9.6 ± 0.1	-11.7 ± 0.3	-2.0 ± 0.3
BRD4(1)	II	2OSS	-9.7 ± 0.1	-11.3 ± 0.3	-1.5 ± 0.3
BRD4(2)	II	2OUO	-9.8 ± 0.1	-10.0 ± 0.3	-0.2 ± 0.3
BRDT(1)	II	4KCX	-9.8 ± 0.1	-9.3 ± 0.3	0.4 ± 0.3
CREBBP	III	4NYX	-7.6 ± 0.1	-10.8 ± 0.2	-3.1 ± 0.3
EP300	III	3I3J	-6.8 ± 0.1	-9.8 ± 0.3	-3.0 ± 0.3
BRD1	IV	5AME	-7.6 ± 0.1	-10.8 ± 0.2	-3.2 ± 0.3
BRD9	IV	4XY8	-9.7 ± 0.1	-12.0 ± 0.2	-2.3 ± 0.3
BRPF1B	IV	4LC2	-8.6 ± 0.1	-11.5 ± 0.3	-2.9 ± 0.3
BAZ2A	V	4QBM	-7.2 ± 0.1	-11.0 ± 0.3	-3.9 ± 0.3
TIF1	V	4YBM	-6.7 ± 0.1	-5.4 ± 0.3	1.3 ± 0.3
TAF1(1)	VII	3UV5	-6.9 ± 0.1	-8.5 ± 0.3	-1.6 ± 0.3
TAF1(2)	VII	3UV4	-10.2 ± 0.1	-10.7 ± 0.3	-0.4 ± 0.3
TAF1L(2)	VII	3HMH	-9.7 ± 0.1	-10.5 ± 0.2	-0.8 ± 0.3
PB1(5)	VIII	3MB4	-6.4 ± 0.1	-7.0 ± 0.2	-0.6 ± 0.3
SMARCA4	VIII	2GRC	-6.2 ± 0.1	-6.7 ± 0.3	-0.5 ± 0.3
Statistics					
Mean Unsigned Error					1.76 [1.66, 1.90]
Root Mean Square Error					2.13 [2.03, 2.26]
Pearson's r					0.48 [0.41, 0.53]
Spearman's ρ					0.50 [0.41, 0.62]

The low number of samples from each bromodomain family hampers a meaningful statistical analysis, and in fact a one-way analysis of variance reveals that the error distribution of each subfamily does not differ in a statistically significant manner from any other subfamily. However, there seem to be a pattern with calculations involving bromodomains from subfamilies II, VII, and VIII providing overall better results than other subfamilies.

Note that the errors shown for the binding free energy estimates reflect the statistical uncertainty of the free energy estimator only. The error due to limited

sampling is not taken into account, and this likely results in an underestimation of the true overall uncertainty of the binding free energies. The uncertainty for decoupling the ligand from the solvent was assessed through five calculations started from different ligand conformations, thereby providing an estimate of uncertainty that includes information about finite sampling. The sample standard deviation of the decoupling free energy from these five calculations was 0.23 kcal/mol, about four-fold the size of the error provided by the statistical estimator (0.05 kcal/mol) for each individual calculation. In order to gauge the magnitude of the uncertainty due to sampling for the complex simulations, the calculation for BRD4(1) was repeated three times, obtaining a standard deviation for decoupling the ligand from the protein of 0.54 kcal/mol, about four to five times the statistical uncertainty given by the estimator for each individual simulation (0.11 kcal/mol, 0.15 kcal/mol, and 0.16 kcal/mol, respectively, for the three repeats). Based on the overall estimates of the binding affinities (-11.28 ± 0.25 kcal/mol, -10.36 ± 0.27 kcal/mol, -11.46 ± 0.28 kcal/mol), the sample standard deviation obtained was 0.59 kcal/mol, about two-fold the uncertainties of each single repeat, which were derived taking into account the five calculations for the free ligand in solution but a single one for the ligand in the complex. For all other proteins, only a single repeat of the complex simulation was used due to the computational cost of these calculations, however, based on the discussion above, I would suggest to bear in mind that real uncertainty in the calculations is likely about the double of the ones reported.

4.5.3 Effect of partial charges

As previously mentioned, the calculations returned systematically overestimated predictions of binding affinity for bromosporine. Interested in identifying the possible source of this issue, I decided to evaluate the effect of the charge model on the results. In fact, while in Chapter 3 the AM1-BCC semi-empirical charge derivation scheme was used, here RESP charges derived at the HF/6-31G* level of theory were employed in the first instance. Furthermore, as RESP charges depend on the input ligand geometry, the charges for bromosporine were derived by

Table 4.4: Bromosporine affinity data and calculated free energies with the AM1-BCC charge model. ITC data taken from Picaud *et al.* [159]. All ΔG uncertainties are one standard deviation. ITC uncertainties are an estimate based on the standard deviation of the ΔG values observed in the ABRF-MIRG'02 inter-laboratory assessment [174]. Values for the difference $\Delta G_{calc} - \Delta G_{exp}$ might appear inconsistent due to rounding. In square brackets the 95% confidence intervals of the statistics based on percentile bootstrap are reported.

Protein	Family	ΔG_{exp} (kcal/mol)	$\Delta\Delta G_{charges}$ (kcal/mol)	ΔG_{calc}^{AM1BCC} (kcal/mol)	$\Delta G_{calc}^{AM1BCC} - \Delta G_{exp}$ (kcal/mol)
CECR2	I	-10.7 ± 0.1	-0.0 ± 0.0	-11.5 ± 0.3	-0.8 ± 0.3
FALZ	I	-6.8 ± 0.1	0.1 ± 0.0	-9.7 ± 0.2	-2.9 ± 0.3
PCAF	I	-7.0 ± 0.1	-0.3 ± 0.0	-11.0 ± 0.3	-4.0 ± 0.3
BRD2(1)	II	-9.2 ± 0.1	-0.7 ± 0.0	-8.4 ± 0.3	0.8 ± 0.3
BRD2(2)	II	-9.6 ± 0.1	-1.6 ± 0.0	-11.3 ± 0.3	-1.6 ± 0.3
BRD3(1)	II	-9.1 ± 0.1	-1.0 ± 0.0	-8.2 ± 0.3	0.9 ± 0.3
BRD3(2)	II	-9.6 ± 0.1	-1.0 ± 0.0	-12.7 ± 0.3	-3.0 ± 0.3
BRD4(1)	II	-9.7 ± 0.1	-1.0 ± 0.0	-12.3 ± 0.3	-2.5 ± 0.3
BRD4(2)	II	-9.8 ± 0.1	-2.0 ± 0.0	-12.0 ± 0.3	-2.3 ± 0.3
BRDT(1)	II	-9.8 ± 0.1	-0.6 ± 0.0	-9.9 ± 0.3	-0.1 ± 0.3
CREBBP	III	-7.6 ± 0.1	-1.3 ± 0.0	-12.2 ± 0.2	-4.5 ± 0.3
EP300	III	-6.8 ± 0.1	-1.7 ± 0.0	-11.5 ± 0.3	-4.7 ± 0.3
BRD1	IV	-7.6 ± 0.1	-0.4 ± 0.0	-11.2 ± 0.2	-3.6 ± 0.3
BRD9	IV	-9.7 ± 0.1	0.8 ± 0.0	-11.2 ± 0.2	-1.5 ± 0.3
BRPF1B	IV	-8.6 ± 0.1	0.4 ± 0.0	-11.1 ± 0.3	-2.5 ± 0.3
BAZ2A	V	-7.2 ± 0.1	-0.2 ± 0.0	-11.2 ± 0.3	-4.1 ± 0.3
TIF1	V	-6.7 ± 0.1	-2.1 ± 0.0	-7.5 ± 0.3	-0.8 ± 0.3
TAF1(1)	VII	-6.9 ± 0.1	-0.9 ± 0.0	-9.4 ± 0.3	-2.4 ± 0.3
TAF1(2)	VII	-10.2 ± 0.1	0.0 ± 0.0	-10.6 ± 0.3	-0.4 ± 0.3
TAF1L(2)	VII	-9.7 ± 0.1	-0.3 ± 0.0	-10.8 ± 0.2	-1.1 ± 0.3
PB1(5)	VIII	-6.4 ± 0.1	-0.6 ± 0.0	-7.6 ± 0.2	-1.2 ± 0.3
SMARCA4	VIII	-6.2 ± 0.1	-1.0 ± 0.0	-7.7 ± 0.3	-1.5 ± 0.3
Statistics					
Mean Unsigned Error					2.15 [2.04, 2.28]
Root Mean Square Error					2.54 [2.44, 2.68]
Pearson's r					0.46 [0.38, 0.52]
Spearman's ρ					0.48 [0.35, 0.58]

fitting to the molecular electrostatic potentials of five different ligand conformations, corresponding to the structure of the five docking poses. In order to calculate the difference in binding free energies between the two charge models ($\Delta\Delta G_{charges}$), a relative alchemical cycle was used, where the atoms partial charges were transformed from RESP to AM1-BCC both for the ligand in solution and the protein-ligand complexes. In such a way, $\Delta\Delta G_{charges}$ can provide accurate information regarding the effect of the charge model on the results, and it can be added to the absolute binding free energies obtained with the RESP model (ΔG_{calc}^{RESP}) in order to obtain the absolute binding free energies with the AM1-BCC model (ΔG_{calc}^{AM1BCC}).

Table 4.4 and Figure 4.6 (right, "AM1-BCC") show the results obtained for the calculations with the AM1-BCC charge model. It is possible to see how this had

a small effect on the quality of the correlation of the predictions with experiment, as the Pearson and Spearman coefficients remained almost unvaried. On the other hand, AM1-BCC partial charges further aggravated the overestimation issue. In fact, most of the $\Delta\Delta G_{charges}$ values are negative (average value of -0.7 kcal/mol), indicating that this charge model would predict even more tighter binding for bromosporine. This is reflected in the larger MUE and RMSE returned: 2.15 and 2.54 kcal/mol respectively. Therefore, these results do not suggest that the RESP charges used were the cause of the overly negative free energies we were observing.

Here, I tested only HF/6-31G* Merz-Kollman RESP charges and AM1-BCC charges, as they are the two most popular approaches to obtain partial charges for molecules modelled with GAFF [260, 304–306]. However, different sets of charges can be obtain with other approaches too. Charges derived from HF/6-31G* have often been used for aqueous-phase simulations because this wave function is known to return overpolarised charge sets due to the lack of electron correlation, thus mimicking the polarising effects of water while running the calculations *in vacuo* [304–307]. However, other methods can now be used too for the derivation of possibly more accurate charge sets [307]. For instance, the ff03 family of Amber force fields still uses the RESP procedure to derive charges, but uses B3LYP/cc-pVTZ rather than HF/6-31G* calculations [308]. Density functional theory (DFT) methods such as B3LYP/cc-pVTZ account for electron exchange and correlation, and return less polarised charges than HF/6-31G* [308]. The polarising effect of the solvent can instead be accounted for by using a polarisable continuum model (PCM) [309]. A possible additional set of charges to be tested could therefore be RESP charges derived using DFT approaches.

4.5.4 Effect of torsional parameters

As briefly discussed already, the limitations of small molecule force fields due to the poor transferability of torsional parameters are becoming evident. [281, 282] Certain chemical groups, such as sulfonamides, are particularly challenging when considering that also quantum effects such as the interaction of the nitrogen lone pair with

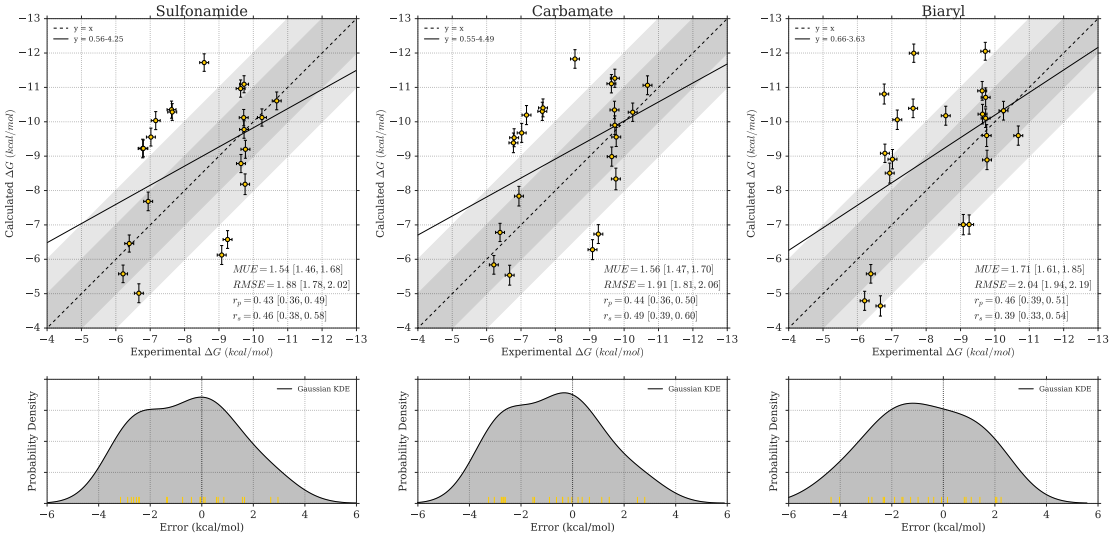


Figure 4.7: Scatter plot of experimental versus calculated affinities for bromosporine with optimised parameters for the sulfonamide, carbamate, and biaryl groups. A linear fit of the data is shown as a black line, and the 1 and 2 kcal/mol error boundaries are shown as shaded grey areas. Below the scatter plots, the error distributions are shown as kernel density estimates.

antibonding orbitals involving sulfur affect the torsional energy around the N-S bond.[310] For this reason, I decided to optimise the dihedrals of the flexible groups in bromosporine by fitting the MM energies to MP2/6-31G* ones using the program *paramfit* provided with the Amber suite. The parameters derived in such a way are specific only for the bromosporine molecule and model we used, and provide a better set of parameters in terms of agreement with QM energies *in vacuo* (although not necessarily with experimental data such as binding affinities). To give an idea of this, the coefficient of determination (R^2) between the GAFF and MP2/6-31G* energies was of 0.83, 0.65, and 0.86 for the conformations considered for the sulfonamide, carbamate, and biaryl torsions, respectively. After fitting additional terms and new torsional parameters, R^2 for the three groups increased to 0.96, 0.99, and 0.98 (Figure B.1). The objective of this exercise was to try to identify some of the causes for the suboptimal performance of the calculation in the case of bromosporine.

The chemical groups for which torsional parameters were optimised were the sulfonamide group, the carbamate group, and again the biaryl torsion. In

order to estimate the effect of this set of parameters, relative binding free energy (RBFE) calculations were carried out, similarly to what was done for the charges. The predicted absolute binding free energies for the optimised models were then derived by adding the $\Delta\Delta G$ obtained from the RBFE calculations, to the ΔG_{calc} values obtained with the initial model. Considering the nonadditive nature of the binding free energy [311], $\Delta\Delta G_{calc}$ values were calculated in series. That is, I first calculated $\Delta\Delta G_{calc}^{sulf}$ (the relative binding free energy difference associated with the new parameters for the sulfonamide), then, starting from the last frame of these simulations I started the free energy calculations to calculate $\Delta\Delta G_{calc}^{carb}$, while keeping the new parameters for the sulfonamide. Then, the same was done for $\Delta\Delta G_{calc}^{biaryl}$, where in these calculations the final model comprised the modified parameters for all the three groups.

The results for these three new bromosporine models are summarised in Figure 4.7 and Table 4.5. As immediately evident, the optimisation of the sulfonamide moiety had a considerable effect on the absolute errors of the calculations. The MUE decreased from 1.76 to 1.54 kcal/mol and the RMSE from 2.13 to 1.88 kcal/mol. In particular, as it is possible to see from the distribution of errors at the bottom of Figure 4.7, the strong overestimation problem was alleviated, so that the errors were distributed more symmetrically around zero. In fact, the resulting mean signed error was in this case of -0.5 kcal/mol, as compared to the previous -1.3 kcal/mol. Unfortunately, this also unexpectedly resulted in a slight deterioration of the correlation between calculated and experimental affinities by 0.04-0.05 in terms of r_p and r_s. The improved fit to QM data did not affect the binding free energy results in the case of the carbamate moiety, with MUE, RMSE, r_p, and r_s having been left mostly unvaried. This might have been expected, however, given the carbamate side chain does not appear to make extensive contacts with the proteins. Finally, the effect of further optimising the parameters for the biaryl torsion in bromosporine was evaluated. However, surprisingly, this new set of parameters resulted in slightly worse binding affinity predictions when measured in terms of MUE, RMSE, and r_s, while it improved only very slightly the linear correlation r_p.

Table 4.5: Bromosporine affinity data and calculated free energies with optimised torsional parameters for the sulfonamide and the carbamate groups and the biaryl torsion. ITC data taken from Picaud *et al.* [159]. All ΔG uncertainties are one standard deviation. ITC uncertainties are an estimate based on the standard deviation of the ΔG values observed in the ABRF-MIRG'02 inter-laboratory assessment [174]. In square brackets the 95% confidence intervals of the statistics based on percentile bootstrap are reported. ΔG_{calc}^{sulf} are the binding free energies for the bromosporine model with optimised torsions for the sulfonamide group; ΔG_{calc}^{carb} are the binding free energies for the model with also the torsions for the carbamate group optimised; ΔG_{calc}^{biaryl} are the binding free energies for the model with the torsions for the biaryl group optimised, in addition to the ones for the sulfonamide and carbamate groups.

Protein	Family	ΔG_{exp} (kcal/mol)	ΔG_{calc}^{sulf} (kcal/mol)	ΔG_{calc}^{carb} (kcal/mol)	ΔG_{calc}^{biaryl} (kcal/mol)
CECR2	I	-10.7 ± 0.1	-10.6 ± 0.3	-11.1 ± 0.3	-9.6 ± 0.3
FALZ	I	-6.8 ± 0.1	-9.2 ± 0.2	-9.5 ± 0.3	-9.1 ± 0.3
PCAF	I	-7.0 ± 0.1	-9.6 ± 0.3	-9.7 ± 0.3	-8.9 ± 0.3
BRD2(1)	II	-9.2 ± 0.1	-6.6 ± 0.3	-6.7 ± 0.3	-7.0 ± 0.3
BRD2(2)	II	-9.6 ± 0.1	-8.8 ± 0.3	-9.0 ± 0.3	-10.2 ± 0.3
BRD3(1)	II	-9.1 ± 0.1	-6.1 ± 0.3	-6.3 ± 0.3	-7.0 ± 0.3
BRD3(2)	II	-9.6 ± 0.1	-11.0 ± 0.3	-11.1 ± 0.3	-10.9 ± 0.3
BRD4(1)	II	-9.7 ± 0.1	-9.8 ± 0.3	-9.9 ± 0.3	-10.7 ± 0.3
BRD4(2)	II	-9.8 ± 0.1	-9.2 ± 0.3	-9.6 ± 0.3	-8.9 ± 0.3
BRDT(1)	II	-9.8 ± 0.1	-8.2 ± 0.3	-8.3 ± 0.3	-9.6 ± 0.3
CREBBP	III	-7.6 ± 0.1	-10.3 ± 0.2	-10.4 ± 0.3	-12.0 ± 0.3
EP300	III	-6.8 ± 0.1	-9.2 ± 0.3	-9.4 ± 0.3	-10.8 ± 0.3
BRD1	IV	-7.6 ± 0.1	-10.4 ± 0.2	-10.3 ± 0.3	-10.4 ± 0.3
BRD9	IV	-9.7 ± 0.1	-11.1 ± 0.2	-11.3 ± 0.3	-10.1 ± 0.3
BRPF1B	IV	-8.6 ± 0.1	-11.7 ± 0.3	-11.8 ± 0.3	-10.2 ± 0.3
BAZ2A	V	-7.2 ± 0.1	-10.0 ± 0.3	-10.2 ± 0.3	-10.1 ± 0.3
TIF1	V	-6.7 ± 0.1	-5.0 ± 0.3	-5.5 ± 0.3	-4.6 ± 0.3
TAF1(1)	VII	-6.9 ± 0.1	-7.7 ± 0.3	-7.8 ± 0.3	-8.5 ± 0.3
TAF1(2)	VII	-10.2 ± 0.1	-10.1 ± 0.3	-10.3 ± 0.3	-10.3 ± 0.3
TAF1L(2)	VII	-9.7 ± 0.1	-10.1 ± 0.2	-10.3 ± 0.3	-12.1 ± 0.3
PB1(5)	VIII	-6.4 ± 0.1	-6.5 ± 0.2	-6.8 ± 0.3	-5.6 ± 0.3
SMARCA4	VIII	-6.2 ± 0.1	-5.6 ± 0.3	-5.8 ± 0.3	-4.8 ± 0.3
Statistics					
Mean Unsigned Error			1.54[1.46, 1.68]	1.56[1.47, 1.70]	1.71[1.61, 1.85]
Root Mean Square Error			1.88[1.78, 2.02]	1.91[1.81, 2.06]	2.04[1.94, 2.19]
Pearson's r			0.43[0.36, 0.49]	0.44[0.36, 0.50]	0.46[0.39, 0.51]
Spearman's ρ			0.46[0.38, 0.58]	0.49[0.39, 0.60]	0.39[0.33, 0.54]

Based on these results, it appears that the sulfonamide parameters were contributing towards the systematic overestimation of the binding free energies for bromosporine. On the other hand, further optimisation of the torsions did not result in improved agreement with ITC data in terms of absolute values, linear correlation, or ranking ability.

4.6 Discussion

Here, I presented two test cases in which the prediction of the different affinities for a single ligand binding to multiple bromodomains was attempted. Excellent agreement with experimental data was achieved for the two RVX compounds, while the bromosporine ligand, which binds to a larger range of proteins, proved to be more challenging. It is important to note how the quality of the ligand parameters might vary largely between different small molecules depending on the chemical groups present. The fact that re-parameterisation of the torsions for the bromosporine model alleviated the systematic overestimation of binding affinities corroborates the hypothesis that transferability of dihedral parameters in small molecules is indeed problematic. Parameterisation approaches where some of the ligand parameters, such as charges and torsions, are derived on an ad-hoc basis for the specific molecule at hand might provide more accurate organic molecule models for the study of dynamics and binding free energies [282]. This is certainly not the only issue with classical force field models, however, one that can be tackled relatively easily without modification of current functional forms, while potentially having a significant positive impact. Overall, small molecule force fields are still relatively young when compared to protein force fields, with much room for improvement. As demonstrated by the recent OPLS3 force field, little adjustments such as off-site charges and careful parametrisation that takes advantage of increasing availability of experimental data can provide better models of protein-ligand interactions within the framework of standard functional forms [288]. In this regard, also work on host-guest systems, for which highly converged calculations can be obtained, might prove valuable for the parameterisation of the next generations of small molecule force fields [312].

In the case of bromosporine, even with the optimisation of the torsional parameters, reasonable but not excellent agreement with ITC data was eventually obtained in terms of correlation (r_p and r_s between 0.4 and 0.5). This was caused by large (> 2 kcal/mol) overestimation or underestimation of the binding free energy in specific cases (e.g. BRD2(1), BRD3(1), CREBBP, EP300, PCAF). The

identification of the causes for this disagreement with ITC proved challenging, as a large number of potential reasons can be conceived. Limited sampling, or more likely the failure to sample relevant states such as certain side chain rotamers, might contribute towards these errors [50, 57]. Identifying specific sampling deficiencies on a case-by-case basis can however be impractical, in particular with the prospect of being able to apply the calculations to larger numbers of protein-ligand pairs. The combination of free energy calculations with enhanced sampling techniques might thus be beneficial, as it does not require the identification of specific degrees of freedom that have been undersampled [47, 313, 314]. In addition, issues with the parameters (including charges and vdW) of certain chemical groups of the ligand might also result in over/under-estimation of interaction energies and consequently affinities. For instance, the average polarisation state represented by fixed-charge force field might result in inaccuracies in cases where changes in polarisation upon binding occur [254, 315, 316]. As different binding pockets have different charge distributions, this might affect the results of the calculations differently across different proteins. Cation-π interactions have been shown to be important for the binding of certain ligands to bromodomains such as CREBBP [129, 142]; this type of protein-ligand interactions too are unlikely to be captured accurately by classical fixed-charge force fields. These are only some of the challenges that might need to be addressed in order to push further the accuracy, reliability, and generality of both relative and absolute binding free energy calculation. Tautomerisation, changes of protonation states, and possible effects due to the presence of the buffer, are other issues that can arise in the prediction of protein-ligand binding [317, 318].

Despite all the challenges lying ahead, overall, the results presented here are encouraging. In terms of absolute errors, the predictions achieved a RMSE below the 1 kcal/mol mark for the RVX ligands and below 2 kcal/mol for bromosporine. The ability to quantitatively reproduce binding free energies, rather than just correlate with them, has been a difficult task for decades and for most computational approaches [25]. Nowadays, physics-based models of chemical and biological matter along with computer simulations seem, however, to be able to provide the necessary

physical accuracy and degree of sampling in order to reproduce one of the most important thermodynamic quantities in biology and pharmacology. It is, nonetheless, wise to exercise some cautious optimism. In fact, the size of the test sets presented here and in the previous chapter are too limited to provide a really broad picture of the expected performance of the calculations. However, it is important to start to assess the accuracy and precision of these calculations even if on a small number of specific systems. As the computer power available continues to increase along with algorithmic improvements, in the near future it will be possible to expand these validation studies to much larger test sets, similarly to what have been just done for RBFE calculations [47]. To give a very rough estimate, here we obtained almost one affinity prediction per day using 42 Intel® Xeon® CPUs (E5-2697 v2, 2.7GHz, 12 cores per CPU) of a Cray® XC30™ supercomputer. Acceleration with GPUs would already allow us to obtain approximately 2 predictions per day. If Moore's law were to hold in the next decade, in 2026 one could expect to obtain over 60 predictions per day, when considering also algorithmic advances [319].

However, in recent years there has been a barrier to increased clock speed due to heat dissipation issues and vendors have rather focussed on increasing the number of cores on chips. The semiconductor industry is also approaching the physical limit for density of transistors on microprocessors. Current high-end microprocessors have circuit features that are already about 14 nm across, and at smaller scales electron behaviour will be affected by quantum uncertainties that would make transistors unreliable. It is thus not possible to take for granted that Moore's law will be able to last for another decade [320, 321]. On the other hand, with the increasing availability of large number of cores on chips, thanks to the above-mentioned vendor strategy, there have been concurrent improvements in the parallel performance of MD codes [321]. In addition, recent years have seen the rise of cloud and distributed computing [322–324]. Since alchemical ABFE calculations involve running multiple and relatively short simulations (rather than one single long one), it will thus be possible to exploit both faster parallel MD codes and the larger number of available cores in order to increase calculations throughput

[321]. With expanded chemical and biological spaces used for testing, the results will provide a better picture of the capabilities and pitfalls of the methodology as well as its most fruitful fields of applicability.

In our opinion, ABFE calculations have the highest potential at the hit discovery and lead development stages, complementing established techniques such as docking as well as emerging ones like RBFE calculations. The objective is not to have perfect predictions and replace experiments, but to shorten and reduce the cost of development cycles. During hit discovery, one can foresee virtual screening protocols where a number of techniques are employed hierarchically. For instance, docking will always be faster than ABFE calculations, and could be used for the initial enrichment of binders within the library screened; this could be followed by intermediate approaches in terms of accuracy and speed, such as implicit solvent calculations like MMPBSA; and finally, for the most accurate rescoring and ranking, ABFE calculations may be used. At the lead development stage, these calculations could be used to assess the likelihood of a ligand with a different scaffold to retain binding, or to assess different ligands for their ability to hit the target selectively, as described in this chapter.

4.7 Conclusions

In this work, it was shown how absolute binding free energy calculations based on molecular dynamics simulations can be applied in order to computationally predict the affinity profile of a ligand across multiple proteins with no prior knowledge of the structure of the complexes or affinities for similar ligands. The prediction of binding affinities and the engineering of selectivity are both still major challenges in drug design; however, with increasing computational power at our hands and with more rigorous and general computational approaches these problems can start to be tackled. It is here shown how a thorough simulation protocol can, in the cases here considered, achieve good correlation and agreement with experimental values for drug-like ligands. It is, however, evident that the accuracy of the calculations is both ligand and protein dependent, and that further investigation is needed in order to

assess the effectiveness of these predictions in different case scenarios and on different systems. While the accuracy of the calculations for other protein families still needs to be established, the performance of the calculations reported here is encouraging.

Everything should be made as simple as possible, but not simpler.

— Albert Einstein

5 Comparison to MMPBSA

Contents

5.1 Introduction 128
5.2 Methods 131
5.2.1 MMPBSA calculations 131
5.2.2 Entropy calculations 132
5.2.3 Data analysis 132
5.3 Case 1: predicting ligand binding to BRD4(1) from X-ray structures 133
5.4 Case 2: predicting ligand binding to BRD4(1) from docking poses 137
5.4.1 Ranking of docking poses 137
5.4.2 Affinity predictions 138
5.5 Case 3: predicting RVX-OH vs RVX-208 binding and selectivity 141
5.5.1 Identification of correct RVX-OH pose 141
5.5.2 Affinity predictions 142
5.6 Case 4: predicting the selectivity profile of bromosporine 146
5.7 Discussion 147
5.8 Conclusion 152

In the previous two chapters, I evaluated the performance of absolute alchemical free energy calculations in a few different scenarios involving bromodomains. In this chapter I will compare the previous results with what is obtained via a computationally less expensive approach that approximates binding free energy as

a sum of a few molecular mechanics terms with the use of implicit solvent.

5.1 Introduction

Alchemical free energy calculations are theoretically rigorous and generally perceived as more accurate than end-point methods, however, they also are computationally much more expensive. While rigorous free energy calculations have a smaller number of empirical constants to adjust in a system-dependent fashion as compared to MMPBSA [230, 325], they also tend to have a more complex setup and the user can fall into many potential pitfalls [326, 327]. Choosing which approach to employ for a specific system can therefore be difficult, as one wonders whether the additional human and computational cost would be rewarded by a more accurate result.

While many notable studies on the performance of MMPBSA and MMGBSA approaches are present in the literature [31, 32, 34, 35, 328–331], only a few have compared end-point methods to more rigorous pathway approaches and, to my knowledge, none to absolute binding free energy calculations. Wang *et al.* [314] have compared the performance of relative alchemical free energy calculations to MMGBSA for a set of 6 cyclin-dependent kinase 2 (CDK2) ligands, reporting the higher performance of their alchemical approach. Homeyer *et al.* [325] evaluated the performance of MMPBSA, linear interaction energy (LIE) and thermodynamic integration (TI) for three sets of 25, 29, and 29 ligands binding to, respectively, factor Xa, CDK2, and the mineralcorticoid receptor. The performance was dependent on the details of the protocol used and the subsets considered, nevertheless they concluded that both MMPBSA and TI could provide valuable predictions when taking into account certain data set-specific features, contrary to LIE. The Pearson correlation coefficients ranged from 0.00 to 0.62 for MMPBSA, and from 0.02 to 0.58 for TI. Recently, Wang *et al.* [47] performed a large retrospective test of relative free energy calculations encompassing 199 ligands and 8 protein systems, where the resulting weighted average correlation for the calculations was 0.75, while it was found to be 0.35 for MMGBSA. A similar study evaluated the performance of the same free energy protocol and MMGBSA calculations on 96 fragments targeting 8

proteins, reporting a weighted average correlation of 0.65 for the pathway and of 0.41 for the end-point approach [332]. All of these studies evaluated the performance of the methods in ranking the affinity of chemically similar ligands. However, there are many scenarios in drug discovery where it would be beneficial to be able to rank any ligand-protein pair, independently of the similarity of the ligands within the set. At the lead discovery stage one might want to rank sets of very different ligands; if a crystal structure of the complex is not available, it would be useful to accurately re-score docking poses in order to identify the most likely orientation of the ligand; if selectivity is of interest, one would like to predict the affinity of the ligand for multiple different protein targets. Despite the challenges of these applications, they can in principle be tackled by end-point methods such as MMPBSA, which approximates binding free energy, and absolute binding free energy calculations (ABFE). However, the performances of the two methodologies in these common drug discovery scenarios have not been compared before. Despite the fact that in theory more rigorous methods should return a better performance, this is not necessarily the case in practice. For instance, for end-point methods it has often been observed that less theoretically rigorous approaches (such as the use of a single rather than separate trajectories, the neglect of the entropic term, or the use of the Generalized Born model) resulted in better performance thanks to improvements in the precision of the calculations [35, 230, 247, 248, 328].

Given what we just discussed above, here we evaluate the performance of MMPBSA on the same test sets presented in the previous two chapters. Thus, the following four case studies will be considered: 1) prediction of the affinities of 11 diverse ligands binding to BRD4(1), assuming knowledge of their crystal structures; 2) prediction of the affinities of 11 diverse ligands binding to BRD4(1), after ranking their poses as returned by docking; 3) prediction of the different selectivity profiles of RVX-OH and the closely related RVX-208 for 7 BET bromodomains; 4) prediction of the selectivity profile of bromosporine, a broad-spectrum BRD inhibitor, for 22 different bromodomains. Thus, a total of 47 affinity data-points will be considered, as Cases 1 and 2 are comparing the calculations to the same set of binding free

energies. We will focus in particular on the ability of the prediction to correlate with experimental values, as measured by the Pearson and Spearman correlation coefficients. In fact, the ability of end-point methods to reproduce absolute values is known to be poor [31, 32, 35, 230, 325, 328, 329]. The trajectories we use here for the MMPBSA calculations are the same that were used for the alchemical free energy calculations. Therefore, the same force field parameters for protein and ligands are employed and both calculations sample the same conformations, so that potential issues related to sampling or the accuracy of the physical model affect both calculations in a way that does not impair comparison of the approaches. In the scenario of a perfectly accurate physical representation of the chemical systems and infinite sampling, the better performance of rigorous calculations as compared to MMPBSA would be the consequence of the approximations of the end-point method. However, given limited sampling of protein and ligand conformations, and inaccuracies in the force fields, the approximations in MMPBSA might result in the cancellation or reduction of errors and faster convergence toward a better correlation with experimental values.

In the first instance, I adopted a standard single-trajectory MMPBSA protocol, neglecting the solute entropic contribution. This contribution is typically estimated through quasi-harmonic or normal mode analysis. However, this term is generally disregarded, as the full sampling of the free energy landscape required by these approaches is computationally demanding and the benefits of including these terms are controversial [31, 35, 246–248, 333]. Nonetheless, Duan *et al.* [249] have recently proposed a computationally simple and efficient approach for the estimation of the entropic term, which they named *interaction entropy*. I therefore decided to evaluate the effect of including this term on the performance of the calculations. In addition, Maffucci and Contini [334] have suggested that the inclusion of an explicit hydration shell during the calculations can lead to a significant improvement in the correlation with experiment. As BRDs are furthermore known to contain structural waters in their pockets, bridging the binding of their inhibitors, I decided to test this approach as well.

5.2 Methods

5.2.1 MMPBSA calculations

All MMPBSA calculations were performed using the set of scripts provided with GMXPBSA 2.1.1 [335]. Equidistant frames were extracted from the first window (unrestrained and fully interacting ligand) of the alchemical free energy calculations every 20 ps, from 5 ns to 15 ns, for a total of 501 snapshots, for the calculations with 15 ns windows in Chapter 4. For the free energy calculations that employed 10 ns windows in Chapter 3, frames were extracted from 2 ns to 10 ns every 16 ps, again for same total number of snapshots. In order to evaluate the effect of including explicit waters, we repeated the calculations on the same frames while retaining the N closest water molecules to the ligand, where N = [10,20,30,40,50], as suggested by Maffucci and Contini [334]. The MD trajectories were processed with MDAnalysis [336] in order to extract the N water molecules closest to any atom in the ligand for all frames considered. During the calculations the explicit water molecules were considered as being part of the protein.

The terms involved in the MMPBSA calculation of the binding free energy (here denoted ΔG_{MMPBSA}) are described in § 2.6.2. Molecular mechanics energies for the vdW term (E_{LJ}) were calculated with Gromacs, while the coulomb tool in APBS 1.3 [272] was employed for the electrostatic term (E_{coul}); note that $E_{bond} = 0$ as the single trajectory method was adopted. The polar (ΔE_{polar}) and nonpolar contributions ($\Delta E_{nonpolar}$) to the free energy of solvation were calculated with APBS 1.3[272]. For the polar solvation energy contribution, the nonlinear Poisson-Boltzmann equation was solved using a value of 80 for the exterior dielectric constant, and a value of 2 for the solute dielectric constant. The temperature was set to 298.15 K and the salt concentration to 0.15 M. The nonpolar term was considered proportional to the solvent accessible surface area (SASA), $\Delta E_{nonpolar} = \gamma \cdot SASA$, where $\gamma = 0.0227$ kJ mol^{-1} Å^{-2}. A probe radius of 1.4 Å was used to define the dielectric boundary. When an explicit ligand hydration shell was included in the

calculations, the water molecules were considered as part of the receptor, as per the original implementation of the method [334].

5.2.2 Entropy calculations

The entropic contribution was estimated as recently proposed by Duan *et al.* [249] and described in § 2.6.2. The interaction entropy was calculated on the 501 frames extracted from the simulations, so to consider the additional N water molecules included as part of the protein. The short- and long-range Lennard-Jones and Coulombic interaction energies were calculated via the *gmx energy* command in Gromacs 5.1 [187], after performing a re-run of the 501-frame trajectories where all the molecules not of interest were stripped from the system.

5.2.3 Data analysis

The ΔG_{MMPBSA} values reported are the mean of the values obtained for the 501 snapshots extracted from the MD trajectory. Where multiple calculations were performed, the uncertainty in the ΔG_{MMPBSA} estimate was taken as the sample standard deviation of the repeated calculations. In all other cases the standard deviation of the mean was estimated by bootstrap [274, 337]. 10^5 bootstrap samples were generated through random resampling with replacement of the 501 free energy values, and the standard deviation of the sampling distribution was taken as the uncertainty of the mean.

Correlation to experimental values was evaluated using the Pearson (r_p) and Spearman (r_p) correlation coefficients. 95% confidence intervals (CIs) for the correlations were obtained by percentile bootstrap [274, 337]. 10^5 samples were built by random sampling from the distributions of measured affinity values assuming normality; then, 10^5 samples of MMPBSA predicted affinities were built in the same fashion based on the means and standard deviation calculated as described above. From the resulting distribution of 10^5 correlation coefficients, the $\alpha/2$ and $1-\alpha/2$ percentiles of the bootstrapped coefficients were taken as the confidence interval, where $\alpha = 0.05$ for a 95% CI. Here I will report the coefficient for the

original samples in front of the 95% CI in square brackets. The data from the alchemical free energy calculations were treated in the same way: 10^5 correlation coefficient samples were obtained by resampling from the normal distributions of the calculated binding free energies.

The distribution of the difference in r_p and r_s between alchemical calculations and MMPBSA was then obtained by subtracting the values of the 10^5 bootstrapped coefficients for both approaches. All distributions are shown as Gaussian kernel density estimates. This approach does not assume any particular distribution of the sampling distributions, and shows the estimated range of possible correlation results for ABFE versus MMPBSA.

The analysis was performed using a set of scripts I wrote in Python 2.7, using the *matplotlib* and *seaborn* libraries for plotting, and *pandas*, *numpy* and *scipy* for data handling and statistics.

5.3 Case 1: predicting ligand binding to BRD4(1) from X-ray structures

In § 3.3 it was shown how ABFE calculations could return a Pearson correlation of 0.87 [0.73, 0.92] and Spearman of 0.84 [0.70, 0.94] when starting from the crystal structures of the protein-ligand complexes. As discussed previously, this represents a best case scenario that is unlikely to present itself in a real drug discovery campaign, however it is a good way of testing the performance of affinity predictions while removing the additional sources of uncertainty that might confound the interpretation of the results. Here I compare the correlations achievable with standard MMPBSA (denoted "W0"), as well as MMPBSA with the inclusion of a number of explicit water molecules (denoted "W10" to "W50"), and with the addition of the entropy estimate (denoted with an additional "e" after the number of explicit water molecules included; e.g. "W10e" indicates calculations where 10 explicit waters were included as well as the entropic term).

Figure C.1 shows the scatter plot of all MMPBSA calculations performed on this set of 11 inhibitors binding to BRD4(1), and Figure 5.1 summarises the

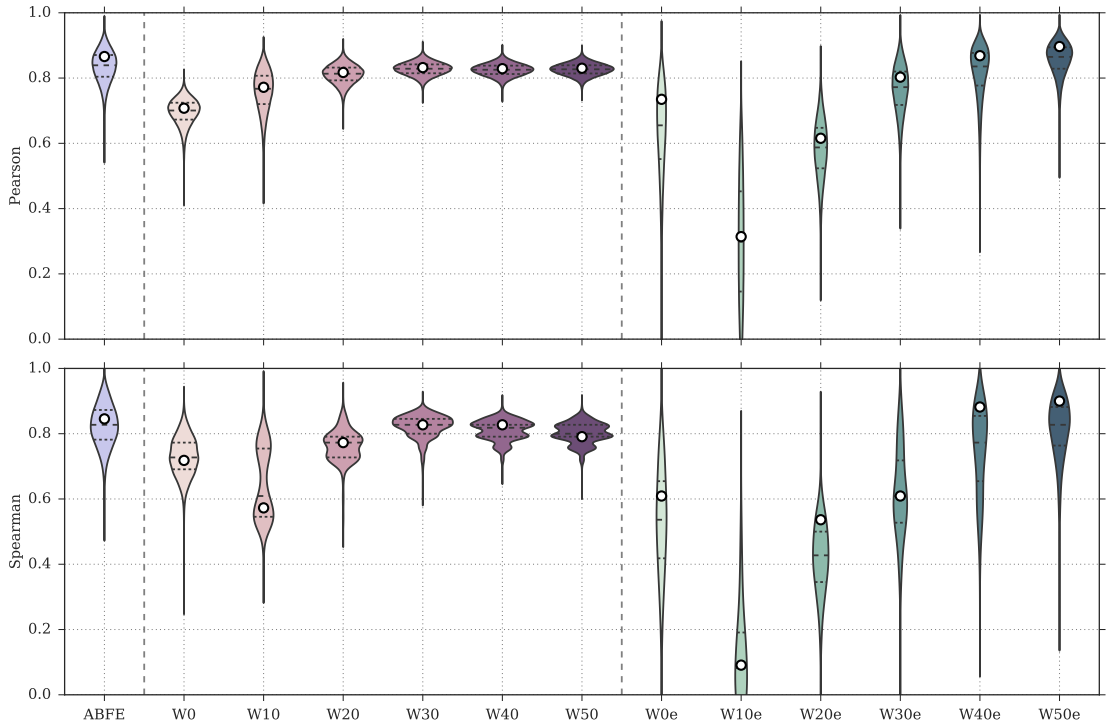


Figure 5.1: Pearson and Spearman correlations for ABFE and MMPBSA for Case 1. The plot shows Gaussian kernel density estimates for the distributions of r_p and r_s as obtained by 10^5 bootstrap samples. A white circle shows the value obtained with the original sample.

correlation obtained by them. White circles show the correlations obtained with the original sample of data, while the violins represent the distribution obtained through bootstrapping, which takes into account the uncertainty of the calculations. Thus, a wide distribution highlights a large uncertainty associated with the correlation values.

Standard MMPBSA ("W0") achieves good agreement with experiment in terms of linear correlation and ranking ability, with $r_p = 0.71$ and $r_s = 0.72$ (Table 5.1). Including an explicit hydration shell for the ligand improves the correlation with experiment. In particular, the addition of 30 explicit water molecules in this case results in the best correlation with experiment for the calculations without the entropy term ($r_p = 0.83$ and $r_s = 0.83$).

It is possible to conceive that the number of waters needed to capture the hydration shell of a ligand might depend on its size, described, for instance, by its SASA. It could be possible to design a set of calculations aimed at identifying a

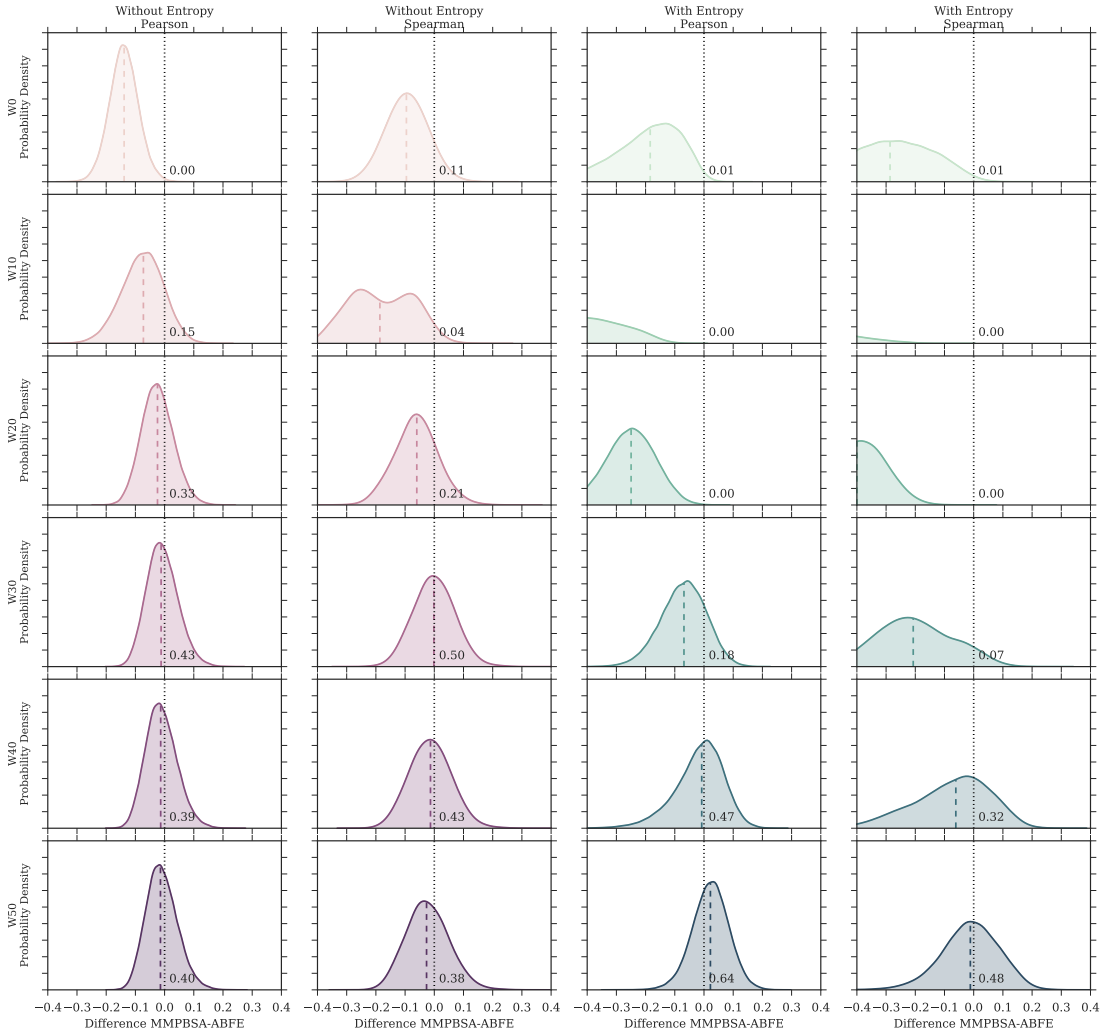


Figure 5.2: Correlation difference between MMPBSA and ABFE calculations. Shown is the Gaussian kernel density estimate for the distribution of the difference of r_p and r_s as obtained by 10^5 bootstrap samples. The median of each distribution is shown as a dashed line. The value of the area under the curve above zero is shown on each chart.

relationship between the size of a ligand (described as SASA, but possibly also by additional variables, such as mass or number of rotatable bonds) and the number of explicit waters needed in order to return the highest correlation to experiment. For instance, one could separate a large number of ligands into different groups according to their size, and test whether there is a significant correlation between size and the optimal number of waters to be treated explicitly. Having such relationship would be very useful since, at the moment, the user does not know in advance how many explicit water molecules it would be most appropriate to include in the calculations.

Figure 5.2 shows the difference in performance between MMPBSA and ABFE calculations, taking into account the uncertainty of the experimental measurements and of the computational predictions. The more to the left (negative values) the shown distribution is, the more ABFE performance is superior to MMPBSA, and vice versa for the distribution being on the right-hand side (positive values). In this figure, it is possible to see how when more explicit waters are included in the calculations, the fraction of the distribution over the zero mark increases, indicating a higher likelihood for the MMPBSA calculations to return better correlation than ABFE, based on their respective uncertainties. The inclusion of any number of waters between 10 and 50 improves the overall correlation, despite a drop in Spearman rank for the sample W10. Nonetheless, as previously mentioned, the best performance is obtained with W30. In this case, r_s effectively corresponds to what obtained with ABFE, and r_p is only marginally worse.

In this test case, introducing the entropy term in the calculations does not bring much benefits to the predictive power of MMPBSA. In most cases (W0 to W40), the inclusion of this term is detrimental for the correlation with experiment. Only for W50 adding the entropy term results in both better r_p and r_s. Surprisingly, this is however the best MMPBSA prediction in the set, with r_s almost equivalent to ABFE, and r_p slightly superior to it.

The uncertainties in the free energies presented in this test case are derived from three repeated calculations, and from the standard deviations in Figure C.1 and the distributions in Figure 5.1 one can see that the entropy estimate carries large uncertainties with it. In fact, the standard deviations from the triplicate MMPBSA calculations without entropy estimation are small, suggesting the calculations are well converged within the time scales simulated and the number of frames considered. However, when entropies are considered, these uncertainties become much larger and this is reflected in the wide range of correlations obtained by the bootstrap samples (Figure 5.1).

5.4 Case 2: predicting ligand binding to BRD4(1) from docking poses

In this section, I compare the MMPBSA results to what obtained in § 3.4. When starting from a set of one to five docking poses for each ligand, ABFE calculations managed to correctly identify the crystallographic binding pose for 9/10 ligands (note that one of the ligands was modelled into the binding pocket), and returned a Pearson correlation of 0.78 [0.68, 0.83] and a Spearman rank of 0.78 [0.55, 0.85].

5.4.1 Ranking of docking poses

As the ligand-protein structures were obtained in this case by docking the ligands into an apo structure of BRD4(1), the first challenge is to re-score the poses in order to identify the most stable ones. As the crystal structures of the protein-ligand complexes are available for 10 out of the 11 ligands considered, it is possible assess the performance of the MMPBSA calculations in identifying the crystallographic poses. It needs however to be borne in mind that this is a small test set for this assessment, and that each ligand has a small number of potential poses associated with it, due to the clustering procedure described in § 3.2.4. Therefore, it is expected that just by chance 5/10 crystallographic poses would be identified; docking correctly ranked the poses for 7/10 ligands.

Figure 5.3 shows the number of correctly ranked poses by all twelve MMPBSA calculations performed. As for ABFE calculations, I carried out MMPBSA calculations starting from all poses proposed by the docking procedure, and took the pose with the lowest energy as the one suggested to be the most stable one. If this pose corresponded to the one closest to the X-ray structure, the pose ranking would be considered successful. From the figure, it is possible to see how standard MMPBSA (W0, in pink) would correctly rank the poses of only 3 ligands. Adding the entropy estimate increases this to 5, which is still not better than chance. When adding explicit waters to the calculations, the situation improves slightly and in the cases of W10, W20, and W30, MMPBSA correctly picks the X-ray poses for 6/10 ligands.

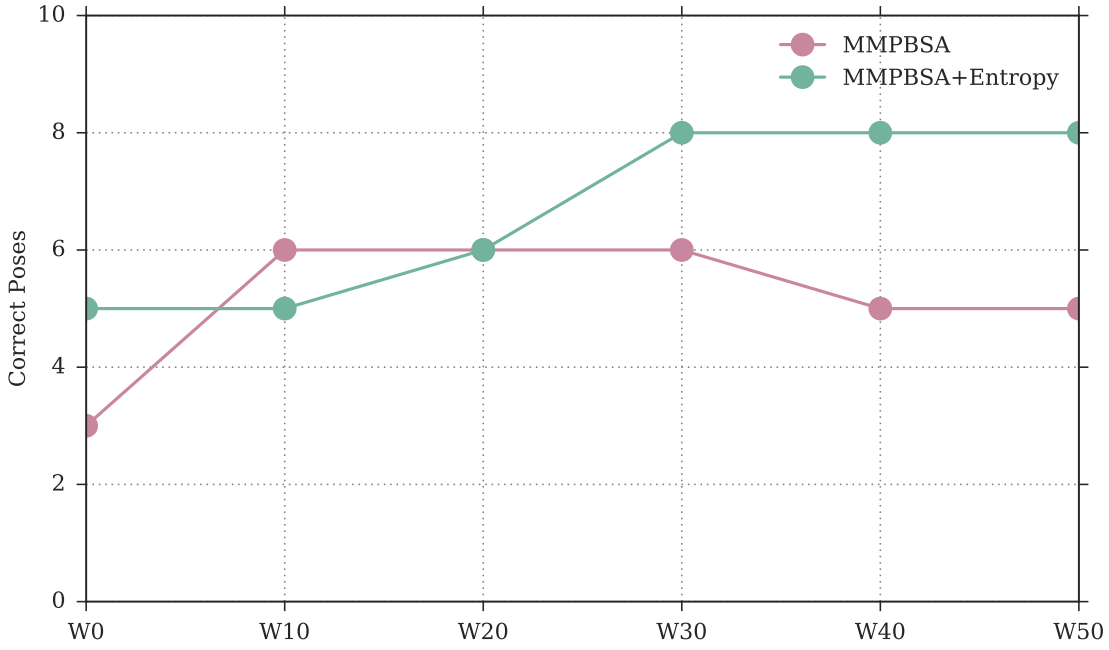


Figure 5.3: Number of correctly ranked the poses for the ligands docked in Case 2.

The best performance is obtained however when the entropy term is added to the ΔG_{MMPBSA}. W30e, W40e, and W50e correctly rank the poses for 8/10 ligands.

Interestingly, no MMPBSA calculation managed to correctly select the X-ray pose of Ligand 1 (Figure 3.1) out of the three docked ones. On the other hand, W10e to W50e correctly identified the best pose for Ligand 6, while this was not correctly predicted by ABFE calculations. Other poses that were challenging for MMPBSA were the ones for Ligands 2 and 7, for which most MMPBSA protocols employed here did not succeed in identifying the X-ray pose.

5.4.2 Affinity predictions

For each set of poses for a specific ligand, the lowest binding free energy was taken as the binding free energy for that ligand. Based on these data, the agreement with experimental affinities was evaluated as in Case 1. The scatter plots of all MMPBSA calculations performed on this set of 11 inhibitors when starting from docked poses are shown in Figure C.2. Figure 5.4 summarises the performance of

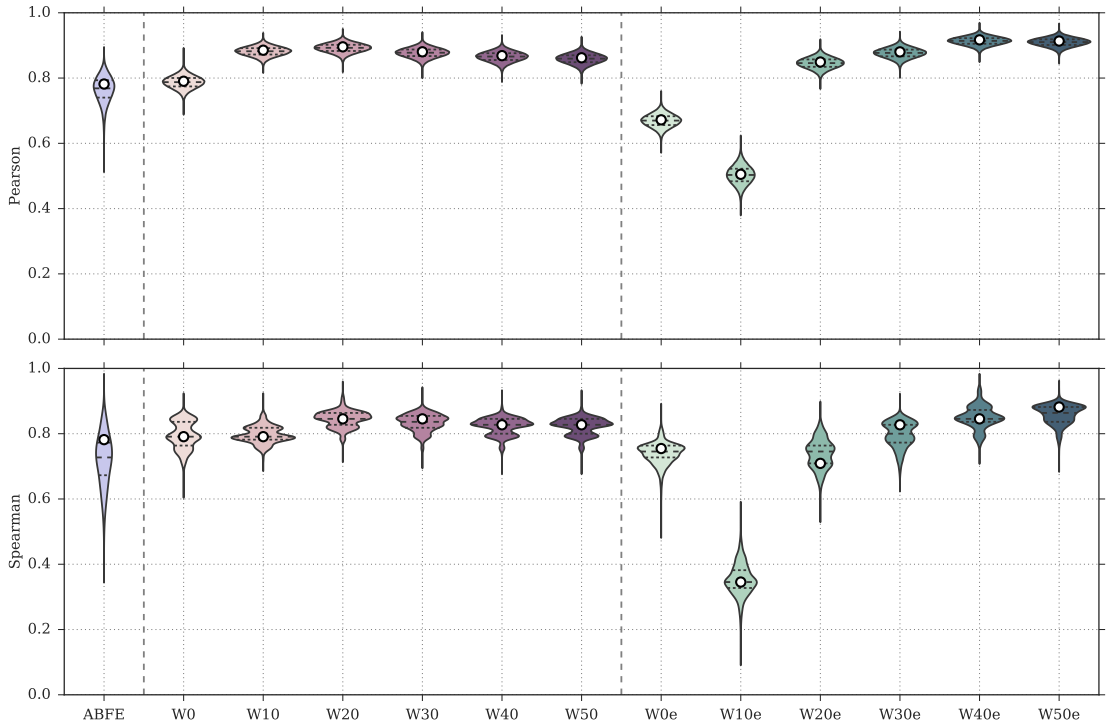


Figure 5.4: Pearson and Spearman correlations for ABFE and MMPBSA for Case 2. The plot shows Gaussian kernel density estimates for the distributions of r_p and r_s as obtained by 10^5 bootstrap samples. A white circle shows the value obtained with the original sample.

the calculations in terms of Pearson and Spearman correlation with experiment, and Figure 5.5 compares the MMPBSA performance to that of ABFE calculations.

From Figure 5.4 it is possible to see that also in this case the inclusion of explicit waters in the calculations improves the correlation with experiment, both in terms of linear correlation and ranking. Without the entropy term, the best performance is achieved by W20, with $r_p = 0.90$ and $r_s = 0.85$. Surprisingly, the fact that only about half of the crystallographic binding modes were identified by the calculations did not affect much the ability to rank ligand affinities, comparing the correlations obtained here and in Case 1. Figure 5.5 show that while for W0 there is still overlap between the distributions of r_p and r_s for ABFE and MMPBSA, for W10 to W50, r_p and r_s these are significantly different with MMPBSA providing higher correlations by $\sim$0.1 points on average.

Adding the entropic term to W0 had a negative effect on the correlation to

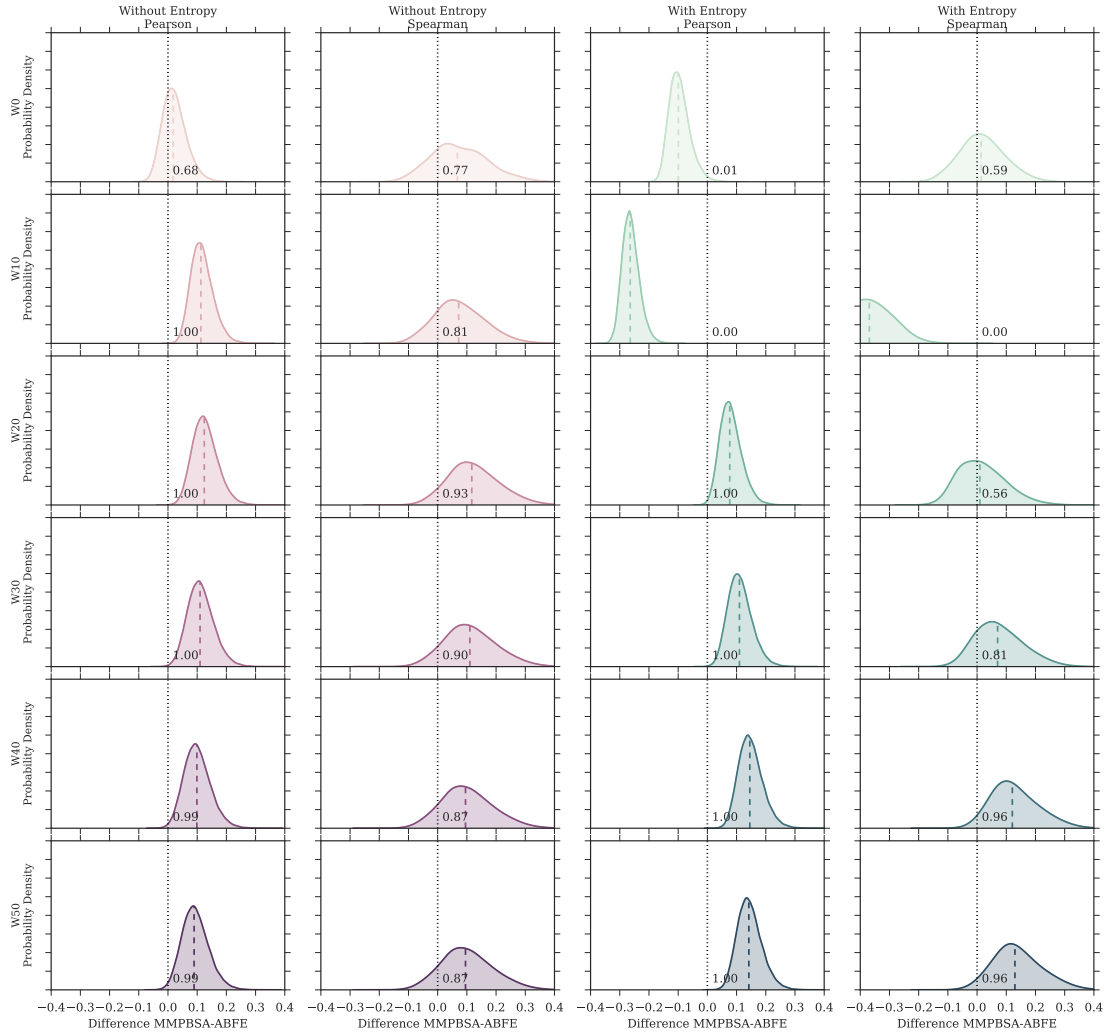


Figure 5.5: Correlation difference between MMPBSA and ABFE calculations for Case 2. Shown is the Gaussian kernel density estimate for the distribution of the difference of r_p and r_s as obtained by 10^5 bootstrap samples. The median of each distribution is shown as a dashed line. The value of the area under the curve above zero is shown on each chart.

experiment, with r_p for MMPBSA now being significantly shifted to lower value as compared to ABFE, and with r_s too having an increased overlap with zero difference. Similarly to what observed for Case 1, W10e shows a plunge in correlation as compared to W0e, which is recovered with increasing inclusion of explicit water molecules. Like for W10 to W50, W30e to W50e show a superior correlation with experiment than ABFE by ∼0.1 points on average, in particular in terms of Pearson correlation. W20e presents superior r_p to ABFE, but comparable r_s. These results

can be somewhat rationalised on the basis of the number of correct X-ray poses recovered from the docking set, as the calculations that correctly identified the highest number of crystallographic poses also returned better correlations. This effect is more visible for the calculations that included the entropy estimate. In fact, W30e to W50e identified the X-ray poses for 8/10 ligands, only one less as compared to ABFE. In addition, this pose is for Ligand 1 and, as discussed in § 3.4.1, while pose 1-a is the closest to the crystal structure, pose 1-c too captures the right binding mode. In fact, the specific position of the tail of the molecule observed in the X-ray structure (PDB-ID 4OGI) is likely affected by crystal packing rather than being a lower energy orientation.

In summary, in this test case, MMPBSA calculations with the inclusion of a number of water molecules between 30 and 50 and with an entropy estimate, resulted in effectively comparable performance to ABFE for the identification of crystallographic binding poses, and a superior performance in terms of correlation with experimental affinity values. Without the entropic term, MMPBSA calculations were superior to ABFE in terms of r_p and r_s, but surprisingly also inferior in the ability to identify the correct lowest energy pose from the set of docking poses.

5.5 Case 3: predicting RVX-OH vs RVX-208 binding and selectivity

This test case, as Case 2, presents both the challenge of ranking poses obtained from docking, and then scoring them in order to recover the correct order of affinities. In contrast to Case 2, however, here more than one ligand binding to multiple BRDs is considered. The background of this test case is explained in more detail in § 4.4. In this instance, ABFE calculations showed a Pearson correlation of 0.75 [0.67, 0.80] and a Spearman rank of 0.78 [0.64, 0.85].

5.5.1 Identification of correct RVX-OH pose

The first challenge in this case scenario is the correct identification of RVX-OH binding pose for the 7 BET bromodomains here considered. In fact, this ligand

displays two different binding modes for the first and second bromodomains of this group of BRDs, as shown in Figure 4.3.

Like I did for ABFE, the MMPBSA calculations were performed for both potential RVX-OH poses for the 7 BRD considered, and the pose returning the lowest energy would be taken as the predicted binding pose for that bromodomain. Given that only two poses are taken into account here, there is a 50% random chance to identify the correct binding mode. Thus, any free energy calculation protocol correctly identifying less than 5 binding modes would not be performing better than random.

As shown in § 4.4, ABFE calculation correctly picked the orientation of RVX-OH in all binding pockets considered. Figure 5.6 shows the number of correctly identified binding orientations by the MMPBSA calculations. W0 suggested that RVX-OH binds to all proteins in the same fashion, adopting the "vertical" binding mode; however, since in the test set there are 4 first bromodomains, to which the ligand binds in the "horizontal" mode, this protocol correctly identifies only 3/7 binding poses. The situation improves markedly when including explicit waters in the calculations: W10 to W20 identified 6/7 binding poses, with the only problematic protein being BRDT(1). W40 and W50 also did not manage to correctly pick the mode of binding for BRD3(2).

The introduction of the entropy term had a large positive impact on the ability of the calculations to predict RVX-OH binding modes. In fact, W0e is the only protocol that correctly identified all binding poses, like ABFE. However, contrary to what observed in Case 2, adding an explicit ligand hydration shell in the calculations resulted in decreasing performance when also the entropic estimate was included. In fact, W10e correctly predicted 6/7 poses, W20e 5/7, and W30e to W50e only 4/7.

5.5.2 Affinity predictions

As in Case 2, for the computation of the correlation coefficients the lowest binding free energy for each set of poses was taken into account. The scatter plots of all 14 data points of this test case are shown in Figure C.3. Figure 5.7 summarises

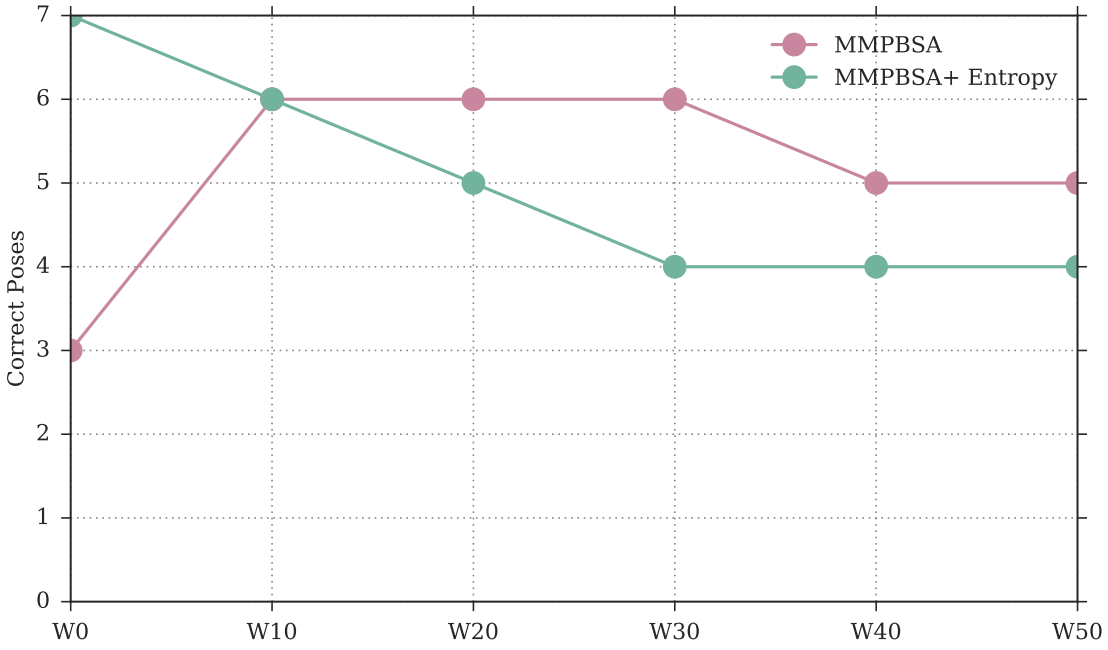


Figure 5.6: Number of correctly ranked the poses for the ligands docked in Case 3.

the performance of the calculations in terms of Pearson and Spearman correlation with experiment, and Figure 5.8 compares the MMPBSA performance to that of ABFE calculations.

Standard MMPBSA (W0) performs in this case very poorly (r_p = 0.05 [-0.06, 0.17] and r_s = -0.05 [-0.14, 0.19]), in accordance to its inability to identify how RVX-OH binds to the first BET bromodomains. The inclusion of a number of explicit water molecules in the calculations again improved the performance, in this case study quite substantially, driving the transition from no to modest correlation. In particular, W20 managed to obtain a Pearson correlation of 0.50 [0.37, 0.60] and a Spearman rank of 0.46 [0.25, 0.60]. Surprisingly, W30 returns low agreement with experimental affinities (r_p = 0.10 [-0.05, 0.25] and r_s = 0.11 [-0.06, 0.36]), despite it correctly identified the same number of RVX-OH binding modes as W20. For W40 and W50 however the correlation drops again to values around zero.

In this test case the inclusion of the entropic term improved the correlation with experiment for all MMPBSA calculations, with and without explicit water molecules. In particular, as one might expect by the fact this was the only set of

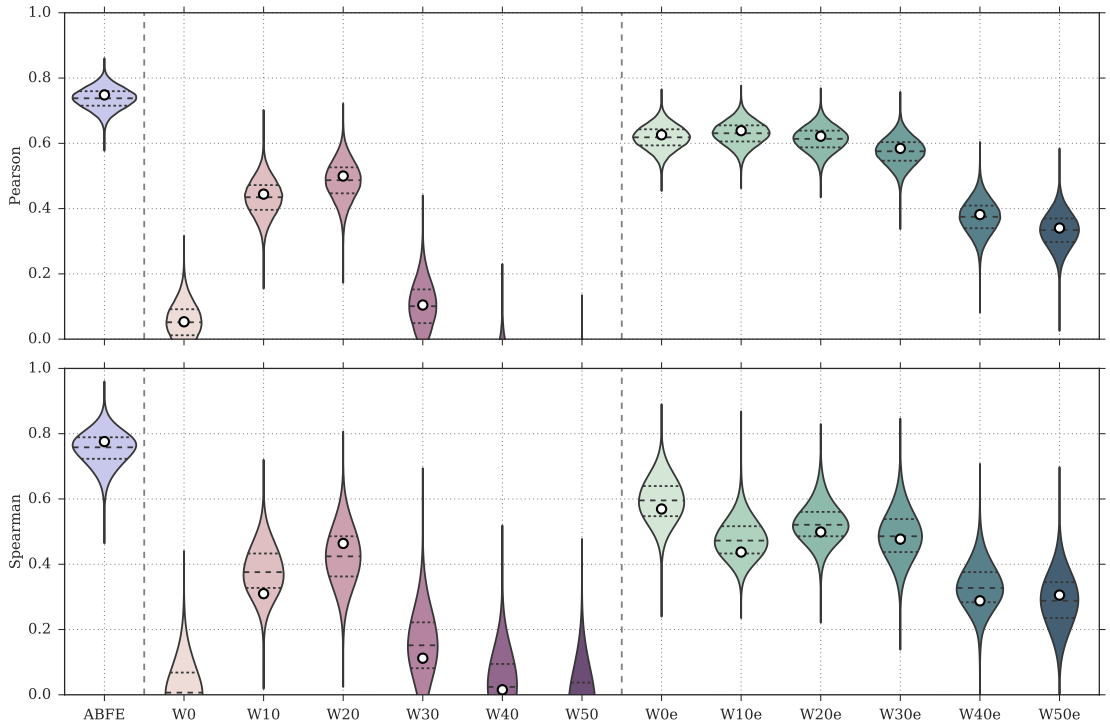


Figure 5.7: Pearson and Spearman correlations for ABFE and MMPBSA for Case 3. The plot shows Gaussian kernel density estimates for the distributions of r_p and r_s as obtained by 10^5 bootstrap samples. A white circle shows the value obtained with the original sample.

calculation that managed to identify all RVX-OH binding modes correctly, W0e provided the best correlation to experiment among all the MMPBSA calculations performed: $r_p = 0.63$ [0.54, 0.69] and $r_s = 0.57$ [0.47, 0.72]. While the number of correctly identified binding modes for RVX-OH decreased for W10e, W20e, and W30e, their Pearson correlation stayed approximately the same; on the other hand, their Spearman rank dropped reasonably. W40e and W50e, on the other hand, showed a significant decrease in both Pearson and Spearman correlation.

From Figure 5.8 it is possible to note how in this case all MMPBSA calculations provided worse agreement with experiment than ABFE. Without the entropy estimate, the best MMPBSA scenario, W20, still returned an r_p than was approximately 0.20-0.30 lower than what obtained with ABFE, and an r_p approximately 0.30 lower. With the inclusion of the entropy estimate, the best performer (W0e) returned an r_p $\sim$0.10 lower than what obtained with ABFE, and an r_s $\sim$0.20 lower.

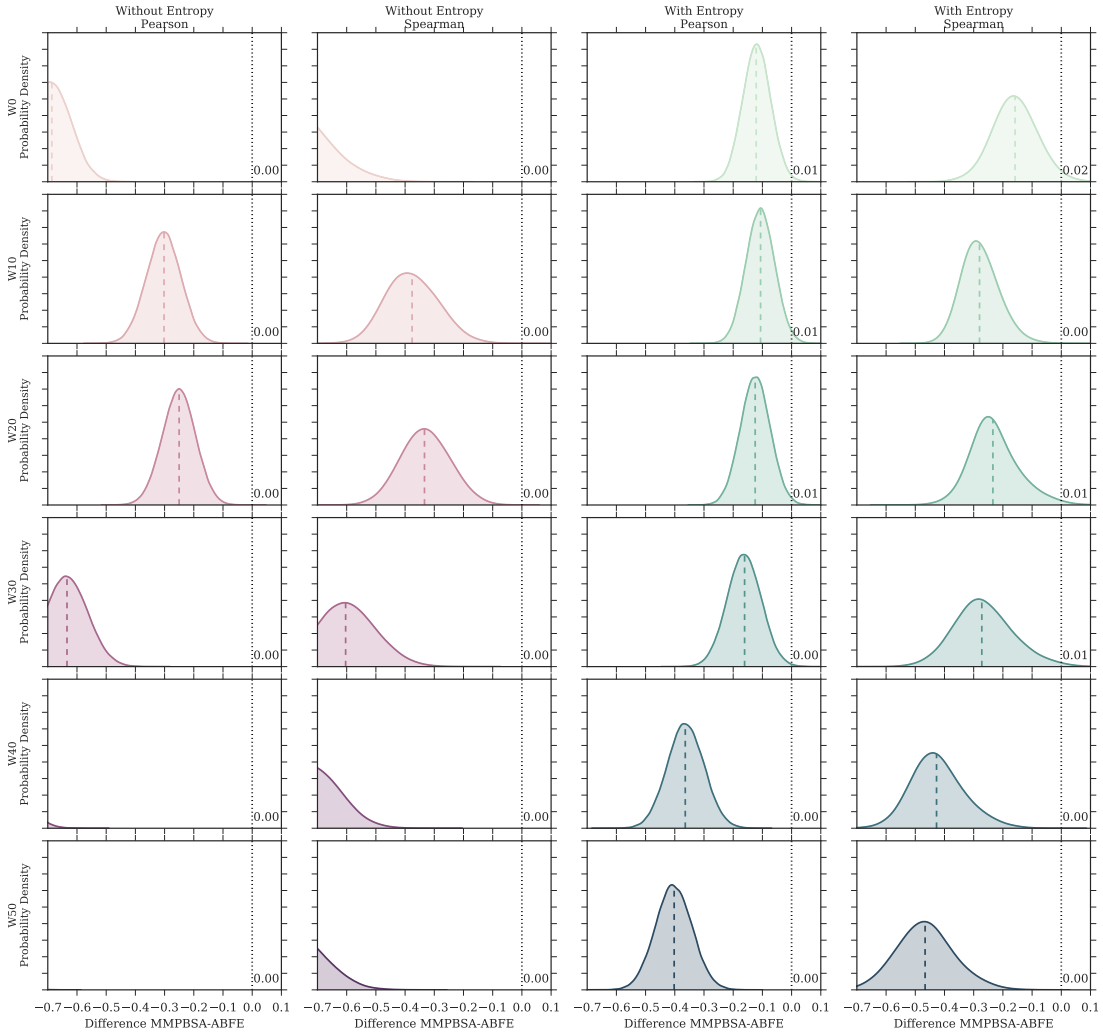


Figure 5.8: Correlation difference between MMPBSA and ABFE calculations for Case 3. Shown is the Gaussian kernel density estimate for the distribution of the difference of r_p and r_s as obtained by 10^5 bootstrap samples. The median of each distribution is shown as a dashed line. The value of the area under the curve above zero is shown on each chart.

Overall, in this case scenario, ABFE calculations provided a better performance than MMPBSA ones, which in many cases struggled to identify the binding mode of RVX-OH, and often returned correlations with experiment close to zero. Compared to Case 2, here it was possible to see a clearer trend where the MMPBSA calculations that correctly ranked a higher number of docked poses also returned better correlation performance, as one would intuitively expect. Furthermore, while in the previous cases the addition of the entropy estimate returned mixed

performances, here it improved both r_p and r_s values for both standard MMPBSA (W0), and MMPBSA with the addition of any number of explicit water molecules considered (W10 to W50).

5.6 Case 4: predicting the selectivity profile of bromosporine

In this test case, MMPBSA calculations are tested for the ability to reproduce the selectivity profile of bromosporine, a broad-spectrum BRD inhibitor. This test case proved challenging for ABFE, which achieved the modest correlations r_p = 0.48 [0.41, 0.53] and r_s = 0.50 [0.41, 0.62], and it was expected to be so also for MMPBSA. Here, only the bromosporine calculations with the first set of parameters used (RESP charges and no optimisation of soft torsions; § 4.5.2) are considered. As compared to test Cases 1 and 2, here it is the ligand that remains constant across calculations, while the protein binding partner changes. The scatter plots of all calculations performed on this set of 22 bromosporine-bromodomain pairs are shown in Figure C.4. Figure 5.9 summarises the performance of the calculations in terms of correlation with experiment, and Figure 5.10 compares the MMPBSA performance to that of ABFE calculations.

In this test case, standard MMPBSA returns lower correlations to experiment as compared to ABFE, with r_p = 0.42 [0.38, 0.46] and r_s = 0.41 [0.36, 0.52]. Furthermore, it is possible to notice from Figure 5.9 that, in contrast to what was observed previously, the inclusion of an explicit hydration shell in the MMPBSA calculations did not lead to an improved correlation with experiment. In particular, the Pearson correlation decreases when adding these waters, both when considering the results with and without the entropic term. The Spearman correlation improves only slightly with explicit waters included when the entropic term is not calculated, and it decreases when entropy is considered.

In this test case, the addition of the entropy term is slightly beneficial only when done for the calculations without any explicit waters included, with r_p and

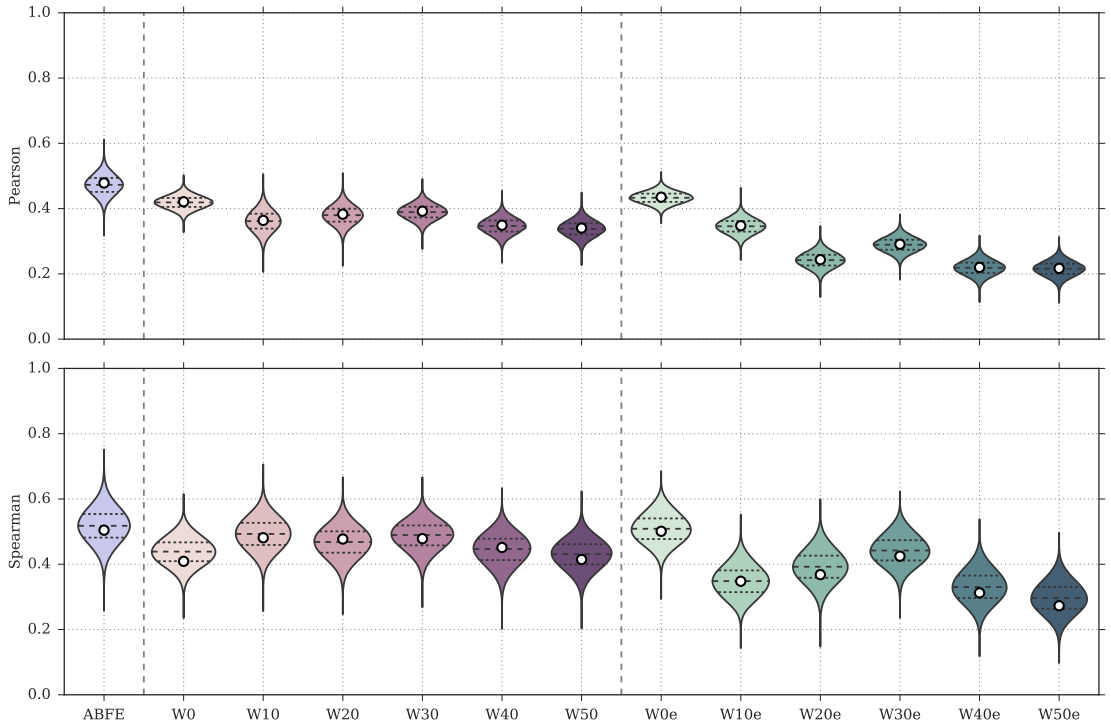


Figure 5.9: Pearson and Spearman correlations for ABFE and MMPBSA for Case 4. The plot shows Gaussian kernel density estimates for the distributions of r_p and r_s as obtained by 10^5 bootstrap samples. A white circle shows the value obtained with the original sample.

r_s increasing to 0.43 [0.40, 0.47] and 0.50 [0.42, 0.60], respectively. For all other cases it negatively affects the performance of the calculations.

From Figure 5.10, it is possible to see that, overall, ABFE calculations proved superior to all end-point protocols here tested, with distributions of r_p shifted toward values between 0.05 and 0.25 higher than MMPBSA, and r_s always returning a better rank of affinities too.

5.7 Discussion

MMPBSA calculations are an alternative to the more computationally demanding absolute alchemical binding free energy calculations, and as such it is useful to compare their performance in a relative fashion with respect to the more rigorous approach. A number of options are available to the user when setting these end-point calculations, with the performance often being system-dependent. Here, in the first

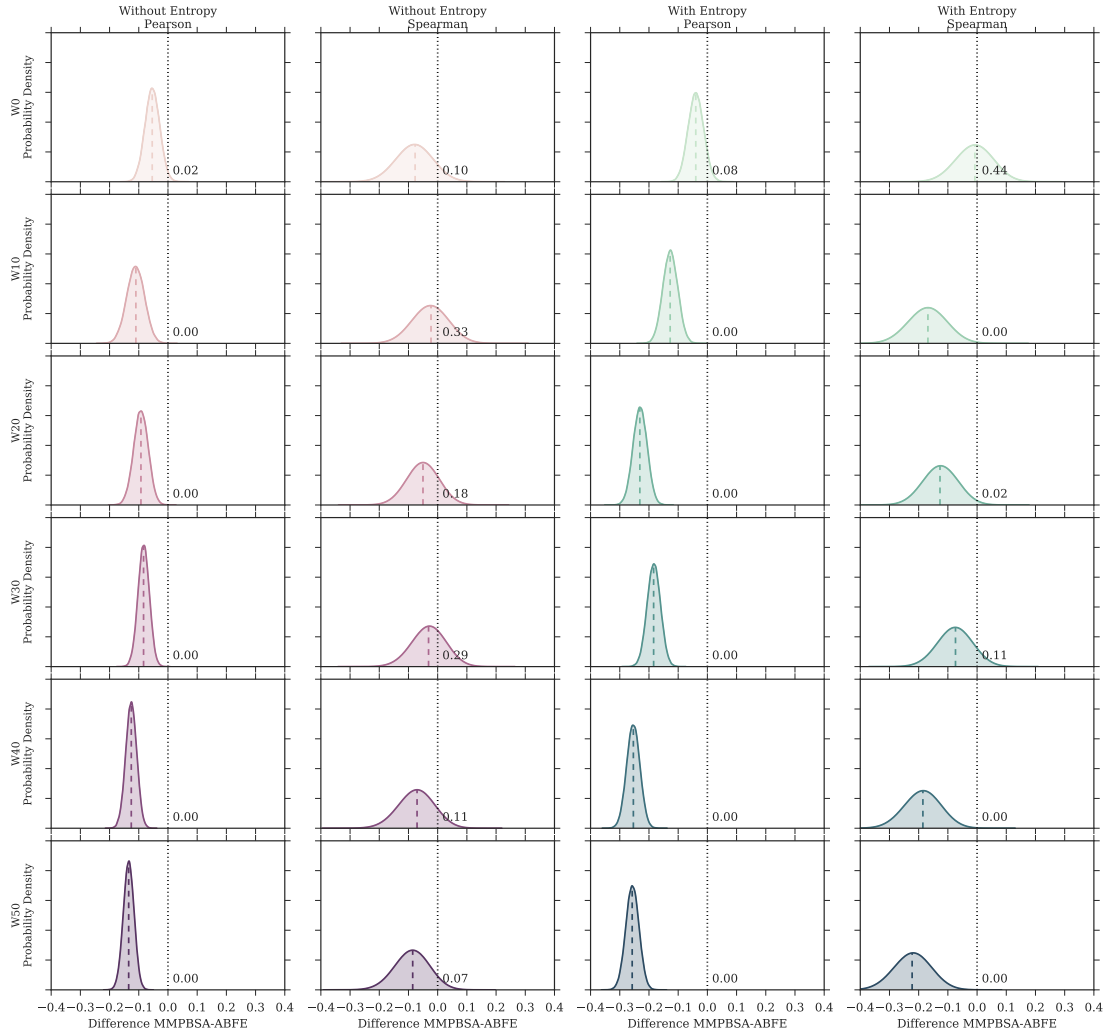


Figure 5.10: Correlation difference between MMPBSA and ABFE calculations for Case 4. Shown is the Gaussian kernel density estimate for the distribution of the difference of r_p and r_s as obtained by 10^5 bootstrap samples. The median of each distribution is shown as a dashed line. The value of the area under the curve above zero is shown on each chart.

instance I adopted one of the most widely employed MMPBSA protocols, that is, the single-trajectory approach with neglect of the entropic term. However, I also evaluated the effect of the inclusion of an explicit ligand hydration shell of different size, and the effect of including an estimate of the entropic term. Thus, twelve MMPBSA sets of calculations have been tested on frames extracted from the ABFE calculations, so that the same conformations have been used for both approaches, as well as the same physical model, and potential sampling issues would affect both

equally. The affinity predictions were evaluated based on four case scenarios, which included 47 experimental affinity data points, where either the protein target of the ligand of interest were kept constant. These are situations within the domain of applicability of MMPBSA, and ABFE calculations. On the other hand, due to large differences within the set of either ligands or proteins considered, relative binding free energy calculations (RBFE) would be impractical in these cases.

Table 5.1 reports the Pearson and Spearman correlations for ABFE and all MMPBSA setups here considered; the performances for all four case studies as well as their weighted averages are reported. Overall, across all tests, ABFE calculations appear to be more robust in obtaining good correlation with experimental affinities (weighted averages for r_p and r_s of 0.64 [0.56, 0.69] and 0.66 [0.53, 0.75] respectively). Nonetheless, some MMPBSA protocols achieved overall correlations that were only about 15% worse than ABFE, despite needing about 5% of the compute time. The best performing MMPBSA protocols overall were W20 and W0e. W20 achieved weighted averages for $r_p = 0.53$ [0.45, 0.59] and $r_s = 0.55$ [0.42, 0.64], and W0e obtained $r_p = 0.55$ [0.46, 0.61] and $r_s = 0.56$ [0.44, 0.68]. Interestingly, W20 performed particularly well in Cases 1 and 2, while W0e was superior for Cases 3 and 4.

It is important to mention that one of the advantages of ABFE calculations as compared to MMPBSA is the ability to agree with experimental binding free energy also in absolute terms. In fact, the weighted average RMSE for the four cases was of 1.5 kcal/mol for ABFE, while the best weighted average RMSE for MMPBSA was of 5.3 kcal/mol, obtained with the W0e protocol. Without entropy estimate, the standard W0 protocol returned an overall RMSE of 15.7 kcal/mol. It also has to be kept in mind however that while it might be possible to identify retrospectively a specific MMPBSA approach that performs equally or better than ABFE in a specific test case, it is not possible to do so prospectively, as a specific approach needs to be chosen a priori. Even if multiple protocols were to be run in a prospective campaign, it would not be possible to identify the one most likely to return the best predictions, unless one has previous knowledge of the system. This

Table 5.1: Pearson and Spearman correlations for all MMPBSA calculations. "wAvg" is the weighted average; the predictions in Case 1 and Case 2 are assigned a weight of 0.5 (for an overall weight of 5.5 each) because they evaluate the same binding affinities twice, while both Case 3 and Case 4 are assigned a weight of 1 (for overall weights of 14 and 22, respectively) because they represent correlations to separate sets of affinities.

Pearson Correlation													
Case	ABFE	W0	W10	W20	W30	W40	W50	W0e	W10e	W20e	W30e	W40e	W50e
1	0.87	0.71	0.77	0.82	0.83	0.83	0.83	0.73	0.31	0.62	0.80	0.87	0.90
2	0.78	0.79	0.89	0.90	0.88	0.87	0.86	0.67	0.51	0.85	0.88	0.92	0.91
3	0.75	0.05	0.44	0.50	0.10	−0.14	−0.23	0.63	0.64	0.62	0.58	0.38	0.34
4	0.48	0.42	0.36	0.38	0.39	0.35	0.34	0.43	0.35	0.24	0.29	0.22	0.22
wAvg	0.64	0.39	0.50	0.53	0.42	0.32	0.29	0.55	0.45	0.47	0.51	0.43	0.41
Spearman Correlation													
1	0.85	0.72	0.57	0.77	0.83	0.83	0.79	0.61	0.09	0.54	0.61	0.88	0.90
2	0.78	0.79	0.79	0.85	0.85	0.83	0.83	0.75	0.35	0.71	0.83	0.85	0.88
3	0.78	−0.05	0.31	0.46	0.11	0.02	−0.05	0.57	0.44	0.50	0.48	0.29	0.31
4	0.50	0.41	0.48	0.48	0.48	0.45	0.41	0.50	0.35	0.37	0.42	0.31	0.27
wAvg	0.66	0.35	0.48	0.55	0.45	0.41	0.37	0.56	0.34	0.47	0.51	0.43	0.43

95% Confidence Intervals

Pearson Correlation													
Case	ABFE	W0	W10	W20	W30	W40	W50	W0e	W10e	W20e	W30e	W40e	W50e
1	[0.73, 0.92]	[0.61, 0.76]	[0.62, 0.86]	[0.75, 0.86]	[0.79, 0.87]	[0.79, 0.86]	[0.79, 0.86]	[0.31, 0.85]	[-0.08, 0.66]	[0.39, 0.75]	[0.60, 0.89]	[0.61, 0.93]	[0.74, 0.94]
2	[0.68, 0.83]	[0.75, 0.82]	[0.85, 0.91]	[0.86, 0.92]	[0.85, 0.91]	[0.83, 0.89]	[0.83, 0.89]	[0.63, 0.71]	[0.45, 0.56]	[0.81, 0.88]	[0.85, 0.91]	[0.89, 0.94]	[0.88, 0.93]
3	[0.67, 0.80]	[-0.06, 0.17]	[0.32, 0.54]	[0.37, 0.60]	[-0.05, 0.25]	[-0.29, 0.01]	[-0.37, -0.08]	[0.54, 0.69]	[0.56, 0.70]	[0.54, 0.69]	[0.49, 0.66]	[0.27, 0.47]	[0.23, 0.44]
4	[0.41, 0.53]	[0.38, 0.46]	[0.30, 0.43]	[0.32, 0.44]	[0.34, 0.44]	[0.30, 0.40]	[0.29, 0.39]	[0.40, 0.47]	[0.30, 0.39]	[0.19, 0.29]	[0.24, 0.33]	[0.17, 0.26]	[0.17, 0.26]
wAvg	[0.56, 0.69]	[0.32, 0.45]	[0.41, 0.57]	[0.45, 0.59]	[0.34, 0.49]	[0.24, 0.39]	[0.21, 0.36]	[0.46, 0.61]	[0.35, 0.53]	[0.39, 0.53]	[0.43, 0.56]	[0.34, 0.48]	[0.34, 0.47]
Spearman Correlation													
1	[0.70, 0.94]	[0.62, 0.84]	[0.50, 0.82]	[0.67, 0.85]	[0.74, 0.87]	[0.75, 0.85]	[0.74, 0.85]	[0.18, 0.82]	[-0.17, 0.48]	[0.19, 0.61]	[0.35, 0.89]	[0.46, 0.94]	[0.58, 0.95]
2	[0.55, 0.85]	[0.72, 0.85]	[0.75, 0.85]	[0.78, 0.89]	[0.78, 0.87]	[0.75, 0.85]	[0.75, 0.85]	[0.67, 0.80]	[0.25, 0.45]	[0.65, 0.82]	[0.72, 0.85]	[0.78, 0.93]	[0.79, 0.89]
3	[0.64, 0.85]	[-0.15, 0.19]	[0.24, 0.53]	[0.25, 0.60]	[-0.06, 0.36]	[-0.20, 0.23]	[-0.28, 0.18]	[0.47, 0.72]	[0.37, 0.60]	[0.41, 0.65]	[0.35, 0.64]	[0.20, 0.47]	[0.13, 0.45]
4	[0.41, 0.62]	[0.36, 0.52]	[0.40, 0.58]	[0.37, 0.56]	[0.40, 0.57]	[0.35, 0.53]	[0.34, 0.51]	[0.42, 0.60]	[0.26, 0.44]	[0.30, 0.49]	[0.36, 0.53]	[0.24, 0.43]	[0.20, 0.39]
wAvg	[0.53, 0.75]	[0.28, 0.50]	[0.40, 0.63]	[0.42, 0.64]	[0.34, 0.58]	[0.28, 0.52]	[0.25, 0.49]	[0.44, 0.68]	[0.24, 0.50]	[0.36, 0.59]	[0.39, 0.64]	[0.31, 0.56]	[0.29, 0.53]

system-dependence can be a nuisance as one is not certain of the MMPBSA protocol to adopt. This uncertainty is not present to the same extent in ABFE calculations.

The present study supports the hypothesis of improved correlation with experiment when an explicit ligand hydration shell is included in the MMPBSA calculations [334]. However, for the bromodomain test systems we observed that correlation improves after including a small number of explicit waters (up to 20) and then level outs or decreases at higher numbers (30 to 50). In fact, while W10 to W30 overall provided better performance as compared to W0, W40 and W50 returned correlations comparable or inferior to W0.

The addition of the entropic term to W0 resulted in an improved overall performance, comparable to what obtained with the W20 protocol. This despite the fact that as shown in Case 1 the estimate appears to be sensitive to the simulated ensemble, and the results from repeated calculations showed large sample standard deviations when entropy was included. In addition, as expected, the inclusion of entropy largely improved the agreement with experiment in absolute terms. The better correlation of W0e over W0 is however entirely due to Case 3, where MMPBSA without the entropy estimate fails. If only Case 1, 2 and 4 are considered, the performances of W0 and W0e are close in terms of correlations, while W0e still clearly improves the RMSE of the predictions.

The addition of the entropy estimate to the MMPBSA results that also included explicit waters resulted in a sharp drop in performance for low numbers of water molecules. This might be caused by the fact that these waters, since they were considered part of the receptor, introduced noise in the protein-ligand interaction energies used to compute the entropy as described in the Methods. In fact, since for each frame the N waters closest to the ligand were included, when the number of explicit waters retained is low the structure of this hydration shell might change considerably frame by frame. On the other hand, with a larger number of explicit waters the structure of the hydration shell is more consistent across all frames, with the variability coming from waters further away from the ligand atoms and thus contributing less to the energy fluctuations. Indeed, on average, the magnitude of the

entropic term also tended to decrease between W10e and W50e, in agreement with a decrease of the magnitude of energy fluctuations around their average value. Overall, however, for the results with the entropy estimate, including the ligand hydration shell (W10e to W50e) did not provide better performance as compared to MMPBSA without this layer of explicit waters (W0e), possibly for the effect described above.

From the present study it is difficult to identify the better approach for the rescoring of docking poses. In Case 2, the highest numbers of correctly ranked poses were obtained with W30e, W40e, and W50e, while in Case 3 with W0e. In both case scenarios, the inclusion of the entropic estimate to standard MMPBSA (W0) resulted in an improved performance. However, given the small set of docking poses here considered, this evidence is anecdotal. For the MMPBSA calculations that took into account the ligand hydration shell, the addition of entropy resulted in worse recovery rates of the best ligand poses in Case 3, but better recovery rates in Case 2.

5.8 Conclusion

In summary, based upon the work described in this chapter, the following observations can be made:

1. Overall, ABFE calculations appear to be more robust than MMPBSA ones, both in providing good correlations with experimental affinities and in the ranking of docking poses, across different case scenarios.

2. Nonetheless, some MMPBSA protocol (namely as W20 and W0e) still achieved good correlation with experimental affinities at a fraction of the computational cost of ABFE.

3. For bromodomains, the inclusion of a small explicit ligand hydration shell seems to result in improved correlation with experiment (when the entropic term is neglected).

4. The inclusion of the entropic term appears to be beneficial for the correlation to experiment in specific cases. In particular, the entropy estimate recovered the lost correlation in Case 4, while it had limited impact in all other cases.

5. As expected, MMPBSA does not generally provide quantitative agreement with experimental binding free energies in absolute term, contrary to ABFE. The incorporation of the entropic term largely improved the average absolute errors, yet the RMSE achieved by MMPBSA was still almost four times larger than the one achieved by ABFE.

Thousands have lived without love, not one without water.

— Wystan H. Auden, in *First Things First*, 1956

6

Water stability in bromodomain binding pockets

Contents

6.1 Introduction 156
6.2 Methods 158
6.2.1 Systems preparation 158
6.2.2 Water hydration free energy 159
6.2.3 Grand Canonical simulations 160
6.3 Water placement 163
6.4 Stability of the water network 166
6.5 Stability of individual waters 170
6.6 Discussion 176
6.7 Conclusion 178

As previously mentioned, bromodomains exhibit a conserved network of water molecules in their binding pockets. Recently, this network has been shown to be displaceable for at least some BRDs [80]. In this chapter I will present a study in which Monte Carlo simulations in the grand canonical (GC) ensemble have been used to estimate the stability of the water network in the cavities of apo bromodomains.

I warn the reader that the results presented here have been obtained with a development version of the simulation code, which contained a bug. Potential future publications arising from this work will therefore contain results obtained with more

recent bug-free versions of the same code, which will be quantitatively different from the ones shown here. Nonetheless, it is expected that the trends and the qualitative observations of this chapter will remain unaltered.

6.1 Introduction

The role of water in protein binding sites has long been recognised in the field of drug design [21, 338, 339]. Upon binding, a ligand may displace water molecules from a binding site. Typically, these waters are those that rapidly exchange with the bulk solvent, and are highly disordered. These are easily displaceable water molecules that would rarely be identified by X-ray crystallography [21]. On the other hand, some water molecules are tightly bound to the macromolecule and are observed in multiple experimentally determined structures of the same protein. These waters might contribute to the stability of certain region of the protein, or might be needed to bridge the binding of natural substrates. From a medicinal chemistry perspective, it is thought that the displacement of stable water molecules is entropically favourable, and thus often sought after, thanks to the increased rotational and translational freedom of water in bulk as compared to the bound state. However, the entropy gain comes with an enthalpy penalty due to the loss of water-protein interactions. As displacement of structural waters is thought to lead to tighter binding and, in some cases, could drive selectivity, it is a commonly pursued strategy.

Water molecules in bromodomain binding pockets have been observed crystallographically in most X-ray structures and, as such, they have been targeted for displacement, yet with relatively little success. As described in § 1.2.1, BRD binding pockets display a conserved water network formed by 4 water molecules at the bottom of the cavity, bridging the binding of acetylated lysine. These four water sites will here be referred to as W1 to W4, as shown in Figure 6.1. An additional highly buried water molecule is often observed experimentally, however, despite being part of the network, this water lies outside of the plane of the other 4 waters and in a particularly buried position, which makes its displacement extremely challenging, and not possible without the displacement of the other waters first (Figure 1.4).

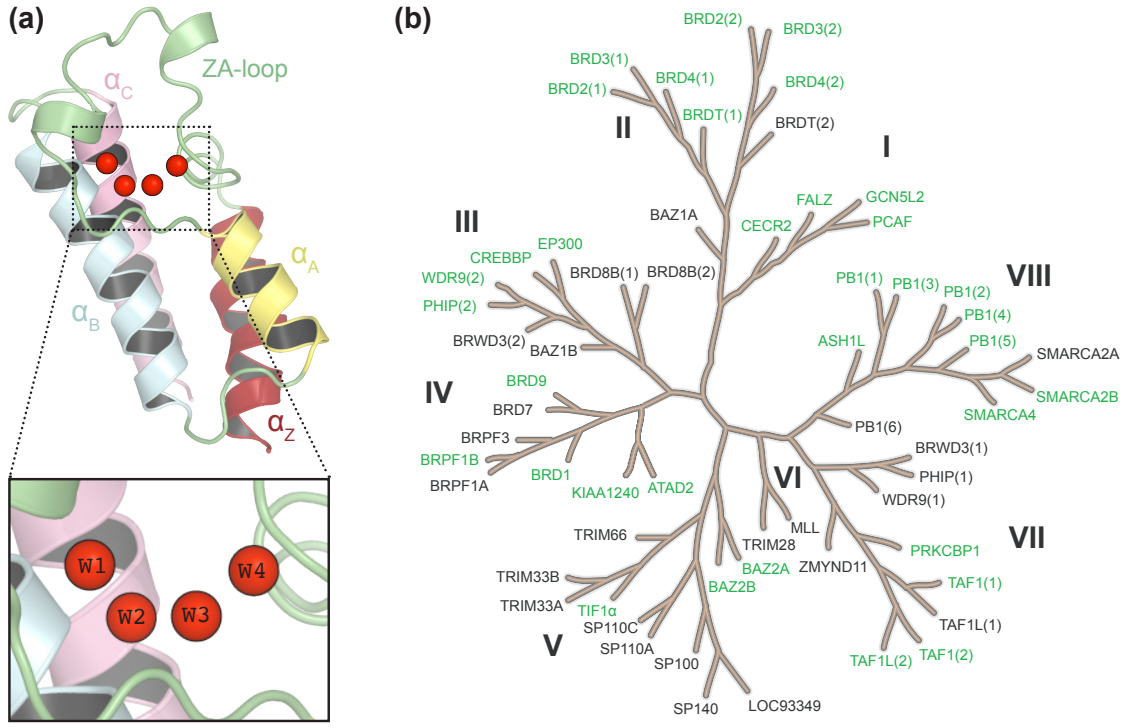


Figure 6.1: Water network structure and proteins considered in the study. (a) Bromodomain fold with conserved water molecules highlighted as red spheres. See also Figure 1.4 for reference. The four conserved water sites are referred to as W1 to W4, and their location within the network topology is shown in the zoomed panel. (b) Phylogenetic tree of the human bromodomain family, where in green are the BRDs considered in the study.

While conserved waters in BRDs had remained undisplaced for many years, despite the medicinal chemistry efforts [340], a number of studies reporting the complete or partial displacement of the 4-water network have recently emerged. Most notably, Fedorov *et al.* [80] reported the displacement of all four conserved waters by salicylic acid from the binding pockets of a number of family VIII bromodomains. The fragment led to the development of a nanomolar ligand, selective for the SMARCA and PB1 bromodomains. Gestenberger *et al.* [156] started from the same fragment and developed another potent chemical probe with a similar selectivity profile. At the same time, Sutherell *et al.* [157] reported the discovery of another series of ligands, also targeting family VIII and displacing all four waters from their pocket, based on a 2,3-dihydropyrrolo[1,2-a]quinazolin-5(1H)-one scaffold. It is here hypothesised that the weaker stability of the water network in SMARCA and PB1

bromodomains contributed to the selectivity of the developed inhibitors. These are, thus far, the only instances of complete displacement of the 4-water network reported. Two other studies have noticed the displacement of a single water molecule by the ligand in specific BRDs. Cox *et al.* [158] noticed the displacement of W1 from the binding pocket of PHIP(2) by a thiourea fragment, while Zhu and Caflish [146] reported the displacement of W1 from BRPF1 by an isoquinolinone fragment.

It has therefore become clear that some or all the structured water molecules found in BRD binding pockets might be displaceable from certain bromodomains. However, the rational identification of poorly stable waters and water networks is very challenging by simple inspection of the crystal structures. Even more arduous is the quantitative estimation of water stability, so that one cannot differentiate a stable water molecule carrying a large free energy penalty for its displacement from a relatively unstable one. In this work, I try to address this issue by using some of the latest computational methods available to study the free energy penalty associated with the displacement of water molecules from all bromodomains with structural information available.

Here, I present the results of a comprehensive study on bromodomain water stability where I estimated the binding free energy of the water sites and networks found in 35 bromodomains with X-ray structure available, using three different water models (TIP3P, TIP4P, and SPC) and free energy calculations based on Monte Carlo simulations. The binding free energy provides a quantitative estimate of water stability, and it is the negative of the free energy penalty associated with the displacement (i.e. dissociation) of water molecules.

6.2 Methods

6.2.1 Systems preparation

The three-dimensional coordinates of the 35 bromodomains studied here were taken from the PDB (all PDB-IDs are provided in Table D.1) [256]. Only structures determined by X-ray crystallography and without missing non-terminal residues were selected. If multiple structures for a bromodomain were available, a single

structure was chosen while taking into account its resolution and model fit. If multiple chains were present in the deposited structure, only chain A was retained. A way to improve the preparation of the models would be to select a chain based on some quality metric, such as the completeness of its structure or its temperature factors. All organic molecules and ions were removed, while water molecules were kept. Missing side-chain atoms were modelled using the WHAT-IF server [257, 258]. All structures were superimposed to the BRD4(1) structure based on the position of the C_α atoms using PyMOL v1.4. This allowed the definition of a single grand canonical region (i.e. GCMC box), which was then used for the free energy calculations of all bromodomains. The grand canonical region is that volume of the simulation box which is simulated in the grand canonical ensemble, and in which the number of water molecules is allowed to fluctuate. Proteins were protonated with the program *reduce* provided with the Amber14 package [341], which also optimises the orientations of Asn/Gln side chain amides, and His rings. Values of pK_a were determined with *propka v3.1* [342, 343] for Asp, Glu, Lys and Cys residues, and they were (de)protonated only in the presence of a large difference (>1.0) between predicted pK_a and the target pH of 7.0. Finally, a short minimisation (100 steps) *in vacuo* using a steepest descent algorithm was carried out with the *sander* program (Amber14) [344] in order to remove potential clashes.

6.2.2 Water hydration free energy

The hydration free energy of the TIP3P, TIP4P, and SPC water models were calculated using a double-decoupling protocol with the ProtoMS (v3.2, development branch) simulation package [216]. A dual topology was used for the water molecule to be decoupled, which was simulated in a solvated cubic box of approximately 22 x 22 x 22 Å. The system was simulated in the isothermal-isobaric ensemble, at 1 atm and 298 K, with periodic boundary conditions. A 10 Å cut off was applied to the non-bonded interactions. 5 million (M) equilibration moves were performed before starting to collect data from 500M Monte Carlo moves. Exchanges between neighbouring replicas were allowed every 0.2M steps. Data for analysis were saved

to file every 0.1M moves, while coordinates every 10M moves. The calculations were repeated five times for each water model. Hydration free energies were estimated using the Multiple Bennet Acceptance Ratio (MBAR) [240] using the *calc_dg.py* script among the ProtoMS v3.2 tools, and taking into account the data from all five calculations. Using this protocol, the hydration free energies for the TIP3P, TIP4P, and SPC water models were estimated to be respectively -6.18 ± 0.04 kcal/mol, -6.16 ± 0.04 kcal/mol, and -6.26 ± 0.02 kcal/mol, in good agreement with the experimental value of -6.32 kcal/mol [345].

6.2.3 Grand Canonical simulations

The simulation package ProtoMS (v3.2, development branch) was used for Monte Carlo simulations and data analysis [216]. Proteins were modelled with the Amber ff14SB force field [346], and water with the TIP3P, TIP4P, and SPC models [179, 180]. The grand canonical box was defined manually so to encompass the binding site of the 4-water network of interest. The same box was used for all proteins, as they had been previously aligned for this purpose. Protein residues that were further than between 16 and 20 Å away from the grand canonical region were removed, with the exact distance chosen so as to retain whole residues. The systems were solvated up to a radius of 30 Å around the grand canonical region. All simulations were carried out at 298 K, and non-bonded interaction were cut off at 10 Å. Before initiating the GCMC simulations, the systems were equilibrated using 50M Monte Carlo moves in the canonical ensemble, in order to allow the water to equilibrate around the protein; at this stage, MC moves were generated only for the solvent molecules. Water molecules present in the grand canonical region were then removed, and this set of coordinates represented the starting point of the GCMC simulations.

A set of 32 simulations were performed at a range of Adams values from -32 to -1 at unit increments (see § 2.4.1 and § 2.6.3 for the definition of the Adams parameter). For each window, 15M equilibration moves were performed only on the grand canonical solute, with insertion, deletion, and translation/rotation moves generated at the same ratio. An additional 5M equilibration moves followed where

the protein side chains and solvent molecules were also sampled, before starting the production simulation of 50M moves. Half of the MC moves were dedicated to the grand canonical water molecules, and the other half was split between protein residues and solvent in proportion to the number of solvent molecules and protein residues, according to the ratio of 1:5. Exchanges between neighbouring replicas were allowed every 0.5M steps. Data for analysis, such as energies, number of GC solutes present, and coordinates were saved to file every 0.1M moves. This procedure was repeated three times for each protein and water model studied, i.e. for each bromodomain I carried out 9 calculations, 3 for each water model here considered (TIP3P, TIP4P, and SPC).

The simulations in the grand canonical ensemble had been performed using a development version of the ProtoMS code, which was later found to contain a bug resulting in the computation of an incorrect probability for the insertion/deletion moves of the grand canonical water molecules. This bug has now been corrected, and an official release of the code is forthcoming. All these calculations will be re-run in the near future with the latest version of the software. However, preliminary calculations indicate that while the results will be quantitatively different, they will also be qualitatively similar to what presented in this chapter; that is, despite the bug, the trends observed here should be conserved also when re-running the calculations with the new bug-free code.

Data analysis

Grand Canonical Integration (GCI) was performed using the *calc_gci.py* script that is part of the ProtoMS (v3.2, development branch) tools [68, 216], analysing the data from the three repeats together. The amount of data to discard as equilibration was determined using the script *calc_series.py*. The same amount of data was discarded from all windows after determining the average number of moves needed to maximise the number of uncorrelated samples.

The sub-volume of the grand canonical region that represents the 4-water network I was interested in studying was defined as follows. First, water location

in the binding site was determined by clustering the grand canonical waters from the simulations at the Adams value corresponding to the state with lowest free energy. The chemical potential representing the minimum binding free energy state can be identified by plotting the average number of waters inserted in the grand canonical region versus the free energy, as described by Ross *et al.* [68]. An average linkage hierarchical clustering procedure with a cut off of 2 Å was performed with the *calc_clusters.py* script to identify the location with highest water occupancies. From these cluster centres, the 4 water sites forming the typical network found in bromodomain pockets were identified by visual inspection. The identification of this network could be automated by using machine learning techniques, trained to recognise the geometrical patterns that define the topology of the network. These four sites were used to define the centres of four spheres with radius 1.4 Å. The average number of grand canonical waters found within the volume of these 4 spheres was obtained by post-processing the frames stored. With this information it was possible to perform the grand canonical integration procedure taking into account only the contribution from this sub-volume. In this case, we know *a priori* the equilibrium number of waters in the sites, so that GCI was performed for up to 4 waters, and the free energy of transferring 4 waters from bulk to gas phase was considered. Note that from a statistical mechanics viewpoint this does not rigorously represent the water free energy of binding to these sites, as this would require running GCMC calculations where the grand canonical waters are allowed to be inserted/removed only within the volume of the four spheres. Nonetheless, it represents a close approximation, corresponding to the free energy of binding four water molecules to the four sites in the presence of more stable grand canonical waters (e.g. the fifth highly structural water molecule in BRD pockets), and in the absence of less stable ones (e.g. solvent exposed water molecules that exchange frequently with bulk).

The free energy of binding of a water molecule to each of the four sites was similarly derived by post-processing the trajectories, and then using the logistic

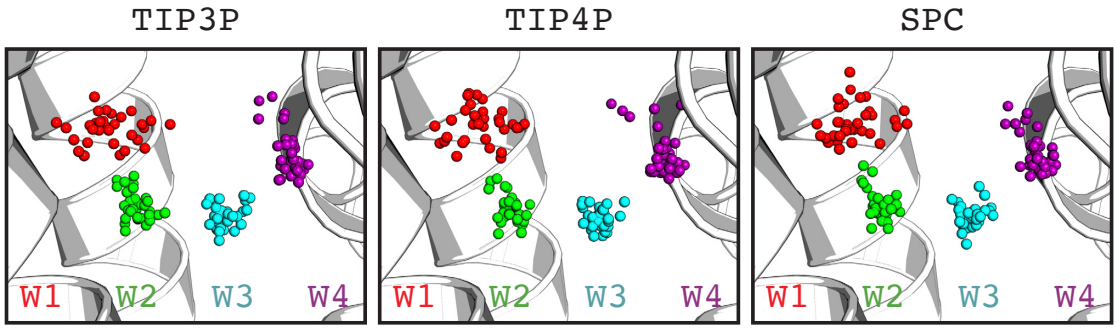


Figure 6.2: Water networks as predicted by the simulations for all the bromodomains considered. The cartoon of BRD4(1) is shown only as a reference.

function proposed by Ross *et al.* [68] for the transfer of a single molecule from ideal gas to grand canonical region.

6.3 Water placement

As discussed in § 2.4.1 and § 2.6.3, GCMC calculations allow the estimation of both the position of water sites as well as their binding free energy. Bromodomains crystal structures are often obtained at high resolution, so that the location of conserved water molecules is often known. Nonetheless, there are instances of BRDs without this network of waters being resolved by crystallography, in particular for atypical BRDs. More importantly, the ability of the simulations to reproduce the location of the water network as found in the experimental structures can be considered as a validation for the methodology. In this study, we will also use the agreement of the water placement between the simulations and the X-ray structure as a proxy for the reliability of the binding free energies returned via grand canonical integration (GCI). In fact, when calculating the mean from the results of the three water models, the free energy for each water site and water network will be weighted according to the quality of the water placement.

Overall, the Monte Carlo simulations managed to recover the location and topology of the water network well in most proteins. Simulations using the TIP3P model identified the presence of a 4-water network at the bottom of the cavity for 33 out of 35 bromodomains: in both CECR2 and FALZ, water site W1 was

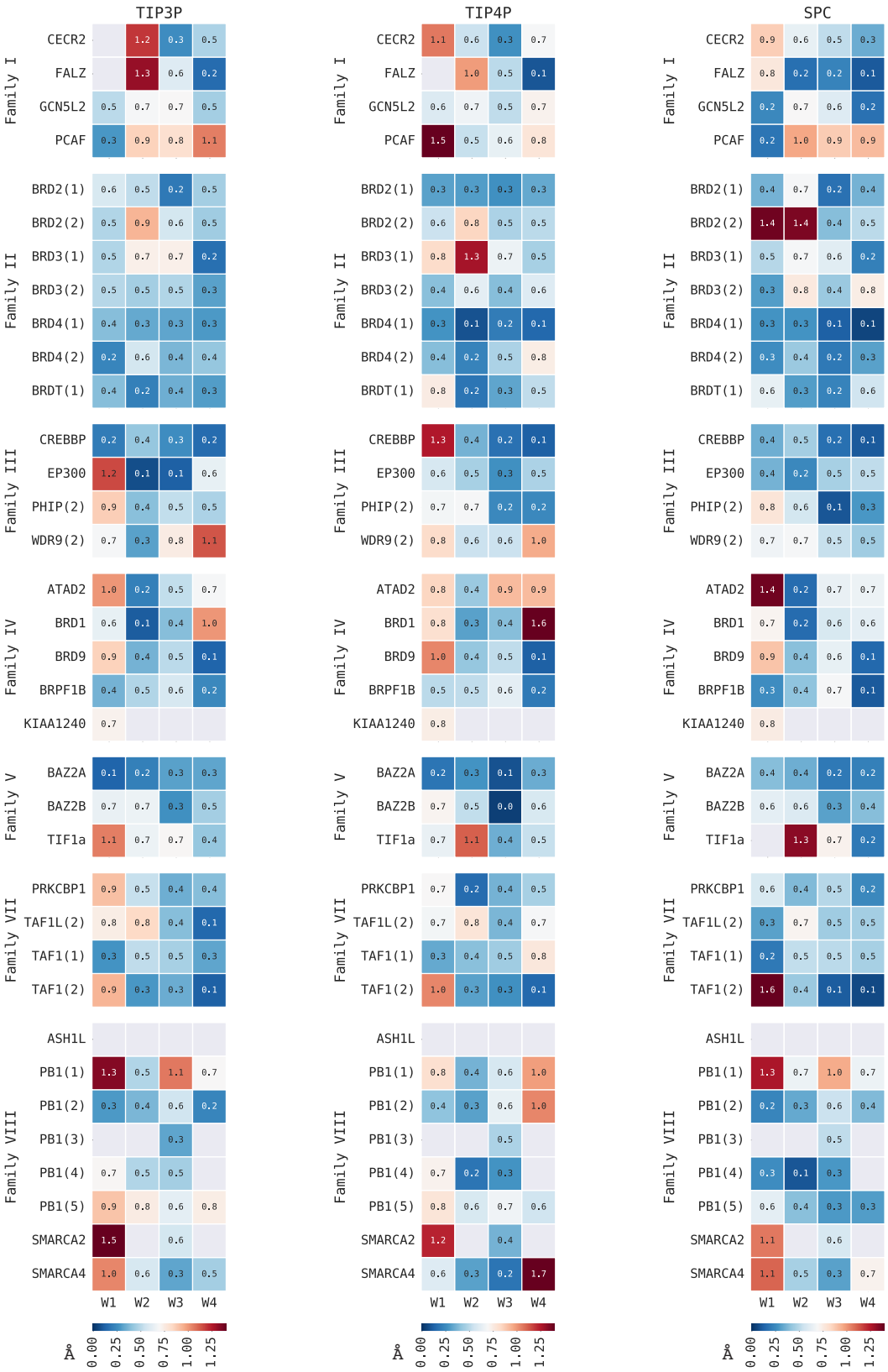


Figure 6.3: Water placement for the three water models. Shown are the errors in Å as compared to the location of the waters in the X-ray structures; empty cells are due to either unresolved waters in the X-ray structures or failure of the calculations to identify a conserved water site.

not suggested to be present. With the TIP4P model the network was identified 34/35 times, as the site W1 was not identified for FALZ. With the SPC model too the network was identified 34/35 times, however in this case it was the W1 site in TIF1α to be missing. This translates into a very high recovery rate of >94% for all three water models. For each of the four sites, in each BRD, the error of the water placement was measured as the distance between the experimental and predicted water locations. All these errors are shown in Figure 6.3. For the whole network, the RMSD from the X-ray structure can be calculated as:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2} = \sqrt{\frac{\delta_{W1}^2 + \delta_{W2}^2 + \delta_{W3}^2 + \delta_{W4}^2}{4}} \tag{6.1}$$

where δ_i is the distance between the predicted position of water i and its position in the X-ray structure, and with $N = 4$ as the network is formed by 4 water molecules (with the few exceptions mentioned above for CECR2, FALZ, and TIF1α where $N = 3$). Overall, all three water models performed well: TIP3P gave an average RMSD across all 35 proteins of 0.59 Å, TIP4P of 0.62 Å, and SPC of 0.56 Å. As possible to see in Figure D.1, the distributions of RMSD values are similar for all three models, with the majority of waters being placed within 1 Å of the experimental location. Also when considering the four sites independently, the different water models perform similarly in terms of placement (Figure D.1). From Figure 6.3 it is however possible to note that, while all three water models exhibit similar overall trends in terms of quality of the placement, in a few cases a certain model performs slightly better or worse than the others. For instance, the position of W4 in PB1(2) agrees well with experiment for the TIP3P and SPC models, while it does not for the TIP4P model. For W4 and BRD1, the SPC model appears to be able to better locate the position of the water site than TIP3P and TIP4P. On the other hand, the location of W1 in BRD2(2) is predicted well by TIP3P and TIP4P, but not by SPC.

6.4 Stability of the water network

Starting from the location of the waters forming the network in each protein, I calculated the stability of this network (ΔG_{netw}) as described in § 2.6.3 and § 6.2.3 using GCI. Figure 6.4 provides a summary of the result obtained. As previously mentioned, the binding free energies shown are the weighted mean and standard deviation from the results obtained with the TIP3P, TIP4P, and SPC water models, where the RMSD values were used for weighting. More specifically, the results for each water model were weighted using the inverse of the RMSD; essentially considering as more reliable the free energies that were obtained from the model(s) with a water placement in better agreement with the X-ray structures:

$$w^i_{netw} = \frac{1}{RMSD^i} \tag{6.2}$$

$$\overline{\Delta G}_{netw} = \frac{\sum\limits_{i \in M} \Delta G^i_{netw} \cdot w^i_{netw}}{\sum\limits_{i \in M} w^i_{netw}} \tag{6.3}$$

$$\sigma_{netw} = \sqrt{\frac{\sum\limits_{i \in M} w^i_{netw} \cdot (\Delta G^i_{netw} - \overline{\Delta G}_{netw})^2}{\sum\limits_{i \in M} w^i_{netw}}} \tag{6.4}$$

where w^i_{netw} is the weight of water model i, which is part of the set of water models $M = \{TIP3P, TIP4P, SPC\}$, and σ is the standard deviation. From here on, I will refer to the average $\overline{\Delta G}_{netw}$ simply as ΔG_{netw}. I am aware of the fact that the position of waters in the crystallographic structures themselves might be unreliable. However, this approach takes this into account indirectly to a certain extent. In fact, if all water models disagree with the experimental location of a water to a similar degree, their free energies will have the same weights and will be treated equally, just as if they were all in agreement with the X-ray structure.

Figure 6.4a ranks the bromodomains considered in this study by the stability of their water network. It is possible to note how, as expected by the experimental evidence available, most of BRDs from family VIII rank low in this chart. Members of family II, family IV, and family VII tend to be ranked quite high instead. However,

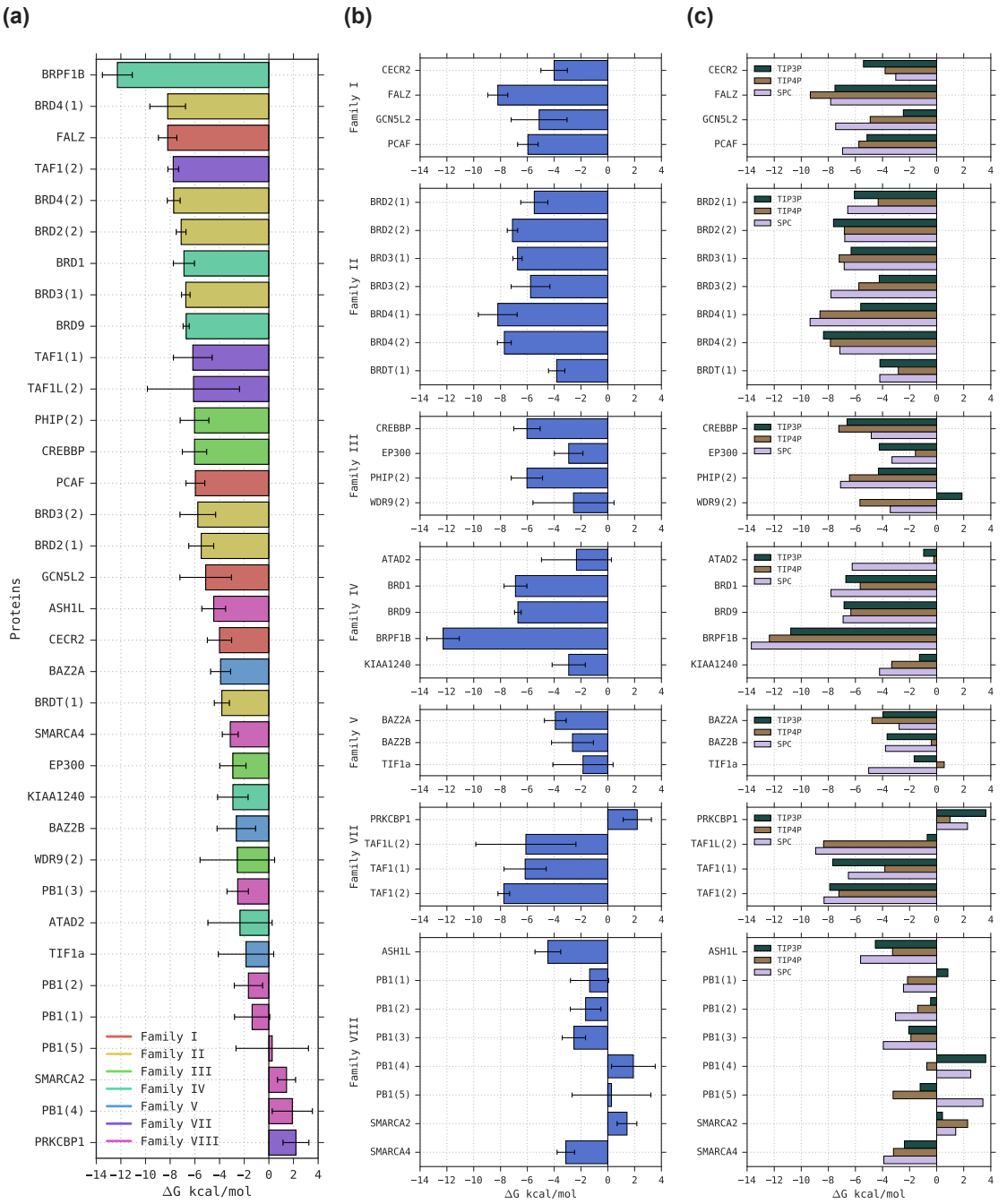


Figure 6.4: Stability of the water network measured as its binding free energy. (a) The BRDs here considered are sorted by predicted network stability and colour-coded by family; the bars and their errors show the weighted means and standard deviations of the three water models. (b) The BRDs are divided by family; here too the weighted means and standard deviations are shown. (c) The BRDs are divided by family and the results for each water model are shown.

with the exception of family VIII, most families do not cluster around a particular free energy value, suggesting that in many cases there might be enough difference in the water-network stability to potentially be exploited for the engineering of intra-family selectivity for specific bromodomains. This can be observed more easily in panel (b) of Figure 6.4, where the results have been divided by family. Here too it is possible to see that overall the water network for BRDs in family VIII is not particularly stable, with predicted binding free energies mostly smaller than -3 kcal/mol. The most stable network is surprisingly predicted to be the one for ASH1L (-4.5 kcal/mol), an atypical bromodomain where the conserved Asn residue is replaced by a Tyr. The tyrosine residue is present also in PB1(1), for which the predicted stability of the network is however lower, and around -1.4 kcal/mol. Family I and family II are predicted to have very stable networks for all of their BRDs, with CECR2 and BRDT(1) having the least tightly bound waters (ΔG_{netw} of about -4 kcal/mol). In Family III, EP300 and WDR9(2) appear to possibly have slightly more unstable networks overall as compared to CREBBP and PHIP(2). However, the prediction for WDR9(2) carries a large uncertainty. Family IV shows instead an interesting difference in stability between the closely related proteins ATAD2 and KIAA1240, and the other BRDs in the family. The predicted binding free energies for ATAD2 and KIAA1240 are in fact of -2.3 kcal/mol and -2.9 kcal/mol, respectively, while for BRD1, BRD9, and BRPF1B, are well above -6 kcal/mol. Family V appears to have a relatively unstable network of waters as compared to the other BRD families. BAZ2A has the largest predicted ΔG_{netw} of -3.9 kcal/mol, while in BAZ2B and TIF1α the predicted free energy is of -2.6 kcal/mol and -1.9 kcal/mol. In family VII, it is interesting to note how the network in PRKCBP1 was predicted to be unstable, with a positive ΔG_{netw}, contrary to the other proteins in the family.

Figure 6.4c shows the differences in predicted ΔG_{netw} between the three water models. This is also reflected in the uncertainty shown in panels (a) and (b) of the same figure, however, with the contributions weighted according to the quality of the water placement. Overall, the three models showcase similar trends. In

fact, the ΔG_{netw} values for all three models correlate with each other well, with a linear correlation of 0.70 between TIP3P and TIP4P, 0.74 between TIP3P and SPC, and 0.76 between TIP4P and SPC. In terms of absolute difference, the RMSD between ΔG_{netw} values is about 1.0-1.2 kcal/mol. This can also be seen by the distribution of the network free energies in Figure D.1. However, while the values of ΔG_{netw} obtained with the TIP3P and TIP4P models have an approximately normal distribution with maxima around -4 kcal/mol, the ones obtained with the SPC model have a bimodal distribution, with the global maximum at about -5 kcal/mol, and a local one around $+3$ kcal/mol. The minimum at zero suggests that, in this study, the model tends to avoid values close to zero and return slightly larger ΔG_{netw} more frequently than TIP3P and TIP4P. Cases in which the three models largely disagree, resulting in standard deviations over 2 kcal/mol, are GCN5L2, where the models suggest stabilities from -2 kcal/mol to -8 kcal/mol; WDR9(2), where the TIP3P model return a positive ΔG_{netw} value of $+1.9$ kcal/mol, while TIP4P and SPC return fairly large and negative free energies (-5.7 and -3.4 kcal/mol); TAF1L(2), where TIP4P and SPC return large and negative ΔG_{netw} (-8.3 and -8.9 kcal/mol), while TIP3P predicts the network to be poorly stable, with $\Delta G_{netw} = -0.7$ kcal/mol; PB1(5), where the TIP4P model gives a ΔG_{netw} of -3.2 kcal/mol, TIP3P of -1.2 kcal/mol, and SPC of $+3.4$ kcal/mol; ATAD2, where both TIP3P and TIP4P suggest the network to be poorly stable (ΔG_{netw} of -1.0 and -0.2 kcal/mol, respectively), while SPC suggests the waters to be tightly bound (ΔG_{netw} of -6.2 kcal/mol); and TIF1α, where, similar to ATAD2, TIP3P and TIP4P predict loosely bound waters (ΔG_{netw} of -1.7 and $+0.6$ kcal/mol) while SPC predicts them to be stable ($\Delta G_{netw} = -5.0$ kcal/mol).

Note that, in theory, a positive hydration free energy value means that the hydration of the volume considered is unfavourable, so that at equilibrium one might expect this volume to be free of water molecules. This is true for the GCI of the whole grand canonical region, for which the calculated hydration free energy is negative. However, because of the re-evaluation of the sub-volumes of this region in order to focus only on the four water sites of interest, the thermodynamic end

states of these calculations are not as well defined as for a standard GCI calculation. Therefore, ΔG_{netw} represents only the free energy contribution of the four water sites to the hydration free energy of the whole grand canonical region. While this contribution (ΔG_{netw}) might be positive, the contribution of all other waters in the GC region make the overall hydration free energy negative. The ΔG_{netw} here shown can also be interpreted as the free energy of binding (or displacing) the waters or water network when in absence of other water molecules in the GC region that are predicted to be less tightly bound than the ones considered. Thus, in this analysis it is assumed that part of the cavity has already been dehydrated. This is reasonable when considering a ligand will likely displace other waters before reaching the most stable and buried ones. Nonetheless, it is also somewhat artificial, as there is no control over exactly which parts of the cavity are dry, while a potential ligand will displace a well-defined volume of water. Overall, this results in positive ΔG values (e.g. the ΔG_{netw} in PB1(5), SMARCA2, PB1(4), and PRKCPB1), which do not mean that the sites are suggested to be dehydrated at equilibrium, but that there is a free energy gain upon displacement when other waters (primarily waters exposed to bulk) have already been removed. It is important to note that approaches such as double-decoupling are not completely realistic either when dealing with water binding. In fact, they calculate the binding free energy considering the displacement of a single water molecule, while it being still surrounded by other solvent. In this case the issue faced is the opposite: it is assumed that hydration environment of the cavity is left unchanged by the presence of a possible ligand.

6.5 Stability of individual waters

In the previous section, the stability of the four-water network as a whole was investigated. Here, the stability of individual water sites will be discussed instead. Figure 6.5 summarises the results for all four water sites for all bromodomains here studied. The heatmap on the left-hand side of the figure shows the weighted mean of the binding free energies for the water site W1 to W4 obtained from the three water models; the map on the right-hand side shows instead the standard

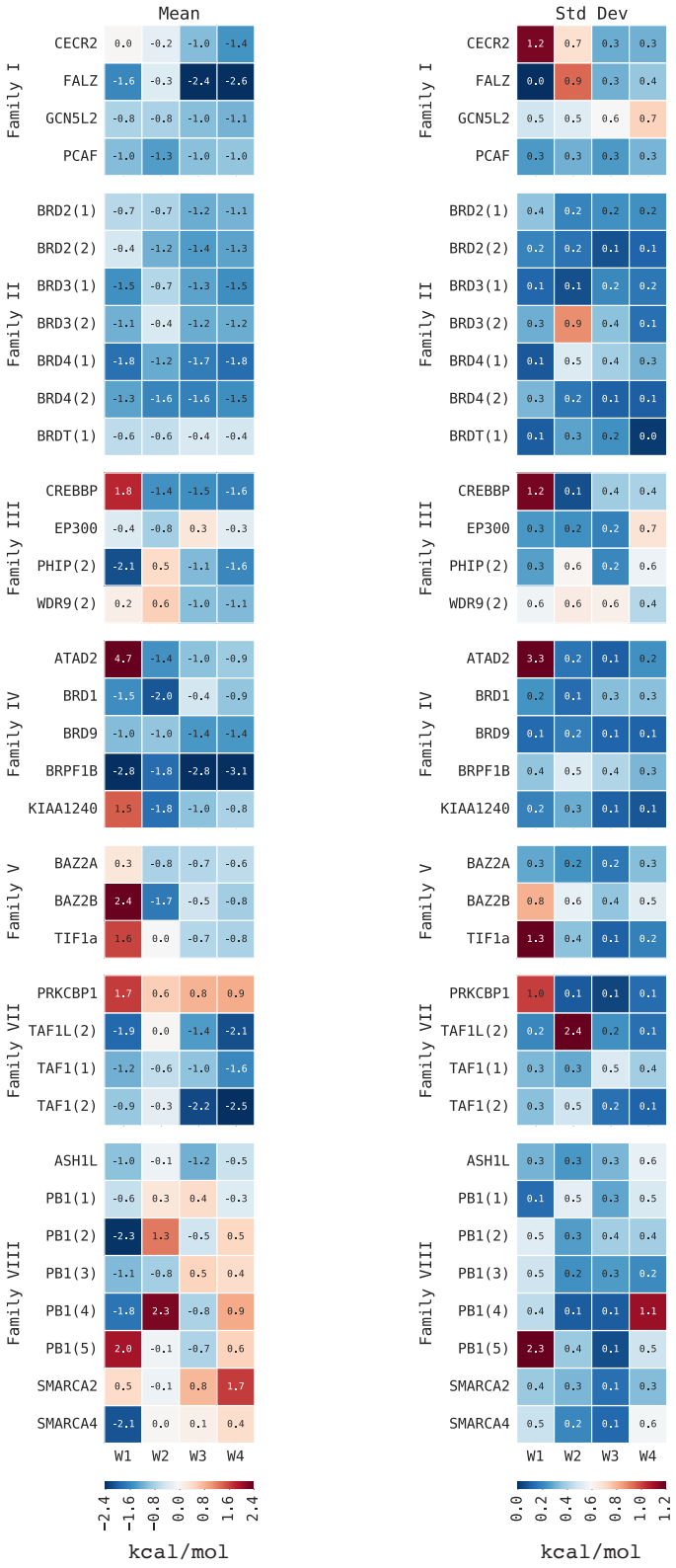


Figure 6.5: Stability of individual water sites for all bromodomains considered. The heatmap on the left shows the predicted binding free energies as the weighted mean of the three water models for each of the four water sites. The heatmap on the right shows the uncertainty in the estimate as the weighted standard deviation from the free energy values obtained with the different water models.

deviations, which give an idea of the uncertainty in the calculations. In this case also the weights used are based on the quality of the water placement:

$$w^i_{W_x} = \frac{1}{\delta^i_{W_x}} \tag{6.5}$$

$$\overline{\Delta G}_{W_x} = \frac{\sum\limits_{i \in M} \Delta G^i_{W_x} \cdot w^i}{\sum\limits_{i \in M} w^i} \tag{6.6}$$

$$\sigma_{W_x} = \sqrt{\frac{\sum\limits_{i \in M} w^i \cdot (\Delta G^i_{W_x} - \overline{\Delta G}_{W_x})^2}{\sum\limits_{i \in M} w^i}} \tag{6.7}$$

where $w^i_{W_x}$ is the weight for water site W_x (with $W_x \in \{W1, W2, W3, W4\}$) and water model i, δ is the distance between the predicted and crystallographic position of the water site, $M = \{TIP3P, TIP4P, SPC\}$, and σ is the standard deviation. From here on, I will refer to the average $\overline{\Delta G}_{W_x}$ simply as ΔG_{W_x}.

It is possible to note how in family I and II, the waters seem to be stable not only as a network but also as individual sites. W1 in CECR2 is the only site predicted to be potentially poorly stable ($\Delta G_{W1} = 0.0$ kcal/mol), although the relatively large uncertainty for this site undermines the confidence in this hypothesis. In family III, it is interesting to see how W1 in CREBBP is predicted to be unstable. However, this prediction also has a relatively large uncertainty; in this case the free energy is significantly shifted toward a positive value, with $\Delta G_{W1} = +1.8$ kcal/mol. Note that in general the free energy values with largest uncertainty are the ones associated with waters predicted to be unstable. In family IV, most waters appear to be stable, with the exception of W1 in ATAD2 and KIAA1240. Again, the ΔG_{W1} in ATAD2 has a large uncertainty associated with it, however, the free energy value is very large and positive (+4.7 kcal/mol). In addition, the fact that the closely related protein KIAA1240 shows an unstable water at the W1 site, corroborates the hypothesis that this site might be targetable for displacement more easily than other sites or bromodomains. Structurally, ATAD2 and KIAA1240 are slightly different from the other BRDs of family IV, mainly due to a four-residue insertion in the BC loop. However, when looking at the structural details near the binding pocket, it is

hard to identify the potential source of instability for W1. The only difference that is clearly visible in proximity of W1 is the presence of an isoleucine at the top of the α_C helix in ATAD2 and KIAA1240, which is a phenylalanine or tyrosine in the other BRDs of family IV. One could speculate that the bulkier residues provide a more occluded space for a water to occupy site W1. In fact, a few other BRDs outside family IV and with unstable W1 carry an isoleucine residue at this position too: PB1(5), SMARCA2, and BAZ2B. On the other hand, there are also BRDs where an isoleucine (or another small residue, such as valine or leucine) in this position does not result in a poorly stable W1; e.g. BRD2(1), BRD3(1), BRD4(1), BRDT(1). This factor alone cannot therefore explain the instability observed for W1 in ATAD2 and KIAA1240. Family V shows a clear trend of unstable water at the W1 site. Again, the uncertainty here is larger than average, yet the consistency within the family and the fact the ΔG_{W1} are significantly above zero suggest that W1 is possibly not stable in these proteins. In family VII, PRKCBP1 displays poor stability for all four of the water sites. This results in the whole network to be unstable, as indicated by the positive ΔG_{netw} discussed previously. Family VIII is interesting as all water sites appear to be unstable in some of the proteins. Firstly, it is possible to note a trend of instability for the site W4. This is in agreement with our previous MD study of water thermodynamics in PB1(5) versus BRD4(1), in which it was hypothesised that the instability at this position is caused by a $\sim 90°$ rotation of a backbone carbonyl group [80]. W2 as well does not appear to be particularly stable in most BRDs in this family, with exception of PB1(3). W3 is instead predicted to be unstable in PB1(1), PB1(3), SMARCA2, and SMARCA4, with good precision. Contrary to what observed in other families, W1 seems instead to be tightly bound to most proteins of family VIII, with the exception of PB1(5) and SMARCA2.

Figure 6.6 summarises the same data discussed above for the four individual sites (W1, W2, W3, W4) but focuses on the distribution of their free energies. From the plot showing the data for all proteins (top left), it is possible to notice how in most cases waters in all four sites are stable, with a binding free energy between 0.0 kcal/mol and -2.0 kcal/mol. In a few instances, these waters are very

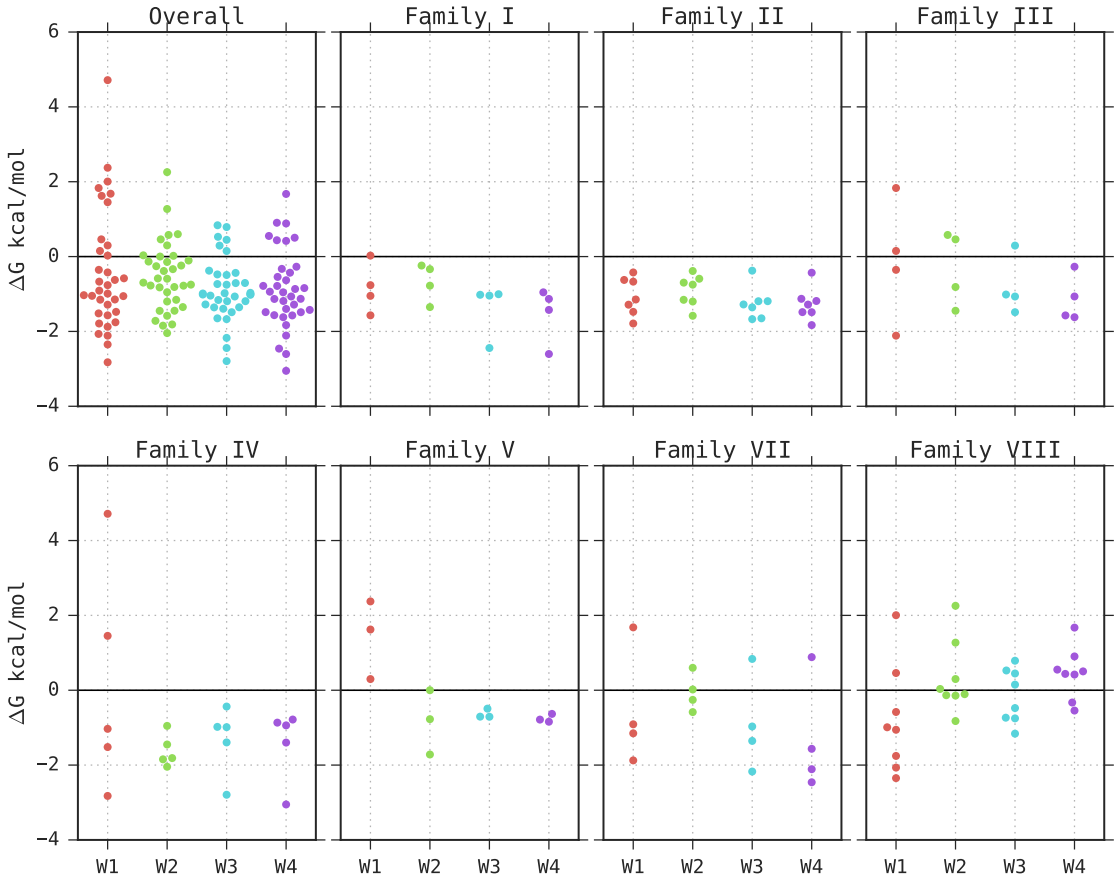


Figure 6.6: Swarm plots showing the distribution of the stabilities of individual water sites by family.

tightly bound with $\Delta G_W < -2$ kcal/mol. In a number of other cases, however, ΔG_W is positive, indicating the presence of potentially unstable waters. As noted above, family I and II display well-balanced water stability across all four sites. In family I, water sites W3 and W4 appear however to be slightly more stable than W1 and W2. In family III there is a larger spread of binding free energy values instead. Most waters seem stable, however, W1, W2, and W3 can all be found above the zero mark in at least one instance. In family IV, all water sites appear stable, with the exception of W1 in ATAD2 and KIAA1240. The instability of W1 is thus probably the cause of the lower stability for the whole network in these two proteins. Similarly, in family V, waters at the sites W2, W3, and W4 are predicted to be stable, contrary to the waters at W1. In this scenario where a single water site seems particularly unstable, trying to rationally target a single water

for displacement might be an alternative strategy to the displacement of the whole network. In family VII, as previously discussed, the waters in most protein cavities appear to be stable overall, apart from PRKCBP1, which has unstable waters at all four sites. Moreover, in this family, W2 appears to be overall the least stable. However, targeting W2 alone for displacement might be challenging considering it sits below W1; from the viewpoint of ligand geometry it might therefore be easier to target W1, or W1 and W2 together. From Figure 6.6 it is also possible to see again how the waters in the pockets of family VIII BRDs are overall the least stable. In particular, W4 is unstable in most cases, with many positive free energy values. Similarly, the binding free energy of waters at site W2 is mostly positive or close to zero. For W3 there appears to be an even split, with four proteins displaying $\Delta G_{W3} > 0$ and four $\Delta G_{W3} < 0$, all of them being however within 1.0 kcal/mol of the zero mark. Waters at site W3 are the only ones that are predicted to be stable in most proteins, with three instances of $\Delta G_{W3} \approx -2$ kcal/mol, three of $\Delta G_{W3} \approx -1$ kcal/mol, and two of $\Delta G_{W3} > 0$ kcal/mol.

As mentioned in the introduction of this chapter, the water at site W1 has recently been shown to be displaceable by specific ligands from the binding pockets of the bromodomains BRPF1B and PHIP(2). As possible to see in Figure 6.6, W1 does have a wider distribution of stabilities as compared to the other sites, so that it appears to be the most displaceable from these families. However, when looking in more detail at the binding free energy for W1 for these two bromodomains in Figure 6.5, it is possible to see how W1 is predicted to be stable. Therefore, in this case the predictions seem to disagree with what is known from the experimental evidence. The two possible simple explanations for this fact are the following. The first is that in this case the simulations fail to capture the relative instability of this site, due to either limited sampling or inadequate physical representation of intermolecular interactions. The second possibility is that W1 is indeed stable, however, the right ligand can still offset the displacement penalty in order to bind the pocket. If the other three waters are then stabilised by the presence of the ligand, part of the penalty might also be recovered. In fact, in the fragment screen

that discovered a number of BRPF1B binders [146], only one (PDB-ID 5C87) out of nineteen ligands displaced W1, and did not result in higher affinity. In the fragment screen that identified four hits for the atypical bromodomain PHIP(2) [158], only one (a thiourea; PDB-ID 5ENB) displaced W1; in this case too, the displacement did not result in particularly higher affinities as compared to the other hits.

It must be noted that the presence of a small molecule will change the environment around the hydration site in a ligand-specific manner, so that the stability of the waters might change depending on the chemical details of the organic molecules. This aspect will be further discussed in the next section.

6.6 Discussion

In this chapter, I estimated the stability of the water network found in bromodomains through Monte Carlo simulations and free energy calculations in the GC ensemble. In order to try limit the deficiencies of specific water models, multiple models have been used, and their ability to predict the position of the structured waters was taken as a proxy for the reliability of the free energy estimates. Here, I focused on the four waters that form the most conserved network in bromodomains. Additional waters can be found in the X-ray structures of different bromodomains, however, these are typically less buried waters and are not conserved across all the proteins. While being conserved, the four water sites studied here are also targetable for displacement when considering their position in the protein, as they sit at the bottom of the binding pockets where a ligand or fragment could be grown into. One additional water, which is more buried into the protein than the four discussed here, is observed in most X-ray structures and has been predicted to be present by the current calculations in all bromodomains studied. This water site, however, is out of the plane of the four-water network and seems to be out of reach for most ligand geometries. In addition, the binding free energy for this water was predicted to be particularly large, suggesting that this water is truly structural and hardly displaceable. For instance, in the TIP4P model, the average binding free energy for this water site was of -4.7 kcal/mol (range from -1.1 to -8.8

kcal/mol). Based on these observations, the development of a ligand that can displace this water site would be remarkable.

As already mentioned, it is important to keep in mind that eventually it is the specific nature of the ligand that will determine whether it can bind a pocket and displace conserved waters. Its geometry will first determine whether it can occupy a certain volume of space (shape complementarity). Its presence will also affect the environment of the binding pockets so that the stability of the waters might be affected as well. Finally, the interactions it can make with the protein (and the undisplaced waters, if any) will determine whether the free energy penalty often associated with the displacement of conserved waters can be compensated. Of course, the smaller the free energy penalty associated with the displacement, the higher the affinity the ligand might be able to reach. In this study, by considering the apo proteins only, any effect due to the details of the chemical nature of the ligand has been disregarded. On one hand, the results can be considered to be more general, while they are also less accurate for any specific ligand. Thus, the data shown here provide guidance on which waters and water networks are expected to be more easily displaceable (by a generic molecule) given only the nature of the proteins binding pockets. Different ligands will then have different abilities to target certain bromodomains and water sites. It is expected that it will be easier to discover a ligand that displaces the waters that are predicted to be less tightly bound. Nonetheless, it might be possible to develop a small molecule that displaces tightly bound waters too, whilst on the other hand it is also possible to design a small molecule that does not displace waters despite them being loosely bound.

In addition to what is discussed above, limited sampling and the approximations of the physical models can affect the results too. For these reasons, consensus observations where the results are conserved across the whole family might be more reliable than observations about individual bromodomains. For instance, the hypothesis that W4 is loosely bound across all BRDs of family VIII is more likely to be reliable than the hypothesis that W1 is unstable only in PB1(5) (see Figure 6.5). At the same time, however, specific disparities within a family and for specific

bromodomains might be very interesting from a medicinal chemistry viewpoint. An example is the observation that the water network is predicted to be stable in family VII apart for PRKCBP1 (Figure 6.4). Unfortunately, there is currently no means to quantitatively validate predictions on water stability: no experimental technique can measure water binding free energy at physiological conditions. It is only possible to assess the ability of the simulations to correctly suggest the location of conserved water sites, by comparing the predictions to X-ray structures, and to qualitatively compare the computational results to past experience. The calculations managed to identify the position of the four-water network accurately, with an average RMSD as compared to X-ray structures well below 1 Å for all three water models tested. In addition, the calculations correctly identified the water network in family VIII BRDs to be poorly stable, in agreement with the experimental observation that the network is displaceable from these proteins. However, the predictions can only really be validated prospectively by targeting specific BRDs for water displacement and observing whether differential binding affinities may be attributable only to the free energy penalty associated with this solvent displacement. In any case, this would still be a qualitative comparison, since it is not possible to experimentally decompose binding affinity into contributions arising from different events. Here, I therefore used the ability of the calculations to recapitulate the water positions only as a proxy for the reliability of the free energy estimates. Overall, these predictions are expected to provide better guidance than human judgement based only on visual inspection of protein structures, and try to provide a quantitative measure of the difficulty of displacing waters from bromodomains.

6.7 Conclusion

In summary, here I presented a study of water stability in bromodomain binding pockets that made use of a modern free energy technique based on Monte Carlo simulations in the grand canonical ensemble. The results shown provide overall guidance to medicinal chemists as to which waters and water networks might be more easily targetable for displacement. A number of interesting observations have

been made regarding specific bromodomain families or proteins. In particular, it was noticed how the four-water network appears poorly stable in family VIII, in agreement with experimental evidence, and how the next least stable network is observed in family V. In addition, the whole network appears to be potentially displaceable also from the bromodomains PRKCPB1, ATAD2 and KIAA1240. In terms of specific water sites within the network, W1 was predicted to be the most easily targetable site overall, in particular in family V, but also in families III and IV. An exception seems however to be family VIII, where W4 was predicted to be the least stable site. Outside family VIII, the least tightly bound sites appear to be W1 in CREBBP (family III), W1 in ATAD2 and KIAA1240 (family IV), and W1 in BAZ2A, BAZ2B and TIF1α (family V). Overall, these data will be useful for guiding the rational targeting of conserved waters in BRD pockets with small molecules.

A conclusion is the place where you got tired thinking.

— Martin H. Fischer

7 Summary

In this thesis I presented a series of computational studies on protein-ligand binding. In particular, the work focussed on the use of free energy calculations based on molecular dynamics simulations for application to drug design. The approaches have been more specifically tested and applied to the epigenetic target family of bromodomains. Such physics-based approaches rely on general models of molecular interactions that try to be as transferable as possible so that target-specific training sets are not needed. In addition, the full dynamics of the all-atom system is sampled, so that entropic contributions and the discrete nature of the solvent are naturally taken into account.

In Chapter 3 I tested the ability of ABFE calculations based on a standard implementation of alchemical transformation to predict the binding free energy of a diverse set of 11 ligands binding to BRD4(1). For this system, ABFE calculations based on molecular dynamics were able to return excellent agreement with ITC data both in terms of correlations (Pearson and Spearman correlations of ~ 0.8) and absolute errors (RMSE < 1.4 kcal/mol), both when starting from X-ray and docked ligand poses. The results suggested that these calculations could be a means to accurate, though computationally expensive, affinity predictions. The novelty of this study as compared to previous ones is that, first, the binding a diverse set of drug-like molecules, rather than small and rigid fragments, have

been considered, and second, the target was a pharmacologically relevant protein currently investigated for its therapeutic potential, rather than an engineered cavity. In addition, thanks to continuous advances in hardware and software, it was possible to reach a simulated time of about 700 ns per calculation, which resulted in good convergence of the calculations. The study showed how ABFE calculations can be used as an accurate rescoring approach for the evaluation of docking poses. Finally, it was noted how torsional parameters in small molecule force fields may negatively affect the results of ABFE calculations, and how this can be tackled by the derivation of new parameters fitted to QM derived energies.

In Chapter 4 the validation of the free energy protocol was extended to multiple bromodomains. The work aimed to test the performance of the predictions on a larger range of proteins, and to assess whether the calculations could be applied for the prediction of ligand selectivity. In this case, no knowledge of the structure of the complexes was assumed from the start of the study, and hypotheses about ligand poses were generated via docking. Moreover, in this work, particular attention was paid to the parameters of the small molecules, as it was noticed that MM torsional energies can sometimes deviate substantially from QM ones. In this study it was observed how the performance of the calculations was satisfying for the ligands RVX-208 and RVX-OH (correlations of > 0.7 and RMSE < 1.0 kcal/mol), while it was more modest for bromosporine (correlations of ~ 0.5 and RMSE ≈ 2.0 kcal/mol). Nonetheless, taken together these results suggest that ABFE calculations might indeed be applied successfully for the prediction of ligand selectivity, despite some challenges. In addition, the work showed how simple refinement of torsional parameters could provide more accurate affinity predictions, suggesting there is still scope for the improvement of classical fixed-charge force fields of organic molecules.

Computational predictions of binding affinity and ligand selectivity remain challenging. However, with ever improving theoretical frameworks and physical models, and increasing computational power, these questions can start being addressed with rigorous and widely applicable approaches. Further investigation will be needed in order to assess the effectiveness of such predictions in different

case scenarios and on different systems. In fact, the accuracy of the calculations on protein families other than bromodomains still needs to be established. Many popular drug targets are much larger and flexible than the proteins here considered (e.g. kinases and GPCRs). In addition, the ligands considered represent only a tiny fraction of the available drug-like chemical space. As the computational and human cost of setting up and running such (now intensive) calculations decreases, it will become increasingly possible to test the performance of the predictions on larger sets of proteins and ligands.

ABFE calculations, while proved to be accurate in our hands, are also (and will probably always be) computationally more expensive than alternative approximate approaches based on the post-processing of MD trajectories in implicit solvent. Moreover, approaches like MMPBSA have already been widely adopted by the academic community. Such techniques make use of a number of approximations for computational efficiency, yet they have been shown to be useful in ranking protein-ligand complexes by their relative affinities. Thus, in Chapter 5, I compared the performance of a number of complementary MMPBSA protocols to ABFE calculations, on the same sets studied in Chapters 3 and 4. Given the scarcity of ABFE studies in the literature on more than a handful of ligands, also comparisons to more approximate methods are lacking. This work suggested that ABFE calculations are indeed more reliable than MMPBSA calculations overall, with a weighted average improvement in terms of correlation to ITC data of $\sim$0.10 points for the test sets considered in this thesis. While this might have been an expected outcome, it is useful to be able to quantify the accuracy gains with respect to the additional costs. In fact, some improved MMPBSA protocols achieved overall correlations that were only about 15% worse than ABFE, despite needing about 5% of the compute time. Nonetheless, ABFE calculations proved to be distinctly superior to MMPBSA in terms of absolute agreement with experimental free energy values (weighted average RMSE of 1.5 kcal/mol versus 5.3 kcal/mol for ABFE versus the best MMPBSA protocol).

In the last result chapter of this thesis (Chapter 6) I described a different but related study on bromodomains, in which the binding of water molecules to the proteins' cavities has been evaluated using free energy calculations based on Monte Carlo simulations in the grand canonical ensemble. Structural water molecules are ubiquitous in protein structures and are often targeted for displacement by medicinal chemists in an attempt to achieve higher potency or selectivity. A network of four conserved waters is also observed in most bromodomain structures and there is an interest in knowing from which BRDs such waters might be easily displaced. Nonetheless, the rigorous evaluation of water displaceability is far from trivial and a number of technical challenges makes it just as, if not more, challenging than predicting ligand binding free energies. A recently proposed approach, named Grand Canonical Integration (GCI), offers a solution to some of these challenges. In the work here presented, the stability of the water network in 35 of the 61 known human bromodomains was investigated using GCI. Three classical water models were used in order to alleviate model-specific biases, and the calculations were also repeated in triplicates in order to assess the reproducibility of the results. A number of differences between the proteins studied emerged in terms of stability of their water network and individual water sites. Firstly, in agreement to what is known experimentally, the four-water network appeared to be poorly stable in BRDs belonging to family VIII. However, also the network in the BRDs of family V, and in the bromodomains PRKCPB1, ATAD2 and KIAA1240, did not appear to be particularly tightly bound to the proteins and thus potentially displaceable. When looking at specific water sites within the network, one water site specifically emerged as being the most easily targetable overall, and in particular in CREBBP (family III), ATAD2 and KIAA1240 (family IV), and BAZ2A, BAZ2B and TIF1α (family V). While it is not possible to directly and quantitatively assess the accuracy of such predictions, the results presented in Chapter 6 might prove useful in providing some guidance to medicinal chemists on which waters and water networks might be more easily targetable for displacement in bromodomains.

Thus, overall, in this thesis I presented a mixture of validation and application of different free energy methods based on statistical mechanics principles and that are founded on physics-based computer simulations of biomolecules. The data shown and discussed can be useful both for computational and theoretical chemists for the advancement as well as application of such approaches, and for medicinal chemists and structural biologists interested more specifically in ligand binding to bromodomains.

Appendices

A

Supplementary material for Chapter 3

Table A.1: Physicochemical properties of the ligands considered in this study. MW is the molecular weight in Daltons; Net q is the net charge; No. Rot. is the number of rotatable bonds; HBA is the number of hydrogen bond acceptors; HBD is the number of hydrogen bond donors; cLogP is the calculated octanol/water partition coefficient (computed with MarvinSketch); † cLogP of the nonionic specie.

Ligand	MW (Da)	No.Atoms	Net q	No.Rot.	HBA	HBD	cLogP
1	522.7	78	+1	8	10	2	$-0.4(3.1)^{\dagger}$
2	457.0	56	0	5	6	0	5.3
3	326.4	42	0	3	6	2	3.0
4	525.7	74	+1	11	9	3	$1.4(4.9)^{\dagger}$
5	295.3	39	0	3	4	2	3.0
6	323.8	37	0	1	5	0	3.8
7	370.4	49	0	6	7	2	2.5
8	308.8	35	0	1	4	0	3.0
9	261.3	38	0	4	4	1	2.2
10	241.1	22	0	0	3	1	1.8
11	242.2	26	0	2	3	2	2.3

Table A.2: Full breakdown of free energy results starting from the crystal structures. All values are in kcal/mol. The meaning of the different free energy terms within the thermodynamic cycle is explained in § 2.6.1 and Figure 2.5. $\Delta\Delta G_{exp-lr}$ is the contribution to the binding free energy due to the long-range dispersion correction. $\Delta G_{elec\ corr}$ is the contribution to the binding free energy due to the correction for electrostatic finite size effects and it has been applied to charged ligands only. The reason of these two corrections is explained in § 2.3.5 and § 2.3.5. All ΔG terms shown here have been rounded to the nearest tenth for ease of reading, despite having been calculated at higher precision.

Ligand	$\Delta G^{solv}_{elec+vdw}$	ΔG^{solv}_{restr}	$\Delta G^{prot}_{elec+vdw+restr}$	$\Delta\Delta G_{exp-lr}$	$\Delta G_{elec\ corr}$
1	189.7 ± 0.2	6.6	-206.6 ± 0.2	-0.7	0.6
	189.7 ± 0.1	6.6	-206.7 ± 0.1	-0.6	0.6
	190.5 ± 0.2	6.6	-207.5 ± 0.3	-0.7	0.6
2	59.2 ± 0.2	6.6	-74.8 ± 0.1	-0.7	
	59.4 ± 0.2	6.6	-74.5 ± 0.4	-0.7	
	59.6 ± 0.1	6.6	-74.5 ± 0.1	-0.7	
3	219.3 ± 0.1	6.7	-235.1 ± 0.2	-0.6	
	219.1 ± 0.1	6.7	-234.7 ± 0.1	-0.6	
	219.5 ± 0.2	6.7	-233.6 ± 0.1	-0.6	
4	444.2 ± 0.3	6.8	-459.8 ± 0.2	-0.8	0.7
	445.1 ± 0.2	6.8	-460.9 ± 0.4	-0.8	0.7
	444.1 ± 0.3	6.8	-460.7 ± 0.2	-0.8	0.7
5	57.8 ± 0.1	7.0	-72.7 ± 0.1	-0.5	
	58.0 ± 0.1	7.0	-73.3 ± 0.1	-0.5	
	57.9 ± 0.0	7.0	-73.0 ± 0.1	-0.5	
6	24.2 ± 0.1	6.6	-40.6 ± 0.2	-0.6	
	25.4 ± 0.1	6.6	-40.5 ± 0.2	-0.6	
	23.7 ± 0.1	6.6	-40.0 ± 0.1	-0.6	
7	201.5 ± 0.1	6.6	-213.0 ± 0.2	-0.5	
	201.4 ± 0.2	6.6	-213.8 ± 0.2	-0.5	
	201.7 ± 0.1	6.6	-213.8 ± 0.2	-0.5	
8	4.1 ± 0.1	6.9	-18.5 ± 0.1	-0.5	
	4.4 ± 0.2	6.9	-18.3 ± 0.2	-0.5	
	3.9 ± 0.1	6.9	-18.2 ± 0.2	-0.5	
9	57.6 ± 0.1	6.7	-71.8 ± 0.1	-0.4	
	57.3 ± 0.1	6.7	-71.4 ± 0.1	-0.4	
	57.4 ± 0.1	6.7	-71.8 ± 0.2	-0.4	
10	89.7 ± 0.1	7.2	-102.6 ± 0.1	-0.4	
	89.6 ± 0.1	7.2	-102.8 ± 0.1	-0.4	
	89.8 ± 0.1	7.2	-102.4 ± 0.1	-0.4	
11	57.2 ± 0.1	6.7	-68.3 ± 0.2	-0.4	
	57.2 ± 0.1	6.7	-68.8 ± 0.1	-0.4	
	55.7 ± 0.1	6.7	-66.7 ± 0.1	-0.4	
	55.8 ± 0.1	6.7	-67.1 ± 0.0	-0.4	

Table A.3: Full breakdown of free energy results starting from the docking poses. All values are in kcal/mol. The meaning of the different free energy terms within the thermodynamic cycle is explained in § 2.6.1 and Figure 2.5. $\Delta\Delta G_{exp-lr}$ is the contribution to the binding free energy due to the long-range dispersion correction. $\Delta G_{elec\ corr}$ is the contribution to the binding free energy due to the correction for electrostatic finite size effects and it has been applied to charged ligands only. The reason of these two corrections is explained in § 2.3.5 and § 2.3.5. All ΔG terms shown here have been rounded to the nearest tenth for ease of reading, despite having been calculated at higher precision.

Ligand	Pose	$\langle\Delta G^{solv}_{elec+vdw}\rangle$	$\Delta G^{solv}_{elec+vdw}$	ΔG^{solv}_{restr}	$\Delta G^{prot}_{elec+vdw+restr}$	$\Delta\Delta G_{exp-lr}$	$\Delta G_{elec\ corr}$
1	a		195.6 ± 0.2	6.5	-213.1 ± 0.1	-0.7	0.6
	b	196.1 ± 0.7	197.1 ± 0.3	7	-207.8 ± 0.3	-0.6	0.7
	c		195.7 ± 0.4	6.5	-213.7 ± 0.3	-0.7	0.8
2	a	59.8 ± 0.2	60.0 ± 0.1	6.6	-75.9 ± 0.3	-0.6	
	b		59.5 ± 0.0	7.3	-73.2 ± 0.4	-0.5	
3	a		207.5 ± 0.1	6.7	-219.8 ± 0.1	-0.5	
	b		207.6 ± 0.1	6.6	-220.1 ± 0.3	-0.4	
	c	207.5 ± 0.2	207.6 ± 0.1	6.9	-224.6 ± 0.1	-0.6	
	d		207.3 ± 0.1	6.8	-224.3 ± 0.1	-0.5	
	e		207.3 ± 0.1	6.8	-221.1 ± 0.1	-0.5	
4	a	443.7 ± 0.3	443.7 ± 0.3	6.7	-461.6 ± 0.3	-0.8	0.6
5	a	56.5 ± 0.1	56.5 ± 0.1	6.6	-68.5 ± 0.1	-0.4	
	b		56.6 ± 0.1	6.7	-71.1 ± 0.1	-0.5	
6	a	24.0 ± 0.1	24.0 ± 0.1	6.6	-40.4 ± 0.3	-0.5	
	b		24.1 ± 0.1	6.8	-40.8 ± 0.2	-0.6	
7	a	187.1 ± 0.1	187.1 ± 0.1	6.7	-199.9 ± 0.2	-0.5	
	b		187.2 ± 0.1	6.7	-198.5 ± 0.1	-0.5	
8	a	4.3 ± 0.1	4.3 ± 0.1	6.6	-20.6 ± 0.1	-0.5	
	b		4.3 ± 0.1	6.8	-17.2 ± 0.1	-0.5	
9	a		56.4 ± 0.1	6.6	-70.4 ± 0.1	-0.4	
	b	56.5 ± 0.1	56.5 ± 0.1	6.6	-70.0 ± 0.1	-0.4	
	c		56.5 ± 0.1	6.8	-69.0 ± 0.1	-0.3	
10	a	89.8 ± 0.0	89.8 ± 0.0	7.1	-102.7 ± 0.1	-0.4	
11	a	57.0 ± 0.1	57.1 ± 0.0	6.6	-68.6 ± 0.1	-0.4	
	b		56.9 ± 0.1	6.6	-68.7 ± 0.1	-0.3	

B

Supplementary material for Chapter 4

Table B.1: Information on the protein models used in the calculations.

Protein	Family	PDB-ID	Resolution (Å)	R-Value Free	R-Value Work
CECR2	I	3NXB	1.83	0.214	0.176
FALZ	I	3UV2	1.58	0.209	0.162
PCAF	I	3GG3	2.25	0.245	0.207
BRD2(1)	II	4ALG	1.60	0.202	0.176
BRD2(2)	II	4MR5	1.63	0.170	0.136
BRD3(1)	II	3S91	2.06	0.233	0.195
BRD3(2)	II	3S92	1.36	0.186	0.131
BRD4(1)	II	2OSS	1.35	0.174	0.149
BRD4(2)	II	2OUO	1.89	0.216	0.182
BRDT(1)	II	4KCX	2.00	0.253	0.209
CREBBP	III	4NYX	1.10	0.135	0.119
EP300	III	3I3J	2.33	0.275	0.229
BRD1	IV	5AME	1.58	0.204	0.181
BRD9	IV	4XY8	1.70	0.186	0.148
BRPF1B	IV	4LC2	1.65	0.220	0.174
BAZ2A	V	4QBM	1.65	0.205	0.185
TIF1	V	4YBM	1.46	0.196	0.163
TAF1(1)	VII	3UV5	2.03	0.230	0.190
TAF1(2)	VII	3UV4	1.89	0.225	0.191
TAF1L(2)	VII	3HMH	2.05	0.252	0.181
PB1(5)	VIII	3MB4	1.66	0.219	0.197
SMARCA4	VIII	2GRC	1.50	0.264	0.242

Table B.2: Full breakdown of free energy components for the calculations of RVX-OH and RVX-208. All free energies are in kcal/mol, and all uncertainties are one standard deviation. The meaning of the different free energy terms within the thermodynamic cycle is explained in § 2.6.1 and Figure 2.5.

Protein	Ligand	Pose	$\Delta G^{solv}_{elec+vdw}$	ΔG^{solv}_{restr}	$\Delta G^{prot}_{elec+vdw+restr}$	$\Delta\Delta G_{exp-lr}$
BRD2(1)	RVX-OH	H	82.189 ± 0.025	6.901	-98.523 ± 0.110	-0.403
BRD2(2)	RVX-OH	H	82.189 ± 0.025	7.340	-96.732 ± 0.118	-0.361
BRD3(1)	RVX-OH	H	82.189 ± 0.025	7.205	-95.696 ± 0.114	-0.383
BRD3(2)	RVX-OH	H	82.189 ± 0.025	7.344	-95.850 ± 0.170	-0.309
BRD4(1)	RVX-OH	H	82.189 ± 0.025	6.922	-98.510 ± 0.077	-0.501
BRD4(2)	RVX-OH	H	82.189 ± 0.025	7.254	-94.052 ± 0.104	-0.427
BRDT(1)	RVX-OH	H	82.189 ± 0.025	7.247	-96.707 ± 0.119	-0.408
BRD2(1)	RVX-OH	V	82.189 ± 0.025	6.587	-94.963 ± 0.097	-0.332
BRD2(2)	RVX-OH	V	82.189 ± 0.025	6.569	-96.417 ± 0.056	-0.251
BRD3(1)	RVX-OH	V	82.189 ± 0.025	6.596	-94.386 ± 0.101	-0.118
BRD3(2)	RVX-OH	V	82.189 ± 0.025	6.522	-95.695 ± 0.090	-0.384
BRD4(1)	RVX-OH	V	82.189 ± 0.025	6.573	-95.019 ± 0.101	-0.342
BRD4(2)	RVX-OH	V	82.189 ± 0.025	6.585	-96.869 ± 0.078	-0.405
BRDT(1)	RVX-OH	V	82.189 ± 0.025	6.610	-94.821 ± 0.079	-0.324
BRD2(1)	RVX-208	V	89.772 ± 0.023	6.521	-103.147 ± 0.070	-0.385
BRD2(2)	RVX-208	V	89.772 ± 0.023	6.584	-103.502 ± 0.078	-0.380
BRD3(1)	RVX-208	V	89.772 ± 0.023	6.573	-101.583 ± 0.132	-0.288
BRD3(2)	RVX-208	V	89.772 ± 0.023	6.562	-104.737 ± 0.126	-0.312
BRD4(1)	RVX-208	V	89.772 ± 0.023	6.546	-102.727 ± 0.120	-0.387
BRD4(2)	RVX-208	V	89.772 ± 0.023	6.575	-105.663 ± 0.087	-0.365
BRDT(1)	RVX-208	V	89.772 ± 0.023	6.595	-102.617 ± 0.080	-0.228

Table B.3: Full breakdown of free energy components for the calculations of bromosporine. All free energies are in kcal/mol, and all uncertainties are one standard deviation. The meaning of the different free energy terms within the thermodynamic cycle is explained in § 2.6.1 and Figure 2.5.

Protein	Family	$\Delta G^{solv}_{elec+vdw}$	ΔG^{solv}_{restr}	$\Delta G^{prot}_{elec+vdw+restr}$	$\Delta\Delta G_{exp-lr}$
CECR2	I	167.545 ± 0.227	6.725	-185.349 ± 0.123	-0.405
FALZ	I	167.545 ± 0.227	6.734	-183.669 ± 0.089	-0.412
PCAF	I	167.545 ± 0.227	6.852	-184.576 ± 0.126	-0.494
BRD2(1)	II	167.545 ± 0.227	6.736	-181.497 ± 0.122	-0.462
BRD2(2)	II	167.545 ± 0.227	6.808	-183.664 ± 0.133	-0.349
BRD3(1)	II	167.545 ± 0.227	6.850	-181.284 ± 0.155	-0.338
BRD3(2)	II	167.545 ± 0.227	6.534	-185.243 ± 0.105	-0.497
BRD4(1)	II	167.545 ± 0.227	6.879	-185.162 ± 0.114	-0.537
BRD4(2)	II	167.545 ± 0.227	6.919	-184.046 ± 0.127	-0.436
BRDT(1)	II	167.545 ± 0.227	6.851	-183.287 ± 0.196	-0.422
CREBBP	III	167.545 ± 0.227	6.844	-184.771 ± 0.087	-0.440
EP300	III	167.545 ± 0.227	6.937	-183.897 ± 0.147	-0.382
BRD1	IV	167.545 ± 0.227	6.918	-184.916 ± 0.100	-0.368
BRD9	IV	167.545 ± 0.227	6.878	-186.011 ± 0.097	-0.446
BRPF1B	IV	167.545 ± 0.227	6.892	-185.495 ± 0.113	-0.423
BAZ2A	V	167.545 ± 0.227	6.829	-184.998 ± 0.129	-0.414
TIF1	V	167.545 ± 0.227	6.760	-179.415 ± 0.153	-0.269
TAF1(1)	VII	167.545 ± 0.227	6.832	-182.439 ± 0.147	-0.437
TAF1(2)	VII	167.545 ± 0.227	6.577	-184.492 ± 0.104	-0.286
TAF1L(2)	VII	167.545 ± 0.227	6.829	-184.453 ± 0.065	-0.402
PB1(5)	VIII	167.545 ± 0.227	6.738	-180.942 ± 0.101	-0.338
SMARCA4	VIII	167.545 ± 0.227	6.754	-180.654 ± 0.117	-0.365

Table B.4: Optimised torsional parameters for bromosporine. f is the factor by which the torsional barrier is divided; k_ϕ is the force constant in kcal/mol; ϕ_0 is the phase in degrees; n is the periodicity. As per the atom types definitions in GAFF [259]: "ca" corresponds to a sp^2 aromatic carbon, "cc" and "cd" to an sp^2 carbon in a conjugated ring system, "n" to an amide nitrogen, "nd" to an sp^2 nitrogen in a conjugated ring system, "c3" to an sp^3 carbon, "s6" to a hypervalent sulfur with four substituents, "nh" to an amine nitrogen connected to an aromatic ring, "hn" to a hydrogen on a nitrogen, "o" to an sp^2 oxygen in a ketone or carboxylate group, "c" to an sp^2 carbon in a ketone or thioketone group, and "os" to an sp^3 oxygen in an ether or ester group.

Chemical Group	Dihedral	f	k_ϕ (kcal/mol)	ϕ_0 (°)	n
Sulfonamide	c3-s6-nh-ca	1	0.1126	115.8277	4
	c3-s6-nh-ca	1	0.5594	15.0293	3
	c3-s6-nh-ca	1	1.5296	5.9420	2
	c3-s6-nh-ca	1	−1.1272	−20.9470	1
	ca-ca-nh-s6	1	0.0772	237.9864	4
	ca-ca-nh-s6	1	−1.6048	207.5177	3
	ca-ca-nh-s6	1	−1.1204	−11.8563	2
	ca-ca-nh-s6	1	2.7114	−44.7267	1
	hn-nh-s6-o	1	−1.8576	28.7191	4
	hn-nh-s6-o	1	0.1574	62.1178	3
	hn-nh-s6-o	1	1.0255	175.3778	2
	hn-nh-s6-o	1	1.5400	4.1667	1
Carbamate	c3-c3-os-c	1	−0.4381	16.3277	3
	c3-c3-os-c	1	−0.0294	−160.9044	2
	c3-c3-os-c	1	0.3266	−28.5958	1
	o -c -os-c3	1	−0.2817	349.4914	3
	o -c -os-c3	1	−1.7625	−1.3486	2
	o -c -os-c3	1	−1.4545	184.3018	1
	os-c -n -hn	1	−4.7008	−28.8849	2
	os-c -n -hn	1	−1.2283	221.7737	1
	o -c -n -hn	1	2.8416	227.0899	2
	o -c -n -hn	1	−3.8285	188.1523	1
	cd-n -c -os	1	3.5869	179.3249	2
	cd-n -c -os	1	7.6446	−33.4048	1
	cd-n -c -o	1	−0.3087	179.2737	2
	cd-n -c -o	1	11.5708	−30.4522	1
	cd-cd-n -hn	1	−2.6948	13.3493	2
	cd-cd-n -hn	1	4.4903	105.3571	1
	cc-cd-n -hn	1	1.2439	139.5067	2
	cc-cd-n -hn	1	7.4016	80.0095	1
	cd-cd-n -c	1	0.8501	−70.9776	2
	cd-cd-n -c	1	7.0133	10.7138	1
	cc-cd-n -c	1	2.2964	−179.1856	2
	cc-cd-n -c	1	−8.3309	215.3574	1
Byaryl	ca-ca-cc-cc	1	0.8874	−1.8464	4
	ca-ca-cc-cc	1	−9.0752	206.9416	3
	ca-ca-cc-cc	1	−3.2711	−11.0960	2
	ca-ca-cc-cc	1	11.3523	−153.6517	1
	ca-ca-cc-nd	1	0.5543	190.6325	4
	ca-ca-cc-nd	1	−5.2506	221.8151	3
	ca-ca-cc-nd	1	−0.9219	176.2785	2
	ca-ca-cc-nd	1	5.3797	189.7465	1

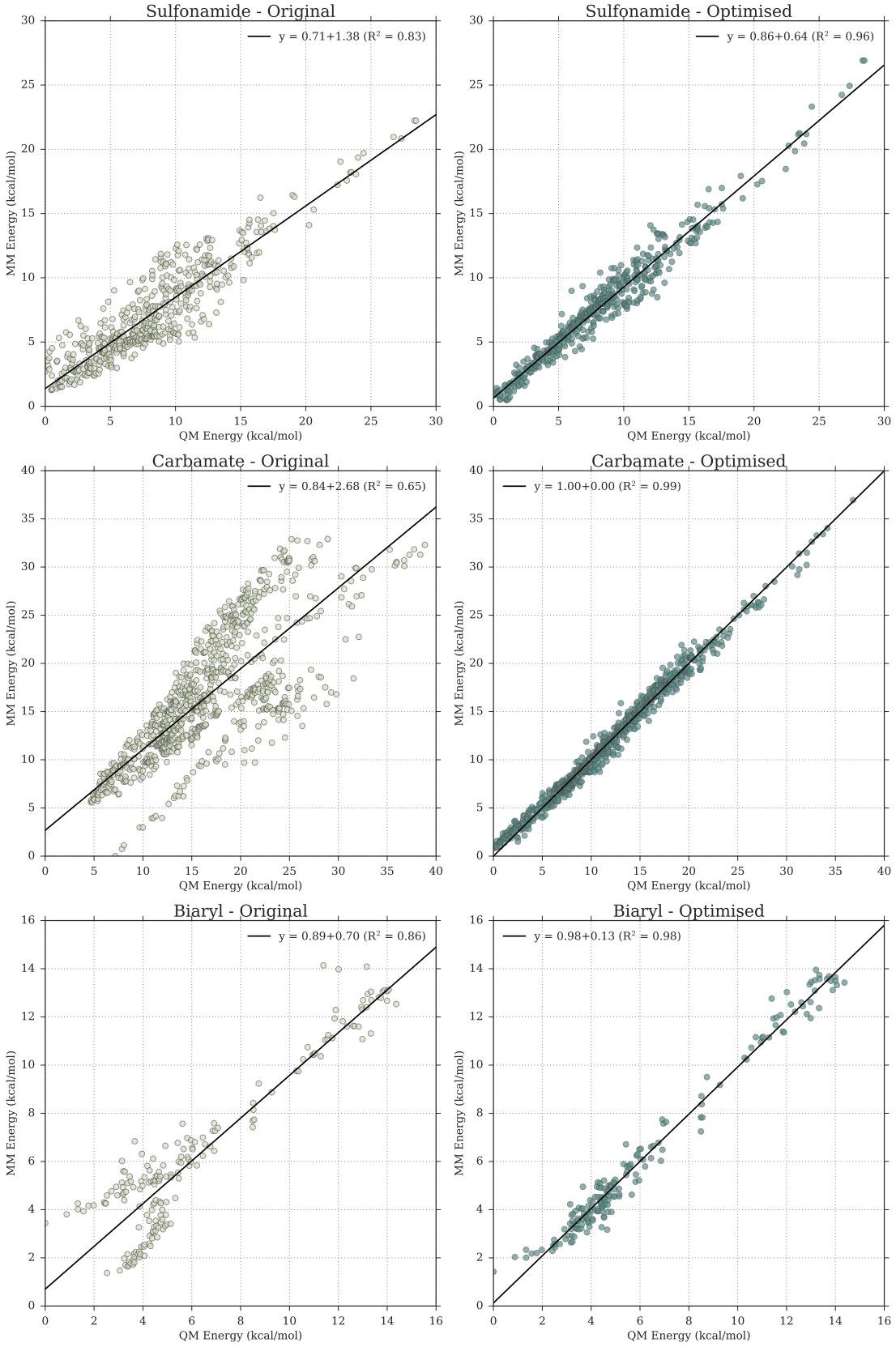


Figure B.1: Quality of fit between MM and QM energies for the optimised bromosporine groups.

C

Supplementary material for Chapter 5

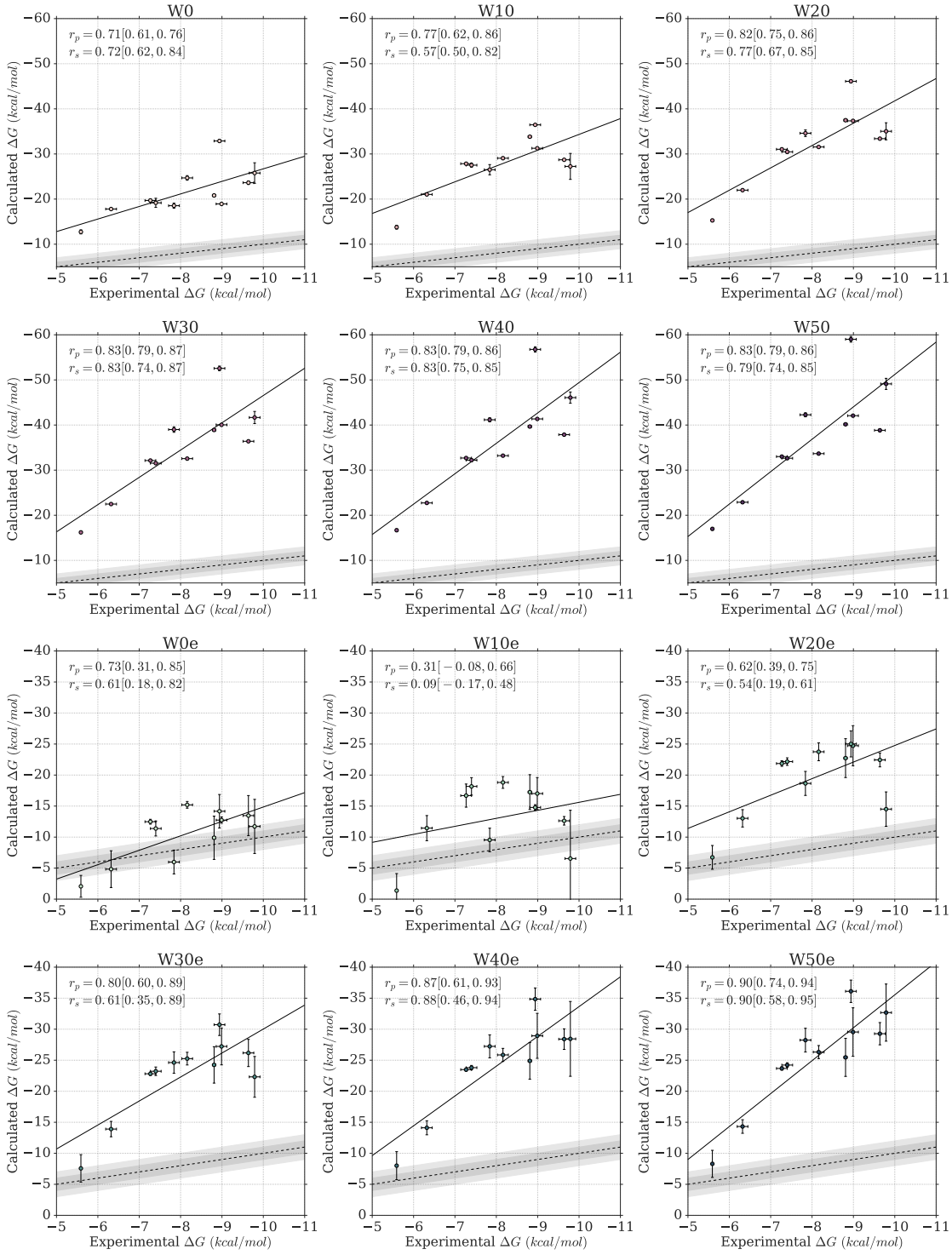


Figure C.1: Scatter plot of experimental versus calculated affinities for Case 1. Shown are the predicted binding free energies with MMPBSA, with and without the inclusion of explicit waters (W0 *vs* W10-W50), and with and without the inclusion of the entropic term (e.g. W0 *vs* W0e). A dashed line represent the diagonal $y = x$, that is the perfect agreement between predictions and experiment, and the shaded areas the 1 and 2 kcal/mol errors respectively. On each individual plot, the Pearson and Spearman coefficients are reported along with their 95% CIs.

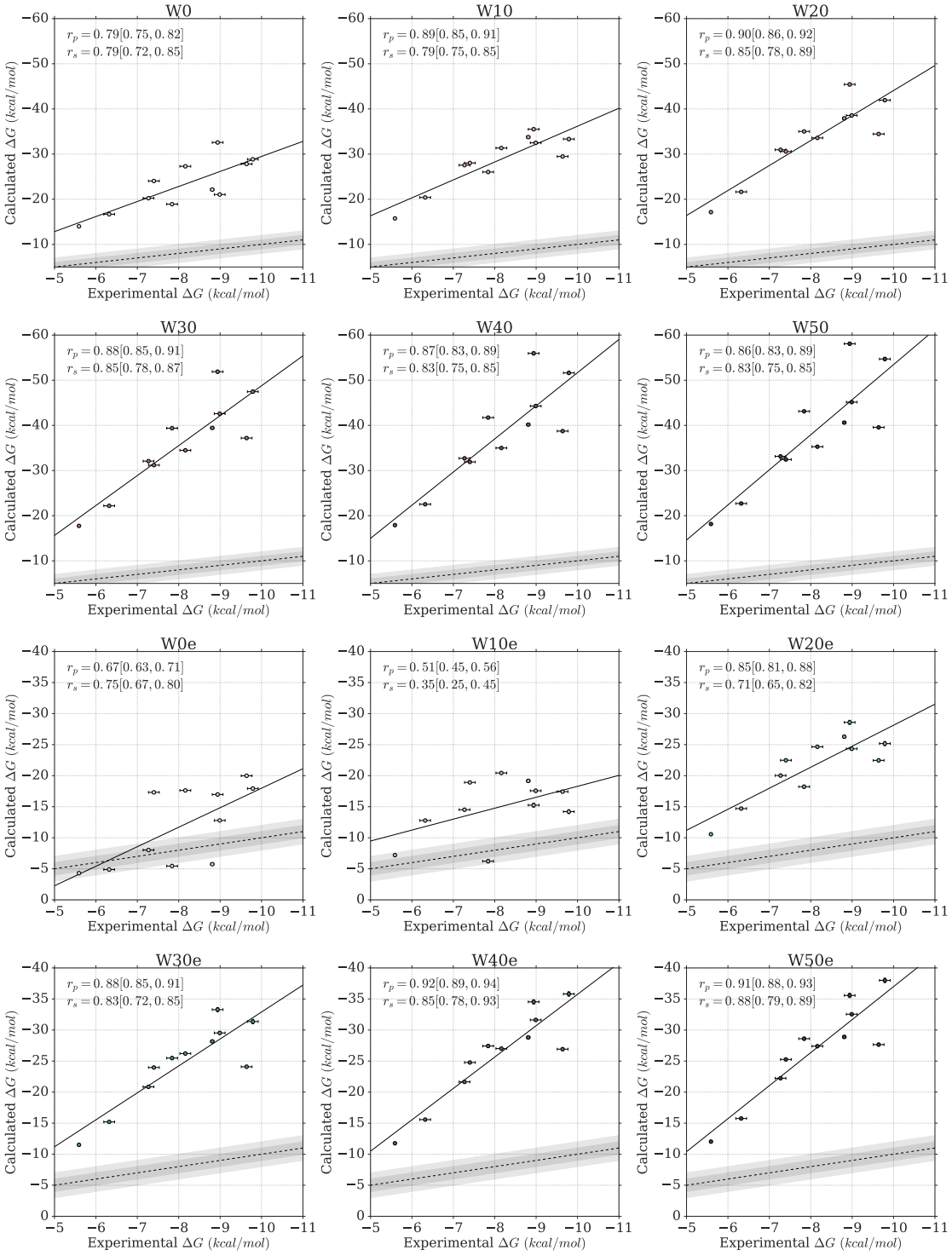


Figure C.2: Scatter plot of experimental versus calculated affinities for Case 2. Shown are the predicted binding free energies with MMPBSA, with and without the inclusion of explicit waters (W0 *vs* W10-W50), and with and without the inclusion of the entropic term (e.g. W0 *vs* W0e). A dashed line represent the diagonal $y = x$, that is the perfect agreement between predictions and experiment, and the shaded areas the 1 and 2 kcal/mol errors respectively. On each individual plot, the Pearson and Spearman coefficients are reported along with their 95% CIs.

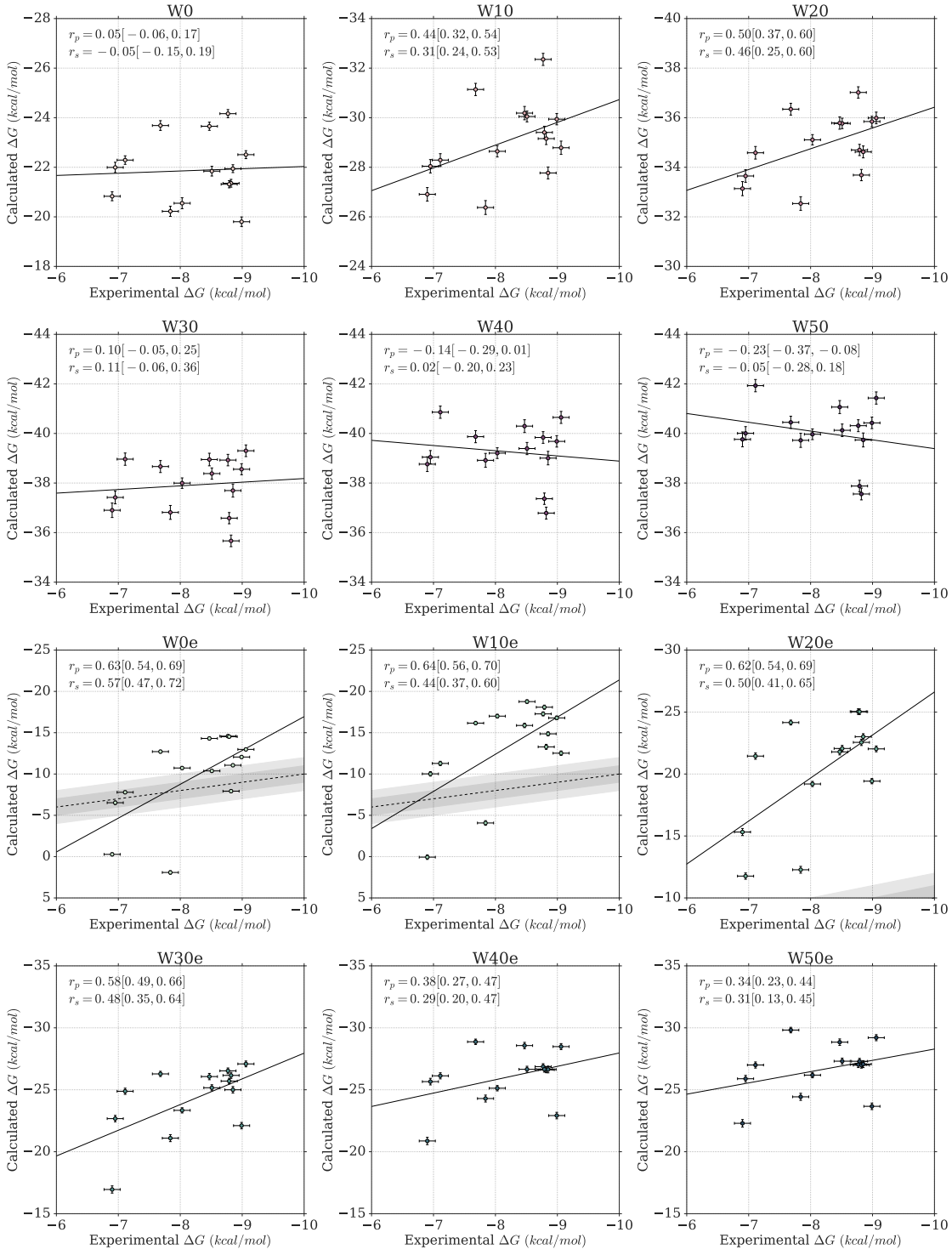


Figure C.3: Scatter plot of experimental versus calculated affinities for Case 3. Shown are the predicted binding free energies with MMPBSA, with and without the inclusion of explicit waters (W0 *vs* W10-W50), and with and without the inclusion of the entropic term (e.g. W0 *vs* W0e). A dashed line represent the diagonal $y = x$, that is the perfect agreement between predictions and experiment, and the shaded areas the 1 and 2 kcal/mol errors respectively. On each individual plot, the Pearson and Spearman coefficients are reported along with their 95% CIs.

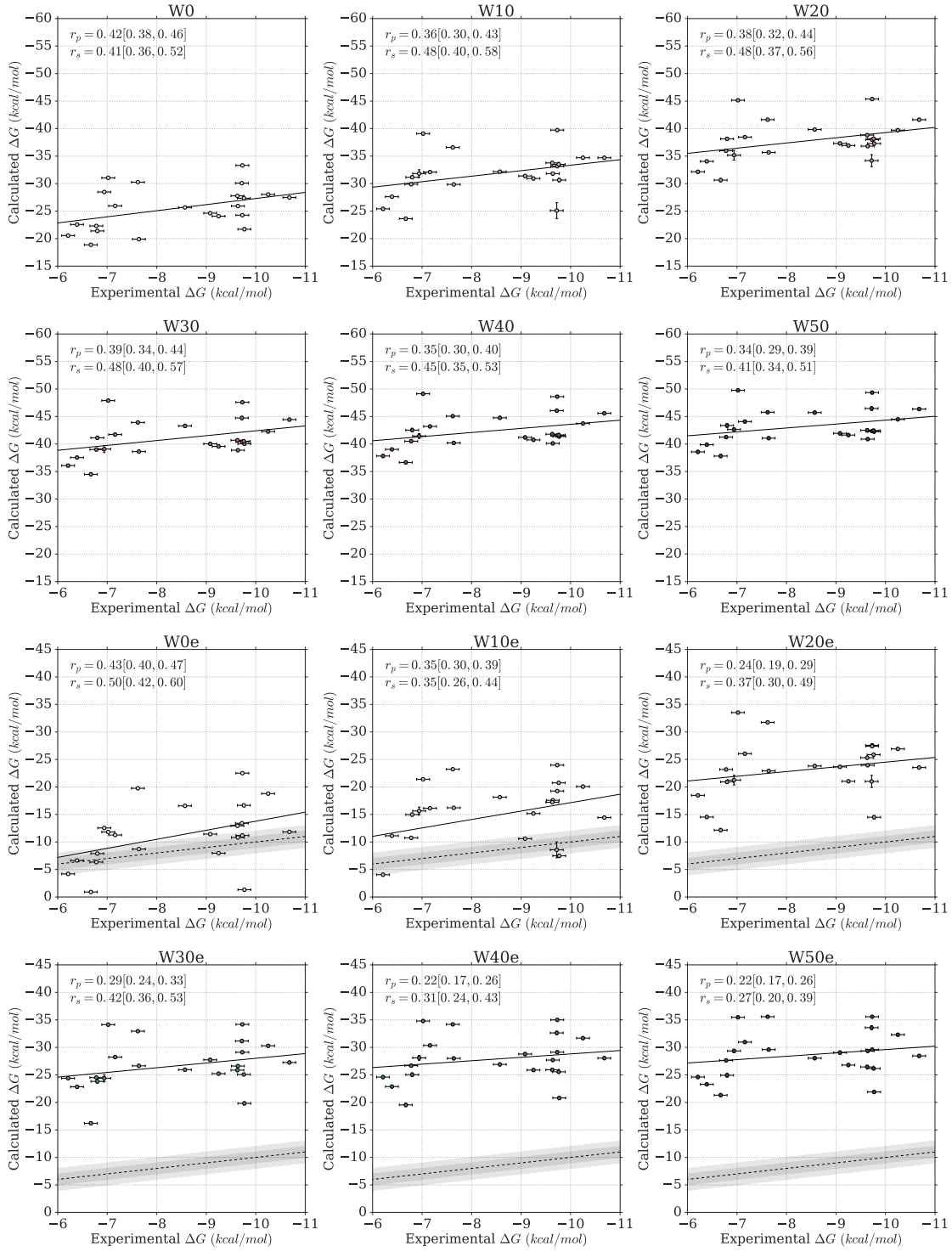


Figure C.4: Scatter plot of experimental versus calculated affinities for Case 4. Shown are the predicted binding free energies with MMPBSA, with and without the inclusion of explicit waters (W0 *vs* W10-W50), and with and without the inclusion of the entropic term (e.g. W0 *vs* W0e). A dashed line represent the diagonal $y = x$, that is the perfect agreement between predictions and experiment, and the shaded areas the 1 and 2 kcal/mol errors respectively. On each individual plot, the Pearson and Spearman coefficients are reported along with their 95% CIs.

Table C.1: All free energy results for Case 1.

Ligand (ID)	ITC	ABFE	W0	W10	W20	W30	W40	W50	W0e	W10e	W20e	W30e	W40e	W50e
Alprazolam (8)	-7.4 ± 0.1	-7.8 ± 0.3	-19.2 ± 1.0	-27.5 ± 0.5	-30.5 ± 0.4	-31.5 ± 0.3	-32.2 ± 0.3	-32.6 ± 0.2	-11.4 ± 1.2	-18.2 ± 1.4	-22.2 ± 0.6	-23.2 ± 0.7	-23.8 ± 0.4	-24.2 ± 0.4
BI-2356 (1)	-9.8 ± 0.1	-10.4 ± 0.6	-25.8 ± 2.3	-27.2 ± 2.9	-35.0 ± 1.9	-41.7 ± 1.3	-46.1 ± 1.2	-49.1 ± 1.2	-11.7 ± 4.4	-6.5 ± 7.8	-14.5 ± 2.8	-22.3 ± 3.3	-28.4 ± 6.0	-32.7 ± 4.6
BzT-7 (6)	-8.2 ± 0.1	-9.9 ± 0.8	-24.7 ± 0.5	-29.1 ± 0.3	-31.6 ± 0.2	-32.6 ± 0.2	-33.2 ± 0.2	-33.7 ± 0.2	-15.2 ± 0.5	-18.8 ± 0.9	-23.8 ± 1.4	-25.3 ± 1.0	-25.9 ± 1.1	-26.3 ± 1.1
Isox-1 (9)	-7.3 ± 0.0	-7.7 ± 0.4	-19.7 ± 0.3	-27.8 ± 0.2	-31.0 ± 0.2	-32.1 ± 0.2	-32.7 ± 0.2	-33.0 ± 0.2	-12.5 ± 0.4	-16.7 ± 1.9	-21.9 ± 0.4	-22.8 ± 0.3	-23.5 ± 0.3	-23.7 ± 0.3
Isox-2 (5)	-8.8 ± 0.1	-8.6 ± 0.3	-20.8 ± 0.2	-33.8 ± 0.2	-37.5 ± 0.2	-38.9 ± 0.2	-39.7 ± 0.1	-40.2 ± 0.1	-9.9 ± 3.5	-17.2 ± 2.8	-22.7 ± 3.1	-24.2 ± 2.9	-24.9 ± 2.9	-25.5 ± 3.1
Isox-cf3 (11)	-5.6 ± 0.0	-5.4 ± 0.2	-12.7 ± 0.4	-13.8 ± 0.4	-15.3 ± 0.3	-16.2 ± 0.2	-16.7 ± 0.3	-17.0 ± 0.3	-2.2 ± 1.6	-1.6 ± 2.8	-6.7 ± 1.9	-7.7 ± 2.1	-8.1 ± 2.2	-8.4 ± 2.1
JQ1 (2)	-9.6 ± 0.1	-9.5 ± 0.4	-23.6 ± 0.4	-28.7 ± 0.0	-33.4 ± 0.2	-36.4 ± 0.2	-37.9 ± 0.3	-38.8 ± 0.3	-13.5 ± 3.2	-12.6 ± 0.7	-22.4 ± 1.1	-26.2 ± 2.2	-28.4 ± 1.7	-29.3 ± 1.8
Quinaz (10)	-6.3 ± 0.1	-5.9 ± 0.2	-17.8 ± 0.2	-21.0 ± 0.2	-22.0 ± 0.2	-22.5 ± 0.2	-22.7 ± 0.2	-22.9 ± 0.2	-4.8 ± 3.0	-11.4 ± 2.0	-13.0 ± 1.4	-13.9 ± 1.2	-14.1 ± 1.1	-14.3 ± 1.1
RVX-208 (7)	-7.8 ± 0.1	-5.9 ± 0.5	-18.5 ± 0.6	-26.5 ± 1.1	-34.6 ± 0.7	-39.0 ± 0.6	-41.2 ± 0.5	-42.3 ± 0.4	-6.0 ± 1.9	-9.5 ± 1.9	-18.7 ± 1.9	-24.6 ± 1.7	-27.2 ± 1.9	-28.2 ± 1.9
RVX-OH (3)	-9.0 ± 0.1	-9.2 ± 0.5	-18.9 ± 0.1	-31.3 ± 0.2	-37.3 ± 0.1	-40.0 ± 0.1	-41.4 ± 0.1	-42.1 ± 0.1	-12.7 ± 0.5	-17.0 ± 2.6	-24.7 ± 3.2	-27.2 ± 3.0	-28.9 ± 3.6	-29.5 ± 3.9
TG-101348 (4)	-8.9 ± 0.1	-9.4 ± 0.8	-32.9 ± 0.3	-36.5 ± 0.2	-46.1 ± 0.3	-52.6 ± 0.5	-56.8 ± 0.6	-59.0 ± 0.6	-14.2 ± 2.7	-14.8 ± 0.5	-25.0 ± 2.1	-30.7 ± 1.7	-34.9 ± 1.8	-36.1 ± 1.8

Table C.2: All free energy results for Case 2.

Ligand (ID)	ITC	ABFE	W0	W10	W20	W30	W40	W50	W0e	W10e	W20e	W30e	W40e	W50e
Alprazolam (8)	-7.4 ± 0.1	-10.2 ± 0.2	-24.0 ± 0.1	-28.0 ± 0.2	-30.6 ± 0.2	-31.2 ± 0.2	-31.9 ± 0.2	-32.4 ± 0.2	-17.3 ± 0.1	-18.9 ± 0.2	-22.5 ± 0.2	-24.0 ± 0.2	-24.8 ± 0.2	-25.3 ± 0.2
BI-2356 (1)	-9.8 ± 0.1	-10.9 ± 0.8	-28.8 ± 0.2	-33.3 ± 0.3	-41.9 ± 0.3	-47.5 ± 0.3	-51.6 ± 0.4	-54.7 ± 0.4	-17.9 ± 0.2	-14.2 ± 0.3	-25.2 ± 0.3	-31.4 ± 0.3	-35.8 ± 0.4	-38.0 ± 0.4
BzT-7 (6)	-8.2 ± 0.1	-10.6 ± 0.3	-27.3 ± 0.2	-31.3 ± 0.2	-33.6 ± 0.2	-34.5 ± 0.2	-35.0 ± 0.2	-35.3 ± 0.2	-17.6 ± 0.2	-20.4 ± 0.2	-24.7 ± 0.2	-26.2 ± 0.2	-27.0 ± 0.2	-27.4 ± 0.2
Isox-1 (9)	-7.3 ± 0.0	-7.7 ± 0.1	-20.2 ± 0.2	-27.5 ± 0.2	-31.0 ± 0.2	-32.1 ± 0.2	-32.7 ± 0.2	-33.1 ± 0.2	-8.0 ± 0.2	-14.5 ± 0.2	-20.0 ± 0.2	-20.9 ± 0.2	-21.6 ± 0.2	-22.2 ± 0.2
Isox-2 (5)	-8.8 ± 0.1	-8.3 ± 0.2	-22.1 ± 0.2	-33.8 ± 0.2	-37.9 ± 0.2	-39.4 ± 0.2	-40.2 ± 0.2	-40.6 ± 0.2	-5.7 ± 0.2	-19.2 ± 0.2	-26.3 ± 0.2	-28.2 ± 0.2	-28.8 ± 0.2	-28.9 ± 0.2
Isox-cf3 (11)	-5.6 ± 0.0	-5.4 ± 0.1	-14.0 ± 0.2	-15.8 ± 0.2	-17.1 ± 0.2	-17.7 ± 0.2	-17.9 ± 0.2	-18.2 ± 0.2	-4.3 ± 0.1	-7.2 ± 0.2	-10.6 ± 0.2	-11.5 ± 0.2	-11.8 ± 0.2	-12.0 ± 0.2
JQ1 (2)	-9.6 ± 0.1	-10.1 ± 0.4	-27.8 ± 0.2	-29.5 ± 0.2	-34.5 ± 0.2	-37.2 ± 0.2	-38.7 ± 0.2	-39.6 ± 0.2	-20.0 ± 0.2	-17.4 ± 0.2	-22.5 ± 0.2	-24.1 ± 0.2	-26.9 ± 0.2	-27.6 ± 0.2
Quinaz (10)	-6.3 ± 0.1	-6.2 ± 0.1	-16.7 ± 0.2	-20.4 ± 0.2	-21.6 ± 0.2	-22.2 ± 0.2	-22.5 ± 0.2	-22.7 ± 0.2	-4.9 ± 0.2	-12.8 ± 0.2	-14.7 ± 0.2	-15.2 ± 0.2	-15.6 ± 0.2	-15.8 ± 0.2
RVX-208 (7)	-7.8 ± 0.1	-6.6 ± 0.3	-18.9 ± 0.1	-26.0 ± 0.3	-35.0 ± 0.2	-39.4 ± 0.2	-41.7 ± 0.2	-43.1 ± 0.2	-5.4 ± 0.1	-6.2 ± 0.3	-18.2 ± 0.2	-25.5 ± 0.2	-27.4 ± 0.2	-28.6 ± 0.2
RVX-OH (3)	-9.0 ± 0.1	-10.8 ± 0.2	-21.0 ± 0.2	-32.5 ± 0.2	-38.6 ± 0.3	-42.6 ± 0.3	-44.3 ± 0.3	-45.1 ± 0.3	-12.8 ± 0.2	-17.6 ± 0.2	-24.3 ± 0.2	-29.5 ± 0.2	-31.6 ± 0.2	-32.5 ± 0.2
TG-101348 (4)	-8.9 ± 0.1	-9.0 ± 0.8	-32.6 ± 0.2	-35.5 ± 0.3	-45.4 ± 0.3	-51.9 ± 0.3	-55.9 ± 0.3	-58.1 ± 0.3	-17.0 ± 0.2	-15.3 ± 0.3	-28.6 ± 0.3	-33.3 ± 0.3	-34.5 ± 0.3	-35.6 ± 0.3

Table C.3: All free energy results for Case 3.

Protein	Ligand	ITC	ABFE	W0	W10	W20	W30	W40	W50	W0e	W10e	W20e	W30e	W40e	W50e
BRD2(1)	RVX-OH	-8.5 ± 0.1	-9.8 ± 0.1	-21.8 ± 0.2	-30.0 ± 0.2	-35.8 ± 0.2	-38.4 ± 0.2	-39.4 ± 0.2	-40.1 ± 0.3	-10.4 ± 0.2	-18.8 ± 0.2	-22.1 ± 0.2	-25.2 ± 0.2	-26.7 ± 0.2	-27.3 ± 0.2
BRD2(2)	RVX-OH	-8.8 ± 0.1	-7.9 ± 0.1	-24.2 ± 0.2	-32.4 ± 0.2	-37.0 ± 0.2	-38.9 ± 0.2	-39.8 ± 0.2	-40.3 ± 0.2	-14.6 ± 0.2	-17.3 ± 0.2	-25.0 ± 0.2	-26.5 ± 0.2	-26.9 ± 0.2	-27.0 ± 0.2
BRD3(1)	RVX-OH	-8.0 ± 0.1	-6.7 ± 0.1	-20.5 ± 0.2	-28.6 ± 0.2	-35.1 ± 0.2	-38.0 ± 0.2	-39.2 ± 0.2	-40.0 ± 0.2	-10.7 ± 0.2	-17.0 ± 0.2	-19.2 ± 0.2	-23.3 ± 0.2	-25.1 ± 0.2	-26.2 ± 0.2
BRD3(2)	RVX-OH	-8.8 ± 0.1	-7.4 ± 0.1	-21.4 ± 0.2	-29.2 ± 0.2	-33.7 ± 0.2	-35.7 ± 0.2	-36.8 ± 0.2	-37.6 ± 0.2	-7.9 ± 0.2	-13.3 ± 0.3	-22.6 ± 0.3	-26.2 ± 0.2	-26.6 ± 0.2	-27.0 ± 0.2
BRD4(1)	RVX-OH	-9.0 ± 0.1	-9.9 ± 0.1	-19.8 ± 0.2	-29.9 ± 0.2	-35.8 ± 0.2	-38.6 ± 0.2	-39.7 ± 0.2	-40.4 ± 0.2	-12.1 ± 0.2	-16.8 ± 0.2	-19.4 ± 0.2	-22.1 ± 0.3	-22.9 ± 0.3	-23.7 ± 0.3
BRD4(2)	RVX-OH	-8.8 ± 0.1	-8.5 ± 0.1	-21.3 ± 0.1	-29.4 ± 0.2	-34.7 ± 0.2	-36.6 ± 0.2	-37.4 ± 0.2	-37.9 ± 0.2	-14.5 ± 0.1	-18.1 ± 0.2	-25.0 ± 0.2	-25.7 ± 0.2	-26.7 ± 0.2	-27.3 ± 0.2
BRDT(1)	RVX-OH	-7.7 ± 0.1	-7.7 ± 0.1	-23.7 ± 0.2	-31.1 ± 0.2	-36.3 ± 0.2	-38.7 ± 0.2	-39.9 ± 0.2	-40.4 ± 0.2	-12.7 ± 0.2	-16.2 ± 0.2	-24.2 ± 0.2	-26.3 ± 0.2	-28.9 ± 0.2	-29.8 ± 0.2
BRD2(1)	RVX-208	-6.9 ± 0.1	-7.2 ± 0.1	-20.8 ± 0.2	-26.9 ± 0.3	-33.1 ± 0.3	-36.9 ± 0.3	-38.8 ± 0.3	-39.8 ± 0.3	-0.3 ± 0.2	0.1 ± 0.3	-15.3 ± 0.3	-17.0 ± 0.3	-20.9 ± 0.3	-22.3 ± 0.3
BRD2(2)	RVX-208	-8.5 ± 0.1	-7.5 ± 0.1	-23.7 ± 0.2	-30.2 ± 0.3	-35.8 ± 0.2	-38.9 ± 0.3	-40.3 ± 0.3	-41.1 ± 0.3	-14.3 ± 0.2	-15.9 ± 0.3	-21.8 ± 0.2	-26.1 ± 0.3	-28.6 ± 0.3	-28.8 ± 0.3
BRD3(1)	RVX-208	-7.1 ± 0.1	-5.5 ± 0.1	-22.3 ± 0.2	-28.3 ± 0.3	-34.6 ± 0.3	-39.0 ± 0.2	-40.9 ± 0.2	-41.9 ± 0.2	-7.8 ± 0.2	-11.3 ± 0.3	-21.4 ± 0.3	-24.9 ± 0.2	-26.1 ± 0.2	-27.0 ± 0.2
BRD3(2)	RVX-208	-8.8 ± 0.1	-8.7 ± 0.1	-21.9 ± 0.2	-27.8 ± 0.2	-34.6 ± 0.2	-37.7 ± 0.3	-39.0 ± 0.3	-39.7 ± 0.3	-11.1 ± 0.2	-14.9 ± 0.2	-23.0 ± 0.2	-25.0 ± 0.3	-26.6 ± 0.3	-27.1 ± 0.3
BRD4(1)	RVX-208	-7.8 ± 0.1	-6.8 ± 0.1	-20.2 ± 0.2	-26.4 ± 0.3	-32.5 ± 0.3	-36.8 ± 0.3	-38.9 ± 0.3	-39.7 ± 0.3	1.9 ± 0.2	-4.1 ± 0.3	-12.3 ± 0.3	-21.1 ± 0.3	-24.3 ± 0.3	-24.4 ± 0.3
BRD4(2)	RVX-208	-9.1 ± 0.1	-9.7 ± 0.1	-22.5 ± 0.2	-28.8 ± 0.3	-36.0 ± 0.2	-39.3 ± 0.2	-40.6 ± 0.2	-41.4 ± 0.2	-13.0 ± 0.2	-12.5 ± 0.3	-22.0 ± 0.2	-27.1 ± 0.2	-28.5 ± 0.2	-29.2 ± 0.2
BRDT(1)	RVX-208	-7.0 ± 0.1	-6.5 ± 0.1	-22.0 ± 0.2	-28.0 ± 0.3	-33.6 ± 0.3	-37.4 ± 0.3	-39.0 ± 0.3	-40.0 ± 0.3	-6.5 ± 0.2	-10.0 ± 0.3	-11.8 ± 0.3	-22.7 ± 0.3	-25.7 ± 0.3	-25.9 ± 0.3

Table C.4: All free energy results for Case 4.

Protein	ITC	ABFE	W0	W10	W20	W30	W40	W50	W0e	W10e	W20e	W30e	W40e	W50e
BAZ2A	-7.2 ± 0.1	-11.0 ± 0.3	-26.0 ± 0.2	-32.1 ± 0.2	-38.5 ± 0.2	-41.7 ± 0.2	-43.2 ± 0.2	-44.1 ± 0.3	-11.3 ± 0.2	-16.1 ± 0.2	-26.0 ± 0.2	-28.2 ± 0.2	-30.4 ± 0.2	-31.0 ± 0.3
BRD1	-7.6 ± 0.1	-10.8 ± 0.2	-30.3 ± 0.2	-36.6 ± 0.2	-41.6 ± 0.2	-43.9 ± 0.2	-45.1 ± 0.2	-45.8 ± 0.2	-19.8 ± 0.2	-23.2 ± 0.2	-31.7 ± 0.2	-32.9 ± 0.2	-34.2 ± 0.2	-35.6 ± 0.2
BRD2(1)	-9.2 ± 0.1	-7.7 ± 0.3	-24.1 ± 0.2	-30.9 ± 0.2	-36.9 ± 0.2	-39.6 ± 0.3	-40.8 ± 0.3	-41.6 ± 0.3	-8.0 ± 0.2	-15.2 ± 0.2	-21.0 ± 0.2	-25.2 ± 0.3	-25.9 ± 0.3	-26.8 ± 0.3
BRD2(2)	-9.6 ± 0.1	-9.7 ± 0.3	-25.9 ± 0.2	-31.8 ± 0.3	-36.8 ± 0.2	-38.9 ± 0.2	-40.1 ± 0.2	-40.9 ± 0.2	-10.8 ± 0.2	-17.6 ± 0.3	-23.9 ± 0.2	-26.6 ± 0.2	-27.7 ± 0.2	-29.3 ± 0.2
BRD3(1)	-9.1 ± 0.1	-7.2 ± 0.3	-24.6 ± 0.2	-31.4 ± 0.2	-37.3 ± 0.2	-40.0 ± 0.3	-41.2 ± 0.3	-41.9 ± 0.3	-11.4 ± 0.2	-10.6 ± 0.2	-23.6 ± 0.2	-27.7 ± 0.3	-28.8 ± 0.3	-29.1 ± 0.3
BRD3(2)	-9.6 ± 0.1	-11.7 ± 0.2	-27.8 ± 0.2	-33.8 ± 0.3	-38.8 ± 0.3	-40.7 ± 0.3	-41.8 ± 0.3	-42.5 ± 0.3	-12.9 ± 0.2	-17.2 ± 0.3	-25.3 ± 0.3	-25.9 ± 0.3	-25.9 ± 0.3	-26.4 ± 0.3
BRD4(1)	-9.7 ± 0.1	-11.3 ± 0.3	-24.3 ± 0.3	-33.2 ± 0.2	-38.0 ± 0.2	-40.4 ± 0.2	-41.5 ± 0.2	-42.4 ± 0.2	-11.2 ± 0.3	-19.3 ± 0.2	-27.6 ± 0.2	-29.1 ± 0.2	-29.1 ± 0.2	-29.6 ± 0.2
BRD4(2)	-9.8 ± 0.1	-10.0 ± 0.3	-21.7 ± 0.3	-30.6 ± 0.3	-37.3 ± 0.3	-40.0 ± 0.3	-41.4 ± 0.3	-42.3 ± 0.3	-1.3 ± 0.3	-7.5 ± 0.3	-14.5 ± 0.3	-19.8 ± 0.3	-20.8 ± 0.3	-21.9 ± 0.3
BRD9	-9.7 ± 0.1	-12.0 ± 0.2	-33.3 ± 0.2	-39.7 ± 0.2	-45.4 ± 0.2	-47.6 ± 0.2	-48.6 ± 0.2	-49.3 ± 0.2	-22.5 ± 0.2	-24.0 ± 0.2	-27.4 ± 0.2	-34.2 ± 0.2	-35.0 ± 0.2	-35.6 ± 0.2
BRDT(1)	-9.8 ± 0.1	-9.3 ± 0.3	-27.3 ± 0.2	-33.5 ± 0.2	-38.2 ± 0.2	-40.5 ± 0.2	-41.6 ± 0.3	-42.4 ± 0.3	-16.7 ± 0.2	-20.7 ± 0.2	-25.9 ± 0.2	-25.1 ± 0.2	-25.6 ± 0.3	-26.2 ± 0.3
BRPF1B	-8.6 ± 0.1	-11.5 ± 0.3	-25.7 ± 0.2	-32.2 ± 0.2	-39.8 ± 0.3	-43.3 ± 0.3	-44.8 ± 0.3	-45.7 ± 0.3	-16.6 ± 0.2	-18.1 ± 0.2	-23.8 ± 0.3	-25.9 ± 0.3	-26.9 ± 0.3	-28.0 ± 0.3
CECR2	-10.7 ± 0.1	-11.5 ± 0.3	-27.5 ± 0.2	-34.7 ± 0.2	-41.6 ± 0.3	-44.4 ± 0.3	-45.6 ± 0.3	-46.4 ± 0.3	-11.8 ± 0.2	-14.4 ± 0.2	-23.5 ± 0.3	-27.3 ± 0.3	-28.0 ± 0.3	-28.5 ± 0.3
CREBBP	-7.6 ± 0.1	-10.8 ± 0.2	-19.9 ± 0.2	-29.9 ± 0.2	-35.7 ± 0.2	-38.6 ± 0.2	-40.2 ± 0.2	-41.1 ± 0.2	-8.7 ± 0.2	-16.2 ± 0.2	-22.9 ± 0.2	-26.6 ± 0.2	-28.0 ± 0.2	-29.6 ± 0.2
EP300	-6.8 ± 0.1	-9.8 ± 0.3	-22.3 ± 0.2	-29.9 ± 0.3	-36.0 ± 0.2	-39.0 ± 0.2	-40.5 ± 0.2	-41.2 ± 0.2	-6.3 ± 0.2	-10.8 ± 0.3	-23.2 ± 0.2	-24.5 ± 0.2	-26.7 ± 0.2	-27.6 ± 0.2
FALZ	-6.8 ± 0.1	-9.8 ± 0.2	-21.4 ± 0.2	-31.2 ± 0.3	-38.1 ± 0.3	-41.1 ± 0.3	-42.5 ± 0.3	-43.4 ± 0.3	-7.9 ± 0.2	-15.0 ± 0.3	-20.9 ± 0.3	-23.8 ± 0.3	-25.0 ± 0.3	-24.9 ± 0.3
PB1(5)	-6.4 ± 0.1	-7.0 ± 0.2	-22.6 ± 0.2	-27.6 ± 0.2	-34.0 ± 0.2	-37.5 ± 0.2	-39.0 ± 0.2	-39.9 ± 0.2	-6.6 ± 0.2	-11.1 ± 0.2	-14.5 ± 0.2	-22.8 ± 0.2	-22.9 ± 0.2	-23.3 ± 0.2
PCAF	-7.0 ± 0.1	-10.7 ± 0.3	-31.1 ± 0.2	-39.1 ± 0.3	-45.2 ± 0.2	-47.9 ± 0.2	-49.1 ± 0.2	-49.8 ± 0.2	-11.8 ± 0.2	-21.4 ± 0.3	-33.5 ± 0.2	-34.1 ± 0.2	-34.8 ± 0.2	-35.5 ± 0.2
SMARCA4	-6.2 ± 0.1	-6.7 ± 0.3	-20.6 ± 0.2	-25.4 ± 0.3	-32.2 ± 0.3	-36.1 ± 0.3	-37.8 ± 0.3	-38.6 ± 0.3	-4.2 ± 0.2	-4.1 ± 0.3	-18.5 ± 0.3	-24.4 ± 0.3	-24.6 ± 0.3	-24.6 ± 0.3
TAF1(1)	-6.9 ± 0.1	-8.5 ± 0.3	-28.5 ± 0.2	-31.8 ± 0.7	-35.2 ± 0.9	-39.1 ± 0.6	-41.5 ± 0.4	-42.6 ± 0.3	-12.5 ± 0.2	-15.6 ± 0.7	-21.3 ± 0.9	-24.4 ± 0.6	-28.1 ± 0.4	-29.4 ± 0.3
TAF1(2)	-10.2 ± 0.1	-10.7 ± 0.2	-28.0 ± 0.2	-34.7 ± 0.3	-39.7 ± 0.3	-42.3 ± 0.3	-43.7 ± 0.3	-44.5 ± 0.3	-18.8 ± 0.2	-20.1 ± 0.3	-26.9 ± 0.3	-30.3 ± 0.3	-31.7 ± 0.3	-32.3 ± 0.3
TAF1L(2)	-9.7 ± 0.1	-10.5 ± 0.2	-30.1 ± 0.2	-25.1 ± 1.4	-34.2 ± 1.1	-44.7 ± 0.3	-46.1 ± 0.3	-46.5 ± 0.4	-13.4 ± 0.2	-8.6 ± 1.4	-21.0 ± 1.1	-31.2 ± 0.3	-32.6 ± 0.3	-33.6 ± 0.4
TIF1	-6.7 ± 0.1	-5.4 ± 0.3	-18.9 ± 0.2	-23.6 ± 0.3	-30.6 ± 0.3	-34.5 ± 0.3	-36.7 ± 0.3	-37.8 ± 0.3	-0.9 ± 0.2	3.1 ± 0.3	-12.2 ± 0.3	-16.2 ± 0.3	-19.5 ± 0.3	-21.3 ± 0.3

D

Supplementary material for Chapter 6

Table D.1: Structures used for the GCMC calculation and overall results. All free energies are in kcal/mol.

Protein	Family	PDB	ΔG_{netw}	ΔG_{W1}	ΔG_{W2}	ΔG_{W3}	ΔG_{W4}
CECR2	I	3NXB	-4.0 ± 1.0	-1.0 ± 0.3	-0.1 ± 0.3	-1.2 ± 0.3	-0.5 ± 0.6
FALZ	I	3UV2	-8.2 ± 0.7	4.7 ± 3.3	-1.4 ± 0.2	-1.0 ± 0.1	-0.9 ± 0.2
GCN5L2	I	3D7C	-5.1 ± 2.1	0.3 ± 0.3	-0.8 ± 0.2	-0.7 ± 0.2	-0.6 ± 0.3
PCAF	I	3GG3	-6.0 ± 0.8	2.4 ± 0.8	-1.7 ± 0.6	-0.5 ± 0.4	-0.8 ± 0.5
BRD2(1)	II	4ALG	-5.5 ± 1.0	-1.0 ± 0.3	-0.1 ± 0.3	-1.2 ± 0.3	-0.5 ± 0.6
BRD2(2)	II	5BT5	-7.1 ± 0.4	4.7 ± 3.3	-1.4 ± 0.2	-1.0 ± 0.1	-0.9 ± 0.2
BRD3(1)	II	2NXB	-6.7 ± 0.3	0.3 ± 0.3	-0.8 ± 0.2	-0.7 ± 0.2	-0.6 ± 0.3
BRD3(2)	II	3S92	-5.8 ± 1.4	2.4 ± 0.8	-1.7 ± 0.6	-0.5 ± 0.4	-0.8 ± 0.5
BRD4(1)	II	2OSS	-8.2 ± 1.4	-1.5 ± 0.2	-2.0 ± 0.1	-0.4 ± 0.3	-0.9 ± 0.3
BRD4(2)	II	2OUO	-7.7 ± 0.5	-0.7 ± 0.4	-0.7 ± 0.2	-1.2 ± 0.2	-1.1 ± 0.2
BRDT(1)	II	4KCX	-3.8 ± 0.6	-1.6 ± 0.2	-1.2 ± 0.2	-1.4 ± 0.1	-1.3 ± 0.1
CREBBP	III	4NYX	-6.0 ± 1.0	-1.0 ± 0.3	-0.1 ± 0.3	-1.2 ± 0.3	-0.5 ± 0.6
EP300	III	5BT3	-2.9 ± 1.1	4.7 ± 3.3	-1.4 ± 0.2	-1.0 ± 0.1	-0.9 ± 0.2
PHIP(2)	III	3MB3	-6.0 ± 1.2	0.3 ± 0.3	-0.8 ± 0.2	-0.7 ± 0.2	-0.6 ± 0.3
WDR9(2)	III	3Q2E	-2.6 ± 3.0	2.4 ± 0.8	-1.7 ± 0.6	-0.5 ± 0.4	-0.8 ± 0.5
ATAD2	IV	4QUT	-2.3 ± 2.6	-1.0 ± 0.3	-0.1 ± 0.3	-1.2 ± 0.3	-0.5 ± 0.6
BRD1	IV	5AME	-6.9 ± 0.9	4.7 ± 3.3	-1.4 ± 0.2	-1.0 ± 0.1	-0.9 ± 0.2
BRD9	IV	4XY8	-6.7 ± 0.2	0.3 ± 0.3	-0.8 ± 0.2	-0.7 ± 0.2	-0.6 ± 0.3
BRPF1B	IV	5C7N	-12.3 ± 1.2	2.4 ± 0.8	-1.7 ± 0.6	-0.5 ± 0.4	-0.8 ± 0.5
KIAA1240	IV	3LXJ	-2.9 ± 1.2	-1.5 ± 0.2	-2.0 ± 0.1	-0.4 ± 0.3	-0.9 ± 0.3
BAZ2A	V	4QBM	-3.9 ± 0.8	-1.0 ± 0.3	-0.1 ± 0.3	-1.2 ± 0.3	-0.5 ± 0.6
BAZ2B	V	4QC3	-2.6 ± 1.6	4.7 ± 3.3	-1.4 ± 0.2	-1.0 ± 0.1	-0.9 ± 0.2
TIF1α	V	4YBM	-1.9 ± 2.2	0.3 ± 0.3	-0.8 ± 0.2	-0.7 ± 0.2	-0.6 ± 0.3
PRKCBP1	VII	4COS	2.2 ± 1.0	-1.0 ± 0.3	-0.1 ± 0.3	-1.2 ± 0.3	-0.5 ± 0.6
TAF1L(2)	VII	3HMH	-6.1 ± 3.7	4.7 ± 3.3	-1.4 ± 0.2	-1.0 ± 0.1	-0.9 ± 0.2
TAF1(1)	VII	3UV5	-6.2 ± 1.6	0.3 ± 0.3	-0.8 ± 0.2	-0.7 ± 0.2	-0.6 ± 0.3
TAF1(2)	VII	3UV4	-7.8 ± 0.4	2.4 ± 0.8	-1.7 ± 0.6	-0.5 ± 0.4	-0.8 ± 0.5
ASH1L	VIII	3MQM	-4.5 ± 1.0	-1.0 ± 0.3	-0.1 ± 0.3	-1.2 ± 0.3	-0.5 ± 0.6
PB1(1)	VIII	3IU5	-1.4 ± 1.4	4.7 ± 3.3	-1.4 ± 0.2	-1.0 ± 0.1	-0.9 ± 0.2
PB1(2)	VIII	3LJW	-1.7 ± 1.1	0.3 ± 0.3	-0.8 ± 0.2	-0.7 ± 0.2	-0.6 ± 0.3
PB1(3)	VIII	3K2J	-2.5 ± 0.9	2.4 ± 0.8	-1.7 ± 0.6	-0.5 ± 0.4	-0.8 ± 0.5
PB1(4)	VIII	3TLP	1.9 ± 1.6	-1.5 ± 0.2	-2.0 ± 0.1	-0.4 ± 0.3	-0.9 ± 0.3
PB1(5)	VIII	3MB4	0.3 ± 2.9	-0.7 ± 0.4	-0.7 ± 0.2	-1.2 ± 0.2	-1.1 ± 0.2
SMARCA2	VIII	4QY4	1.4 ± 0.7	-1.6 ± 0.2	-1.2 ± 0.2	-1.4 ± 0.1	-1.3 ± 0.1
SMARCA4	VIII	2GRC	-3.1 ± 0.6	-1.5 ± 0.1	-0.7 ± 0.1	-1.3 ± 0.2	-1.5 ± 0.2

Table D.2: Results obtained with the TIP3P water model. All free energies (ΔG) are in kcal/mol, and all deviations from the crystal water positions (δ), as well as their RMSD, are in Å.

Protein	Family	ΔG_{netw}	ΔG_{W1}	ΔG_{W2}	ΔG_{W3}	ΔG_{W4}	δ_{W1}	δ_{W2}	δ_{W3}	δ_{W4}	RMSD
CECR2	I	−5.4		−1.1	−1.4	−1.9		1.2	0.3	0.5	0.8
FALZ	I	−7.5		−1.9	−2.2	−2.3		1.3	0.6	0.2	0.8
GCN5L2	I	−2.5	−0.8	0.0	−0.2	0.1	0.5	0.7	0.7	0.5	0.6
PCAF	I	−5.2	−1.3	−0.9	−0.7	−0.6	0.3	0.9	0.8	1.1	0.8
BRD2(1)	II	−6.1	−0.9	−0.7	−1.3	−1.3	0.6	0.5	0.2	0.5	0.5
BRD2(2)	II	−7.6	−1.7	−1.3	−1.5	−1.4	0.5	0.9	0.6	0.5	0.6
BRD3(1)	II	−6.3	−1.4	−0.8	−1.1	−1.3	0.5	0.7	0.7	0.2	0.6
BRD3(2)	II	−4.2	−0.7	0.6	−0.8	−1.2	0.5	0.5	0.5	0.3	0.4
BRD4(1)	II	−5.6	−1.8	−0.1	−0.9	−1.0	0.4	0.3	0.3	0.3	0.3
BRD4(2)	II	−8.4	−1.7	−1.4	−1.8	−1.6	0.2	0.6	0.4	0.4	0.4
BRDT(1)	II	−4.2	−0.7	−0.9	−0.3	−0.4	0.4	0.2	0.4	0.3	0.3
CREBBP	III	−6.6	2.4	−1.4	−2.0	−2.2	0.2	0.4	0.3	0.2	0.3
EP300	III	−4.2	−0.9	−0.9	0.5	−1.0	1.2	0.1	0.1	0.6	0.7
PHIP(2)	III	−4.3	−2.3	0.8	−0.6	−0.7	0.9	0.4	0.5	0.5	0.6
WDR9(2)	III	1.9	0.3	1.2	0.0	−0.4	0.7	0.3	0.8	1.1	0.8
ATAD2	IV	−1.0	5.4	−1.3	−1.0	−0.9	1.0	0.2	0.5	0.7	0.7
BRD1	IV	−6.7	−1.6	−1.9	−0.2	−1.1	0.6	0.1	0.4	1.0	0.6
BRD9	IV	−6.8	−1.2	−0.7	−1.5	−1.5	0.9	0.4	0.5	0.1	0.5
BRPF1B	IV	−10.8	−2.9	−1.3	−2.2	−2.5	0.4	0.5	0.6	0.2	0.4
KIAA1240	IV	−1.3	1.6	−1.4	−1.1	−0.7	0.7				0.7
BAZ2A	V	−4.0	0.5	−0.6	−0.9	−0.9	0.1	0.2	0.3	0.3	0.3
BAZ2B	V	−3.7	3.1	−1.6	−1.4	−1.4	0.7	0.7	0.3	0.5	0.5
TIF1a	V	−1.7	−0.1	−0.2	−0.9	−1.1	1.1	0.7	0.7	0.4	0.7
PRKCBP1	VII	3.6	3.4	0.7	0.9	0.8	0.9	0.5	0.4	0.4	0.6
TAF1L(2)	VII	−0.7	−2.3	3.7	−1.2	−2.2	0.8	0.8	0.4	0.1	0.6
TAF1(1)	VII	−7.7	−1.4	−0.9	−1.6	−1.8	0.3	0.5	0.5	0.3	0.4
TAF1(2)	VII	−7.9	−1.3	0.2	−2.2	−2.6	0.9	0.3	0.3	0.1	0.5
ASH1L	VIII	−4.5	−1.4	0.2	−1.5	−0.4					
PB1(1)	VIII	0.8	−0.4	1.1	0.9	−0.1	1.3	0.5	1.1	0.7	1.0
PB1(2)	VIII	−0.5	−3.0	1.7	−0.4	0.7	0.3	0.4	0.6	0.2	0.4
PB1(3)	VIII	−2.1	−1.0	−0.6	0.5	0.7			0.3		0.3
PB1(4)	VIII	3.6	−2.5	2.4	−0.9	1.9	0.7	0.5	0.5		0.6
PB1(5)	VIII	−1.2	0.1	0.4	−0.6	0.7	0.9	0.8	0.6	0.8	0.8
SMARCA2	VIII	0.4	0.6	−0.4	0.6	1.2	1.5		0.6		1.1
SMARCA4	VIII	−2.4	−2.4	−0.1	0.2	0.8	1.0	0.6	0.3	0.5	0.7

Table D.3: Results obtained with the TIP4P water model. All free energies (ΔG) are in kcal/mol, and all deviations from the crystal water positions (δ), as well as their RMSD, are in Å.

Protein	Family	ΔG_{netw}	ΔG_{W1}	ΔG_{W2}	ΔG_{W3}	ΔG_{W4}	δ_{W1}	δ_{W2}	δ_{W3}	δ_{W4}	RMSD
CECR2	I	−3.8	−1.3	0.5	−0.7	−1.2	1.1	0.6	0.3	0.7	0.7
FALZ	I	−9.3		−2.2	−3.0	−3.1		1.0	0.5	0.1	0.7
GCN5L2	I	−4.9	0.1	−1.1	−1.1	−1.0	0.6	0.7	0.5	0.7	0.6
PCAF	I	−5.8	−0.3	−1.6	−1.1	−0.9	1.5	0.5	0.6	0.8	0.9
BRD2(1)	II	−4.3	−0.2	−0.6	−0.8	−0.8	0.3	0.3	0.3	0.3	0.3
BRD2(2)	II	−6.8	−1.3	−1.2	−1.3	−1.3	0.6	0.8	0.5	0.5	0.6
BRD3(1)	II	−7.2	−1.4	−0.6	−1.6	−1.8	0.8	1.3	0.7	0.5	0.9
BRD3(2)	II	−5.8	−1.0	−0.9	−1.0	−1.0	0.4	0.6	0.4	0.6	0.5
BRD4(1)	II	−8.6	−1.7	−1.3	−1.8	−1.9	0.3	0.1	0.2	0.1	0.2
BRD4(2)	II	−7.8	−1.2	−1.7	−1.6	−1.6	0.4	0.2	0.5	0.8	0.5
BRDT(1)	II	−2.8	−0.5	−0.3	−0.1	−0.4	0.8	0.2	0.3	0.5	0.5
CREBBP	III	−7.2	−1.7	−1.4	−1.1	−1.1	1.3	0.4	0.2	0.1	0.7
EP300	III	−1.6	−0.2	−0.2	0.1	0.6	0.6	0.5	0.3	0.5	0.5
PHIP(2)	III	−6.4	−1.8	1.0	−1.3	−2.2	0.7	0.7	0.2	0.2	0.5
WDR9(2)	III	−5.7	−0.7	−0.1	−1.6	−1.6	0.8	0.6	0.6	1.0	0.7
ATAD2	IV	−0.2	7.5	−1.4	−0.8	−0.5	0.8	0.4	0.9	0.9	0.8
BRD1	IV	−5.7	−1.2	−2.2	−0.2	−0.2	0.8	0.3	0.4	1.6	0.9
BRD9	IV	−6.3	−0.9	−1.1	−1.2	−1.2	1.0	0.4	0.5	0.1	0.6
BRPF1B	IV	−12.4	−2.2	−1.6	−3.1	−3.3	0.5	0.5	0.6	0.2	0.5
KIAA1240	IV	−3.3	1.5	−2.1	−0.9	−0.8	0.8				0.8
BAZ2A	V	−4.8	−0.1	−1.1	−0.8	−0.9	0.2	0.3	0.1	0.3	0.2
BAZ2B	V	−0.4	2.8	−2.4	−0.3	−0.1	0.7	0.5	0.0	0.6	0.5
TIF1a	V	0.6	2.7	0.6	−0.7	−0.8	0.7	1.1	0.4	0.5	0.7
PRKCBP1	VII	1.0	0.8	0.5	0.8	0.7	0.7	0.2	0.4	0.5	0.5
TAF1L(2)	VII	−8.3	−1.8	−1.7	−1.2	−1.8	0.7	0.8	0.4	0.7	0.7
TAF1(1)	VII	−3.8	−0.6	−0.2	−0.3	−0.8	0.3	0.4	0.5	0.8	0.5
TAF1(2)	VII	−7.2	−0.7	−0.2	−1.9	−2.4	1.0	0.3	0.3	0.1	0.6
ASH1L	VIII	−3.3	−0.7	−0.3	−0.7	0.0					
PB1(1)	VIII	−2.1	−0.7	−0.2	0.2	0.3	0.8	0.4	0.6	1.0	0.7
PB1(2)	VIII	−1.4	−1.6	1.1	0.0	0.9	0.4	0.3	0.6	1.0	0.7
PB1(3)	VIII	−1.9	−0.5	−1.1	0.8	0.3			0.5		0.5
PB1(4)	VIII	−0.7	−1.2	2.4	−0.8	−0.7	0.7	0.2	0.3		0.5
PB1(5)	VIII	−3.2	0.1	−0.6	−0.7	−0.2	0.8	0.6	0.7	0.6	0.7
SMARCA2	VIII	2.3	0.9	−0.3	0.9	2.0	1.2		0.4		0.9
SMARCA4	VIII	−3.2	−1.6	0.3	0.2	−0.9	0.6	0.3	0.2	1.7	0.9

Table D.4: Results obtained with the SPC water model. All free energies (ΔG) are in kcal/mol, and all deviations from the crystal water positions (δ), as well as their RMSD, are in Å.

Protein	Family	ΔG_{netw}	ΔG_{W1}	ΔG_{W2}	ΔG_{W3}	ΔG_{W4}	δ_{W1}	δ_{W2}	δ_{W3}	δ_{W4}	RMSD
CECR2	I	−3.0	1.1	−0.7	−1.1	−1.2	0.9	0.6	0.5	0.3	0.6
FALZ	I	−7.8	−1.6	0.2	−2.3	−2.3	0.8	0.2	0.2	0.1	0.4
GCN5L2	I	−7.5	−1.1	−1.2	−1.7	−1.7	0.2	0.7	0.6	0.2	0.5
PCAF	I	−7.0	−1.0	−1.4	−1.3	−1.4	0.2	1.0	0.9	0.9	0.8
BRD2(1)	II	−6.6	−1.0	−1.1	−1.3	−1.3	0.4	0.7	0.2	0.4	0.4
BRD2(2)	II	−6.8	−1.8	−0.9	−1.3	−1.1	1.4	1.4	0.4	0.5	1.0
BRD3(1)	II	−6.8	−1.6	−0.7	−1.2	−1.5	0.5	0.7	0.6	0.2	0.5
BRD3(2)	II	−7.8	−1.5	−1.4	−1.7	−1.4	0.3	0.8	0.4	0.8	0.6
BRD4(1)	II	−9.3	−1.9	−1.8	−1.9	−1.9	0.3	0.3	0.1	0.1	0.2
BRD4(2)	II	−7.2	−0.9	−1.4	−1.6	−1.4	0.3	0.4	0.2	0.3	0.3
BRDT(1)	II	−4.2	−0.7	−0.7	−0.6	−0.5	0.6	0.3	0.2	0.6	0.5
CREBBP	III	−4.8	1.9	−1.6	−1.5	−1.6	0.4	0.5	0.2	0.1	0.3
EP300	III	−3.3	−0.3	−0.7	0.1	−0.6	0.4	0.2	0.5	0.5	0.4
PHIP(2)	III	−7.1	−2.4	−0.4	−1.1	−1.4	0.8	0.6	0.1	0.3	0.5
WDR9(2)	III	−3.4	0.7	0.0	−1.2	−1.1	0.7	0.7	0.5	0.5	0.6
ATAD2	IV	−6.2	−0.7	−1.6	−1.1	−1.1	1.4	0.2	0.7	0.7	0.8
BRD1	IV	−7.8	−1.8	−2.2	−1.0	−1.1	0.7	0.2	0.6	0.6	0.6
BRD9	IV	−6.9	−0.9	−1.1	−1.5	−1.4	0.9	0.4	0.6	0.1	0.6
BRPF1B	IV	−13.7	−3.1	−2.5	−3.1	−3.2	0.3	0.4	0.7	0.1	0.4
KIAA1240	IV	−4.2	1.2	−1.9	−0.9	−0.8	0.8				0.8
BAZ2A	V	−2.8	0.5	−0.7	−0.5	−0.2	0.4	0.4	0.2	0.2	0.3
BAZ2B	V	−3.8	1.3	−1.1	−0.9	−0.9	0.6	0.6	0.3	0.4	0.5
TIF1a	V	−5.0	−1.7	−0.3	−0.6	−0.6		1.3	0.7	0.2	0.9
PRKCBP1	VII	2.3	1.3	0.7	0.9	1.0	0.6	0.4	0.5	0.2	0.5
TAF1L(2)	VII	−8.9	−1.8	−1.6	−1.7	−2.0	0.3	0.7	0.5	0.5	0.5
TAF1(1)	VII	−6.5	−1.3	−0.7	−1.0	−1.6	0.2	0.5	0.5	0.5	0.4
TAF1(2)	VII	−8.3	−0.5	−1.0	−2.3	−2.4	1.6	0.4	0.1	0.1	0.8
ASH1L	VIII	−5.6	−0.8	−0.3	−1.3	−1.3					
PB1(1)	VIII	−2.5	−0.6	0.2	0.4	−0.9	1.3	0.7	1.0	0.7	0.9
PB1(2)	VIII	−3.0	−2.2	1.1	−0.9	0.0	0.2	0.3	0.6	0.4	0.4
PB1(3)	VIII	−3.9	−1.7	−0.8	0.2	0.2			0.5		0.5
PB1(4)	VIII	2.5	−1.7	2.2	−0.6	1.4	0.3	0.1	0.3		0.2
PB1(5)	VIII	3.4	4.8	−0.1	−0.8	0.9	0.6	0.4	0.3	0.3	0.4
SMARCA2	VIII	1.4	0.0	0.3	0.8	1.9	1.1		0.6		0.9
SMARCA4	VIII	−3.9	−2.6	−0.2	−0.1	0.3	1.1	0.5	0.3	0.7	0.7

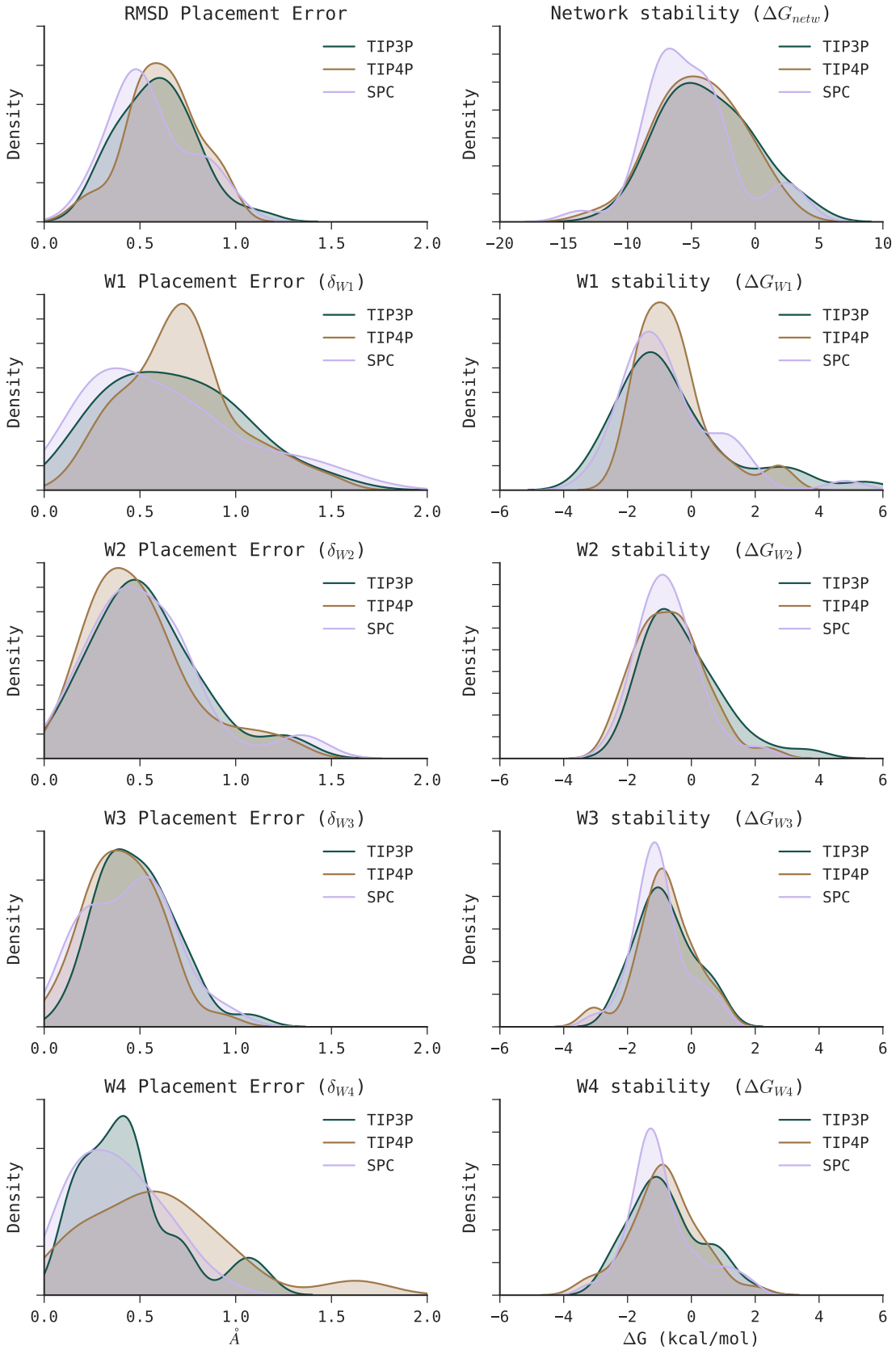


Figure D.1: Distributions of placement errors and free energies for the three water models. Shown are gaussian kernel density estimates of the samples.

Nos esse quasi nanos, gigantium humeris insidentes.

— John of Salisbury, in *Metalogicon*, 1159

References

[1] P. G. Polishchuk, T. I. Madzhidov, and A. Varnek. "Estimation of the size of drug-like chemical space based on GDB-17 data". In: *Journal of Computer-Aided Molecular Design* 27.8 (2013), pp. 675–679.

[2] Birgit T. Priest and Gul Erdemli. "Phenotypic Screening in the 21st Century". In: *Frontiers in Pharmacology* 5.264 (2014).

[3] Amy C. Anderson. "The Process of Structure-Based Drug Design". In: *Chemistry & Biology* 10.9 (2003), pp. 787–797.

[4] Joseph A. DiMasi, Henry G. Grabowski, and Ronald W. Hansen. "Innovation in the pharmaceutical industry: New estimates of R&D costs". In: *Journal of Health Economics* 47 (2016), pp. 20–33.

[5] Ted T. Ashburn and Karl B. Thor. "Drug repositioning: identifying and developing new uses for existing drugs". In: *Nature Review Drug Discovery* 3.8 (2004), pp. 673–683.

[6] William L. Jorgensen. "The Many Roles of Computation in Drug Discovery". In: *Science* 303.5665 (2004), pp. 1813–1818.

[7] David S. Goodsell, Garrett M. Morris, and Arthur J. Olson. "Automated docking of flexible ligands: Applications of autodock". In: *Journal of Molecular Recognition* 9.1 (1996), pp. 1–5.

[8] Garrett M. Morris et al. "Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function". In: *Journal of Computational Chemistry* 19.14 (1998), pp. 1639–1662.

[9] Garrett M. Morris et al. "AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility". In: *Journal of Computational Chemistry* 30.16 (2009), pp. 2785–2791.

[10] Ruth Huey et al. "A semiempirical free energy force field with charge-based desolvation". In: *Journal of Computational Chemistry* 28.6 (2007), pp. 1145–1152.

[11] Fredrik Österberg et al. "Automated docking to multiple target structures: Incorporation of protein mobility and structural water heterogeneity in AutoDock". In: *Proteins: Structure, Function, and Bioinformatics* 46.1 (2002), pp. 34–40.

[12] Valère Lounnas et al. "Current progress in structure-based rational drug design marks a new mindset in drug discovery". In: *Computational and Structural Biotechnology Journal* 5.6 (2013), pp. 1–14.

[13] Soma Mandal, Mee'nal Moudgil, and Sanat K. Mandal. "Rational drug design". In: *European Journal of Pharmacology* 625.1–3 (2009), pp. 90–100.

[14] Jacob Durrant and J. Andrew McCammon. “Molecular dynamics simulations and drug discovery”. In: *BMC Biology* 9.1 (2011), pp. 71–71.

[15] J. Andrew McCammon, Bruce R. Gelin, and Martin Karplus. “Dynamics of folded proteins”. In: *Nature* 267.5612 (1977), pp. 585–590.

[16] Christopher D. Wassman et al. “Computational identification of a transiently open L1/S3 pocket for reactivation of mutant p53”. In: *Nature Communications* 4 (2013), 1407 EP.

[17] Rommie E. Amaro et al. “Discovery of drug-like inhibitors of an essential RNA-editing ligase in Trypanosoma brucei”. In: *Proceedings of the National Academy of Sciences* 105.45 (2008), pp. 17278–17283.

[18] Vittorio Limongelli et al. “Molecular basis of cyclooxygenase enzymes (COXs) selective inhibition”. In: *Proceedings of the National Academy of Sciences* 107.12 (2010), pp. 5411–5416.

[19] Giovanni Grazioso et al. “Investigating the mechanism of substrate uptake and release in the glutamate transporter homologue Glt Ph through metadynamics simulations”. In: *Journal of the American Chemical Society* 134.1 (2012), pp. 453–463.

[20] Corentin Herbert et al. “Molecular mechanism of SSR128129E, an extracellularly acting, small-molecule, allosteric inhibitor of fgf receptor signaling”. In: *Cancer Cell* 23.4 (2013), pp. 489–501.

[21] Caterina Bissantz, Bernd Kuhn, and Martin Stahl. “A Medicinal Chemist’s Guide to Molecular Interactions”. In: *Journal of Medicinal Chemistry* 53.14 (2010), pp. 5061–5084.

[22] John D. Chodera and David L. Mobley. “Entropy-Enthalpy Compensation: Role and Ramifications in Biomolecular Ligand Recognition and Design”. In: *Annual Review of Biophysics* 42.1 (2013), pp. 121–142.

[23] Jeff Wereszczynski and J. Andrew McCammon. “Statistical mechanics and molecular dynamics in evaluating thermodynamic properties of biomolecular recognition.” In: *Quarterly Reviews of Biophysics* 45.1 (2012), pp. 1–25.

[24] William L. Jorgensen. “Efficient Drug Lead Discovery and Optimization”. In: *Accounts of Chemical Research* 42.6 (2009), pp. 724–733.

[25] Christophe Chipot. “Frontiers in free-energy calculations of biological systems”. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 4.1 (2014), pp. 71–89.

[26] Gerhard Hummer and Attila Szabo. “Free energy reconstruction from nonequilibrium single-molecule pulling experiments”. In: *Proceedings of the National Academy of Sciences* 98.7 (2001), pp. 3658–3661.

[27] C. Jarzynski. “Nonequilibrium Equality for Free Energy Differences”. In: *Physical Review Letters* 78.14 (1997), pp. 2690–2693.

[28] C. Jarzynski. “Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach”. In: *Physical Review E* 56.5 (1997), pp. 5018–5035.

[29] Michael S. Lee and Mark A. Olson. "Calculation of Absolute Protein-Ligand Binding Affinity Using Path and Endpoint Approaches". In: *Biophysical Journal* 90.3 (2006), pp. 864–877.

[30] Hyung-June Woo and Benoît Roux. "Calculation of absolute protein–ligand binding free energy from computer simulations". In: *Proceedings of the National Academy of Sciences* 102.19 (2005), pp. 6825–6830.

[31] Tingjun Hou et al. "Assessing the performance of the molecular mechanics/Poisson Boltzmann surface area and molecular mechanics/generalized Born surface area methods. 1. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations". In: *Journal of Chemical Information and Modeling* 51.5 (2011), pp. 69–82.

[32] Tingjun Hou et al. "Assessing the performance of the molecular mechanics/Poisson Boltzmann surface area and molecular mechanics/generalized Born surface area methods. II. The accuracy of ranking poses generated from docking". In: *Journal of Computational Chemistry* 32.5 (2011), pp. 866–877.

[33] Rashmi Kumari, Rajendra Kumar, and Andrew Lynn. "g_mmpbsa–A GROMACS Tool for High-Throughput MM-PBSA Calculations". In: *Journal of Chemical Information and Modeling* 54.7 (2014), pp. 1951–1962.

[34] David W. Wright et al. "Computing Clinically Relevant Binding Free Energies of HIV-1 Protease Inhibitors." In: *Journal Of Chemical Theory and Computation* 10.3 (2014), pp. 1228–1241.

[35] Lei Xu et al. "Assessing the Performance of MM/PBSA and MM/GBSA Methods. 3. The Impact of Force Fields and Ligand Charge Models". In: *The Journal of Physical Chemistry B* 117.28 (2013), pp. 8408–8421.

[36] Francesco Luigi Gervasio, Alessandro Laio, and Michele Parrinello. "Flexible Docking in Solution Using Supporting Information ." In: *Journal of the American Chemical Society* 16 (2004), pp. 1–5.

[37] Ludovico Sutto, Simone Marsili, and Francesco Luigi Gervasio. "New advances in metadynamics". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 2.5 (2012), pp. 771–779.

[38] Andrea Cavalli et al. "Investigating Drug − Target Association and Dissociation Mechanisms Using Metadynamics-Based Algorithms". In: *Accounts of Chemical Research* 48 (2015), pp. 277–285.

[39] Vittorio Limongelli, Massimiliano Bonomi, and Michele Parrinello. "Funnel metadynamics as accurate binding free-energy method". In: *Proceedings of the National Academy of Sciences* 110.16 (2013), pp. 6358–6363.

[40] Laura Troussicot et al. "Funnel-Metadynamics and Solution NMR to Estimate Protein–Ligand Affinities". In: *Journal of the American Chemical Society* 137.3 (2015), pp. 1273–1281.

[41] G. Saladino et al. "Assessing the Performance of Metadynamics and Path Variables in Predicting the Binding Free Energies of p38 Inhibitors". In: *Journal of Chemical Theory and Computation* 8.4 (2012), pp. 1165–1170.

[42] Neva Bešker and Francesco L. Gervasio. "Using Metadynamics and Path Collective Variables to Study Ligand Binding and Induced Conformational Transitions". In: *Computational Drug Discovery and Design*. Ed. by Riccardo Baron. New York, NY: Springer New York, 2012, pp. 501–513.

[43] Michael R. Shirts, David L. Mobley, and John D. Chodera. "Alchemical Free Energy Calculations: Ready for Prime Time?" In: vol. 3. Annual Reports in Computational Chemistry. Elsevier, 2007, pp. 41–59.

[44] David L. Mobley and Pavel V. Klimovich. "Perspective: Alchemical free energy calculations for drug discovery". In: *The Journal of Chemical Physics* 137.23 (2012), 230901 EP.

[45] Mariela Bollini et al. "Computationally-Guided Optimization of a Docking Hit to Yield Catechol Diethers as Potent Anti-HIV Agents". In: *Journal of Medicinal Chemistry* 54.24 (2011), pp. 8582–8591.

[46] William L. Jorgensen et al. "Efficient Discovery of Potent Anti-HIV Agents Targeting the Tyr181Cys Variant of HIV Reverse Transcriptase". In: *Journal of the American Chemical Society* 133.39 (2011), pp. 15686–15696.

[47] Lingle Wang 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". In: *Journal of the American Chemical Society* 137.7 (2015), pp. 2695–2703.

[48] David L. Mobley et al. "Blind prediction of HIV integrase binding from the SAMPL4 challenge". In: *Journal of Computer-Aided Molecular Design* (2014), pp. 1–19.

[49] Nathan M. Lim et al. "Sensitivity in binding free energies due to protein reorganization". In: *Journal of Chemical Theory and Computation* 12.9 (2016), pp. 4620–4631.

[50] David L. Mobley, John D. Chodera, and Ken A. Dill. "Confine-and-Release Method: Obtaining Correct Binding Free Energies in the Presence of Protein Conformational Change". In: *Journal of Chemical Theory and Computation* 3.4 (2007), pp. 1231–1235.

[51] Kai Wang et al. "Identifying ligand binding sites and poses using GPU-accelerated Hamiltonian replica exchange molecular dynamics". In: *Journal of Computer-Aided Molecular Design* 27.12 (2013), pp. 989–1007.

[52] Shuai Liu, Lingle Wang, and David L. Mobley. "Is Ring Breaking Feasible in Relative Binding Free Energy Calculations?" In: *Journal of Chemical Information and Modeling* 55.4 (2015), pp. 727–735.

[53] Michael R. Shirts, David L. Mobley, and Scott P. Brown. "Free-energy calculations in structure-based drug design". In: Cambridge University Press, 2010, pp. 61–86.

[54] M. O. Jensen et al. "Mechanism of Voltage Gating in Potassium Channels". In: *Science* 336.6078 (2012), pp. 229–233.

[55] Ron O. Dror et al. "Structural basis for nucleotide exchange in heterotrimeric G proteins". In: *Science* 348.6241 (2015), pp. 1361–1365.

[56] John S. Burg et al. “Structural basis for chemokine recognition and activation of a viral G protein-coupled receptor”. In: *Science* 347.6226 (Mar. 2015), pp. 1113–1117.

[57] David L. Mobley et al. “Predicting Absolute Ligand Binding Free Energies to a Simple Model Site”. In: *Journal of Molecular Biology* 371.4 (2007), pp. 1118–1134.

[58] Sarah E. Boyce et al. “Predicting Ligand Binding Affinity with Alchemical Free Energy Methods in a Polar Model Binding Site”. In: *Journal of Molecular Biology* 394.4 (2009), pp. 747–763.

[59] Hideaki Fujitani et al. “Direct calculation of the binding free energies of FKBP ligands”. In: *The Journal of Chemical Physics* 123.8 (2005), 084108 EP.

[60] Gregory A. Ross, Garrett M. Morris, and Philip C. Biggin. “Rapid and Accurate Prediction and Scoring of Water Molecules in Protein Binding Sites”. In: *PLoS ONE* 7.3 (2012), 32036 EP.

[61] David Chandler, John D. McCoy, and Sherwin J. Singer. “Density functional theory of nonuniform polyatomic systems. I. General formulation”. In: *The Journal of Chemical Physics* 85.10 (1986), 5971 EP.

[62] Andriy Kovalenko and Fumio Hirata. “Self-consistent description of a metal–water interface by the Kohn–Sham density functional theory and the three-dimensional reference interaction site model”. In: *The Journal of Chemical Physics* 110.20 (1999), pp. 10095–10112.

[63] Themis Lazaridis. “Inhomogeneous Fluid Approach to Solvation Thermodynamics. 1. Theory”. In: *The Journal of Physical Chemistry B* 102.18 (1998), pp. 3531–3541.

[64] Themis Lazaridis. “Inhomogeneous Fluid Approach to Solvation Thermodynamics. 2. Applications to Simple Fluids”. In: *The Journal of Physical Chemistry B* 102.18 (1998), pp. 3542–3550.

[65] Tom Young et al. “Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding.” In: *Proceedings of the National Academy of Science* 104.3 (2007), pp. 808–13.

[66] Crystal N. Nguyen, Tom Kurtzman Young, and Michael K. Gilson. “Grid inhomogeneous solvation theory: hydration structure and thermodynamics of the miniature receptor cucurbit[7]uril.” In: *The Journal of Chemical Physics* 137.4 (2012), 044101 EP.

[67] Georgios Gerogiokas et al. “Prediction of Small Molecule Hydration Thermodynamics with Grid Cell Theory”. In: *Journal of Chemical Theory and Computation* 10.1 (2014), pp. 35–48.

[68] Gregory A. Ross, Micheal S. Bodnarchuk, and Jonathan W. Essex. “Water sites, networks, and free energies with grand canonical Monte Carlo”. In: *Journal of the American Chemical Society* 137.47 (2015), pp. 14930–14943.

[69] Panagis Filippakopoulos and Stefan Knapp. “Targeting bromodomains: epigenetic readers of lysine acetylation”. In: *Nature Review Drug Discovery* 13.5 (2014), pp. 337–356.

[70] Panagis Filippakopoulos et al. “Histone Recognition and Large-Scale Structural Analysis of the Human Bromodomain Family”. In: *Cell* 149.1 (2012), pp. 214–231.

[71] Panagis Filippakopoulos and Stefan Knapp. “The bromodomain interaction module”. In: *FEBS Letters* 586.17 (2012), pp. 2692–2704.

[72] Steven G. Smith and Ming-Ming Zhou. “The Bromodomain: A New Target in Emerging Epigenetic Medicine”. In: *ACS Chemical Biology* 11.3 (2016), pp. 598–608.

[73] John W. Tamkun et al. “Brahma: A regulator of Drosophila homeotic genes structurally related to the yeast transcriptional activator SNF2/SWI2”. In: *Cell* 68.3 (1992), pp. 561–572.

[74] Susan R. Haynes et al. “The bromodomain: a conserved sequence found in human, Drosophila and yeast proteins”. In: *Nucleic Acids Research* 20.10 (1992), p. 2603.

[75] Christophe Dhalluin et al. “Structure and ligand of a histone acetyltransferase bromodomain”. In: *Nature* 399.6735 (1999), pp. 491–496.

[76] Raymond H. Jacobson et al. “Structure and Function of a Human TAFII250 Double Bromodomain Module”. In: *Science* 288.5470 (2000), pp. 1422–1425.

[77] David J. Owen et al. “The structural basis for the recognition of acetylated histone H4 by the bromodomain of histone acetyltransferase Gcn5p”. In: *The EMBO Journal* 19.22 (2000), pp. 6141–6149.

[78] Vanessa Brès, Sunnie M Yoh, and Katherine A Jones. “The multi-tasking P-TEFb complex”. In: *Current Opinion in Cell Biology* 20.3 (2008), pp. 334–340.

[79] Roberto Sanchez and Ming-Ming Zhou. “The role of human bromodomains in chromatin biology and gene transcription”. In: *Current Opinion in Drug Discovery & Development* 12.5 (2009), pp. 659–665.

[80] Oleg Fedorov et al. “Selective targeting of the BRG/PB1 bromodomains impairs embryonic and trophoblast stem cell maintenance”. In: *Science Advances* 1.10 (2015).

[81] Jeanne Moriniere et al. “Cooperative binding of two acetylation marks on a histone tail by a single bromodomain”. In: *Nature* 461.7264 (2009), pp. 664–668.

[82] Alexey V. Ivanov et al. “PHD Domain-Mediated E3 Ligase Activity Directs Intramolecular Sumoylation of an Adjacent Bromodomain Required for Gene Silencing”. In: *Molecular Cell* 28.5 (2007), pp. 823–837.

[83] Lei Zeng et al. “Structural insights into human KAP1 PHD finger-bromodomain and its role in gene silencing”. In: *Nature Structural & Molecular Biology* 15.6 (2008), pp. 626–633.

[84] Ivana Celic et al. “The Sirtuins Hst3 and Hst4p Preserve Genome Integrity by Controlling Histone H3 Lysine 56 Deacetylation”. In: *Current Biology* 16.13 (2006), pp. 1280–1289.

[85] Susanne Muller, Panagis Filippakopoulos, and Stefan Knapp. “Bromodomains as therapeutic targets”. In: *Expert Reviews in Molecular Medicine* 13 (2011).

[86] Nicolas Reynoird et al. “Oncogenesis by sequestration of CBP/p300 in transcriptionally inactive hyperacetylated chromatin domains”. In: *The EMBO Journal* 29.17 (2010), pp. 2943–2952.

[87] Panagis Filippakopoulos et al. “Selective inhibition of BET bromodomains”. In: *Nature* 468.7327 (2010), pp. 1067–1073.

[88] Kenji Deguchi et al. “MOZ-TIF2-induced acute myeloid leukemia requires the MOZ nucleosome binding motif and TIF2-mediated recruitment of CBP”. In: *Cancer Cell* 3.3 (2003), pp. 259–271.

[89] Olatoyosi M Sobulo et al. “MLL is fused to CBP, a histone acetyltransferase, in therapy-related acute myeloid leukemia with a t(11;16)(q23;p13.3)”. In: *Proceedings of the National Academy of Sciences* 94.16 (1997), pp. 8732–8737.

[90] Laura Pasqualucci et al. “Inactivating mutations of acetyltransferase genes in B-cell lymphoma”. In: *Nature* 471.7337 (2011), pp. 189–195.

[91] Jan Bouchal et al. “Transcriptional coactivators p300 and CBP stimulate estrogen receptor-beta signaling and regulate cellular events in prostate cancer”. In: *The Prostate* 71.4 (2011), pp. 431–437.

[92] Marco Ciró et al. “ATAD2 Is a Novel Cofactor for MYC, Overexpressed and Amplified in Aggressive Tumors”. In: *Cancer Research* 69.21 (2009), pp. 8491–8498.

[93] Alexey S. Revenko et al. “Chromatin Loading of E2F-MLL Complex by Cancer-Associated Coregulator ANCCA via Reading a Specific Histone Mark”. In: *Molecular and Cellular Biology* 30.22 (2010), pp. 5260–5272.

[94] Nathaniel T. Leachman et al. “ATAD2B is a phylogenetically conserved nuclear protein expressed during neuronal differentiation and tumorigenesis”. In: *Development, Growth & Differentiation* 52.9 (2010), pp. 747–755.

[95] Carolin Grunwald et al. “Expression of multiple epigenetically regulated cancer/germline genes in nonsmall cell lung cancer”. In: *International Journal of Cancer* 118.10 (2006), pp. 2522–2528.

[96] Matthew J Scanlan et al. “Expression of cancer-testis antigens in lung cancer: definition of bromodomain testis-specific gene (BRDT) as a new CT gene, CT9”. In: *Cancer Letters* 150.2 (2000), pp. 155–164.

[97] Wei-Gang Tong et al. “Phase I and Pharmacologic Study of SNS-032, a Potent and Selective Cdk2, 7, and 9 Inhibitor, in Patients With Advanced Chronic Lymphocytic Leukemia and Multiple Myeloma”. In: *Journal of Clinical Oncology* 28.18 (2010), pp. 3015–3022.

[98] Aristeidis Chaidos et al. “Potent antimyeloma activity of the novel bromodomain inhibitors I-BET151 and I-BET762”. In: *Blood* 123.5 (2014), pp. 697–705.

[99] Jennifer A. Mertz et al. “Targeting MYC dependence in cancer by inhibiting BET bromodomains”. In: *Proceedings of the National Academy of Sciences* 108.40 (2011), pp. 16669–16674.

[100] Edwige Nicodeme et al. “Suppression of inflammation by a synthetic histone mimic”. In: *Nature* 468.7327 (2010), pp. 1119–1123.

[101] Manolis Pasparakis. “Regulation of tissue homeostasis by NF-κB signalling: implications for inflammatory diseases”. In: *Nature Review Immunology* 9.11 (2009), pp. 778–788.

[102] Hiba Mahdi et al. “Specific interaction between genotype, smoking and autoimmunity to citrullinated [alpha]-enolase in the etiology of rheumatoid arthritis”. In: *Nature Genetics* 41.12 (2009), pp. 1319–1324.

[103] Zichong Li et al. “The BET bromodomain inhibitor JQ1 activates HIV latency through antagonizing Brd4 inhibition of Tat-transactivation”. In: *Nucleic Acids Research* 41.1 (2013), pp. 277–287.

[104] Panpan Lu et al. “The BET inhibitor OTX015 reactivates latent HIV-1 through P-TEFb”. In: *Scientific Reports* 6.24100 (2016).

[105] Jian Zhu et al. “Reactivation of Latent HIV-1 by Inhibition of BRD4”. In: *Cell Reports* 2.4 (2012), pp. 807–816.

[106] Melanie Ott, Matthias Geyer, and Qiang Zhou. “The Control of HIV Transcription: Keeping RNA Polymerase II on Track”. In: *Cell Host & Microbe* 10.5 (2011), pp. 426–435.

[107] Katarzyna Kaczmarek Michaels et al. “RNAP II processivity is a limiting step for HIV-1 transcription independent of orientation to and activity of endogenous neighboring promoters”. In: *Virology* 486 (2015), pp. 7–14.

[108] Sona Budhiraja and Andrew P Rice. “Reactivation of latent HIV: do all roads go through P-TEFb?” In: *Future Virology* 8.7 (2013), pp. 649–659.

[109] Dwayne A. Bisgrove et al. “Conserved P-TEFb-interacting domain of BRD4 inhibits HIV transcription”. In: *Proceedings of the National Academy of Sciences* 104.34 (2007), pp. 13690–13695.

[110] Daniela Boehm et al. “BET bromodomain-targeting compounds reactivate HIV from latency via a Tat-independent mechanism”. In: *Cell Cycle* 12.3 (2013), pp. 452–462.

[111] Mondira Kundu et al. “Interaction between Cell Cycle Regulator, E2F-1, and NF-κB Mediates Repression of HIV-1 Gene Transcription”. In: *Journal of Biological Chemistry* 272.47 (1997), pp. 29468–29474.

[112] Xin Wang et al. “Bromodomain Protein Brd4 Plays a Key Role in Merkel Cell Polyomavirus DNA Replication”. In: *PLOS Pathogens* 8.11 (2012), pp. 1–16.

[113] Xin Wang et al. “Recruitment of Brd4 to the Human Papillomavirus Type 16 DNA Replication Complex Is Essential for Replication of Viral DNA”. In: *Journal of Virology* 87.7 (2013), pp. 3871–3884.

[114] Lei Zeng et al. “Selective Small Molecules Blocking HIV-1 Tat and Coactivator PCAF Association”. In: *Journal of the American Chemical Society* 127.8 (2005), pp. 2376–2377.

[115] Chongfeng Pan et al. “Structure-Guided Optimization of Small Molecules Inhibiting Human Immunodeficiency Virus 1 Tat Association with the Human Coactivator p300/CREB Binding Protein-Associated Factor”. In: *Journal of Medicinal Chemistry* 50.10 (2007), pp. 2285–2288.

[116] Michael Brand et al. “Small Molecule Inhibitors of Bromodomain–Acetyl-lysine Interactions”. In: *ACS Chemical Biology* 10.1 (2015), pp. 22–39.

[117] Olivier Mirguet et al. “Discovery of Epigenetic Regulator I-BET762: Lead Optimization to Afford a Clinical Candidate Inhibitor of the BET Bromodomains”. In: *Journal of Medicinal Chemistry* 56.19 (2013), pp. 7501–7515.

[118] T. Tahara et al. *Therapeutic agent for osteoporosis and diazepine compound*. 1998.

[119] K. Adachi et al. *Thienotriazolodiazepine compound and a medicinal use thereof.* 2006.

[120] S. Miyoshi et al. *Antitumor agent.* 2009.

[121] J. Kay Noel et al. "Abstract C244: Development of the BET bromodomain inhibitor OTX015." In: *American Association for Cancer Research* 12.11 Supplement (2014), p. C244.

[122] David S. Hewings et al. "3,5-Dimethylisoxazoles Act As Acetyl-lysine-mimetic Bromodomain Ligands". In: *Journal of Medicinal Chemistry* 54.19 (2011), pp. 6761–6770.

[123] David S. Hewings et al. "Optimization of 3,5-Dimethylisoxazole Derivatives as Potent Bromodomain Ligands". In: *Journal of Medicinal Chemistry* 56.8 (2013), pp. 3217–3227.

[124] Mark A. Dawson et al. "Inhibition of BET recruitment to chromatin as an effective treatment for MLL-fusion leukaemia". In: *Nature* 478.7370 (2011), pp. 529–533.

[125] Paul Bamborough et al. "Fragment-Based Discovery of Bromodomain Inhibitors Part 2: Optimization of Phenylisoxazole Sulfonamides". In: *Journal of Medicinal Chemistry* 55.2 (2012), pp. 587–596.

[126] Olivier Mirguet et al. "From ApoA1 upregulation to BET family bromodomain inhibition: Discovery of I-BET151". In: *Bioorganic & Medicinal Chemistry Letters* 22.8 (2012), pp. 2963–2967.

[127] Jonathan Seal et al. "Identification of a novel series of BET family bromodomain inhibitors: Binding mode and profile of I-BET151 (GSK1210151A)". In: *Bioorganic & Medicinal Chemistry Letters* 22.8 (2012), pp. 2968–2972.

[128] Duncan Hay et al. "The design and synthesis of 5- and 6-isoxazolylbenzimidazoles as selective inhibitors of the BET bromodomains". In: *Medicinal Chemistry Communications* 4 (2013), pp. 140–144.

[129] Duncan A. Hay et al. "Discovery and optimization of small-molecule ligands for the CBP/p300 bromodomains". In: *Journal of the American Chemical Society* 136.26 (2014), pp. 9308–9319.

[130] Victor S. Gehling et al. "Discovery, Design, and Optimization of Isoxazole Azepine BET Inhibitors". In: *ACS Medicinal Chemistry Letters* 4.9 (2013), pp. 835–840.

[131] Brian K. Albrecht et al. "Identification of a Benzoisoxazoloazepine Inhibitor (CPI-0610) of the Bromodomain and Extra-Terminal (BET) Family as a Candidate for Human Clinical Trials". In: *Journal of Medicinal Chemistry* 59.4 (2016), pp. 1330–1339.

[132] Dana Bailey et al. "RVX-208: A Small Molecule That Increases Apolipoprotein A-I and High-Density Lipoprotein Cholesterol In Vitro and In Vivo". In: *Journal of the American College of Cardiology* 55.23 (2010), pp. 2580–2589.

[133] Dean Gilham et al. "RVX-208, a BET-inhibitor for treating atherosclerotic cardiovascular disease, raises ApoA-I/HDL and represses pathways that contribute to cardiovascular disease". In: *Atherosclerosis* 247 (2016), pp. 48–57.

[134] Sarah Picaud et al. “RVX-208, an inhibitor of BET transcriptional regulators with selectivity for the second bromodomain”. In: *Proceedings of the National Academy of Sciences* 110.49 (2013), pp. 19754–19759.

[135] Dragana Nikolic et al. “An evaluation of RVX-208 for the treatment of atherosclerosis”. In: *Expert Opinion on Investigational Drugs* 24.10 (2015), pp. 1389–1398.

[136] Paul V. Fish et al. “Identification of a Chemical Probe for Bromo and Extra C-Terminal Bromodomain Inhibition through Optimization of a Fragment-Derived Hit”. In: *Journal of Medicinal Chemistry* 55.22 (2012), pp. 9831–9837.

[137] Sarah Picaud et al. “PFI-1, a highly selective protein interaction inhibitor, targeting BET Bromodomains”. In: *Cancer Research* 73.11 (2013), pp. 3336–3346.

[138] Pietro Ciceri et al. “Dual kinase-bromodomain inhibitors for rationally designed polypharmacology”. In: *Nature Chemical Biology* 10.4 (2014), pp. 305–312.

[139] Stuart W. J. Ember et al. “Acetyl-lysine Binding Site of Bromodomain-Containing Protein 4 (BRD4) Interacts with Diverse Kinase Inhibitors”. In: *ACS Chemical Biology* 9.5 (2014), pp. 1160–1171.

[140] Sachchidanand et al. “Target Structure-Based Discovery of Small Molecules that Block Human p53 and CREB Binding Protein Association”. In: *Chemistry & Biology* 13.1 (2006), pp. 81–90.

[141] Jagat C. Borah et al. “A Small Molecule Binding to the Coactivator CREB-Binding Protein Blocks Apoptosis in Cardiomyocytes”. In: *Chemistry & Biology* 18.4 (Apr. 2011), pp. 531–541.

[142] Timothy P. C. Rooney et al. “A Series of Potent CREBBP Bromodomain Ligands Reveals an Induced-Fit Pocket Stabilized by a Cation-π Interaction”. In: *Angewandte Chemie International Edition* 53.24 (2014), pp. 6126–6130.

[143] Sarah Picaud et al. “Generation of a Selective Small Molecule Inhibitor of the CBP/p300 Bromodomain for Leukemia Therapy”. In: *Cancer Research* 75.23 (2015), pp. 5106–5119.

[144] Emmanuel H. Demont et al. “1,3-Dimethyl Benzimidazolones Are Potent, Selective Inhibitors of the BRPF1 Bromodomain”. In: *ACS Medicinal Chemistry Letters* 5.11 (2014), pp. 1190–1195.

[145] Wylie S. Palmer et al. “Structure-Guided Design of IACS-9571, a Selective High-Affinity Dual TRIM24-BRPF1 Bromodomain Inhibitor”. In: *Journal of Medicinal Chemistry* 59.4 (2016), pp. 1440–1454.

[146] Jian Zhu and Amedeo Caflisch. “Twenty Crystal Structures of Bromodomain and PHD Finger Containing Protein 1 (BRPF1)/Ligand Complexes Reveal Conserved Binding Motifs and Rare Interactions”. In: *Journal of Medicinal Chemistry* 59.11 (2016), pp. 5555–5561.

[147] Mary J. Harner et al. “Fragment-Based Screening of the Bromodomain of ATAD2”. In: *Journal of Medicinal Chemistry* 57.22 (2014), pp. 9687–9692.

[148] Emmanuel H. Demont et al. “Fragment-Based Discovery of Low-Micromolar ATAD2 Bromodomain Inhibitors”. In: *Journal of Medicinal Chemistry* 58.14 (2015), pp. 5649–5673.

[149] Paul Bamborough et al. “Structure-Based Optimization of Naphthyridones into Potent ATAD2 Bromodomain Inhibitors”. In: *Journal of Medicinal Chemistry* 58.15 (2015), pp. 6151–6178.

[150] Ludovic Drouin et al. “Structure Enabled Design of BAZ2-ICR, A Chemical Probe Targeting the Bromodomains of BAZ2A and BAZ2B”. In: *Journal of Medicinal Chemistry* 58.5 (2015), pp. 2553–2559.

[151] Peiling Chen et al. “Discovery and Characterization of GSK2801, a Selective Chemical Probe for the Bromodomains BAZ2A and BAZ2B”. In: *Journal of Medicinal Chemistry* 59.4 (2016), pp. 1410–1424.

[152] Peter G. K. Clark et al. “LP99: Discovery and Synthesis of the First Selective BRD7/9 Bromodomain Inhibitor”. In: *Angewandte Chemie International Edition* 54.21 (2015), pp. 6217–6221.

[153] Duncan A. Hay et al. “Design and synthesis of potent and selective inhibitors of BRD7 and BRD9 bromodomains”. In: *Medicinal Chemistry Communications* 6.7 (2015), pp. 1381–1386.

[154] Sarah Picaud et al. “9H-Purine Scaffold Reveals Induced-Fit Pocket Plasticity of the BRD9 Bromodomain”. In: *Journal of Medicinal Chemistry* 58.6 (2015), pp. 2718–2736.

[155] Natalie H. Theodoulou et al. “Discovery of I-BRD9, a Selective Cell Active Chemical Probe for Bromodomain Containing Protein 9 Inhibition”. In: *Journal of Medicinal Chemistry* 59.4 (2016), pp. 1425–1439.

[156] Brian S. Gerstenberger et al. “Identification of a Chemical Probe for Family VIII Bromodomains through Optimization of a Fragment Hit”. In: *Journal of Medicinal Chemistry* 59.10 (2016), pp. 4800–4811.

[157] Charlotte Louise Sutherell et al. “Identification and development of 2,3-dihydropyrrolo[1,2- a]quinazolin-5(1 H)-one inhibitors targeting bromodomains within the Switch/Sucrose Non-Fermenting complex”. In: *Journal of Medicinal Chemistry* 59.10 (2016), pp. 5095–5101.

[158] Oakley B. Cox et al. “A Poised Fragment Library Enables Rapid Synthetic Expansion Yielding the First Reported Inhibitors of PHIP(2), an Atypical Bromodomain”. In: *Chemical Science* 7 (2016), pp. 2322–2330.

[159] Sarah Picaud et al. “Promiscuous targeting of bromodomains by Bromosporine identifies BET proteins as master regulators of primary transcription response in leukemia”. In: *Science Advances* (2016), in press.

[160] Frank H Niesen, Helena Berglund, and Masoud Vedadi. “The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability”. In: *Nature Protocols* 2.9 (2007), pp. 2212–2221.

[161] Kenji Schorpp et al. “Identification of Small-Molecule Frequent Hitters from AlphaScreen High-Throughput Screens”. In: *Journal of Biomolecular Screening* 19.5 (2014), pp. 715–726.

[162] Martin Philpott et al. “Bromodomain-peptide displacement assays for interactome mapping and inhibitor discovery.” In: *Molecular bioSystems* 7.10 (2011), pp. 2899–2908.

[163] Rajesh Ghai, Robert J. Falconer, and Brett M. Collins. "Applications of isothermal titration calorimetry in pure and applied research—survey of the literature from 2010". In: *Journal of Molecular Recognition* 25.1 (2012), pp. 32–52.

[164] Stephanie Leavitt and Ernesto Freire. "Direct measurement of protein binding energetics by isothermal titration calorimetry". In: *Current Opinion in Structural Biology* 11.5 (2001), pp. 560–566.

[165] Patricia C. Weber and F. Raymond Salemme. "Applications of calorimetric methods to drug discovery and the study of protein interactions". In: *Current Opinion in Structural Biology* 13.1 (2003), pp. 115–121.

[166] Matthew W. Freyer and Edwin A. Lewis. "Isothermal Titration Calorimetry: Experimental Design, Data Analysis, and Probing Macromolecule/Ligand Binding and Kinetic Interactions". In: vol. 84. Methods in Cell Biology. Academic Press, 2008, pp. 79–113.

[167] John E. Ladbury, Gerhard Klebe, and Ernesto Freire. "Adding calorimetric data to decision making in lead discovery: a hot tip". In: *Nature Review Drug Discovery* 9.1 (2010), pp. 23–27.

[168] A.L. Lavoisier and P.S. Laplace. "Memoire sur la Chaleur". In: *Memoir Acadademie Royale Scientifique* (1784), pp. 355–408.

[169] J. J. Christensen et al. "Entropy Titration. A Calorimetric Method for the Determination of ΔG, ΔH, and ΔS from a Single Thermometric Titration". In: *The Journal of Physical Chemistry* 70.6 (1966), pp. 2003–2010.

[170] L. D. Hansen, J. J. Christensen, and R. M. Izatt. "Entropy titration. A calorimetric method for the determination of ΔG°(K), ΔH° and ΔS°". In: *Chem. Commun. (London)* (3 1965), pp. 36–38.

[171] Norman V. Beaudette and Neal Langerman. "An improved method for obtaining thermal titration curves using micromolar quantities of protein". In: *Analytical Biochemistry* 90.2 (1978), pp. 693–704.

[172] R. L. Biltonen and Neal and Langerman. "Microcalorimetry for biological chemistry: Experimental design, data analysis, and interpretation". In: *Methods in Enzymology*. Vol. 61. Academic Press, 1979, pp. 287–318.

[173] Jonathan B. Chaires. "Calorimetry and Thermodynamics in Drug Design". In: *Annual Review of Biophysics* 37.1 (2008), pp. 135–151.

[174] D. G. Myszka et al. "The ABRF-MIRG'02 Study: Assembly State, Thermodynamic, and Kinetic Analysis of an Enzyme/Inhibitor Interaction". In: *Journal of Biomolecular Technology* 14.4 (2003), pp. 247–269.

[175] A.R. Leach. *Molecular Modelling: Principles and Applications*. Pearson Education. Prentice Hall, 2001.

[176] M. Tuckerman. *Statistical Mechanics: Theory and Molecular Simulation*. Oxford Graduate Texts. OUP Oxford, 2010.

[177] C. Chipot and A. Pohorille. *Free Energy Calculations: Theory and Applications in Chemistry and Biology*. Springer Series in Chemical Physics. Springer Berlin Heidelberg, 2007.

[178] Michael Levitt. "The birth of computational structural biology". In: *Nat Struct Mol Biol* 8.5 (2001), pp. 392–393.

[179] William L. Jorgensen et al. “Comparison of simple potential functions for simulating liquid water”. In: *The Journal of Chemical Physics* 79.2 (1983), pp. 926–935.

[180] W. F. van Gunsteren H. J. C. Berendsen J. P. M. Postma and J. Hermans. “Intermolecular Forces”. In: ed. by B. Pullman. Dordrecht: Reidel, 1981. Chap. Interaction models for water in relation to protein hydration, pp. 331–342.

[181] William L. Jorgensen and Jeffry D. Madura. “Temperature and size dependence for Monte Carlo simulations of TIP4P water”. In: *Molecular Physics* 56.6 (1985), pp. 1381–1392.

[182] 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”. In: *The Journal of Chemical Physics* 112.20 (2000), pp. 8910–8922.

[183] Michael R. Shirts and Vijay S. Pande. “Solvation free energies of amino acid side chain analogs for common molecular mechanics water models”. In: *The Journal of Chemical Physics* 122.13 (2005), 134508 EP.

[184] Saeed Izadi, Ramu Anandakrishnan, and Alexey V. Onufriev. “Building Water Models: A Different Approach”. In: *The Journal of Physical Chemistry Letters* 5.21 (2014), pp. 3863–3871.

[185] Frank H. Stillinger and Aneesur Rahman. “Improved simulation of liquid water by molecular dynamics”. In: *The Journal of Chemical Physics* 60.4 (1974), pp. 1545–1557.

[186] Loup Verlet. “Computer “Experiments” on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules”. In: *Phys. Rev.* 159.1 (1967), pp. 98–103.

[187] Mark James Abraham et al. “GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers”. In: *SoftwareX* 2 (2015), pp. 1–7.

[188] R.W Hockney, S.P Goel, and J.W Eastwood. “Quiet high-resolution computer models of a plasma”. In: *Journal of Computational Physics* 14.2 (1974), pp. 148–158.

[189] Herman J. C. Berendsen and Wilfred F. Van Gunsteren. “Practical algorithms for dynamic simulations”. In: *Molecular-dynamics simulation of statistical-mechanical systems* (1986), pp. 43–65.

[190] William C. Swope et al. “A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters”. In: *The Journal of Chemical Physics* 76.1 (1982), pp. 637–649.

[191] D. Beeman. “Some multistep methods for use in molecular dynamics calculations”. In: *Journal of Computational Physics* 20.2 (1976), pp. 130–139.

[192] W. F. Van Gunsteren and H. J. C. Berendsen. “A Leap-frog Algorithm for Stochastic Dynamics”. In: *Molecular Simulation* 1.3 (1988), pp. 173–185.

[193] N. Goga et al. “Efficient Algorithms for Langevin and DPD Dynamics”. In: *Journal of Chemical Theory and Computation* 8.10 (2012), pp. 3637–3649.

[194] Shuichi Miyamoto and Peter A. Kollman. “Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models”. In: *Journal of Computational Chemistry* 13.8 (1992), pp. 952–962.

[195] Berk Hess et al. “LINCS: A linear constraint solver for molecular simulations”. In: *Journal of Computational Chemistry* 18.12 (1997), pp. 1463–1472.

[196] Berk Hess. “P-LINCS: A parallel linear constraint solver for molecular simulation”. In: *Journal of Chemical Theory and Computation* 4.1 (2008), pp. 116–122.

[197] L.V. Woodcock. “Isothermal molecular dynamics calculations for liquid salts”. In: *Chemical Physics Letters* 10.3 (1971), pp. 257–261.

[198] H. J. C. Berendsen et al. “Molecular dynamics with coupling to an external bath”. In: *The Journal of Chemical Physics* 81.8 (1984), pp. 3684–3690.

[199] Hans C. Andersen. “Molecular dynamics simulations at constant pressure and/or temperature”. In: *The Journal of Chemical Physics* 72.4 (1980), pp. 2384–2393.

[200] Shuichi Nosé. “A molecular dynamics method for simulations in the canonical ensemble”. In: *Molecular Physics* 52.2 (1984), pp. 255–268.

[201] Stephen C. Harvey, Robert K.Z. Tan, and Thomas E. Cheatham. “The flying ice cube: Velocity rescaling in molecular dynamics leads to violation of energy equipartition”. In: *Journal of Computational Chemistry* 19.7 (1998), pp. 726–740.

[202] M. Parrinello and A. Rahman. “Polymorphic transitions in single crystals: A new molecular dynamics method”. In: *Journal of Applied Physics* 52.12 (1981), pp. 7182–7190.

[203] David Van Der Spoel et al. “GROMACS: Fast, flexible, and free”. In: *Journal of Computational Chemistry* 26.16 (2005), pp. 1701–1718.

[204] Sander Pronk et al. “GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit”. In: *Bioinformatics* 29.7 (2013), pp. 845–854.

[205] Michael R. Shirts et al. “Accurate and Efficient Corrections for Missing Dispersion Interactions in Molecular Simulations”. In: *The Journal of Physical Chemistry B* 111.45 (2007), pp. 13052–13063.

[206] Christian L. Wennberg et al. “Direct-Space Corrections Enable Fast Accurate Lorentz-Berthelot Combination Rule Lennard-Jones Lattice Summation”. In: *Journal of Chemical Theory and Computation* 11.12 (2015), pp. 5737–5746.

[207] Michael Bergdorf, Christine Peter, and Philippe H. Hünenberger. “Influence of cut-off truncation and artificial periodicity of electrostatic interactions in molecular simulations of solvated ions: A continuum electrostatics study”. In: *The Journal of Chemical Physics* 119.17 (2003), pp. 9129–9144.

[208] P. P. Ewald. “Die Berechnung optischer und elektrostatischer Gitterpotentiale”. In: *Annalen der Physik* 369.3 (1921), pp. 253–287.

[209] Tom Darden, Darrin York, and Lee Pedersen. “Particle mesh Ewald: An N log(N) method for Ewald sums in large systems”. In: *The Journal of Chemical Physics* 98.12 (1993), pp. 10089–10092.

[210] Ulrich Essmann et al. “A smooth particle mesh Ewald method”. In: *The Journal of Chemical Physics* 103.19 (1995), pp. 8577–8593.

[211] J.W. Eastwood, R.W. Hockney, and D.N. Lawrence. "P3M3DP—The three-dimensional periodic particle-particle/ particle-mesh program". In: *Computer Physics Communications* 19.2 (1980), pp. 215–261.

[212] Brock A. Luty et al. "A Comparison of Particle-Particle, Particle-Mesh and Ewald Methods for Calculating Electrostatic Interactions in Periodic Molecular Systems". In: *Molecular Simulation* 14.1 (1994), pp. 11–20.

[213] Gabriel J. Rocklin et al. "Calculating the binding free energies of charged species based on explicit-solvent simulations employing lattice-sum methods: An accurate correction scheme for electrostatic finite-size effects". In: *The Journal of Chemical Physics* 139.18 (2013), 184103 EP.

[214] Nicholas Metropolis and S. Ulam. "The Monte Carlo Method". In: *Journal of the American Statistical Association* 44.247 (1949), pp. 335–341.

[215] Nicholas Metropolis et al. "Equation of State Calculations by Fast Computing Machines". In: *The Journal of Chemical Physics* 21.6 (1953), pp. 1087–1092.

[216] *ProtoMS v3.2*. 2016.

[217] D.J. Adams. "Grand canonical ensemble Monte Carlo for a Lennard-Jones fluid". In: *Molecular Physics* 29.1 (1975), pp. 307–311.

[218] Tami J. Marrone et al. "Solvation studies of DMP323 and A76928 bound to HIV protease: Analysis of water sites using grand canonical Monte Carlo simulations". In: *Protein Science* 7.3 (1998), pp. 573–579.

[219] R. H. Swendsen and J. S. Wang. "Replica monte carlo simulation of spin glasses". In: *Physical Review Letters* 57.21 (1986), pp. 2607–2609.

[220] Yuko Okamoto. "Generalized-ensemble algorithms: Enhanced sampling techniques for Monte Carlo and molecular dynamics simulations". In: *Journal of Molecular Graphics and Modelling* 22.5 (2004), pp. 425–439.

[221] Rafael C. Bernardi, Marcelo C. R. Melo, and Klaus Schulten. "Enhanced sampling techniques in molecular dynamics simulations of biological systems". In: *Biochimica et Biophysica Acta* 1850.5 (2014), pp. 872–877.

[222] David J. Earl and Michael W. Deem. "Parallel Tempering: Theory, Applications, and New Perspectives". In: *Physical Chemistry Chemical Physics* 7 (2005), pp. 3910–3916.

[223] Jose D. Faraldo-Gomez and Benoît Roux. "Characterization of conformational equilibria through Hamiltonian and temperature replica-exchange simulations: Assessing entropic and environmental effects". In: *Journal of Computational Chemistry* 28.10 (2007), pp. 1634–1647.

[224] Cameron Abrams and Giovanni Bussi. "Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration". In: *Entropy* 16.1 (2014), pp. 163–199.

[225] John D. Chodera and Michael R. Shirts. "Replica exchange and expanded ensemble simulations as Gibbs sampling: Simple improvements for enhanced mixing". In: *The Journal of Chemical Physics* 135.19 (2011).

[226] Michael K. Gilson et al. "The statistical-thermodynamic basis for computation of binding affinities: a critical review". In: *Biophysical Journal* 72.3 (1997), pp. 1047–1069.

[227] Michael R. Shirts. “Best Practices in Free Energy Calculations for Drug Design”. In: *Computational Drug Discovery and Design*. Ed. by Riccardo Baron. Vol. 819. Methods in Molecular Biology. Springer New York, 2012, pp. 425–467.

[228] John D. Chodera et al. “Alchemical free energy methods for drug discovery: progress and challenges”. In: *Current Opinion in Structural Biology* 21.2 (2011), pp. 150–160.

[229] David L. Mobley and Ken A. Dill. “Binding of Small-Molecule Ligands to Proteins: “What You See”Is Not Always “What You Get””. In: *Structure* 17.4 (2009), pp. 489–498.

[230] Samuel Genheden and Ulf Ryde. “The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities”. In: *Expert Opinion on Drug Discovery* 10.5 (2015), pp. 449–461.

[231] Samuel Genheden and Ulf Ryde. “Comparison of end-point continuum-solvation methods for the calculation of protein–ligand binding free energies”. In: *Proteins: Structure, Function, and Bioinformatics* 80.5 (2012), pp. 1326–1342.

[232] Gavin E. Crooks. “Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences”. In: *Physical Review E* 60 (3 1999), pp. 2721–2726.

[233] Stefan Boresch et al. “Absolute Binding Free Energies: A Quantitative Approach for Their Calculation”. In: *The Journal of Physical Chemistry B* 107.35 (2003), pp. 9535–9551.

[234] Robert W. Zwanzig. “High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases”. In: *The Journal of Chemical Physics* 22.8 (1954), pp. 1420–1426.

[235] Nandou Lu, Jayant K. Singh, and David A. Kofke. “Appropriate methods to combine forward and reverse free-energy perturbation averages”. In: *The Journal of Chemical Physics* 118.7 (2003), pp. 2977–2984.

[236] 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”. In: *The Journal of Chemical Physics* 122.14 (2005), 144107 EP.

[237] Charles H. Bennett. “Efficient estimation of free energy differences from Monte Carlo data”. In: *Journal of Computational Physics* 22.2 (1976), pp. 245–268.

[238] Michael R. Shirts et al. “Equilibrium Free Energies from Nonequilibrium Measurements Using Maximum-Likelihood Methods”. In: *Phys. Rev. Lett.* 91 (14 2003), 140601 EP.

[239] F. Marty Ytreberg, Robert H. Swendsen, and Daniel M. Zuckerman. “Comparison of free energy methods for molecular systems”. In: *The Journal of Chemical Physics* 125.18, 184114 (2006).

[240] Michael R. Shirts and John D. Chodera. “Statistically optimal analysis of samples from multiple equilibrium states”. In: *The Journal of Chemical Physics* 129.12 (2008), 124105 EP.

[241] Nadine Homeyer and Holger Gohlke. "Free Energy Calculations by the Molecular Mechanics Poisson-Boltzmann Surface Area Method". In: *Molecular Informatics* 31.2 (2012), pp. 114–122.

[242] Nad'a Špačková et al. "Molecular Dynamics Simulations and Thermodynamics Analysis of DNA-Drug Complexes. Minor Groove Binding between 4',6-Diamidino-2-phenylindole and DNA Duplexes in Solution". In: *Journal of the American Chemical Society* 125.7 (2003), pp. 1759–1769.

[243] David A. Pearlman. "Evaluating the Molecular Mechanics Poisson-Boltzmann Surface Area Free Energy Method Using a Congeneric Series of Ligands to p38 MAP Kinase". In: *Journal of Medicinal Chemistry* 48.24 (2005), pp. 7796–7807.

[244] M. Lepšík et al. "Efficiency of a second-generation HIV-1 protease inhibitor studied by molecular dynamics and absolute binding free energy calculations". In: *Proteins: Structure, Function, and Bioinformatics* 57.2 (2004), pp. 279–293.

[245] Giulio Rastelli et al. "Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA". In: *Journal of Computational Chemistry* 31.4 (2010), pp. 797–810.

[246] Tianyi Yang et al. "Virtual screening using molecular simulations". In: *Proteins: Structure, Function and Bioinformatics* 79.6 (2011), pp. 1940–1951.

[247] Hannes G. Wallnoefer, Klaus R. Liedl, and Thomas Fox. "A challenging system: Free energy prediction for factor Xa". In: *Journal of Computational Chemistry* 32.8 (2011), pp. 1743–1752.

[248] Aaron Weis et al. "Ligand affinities predicted with the MM/PBSA method: Dependence on the simulation method and the force field". In: *Journal of Medicinal Chemistry* 49.22 (2006), pp. 6596–6606.

[249] Lili Duan, Xiao Liu, and John Z. H. Zhang. "Interaction entropy: A new paradigm for highly efficient and reliable computation of protein-ligand binding free energy". In: *Journal of the American Chemical Society* 138.17 (2016), pp. 5722–5728.

[250] Dennis A. Holt et al. "Design, synthesis, and kinetic evaluation of high-affinity FKBP ligands and the X-ray crystal structures of their complexes with FKBP12". In: *Journal of the American Chemical Society* 115.22 (1993), pp. 9925–9938.

[251] Michael R. Shirts. "Calculating precise and accurate free energies in biomolecular systems". PhD thesis. 2005.

[252] Jiyao Wang, Yuqing Deng, and Benoît Roux. "Absolute Binding Free Energy Calculations Using Molecular Dynamics Simulations with Restraining Potentials". In: *Biophysical Journal* 91.8 (2006), pp. 2798–2814.

[253] Surjit B. Dixit and Christophe Chipot. "Can Absolute Free Energies of Association Be Estimated from Molecular Mechanical Simulations? The Biotin−Streptavidin System Revisited". In: *The Journal of Physical Chemistry A* 105.42 (2001), pp. 9795–9799.

[254] Dian Jiao et al. "Calculation of protein–ligand binding free energy by using a polarizable potential". In: *Proceedings of the National Academy of Sciences* 105.17 (2008), pp. 6290–6295.

[255] Chunfeng Zhao, David A. Caplan, and Sergei Yu. Noskov. “Evaluations of the Absolute and Relative Free Energies for Antidepressant Binding to the Amino Acid Membrane Transporter LeuT with Free Energy Simulations”. In: *Journal of Chemical Theory and Computation* 6.6 (2010), pp. 1900–1914.

[256] Helen M. Berman et al. “The Protein Data Bank”. In: *Nucleic Acids Research* 28.1 (2000), pp. 235–242.

[257] G. Vriend. “WHAT IF: A molecular modeling and drug design program”. In: *Journal of Molecular Graphics* 8.1 (1990), pp. 52–56.

[258] G. Chinea et al. “The use of position-specific rotamers in model building by homology”. In: *Proteins: Structure, Function, and Bioinformatics* 23.3 (1995), pp. 415–421.

[259] Junmei Wang et al. “Development and testing of a general amber force field”. In: *Journal of Computational Chemistry* 25.9 (2004), pp. 1157–1174.

[260] Araz Jakalian et al. “Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method”. In: *Journal of computational chemistry* 21.2 (2000), pp. 132–146.

[261] Araz Jakalian, David B. Jack, and Christopher I. Bayly. “Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation”. In: *Journal of Computational Chemistry* 23.16 (2002), pp. 1623–1641.

[262] Junmei Wang et al. “Automatic atom type and bond type perception in molecular mechanical calculations”. In: *Journal of Molecular Graphics and Modelling* 25.2 (2006), pp. 247–260.

[263] Hannes H. Loeffler, Julien Michel, and Christopher Woods. “FESetup: Automating Setup for Alchemical Free Energy Simulations”. In: *Journal of Chemical Information and Modeling* 55.12 (2015), pp. 2485–2490.

[264] Alan da Silva and Wim Vranken. “ACPYPE - AnteChamber PYthon Parser interfacE”. In: *BMC Research Notes* 5.1 (2012), pp. 367–367.

[265] Kresten Lindorff-Larsen et al. “Improved side-chain torsion potentials for the Amber ff99SB protein force field”. In: *Proteins: Structure, Function, and Bioinformatics* 78.8 (2010), pp. 1950–1958.

[266] J. Clayden, N. Greeves, and S. Warren. *Organic Chemistry*. Oxford University Press, 2012.

[267] John Mongan and David A Case. “Biomolecular simulations at constant pH”. In: *Current Opinion in Structural Biology* 15.2 (Apr. 2005), pp. 157–163.

[268] Garrett B. Goh et al. “Constant pH molecular dynamics of proteins in explicit solvent with proton tautomerism”. In: *Proteins: Structure, Function, and Bioinformatics* 82.7 (2014), pp. 1319–1331.

[269] Juyong Lee, Benjamin T. Miller, and Bernard R. Brooks. “Computational scheme for pH-dependent binding free energy calculation with explicit solvent”. In: *Protein Science* 25.1 (2016), pp. 231–243.

[270] Berk Hess et al. “GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation”. In: *Journal of Chemical Theory and Computation* 4.3 (2008), pp. 435–447.

[271] Thomas C. Beutler et al. “Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations”. In: *Chemical Physics Letters* 222.6 (1994), pp. 529–539.

[272] Nathan A. Baker et al. “Electrostatics of nanosystems: Application to microtubules and the ribosome”. In: *Proceedings of the National Academy of Sciences* 98.18 (2001), pp. 10037–10041.

[273] Pavel V. Klimovich, Michael R. Shirts, and David L. Mobley. “Guidelines for the analysis of free energy calculations”. In: *Journal of Computer-Aided Molecular Design* 29.5 (2015), pp. 397–411.

[274] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. Vol. 5. Chapman and Hall/CRC, 1994.

[275] David L. Mobley, John D. Chodera, and Ken A. Dill. “On the use of orientational restraints and symmetry corrections in alchemical free energy calculations”. In: *The Journal of Chemical Physics* 125.8 (2006), 084902 EP.

[276] Viktor Hornak et al. “Comparison of multiple Amber force fields and development of improved protein backbone parameters”. In: *Proteins: Structure, Function, and Bioinformatics* 65.3 (2006), pp. 712–725.

[277] Paul R. Gerber and Klaus Müller. “MAB, a generally applicable molecular force field for structure modelling in medicinal chemistry”. In: *Journal of Computer-Aided Molecular Design* 9.3 (1995), pp. 251–268.

[278] Bruce L. Bush, Christopher I. Bayly, and Thomas A. Halgren. “Consensus bond-charge increments fitted to electrostatic potential or field of many compounds: Application to MMFF94 training set”. In: *Journal of Computational Chemistry* 20.14 (1999), pp. 1495–1516.

[279] Panagis Filippakopoulos et al. “Benzodiazepines and benzotriazepines as protein interaction inhibitors targeting bromodomains of the BET family”. In: *Bioorganic & Medicinal Chemistry* 20.6 (2012), pp. 1878–1886.

[280] Lewis R. Vidler et al. “Discovery of novel small-molecule inhibitors of BRD4 using structure-based virtual screening.” In: *Journal of medicinal chemistry* 56.20 (2013), pp. 8073–88.

[281] Markus K. Dahlgren et al. “Characterization of Biaryl Torsional Energetics and its Treatment in OPLS All-Atom Force Fields”. In: *Journal of Chemical Information and Modeling* 53.5 (2013), pp. 1191–1199.

[282] Lei Huang and Benoît Roux. “Automated Force Field Parameterization for Non-Polarizable and Polarizable Atomic Models Based on Ab Initio Target Data.” In: *Journal of chemical theory and computation* 9.8 (2013), pp. 3543–3556.

[283] Peter Politzer et al. “An overview of halogen bonding”. In: *Journal of Molecular Modeling* 13.2 (2007), pp. 305–311.

[284] Lydia C. Gilday et al. “Halogen Bonding in Supramolecular Chemistry”. In: *Chemical Reviews* 115.15 (2015), pp. 7118–7195.

[285] Peter Politzer, Jane S. Murray, and Timothy Clark. “Halogen bonding: an electrostatically-driven highly directional noncovalent interaction”. In: *Physical Chemistry Chemical Physics* 12.28 (2010), pp. 7748–7757.

[286] Mahmoud A. A. Ibrahim. “Molecular mechanical study of halogen bonding in drug discovery”. In: *Journal of Computational Chemistry* 32.12 (2011), pp. 2564–2574.

[287] Michal Kolář and Pavel Hobza. “On Extension of the Current Biomolecular Empirical Force Field for the Description of Halogen Bonds”. In: *Journal of Chemical Theory and Computation* 8.4 (2012), pp. 1325–1333.

[288] Edward Harder et al. “OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins.” In: *Journal of Chemical Theory and Computation* 12.1 (2015), pp. 281–296.

[289] M. J. Frisch et al. *Gaussian 09*. Gaussian Inc. Wallingford CT 2009.

[290] Gabriel J. Rocklin, David L. Mobley, and Ken A. Dill. “Calculating the Sensitivity and Robustness of Binding Free Energy Calculations to Force Field Parameters”. In: *Journal of Chemical Theory and Computation* 9.7 (2013), pp. 3072–3083.

[291] David L. Mobley et al. “Predictions of Hydration Free Energies from All-Atom Molecular Dynamics Simulations”. In: *The Journal of Physical Chemistry B* 113.14 (2009), pp. 4533–4537.

[292] Stephen J. Fox et al. “Free Energies of Binding from Large-Scale First-Principles Quantum Mechanical Calculations: Application to Ligand Hydration Energies”. In: *The Journal of Physical Chemistry B* 117.32 (2013), pp. 9478–9485.

[293] David J. Huggins, Woody Sherman, and Bruce Tidor. “Rational approaches to improving selectivity in drug design.” In: *Journal of medicinal chemistry* 55.4 (2012), pp. 1424–44.

[294] Jens Uwe Peters. “Polypharmacology - Foe or friend?” In: *Journal of Medicinal Chemistry* 56.22 (2013), pp. 8955–8971.

[295] Mindy I. Davis et al. “Comprehensive analysis of kinase inhibitor selectivity”. In: *Nat Biotech* 29.11 (2011), pp. 1046–1051.

[296] Joost C. M. Uitdehaag et al. “A guide to picking the most selective kinase inhibitor tool compounds for pharmacological validation of drug targets”. In: *British Journal of Pharmacology* 166.3 (2012), pp. 858–876.

[297] Stefan Henrich et al. “Comparative binding energy analysis for binding affinity and target selectivity prediction”. In: *Proteins: Structure, Function and Bioinformatics* 78.1 (2010), pp. 135–153.

[298] Ci Christopher I. Bayly et al. “A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model”. In: *Journal of Physical Chemistry* 97.40 (1993), pp. 10269–10280.

[299] Enguerran Vanquelef et al. “R.E.D. Server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments”. In: *Nucleic Acids Research* 39.suppl 2 (2011), W511–W517.

[300] Robin M. Betz and Ross C. Walker. “Paramfit: Automated optimization of force field parameters for molecular dynamics simulations”. In: *Journal of Computational Chemistry* 36.2 (2015), pp. 79–87.

[301] Szilárd Páll and Berk Hess. “A flexible algorithm for calculating pair interactions on SIMD architectures”. In: *Computer Physics Communications* 184.12 (2013), pp. 2641–2650.

[302] Sreeja Parameswaran and David L. Mobley. "Box size effects are negligible for solvation free energies of neutral solutes." In: *Journal of Computer-Aided Molecular Design* 28.8 (2014), pp. 825–829.

[303] Schrödinger LLC. *The PyMOL Molecular Graphics System, Version 1.4.*

[304] U. Chandra Singh and Peter A. Kollman. "An approach to computing electrostatic charges for molecules". In: *Journal of Computational Chemistry* 5.2 (1984), pp. 129–145.

[305] Brent H. Besler, Kenneth M. Merz, and Peter A. Kollman. "Atomic charges derived from semiempirical methods". In: *Journal of Computational Chemistry* 11.4 (1990), pp. 431–439.

[306] Wendy D. Cornell et al. "Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation". In: *Journal of the American Chemical Society* 115.21 (1993), pp. 9620–9631.

[307] David L. Mobley et al. "Comparison of Charge Models for Fixed-Charge Force Fields: Small-Molecule Hydration Free Energies in Explicit Solvent". In: *The Journal of Physical Chemistry B* 111.9 (2007), pp. 2242–2254.

[308] Yong Duan et al. "A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations". In: *Journal of Computational Chemistry* 24.16 (2003), pp. 1999–2012.

[309] Jacopo Tomasi, Benedetta Mennucci, and Roberto Cammi. "Quantum Mechanical Continuum Solvation Models". In: *Chemical Reviews* 105.8 (2005), pp. 2999–3094.

[310] Eric Hansen et al. "Anomeric Effects in Sulfamides". In: *The Journal of Physical Chemistry A* 120.20 (2016), pp. 3677–3682.

[311] Gaetano Calabrò et al. "Elucidation of Nonadditive Effects in Protein–Ligand Binding Energies: Thrombin as a Case Study". In: *The Journal of Physical Chemistry B* 120.24 (2016), pp. 5340–5350.

[312] Lauren Wickstrom et al. "Parameterization of an effective potential for protein–ligand binding from host–guest affinity data". In: *Journal of Molecular Recognition* 29.1 (2016), pp. 10–21.

[313] Lingle Wang, B. J. Berne, and Richard A. Friesner. "On achieving high accuracy and reliability in the calculation of relative protein–ligand binding affinities". In: *Proceedings of the National Academy of Sciences* 109.6 (2012), pp. 1937–1942.

[314] Lingle Wang et al. "Modeling local structural rearrangements using FEP/REST: Application to relative binding affinity predictions of CDK2 inhibitors". In: *Journal of Chemical Theory and Computation* 9.2 (2013), pp. 1282–1293.

[315] Christopher M. Baker. "Polarizable force fields for molecular dynamics simulations of biomolecules". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 5.2 (2015), pp. 241–254.

[316] Li L. Duan, Guo Q. Feng, and Qing G. Zhang. "Large-scale molecular dynamics simulation: Effect of polarization on thrombin-ligand binding energy". In: *Scientific Reports* 6 (2016), 31488 EP.

[317] Yuan Hu et al. "The importance of protonation and tautomerization in relative binding affinity prediction: a comparison of AMBER TI and Schrödinger FEP". In: *Journal of Computer-Aided Molecular Design* 30.7 (2016), pp. 533–539.

[318] Andrew T. Fenley et al. “Bridging Calorimetry and Simulation through Precise Calculations of Cucurbituril-Guest Binding Enthalpies”. In: *Journal of Chemical Theory and Computation* 10.9 (2014), pp. 4069–4078.

[319] Alberto Perez et al. “Advances in free-energy-based simulations of protein folding and ligand binding”. In: *Current Opinion in Structural Biology* 36 (2016), pp. 25–31.

[320] M. Mitchell Waldrop. “The chips are down for Moore’s law”. In: *Nature* 530 (2016), pp. 144–147.

[321] Per Larsson, Berk Hess, and Erik Lindahl. “Algorithm improvements for molecular dynamics simulations”. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.1 (2011), pp. 93–108.

[322] Michael Shirts and Vijay S. Pande. “Screen Savers of the World Unite!” In: *Science* 290.5498 (Dec. 2000), pp. 1903–1904.

[323] Michael C. Schatz, Ben Langmead, and Steven L. Salzberg. “Cloud computing and the DNA data race”. In: *Nature Biotechnology* 28.7 (2010), pp. 691–693.

[324] Joseph L. Hellerstein, Kai J. Kohlhoff, and David E. Konerding. “Science in the cloud: Accelerating discovery in the 21st century”. In: *IEEE Internet Computing* 16.4 (2012), pp. 64–68.

[325] Nadine Homeyer et al. “Binding Free Energy Calculations for Lead Optimization: Assessment of Their Accuracy in an Industrial Drug Design Context”. In: *Journal of Chemical Theory and Computation* 10.8 (2014), pp. 3331–3344.

[326] Julien Michel and Jonathan W. Essex. “Prediction of protein–ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations”. In: *Journal of Computer-Aided Molecular Design* 24.8 (2010), pp. 639–658.

[327] Niels Hansen and Wilfred F. van Gunsteren. “Practical Aspects of Free-Energy Calculations: A Review”. In: *Journal of Chemical Theory and Computation* 10.7 (2014).

[328] Huiyong Sun et al. “Assessing the performance of MM/PBSA and MM/GBSA methods. 4. Accuracies of MM/PBSA and MM/GBSA methodologies evaluated by various simulation protocols using PDBbind data set.” In: *Physical Chemistry Chemical Physics* 16.31 (2014), pp. 16719–16729.

[329] Huiyong Sun et al. “Assessing the Performance of MM/PBSA and MM/GBSA Methods. 5. Improved Docking Performance by Using High Solute Dielectric Constant MM/GBSA and MM/PBSA Rescoring”. In: *Physical Chemistry Chemical Physics* 16 (2014), pp. 22035–22045.

[330] S. Kashif Sadiq et al. “Accurate ensemble molecular dynamics binding free energy ranking of multidrug-resistant HIV-1 proteases”. In: *Journal of Chemical Information and Modeling* 50.5 (2010), pp. 890–905.

[331] Paulette a. Greenidge et al. “MM/GBSA binding energy prediction on the PDBbind data set: Successes, failures, and directions for further improvement”. In: *Journal of Chemical Information and Modeling* 53.1 (2013), pp. 201–209.

[332] Thomas B. Steinbrecher et al. “Accurate Binding Free Energy Predictions in Fragment Optimization”. In: *Journal of Chemical Information and Modeling* 55.11 (2015), pp. 2411–2420.

[333] Samuel Genheden et al. "The Normal-Mode Entropy in the MM/GBSA Method: Effect of System Truncation, Buffer Region, and Dielectric Constant". In: *Journal of Chemical Information and Modeling* 52.8 (2012), pp. 2079–2088.

[334] Irene Maffucci and Alessandro Contini. "Explicit ligand hydration shells improve the correlation between MM-PB/GBSA binding energies and experimental activities". In: *Journal of Chemical Theory and Computation* 9.6 (2013), pp. 2706–2717.

[335] C. Paissoni et al. "GMXPBSA 2.1: A GROMACS tool to perform MM/PBSA and computational alanine scanning". In: *Computer Physics Communications* 186 (2015), pp. 105–107.

[336] Naveen Michaud-Agrawal et al. "MDAnalysis: A toolkit for the analysis of molecular dynamics simulations". In: *Journal of Computational Chemistry* 32.10 (2011), pp. 2319–2327.

[337] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and their Application*. Cambridge University Press, 1997.

[338] Julien Michel, Julian Tirado-Rives, and William L. Jorgensen. "Energetics of Displacing Water Molecules from Protein Binding Sites: Consequences for Ligand Optimization". In: *Journal of the American Chemical Society* 131.42 (2009), pp. 15403–15411.

[339] Zheng Li and Themis Lazaridis. "The effect of water displacement on binding thermodynamics: Concanavalin A". In: *Journal of Physical Chemistry B* 109.1 (2005), pp. 662–670.

[340] Friederike Vollmuth and Matthias Geyer. "Interaction of Propionylated and Butyrylated Histone H3 Lysine Marks with Brd4 Bromodomains". In: *Angewandte Chemie International Edition* 49.38 (2010), pp. 6768–6772.

[341] J. M. Word et al. "Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation." In: *Journal of Molecular Biology* 285.4 (1999), pp. 1735–1747.

[342] Mats H. M. Olsson et al. "PROPKA3: Consistent treatment of internal and surface residues in empirical p K a predictions". In: *Journal of Chemical Theory and Computation* 7.2 (2011), pp. 525–537.

[343] Chresten R. Søndergaard et al. "Improved treatment of ligands and coupling effects in empirical calculation and rationalization of p K a values". In: *Journal of Chemical Theory and Computation* 7.7 (2011), pp. 2284–2295.

[344] J. T. Berryman R. M. Betz Q. Cai D. S. Cerutti T. E. Cheatham I. I. I. T. A. Darden R. E. Duke H. Gohlke A. W. Goetz S. Gusarov N. Homeyer P. Janowski J. Kaus I. Kolossváry A. Kovalenko T. S. Lee S. LeGrand T. Luchko R. Luo B. X. Wu D.A. Case V. Babin and P. A. Kollman. *AMBER 14*. 2014.

[345] A. Ben-Naim and Y. Marcus. "Solvation thermodynamics of nonionic solutes". In: *The Journal of Chemical Physics* 81.4 (1984), pp. 2016–2027.

[346] J. A. Maier et al. "ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB". In: *Journal of Chemical Theory and Computation* 11 (2015), pp. 3696–3713.