UNIVERSITY OF SOUTHAMPTON

Faculty of Engineering and Physical Sciences
School of Chemistry

Evaluating and Improving the Robustness of Alchemical Binding Free Energy Calculations Using Adaptive Enhanced Sampling Methods

by

Miroslav Dikov Suruzhon

ORCiD: 0000-0002-6794-1679

A thesis for the degree of
Doctor of Philosophy

May 2022

University of Southampton

Abstract

Faculty of Engineering and Physical Sciences
School of Chemistry

Doctor of Philosophy

Evaluating and Improving the Robustness of Alchemical Binding Free Energy Calculations Using Adaptive Enhanced Sampling Methods

by Miroslav Dikov Suruzhon

Alchemical protein–ligand binding free energy calculations are currently a topic in computational chemistry which requires expert knowledge and a multitude of initial choices and parameters set by the researcher. While the impact of many of these decisions on the resulting free energy values has been explored in recent years, the influence of the initial protein crystal structure, as well as the protonation, tautomeric and rotameric states of the amino acid side chains on the free energy values have been underexplored. To perform these studies, a Python library (ProtoCaller) supporting an arbitrary level of automation for setting up and running binding free energy calculations is first developed and presented. Afterwards, it is shown that the choice of initial protein crystal structure can significantly impact the resulting free energy values at short timescales, while ligand rare events can induce discrepancies at longer timescales. Similarly, different initial histidine protonation, tautomeric and rotameric states can also result in free energy discrepancies, showcasing the need for enhanced sampling methods on protein and ligand degrees of freedom.

To address these sampling problems, an alchemical variant of the sequential Monte Carlo (SMC) enhanced sampling method is presented and validated on a range of test cases. This methodology is then augmented with long-timescale sampling provided by simulated tempering (ST), whose initial parameters are obtained from a preliminary exploratory SMC simulation and are afterwards refined over time in an adaptive fashion. The resulting method—fully adaptive simulated tempering (FAST)—is completely automatable and does not require any system-dependent parameters, making it generally applicable to the ligand sampling problem. Finally, FAST is applied to relative protein–ligand binding free energy calculations, enabling their full automation in combination with adaptive enhanced sampling. This methodology improves free energy reproducibility by decreasing the number of initial choices made by the researcher and can also be readily generalised to other sampling scenarios, making it a highly relevant contribution to the field.

Contents

List of Figures

List of Tables

Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:
   M. Suruzhon, T. Senapathi, M. S. Bodnarchuk, R. Viner, I. D. Wall, C. B. Barnett, K. J. Naidoo and J. W. Essex, *Journal of Chemical Information and Modeling*, 2020, **60**, 1917–1921.
   M. Suruzhon, M. S. Bodnarchuk, A. Ciancetta, R. Viner, I. D. Wall and J. W. Essex, *Journal of Chemical Theory and Computation*, 2021, **17**, 1806–1821.
   M. Suruzhon, M. L. Samways and J. W. Essex, in *Free Energy Methods in Drug Discovery: Current State and Future Directions*, ed. K. A. Armacost and D. C. Thompson, American Chemical Society, Washington, DC, 2021, pp. 109–125.

Signed:..... Date:.....

Acknowledgements

I wish to thank Prof. Jonathan Essex for the continuous supervision and support during my PhD, as well as giving me the freedom to pursue my own scientific interests and ideas. I also thank my industrial supervisors and collaborators for the helpful discussions and useful feedback: Dr Michael Bodnarchuk, Dr Russell Viner, Dr Ian Wall and Prof. Antonella Ciancetta. In addition, I would like to thank Prof. Kevin Naidoo, Dr Christopher Barnett and Dr Tharindu Senapathi for the collaborative effort during the development of ProtoCaller (Chapter 4).

Many thanks to all current and past members of the Essex group with whom I worked during my PhD for the continuous help, scientific discussions and friendly atmosphere. I would also like to give special thanks to Dr Marley Samways for his input during the writing of Chapter 3, and to Khaled Maksoud and Chi Cheng for the many interesting scientific conversations which gave me significant inspiration in my research.

I thank everyone who has been part of the Theory and Modelling in Chemical Sciences (TMCS) centre for doctoral training for giving me the opportunity and providing me with the foundational knowledge to embark on a computational chemistry PhD project. This also includes many of the students in the programme, mainly from the 2017–2018 cohort, from whom I have learned a lot.

Finally, I wish to thank all my friends, teachers and mentors who have supported me over the years—the list is too long to do everyone justice. In particular, I dedicate this thesis to my late grandmother, Olga, and to my mother, Svetlana, for shaping me into the person I am today.

Acronyms

λD λ-dynamics

AASMC adaptive alchemical sequential Monte Carlo

AFE alchemical free energy

AM1 Austin model 1

AMBER assisted model building with energy refinement

AMOEBA atomic multipole optimized energetics for biomolecular applications

BAR Bennett acceptance ratio

BCC bond charge correction

CCMA constant constraint matrix approximation

CDK2 cyclin-dependent kinase 2

CI confidence interval

CMA-ES covariance matrix adaptation evolution strategy

COM centre-of-mass

CPU central processing unit

CSC classical soft-core potential

CV collective variable

DFT density functional theory

DHFR dihydrofolate reductase

DQMC diffusion quantum Monte Carlo

EDS enveloping distribution sampling

ESS effective sample size

FAST fully adaptive simulated tempering

FEP free energy perturbation

ff14SB force field 14 Stony Brook

ff19SB force field 19 Stony Brook

ff99SB force field 99 Stony Brook

FXa coagulation factor Xa

GAFF general AMBER force field

GBSA generalised Born surface area

GCMC grand canonical Monte Carlo

GMM Gaussian mixture model

GPU graphics processing unit

GROMACS Groningen machine for chemical simulations

GSC Gaussian soft-core potential

HF Hartree-Fock theory

HMC Hamiltonian Monte Carlo

H-REMD Hamiltonian replica exchange

Hsp90 heat shock protein 90

InChI international chemical identifier

IST irreversible simulated tempering

ITS integrated tempering sampling

JAWS just add water molecules

KL Kullback–Leibler

L-BFGS limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm

LINCS linear constraint solver

LJ Lennard-Jones

MAD mean absolute deviation

MALA Metropolis-adjusted Langevin algorithm

MBAR multistate Bennett acceptance ratio

MC Monte Carlo

MCMC Markov chain Monte Carlo

MCS maximum common substructure

MD molecular dynamics

MM molecular mechanics

MMFF Merck molecular force field

MP2 second-order Møller-Plesset perturbation theory

MSM Markov state model

$NADP^+$ nicotinamide adenine dinucleotide phosphate

NADPH reduced nicotinamide adenine dinucleotide phosphate

NAMD nanoscale molecular dynamics

NCMC nonequilibrium candidate Monte Carlo

NMR nuclear magnetic resonance

OPC optimal point-charge

OPLS optimised potentials for liquid simulations

OPLS-AA optimised potentials for liquid simulations all-atom

PBSA Poisson–Boltzmann surface area

PCA principal component analysis

PDB Protein Data Bank

PLUMED plugin for molecular dynamics

PME particle mesh Ewald

PTP1B protein tyrosine phosphatase 1B

QUBE quantum mechanical bespoke force field

REMD replica exchange molecular dynamics

RESP restrained electrostatic potential

REST replica exchange with solute tempering

REST2 replica exchange with solute scaling

RMSD root-mean-square deviation

SAMPL statistical assessment of the modeling of proteins and ligands

SIR sequential importance resampling

SMC sequential Monte Carlo

SMILES simplified molecular-input line-entry system

SPC simple point-charge

ST simulated tempering

TGF-β transforming growth factor beta

TI thermodynamic integration

TICA time-lagged independent component analysis

TIP3P transferable intermolecular potential with 3 points

TIP4P transferable intermolecular potential with 4 points

UWHAM unbinned weighted histogram analysis method

Chapter 1

Introduction

1.1 Motivation

Drug discovery is a long and expensive process, which can involve the screening of hundreds of thousands of compounds over the course of 12–15 years.[9] Having access to reliable computational methods which can eliminate a large amount of this effort is therefore a highly desirable endeavour. Because of this, computational methods spanning multiple levels of sophistication have become increasingly more popular in drug discovery in the recent decades.[10]

There are various classes of computational methods, providing different levels of accuracy. At the highest levels one has "rigorous" methods, i.e. methods based on physical principles rather than statistical models. This thesis concentrates on one particular such class of methods—alchemical free energy (AFE) calculations. AFE methods are considerably more computationally expensive than most alternatives and can only handle a small number of potential drug candidates, making them primarily useful for the later stages of drug discovery.

The theoretical groundwork which preceded the first computationally performed AFE calculations was first developed by Kirkwood[11] in 1935, who introduced the idea of a coupling parameter, which interpolates between different thermodynamic states of interest and can be used to obtain the free energy difference between them. In 1954, Zwanzig[12] further extended this work by developing a method, known as the Zwanzig equation or free energy perturbation (FEP), which transformed the problem of calculating a ratio of partition functions into a sampling problem. This methodology could then be used to exactly obtain the free energy difference between any two thermodynamics states by means of computational sampling, as long as their Hamiltonian is known. The acceptance ratio method introduced by Bennett[13] in 1976 further improved on this idea by introducing a statistically optimal estimator of the free energy between two probability distributions.

As computational power increased in the next decades, the calculation of free energies of larger systems by computational means became feasible. The first application of FEP on solvation free energy calculations of organic molecules was published in 1985 by Jorgensen and Ravimohan [14], where the relative solvation free energies of methanol and ethane were calculated. The success of this result motivated subsequent retrospective [15] and prospective [16] work on the calculation of relative protein–ligand binding free energies by computational means. Since then, the applications of alchemical free energy methods have been increasing and the field has been maturing. However, despite the numerous improvements over the years and the theoretical rigour of AFE methods, they still suffer from issues regarding accuracy, precision and reproducibility, which undermine their practical utility.

Accuracy is related to the deviation between the calculated and the true free energy. Accuracy problems are inevitable, due to the quantum nature of chemistry, meaning that classical models are inherently limited in their description of chemical systems. On the other hand, quantum methods are currently too computationally expensive to have a large impact in drug discovery. Therefore, the only way the accuracy problem can be currently tackled is through the development of more sophisticated models using recent developments in machine learning. Developing such improved models often requires large-scale calculations and extensive expertise in computational chemistry, making it one of the most challenging areas of the field.

Precision is directly linked to the ability of the computational method to sample all relevant conformations of the molecular system. The precision problem arises due to the enormous complexity of even small biomolecular systems, where multiple energy minima separated by high kinetic barriers can co-exist and introduce variability and bias in any estimated observables. This problem is commonly addressed by various enhanced sampling methods, which help surmount these kinetic barriers using prior knowledge about the important degrees of freedom of the system of interest. Current enhanced sampling research is primarily concerned with the development of methods which are robust, efficient and sufficiently general at the same time, as well as the determination of the important degrees of freedom *a priori*.

Irreproducibility is another major problem in the field. Each biomolecular simulation involves many choices and decisions which are completely unrelated to the actual model but can nevertheless impact the results of the calculation by introducing difficult to detect biases. A poor level of reproducibility is concerning, as it calls into question the validity of a large fraction of the published literature, meaning that evaluating the impact of these subjective decisions on AFE calculations is extremely important. However, this area of research has been comparatively underexplored until recently.

Solving these three fundamental problems is the ultimate goal of this field and the next section will describe how this thesis relates to them.

1.2 Thesis Overview

The primary aim of this thesis is to investigate some aspects of the reproducibility problem and solve them if possible. This would not only improve the confidence in AFE methods and the quality of their results, but it would also point towards issues that need to be considered when developing new methods. Although easy to state, reproducibility is difficult to explore, mainly due to practical reasons: scientific software is specialised and often developed by completely different communities, making the software landscape largely scattered and incohesive. This makes such large-scale studies technically, albeit not conceptually, challenging.

Despite the technical challenges, there has recently been an increased interest in measuring the impact of various decisions made by the researcher on the resulting free energy calculations. The relevant developments in this area, as well as some important factors that have not yet been considered in the literature, are reviewed in Chapter 3. While these factors provide a good starting point for a research project, any such large-scale study needs to be facilitated by an appropriate computational tool.

The first part of the project was therefore to develop a Python library, which acts as an interface between different specialised pieces of software. The aim of this development was to create a framework which is robust and permits arbitrary degrees of customisation and flexibility in the scientific problems that can be studied by it. The development and structure of this library, ProtoCaller, is described in Chapter 4.

In Chapter 5, ProtoCaller is then used to perform a large-scale study on the impact of the initial crystal structure on binding free energy calculations. Although recent literature has hinted at the significance of the problem, there had been no comprehensive study addressing this question before this work. The practical implications of this study are significant, as they call into question many of the assumptions that are frequently made in alchemical free energy (AFE) calculations.

The study in Chapter 5 is afterwards continued in Chapter 6, where the protonation, tautomeric and rotameric states of the histidine side chains of two different protein systems will be investigated. Although it is commonly recognised that these can have a significant impact on the resulting free energies, the magnitude of this effect had not been known before this study. Another reason for undertaking this problem were some of the observations made in Chapter 6, where it was discovered that the initial crystal structure can also influence the setup process, including protonation and tautomeric side-chain states.

It will be shown in Chapter 5 that long-timescale sampling, as well as thorough exploration of the ligand degrees of freedom is vital for performing reliable and reproducible free energy calculations. This conclusion calls for the use of enhanced sampling methods in the context of free energy calculations. As the literature on this topic is vast, it warrants a separate review. In Chapter 7, tempering methods, which are a subclass of enhanced sampling methods, are presented in a concise fashion and comparison between the different algorithms will be drawn. These considerations will evaluate the utility of each tempering method with the goal of improving sampling of select few degrees of freedom over long timescales in the context of AFE calculations.

In Chapter 8, an adaptive alchemical version of the sequential Monte Carlo (SMC) method is presented and evaluated over a range of systems. The main advantage of adaptive alchemical sequential Monte Carlo (AASMC) is its ability to perform explorative simulations with minimal system-specific knowledge. Although Chapter 8 demonstrates that SMC is a valid sampling algorithm, the approach taken in this thesis was to only use SMC in the preliminary stages of the AFE calculation.

The combination of AASMC and another enhanced sampling method, simulated tempering (ST), is investigated in Chapter 9. In this chapter, AASMC is used as a preliminary exploratory method which provides various initial parameters for ST in a system-independent manner. This procedure is then refined as the simulation progresses, resulting in a fully adaptive simulated tempering (FAST) algorithm, which minimises the requirements for prior system-specific knowledge. FAST is then validated on a range of protein–ligand systems, and shown to significantly improve sampling in an automated way.

The theoretical framework presented in Chapter 9 is afterwards extended in Chapter 10 to the case of relative protein–ligand AFE calculations and these are also combined with enhanced sampling. This chapter shows that the resulting method solves all of the issues presented in Chapter 5, thereby improving the robustness of AFE calculations. The implications of this thesis, as well as possible future developments, are finally discussed in Chapter 11.

We now proceed to Chapter 2, which contains a brief introduction to the relevant theoretical methods common to all of the following chapters.

Chapter 2

Theoretical Background

2.1 Statistical Mechanics

2.1.1 Ensembles

Statistical mechanics is the branch of physics which studies emergent macroscopic properties arising from the behaviour of all constituent microscopic particles. Central to statistical mechanics is the idea that the internal energy (E) of a system is an extensive property, i.e. one which scales linearly with system size on a macroscopic scale. One can then express E as a multilinear function with respect to all extensive macroscopic observables of the system: the entropy (S, defined later), the volume (V), the number of particles (N) and others, depending on the system studied. With the help of the fundamental laws of thermodynamics, one can then include the following intensive conjugate observables: temperature (T), pressure (P) and chemical potential (μ) to postulate the relation:

$$E = TS - PV + \sum_i \mu_i N_i \tag{2.1}$$

where i iterates over all types of indistinguishable particles, whose number will henceforth be denoted as a vector $\vec{N}$. If we express the internal energy as a function of all extensive variables, the following scaling behaviour holds for an arbitrary scaling factor λ:

$$\lambda E(S, V, \vec{N}) = E(\lambda S, \lambda V, \lambda \vec{N}) \tag{2.2}$$

from which we can obtain the total differential of the internal energy:

$$dE(S,V,\vec{N}) = T(S,V,\vec{N})dS - P(S,V,\vec{N})dV + \sum_i \mu_i(S,V,\vec{N})dN_i \tag{2.3}$$

This relationship gives us a conditional conservation law: in a system with constant entropy, volume and number of particles, the total internal energy is conserved. This system is denoted as the microcanonical (NVE) ensemble.

In general, one can derive a multitude of statistical ensembles by expressing the relevant conserved quantity (the thermodynamic potential) as a function of any combination of intensive and extensive variables, provided that at least one variable is extensive. For our intents and purposes, we will concentrate on the canonical ensemble, where we keep the number of particles, system volume and temperature constant (NVT) and the isothermal–isobaric ensemble (NPT), where instead of the volume, we keep the pressure constant, much like in chemical reactions under laboratory conditions.

In the canonical ensemble the relevant conserved thermodynamic potential is the Helmholtz free energy F:

$$\begin{aligned} F &= E - TS = -PV + \sum_i \mu_i N_i \\ dF(T,V,\vec{N}) &= -S(T,V,\vec{N})dT - P(T,V,\vec{N})dV + \sum_i \mu_i(T,V,\vec{N})dN_i \end{aligned} \tag{2.4}$$

and in the isothermal–isobaric ensemble the conserved quantity is the Gibbs free energy G:

$$\begin{aligned} G &= E - TS + PV = \sum_i \mu_i N_i \\ dG(T,P,\vec{N}) &= -S(T,P,\vec{N})dT + V(T,P,\vec{N})dP + \sum_i \mu_i(T,P,\vec{N})dN_i \end{aligned} \tag{2.5}$$

After we have defined an ensemble, we are most interested in its underlying probability distribution defined over the phase space, spanned by the atom coordinates $\vec{x}$ and momenta $\vec{p}$. A central idea in statistical mechanics is that all states of equal energy are equally probable. This leads to the corollary that for a very large number of particles ($O(10^{23})$), the probability distribution leading to the highest state degeneracy will dominate over all other possible probability distributions. This measure of degeneracy is denoted as "entropy" in physics and as "information" in statistics and is proportional to minus the logarithm of the number of possible states W:

$$S = -k_B \ln W \tag{2.6}$$

where k_B is the Boltzmann constant. One can then postulate that a large number of microscopic systems make up a microcanonical system (the "universe"), which imposes constraints on the average number of particles, volume and internal energy. The distribution which maximises the entropy given these constraints in the case of the canonical ensemble is the Boltzmann distribution $\pi_{NVT}(\vec{x}, \vec{p})$:

$$\pi_{NVT}(\vec{x}, \vec{p}) = \frac{e^{-\beta E(\vec{x},\vec{p})}}{Z_{NVT}} \tag{2.7}$$

where $\beta \equiv 1/k_B T$ and Z_{NVT} is a normalisation constant, also known as the partition function:

$$Z_{NVT} = \frac{1}{h^{3N}} \int e^{-\beta E(\vec{x},\vec{p})} d\vec{x} d\vec{p} \tag{2.8}$$

where we assume distinguishable particles and h denotes Planck's constant. Since the above definitions can be readily extended to the NPT ensemble using the substitution $-\beta E(\vec{x}, \vec{p}) \to -\beta(E(\vec{x}, \vec{p}) + PV)$, we will henceforth concentrate on the NVT ensemble without loss of generality. In the following discussion we will also omit all ensemble subscripts.

Using the underlying probability distribution one can then calculate any observable as the average over the probability distribution (ensemble average). For example, the total macroscopic internal energy of the system E can be expressed as a weighted sum of the microscopic energies:

$$E \equiv \langle E(\vec{x}, \vec{p}) \rangle = \int E(\vec{x}, \vec{p}) \pi(E(\vec{x}, \vec{p})) d\vec{x} d\vec{p} = \int E(\vec{x}, \vec{p}) \frac{e^{-\beta E(\vec{x},\vec{p})}}{Z} d\vec{x} d\vec{p} \tag{2.9}$$

There is also a fundamental relationship between the relevant thermodynamic potential and the partition function of the equilibrium distribution defined by that potential:

$$F = -\beta^{-1} \ln Z \tag{2.10}$$

Even though the partition function and/or the free energy can in principle supply us with all relevant system information, the calculation of these quantities is impossible analytically, apart from the simplest cases, and mostly unfeasible numerically, although some efficient approaches exist for small systems.[17] Therefore, we will mainly be interested in free energy differences between e.g. two Hamiltonians A and B:

$$\Delta F_{AB} = -\beta^{-1} \ln \frac{Z_B}{Z_A} \tag{2.11}$$

2.1.2 Free Energy Estimators

All commonly used approaches to numerically solve Equation 2.11 convert the intractable problem of calculating the ratio of two partition functions into a sampling problem. In general, there are three classes of free energy calculation methods: free energy perturbation (FEP), thermodynamic integration (TI) and nonequilibrium free energy calculations. We will henceforth focus on FEP methods. We will also be working with the dimensionless potential energy $u(\vec{x})$, as by doing so we can generalise the treatment of all relevant thermodynamic ensembles, as well as dispose of the kinetic energy contribution to the free energy difference, which can be calculated analytically and is of no numerical interest.

The simplest way of evaluating the dimensionless free energy difference Δf_{AB} is by transforming the free energy calculation into a sampling problem using the trivial relation (Zwanzig equation):[12]

$$\begin{aligned} \Delta f_{AB} &= -\ln \frac{Z_B}{Z_A} = -\ln \frac{\int e^{-(u_B(\vec{x})-u_A(\vec{x}))} e^{-u_A(\vec{x})} d\vec{x}}{Z_A} \\ &\equiv -\ln \left\langle e^{-\Delta u_{AB}(\vec{x})} \right\rangle_A = \ln \left\langle e^{\Delta u_{AB}(\vec{x})} \right\rangle_B \end{aligned} \tag{2.12}$$

with $\Delta u_{AB}(\vec{x}) \equiv u_B(\vec{x}) - u_A(\vec{x})$. In practice, estimating Equation 2.12 is difficult unless $\Delta u_{AB}(\vec{x})$ exhibits low variance. Therefore, even though the final two relations in Equation 2.12 are formally equivalent, they will often give very different results in practice. There are two modifications to the method to remedy this: the first is to split a perturbation into N smaller perturbations dependent on a coupling parameter λ, such that $u_{\lambda_1}(\vec{x}) \equiv u_A(\vec{x})$ and $u_{\lambda_N}(\vec{x}) \equiv u_B(\vec{x})$:[14]

$$\Delta f_{AB} = \sum_{i=1}^{N} \Delta f_{\lambda_i \lambda_{i+1}} \tag{2.13}$$

and the second is to use a statistically optimal estimator. The minimum variance two-state estimator is known as the Bennett acceptance ratio (BAR)[13] and its multistate generalisation is commonly referred to as MBAR[18] or unbinned weighted histogram analysis method (UWHAM).[19] MBAR is not only asymptotically optimal, but is also a Rao–Blackwell estimator,[20] which maximises the likelihood of the observations given the statistical model.[18]

An overview of MBAR will be presented in Section 9.2.2. Although MBAR is a theoretically more efficient estimator than BAR, it requires MN^2 energy evaluations for N λ-windows and M frames per window. In contrast, BAR scales as $O(MN)$, which means that depending on the number of the intermediate states, BAR can therefore be more desirable, since MBAR typically results in similar values to those obtained with BAR.[21]

2.2 Energy Models

The main focus of this thesis will be solvated proteins which typically contain between 10^5 and 10^7 atoms. It is not possible to use high-level *ab initio* electronic structure methods in such large systems with the current level of computational power, meaning that electronic effects have to be handled approximately. The only quantum mechanical method that can be employed on small whole proteins is density functional theory (DFT)[22] but it is still a highly expensive method for single-point energy evaluations. Since free energy calculations involve ensemble averages, we need to be able to not only evaluate thousands of single-point energies but also have a procedure to generate new samples. The latter usually involves force calculations which increase the computational complexity of the problem even further.

Therefore, we will opt for classical methods where electronic effects are coarse-grained and accounted for by empirical parameters. This is justified for insulators, such as most proteins, where electron mobility is expected to be low and the ground electronic state is expected to dominate most of the physics. The most widespread type of classical models is the atomistic force field, where only nuclear motions are accounted for. While most atomistic force fields have similar features, in this discussion we will concentrate on one particular model, the AMBER force field,[23] which will be exclusively considered hereafter.

2.2.1 Functional Form

One of the approximations of the AMBER force field is the splitting of the energy function in two-body terms with the exceptions of bonded systems, where three- and four-body effects are also accounted for. The total energy function can be expressed as a sum of bonded, angular, torsional, van der Waals and electrostatic terms:

$$U_{AMBER} = U_{bond} + U_{angle} + U_{torsions} + U_{vdW} + U_{el} \tag{2.14}$$

The bonded and angular terms are harmonically centred around their equilibrium values:

$$\begin{aligned} U_{bond}(\vec{x}) &= \sum_{bonds} k_r(r - r_{eq})^2 \\ U_{angle}(\vec{x}) &= \sum_{angles} k_\theta(\theta - \theta_{eq})^2 \end{aligned} \tag{2.15}$$

where r is the bond distance, θ is the bond angle, x_{eq} represents the corresponding equilibrium value and k_x denotes the respective force constant. The torsional terms consist of proper dihedrals and improper dihedrals, which describe out-of-plane motions. They have a periodic form:

$$U_{torsions} = \sum_{torsions} k_n(1 + \cos(n\phi - \gamma)) \tag{2.16}$$

Each torsion has a multiplicity n, an offset γ and an amplitude k_n corresponding to the dihedral angle ϕ. The short-range nonbonded interactions are described by the Lennard-Jones (LJ) potential:

$$U_{vdW} = \sum_{i<j} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{2.17}$$

where σ is related to the equilibrium inter-atomic distance and ϵ represents the depth of the energy well. The following combination rules (Lorentz–Berthelot) are in place in the AMBER force field for two atoms A and B with LJ parameters (σ_A, ϵ_A) and (σ_B, ϵ_B), respectively:

$$\begin{aligned} \epsilon_{AB} &= \sqrt{\epsilon_A \epsilon_B} \\ \sigma_{AB} &= \frac{1}{2}(\sigma_A + \sigma_B) \end{aligned} \tag{2.18}$$

Finally, the AMBER force field approximates long-range electrostatic interactions using fixed partial charges q centred on each atom:

$$U_{el} = \sum_{i<j} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{2.19}$$

Here ϵ_0 denotes the permittivity of free space. All nonbonded interactions are ignored for pairs of atoms with bonded and angular interactions, since these are described by the harmonic terms outlined above. In the case of 1,4 interactions, the following "fudge" factors are used to scale the nonbonded terms in order to partially introduce steric repulsion and classical electrostatic interactions to the periodic torsional potential:

$$\begin{aligned} f_{vdW} &= 0.5 \\ f_{el} &\approx 0.833 \end{aligned} \tag{2.20}$$

2.2.2 Force Field Types

2.2.2.1 Proteins

One of the most widely-used standard amino acid models is ff99SB,[24] which has been superseded by its more recent version, ff14SB[25] and most recently, ff19SB.[26] Here and henceforth we will concentrate on ff14SB, which has been shown to achieve a marked improvement over ff99SB by employing a large set (∼15,000) of training data in vacuum with geometries optimised at low-level Hartree-Fock theory (HF) and single-point energies calculated with second-order Møller-Plesset perturbation theory (MP2). Since the functional form of the force field is inherently approximate, there is an accuracy limit on the model and it has been argued[26] that diverse training data is more important in developing a practically useful force field than more accurate training energies, making ff14SB a suitable protein force field for the systems considered in the following chapters.

2.2.2.2 Small Molecules

Similarly to ff14SB, a general AMBER force field (GAFF)[27] has been developed to describe a diverse set of small molecules. The fitting procedure and the functional form are comparable to ff14SB. The main differences are the need for an extensive set of atom types, representing different chemical environments (which set has been extended in the more recent version, GAFF2), as well as the on-demand calculation of partial charges. Dihedral terms which are not present in GAFF are usually replaced by their closest matches and associated with a quality metric to assess whether additional quantum mechanical calculation is required to determine these. Similarly, partial charges are obtainable from a quantum calculation using different approaches, such as RESP.[28] Alternatively, one can use semi-empirical charge derivation methods, such as AM1-BCC.[29,30] It is important to note that the latter is conformation-dependent and the initial choice of coordinates will affect the energy model.

2.2.2.3 Water

Water is obviously a crucial component of any biomolecular simulation and much effort has been spent in developing good water models. Unfortunately, since water exhibits unusual macroscopic behaviour due to its extensive hydrogen bonding, it is

difficult to create a reliable water model. Most widely used are 3- and 4-point water models, such as TIP3P,[31] TIP4P-Ew,[32] SPC[33] and OPC.[34] 3-point water models are still widely used in the literature, mostly because of the extra computational speed and in our study we will use TIP3P, which is compatible with both ff14SB and GAFF/GAFF2. The TIP3P model only defines three constrained O–H and H–H distances alongside a single partial charge parameter (the other two are obtainable by symmetry and net neutrality) and two LJ parameters for the whole molecule, centred on the oxygen atom.

2.2.2.4 Others

There are also LJ parameters for simple cations and anions, associated with TIP3P. This is unfortunately not true for more complicated ions, for which complete reparametrisation is needed. This is particularly difficult for transition metal centres, since the concept of bonding is not as well-defined in these systems as in organic molecules, meaning that the standard functional form of the force field is often inadequate in such cases. In this thesis, systems with transition metal centres will not be considered. Finally, many proteins have a range of purely organic cofactors, for which a general force field is not adequate, partly due to the frequent presence of difficult-to-parametrise phosphate groups. Therefore, manual parametrisation is preferred and there is an online database[35] containing many such efforts which are compatible with the AMBER force field. This is the resource that will be used in the following chapters for the systems containing common cofactors.

2.2.3 Evaluation

2.2.3.1 Periodic Boundary Conditions

As mentioned above, we can only simulate a number of atoms many orders of magnitude less than Avogadro's number with current computational power—a system size which is much smaller than the full macroscopic system of interest. The boundaries of the simulated system are thus non-physical and as such should not exert any forces on it, which would be non-negligible at this lengthscale. A common practice which circumvents this problem is to create an infinite 3-dimensional "crystal" with every unit cell being the simulation of interest. This approximation allows a net particle flow over the boundaries without any changes in the total number of particles—a property which is extremely important for simulations studying ensembles with constant number of particles. In practice, this simply means that when a particle crosses a boundary, it is immediately moved to the other side of the unit cell. The most commonly used unit cell shape is the cuboid and even though

alternative unit cell shapes, such as dodecahedral, are also used in biomolecular simulations for performance purposes, cubic cells will be exclusively utilised in the following chapters.

2.2.3.2 Cut-Offs

In practice, the force field is a very complex model, and this complexity makes it prohibitively expensive to evaluate it and its derivatives. The main reason for this is the unfavourable $O(N^2)$ scaling of the multiple nonbonded interaction terms. Additionally, there are an infinite number of nonbonded terms if there are periodic boundary conditions, so the most common practice is to introduce short-range cut-offs r_c in the range between 0.8 and 1.2 nm, beyond which any interactions are ignored.

2.2.3.3 Long-Range Corrections

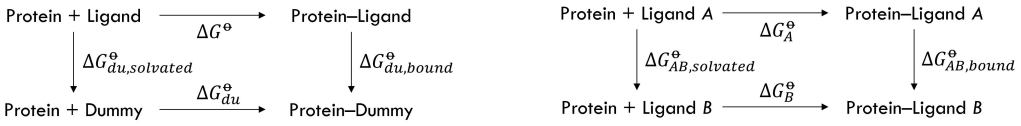
The introduced cut-off distances r_c are only useful for short-range interactions and introduce long-range errors in both LJ and Coulomb terms. This is not particularly problematic for LJ dispersion interactions, since they decay as r^{-6}, and any contributions to the energy beyond the cut-off ΔU_{disp} can be approximated in a mean-field fashion:

$$\Delta U_{disp} \propto \int_{r_c}^{\infty} 4\pi r^2 \frac{1}{r^6} dr \propto \frac{1}{r_c^3} \tag{2.21}$$

This treatment is not possible for periodic electrostatic interactions, since it is readily seen by comparison with Equation 2.21 that the mean-field integral diverges for $U_{el} \propto \frac{1}{r}$. Nevertheless, one can split the whole electrostatic sum into real-space and Fourier-space terms, each with an associated cut-off (Ewald summation[36]). However, this sum still scales unfavourably as $O(N^2)$ and its fast Fourier transform version, particle mesh Ewald (PME)[37] is frequently used in practice instead. PME scales as $O(N \log N)$ making it much faster than regular Ewald. It is still, however, the slowest force field component to evaluate.

2.3 Performing Free Energy Calculations

In Section 2.1.2 it was shown that free energies are commonly calculated by calculating thermodynamic averages of certain quantities over several simulations. However, there are some practical subtleties when performing these calculations and they will be described in this section.



(A) Absolute binding free energy cycle (B) Relative binding free energy cycle

FIGURE 2.1: Constructing a thermodynamic cycle describing protein–ligand binding. (2.1a): each atom in the ligand is annihilated to a "dummy". Since this dummy does not interact with its surroundings, $\Delta G^{\ominus}_{du} = 0$, meaning that $\Delta G^{\ominus} = \Delta G^{\ominus}_{du,bound} - \Delta G^{\ominus}_{du,solvated}$. (2.1b): calculating the relative free energy $\Delta\Delta G_{AB}$ by ligand transformation. This thermodynamic cycle shows that $\Delta\Delta G^{\ominus}_{AB} \equiv \Delta G^{\ominus}_{B} - \Delta G^{\ominus}_{A} = \Delta G^{\ominus}_{AB,bound} - \Delta G^{\ominus}_{AB,solvated}$.

2.3.1 Calculating Binding Free Energies

The main interest of the work presented in the following chapters will be the calculation of equilibrium protein-ligand binding free energies. The standard Gibbs binding free energy $\Delta G^{\ominus}_{A}$ is a quantity which is directly related to the equilibrium binding constant K_A of some ligand A to the protein:

$$\Delta G^{\ominus}_{A} = -\frac{1}{\beta} \ln K_A \tag{2.22}$$

In practice, the standard relative Gibbs binding free energy between two ligands A and B, $\Delta\Delta G^{\ominus}_{AB}$, is more desirable for direct calculation, due to its better convergence properties:

$$\Delta\Delta G^{\ominus}_{AB} = \Delta G^{\ominus}_{B} - \Delta G^{\ominus}_{A} \tag{2.23}$$

Direct calculation of binding constants is difficult, since binding events are usually extremely rare and practically intractable using current computational speed. A technique which is overwhelmingly popular is to make use of the fact that the free energy is a state function and any closed thermodynamic cycles have free energy changes summing up to zero (Figure 2.1). In this way, one can express the relative free energy $\Delta\Delta G^{\ominus}_{AB}$ as:

$$\Delta\Delta G^{\ominus}_{AB} = \Delta G^{\ominus}_{AB,bound} - \Delta G^{\ominus}_{AB,solvated} \tag{2.24}$$

Where $\Delta G^{\ominus}_{AB,bound}$ and $\Delta G^{\ominus}_{AB,solvated}$ refer to the ligand transformation from A to B when bound to the protein and in pure water, respectively. One downside of relative free energy calculations is that the binding mode used for calculating $\Delta G^{\ominus}_{AB,bound}$ should be known in advance, since binding mode changes do not typically occur

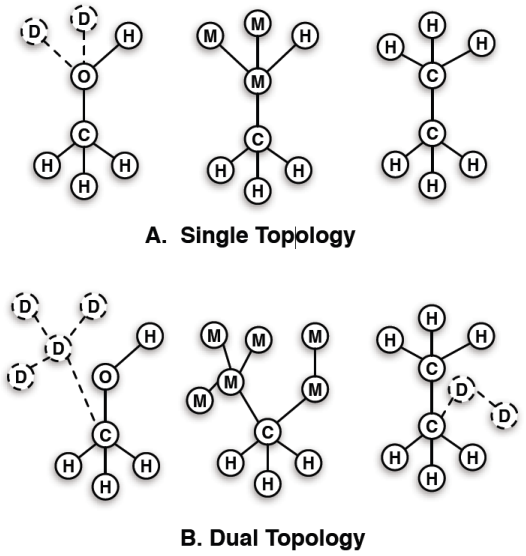


FIGURE 2.2: Illustration of the single and dual topology approaches. "D" refers to a dummy atom, while "M" indicates a partially interacting atom in its intermediate state. Image taken from [1].

under the usual timescales used for alchemical free energy (AFE) calculations. Equation 2.24 is then reduced to two free energy calculations coupling two different Hamiltonians ("bound leg" and "solvated leg"), and these free energies can be readily estimated using intermediate alchemical states and the techniques described in Section 2.1.2. Although the design of these intermediate states does not theoretically affect the asymptotic free energy value, this choice can significantly impact the efficiency of the calculation. Some common practices for defining the intermediate states will be described in the next sections.

2.3.2 Topology

When interpolating the two Hamiltonians H_A and H_B, one usually distinguishes between two types of protocols: single topology and dual topology (Figure 2.2). In the single topology protocol, all atoms are mapped onto a maximum common substructure (MCS) and the atoms that do not match are perturbed from/into noninteracting dummy atoms. This protocol is employed in GROMACS.[38] In the dual topology protocol, both molecules are simulated at the same time and they do not interact with one another, whilst being restrained to each other. This approach is used in NAMD.[39] One can also effectively combine both methodologies by simulating the two molecules separately whilst constraining a subset of their atoms to have common coordinates. This is the approach used in AMBER.[40]

In practice, single topology protocols result in lower estimator variance, since both ligands sample the same part of phase space. However, this apparent convergence

could be misleading, as will be shown in most of the following chapters. On the other hand, dual topology is most useful for perturbations involving ring breaking, where single topology approaches are either problematic[41] or require substantial modifications.[42]

When using single topology mapping, one can not only interpolate between the energy functions of both endstates, but also between the force field parameters thereof. The latter is the method used in GROMACS. Although the interpolation can in principle use any powers of the λ variable, linear coupling is usually preferred in practice, although the next section will discuss an important exception to this rule.

2.3.3 Perturbing Potentials

As discussed in the previous section, an obvious way to combine the two terminal potential energy functions U_0 and U_1 into an intermediate U_λ is:

$$U_\lambda(r) = (1-\lambda)U_0(r) + \lambda U_1(r) \tag{2.25}$$

However, this simplistic approach can lead to numerical instabilities in the force field components that admit singularities (in our case, electrostatics and van der Waals interactions). Since any intermediate interpolation is arbitrary, one can choose to use shifted distances which vary smoothly with λ and do not result in any singularities at the intermediate states. This intermediate potential is called a soft-core potential[43] and a commonly used functional form is:[44]

$$\begin{aligned} U_\lambda(r) &= (1-\lambda)^a U_0(r_0) + \lambda^a U_1(r_1) \\ r_0 &= (\alpha\sigma_A^6\lambda^b + r^c)^{\frac{1}{c}} \\ r_1 &= (\alpha\sigma_B^6(1-\lambda)^b + r^c)^{\frac{1}{c}} \end{aligned} \tag{2.26}$$

where α, a, b and c are tunable parameters. For LJ terms, it is commonly chosen that $a = 1$, $b = 1$ and $c = 6$. On the other hand, electrostatic interactions are usually scaled separately from LJ interactions and in such a setting it is more efficient to decouple them linearly. Soft-core potentials are thus mainly used for perturbing van der Waals interactions only.

2.4 Sampling

In order to obtain ensemble averages, one needs to find a way to sample from the corresponding probability distribution. In general, this is a highly disconnected

multimodal distribution, which means that global sampling is a challenge. In practice, especially for large systems, one can only achieve local sampling. The main traditionally employed methods to locally sample points in phase space are: derivative-based approaches (e.g. molecular dynamics [MD]), derivative-free approaches (e.g. Markov chain Monte Carlo [MCMC]) and hybrids (e.g. Hamiltonian Monte Carlo [HMC][45]). These will be briefly outlined in the next sections.

Even if global sampling is not possible for large systems, one can enhance the local sampling over certain degrees of freedom of interest. Enhanced sampling is a broad area of research and many algorithms have been published in the literature to overcome the kinetic barriers associated with the highly disconnected multimodal distributions. A brief summary of different approaches to enhancing the basic MCMC algorithm will be considered in Section 2.4.3, while a more detailed discussion of enhanced sampling methods, in particular tempering methods, will be presented in Chapter 7.

2.4.1 Molecular Dynamics (MD)

2.4.1.1 Equations of Motion

Central to MD is the ergodic hypothesis,[46] which postulates that for a sufficiently chaotic system, the expectation value of an observable $O(\vec{x}, \vec{p})$ over the multidimensional probability distribution $\pi(\vec{x}, \vec{p})$ can be mapped onto a one-dimensional time integral:

$$\langle O(\vec{x}, \vec{p}) \rangle \equiv \int O(\vec{x}, \vec{p}) \pi(\vec{x}, \vec{p}) d\vec{x} d\vec{p} = \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^\tau O(\vec{x}(t), \vec{p}(t)) dt \tag{2.27}$$

The time evolution is then achieved using the classical equations of motion:

$$\begin{aligned} \frac{dx_i}{dt} &= v_i \\ \frac{dv_i}{dt} &= \frac{F_i(\vec{x})}{m_i} \end{aligned} \tag{2.28}$$

where $\vec{v}$ is the velocity vector, $\vec{F}(\vec{x})$ is the force vector and m_i is the mass associated with the i-th degree of freedom. The above equation can be formalised as an operator acting on an initial state in phase space $\vec{z}(0) \equiv (\vec{x}(0), \vec{p}(0))$:

$$\vec{z}(t) = e^{i\hat{\mathcal{L}}t} \vec{z}(0) \tag{2.29}$$

with $\hat{\mathcal{L}}$ being the Liouville operator, defined as:

$$i\hat{\mathcal{L}} = \sum_{i=1}^{3N} \left[v_i \frac{\partial}{\partial x_i} + \frac{F_i(\vec{x})}{m_i} \frac{\partial}{\partial v_i} \right] \tag{2.30}$$

where we have converted all momenta to velocities in order to compare with Equation 2.28.

Since Equation 2.28 is a system of $6N$ first-order differential equations, solving it analytically is not possible, apart from the simplest cases. However, numerical methods have been quite successful at generating an approximate solution to the above equations and the Liouvillian formalism facilitates the derivation of otherwise unwieldy approximations in discrete time.

2.4.1.2 Numerical Integration

If we split the Liouville operator into a position and a velocity term:

$$i\hat{\mathcal{L}} = i\hat{\mathcal{L}}_{\vec{x}} + i\hat{\mathcal{L}}_{\vec{v}} \tag{2.31}$$

we can discretise Equation 2.29 using a symmetric Trotter expansion:[47,48]

$$e^{i\hat{\mathcal{L}}t}\vec{z}(0) = \left[e^{i\hat{\mathcal{L}}_{\vec{v}}\frac{\Delta t}{2}} e^{i\hat{\mathcal{L}}_{\vec{x}}\Delta t} e^{i\hat{\mathcal{L}}_{\vec{v}}\frac{\Delta t}{2}} + O(\Delta t^3) \right]^{\frac{t}{\Delta t}} \vec{z}(0) \tag{2.32}$$

If we ignore the $O(\Delta t^3)$ terms and approximate the action of each operator on our state using the Euler method, we obtain the time-reversible volume-preserving velocity Verlet algorithm,[49] which is commonly used in MD simulations. Most existing integrators can be derived by either keeping the higher-order terms or by using alternative Liouvillian splitting procedures. More generally, we can denote the velocity Verlet integrator as a "second-order VRV" integrator, where we use the Liouvillian splitting as a unique label. This notation is useful for providing a concise description of more sophisticated integration schemes.

The magnitude of the timestep is determined by the numerical stability of the fastest motions in the system. In chemical systems, the fastest degrees of freedom are the hydrogen atoms, which can usually be reliably simulated without any numerical instabilities with a 1 fs timestep. Another common procedure is to integrate out the hydrogen degrees of freedom by constraining their associated bonds, thereby allowing a larger timestep of typically 2 fs. In this case, one needs a constraint algorithm, such as SHAKE,[50] LINCS[51] or CCMA[52] to efficiently solve the system of nonlinear equations which arises from the constrained equations of motion. For simple cases,

such as water molecules, it is possible to solve the constraint equations analytically at a lower computational cost. This approach is known as the SETTLE[53] algorithm.

2.4.1.3 Thermostats

The above integration procedures generate approximately energy-conserving motion. In a system with heat dissipation we need to use thermostats in order to sample from a canonical distribution. One approach is velocity rescaling (e.g. Berendsen thermostat[54]), where the kinetic energy is rescaled exponentially quickly to the desired temperature. This approach is not rigorous,[55] but is nonetheless efficient for equilibration. Another approach is the use of additional degrees of freedom which represent the thermostat (e.g. Nosé-Hoover thermostat[56]). These extended Lagrangian approaches are however only ergodic in the limit of an infinite amount of additional degrees of freedom.[57] Arguably the most reliable family of thermostat methods sample velocities from their equilibrium distribution and mix them with the current velocities (e.g. Langevin thermostat[58]). This approach is non-deterministic and completely rigorous. In this case, one needs extended equations of motion with an extra stochastic step (also abbreviated as "O"), which can be derived using Equation 2.32. One commonly used form of an integrator incorporating a Langevin thermostat is the second-order VRORV integrator (also known as BAOAB), which has been shown to be the most accurate second-order splitting in practice.[59]

2.4.1.4 Barostats

Similarly, pressure control is needed if one samples from an isothermal–isobaric ensemble. This is achieved using barostats. As with thermostats, there are coordinate-rescaling barostats (e.g. the Berendsen barostat[54]) which are efficient but not rigorous, extended Lagrangian barostats (e.g. the Parrinello–Rahman barostat[60]), which are deterministic in nature, and Monte Carlo barostats, which are stochastic and rigorous.

2.4.2 Markov Chain Monte Carlo (MCMC)

An alternative method for generating samples according to an arbitrary multidimensional distribution is MCMC. A Markov chain is a stochastic model defined over a space of discrete and continuous states. Sampling over these spaces is only dependent on the most recent sample, making MCMC a memoryless method. If we denote the current configuration by $\vec{x}$ and the subsequent configuration by $\vec{x'}$, a suitable normalised transition kernel $T(\vec{x'}|\vec{x})$ preserves the target probability distribution $\pi(\vec{x})$ only if the following condition ("balance") is met:

$$\begin{aligned} \pi(\vec{x'}) &= \int \pi(\vec{x}) T(\vec{x'}|\vec{x}) d\vec{x} \\ 1 &= \int T(\vec{x'}|\vec{x}) d\vec{x'} \end{aligned} \tag{2.33}$$

The balance condition combined with the requirement that all states are connected in a finite number of transitions ("ergodicity") are the two sufficient and necessary requirements to sample from $\pi(\vec{x})$. While balance is a completely general condition, it is in most cases difficult to describe a suitable transition kernel $T(\vec{x'}|\vec{x})$ for an arbitrarily complex probability distribution in a simple way. Consequently, a more stringent but universally applicable condition, called detailed balance, is usually used instead:

$$\pi(\vec{x}) T(\vec{x'}|\vec{x}) = \pi(\vec{x'}) T(\vec{x}|\vec{x'}) \qquad \forall \vec{x}, \vec{x'} \tag{2.34}$$

It can be easily seen that detailed balance trivially satisfies Equation 2.33 for any distribution $\pi(\vec{x})$. In order to enforce detailed balance, one usually splits $T(\vec{x'}|\vec{x})$ into an arbitrary proposal probability $p_{prop}(\vec{x'}|\vec{x})$ and a residual acceptance probability $p_{acc}(\vec{x'}|\vec{x})$, such that:

$$\begin{aligned} \pi(\vec{x}) p_{prop}(\vec{x'}|\vec{x}) p_{acc}(\vec{x'}|\vec{x}) &= \pi(\vec{x'}) p_{prop}(\vec{x}|\vec{x'}) p_{acc}(\vec{x}|\vec{x'}) \\ \frac{p_{acc}(\vec{x'}|\vec{x})}{p_{acc}(\vec{x}|\vec{x'})} &= \frac{\pi(\vec{x'}) p_{prop}(\vec{x}|\vec{x'})}{\pi(\vec{x}) p_{prop}(\vec{x'}|\vec{x})} \end{aligned} \tag{2.35}$$

A common choice for $p_{acc}(\vec{x}|\vec{x'})$ which satisfies Equation 2.35 while maximising the transition probability is the Metropolis acceptance criterion:[61]

$$p_{acc}(\vec{x}|\vec{x'}) = \min \left[1, \frac{\pi(\vec{x'}) p_{prop}(\vec{x}|\vec{x'})}{\pi(\vec{x}) p_{prop}(\vec{x'}|\vec{x})} \right] \tag{2.36}$$

Most of the efforts in MCMC research are centred around obtaining an appropriate form of $p_{prop}(\vec{x}|\vec{x'})$, i.e. a way to generate new configurations which results in both high acceptance and fast sample decorrelation. Different schemes to achieve this exist, some of which will be outlined in Section 2.4.3.

A key strength of MCMC in comparison to MD is its ability to not only sample continuous state spaces, but discrete ones as well. In this case, the above formalism still holds, except that the transition kernels can be thought of as finite-dimensional matrices, rather than continuous functions. The utility of discrete state spaces will be

described in more detail in Chapter 7, where it is shown how various enhanced sampling methods benefit from this framework.

Another useful application of MCMC is its ability to sample an arbitrary subset of the underlying variables conditionally on the other variables (Gibbs sampling[62]). This can in practice be combined with other sampling methods, such as MD. Examples of this specialised application of MCMC is the implementation of a Monte Carlo barostat which only updates the box size conditionally on the system coordinates, as well as temperature-based enhanced sampling methods, which update the temperature T independently of the system coordinates (Chapter 7). Indeed, Gibbs sampling is often preferred in high-dimensional systems due to its comparatively high acceptance rates.

The greatest advantage of MCMC is that it is an extremely general method which is simple to understand and implement. However, one significant drawback is the fact that detailed balance has to be almost invariably used in practice. Since detailed balance is a sufficient but not necessary condition to preserve the equilibrium distribution, it is usually very restrictive, leading to high rejection rates and therefore wasted computational effort. On the other hand, such waste is not a feature of MD.

2.4.3 Enhanced Monte Carlo

As mentioned in Section 2.4.2, the proposal distribution $p_{prop}(\vec{x}'|\vec{x})$ can be engineered to give rise to various Monte Carlo algorithms. For example, a transition kernel consisting of untempered MD evolution over a time τ gives rise to the HMC algorithm, while a tempered kernel results in the Metropolis-adjusted Langevin algorithm (MALA).[63] These methods can be seen as the exact version of MD, since the latter has an associated discretisation error proportional to the timestep.

Similarly, one can devise transition kernels for surmounting specific kinetic barriers, including those associated with certain translational, rotational and/or torsional degrees of freedom. One such method is nonequilibrium candidate Monte Carlo (NCMC),[64] which is a generalisation of MALA to a time-dependent sequence of different Hamiltonians. The suitable design of these can then enhance the sampling of certain degrees of freedom.[65] Another transition kernel more specific to simulations in the grand canonical (μVT) ensemble is used in grand canonical Monte Carlo (GCMC).[66] This methodology can be utilised in exploring the translational and rotational degrees of freedom of kinetically trapped binding site water molecules.[67]

2.5 Markov State Models (MSMs)

Markov chains can not only be used to sample discrete and continuous distributions, but they can also be utilised as a modelling tool to create a kinetic profile of certain rare events of interest in a molecular simulation. This approach is known as Markov state modelling.

A Markov state model (MSM) is defined by a collection of macrostates of interest (e.g. a folded and an unfolded protein state) and the transition matrix describing conditional transition probabilities p_{trans} after a lag time τ, such that:

$$T_{ij}(\tau) = p_{trans}(j|i,\tau) \tag{2.37}$$

for some states i and j. In this way, an MSM assumes that any transitions between different macrostates are probabilistic and memoryless.

The assignment of a unique macrostate to each of the observed microstates depends on what types of rare events are of interest. For example, the kinetics of a *cis–trans* isomerisation of a particular double bond can be described by assigning to each sampled structure a corresponding label: either *cis* or *trans*. This assignment can either be performed by manually determining a dihedral angle boundary, or by using various clustering methods.[68] In other cases, such as protein folding, defining a degree of freedom for subsequent clustering is not as straightforward as measuring a single dihedral angle, and more general dimensionality reduction techniques need to be used instead, such as principal component analysis (PCA)[69] or time-lagged independent component analysis (TICA).[70]

Once the macrostates of interest have been defined, the next step is the determination of the transition matrix $\mathbf{T}(\tau)$ given a particular lag time τ. Although this can in principle be done by simply counting the number of transitions between states, this methodology is not reliable when the number of transitions is low, which is often the case if one is interested in the kinetics of rare events.[71] Instead, maximum likelihood approaches are more commonly used, where the likelihood of observing a particular trajectory of macrostates given a transition matrix $\mathbf{T}(\tau)$ is maximised by varying the individual elements of $\mathbf{T}(\tau)$.

However, a simple maximum likelihood approach is highly dependent on the clustering procedure and is consequently sensitive to any discretisation errors. This problem can be circumvented by employing hidden MSMs,[72] which consist of hidden states and observed states, where there is a finite probability that each hidden state will be observed as a different observed state. In this way, discretisation errors can be handled more robustly by modelling them as a noisy observation process. Another

desirable quality of an MSM is the ability to estimate a confidence interval of its observables. This can be done by generating a population of different MSMs using a Bayesian approach[73] and can also be performed in the case of hidden MSMs.[74]

As long as the estimated transition matrix is regular (i.e. there exists a positive power thereof, such that all of its elements are strictly positive), it has an associated stationary distribution $\vec{\pi}_0$,[75] such that:

$$\vec{\pi}_0 \mathbf{T}(\tau) = \vec{\pi}_0 \tag{2.38}$$

Since any initial distribution $\vec{p}_0$ can be expressed as a weighted sum of the left eigenvectors of $\mathbf{T}(\tau)$, such that $\vec{p}_0 = \sum_i c_i \vec{\pi}_i$, one can then express the evolution of $\vec{p}_0$ after $t \equiv n\tau$ timesteps:

$$\vec{p}_0 \mathbf{T}^n(\tau) = \sum_i c_i \vec{\pi}_i \mathbf{T}^n(\tau) = \sum_i c_i \vec{\pi}_i \lambda_i(\tau)^n \tag{2.39}$$

Since the magnitude of each eigenvalue apart from λ_0 is less than unity,[76] this means that the contribution of the i-th eigenvector ($i > 0$) decays exponentially over time with a decay factor of $\lambda_i(\tau)^n \equiv \lambda_i(\tau)^{t/\tau}$. One can relate this decay factor to the one corresponding to a general exponential decay process, $e^{-t/t_{decay}}$:[77]

$$t_{decay,i}(\tau) = -\frac{\tau}{\ln \lambda_i(\tau)} \tag{2.40}$$

$t_{decay,i}$ is commonly referred to as the i-th implied timescale. For $i = 1$, which we define as corresponding to the second largest eigenvalue, one can obtain the slowest implied timescale, which gives a measure of the mixing rate of the Markov chain:

$$t_{slowest}(\tau) = -\frac{\tau}{\ln \lambda_{slowest}(\tau)} \tag{2.41}$$

For a perfectly Markovian system, $\lambda(k\tau) = \lambda(\tau)^k$, meaning that the estimated implied timescales should be independent of the choice of lag time.[78] This expected property can be used to validate the MSM. In practice, this is commonly done by plotting the implied timescale of interest over a range of lag times and analysing the deviation of the resulting plot from an expected horizontal line.[79]

We now turn to Chapter 3, where we review the recent literature investigating the reproducibility of AFE calculations.

Chapter 3

Reproducibility of Alchemical Free Energy Calculations: A Review

3.1 Introduction

As described in Chapter 2, the two essential elements when performing a molecular simulation are the energy model (force field) and the sampling method used to generate structures according to the thermodynamic ensemble partially defined by the force field Hamiltonian. These two constituents are also the main focus of study when developing improved algorithms—the former directly affecting the accuracy, and the latter determining the precision of the results obtained from the simulation.

Compared to force field development and enhanced sampling research, an often-overlooked topic in the field of molecular simulation is that of reproducibility. Reproducibility is defined as the extent to which two unrelated research groups can obtain statistically equivalent results given the same research problem and methodology. We will also note that this is a more general condition than repeatability, which is related to the ability of the same researcher to obtain the same result by exactly repeating the same methodology using the same apparatus. However, since molecular dynamics (MD) simulation studies rely on many implicit choices and parameters set by the researcher (e.g. initial coordinates and force field parameters), these are only partially described in practice, meaning that the vast majority of MD calculations are inherently difficult to reproduce. These decisions are not only method-specific but are also related to how the researcher chooses to represent the real-world physical system as a computational model and what the purpose of the simulation is. Therefore, we will here and henceforth extend the concept of reproducibility to the ability to obtain statistically similar results using two different protocols.

Evaluating the reproducibility of binding free energy calculations is critical to their viability in commercial drug discovery studies. Even though guidelines and best practices for alchemical free energy (AFE) calculations are routinely published,[1,80,81] the process remains highly involved with multiple arbitrary choices made by the researcher. If the resulting free energy values are highly sensitive to these choices, this presents a significant problem for the utility of AFE methods.

Each binding free energy simulation consists of two stages: system setup and simulation. The former involves the selection of starting coordinates, the choice of force field, and the assignment of tautomeric and protonation states of the titratable amino acids and the ligand. The latter is dependent on the choice of simulation length, number of repeats, alchemical protocol and its associated parameters, the sampling method used and its associated parameters, and the final free energy analysis method. All of these choices can impact the precision and/or the accuracy of the resulting free energy values.

For instance, system setup affects both precision and accuracy, since both the initial coordinates and the Hamiltonian are defined during this stage. On the other hand, the simulation stage mainly affects the simulation precision, although some of the arbitrary parameters that can be chosen during this stage affect the accuracy of the energy function evaluation. In recent years there has been a heightened interest in measuring the extent to which various choices during either of these stages affect the resulting binding free energy values. These will be reviewed in this chapter.

3.2 Simulation Setup

3.2.1 Initial Coordinates

3.2.1.1 Protein Crystal Structure

Owing to the limitations of current computing power, *ab initio* prediction of the three-dimensional structure of a protein–ligand complex remains a computationally challenging problem and is currently difficult to perform in the context of drug discovery. As such, the choice of initial coordinates is crucial for ensuring the physical validity of the MD simulation and needs to be obtained in a reliable and reproducible way.

Although it is common practice to use a structure experimentally resolved by X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy to provide the initial coordinates, problems arise when no such structure is available. In this case one must choose a somewhat unrelated complex to derive the initial coordinates. This can

for example be achieved by homology modelling.[82] However, despite the introduction of automated homology modelling workflows,[83] these methods remain sensitive to choices in the computational protocol, making them difficult to reproduce.[84]

Another difficulty specific to AFE calculations is the choice of the initial structure of the intermediate λ windows. For an alchemical perturbation of two ligands $A \rightarrow B$, it is not clear which set of initial coordinates corresponding to these two complexes should be used to model the intermediate states, making the choice thereof largely arbitrary and not reproducible. In addition, even though two free energy calculations with different initial coordinates and identical topology should eventually converge to the same value with infinite sampling, the initial protein coordinates can also influence most of the setup process, including initial protonation states and water placement (both discussed in later sections).

The above considerations show that obtaining initial protein structures is a complex process which not only directly impacts the sampling quality, but can also affect the asymptotic free energy value. Since different researchers will generally have different approaches to choosing an initial set of coordinates, it is important to be aware of the extent to which this lack of reproducibility can impact the resulting free energies in a practical setting. However, the magnitude of this effect has not been explored in the literature.

Even more challenging are proteins with slow conformational modes. In these cases, the initially resolved crystal structure is no longer sufficient for obtaining representative sampling, and long-timescale or enhanced sampling becomes essential for these systems. This was shown in the case of T4-lysozyme by Lim *et al.*[85], who reported a root-mean-square deviation (RMSD) of 4 kcal/mol between the two sets of binding free energy values obtained from the open and closed conformations. In these cases temperature-based enhanced sampling is a useful tool in overcoming relevant high kinetic barriers and the authors reported a marked increase in consistency when protein replica exchange with solute tempering (pREST)[86] was used alongside the calculation, reducing the RMSD to 0.57 kcal/mol.

3.2.1.2 Ligand Binding Mode

Similarly to protein folding, ligand binding/unbinding is a slow event which requires timescales on the order of milliseconds to seconds. Although the binding kinetics of small fragments can be studied in real time with long MD simulations[87] ($\sim$10 μs), this approach is in general not feasible for high-throughput AFE calculations of larger and more flexible ligands. In these cases, the initial binding mode needs to be known in advance, preferably through experimental determination.

In some cases, the binding modes of both ligands A and B in the alchemical transformation $A \rightarrow B$ are known. In such a scenario, it is not immediately clear which complex should be used to provide the initial coordinates for the intermediate λ windows. This choice has been partially studied by Pérez-Benito *et al.*[88], who demonstrated that the free energy discrepancies which arise from using the initial coordinates of the protein bound either to ligand A or ligand B can often surpass 1 kcal/mol. Using one of the structures in favour of the other can even change the sign of the free energy value in cases where its magnitude is large.

In other cases, such as large-scale drug discovery studies, the binding mode is not always known in advance. In this scenario, the starting ligand coordinates are often obtained by docking.[89] Because of the large number of docking methods, this type of workflow is poorly reproducible. Indeed, a study by Cappel *et al.*[90] on five different protein–ligand systems demonstrated that different methods of obtaining initial binding modes can in some cases result in strikingly different correlations against experimentally obtained ΔG° values, and they found that a docking procedure based on maximum common substructure (MCS) constraints exhibits most consistent performance on average. Furthermore, it has been shown by Granadino-Roldán *et al.*[91] that manual interventions after docking can often help define a more physically relevant binding mode and thereby improve the accuracy of the resulting binding free energies. Such an approach is even more difficult to reproduce reliably and directly hinders the automatability of binding free energy calculations.

If several plausible binding modes exist, or if it is known that the ligand binds in several different conformations, it is possible to combine the results of alchemical free energy calculations from several binding modes.[92,93] However, this methodology requires prior knowledge of the binding modes, as well as a separate simulation for each one of them, making this method computationally expensive.

An alternative approach to evaluating the contributions of several binding modes is through enhanced sampling simulations. Examples of methods facilitating binding mode conversion are replica exchange with solute scaling (REST2),[94,95] nonequilibrium candidate Monte Carlo (NCMC)[64,65,96] and metadynamics.[97,98] Of these, one of the most popular methods for combining ligand sampling and relative binding free energy calculations is the FEP/REST[95] method, where sampling of internal ligand degrees of freedom is concurrently performed with alchemical perturbation between the two ligands. In this way, enhanced sampling provides a much more automatable and reproducible workflow and this is one of the reasons FEP/REST has been widely used in drug discovery studies.[99]

3.2.1.3 Binding Site Hydration

Another kinetically hindered event is water molecule diffusion in a closed binding site. Owing to the importance of interfacial water in mediating protein–ligand interactions,[100,101] the initial water molecule coordinates can have significant effects on the resulting binding free energy. For example, it was shown by Wahl and Smieško[102] that the choice of initial coordinates belonging to the protein complexed with either ligand *A* or *B* can have significant effects (sometimes more than 3 kcal/mol) on the calculated relative binding free energy values and this effect was largely explained by hydration differences between the two ligands. This has also been observed by Ross *et al.*[103], where removal of a single water molecule had an average RMSD of 2.32 kcal/mol between the two sets of relative free energy values.

One way to assign binding site water molecules is to use experimental crystallographic data.[104] However, such data might not be reliable due to low resolution[105,106] or insufficient experimental reproducibility of the crystallographic data.[107] One way to augment the initial crystallographic structure in an automated way is through the just add water molecules (JAWS) method.[108] This can then significantly improve calculated binding free energy values.[109] Similarly to the previous section, the initial water placement can also be improved by manual intervention, which can result in better comparison to experiment.[91] As already discussed, however, such approaches are not amenable to automation and alternative enhanced sampling methods are more desirable in a high-throughput practical setting.

One of the most widely used methods to enhance binding site water sampling in the context of MD simulations is grand canonical Monte Carlo (GCMC).[67] It has been employed in many studies to reduce biases posed by the choice of initial water molecule coordinates.[103,104,110–112] For example, it reduced many of the discrepancies arising in the studies by Wahl and Smieško[102] and Ross *et al.*[103] described above. Another approach involves nonequilibrium switching using NCMC.[113–115] However, it has been suggested that GCMC is more efficient at water sampling that NCMC.[115]

3.2.1.4 Protonation States

Since hydrogen atoms are not detected in an X-ray crystallography experiment, they have to be modelled by the researcher. While for most neutral amino acid residues this is a relatively straightforward task, the acidic and basic ones can potentially exist in multiple tautomeric states, which makes this assignment more challenging. Arguably the most difficult amino acid residue to model is histidine, which can exist in two neutral states (δ and ϵ), as well as a third protonated state. Moreover, it has a side chain pK_a of 6.5, which is very close to physiological pH conditions,[116] meaning that in many cases all of the tautomeric forms are physically relevant and modelling only

one of them is an inherent approximation. While it is known that the choice of an initial histidine protonation state can significantly affect docking results,[117] no such study has been performed in the context of binding free energy calculations. Nevertheless, it is common knowledge among practitioners that such choices can significantly affect the obtained binding free energy values and these need to be carefully assigned, especially if these amino acids are in the binding site.

A popular approach to ameliorate the amino acid protonation state problem is to run constant-pH simulations, where multiple protonation and/or tautomeric states are simulated at the same time. Various methodologies to achieve this have been explored, including propagation of a fictional protonation state coordinate,[118–120] nonequilibrium switching moves between different protonation states,[121–124] Monte Carlo moves between different protonation states[125] and concurrent simulation of several protonation states.[126] These have also been used in the context of host–guest[127,128] and protein–ligand[129,130] binding free energy calculations.

Ligand tautomers can also affect the obtained binding free energy value. For example, Hu *et al.*[131] showed that the protonation state of the ligand can significantly affect correlation against experiment. Similar observations were later made by de Oliveira *et al.*[132], who found binding free energy discrepancies of ∼0.5–1.0 kcal/mol between neutral and charged states of ligands bound to kinesin spindle protein and coagulation factor Xa (FXa). In the latter publication, the authors showed that it is possible to combine the relative binding free energies of different protonation states and/or tautomers by calculating the expected pK_a values of the protonation site of interest using an *ab initio* approach. Even though their results showed consistent improvement against experiment, as well as increased reproducibility with respect to the initial tautomer choice, the computational cost of this approach increases exponentially with the number of titratable sites and is therefore only reserved for critical protonation sites in practice.

3.2.2 Force Fields

3.2.2.1 Protein/Ligand Force Field

Although the choice of protein force field can in principle impact the calculated binding free energy, the ligand force field is commonly considered to be the more important factor, owing to the larger number of atom types that need to be parametrised compared to the twenty standard amino acids. In addition, since the ligand is the part of the system which is perturbed in an AFE calculation, the resulting free energies are likely to be most sensitive to its parameters. Combined with the historical practice of using compatible protein/ligand force fields (e.g. ff14SB[25]/GAFF[27] and CHARMM36[133]/CGenFF[134,135]), this means that the choice of

protein force field is usually determined by the choice of ligand force field in practice. Moreover, while the effect of the ligand force field on free energy calculations can be separated from the protein force field in e.g. a solvation free energy calculation, this cannot be done in a binding free energy calculation, where it will be nonetheless assumed that the ligand force field is the main source of variability.

When comparing different force fields, it is rarely the case that all of them are implemented in a single MD engine. This inevitably introduces another layer of difficulty, as multiple MD engines need to be used to perform the study. In these cases it is also difficult to decouple the effect of the force field parameters from other implementation differences between different MD engines. Despite this difficulty, studies comparing different ligand force fields have been performed. For example, Vassetti *et al.* [136] showed that solvation free energies calculated using either OPLS-AA [137] or GAFF2 performed comparably against experiment on average but there still were significant differences between each of the calculated values, with a mean absolute difference of ∼3.5 kcal/mol. Similarly, GAFF was also compared against OPLS3e, [138] where the latter displayed better agreement with experiment and the correlation between results from the two force fields was system-dependent, ranging from ∼0.6 to ∼0.9. [88]

It transpires that sensitivity towards force field parameters also extends to minute differences in the parameters of the same force field. For example, Rocklin *et al.* [139] demonstrated that small variations in the nonbonded parameters, such as charge differences of more than 0.02 e, can result in significant free energy changes of more than 1 kcal/mol. These results are especially relevant to charge derivation methods, which are notorious for their dependence on the ligand conformation. In view of this, Manzoni and Ryde [140] compared different charge derivation methods for ligands bound to galectin-3C using different starting ligand geometries and found that these can result in binding free energy discrepancies of more than 1 kcal/mol. Comparison to experiment was also inconsistent, with the RESP method [28] generating the datasets with both highest and lowest correlation against experimental values.

Addressing the significant sensitivity of binding and solvation free energies to the force field functional form and its parameters is not trivial, owing to the inherent limitations of choosing a particular approximate functional form over another. Nevertheless, reproducibility between different force fields is expected to be higher if they are constantly updated to improve their performance against experiment in edge cases. This can be achieved by using bespoke force fields with a suitable level of quantum theory, such as QUBE, [141] or force fields which are constantly updated with extra parameters to handle edge cases, such as OPLS3 [138,142,143] and OpenFF. [144,145] It is even more challenging to increase the reproducibility of the charge derivation methods, given the inherent limitations of atom-centred point charges. These can be partially circumvented by using force fields with a more sophisticated treatment of the

electrostatic interactions, such as AMOEBA,[146] but these currently remain of limited utility to drug discovery due to their high computational cost.

3.2.2.2 Water Model

As already discussed, the description of the binding site water molecules can have a significant impact on the obtained free energy values. Therefore, it is expected that the choice of a water model can also affect AFE calculations. There are two major categories of water models: implicit[147] (e.g. MM-GBSA, MM-PBSA[148]) and explicit (e.g. SPC,[33] TIP3P,[31] TIP4P-Ew[32]). While implicit water models result in significantly faster calculations, they are not straightforward to use in a binding site setting, where there is a small number of solvent molecules, rather than a continuum. In such cases, explicit solvent models are more appropriate.[101]

It has been shown by Michel *et al.*[149] that binding free energy calculations can result in significantly different values depending on the use of an implicit or an explicit solvent model. Surprisingly, the implicit water model is shown to perform at least as well as the explicit one for cyclin-dependent kinase 2 (CDK2) and neuraminidase. Similarly, Aldeghi *et al.*[150] demonstrated that different explicit solvent models can result in significantly different binding free energy values. These findings are in agreement with a previous study by Izadi *et al.*[151], where the authors reported significant differences in the predicted electrostatic free energies across different implicit and explicit solvent models. These differences can be extremely high, in some cases reaching $\sim$9 kcal/mol.

These findings suggest that there is no "gold standard" for a water model and more sophisticated models need to be developed. There has been some effort in this direction,[152,153] such as the OPC[34] and Bind3P[154] water models and the performance of these models among traditionally used water models has been recently investigated by Çınaroğlu and Biggin[155], who found that Bind3P in conjunction with the Parsley[145] force field produce the most accurate binding enthalpy values for a model host–guest system. Despite these encouraging preliminary results, until one of the proposed models proves consistently more accurate than the other for a variety of systems, water models will remain one of the main weaknesses of MD-based free energy calculations in terms of reproducibility.

3.3 Simulation Details

3.3.1 Sampling Time

As mentioned above, many of the initial choices during system preparation (particularly those of initial coordinates), should not, in principle, affect the true ensemble average. However, biologically relevant timescales (milliseconds to seconds) are beyond the reach of most modern computing capabilities. Moreover, alchemical free energy calculations often need multiple simulations to obtain a converged free energy estimate, meaning that one can only dedicate a fraction of the allocated computer time to a single λ window. Consequently, the length of each λ window is typically chosen in practice to be in the range of 1–5 ns, especially in commercial applications.[156,157]

While it is obvious that longer simulation times provide the researcher with more highly decorrelated structures and access to molecular motions that are inaccessible at shorter timescales, drug discovery applications benefit most from high throughput, since even short alchemical calculations are expensive. Therefore, computational free energy studies have historically focused on direct comparison to experiment rather than measuring the short-timescale bias with respect to the true ensemble average predicted by the force field.

A large-scale study by Fratev and Sirimulla[158] investigated the quality of the free energy values with respect to equilibration time and the simulation time. A key point in their paper is that there is a practical trade-off in terms of both equilibration and simulation time, with short simulations comparing unfavourably to experiment and long simulations resulting in low throughput. They found that an optimal procedure involves a pre-REST equilibration protocol of two independent 10 ns runs followed by 8 ns sampling time per λ window. This protocol results in an approximately two-fold decrease in mean absolute deviation (MAD) with respect to experiment across all five protein systems studied. Their results suggest that while apparent convergence in the sampling stage is not difficult to obtain, prolonged equilibration is crucial for exploring crucial slow modes of motion. Therefore, proteins with higher levels of structural flexibility benefit more from these extended protocols.

Nevertheless, other studies have found that extended protocols do not necessarily result in better agreement with experiment. For instance, Wan *et al.*[159] showed that a tenfold increase in sampling from 4 to 40 ns can significantly reduce correlation with experiment, while any improvement is on average negligible regardless of the sampling method used. In many cases the authors observed a significant shift in predicted free energies after extending the length of the simulation, often reaching 1 kcal/mol, suggesting improved sampling offset by an insufficiently accurate force field model.

3.3.2 Free Energy Estimator

The choice of the free energy estimator can also have an impact on the obtained free energy value. The most widely used equilibrium free energy estimators are the Zwanzig equation (FEP),[12] the Bennett acceptance ratio (BAR)[13] and its multistate generalisation (MBAR),[18] and thermodynamic integration (TI). For a perturbation involving N total λ windows, FEP requires a minimum of $N-1$ simulations, unlike all other methods, which require the full set of N simulations. TI, on the other hand, needs only a minimum of N energy evaluations per unit time across all λ windows, compared to $2N-2$ for FEP, $3N-2$ for BAR and N^2 for MBAR. Consequently, even though MBAR has been proven to be the asymptotically statistically optimal estimator for N λ windows (reducing to BAR at $N=2$),[18] BAR and TI are still commonly used in the literature because of their lower computational requirements, while simultaneously providing sufficiently good accuracy in many cases.

Although it is usually assumed that the above estimators will eventually converge to the same value with infinite sampling, this is not the case for TI, which also requires infinitely many intermediate λ windows for asymptotic convergence. While these conceptual differences are not necessarily practically significant, discrepancies between the estimated free energy values do arise in some cases. For example, a study by de Ruiter *et al.*[160] has demonstrated that there can be strikingly large differences between BAR and TI estimates in protein-ligand binding free energy calculations, in some cases reaching 3 kcal/mol. Moreover, TI is also dependent on the integration procedure used, which can result in differences of 1 kcal/mol.[160] Nevertheless, the authors observed that increasing the number of λ windows to 21 naturally makes most of these discrepancies negligible, showing that while TI is more sensitive to the shape of the free energy profile, it is still systematically improvable in practice.

These observations are in accordance with an earlier publication by Shirts and Pande[161], which also showed that BAR is expected to significantly outperform TI and FEP in most practical use cases. However, more recent developments have shown that TI can perform sufficiently well in practice with a carefully designed protocol.[162] Nevertheless, BAR and MBAR remain the free energy estimators that do not require any additional user input (c.f. choosing FEP direction or TI integration method), making them the most reproducible free energy methods.

3.3.3 Independent Repeats

Obtaining binding free energy estimates is an inherently stochastic process, meaning that any estimated values have an associated variance which is a measure of repeatability. Even though it is conceptually possible to estimate this variance from a single simulation using effectively decorrelated samples,[18,163] a more reliable, albeit

more computationally expensive, approach is to run several repeats where the only difference is the initial seed for the pseudo-random number generator.

It has been suggested by Knapp *et al.* [164] that running multiple repeats significantly improves the certainty and hence the repeatability of the free energy estimation procedure. In this way, it has been argued that multiple short simulations can be more preferable than one long simulation. This view has recently started gaining support from other authors. [165,166]

However, there is always a practical trade-off between the number of repeats and the resources allotted to a single binding free energy calculation. While it is important to obtain an estimate of the free energy variance, it is also highly desirable to obtain effective decorrelation from the initial coordinates and sample binding rare events, which can only be achieved with enhanced sampling and/or longer timescales. It is therefore not obvious where the optimal balance lies between longer simulations and more repeats—the definition of "optimal" is also somewhat ambiguous in this scenario. In any case, several replicate simulations should be performed in practice to measure the repeatability of the results.

3.3.4 Soft-Core Potential

When performing alchemical free energy calculations, the choice of the functional form of the energy coupling between the endstates is arbitrary. However, it is desirable to choose a functional form which minimises the free energy variance over λ space. In practice, this is commonly achieved using soft-core potentials, which soften the potential energy singularities of the alchemical atoms. Soft-core potentials are most commonly used with van der Waals interactions, but can also be used with electrostatic interactions. The choice of the soft-core potential, its parameters, and the protocol of perturbing the bonded, van der Waals and electrostatic interactions can all affect the free energy estimate in non-obvious ways. Arguably the most widely used soft-core potential uses an effective radius $r_{ij,eff}$ between two atoms i and j, which is related to the real radius r_{ij} in the following way:

$$r_{ij,eff} = (\sigma_{ij}^c \alpha \lambda^b + r_{ij}^c)^{\frac{1}{c}} \tag{3.1}$$

Here α is a continuous parameter, b and c are discrete parameters, and σ_{ij} is a force field parameter (the average particle "size"). [44] The nature of the optimal parameters has also been investigated. For example, Steinbrecher *et al.* [167] identified an acceptable range of soft-core potential values, while demonstrating that the exact value does not significantly affect the free energy estimate itself, only its variance. However, de Ruiter *et al.* [168] observed significant differences between some soft-core parameter

combinations, in some cases reaching 1 kcal/mol discrepancies. Nevertheless, these are also dependent on the number of intermediate λ windows, meaning that with an insufficient amount of intermediate states, discrepancies of 2 kcal/mol can be observed using BAR and more than 10 kcal/mol using TI.

The viability of simultaneously performing the van der Waals and electrostatic perturbations has also been investigated in several studies. For instance, Steinbrecher *et al.* [167] showed that both the sequentially perturbed (split) and the concurrently decoupled (unified) protocols result in consistent free energy estimates. Although Loeffler *et al.* [169] obtained results which were largely consistent with this notion, they observed in some cases free energy discrepancies of up to 1.5 kcal/mol. Even though these differences appeared well-converged, the authors noted that the split protocol is often more desirable than the unified, since electrostatic soft-core potentials often introduce irregularities in the free energy profile, meaning that the free energies are more easily reproducible.

3.3.5 Other

Alchemical free energy calculations are often modelled on the basis of either a canonical (NVT), or an isothermal-isobaric (NPT) ensemble. As such, the choice and the parameters of the corresponding integrators, thermostats and barostats can also potentially affect the resulting free energy estimates. For example, a deficiency of the Berendsen barostat was found during the course of the SAMPL6 challenge, resulting in non-negligible sampling artifacts.[170] More generally, it is well-known that velocity- and pressure-rescaling algorithms do not correctly sample from the corresponding thermodynamic distributions[55] and they should be avoided in MD simulations. In addition, different Liouville splittings of the integrator can give rise to significantly different sampling distributions.[59] Although these can be corrected by adding a Metropolisation step[171] which ensures stationarity of the Boltzmann distribution, many practitioners still use rescaling algorithms or integrators without Metropolisation, thereby hindering reproducibility.

Finally, the MD engine of choice can also have a non-negligible impact on the calculated free energies. Quantifying the extent of this impact is nontrivial, however, not least because different MD engines implement different thermostats, barostats, integrators, soft-core potentials, etc. For instance, a study by Loeffler *et al.* [169] found a reproducibility limit of 0.2 kcal/mol between relative free energies of solvation and similar discrepancies of up to 1 kcal/mol were further observed in a host-guest system during the SAMPL6 challenge.[170] Although narrowing down the reason for these inconsistencies remains largely speculative, they are at least partially explained by code issues, some of which have notably been fixed since.[172] It is therefore important to keep oneself up to date with major bug fixes, as well as test any freshly installed

code on basic benchmarks. On the developer side, this also means that any major bug fixes should come with additional unit and integration tests to prevent the accidental reintroduction of the bug.

3.3.6 Summary

The majority of the recent studies investigating the impact of various choices on the obtained free energy value from an AFE calculation show that in many cases reproducibility can be significantly hindered depending on these choices. Although these discrepancies can be, in many cases, sufficiently improved by using enhanced sampling methods, approximations introduced by other elements of the simulation, such as force field parameters, energy evaluation and integration, method-specific parameters and implementation details, remain difficult to address definitively.

There are at least two important areas which have not been sufficiently explored by the above studies: the effect of the initial PDB structure and the amino acid protonation states on AFE calculations. The impact of both of these decisions is significant and warrants a large-scale study. Insights from these studies will also suggest ways to address any reproducibility issues.

Before conducting these studies, a Python library, ProtoCaller, was created to facilitate the preparation and running of multiple simulations in a semi-automated manner, where all parts of the workflow can be automated to an arbitrary degree. This allowed maximum control over the whole process, while minimising the chance of random human error. ProtoCaller will be described in Chapter 4, while the crystal structure and protonation state studies will be discussed in detail in Chapters 5 and 6.

Chapter 4

ProtoCaller: Robust Automation of Binding Free Energy Calculations

4.1 Introduction

A prerequisite for performing reliable and robust free energy calculations is their automatability. This is particularly true for large-scale studies which investigate the influence of a small subset of factors on the overall behaviour of the system. In such cases, it is even more important that random human errors are not introduced into the setup process.

Unfortunately, it is not always feasible to completely automate the setup of every protein–ligand system in an unbiased way. Each system setup requires multiple steps with varying degrees of user intervention. This means that system preparation is arguably more time-consuming than data generation and has been suggested to be a crucial step in determining the resulting free energy.[88] Therefore, one needs a "grey-box" approach, where it is possible to achieve full automation of the setup process, while controlling an arbitrary number of intermediate steps and parameters depending on the user's needs.

Another issue is software interoperability. Linking together different specialised pieces of software is an undesirable but necessary task which is usually solved by using in-house scripts or commercial software. This is also prone to human errors and can quickly become unmanageable when one starts interfacing with different pieces of software.

Several tools exist which tackle these issues. Notably, YANK[173] provides a fully-automated workflow from system setup to computation of absolute free energies. Protein Preparation Wizard[174] and HTMD[175] also handle system preparation for relative free energies in a robust way, providing seamless links to

commercial molecular dynamics (MD) engines. Finally, alchemical setup,[176] FESetup[177] and pmx[178] are open-source tools which automate system setup for relative free energy calculations on multiple molecular dynamics (MD) engines.

This chapter describes ProtoCaller, an open-source conda-installable Python library which attempts to solve the above challenges by providing a customisable unified interface to all of the steps of the free energy workflow. It utilises freely available specialised libraries to set up and perform relative free energy calculations in an open-source MD engine, GROMACS.[38] Moreover, its modular nature means that the user could feasibly tailor it to their needs, even if it is not currently directly supported by the software. For example, ProtoCaller provides sufficient flexibility for performing calculations outside its originally intended scope, such as absolute and relative solvation free energies and simulations in MD engines supported by BioSimSpace,[179] such as Sire[180] (explained in more detail later in the text).

The next section will describe the workflow in more detail, whilst giving an overview of the main algorithms and procedures used in the library.

4.2 ProtoCaller

4.2.1 Protein Preparation

The first step of the workflow (Figure 4.1) is choosing the protein crystal structure from the Protein Data Bank (PDB).[181] Since there are often a large number of relevant crystal structures, the typical guideline is to choose the protein crystal structure which has a bound ligand with a structure closest to the ligand of interest. This part of the workflow determines the initial ligand binding orientation, making it the most crucial step. In ProtoCaller, one can either use a plain PDB code or provide a user-specified PDB file.

Experimental crystal structures are obtained using sophisticated models bridging theory and experiment. Because of this, it is very likely that some parts of the structure will be less reliable than others. For example, it is common that there will be missing atoms and residues which require modelling. In ProtoCaller, several tools which add missing atoms and/or residues have been linked: Modeller,[182] pdbfixer,[183] CHARMM-GUI[184] and PDB2PQR,[185] providing the user with the ability to choose the most appropriate tool for their system. PDB2PQR also removes steric clashes in the crystal structure and optimises hydrogen bonding by rotating His, Asn and Gln residues.

To build an interface to all of the above pieces of software, an extensive PDB parser was developed, which allows the user to manipulate PDB files in an object-oriented

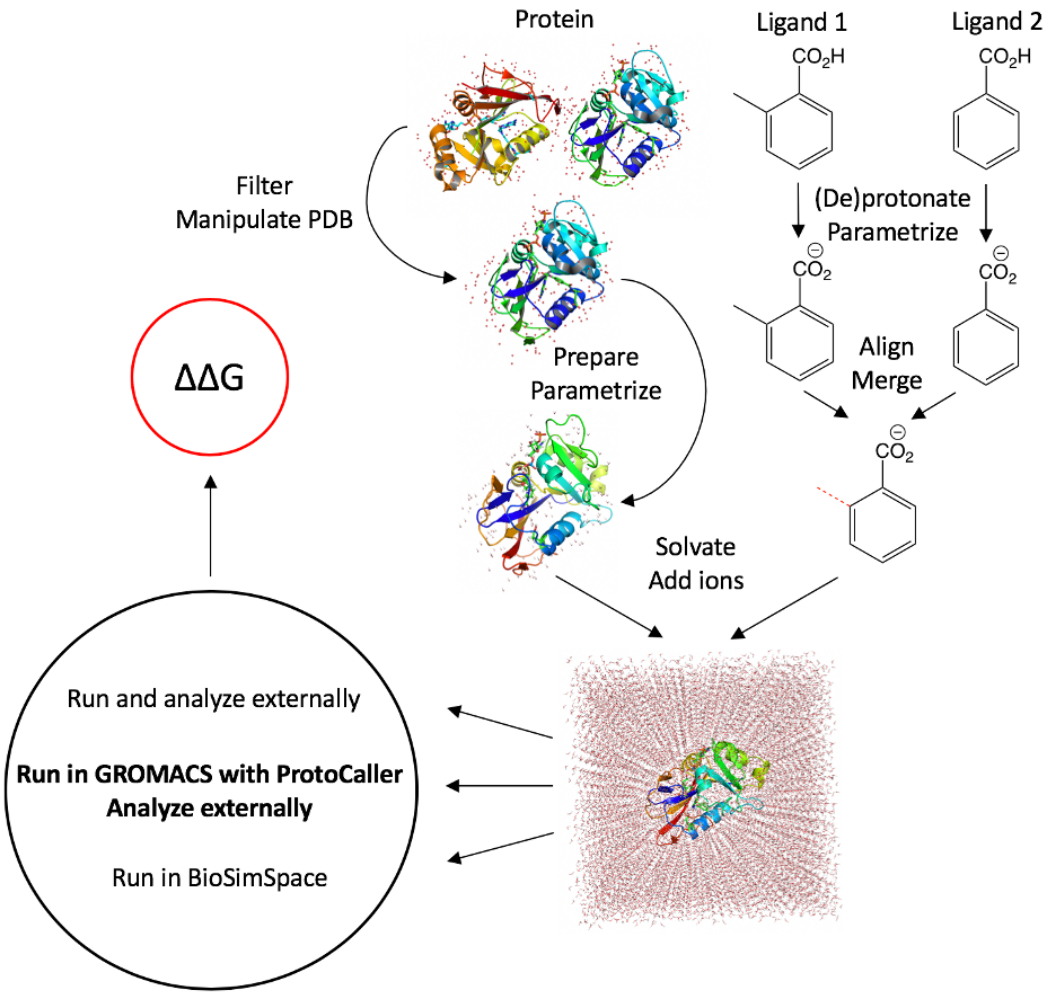


FIGURE 4.1: The full workflow for obtaining a relative binding free energy. Each of these steps utilises one or more specialised tools. In this scheme "prepare" stands for the addition of missing residues and atoms, protonation and removing steric clashes.

fashion inside Python. This parser allows seamless communication between different modules inside the software without any loss of relevant header data. This can also be extremely useful when one deals with incomplete PDB files or simply wants to introduce changes in the file in a way which is consistent with ProtoCaller.

The next step is the protonation of the protein. This is particularly difficult, since X-ray crystallography does not usually locate hydrogen nuclei and they must be modelled. Moreover, while in most cases the amino acid protonation state is straightforward to determine, the acidic and basic amino acids can exist in several protonation states. This is especially problematic for histidine, which has three different protonation states, all of which may be relevant at equilibrium. Finally, a good algorithm should be able to distinguish scenarios where there are exceptions, such as a protonated aspartic acid in an Asp–Asp dyad.[186] There are several approaches which attempt to deal with protein protonation. The program used to protonate the protein in

ProtoCaller is PROPKA3.1,[187] as utilised by PDB2PQR because of its extensive functionality in protein preparation and ease of incorporation into the workflow.

4.2.2 Ligand Protonation

Similarly, the protonation state of the ligands in the system must be determined as well. Ideally, one should consider the protein–ligand complex as one entity while performing protonation. However, in practice, the assignment of protonation states can be sensitive to the ligand binding mode which is often unknown *a priori*. In ProtoCaller, ligand protonation is thus performed separately from the protein using Open Babel.[188] Alternatively, one can provide an already protonated ligand as an external file compatible with Open Babel or ParmEd.[189] If the ligands are sufficiently simple such that there are not multiple relevant conformations, one can also use simplified molecular-input line-entry system (SMILES)[190] and international chemical identifier (InChI)[191] strings to define ligands in ProtoCaller and generate starting conformations using Open Babel. These initial conformations are crucial during parametrisation, since the charge derivation method is conformation-dependent.

4.2.3 Parametrisation

The next step is force field parametrisation. In ProtoCaller there is currently only support for the AMBER force field.[23] Since parametrisation using AmberTools is straightforward to implement, protein, ligand, water and simple ion parametrisations are directly performed in a relevant wrapper using any supported force field (ff99SB,[24] ff14SB,[25] GAFF,[27] GAFF2 and TIP3P[31] have currently been tested). Atomic charges are derived using the semi-empirical AM1-BCC[29,30] method where any subsequent floating point errors in the total charges are equally distributed between the atoms. In addition, some common cofactor parameters obtained from the AMBER parameter database[35] are also available in ProtoCaller. However, there is currently no support for modified/nonstandard amino acid residues or systems containing transition metals, since these require user intervention in any case and are unfortunately not currently automatable.

4.2.4 Mapping and Alignment

Determining the ligand binding pose is a nontrivial and critical task, even when there is a native ligand with a similar structure. This part of the workflow consists of two steps: determining the maximum common substructure (MCS) and physically aligning the ligand to the reference crystallographic binding pose.

The MCS algorithm, more commonly known as the maximum common subgraph algorithm, is an NP-complete[192] problem whose exact solution requires exponential time with respect to the number of graph nodes. Therefore, approximate solutions are needed. An open-source implementation is available in RDKit[193] and is the one used in ProtoCaller. However, some modifications in addition to this code were made in ProtoCaller to obtain a physically relevant MCS. First, ProtoCaller currently prohibits mapping between two rings of different sizes due to the nature of the subsequent alignment algorithm and the difficulties in opening a ring using a single-topology protocol.[41] Second, mapping of an acyclic chain to a reference ring is allowed, but the reverse is not, due to the hard position constraints imposed by the subsequent alignment algorithm. Finally, if there are chiral centres of different chiralities, only the longest MCS segment between two such atoms is considered a valid common substructure, which may or may not be the maximum one. The reason for this is that MCS is purely a graph theoretical algorithm with no regard to stereoisomers and the results from the algorithm have to be pruned to be physically relevant.

The alignment process is based on constraining the MCS atoms to the reference positions (e.g. crystal ligand binding orientation) and performing a force field minimisation using MMFF in RDKit[194] on the rest of the atoms whilst preserving all bond angles outside of the MCS to their prior values. Afterwards, all of the external rotatable bonds of the target molecule are rotated until an optional target metric is minimised. The current heuristic metric used in ProtoCaller is a generalised squared error ϵ, computed in the following way:

$$\epsilon = \sum_{i=1}^{N_{ref}} \sum_{j=1}^{N_{mol}} |\vec{x}_{ref,i} - \vec{x}_{mol,j}|^2 - \sum_{i \neq j}^{N_{mol}} \frac{1}{|\vec{x}_{mol,i} - \vec{x}_{mol,j}|^{12}} \tag{4.1}$$

Where N_{ref} and N_{mol} refer to the number of atoms in the reference molecule and the aligned molecule respectively, and $\vec{x}_{ref}$ and $\vec{x}_{mol}$ refer to the position vectors in both molecules. Here the first term ensures a good spatial match between the two molecules, while the second is a repulsion term which penalises steric clashes arising from the clustering of target atoms. In ProtoCaller, the second term is ignored above 1 Å, due to its negligible contributions to the sum and all MCS atoms are ignored in the first term due to the already imposed hard constraints. Finally, if there are several MCSs of the same length, those that maximise matching between the same atom types are prioritised over the others.

Of these, the one with the lowest ϵ is then chosen as the optimal alignment. It has to be noted that this metric should be used with caution and is only relevant when there are small deviations from the reference crystallographic binding pose. Otherwise, a custom binding orientation set by the user is highly recommended.

4.2.5 Solvation and Simulation

Afterwards, solvation, neutralisation and NaCl addition to the system is performed using GROMACS. It has to be noted that in each case the resulting structures have to be minimised before any simulations, since there might be some structural distortions introduced by the addition of missing amino acids or the constrained ligand mapping. The resulting output provides files for the complete free energy cycle, which can be run externally using GROMACS. Alternatively, there are several routines in ProtoCaller which provide presets for some typical protocols for performing an MD simulation in GROMACS at some user-specified simulation parameter values. One can afterwards use external tools to analyse the resulting energy files and obtain a free energy, such as the native Bennett acceptance ratio (BAR)[13] implementation in GROMACS (gmx bar) or the alchemical analysis script[195] available online.[196]

4.3 Conclusion

ProtoCaller enables the controlled automation of a large number of free energy calculations in GROMACS. This makes it a central tool to performing the studies described in Chapters 5 and 6.

Chapter 5

Sensitivity of Binding Free Energy Calculations with Respect to Initial Crystal Structure

5.1 Introduction

As discussed in Chapter 3, the choice of initial protein crystal structure is arguably the most impactful decision made by the computational chemist, as it can potentially affect the whole setup process, as well as the subsequent sampling. Although the magnitude of its effect on alchemical protein–ligand binding free energy calculations has been hinted at in previous studies,[85,88,102] there has not yet been a definitive study which addresses this problem systematically on a large scale.

This chapter examines the effect of the initial protein crystal structure on a range of ligand–ligand perturbations systematically and in depth. The crystal structures have been chosen so that they vary in their resolution, year of deposition, bound ligands and research groups. The ligand perturbations have generally been kept simple in order to obtain apparent convergence at short timescales. The test systems (dihydrofolate reductase [DHFR], protein tyrosine phosphatase 1B [PTP1B] and coagulation factor Xa [FXa]) have been chosen so that the following commonly encountered features are covered at least once: cofactors, auxiliary ions, disulfide bonds, multiple protein chains, and missing residues and atoms. Furthermore, these proteins are relatively small and shown in previous computational studies[2,156,197,198] to compare favourably to experiment, meaning that major force field and efficiency issues are not expected.

In addition, all of the calculations will be performed in two ways: with and without an extra 20 ns equilibration. This will enable the verification of the increasingly common

notion that multiple short simulations are preferable to a single long one.[164–166] Analysis of cycle closure errors and comparison to experiment will also be performed whenever possible.

5.2 Methods

5.2.1 System Preparation

All system preparation in this study was performed using ProtoCaller (Chapter 4). The following X-ray crystal structures were used: 1OHJ,[199] 2W3M,[200] 3GHW,[201] 4DDR,[202] 4M6J,[203] 5HPB,[204] 6A7E[205] and 6DAV[206] for DHFR; 1BZJ,[207] 1NWE,[208] 2AZR,[209] 2F71,[210] 2H4K,[211] 2NTA,[212] 2QBP[3] and 2ZN7[213] for PTP1B; 1EZQ,[214] 1KSN,[215] 1LQD,[4] 1NFW,[216] 2CJI,[217] 2J38,[218] 2J95[219] and 4Y71[220] for FXa. All protein crystal structures were obtained from the Protein Data Bank (PDB).[181] Some relevant metrics by which these crystal structures differ are shown in Appendix A (Tables A.1 to A.3). Most importantly, all of the above structures have a root-mean-square deviation (RMSD) of less than 1 Å after alignment to a reference structure (described later) and only one of them (4M6J) has an RMSD larger than 0.5 Å, indicating that the initial structures used can all be considered very similar to one another. In addition, some of the structures (4M6J, 1NWE, 2NTA, 1EZQ, 1LQD, 1NFW, 2J38, 4Y71) exhibit slight differences in their reported protein sequence compared to the others and these differences were kept. None of these differences were near the binding site (within 1.2 nm of the ligand centre of mass).

Where applicable, non-terminal missing residues were added using Modeller[182] and missing atoms were modelled using PDB2PQR.[185] All crystal structures were protonated with PDB2PQR, where all histidine residues were arbitrarily forced to the ϵ tautomer for FXa for consistency, due to the presence of multiple histidines near the binding site and the fact that the protonation state and tautomer assignment in PDB2PQR is dependent on the initial protein coordinates. No equivalent modifications to tautomer preference were made for DHFR and PTP1B, in some cases resulting in histidine tautomer differences across structures distal from the binding site (more than 1.2 nm away from the ligand centre of mass). All amino acids were assigned their default expected protonation states at pH $= 7$. All crystal structure waters were retained. In two of the DHFR structures, there were two copies of the protein in the asymmetric unit and in these cases only the first chain of the PDB file was used in the subsequent simulations.

For each crystal structure, the following number of ligand–ligand perturbations were performed: 8 for DHFR, 9 for PTP1B and 9 for FXa. The ligand scaffolds and thermodynamic cycles can be seen in Figure 5.2. Some of the ligands have been forced

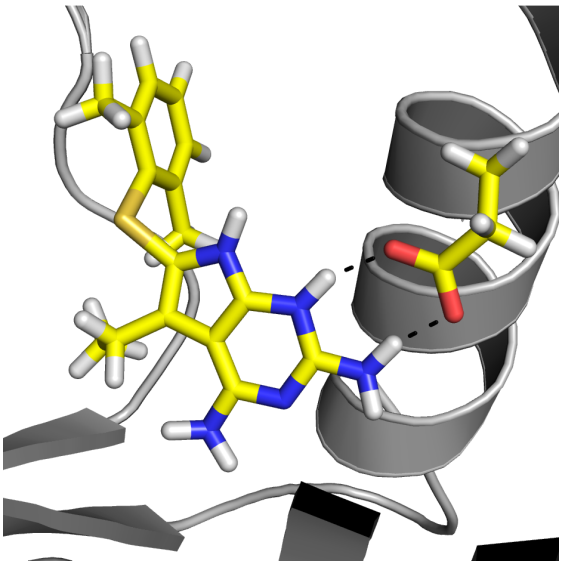


FIGURE 5.1: Hydrogen bond interactions between one of the protonated DHFR ligands and Glu30 (PDB code: 5HPB).

into an unnatural protonation state so that no charge perturbation was needed. Such ligands have only been treated as intermediates and have not been compared to experiment. The crystal structures whose native ligands most closely matched the scaffolds of the ligands of interest were used for alignment: 5HPB, 2QBP and 1LQD for DHFR, PTP1B and FXa, respectively. This initial ligand conformation was enforced across all other crystal structures based on protein–protein backbone alignment using PyMOL[8] and it was also used during the ligand parametrisation stage in order to obtain the same partial charges for the same ligand across different crystal structures. Ligand protonation states were determined manually and were in most cases in agreement with Open Babel.[188] The most notable difference is the addition of an extra proton to the DHFR ligand at one of the pyrimidine nitrogen atoms (Figure 5.1), which has been previously suggested[221] to take part in ligand-carboxylate interactions involving a closely related ligand, methotrexate.

All of the following simulations used the AMBER force field. The amino acid residues were parametrised with ff14SB,[25] and GAFF2[27] was used for the ligands. The TIP3P water model[31] and its associated calcium parameters were also used alongside the additional NADPH and NADPH parameters,[222] available online.[35] All ligand charges were parametrised using AM1-BCC.[29,30]

All ligand perturbations were performed using a single-topology mapping along an alchemical coordinate (λ) and alignment to the reference ligands was performed using ProtoCaller's default protocol (Chapter 4) by constraining the positions of the maximum common substructure (MCS) atoms to the reference ones. Whenever applicable, all perturbations were in the direction of atom annihilation. All dummy atom equilibrium distances were scaled to half of their initial values in an attempt to

achieve better phase space overlap between λ windows. All systems were solvated in cubic periodic boxes with a length of 7 nm in the bound leg and 4 nm in the solvated leg. NaCl was also added with an ionic concentration of ~0.154 M.

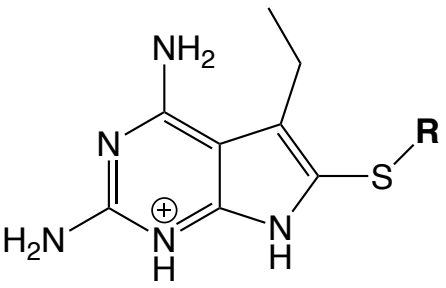
5.2.2 Simulation

All simulations were performed in GROMACS 2018.4.[38] The perturbations were carried out over 40 λ windows with 10 equally-spaced perturbations of the electrostatic terms followed by simultaneous scaling of the Lennard-Jones (LJ) and bonded terms during the other 30 windows. The latter λ windows were also equally spaced (rounded to two significant figures), except for the final values, which were more closely spaced in an attempt to increase phase space overlap: 0.95, 0.97, 0.98, 0.99, 0.999 and 1. All interaction parameters were scaled linearly with respect to λ, except for the LJ interactions, which were perturbed using a soft-core potential[43] with a parameter value $\alpha = 0.5$.

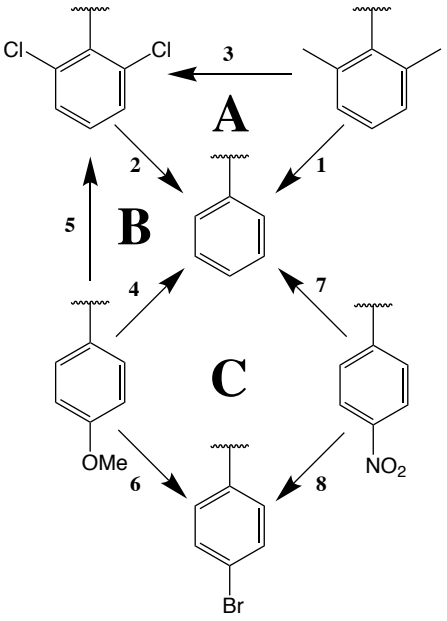
Each λ window involved 25,000-step steepest descent minimisation, 50 ps of NVT equilibration followed by 50 ps of NPT equilibration using a 1 fs timestep and 4 ns NPT production with a 2 fs timestep. In all cases the calculations were repeated after an additional initial 20 ns NPT equilibration at $\lambda = 0$, with this coordinate set being used to prepare simulations at all λ values. Both the 100 ps and the 20 ns protocols were repeated in triplicates, where the only difference was the pseudorandom number seed used for velocity initialisation from the Maxwell–Boltzmann distribution.

All simulations were run at 298 K using the Langevin thermostat[58] with $\tau_T = 1$ ps^{-1}. Equilibration pressure control at 1 bar was achieved with the Berendsen barostat,[54] whereas the production barostat of choice was the Parrinello–Rahman[60] barostat with $\tau_P = 1$ ps^{-1}. In all cases bonds involving hydrogen were constrained using the 4$^{\text{th}}$ order LINCS algorithm.[51] Long-range electrostatics were calculated using particle mesh Ewald (PME)[37] with real space cut-off at 1.2 nm. LJ interactions also had a cut-off at 1.2 nm with a relevant energy and pressure dispersion correction. A Verlet cut-off scheme was used for neighbour list updates every 20 integration steps.

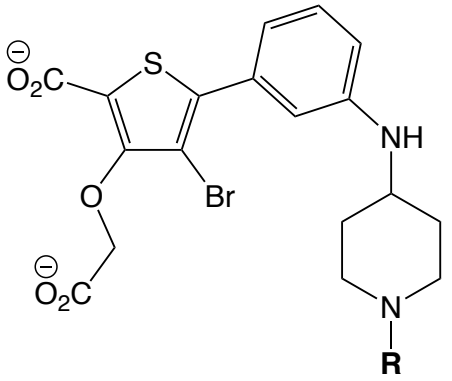
Energy difference (ΔH) readings were taken every 10 ps and were analysed using the Bennett acceptance ratio (BAR)[13] implementation in Python.[195,196] Since in many cases the perturbations involved constrained atoms, the multistate Bennett acceptance ratio (MBAR)[18] approach was not feasible, since LINCS constraint contributions to free energy differences in GROMACS 2018.4 are extrapolated from $\frac{\partial H}{\partial \lambda}$ values and are thus not suitable for non-neighbouring λ windows.



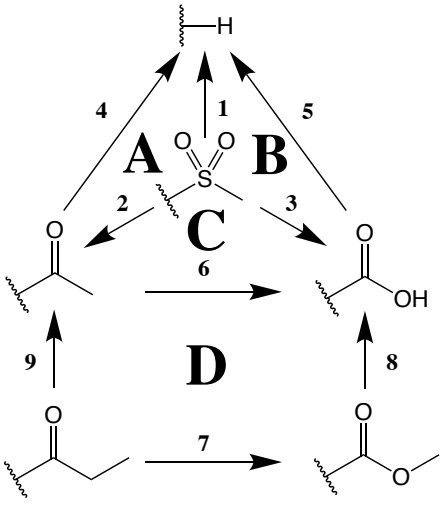
(A) DHFR ligand scaffold



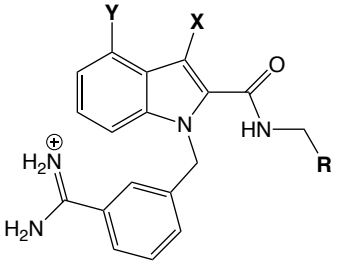
(B) DHFR perturbations



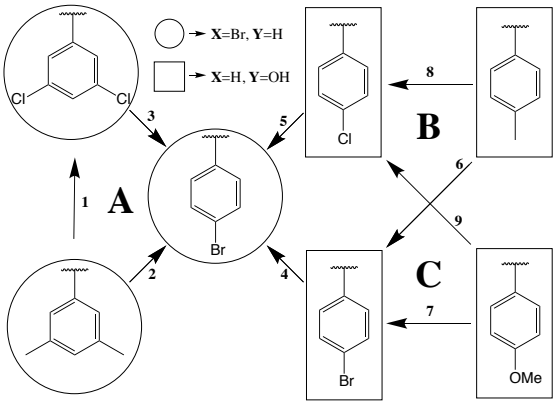
(C) PTP1B ligand scaffold



(D) PTP1B perturbations



(E) FXa ligand scaffold



(F) FXa perturbations

FIGURE 5.2: Ligand scaffolds and perturbations for all three systems. The perturbed ligand pairs are denoted with numbers and thermodynamic cycles are labelled with letters. In all cases perturbations of **R** are shown. In Figure 5.2f, the circular ligands are substituted with **X**=Br and **Y**=H, and the rectangular ligands contain **X**=H and **Y**=OH.

5.2.3 Analysis

When analysing results from different replicates, the errors have been assumed to be approximately normally distributed, and in these cases the sample mean and its associated standard error were used to describe the data. Normality was not assumed in all other cases, e.g. when comparing different crystal structures and equilibration times. In these cases robust statistics, such as the sample median and the interquartile range, were used.

In accordance with the non-assumption of normality, non-parametric rank-based tests were used for all comparisons, namely the Mann–Whitney U test[7] for comparing two populations and the Kruskal–Wallis test[223] for comparing multiple populations. In both tests the null hypothesis is that the mean ranks of the populations are the same and the resulting p-values indicate the probability of observing the data given that the null hypothesis is correct. Since in all cases the compared populations were sampled from practically the same distribution, meaning that the null hypothesis was satisfied, the p-values were not used as a tool for statistical inference but rather as an approximate measure of apparent sampling quality. Where applicable, correlation between populations was also measured in a rank-based fashion, using Kendall's τ.[224]

Appendix B presents data involving 145 independent Kruskal–Wallis tests. Although p-values for each test have been reported, no attempt has been made to demonstrate any statistical significance, since the large number of tests increases the probability of a type I error. These values have therefore been presented purely for information purposes.

5.3 Results and Analysis

5.3.1 Variance Between Structures after 100 ps and 20 ns Equilibration

5.3.1.1 Dihydrofolate Reductase (DHFR)

The calculated $\Delta\Delta G^{\ominus}$ values across all perturbed pairs and crystal structures are shown in Figures 5.3 and 5.4 after 100 ps and an additional 20 ns equilibration, respectively. It can be seen that the "largest" (i.e. perturbing the highest number of atoms) perturbation (pair 5) has the highest total variance with 90% of the samples spanning a confidence interval (CI) of $\sim$2.0 kcal/mol. Correspondingly, the "smallest" perturbation (pair 8) has the lowest spread of $\Delta\Delta G^{\ominus}$ values with CI spanning $\sim$0.5 kcal/mol at the 90% level with all other perturbations exhibiting an approximately similar variance of $\sim$1.0 kcal/mol at this CI. This correlation between perturbation size and total data spread is hardly surprising, since the same computational time was

dedicated to perturbations of varying difficulty. More notably, inter-replicate variance is generally low with no crystal structures exhibiting consistently higher variances. Therefore, the total variance is mostly explained by the inter-structure variance, meaning that the use of simulation repeats starting from the same coordinates is not capturing the variance observed when using different but acceptable crystal structures. This is exemplified by the low p-values shown in Figure 5.3 obtained by the non-parametric Kruskal–Wallis test. It can be noted that 6DAV is a consistent outlier in most of the cases—an unsurprising observation which can be readily explained by the different cofactor in the crystal structure ($NADP^+$, instead of NADPH). However, most inter-crystal differences are significant at the 10% level even if we discard 6DAV—a value which is still concerning, since in principle the choice of initial crystal structure should have little to no effect on the free energy values.

These observations change drastically after 20 ns pre-equilibration at $\lambda = 0$. In this case we observe increased total variances with pairs 5 and 7 exhibiting a spread of $\Delta\Delta G^{\ominus}$ values over $\sim$3.0 kcal/mol at 90% CI. Even the perturbations with the smallest variances span a range of $\sim$1.0 kcal/mol at this CI. This time much larger inter-replicate standard errors are observed with the largest one being 6DAV in pair 2 with $\sigma_{\overline{\Delta\Delta G^{\ominus}}} \approx 1.5$ kcal/mol. Although it is unsurprising that increased decorrelation between replicates results in higher variance, the magnitude of this increase after only 20 ns is remarkable. Although 6DAV is still a rather consistent outlier, it is much less so, resulting in heightened p-values, meaning that there is no significant difference between the crystal structures. Most p-values are now insignificant at the 10% level, again demonstrating the increased loss of memory of the starting crystal structure.

5.3.1.2 Protein Tyrosine Phosphatase 1B (PTP1B)

The corresponding data for PTP1B can be seen in Figures 5.5 and 5.6. Similarly to DHFR, here we observe a total data spread ranging between $\sim$0.5 kcal/mol at the 90% CI for the simpler perturbations (pairs 7 and 8) up to $\sim$1.5 kcal/mol for the most difficult perturbation (pair 3). Inter-replicate variance is generally higher than for DHFR for most perturbations with no structures being consistent outliers. This is illustrated by the generally high p-values for 4 of the pairs. However, the rest of the pairs exhibit consistent significant differences at the 10% level, similar to the results obtained with DHFR. Perhaps the most curious perturbation is pair 9, which has a low p-value and a very low inter-replicate variance, exhibiting a total spread of values of over $\sim$1.0 kcal/mol. This behaviour would not have been expected if we had only used a single crystal structure which exhibits apparent convergence.

The corresponding results after a longer equilibration time are similar to those obtained for DHFR. It can be seen in Figure 5.6 that the total spread of $\Delta\Delta G^{\ominus}$ values is generally much larger. The most remarkable example of this is pair 5 with a total

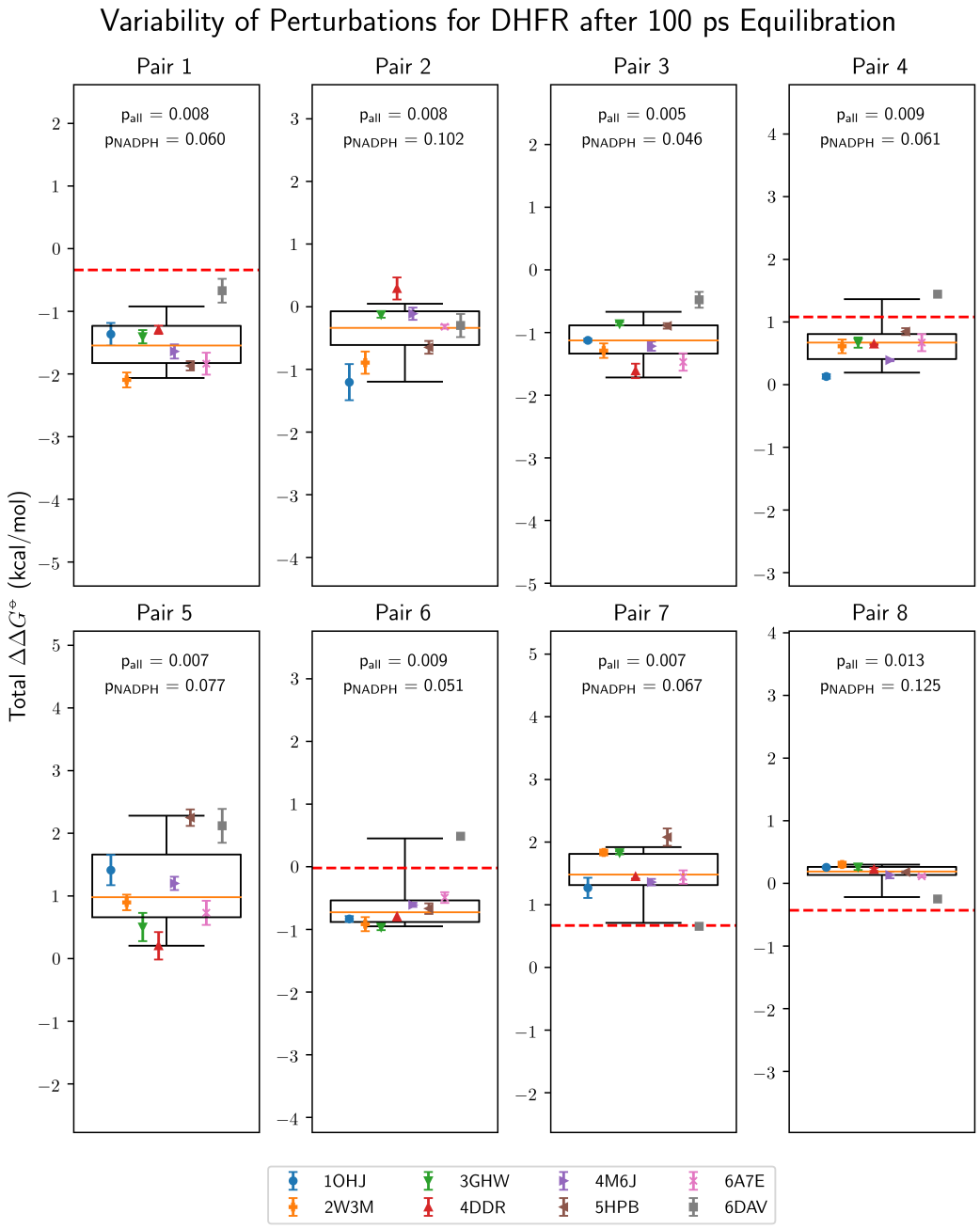


FIGURE 5.3: Box plots of the $\Delta\Delta G^{\ominus}$ values per perturbation for each of the DHFR crystal structures after 100 ps total equilibration time. Each point represents the average of three repeats and the associated error bar is its standard error of the mean. The boxes contain all values between 25th and 75th percentile and the whiskers are based on the 5th and 95th percentile. The p-values have been obtained from the Kruskal–Wallis test on all samples (p_{all}) and on all samples except for 6DAV (p_{NADPH}). The solid orange line shows the median value and the dashed red line denotes the measured experimental value,[2] if available.

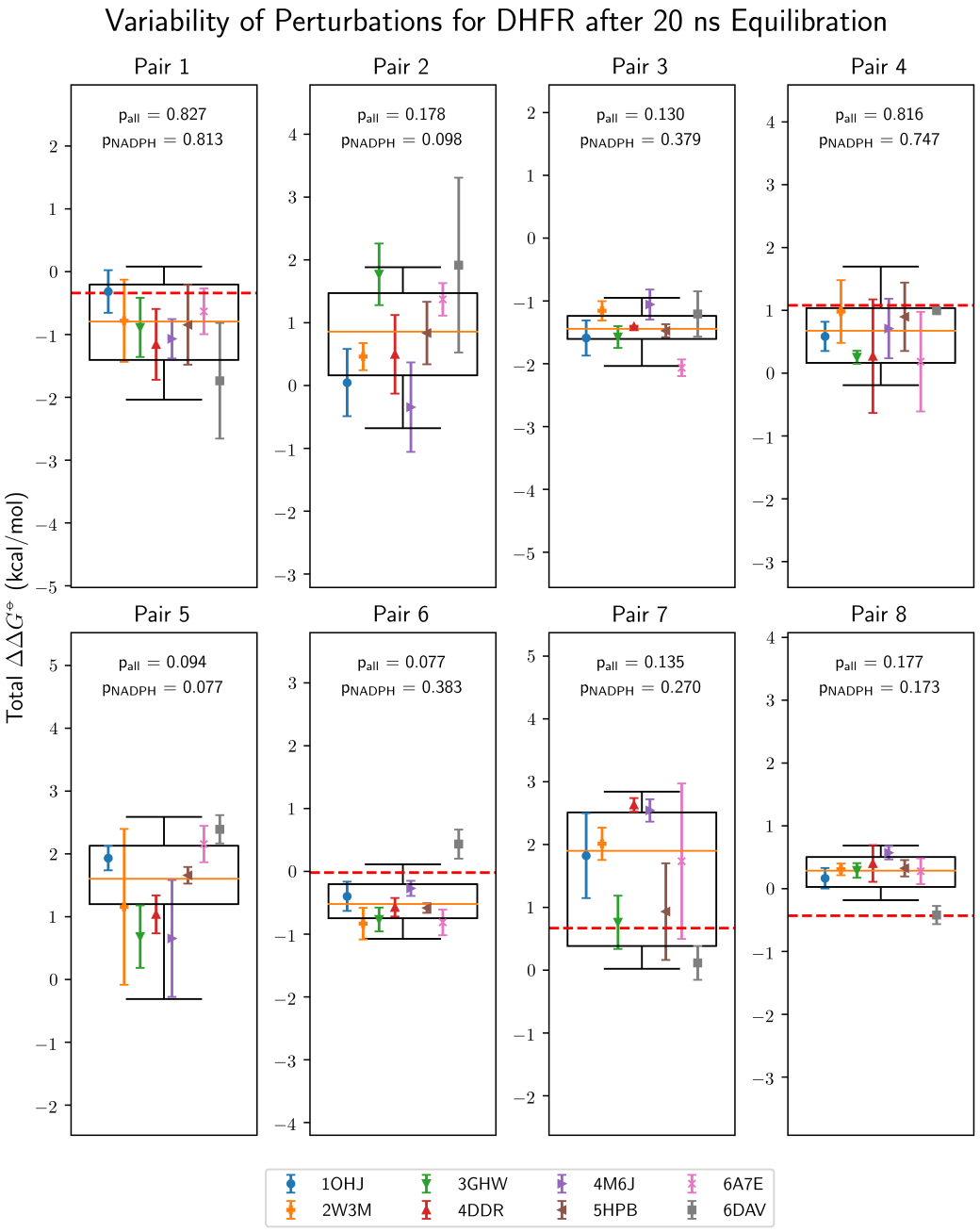


FIGURE 5.4: Box plots of the $\Delta\Delta G^{\ominus}$ values per perturbation for each of the DHFR crystal structures after 20 ns total equilibration time. Each point represents the average of three repeats and the associated error bar is its standard error of the mean. The boxes contain all values between 25th and 75th percentile and the whiskers are based on the 5th and 95th percentile. The p-values have been obtained from the Kruskal–Wallis test on all samples (p_{all}) and on all samples except for 6DAV (p_{NADPH}). The solid orange line shows the median value and the dashed red line denotes the measured experimental value,[2] if available.

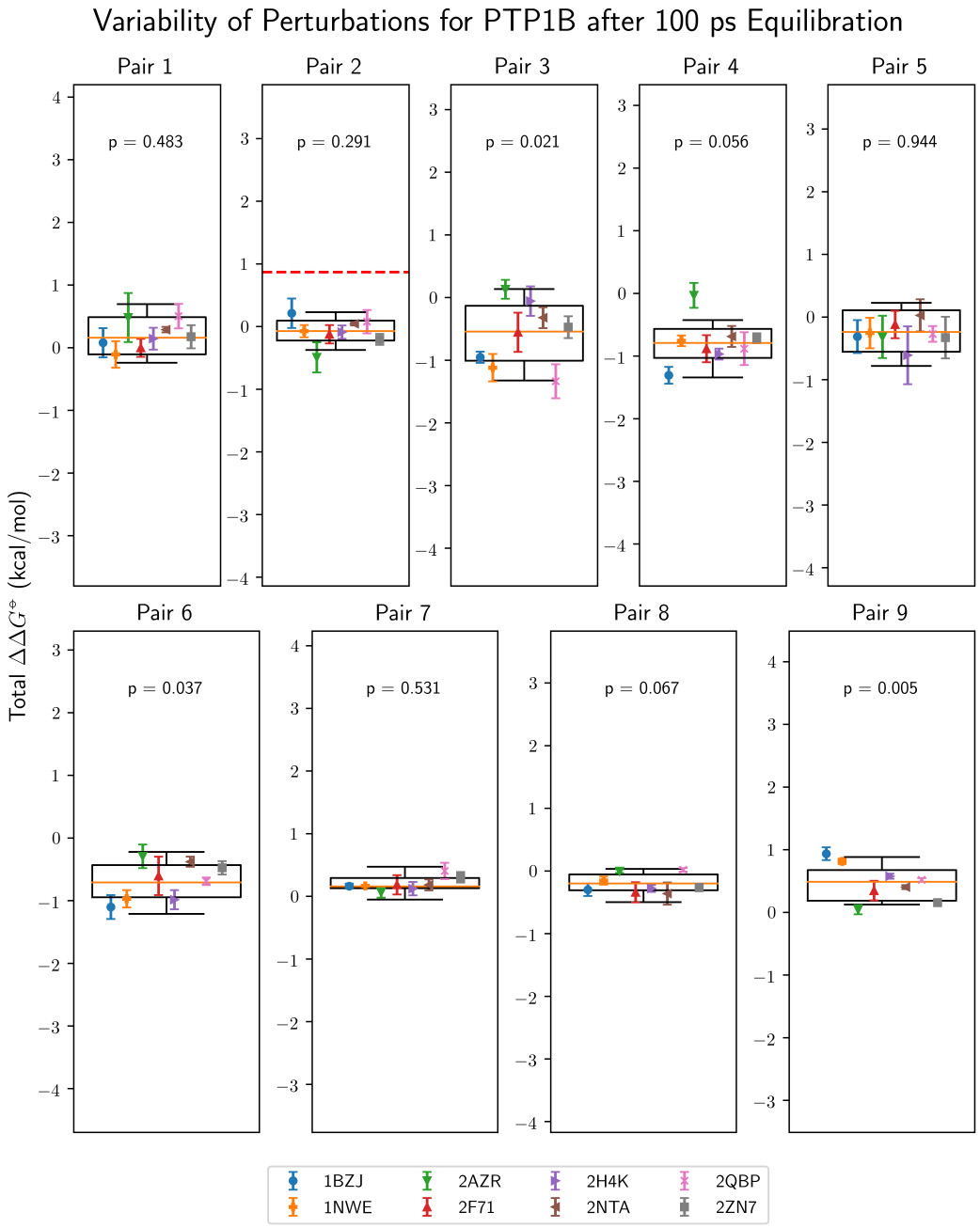


FIGURE 5.5: Box plots of the $\Delta\Delta G^{\circ}$ values per perturbation for each of the PTP1B crystal structures after 100 ps total equilibration time. Each point represents the average of three repeats and the associated error bar is its standard error of the mean. The boxes contain all values between 25th and 75th percentile and the whiskers are based on the 5th and 95th percentile. The p-values have been obtained from the Kruskal–Wallis test on all samples. The solid orange line shows the median value and the dashed red line denotes the measured experimental value,[3] if available.

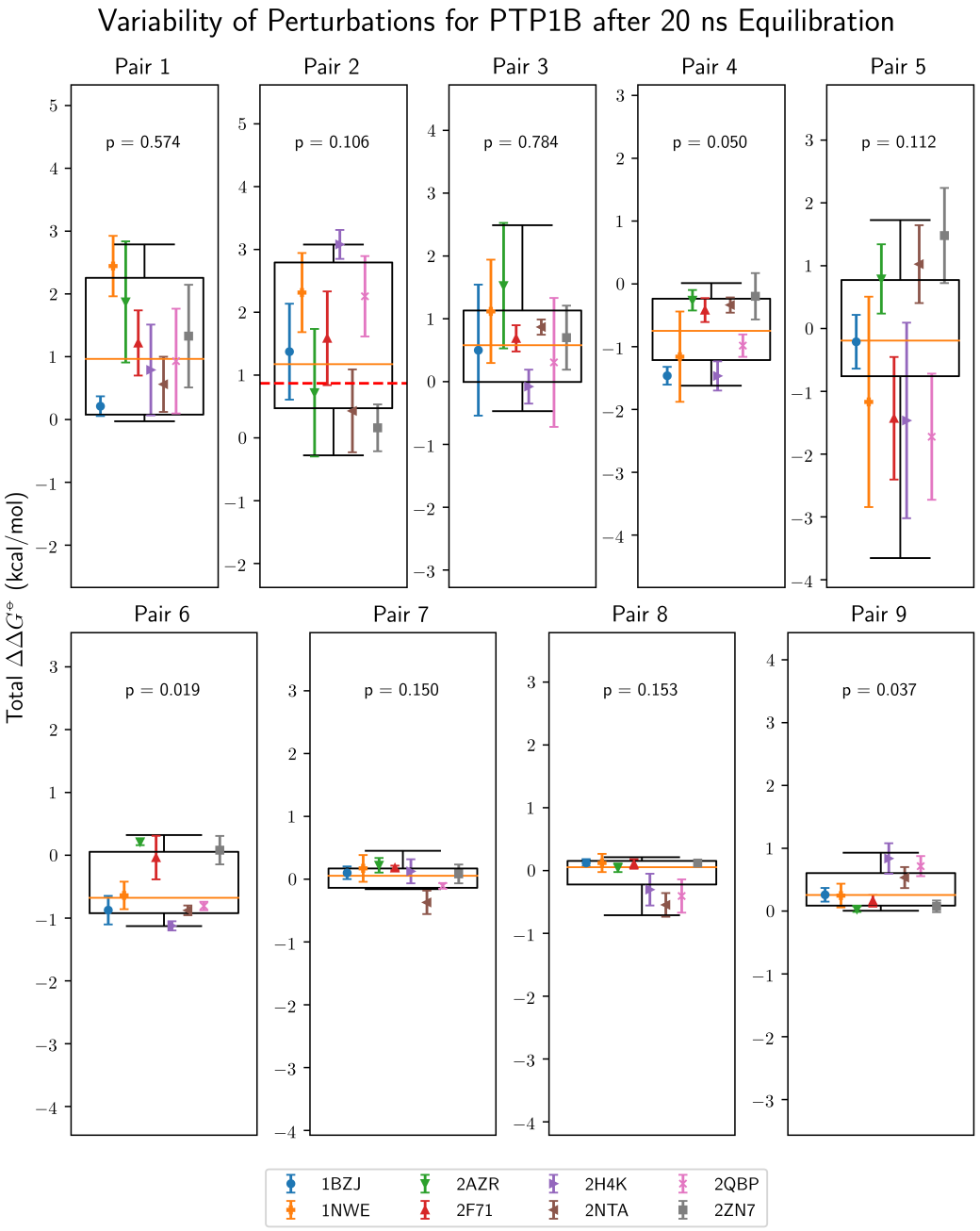


FIGURE 5.6: Box plots of the $\Delta\Delta G^{\circ}$ values per perturbation for each of the PTP1B crystal structures after 20 ns total equilibration time. Each point represents the average of three repeats and the associated error bar is its standard error of the mean. The boxes contain all values between 25th and 75th percentile and the whiskers are based on the 5th and 95th percentile. The p-values have been obtained from the Kruskal–Wallis test on all samples. The solid orange line shows the median value and the dashed red line denotes the measured experimental value,[3] if available.

value range of more than 5.0 kcal/mol at 90% CI. Only the last three pairs exhibit spread of less than 1.0 kcal/mol, whereas all perturbations involving the sulfonamide derivative have an uncertainty of ∼3.0 kcal/mol at 90% CI. In all cases the inter-replicate variance is also markedly higher than the shorter runs with the highest $\sigma_{\overline{\Delta\Delta G^{\ominus}}} > 1.5$ kcal/mol (1NWE, pair 5). Moreover, the new p-values are also on average higher in comparison, once again showing that longer decorrelation times result in reduced distinguishability between different crystal structures. Nevertheless, pairs 4, 6 and 9 exhibit significant differences at the 10% level, implying that the initial crystal structures still influence the obtained free energies. These observations futher demonstrate that any apparent convergence at shorter timescales is usually deceiving.

5.3.1.3 Coagulation Factor Xa (FXa)

The results for FXa are shown in Figures 5.7 and 5.8. In this case we generally observe smaller inter-structure variances than the previous systems with the largest spread being ∼1.0 kcal/mol for the largest perturbation (pair 3) at 90% CI. In some cases this spread is less than 0.2 kcal/mol, indicating good apparent agreement between initial crystal structures. However, inter-replicate variance is generally even lower, resulting in all but two perturbation pairs being significantly different at the 10% level. For example, in pair 3 one can observe a maximum difference of ∼1.0 kcal/mol for two of the crystal structures with little apparent variance, once again highlighting the impact of the choice of initial crystal structure at shorter timescales.

Similarly to previous data, FXa exhibits larger inter-structure variance across all perturbations after prolonged equilibration ranging from ∼0.5 to ∼1.0 kcal/mol. Especially curious is pair 9, which shows dramatic relative increase in variance compared to the short-equilibration results. However, the absolute spread in free energy values is still unremarkable in light of the previous test cases. In all but two perturbations, the differences between initial crystal structures are insignificant at the 10% level, once again demonstrating decreasing dependency of $\Delta\Delta G^{\ominus}$ on the choice of initial crystal structure over time, as expected.

5.3.2 Comparison Between $\Delta\Delta G^{\ominus}$ after 100 ps and 20 ns Equilibration

In addition to the previous analysis, we can compare the distribution of $\Delta\Delta G^{\ominus}$ values across all crystal structures and replicates after 100 ps and 20 ns equilibration. Comparison between median $\Delta\Delta G^{\ominus}$ for each pair and system is shown in Figure 5.9 with associated Mann–Whitney U test p-values in Table 5.1. In some cases we see remarkable median differences in the calculated free energies of approximately 1.0–1.5 kcal/mol (DHFR: pair 2; PTP1B: pairs 1, 2 and 3), whereas most other values, including all of the FXa perturbations, appear to approximately agree by visual

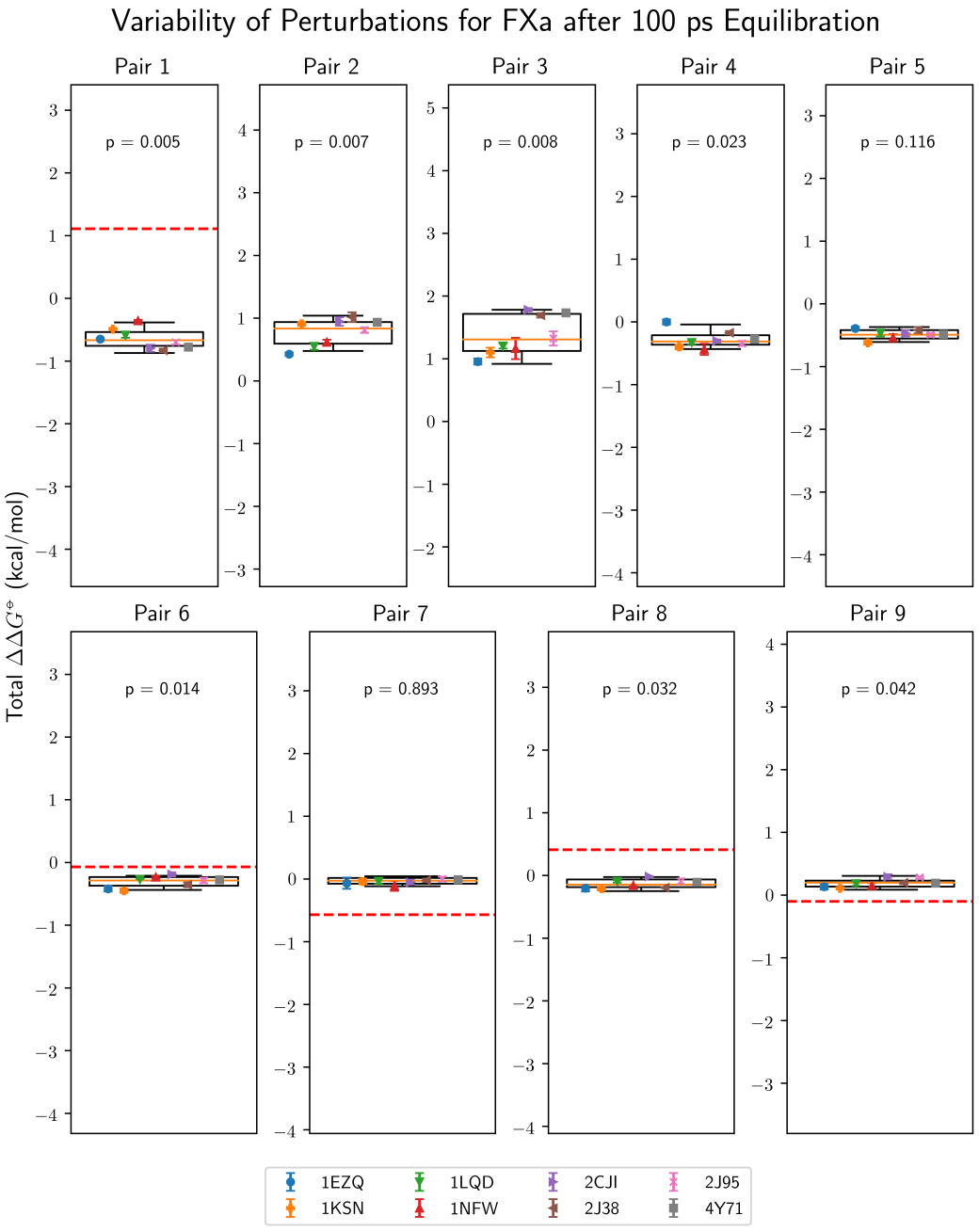


FIGURE 5.7: Box plots of the $\Delta\Delta G^{\ominus}$ values per perturbation for each of the FXa crystal structures after 100 ps total equilibration time. Each point represents the average of three repeats and the associated error bar is its standard error of the mean. The boxes contain all values between 25th and 75th percentile and the whiskers are based on the 5th and 95th percentile. The p-values have been obtained from the Kruskal–Wallis test on all samples. The solid orange line shows the median value and the dashed red line denotes the measured experimental value,[4] if available.

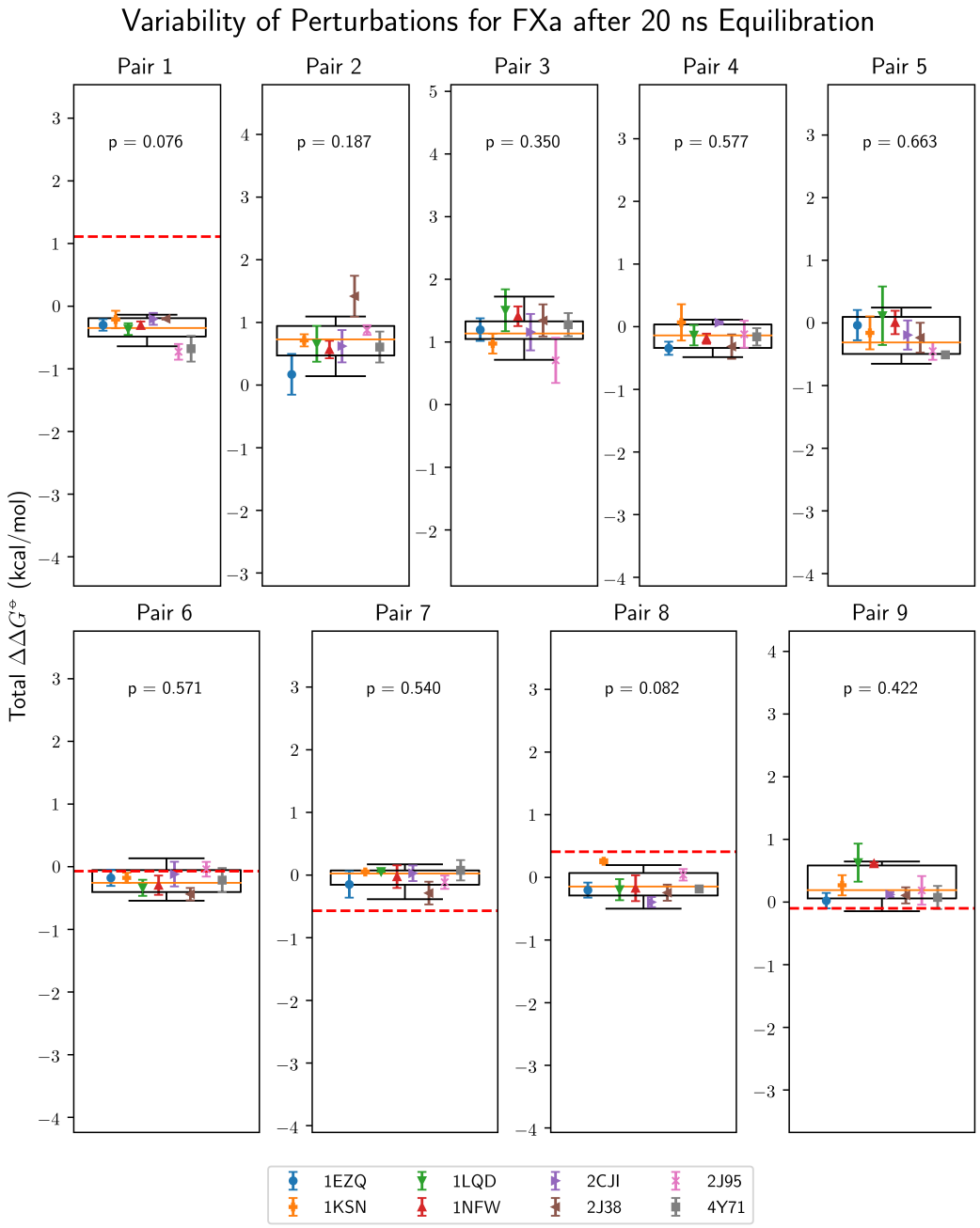
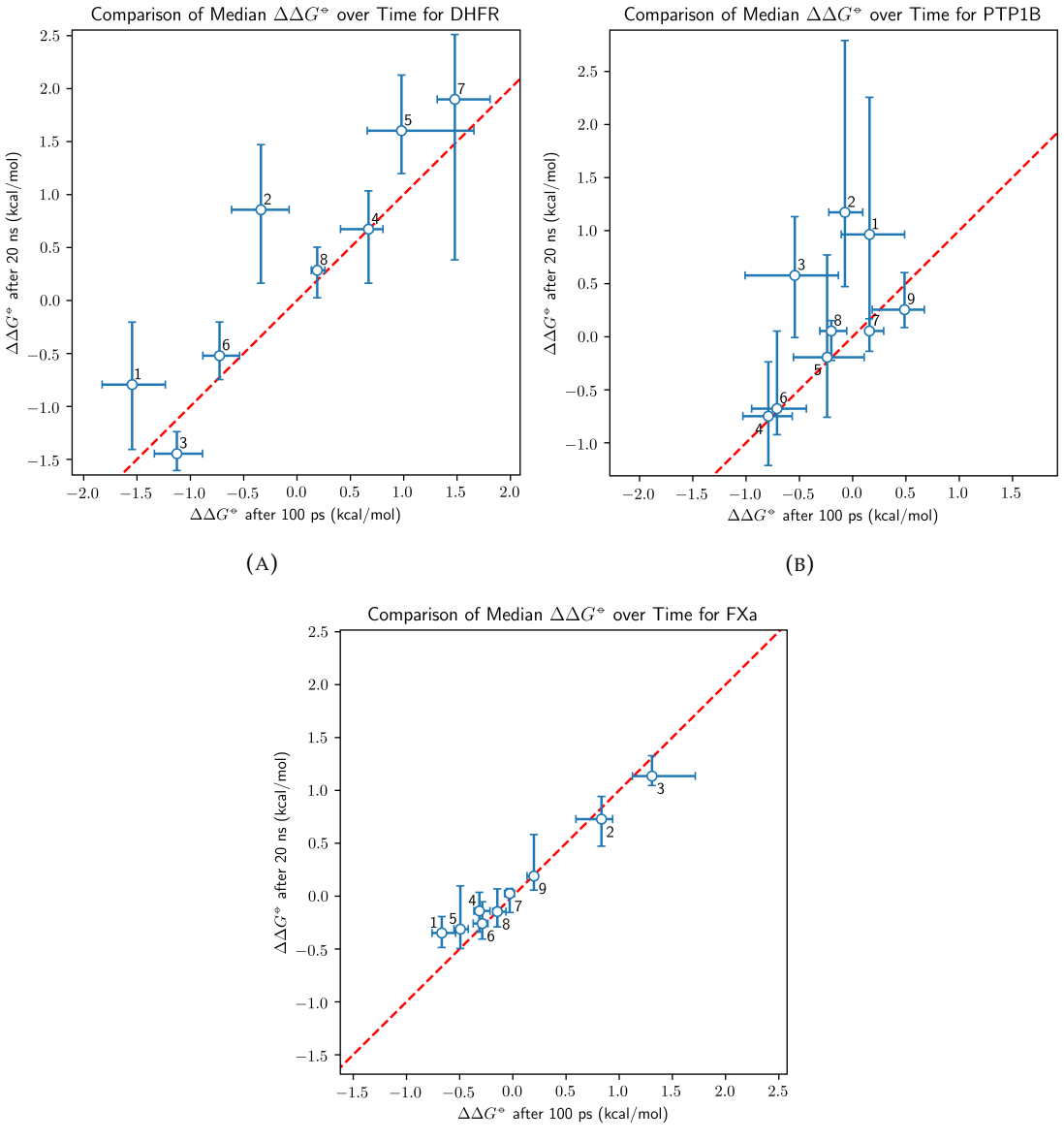


FIGURE 5.8: Box plots of the $\Delta\Delta G^{\ominus}$ values per perturbation for each of the FXa crystal structures after 20 ns total equilibration time. Each point represents the average of three repeats and the associated error bar is its standard error of the mean. The boxes contain all values between 25th and 75th percentile and the whiskers are based on the 5th and 95th percentile. The p-values have been obtained from the Kruskal–Wallis test on all samples. The solid orange line shows the median value and the dashed red line denotes the measured experimental value,[4] if available.

inspection. However, the Mann–Whitney U test indicates significant differences at 2% CI for: DHFR pairs 1, 2 and 3; PTP1B pairs 1, 2, 3, 7 and 8; FXa pairs 1 and 5, which constitute more than a third of all perturbations. This indicates that even after comparing across protein crystal structures and repeats we observe significant time-dependent sampling changes. Nevertheless, it has to be noted that these differences could to some extent arise from the sampling bias introduced by prolonged equilibration at only a single λ value and one should ideally compare datasets where all λ values have been independently equilibrated for 20 ns. In this study, this was not feasible due to computational resource limitations.



(A) (B)

(C)

FIGURE 5.9: Comparison of median $\Delta\Delta G^{\ominus}$ values across all initial crystal structures and replicates after short (100 ps) and long (20 ns) equilibration for DHFR (Figure 5.9a), PTP1B (Figure 5.9b) and FXa (Figure 5.9c). All error bars indicate 25%–75% CI. The dashed red line represents the line $y = x$.

	Pair Number								
System	1	2	3	4	5	6	7	8	9
DHFR	0.003	0.000	0.007	0.992	0.140	0.164	0.370	0.177	N/A
PTP1B	0.011	0.000	0.000	0.757	0.370	0.445	0.016	0.016	0.115
FXa	0.000	0.503	0.124	0.045	0.015	0.288	0.307	0.861	0.829

TABLE 5.1: P-values calculated using the two-sided Mann–Whitney U test[7] after comparison of $\Delta\Delta G$ values obtained across different initial crystal structures and replicates between 100 ps and 20 ns equilibration.

5.3.3 Cycle Closure Errors

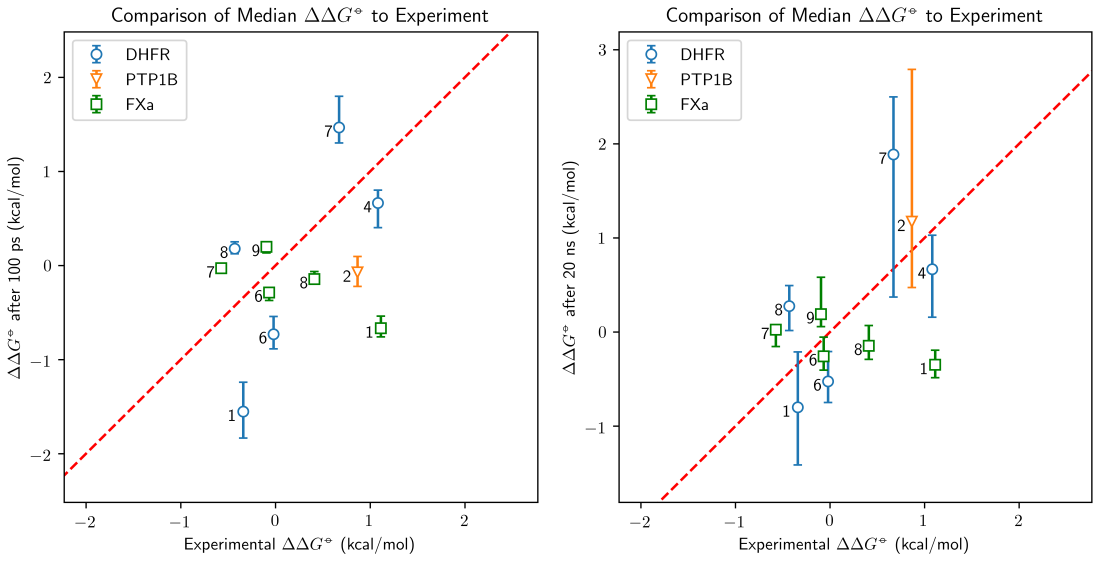
Since the Gibbs free energy G is a state function, any combination of perturbations which returns to the initial state must yield a net free energy change of zero. Any deviation from this value indicates insufficient sampling and lack of convergence. It can be seen (Table 5.2) that in most cases cycle closure errors indicate apparent convergence (less than 1.0 kcal/mol) after a short equilibration both on a crystal-by-crystal basis and on average, with the notable exception of all cycles involving PTP1B and the sulfonamide ligand derivative (cycles A, B and C). However, this apparent convergence is not observed after longer equilibration with some cycle closure errors surpassing the 1.0 kcal/mol barrier for some crystal structures. Nevertheless, all average cycle closures are within 1.0 kcal/mol with the exception of DHFR, cycle B, which has a magnitude of 1.67 kcal/mol, despite exhibiting apparent convergence at shorter equilibration times. This is a striking observation, since one would expect that a net equilibration time of $\sim$0.5 µs and a total sampling time of $\sim$4 µs per perturbation to exhibit unconditional convergence, especially for these rather straightforward perturbations. These results show that any apparent convergence at shorter timescales can be deceiving even for simple systems and low cycle closure errors do not necessarily imply sufficient sampling.

5.3.4 Comparison to Experiment

While not the focus of this study, which is concerned with reproducibility and precision, rather than accuracy, it is nevertheless informative to compare the above results to experimental $\Delta\Delta G^{\ominus}$ values. Here we only compare direct perturbations against experiment, as opposed to thermodynamic chains. It is shown in Figure 5.10 that the extensively equilibrated median binding free energies generally move slightly closer to experimental values over time. The relative ranking, represented by Kendall's τ,[224] changes insignificantly from 0.09 to 0.13, indicating very weak correlation to experimental data. The mean absolute deviation (MAD) also improves weakly with more equilibration from 0.73 to 0.61 kcal/mol. Both of these metrics are influenced by the low experimental free energy magnitudes ($\sim$1.0 kcal/mol), meaning

		Cycle Closure Errors (kcal/mol)					
		Minimum		Maximum		Average	
System	Cycle	100 ps	20 ns	100 ps	20 ns	100 ps	20 ns
DHFR	A	0.03	0.07	0.97	2.44	0.02	0.30
	B	0.08	0.40	0.76	3.35	0.08	1.67
	C	0.01	0.03	0.38	1.39	0.06	0.24
PTP1B	A	0.72	0.05	1.31	1.37	1.06	0.47
	B	0.58	0.08	2.11	2.50	1.05	0.80
	C	0.01	0.00	1.02	2.03	0.18	0.28
	D	0.03	0.02	0.60	0.84	0.20	0.11
FXa	A	0.02	0.01	0.30	0.90	0.06	0.12
	B	0.01	0.21	0.17	0.53	0.03	0.05
	C	0.01	0.01	0.19	0.84	0.03	0.26

TABLE 5.2: Absolute cycle closure errors for all systems after 100 ps and 20 ns equilibration. The cycles have been calculated per structure as the average of three replicates and denoted according to Figure 5.2. The three columns represent the cycle closure errors from the best- and worst-performing crystal structures, as well as the average cycle closure errors between all structures.



(A) (B)

FIGURE 5.10: Comparison of median $\Delta\Delta G^{\ominus}$ values across all initial crystal structures and replicates for some of the denoted pairs after short (100 ps, Figure 5.10a) and long (20 ns, Figure 5.10b) against experiment.[2–4] The associated error bars indicate 25%–75% CI and the dashed red line represents the line $y = x$.

that the MAD is more likely to appear favourable and the relative ranking is dominated by noise. Owing to the size of the dataset, the low magnitude of the experimental free energy values and the high variability between different replicates, it can be concluded that the achieved improvement in comparison to experiment over time is not substantial for these test cases.

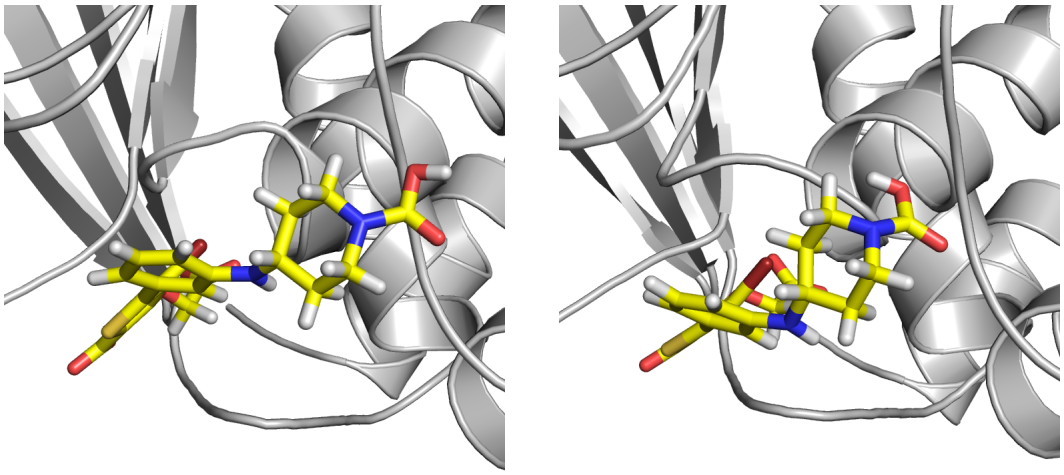
5.3.5 The Origin of Long-Timescale Variance

Owing to the complex nature of biomolecular systems, it is difficult to narrow down the reason for the observed increase in inter-replicate variance with simulation time. One obvious way to compare different replicates is to quantify the changes which occur during the 20 ns equilibration based on the final trajectory frame. The most apparent differences by visual inspection indicate the presence of various rare events with transition times larger than several nanoseconds, such as rotations of ligand torsions with high kinetic barriers. The most striking example is the rotation of the acidic hydrogen in PTP1B, pair 5 (Figure 5.11). Comparison of free energy calculations starting from the two different rotamers reveal a median difference of more than 4.0 kcal/mol, indicating that this rotation is likely the primary reason for the extraordinary variance, observed in Figure 5.6. Another conspicuous example of rare events determining the outcome of a free energy calculation is the sulfonamide bond rotation in the first three PTP1B perturbations (Figure 5.12), with each rotamer exhibiting an average of approximately 1.0–2.0 kcal/mol difference to the other rotamers. Detailed analysis of these and all other rotamers can be found in Appendix B (Figures B.1 to B.4), revealing the prevalence of this trend in many of the perturbations involving DHFR and PTP1B. This analysis also shows that such ligand flexibility is observed to a much lesser extent for FXa, thereby explaining the comparatively low free energy variance even after extended equilibration.

Naturally, it is expected that the protein backbone also has an impact on the increased free energy variance over time. However, analysing such a high-dimensional dataset requires an immense amount of data points in the form of $\Delta\Delta G^{\circ}$ values. With only 24 $\Delta\Delta G^{\circ}$ values per perturbation, establishing a statistically significant connection between protein internal degrees of freedom and calculated free energies is not feasible and we are going to attribute most of the long-timescale variability to slow ligand motions—a conclusion, which is supported by all of the data presented above.

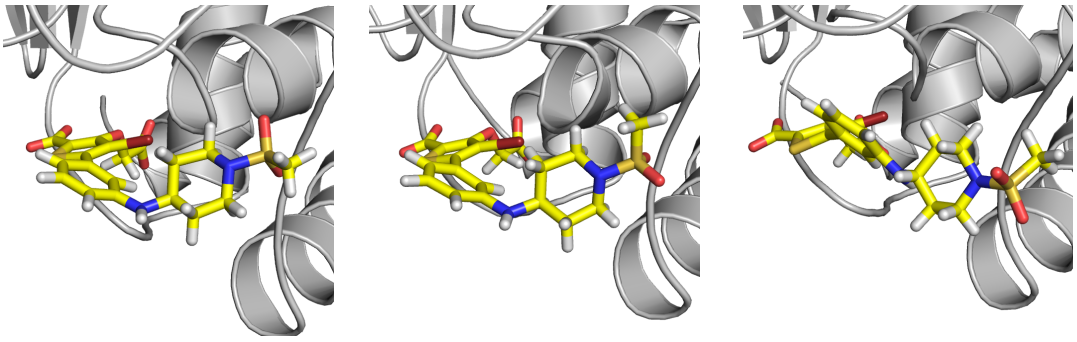
5.4 Discussion

There are several important lessons to be learned from the above results. Most strikingly, they show that at short timescales different protein crystal structures can disagree significantly over the free energy change and these differences can be more than 1.0 kcal/mol in magnitude. Such short-timescale simulations are commonly used in practice, most notably in commercial implementations,[156,157] making these results highly relevant to state-of-the-art applications of alchemical methods. More worryingly, most of these results appear well-converged, as evidenced by the low inter-replicate variance and the satisfactory cycle closure errors. The issue of using a single crystal structure is now apparent: this choice can covertly affect the relative



(A) (B)

FIGURE 5.11: Acidic hydrogen rotation in pair 5 observed in extended PTP1B simulations. Images generated from the final trajectory frame of the extended equilibration for 1BZJ (Figure 5.11a) and 1NWE (Figure 5.11b).



(A) (B) (C)

FIGURE 5.12: Sulfonamide rotation in pair 3 observed in extended PTP1B simulations. Images generated from the final trajectory frame of the extended equilibration for 1BZJ (Figure 5.12a), 2AZR (Figure 5.12b) and 2H4K (Figure 5.12c).

ranking of compounds, even when the free energy changes appear to be too large for this to be likely to occur.

The above analysis also shows that these inter-structure differences are largely reduced after a prolonged equilibration time and the inter-triplicate differences become more representative of their true uncorrelated values. However, even after ∼24 ns of dynamics one can often distinguish between the different initial structures, meaning that this initial choice has long-term effects on the free energy calculation. Nevertheless, these results show that the proposition that multiple short simulations are preferable to a single long simulation does not necessarily capture the full nature of the problem: replicates are necessary but insufficient for convergence, while many important long-timescale motions are practically inaccessible at shorter timescales, regardless of the number of replicates. Therefore, one needs both multiple replicates

for statistical confidence and longer simulations for physical validity—a requirement which is rarely practically feasible with current computational capabilities.

One obvious way to practically circumvent this problem is to run short simulations over more than one crystal structure. While the average of the resulting free energy values would likely be biased, at least the researcher would be aware of the minimal uncertainty in their results. However, this approach would result in reduced quality of ligand sampling, since pure molecular dynamics is not good at exploring multiple binding modes at short timescales.[225] Alternatively, one could run repeats over several binding poses determined by e.g. docking, but the same problem of determining the correct weights of each binding pose persists, resulting in biased free energy differences.

Another possibly more preferable working alternative is running longer simulations on one crystal structure with enhanced sampling of the ligand degrees of freedom. The results from longer timescales are typically more sensitive to ligand conformational changes and the initial crystal structure becomes less influential on the free energy changes over time. A protocol combining alchemical free energy (AFE) calculations and replica exchange with solute scaling (REST2)[94] has long been used in commercial implementations, such as FEP+[156] and has recently been used over longer timescales.[226–228] Although it is expected that this approach would lead to a much higher variance due to the number of ligand degrees of freedom and decreased phase space overlap between neighbouring states due to the Hamiltonian rescaling, this variance would be more representative of the true result and this approach would be much less likely to exhibit false convergence. In all cases it is highly recommended to run at least triplicate simulations.

We also observed that larger perturbations result in much more variable free energy estimates, a largely expected result. However, even the simplest of perturbations should be treated with care. More specifically, cycle closure errors can indicate false convergence and should therefore only be used to demonstrate insufficient sampling. Indeed, it was shown that extensive sampling usually results in higher and more realistic cycle closure errors, meaning that this criterion is necessary but not sufficient for convergence.

Since it was unclear from the above data whether prolonged equilibration affects comparison to experiment significantly, one might argue that better sampling might not be necessarily cost-efficient for applications. While this is possible considering the accuracy provided by current force fields, it has to be remembered that all computational time saved from less sampling, results in reduced physical and statistical confidence. Therefore, one should at the very least use timescales and enhanced sampling techniques providing sufficient ligand conformational sampling

whenever possible, so that the binding conformation is not completely dependent on the ligand alignment method and/or the researcher's intuition.

5.5 Conclusion

This chapter has shown the influence of initial crystal structure and extended equilibration time on the binding free energy values of three different systems: DHFR, PTP1B and FXa. The results indicate that at short timescales, initial crystal structure differences consistently result in statistically, although not necessarily numerically, significant changes in $\Delta\Delta G^{\ominus}$, sometimes reaching differences of over 1.0 kcal/mol. Furthermore, large perturbations result in higher sensitivity to the initial structure at short timescales.

At longer timescales, there is a marked increase in the inter-replicate variance in $\Delta\Delta G^{\ominus}$, which makes the results from different initial crystal structures appear more similar. In many cases, the slow changes in ligand conformation, which become more common at these timescales, are a significant contributor to this variance. The extent to which the protein degrees of freedom impact these results is not clear and is to be investigated in future work. Sometimes this prolonged sampling can significantly change the expected free energy value. Nevertheless, the extra sampling results in no significant improvement against experiment. In addition, it has been shown that thermodynamic cycle closure values can often indicate false convergence at short timescales, meaning that long-timescale enhanced sampling is needed even for simple perturbations.

This chapter has emphasised the importance of long-timescale dynamics and enhanced sampling in AFE calculations, as well as performing multiple repeats with the same initial configurations. Therefore, an optimal protocol needs to find the balance between the number of repeats, simulation length, and the number of λ windows in the general case. One such protocol will be developed in Chapters 8 to 10.

Chapter 6

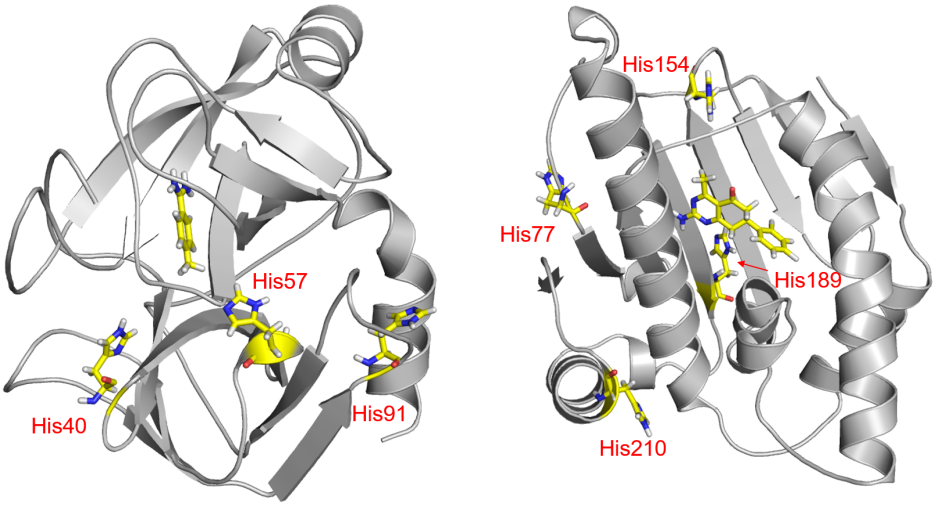
Sensitivity of Binding Free Energy Calculations to Histidine Tautomers and Rotamers

6.1 Introduction

It was earlier discussed in Section 5.2.1 that the choice of initial protein crystal structure can significantly affect the whole simulation setup process, most notably the assignment of amino acid side-chain protonation states. This uncertainty in the relevant protonation state is amplified in histidine, which has three accessible protonation/tautomeric (PT) states at physiological pH. Although these can often be reliably assigned by investigating the protein hydrogen bonding network, these considerations are not always possible or unambiguous. Moreover, such approaches hinder the automation of free energy calculations and this setback becomes more relevant as advances in computational power increase the throughput of alchemical binding free energy calculations.

The discussion in Section 3.2.1.4 also highlighted some recent studies which investigate the influence of ligand PT states on alchemical free energy (AFE) calculations and the impact of histidine PT states on docking results. However, the impact of histidine PT states on AFE calculations is not currently known and will therefore be the focus of this chapter.

In this chapter, the influence of changing a single histidine state to one of its tautomeric states—δ (HID), ϵ (HIE) and its protonated state (HIP)—on the resulting protein–ligand binding free energies will be explicitly explored. In addition, different initial imidazole ring rotamers will be investigated in order to shed light on how their



(A) Trypsin (B) Hsp90

FIGURE 6.1: The trypsin (Figure 6.1a) and Hsp90 (Figure 6.1b) histidine residues studied in this work. Ligands and histidines shown as stick.

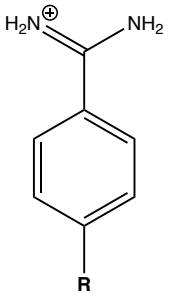
kinetics affect the resulting free energy values. The two systems that will be considered hereafter are trypsin and heat shock protein 90 (Hsp90).

Trypsin is an ideal system for this study, since it has three histidine residues (His40, His57 and His91, Figure 6.1a) at different distances from the binding site (approximately 0.9, 0.4 and 1.6 nm, respectively, Table 6.1), with one of them being part of a catalytic triad responsible for peptide bond hydrolysis. It has also been thoroughly explored in the literature, including in free energy studies. In this work a crystal structure and a perturbation network closely related to a previous study[229] will be used.

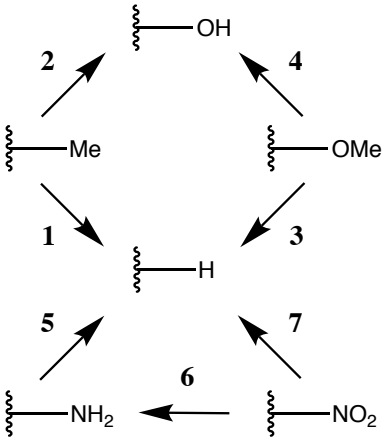
In contrast to trypsin, Hsp90 is a more challenging system to study, due to the higher mobility of its protein backbone.[230] Indeed, several types of tertiary structures have been observed in combination with different ligands, in some cases resulting in a tertiary structure change between two closely related ligands.[6] This behaviour makes Hsp90 a good target for AFE calculation challenges. In this work we will adhere to a single tertiary structure and a perturbation network of closely related ligands, which have again been explored in a previous study.[231] Hsp90 has four histidine residues: His77, His154, His189 and His210 (Figure 6.1b) with distances from the binding site ranging from 0.7 to 2.3 nm (Table 6.1).

6.2 Methods

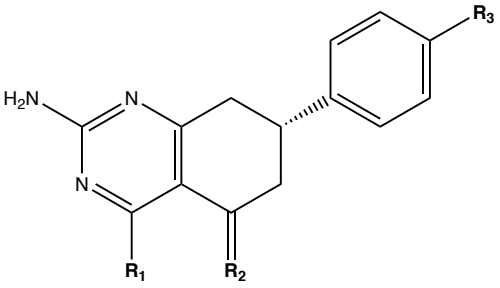
6.2.1 System Preparation



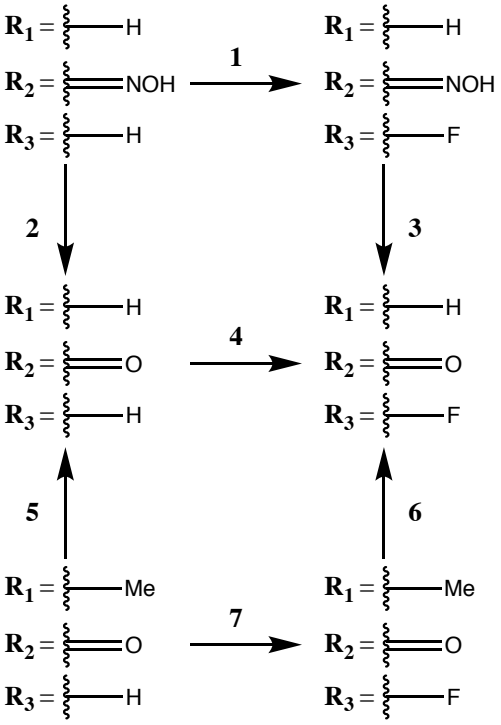
(A) Trypsin ligand scaffold



(B) Trypsin perturbations



(C) Hsp90 ligand scaffold



(D) Hsp90 perturbations

FIGURE 6.2: Ligand scaffolds and perturbations for trypsin and Hsp90.

All of the system preparation in this study was performed in ProtoCaller (Chapter 4). The crystal structures used were 3PTB[232] for Trypsin and 3FT8[6] for Hsp90 and were obtained from the Protein Data Bank.[181] The non-terminal missing residues in 3FT8 were added using pdbfixer[233] and protonation to the default tautomeric states at pH = 7 was performed using PDB2PQR.[185] Because of the stochasticity of the residue

		Histidine		
System	Perturbation	His40	His57	His91
Trypsin	Pairs 1–7	0.9	0.4	1.6

		Histidine			
System	Perturbation	His77	His154	His189	His210
Hsp90	Pairs 2, 3	1.7	0.8	1.9	2.2
	Pairs 5, 6	1.4	0.7	1.9	2.2
	Pairs 1, 4, 7	2.3	1.6	1.8	2.0

TABLE 6.1: Shortest distances in nm between each histidine residue and each perturbed ligand group for both trypsin and Hsp90.

addition procedure, it was only performed once for each system and the resulting structure was used as a starting point for the subsequent setups. In all cases the initial crystallographic water molecules were retained. Even though Hsp90 is known to be biologically active as a homodimer,[234] only the first chain of its PDB structure was used for the following simulations in the interest of computational performance.

The histidine protonation, tautomeric and rotameric (PTR) states assigned by PDB2PQR were in both cases taken as a reference. Afterwards, a series of structures was generated based on these references where the only difference was a single histidine PT (HID, HIE and HIP) and/or rotameric state (native, or with ring rotation of 180°). This resulted in an additional five structures per histidine, or 16 structures in total for Trypsin and 21 structures for Hsp90. Each of these structures was used in a series of ligand–ligand perturbations (7 for both Trypsin and Hsp90), as shown in Figure 6.2. The binding mode of each ligand was manually generated from the reference binding mode of the initial crystal structure and was kept consistent across all simulations. Finally, the structure generating procedure for each combination as well as the subsequent simulation were repeated three times, in order to account for stochastic effects during the solvation procedure and the initial velocity generation.

The system parametrisation, ligand alignment and subsequent complex solvation followed the same system preparation procedure as described in Chapter 5.

6.2.2 Simulation

Each of the simulations described below followed the simulation protocol with a short 100 ps equilibration, described in Chapter 5.

6.2.3 Analysis

In the following discussion, no assumptions have been made about the underlying distributions of the resulting free energy values. In all cases, robust estimators such as the sample median, confidence interval and the mean absolute deviation (MAD) have been reported. The non-parametric Kruskal–Wallis test[223] has been used for statistical comparisons between all free energy distributions.

The Kruskal–Wallis test verifies the null hypothesis that the mean ranks of the compared populations are the same. The associated p-values therefore indicate the probability of observing the data given the correctness of the null hypothesis. In order to combine p-values testing the same null hypothesis on different datasets, Fisher's method[235] has been used to report an aggregate p-value (hereafter referred to as "Fisher average"). In these cases, the separate p-values have only been reported for information purposes, since a large number of independent tests on the same hypothesis runs the risk of a type I error (false rejection of the null hypothesis).

Three different types of p-values have been considered during the analysis: PT, rotamer and total p-values (p_{prot}, p_{rot} and p_{tot}). p_{prot} was calculated as the Fisher average of the Kruskal–Wallis p-values obtained from comparing the free energy distributions at each pair of histidine PT states (HID–HIE, HIE–HIP and HID–HIP). In this case, both initial rotamer states for each PT state were used together for each comparison. p_{rot}, on the other hand, was obtained as the Fisher average of the Kruskal–Wallis p-values resulting from comparing the free energy distributions corresponding to each pair of different initial rotamers (i.e. one comparison each for HID, HIE and HIP). p_{tot} is simply the Kruskal–Wallis p-value obtained from comparing all six groups of free energies.

Similarly, three different types of absolute deviations were calculated: PT, rotamer and total. Each of these absolute deviation distributions was obtained by considering the absolute difference of every possible combination of free energy values between different replicates from the different groups described in the previous paragraph. A reference, or inter-replicate absolute deviation distribution was also calculated from the unsigned free energy differences between all separate repeats (i.e. between repeats 1 and 2, 2 and 3, and 1 and 3). The inter-replicate MAD was then used as a baseline for comparison to the other MADs to determine their effect size relative to the sampling noise.

6.2.4 Markov State Models (MSMs)

Markov state models (MSMs)[71] were also used to analyse the interconversion kinetics between different histidine rotamers (see Section 2.5 for some background on MSMs).

To generate these, all trajectories at $\lambda = 0$ and $\lambda = 1$ across all simulations dedicated to a particular histidine residue were used to determine an averaged kinetic profile across different protonation states and ligands. This resulted in a total of 252×4 ns trajectories (7 ligands, 6 histidine states, 3 repeats, 2 λ values) with a resolution of 5 ps being used for estimating the rotation kinetics of each histidine residue.

The χ_1 dihedral angles (CA-CB-CG-CD2) were measured for each histidine using MDTraj 1.9.3[236] and were subsequently clustered using Gaussian mixture models (GMMs),[237] as implemented in scikit-learn 0.24.2[238] using the default settings. The number of Gaussian components used was manually determined based on the observed number of modes for the dihedral distribution of each histidine residue: 3 in the case of Hsp90 His77 and His210 and 2 in all other cases.

The discretised clusters were then used to fit a Bayesian hidden MSM,[72,74] as implemented in PyEMMA 2.5.9.[239] The estimator fitting procedure was performed using the default settings to generate 100 different transition matrices over a range of different lag times τ in the range of 5–100 ps. The distributions of the slowest implied timescales $t_{slowest}(\tau)$ from each of these lag times were subsequently obtained using the formula:[77]

$$t_{slowest}(\tau) = -\frac{\tau}{\ln \lambda_{slowest}(\tau)} \tag{6.1}$$

where $\lambda_{slowest}$ is the second largest eigenvalue of the corresponding stochastic matrix. One of the τ values exhibiting satisfactory convergence was afterwards chosen and the corresponding mean and sample standard deviation of $t_{slowest}(\tau)$ at this lag time were calculated and will be reported later in the text.

6.3 Results and Analysis

6.3.1 Free Energy Calculations at Different Histidine States

6.3.1.1 Trypsin

The calculated $\Delta\Delta G^{\ominus}$ values across different trypsin histidine states for His40, His57 and His91 are presented in Figures 6.3 to 6.5. It can be seen that the highest $\Delta\Delta G^{\ominus}$ variability is observed for His57. For most perturbations in this case the maximum discrepancies between different histidine states are between 0.5 and 1.0 kcal/mol. However, the two perturbations involving a nitro group (pairs 6 and 7) show maximum deviations of $\sim$1.5 kcal/mol. Interestingly, the other perturbation involving an amine group (pair 5) shows the lowest discrepancies, even compared to less polar

perturbations, such as pair 1. On average, the MAD between different PT state with the same rotamer is 0.48 kcal/mol, while the MAD between different rotamers is 0.38 kcal/mol (Table 6.2). Combined with the fact that both the PT and rotamer Fisher averaged Kruskal–Wallis p-values (p_{prot} and p_{rot}) are much smaller than 0.001, it is straightforward to conclude that this result is both numerically and statistically significant.

Although these observations are hardly surprising for a binding site histidine, similar behaviour is observed for His40, even though it is $\sim$0.9 nm away from the ligand. In this case, most maximum free energy deviations between different histidine states are $\sim$0.5 kcal/mol, reaching $\sim$1.0 kcal/mol for pair 7. Although the influence of this histidine is clearly dampened in comparison with His57, the discrepancies are still highly significant with $p_{prot} \ll 0.001$ and $p_{rot} \approx 0.001$. At a PT MAD of 0.26 kcal/mol and a rotamer MAD of 0.15 kcal/mol, it can be seen that the free energy discrepancies between different rotameric states are on average less pronounced compared to His57.

The observed free energy discrepancies are much smaller for His91, which is $\sim$1.6 nm away from the binding site, with both PT and rotamer MADs being 0.13 and 0.12 kcal/mol, respectively. However, statistically significant differences between different PT states can still be observed, reaching an average p-value of $<$0.001, with the lowest p-values being observed for pairs 1 and 2. On the other hand, the statistical significance of the influence of the initial His91 rotamers is substantially diminished, reaching an average p-value of 0.154.

6.3.1.2 Heat Shock Protein 90 (Hsp90)

The corresponding $\Delta\Delta G^{\ominus}$ values for Hsp90 after changing His77, His154, His189 and His210 are shown in Figures 6.6 to 6.9. In contrast to Trypsin, the discrepancies between the free energies obtained from different PTR states are less pronounced: all PTR MADs are in the range 0.13–0.19 kcal/mol (Table 6.2).

Despite being the second closest histidine residue to the binding site, His77 exhibits lower PTR MADs than the other histidines at 0.14 and 0.13 kcal/mol, respectively. While the low magnitude of these can be explained by the relatively large distances from the binding site (Table 6.1), it is not clear why the absolute deviations are on average lower than those observed for residues as far as 2.3 nm from the active site. Nevertheless, this difference is negligibly small and is likely the effect of statistical noise. Interestingly, the Fisher averaged rotamer p-value for His77 is the most significant of all histidine residues, despite the lower MAD. This is mainly reflected in pairs 4 and 5, which show significant rotamer p-values at the 5% level.

His154, being the closest residue to the binding site, shows a much less pronounced sensitivity to the different histidine states compared to trypsin, with the largest

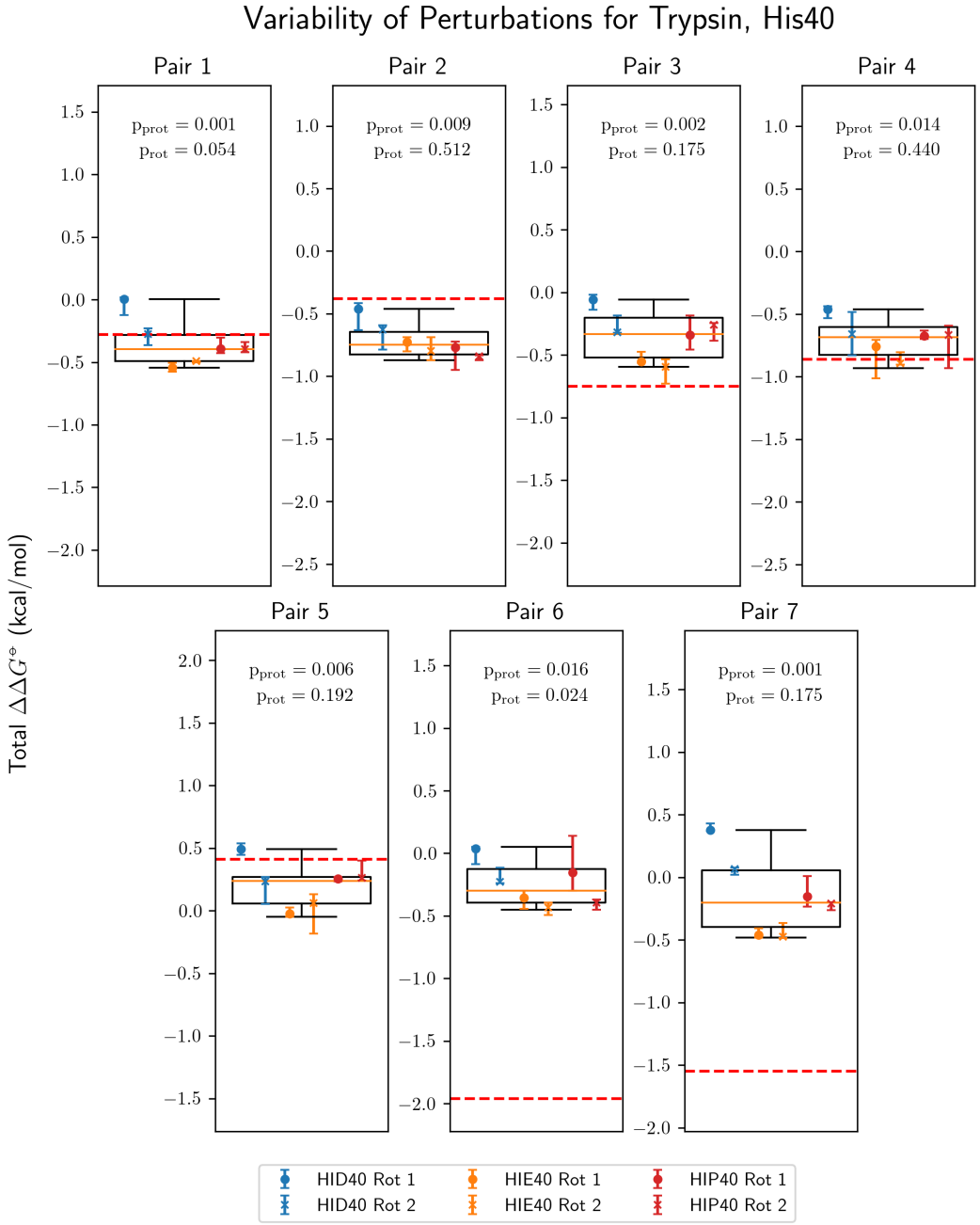


FIGURE 6.3: Box plots of the $\Delta\Delta G^{\ominus}$ values for different Trypsin His40 PTR states at each ligand perturbation. Each point represents the median of three repeats, and the associated error bar extends between the other two repeats. The boxes contain all values from the merged dataset between the 25th and 75th percentile and the whiskers extend to the 5th and 95th percentile. The solid orange line represents the median of the whole dataset, while the dashed red line shows the experimentally obtained free energy value,[5] if available. The associated p-values obtained from a Kruskal–Wallis test between different protonation states have been reported as p_{prot}, while the Fisher averages of the three pairwise rotamer Kruskal–Wallis p-values have been annotated as p_{rot}.

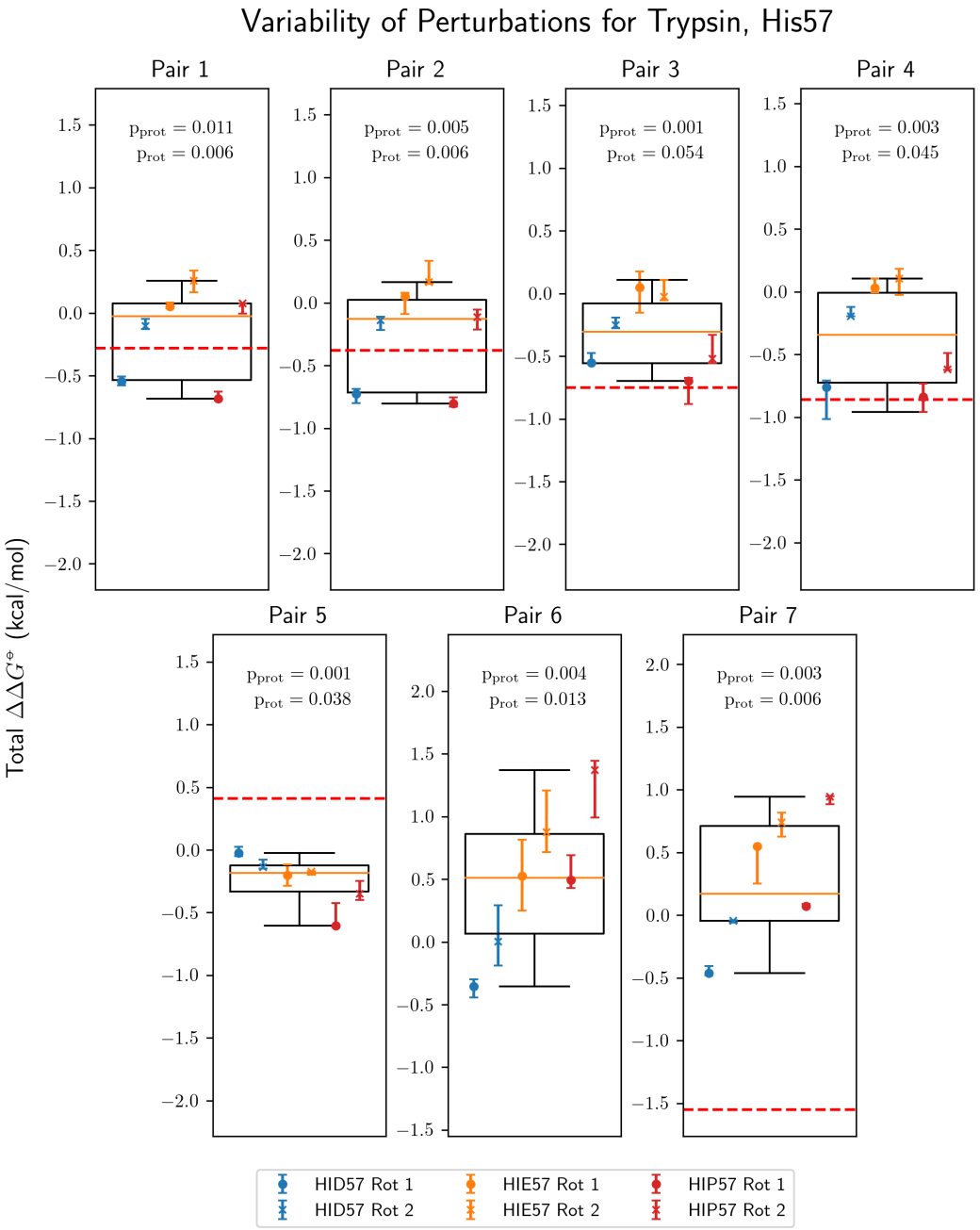


FIGURE 6.4: Box plots of the $\Delta\Delta G^{\ominus}$ values for different Trypsin His57 PTR states at each ligand perturbation. Each point represents the median of three repeats, and the associated error bar extends between the other two repeats. The boxes contain all values from the merged dataset between the 25th and 75th percentile and the whiskers extend to the 5th and 95th percentile. The solid orange line represents the median of the whole dataset, while the dashed red line shows the experimentally obtained free energy value,[5] if available. The associated p-values obtained from a Kruskal–Wallis test between different protonation states have been reported as p_{prot}, while the Fisher averages of the three pairwise rotamer Kruskal–Wallis p-values have been annotated as p_{rot}.

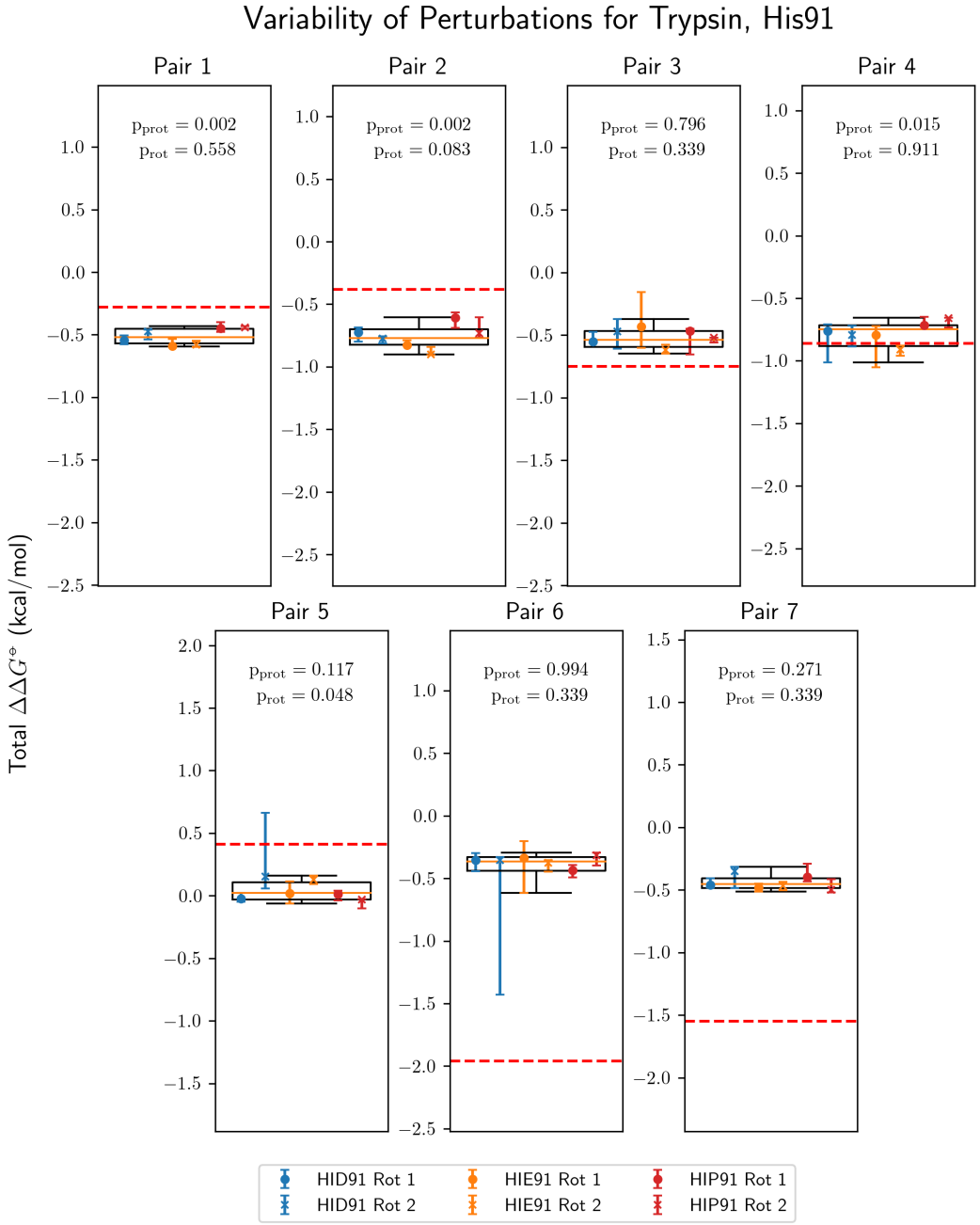


FIGURE 6.5: Box plots of the $\Delta\Delta G^{\ominus}$ values for different Trypsin His91 PTR states at each ligand perturbation. Each point represents the median of three repeats, and the associated error bar extends between the other two repeats. The boxes contain all values from the merged dataset between the 25th and 75th percentile and the whiskers extend to the 5th and 95th percentile. The solid orange line represents the median of the whole dataset, while the dashed red line shows the experimentally obtained free energy value,[5] if available. The associated p-values obtained from a Kruskal–Wallis test between different protonation states have been reported as p_{prot}, while the Fisher averages of the three pairwise rotamer Kruskal–Wallis p-values have been annotated as p_{rot}.

		Average p-value			MAD (kcal/mol)		
System	Residue	PT	R	Total	PT	R	Total
Trypsin	His40	<0.001	0.014	<0.001	0.26 (0.91)	0.15 (0.59)	0.24 (0.92)
	His57	<0.001	<0.001	<0.001	0.48 (1.60)	0.38 (1.00)	0.49 (1.90)
	His91	<0.001	0.154	0.003	0.13 (1.10)	0.12 (1.10)	0.13 (1.20)
Hsp90	His77	0.237	0.043	0.118	0.14 (0.46)	0.13 (0.48)	0.14 (0.50)
	His154	0.963	0.081	0.257	0.18 (0.69)	0.19 (0.70)	0.12 (0.70)
	His189	0.548	0.420	0.417	0.18 (0.81)	0.17 (0.58)	0.11 (0.81)
	His210	0.057	0.124	0.042	0.17 (0.71)	0.16 (0.44)	0.12 (0.71)

TABLE 6.2: The average p_{prot}, p_{rot} and p_{tot} values obtained by Fisher's method for each of the histidine residues and the corresponding mean absolute deviations in kcal/mol with the maximum absolute deviations given in parentheses.

median discrepancy being observed in pair 3, where the two HIE rotamers differ by 0.33 kcal/mol. This is the main reason His154 has the second lowest cumulative rotamer p-value of 0.081. However, its cumulative PT p-value is the highest in the dataset (0.963), implying that the observed free energy differences between the different PT states are insignificant.

His189 shows the least significant results compared to the other histidine residues, with $p_{prot} = 0.548$ and $p_{rot} = 0.420$. Although the highest median free energy discrepancy is 0.30 kcal/mol (pair 2, HIE), it is overshadowed by the large inter-replicate variance, thereby losing statistical significance. In most cases, it can be seen that the free energy discrepancies are numerically negligible, a result which is in agreement with the long distance from the binding site (1.8–1.9 nm, Table 6.1).

Despite being the furthest histidine from the binding site (2.0–2.2 nm, Table 6.1), His210 can exhibit significant median discrepancies between different rotamers, reaching 0.36 kcal/mol (pair 3 HIE). Furthermore, the free energy differences between pair 5 HID and HIE are observed to be 0.15 kcal/mol, and this result along with pairs 1 and 2 contribute to its comparatively lowest cumulative PT p-value, reaching 0.057. Interestingly, the rotamer p-value (0.124) is comparable to that of trypsin His91, despite the latter being 0.4–0.6 nm closer to the binding site.

6.3.2 Histidine Mobility

It is informative to relate the free energy discrepancies between different histidine rotamers to the mobility of the histidine residues, since it is expected that the residues with faster kinetics would result in free energy values that are less sensitive to their initial rotamer. To achieve this, a Bayesian hidden MSM was used after preliminary clustering in dihedral space using GMMs. Hidden MSMs are crucial for obtaining meaningful kinetics, because they are highly robust to clustering errors.[72] Since the latter can be prevalent during the clustering of dihedral angles, a regular MSM

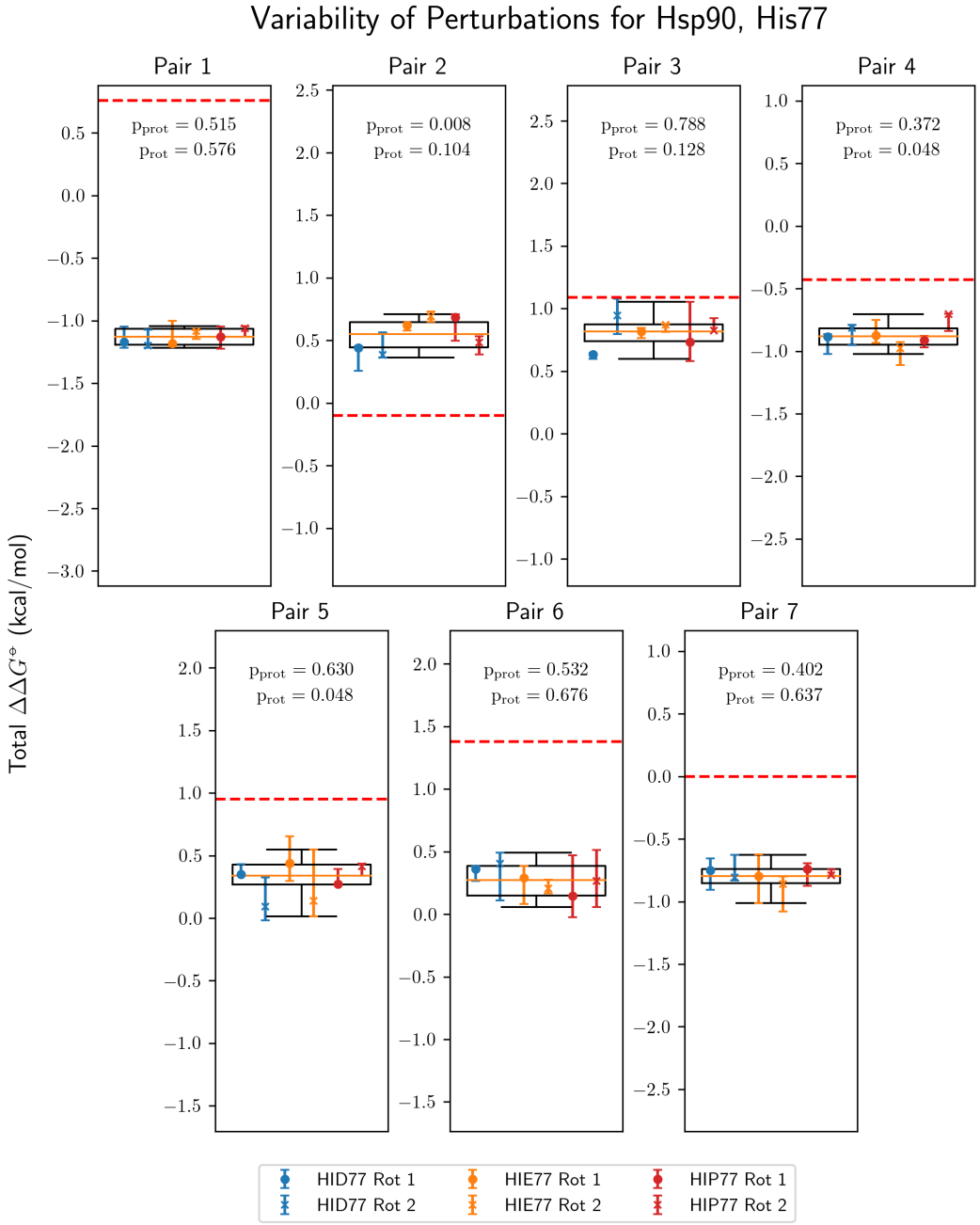


FIGURE 6.6: Box plots of the $\Delta\Delta G^{\ominus}$ values for different Hsp90 His77 PTR states at each ligand perturbation. Each point represents the median of three repeats, and the associated error bar extends between the other two repeats. The boxes contain all values from the merged dataset between the 25th and 75th percentile and the whiskers extend to the 5th and 95th percentile. The solid orange line represents the median of the whole dataset, while the dashed red line shows the experimentally obtained free energy value,[6] if available. The associated p-values obtained from a Kruskal–Wallis test between different protonation states have been reported as p_{prot}, while the Fisher averages of the three pairwise rotamer Kruskal–Wallis p-values have been annotated as p_{rot}.

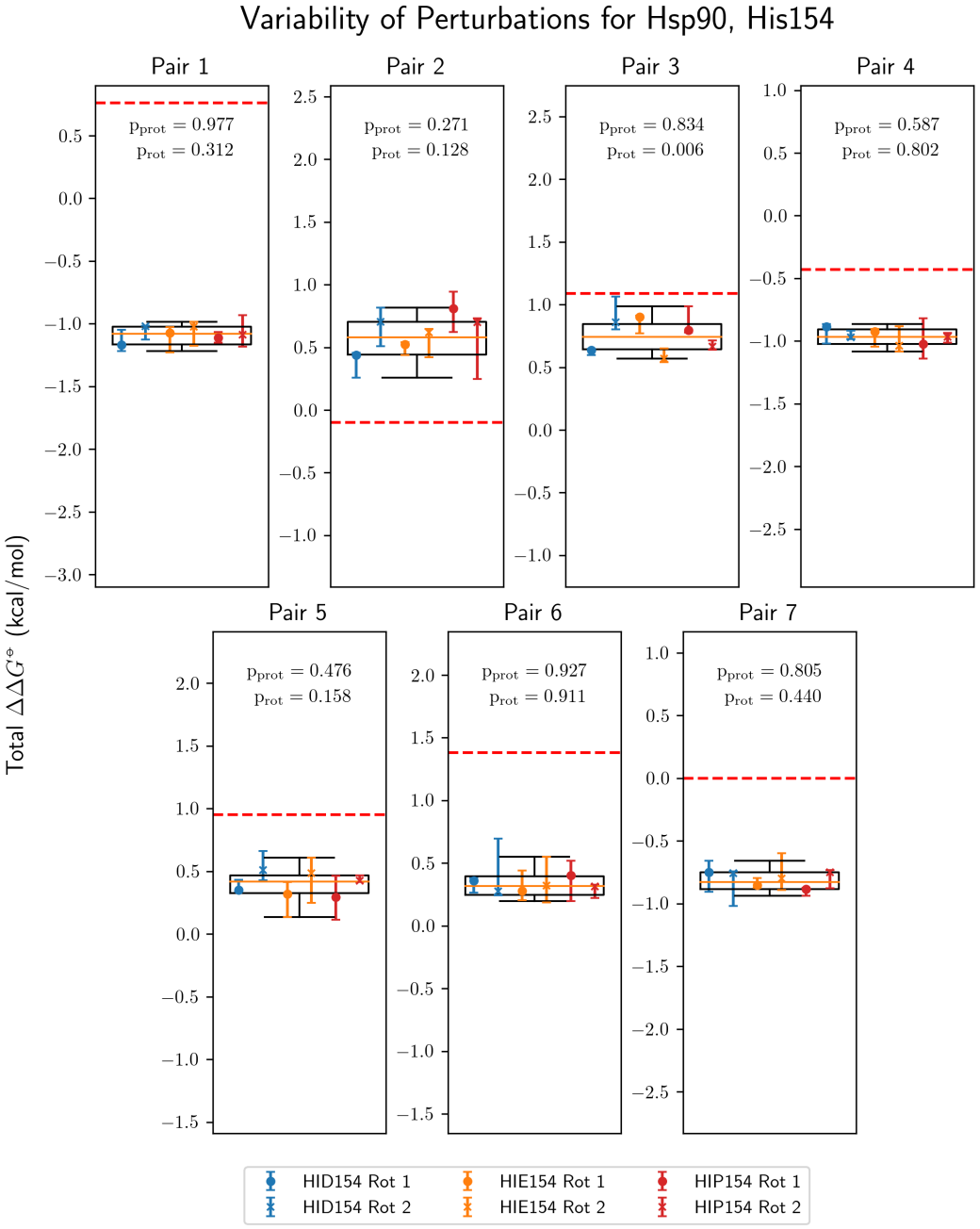


FIGURE 6.7: Box plots of the $\Delta\Delta G^{\ominus}$ values for different Hsp90 His154 PTR states at each ligand perturbation. Each point represents the median of three repeats, and the associated error bar extends between the other two repeats. The boxes contain all values from the merged dataset between the 25th and 75th percentile and the whiskers extend to the 5th and 95th percentile. The solid orange line represents the median of the whole dataset, while the dashed red line shows the experimentally obtained free energy value,[6] if available. The associated p-values obtained from a Kruskal–Wallis test between different protonation states have been reported as p_{prot}, while the Fisher averages of the three pairwise rotamer Kruskal–Wallis p-values have been annotated as p_{rot}.

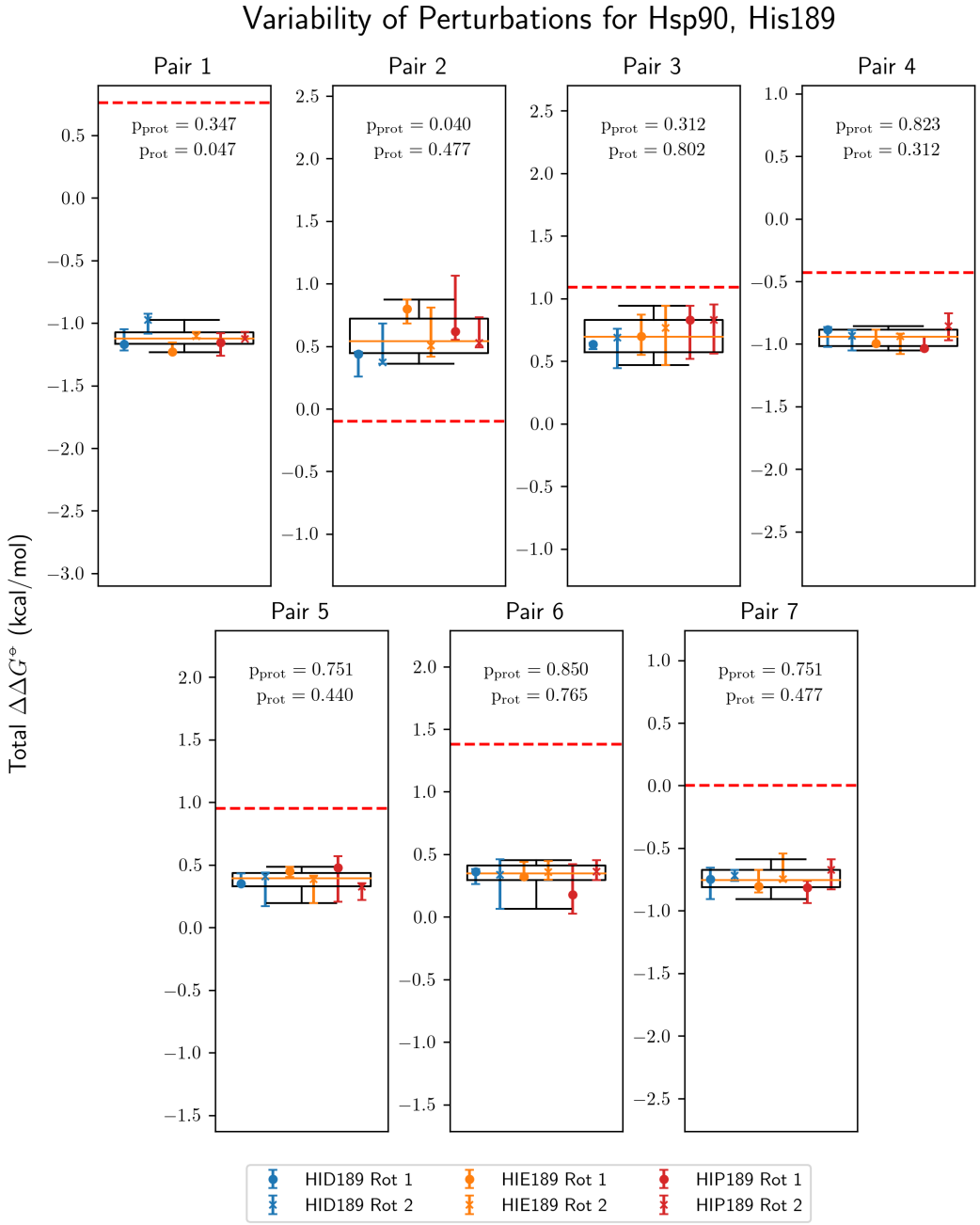


FIGURE 6.8: Box plots of the $\Delta\Delta G^{\ominus}$ values for different Hsp90 His189 PTR states at each ligand perturbation. Each point represents the median of three repeats, and the associated error bar extends between the other two repeats. The boxes contain all values from the merged dataset between the 25th and 75th percentile and the whiskers extend to the 5th and 95th percentile. The solid orange line represents the median of the whole dataset, while the dashed red line shows the experimentally obtained free energy value,[6] if available. The associated p-values obtained from a Kruskal–Wallis test between different protonation states have been reported as p_{prot}, while the Fisher averages of the three pairwise rotamer Kruskal–Wallis p-values have been annotated as p_{rot}.

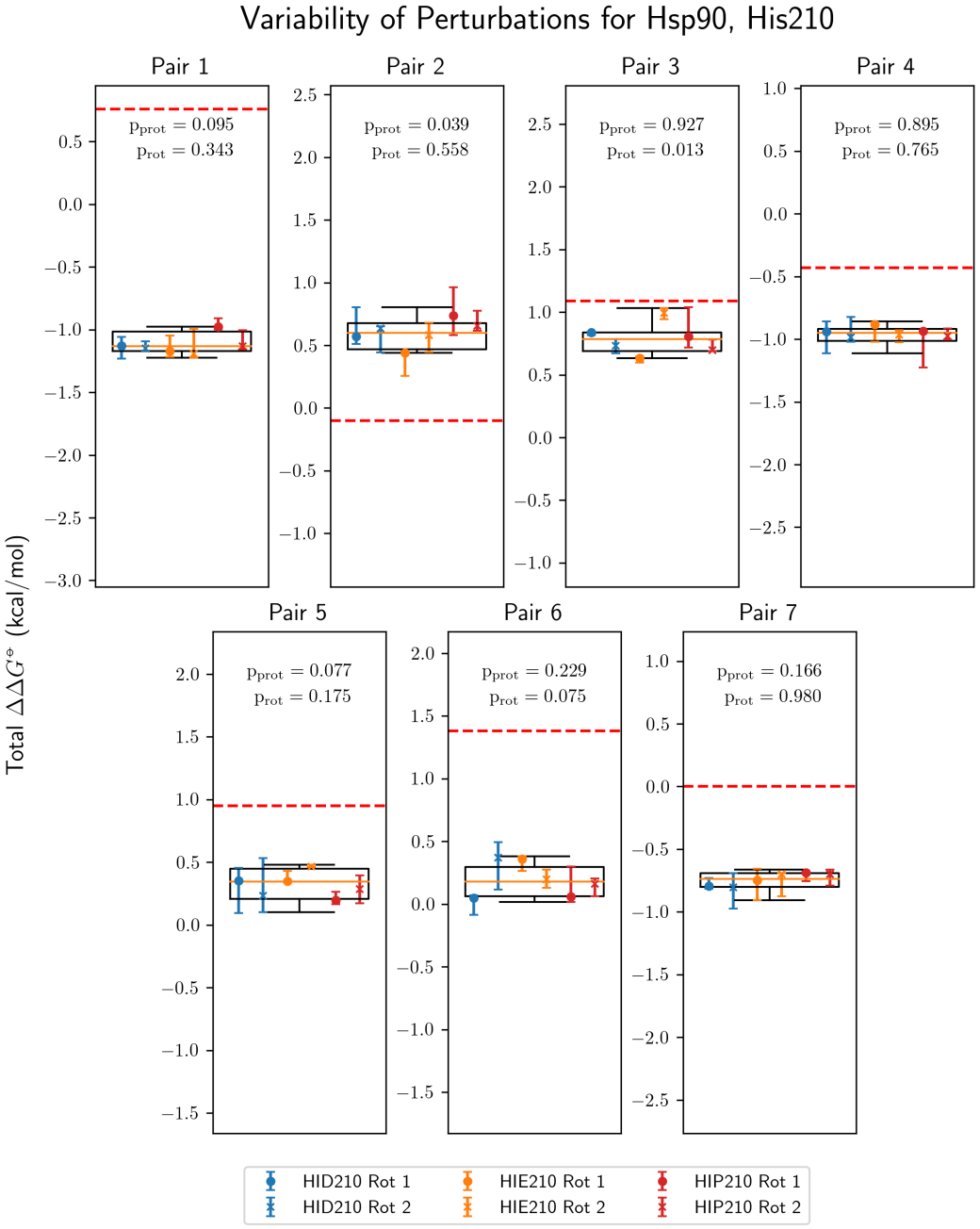


FIGURE 6.9: Box plots of the $\Delta\Delta G^{\ominus}$ values for different Hsp90 His210 PTR states at each ligand perturbation. Each point represents the median of three repeats, and the associated error bar extends between the other two repeats. The boxes contain all values from the merged dataset between the 25th and 75th percentile and the whiskers extend to the 5th and 95th percentile. The solid orange line represents the median of the whole dataset, while the dashed red line shows the experimentally obtained free energy value,[6] if available. The associated p-values obtained from a Kruskal–Wallis test between different protonation states have been reported as p_{prot}, while the Fisher averages of the three pairwise rotamer Kruskal–Wallis p-values have been annotated as p_{rot}.

without hidden states is more prone to underestimating the length of the implied timescales.[72] In addition, a Bayesian MSM is vital for outputting a population of MSMs which enable one to obtain the confidence interval (CI) of the estimated implied timescales.[73,74]

As described in Section 6.2.4, the implied timescales were calculated over a range of lag times τ in the range between 5–100 ps. Since the obtained implied timescales should be invariant with respect to a change in lag time, calculating them over a range of τ values enables us to ascertain the validity and the convergence of the MSM.

The results from the MSM analysis show well-converged estimates of the slowest implied timescale over a range of lag times (Figure 6.10). In the following discussion, one of the sufficiently converged lag times will be arbitrarily chosen to report the implied timescales. These implied timescales will enable us to assess the mobility of each histidine residue—the higher the implied timescale, the lower the mobility.

The least mobile trypsin histidine residue is His40 (Figure 6.10a), which has an implied timescale of 45 ± 16 ns. Since this timescale is an order of magnitude larger than the simulation length of each of the λ windows, this means that in this case it is practically certain that the choice of initial histidine residue will achieve maximum possible bias of the resulting free energies. The magnitude of this bias, however, is of course dependent on the nature of the perturbation and the distance between the ligand and the histidine residue. As shown in Table 6.2, this translates to a rotamer MAD of 0.15 kcal/mol over the range of all perturbations considered in this study.

The binding site His57 and the distal His91 show an order-of-magnitude higher mobility with implied timescales of 4.1 ± 0.4 ns and 3.4 ± 0.4 ns, respectively (Figure 6.10a). Since these timescales are comparable to the simulation time per λ window, it is again expected that the resulting free energies will be significantly affected, albeit to a lesser extent, by the choice of initial histidine conformation. However, Table 6.2 shows that His57 and His91 have significantly different rotamer MADs, with the more mobile His57 having larger discrepancies than the less mobile His40. Therefore, the main predictor of the magnitude of these discrepancies in the case of trypsin is likely the ligand–histidine distance, rather than the histidine mobility.

His154 is the least mobile histidine residue in Hsp90 with an implied timescale of 475 ± 167 ns (Figure 6.10b). While this estimate is highly unreliable and implied timescale estimation at many of the intermediate lag times was not possible due to undersampling, it is clear that this timescale is several orders of magnitude higher than the sampling per λ window, making His154 the least mobile histidine residue in this study. Although this is reflected by His154 having the highest rotamer MAD (Table 6.2), the statistical significance of the result is lower than that of His77, which is also further away from the binding site.

His210 and His189 are residues with limited mobility, similarly to trypsin His57 and His91, with implied timescales of 3.8 ± 0.4 ns and 3.4 ± 0.3 ns, respectively (Figure 6.10b). Despite these timescales being comparable to the simulation time per λ value, their rotamer MADs are only 0.02–0.03 kcal/mol lower than His154 (Table 6.2), showing that in this case mobility is again not a decisive factor in predicting the magnitude of the free energy discrepancies between different histidine rotamers.

His77 is the most mobile residue in the whole study with an implied timescale of 0.42 ± 0.03 ns, or an order of magnitude lower than the sampling time per λ window (Figure 6.10b). Despite this high mobility, it exhibits the most significant free energy discrepancies with a Kruskal–Wallis p-value of 0.043 (Table 6.2). Nevertheless, it has the lowest rotamer MAD in Hsp90, indicating that the increased mobility likely has some role in decreasing it. However, this result shows that even timescales which are an order of magnitude shorter than the sampling time per λ window are not sufficient to completely remove the initial rotamer bias, which is observed even at distances larger than 1.4 nm (Table 6.1).

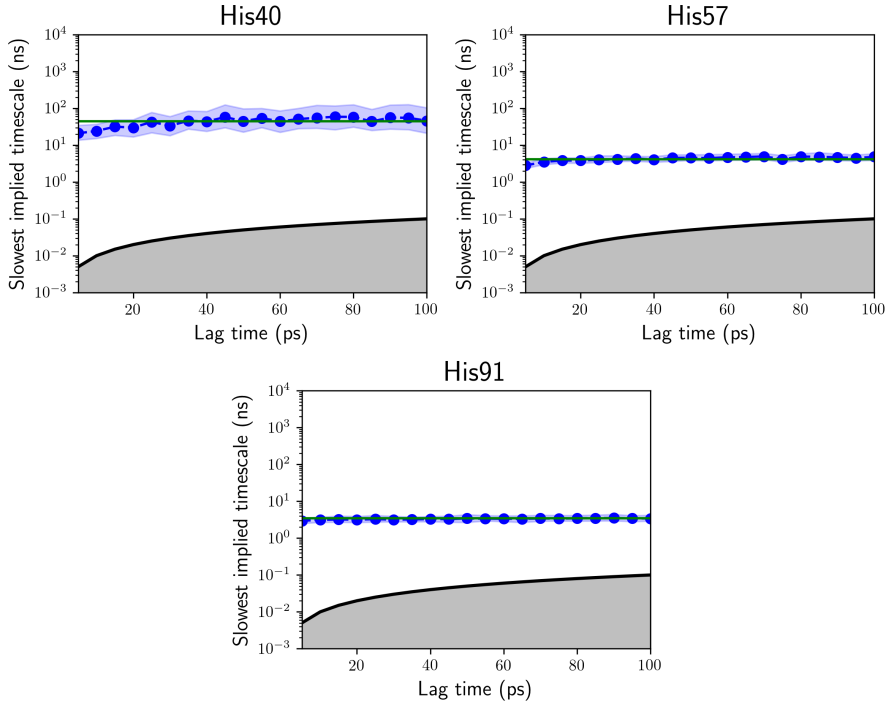
6.3.3 Total Variability per Perturbation

It is informative to explore the cumulative influence of all histidine residues on each of the alchemical perturbations. This will provide us with some knowledge on how sensitive different perturbations are to changes in the histidine PTR states.

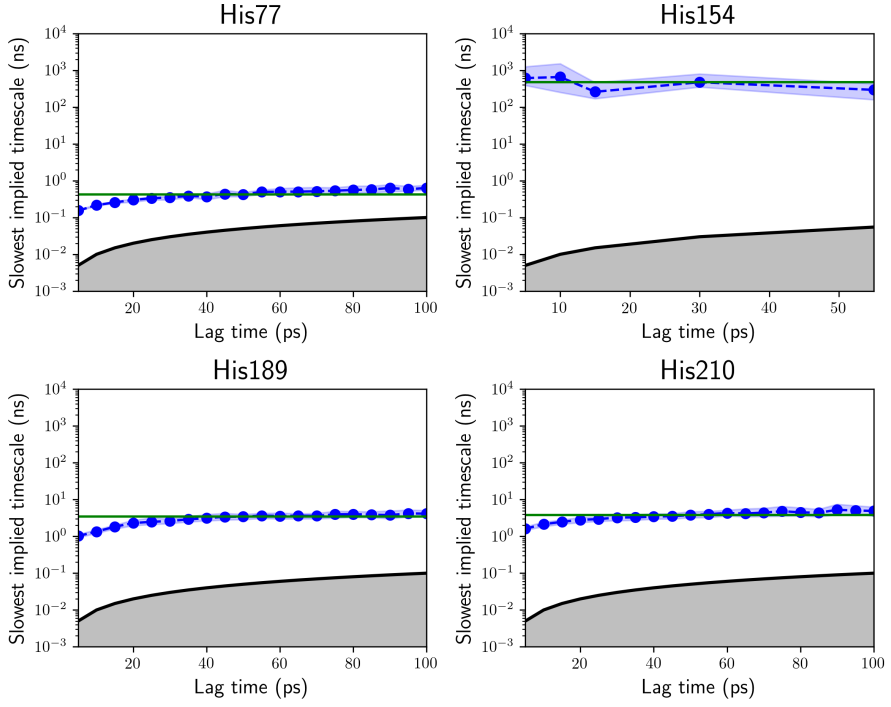
As shown in Figure 6.11a, all of the trypsin perturbations exhibit statistically significant differences between the free energies obtained at different histidine PTR states with Kruskal–Wallis p-values of at most 0.001. Therefore, most of the differences in the total ligand variances can be attributed to ligand–histidine interactions.

The trypsin perturbation with highest variability is pair 6, with a 5–95 percentile difference of $\sim$1.5 kcal/mol (Figure 6.11a). Pair 7 is a close second with a $\sim$1.3 kcal/mol uncertainty at the 90% CI. All other perturbations have 5–95 percentile ranges between $\sim$0.5 and $\sim$1.0 kcal/mol, which gives a measure of the PTR reproducibility limit for trypsin. Since pairs 6 and 7 are the only perturbations involving the alchemical change of a nitro group, it can be proposed that the partial charges of the latter interact particularly strongly with the histidines, making them more sensitive to PTR variations. Even though the rest of the perturbations cover a range of partial charge changes, with pair 5 involving a charge change of $\sim$1 e, they all have lower sensitivity to the histidine residues and no clear pattern can be inferred in these cases.

On the other hand, the Hsp90 perturbations (Figure 6.11b) exhibit much less pronounced differences between the free energies obtained at different histidine PTR states, with only pair 3 having a p-value less than 0.05. Since pair 2 and pair 3 have the



(A) Trypsin



(B) Hsp90

FIGURE 6.10: Rotation kinetics of the three trypsin histidine residues (Figure 6.10a) and the four Hsp90 histidine residues (Figure 6.10b) estimated with a hidden state Bayesian MSM after clustering with GMMs. The y axis represents the slowest implied timescale at a range of lag times. The blue line represents the mean over 100 different MSMs, while the shaded blue area represents a 95% confidence interval. The shaded grey area indicates the precision limit of the MSM given the lag time used for estimation. The green line represents the implied timescale of one of the lag times. The latter was manually chosen as a sufficiently converged representative of the series of lag times.

most significant discrepancies and they are also the only perturbations involving the perturbation of a carbonyl oxygen to an oxime, it is reasonable to deduce that the charge distribution change induced by this perturbation interacts significantly with the histidine residues. However, these effects are still much weaker compared to the trypsin results.

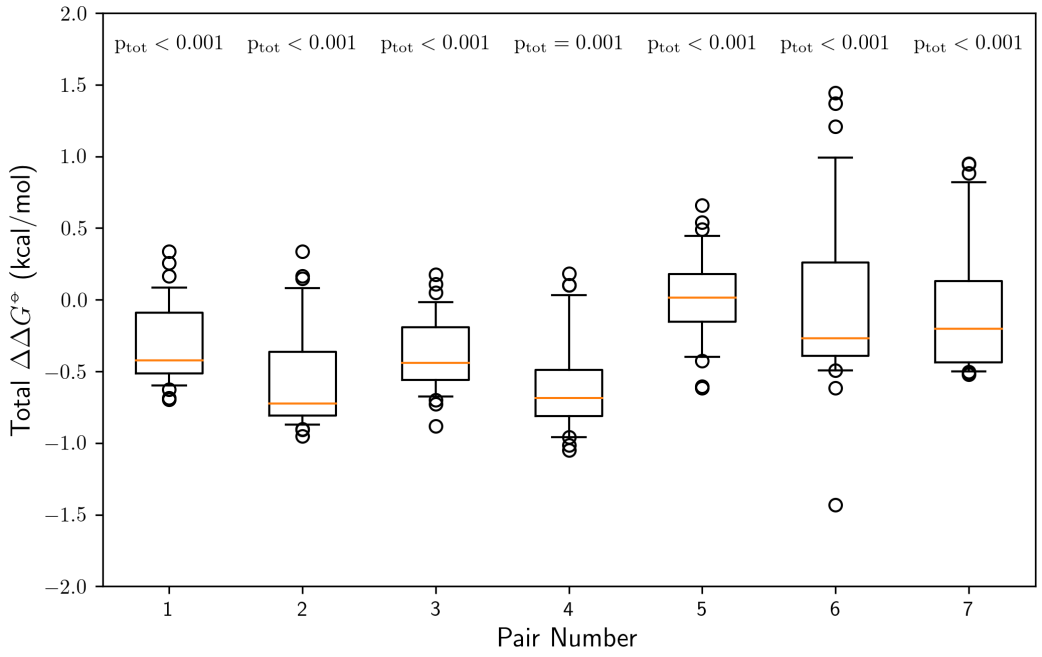
The total variabilities of each of the Hsp90 perturbations are also markedly lower than for trypsin, with 5–95 percentile ranges between $\sim$0.25 and $\sim$0.5 kcal/mol. Most perturbations have comparable variability, with the lowest uncertainty exhibited by pairs 1, 4, 7. Since these are also the pairs involving a hydrogen to a fluorine perturbation, it can be proposed that the reason for this decrease in variability is the greater simplicity of the perturbation compared to the other pairs.

6.3.4 Free Energy Discrepancies as a Function of Ligand–Histidine Distance

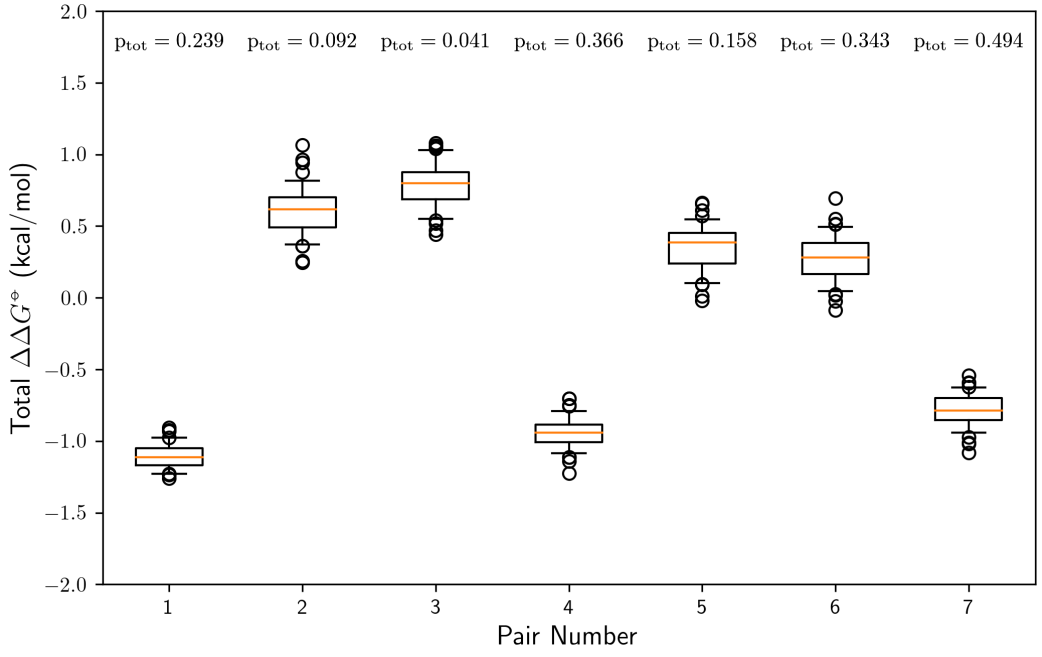
It is intuitively expected that the larger the distance between the ligand and the histidine residue, the smaller the effect of the latter on the estimated free energy values will be. On the other hand, it is also expected that perturbations involving large partial charge changes will have stronger long-range interactions and will be more sensitive to the histidine protonation state. In this section we evaluate the validity of these notions.

To investigate the influence of the ligand–histidine distance on the observed free energies, we will define it as the shortest distance between any of the histidine atoms and any of the atoms of the relevant perturbed ligand group. As shown in Table 6.1, we will consider three different perturbation groups in Hsp90, since there are three different ligand sites that are perturbed. The absolute deviations will be calculated as the absolute differences between all different combinations of samples from the corresponding groups. We will also distinguish between four types of MADs: the PT, rotamer, total and inter-replicate MAD. The latter will be used as a reference deviation limit to which we will compare the other three MADs. The raw calculated PT, rotamer and total MADs will be referred to as “absolute”, while the same with the inter-replicate MAD subtracted from them will be denoted as “relative”.

As shown in Figure 6.12a, the MADs corresponding to the trypsin binding site His57 are unsurprisingly significantly higher than the rest of the histidine residues with the PT MAD being 0.48 kcal/mol, or 0.10 kcal/mol higher than the rotamer MAD. These MADs are significantly different to the inter-replicate MAD and the relative total MAD amounts to an extra 0.38 kcal/mol of variability added by the binding site histidine (Figure 6.12b).



(A) Trypsin



(B) Hsp90

FIGURE 6.11: Total free energy variability for each of the trypsin (Figure 6.11a) and Hsp90 (Figure 6.11b) perturbations over all histidine PTR states. The orange lines represent the median, the boxes include the interquartile range, while the whiskers extend to the 5^{th} and 95^{th} percentiles. Outliers beyond the whisker range are shown as empty circles.

At higher ligand–histidine distances, the absolute free energy discrepancies decrease considerably and stay relatively constant at ∼0.20 kcal/mol, with the trypsin His40 residue reaching 0.26 kcal/mol. Except for trypsin His40 and His57, the rotamer MAD is comparable to the PT MAD in all cases. Surprisingly, the Hsp90 His154 perturbations in pairs 2, 3, 5 and 6 show lower MADs than trypsin His40, despite the shorter ligand–histidine distance.

In Figure 6.12a one can distinguish between two "bands" of MADs at distances over 1.0 nm. The lower MADs correspond to interactions between the Hsp90 histidines with pairs 1, 4 and 7, while the other data points represent all other interactions. However, these differences are largely smoothed when the inter-replicate MAD is subtracted (Figure 6.12b), indicating that the difference in behaviour is related to the lower inter-replicate MAD for pairs 1, 4, 7, as discussed in the previous section.

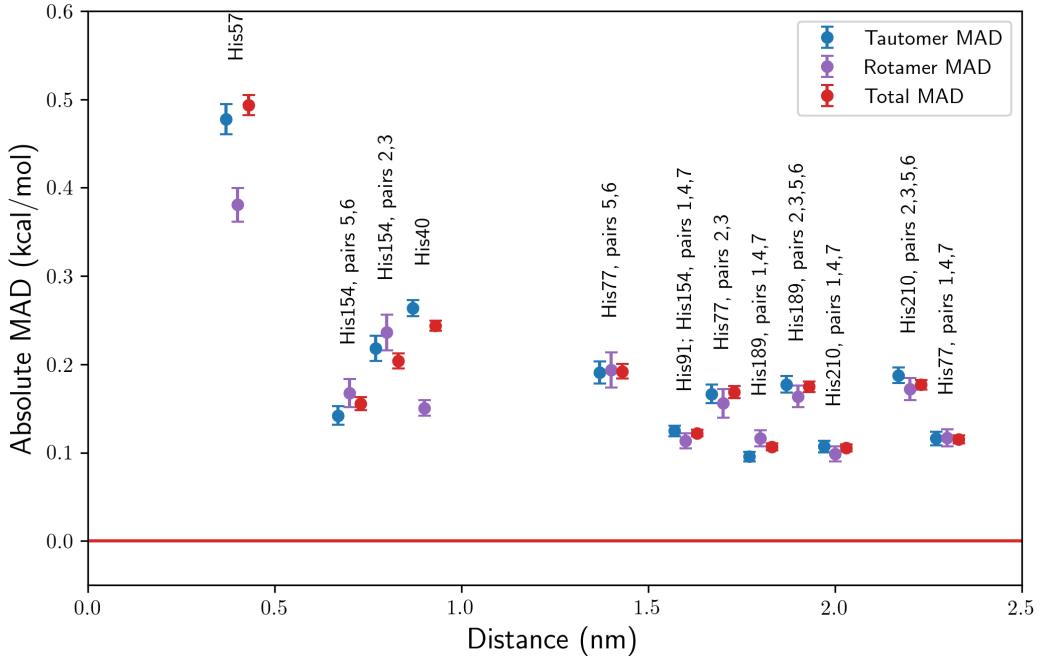
Compared to the inter-replicate MADs, the variability introduced by the histidines at distances over 1.0 nm is in many cases insignificant with most relative MADs being close to zero (Figure 6.12b). The only exceptions are observed at 1.7 nm (His77, pairs 2 and 3) and 2.2 nm (His210, pairs 2, 3, 5 and 6), with MADs of ∼0.05 kcal/mol. Since pairs 2 and 3 involve the carbonyl to oxime perturbations, it can be seen that the partial charge changes associated with these perturbations can result in observable long-range histidine effects on the free energies. Nevertheless, the magnitude of these effects is still practically negligible at such long distances.

6.4 Discussion

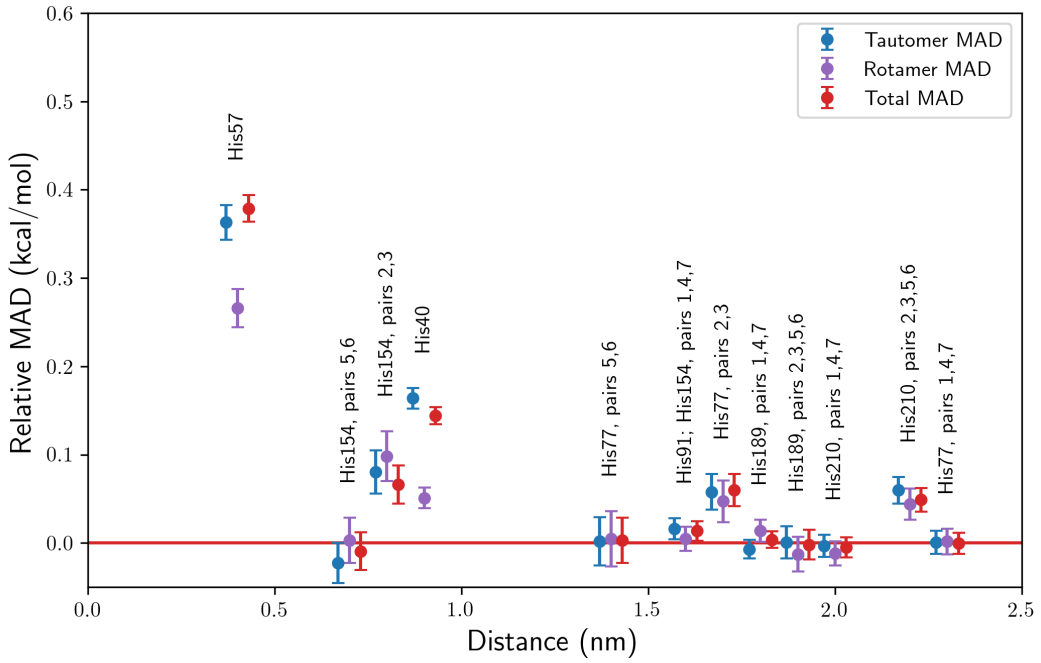
The above results show two very different types of behaviour. On one hand, trypsin produces well-converged results which are significantly affected by the histidine PTR states even at long ligand–histidine distances. On the other hand, Hsp90 results in more consistent free energies with most free energy discrepancies being insignificant. It follows that one could regard these two systems as two possible extremes of free energy sensitivity towards the initial PTR states.

It was shown that the PTR states of an active site histidine (trypsin His57) can significantly affect the observed free energy values, with most discrepancies being in the range of ∼0.5 to ∼1.0 kcal/mol. However, perturbations involving nitro groups can increase these discrepancies to ∼1.5 kcal/mol. This can therefore be regarded as a "worst" case reproducibility limit.

In practice, however, it is common knowledge to expect such sensitivity and binding site protonation states are often carefully chosen. In this case, a more realistic reproducibility limit is ∼0.5 kcal/mol with a worst case scenario of ∼1.0 kcal/mol for a ligand containing a nitro group interacting with a nearby immobile histidine, as



(A) Absolute MAD



(B) Relative MAD

FIGURE 6.12: The MADs of the free energy values as a function of distance (Figure 6.12a). The three types of MADs have been slightly separated in the x axis for visualisation purposes. In Figure 6.12b, the inter-replicate MADs were subtracted from the corresponding data points. The error bars represent the standard error of the mean (Figure 6.12a) or the mean difference (Figure 6.12b). Values below the red line at $x = 0$ can be regarded as statistical noise.

observed in trypsin His40. These reproducibility issues persist at distances as long as 1.6 nm and can in these cases reach discrepancies of ∼0.1 kcal/mol.

All of these differences are naturally relative to the observed inter-replicate variance, which is in turn highly system- and protocol-dependent. As observed in Hsp90, the discrepancies between different repeats can often overshadow the ones caused by alternative histidine states. In addition, this inter-replicate variance can also be affected by more mobile histidine residues. Of course, it is expected that other comparatively slower events can also affect this variability, meaning that this consideration is not exclusive to the rotation of histidine residues, but can include other degrees of freedom as well.

The above results suggest that there are four factors that can have an impact on the magnitude of the free energy discrepancies: the magnitude of the charge changes, the ligand–histidine distance, the histidine mobility, and the sampling variance measured by the different simulation repeats. Of these, it can be argued that the latter is the most decisive factor, since it sets a repeatability limit for the system which may or may not be higher than the expected variance induced by the different histidine states. It is also the most difficult factor to predict, since inter-replicate variance is system-dependent and can also be affected by all of the other factors.

The second most important factor is arguably the ligand–histidine distance, since it was shown that the magnitude of the free energy discrepancies rapidly decreases with increasing distance and these discrepancies are mostly negligible after $\sim$ 1.0 nm (Figure 6.12). However, at short distances the initial histidine state can significantly affect the observed free energy changes regardless of the type of perturbation.

Nevertheless, perturbations involving groups with high partial charges can exhibit long-range sensitivity towards the initial histidine PTR states, even if the discrepancies are not necessarily numerically significant, making the nature of the ligand–ligand perturbation the third most important factor. Although perturbations involving large partial charge changes have been shown to have comparatively higher sensitivity towards histidine PTR states even at short distances (nitro groups in trypsin and oximes in Hsp90), the charge differences are not always predictive of the magnitude of the variance. For example, the perturbation from an amino group to a hydrogen atom exhibited the lowest variance of all trypsin perturbations involving a variation in the active site histidine, including the perturbation of a methyl group to a hydrogen atom, despite the former involving a highest partial charge change of ∼1 e, and the latter only involving charge changes up to ∼0.25 e.

Finally, the histidine mobility was the least important factor in this study. One possible reason for this observation is that only one histidine residue had a slowest implied timescale significantly lower than the simulation time per λ window (Hsp90 His77), meaning that the other histidine residues could not reliably decrease any free

energy discrepancies arising from different initial rotamers. However, even the most mobile histidine residue showed observable free energy discrepancies in some cases. This can be explained by the large number of λ windows (40), since a complete removal of the initial rotamer bias can only be achieved if the histidine rotation is consistently sampled in each of the intermediate λ windows.

It is expected that the above effects will amplify with a higher number of titratable residues and/or possible residue rotamers, meaning that the initial system setup is a decisive element in the free energy calculation. Since system preparation can be highly sensitive to the initial crystal structure (Chapter 5), this poses a significant problem for reproducible free energy calculations. Therefore, the above results strongly suggest that enhanced sampling methods are an indispensable element of a robust free energy workflow.

Since different PT states correspond to different Hamiltonians, one needs an expanded ensemble enhanced sampling method to explore the Hamiltonian space of the different protonation states. As explicitly handling every possible combination of amino acid PT states is not feasible, stochastic approaches, such as constant-pH[121] methods will be valuable in removing any implicit free energy biases posed by an incorrect protonation state. These have been repeatedly shown[118,121,123,124,127] to provide significant improvement in sampling and the results presented above further showcase the need for their mainstream use in free energy applications.

The enhanced sampling of amino acid rotamers is also important, as in the above cases two of the seven histidines had unfeasibly long implied timescales of rotation, while four required several nanoseconds on average for mixing. On the other hand, only one of the seven histidines exhibited mobility on timescales shorter than 1 ns. This highlights one significant weakness of alchemical free energy calculations, as traditionally performed: the relatively short duration of the separate λ windows is not sufficient to reliably explore all of these motions in all windows, resulting in significant bias even after 4 ns of sampling per window. One way to mitigate this problem in a general way is by employing single-trajectory methods, such as simulated tempering (ST),[240] λ-dynamics (λD),[241] enveloping distribution sampling (EDS)[242] or integrated tempering sampling (ITS),[243] which will reliably sample many of these motions over long timescales without any extra input. Alternatively, if there are known residue side chains whose rotations are known to be important but happen on very long time scales, targeted enhanced sampling methods can help in exploring these. One such method is nonequilibrium candidate Monte Carlo (NCMC),[64] which has been shown to significantly improve side-chain kinetics of sterically hindered residues.[244]

6.5 Conclusion

The influence of histidine PTR states on the calculated relative binding free energy values was evaluated over a range of alchemical perturbations on two protein systems: trypsin and Hsp90. The trypsin complexes exhibited a higher sensitivity to the initial histidine states, with the binding site histidine residue inducing free energy discrepancies of up to $\sim$1.5 kcal/mol for a perturbation involving a nitro group. These discrepancies can be observed at distances as long as 1.6 nm, albeit to a much less significant extent ($\sim$0.1 kcal/mol).

Hsp90, on the other hand, showed little sensitivity to the alternative histidine states, compared to trypsin. This was likely caused by a combination of different factors, such as the increased sampling variance, the large ligand–histidine distances and the smaller partial charge changes in most of the perturbations. However, even in these cases discrepancies of up to $\sim$0.5 kcal/mol can be observed, indicating that both histidine protonation, tautomeric and rotameric states can have significant impact on the observed free energies. These are expected to amplify with the number of side chains with uncertain PTR states.

We have also discussed the necessity for the mainstream adoption of enhanced sampling methods to improve the robustness of alchemical binding free energy calculations. It is expected that constant-pH methods can ameliorate the above issues, while single-trajectory sampling methods, such as ST, λD, EDS and ITS, can help explore rotamers over long timescales without any system-specific user input. In the case where a small subset of the side chains exhibits slow kinetics, while participating in important interactions with the ligand, more targeted enhanced sampling methods can be useful, such as NCMC.

Chapter 7

Parameter-Based Enhanced Sampling Methods: A Review

7.1 Introduction

The results in Chapters 5 and 6 show that rare event exploration is crucial for performing reliable alchemical free energy (AFE) calculations. Since regular sampling methods, such as molecular dynamics (MD), are insufficient for exploring such slow motions at short timescales, enhanced sampling methods are needed.

Central to the idea of each enhanced sampling method is the selection of a nonlinear function of the system coordinates $\vec{s}(\vec{x})$, which has a much lower dimensionality (typically 1–3 dimensions) compared to the $3N$ degrees of freedom for a system containing N atoms. This function is said to define several "collective variables" (CVs), which represent the slow degrees of freedom that are of central interest.[245]

The choice of a CV is specific to the purpose of the simulation and to the studied system.[246] Sometimes this choice is relatively straightforward (e.g. passing of a small molecule through a cell membrane[247]), other times it is highly non-trivial and system-specific (e.g. protein folding[245,246]). In the latter cases, the prospective application of enhanced sampling methods on new systems is greatly reduced due to the significant prior knowledge needed to determine the relevant CVs. Therefore, general non-specific CVs are highly desirable when such knowledge is not available.

The most natural way to define a general CV is to recognise that at higher temperatures, all kinetic barriers are universally smoothed, meaning that rare events are more likely to happen. Therefore, if we extend the notion of a CV to coordinate-independent parameters as well, we could regard temperature as the most general CV. This realisation leads us to a class of enhanced sampling methods, which will be referred to as "tempering" methods. Even more generally, one can introduce

an arbitrary number of parameters $\vec{\lambda}$ into the dimensionless potential energy function $u(\vec{x})$, such that some values of these parameters correspond to a subset of the kinetic barriers being smoothed, while others recover the original parameter-free potential energy function of interest.

In this chapter, enhanced sampling methods, which require coordinate-based CVs ("restraint-based" methods), and those utilising parameter-based CVs (tempering methods), will be distinguished. This distinction stems from the fact that since the parameters are introduced purely out of convenience, they could be defined in any way, as long as one set of parameters recovers the original distribution. This allows us to explore discrete sequences of parameter values, in contrast to restraint-based methods, which are restricted to continuous spaces. This ability of tempering methods to explore discretised CV spaces is directly relevant to alchemical free energy calculations, which use a set of discrete values of an alchemical variable λ to define the transformation between two ligands and calculate the free energy difference between them. This means that tempering methods can be used to address the sampling problem and the free energy calculation problem in the same manner.

Some of the more well-known restraint-based methods are: multicanonical sampling,[248] Wang–Landau sampling,[249] metadynamics,[97,250,251] umbrella sampling,[252] accelerated MD,[253,254] adaptive biasing force,[255] targeted MD,[256] steered MD,[257] conformational flooding[258] and blue moon sampling.[259] Although many of these methods have significant similarities to and/or could be used in conjunction with tempering methods, they will not be discussed in this chapter, which focuses on parameter-based CVs, and a detailed review of these methods can be found elsewhere.[260]

The aim of this chapter is to present a wide range of tempering enhanced sampling methods and compare their similarities and differences, as well as their advantages and disadvantages. This will allow us to rationalise the choices made in the following method development chapters and highlight the relevance of the methodologies presented there.

7.2 Tempering Methods

The two most important distributions defined in a tempering method are the distribution of interest and the distribution with maximally smoothed kinetic barriers. The former provides us with the information that we seek from the computational experiment, while the latter increases the quality of sampling. The introduction of samples from one distribution into another can be visualised through the framework of importance sampling, where any expectation value of an observable $\langle O \rangle$ over a

particular distribution $p(\vec{x})$ can be related to any other distribution $q(\vec{x})$ in the following manner:

$$\langle O \rangle = \int O(\vec{x}) p(\vec{x}) d\vec{x} = \int O(\vec{x}) \frac{p(\vec{x})}{q(\vec{x})} q(\vec{x}) d\vec{x} \tag{7.1}$$

If we regard $\frac{p(\vec{x})}{q(\vec{x})}$ as effective weights, this essentially means that we can generate samples from any arbitrary $q(\vec{x})$ and we can determine the expectation of any observable over any other distribution $p(\vec{x})$, as long as we reweight the samples by a factor proportional to $\frac{p(\vec{x})}{q(\vec{x})}$.

The main issue with importance sampling is its quickly diminishing efficiency as $p(\vec{x})$ and $q(\vec{x})$ diverge in terms of their probability densities. In such cases, the variance of the weights becomes very high, meaning that only a small subset of the samples is used in the final estimate. This in turn results in high uncertainty of the estimated average.

There are three possible ways to circumvent this problem. The first involves defining a reversible transformation $\vec{T}(\vec{x})$ which maps each sample generated by $q(\vec{x})$ onto a similarly favourable sample from $p(\vec{x})$. Unfortunately, this method can only be applied to very specific or trivial cases and is not generally possible in most practical situations. The second approach involves smoothing both probability distributions using another parameter, which results in sufficient overlap. The final approach is the introduction of intermediate distributions, such that each neighbour pair achieves sufficient overlap, enabling the use of sequential importance sampling. This is the approach used by most of the following methods.

7.2.1 Replica Exchange Molecular Dynamics (REMD)

The most widely used tempering method is parallel tempering,[261] more widely known in the field of computational chemistry as replica exchange molecular dynamics (REMD). In REMD, the following expanded probability distribution $\pi_{mix}(\vec{x})$ is explored:

$$\pi_{mix}(\vec{x}) = \prod_{i=1}^{N} \pi(\vec{\lambda}_i, \vec{x}) \tag{7.2}$$

where $\vec{x}$ is the set of coordinates and $\vec{\lambda}$ is an arbitrarily large number of parameters with N different combinations of values. In addition, it will be here and henceforth assumed that $\pi(\vec{\lambda}_1, \vec{x})$ is the distribution providing the most sampling, while $\pi(\vec{\lambda}_N, \vec{x})$ is the distribution of interest. In practice, this is done by collecting samples from each

of the distributions in parallel, resulting in N concurrent simulations. Each simulation samples from its corresponding distribution using MD.

Sample enrichment is achieved by periodically attempting to swap a number of replica pairs. This can be done in many ways,[262] one of the most widely used of which is attempting to swap all even replicas with all odd replicas in a reversible manner obeying detailed balance. In this case, the acceptance criterion $p_{acc}(\vec{x}_j, \vec{x}_i | \vec{x}_i, \vec{x}_j)$ for swapping two sets of coordinates $\vec{x}_i$ and $\vec{x}_j$ sampled at $\vec{\lambda}_i$ and $\vec{\lambda}_j$, respectively, is:

$$p_{acc}(\vec{x}_j, \vec{x}_i | \vec{x}_i, \vec{x}_j) = \min \left[1, \frac{\pi(\vec{\lambda}_i, \vec{x}_j)\pi(\vec{\lambda}_j, \vec{x}_i)}{\pi(\vec{\lambda}_i, \vec{x}_i)\pi(\vec{\lambda}_j, \vec{x}_j)} \right] \tag{7.3}$$

In contrast to many of the methods discussed later, this acceptance criterion does not require any knowledge of normalisation constants, since they cancel out. Therefore, one only needs to calculate the relative probabilities, given by the instantaneous energies, and no further estimations are needed. This ease of use has contributed to the widespread adoption of this method.

The most general type of REMD is Hamiltonian replica exchange (H-REMD),[263,264] where any set of parameters $\vec{\lambda}$ can be used to define the intermediate distributions. More specific variations of H-REMD involve the original temperature-based method,[265] methods involving the effective local temperature of a subset of the system: replica exchange with solute tempering (REST)[86] and replica exchange with solute scaling (REST2),[94] alchemical transformations,[266] protonation states,[267] and water networks.[268] Various combinations of these methods exist, including concurrent alchemical and tempering transformations[95,269] and simultaneous alchemical and water network exploration.[111] Alchemical free energy (AFE) methods in particular benefit greatly from REMD, since free energy calculations already involve simulations at multiple intermediate states and an additional exchange procedure incurs a negligible computational overhead in return for a significant sampling enhancement.

7.2.2 Simulated Tempering (ST)

Simulated tempering (ST)[240,270] is an expanded ensemble method, which can be regarded as the single-replica version of REMD. Here, the mixture distribution $\pi_{mix}(\vec{x})$ is a weighted sum of the underlying probability distributions, rather than a product::

$$\pi_{mix}(\vec{x}) = \sum_{i=1}^{N} w_i \pi(\vec{\lambda}_i, \vec{x}) \tag{7.4}$$

where w_i is the corresponding weight of the i-th distribution. While the assignment of these weights is arbitrary, they are commonly set to unity, so that all of the distributions are sampled with equal probability.

The single replica then traverses this set of $\vec{\lambda}$ states. Exploration in parameter space is commonly done in a Gibbs sampling fashion,[262] where a change is attempted after a fixed amount of MD steps. As in REMD, multiple ways of choosing the proposal probabilities $p_{prop}(\vec{\lambda}_j|\vec{\lambda}_i)$ are possible[262] but a common way of defining the Markov chain is by only attempting transitions between nearest neighbours, with equal proposal probabilities:

$$p_{prop}(\vec{\lambda}_j|\vec{\lambda}_i) = \begin{cases} \frac{1}{2}\delta_{|i-j|,1} & \vec{\lambda}_i \notin \{1, N\} \\ \delta_{|i-j|,1} & \vec{\lambda}_i \in \{1, N\} \end{cases} \tag{7.5}$$

with δ being the Kronecker delta. The acceptance criterion $p_{acc}(\vec{\lambda}_j|\vec{\lambda}_i, \vec{x})$ is then related to the importance sampling weight of the configuration $\vec{x}$, and is commonly chosen to satisfy detailed balance:

$$p_{acc}(\vec{\lambda}_j|\vec{\lambda}_i, \vec{x}) = \min\left[1, \frac{w_j \pi(\vec{\lambda}_j, \vec{x}) p_{prop}(\vec{\lambda}_i|\vec{\lambda}_j)}{w_i \pi(\vec{\lambda}_i, \vec{x}) p_{prop}(\vec{\lambda}_j|\vec{\lambda}_i)}\right] \tag{7.6}$$

Unlike the acceptance criterion used in REMD, it can be seen that the normalisation constant ratios, or equivalently, the free energy differences between the two distributions need to be known. This apparent setback of the method has contributed to its underutilisation compared to REMD.[271] Since such knowledge is rarely available in advance, commonly used approaches obtain these free energies either by approximate heuristics,[271,272] maximum likelihood free energy estimators,[273,274] or on-the-fly learning.[273,275]

While the original ST algorithm uses the total system temperature as a parameter,[240,270] the method has been further generalised to an arbitrary Hamiltonian parameter dependence similarly to H-REMD.[276] ST has been used in a variety of applications, including studying phase transitions,[277] exploring molecular conformations[271,274,278] and performing solvation free energy calculations.[274,279]

7.2.3 Integrated Tempering Sampling (ITS)

Integrated tempering sampling (ITS)[243] is another expanded ensemble method which is conceptually very similar to ST. One of the main differences is that in ITS the sampled distribution $\pi_{mix}(\vec{x})$ is not a sum of discrete distributions, but is instead an integral over a continuum of distribution parameters:

$$\pi_{mix}(\vec{x}) = \int w(\vec{\lambda})\pi(\vec{\lambda}, \vec{x})d\vec{\lambda} \tag{7.7}$$

This makes ITS a generalisation of ST.[280] Nevertheless, to calculate the integral in Equation 7.7, the mixture distribution is still discretised in practice (Equation 7.4),[243,281,282] although the number of intermediates is usually higher than those typically used in ST simulations. The main difference with ST then lies in the way this distribution is explored. While in ST this is done by a Monte Carlo (MC) walk in $\vec{\lambda}$ space, ITS treats $\pi_{mix}(\vec{x})$ as a distribution with an effective potential energy $U_{eff}(\vec{\lambda}, \vec{x})$, such that:

$$U_{eff}(\vec{x}) = -\frac{1}{\beta} \ln \pi_{mix}(\vec{x}) \tag{7.8}$$

This potential energy function can then be straightforwardly used in the classical equations of motion (Equation 2.28). In this setting, all of the underlying distributions contribute their associated forces to the effective force $\vec{F}_{eff}(\vec{x})$ at each timestep.

As with ST, knowledge of the relative free energies of the underlying distributions is critical in ensuring satisfactory performance of the method. These can be obtained with estimation approaches similar to those used in ST.[243,281,282]

Similarly to REMD and ST, there are a few published variations of ITS depending on the way the parameters are used to define the intermediate states. The most general type approach is Hamiltonian ITS,[282] where an arbitrary set of parameters $\vec{\lambda}$ can be used to define the interpolation between the two final distributions. The more specific versions are the original tempering method[243] and its extension where only parts of the system are tempered[281] (selective ITS).

7.2.4 λ-Dynamics (λD)

λ-Dynamics,[241] introduced 12 years before ITS, is another method which explores a mixture distribution $\pi_{mix}(\vec{x})$ over a continuous set of states (Equation 7.7). The main conceptual difference with ITS is the exploration of $\vec{\lambda}$ space: instead of extracting an effective potential energy function from the mixture distribution, $\vec{\lambda}$ is treated as a set of fictitious particles with their own masses, which can then be evolved using classical equations of motion (Equation 2.28). However, the continuity in parameter space means that there is a vanishing probability of $\vec{\lambda}$ being in any single state. Therefore, the resulting distribution of $\vec{\lambda}$ values needs to be approximately separated into discrete bins before analysis and/or free energy estimation.

Similarly to ST and ITS, uniform exploration in $\vec{\lambda}$ space needs to be facilitated by adaptive weight estimation. In the context of λD, this has been achieved by defining a

family of parameter-dependent restraint-like continuous functions in $\vec{\lambda}$ space and subsequently optimising them.[283,284]

Different variations of λD have been published, some of them similar to other tempering methods. While the original version of the method samples $\vec{\lambda}$ space using classical equations of motion,[241] some of more recent publications use a Gibbs sampler,[285,286] with the latter publication sampling discrete $\vec{\lambda}$ space in a comparable manner to ST. λD has also been generalised to multiple $\vec{\lambda}$ variables,[287] combined with REMD,[283,288] and applied in large-scale free energy studies.[289]

7.2.5 Enveloping Distribution Sampling (EDS)

Enveloping distribution sampling (EDS)[242] is another expanded ensemble method, originally introduced to sample the two endstates of a relative alchemical free energy calculation, although it has subsequently been applied in the context of constant-pH calculations as well.[127] In this setting, the same mixture probability distribution as ST is sampled (Equation 7.4), except that in the original formulation of EDS no intermediates between the two endpoint distributions were used. This two-state reference mixture distribution has previously been used to estimate free energy values.[13,290]

Instead of using intermediate distributions to increase phase space overlap, as used in the first publication of the method,[242] more recent publications have introduced a temperature-like smoothing parameter s to improve the overlap between the end states. One can then surmount the kinetic barriers in the mixture distribution either by choosing a single value of s which provides high smoothing[291,292] or by sampling the s coordinate using REMD.[126,293] In the former case, the simulations need to be appropriately reweighted, since the $\pi_{mix}(\vec{x})$ no longer corresponds to a sum of the physical distributions of interest. The latter approach, on the other hand, is functionally equivalent to using intermediate alchemical states, since multiple values of s are used to achieve sufficient overlap.

As with other expanded ensemble methods exploring sums of distributions, the relative free energy values of the distributions need to be estimated to ensure good performance of the method. Several approaches estimating these, as well as the smoothing parameter schedule, have been published, including the utilisation of an orthogonal REMD protocol.[291–293]

Similarly to the ITS method, first introduced one year after EDS, exploration in expanded ensemble space is not performed using MC but instead by defining an effective potential $U_{eff}(\vec{x})$, as described in Equation 7.8. This makes EDS and discrete-space ITS conceptually equivalent, where the main difference is the way intermediate states are introduced.

7.2.6 Nonequilibrium Candidate Monte Carlo (NCMC)

Nonequilibrium candidate Monte Carlo (NCMC)[64] is a sequential importance sampling method for proposing coordinate transformations $p_{prop}(\vec{x}'|\vec{x})$ based on a nonequilibrium MD kernel with time-dependent transformations in parameter space. The underlying mixture distribution explored by NCMC is the same as for λ-dynamics in the continuous case (Equation 7.7), or ST in the discrete case (Equation 7.4), the latter of which is used in computational implementations, and we are going to focus on it exclusively. However, in contrast to the previous methods, where changes in parameter space are attempted successively, in NCMC one or more predetermined sequences of $\vec{\lambda}$ states with intermittent coordinate decorrelation MD kernels are instead sampled and accepted based on their relative probability of occurring. Therefore, NCMC does not preserve the balance condition, since it does not explore the whole ensemble of possible paths in the Markov chain of connected states, but rather only a particular set of paths in the Markov chain.

The acceptance probability of a particular discrete trajectory consisting of a sequence of M parameters $\mathbf{\Lambda} \equiv (\vec{\lambda}_{i_1}, ..., \vec{\lambda}_{i_M})$ with corresponding coordinates $\mathbf{X} \equiv (\vec{x}_{i_1}, ..., \vec{x}_{i_{M-1}})$ is then chosen to satisfy detailed balance:

$$p_{acc}(\mathbf{\Lambda}, \mathbf{X}) = \min\left[1, \frac{w_{i_M}}{w_{i_1}} \frac{p_{prop}(\tilde{\mathbf{\Lambda}}|\vec{\lambda}_{i_M})}{p_{prop}(\mathbf{\Lambda}|\vec{\lambda}_{i_1})} \frac{\pi(\vec{\lambda}_{i_2}, \vec{x}_{i_1})}{\pi(\vec{\lambda}_{i_1}, \vec{x}_{i_1})} ... \frac{\pi(\vec{\lambda}_{i_M}, \vec{x}_{i_{M-1}})}{\pi(\vec{\lambda}_{i_{M-1}}, \vec{x}_{i_{M-1}})}\right] \tag{7.9}$$

with $\tilde{\mathbf{\Lambda}} \equiv (\vec{\lambda}_{i_M}, ..., \vec{\lambda}_{i_1})$ being the reversed parameter trajectory. It has been assumed in Equation 7.9 that the decorrelated coordinates are proposed by a sequence of distribution-preserving MD kernels, and the momenta are either flipped or resampled from the stationary distribution after acceptance to preserve detailed balance. With this acceptance criterion, only the stationary probability distributions at the protocol endpoints $\pi(\vec{\lambda}_{i_1}, \vec{x})$ and $\pi(\vec{\lambda}_{i_M}, \vec{x})$ are preserved, as opposed to the intermediate distributions which are not sampled correctly (hence "nonequilibrium"). Similarly to the previously discussed methods, Equation 7.9 clearly shows that the term inside the $\min[\cdot]$ function is proportional to the product of $M-1$ importance sampling weights and reduces to Equation 7.6 when $M = 2$. It can also be seen that only the normalisation constants of the two endpoint distributions are needed, since the rest cancel out. In the special case of a symmetric protocol $\mathbf{\Lambda} = \tilde{\mathbf{\Lambda}}$, these two normalisation constants cancel out as well.

NCMC has been explored in various contexts, including ligand sampling,[65,96] amino acid side chain sampling,[244,294] binding site water sampling[113–115] and constant-pH simulations.[123,124]

7.2.7 Sequential Monte Carlo (SMC)

Sequential Monte Carlo (SMC) is another sequential importance sampling method, which, unlike NCMC, is ensemble based, i.e. it realises multiple trajectories (walkers) at the same time. Indeed, NCMC can be viewed as a single-walker special case of SMC. The main difference is that SMC is not usually used as an MC proposal method, but rather as a distribution approximation method using the ensemble of walkers. This is done by assigning to the k-th walker a trajectory-dependent weight w_k:

$$w_k \propto \frac{\pi(\vec{\lambda}_{i_2,k}, \vec{x}_{i_1,k})}{\pi(\vec{\lambda}_{i_1,k}, \vec{x}_{i_1,k})} \ldots \frac{\pi(\vec{\lambda}_{i_M,k}, \vec{x}_{i_{M-1},k})}{\pi(\vec{\lambda}_{i_{M-1},k}, \vec{x}_{i_{M-1},k})} \tag{7.10}$$

As in NCMC, these total trajectory weights are simply proportional to the product of $M-1$ consecutive importance sampling weights. They can be afterwards used to estimate weighted averages over the final distribution in the sequence $\pi(\vec{\lambda}_{i_M}, \vec{x})$.

In contrast to NCMC, the simultaneous propagation of K walkers over parameter space permits a special type of weight variance reduction, resulting in the most commonly used variant of SMC: sequential importance resampling (SIR).[295] In this setting, the walkers are resampled proportionally to their weights after each consecutive distribution, resulting in uniform final weights. In this way, SIR automatically assigns more computational time to more probable trajectories.

In physics, SMC has been used to solve solid-state sampling problems,[296,297] perform nonequilibrium free energy calculations,[298–300] facilitate polymer growing and protein folding[301–303] and explore peptide conformers.[304,305] However, SMC-like algorithms have been widely used in many other fields, such as statistics,[306] meteorology,[307] geology[308] and robotics.[309] SMC variants are known under many different names, such as: particle filtering,[310–312] population annealing,[296,297,304,305] diffusion quantum Monte Carlo (DQMC)[313] and Rosenbluth sampling.[314,315] The latter appears to be the first published variant of SMC and one of the first enhanced sampling methods in the literature.

7.3 Discussion

7.3.1 Critical Comparison between ST, ITS, λD and EDS

The methods outlined above are similar in many ways and this similarity will help us determine the advantages and disadvantages of each one of them. For instance, let us consider ST, ITS, λD and EDS. All of these methods involve single-replica simulations over an expanded ensemble. As such, all of them need a way to determine the relative

	State exploration	
Parameter space	Consecutive	Simultaneous
Discrete	ST	EDS
Continuous	λD	ITS

TABLE 7.1: Comparison between single-replica expanded ensemble methods.

normalisation constants between the different components of the distribution in order to achieve good performance. The main differences lie in the presumed continuity of parameter space and the way the expanded ensemble is explored (Table 7.1): ST and EDS explore a discrete sum of underlying distributions, while λD and ITS perform this exploration over a continuous space of parameters; similarly, ST and λD have a well-defined value of the parameter at each snapshot of the simulation, while ITS and EDS perform the integration over an effective potential energy, thereby being in several states at the same time.

While continuous functions are easier to handle from a theoretical perspective, one significant disadvantage of using them is the extra difficulty in analysing the resulting distributions. Since one is often interested in a particular point in parameter space, which recovers the target distribution of interest, the probability of this point being observed is vanishingly small in the continuous setting. Therefore, approximate histogram methods need to be used in practice for analysis. However, it is not obvious what an appropriate choice for a bin size is: too many bins will result in higher variance and lower bias in the number of samples per bin, while the opposite is true for large bin sizes. This can also result in a lack of consistency and impact reproducibility. For example, in the earlier λD publications any $\lambda > 0.8$ was assumed to be equivalent to $\lambda = 1$,[283–285,287] while a more recent publication has considered a much more stringent discretisation cut-off—$\lambda > 0.99$.[284] Although this inconsistency can be alleviated in some applications by using generalised Rao-Blackwell estimators coupled with a Gibbs-sampling approach in continuous $\vec{\lambda}$ space,[285] none of these issues and approximations are present when using discrete states and no extra post-processing is needed on the resulting parameter distributions, since the parameter discretisation is performed *a priori*, rather than *a posteriori*. As already mentioned above, these deficiencies have been recognised in the computational implementation of ITS[243,281,282] and the recent changes to λD,[286] both of which utilise a discrete state space, much like ST and EDS.

An additional drawback of λD is the requirement for a suitable potential energy function which results in dynamics, giving rise to the desired distribution in $\vec{\lambda}$ space. It is not clear in advance what a good choice for such a function is and although multi-layered adaptation procedures have been published to address this problem,[283,284] it remains difficult to engineer the motion in parameter space *a priori*.

This is not the case when MC is instead used for sampling in parameter space, as in ST or REMD.

This lack of control over the transitions in parameter space is also present in the methods which sample from the whole mixture distribution using an effective potential (ITS and EDS), since they do not utilise MC to achieve the sampling. An advantage of this methodology, however, is the ability to include multiple intermediates without slowing down the transitions in parameter space. In contrast, MC-based methods (ST, REMD) decrease in efficiency at least linearly with respect to the number of intermediate states.[316,317] Although this can be alleviated by attempting non-neighbour transitions,[262] EDS and ITS achieve this behaviour automatically without any extra modifications.

A more significant disadvantage of ITS and EDS is the need to compute the contributions from each one of the underlying distributions, even if these contributions are negligible. This results in a linearly increasing cost in force computation with respect to the number of underlying distributions. To mitigate this, one needs to use specific types of parameter schedules which can decrease this complexity.[282,292] However, this also means that these methods are practically less generally applicable than the single-state approaches of ST and λD. This consideration is particularly relevant for alchemical perturbations, where nonlinear soft-core potentials are routinely used and simple linear schedules are rarely feasible.

A unique feature of EDS which circumvents the above issue is the lack of alchemical intermediates in the mixture distribution, instead replacing them with a scaling factor s. However, in most practical cases, this does not remove the need for intermediate states, but rather changes the way they are defined. This approach is advantageous when exploring multiple simultaneous alchemical perturbations, where instead of a vector of alchemical parameters $\vec{\lambda}$, there is only one control parameter s. However, a disadvantage of mapping all of the $\vec{\lambda}$ parameters onto a single one is the different sensitivity of each of them to s. The unevenly distributed kinetic barriers in $\vec{\lambda}$ space can therefore decrease the utility of this mapping.

A first conclusion of this chapter is therefore that ST with a sufficient number of intermediate states is likely to be more advantageous for sampling based on alchemical changes compared to λD, ITS and EDS, due to the discrete nature of its parameter space and its lack of need for effective potential energy functions which limit the use of soft-core potentials. However, an efficient ST algorithm still requires a suitable definition of its underlying states and proposal probabilities in order to surpass the other three methods in the alchemical setting. In the next section, we will concentrate on ST and compare it to the rest of the methods outlined in this chapter. Nevertheless, many of the following considerations for ST apply to λD, EDS and ITS as well.

	No. importance sampling iterations	
No. replicas/walkers	One	Multiple
One	ST	NCMC
Multiple	REMD	SMC

TABLE 7.2: Comparison between tempering methods.

7.3.2 Critical Comparison between REMD, ST, NCMC and SMC

While REMD, ST, NCMC and SMC are at first glance four significantly different methods, these differences can be essentially summarised by only two key decisions made in these algorithms: how many replicas of the procedure are run in parallel, and how many consecutive importance sampling iterations are performed before a proposal (Table 7.2). Here the merits and the implications of each one of these choices are discussed.

The number of consecutive importance sampling iterations is the most significant difference between the above methods. We can distinguish between single-step importance sampling methods (REMD and ST) and sequential importance sampling methods (NCMC and SMC). While the former two methods always sample at equilibrium, the latter two approaches only explore a particular path in the expanded ensemble Markov chain, making them inherently nonequilibrium in nature. Consequently, sequential importance sampling methods are expected to perform comparatively better to the other methods when there are kinetic traps in parameter space arising from unfavourable intermediate states. This is due to the larger expected value of the acceptance criterion for that particular path, stemming from the nonlinearity of the $\min[\cdot]$ function:

$$\min\left[1, \frac{\pi(\vec{\lambda}_{i_2}, \vec{x}_{i_1})}{\pi(\vec{\lambda}_{i_1}, \vec{x}_{i_1})}\right] ... \min\left[1, \frac{\pi(\vec{\lambda}_{i_M}, \vec{x}_{i_{M-1}})}{\pi(\vec{\lambda}_{i_{M-1}}, \vec{x}_{i_{M-1}})}\right] \leq \min\left[1, \frac{\pi(\vec{\lambda}_{i_2}, \vec{x}_{i_1})}{\pi(\vec{\lambda}_{i_1}, \vec{x}_{i_1})} ... \frac{\pi(\vec{\lambda}_{i_M}, \vec{x}_{i_{M-1}})}{\pi(\vec{\lambda}_{i_{M-1}}, \vec{x}_{i_{M-1}})}\right] \quad (7.11)$$

However, it is evident that the expectation value of the acceptance criterion still decreases exponentially quickly with respect to the number of required intermediate states. This can be straightforwardly verified in the special case of the importance sampling weights being log-normally distributed (Appendix C). Since the average amount of computational time required to accept a proposal is proportional to $\frac{1}{\langle p_{acc} \rangle}$, this means that the efficiency of NCMC and SMC decreases exponentially with the number of intermediate states. While in SMC this is partially alleviated by resampling, no such procedure is possible for NCMC, making it the most sensitive method on this list to decreasing endpoint distribution overlap. This sensitivity is particularly pronounced in dense explicit-solvent systems, such as the ones simulated

in biomolecular studies.[294] In contrast, the relaxation times of the Markov chains explored by REMD and ST are known to scale in the range between $O(N)$[316,317] and $O(N^2)$,[318] depending on the Markov chain used, making single-step importance sampling methods much more favourable for exploring a large number of states.

Another disadvantage of sequential importance sampling methods is the fact that rejection (NCMC) or weight degeneracy/pruning (SMC) of unfavourable trajectories is irreversible, meaning that any computational effort invested towards these proposals is effectively wasted to no sampling benefit. In contrast, REMD and ST use all of the available computational time for time-dependent dynamics and there are no rejected states. Since local sampling without enhancement using MD is still more desirable than no sampling, this means that the sequential importance sampling methods could in practice prove less efficient than MD depending on the nature of the system and the way the parameter space is defined.[244,319] This problem is particularly evident when determining the number of MD steps needed for coordinate decorrelation, since too low a number of steps results in insufficient relaxation and thus poor acceptance rate, while too many steps reduce the efficiency of the algorithm.[319] This means that sequential importance sampling methods require careful parameter tuning, whose optimal values are not known *a priori*. In contrast, using short coordinate decorrelation times with single-step importance sampling methods is always more beneficial for mixing[320] and is in practice only limited by the speed of the energy evaluation (usually much faster than force evaluation) and the algorithm implementation details.

While sequential importance sampling methods appear to be significantly less efficient for sampling in the general case compared to the single-step importance sampling methods, there are still merits to using them. For instance, NCMC only needs the normalisation constants for the endpoint distributions, meaning that no free energy estimation is needed for the intermediate states. This is particularly useful in e.g. constant-pH[123] calculations, where the pH of the intermediate states is not a well-defined quantity and partitioning the pH-dependent sequential importance sampling acceptance criterion into multiple single-step importance sampling steps is not trivial and could result in decreased efficiency (c.f. Equation 7.11).

Another potential advantage of NCMC emerges in the commonly used special case of a symmetric protocol with the initial and final distributions being the same, leading to their normalisation constants cancelling out in the acceptance criterion. In this case, if a rare event occurring on a longer timescale than the switching time is observed, this can lead to an effective shift in all intermediate distributions and their corresponding free energy values. While such a transition could greatly reduce the round-trip rate of both ST and REMD, the forward and the reverse NCMC protocols would sample from the same minimum, thereby decreasing the variance of the trajectory weight. Nevertheless, NCMC has not yet been applied to such a scenario in the literature.

The other main difference between the four methods is how many parallel replicas/walkers are used. While ST and NCMC only require a single replica, REMD and SMC are based on multiple replicas. An advantage of the latter two methods is therefore that they are readily parallelisable on e.g. CPU clusters, while the former two methods benefit less from parallelism. However, a significant disadvantage of multiple-replica methods is the fact that the simulation time dedicated to a single replica decreases proportionally to the number of replicas. Since multi-replica simulations are in practice initialised from a common starting structure, this means that they are inherently more biased compared to single-replica methods, due to the decreased decorrelation with respect to their initial coordinates. Indeed, this was already demonstrated in Chapter 5, where it was shown that multiple short simulations can result in free energy estimates significantly biased to the starting structure.

In this regard, ST appears more desirable than REMD. Indeed, it is well-known that ST is more efficient than REMD at exploring the same Markov chain.[321,322] A drawback of ST, however, is the need for adaptive free energy estimation of each of the intermediate states. This means that ST is only more efficient than REMD if it operates in the optimal setting, which is in practice unknown *a priori*.

Although adaptive free energy estimations can be reliably integrated into ST, as already discussed, the most significant weakness of all tempering methods is the lack of prior knowledge of a suitable $\vec{\lambda}$ protocol. This is especially detrimental for ST, since it is not guaranteed to sample all of its underlying distributions uniformly at finite timescales, even with adaptive weights, meaning that kinetic barriers in $\vec{\lambda}$ space can make this non-uniformity even more severe. In contrast, REMD always outputs a predetermined number of samples, even if no sampling enhancement has been achieved. This makes REMD a more reliable method if a suboptimal parameter protocol is used.

Adaptively determining an optimal parameter protocol is therefore a highly desirable task and has been the focus of many studies, mostly concentrated on REMD.[317,323–328] Nevertheless, protocol optimisation is inherently more limited in the case of REMD, since the number of intermediate states cannot be changed without any loss of sample diversity. For example, if the number of intermediates is reduced by the optimisation algorithm, one needs to irreversibly remove one replica from the simulation. On the other hand, if an extra intermediate is introduced, it needs to be initialised from one of the other intermediates, leading to correlated structures. In contrast, ST does not suffer from this difficulty, since only one state is explored at a time and all other states can be optimised without any drawbacks. Despite the more natural integration of adaptive algorithms with ST and their higher impact on its performance relative to REMD, they have been underexplored in the literature and will therefore be the main focus of Chapters 9 and 10.

Depending on the optimisation procedure, it is often advantageous to have a relatively efficient protocol as a starting point before optimisation. In all of the tempering methods discussed above, these need to be manually selected by the researcher beforehand. This can in turn lead to a decrease in efficiency and reproducibility and it is much more desirable to have a method which generates such a protocol without any prior knowledge. SMC is unique in this regard, because it is the only method on this list which can sequentially generate these protocols in a black-box manner, based on distribution-independent overlap metrics (discussed in more detail in Chapter 8). Therefore, while not as efficient at sampling as the single-step importance sampling methods, SMC is ideal for initialising any of the other tempering methods without any prior system knowledge. These can then be refined by on-the-fly optimisation algorithms.

The second conclusion of this chapter is that, in contrast to REMD, SMC and NCMC, ST can significantly decrease the initial-structure bias of molecular simulations with no loss of sampling time, while exhibiting a complexity of at most $O(N^2)$ with respect to the number of intermediate states. This makes it a potentially highly desirable method to use in enhanced sampling and free energy calculations. However, an appropriate protocol optimisation procedure needs to be developed to ensure its robustness and competitiveness with REMD. In the following chapters, this will be done in a two-step manner, where an initial exploratory SMC procedure generates an appropriate initial guess (Chapter 8), which is subsequently refined over time (Chapter 9). In this way, the number of parameters set by the scientist is substantially decreased, thereby resulting in robust automatable enhanced sampling (Chapter 9) and free energy calculation (Chapter 10) workflows.

Chapter 8

Enhanced Ligand Sampling by Adaptive Alchemical Sequential Monte Carlo

8.1 Introduction

As discussed in Chapter 7, sequential Monte Carlo (SMC) is the most suitable tempering method for performing exploratory simulations, mainly due to its directed ensemble approach. In particular, the adaptive tempered version of SMC was recently shown to be efficient at exploring peptide conformations using molecular force field models.[304,305] In this chapter, we will extend this methodology to an alchemical setting, where instead of uniformly increasing the temperature of the whole system, a small subset of the molecular interactions will be completely decoupled instead. This approach is particularly suitable for exploring specific molecular degrees of freedom of interest and has been utilised in other methods, such as Hamiltonian replica exchange (H-REMD)[263,329] and nonequilibrium candidate Monte Carlo (NCMC).[64,65] In the next section, one of the most popular SMC algorithms—sequential importance resampling (SIR)[295]—will first be presented in more detail, before extending it to the context of adaptive alchemical sampling and testing it on a variety of examples.

8.2 Fundamentals of SIR

The fundamental assumption behind SIR is that one starts from a distribution which is trivial to sample from (e.g. a uniform distribution). In most practical examples, where the distributions have many correlated dimensions, this is not possible and the initial distribution is chosen so that transitions between a subset of the modes are more likely

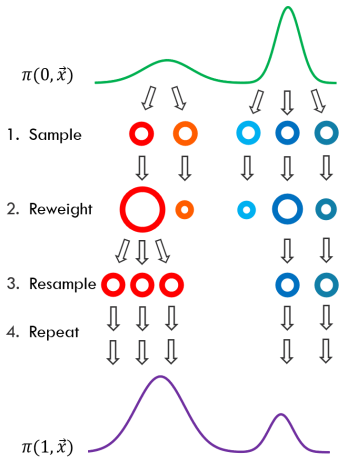


FIGURE 8.1: The three stages of each SIR iteration: sampling, reweighting and resampling. Each unique walker is shown with a different colour and the size of the walker represents its weight. Here $\pi(0,\vec{x})$ and $\pi(1,\vec{x})$ represent the initial and final distributions, respectively.

than in the distribution of interest. Afterwards, a population of samples is propagated over a number of intermediate distributions which connect the initial distribution to the final distribution of interest.

The main focus of this chapter are Boltzmann-like distributions of the form:

$$\pi(\lambda,\vec{x}) = e^{-u(\lambda,\vec{x})+f(\lambda)} \tag{8.1}$$

where $\vec{x}$ are the system coordinates; λ is an adjustable parameter, such that $0 \leq \lambda \leq 1$; $u(\lambda,\vec{x})$ is the dimensionless potential energy of the system, which can also contain extra terms, such as a pressure-volume term in the case of an isothermal-isobaric ensemble; and $f(\lambda)$ is the dimensionless free energy which normalises the distribution. The coupling parameter λ is defined to be 0 at the initial distribution and 1 at the final distribution of interest.

Each SIR iteration consists of three steps (Figure 8.1): sampling, reweighting and resampling. Any valid samplers can be used in the first step, such as Markov chain Monte Carlo (MCMC) or Langevin molecular dynamics (MD), to generate a population of N locally decorrelated samples (walkers). The second step determines the relative transition probability of the j-th walker $p(\lambda_{i+1}|\lambda_i,\vec{x}_j)$ between the current distribution $\pi(\lambda_i,\vec{x})$ and the next distribution in the sequence $\pi(\lambda_{i+1},\vec{x})$, $0 \leq \lambda_i < \lambda_{i+1} \leq 1$. These relative transition probabilities are normalised and converted into importance sampling weights $w_j(\lambda_{i+1}|\lambda_i)$, which are then assigned to each walker:

$$p(\lambda_{i+1}|\lambda_i, \vec{x}_j) \propto \frac{\pi(\lambda_{i+1}, \vec{x}_j)}{\pi(\lambda_i, \vec{x}_j)} = \frac{e^{u(\lambda_i, \vec{x}_j) - u(\lambda_{i+1}, \vec{x}_j)}}{\sum_{j=1}^{N} e^{u(\lambda_i, \vec{x}_j) - u(\lambda_{i+1}, \vec{x}_j)}} \equiv w_j(\lambda_{i+1}|\lambda_i) \tag{8.2}$$

The final step of an SIR iteration consists of weighted resampling with replacement based on these weights to generate a new set of equally-weighted N walkers. This results in the high-weight walkers being copied multiple times and the low-weight walkers being annihilated. This three-step procedure is then repeated for each consecutive distribution until the final distribution has been reached.

One can readily see what sets SIR apart from other enhanced sampling methods: the "survival of the fittest" approach combined with the lack of reversibility, and the fact that the method does not satisfy the rather restrictive detailed balance condition, mean that SIR only explores the best paths and that one can "peek into the future" and adapt the hyperparameters of the method based on this knowledge. This notwithstanding, SIR satisfies a more general stationarity condition, balance,[330] and is known to be completely rigorous in terms of preserving the target distribution $\pi(1, \vec{x})$ in the limit of infinite walkers and infinite sampling at $\pi(0, \vec{x})$.[331]

It can be shown that the expectation value of the unnormalised weights $\tilde{w}_j(\lambda_{i+1}|\lambda_i) \equiv e^{u(\lambda_i, \vec{x}_j) - u(\lambda_{i+1}, \vec{x}_j)}$ of the samples generated from $\pi(\lambda_i, \vec{x})$ is an unbiased estimator of the partition function ratio $\frac{Z(\lambda_{i+1})}{Z(\lambda_i)} = e^{f(\lambda_i) - f(\lambda_{i+1})}$ (Zwanzig equation[12]). This means that $\frac{Z(1)}{Z(0)}$ can also be estimated in an unbiased way from the products of the consecutive expectation values of the unnormalised weights. If one is interested in obtaining unbiased expectation values over separate SIR runs, then the final samples from each run need to be reweighted by the total estimated $\widehat{\frac{Z(1)}{Z(0)}}$ for this run,[332] which can be interpreted as the collective relative weight of the final samples. In effect, the samples are weighted by their free energies, as reflected in the partition function ratio. In this case, the unbiased expectation value $\langle O \rangle$ of an observable O over K independent SIR simulations each having M walkers is:

$$\langle O \rangle = \frac{\sum_{k=1}^{K} \widehat{\frac{Z(1)}{Z(0)}}_k \frac{1}{M} \sum_{i=1}^{M} O_{ik}}{\sum_{k=1}^{K} \widehat{\frac{Z(1)}{Z(0)}}_k} \tag{8.3}$$

where O_{ik} is the observable evaluated on the i-th walker in the k-th simulation and $\widehat{\frac{Z(1)}{Z(0)}}_k$ is the estimated collective walker weight of the k-th simulation.

It is known that this sample reweighting procedure is not in general unbiased for adaptive SIR, where the strides in λ space depend on the weights at each step.[333] Although this condition can be circumvented by running adaptive SIR once and using the derived protocol for all consecutive repeats,[334] this approach is not practical for running simulation repeats in parallel, and in this study we will apply the reweighting

procedure during analysis regardless and demonstrate its sufficient precision in a wide range of test cases.

8.3 Adaptive Alchemical Sequential Monte Carlo

This section highlights some important considerations about performing SMC on a protein-ligand system, as well as several changes to the base method. Some of these modifications allow us to substitute the system-dependent hyperparameters (e.g. the exact sequence of optimal intermediate distributions) with system-independent hyperparameters (e.g. adaptively choosing the intermediate distributions based on constant distribution overlap). Here and henceforth, the method presented in this work will be referred to as adaptive alchemical sequential Monte Carlo (AASMC).

8.3.1 Alchemical Perturbation versus Tempering

Enhancing sampling in temperature space is valuable when one wants to treat all degrees of freedom equally. However, this approach becomes less feasible for large systems and enhancing specific degrees of freedom is often more desirable whenever possible. In this work we consider systems where some degrees of freedom are of greater interest than others. For example, when calculating solvation or protein-ligand binding free energies, the small molecule rotamers are expected to influence the result more than any other degrees of freedom. Therefore, the molecular torsions together with centre-of-mass (COM) translation and rotation constitute arguably the most important degrees of freedom for most small molecules. These are also the degrees of freedom which have multiple minima, often separated by high-energy barriers.

In these cases, one can use an alchemical approach with a coupling parameter λ, where $\lambda = 0$ denotes all relevant interactions turned off, and $\lambda = 1$ represents the target potential energy function of the system (Figure 8.2). In this regime, one can readily use any knowledge from the alchemical free energy (AFE) literature. Most notably, an often employed method for deriving the functional form of the intermediate distributions is to introduce a soft-core potential,[44] which disposes of certain singularities in the potential energy function, thereby improving the statistical efficiency of any estimators dependent on the intermediate λ states. This will be invaluable for the systems discussed later, allowing us to make high-energy insertions and rotations without much of a performance penalty.

There are two common ways to turn on the potential energy interactions: the first is to use the soft-core potential only on the Lennard-Jones (LJ) part of the perturbation, followed by a linear coupling of the electrostatics (“split” protocol); the second

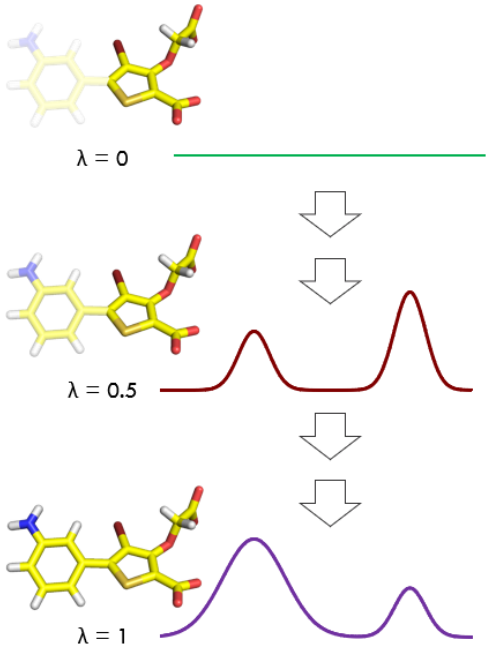


FIGURE 8.2: Exploring conformational degrees of freedom with SMC using an alchemical parameter λ. At $\lambda = 0$, all of the nonbonded interactions involving the 3-aminophenyl group are fully decoupled and the distribution of the torsional angle is uniform. At $\lambda = 0.5$, the 3-aminophenyl group is partially coupled and at $\lambda = 1$ it is fully interacting, in both cases resulting in two main modes/states.

method involves concurrent introduction of all relevant potential terms ("unified protocol"), meaning that a soft-core functional form needs to be used for both LJ and electrostatic interactions. It is expected that a unified protocol is generally less desirable due to the presence of soft-core electrostatic terms, meaning that overlapping positive and negative charges are highly energetically favourable and such unphysical structures can dominate the sampling. On the other hand, the split protocol is expected to produce structures biased towards steric favourability, since most of the resampling is expected to take place before introducing the electrostatics. In this work we will explore and evaluate both protocols.

8.3.2 Adaptively Determining λ_{i+1}

One can use the knowledge obtained from the distribution of the transition probability weights to assess the quality of configurational space overlap between the current distribution $\pi(\lambda_i, \vec{x})$ and the next distribution in the sequence $\pi(\lambda_{i+1}, \vec{x})$. In general, one can use any measure of distribution overlap to achieve this. In the SMC literature, an overwhelmingly popular metric is the effective sample size (ESS) estimator R_{ESS}:[335]

$$R_{ESS}(\lambda_{i+1}|\lambda_i) = \frac{1}{\sum_{j=1}^{N} w_j(\lambda_{i+1}|\lambda_i)^2} \tag{8.4}$$

A general problem with most overlap metrics is the difficulty in defining what value range can be considered "good". Although ESS-based measures can be interpreted intuitively more readily than other measures, it has been suggested[336] that R_{ESS} is not necessarily a reliable estimator for the true ESS and should only be seen as a rough heuristic. Instead, one can use a much more conservative measure R_{min}, which acts as a lower bound for the true ESS:[336]

$$R_{min} = \frac{1}{\max[w_1(\lambda_{i+1}|\lambda_i), ..., w_N(\lambda_{i+1}|\lambda_i)]} \tag{8.5}$$

After defining the desired system-independent value of this measure, one can iteratively[305,333,337] determine the next value in the sequence (λ_{i+1}) which results in an overlap metric closest to this threshold with a basic root finding algorithm, such as bisection. Although each iteration of this adaptive algorithm requires energy evaluations of each walker, they are in practice much faster to perform than generating new samples using dynamics, and the speed of this step will likely be limited by the computational implementation.

The utility of adaptively determining the λ protocol in this way is the guaranteed constant overlap between sequential distributions and the independence of the resulting protocol on the nature of the distributions. Furthermore, if one uses the same overlap metric and value, more dissimilar initial and final distributions will automatically result in a higher number of intermediate distributions without any additional system-specific input.

8.3.3 Adaptively Determining Optimal Sampling Time

An overwhelmingly common way to generate new configurations in biomolecular simulations is MD. This method will be very useful for generating locally decorrelated samples at each λ value. However, the decorrelation time is typically dependent on the system and the nature of the alchemical perturbation. Although it is common practice to choose a value between 1 and 10 ps to achieve local decorrelation, making this step adaptive as well could help maintain the balance between obtaining valid locally decorrelated samples independent of the system and spending as little computational effort as possible.

Since in our typical systems of interest the equilibrium probability of observing a particular configuration $\vec{x}_j$ at some λ is solely a function of $u(\lambda, \vec{x}_j)$, a natural way to measure sample decorrelation is to measure the Pearson correlation coefficient r_τ between the potential energies of all initial walkers $(u(\lambda, \vec{x}_{j,0}); j = 1, ..., N)$ and the walkers decorrelated for τ units $(u(\lambda, \vec{x}_{j,\tau}); j = 1, ..., N)$. Afterwards, the sampling step can only be terminated if r_τ is within some acceptable range,[337] e.g. $|r_\tau| \leq 0.1$. In

practice this step also requires an energy evaluation for every walker and a conceivable implementation could for instance involve evaluating these energies every 1 ps, so as to minimise computational overhead.

8.3.4 Sampling at $\lambda = 0$

SMC only converges to the correct distribution at $\lambda = 1$ if the initial distribution at $\lambda = 0$ has been sampled exhaustively. In a protein-ligand system, this means running long-timescale protein dynamics—a task, which itself often requires other sophisticated enhanced sampling methods to produce satisfactory results. An additional problem is the fact that a very small fraction of the generated structures at $\lambda = 0$ will typically be relevant at $\lambda = 1$, due to diminishing phase space overlap. In this work, we will not be concerned with long-timescale dynamics and we will instead explore ligand conformers from a limited set of locally decorrelated equilibrated starting structures. The aim behind this approximation is being able to quickly estimate equilibrium populations biased to the initial structure either as a qualitative tool or as a way to provide information to more expensive methods, such as AFE calculations. Moving beyond this approximation requires a more sophisticated SMC algorithm which can achieve adequate sampling over time and is thus beyond the scope of this work.

Since this initial stage of SIR is the only checkpoint which generates sample diversity, it is important to take advantage of this. In the test cases we are going to consider, there are three types of sampling moves, for which we know the underlying distribution: torsional rotation, COM rotation, and translation. In all of these cases, we can generate samples typically 1–2 orders of magnitude more than our desired number of walkers, due to the fact that all translational and rotational distributions of the noninteracting atoms in these cases are uniform and therefore trivial to sample.

8.3.4.1 Torsional Rotation

If one removes all nonbonded interactions from at least one side of the torsional bond along with all dihedral terms centred around it, then the initial distribution with respect to the dihedral angle ϕ is uniform and one can generate configurations by simply drawing random numbers between 0 and 2π. One can use any valid sampling method to achieve this and in this work we opt for a low-discrepancy alternative to pseudo-random number generation, which consists of generating equally-spaced samples between 0 and 2π with a pseudo-randomly generated offset. In this way, we can be more certain in the representativeness and quality of our samples.

8.3.4.2 COM Rotation

COM rotation requires all nonbonded interactions between the molecule and the environment to be turned off and it needs three degrees of freedom to be defined: two spherical coordinate angles on the unit sphere, defining the axis of rotation (θ and ϕ) and the amount of rotation ψ around that axis. To generate uniform rotations on the unit sphere, both ϕ and ψ need to be uniformly distributed between 0 and 2π, while $\theta = \arccos(2X - 1)$ for a uniformly distributed variable $X \in [0, 1)$. As in the previous example, one can use different sampling methods to generate the uniformly distributed variables and although one can couple the different degrees of freedom to reduce the multidimensional sample discrepancy (i.e. sample "clumping"), in this study we opt for shuffled one-dimensional grid-based samples with a pseudo-random offset for each degree of freedom. Further research will be needed to test alternative low-discrepancy sampling methods for COM rotation.

8.3.4.3 COM Translation

Much like COM rotation, COM translation requires the molecule of interest to be decoupled from its environment. The simplest case is COM translation within a cuboidal region, in which case only three uniformly distributed random numbers between -1 and 1 are needed to define the new reduced coordinates, which can be then scaled to the dimensions of the region of interest. Alternatively, one can uniformly generate points within a sphere with radius R. To achieve this, we can generate the spherical angles θ and ϕ in the same way as in the previous section, while the radius can be expressed as $r = \sqrt[3]{X}$ for a uniformly distributed variable $X \in [0, 1)$. Final scaling by R results in uniform spherical sampling. Similar considerations about low-discrepancy sampling apply here and we again opt for the same routine for uniform sample generation as in the previous section.

8.3.4.4 Coupled Moves

Since in all of our examples we generate random samples for each degree of freedom independently of the others, this procedure is readily extendable to multiple degrees of freedom. However, the presence of more than a few degrees of freedom can quickly lead to a combinatorial explosion, thereby reducing sampling efficiency, and in this case one should consider multidimensional low-discrepancy sampling alternatives. However, this approach is beyond the scope of this work and we will not be utilising it.

Algorithm 1 AASMC

Input
  $\vec{x}_0$ initial system coordinates
  N number of walkers
  $r_{\tau,target}$ target decorrelation metric
  R_{target} target resampling metric
Output
  $(\vec{x}_1,..., \vec{x}_N)$ the walker coordinates at $\lambda = 1$
procedure AASMC($\vec{x}_0, N, r_{\tau,target}, R_{target}$)
  $\lambda = 0$ ▷ decouple relevant interactions
  $(\vec{x}_1,..., \vec{x}_N) \leftarrow$ Equilibrate($\vec{x}_0$) ▷ spawn N walkers
  $(\vec{x}_1,..., \vec{x}_N) \leftarrow$ GenerateConformers($(\vec{x}_1,..., \vec{x}_N)$) ▷ as in Section 8.3.4
  while $\lambda < 1$ **do**
    $(\vec{x}_{1,0},..., \vec{x}_{N,0}) \leftarrow (\vec{x}_1,..., \vec{x}_N)$ ▷ store initial coordinates
    $r_\tau = 1$ ▷ initial decorrelation metric value
    while $r_\tau > r_{\tau,target}$ **do**
      $(\vec{x}_1,..., \vec{x}_N) \leftarrow$ Sample($(\vec{x}_1,..., \vec{x}_N), \lambda, \tau$) ▷ τ is the sampling time
      $r_\tau \leftarrow$ DecorrelationMetric($(\vec{x}_{1,0},..., \vec{x}_{N,0})$, $(\vec{x}_1,..., \vec{x}_N)$) ▷ as in Section 8.3.3
    $R = 0$ ▷ initial resampling metric value
    while $|R - R_{target}| > \epsilon$ **do** ▷ ϵ determines the precision
      $\lambda_{next} \leftarrow$ ProposeLambda() ▷ using bisection, starting from $\lambda = 1$
      $\vec{w} \leftarrow$ Reweight(λ_{next}, λ,$(\vec{x}_1,..., \vec{x}_N)$) ▷ as in Equation 8.2
      $R \leftarrow$ ResamplingMetric($\vec{w}$) ▷ as in Section 8.3.2
    $\lambda \leftarrow \lambda_{next}$
    $(\vec{x}_1,..., \vec{x}_N) \leftarrow$ Resample($(\vec{x}_1,..., \vec{x}_N)$, $\vec{w}$) ▷ as in Section 8.3.5
  return $(\vec{x}_1,..., \vec{x}_N)$

8.3.5 Using a Conservative Resampling Method

One drawback of SIR is that any loss of walker diversity is irreversible and in many cases all of the final samples can be traced to just a few initial samples.[338] It is important, therefore, to minimise unnecessary diversity loss during the resampling step.

The most obvious way to perform weighted resampling is multinomial resampling with replacement. In this case one draws each new walker independently from the others. This is problematic, since there is always a finite, albeit small, probability that the same sample will be resampled in all cases, resulting in sampling that is potentially not representative of the true weights.

More conservative resampling methods have been proposed, the most deterministic and widely used of which being systematic resampling.[312] In this case, it is guaranteed that the number of new samples corresponding to each weight $w_j(\lambda_{i+1}|\lambda_i)$ (derived from Equation 8.2) is between the rounded down and and the rounded up fractional number of walkers $Mw_j(\lambda_{i+1}|\lambda_i)$, where M is the number of walkers in the next iteration. For example, if the normalised weight of a particular walker is

determined to be 0.27 and the total number of walkers in the next iteration is 10, then the fractional number of copies allotted to this walker is 2.7, meaning that systematic resampling will have a 70% probability of copying this walker three times and 30% probability of copying it twice. Because of this certainty, systematic resampling is highly reliable and will be the algorithm of choice in this study.

8.3.6 An AASMC Workflow in Practice

The first step in describing the problem of interest is identifying the relevant degrees of freedom to be explored, which in turn define a set of interactions to be decoupled at $\lambda = 0$. One then supplies an initial structure, the desired number of walkers, as well as target values for the correlation and decorrelation metrics to the procedure, resulting in an ensemble of structures generated at $\lambda = 1$ (Algorithm 1). While the choice of these hyperparameters is somewhat arbitrary and dependent on the available computational resources, they can be used on a variety of systems and this is the approach which will be taken in this work.

8.4 Methods

8.4.1 System Setup and Simulation

All of the following AASMC simulations have been run using OpenMM 7.4.2,[233] OpenMMTools[339] 0.19.0 and OpenMMSLICER 1.0.0, a plugin for OpenMM developed during the course of this study. All proteins were protonated with PDB2PQR[185] and subsequently parametrised with the ff14SB[25] protein force field. GAFF2[27] with AM1-BCC charges[29,30] was used for all small molecules. All systems were solvated in cubic boxes of TIP3P[31] water with a length of 3 nm for the solvated ligand systems or 7 nm for the protein-ligand systems. Each system was run independently in 6 replicates from the same initial coordinates. Each run consisted of an initial minimisation, followed by 100 ps of equilibration at $\lambda = 0$ before the AASMC run. During this equilibration, all protein backbone atoms were harmonically restrained with force constants of 5 kcal/mol/Å^2. 500 walkers were used for each replicate with 100 initial conformers generated per walker, where all rotatable bonds between alchemical atoms were rotated in addition to the main alchemical moves. An energy decorrelation condition of $|r_{\tau,target}| \leq 0.1$ alongside a minimum relative configurational space overlap of $\frac{R_{min,target}}{N_{walkers}} \geq \frac{1}{5}$ was consistently used throughout the simulations. These values were arbitrarily chosen with the goal of providing a reasonable balance between computational cost and sampling quality. Systematic resampling was performed in all cases and all velocities were resampled from the Maxwell-Boltzmann distribution after each iteration.

All short-range nonbonded interactions had a cut-off of 1.2 nm, while long-range electrostatics were calculated with particle mesh Ewald (PME).[37] A BAOAB[59] Langevin integrator at 298 K with a 2 fs timestep and a collision rate of 1 ps^{-1} was used, where all water molecules were constrained using the SETTLE[53] algorithm and all other bonds containing hydrogen atoms were constrained with the SHAKE[50] and CCMA[52] algorithms. A Monte Carlo barostat was used for pressure control at 1 atm with rescaling attempts every 50 fs. LJ and electrostatic interactions were either switched on simultaneously (unified protocol), or consecutively from $\lambda = 0$ to $\lambda = 0.8$ and from $\lambda = 0.8$ to $\lambda = 1$, respectively (split protocol). A soft-core potential was used for the LJ interactions in both cases and for the electrostatics during the unified protocol with $\alpha = 0.5$, using the following functional form:

$$\begin{aligned} r_{ij,eff} &= (\alpha(1-\lambda)\sigma_{ij}^6 + r_{ij}^6)^{\frac{1}{6}} \qquad \text{(sterics)} \\ r_{ij,eff} &= (\alpha(1-\lambda)\sigma_{ij}^2 + r_{ij}^2)^{\frac{1}{2}} \qquad \text{(electrostatics)} \end{aligned} \tag{8.6}$$

where all inter-atom distances r_{ij} in the potential energy terms involving alchemically modified atoms are replaced with $r_{ij,eff}$ in the potential energy function and σ_{ij} is the "particle size" parameter defined by the LJ potential for the ij-th particle pair. In all cases nonbonded interactions were completely annihilated, rather than decoupled from the environment at $\lambda = 0$.

AASMC was then validated against established methods in one of two ways. The first approach involved a H-REMD simulation in λ space between 0 and 1 with multiple intermediates defined similarly to AASMC. The resulting conformational populations were afterwards obtained from the averaged samples at $\lambda = 1$. The second approach involved AFE calculations, which were only performed when there were only two expected rotamers separated by a high kinetic barrier. In this setting, two separate perturbations were performed in a single-topology fashion from both initial conformations, where the only difference was the rotation of the relevant torsion by 180 degrees, to the nearest common physical intermediate (i.e. to propene in the case of butene and to a phenyl group in place of a substituted phenyl group). The corresponding dihedral terms were not scaled during the AFE calculations, so that no unwanted transitions between the rotamers of interest would be observed. The population ratio between both rotamers $\frac{p_{state1}}{p_{state2}}$ was then calculated using the formula $k_B T \ln \frac{p_{state1}}{p_{state2}} = \Delta G^{\ominus}_{state1 \rightarrow intermediate} - \Delta G^{\ominus}_{state2 \rightarrow intermediate}$.

Both AFE and Hamiltonian replica exchange (H-REMD) calculations were performed in sextuplicate in GROMACS[38] 2018.4, patched with PLUMED[340] 2.4.3 using ProtoCaller 1.1 (Chapter 4) from the same initial structures as those used for the AASMC runs (and in the case of AFE, the relevant manually generated rotameric states). In all cases, the alchemically decoupled groups in the H-REMD simulations were the same as those in the AASMC simulations. The only exception were the

T4-lysozyme/3,5-difluoroaniline simulations, where a single ligand carbon atom remained coupled at $\lambda = 0$ to prevent diffusion away from the (closed) binding site without the need of extra restraint potentials. In some cases several batches of H-REMD simulations were run from different starting conformations to investigate initial structure biasing. These will be indicated later in the text.

The split alchemical protocol was used during both AFE and H-REMD calculations, with 30 initial λ windows used for co-perturbing the soft-core sterics and the bonded interactions, and 10 subsequent windows for the electrostatics. All λ values were equally spaced to two significant figures, except for the initial values, which were more closely spaced in an attempt to increase phase space overlap: 0.001, 0.01, 0.02, 0.03 and 0.05. The Bennett acceptance ratio (BAR)[13] was used for free energy analysis with snapshots every 5 ps.

The AFE protocol involved an initial 25,000-step steepest descent minimisation, followed by a 50 ps NVT equilibration and a 50 ps NPT equilibration before a 4 ns NPT production. The Berendsen barostat[54] was used for equilibration in all cases, while the Parrinello-Rahman barostat was used for the production runs.[60] The LINCS algorithm[51] was used to constrain the non-water hydrogen atoms during both stages, while the rest of the simulation settings matched the ones from the AASMC runs. In the H-REMD simulations, the above equilibration schedule was only performed at $\lambda = 1$ and the resulting volume was fixed for all replicas. This was followed by an additional minimisation and equilibration only in the NVT ensemble and subsequent 4 ns simulations at constant volume. During both H-REMD equilibration and production, adjacent replica swaps were attempted every 1 ps.

8.4.2 Analysis

All of the measured populations in this study were weighted by the estimated partition function ratio $\widehat{\frac{Z(1)}{Z(0)}}$ for the relevant simulation, as previously described in Equation 8.3. These were used to report weighted averages and weighted sample standard deviations. On the other hand, the estimated dimensionless free energies and the simulation times have been reported as unweighted averages with unweighted standard deviations in the main text.

To appropriately analyse the relevant kinetically separated states, clustering on the degrees of freedom of interest was performed. In most cases, this was achieved using manually defined cluster boundaries determined from the observed multimodal distributions of the angle of interest. The only exception is the ligand common core clustering analysis performed for TGF-β, where all trajectories at $\lambda = 1$ from the AASMC and H-REMD simulations were pooled together and aligned against the protein backbone α-carbon atoms of the initial structure using MDTraj[236] and

MDAnalysis.[341,342] Afterwards, the three Euler angles providing the best alignment of the common core ligand atoms against their initial coordinates were calculated using the `align_vectors` routine implemented in SciPy.[343] The sines and cosines of these three Euler angles (six degrees of freedom in total) were used to perform agglomerative clustering with default settings, as implemented in scikit-learn.[238] This analysis resulted in two clusters, whose populations will be reported later in the text alongside two representative structures corresponding to each cluster.

Where applicable, the number of round trips of the H-REMD simulations have been reported. These have been calculated as the total number of round trips of all replicas, where a round trip denotes the transition from $\lambda = 1$ to $\lambda = 0$ and back of a single replica.

8.5 Results

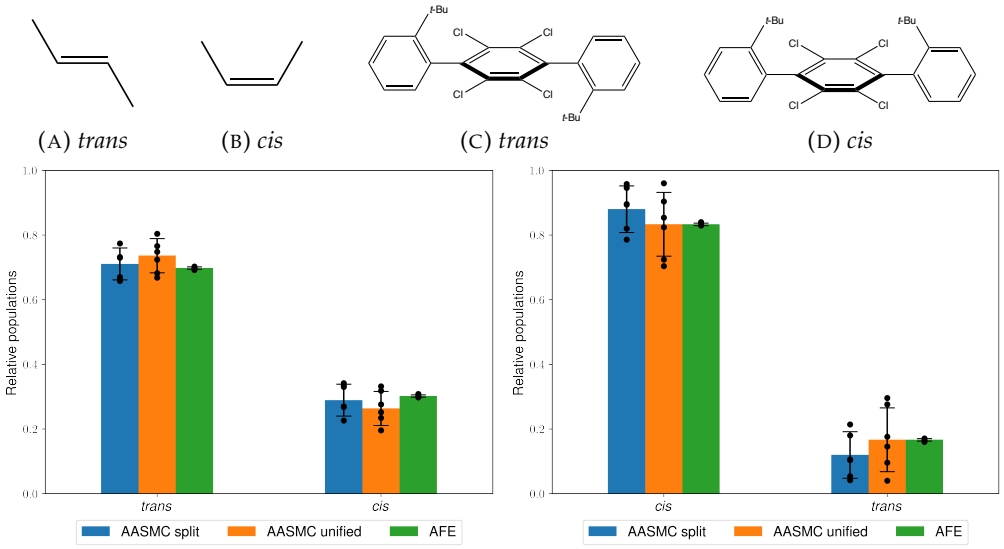
8.5.1 Butene in Water

One of the simplest systems involving a high kinetic barrier is the *cis-trans* isomerisation of butene solvated in water (Figure 8.3). Although not of significant practical interest, this test case is a good demonstration of AASMC's capabilities in an ideal setting. To explore this kinetic barrier, all atoms on one side of the double bond, together with all corresponding dihedral terms, were decoupled from their environment at $\lambda = 0$. This enabled us to directly sample this dihedral angle from the uniform distribution at $\lambda = 0$.

The results from AASMC using both the unified and the split protocols are presented in Figure 8.3e. Both protocols compare favourably to the converged 160 ns AFE results (70% ± 0% *trans*) with the split protocol resulting in 71% ± 5% and the unified protocol yielding an average of 74% ± 5%. In addition, both protocols result in similar performance, with 16 ± 1 ns total computational for the adaptive split protocol and 15 ± 1 ns for the adaptive unified protocol. Finally, both protocols result in comparable standard deviations of $-\ln \frac{Z_1}{Z_0}$ (here and henceforth referred to as "dimensionless free energy") with values of 4.79 ± 0.28 and 4.89 ± 0.15, for the split and the unified protocol respectively, indicating good convergence in both cases.

8.5.2 Terphenyl in Water

A much more challenging test case with an insurmountable kinetic barrier is the terphenyl derivative shown in Figure 8.3. It is expected that only alchemical methods can handle such a system, since approaching the kinetic barrier with all interactions



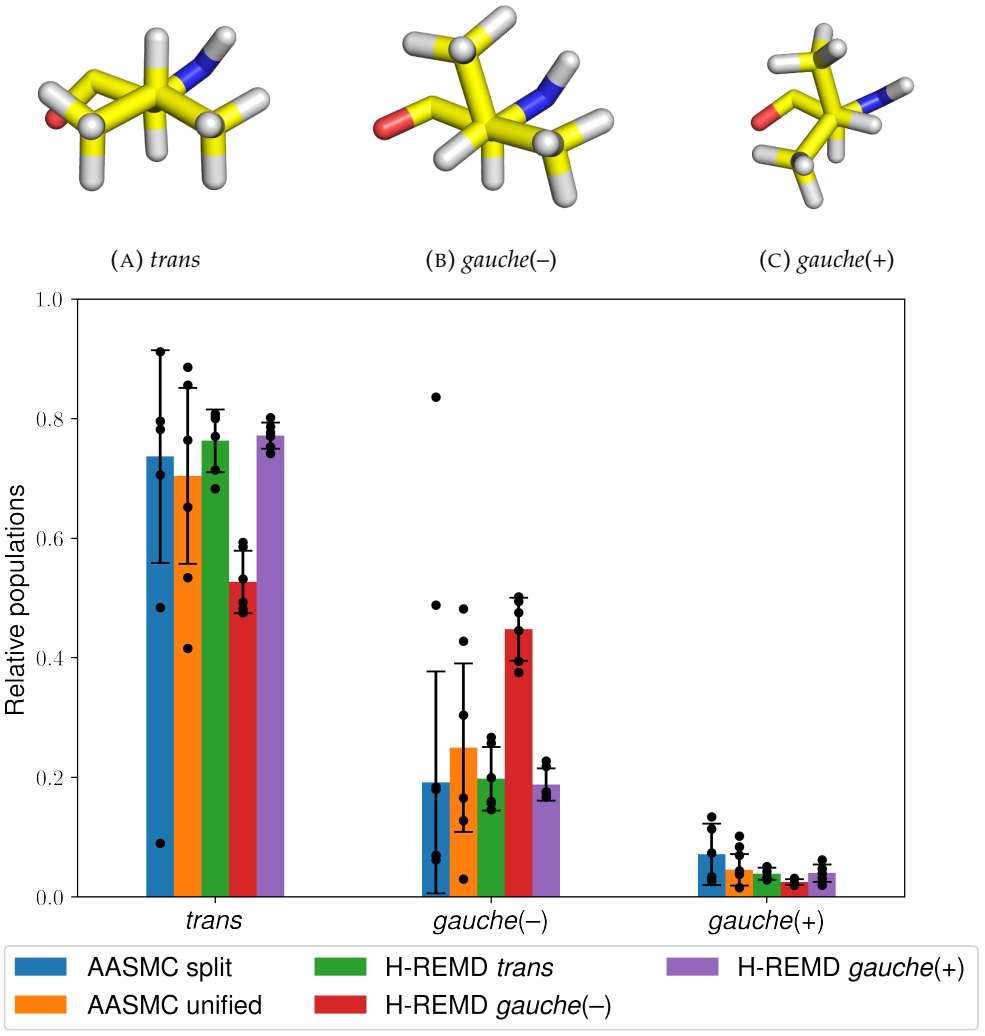
(A) *trans* (B) *cis* (C) *trans* (D) *cis*

(E) Butene populations (F) Terphenyl populations

FIGURE 8.3: The two butene stereoisomers (Figures 8.3a and 8.3b) and the two isomers of the terphenyl derivative (Figures 8.3c and 8.3d) with populations measured by AFE and AASMC (Figures 8.3e and 8.3f) using the split and unified protocols. The heights of the bars represent the mean values weighted by the estimated partition function ratio $\widehat{\frac{Z(1)}{Z(0)}}$ and the error bars represent one weighted standard deviation based on 6 independent runs (shown as individual data points).

turned on will result in large repulsive forces and numerical instability. Moreover, alchemically decoupling the tert-butylphenyl substituent is also likely to be challenging, making this system a good example of a difficult enhanced sampling problem in solution. Similarly to the previous test case, one of the tert-butylphenyl substituents, as well as all dihedral terms corresponding to the rotatable bond, were completely decoupled at $\lambda = 0$ to facilitate sampling.

Figure 8.3f demonstrates that both the split and the unified protocols yield similar results for the main *cis* conformer: $87\% \pm 7\%$ and $83\% \pm 9\%$ compared to $83\% \pm 0\%$ using 160 ns AFE. Moreover, both methods estimate the dimensionless free energy very precisely: 35.60 ± 0.23 for the split protocol and 35.64 ± 0.19 for the unified protocol, indicating good sampling consistency between the AASMC alchemical protocols and repeats. Finally, both methods show similar performance, with the split protocol being slightly slower on average (42 ± 2 ns) than the unified protocol (37 ± 3 ns). The longer average simulation times compared to the butene perturbation show that the adaptive protocol with the same hyperparameters automatically allocates more computational time to a more difficult problem, as expected.



(A) *trans* (B) *gauche*(–) (C) *gauche*(+)

(D) T4-lysozyme Val111 populations

FIGURE 8.4: The three Val111 rotamers (Figures 8.4a to 8.4c) in T4-lysozyme/*p*-xylene and the relative populations of all states using split and unified AASMC and H-REMD from the three different initial rotamers (Figure 8.4d). The heights of the bars represent the mean values weighted by the estimated partition function ratio $\widehat{\frac{Z(1)}{Z(0)}}$ and the error bars represent one weighted standard deviation based on 6 independent runs (shown as individual data points).

8.5.3 T4-lysozyme/*p*-xylene

A seemingly simple case which nevertheless showcases the inability of regular MD to provide adequate sampling is the exploration of the active site Val111 rotamers (Figures 8.4a to 8.4c) in a model T4-lysozyme L99A with bound *p*-xylene (PDB ID: 187L[344]). It has previously been shown[244] that MD results in highly insufficient rotamer transitions even at 1 μs, suggesting that enhanced sampling is indispensable for this system. We can handle this system similarly to the previous test cases by completely decoupling the Val111 isopropyl group and the corresponding dihedral

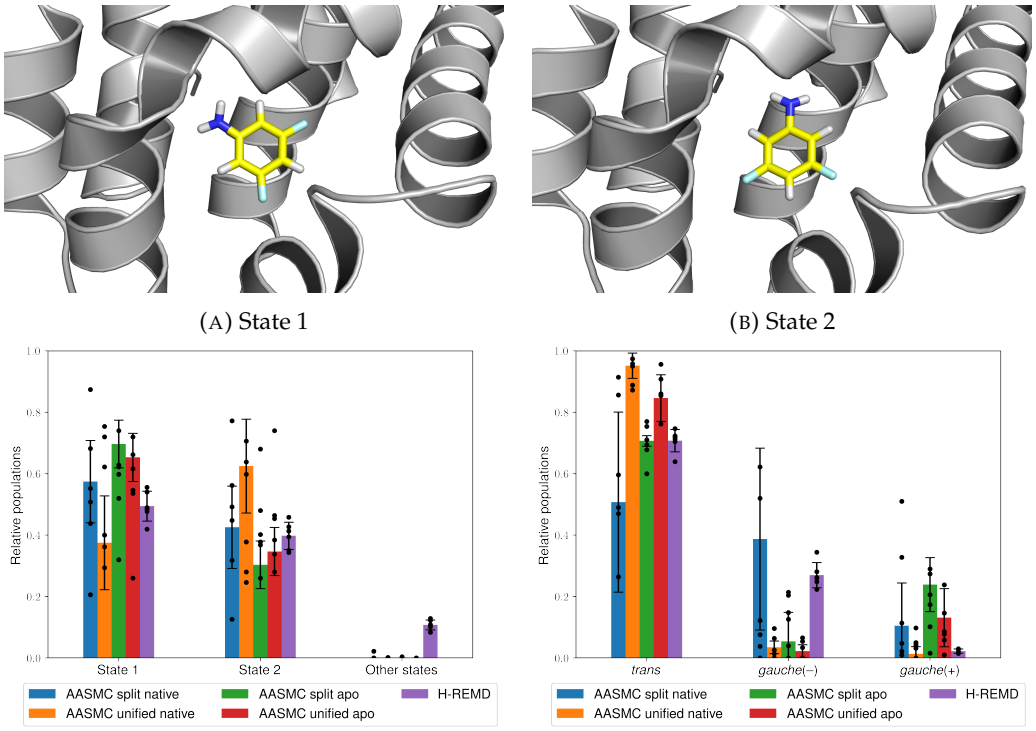
term to facilitate movement at $\lambda = 0$. In this setting, the sampling of *p*-xylene was not enhanced.

The resulting AASMC protocols are highly efficient, requiring an average of 25 ± 2 ns and 21 ± 2 ns per repeat for the split and the unified protocol, respectively, while exploring all relevant Val111 rotamers. Although the split protocol results in higher variance than the unified protocol (Figure 8.4d), both methods result in similar torsional populations and are qualitatively consistent with one another. This is also demonstrated by the relatively high precision of the dimensionless free energy: -43.09 ± 0.85 and -44.32 ± 0.72 for the split and the unified protocol, respectively.

To test the accuracy of the results, they were compared against 6 H-REMD simulations from each initial Val111 conformer (18 simulations in total) with 160 ns per repeat, or 4 ns per replica. As shown in Figure 8.4, even after an average of 252 ± 13 round trips per repeat, there is a significant bias in the populations depending on the starting conformation. This discrepancy can be partially attributed to the fact that the H-REMD implementation used does not explicitly draw the decoupled dihedral from the uniform distribution at $\lambda = 0$, but instead relies purely on integrator decorrelation to achieve this, meaning that any Val111 state transitions are effectively slowed down even when there are no kinetic barriers. In contrast, the AASMC simulations are not biased towards the initial Val111 conformation, since all simulations start from a completely decoupled state. Nevertheless, the relative ranking of the populations is consistent between different starting structures, as well as with the AASMC simulations using either the split or the unified protocol. Although the predicted dominant rotamer (*trans*) does not correspond to that in the crystal structure (*gauche*(–)), the agreement between both enhanced sampling methods suggests that this discrepancy is most likely related to the force field quality and/or long-timescale populations shifts due to e.g. protein rare events, which are beyond the scope of this work.

8.5.4 T4-lysozyme/3,5-difluoroaniline

A more difficult test case is coupling the Val111 motion with translational and rotational movements of the ligand. An example ligand is 3,5-difluoroaniline bound to a L99A/M102Q T4-lysozyme mutant. In this case the ligand was completely decoupled in addition to the Val111 isopropyl group and uniformly moved at $\lambda = 0$ within a sphere with a radius of 0.5 nm centred on its initial COM, suggested by the crystal structure (PDB ID: 1LGX[345]). Since there were two competing ligand binding modes in the electron density, the one with the higher experimentally determined occupancy was chosen for the initial COM evaluation.



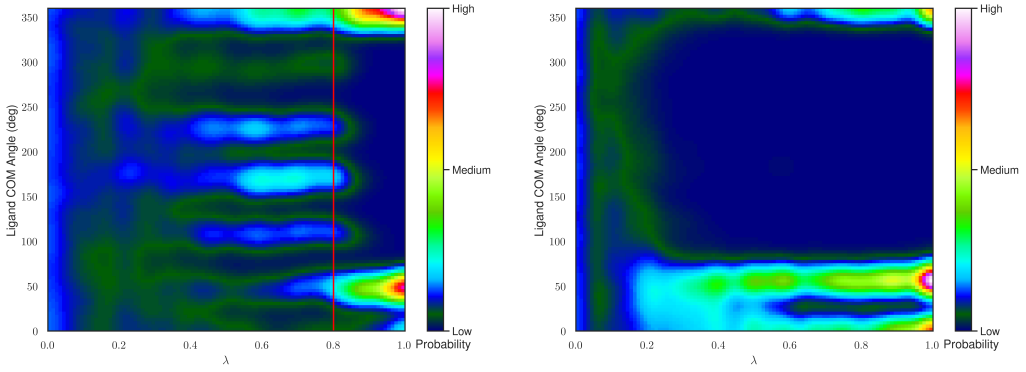
(A) State 1

(B) State 2

(C) T4-lysozyme ligand populations

(D) T4-lysozyme Val111 populations

FIGURE 8.5: The two 3,5-difluoroaniline binding modes (Figures 8.5a and 8.5b) bound to T4-lysozyme, the relative populations of both ligand states using the split and unified AASMC protocols and H-REMD (Figure 8.5c) and the Val111 states from the same simulations (Figure 8.5d). The heights of the bars represent the mean values weighted by the estimated partition function ratio $\widehat{\frac{Z(1)}{Z(0)}}$ and the error bars represent one weighted standard deviation based on 6 independent runs (shown as individual data points).



(A) Split protocol

(B) Unified protocol

FIGURE 8.6: Heat maps of 3,5-difluoroaniline COM angle populations relative to the initial dominant conformer using the split (Figure 8.6a) and unified (Figure 8.6b) AASMC protocols taken from a single representative repeat. The data at discrete λ values have been smoothed in both cases for visual purposes. The solid red line in Figure 8.6a indicates the alchemical intermediate with fully coupled sterics and fully decoupled electrostatics.

The AASMC simulations required an average of 48 ± 2 ns simulation time for the split protocol and 40 ± 2 ns for the unified protocol. Both protocols resulted in two main binding modes for the ligand, which are shown in Figure 8.5. The states are approximately equally probable, with acceptable agreement between the split protocol ($57\% \pm 13\% : 43\% \pm 13\%$), the unified protocol ($38\% \pm 15\% : 62\% \pm 15\%$) and 160 ns H-REMD ($49\% \pm 5\% : 40\% \pm 4\%$). With the exception of the unified protocol, these results are also qualitatively consistent with experiment (60%:40%). The dimensionless free energies are also comparable between both protocols, averaging -266.32 ± 3.46 for the split protocol and -266.68 ± 2.27 for the unified protocol.

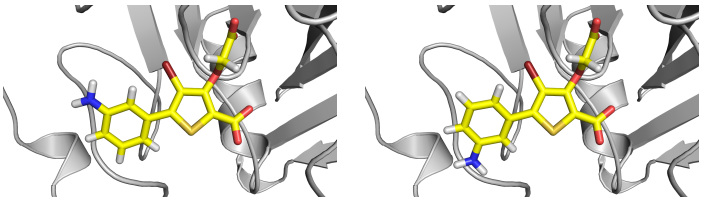
It is interesting to note the sampling differences between both AASMC protocols during the intermediate λ values, as measured by the angle of rotation of the difluoroaniline ring around its centre of mass relative to its initial state (COM angle). As shown in Figure 8.6a, the split protocol explores six different binding modes with approximately equal probabilities during the steric coupling step, before collapsing into the two main binding modes during the electrostatic coupling step. In contrast, the unified protocol (Figure 8.6b) collapses almost immediately into the two main binding modes, indicating that in this case there is higher monotonicity in the population changes over λ.

The same AASMC protocols were performed on the same mutant using a different crystal structure (PDB ID: 1LGU[345]), where only mercaptoethanol (part of the crystallisation liquor) was bound, making this crystal structure the closest experimentally available structure to an apo form for this mutant. Little difference in the results was observed using both the split ($70\% \pm 8\% : 30\% \pm 8\%$) and the unified ($65\% \pm 8\% : 35\% \pm 8\%$) protocols, indicating that the method is not strongly dependent on the initial crystal structure in this case and the results are therefore not biased in an obvious way.

Larger differences were observed for the Val111 rotamers, where there were discrepancies between the populations from both AASMC protocols and H-REMD. Since both the native and the apo structures exhibit significant differences between both protocols, it can be concluded that the split and the unified protocol results are not consistent with each other in this case. This can be attributed to the different ways in which the different λ schedules affect the time-dependent dynamics of each walker. Since the simulation time for each walker remains very short, the lack of long-timescale sampling can therefore result in biased populations.

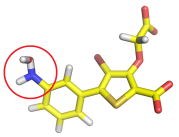
8.5.5 Protein Tyrosine Phosphatase 1B (PTP1B)

Another commonly encountered problem is handling dihedral rotations of flexible bound ligands, such as the thiophene derivative bound to PTP1B (PDB ID: 2QBS[3]),

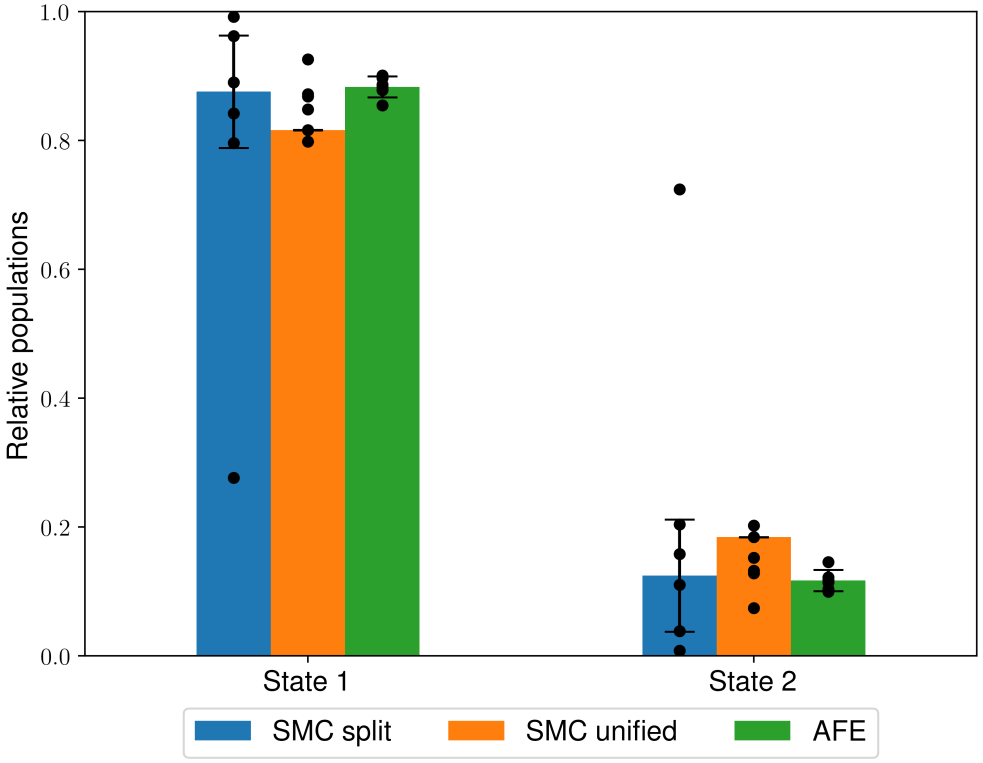


(A) State 1

(B) State 2



(C) An unphysical state



(D) PTP1B ligand populations

FIGURE 8.7: The two thiophene derivative rotamers bound to PTP1B (Figures 8.7a and 8.7b), the unphysical interactions between the amino group and a solvent water molecule commonly observed during the unified protocol (Figure 8.7c, circled in red) and the relative populations of both states using AASMC and AFE calculations (Figure 8.7d). The heights of the bars represent the mean values weighted by the estimated partition function ratio $\widehat{\frac{Z(1)}{Z(0)}}$ and the error bars represent one weighted standard deviation based on 6 independent runs (shown as individual data points).

shown in Figure 8.7. In this case there are two main states of interest (Figures 8.7a and 8.7b) and we can explore this rotation by completely decoupling the 3-aminophenyl group and the relevant dihedral terms at $\lambda = 0$.

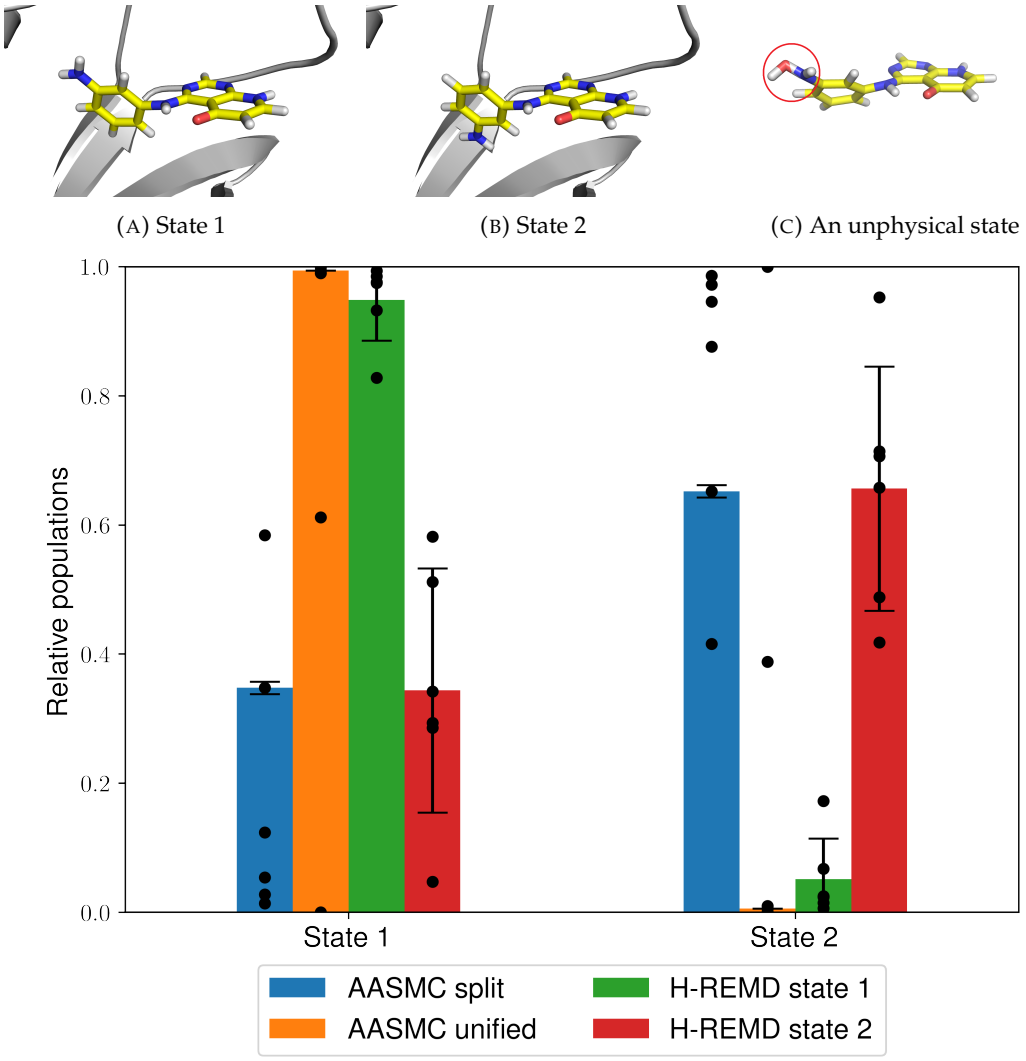
Similarly to the previous torsional rotation cases, there is a good agreement between the dominant conformer in the split protocol ($88\% \pm 9\%$), the unified protocol ($81\% \pm 0\%$), AFE ($88\% \pm 2\%$) and the experimental crystal structure. However, in this case the split protocol results in much higher unweighted standard deviation (26%), mostly caused by a single outlier. Although the split protocol performs apparently worse than the unified protocol, the latter exhibits extremely poor and variable

dimensionless free energy differences: 231.73 ± 20.56, compared to -85.51 ± 2.30 for the former. Since the dimensionless free energies correspond to the negative logarithm of the average relative weight of all walkers sampled from a particular simulation, the unified protocol has a negligible total weight compared to the split protocol due to its strikingly high dimensionless free energy. Therefore, even though the dihedral profiles yielded by the unified protocol appear consistent, the sampling is nevertheless remarkably poor. This can be explained by an energetically favourable overlap between one of the ligand nitrogen atoms and a water hydrogen atom, coupled with an interaction between the aniline hydrogen and the water oxygen (Figure 8.7c). These unphysical interactions are not forbidden and quite favourable, since introducing a soft-core potential to both sterics and electrostatics removes all potential energy singularities at the atom centres. Although these interactions vanish at $\lambda = 1$, they persist for most of the λ schedule, meaning that in this case the split protocol is much more preferable. This conclusion is also supported by the average simulation times: 42 ± 2 ns for the split protocol and 55 ± 1 ns for the unified protocol, indicating that these unphysical states hinder the short-timescale dynamics as well.

8.5.6 Transforming Growth Factor Beta (TGF-β)

The final test case combines a torsional rotation of a flexible ligand bound to transforming growth factor beta (TGF-β) and a nearby Ser82 side-chain rotation. In this case we have used the initial protein coordinates of TGF-β bound to a ligand containing a related symmetric 4-aminophenyl substituent (PDB ID: 4X2G[346]) combined with the initial binding mode of the 3-aminophenyl-substituted ligand of interest (PDB ID: 4X2J[346]), so that the potential bias towards a particular conformer in the initial PDB file has been minimised. It is known from the PDB file that there are two approximately equally-populated alternative conformations of the ligand (Figures 8.8a and 8.8b) and the nearby Ser82 residue (Figures 8.9a and 8.9b). As with the previous examples, this system was handled by decoupling the 3-aminophenyl ligand group concurrently with the Ser82 hydroxymethyl group.

Similarly to PTP1B, the unified protocol has sampling difficulties related to favourable unphysical interactions between an alchemically modified amine group and a water molecule (Figure 8.8c), resulting in large discrepancies between the dimensionless free energies: -225.51 ± 6.25 for the split protocol, compared to 200.65 ± 49.12 for the unified protocol, showing once again that this type of interaction results in populations with a negligible total weight compared to those obtained from the split protocol. Another point of similarity to the previous test case is the higher average simulation time that is needed by the unified protocol: 90 ± 3 ns versus 60 ± 6 ns for the split protocol.

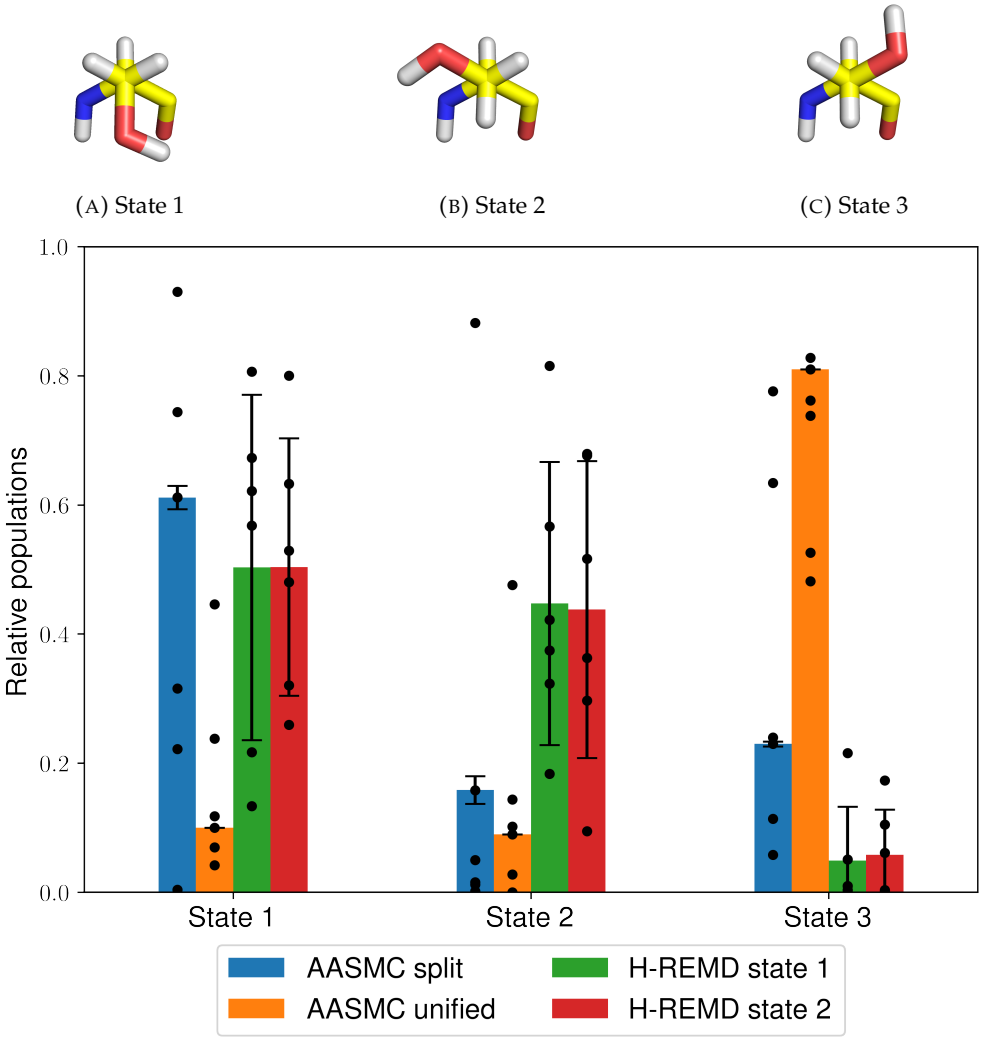


(A) State 1 (B) State 2 (C) An unphysical state

(D) TGF-β ligand populations

FIGURE 8.8: The two TGF-β ligand rotamers (Figures 8.8a and 8.8b), the unphysical interactions between the amino group and a solvent water molecule commonly observed during the unified protocol (Figure 8.8c, circled in red) and the relative populations of both states using the split and unified protocols and H-REMD starting from either of the states (Figure 8.8d). The heights of the bars represent the mean values weighted by the estimated partition function ratio $\frac{\widehat{Z(1)}}{\widehat{Z(0)}}$ and the error bars represent one weighted standard deviation based on 6 independent runs (shown as individual data points).

In both cases, however, there is a marked increase in the relative weight variance compared to the previous test cases, indicating poor convergence. This is also confirmed by the ligand dihedral profiles (Figure 8.8d), which show significant quantitative and qualitative differences between the results of both protocols. This observation is reflected by the low efficiency of the 160 ns H-REMD runs, with an average of only 7 ± 4 round trips per repeat. Despite the low number of round trips and the slow convergence, the data from the H-REMD simulations starting from both ligand rotamers suggest that the first conformer (Figure 8.8a) is likely more stable than the other, implying that the unified protocol is surprisingly qualitatively consistent



(A) State 1 (B) State 2 (C) State 3

(D) TGF-β Ser82 populations

FIGURE 8.9: The three TGF-β Ser82 rotamers (Figures 8.9a to 8.9c) and the relative populations of both states using the split and unified protocols and H-REMD starting from either of the states (Figure 8.9d). The heights of the bars represent the mean values weighted by the estimated partition function ratio $\frac{\widehat{Z(1)}}{Z(0)}$ and the error bars represent one weighted standard deviation based on 6 independent runs (shown as individual data points).

with H-REMD. The two AASMC protocols and H-REMD do not agree on the Ser82 populations, however (Figure 8.9d), meaning that in both cases there is evidence for insufficient sampling.

Clustering analysis of the ligand common core at $\lambda = 1$ using agglomerative clustering (as described in Section 8.4) reveals the presence of two distinct, albeit apparently similar, clusters, which correspond to a concerted translational and rotational motion of the ligand common core (Figures 8.10a and 8.10b). It can be seen that the first cluster is overrepresented in the unified protocol structures, as well as the H-REMD simulations starting from the dominant state (Figure 8.10c). However, the

second cluster is the one resulting in the highest total relative weights for both the split and unified AASMC runs. Since both clusters are more equally sampled during the comparatively longer H-REMD simulations, this behaviour indicates an insufficient level of decorrelation of the AASMC results from the initial structure, resulting in significant biasing of the observed ligand dihedral populations. Moreover, these ligand transitions present an orthogonal rare event which is not adequately sampled even at longer timescales and thus results in increased population variance for both AASMC and H-REMD.

8.6 Discussion

The above results show that AASMC is extremely efficient at exploring ligand conformers in solution, even for alchemical changes that would be considered difficult to perform in practice. This is not surprising, since this is the ideal setting for the method: the ligand degrees of freedom are the only ones which require extensive sampling, while the environment does not need much long-timescale sampling to respond to the ligand motions. Therefore, AASMC can be a valuable tool in exploring the degrees of freedom of solvated small molecules and is likely one of the most robust ways to achieve this.

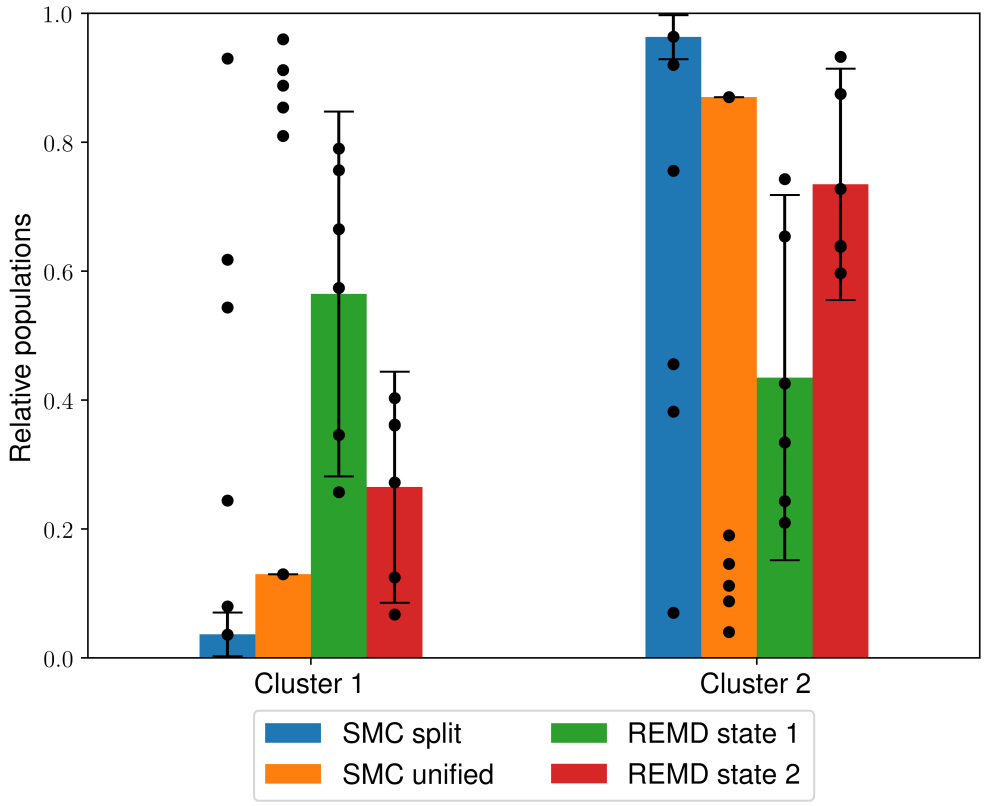
The T4-lysozyme test cases show that a closed binding pocket exhibiting little flexibility also constitutes a favourable application of the method. We have shown that AASMC is unaffected by high kinetic barriers and relatively unbiased towards the initial ligand structure, while providing efficient protocols which require no system-specific parameters. These results appear to hold even when exploring coupled motions between a side chain and a ligand.

Similar observations have been made for PTP1B, where the ligand is strongly bound to the protein and the rotatable group of interest faces the solvent. In this case, the efficiency of AASMC is similar to the one observed in the solvated ligand systems. However, the resulting unweighted population variances from all protein test cases are much higher compared to the first two test cases and this trend carries to the dimensionless free energies. This is expected, since protein-ligand systems present a much more challenging sampling problem compared to solvated ligand systems.

TGF-β presents a more challenging system, where rare motions of the unmodified part of the ligand contribute to a much higher observed dihedral population variance, compared to the previous test systems. This behaviour is observed for both AASMC and H-REMD, meaning that exploring long-timescale motions for this system is crucial and the short-timescale AASMC runs are not sufficient in this case. It is therefore important to be able to identify such potentially problematic systems *a priori* and this should be addressed in future work.

(A) Cluster 1

(B) Cluster 2



(C) TGF-β ligand cluster populations

FIGURE 8.10: The two TGF-β ligand clusters, (Figures 8.10a and 8.10b, common core circled in red) and their relative populations using the split and unified protocols and H-REMD starting from either of the states (Figure 8.10c). The heights of the bars represent the mean values weighted by the estimated partition function ratio $\frac{\widehat{Z(1)}}{Z(0)}$ and the error bars represent one weighted standard deviation based on 6 independent runs (shown as individual data points).

The above test cases also show the advantages and disadvantages of the split and unified force field scaling protocols. It has been demonstrated that the unified protocol can result in unpredictable performance and can suffer from unphysical interactions between atoms with opposite charges, resulting in them collapsing on top of one another. This means that while the unified protocol can in many cases be more efficient than the split protocol, it is also less robust. The split protocol, on the other hand, has been shown to be extremely consistent both in terms of sampling time and

free energy estimation, but often results in a larger unweighted variance of the sampled populations. It is not yet clear how the above protocols will perform in a system which exhibits a drastic shift in rotamer populations when the electrostatic interactions are switched on, but the above results strongly suggest that the split protocol is by far the safer and more conservative choice for most systems.

All of the above results paint a clear picture of the current advantages and limitations of AASMC. AASMC excels in cases where one is interested in few degrees of freedom and where the populations of interest remain relatively unchanged over long timescales. In such systems, one can expect high performance with minimal user input, meaning that very different systems can be run with the same hyperparameters without external intervention. Another advantage of AASMC is the lack of need for supplying an initial conformation of the degree of freedom of interest, thereby providing an unbiased estimate of the population over this degree of freedom. In contrast, while H-REMD results in populations with apparently lower variance than AASMC, it also exhibits long-timescale bias towards the initial conformation. Taking this bias into account then results in a similar performance to AASMC. Moreover, the collective estimated AASMC simulation weights $\widehat{\frac{Z(1)}{Z(0)}}$ provide a straightforward way to measure sampling quality, while investigating bias is not as obvious, meaning that AASMC is more useful for performing exploratory simulations. On the other hand, AFE calculations result in significantly lower variance than both AASMC and H-REMD but their main disadvantage is that a separate simulation is required for each cluster of interest. These need to be known in advance and this knowledge is not always available in practice.

AASMC is therefore a valuable qualitative exploratory tool which can quickly provide initial structures which are unbiased over a particular degree of freedom, as well as generate an efficient λ schedule for another more computationally expensive method, such as AFE calculations or H-REMD simulations. The latter methods, on the other hand, can sample over arbitrarily long timescales, thereby being systematically improvable while simultaneously reducing their bias towards the initial protein crystal structure over time. This decorrelation is crucial in, for example, binding free energy calculations, where the initial protein crystal structure can significantly impact the calculated free energies (Chapter 5). Therefore, one can use the strengths of both AASMC and long-timescale methods to minimise the dependence of the sampled conformations on the choice of initial protein and ligand coordinates.

Owing to the shortness of its simulations, AASMC is so far impractical for sampling long timescale motions and can suffer from initial structure templating, as well as kinetic trapping which occurs during the alchemical steps. The latter is a significant challenge not only for AASMC but for H-REMD as well, and it can be triggered by orthogonal rare events, such as slow motions of the unmodified part of the ligand,

which makes this behaviour difficult to predict. These problems will therefore require key modifications to the AASMC method and will be the subject of future work.

8.7 Conclusion

In this chapter, the adaptive alchemical sequential Monte Carlo (AASMC) method was presented. AASMC is a directed and irreversible ensemble-based algorithm and can be used for sampling rare events using adaptive sequential importance resampling. AASMC combines adaptive sequential Monte Carlo methods with knowledge from the alchemical free energy (AFE) literature and is thus ideally suited for protein-ligand and related systems where the requirement for system-specific method parameters would be highly undesirable.

The performance of AASMC was demonstrated on a variety of test cases where regular molecular dynamics (MD) is unable to provide adequate sampling, and the relative efficiencies of a split perturbation protocol and a unified scheme were compared. It was shown that that AASMC performs best when the results are largely independent of long-timescale motions and other important orthogonal kinetic barriers. In these cases, AASMC provides efficient sampling and is unaffected by the exact nature and size of the system. The most consistent and robust results are also observed when the split protocol is used, which makes it more desirable in the general case.

The above results show that AASMC is good at generating unbiased conformations over a selection of degrees of freedom. Moreover, it provides a good metric for convergence, the estimated collective weight $\widehat{\frac{Z(1)}{Z(0)}}$, which can be used to assess the sampling quality over different simulation repeats. In this setting, methods such as H-REMD are less useful, due to their long-timescale bias, which is often difficult to detect. Similarly, AFE calculations require prior knowledge for the conformers of interest and their cost scales rapidly with the number of possible states for each degree of freedom. This makes AASMC a good method for performing exploratory simulations with minimal input. However, in one of the test cases, AASMC exhibits large variance and poor convergence. This suboptimal performance can be attributed to high dependence on the initial coordinates, meaning that the method needs to be extended to long-timescale sampling. The extension of AASMC to longer timescales will be the subject of Chapter 9.

Chapter 9

Enhancing Torsional Sampling Using Fully Adaptive Simulated Tempering

9.1 Introduction

Since sequential importance sampling methods are expected to be less efficient at sampling than single-step importance sampling methods (Chapter 7), the optimal approach for extending adaptive alchemical sequential Monte Carlo (AASMC) to long timescales is to combine it with methods such as replica exchange molecular dynamics (REMD) or simulated tempering (ST). Such a hybrid method would then have all of the advantages of both methods and eliminate most of their disadvantages.

It was discussed in Chapter 7 that REMD is one of the most widely used enhanced sampling algorithms in computational chemistry. However, despite its widespread use, the standard REMD algorithm suffers from a number of drawbacks. The most significant of these is the requirement of having a number of parallel simulations, which are in practice instantiated with shortly equilibrated (i.e. highly correlated) structures. This means that given the same total simulation time, more replicas result in worse exploration of long-timescale motions, thereby diminishing the probability of unexpected rare events. The second issue is the fact that the original REMD algorithm is reversible,[347] which results in suboptimal diffusive motion in temperature space with relaxation time complexity of $O(N^2)$ with respect to the number of intermediate temperatures.[318] Finally, REMD, as all tempering methods, can be greatly hindered by suboptimal temperature spacing, meaning that optimal temperature protocols are needed.

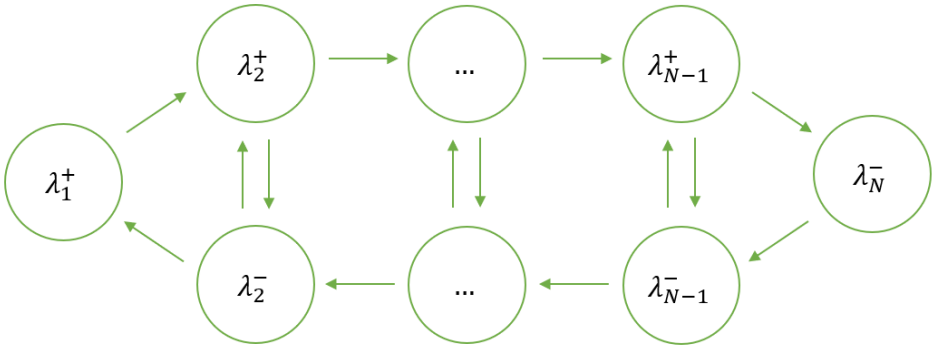
The first issue is circumvented by the serial version of REMD, ST.[240,270] This makes ST the method of choice in this chapter. In ST, only one replica is simulated and it is periodically moved in temperature space. Because all of the computational time is devoted to a single structure, ST achieves higher decorrelation from the supplied initial coordinates compared to REMD. However, unlike REMD, an efficient ST algorithm needs precise free energy estimates, even if these are not formally needed to correctly sample from each temperature. Although there have been several approaches which solve this problem,[271,272,275,279] ST remains sensitive to suboptimal free energy estimates and temperature protocols, since it is not guaranteed to spend equal amounts of time at each temperature after a finite number of timesteps. This has led to ST being underutilised in molecular simulations compared to REMD,[271] although exceptions exist.[274,278,348,349]

The suboptimal $O(N^2)$ complexity of tempering methods can be remedied by modifying the topology of the Markov chain,[262] or, more specifically, by using irreversible Markov chain Monte Carlo (MCMC) methods, which are known to provide faster state mixing than their reversible counterparts.[318,350] A straightforward way to make reversible algorithms irreversible is to introduce an additional "lifting coordinate"[351] and enforce an antisymmetric balance condition, known as skew detailed balance.[352] One can then devise a suitable expression for the acceptance criterion which minimises the diffusive motion in temperature space as much as possible.[353] This methodology can be readily applied to a variety of MCMC methods,[350] and irreversible REMD and ST algorithms have been previously published.[316,317,328] The introduction of this irreversibility can substantially improve the motion in temperature space, which can reach a scaling of $O(N)$ with respect to the number of intermediate distributions.[316]

This chapter will describe a general numerical methodology which solves the third issue described above—obtaining an optimal temperature protocol. While analytical results exist for reversible REMD,[354–356] irreversible REMD[317,328] and reversible ST,[356–358] a simple analytical expression is not known for irreversible simulated tempering (IST). Such a methodology is highly desirable not only because of its general applicability, but also because of the well-known fact that a well-performing ST algorithm is more efficient than a well-performing REMD algorithm.[321,322] One can then use this procedure to maximise the sampling quality of the tempering method.

The protocol optimisation process has two layers of adaptation: a preliminary offline protocol estimation using AASMC, as described in Chapter 8, followed by an iterative online protocol refinement during the ST procedure. The former approach is similar to the methodology considered by Syed *et al.*,[317] while the online optimisation algorithm is the main contribution of this work.

(A) ST Markov chain



(B) IST Markov chain

FIGURE 9.1: The Markov chains corresponding to simulated tempering (ST) (Figure 9.1a) and irreversible simulated tempering (IST) (Figure 9.1b).

As in Chapter 8, the primary focus of this chapter is the applicability of enhanced sampling to the exploration of small molecule rare events. To achieve this, a subset of the nonbonded interactions will be completely decoupled in a simulated scaling manner,[276] meaning that instead of temperature space, we will be working in alchemical (λ) space, where the decoupled system will be denoted by $\lambda = 0$ and the fully-interacting system will be denoted by $\lambda = 1$. Nevertheless, the adaptive procedure presented below is generally applicable to any family of distributions parametrised by a single parameter.

9.2 Theoretical Background

The following discussion builds upon the basic ST algorithm presented in Chapter 7 and employs the same notation as defined there.

9.2.1 Irreversible Simulated Tempering (IST)

Irreversible simulated tempering (IST)[316] is a generalisation of ST which relaxes the condition of detailed balance (Equation 2.34) to that of skew detailed balance.[352] To achieve this, an extra variable σ ("lifting coordinate") is introduced,[351] thereby creating a mirror image of the irreversible Markov chain of choice (Figure 9.1). In this way, the Markov chain increases its state space from N_λ states to $2N_\lambda - 2$ states. With the skew detailed balance condition, the underlying Markov chain can be designed in any way, as long as all of its states are connected. In IST, the extra variable can be

thought of as the direction of the flow—"+1" in the direction $\lambda_1 \rightarrow \lambda_{N_\lambda}$ and "−1" in the opposite direction (Figure 9.1b). This extra variable is strictly $+1$ at $\lambda = \lambda_1$, -1 at $\lambda = \lambda_{N_\lambda}$ and can take any value at the intermediate λ values. This changes the proposal probability $p_{prop}(\lambda_j|\lambda_i)$ to:

$$p_{prop}(\lambda_j, \sigma_j|\lambda_i, \sigma_i) = \begin{cases} \delta_{\sigma_i\sigma_j} & \lambda_j = \lambda_{i+\sigma_i}, \lambda_j \notin \{1, N_\lambda\} \\ 1 - \delta_{\sigma_i\sigma_j} & \lambda_j = \lambda_{i+\sigma_i}, \lambda_j \in \{1, N_\lambda\} \\ 0 & \text{otherwise} \end{cases} \tag{9.1}$$

In IST, the skew detailed balance condition enforces the equality:

$$\pi(\lambda_i, \sigma_i, \vec{x})T(\lambda_{i+\sigma_i}, \sigma_i|\lambda_i, \sigma_i, \vec{x}) = \pi(\lambda_{i+\sigma_i}, -\sigma_i, \vec{x})T(\lambda_i, -\sigma_i|\lambda_{i+\sigma_i}, -\sigma_i, \vec{x}) \tag{9.2}$$

with the transition probability $T(\cdot)$ being the product of the proposal probability $p_{prop}(\cdot)$ and the acceptance probability $p_{acc}(\cdot)$. In this setting, the probability of evolving in λ space $p_{acc}(\lambda_j, \sigma_j|\lambda_i, \sigma_i, \vec{x})$ is similar to the reversible case (Equation 7.6):

$$p_{acc}(\lambda_j, \sigma_j|\lambda_i, \sigma_i, \vec{x}) = \min\left[1, \frac{w_j\pi(\lambda_j, \vec{x})p_{prop}(\lambda_i, \sigma_i|\lambda_j, \sigma_j)}{w_i\pi(\lambda_i, \vec{x})p_{prop}(\lambda_j, \sigma_j|\lambda_i, \sigma_i)}\right] \tag{9.3}$$

where we have omitted the lifting variable from the probability distributions, since it is purely a dummy variable which does not change their functional form. In order to satisfy skew detailed balance, the probability flow in σ space needs to counterbalance the flow in λ space. This leads to the lifting coordinate being flipped with probability $p_{acc}(-\sigma_i|\lambda_i, \sigma_i, \vec{x})$:

$$p_{acc}(-\sigma_i|\lambda_i, \sigma_i, \vec{x}) = \max\left[0, \textstyle\sum_{\sigma_k\in\{-1,1\}} T(\lambda_{i-\sigma_i}, \sigma_k|\lambda_i, -\sigma_i, \vec{x}) - T(\lambda_{i+\sigma_i}, \sigma_k|\lambda_i, \sigma_i, \vec{x})\right] \tag{9.4}$$

The term inside the max$[\cdot]$ function is simply the difference between the forward transition probability at the current state and the backward transition probability at the mirrored state. This acceptance criterion is not the only one that satisfies skew detailed balance but is the one that minimises the probability of changing directions.[353] Since the rationale for using IST is precisely the minimisation of diffusive motion in λ space, this is the acceptance criterion that will be used hereafter. Finally, the probability of not accepting any λ or σ transitions $p_{rej}(\lambda_i, \sigma_i, \vec{x})$ is:

$$p_{rej}(\lambda_i, \sigma_i, \vec{x}) = 1 - p_{acc}(\lambda_j, \sigma_j|\lambda_i, \sigma_i, \vec{x}) - p_{acc}(-\sigma_i|\lambda_i, \sigma_i, \vec{x}) \tag{9.5}$$

In practice, one of $p_{acc}(\lambda_j, \sigma_j|\lambda_i, \sigma_i, \vec{x})$, $p_{acc}(-\sigma_i|\lambda_i, \sigma_i, \vec{x})$ and $p_{rej}(\lambda_i, \sigma_i, \vec{x})$ is chosen using a pseudo-random number uniformly distributed between 0 and 1, leading to either a transition in λ space, a σ flip, or no change. Similarly to ST and REMD, the transitions in λ and σ space are performed completely independently of the evolution in coordinate ($\vec{x}$) and momentum ($\vec{p}$) space and can in principle be attempted at any point in the simulation.

9.2.2 Multistate Bennett Acceptance Ratio (MBAR)

The multistate Bennett acceptance ratio (MBAR)[18] is a maximum likelihood free energy estimation method, which is also known to be statistically optimal, in the sense of minimising the asymptotic estimator variance. MBAR estimates different thermodynamic observables by creating a self-consistent expanded ensemble model $\hat{\pi}_{mix}(\vec{\lambda}, \vec{x})$ (here and henceforth, the hat operator denotes an estimated quantity):

$$\begin{aligned} \hat{\pi}_{mix}(\vec{\lambda}, \vec{x}) &= \sum_{i=1}^{N_\lambda} \hat{w}_i \hat{\pi}(\lambda_i, \vec{x}) \\ \hat{w}_i &= \frac{N_i}{\sum_{k=1}^{N_\lambda} N_k} \\ \hat{\pi}(\lambda_i, \vec{x}) &= e^{-u(\lambda_i, \vec{x}) + \hat{f}(\lambda_i)} \\ \hat{f}(\lambda_i) &= -\ln \left\langle \frac{e^{-u(\lambda_i, \vec{x})}}{\hat{\pi}_{mix}(\vec{\lambda}, \vec{x})} \right\rangle_{\hat{\pi}_{mix}(\vec{\lambda})} \end{aligned} \tag{9.6}$$

where N_i is the total number of samples from λ_i, $\hat{w}_i$ is the estimated expanded ensemble weight of the i-th state and $\hat{f}(\lambda_i)$ is the estimated free energy at λ_i, usually chosen to be relative to $\hat{f}(\lambda_1)$. These equations can be solved iteratively using different approaches.[18,20] One can then estimate the expectation value of any observable of interest O at any intermediate distribution $\pi(\lambda_k)$ using importance sampling, even if this distribution is not explicitly sampled in the observed data:

$$\left\langle O(\vec{\lambda}, \vec{x}) \right\rangle_{\pi(\lambda_k)} \approx \left\langle O(\vec{\lambda}, \vec{x}) \frac{\hat{\pi}(\lambda_k, \vec{x})}{\hat{\pi}_{mix}(\vec{\lambda}, \vec{x})} \right\rangle_{\hat{\pi}_{mix}(\vec{\lambda})} \tag{9.7}$$

where the expectation $\langle \cdot \rangle_{\hat{\pi}_{mix}(\vec{\lambda})}$ is obtained by averaging the integrand over each of the total samples. In this work, the samples will be obtained at the same time as proposing a move in λ space, or every τ_{sample} units, where τ_{sample} is the sampling time between λ change proposals. Afterwards, the resulting MBAR estimator will be used to predict the expectation values of the acceptance criteria described above and estimate transition matrices, as discussed in section Section 9.2.3.

9.2.3 On-the-Fly Protocol Adaptation

The main contribution of this work is the development of a general adaptive on-the-fly procedure to continuously estimate the optimal protocol $\vec{\lambda}_{opt} \equiv (\lambda_{1,opt}, ..., \lambda_{N_\lambda,opt})$. We will define $\vec{\lambda}_{opt}$ to be the protocol which minimises the predicted expected round-trip time $\hat{\tau}_{round,pred}(\vec{\lambda})$ between $\lambda_1 = 0$ and $\lambda_{N_\lambda} = 1$ (Figure 9.1). To obtain this, we first use our MBAR model to estimate an expected transition matrix $\widehat{\mathbf{T}}(\vec{\lambda})$ connecting all the states in λ and σ space. In the case of IST, this translates to a $(2N_\lambda - 2)$ by $(2N_\lambda - 2)$ matrix (Figure 9.1):

$$\widehat{T}_{ij}(\vec{\lambda}) \approx \left\langle \widehat{T}(\lambda_j, \sigma_j | \lambda_i, \sigma_i, \vec{x}) \frac{\hat{\pi}(\lambda_i, \vec{x})}{\hat{\pi}_{mix}(\vec{\lambda}, \vec{x})} \right\rangle_{\hat{\pi}_{mix}(\vec{\lambda})} \tag{9.8}$$

where each transition probability is calculated using the MBAR free energy estimates according to Equations 9.3 to 9.5. In all cases $\hat{\pi}_{mix}$ is estimated from all previous samples at all λ values, as described in Equation 9.6.

One can then straightforwardly obtain $\hat{\tau}_{round,pred}(\vec{\lambda})$ by expressing it as the sum of the mean first passage times $\hat{\tau}_{\lambda=0\to\lambda=1} + \hat{\tau}_{\lambda=1\to\lambda=0}$. More generally, the mean first passage time $\hat{\tau}_{ij}$ from state i to state j can be obtained from the equation:[359]

$$\begin{aligned} \hat{\tau}_{ij} &= [(\mathbf{I} - \widehat{\mathbf{T}}_{jj})^{-1}\mathbf{1}]_i \tau_{sample} \qquad i < j \\ \hat{\tau}_{ij} &= [(\mathbf{I} - \widehat{\mathbf{T}}_{jj})^{-1}\mathbf{1}]_{i-1} \tau_{sample} \qquad i > j \end{aligned} \tag{9.9}$$

where $\widehat{\mathbf{T}}_{jj}$ is the transition matrix with the j-th column and row removed, $\mathbf{I}$ is the identity matrix, $\mathbf{1}$ is a column vector of ones and $[\cdot]_i$ denotes the i-th vector element.

The key assumption behind this methodology is the instantaneous decorrelation of the phase space coordinates—a necessary assumption which is not usually satisfied in real-world applications. Nevertheless, as will be shown later, it is a very useful assumption which works remarkably well in practice, since even dense macromolecular systems often exhibit apparent local energy decorrelation at relatively short timescales (1–10 ps).

The minimisation of $\hat{\tau}_{round,pred}(\vec{\lambda})$ with respect to $\vec{\lambda}$ requires an appropriate derivative-free optimisation method. In this work we opt for the CMA-ES algorithm,[360] which is a global optimisation algorithm that is well-known for its robust performance at the number of dimensions relevant to alchemical transformations (typically ≤ 50 in practice).[361] In order to keep the sensitivity of each λ value towards variation relatively constant, we will map the λ protocol onto an equally-spaced sequence $\in [0, 1]$ using a piecewise linear interpolation function and perform the optimisation in this transformed space. The optimisation procedure

always keeps three λ values unchanged: the current λ value, 0 and 1. The number of the optimisable λ variables will be denoted throughout the text as $\tilde{N}_\lambda$.

Once $\vec{\lambda}_{opt}(\tilde{N}_\lambda)$ corresponding to a particular number of optimisable values $\tilde{N}_\lambda$ is estimated, the final step is to optimise $\tilde{N}_\lambda$. In this work, this will be done using discrete brute-force optimisation, where $\vec{\lambda}_{opt}(\tilde{N}_\lambda)$ is first calculated at each of $\min[0, \tilde{N}_\lambda - 1]$, $\tilde{N}_\lambda$ and $\tilde{N}_\lambda + 1$ dimensions, using the procedure described above. Afterwards, $\hat{\tau}_{round,pred}(\vec{\lambda}_{opt}(\tilde{N}_\lambda))$ is evaluated at each of these dimensions until a local minimum is found, in which case the minimisation procedure terminates. In all cases, the initial protocol guess $\vec{\lambda}_0(\tilde{N}_\lambda)$ will either be interpolated in transformed space from the previous minimisation result with the closest (preferably larger) number of dimensions, or taken from the initial AASMC run.

Even though the procedure outlined above is theoretically exact and likely to return a globally optimal protocol at infinite sampling, there are still some practical considerations in order for this method to be viable in practice. These will be described below.

9.2.4 Interpolation

The functional form of the alchemical interpolation scheme between the two endpoint distributions $\pi(0, \vec{x})$ and $\pi(1, \vec{x})$ does not theoretically influence the sampling at the distribution of interest. Practically, however, the interpolation procedure needs to be carefully chosen to ensure good phase space overlap between all intermediate λ windows. To achieve this, soft-core potentials are commonly used for interpolation of Lennard-Jones (LJ) interactions in place of simple linear decoupling. While this results in more efficient sampling, energy evaluations at each λ window have to be performed using the whole Hamiltonian and cannot be simply interpolated from the energies at the endpoints. On the other hand, the MBAR estimator uses the evaluated energy of each sample at each previously sampled λ value and each iteration of the above protocol optimisation algorithm requires energy evaluations at arbitrary λ values chosen by the minimiser as well. Therefore, one would incur unfeasibly high computational expense if full Hamiltonian evaluation was performed throughout these algorithms.

In this chapter, two different approaches will be compared. The first is always exact and uses a recently developed linearly interpolatable soft-core potential,[362] where only three expensive energy evaluations are needed per sample, which can be afterwards stored in memory and readily interpolated as needed. The second approach utilises a commonly used soft-core potential,[44] where expensive energy evaluations will be made at each λ value of the protocol used to generate this structure. These will then be interpolated linearly in order to approximate intermediate energies for the MBAR

estimator and protocol minimiser. Since protocol convergence is asymptotically guaranteed (discussed in Section 9.2.7), the energy interpolation error resulting from this procedure at the protocol λ values will always tend to zero. This also means that the fraction of samples at the past suboptimal λ values diminishes over time, meaning that the second approach also results in asymptotically exact free energy values of the converged λ protocol, despite being approximate at finite sampling. However, the energy and free energy errors at all other λ values will remain finite, meaning that the protocol optimisation procedure is always approximate in this setting and is therefore not guaranteed to converge to the true asymptotically optimal protocol.

9.2.5 Improving MBAR Estimation

Although MBAR is an asymptotically optimal estimator, it is known to produce biased estimates at finite sampling.[363] These can then adversely influence the adaptation process, resulting in local trapping of the free energy and/or protocol estimates and therefore highly suboptimal efficiency. To help tackle this problem, one can use an ensemble of MBAR estimators, using bootstrap aggregation ("bagging")—a technique often used in machine learning applications to increase the robustness of the estimation.[364] In the current setting, the process consists of simply bootstrapping all available trajectory frames $N_{bootstrap}$ times and fitting an MBAR estimator to each bootstrapped dataset, resulting in $N_{bootstrap}$ different, but equally valid, estimators. One can then use the average of the predictions from these models to obtain any observable of interest. In this work, bagging will be used when calculating the Metropolis acceptance criterion and the transition matrix, both of which are dependent on all estimated free energy values $\vec{f}$ provided by the corresponding MBAR model:

$$\begin{aligned} \hat{p}_{acc}(\lambda_j, \sigma_j, \hat{f}_j | \lambda_i, \sigma_i, \hat{f}_i, \vec{x}) &\mapsto \left\langle \hat{p}_{acc}(\lambda_j, \sigma_j, \hat{f}_j | \lambda_i, \sigma_i, \hat{f}_i, \vec{x}) \right\rangle_{bootstrap} \\ \widehat{\mathbf{T}}(\vec{\lambda}, \vec{\hat{f}}) &\mapsto \left\langle \widehat{\mathbf{T}}(\vec{\lambda}, \vec{\hat{f}}) \right\rangle_{bootstrap} \end{aligned} \tag{9.10}$$

In order for bootstrapping to generate a correct distribution of the estimators of interest, one needs to supply it with decorrelated samples. Although, as previously discussed, the instant decorrelation assumption is often sufficiently satisfied in practice for timescales on the order of 1–10 ps, it is still essential to obtain a reliable estimate of the effective decorrelation time τ_{decorr} for more challenging systems. While a method for estimating τ_{decorr} has been previously published,[163] its applicability to FAST is limited due to the constant changes in the λ protocol, meaning that an alternative approach needs to be taken.

To obtain an estimate of τ_{decorr}, we first note that the round trip time $\hat{\tau}_{round,pred}(\vec{\lambda})$ predicted by the transition matrix $\widehat{\mathbf{T}}(\vec{\lambda})$ is directly proportional to the sampling time τ_{sample} between λ proposals (Equation 9.9), where it is assumed that τ_{sample} provides

complete decorrelation at each λ value. However, since the transition matrix $\widehat{\mathbf{T}}(\vec{\lambda})$ is independent of τ_{sample}, one can also regard the true observed round trip time $\hat{\tau}_{round,true}$ as being predicted by the same transition matrix $\widehat{\mathbf{T}}(\vec{\lambda})$, with the only difference being the effective sampling time between λ proposals. We now propose that this effective sampling time be equal to an effective decorrelation time τ_{decorr}, which can be estimated using the following equation:

$$\hat{\tau}_{decorr} = \frac{\hat{\tau}_{round,true}}{\hat{\tau}_{round,pred}(\vec{\lambda})}\tau_{sample} \tag{9.11}$$

It should be noted that the ratio $\frac{\hat{\tau}_{round,true}}{\hat{\tau}_{round,pred}(\vec{\lambda})}$ may not necessarily be independent of τ_{sample} in practice. Therefore, $\hat{\tau}_{decorr}$ is best viewed not as a physical autocorrelation time, but rather as an effective deviation from the instantly decorrelated transition matrix model. In this chapter, the rounded dimensionless $N_{decorr} \equiv \max[1, \lfloor \frac{\hat{\tau}_{decorr}}{\tau_{sample}} \rceil]$ will be used to remove the correlated samples. This will be achieved by starting from an initial sample pseudo-randomly chosen from the most recent N_{decorr} samples and then keeping only every N_{decorr}-th previous sample. The resulting effective number of samples $N_{samples,eff}$ will afterwards be bootstrapped and used for MBAR estimation. Finally, if no round trips have yet been observed, $\hat{\tau}_{decorr}$ will be estimated from the expected transition time of the longest transition so far observed, instead of the round-trip time.

9.2.6 Computational Footprint of FAST

Since the computational power required to handle both the estimation of the transition matrix and the free energies scales linearly with respect to the total number of samples $N_{samples}$, performing these calculations at a fixed frequency will result in a total computational cost of $O(N_{samples}^2)$. To alleviate this, these calculations will in practice be performed at an exponentially diminishing frequency, reducing the complexity to $O(N_{samples} \log N_{samples})$. If the implementation is parallelised, written in a compiled language, and/or run in the background on the central processing unit (CPU) while the MD simulation is run on the graphics processing unit (GPU), the computational overhead from adaptation can become negligible with suitably chosen frequency parameters. In this work, the free energies will be calculated every $\lfloor 1 + 0.01 N_{samples,eff} \rceil$ steps, while protocol optimisation will be performed every $\lfloor 100 + 0.1 N_{samples,eff} \rceil$ steps, meaning that the number of steps between subsequent optimisations increases linearly with respect to the effective number of samples in both cases.

On the other hand, the memory consumption of the energy matrices required for the MBAR calculations always increases linearly over time, meaning that depending on

the system and the simulation length one might run into memory limitations. In this work, however, matrices with $\sim 1.6 \times 10^5$ samples (over 160 ns) were routinely handled with memory usage of less than one gigabyte, which is well within the capability of an average computer. Therefore, these considerations are reserved for more computationally intensive cases, where memory requirements could potentially be alleviated by limiting the number of samples used for adaptation, using stochastic approximations of MBAR,[365–367] and/or offloading the matrices to the hard drive using specialised libraries.[368]

The CPU and memory requirements of the free energy estimation procedure are not only dependent on $N_{samples}$, but also on the total λ value history, which also grows over time (linearly or logarithmically, depending on the adaptation frequency). This means that exploring a space of continuous (or very high-precision) λ values will likely result in unfeasibly high computational requirements and in this chapter all λ values will be preliminarily rounded to two decimal places, meaning that only a maximum of 101 λ values can be present in the energy matrix. In most practical cases, it is expected that two to three significant figures are completely sufficient for achieving near-optimal performance, while using a relatively low amount of memory.

9.2.7 Convergence

The algorithm described above is highly adaptive and includes the following non-Markovian steps:

- Observable estimation from the MBAR model (free energies and acceptance rates)
- Protocol minimisation
- Estimation of τ_{decorr}

Using general results from the literature, it can be shown that the algorithm is asymptotically convergent, despite the multiple layers of adaptation. To demonstrate this, we first state the two sufficient conditions for asymptotic convergence: containment and diminishing adaptation.[369] The former condition means that if adaptation is stopped at any point, the convergence to the corresponding stationary distribution is guaranteed. This condition is readily satisfied, since all of the above procedures produce finite quantities and non-adaptive IST is still an ergodic sampling algorithm which satisfies skew detailed balance (Equation 9.2), even if suboptimal weights and/or λ values are used. The second condition can be trivially enforced by using all generated samples for the adaptation, since the variance of any sample-dependent quantity diminishes at infinite sampling. If $\hat{\tau}_{decorr}$ is used to remove correlated samples, it also needs to converge to a finite value in order for the above

Algorithm 2 FAST

Input
  $\{\vec{x}_{history}\}$ initial set of all previously sampled system coordinates and λ values
  $\vec{\lambda}$ initial λ protocol
  N_{iter} number of FAST iterations
Output
  $\{\vec{x}_{history}\}$ the final set of all system coordinates and λ values
procedure FAST($\{\vec{x}_{history}\}$, $\vec{\lambda}$, N_{iter})
  $t_{MBAR} = 0$ ▷ Number of iterations before MBAR update
  $t_{opt} = 100$ ▷ Number of iterations before protocol update
  $i = 0$
  while $i < N_{iter}$ **do**
    if $t_{MBAR} >= i$ **then**
      $\vec{f}$, $model$, $t_{MBAR} \leftarrow$ MBAR($\{\vec{x}_{history}\}$) ▷ as in Section 9.2.2
    if $t_{opt} >= i$ **then**
      $\vec{\lambda}$, $t_{opt} \leftarrow$ OptimiseProtocol($model$, $\vec{\lambda}$, $\{\vec{x}_{history}\}$) ▷ as in Section 9.2.3
    $\vec{x} \leftarrow$ Sample($\vec{x}$, λ_{curr}, τ_{sample}) ▷ τ_{sample} is the sampling time
    $\{\vec{x}_{history}\} = \{\vec{x}_{history}\} \cup \{(\vec{x}, \lambda_{curr})\}$
    λ_{curr}, $\sigma_{curr} \leftarrow$ AttemptISTMove(λ_{curr}, σ_{curr}, $\vec{\lambda}$, $\vec{f}$) ▷ as in Section 9.2.1
    $i = i + 1$
return $\{\vec{x}_{history}\}$

assertion to hold. As can be seen from Equation 9.11, this convergence is guaranteed as long as the expected round-trip time is finite—a condition which is also satisfied for the IST Markov chain considered hereafter due to its ergodicity (Figure 9.1b).

9.2.8 Summary of the Method

The full FAST algorithm is shown in Algorithm 2 and Figure 9.2. The only required input is a set of samples generated at a range of λ values with sufficiently good overlap. These can be readily generated by an adaptive SMC algorithm, such the AASMC algorithm presented in Chapter 8. Although AASMC requires a number of input parameters, they are all system-independent and only affect the efficiency of the initial stages of the simulation, since FAST eventually converges to an asymptotically optimal protocol regardless of the initial input. In addition, all of the parameters required by the FAST algorithm are related to free energy and/or protocol estimation frequency and quality. Therefore, apart from the degrees of freedom to enhance, all input given to FAST is effectively system-independent, making FAST a nearly black-box enhanced sampling method.

In the following discussion, we will validate FAST on a range of protein-bound and solvated ligand systems, where we will enhance the motions of certain torsional

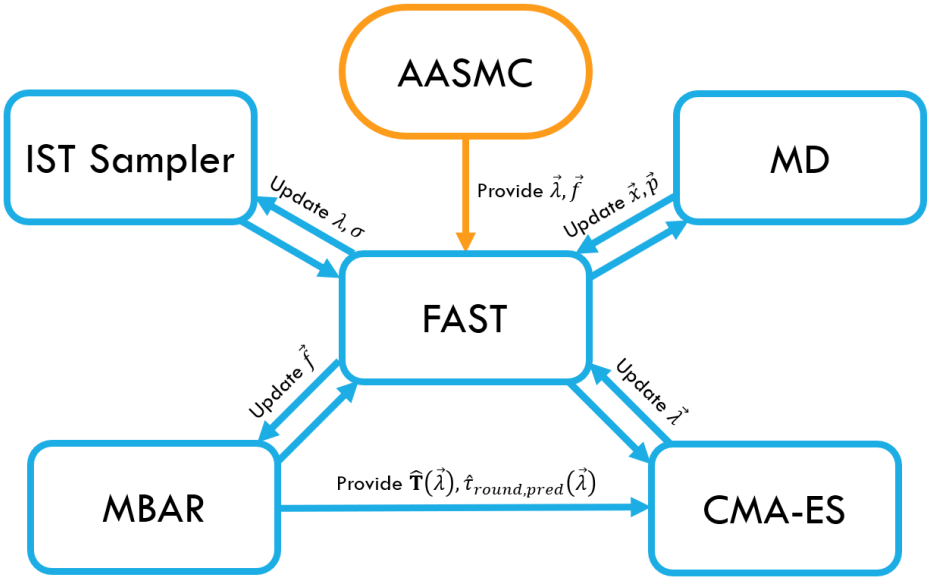


FIGURE 9.2: A summary of the FAST workflow.

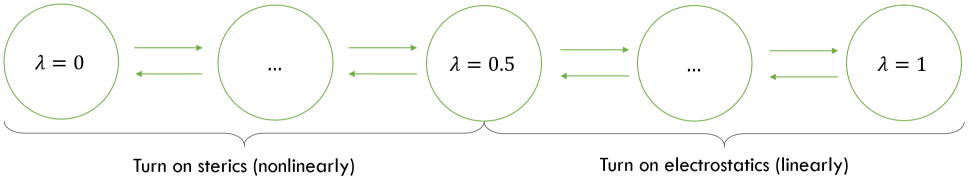
degrees of freedom by decoupling one side of the rotatable bond of interest from the rest of the system at $\lambda = 0$, analogous to the approach taken in Chapter 8. In all cases, the same AASMC and FAST parameters will be used, to demonstrate the wide applicability of the method with minimal prior knowledge.

9.3 Methods

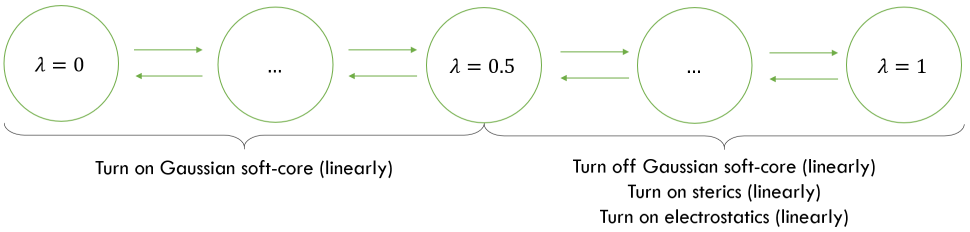
9.3.1 System Setup and Simulation

The FAST protocol was validated on four of the systems considered in Chapter 8. All of these following systems were initially protonated using PDB2PQR[185] without any additional pK_a calculations. The ff14SB[25] force field was used for the proteins, while GAFF2[27] with AM1-BCC[29,30] charges was used for the ligands. All systems were solvated in a TIP3P[31] periodic cubic box with a length of 4 nm (solvated) or 7 nm (bound). Sodium chloride counterions with TIP3P–compatible parameters were also added for neutralisation up to a concentration of ∼0.154 M.

All FAST simulations were performed using OpenMM 7.4.2,[233] OpenMMTools[339] 0.19.0 and OpenMMSLICER 2.0.0. All systems were initially minimised and afterwards equilibrated at $\lambda = 0$ (decoupled state) for 100 ps. During the equilibration, harmonic restraints were used for the protein backbone with a force constant of 5 kcal/mol/Å^2. The equilibrated structure was used as a starting point for the AASMC process, which was in turn used to obtain an initial λ schedule and free



(A) CSC decoupling scheme



(B) GSC decoupling scheme

FIGURE 9.3: The main stages of the classical soft-core potential (CSC) (Figure 9.3a) and the Gaussian soft-core potential (GSC) alchemical schemes (Figure 9.3b).

energy estimates for FAST. Each step in the AASMC algorithm is described in detail in Chapter 8.

50 walkers were spawned from the equilibrated structure at the beginning of the AASMC procedure. 100 conformers were generated per walker at $\lambda = 0$ and a decorrelation time of 1 ps was used for each walker before reweighting and resampling. The expected sample size estimator based on the Metropolis–Hastings acceptance criterion $\hat{R}_{MH} = 0.5$ was used for each AASMC iteration to determine the next λ value in the sequence:

$$\hat{R}_{MH} = \frac{1}{N}\sum_{i}^{N}\min[Nw_i, 1] \tag{9.12}$$

where the sum is over all N walkers and w_i is the corresponding weight of each walker. The reason for choosing $\hat{R}_{MH}$ over other resampling metrics in this case was that a constant Metropolis–Hastings criterion expectation value is known to provide an optimal λ spacing in the case of reversible uniformly-weighted ST,[357] making it a good starting point for IST as well. Finally, systematic resampling was used during the resampling stage.[370]

After the final AASMC iteration, one walker was chosen at random and was subsequently used as a starting point for an irreversible simulated tempering procedure for a total of 160 ns. Movement in λ space was attempted every 1 ps. New samples for the MBAR[18] estimation were also drawn every 1 ps immediately before attempting a move in λ space. An ensemble of $\min[10, N_{decorr}]$ MBAR models using decorrelated bootstrapped samples was generated with diminishing frequency every

$\lfloor 1 + 0.01 N_{samples,eff} \rceil$ steps using a robust L-BFGS solving algorithm[20] implemented in SciPy/Numba.[343,371,372] Protocol optimisation was performed every $\lfloor 100 + 0.1 N_{samples,eff} \rceil$ steps using a CMA-ES[360] implementation in Python[373] using the default settings up to a maximum of $10\tilde{N}_\lambda$ evaluations and initial minimiser standard deviation $\sigma_0 = \frac{1}{\max[1, N_\lambda - 2]}$. During the optimisation of the number of lambda values N_λ, the initial guesses for the λ protocols of a particular length were either taken from previous optimisations, if available, or generated using linear interpolation in the transformed space from the λ protocol of the closest length. The resulting λ values were rounded to two decimal places.

Each simulation was run in sextuplicates with two different soft-core schedules. The first schedule used the classical soft-core potential (CSC)[44] for the Lennard-Jones (LJ) interactions with parameters $a = 1$, $b = 1$, $c = 6$ and $\alpha = 0.5$. In this case, all LJ interactions were introduced before subsequently turning on all electrostatic interactions (Figure 9.3a). The second set of simulations utilised the recently published Gaussian soft-core potential (GSC),[362] using the protocol and parameters described in the original publication (Figure 9.3b).

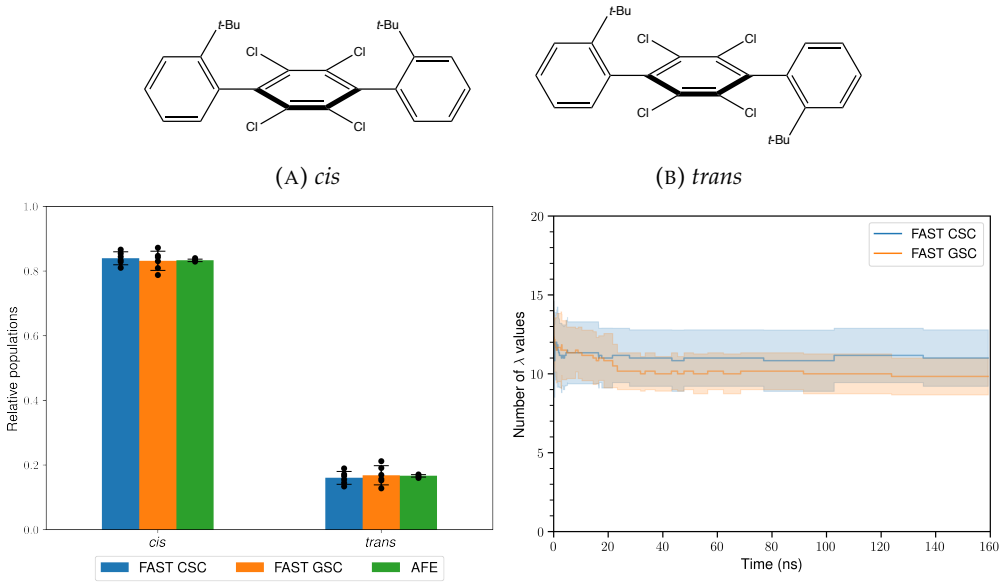
A BAOAB Langevin integrator[59] was used during all simulations at 298 K with a 2 fs timestep and a collision rate of 1 ps^{-1}. All water molecules were constrained using the SETTLE algorithm,[53] while all other bonds involving hydrogen were constrained with the SHAKE[50] and CCMA[52] algorithms. A Monte Carlo barostat was used for pressure control at 1 atm with rescaling attempts every 50 fs. A cut-off of 1.2 nm was used for all short-range nonbonded interactions. Long-range electrostatics were calculated with particle mesh Ewald (PME).[37]

The FAST simulations were then validated against Hamiltonian replica exchange (H-REMD) and alchemical free energy (AFE) calculations in GROMACS 2018.4 following the same protocol as described in Chapter 8.

9.3.2 Analysis

All of the following results and metrics have been reported in terms of their arithmetic averages and standard deviations. The only exception is the ratio of samples between $\lambda = 1$ and $\lambda = 0$, $\frac{N_1}{N_0}$, which has been used as a way to measure the deviation of the number of samples from the ideal $\frac{N_1}{N_0} = 1$. In this case, the geometric mean and standard deviation have been considered instead, in order to more faithfully represent the variability when $N_1 \gg N_0$ and $N_0 \gg N_1$. In all cases, these statistics have been reported as $mean \pm deviation$, although one needs to keep in mind that the geometric standard deviation is multiplicative and the $\pm$ sign has only been used for consistency.

Clustering of the resulting conformer populations was performed as described in Chapter 8. The only exception is the ligand common core clustering analysis



(A) *cis* (B) *trans*

(C) Relative rotamer populations (D) Number of λ values over time

FIGURE 9.4: The two terphenyl derivative rotamers (Figures 9.4a and 9.4b), the mean relative populations of these states obtained using FAST, H-REMD and AFE calculations after 6 runs (Figure 9.4c) and the average number of λ values over time (Figure 9.4d). The error bars represent one standard sample deviation.

considered in TGF-β, where each FAST trajectory was pre-aligned to its initial coordinates based on the protein backbone α carbon atoms, before extracting the centre of geometry of the ten heavy triazanaphthalene ring atoms of the ligand with MDTraj[236] and MDAnalysis.[341,342] Agglomerative clustering with default settings in scikit-learn[238] was then performed on these geometric centres using every tenth frame of the original trajectory, resulting in two clusters. These were then used to train a k-nearest neighbours classifier,[374] which was used to extend the clustered data over the whole trajectory. This approximation was needed because of the large memory requirements of the agglomerative clustering algorithm if the whole set of trajectory frames had been used.

Any other aspects of the analysis follow the same methods as described in Chapter 8.

9.4 Results

9.4.1 Terphenyl in Water

A relatively simple system with an insurmountable kinetic barrier is the terphenyl derivative shown in Figure 9.4, making it a good test case for alchemical methods. To sample both conformers, one of the 2-tert-butylphenyl groups was completely decoupled at $\lambda = 0$, as performed in Chapter 8. This large alchemical change makes this system one of the more challenging test cases involving a solvated system.

System	Potential	Replicate 1	2	3	4	5	6	Total Average	St. dev.
Terphenyl	CSC	2.31	2.59	2.44	2.51	2.49	2.51	2.48	0.09
	GSC	2.36	2.63	2.57	2.42	2.50	2.62	2.52	0.11
T4-lysozyme	CSC	5.89	6.38	6.59	6.45	5.97	6.09	6.23	0.28
	GSC	4.14	3.61	4.03	2.43	3.81	3.96	3.66	0.63
PTP1B	CSC	1.64	2.23	2.34	2.63	2.31	2.56	2.29	0.35
	GSC	2.49	1.89	2.21	2.11	2.62	2.12	2.24	0.27
TGF-β	CSC	0.04	0.08	0.06	0.22	0.11	0.03	0.09	0.07
	GSC	0.04	0.06	0.02	0.01	0.02	0.05	0.03	0.02

TABLE 9.1: The number of round trips per nanosecond for each of the systems across different replicates and soft-core potentials. The averages and the corresponding standard deviations are also given.

System	Potential	Replicate 1	2	3	4	5	6	Total Average	St. dev.
Terphenyl	CSC	0.93	1.11	2.55	1.92	1.66	2.20	1.63	1.48
	GSC	0.86	0.95	0.89	1.03	1.13	1.01	0.97	1.11
T4-lysozyme	CSC	2.13	1.73	1.06	1.72	1.58	1.52	1.59	1.26
	GSC	1.19	0.91	0.74	1.35	0.75	0.79	0.93	1.29
PTP1B	CSC	2.38	0.83	1.22	1.36	0.93	2.11	1.36	1.53
	GSC	0.65	0.90	0.96	0.79	0.94	1.48	0.92	1.32
TGF-β	CSC	0.24	9.21	56.77	4.94	3.16	7.16	4.91	5.94
	GSC	195.14	4.50	15.60	0.18	276.28	0.12	6.59	27.84

TABLE 9.2: The fraction of total samples between $\lambda = 1$ and $\lambda = 0$ for each of the systems across different replicates and soft-core potentials. The geometric averages and the corresponding geometric standard deviations are also given.

System	Potential	Replicate 1	2	3	4	5	6	Total Average	St. dev.
Terphenyl	CSC	1.38	1.34	1.40	1.31	1.40	1.38	1.37	0.04
	GSC	1.57	1.47	1.40	1.46	1.52	1.39	1.47	0.07
T4-lysozyme	CSC	1.64	1.64	1.56	1.59	1.70	1.63	1.63	0.05
	GSC	2.08	2.23	2.13	3.15	2.17	2.13	2.31	0.28
PTP1B	CSC	2.78	2.15	1.96	1.85	2.03	1.80	2.10	0.36
	GSC	2.05	2.37	2.13	2.22	1.83	2.36	2.24	0.27
TGF-β	CSC	80.51	36.80	47.29	13.36	27.42	92.78	49.69	24.63
	GSC	78.06	46.94	161.32	249.92	173.82	60.98	128.51	66.51

TABLE 9.3: The final measured $\hat{\tau}_{decorr}$ in ps for each of the systems across different replicates and soft-core potentials. The averages and the corresponding standard deviations are also given.

The relative rotamer populations, shown in Figure 9.4c, are well-converged for both the CSC and the GSC soft-core potentials, with the dominant conformer being the *cis* state at 84% ± 2% and 83% ± 3%, respectively. These results are in excellent agreement with the populations obtained from AFE calculations (83% ± 0%), indicating that both FAST protocols are able to handle this system without any issues.

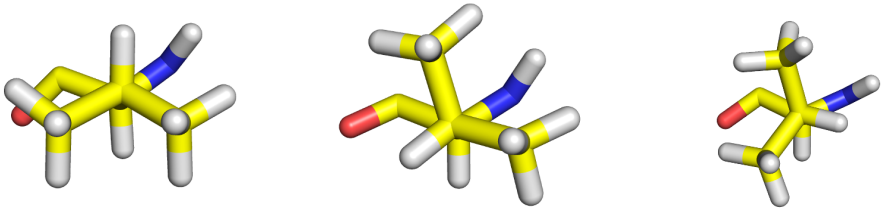
As shown in Table 9.1, both the CSC and the GSC protocols result in a similar number of round trips—∼2.5 per nanosecond. This is reflected by the final observed protocol lengths, with an average of 11 ± 2 total λ values for CSC and 10 ± 1 for GSC. These final protocols are approximately 4 λ windows shorter on average than the protocol obtained by the initial AASMC run (∼14 for both CSC and GSC). This not only showcases the increased efficiency of sampling in λ space but also demonstrates the independence of this method on the initial protocol generated by AASMC.

These short final protocol lengths present a somewhat surprising result, because one would expect to need a higher number of intermediate windows for an alchemical perturbation of this size. Indeed, many reported free energy protocols use a higher number of intermediate λ windows for arguably simpler alchemical changes.[140,162,169] The advantage of using FAST over conventional wisdom is therefore not only the increased robustness and reproducibility of the method compared to manual tuning, but also the increased relative amount of sampling time at $\lambda = 1$.

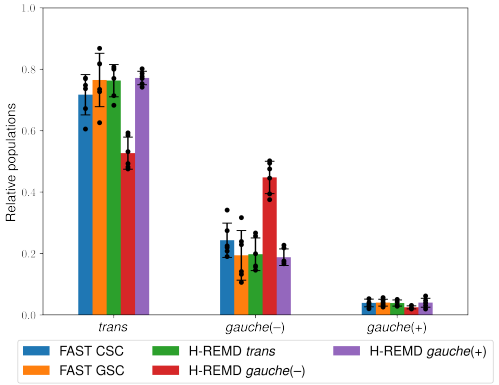
The main potential weakness of ST-based methods is the non-uniform sampling in λ space. In this chapter we will consider the sampling ratio between $\lambda = 1$ and $\lambda = 0$, $\frac{N_1}{N_0}$, and the final relative effective decorrelation time, $\hat{\tau}_{decorr}$ (Equation 9.11), to gauge the sampling reliability of FAST. As shown in Table 9.2, the CSC protocol results in a less uniform sampling ratio between the two terminal λ values with $\frac{N_1}{N_0} = 1.63 \pm 1.48$, compared to $\frac{N_1}{N_0} = 0.97 \pm 1.11$ for GSC. $\hat{\tau}_{decorr}$ is on the other hand comparable between both potentials, with an average value of 1.37 ± 0.04 ps for CSC and 1.47 ± 0.07 ps for GSC (Table 9.3). Nevertheless, both protocols result in satisfactory sampling ratios and effective decorrelation times, making FAST suitable for the conformational exploration of this system.

9.4.2 T4-lysozyme

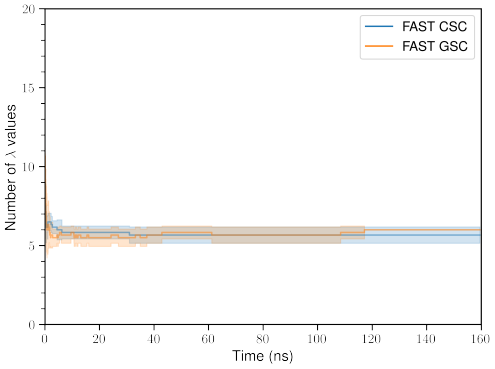
Another application of FAST is protein side-chain exploration. One such test case is the Val111 rotation in T4-lysozyme L99A with bound p-xylene (PDB ID: 187L[344]), which has been previously explored with other enhanced sampling methods[95,244,269] and was also investigated in Chapter 8. Although this is a relatively simple test case, it is a good way to compare the maximum efficiency of both soft-core potentials in this setting.



(A) *trans* (B) *gauche*(–) (C) *gauche*(+)



(D) Relative rotamer populations



(E) Number of λ values over time

FIGURE 9.5: The three T4-lysozyme Val111 rotamers (Figures 9.5a to 9.5c), the mean relative populations of these states obtained using FAST and H-REMD after 6 runs (Figure 9.5d) and the average number of λ values over time (Figure 9.5e). The error bars represent one standard deviation.

There are three characteristic conformers for Val111, shown in Figures 9.5a to 9.5c. These have previously proven difficult to sample with regular MD,[244] and therefore enhanced sampling methods are needed. In this setting we can achieve this sampling simply by completely alchemically decoupling the Val111 isopropyl group at $\lambda = 0$ with both FAST and H-REMD, as performed in Chapter 8.

As shown in Figure 9.5d, both the CSC and GSC protocols result in statistically equivalent populations after 160 ns of sampling, with the dominant conformer being the *trans* state at 72% $\pm$ 7% and 77% $\pm$ 9%, respectively. These results show that the CSC protocol is slightly better converged in this case, similarly to the previous system.

Although both FAST protocols result in variances that are apparently higher than the H-REMD ones, the initial H-REMD state can significantly affect the final populations even after 160 ns of cumulative sampling over 40 λ values. For instance, the *trans* conformer populations are 76% $\pm$ 5% and 53% $\pm$ 5%, if one starts from the *trans* and *gauche*(–) states, respectively. Therefore, H-REMD results in a significant bias towards the initial supplied conformer, meaning that there is insufficient decorrelation from the initial coordinates. This problem is in contrast not observed when FAST is used, since all of the sampling time is dedicated to a single replica. It follows that despite the lower apparent H-REMD variance, there is a higher bias in the resulting

population, and if all H-REMD simulations are considered together, their cumulative standard deviation becomes 12%, which is higher than either of the FAST protocols.

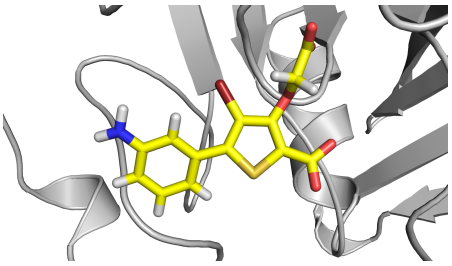
The resulting round trips per nanosecond are 6.23 ± 0.25 for CSC and 3.88 ± 0.38 for GSC, as shown in Table 9.1, suggesting that the CSC protocol is significantly more efficient compared to GSC. In comparison, the H-REMD protocol results in only 1.57 ± 0.08 round trips, indicating that the worse-performing FAST protocol is still more than twice as efficient as the unoptimised H-REMD protocol. This is evidenced by the low final number of λ values: $\sim$6 in both cases compared to 9 ± 0 and 11 ± 1 initial λ values for CSC and GSC, respectively (Figure 9.5e), again showing that the FAST procedure is largely independent of the initial protocol estimated by AASMC.

Similarly to the previous test case, the GSC protocol results in more uniform $\frac{N_1}{N_0}$ ratios with an average of 0.93 ± 1.29 compared to 1.59 ± 1.26 for CSC (Table 9.2). However, a higher $\hat{\tau}_{decorr}$ is observed for GSC: 2.31 ± 0.28 ps compared to 1.63 ± 0.05 ps for CSC (Table 9.3). It can be therefore concluded that FAST with CSC is more efficient for this test case than FAST with GSC, while H-REMD has a very low comparative efficiency to both of the FAST protocols.

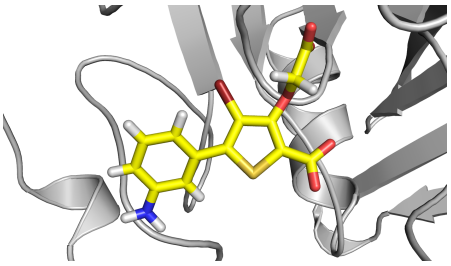
9.4.3 Protein Tyrosine Phosphatase 1B (PTP1B)

A practically important use case for enhanced sampling methods is bound ligand conformer sampling. One such test case is PTP1B bound to a thiophene derivative (PDB ID: 2QBS[3]). The rotation of the ligand 3-aminophenyl ring is a rare event whose exploration would be desirable in e.g. binding free energy calculations. This results in two alternative ligand conformers, shown in Figures 9.6a and 9.6b. Similarly to Chapter 8, here we achieve this exploration by completely turning off the 3-aminophenyl ring at $\lambda = 0$.

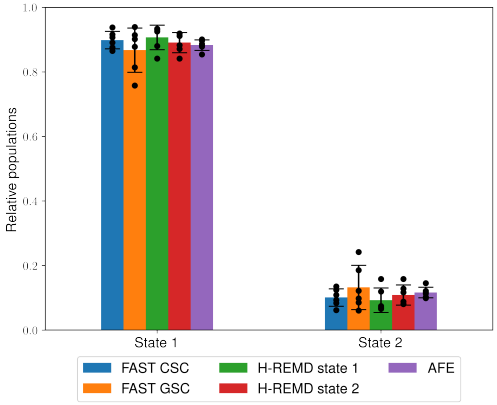
As in the previous test cases, the CSC and GSC protocols result in statistically equivalent populations (Figure 9.6c): $90\% \pm 3\% : 10\% \pm 3\%$ and $87\% \pm 7\% : 13\% \pm 7\%$ with higher variance observed for the GSC protocol. Similar populations are observed for H-REMD starting from state 1 and state 2: $91\% \pm 4\% : 9\% \pm 4\%$ and $89\% \pm 3\% : 11\% \pm 3\%$, respectively, showing that in this case H-REMD provides results of equivalent quality to the CSC protocol. All of these populations agree with the AFE calculations, which result in populations of $88\% \pm 2\% : 12\% \pm 2\%$. Similarly to the terphenyl test case, AFE results in lower variance compared to the FAST—an expected behaviour for a low number of conformers. As this number increases, however, AFE calculations become increasingly more impractical. Furthermore, obtaining conformational populations using AFE methods requires prior knowledge of the conformers of interest—knowledge, which is not required by FAST and H-REMD.



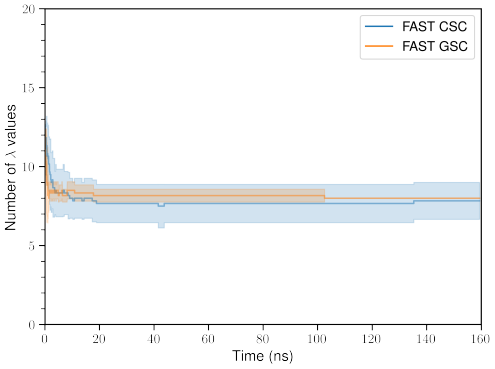
(A) State 1



(B) State 2



(C) Relative rotamer populations

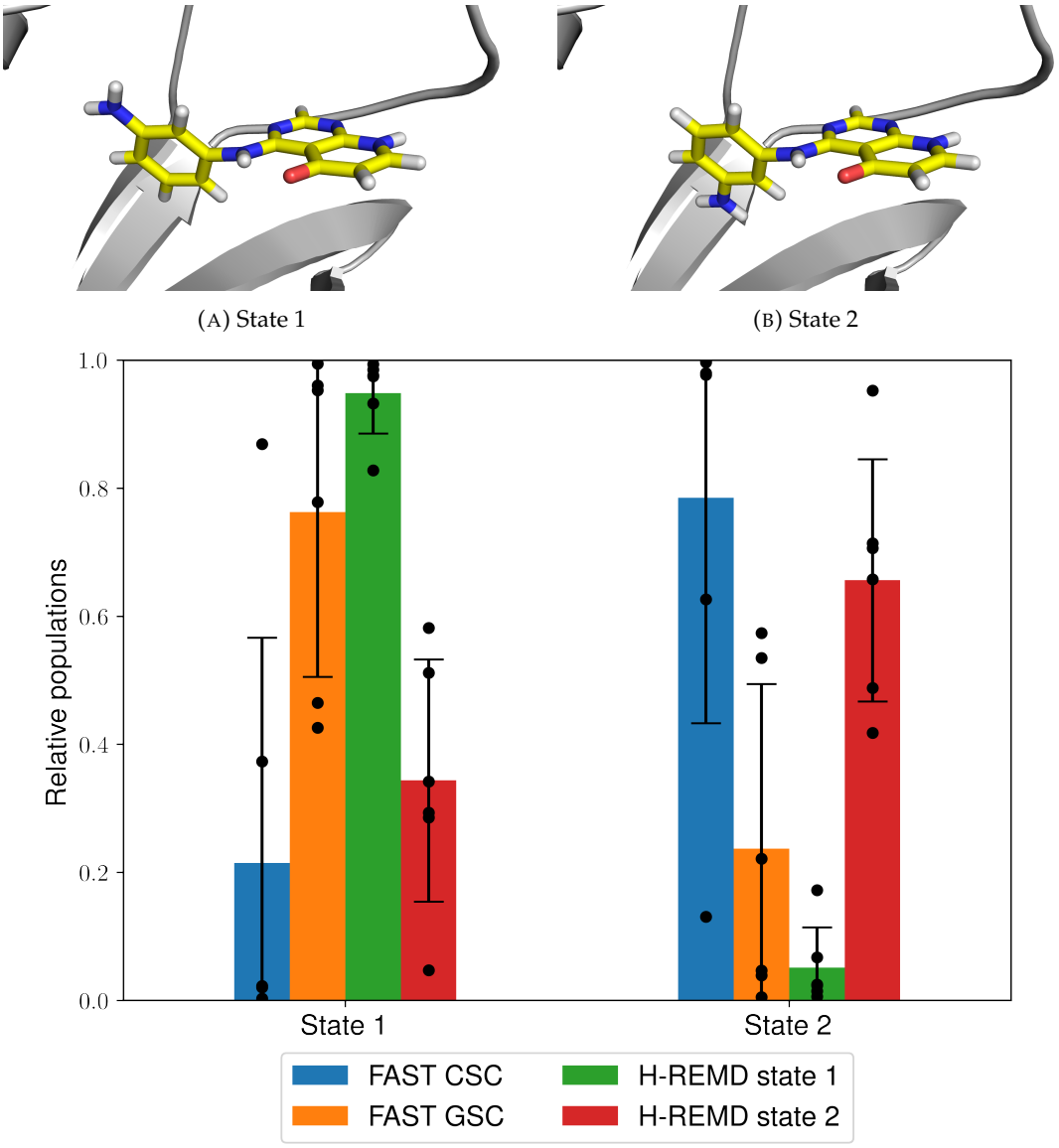


(D) Number of λ values over time

FIGURE 9.6: The two thiophene derivative rotamers bound to PTP1B (Figures 9.5a to 9.5c), the mean relative populations of these states obtained using FAST, H-REMD and AFE calculations after 6 runs (Figure 9.6c) and the average number of λ values over time (Figure 9.6d). The error bars represent one standard sample deviation.

Both FAST protocols result in a similar number of round trips per nanosecond: 2.29 ± 0.35 for CSC and 2.24 ± 0.27 for GSC (Table 9.1). These can be compared to 0.64 ± 0.08 for H-REMD, meaning that both FAST protocols result in a nearly fourfold increase in efficiency. This behaviour is again explained by both CSC and GSC resulting in an unexpectedly low total number of λ windows (Figure 9.6d): 8 in both cases. This presents a significant improvement over the initial 13 (CSC) and 12 (GSC) λ values obtained by AASMC and shows that decoupling a whole phenyl ring does not necessarily require a large number of intermediate steps, as long as the decoupling is performed optimally.

As in the previous systems, the $\frac{N_1}{N_0}$ ratio is less optimal for the CSC protocol, with a mean value of 1.36 ± 1.53, compared to 0.92 ± 1.32 for GSC (Table 9.2). The average $\hat{\tau}_{decorr}$ is statistically equivalent in both cases: 2.10 ± 0.36 ps for CSC versus 2.16 ± 0.21 ps for GSC. (Table 9.3). Therefore, both protocols have comparable performance for this system, with the CSC protocol resulting in lower population variance and the GSC protocol having more consistent $\frac{N_1}{N_0}$ ratios.



(A) State 1

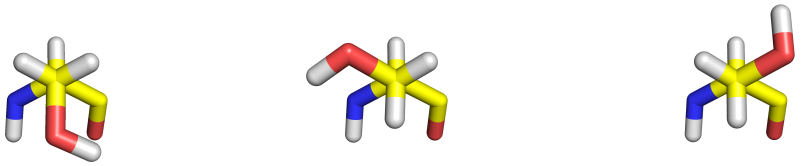
(B) State 2

(C) Ligand populations

FIGURE 9.7: The two TGF-β ligand rotamers (Figures 9.7a and 9.7b) and the mean relative populations of these states obtained using the FAST and H-REMD after 6 runs (Figure 9.7c). The error bars represent one standard sample deviation.

9.4.4 Transforming Growth Factor Beta (TGF-β)

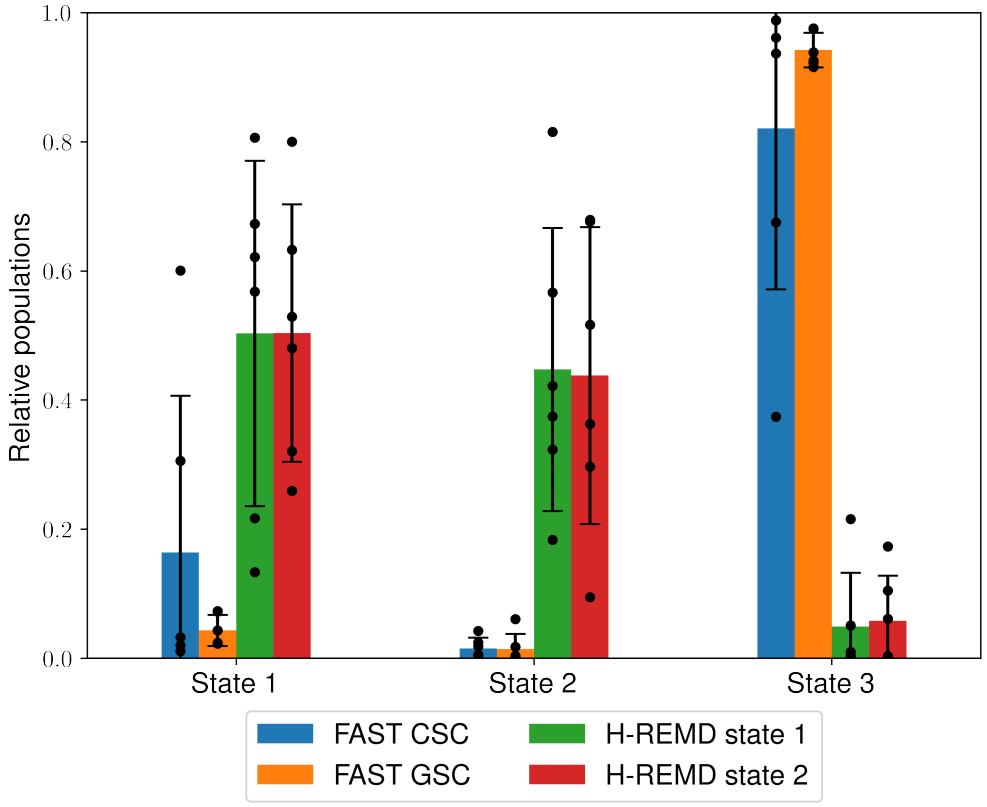
The final test case combines the exploration of a torsional degree of freedom of a ligand bound to transforming growth factor beta (TGF-β) and the nearby Ser82 rotamers. It is experimentally known (PDB ID: 4X2J[346]) that the 4-aminophenyl group of the ligand occupies two alternative states with approximately equal occupancy (Figures 9.7a and 9.7b) and that the Ser82 group has two alternative conformations (Figures 9.8a and 9.8b). However, a related PDB structure of a more symmetric 3-aminophenyl ligand derivative (PDB ID: 4X2G[346]) was instead used to keep the procedure consistent with the results in Chapter 8. As in Chapter 8, sampling



(A) State 1

(B) State 2

(C) State 3



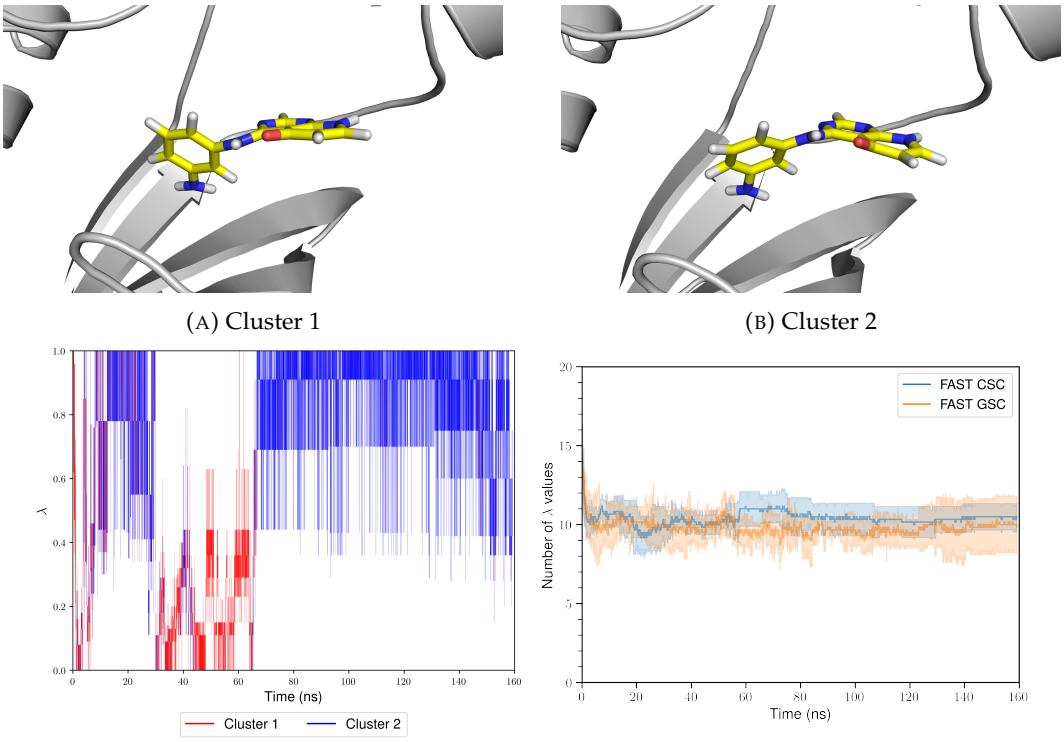
(D) Ser82 populations

FIGURE 9.8: The three TGF-β Ser82 rotamers (Figures 9.7a and 9.7b) and the mean relative populations of these states obtained using FAST and H-REMD after 6 runs (Figure 9.7c). The error bars represent one standard sample deviation.

enhancement was achieved by decoupling both the 3-aminophenyl ligand group and the Ser82 hydroxymethyl group at $\lambda = 0$.

Both FAST protocols result in highly variable ligand populations (Figure 9.7c), with state 1 being occupied at 21% $\pm$ 35% using the CSC protocol and at 76% $\pm$ 26% using the GSC protocol. This high uncertainty is partially observed in the H-REMD runs, where the simulations starting from state 1 stayed in it 95% $\pm$ 6% of the time, compared to 34% $\pm$ 19% if state 2 is used as an intial state. Similarly to the T4-lysozyme test case, we can see that initial structure biasing is an issue when H-REMD is used and the cumulative variance of all H-REMD results over both initial conformers is 34%, which is comparable to the CSC results.

The Ser82 populations are better converged for GSC than CSC (Figure 9.8d), with state



(A) Cluster 1

(B) Cluster 2

(C) TGF-β ligand cluster populations

(D) Number of λ values over time

FIGURE 9.9: The two TGF-β ligand clusters (Figures 9.9a and 9.9b), their transitions over time (Figure 9.9c) and the average number of λ values over time (Figure 9.9d).

3 being occupied at $94 \pm 3\%$ and $82\% \pm 25\%$, respectively. Although these results are consistent between both protocols, they differ significantly from the H-REMD simulations, where a higher variance is observed when both ligand state 1 and state 2 are simulated with Ser82 state 1.

These insufficiently converged results can be related to the low number of round trips: only 0.09 ± 0.07 per nanosecond using the CSC protocol compared to 0.03 ± 0.02 with the GSC protocol (Table 9.1). The H-REMD simulations result in similarly low efficiency with 0.05 ± 0.03 round trips per nanosecond. Since the effective correlation time was estimated to be 50 ± 25 ps (CSC) and 129 ± 67 ps (GSC), as shown in (Table 9.3), it is clear that the instantaneous decorrelation assumption breaks down in this case and kinetic barriers in λ space decrease the efficiency of the simulations.

Clustering analysis of the ligand triazanaphthalene ring centre of geometry relative to the initial structure reveals the presence of two main ligand clusters, shown in Figures 9.9a and 9.9b. When plotted over time for one of the FAST CSC simulations (Figure 9.9c), it is revealed that the trappings in λ space are correlated with the observed cluster: cluster 1 is more favourable at the lower λ values, while cluster 2 is preferred at the fully coupled λ values. This behaviour readily explains the low number of round trips, and it can be concluded a slow orthogonal rare event indeed limits the mobility in λ space and results in a high effective decorrelation time.

Unfortunately, determining orthogonal rare events *a priori* is not a straightforward task, meaning that systems exhibiting such behaviour are likely to prove challenging for FAST, since the low number of round trips inevitably implies low uniformity of sampling. This is evidenced by the highly inconsistent $\frac{N_1}{N_0}$ ratios at 4.91 ± 5.94 for CSC and 6.59 ± 27.84 for GSC (Table 9.2). Nevertheless, the optimisation procedure still results in a relative increase of efficiency, where the initial AASMC λ values are consistently decreased from $\sim$16 to 11 for CSC and 10 for GSC (Figure 9.9d). However, this increase is completely overshadowed by the slow binding mode change.

9.5 Discussion

The above results show that FAST is an efficient general-purpose enhanced sampling method of specific internal degrees of freedom, which readily extends the functionality of AASMC to longer timescales. In all of the above test cases, FAST significantly decreases the size of the initial alchemical protocols provided by AASMC, resulting in a higher proportion of samples being drawn from $\lambda = 1$. This results in a better reproducibility of the method, since manual protocol tuning is not required.

However, the choice of the functional form of the interpolation procedure is still a factor which can impact reproducibility. We have shown that although both the CSC and GSC protocols result in similar populations, they exhibit different efficiencies, with GSC consistently resulting in higher dihedral population variance. Moreover, the T4-lysozyme test case demonstrates that GSC can result in a significantly lower round trip rate than CSC. The reason for this is likely suboptimal long-range phase space overlap, which can be explained by the fact that the GSC potential does not always accurately reproduce the real LJ potential, resulting in higher kinetic barriers in λ space. This is also evidenced by the PTP1B test case, where the GSC protocol produced populations with higher variance than the H-REMD protocol, even though the number of round trips in the former setting was almost four times higher than the latter.

Interestingly, however, GSC consistently produces more optimal $\frac{N_1}{N_0}$ ratios (i.e. closer to unity) with lower standard deviations than CSC, where the latter consistently produced samples more highly biased towards $\lambda = 1$. This is a surprising result, since with infinite sampling and converged free energy values one would expect these ratios to approach unity and there is no obvious reason why the sampling should be biased in one direction in favour of another. Nevertheless, it appears that GSC is more reliable in this regard, presumably due to its often smoother free energy profiles.[362]

It has also been demonstrated τ_{decorr} appears to be a very useful metric for determining unexpected kinetic barriers. For instance, it expectedly produces values close to 1 ps in the solvated terphenyl test case, meaning that there is low effective

correlation in a homogeneous environment. On the other hand, τ_{decorr} is extremely high in the case of TGF-β, which immediately hints at orthogonal slow degrees of freedom which impact the sampling negatively. Therefore, τ_{decorr} can be monitored in real time to gauge the performance of the FAST sampler if needed.

TGF-β is a particularly interesting test case, since it results in a significantly higher variance between different repeats compared to the other systems. Even though the nature of the transformation is similar to the other test cases, a substantial increase in τ_{decorr} indicates that local exploration of phase space and λ space is not as efficient as in the other test cases. As shown in the previous section, this is readily explained by the several alternative binding modes the ligand adopts throughout the simulations. Some of these modes are favourable only in a particular range of λ windows, resulting in significant kinetic trapping and drastic decrease in sampling efficiency. Moreover, any kinetic trapping due to binding modes away from $\lambda = 1$ indicates that these new modes are not physically relevant and only decrease sampling efficiency to no benefit. This demonstrates the undesirable impact of alchemical decoupling on sampling—it can significantly affect the relative populations in an unexpected way.

Despite the high robustness of FAST on a range of systems, the above test cases show the main weaknesses of the method: unexpected kinetic barriers in λ space, as well as slow orthogonal degrees of freedom can significantly affect the sampling efficiency. However, this is a problem which is not unique to FAST, but is more generally relevant to all alchemical/tempering methods using a family of intermediate distributions. Since the slow degrees of freedom are not always known in advance, it will be therefore useful to develop a more general framework to improve long-range phase space overlap either by optimising the functional form of the soft-core potential, or by using e.g. restraint potentials which can help smooth the kinetic barriers in λ space. Future work addressing these can therefore help alleviate suboptimal effective decorrelation times.

9.6 Conclusion

A fully adaptive version of irreversible simulated tempering has been presented (FAST), where the intermediate distribution protocol is adaptively optimised in real time alongside the relative weights of the distributions. Validation on a variety of systems containing small molecules shows that this method is highly efficient and requires little prior knowledge.

We have also compared two soft-core interpolation methods: classical soft-core potential (CSC) and Gaussian soft-core potential (GSC). In all of the test cases CSC resulted in lower final distribution variance. Moreover, CSC exhibited higher round trip rates in most cases, as well as lower effective decorrelation times. Nevertheless,

more consistent sampling across λ values was observed for GSC, while CSC consistently produced more samples at the distribution of interest ($\lambda = 1$) during the 160 ns of total simulation time.

Unforeseen kinetic barriers in the alchemical/temperature space and phase space have been shown to be the main weakness of FAST, and, more generally, alchemical/tempering methods. These have been observed in a protein-ligand system with a slow binding mode transition (TGF-β). Future research will need to improve the robustness of FAST towards orthogonal slow modes.

We now proceed to Chapter 10, where we will extend FAST to binding free energy calculations with enhanced ligand sampling.

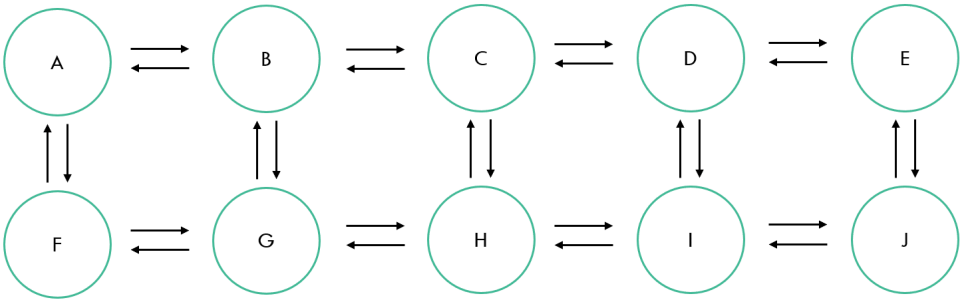
Chapter 10

Fully Automatable Relative Binding Free Energy Calculations with Enhanced Sampling using FAST/MBAR

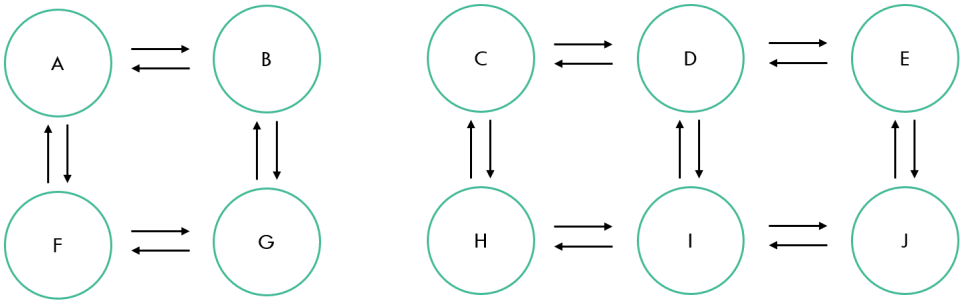
10.1 Introduction

The fully adaptive framework presented in Chapter 9 has been described in the context of sampling over several distributions, where one is typically interested in only one reference distribution, with the other distributions providing increased mobility of the degrees of freedom of interest. However, FAST is a very general procedure which can be readily applied in a broader context to any ergodic discrete Markov chain of states.

This chapter will show how this framework generalises to the context of relative binding free energy calculations. Afterwards, these calculations will be combined with the enhanced sampling protocol presented in Chapter 9, resulting in a hybrid algorithm: FAST/MBAR. This will allow us to not only explicitly address the sampling of certain degrees of freedom during free energy calculations, but also to minimise the correlation of the resulting free energy difference to the initial crystal structure. To do this, we will start with several preliminary theoretical considerations which extend the methodology presented in Chapter 9 before applying the FAST/MBAR algorithm on test cases of interest.



(A) Ergodic Markov chain



(B) Non-ergodic Markov chain

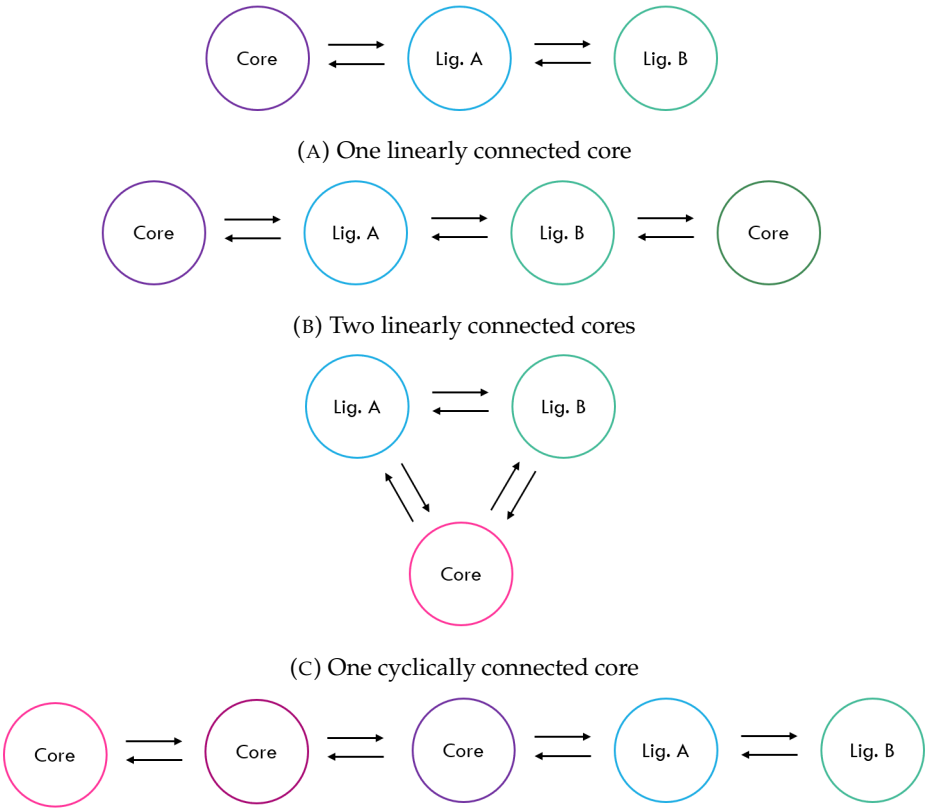
FIGURE 10.1: An example of an ergodic/connected (Figure 10.1a) and a non-ergodic/disconnected (Figure 10.1b) Markov chain.

10.2 Theoretical Considerations

10.2.1 Constructing the Markov Chain

The FAST method, described in Chapter 9, can be regarded as a black-box procedure which automatically determines an interpolative mapping between two user-specified fixed states. In the context of enhanced sampling, these two states were defined to represent the Hamiltonian of interest and a Hamiltonian where a subset of the interactions is completely decoupled. However, these states could also correspond to two different ligands, as in the context of relative binding free energy calculations. Furthermore, this framework can be readily generalised to the case of N states of interest connected in an arbitrary way, the only requirement being that there are no disconnected "islands" of states, i.e. the Markov chain is ergodic (Figure 10.1).

There are multiple ways to define the fixed states and their connectivity in a suitable manner for a relative binding free energy calculation with enhanced sampling (Figure 10.2). The simplest way to achieve this is by adding a single reference decoupled state and connecting it to one of the ligands (Figure 10.2a). However, one could conceivably add another reference state and connect it to the other ligand as well (Figure 10.2b). Another scenario arises when both of these reference states have the same Hamiltonian, in which case the corresponding Markov chain can form a



(A) One linearly connected core

(B) Two linearly connected cores

(C) One cyclically connected core

(D) Multiple linearly connected cores

FIGURE 10.2: Examples of different Markov chains combining enhanced sampling and ligand–ligand perturbations.

closed loop (Figure 10.2c). Finally, one can partition the enhanced sampling of many degrees of freedom into several stages of increasing decoupling (Figure 10.2d).

In this chapter, we will opt for the simplest type of Markov chain, where a single connection is made between the decoupled state and the ligand containing dummy atoms, while the two perturbed ligands will be connected in a single-topology fashion (Figure 10.2a). We will not consider cases where both ligands have dummy atoms, as these can be always decomposed into two separate alchemical perturbations. In this way, if the simulation starts at the decoupled state, the phase space volume relevant to the next states is always strictly decreasing, since there are more configurational states accessible to the noninteracting dummy atoms than to the fully coupled ones. This is especially relevant for the initial adaptive alchemical sequential Monte Carlo (AASMC) procedure, where it is important to minimise the sample diversity loss as the simulation progresses. Moreover, this type of Markov chain minimises the number of state connections, thereby minimising the computational effort dedicated to adaptation and energy evaluation over all intermediate states.

Even though we only focus on a maximum of three fixed states, this framework can be conceptually applied to an arbitrarily large amount of states. These could conceivably

represent different alchemical regions, effective temperatures, ligands, protonation states, or hydration levels. While all of these considerations are of potential interest for future work, they will not be addressed in this chapter.

10.2.2 Perturbing Harmonic Bonds

The FAST protocol presented in Chapter 9 only involved the perturbation of nonbonded interactions and dihedral terms. While these considerations are sufficient in the context of enhanced sampling, a single-topology relative binding free energy calculation often requires at least one change in the harmonic bond and angle terms as well. Of particular interest to us will be the harmonic bond terms U_{bond}, which are of the form:

$$U_{bond}(r) = k_r(r - r_{eq})^2 \tag{10.1}$$

where r is the bond length, r_{eq} is the equilibrium bond length and k_r is the hamonic force constant.

In practice, oscillations about the equilibrium bond length r_{eq} have low amplitudes caused by stiff force constants. This results in highly localised distributions, reminiscent of their asymptotic limit, the Dirac delta distribution[375] (corresponding to fully constrained bond lengths). While perturbing the force field parameters is still a valid procedure in this scenario, the high localisation of the distributions means that a large number of intermediate states will be needed to interpolate between the two endpoint equilibrium bond lengths. This can be particularly problematic when small atoms are perturbed to large atoms. For example, changing a carbon-bonded hydrogen to a bromine atom involves a bond length change as large as 0.85Å[27] and will as such require multiple intermediate states for sufficient overlap, thereby reducing the efficiency of the FAST procedure.

Fortunately, one can take advantage of the high stiffness of k_r, since in practice it means that the remainder of the other much weaker forces do not significantly affect the probability distribution of the bond length. This is a prime setting for using specifically tailored MC moves to transform the bond distance independently of the other degrees of freedom. Such transformations, if possible, are much more desirable, since they can provide a direct mapping between two distributions (e.g. two normal distributions, as shown in Appendix D) without much loss in information.

Although the probability distributions defined by the harmonic bond terms are not as simple as the example in Appendix D, we can still use this intuition to construct a reversible transformation which will enable us to reduce the number of alchemical intermediates. In this context, we define reversible transformations $\vec{T}_{i\to j}(\vec{x}_i)$ of a

sample $\vec{x}_i$ drawn from a probability distribution $\pi(\lambda_i, \vec{x})$ to obey the following identity with respect to another probability distribution $\pi(\lambda_j, \vec{x})$:

$$\begin{aligned} \vec{x}_j &= \vec{T}_{i\to j}(\vec{x}_i) \\ \vec{x}_i &= \vec{T}_{j\to i}(\vec{x}_j) \end{aligned} \tag{10.2}$$

This condition simply means that there is a one-to-one mapping between $\vec{x}_i$ and $\vec{x}_j$, given by $\vec{T}(\cdot)$. The subscripts will be omitted in the following discussion for conciseness.

In practice, this requirement confines us to simple linear transformations in configuration space, such as translation and rotation. For such transformations, we can impose a type of detailed balance which relates concerted moves in configuration and λ space:

$$\pi(\lambda_i, \vec{x}) p_{acc}(\lambda_j, \vec{T}(\vec{x})|\lambda_i, \vec{x}) d\vec{x} = \pi(\lambda_j, \vec{T}(\vec{x})) p_{acc}(\lambda_i, \vec{x}|\lambda_j, \vec{T}(\vec{x})) d\vec{T}(\vec{x}) \tag{10.3}$$

where we have assumed equal proposal probabilities $p_{prop}(\lambda_j|\lambda_i) = p_{prop}(\lambda_i|\lambda_j)$ for brevity. This equation is also directly applicable to the case of skew detailed balance (Equation 9.2).[352] The corresponding Metropolis acceptance criterion $p_{acc}(\lambda_j, T(\vec{x})|\lambda_i, \vec{x})$ is then:

$$\begin{aligned} p_{acc}(\lambda_j, \vec{T}(\vec{x})|\lambda_i, \vec{x}) &= \min\left[1, \frac{\pi(\lambda_j, \vec{T}(\vec{x}))}{\pi(\lambda_i, \vec{x})} \frac{d\vec{T}(\vec{x})}{d\vec{x}}\right] \\ &= \min\left[1, \frac{\pi(\lambda_j, \vec{T}(\vec{x}))}{\pi(\lambda_i, \vec{x})} |\mathbf{J}_{\vec{T}}(\vec{x})|\right] \end{aligned} \tag{10.4}$$

where $|\mathbf{J}_{\vec{T}}(\vec{x})|$ denotes the Jacobian determinant corresponding to the transformation $\vec{T}(\vec{x})$. The usefulness of Equation 10.3 comes from the fact that the evaluation of the potentially unfavourable $\pi(\lambda_j, \vec{x})$ is circumvented and is instead replaced by the more favourable $\pi(\lambda_j, \vec{T}(\vec{x}))$. For a bond rescaling transformation, this means that the bond length is changed to a more favourable value before evaluating $\pi(\lambda_j, \vec{T}(\vec{x}))$ and is reverted back to the original bond length if the transition attempt is rejected.

In this chapter, the simplest type of bond rescaling transformation will be considered, where the bond length is scaled by a constant scaling factor s, which is determined by the ratio of the parameter-dependent equilibrium bond distances r_{eq} at λ_i and λ_j:

$$s \equiv \frac{r_{eq}(\lambda_j)}{r_{eq}(\lambda_i)} \tag{10.5}$$

In this scenario, it can be shown that $|\mathbf{J}_{\vec{T}}(\vec{x})|$ is equal to s^3 (Appendix E). In practice, one of the bond atoms will be chosen to be stationary, while the other atoms and the rest of the molecule will be moved by this transformation. While this type of scaling is obviously reversible, it is not expected to be useful in the case of ring bond perturbations, since one cannot change the bond length without affecting the other internal degrees of freedom, and thus the other energy terms. Therefore, this procedure will only be used for rescaling bonds outside of rings, while any other degrees of freedom will not be changed.

10.3 Methods

10.3.1 System Preparation and Simulation

All system preparation and simulation methods are identical to those described in Chapter 9, where only the split protocol utilising the classical soft-core potential (CSC)[44] potential (Figure 9.3a) was used for all simulations with parameters $a = 1$, $b = 1$, $c = 6$ and $\alpha = 0.5$. All simulations were performed in triplicate. Two different types of protocols were investigated for all of the systems: FAST with only two fixed states corresponding to the fully-coupled ligands A and B, and FAST with enhanced sampling, with the corresponding Markov chain shown in Figure 10.2a. In the case of non-ring bond perturbations, all λ transitions were facilitated with bond rescaling, as described in Section 10.2.2. During the bond rescaling procedure, the side of the bond connected only to alchemically modified atoms was chosen to be mobile, while the other side was kept static.

The FAST simulation procedure was the same as described in Chapter 9 using OpenMMSLICER 3.0.0. The main difference was that all λ values were discretised to three significant figures instead of two. The reason for this choice was the expected increased number of intermediate λ values for the alchemical changes explored in this chapter. In addition, harmonic restraints with force constants of 5 kcal/mol/Å^2 each were added to all non-perturbed ligand atoms during the equilibration stage of the extended heat shock protein 90 (Hsp90) FAST/MBAR protocol in order to prevent a ligand binding mode change. These procedures were repeated for both the bound and solvated legs of the free energy calculations.

In the case of protein tyrosine phosphatase 1B (PTP1B), one of the ligand perturbations investigated in Chapter 5 was used for FAST/MBAR validation. In this case, an additional set of triplicate FAST/MBAR simulations was performed, where each starting structure was taken from the 20 ns equilibrated structures reported in Chapter 5.

To compare FAST/MBAR to regular methods, triplicate alchemical free energy (AFE) calculations were run for each system in both legs in GROMACS 2018.4 using the same procedure described in Chapter 8. The free energy values corresponding to the PTP1B system with and without equilibration were directly taken from Chapter 5.

10.3.2 Topology

All alchemical transformations were performed in a single-topology fashion. The implementation of these topologies in OpenMM used code from Perses,[300,376] subsequently modified for the purposes of this study and included in OpenMMSLICER. In the implementation used hereafter, five different types of alchemical variables λ are used, each associated with a particular type of interaction: λ_{bonds}, λ_{angles}, $\lambda_{dihedrals}$, $\lambda_{sterics}$ and $\lambda_{electrostatics}$. In all cases, the force field parameters (force constants, equilibrium distances and charges) are linearly interpolated between the two end values as a function of the relevant alchemical variable. The only exception are the dihedral terms, where their corresponding energies, rather than the associated parameters, are linearly interpolated. Finally, the parameters of the nonbonded 1–4 interactions (force constants, equilibrium distances and charge products) are also interpolated in a linear fashion. All dummy atoms assume the equilibrium bonded values of their interacting counterparts, meaning that only the force constants and charges/charge products are scaled in these cases. They are then introduced into the system using a nonbonded soft-core potential with the same functional form as described in [44].

In the following simulations, λ_{bonds}, λ_{angles}, $\lambda_{dihedrals}$ and $\lambda_{sterics}$ were always changed simultaneously and independently of $\lambda_{electrostatics}$ (split protocol, Figure 9.3a). However, a generalised alchemical variable λ will be instead reported in the text for brevity, following the convention in Chapter 9. For the three-state FAST/MBAR protocols, λ corresponds to the common core at $\lambda = 0$ and to the fully coupled ligands A and B at $\lambda = 0.5$ and $\lambda = 1$ respectively. Any interpolation between these states is then performed identically to the two-state protocols.

10.3.3 Analysis

All reported time-dependent free energy distributions were obtained after 10-fold bootstrapping of the decorrelated samples as described in Chapter 9. The effective decorrelation time τ_{decorr} used for removing correlated samples was obtained from its last estimate during the FAST/MBAR run. This procedure was repeated for both the bound and solvated legs. The median of all final bootstrapped free energy values over all replicates of the enhanced solvated leg FAST/MBAR protocol was then consistently subtracted from all bound legs for this system in order to report relative

binding free energy values. Since in all cases the solvated legs displayed an order of magnitude lower variability than the bound legs, all of the observed variance has been attributed to the much more practically interesting bound legs.

To determine the dependence of the calculated $\Delta\Delta G^{\ominus}$ values on different ligand conformers, dihedral clustering was performed on the marginal distributions of each relevant ligand torsion at all λ values. All cluster boundaries were defined manually before extracting only the relevant simulation frames belonging to a particular cluster. If more than one relevant mode was observed, these were then used to obtain bootstrapped cluster-dependent free energies as described in the previous paragraph. The kinetics and the slowest implied timescales obtained from Markov state models (MSMs) followed the same procedure described in detail in Chapter 6. In almost all cases the lag time for reporting the implied timescales was chosen to be 50 ps. The only exception was the kinetic analysis corresponding to the ligand ethyl group rotation in Hsp90, where a lag time of 20 ps was instead used due to the fast substituent motion and poor MSM accuracy at longer lag times.

In two cases (Hsp90 and PTP1B) clustering of the ligand common core was performed to investigate the dynamics of some unenhanced rare events over time. The method used to perform this clustering followed the procedure used on the transforming growth factor beta (TGF-β) ligand common core described in Chapter 9. The atoms used to obtain the centre of geometry were chosen to be the six dihydroxybenzene ring atoms (Hsp90) and the five thiophene ring atoms (PTP1B).

In some cases, the Kruskal–Wallis rank-based test of statistical significance was used to compare several groups of $\Delta\Delta G^{\ominus}$ values. This test verified the null hypothesis that the mean ranks of the samples drawn from each of the groups are the same, and the resulting p-values obtained using SciPy[343] were accordingly reported. All other aspects of the analysis, such as the efficiency metrics followed the same procedures described in Chapter 9.

10.4 Results

10.4.1 Coagulation Factor Xa (FXa)

The first test case for validating FAST/MBAR is coagulation factor Xa (FXa) bound to a medium-sized ligand (PDB ID: 1LQD[4]), where a fluoride group is perturbed to a methyl group (Figures 10.3b and 10.3c). In the enhanced FAST/MBAR protocol, the two torsions closest to the perturbed group, as well as an additional distal hydroxide torsion, were explicitly decoupled as well (Figure 10.3a). The latter torsion was added purely for demonstrative purposes, as it is expected to be straightforward to sample with alchemical methods without much performance penalty.

System	Enhanced	Replicate 1	Replicate 2	Replicate 3	Total Average	Total St. dev.
FXa	No	35.99	35.48	36.77	36.08	0.65
	Yes	0.33	0.38	0.36	0.35	0.02
Thrombin	No	20.41	21.23	17.05	19.56	2.21
	Yes	0.16	0.13	0.20	0.16	0.04
Hsp90	No	3.00	4.51	3.04	3.52	0.86
	Yes	0.12	0.06	0.13	0.10	0.04
PTP1B w/o Equilibration	No	1.51	2.01	1.11	1.55	0.45
	Yes	0.11	0.07	0.07	0.08	0.02
PTP1B w/ Equilibration	No	0.84	2.13	1.66	1.54	0.65
	Yes	0.06	0.04	0.11	0.07	0.04

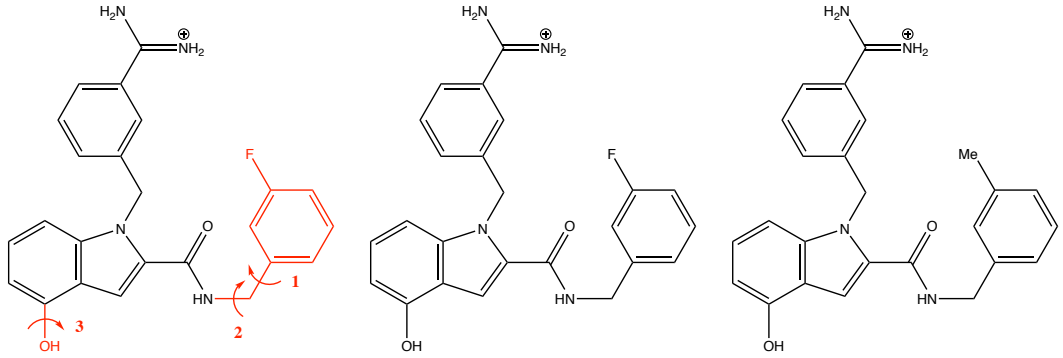
TABLE 10.1: The number of round trips per nanosecond for each of the systems across different replicates and FAST/MBAR protocols. The averages and the corresponding standard deviations are also given.

System	Enhanced	Replicate 1	Replicate 2	Replicate 3	Total Average	Total St. dev.
FXa	No	0.93	0.76	0.76	0.81	1.12
	Yes	3.30	1.26	2.81	2.27	1.68
Thrombin	No	1.47	2.68	1.92	1.96	1.35
	Yes	2.36	6.36	1.83	3.02	1.93
Hsp90	No	2.24	0.89	1.62	1.48	1.60
	Yes	3.56	0.27	3.50	1.51	4.36
PTP1B w/o Equilibration	No	0.41	0.75	1.08	0.69	1.64
	Yes	0.37	0.81	2.09	0.86	2.38
PTP1B w/ Equilibration	No	3.76	0.82	0.58	1.22	2.70
	Yes	1.66	3.85	2.21	2.42	1.53

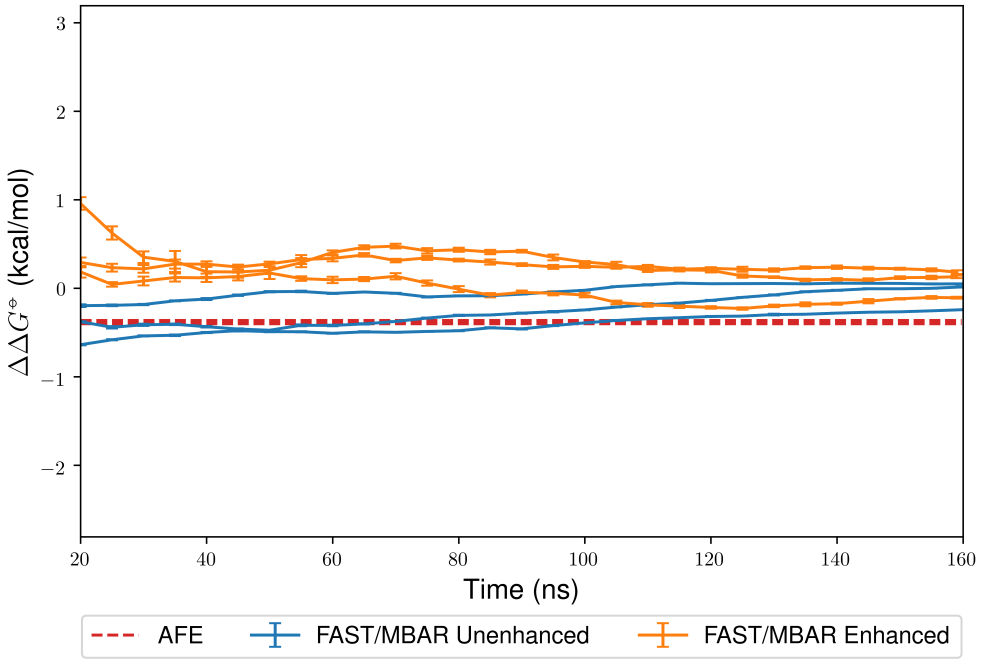
TABLE 10.2: The fraction of total samples between $\lambda = 1$ and $\lambda = 0$ for each of the systems across different replicates and FAST/MBAR protocols. The geometric averages and the corresponding geometric standard deviations are also given.

System	Enhanced	Replicate 1	Replicate 2	Replicate 3	Total Average	Total St. dev.
FXa	No	1.14	1.15	1.14	1.15	0.01
	Yes	5.45	5.45	6.30	5.73	0.49
Thrombin	No	1.32	1.29	1.34	1.32	0.02
	Yes	14.62	13.91	11.13	13.22	1.84
Hsp90	No	3.25	2.34	3.16	2.92	0.50
	Yes	9.38	21.80	8.97	13.38	7.29
PTP1B w/o Equilibration	No	2.71	1.97	3.76	2.81	0.90
	Yes	5.03	8.63	7.85	7.17	1.90
PTP1B w/ Equilibration	No	4.82	1.87	2.50	3.06	1.56
	Yes	2.98	3.49	1.87	2.78	0.83

TABLE 10.3: The final measured $\hat{\tau}_{decorr}$ in ps for each of the systems across different replicates and FAST/MBAR protocols. The averages and the corresponding standard deviations are also given.



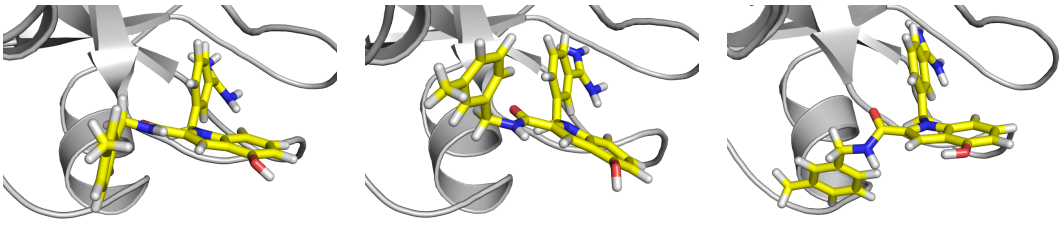
(A) Common core (B) Ligand *A* (C) Ligand *B*



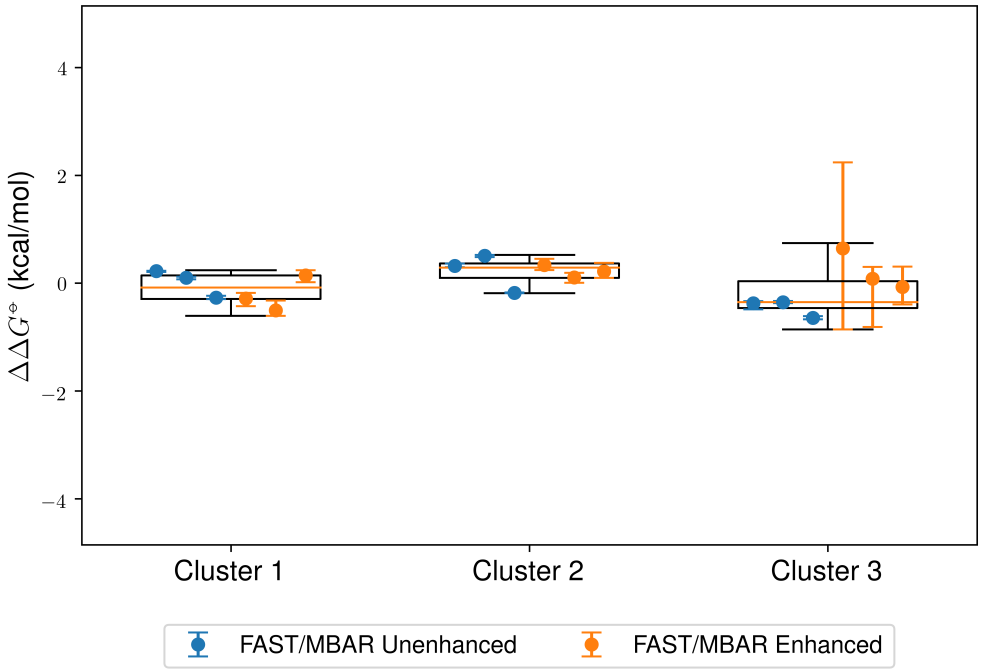
(D) $\Delta\Delta G^{\ominus}$ over time

FIGURE 10.3: The common core of the FXa ligand with dummy atoms and bonds in red (Figure 10.3a), the two ligands constituting the relative free energy perturbation (Figures 10.3b and 10.3c) and the bootstrapped $\Delta\Delta G^{\ominus}$ estimates corresponding to both FAST/MBAR protocols over time (Figure 10.3d). In Figure 10.3d, the median $\Delta\Delta G^{\ominus}$ is plotted over a range of timesteps with an associated error bar determined by the bootstrapped mean absolute deviation (MAD), while $\Delta\Delta G^{\ominus}$ determined by each AFE repeat is shown as a separate red line.

The behaviour of the FAST/MBAR protocol with and without enhanced sampling shows a clear time-dependent trend (Figure 10.3d). At short timescales (20 ns), the median free energy discrepancy between both protocols is 0.67 kcal/mol, while at longer timescales (160 ns), this difference decreases to 0.12 kcal/mol. In addition, the values obtained by AFE calculations are comparable to the FAST/MBAR protocol without extra sampling after 20 ns, with median difference of 0.00 kcal/mol. In this way, the unenhanced FAST/MBAR protocol outputs free energy values consistent with AFE calculations, before slowly converging to the values predicted by the enhanced FAST/MBAR protocol. This suggests that the rare events explicitly sampled



(A) Cluster 1 (B) Cluster 2 (C) Cluster 3



(D) Cluster $\Delta\Delta G^{\circ}$ for torsion 2

FIGURE 10.4: Three different clusters corresponding to torsion 2 in the FXa ligand (Figures 10.4a to 10.4c) and their respective $\Delta\Delta G^{\circ}$ values obtained from both FAST/MBAR protocols (Figure 10.4d). The orange lines represent the median of the total dataset, the boxes include the interquartile range, while the whiskers extend to the 5^{th} and 95^{th} percentiles. Each data point represents the median bootstrapped value, and the error bars extend between the minimum and maximum bootstrapped values.

by the enhanced protocol are explored by the unenhanced one simply by virtue of long-timescale molecular dynamics (MD) sampling.

Analysis of the obtained $\Delta\Delta G^{\circ}$ values from different ligand conformers reveals that torsion 2, shown in Figure 10.3a, results in the highest free energy discrepancies. Its three different modes (Figures 10.4a to 10.4c) result in significant free energy differences (Kruskal–Wallis p-value $\ll 0.001$) with median values of -0.08, 0.29 and -0.35 kcal/mol, respectively (Figure 10.4d), indicating that an unfavourable starting conformer can result in a significant free energy bias if the ligand sampling is poor, as is often the case for short-timescale AFE simulations. MSM analysis of torsion 2 reveals that explicit enhancement results in an approximately 5-fold decrease of its slowest implied timescale of the transition (from 4.29 ± 0.65 ns to 0.86 ± 0.08 ps),

indeed confirming that this torsion is the main source of short-timescale discrepancies between the two protocols. While implied timescales of ∼4 ns are not accessible to the AFE protocol used in this study, they can still be readily explored by the unenhanced FAST/MBAR protocol over 160 ns, which is why the two FAST/MBAR protocols have a significantly better agreement at these longer timescales.

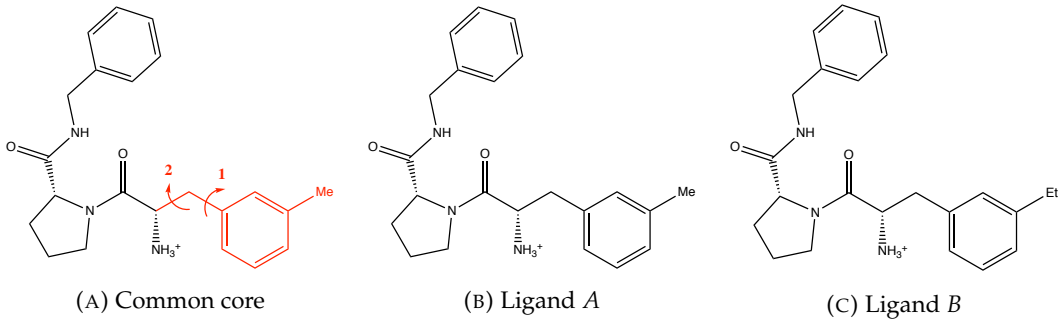
Both FAST/MBAR protocols require a small number of λ values, with the unenhanced protocol converging to only 3 ± 0 final values, compared to 5 ± 0 values generated by the initial AASMC procedure. In contrast, the optimal enhanced FAST/MBAR protocol has 13 ± 3 λ values compared to 21 ± 1 windows generated by AASMC. Similarly to the results in Chapter 9, it can be seen that the adaptive protocol optimisation procedure is independent of the initial AASMC protocol and generates the shortest possible λ schedules, with the perturbation of a fluoride to a methyl group only requiring one intermediate λ window. This also shows that the bond rescaling routine is also extremely efficient at reducing the number of intermediates when the bond length is changed.

The efficiency of the protocols is reflected by the round-trip rates per nanosecond: 36 ± 1 for the unenhanced protocol and 0.35 ± 0.02 for the enhanced protocol, respectively (Table 10.1). Despite the former protocol being 100 times less efficient than the latter, this round-trip rate is still satisfactory over the timescales studied (160 ns). While this 100-fold difference in efficiency can be largely attributed to the longer λ protocols when some of the torsions are explicitly enhanced, the effective decorrelation time $\hat{\tau}_{decorr}$ for the enhanced protocol is still ∼5 times higher: 5.73 ± 0.49 ps compared to 1.15 ± 0.01 ps for the unenhanced one (Table 10.3). Therefore, the addition of extra sampling creates an additional kinetic barrier in λ space which increases τ_{decorr}. This observation is also supported by the higher deviation from unity of the ratio of samples between $\lambda = 1$ and $\lambda = 0$, $\frac{N_1}{N_0}$, for the enhanced protocol: 2.27 ± 1.68 versus 0.81 ± 1.12 for the unenhanced protocol (Table 10.2).

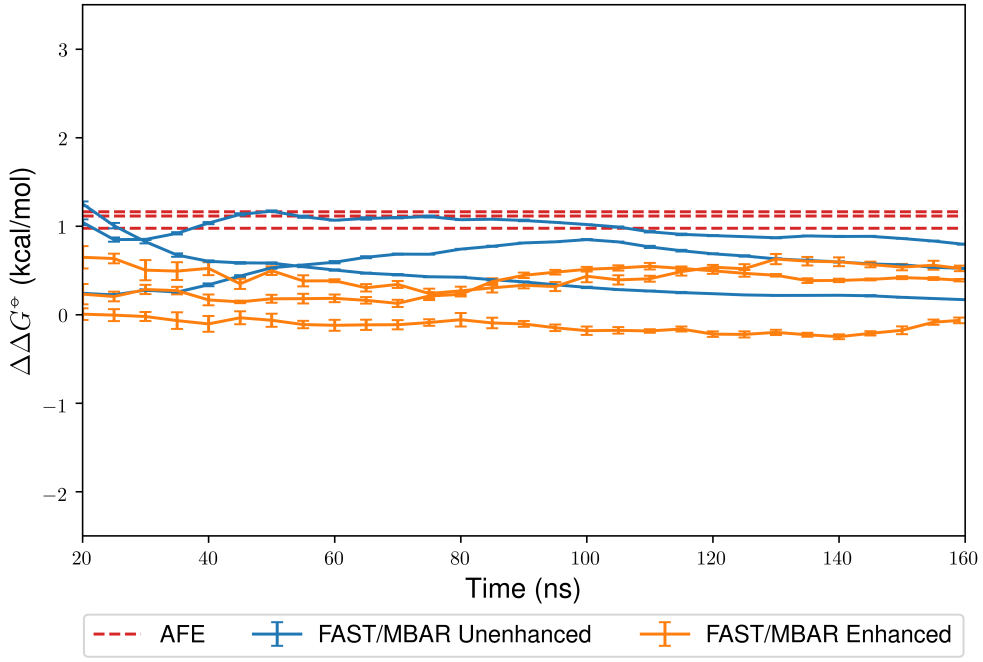
10.4.2 Thrombin

Thrombin bound to a medium-sized ligand is another system which constitutes a practically relevant test case for FAST/MBAR (PDB ID: 2ZFF[377]). Let us consider the simple perturbation of a methyl group to an ethyl group, where the enhanced FAST/MBAR protocol explicitly handles the rotation of the two torsions closest to the alchemical perturbation (Figures 10.5a to 10.5c). In this way we will be able to compare the quality of sampling of the tolyl ring between both FAST/MBAR protocols.

The free energy results from both FAST/MBAR protocols closely follow the behaviour observed in the FXa test case. Here, the median free energy difference between both protocols after only 20 ns of sampling is 0.81 kcal/mol, decreasing to 0.13 kcal/mol



(A) Common core (B) Ligand *A* (C) Ligand *B*

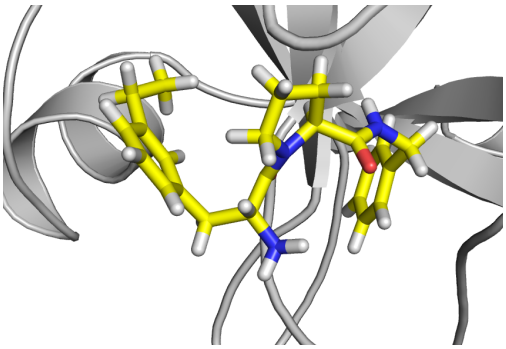


(D) $\Delta\Delta G^{\ominus}$ over time

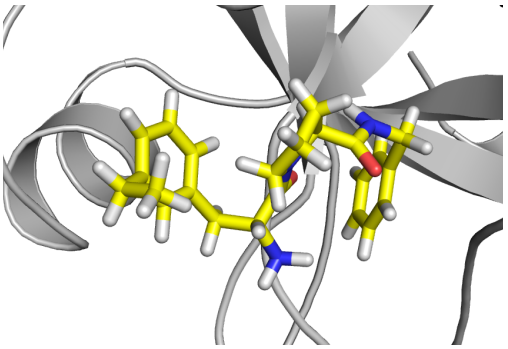
FIGURE 10.5: The common core of the thrombin ligand with dummy atoms and bonds in red (Figure 10.5a), the two ligands constituting the relative free energy perturbation (Figures 10.5b and 10.5c) and the bootstrapped $\Delta\Delta G^{\ominus}$ estimates corresponding to both FAST/MBAR protocols over time (Figure 10.5d). In Figure 10.5d, the median $\Delta\Delta G^{\ominus}$ is plotted over a range of timesteps with an associated error bar determined by the bootstrapped MAD, while $\Delta\Delta G^{\ominus}$ determined by each AFE repeat is shown as a separate red line.

after 160 ns (Figure 10.5d). Similarly, the AFE results agree well with the unenhanced FAST/MBAR protocol after 20 ns with a median difference of 0.07 kcal/mol. In this way, it can again be seen that FAST/MBAR at shorter timescales closely follows the AFE results before eventually approaching the free energy value obtained by the enhanced protocol. This again suggests the presence of a slow transition to a more thermodynamically relevant state which is not captured by short-timescale sampling.

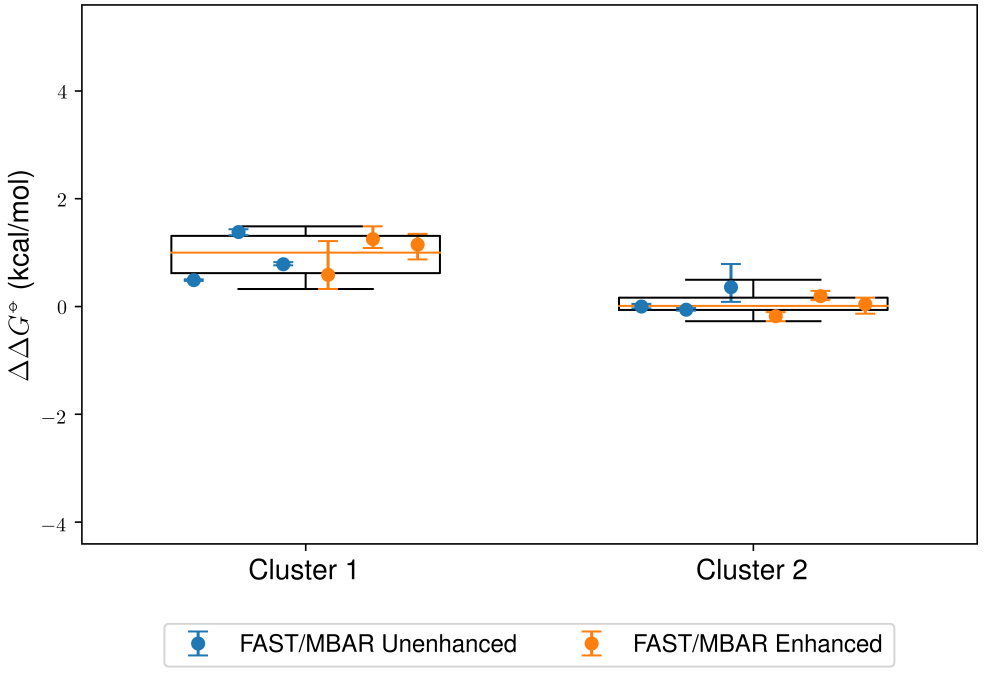
The main source of these free energy discrepancies torsion 1 in Figure 10.5a, which has two modes (Figures 10.6a and 10.6b) with associated median $\Delta\Delta G^{\ominus}$ values of 1.00 and 0.01 kcal/mol, respectively (Figure 10.6c). This significant difference (Kruskal–Wallis p-value $\ll 0.001$) once again shows that the choice of initial conformer introduces



(A) Cluster 1



(B) Cluster 2



(C) Cluster $\Delta\Delta G^{\circ}$ for torsion 1

FIGURE 10.6: Two different clusters corresponding to torsion 1 in the thrombin ligand (Figures 10.6a and 10.6b) and their respective $\Delta\Delta G^{\circ}$ values obtained from both FAST/MBAR protocols (Figure 10.6c). The orange lines represent the median of the total dataset, the boxes include the interquartile range, while the whiskers extend to the 5^{th} and 95^{th} percentiles. Each data point represents the median bootstrapped value, and the error bars extend between the minimum and maximum bootstrapped values.

non-negligible bias and adequate sampling is required for reliable free energy calculations. Similarly to FXa, the kinetics of this torsion are substantially accelerated by explicit sampling, with the slowest implied timescale being 26.04 ± 9.27 ns and 0.94 ± 0.11 ns for the unenhanced and enhanced FAST/MBAR protocols, respectively. Despite the much higher torsional kinetic barrier compared to the one observed for FXa, it is clear that this long implied timescale can still be explored by the unenhanced 160 ns FAST/MBAR protocol, even if the sampling is not as efficient as in the explicitly enhanced protocol. In this way, this test case once again demonstrates the sampling advantage of FAST/MBAR over conventional AFE calculations.

The protocol efficiency of both FAST/MBAR schedules is again demonstrably higher than the initial protocols output by AASMC. For FAST/MBAR without explicit sampling enhancement, the final average protocol length is 4 ± 0 λ windows, compared to 7 ± 1 initial λ values. Similarly, the enhanced FAST/MBAR protocol only needs an average of 12 ± 2 λ windows, as opposed to 18 ± 1 intermediates output by AASMC. This behaviour is comparable to FXa, once again demonstrating the parsimony of the protocol optimisation procedure.

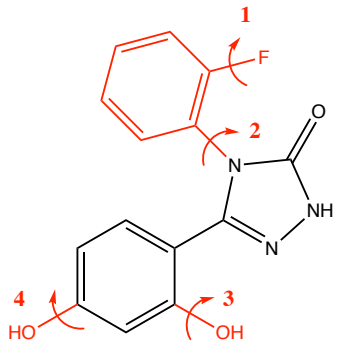
As expected, the shorter protocols of the unenhanced FAST/MBAR algorithm result in high round-trip rates per nanosecond: 20 ± 2, versus 0.16 ± 0.04 for the enhanced FAST/MBAR protocol (Table 10.1), again showing a two orders of magnitude difference in efficiency. Although the average round-trip time of the FAST/MBAR protocol without explicit sampling enhancement is close to the one predicted by the protocol optimisation procedure with an average $\hat{\tau}_{decorr}$ of only 1.32 ± 0.02 ps, the enhanced protocol exhibits a much higher effective correlation with $\hat{\tau}_{decorr}$ of 13.22 ± 1.84 (Table 10.3), indicating the presence of substantial kinetic barriers in λ space introduced by the extra sampling. Unsurprisingly, the less efficient protocol also exhibits a larger $\frac{N_1}{N_0}$ with a geometric mean of 3.02 ± 1.93 compared to 1.96 ± 1.35 for the unenhanced protocol (Table 10.2).

10.4.3 Heat Shock Protein 90 (Hsp90)

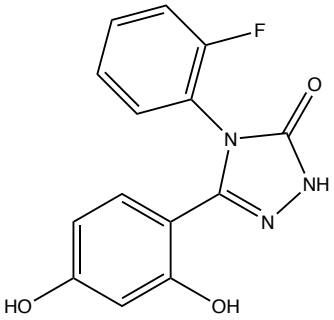
The next test case for FAST/MBAR is Hsp90 bound to a ligand containing three rings (PDB ID: 5J64[378]). Here we consider the perturbation of a fluoride group to an ethyl substituent, with the enhanced FAST/MBAR protocol exploring the torsion closest to the perturbation, as well as the simple dihedral rotations of two hydroxyl groups (Figures 10.7a to 10.7c). This system has been previously considered in the context of nonequilibrium free energy calculations, where it was found that rare binding site events can drastically hinder the sampling efficiency thereof.[379]

In contrast to the previous test cases, the behaviour of the free energy estimates of both FAST/MBAR protocols over time shows poor convergence, with a final free energy MAD of 1.30 and 0.60 for the enhanced and unenhanced protocols, respectively (Figure 10.7d). In this case, the median free energy deviation between both protocols is 0.55 kcal/mol after 20 ns, increasing to 0.70 kcal/mol after 160 ns. Nevertheless, the free energy estimates by AFE calculations yield values closer to the unenhanced protocol (0.21–0.23 kcal/mol) at both timescales—behaviour which is consistent with the previous test cases.

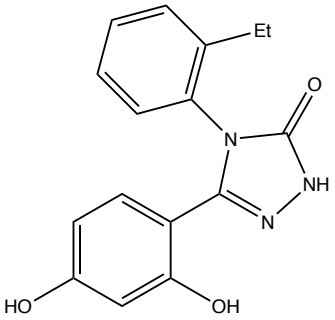
The main source of the increased variability in the estimated free energies is one of the enhanced FAST/MBAR simulations, which results in a final free energy difference of -1.78 kcal/mol. Correlated with this behaviour is the kinetic trapping of this



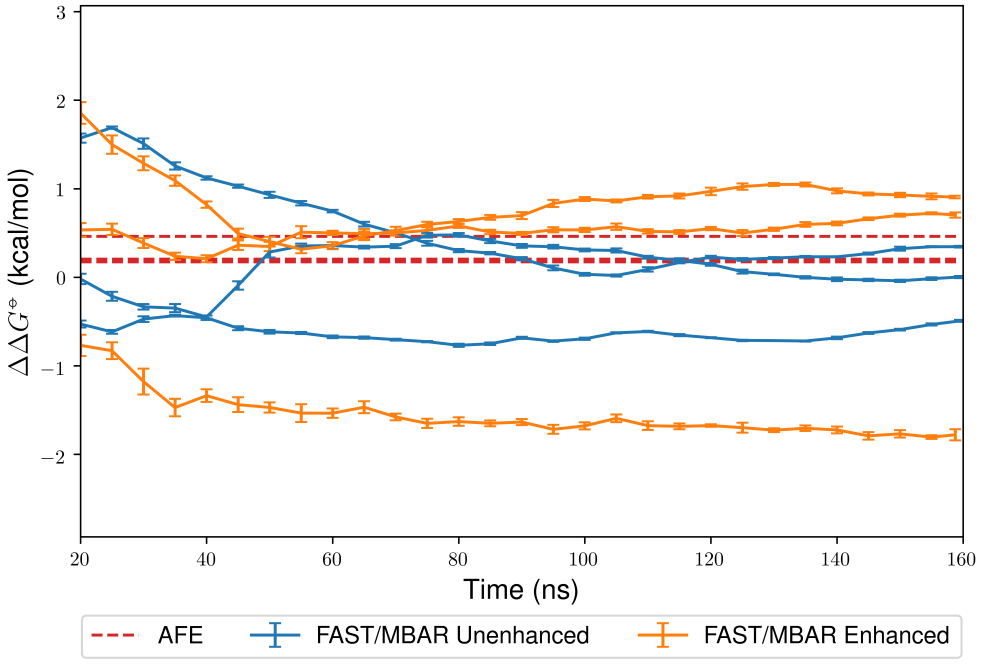
(A) Common core



(B) Ligand *A*



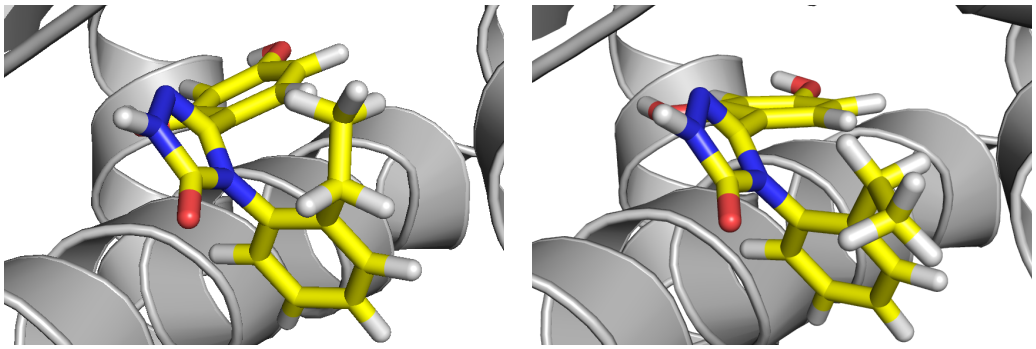
(C) Ligand *B*



(D) $\Delta\Delta G^{\ominus}$ over time

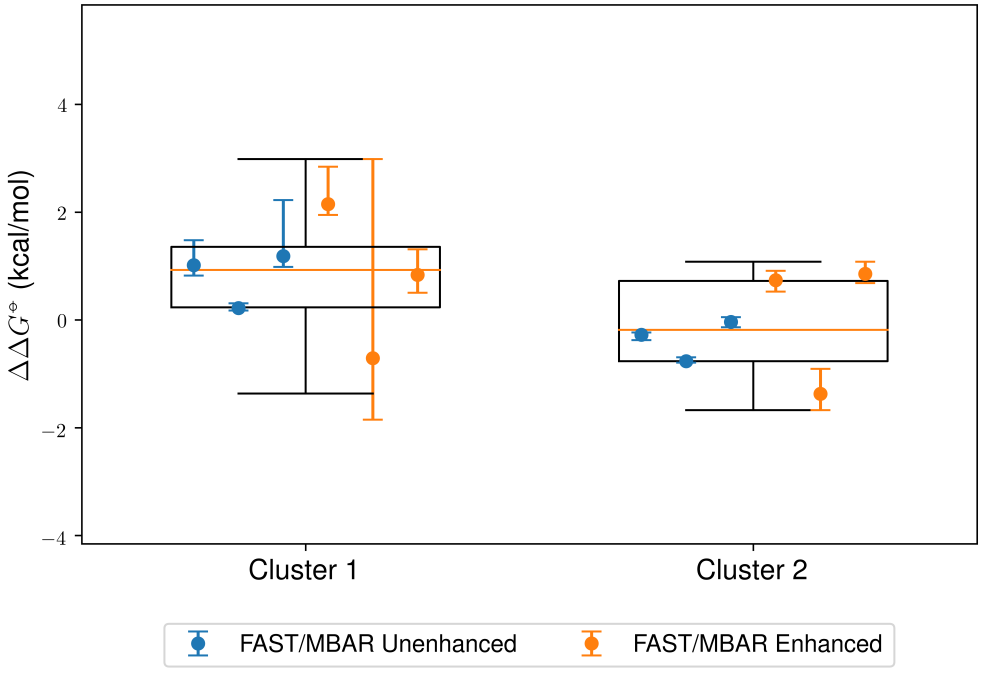
FIGURE 10.7: The common core of the Hsp90 ligand with dummy atoms and bonds in red (Figure 10.7a), the two ligands constituting the relative free energy perturbation (Figures 10.7b and 10.7c) and the bootstrapped $\Delta\Delta G^{\ominus}$ estimates corresponding to both FAST/MBAR protocols over time (Figure 10.7d). In Figure 10.7d, the median $\Delta\Delta G^{\ominus}$ is plotted over a range of timesteps with an associated error bar determined by the bootstrapped MAD, while $\Delta\Delta G^{\ominus}$ determined by each AFE repeat is shown as a separate red line.

simulation in λ space which deteriorates the sampling quality (Figure 10.9). These kinetic barriers are caused by slow transitions of the common core, observed at decoupled λ values ($0 \leq \lambda \leq 0.5$), resulting in three main common-core conformers (Figures 10.9a to 10.9c). Since these binding modes are not observed at higher λ values, it can be deduced that these transitions are promoted by the increased mobility at lower λ values, resulting in kinetically trapped physically irrelevant states which create a bottleneck in the simulation. Since the states relevant for the free energy estimation ($0.5 \leq \lambda \leq 1.0$) are not sampled after $\sim$50 ns, it can be conjectured that the anomalous free energy behaviour of this simulation is primarily caused by these rare



(A) Cluster 1

(B) Cluster 2

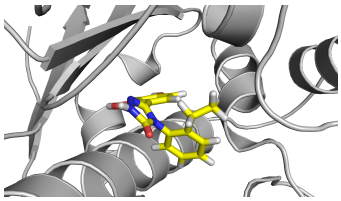


(C) Cluster $\Delta\Delta G^{\ominus}$ for torsion 1

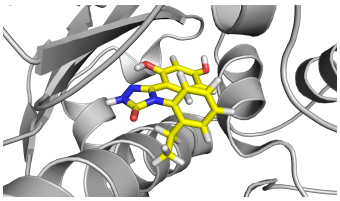
FIGURE 10.8: Two different clusters corresponding to torsion 1 in the Hsp90 ligand (Figures 10.8a and 10.8b) and their respective $\Delta\Delta G^{\ominus}$ values obtained from both FAST/MBAR protocols (Figure 10.8c). The orange lines represent the median of the total dataset, the boxes include the interquartile range, while the whiskers extend to the 5^{th} and 95^{th} percentiles. Each data point represents the median bootstrapped value, and the error bars extend between the minimum and maximum bootstrapped values.

events. These observations are consistent with those in [379], where the authors describe concerted ligand transitions and binding site hydration changes to be the main source of the decreased efficiency of the free energy estimation for this protein–ligand system.

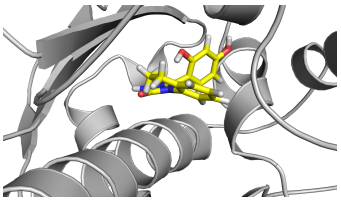
In contrast to the previous test cases, not much mobility in the torsional ligand degrees of freedom is observed during the unenhanced protocol, with only the ethyl group torsion (Figure 10.7a) exhibiting transitions with implied timescales of 0.10 ± 0.02 ns and 0.92 ± 0.26 ns for the unenhanced and the enhanced FAST/MBAR protocols, respectively. Although the two modes of this torsion (Figures 10.8a



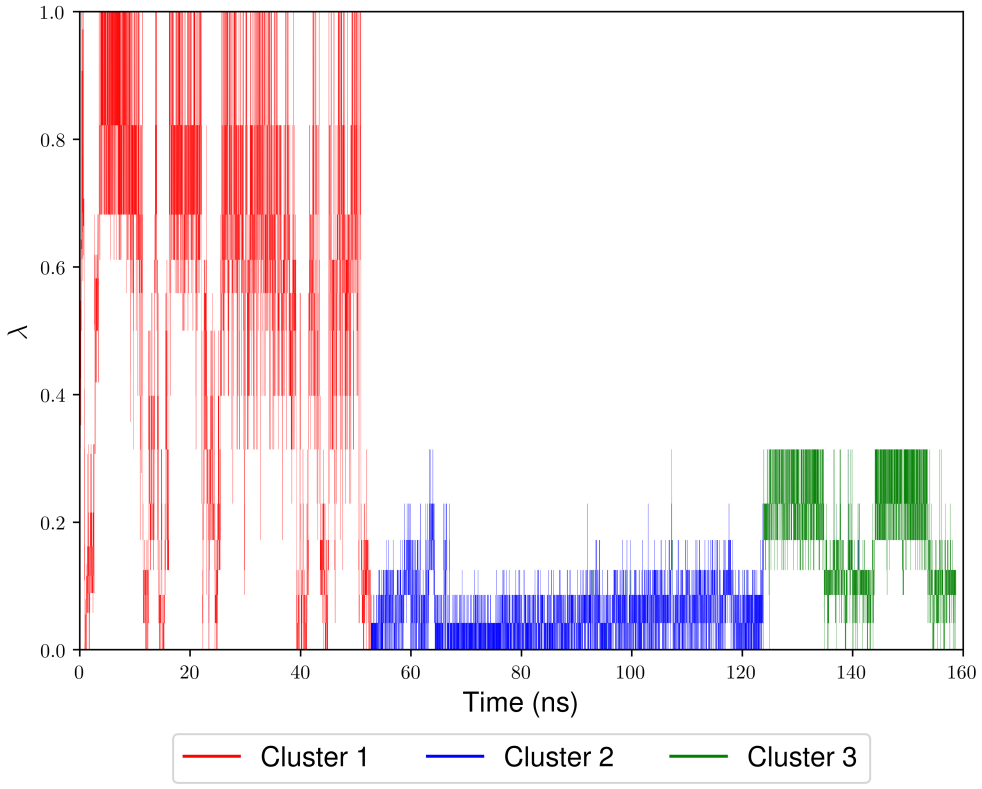
(A) Cluster 1



(B) Cluster 2



(C) Cluster 3



(D) Hsp90 ligand cluster populations

FIGURE 10.9: The three ligand common core clusters (Figures 10.9a to 10.9c) and their transitions in λ space over time (Figure 10.9d) during the outlier simulation in Figure 10.7d.

and 10.8b) have significantly different median $\Delta\Delta G^{\ominus}$ values of 0.93 and -0.18 kcal/mol, respectively (Figure 10.8c), much of the variance is still unexplained by this transition and likely related to various nonbonded interactions between the ligand and the solvated protein. While the enhanced protocol also promotes the rotation of the fluorophenyl ring (Figure 10.7a) and results in observable transitions with an implied timescale of 0.66 ± 0.04 ns, no reliable cluster-based free energy analysis can be performed in this case due to the low populations of this conformer at $\lambda = 1$. Because of these low populations, the extra sampling provided by the enhanced protocol does not significantly improve the quality of the estimated $\Delta\Delta G^{\ominus}$ values over the unenhanced FAST/MBAR protocol for this system.

Despite the observed decrease in the efficiency of FAST/MBAR compared to the previous test cases, the alchemical protocols are of similar quality to FXa and

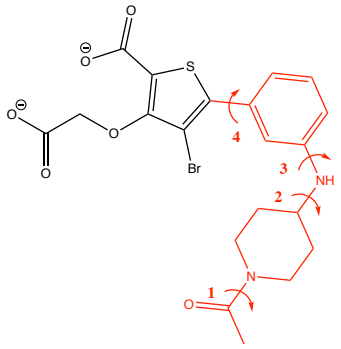
thrombin, with the unenhanced FAST/MBAR protocol containing 5 ± 0 λ values in total, compared to 10 ± 0 λ values output by the AASMC routine. Similarly, the enhanced FAST/MBAR protocol reduces the λ values from 26 ± 1 to 17 ± 4, which behaviour is also comparable to the previous test cases.

The observed round-trip rate per nanosecond is 3.52 ± 0.86 for the unenhanced protocol, compared to an average of 0.10 ± 0.04 for the enhanced one (Table 10.1). These values are consistent with $\hat{\tau}_{decorr}$: 2.92 ± 0.50 ps and 13.38 ± 7.29 ps for the unenhanced and enhanced protocol, respectively (Table 10.3). While the added extra sampling decreases the efficiency substantially, these results are still consistent with the previous test cases. The main exception is the kinetically trapped outlier described above, which has a much higher $\hat{\tau}_{decorr}$ of 21.80 ps. A more interesting difference is the lower efficiency of the unenhanced protocol compared to the previous test cases. Although this difference can be partially explained by the slightly higher complexity of the alchemical perturbation compared to the previous test cases, it is also likely influenced by the kinetic trapping in λ space discussed above. Nevertheless, the $\frac{N_1}{N_0}$ ratios are comparable to the previous systems with values of 1.48 ± 1.60 and 1.51 ± 4.36 for the unenhanced and enhanced FAST/MBAR protocols, respectively.

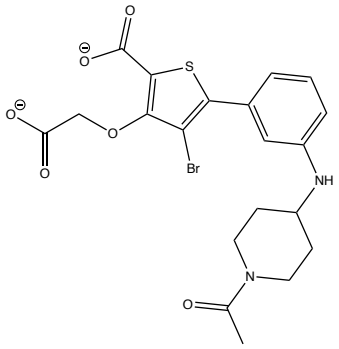
10.4.4 Protein Tyrosine Phosphatase 1B (PTP1B)

The final test case is taken from Chapter 5, where it was demonstrated that significantly different $\Delta\Delta G^{\ominus}$ values can be observed after an AFE calculation with a short (100 ps) or a long (20 ns) equilibration protocol. The perturbation of choice (PTP1B, pair 2, PDB ID: 2QBP[3]), shown in Figures 10.10b and 10.10c, exhibits large median differences between both equilibration protocols (larger than 2 kcal/mol). To investigate the sensitivity of FAST/MBAR to the equilibration length, two groups of simulations were run: a triplicate run starting from the unequilibrated structure and a separate simulation starting from each of the three pre-equilibrated structures generated in Chapter 5. In both scenarios, the enhanced FAST/MBAR protocol was designed to improve the sampling of the four closest torsions to the perturbed groups. To achieve this, the larger part of the ligand was alchemically decoupled from the rest of the system at $\lambda = 0$ (Figure 10.10a).

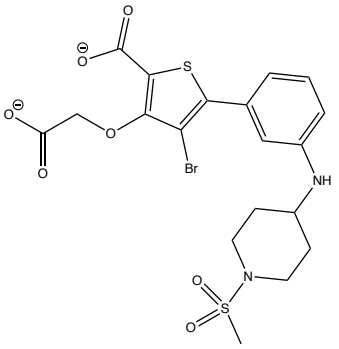
Interestingly, the FAST/MBAR protocol without enhanced sampling results in less variable free energy values for the pre-equilibrated structures, compared to the unequilibrated ones, with a MAD of 0.25 kcal/mol and 0.80 kcal/mol, respectively (Figures 10.10d and 10.10e). Furthermore, this difference in MAD is significant at the 5% level with a Kruskal–Wallis p-value of 0.05 when comparing the two respective populations of inter-replicate absolute deviations. This is a surprising result, since the extra decorrelation provided by the additional equilibration is expected to introduce uncertainty into the free energy estimates, instead of reducing it. Nevertheless,



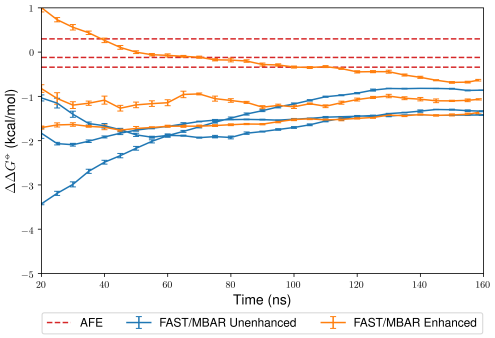
(A) Common core



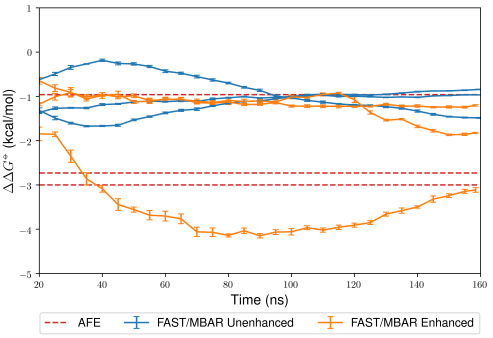
(B) Ligand *A*



(C) Ligand *B*



(D) Unequilibrated $\Delta\Delta G^{\ominus}$ over time

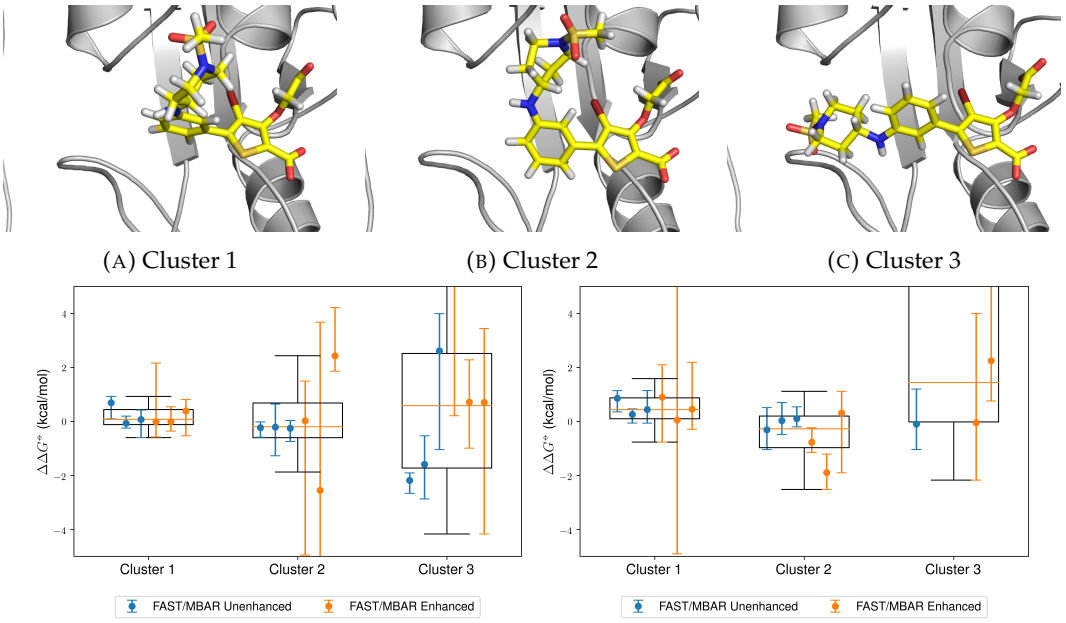


(E) Equilibrated $\Delta\Delta G^{\ominus}$ over time

FIGURE 10.10: The common core of the PTP1B ligand with dummy atoms and bonds in red (Figure 10.10a), the two ligands constituting the relative free energy perturbation (Figures 10.10b and 10.10c) and the bootstrapped $\Delta\Delta G^{\ominus}$ estimates corresponding to both FAST/MBAR protocols over time for the unequilibrated (Figure 10.10d) and the equilibrated (Figure 10.10e) structures. In Figure 10.10d, the median $\Delta\Delta G^{\ominus}$ is plotted over a range of timesteps with an associated error bar determined by the bootstrapped MAD, while $\Delta\Delta G^{\ominus}$ determined by each AFE repeat is shown as a separate red line.

comparison between the results obtained from the unequilibrated and the equilibrated protocol yields a high Kruskal–Wallis p-value of 0.28, due to the large inter-replicate MAD, meaning that the two sets of free energy values are statistically indistinguishable. After 160 ns, both types of unenhanced FAST/MBAR simulations result in comparable MAD of 0.19 kcal/mol (unequilibrated) and 0.21 kcal/mol (equilibrated), respectively and a Kruskal–Wallis p-value of 0.83, again indicating the statistical indistinguishability of both sets of simulations.

The enhanced FAST/MBAR protocol results in similar behaviour, with the unequilibrated structures resulting in higher variability than the equilibrated ones after 20 ns of simulation (0.90 kcal/mol versus 0.40 kcal/mol MAD). The trend is again reversed after 160 ns, with inter-replicate MADs of 0.24 and 0.64 kcal/mol for the unequilibrated and the equilibrated simulations, respectively (Figures 10.10d and 10.10e). However, in this case the difference can be explained by a single outlier, which reaches values as low as -4 kcal/mol during the course of the simulation, or $\sim$2 kcal/mol less than the lowest free energy observed in any of the other simulations.



(A) Cluster 1 (B) Cluster 2 (C) Cluster 3

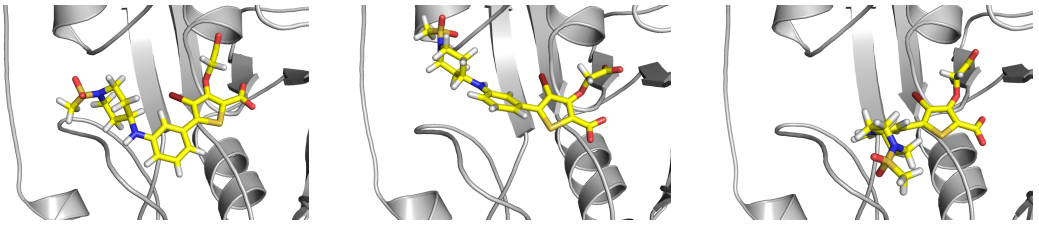
(D) Unequilibrated cluster $\Delta\Delta G^{\ominus}$ for torsion 4 (E) Equilibrated cluster $\Delta\Delta G^{\ominus}$ for torsion 4

FIGURE 10.11: Three different clusters corresponding to torsion 4 in the PTP1B ligand (Figures 10.11a to 10.11c) and their respective $\Delta\Delta G^{\ominus}$ values obtained from both FAST/MBAR protocols for the unequilibrated (Figure 10.11d) and the equilibrated (Figure 10.11e) structures. The orange lines represent the median of the total dataset, the boxes include the interquartile range, while the whiskers extend to the 5^{th} and 95^{th} percentiles. Each data point represents the median bootstrapped value, and the error bars extend between the minimum and maximum bootstrapped values.

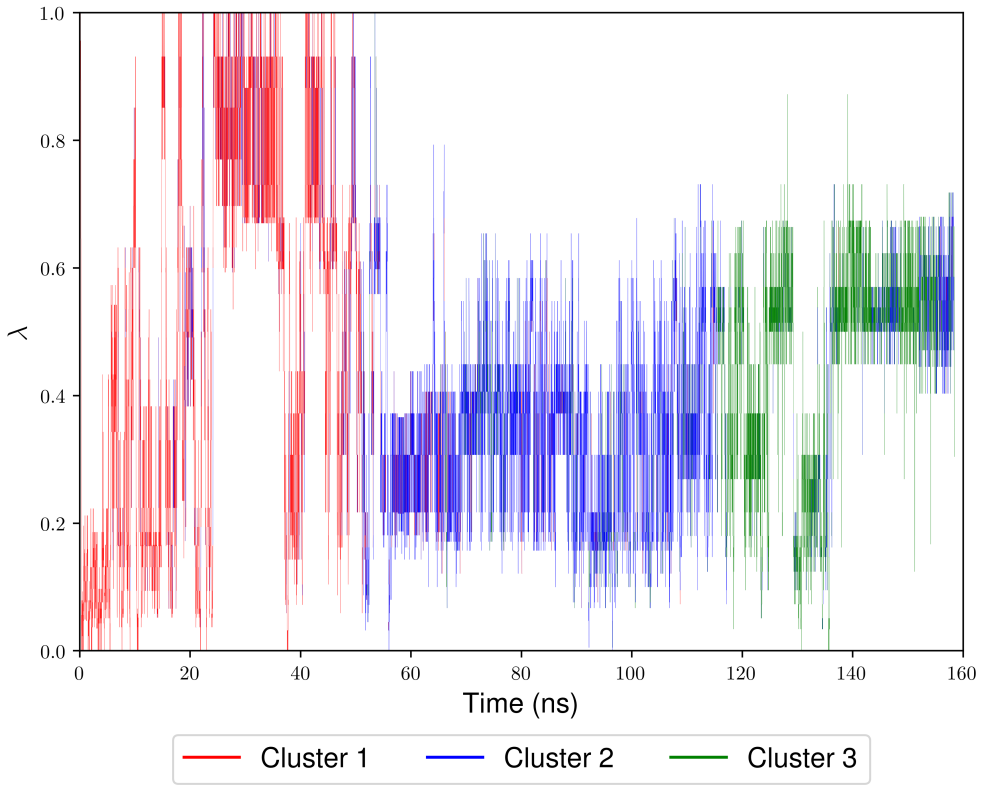
This behaviour is reminiscent of the Hsp90 outlier discussed in the previous section and is analogously expected to be caused by an unenhanced rare event.

As with the Hsp90 test case, this outlier is kinetically trapped in λ space, thereby reducing the reliability of the estimated free energies (Figure 10.12). The kinetic barriers are again correlated with three main common-core conformers (Figures 10.12a to 10.12c), whose transitions are again observed at highly decoupled states. In this case, however, the conformers are significantly sampled at $\lambda = 0.5$, which corresponds to the fully coupled carbonyl derivative. Therefore, these conformers are much more energetically favourable for one of the ligands, thereby creating a bottleneck in the part of the protocol responsible for the relative free energy calculation ($0.5 \leq \lambda \leq 1.0$).

Analysis of the free energy values as a function of the torsional degrees of freedom is difficult for this highly flexible ligand (Figure 10.10a), due to the large number of possible conformers, leading to a large uncertainty in the estimated $\Delta\Delta G^{\ominus}$ values. The degree of freedom showing the highest per-cluster free energy convergence is torsion 4 (Figure 10.10a), which has three associated modes (Figures 10.11a to 10.11c). These three clusters have significantly different respective $\Delta\Delta G^{\ominus}$ values of 0.08, -0.20 and 0.58 kcal/mol for the unequilibrated protocol, and 0.44, -0.27 and 1.44 kcal/mol for the equilibrated protocol (Figures 10.11d and 10.11e). There is large uncertainty in all



(A) Cluster 1 (B) Cluster 2 (C) Cluster 3



(D) PTP1B ligand cluster populations

FIGURE 10.12: The three PTP1B ligand common core clusters (Figures 10.12a to 10.12c) and their transitions in λ space over time (Figure 10.12d) during the outlier simulation in Figure 10.10e.

per-cluster $\Delta\Delta G^{\ominus}$ values, primarily caused by undersampling of this particular degree of freedom (cluster 3) and increased sampling of orthogonal degrees of freedom (clusters 1 and 2, enhanced protocol). Nevertheless, comparison between the median $\Delta\Delta G^{\ominus}$ values of cluster 1 and cluster 2 for both the unequilibrated and equilibrated FAST/MBAR protocols without sampling enhancement yields differences of 0.31 and 0.41 kcal/mol, respectively (Kruskal–Wallis p-value $\ll 0.001$), showing that even this distal degree of freedom can have a significant effect on the obtained free energy values. The slowest and most significantly enhanced degree of freedom, however, is torsion 3, with an implied timescale of 0.53 ± 0.04 ns after enhancement, compared to 9.56 ± 5.39 ns without enhancement (unequilibrated protocol), as determined by the MSM analysis. These results again show that even high kinetic barriers with implied timescales on the order of 10 ns can be readily explored by the unenhanced protocol

over the range of 160 ns, thereby demonstrating the robustness of FAST/MBAR towards the initial conformer used compared to standard AFE calculation methods.

This test case is significantly more challenging than the previous test cases, as evidenced by the high number of required λ windows. Their initial number can be as high as 34 for the enhanced protocol and 13 for the unenhanced protocol. This large number of intermediates is caused by a combination of a challenging decoupling procedure of a large part of the ligand and a relative perturbation involving transformations in five of the bonded terms, arising from the atom type change of the sulfonamide sulfur and nitrogen atoms. Although the FAST protocol optimisation procedure reduces the average number of λ windows to 9–11 for the unenhanced protocol in both the pre-equilibrated and unequilibrated simulations, a more peculiar behaviour is observed for the enhanced protocol. In this case, the λ values of the simulations starting from the equilibrated structures are reduced to 25 ± 2, while all of the unequilibrated simulations retain the same protocol as the initial AASMC procedure throughout the simulations, resulting in longer protocols (33 ± 2 λ values). While this observation is clearly caused by the inability of the optimiser to find a more favourable protocol and it can be partially explained by the high complexity of the protocols, it is not obvious why the optimisation procedure is successful for all of the pre-equilibrated simulations. A possible explanation is that the unequilibrated structures result in more unfavourable AASMC protocols, which makes finding a new minimum easier for the protocol optimisation procedure.

Similarly to the previous test cases, the increased complexity of the protocols is reflected by the reduced round-trip rate per nanosecond: 0.08 ± 0.03 and 1.61 ± 0.45 for the enhanced and unenhanced FAST/MBAR protocols, respectively (Table 10.1). The effective decorrelation time is comparable to the previous test cases with 8.64 ± 4.70 ps for the protocol with extra sampling and 2.75 ± 0.69 ps for the unenhanced protocol (Table 10.3). In all cases the values are comparable between the equilibrated and unequilibrated simulations, with the sole exception of the previously discussed outlier in Figure 10.10e, which exhibits $\hat{\tau}_{decorr}$ of 17.62 ps. As in the previous systems, the $\frac{N_1}{N_0}$ ratios are satisfactory for both protocols, with values of 0.92 ± 2.15 for the unenhanced protocol and 1.44 ± 2.31 for the enhanced protocol and (Table 10.2).

10.5 Discussion

The above test cases demonstrate that the FAST framework is not only capable of enhancing the sampling of specific degrees of freedom, as shown in Chapter 9, but can also be readily employed in the context of binding free energy calculations with or without extra targeted sampling. This makes FAST/MBAR a robust, flexible, highly

reproducible and fully automatable alternative to standard free energy calculation methods.

Comparing the behaviour between FAST/MBAR free energy calculations with and without enhanced sampling is of particular interest, since it can directly show the impact of enhanced sampling on the efficiency and quality of the free energy calculations. The FXa and thrombin results show a somewhat expected trend: at relatively short timescales ($\sim$20 ns), the free energy values obtained from the FAST simulations without extra sampling are similar to the short AFE calculations, while at longer timescales, they tend towards the values predicted by the enhanced protocol. This observation is readily explained by the fact that at longer timescales even higher torsional kinetic barriers can be surmounted with regular MD.

In contrast to FXa and thrombin, the Hsp90 results show large variance in the obtained free energy values and this variance increases over time, particularly for the enhanced protocol. Normally this would be an unexpected result, since the apparent complexity of the ligand is comparable to the other better-behaved test cases. However, it is well-known that Hsp90 can be a problematic system for molecular simulations, due its high mobility and its closed binding site which makes binding site ligand and water sampling significantly more challenging.[114,230,379] Indeed, it was shown above that the main free energy outlier is caused by slow transitions of the common core at highly decoupled λ values—an observation which is reminiscent of the kinetically trapped TGF-β system considered in Chapters 8 and 9. In addition, the extra conformers provided by the enhanced FAST/MBAR are not very stable at the fully coupled ligand states, resulting in increased effective correlation and further reduced efficiency. These two manifestations of kinetic trapping in λ space make Hsp90 a much more challenging system to study than FXa and thrombin.

The PTP1B test case shows that even highly flexible large ligands can be handled by FAST/MBAR. However, turning off the nonbonded interactions of a large part of the ligand can lead to the same problem of common core transition at more decoupled states, as observed in Hsp90 and TGF-β. Despite this challenge, the robustness of FAST/MBAR with respect to the initial coordinates was shown to be very high and FAST/MBAR provides significant improvement over traditional AFE calculations.

It is therefore clear that both FAST/MBAR protocols have advantages and disadvantages and the superiority of one of them over the other is not obvious. On one hand, FAST/MBAR without extra sampling results in a large number of round trips, which in turn yields a higher amount of sampling at the states of interest and therefore higher confidence in the calculated free energies. On the other hand, extra sampling of the relevant degrees of freedom can provide a higher consistency of the free energy values over time, as well as output a better converged distribution of conformers, if these are of interest. Indeed, the above results show that the free energy

differences associated with a particular conformer can differ by more than 1 kcal/mol, meaning that sampling different conformers as efficiently as possible is crucial for minimising the initial-structure bias of the free energy calculation. It is equally as important that the enhanced region is carefully selected, however, as large regions can significantly decrease the efficiency of the algorithm by creating kinetic barriers in λ space. Similarly, enhancing the sampling of an unimportant degree of freedom will simply result in increased computational expense to no benefit. As this information is usually not known in advance, it follows that enhanced sampling of the ligand degrees of freedom should be done sparingly on only several key degrees of freedom, such as torsions that are close to the perturbed group, or ones that are expected to have very high kinetic barriers. The sampling of the other degrees of freedom can then be handled by the MD integrator, especially since the single-trajectory nature of FAST provides maximum decorrelation from the initial coordinates.

The above results also demonstrate the utility of the bond rescaling procedure in handling perturbations of harmonic bond potentials. For example, the addition of an extra methyl group required only 1–2 intermediate states for both FXa and thrombin, while Hsp90 needed 3 intermediate λ windows for the addition of an ethyl group. In cases where the bond rescaling procedure could not be used to handle all bonds, the number of required intermediates can be significantly higher—e.g. between 9 and 11 for PTP1B. In any case, however, the number of these intermediates is always optimal, as determined by the FAST protocol optimisation procedure. In this way, FAST maximises the mobility in λ space, as well as the amount of sampling at the endstates. The only case where the optimisation procedure failed to provide improvement was for the very long explicitly enhanced PTP1B protocols, which had over 30 starting intermediate λ values predicted by the initial AASMC routine. Alternative optimisation algorithms might need to be considered in the future for such challenging cases.

The significant advantages of FAST/MBAR over conventional AFE calculation methods are clear: the completely automated procedure which does not depend on a provided alchemical protocol guarantees the optimality of the free energy calculation, while the single-trajectory nature of the method provides maximum long-timescale decorrelation and thus robustness to the choice of initial coordinates. In contrast, there is always a trade-off between the number of λ values and the simulation time per intermediate in regular AFE calculations—too few intermediates will lead to poor free energy estimation due to insufficient phase space overlap, while too many of them will result in decreased computational time per λ window, and therefore increased correlation and bias to their initial coordinates.

While FAST/MBAR eliminates the most significant weakness of traditional AFE methods, it is important to note that this comes at a price. Firstly, FAST/MBAR does not result in equal sampling over λ space at a finite number of samples, meaning that

the amount of time spent in a particular λ state is not known *a priori*. This also means that FAST/MBAR is particularly sensitive to kinetic barriers in λ space or slow events which can shift the free energy profile over time. Finally, there is an extra computational cost associated with the protocol optimisation procedure, as well as the need to estimate free energy values during the simulation.

Arguably the most significant of these weakness is the sensitivity towards kinetic barriers in λ space, since addressing it would significantly improve the sampling homogeneity in λ space, as the free energy values start converging. Similarly, the extra computational cost can be either tackled by performing protocol and free energy updates less frequently, or by using an implementation which performs these in the background on the central processing unit (CPU), while the simulation is running on the graphics processing unit (GPU). Since GPUs are becoming increasingly more widely used in molecular simulations,[162,233,380,381] this effectively means that the extra CPU overhead posed by FAST could be made practically negligible with a sufficiently optimised implementation.

Future work should therefore focus on improving the robustness of the algorithm towards systems exhibiting high kinetic barriers in λ space. Nevertheless, such barriers are not expected for relatively simple alchemical perturbations, meaning that FAST/MBAR is already a suitable method for performing automated and robust free energy calculations. In addition to its black-box nature, FAST/MBAR is also extremely general and can be readily extended to more sophisticated applications, such as combining free energy calculations with several types of enhanced sampling over different amino acid protonation states. This is another promising avenue for future research which will increase the reliability of free energy calculations performed by FAST/MBAR even further.

10.6 Conclusion

In this chapter the fully adaptive simulated tempering (FAST) method presented in Chapter 9 was extended to the context of relative free energy calculations. The resulting method, FAST/MBAR, is highly efficient and enables the automation of alchemical free energy calculations in a robust way, as demonstrated on four different protein–ligand systems.

It was shown that combining FAST/MBAR with enhanced sampling can decrease the short-timescale bias of the free energy values towards the initial structure at the cost of reduced mobility in λ space. On the other hand, free energy protocols without extra sampling retain this bias at longer timescales. In both cases, however, many ligand rare events can be readily explored due to the single-trajectory nature of the method.

In conjunction with its black-box nature, this makes FAST/MBAR a more reliable alternative to standard AFE calculation methods.

The above results also illustrate how large perturbations and/or orthogonal slow motions can have a detrimental effect on the efficiency of FAST/MBAR in λ space, as well as the convergence of the free energies. Although this consideration is not exclusive to FAST/MBAR, and is instead applicable to all alchemical/tempering methods, it is still an important topic to address in future work. In addition, the flexibility of the FAST framework can be used for enhancing the sampling of many different parts of the system over different protonation states. Combining this with free energy calculations is another area of interest for further research.

Chapter 11

Conclusions and Further Directions

The original aim of this thesis was to investigate the robustness of alchemical binding free energy calculations and their sensitivity to various choices made by the researcher. After exploring the impact of some of these decisions in the first half of the project, however, it became clear that incorporating enhanced sampling algorithms into free energy calculations is not an option, but rather a necessity for performing reproducible free energy calculations. Therefore, the second part of the project focused on the development of maximally automatable and robust enhanced sampling methods which can be naturally incorporated into free energy workflows. In this way, many of the implicit biases made by the researcher can be converted into observable variance, which in turn improves the validity of any successive statistical analysis.

In Chapter 2, the theoretical underpinnings of molecular simulations relevant to all of the following chapters were presented in a concise manner. This chapter focused on the most fundamental concepts in statistical mechanics, molecular dynamics and alchemical free energy (AFE) calculations. Further chapter-specific theoretical considerations were then presented in each of the following chapters, as required.

The review in Chapter 3 presented the current state of the literature regarding the reproducibility of AFE calculations. It showed that a multitude of factors and decisions present in both system setup and the subsequent simulation and analysis can significantly influence the outcome of the calculated free energy values. Many of these discrepancies have been shown to be dampened by the appropriate use of enhanced sampling methods, giving the reader a first glimpse of their importance in molecular dynamics (MD) simulations.

In Chapter 4, a free energy automation software, ProtoCaller, was developed and presented. This library was central to performing the high-throughput free energy benchmarking work in Chapters 5 and 6 by both saving researcher time and eliminating random human errors. Moreover, it allowed the fine control of many parts

of the workflow, thereby enabling the robust experimental design of these studies. In this way, ProtoCaller can be straightforwardly used for other similar studies in the future without much additional manual manipulation.

The study presented in Chapter 5 was the first in the literature to explore the influence of the initial coordinates on protein-ligand binding free energies in such detail and it demonstrated that even highly similar protein crystal structures can require a significant amount of time to decorrelate, thereby influencing any calculated observables. It also unambiguously shows that long-timescale dynamics, primarily in the form of rare ligand transitions, are crucial in binding free energy calculations—a reality which is often ignored in the field for the sake of practicality. As a result, Chapter 5 justifies the use of enhanced sampling methods in free energy calculations and highlights the necessity of a robust automatable enhanced sampling algorithm of specific degrees of freedom.

Chapter 6 expanded on the observation in Chapter 5 that the system preparation procedure is also dependent on the initial crystal structure used. This can have knock-on effects on the automatic assignment of protonation, tautomeric and rotameric (PTR) states of the side chains. The study performed in this chapter shows that different PTR states can visibly affect the free energy calculation even at large distances (longer than 2.0 nm). However, the magnitude of these discrepancies is dependent on four competing factors: inter-replicate variance, side-chain mobility, the polarity of the alchemically perturbed group and ligand–histidine distance. Although the discrepancies arising from different PTR states appear to be less significant than those observed in Chapter 5, it is expected that multiple ambiguous residues can have an amplifying effect on the free energy inconsistencies. Therefore, this chapter once again showed the importance of long-timescale decorrelation in free energy calculations, as well as the utility of expanded ensemble methods in handling multiple Hamiltonians at the same time.

Chapter 7 gave a concise overview of the main classes of tempering-based enhanced sampling methods developed over the years and summarised the basic elements of each of seven different enhanced sampling methods. These were then critically compared by discussing their most significant differences between them and their applicability in the enhanced sampling of specific degrees of freedom in the context of AFE calculations. In this chapter, it was concluded that simulated tempering (ST) and replica exchange molecular dynamics (REMD) are likely to be the most generally applicable enhanced sampling methods, especially in the case of alchemical transformations, while sequential Monte Carlo (SMC) is best used for exploratory sampling simulations without any initial system-specific knowledge. It was also discussed that ST is also much better suited to on-the-fly optimisation than REMD, while also providing maximum long-timescale decorrelation, thereby justifying the focus of the following chapters on SMC and ST.

The adaptive explorative sampling of a relevant subset of the torsional and centre-of-mass (COM) degrees of freedom using adaptive alchemical sequential Monte Carlo (AASMC) was presented in Chapter 8. The method, which is a combination of recent statistical literature and AFE calculation practices, readily outputs a population of states, an intermediate alchemical protocol and a free energy estimate between the intermediate states without requiring any system-specific input. All of its hyperparameters are system-independent and they have proven to be applicable to a range of systems of practical interest. However, as discussed in Chapter 7, the irreversible sequential nature of the method hinders its use as a sampling method and in these settings single-step importance sampling methods, such as REMD and ST are more desirable.

The obvious way to improve on the weaknesses of both SMC and long-timescale sampling methods is to combine them, so that the advantages of both methods synergise to create a robust enhanced sampling method. This idea was explored in Chapter 9, where an initial exploratory AASMC run was shown to give sufficiently good intermediate protocols and free energy estimates, which can be improved in an on-the-fly fashion during the course of the simulation. This procedure is best suited to ST, due to the dynamic nature of the adaptation in alchemical space, resulting in fully adaptive simulated tempering (FAST). The novel approach used to iteratively optimise round-trip times using a model based on the multistate Bennett acceptance ratio (MBAR) results in the asymptotic optimality of the method regardless of the hyperparameters used to initialise the preliminary AASMC run. This makes FAST a practically parameter-free method, where the only impactful decision made by the researcher is the nature and connectivity of the Markov chain explored by the method. This is of course dependent on the purpose of the simulation, meaning that FAST is as close as possible to a “black-box” enhanced sampling method.

The automatable nature of FAST makes it a general-purpose method for exploring an arbitrary Markov chain of states of interest. This notion has far-reaching implications, which were partially explored in Chapter 10, where relative binding free energy calculations were combined with enhanced sampling in an automated fashion using the methodology presented in Chapter 9. This chapter showed that the optimal number of intermediate states can be as low as one intermediate state for the simplest of alchemical perturbations, showing that protocols employing a higher number of intermediates are significantly less efficient. Similarly, more complex perturbations are automatically assigned more intermediate states without any prior information, thereby removing the need of any validation of the alchemical protocol before the simulation. This makes FAST a powerful procedure not only for improving sampling, but also for automatically determining a sufficiently overlapping sequence of distributions which can be used to perform AFE calculations.

Chapter 10 also showed that although enhanced sampling naturally decreases the bias of the initial coordinates both by means of targeted softening of particular energy terms and natural long-timescale decorrelation provided by MD, much of this bias is converted into variance, which can be observed by running multiple repeats. Although this increase in variance is not desirable, it showcases the propensity of AFE calculations, as commonly performed, for exhibiting false convergence. This can have the unwanted effect of obtaining wrong results from any subsequent statistical analysis due to the significant underestimation of the variance. Since variance is much easier to detect than bias without any prior knowledge, it therefore follows that even when convergence is apparently worsened by the addition of enhanced sampling, it is preferable to employ an automatable enhanced sampling method.

The results from Chapter 10 immediately suggest the possibility for combining any of: AFE perturbations of a ligand network, enhanced sampling of ligand and side-chain degrees of freedom, exploring multiple tautomeric and protonation states, and sampling water networks. Apart from the technical challenges involved in implementing these ideas, the main point of consideration for the design of such a general sampling protocol is the construction of the underlying Markov chain, whose structure will have a direct impact on the viability of the method. Since each connection in this Markov chain corresponds to a separate FAST procedure, the complexity of such a method will quickly increase with the number of connected states. Reducing this complexity will likely involve different approaches, such as minimising the number of these connections, as well as employing approximate optimisation procedures, where several connections are forced to employ the same alchemical protocol. Nevertheless, this is a promising venture which will tackle many of the implicit biases during system setup, including the ones investigated in Chapter 6, thereby improving the reproducibility and reliability of AFE calculations.

Chapters 9 and 10 also showed that even an adaptive alchemical protocol can exhibit a very low efficiency, caused by kinetic barriers in alchemical space. These can result from orthogonal slow modes, as well as a suboptimal functional form of the interpolative potential. This warrants further research into the viability of different soft-core potentials, as well as different types of alchemical protocols. In a more general sense, phase space overlap can be further improved by combining the other major class of enhanced sampling methods, namely restraint-based methods, with the FAST procedure. While an appropriate restraint potential should significantly improve the mobility in alchemical space of kinetically trapped systems, it is very likely that these biases will be highly system-dependent. Therefore, further research into adaptive restraint-based methods and their combination with FAST could also help improve the reliability of FAST and tempering methods in general.

Appendix A

Chapter 5: Initial Crystal Structures

A.1 Dihydrofolate Reductase (DHFR)

Metric	PDB ID							
	1OHJ	2W3M	3GHW	4DDR	4M6J	5HPB	6A7E	6DAV
RMSD (Å)	0.568	0.483	0.253	0.230	0.806	0.000	0.210	0.298
Resolution (Å)	2.5	1.6	1.24	2.05	1.201	1.65	1.85	1.55
Year of deposition	1997	2008	2009	2012	2013	2016	2018	2018
Clashscore	38	4	8	9	4	11	4	3
Ramachandran outliers	2.2%	0%	0%	0.5%	0%	0%	0%	0%
Side-chain outliers	14.7%	0.6%	2.4%	1.8%	0%	3.6%	0.6%	0.3%
No. of chains	1	2	1	1	1	1	1	2
Total no. of residues	186	374	186	186	187	186	186	374
Cofactor	NADPH	NADPH	NADPH	NADPH	NADPH	NADPH	NADPH	$NADP^+$

TABLE A.1: The different crystal structures used alongside with some metrics: root-mean-square deviation (RMSD) after alignment to 5HPB using PyMOL,[8] resolution (lower is better), year of deposition, clashscore (lower is better), Ramachandran outliers (lower is better), side-chain outliers (lower is better), number of chains, total number of residues and cofactor used.

A.2 Protein Tyrosine Phosphatase 1B (PTP1B)

	PDB ID							
Metric	1BZJ	1NWE	2AZR	2F71	2H4K	2NTA	2QBP	2ZN7
RMSD (Å)	0.257	0.303	0.196	0.232	0.224	0.223	0.000	0.211
Resolution (Å)	2.25	3.1	2	1.55	2.3	2.1	2.5	2.1
Year of deposition	1998	2003	2005	2005	2006	2006	2007	2008
Clashscore	6	6	1	9	5	9	12	4
Ramachandran outliers	0.3%	0.3%	0.3%	1.0%	0.3%	0.3%	0.7%	0.7%
Side-chain outliers	1.9%	4.7%	0%	1.1%	2.2%	1.5%	6.7%	2.2%
No. of chains	1	1	1	1	1	1	1	1
Total no. of residues	297	298	299	298	299	299	299	299

TABLE A.2: The different crystal structures used alongside with some metrics: RMSD after alignment to 2QBP using PyMOL, resolution (lower is better), year of deposition, clashscore (lower is better), Ramachandran outliers (lower is better), side-chain outliers (lower is better), number of chains and total number of residues.

A.3 Coagulation Factor Xa (FXa)

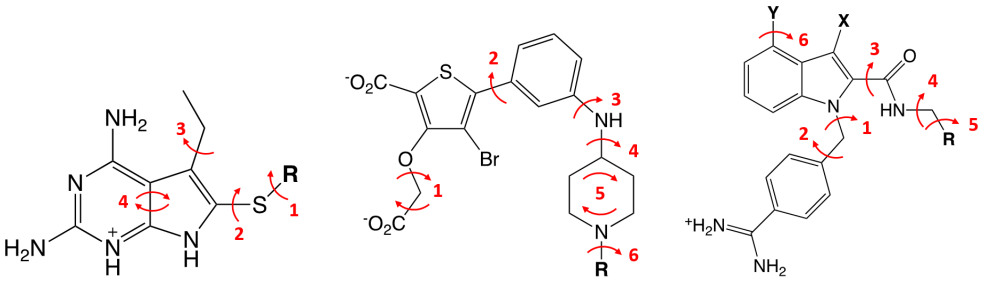
	PDB ID							
Metric	1EZQ	1KSN	1LQD	1NFW	2CJI	2J38	2J95	4Y71
RMSD (Å)	0.249	0.264	0.000	0.260	0.289	0.321	0.304	0.240
Resolution (Å)	2.2	2.1	2.7	2.1	2.1	2.1	2.01	1.8
Year of deposition	2000	2002	2002	2002	2006	2006	2006	2015
Clashscore	17	16	11	10	1	2	1	3
Ramachandran outliers	0%	0.4%	0.7%	0.4%	0%	0%	0%	0%
Side-chain outliers	15.3%	9.1%	8.6%	12%	0.5%	1.7%	1.2%	0.8%
No. of chains	2	2	2	2	2	2	2	2
Total no. of residues	388	388	388	388	388	388	388	388

TABLE A.3: The different crystal structures used alongside with some metrics: RMSD after alignment to 1LQD using PyMOL, resolution (lower is better), year of deposition, clashscore (lower is better), Ramachandran outliers (lower is better), side-chain outliers (lower is better), number of chains and total number of residues.

Appendix B

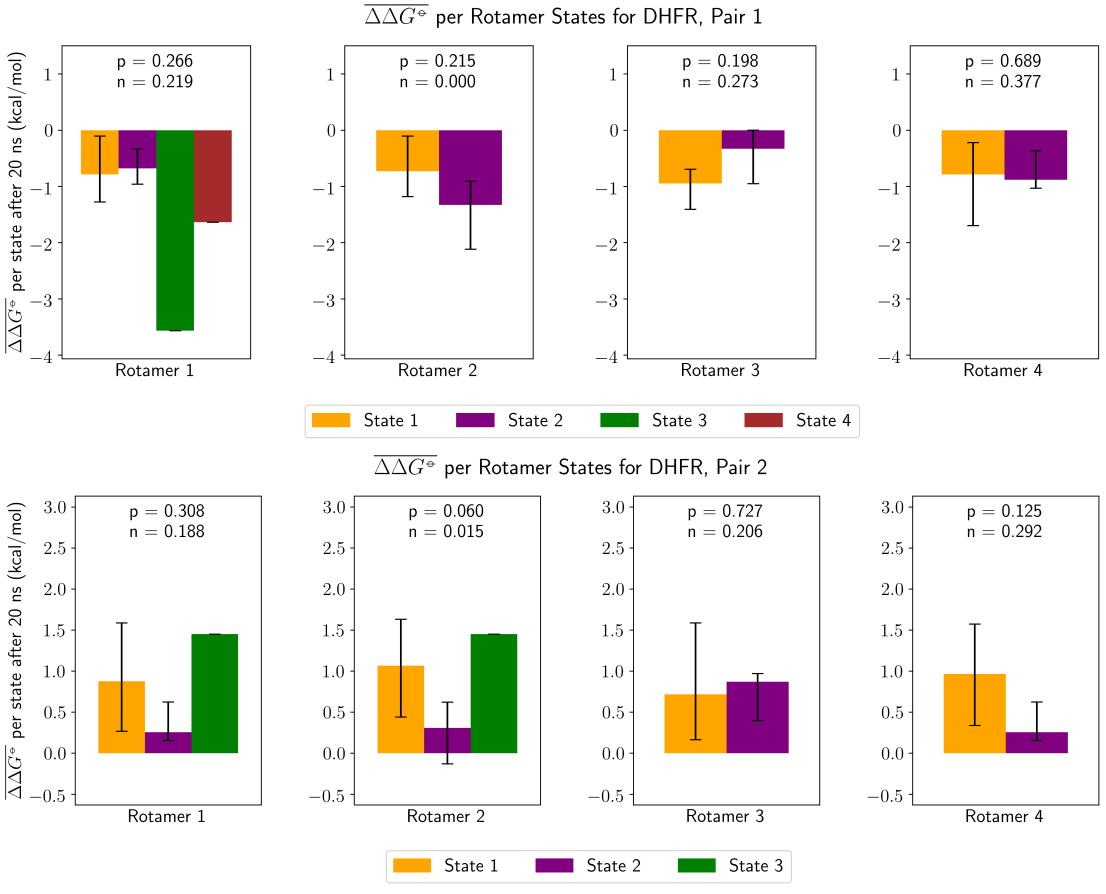
Chapter 5: Long-Timescale Torsional Analysis

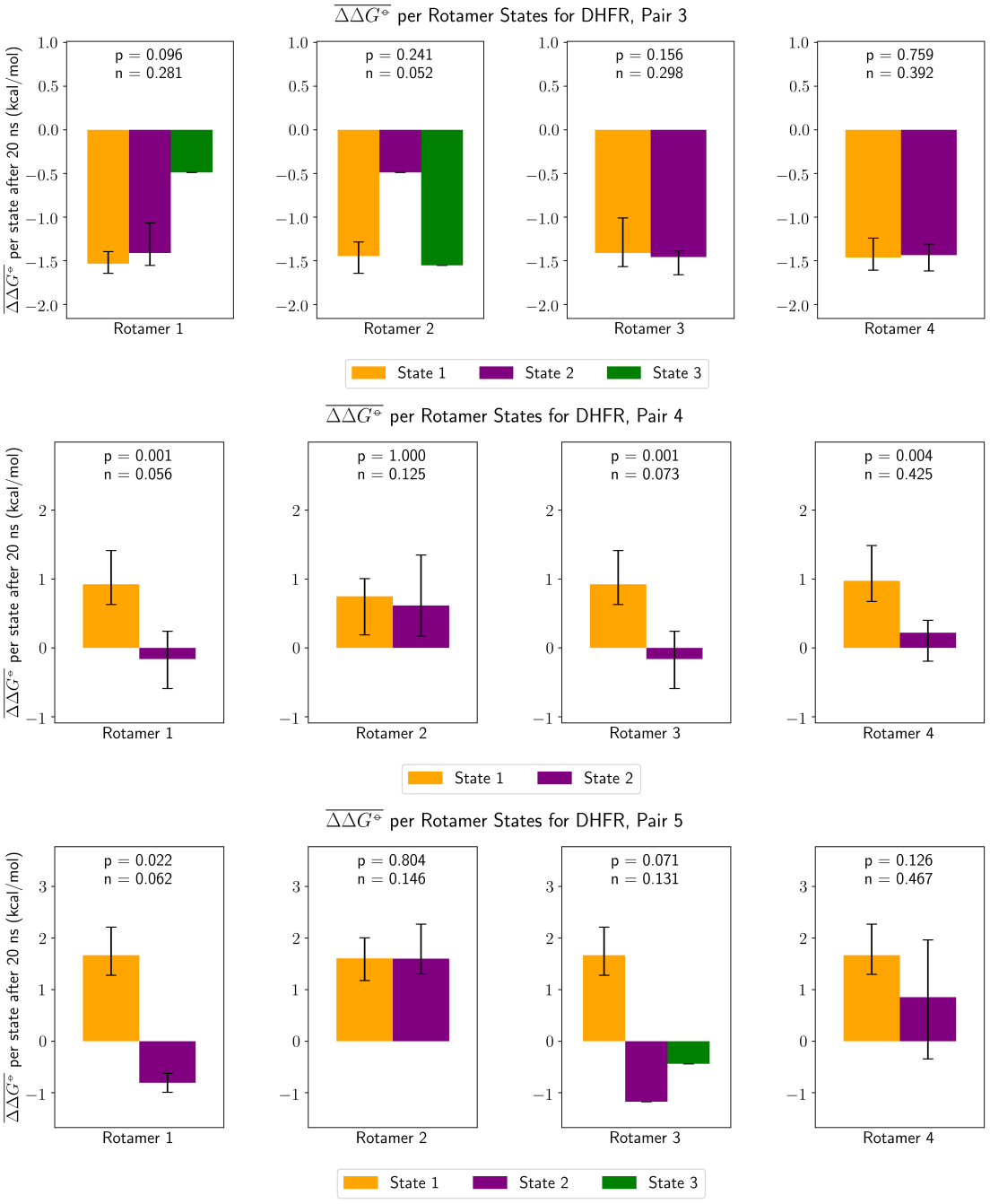
In this section we perform dihedral angle clustering of every relevant rotatable bond of each ligand in Chapter 5 at the final frame of the 20 ns equilibration (Figure B.1). Each of the following bar plots represents a single rotatable bond ("Rotamer") and each bar represents a cluster of the torsional profile ("State"). The height of each bar represents the median free energy across all protein crystal structures and repeats, and the error bars represent the 25th and the 75th percentile of these values. The clustering method in all cases is the mean shift algorithm[382] with a bandwidth automatically determined from the 30th percentile using the `estimate_bandwidth` routine in scikit-learn.[238] The non-parametric Kruskal–Wallis[223] method has been used for testing the null hypothesis that all clusters produce free energy values drawn from the same distribution and the corresponding p-value has been shown on each plot as "p". Any inter-cluster transitions at a transition time of 0.5 ns have also been measured and the average number of transitions per nanosecond has been shown on each plot as "n". In some cases there are observed transitions but all final states correspond to a single cluster. These have been included solely for completion purposes.



(A) DHFR (B) PTP1B (C) FXa

FIGURE B.1: The relevant rotatable bonds for each ligand across all three protein systems. Most of these represent torsional rotations, except for the following: DHFR, No. 4—slight asymmetric ring puckering due to suboptimal force field parameters; PTP1B, No. 1—a concerted twist of two dihedrals inside the binding pocket; PTP1B, No. 5—a ring flip. Rotations within the substituent are referred to as: No. 5 (DHFR); No. 7 (PTP1B); No. 6/7, depending on the presence of a hydroxyl group at substituent **Y** (FXa).




$\overline{\Delta\Delta G^{\circ}}$ per Rotamer States for DHFR, Pair 3
$\overline{\Delta\Delta G^{\circ}}$ per state after 20 ns (kcal/mol)
p = 0.096
n = 0.281
p = 0.241
n = 0.052
p = 0.156
n = 0.298
p = 0.759
n = 0.392
Rotamer 1
Rotamer 2
Rotamer 3
Rotamer 4
State 1
State 2
State 3
$\overline{\Delta\Delta G^{\circ}}$ per Rotamer States for DHFR, Pair 4
$\overline{\Delta\Delta G^{\circ}}$ per state after 20 ns (kcal/mol)
p = 0.001
n = 0.056
p = 1.000
n = 0.125
p = 0.001
n = 0.073
p = 0.004
n = 0.425
Rotamer 1
Rotamer 2
Rotamer 3
Rotamer 4
State 1
State 2
$\overline{\Delta\Delta G^{\circ}}$ per Rotamer States for DHFR, Pair 5
$\overline{\Delta\Delta G^{\circ}}$ per state after 20 ns (kcal/mol)
p = 0.022
n = 0.062
p = 0.804
n = 0.146
p = 0.071
n = 0.131
p = 0.126
n = 0.467
Rotamer 1
Rotamer 2
Rotamer 3
Rotamer 4
State 1
State 2
State 3

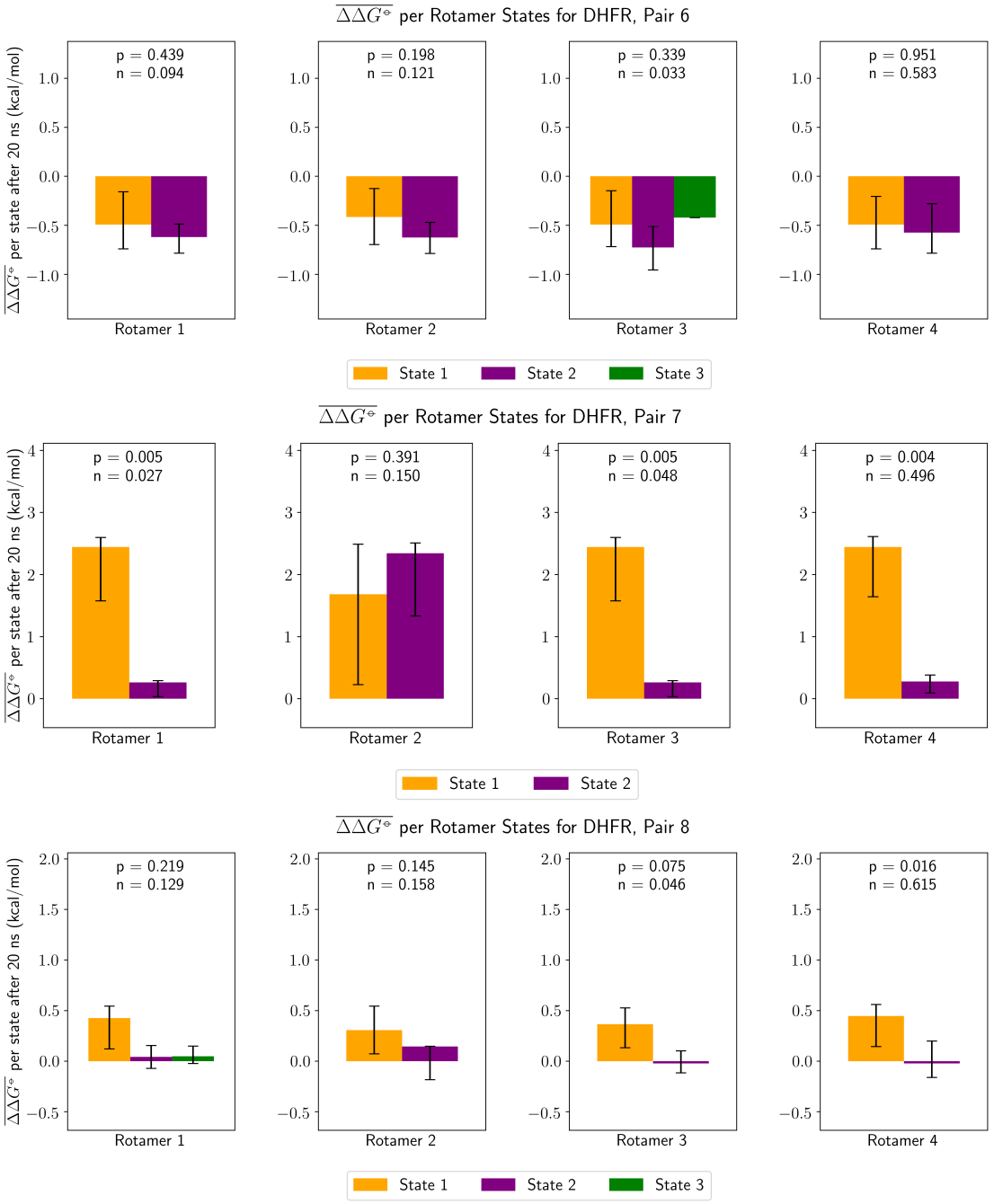
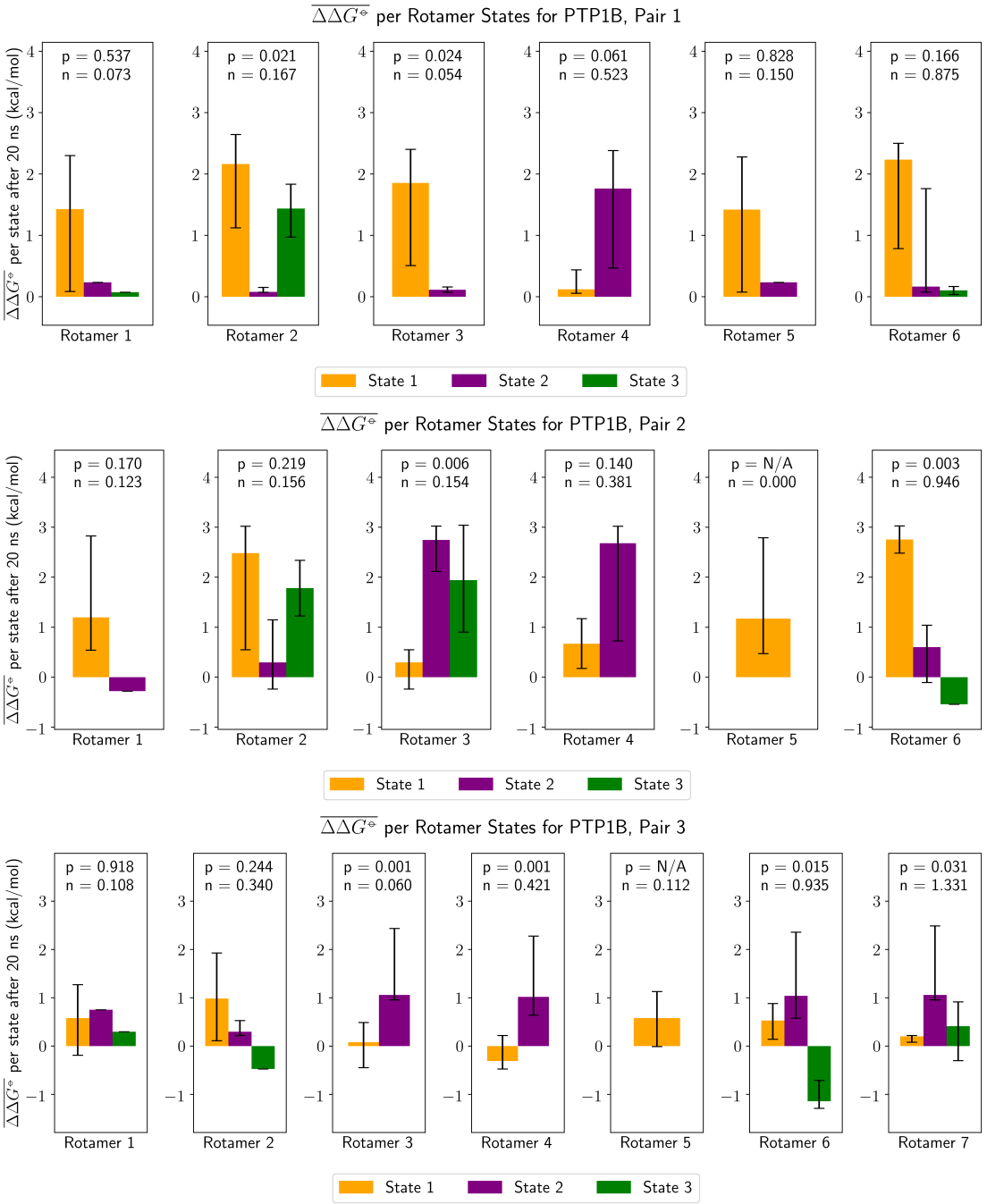
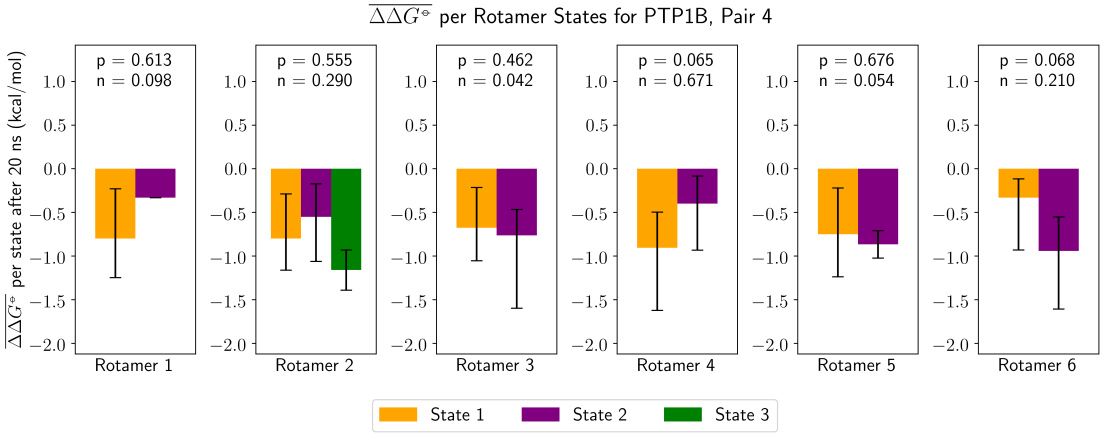
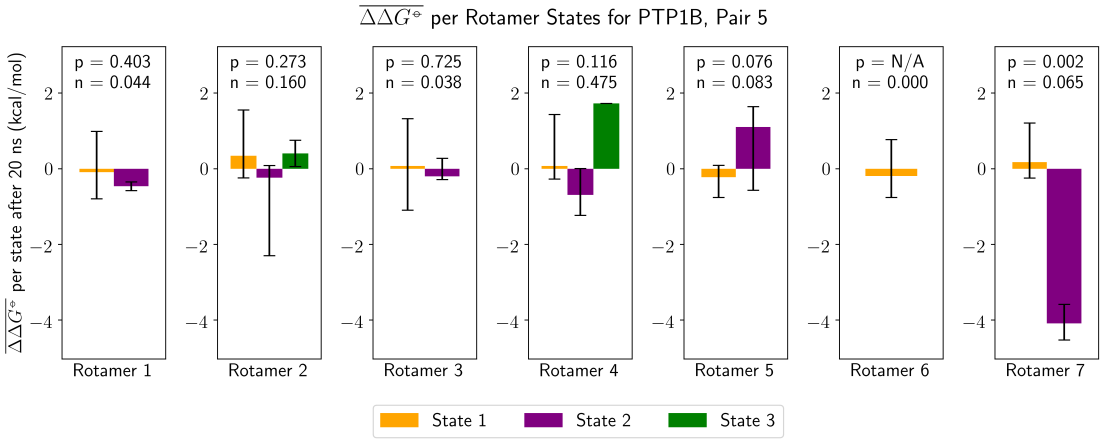
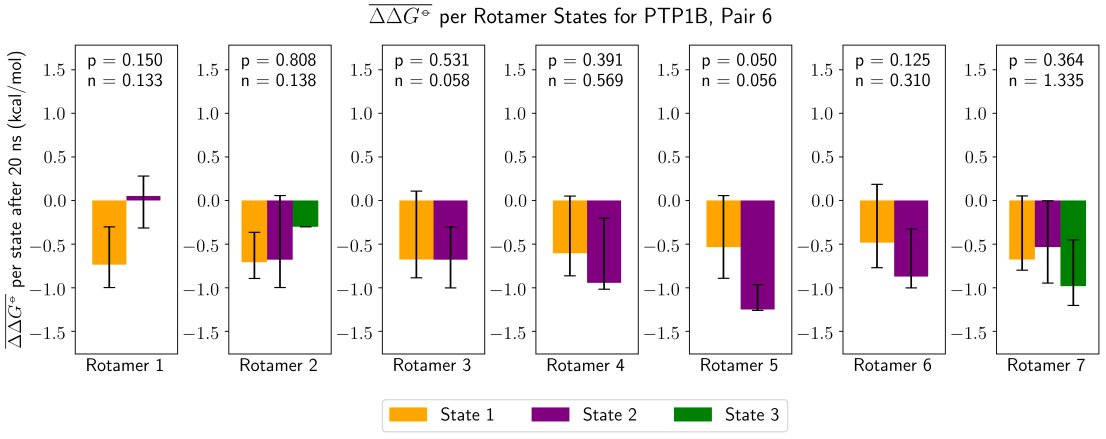


FIGURE B.2: Dihedral profiles of DHFR. All analysis has been performed as described in the main text.


$\overline{\Delta\Delta G^{\circ}}$ per Rotamer States for PTP1B, Pair 1
$\overline{\Delta\Delta G^{\circ}}$ per state after 20 ns (kcal/mol)
p = 0.537
n = 0.073
p = 0.021
n = 0.167
p = 0.024
n = 0.054
p = 0.061
n = 0.523
p = 0.828
n = 0.150
p = 0.166
n = 0.875
Rotamer 1
Rotamer 2
Rotamer 3
Rotamer 4
Rotamer 5
Rotamer 6
State 1
State 2
State 3
$\overline{\Delta\Delta G^{\circ}}$ per Rotamer States for PTP1B, Pair 2
$\overline{\Delta\Delta G^{\circ}}$ per state after 20 ns (kcal/mol)
p = 0.170
n = 0.123
p = 0.219
n = 0.156
p = 0.006
n = 0.154
p = 0.140
n = 0.381
p = N/A
n = 0.000
p = 0.003
n = 0.946
Rotamer 1
Rotamer 2
Rotamer 3
Rotamer 4
Rotamer 5
Rotamer 6
State 1
State 2
State 3
$\overline{\Delta\Delta G^{\circ}}$ per Rotamer States for PTP1B, Pair 3
$\overline{\Delta\Delta G^{\circ}}$ per state after 20 ns (kcal/mol)
p = 0.918
n = 0.108
p = 0.244
n = 0.340
p = 0.001
n = 0.060
p = 0.001
n = 0.421
p = N/A
n = 0.112
p = 0.015
n = 0.935
p = 0.031
n = 1.331
Rotamer 1
Rotamer 2
Rotamer 3
Rotamer 4
Rotamer 5
Rotamer 6
Rotamer 7
State 1
State 2
State 3


$\overline{\Delta\Delta G^{\circ}}$ per Rotamer States for PTP1B, Pair 4
$\overline{\Delta\Delta G^{\circ}}$ per state after 20 ns (kcal/mol)
p = 0.613
n = 0.098
p = 0.555
n = 0.290
p = 0.462
n = 0.042
p = 0.065
n = 0.671
p = 0.676
n = 0.054
p = 0.068
n = 0.210
Rotamer 1
Rotamer 2
Rotamer 3
Rotamer 4
Rotamer 5
Rotamer 6
State 1
State 2
State 3


$\overline{\Delta\Delta G^{\circ}}$ per Rotamer States for PTP1B, Pair 5
$\overline{\Delta\Delta G^{\circ}}$ per state after 20 ns (kcal/mol)
p = 0.403
n = 0.044
p = 0.273
n = 0.160
p = 0.725
n = 0.038
p = 0.116
n = 0.475
p = 0.076
n = 0.083
p = N/A
n = 0.000
p = 0.002
n = 0.065
Rotamer 1
Rotamer 2
Rotamer 3
Rotamer 4
Rotamer 5
Rotamer 6
Rotamer 7
State 1
State 2
State 3


$\overline{\Delta\Delta G^{\circ}}$ per Rotamer States for PTP1B, Pair 6
$\overline{\Delta\Delta G^{\circ}}$ per state after 20 ns (kcal/mol)
p = 0.150
n = 0.133
p = 0.808
n = 0.138
p = 0.531
n = 0.058
p = 0.391
n = 0.569
p = 0.050
n = 0.056
p = 0.125
n = 0.310
p = 0.364
n = 1.335
Rotamer 1
Rotamer 2
Rotamer 3
Rotamer 4
Rotamer 5
Rotamer 6
Rotamer 7
State 1
State 2
State 3

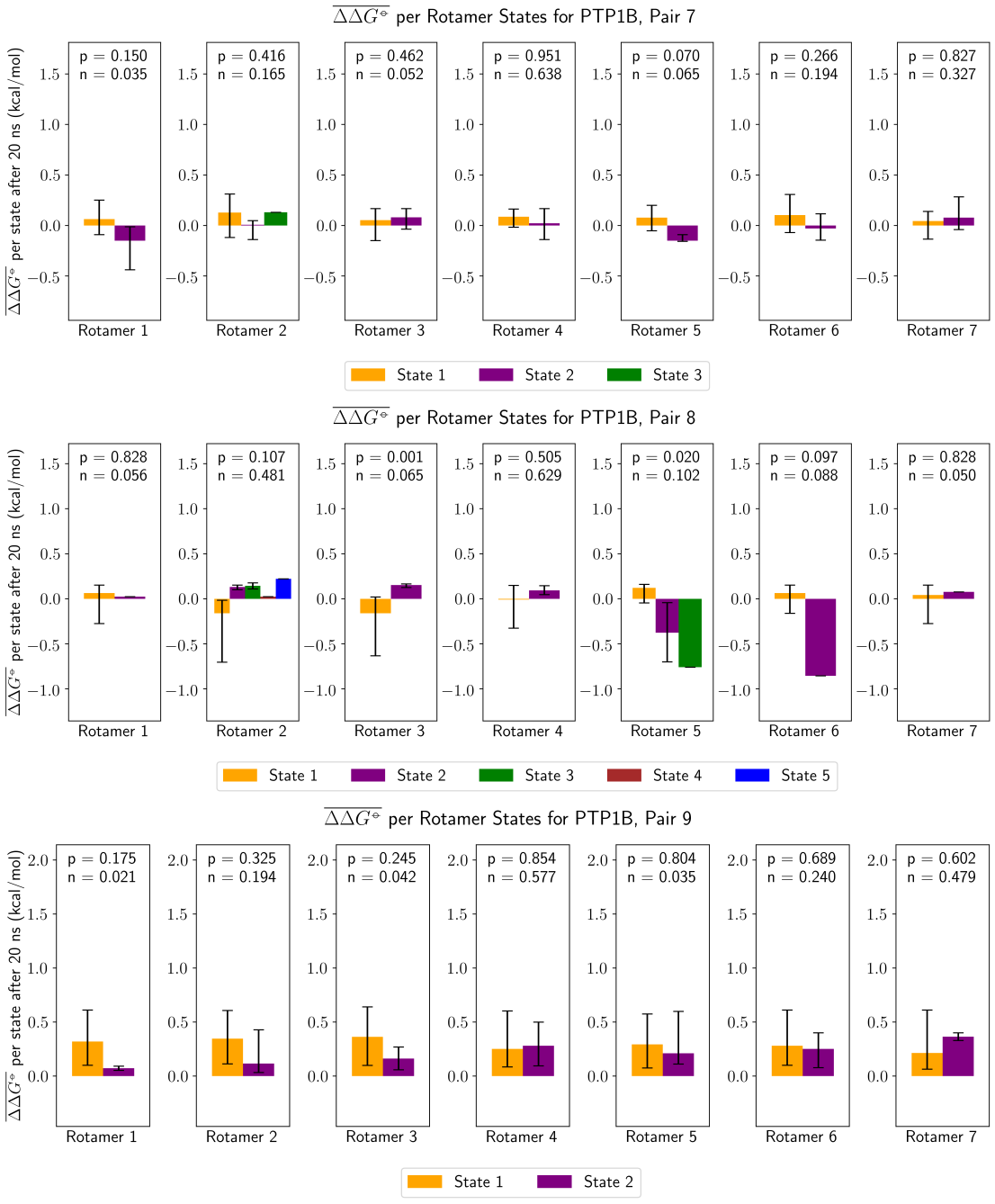
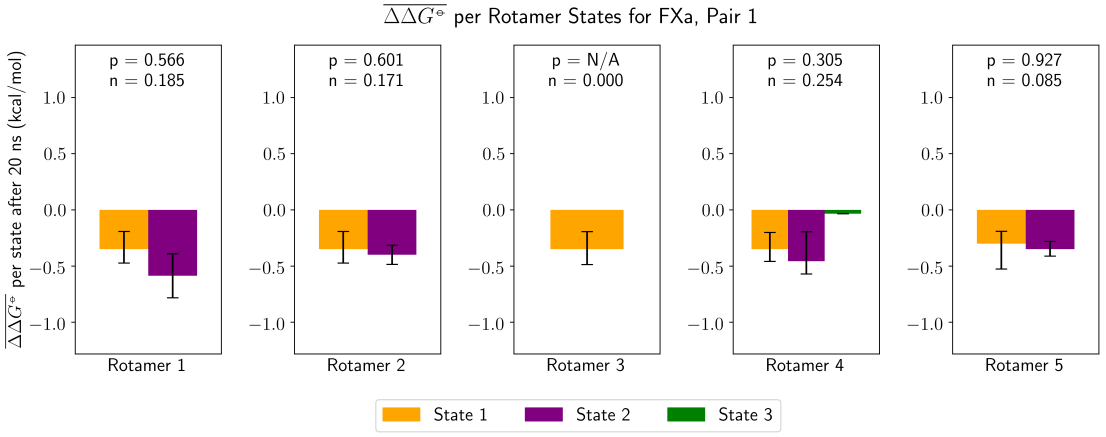
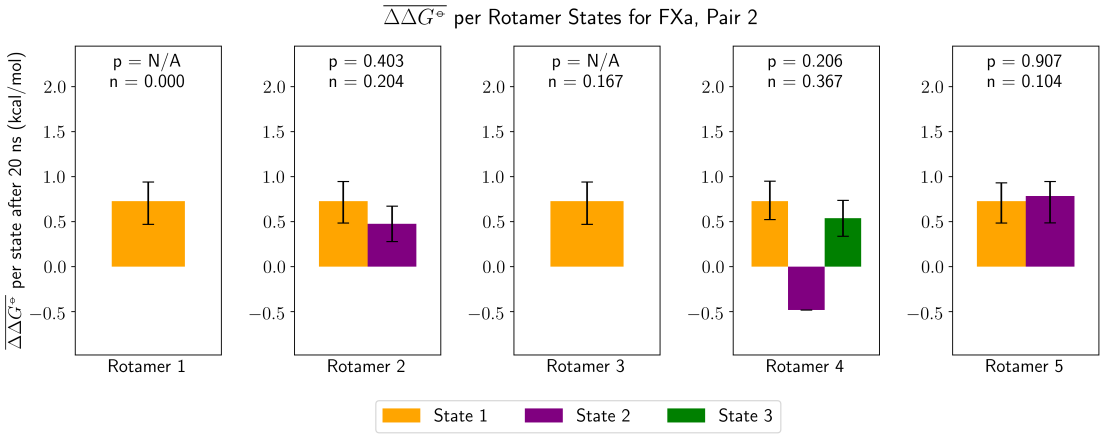
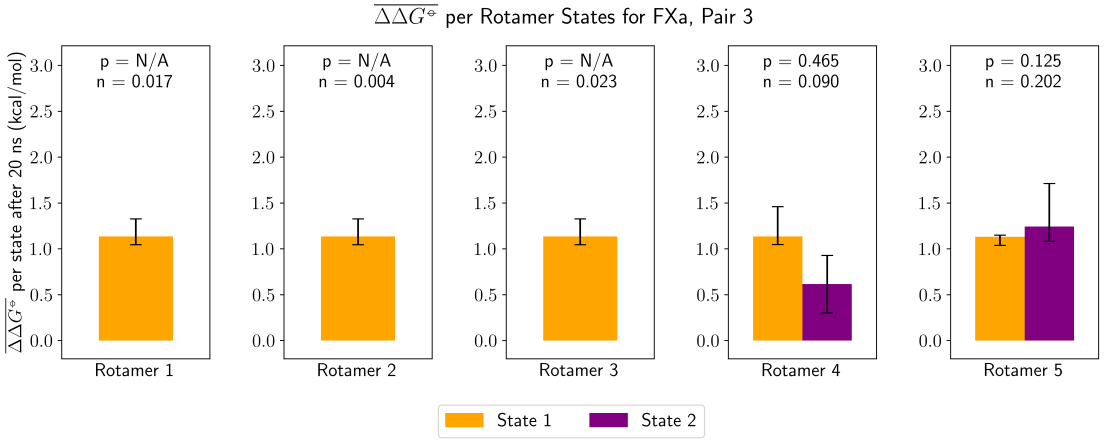
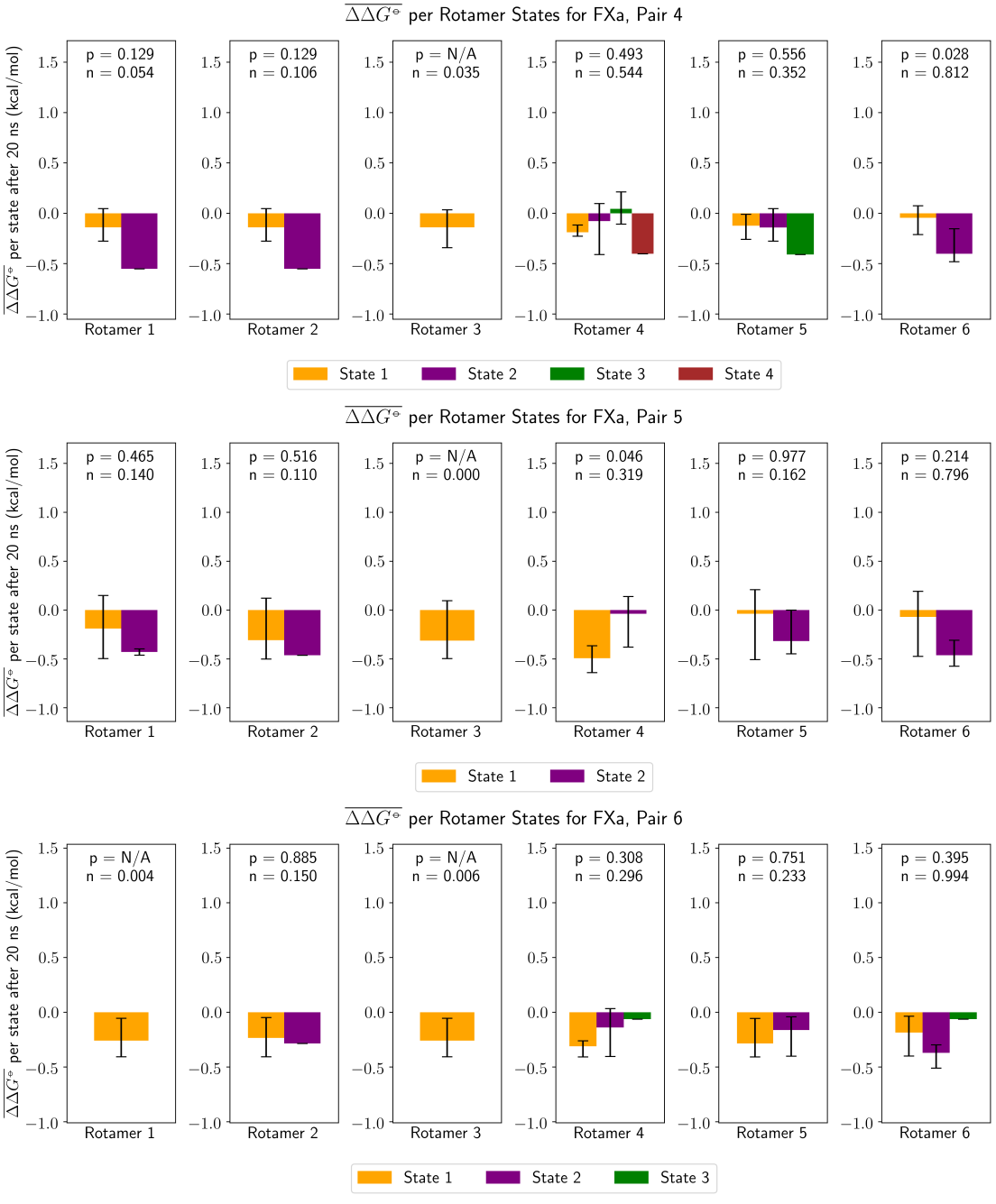


FIGURE B.3: Dihedral profiles of PTP1B. All analysis has been performed as described in the main text.


$\overline{\Delta\Delta G^{\ominus}}$ per Rotamer States for FXa, Pair 1
$\overline{\Delta\Delta G^{\ominus}}$ per state after 20 ns (kcal/mol)
p = 0.566
n = 0.185
p = 0.601
n = 0.171
p = N/A
n = 0.000
p = 0.305
n = 0.254
p = 0.927
n = 0.085
Rotamer 1
Rotamer 2
Rotamer 3
Rotamer 4
Rotamer 5
State 1
State 2
State 3


$\overline{\Delta\Delta G^{\ominus}}$ per Rotamer States for FXa, Pair 2
$\overline{\Delta\Delta G^{\ominus}}$ per state after 20 ns (kcal/mol)
p = N/A
n = 0.000
p = 0.403
n = 0.204
p = N/A
n = 0.167
p = 0.206
n = 0.367
p = 0.907
n = 0.104
Rotamer 1
Rotamer 2
Rotamer 3
Rotamer 4
Rotamer 5
State 1
State 2
State 3


$\overline{\Delta\Delta G^{\ominus}}$ per Rotamer States for FXa, Pair 3
$\overline{\Delta\Delta G^{\ominus}}$ per state after 20 ns (kcal/mol)
p = N/A
n = 0.017
p = N/A
n = 0.004
p = N/A
n = 0.023
p = 0.465
n = 0.090
p = 0.125
n = 0.202
Rotamer 1
Rotamer 2
Rotamer 3
Rotamer 4
Rotamer 5
State 1
State 2


$\overline{\Delta\Delta G^{\ominus}}$ per Rotamer States for FXa, Pair 4
$\overline{\Delta\Delta G^{\ominus}}$ per state after 20 ns (kcal/mol)
p = 0.129
n = 0.054
p = 0.129
n = 0.106
p = N/A
n = 0.035
p = 0.493
n = 0.544
p = 0.556
n = 0.352
p = 0.028
n = 0.812
Rotamer 1
Rotamer 2
Rotamer 3
Rotamer 4
Rotamer 5
Rotamer 6
State 1
State 2
State 3
State 4
$\overline{\Delta\Delta G^{\ominus}}$ per Rotamer States for FXa, Pair 5
$\overline{\Delta\Delta G^{\ominus}}$ per state after 20 ns (kcal/mol)
p = 0.465
n = 0.140
p = 0.516
n = 0.110
p = N/A
n = 0.000
p = 0.046
n = 0.319
p = 0.977
n = 0.162
p = 0.214
n = 0.796
Rotamer 1
Rotamer 2
Rotamer 3
Rotamer 4
Rotamer 5
Rotamer 6
State 1
State 2
$\overline{\Delta\Delta G^{\ominus}}$ per Rotamer States for FXa, Pair 6
$\overline{\Delta\Delta G^{\ominus}}$ per state after 20 ns (kcal/mol)
p = N/A
n = 0.004
p = 0.885
n = 0.150
p = N/A
n = 0.006
p = 0.308
n = 0.296
p = 0.751
n = 0.233
p = 0.395
n = 0.994
Rotamer 1
Rotamer 2
Rotamer 3
Rotamer 4
Rotamer 5
Rotamer 6
State 1
State 2
State 3

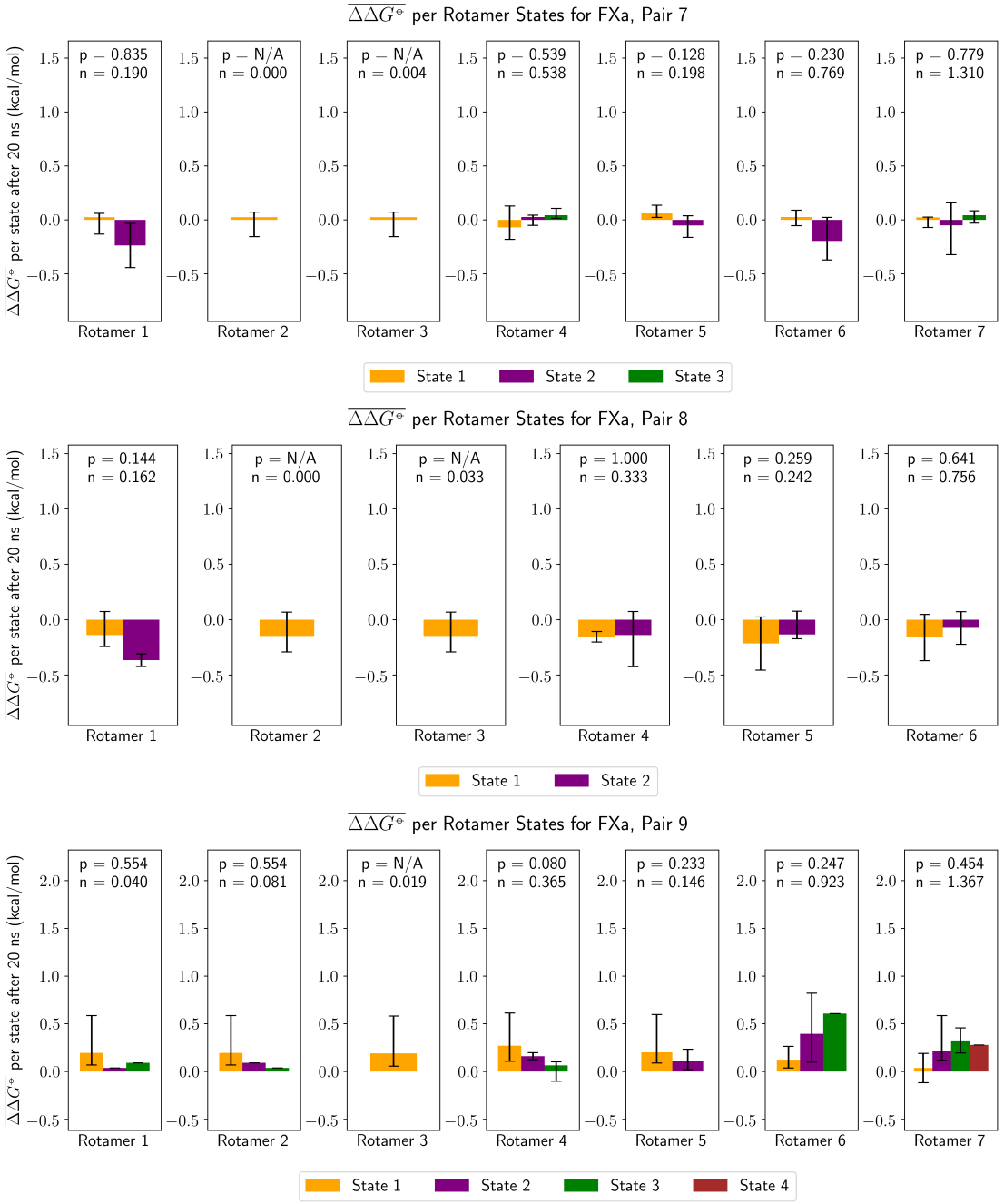


FIGURE B.4: Dihedral profiles of FXa. All analysis has been performed as described in the main text.

Appendix C

Chapter 7: Asymptotic Complexity of Sequential Importance Sampling

Let us assume that the single-step importance sampling ratio $\frac{\pi(\vec{\lambda}_{i_2},\vec{x}_{i_1})}{\pi(\vec{\lambda}_{i_1},\vec{x}_{i_1})}$ can be expressed as a log-normally distributed variable with a mean μ and a variance σ^2. Since its expectation value over $\pi(\vec{\lambda}_{i_1},\vec{x}_{i_1})$ is equal to unity, this imposes the constraint $\mu = -\frac{1}{2}\sigma^2$. This follows from the well-known expression for the average of a log-normally distributed random variable e^X:

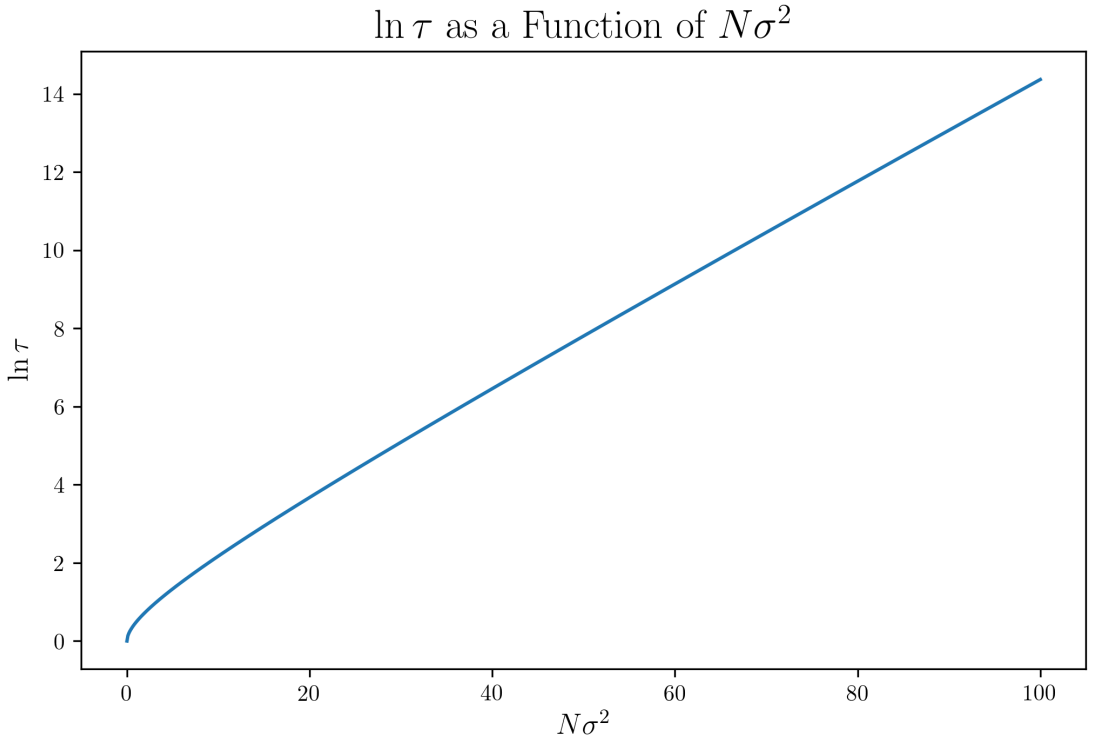
$$\left\langle e^X \right\rangle = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^X e^{-\frac{(X-\mu)^2}{2\sigma^2}} dX = e^{\mu+\frac{1}{2}\sigma^2} \tag{C.1}$$

Assuming independence, the logarithm of the sequential importance sampling ratio of N log-normally distributed single-step importance sampling ratios with the same parameters μ and σ^2 is also a normal variable Y with a variance of $N\sigma^2$ and a mean of $N\mu = -\frac{1}{2}N\sigma^2$. The expectation value of the acceptance criterion $\left\langle \min[1, e^Y] \right\rangle$ is then:

$$\begin{aligned}
\left\langle \min[1, e^Y] \right\rangle &= \frac{1}{\sigma\sqrt{2\pi N}} \int_{-\infty}^{\infty} \min[1, e^Y] e^{-\frac{(Y+\frac{1}{2}N\sigma^2)^2}{2N\sigma^2}} dY \\
&= \frac{1}{\sigma\sqrt{2\pi N}} \int_{-\infty}^{0} e^Y e^{-\frac{(Y+\frac{1}{2}N\sigma^2)^2}{2N\sigma^2}} dY + \frac{1}{\sigma\sqrt{2\pi N}} \int_{0}^{\infty} e^{-\frac{(Y+\frac{1}{2}N\sigma^2)^2}{2N\sigma^2}} dY \\
&= 1 - \operatorname{erf}\left(\frac{\sigma\sqrt{N}}{2\sqrt{2}}\right)
\end{aligned} \tag{C.2}$$

where erf denotes the error function.

The expected transition time per number of transition attempts τ is inversely proportional to $\left\langle \min[1, e^Y] \right\rangle$:



(A)

FIGURE C.1: $\ln \tau$ expressed as a function of $N\sigma^2$.

$$\tau = \frac{1}{\langle \min[1, e^Y] \rangle} \tag{C.3}$$

Expressing $\ln \tau = -\ln \left[1 - \operatorname{erf}\left(\frac{\sigma\sqrt{N}}{2\sqrt{2}}\right)\right]$ as a function of $N\sigma^2$ results in an asymptotically linear relationship (Figure C.1). Therefore, in this simplified case the expected transition time τ increases exponentially with respect to the number of distributions.

Appendix D

Chapter 10: Importance Sampling between Two Normal Distributions

It was stated in Chapter 10 that, when possible, sample transformation can be significantly more efficient in the context of importance sampling than using a series of intermediate distributions without any modifications to the generated samples. A simple example illustrating this argument is the case of two unidimensional normal distributions $\pi_0(x) \equiv \mathcal{N}(\mu_0, \sigma_0)$ and $\pi_1(x) \equiv \mathcal{N}(\mu_1, \sigma_1)$, where we are interested in drawing samples from $\pi_1(x)$ based on samples from $\pi_0(x)$. In this case, one can measure the dissimilarity of these two distributions using the Kullback–Leibler (KL) divergence $K(p_0, \pi_0)$:[383,384]

$$K(\pi_0, \pi_1) = \log \frac{\sigma_1}{\sigma_0} + \frac{\sigma_0^2 + (\mu_0 - \mu_1)^2}{2\sigma_1^2} - \frac{1}{2} \tag{D.1}$$

It is evident that $K(\pi_0, \pi_1)$ is unbounded with respect to the distance between the two means μ_0 and μ_1. In this setting, sequential importance sampling using a smooth perturbation of the parameters will require an increasingly large number of intermediates directly dependent on $|\mu_0 - \mu_1|$. In contrast, we can sample directly from $\pi_1(x)$ given a sample $X_0 \sim \mathcal{N}(\mu_0, \sigma_0)$ without any loss of information if we transform X_0 in the following way:

$$X_1 = \frac{\sigma_1}{\sigma_0}(X_0 - \mu_0 + \mu_1) \tag{D.2}$$

This approach is independent of the parameters μ_0, μ_1, σ_0 and σ_1 and requires no intermediate distributions, making it much more efficient than the regular importance sampling method. Moreover, this transformation is reversible, thereby making it compatible with detailed balance.

Appendix E

Chapter 10: Derivation of the Bond Rescaling Jacobian

The elements of the Jacobian matrix $\mathbf{J}_{\vec{T}}(\vec{x})$ corresponding to a transformation $\vec{T}$ are defined as follows:

$$J_{ij,\vec{T}}(\vec{x}) = \frac{\partial T_i(\vec{x})}{\partial x_j} \tag{E.1}$$

A bond rescaling move between two atoms with coordinates $\vec{r}_0$ and $\vec{r}_1$ by a scaling factor s can be considered as a translation of $\vec{r}_1$ and all atoms bonded to it $\vec{r}_{movable}$ by $(s-1)(\vec{r}_1 - \vec{r}_0)$:

$$\vec{T}(\vec{r}_{movable}) = \vec{r}_{movable} + (s-1)(\vec{r}_1 - \vec{r}_0) \tag{E.2}$$

To calculate the determinant $|\mathbf{J}_{\vec{T}}(\vec{x})|$, we first note that $J_{ij,\vec{T}}(\vec{x}) = 0$ for all row elements $i \neq j$ corresponding to the stationary atoms, since their coordinates remain unchanged. Afterwards, we use the following well-known identity for the determinant of a block matrix $\mathbf{M} = \left(\begin{smallmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{smallmatrix}\right)$:

$$|\mathbf{M}| = |\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}||\mathbf{D}| \tag{E.3}$$

If either **B** or **C** have only zero elements, the determinant reduces to:

$$|\mathbf{M}| = |\mathbf{A}||\mathbf{D}| \tag{E.4}$$

Therefore, all off-diagonal column elements corresponding to the stationary atoms do not contribute to the Jacobian determinant, even if they are not zero. This argument

can be then iteratively applied to conclude that in our setting only the diagonal elements of each atom contribute to the determinant. For the transformation shown in Equation E.2, all diagonal terms are unity except for the three coordinates corresponding to $\vec{r}_1$, where each element is equal to s. Therefore, the Jacobian determinant for a bond rescaling transformation by a scaling factor s is:

$$|\mathbf{J}_{\vec{T}}(\vec{x})| = s^3 \tag{E.5}$$

This procedure can be straightforwardly generalised to multiple concurrent reversible bond rescaling transformations with scaling factors $s_1, ..., s_N$, in which case the Jacobian determinant is:

$$|\mathbf{J}_{\vec{T}}(\vec{x})| = \prod_{i=1}^{N} s_i^3 \tag{E.6}$$

References

[1] M. R. Shirts and D. L. Mobley, in *Biomolecular Simulations: Methods and Protocols*, ed. L. Monticelli and E. Salonen, Humana Press, Totowa, 2013, pp. 271–311.

[2] R. D. Tosso, S. A. Andujar, L. Gutierrez, E. Angelina, R. Rodríguez, M. Nogueras, H. Baldoni, F. D. Suvire, J. Cobo and R. D. Enriz, *Journal of Chemical Information and Modeling*, 2013, **53**, 2018–2032.

[3] D. P. Wilson, Z.-K. Wan, W.-X. Xu, S. J. Kirincich, B. C. Follows, D. Joseph-McCarthy, K. Foreman, A. Moretto, J. Wu, M. Zhu, E. Binnun, Y.-L. Zhang, M. Tam, D. V. Erbe, J. Tobin, X. Xu, L. Leung, A. Shilling, S. Y. Tam, T. S. Mansour and J. Lee, *Journal of Medicinal Chemistry*, 2007, **50**, 4681–4698.

[4] H. Matter, E. Defossa, U. Heinelt, P.-M. Blohm, D. Schneider, A. Müller, S. Herok, H. Schreuder, A. Liesum, V. Brachvogel, P. Lönze, A. Walser, F. Al-Obeidi and P. Wildgoose, *Journal of Medicinal Chemistry*, 2002, **45**, 2749–2769.

[5] M. Mares-Guia, D. L. Nelson and E. Rogana, *Journal of the American Chemical Society*, 1977, **99**, 2331–2336.

[6] J. J. Barker, O. Barker, R. Boggio, V. Chauhan, R. K. Y. Cheng, V. Corden, S. M. Courtney, N. Edwards, V. M. Falque, F. Fusar, M. Gardiner, E. M. N. Hamelin, T. Hesterkamp, O. Ichihara, R. S. Jones, O. Mather, C. Mercurio, S. Minucci, C. A. G. N. Montalbetti, A. Müller, D. Patel, B. G. Phillips, M. Varasi, M. Whittaker, D. Winkler and C. J. Yarnold, *ChemMedChem*, 2009, **4**, 963–966.

[7] H. B. Mann and D. R. Whitney, *The Annals of Mathematical Statistics*, 1947, 50–60.

[8] Schrödinger, LLC, *PyMOL*, 2015, `https://pymol.org`.

[9] J. P. Hughes, S. Rees, S. B. Kalindjian and K. L. Philpott, *British Journal of Pharmacology*, 2011, **162**, 1239–1249.

[10] G. Sliwoski, S. Kothiwale, J. Meiler and E. W. Lowe, *Pharmacological Reviews*, 2014, **66**, 334.

[11] J. G. Kirkwood, *The Journal of Chemical Physics*, 1935, **3**, 300–313.

[12] R. W. Zwanzig, *The Journal of Chemical Physics*, 1954, **22**, 1420–1426.

[13] C. H. Bennett, *Journal of Computational Physics*, 1976, **22**, 245–268.

[14] W. L. Jorgensen and C. Ravimohan, *The Journal of Chemical Physics*, 1985, **83**, 3050–3054.

[15] C. F. Wong and J. A. McCammon, *Journal of the American Chemical Society*, 1986, **108**, 3830–3832.

[16] K. M. Merz and P. A. Kollman, *Journal of the American Chemical Society*, 1989, **111**, 5649–5658.

[17] R. J. Baldock, L. B. Pártay, A. P. Bartók, M. C. Payne and G. Csányi, *Physical Review B*, 2016, **93**, 174108.

[18] M. R. Shirts and J. D. Chodera, *The Journal of Chemical Physics*, 2008, **129**, 124105.

[19] Z. Tan, *Journal of the American Statistical Association*, 2004, **99**, 1027–1036.

[20] X. Ding, J. Z. Vilseck and C. L. Brooks III, *Journal of Chemical Theory and Computation*, 2019, **15**, 799–802.

[21] Z. X. Sun, X. H. Wang and J. Z. Zhang, *Physical Chemistry Chemical Physics*, 2017, **19**, 15005–15020.

[22] S. J. Fox, J. Dziedzic, T. Fox, C. S. Tautermann and C.-K. Skylaris, *Proteins: Structure, Function, and Bioinformatics*, 2014, **82**, 3335–3346.

[23] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman, *Journal of the American Chemical Society*, 1995, **117**, 5179–5197.

[24] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror and D. E. Shaw, *Proteins: Structure, Function, and Bioinformatics*, 2010, **78**, 1950–1958.

[25] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling, *Journal of Chemical Theory and Computation*, 2015, **11**, 3696–3713.

[26] C. Tian, K. Kasavajhala, K. A. A. Belfon, L. Raguette, H. Huang, A. N. Migues, J. Bickel, Y. Wang, J. Pincay, Q. Wu and C. Simmerling, *Journal of Chemical Theory and Computation*, 2020, **16**, 528–552.

[27] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *Journal of Computational Chemistry*, 2004, **25**, 1157–1174.

[28] C. I. Bayly, P. Cieplak, W. Cornell and P. A. Kollman, *The Journal of Physical Chemistry*, 1993, **97**, 10269–10280.

[29] A. Jakalian, B. L. Bush, D. B. Jack and C. I. Bayly, *Journal of Computational Chemistry*, 2000, **21**, 132–146.

[30] A. Jakalian, D. B. Jack and C. I. Bayly, *Journal of Computational Chemistry*, 2002, **23**, 1623–1641.

[31] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *The Journal of Chemical Physics*, 1983, **79**, 926–935.

[32] H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura and T. Head-Gordon, *The Journal of Chemical Physics*, 2004, **120**, 9665–9678.

[33] H. J. C. Berendsen, J. R. Grigera and T. P. Straatsma, *The Journal of Physical Chemistry*, 1987, **91**, 6269–6271.

[34] S. Izadi, R. Anandakrishnan and A. V. Onufriev, *The Journal of Physical Chemistry Letters*, 2014, **5**, 3863–3871.

[35] R. Bryce, *AMBER Parameter Database*, 2001, `https://amber.manchester.ac.uk`, (accessed September 2021).

[36] P. P. Ewald, *Annalen der Physik*, 1921, **369**, 253–287.

[37] T. Darden, D. York and L. Pedersen, *The Journal of Chemical Physics*, 1993, **98**, 10089–10092.

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

[39] M. T. Nelson, W. Humphrey, A. Gursoy, A. Dalke, L. V. Kalé, R. D. Skeel and K. Schulten, *International Journal of High Performance Computing Applications*, 1996, **10**, 251–268.

[40] D. A. Pearlman, D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham, S. DeBolt, D. Ferguson, G. Seibel and P. Kollman, *Computer Physics Communications*, 1995, **91**, 1–41.

[41] S. Liu, L. Wang and D. L. Mobley, *Journal of Chemical Information and Modeling*, 2015, **55**, 727–735.

[42] A. J. Clark, C. Negron, K. Hauser, M. Sun, L. Wang, R. Abel and R. A. Friesner, *Journal of Molecular Biology*, 2019, **431**, 1481–1493.

[43] T. C. Beutler, A. E. Mark, R. C. van Schaik, P. R. Gerber and W. F. van Gunsteren, *Chemical Physics Letters*, 1994, **222**, 529–539.

[44] T. T. Pham and M. R. Shirts, *The Journal of Chemical Physics*, 2011, **135**, 034114.

[45] M. Betancourt, 2017, arXiv:1701.02434.

[46] D. Szász, in *Hard Ball Systems and the Lorentz Gas*, ed. L. A. Bunimovich, D. Burago, N. Chernov, E. G. D. Cohen, C. P. Dettmann, J. R. Dorfman, S. Ferleger, R. Hirschl, A. Kononenko, J. L. Lebowitz, C. Liverani, T. J. Murphy, J. Piasecki, H. A. Posch, N. Simányi, Y. Sinai, D. Szász, T. Tél, H. van Beijeren, R. van Zon, J. Vollmer, L. S. Young and D. Szász, Springer, Berlin, Heidelberg, 2000, pp. 421–446.

[47] M. Tuckerman, B. J. Berne and G. J. Martyna, *The Journal of Chemical Physics*, 1992, **97**, 1990–2001.

[48] J. C. Sexton and D. H. Weingarten, *Nuclear Physics B*, 1992, **380**, 665–677.

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

[50] J.-P. Ryckaert, G. Ciccotti and H. J. Berendsen, *Journal of Computational Physics*, 1977, **23**, 327–341.

[51] B. Hess, H. Bekker, H. J. C. Berendsen and J. G. E. M. Fraaije, *Journal of Computational Chemistry*, 1997, **18**, 1463–1472.

[52] P. Eastman and V. S. Pande, *Journal of Chemical Theory and Computation*, 2010, **6**, 434–437.

[53] S. Miyamoto and P. A. Kollman, *Journal of Computational Chemistry*, 1992, **13**, 952–962.

[54] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. R. Haak, *The Journal of Chemical Physics*, 1984, **81**, 3684–3690.

[55] E. Braun, S. M. Moosavi and B. Smit, *Journal of Chemical Theory and Computation*, 2018, **14**, 5262–5272.

[56] D. J. Evans and B. L. Holian, *The Journal of Chemical Physics*, 1985, **83**, 4069–4074.

[57] G. J. Martyna, M. L. Klein and M. Tuckerman, *The Journal of Chemical Physics*, 1992, **97**, 2635–2643.

[58] N. Goga, A. J. Rzepiela, A. H. de Vries, S. J. Marrink and H. J. C. Berendsen, *Journal of Chemical Theory and Computation*, 2012, **8**, 3637–3649.

[59] J. Fass, D. A. Sivak, G. E. Crooks, K. A. Beauchamp, B. Leimkuhler and J. D. Chodera, *Entropy*, 2018, **20**, 318.

[60] M. Parrinello and A. Rahman, *Journal of Applied Physics*, 1981, **52**, 7182–7190.

[61] N. Metropolis and S. Ulam, *Journal of the American Statistical Association*, 1949, **44**, 335–341.

[62] A. E. Gelfand, *Journal of the American Statistical Association*, 2000, **95**, 1300–1304.

[63] J. E. Besag, *Journal of the Royal Statistical Society*, 1994, **56**, 591–592.

[64] J. P. Nilmeier, G. E. Crooks, D. D. L. Minh and J. D. Chodera, *Proceedings of the National Academy of Sciences of the United States of America*, 2011, **108**, E1009.

[65] S. C. Gill, N. M. Lim, P. B. Grinaway, A. S. Rustenburg, J. Fass, G. A. Ross, J. D. Chodera and D. L. Mobley, *The Journal of Physical Chemistry B*, 2018, **122**, 5579–5598.

[66] D. J. Adams, *Molecular Physics*, 1974, **28**, 1241–1252.

[67] H.-J. Woo, A. R. Dinner and B. Roux, *The Journal of Chemical Physics*, 2004, **121**, 6392–6400.

[68] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding and C.-T. Lin, *Neurocomputing*, 2017, **267**, 664–681.

[69] A. Amadei, A. B. M. Linssen and H. J. C. Berendsen, *Proteins: Structure, Function, and Bioinformatics*, 1993, **17**, 412–425.

[70] L. Molgedey and H. G. Schuster, *Physical Review Letters*, 1994, **72**, 3634–3637.

[71] V. S. Pande, K. Beauchamp and G. R. Bowman, *Protein Folding*, 2010, **52**, 99–105.

[72] F. Noé, H. Wu, J.-H. Prinz and N. Plattner, *The Journal of Chemical Physics*, 2013, **139**, 184114.

[73] P. Metzner, F. Noé and C. Schütte, *Physical Review E*, 2009, **80**, 021106.

[74] J. D. Chodera, P. Elms, F. Noé, B. Keller, C. M. Kaiser, A. Ewall-Wice, S. Marqusee, C. Bustamante and N. S. Hinrichs, 2011, arXiv:1108.1430.

[75] S. Andrilli and D. Hecker, in *Elementary Linear Algebra (Fifth Edition)*, ed. S. Andrilli and D. Hecker, Academic Press, Boston, 2016, pp. 513–605.

[76] S. U. Pillai, T. Suel and S. Cha, *IEEE Signal Processing Magazine*, 2005, **22**, 62–75.

[77] W. C. Swope, J. W. Pitera and F. Suits, *The Journal of Physical Chemistry B*, 2004, **108**, 6571–6581.

[78] W. C. Swope, J. W. Pitera, F. Suits, M. Pitman, M. Eleftheriou, B. G. Fitch, R. S. Germain, A. Rayshubski, T. J. C. Ward, Y. Zhestkov and R. Zhou, *The Journal of Physical Chemistry B*, 2004, **108**, 6582–6594.

[79] E. Suárez, R. P. Wiewiora, C. Wehmeyer, F. Noé, J. D. Chodera and D. M. Zuckerman, *Journal of Chemical Theory and Computation*, 2021, **17**, 3119–3133.

[80] A. Pohorille, C. Jarzynski and C. Chipot, *The Journal of Physical Chemistry B*, 2010, **114**, 10235–10253.

[81] A. S. J. S. Mey, B. K. Allen, H. E. B. Macdonald, J. D. Chodera, D. F. Hahn, M. Kuhn, J. Michel, D. L. Mobley, L. N. Naden, S. Prasad, A. Rizzi, J. Scheen, M. R. Shirts, G. Tresadern and H. Xu, *Living Journal of Computational Molecular Science*, 2020, **2**, 18378.

[82] C. N. Cavasotto and S. S. Phatak, *Drug Discovery Today*, 2009, **14**, 676–683.

[83] A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F. T. Heer, T. A. P. de Beer, C. Rempfer, L. Bordoli, R. Lepore and T. Schwede, *Nucleic Acids Research*, 2018, **46**, W296–W303.

[84] Y. Haddad, V. Adam and Z. Heger, *PLOS Computational Biology*, 2020, **16**, e1007449.

[85] N. M. Lim, L. Wang, R. Abel and D. L. Mobley, *Journal of Chemical Theory and Computation*, 2016, **12**, 4620–4631.

[86] P. Liu, B. Kim, R. A. Friesner and B. J. Berne, *Proceedings of the National Academy of Sciences of the United States of America*, 2005, **102**, 13749.

[87] A. C. Pan, H. Xu, T. Palpant and D. E. Shaw, *Journal of Chemical Theory and Computation*, 2017, **13**, 3372–3377.

[88] L. Pérez-Benito, N. Casajuana-Martin, M. Jiménez-Rosés, H. van Vlijmen and G. Tresadern, *Journal of Chemical Theory and Computation*, 2019, **15**, 1884–1895.

[89] D. L. Mobley and K. A. Dill, *Structure*, 2009, **17**, 489–498.

[90] D. Cappel, S. Jerome, G. Hessler and H. Matter, *Journal of Chemical Information and Modeling*, 2020, **60**, 1432–1444.

[91] J. M. Granadino-Roldán, A. S. J. S. Mey, J. J. Pérez González, S. Bosisio, J. Rubio-Martinez and J. Michel, *PloS one*, 2019, **14**, e0213217–e0213217.

[92] D. L. Mobley, J. D. Chodera and K. A. Dill, *The Journal of Chemical Physics*, 2006, **125**, 084902.

[93] J. W. Kaus, E. Harder, T. Lin, R. Abel, J. A. McCammon and L. Wang, *Journal of Chemical Theory and Computation*, 2015, **11**, 2670–2679.

[94] L. Wang, R. A. Friesner and B. J. Berne, *The Journal of Physical Chemistry B*, 2011, **115**, 9431–9438.

[95] L. Wang, B. J. Berne and R. A. Friesner, *Proceedings of the National Academy of Sciences of the United States of America*, 2012, **109**, 1937.

[96] S. Sasmal, S. C. Gill, N. M. Lim and D. L. Mobley, *Journal of Chemical Theory and Computation*, 2020, **16**, 1854–1865.

[97] A. Laio and M. Parrinello, *Proceedings of the National Academy of Sciences of the United States of America*, 2002, **99**, 12562–12566.

[98] A. J. Clark, P. Tiwary, K. Borrelli, S. Feng, E. B. Miller, R. Abel, R. A. Friesner and B. J. Berne, *Journal of Chemical Theory and Computation*, 2016, **12**, 2990–2998.

[99] L. Wang, J. Chambers and R. Abel, in *Biomolecular Simulations: Methods and Protocols*, ed. M. Bonomi and C. Camilloni, Springer, New York, 2019, pp. 201–232.

[100] J. E. Ladbury, *Chemistry & Biology*, 1996, **3**, 973–980.

[101] A. Rudling, A. Orro and J. Carlsson, *Journal of Chemical Information and Modeling*, 2018, **58**, 350–361.

[102] J. Wahl and M. Smieško, *Journal of Chemical Information and Modeling*, 2019, **59**, 754–765.

[103] G. A. Ross, E. Russell, Y. Deng, C. Lu, E. D. Harder, R. Abel and L. Wang, *Journal of Chemical Theory and Computation*, 2020, **16**, 6061–6076.

[104] M. S. Bodnarchuk, R. Viner, J. Michel and J. W. Essex, *Journal of Chemical Information and Modeling*, 2014, **54**, 1623–1633.

[105] O. Carugo and D. Bordo, *Acta Crystallographica Section D*, 1999, **55**, 479–483.

[106] R. Abel, N. K. Salam, J. Shelley, R. Farid, R. A. Friesner and W. Sherman, *ChemMedChem*, 2011, **6**, 1049–1066.

[107] A. M. Davis, S. J. Teague and G. J. Kleywegt, *Angewandte Chemie International Edition*, 2003, **42**, 2718–2736.

[108] J. Michel, J. Tirado-Rives and W. L. Jorgensen, *The Journal of Physical Chemistry B*, 2009, **113**, 13337–13346.

[109] J. Luccarelli, J. Michel, J. Tirado-Rives and W. L. Jorgensen, *Journal of Chemical Theory and Computation*, 2010, **6**, 3850–3856.

[110] G. A. Ross, M. S. Bodnarchuk and J. W. Essex, *Journal of the American Chemical Society*, 2015, **137**, 14930–14943.

[111] H. E. Bruce Macdonald, C. Cave-Ayland, G. A. Ross and J. W. Essex, *Journal of Chemical Theory and Computation*, 2018, **14**, 6586–6597.

[112] M. L. Samways, H. E. Bruce Macdonald and J. W. Essex, *Journal of Chemical Information and Modeling*, 2020, **60**, 4436–4441.

[113] I. Y. Ben-Shalom, Z. Lin, B. K. Radak, C. Lin, W. Sherman and M. K. Gilson, *Journal of Chemical Theory and Computation*, 2020, **16**, 7883–7894.

[114] T. D. Bergazin, I. Y. Ben-Shalom, N. M. Lim, S. C. Gill, M. K. Gilson and D. L. Mobley, *Journal of Computer-Aided Molecular Design*, 2021, **35**, 167–177.

[115] Y. Ge, D. C. Wych, M. L. Samways, M. E. Wall, J. W. Essex and D. L. Mobley, 2021, bioRxiv:2021.06.14.448350.

[116] S.-M. Liao, Q.-S. Du, J.-Z. Meng, Z.-W. Pang and R.-B. Huang, *Chemistry Central Journal*, 2013, **7**, 44.

[117] M. O. Kim, S. E. Nichols, Y. Wang and J. A. McCammon, *Journal of Computer-Aided Molecular Design*, 2013, **27**, 235–246.

[118] M. S. Lee, F. R. Salsbury Jr. and C. L. Brooks III, *Proteins: Structure, Function, and Bioinformatics*, 2004, **56**, 738–752.

[119] J. Khandogin and C. L. Brooks III, *Biophysical Journal*, 2005, **89**, 141–157.

[120] J. A. Wallace and J. K. Shen, *Journal of Chemical Theory and Computation*, 2011, **7**, 2617–2629.

[121] A. M. Baptista, V. H. Teixeira and C. M. Soares, *The Journal of Chemical Physics*, 2002, **117**, 4184–4200.

[122] H. A. Stern, *The Journal of Chemical Physics*, 2007, **126**, 164112.

[123] Y. Chen and B. Roux, *Journal of Chemical Theory and Computation*, 2015, **11**, 3919–3931.

[124] B. K. Radak, C. Chipot, D. Suh, S. Jo, W. Jiang, J. C. Phillips, K. Schulten and B. Roux, *Journal of Chemical Theory and Computation*, 2017, **13**, 5933–5944.

[125] J. M. Swails, D. M. York and A. E. Roitberg, *Journal of Chemical Theory and Computation*, 2014, **10**, 1341–1352.

[126] J. Lee, B. T. Miller, A. Damjanović and B. R. Brooks, *Journal of Chemical Theory and Computation*, 2014, **10**, 2738–2750.

[127] J. Lee, B. T. Miller and B. R. Brooks, *Protein Science*, 2016, **25**, 231–243.

[128] T. J. Paul, J. Z. Vilseck, R. L. Hayes and C. L. Brooks III, *The Journal of Physical Chemistry B*, 2020, **124**, 6520–6528.

[129] M. O. Kim, P. G. Blachly and J. A. McCammon, *PLOS Computational Biology*, 2015, **11**, e1004341.

[130] M. O. Kim and J. A. McCammon, *Biopolymers*, 2016, **105**, 43–49.

[131] Y. Hu, B. Sherborne, T.-S. Lee, D. A. Case, D. M. York and Z. Guo, *Journal of Computer-Aided Molecular Design*, 2016, **30**, 533–539.

[132] C. de Oliveira, H. S. Yu, W. Chen, R. Abel and L. Wang, *Journal of Chemical Theory and Computation*, 2019, **15**, 424–435.

[133] J. Huang and A. D. MacKerell Jr, *Journal of Computational Chemistry*, 2013, **34**, 2135–2145.

[134] K. Vanommeslaeghe and A. D. MacKerell Jr, *Journal of Chemical Information and Modeling*, 2012, **52**, 3144–3154.

[135] K. Vanommeslaeghe, E. P. Raman and A. D. MacKerell Jr, *Journal of Chemical Information and Modeling*, 2012, **52**, 3155–3168.

[136] D. Vassetti, M. Pagliai and P. Procacci, *Journal of Chemical Theory and Computation*, 2019, **15**, 1983–1995.

[137] G. A. Kaminski, R. A. Friesner, J. Tirado-Rives and W. L. Jorgensen, *The Journal of Physical Chemistry B*, 2001, **105**, 6474–6487.

[138] K. Roos, C. Wu, W. Damm, M. Reboul, J. M. Stevenson, C. Lu, M. K. Dahlgren, S. Mondal, W. Chen, L. Wang, R. Abel, R. A. Friesner and E. D. Harder, *Journal of Chemical Theory and Computation*, 2019, **15**, 1863–1874.

[139] G. J. Rocklin, D. L. Mobley and K. A. Dill, *Journal of Chemical Theory and Computation*, 2013, **9**, 3072–3083.

[140] F. Manzoni and U. Ryde, *Journal of Computer-Aided Molecular Design*, 2018, **32**, 529–536.

[141] A. E. A. Allen, M. J. Robertson, M. C. Payne and D. J. Cole, *ACS Omega*, 2019, **4**, 14537–14550.

[142] E. Harder, W. Damm, J. Maple, C. Wu, M. Reboul, J. Y. Xiang, L. Wang, D. Lupyan, M. K. Dahlgren, J. L. Knight, J. W. Kaus, D. S. Cerutti, G. Krilov, W. L. Jorgensen, R. Abel and R. A. Friesner, *Journal of Chemical Theory and Computation*, 2016, **12**, 281–296.

[143] C. Lu, C. Wu, D. Ghoreishi, W. Chen, L. Wang, W. Damm, G. A. Ross, M. K. Dahlgren, E. Russell, C. D. Von Bargen, R. Abel, R. A. Friesner and E. D. Harder, *Journal of Chemical Theory and Computation*, 2021, **17**, 4291–4300.

[144] D. L. Mobley, C. C. Bannan, A. Rizzi, C. I. Bayly, J. D. Chodera, V. T. Lim, N. M. Lim, K. A. Beauchamp, D. R. Slochower, M. R. Shirts, M. K. Gilson and P. K. Eastman, *Journal of Chemical Theory and Computation*, 2018, **14**, 6076–6092.

[145] Y. Qiu, D. Smith, S. Boothroyd, H. Jang, J. Wagner, C. C. Bannan, T. Gokey, V. T. Lim, C. Stern, A. Rizzi, X. Lucas, B. Tjanaka, M. R. Shirts, M. Gilson, J. Chodera, C. I. Bayly, D. Mobley and L.-P. Wang, 2020, ChemRxiv:13082561.

[146] P. Ren and J. W. Ponder, *The Journal of Physical Chemistry B*, 2003, **107**, 5933–5947.

[147] S. Genheden and U. Ryde, *Expert Opinion on Drug Discovery*, 2015, **10**, 449–461.

[148] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case and T. E. Cheatham, *Accounts of Chemical Research*, 2000, **33**, 889–897.

[149] J. Michel, M. L. Verdonk and J. W. Essex, *Journal of Medicinal Chemistry*, 2006, **49**, 7427–7439.

[150] M. Aldeghi, G. A. Ross, M. J. Bodkin, J. W. Essex, S. Knapp and P. C. Biggin, *Communications Chemistry*, 2018, **1**, 19.

[151] S. Izadi, B. Aguilar and A. V. Onufriev, *Journal of Chemical Theory and Computation*, 2015, **11**, 4450–4459.

[152] B. Guillot, *Journal of Molecular Liquids*, 2002, **101**, 219–260.

[153] J. F. Ouyang and R. P. A. Bettens, *CHIMIA International Journal for Chemistry*, 2015, **69**, 104–111.

[154] J. Yin, N. M. Henriksen, H. S. Muddana and M. K. Gilson, *Journal of Chemical Theory and Computation*, 2018, **14**, 3621–3632.

[155] S. S. Çınaroğlu and P. C. Biggin, *The Journal of Physical Chemistry B*, 2021, **125**, 1558–1567.

[156] L. Wang, Y. Wu, Y. Deng, B. Kim, L. Pierce, G. Krilov, D. Lupyan, S. Robinson, M. K. Dahlgren, J. Greenwood, D. L. Romero, C. Masse, J. L. Knight, T. Steinbrecher, T. Beuming, W. Damm, E. Harder, W. Sherman, M. Brewer, R. Wester, M. Murcko, L. Frye, R. Farid, T. Lin, D. L. Mobley, W. L. Jorgensen, B. J. Berne, R. A. Friesner and R. Abel, *Journal of the American Chemical Society*, 2015, **137**, 2695–2703.

[157] M. Kuhn, S. Firth-Clark, P. Tosco, A. S. J. S. Mey, M. Mackey and J. Michel, *Journal of Chemical Information and Modeling*, 2020, **60**, 3120–3130.

[158] F. Fratev and S. Sirimulla, *Scientific Reports*, 2019, **9**, 16829.

[159] S. Wan, G. Tresadern, L. Pérez-Benito, H. van Vlijmen and P. V. Coveney, *Advanced Theory and Simulations*, 2020, **3**, 1900195.

[160] A. de Ruiter, D. Petrov and C. Oostenbrink, *Journal of Chemical Theory and Computation*, 2021, **17**, 56–65.

[161] M. R. Shirts and V. S. Pande, *The Journal of Chemical Physics*, 2005, **122**, 144107.

[162] T.-S. Lee, B. K. Allen, T. J. Giese, Z. Guo, P. Li, C. Lin, T. D. McGee, D. A. Pearlman, B. K. Radak, Y. Tao, H.-C. Tsai, H. Xu, W. Sherman and D. M. York, *Journal of Chemical Information and Modeling*, 2020, **60**, 5595–5623.

[163] J. D. Chodera, W. C. Swope, J. W. Pitera, C. Seok and K. A. Dill, *Journal of Chemical Theory and Computation*, 2007, **3**, 26–41.

[164] B. Knapp, L. Ospina and C. M. Deane, *Journal of Chemical Theory and Computation*, 2018, **14**, 6127–6138.

[165] J. Chen, J. Wang, B. Yin, L. Pang, W. Wang and W. Zhu, *ACS Chemical Neuroscience*, 2019, **10**, 4303–4318.

[166] D. W. Wright, S. Wan, C. Meyer, H. van Vlijmen, G. Tresadern and P. V. Coveney, *Scientific Reports*, 2019, **9**, 6017.

[167] T. Steinbrecher, I. Joung and D. A. Case, *Journal of Computational Chemistry*, 2011, **32**, 3253–3263.

[168] A. de Ruiter, S. Boresch and C. Oostenbrink, *Journal of Computational Chemistry*, 2013, **34**, 1024–1034.

[169] H. H. Loeffler, S. Bosisio, G. Duarte Ramos Matos, D. Suh, B. Roux, D. L. Mobley and J. Michel, *Journal of Chemical Theory and Computation*, 2018, **14**, 5567–5582.

[170] A. Rizzi, T. Jensen, D. R. Slochower, M. Aldeghi, V. Gapsys, D. Ntekoumes, S. Bosisio, M. Papadourakis, N. M. Henriksen, B. L. de Groot, Z. Cournia, A. Dickson, J. Michel, M. K. Gilson, M. R. Shirts, D. L. Mobley and J. D. Chodera, *Journal of Computer-Aided Molecular Design*, 2020, **34**, 601–633.

[171] J. A. Wagoner and V. S. Pande, *The Journal of Chemical Physics*, 2012, **137**, 214105.

[172] H.-C. Tsai, Y. Tao, T.-S. Lee, K. M. Merz and D. M. York, *Journal of Chemical Information and Modeling*, 2020, **60**, 5296–5300.

[173] K. Wang, J. D. Chodera, Y. Yang and M. R. Shirts, *Journal of Computer-Aided Molecular Design*, 2013, **27**, 989–1007.

[174] G. M. Sastry, M. Adzhigirey, T. Day, R. Annabhimoju and W. Sherman, *Journal of Computer-Aided Molecular Design*, 2013, **27**, 221–234.

[175] S. Doerr, M. J. Harvey, F. Noé and G. De Fabritiis, *Journal of Chemical Theory and Computation*, 2016, **12**, 1845–1852.

[176] P. V. Klimovich and D. L. Mobley, *Journal of Computer-Aided Molecular Design*, 2015, **29**, 1007–1014.

[177] H. H. Loeffler, J. Michel and C. Woods, *Journal of Chemical Information and Modeling*, 2015, **55**, 2485–2490.

[178] V. Gapsys, S. Michielssens, D. Seeliger and B. L. de Groot, *Journal of Computational Chemistry*, 2015, **36**, 348–354.

[179] L. O. Hedges, A. S. Mey, C. A. Laughton, F. L. Gervasio, A. J. Mulholland, C. J. Woods and J. Michel, *Journal of Open Source Software*, 2019, **4**, 1831.

[180] C. J. Woods and J. Michel, *Sire*, 2016, https://siremol.org.

[181] H. M. Berman, *Nucleic Acids Research*, 2000, **28**, 235–242.

[182] B. Webb and A. Sali, *Current Protocols in Bioinformatics*, 2016, **54**, 5.6.1–5.6.37.

[183] P. Eastman, M. S. Friedrichs, J. D. Chodera, R. J. Radmer, C. M. Bruns, J. P. Ku, K. A. Beauchamp, T. J. Lane, L.-P. Wang, D. Shukla, T. Tye, M. Houston, T. Stich, C. Klein, M. R. Shirts and V. S. Pande, *Journal of Chemical Theory and Computation*, 2013, **9**, 461–469.

[184] S. Jo, T. Kim, V. G. Iyer and W. Im, *Journal of Computational Chemistry*, 2008, **29**, 1859–1865.

[185] T. J. Dolinsky, P. Czodrowski, H. Li, J. E. Nielsen, J. H. Jensen, G. Klebe and N. A. Baker, *Nucleic Acids Research*, 2007, **35**, W522–W525.

[186] D. C. Bas, D. M. Rogers and J. H. Jensen, *Proteins: Structure, Function, and Bioinformatics*, 2008, **73**, 765–783.

[187] M. H. Olsson, C. R. Søndergaard, M. Rostkowski and J. H. Jensen, *Journal of Chemical Theory and Computation*, 2011, **7**, 525–537.

[188] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, *Journal of Cheminformatics*, 2011, **3**, 33.

[189] J. Swails, C. Hernandez, D. L. Mobley, H. Nguyen, L. P. Wang and P. Janowski, *ParmEd*, 2010, https://github.com/ParmEd/ParmEd.

[190] D. Weininger, *Journal of Chemical Information and Computer Sciences*, 1988, **28**, 31–36.

[191] S. Heller, A. McNaught, S. Stein, D. Tchekhovskoi and I. Pletnev, *Journal of Cheminformatics*, 2013, **5**, 7.

[192] R. V. Book, M. R. Garey and D. S. Johnson, *Bulletin of The American Mathematical Society*, 1980, **3**, 898–904.

[193] G. Landrum, *RDKit*, 2012, https://www.rdkit.org.

[194] P. Tosco, N. Stiefl and G. Landrum, *Journal of Cheminformatics*, 2014, **6**, 37.

[195] P. V. Klimovich, M. R. Shirts and D. L. Mobley, *Journal of Computer-Aided Molecular Design*, 2015, **29**, 397–411.

[196] K. A. Beauchamp, N. M. Lim and D. L. Mobley, *Alchemical Analysis*, 2015, https://github.com/MobleyLab/alchemical-analysis.

[197] S. Genheden, I. Nilsson and U. Ryde, *Journal of Chemical Information and Modeling*, 2011, **51**, 947–958.

[198] S. Genheden, *Journal of Chemical Information and Modeling*, 2012, **52**, 3013–3021.

[199] V. Cody, N. Galitsky, J. R. Luft, W. Pangborn, A. Rosowsky and R. L. Blakley, *Biochemistry*, 1997, **36**, 13897–13903.

[200] A. K. W. Leung, L. J. Ross, S. Zywno-Van Ginkel, R. C. Reynolds, L. E. Seitz, V. Pathak, W. W. Barrow, E. L. White, W. J. Suling, J. R. Piper and D. W. Borhani, *Structural Basis for Selective Inhibition of Mycobacterium Avium Dihydrofolate Reductase by a Lipophilic Antifolate*, Protein Data Bank, 2009, https://www.rcsb.org/structure/2W3M, (accessed September 2021).

[201] A. Gangjee, W. Li, R. L. Kisliuk, V. Cody, J. Pace, J. Piraino and J. Makin, *Journal of Medicinal Chemistry*, 2009, **52**, 4892–4902.

[202] Y. Yuthavong, B. Tarnchompoo, T. Vilaivan, P. Chitnumsub, S. Kamchonwongpaisan, S. A. Charman, D. N. McLennan, K. L. White, L. Vivas, E. Bongard, C. Thongphanchang, S. Taweechai, J. Vanichtanankul, R. Rattanajak, U. Arwon, P. Fantauzzi, J. Yuvaniyama, W. N. Charman and D. Matthews, *Proceedings of the National Academy of Sciences of the United States of America*, 2012, **109**, 16823.

[203] G. Bhabha, D. C. Ekiert, M. Jennewein, C. M. Zmasek, L. M. Tuttle, G. Kroon, H. J. Dyson, A. Godzik, I. A. Wilson and P. E. Wright, *Nature Structural & Molecular Biology*, 2013, **20**, 1243–1249.

[204] V. Cody and A. Gangjee, *Human Dihydrofolate Reductase Complex With NADPH and 5-Methyl-6-(Phenylthio-4'Trifluoromethyl)Thieno[2,3-d]Pyrimidine-2,4-Diamine*, Protein Data Bank, 2016, https://www.rcsb.org/structure/5HPB, (accessed September 2021).

[205] B. Tarnchompoo, P. Chitnumsub, A. Jaruwat, P. J. Shaw, J. Vanichtanankul, S. Poen, R. Rattanajak, C. Wongsombat, A. Tonsomboon, S. Decharuangsilp, T. Anukunwithaya, U. Arwon, S. Kamchonwongpaisan and Y. Yuthavong, *ACS Medicinal Chemistry Letters*, 2018, **9**, 1235–1240.

[206] S. J. Mayclin, D. M. Dranow, C. Walpole and D. D. Lorimer, *Crystal Structure of Human DHFR Complexed with NADP and N10-Formyltetrahydrofolate*, Protein Data

Bank, 2018, https://www.rcsb.org/structure/6DAV, (accessed September 2021).

[207] M. R. Groves, Z.-J. Yao, P. P. Roller, T. R. Burke and D. Barford, *Biochemistry*, 1998, **37**, 17773–17783.

[208] D. A. Erlanson, R. S. McDowell, M. M. He, M. Randal, R. L. Simmons, J. Kung, A. Waight and S. K. Hansen, *Journal of the American Chemical Society*, 2003, **125**, 5602–5603.

[209] A. F. Moretto, S. J. Kirincich, W. X. Xu, M. J. Smith, Z.-K. Wan, D. P. Wilson, B. C. Follows, E. Binnun, D. Joseph-McCarthy, K. Foreman, D. V. Erbe, Y. L. Zhang, S. K. Tam, S. Y. Tam and J. Lee, *Bioorganic & Medicinal Chemistry*, 2006, **14**, 2162–2177.

[210] S. R. Klopfenstein, A. G. Evdokimov, A.-O. Colson, N. T. Fairweather, J. J. Neuman, M. B. Maier, J. L. Gray, G. S. Gerwe, G. E. Stake, B. W. Howard, J. A. Farmer, M. E. Pokross, T. R. Downs, B. Kasibhatla and K. G. Peters, *Bioorganic & Medicinal Chemistry Letters*, 2006, **16**, 1574–1578.

[211] Z.-K. Wan, J. Lee, W. Xu, D. V. Erbe, D. Joseph-McCarthy, B. C. Follows and Y.-L. Zhang, *Bioorganic & Medicinal Chemistry Letters*, 2006, **16**, 4941–4945.

[212] Z.-K. Wan, B. Follows, S. Kirincich, D. Wilson, E. Binnun, W. Xu, D. Joseph-McCarthy, J. Wu, M. Smith, Y.-L. Zhang, M. Tam, D. Erbe, S. Tam, E. Saiah and J. Lee, *Bioorganic & Medicinal Chemistry Letters*, 2007, **17**, 2913–2920.

[213] Z.-K. Wan, J. Lee, R. Hotchandani, A. Moretto, E. Binnun, D. P. Wilson, S. J. Kirincich, B. C. Follows, M. Ipek, W. Xu, D. Joseph-McCarthy, Y.-L. Zhang, M. Tam, D. V. Erbe, J. F. Tobin, W. Li, S. Y. Tam, T. S. Mansour and J. Wu, *ChemMedChem*, 2008, **3**, 1525–1529.

[214] S. Maignan, J.-P. Guilloteau, S. Pouzieux, Y. M. Choi-Sledeski, M. R. Becker, S. I. Klein, W. R. Ewing, H. W. Pauls, A. P. Spada and V. Mikol, *Journal of Medicinal Chemistry*, 2000, **43**, 3226–3232.

[215] K. R. Guertin, C. J. Gardner, S. I. Klein, A. L. Zulli, M. Czekaj, Y. Gong, A. P. Spada, D. L. Cheney, S. Maignan, J.-P. Guilloteau, K. D. Brown, D. J. Colussi, V. Chu, C. L. Heran, S. R. Morgan, R. G. Bentley, C. T. Dunwiddie, R. J. Leadley and H. W. Pauls, *Bioorganic & Medicinal Chemistry Letters*, 2002, **12**, 1671–1674.

[216] S. Maignan, J.-P. Guilloteau, Y. M. Choi-Sledeski, M. R. Becker, W. R. Ewing, H. W. Pauls, A. P. Spada and V. Mikol, *Journal of Medicinal Chemistry*, 2003, **46**, 685–690.

[217] N. S. Watson, D. Brown, M. Campbell, C. Chan, L. Chaudry, M. A. Convery, R. Fenwick, J. N. Hamblin, C. Haslam, H. A. Kelly, N. P. King, C. L. Kurtis, A. R.

Leach, G. R. Manchee, A. M. Mason, C. Mitchell, C. Patel, V. K. Patel, S. Senger, G. P. Shah, H. E. Weston, C. Whitworth and R. J. Young, *Bioorganic & Medicinal Chemistry Letters*, 2006, **16**, 3784–3788.

[218] S. Senger, M. A. Convery, C. Chan and N. S. Watson, *Bioorganic & Medicinal Chemistry Letters*, 2006, **16**, 5731–5735.

[219] C. Chan, A. D. Borthwick, D. Brown, C. L. Burns-Kurtis, M. Campbell, L. Chaudry, C.-w. Chung, M. A. Convery, J. N. Hamblin, L. Johnstone, H. A. Kelly, S. Kleanthous, A. Patikis, C. Patel, A. J. Pateman, S. Senger, G. P. Shah, J. R. Toomey, N. S. Watson, H. E. Weston, C. Whitworth, R. J. Young and P. Zhou, *Journal of Medicinal Chemistry*, 2007, **50**, 1546–1557.

[220] M. A. Convery, R. J. Young, S. Senger, J. N. Hamblin, C. Chan, J. R. Toomey and N. S. Watson, *Factor Xa Complex with GTC000398*, Protein Data Bank, 2015, https://www.rcsb.org/structure/4Y71, (accessed September 2021).

[221] S. Forli, *Molecules*, 2015, **20**, 18732–18758.

[222] N. Holmberg, U. Ryde and L. Bülow, *Protein Engineering, Design and Selection*, 1999, **12**, 851–856.

[223] W. H. Kruskal and W. A. Wallis, *Journal of the American Statistical Association*, 1952, **47**, 583–621.

[224] M. G. Kendall, *Biometrika*, 1938, **30**, 81–93.

[225] S. Bhakat, E. Åberg and P. Söderhjelm, *Journal of Computer-Aided Molecular Design*, 2018, **32**, 59–73.

[226] H. Keränen, L. Pérez-Benito, M. Ciordia, F. Delgado, T. B. Steinbrecher, D. Oehlrich, H. W. T. van Vlijmen, A. A. Trabanco and G. Tresadern, *Journal of Chemical Theory and Computation*, 2017, **13**, 1439–1453.

[227] A. Saha, A. Y. Shih, T. Mirzadegan and M. Seierstad, *Journal of Chemical Theory and Computation*, 2018, **14**, 5815–5822.

[228] C. Schindler, F. Rippmann and D. Kuhn, *Journal of Computer-Aided Molecular Design*, 2018, **32**, 265–272.

[229] A. de Ruiter and C. Oostenbrink, *Journal of Chemical Theory and Computation*, 2012, **8**, 3686–3695.

[230] K. A. Krukenberg, T. O. Street, L. A. Lavery and D. A. Agard, *Quarterly Reviews of Biophysics*, 2011, **44**, 229–255.

[231] T. B. Steinbrecher, M. Dahlgren, D. Cappel, T. Lin, L. Wang, G. Krilov, R. Abel, R. Friesner and W. Sherman, *Journal of Chemical Information and Modeling*, 2015, **55**, 2411–2420.

[232] M. Marquart, J. Walter, J. Deisenhofer, W. Bode and R. Huber, *Acta Crystallographica Section B Structural Science*, 1983, **39**, 480–490.

[233] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks and V. S. Pande, *PLOS Computational Biology*, 2017, **13**, e1005659.

[234] N. Wayne and D. N. Bolon, *Journal of Biological Chemistry*, 2007, **282**, 35386–35395.

[235] R. A. Fisher, *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh, 1934.

[236] R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane and V. S. Pande, *Biophysical Journal*, 2015, **109**, 1528–1532.

[237] D. Reynolds, in *Encyclopedia of Biometrics*, ed. S. Z. Li and A. Jain, Springer, Boston, 2009, pp. 659–663.

[238] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *Journal of Machine Learning Research*, 2011, **12**, 2825–2830.

[239] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz and F. Noé, *Journal of Chemical Theory and Computation*, 2015, **11**, 5525–5542.

[240] E. Marinari and G. Parisi, *Europhysics Letters (EPL)*, 1992, **19**, 451–458.

[241] X. Kong and C. L. Brooks III, *The Journal of Chemical Physics*, 1996, **105**, 2414–2423.

[242] C. D. Christ and W. F. van Gunsteren, *The Journal of Chemical Physics*, 2007, **126**, 184110.

[243] Y. Q. Gao, *The Journal of Chemical Physics*, 2008, **128**, 064105.

[244] K. H. Burley, S. C. Gill, N. M. Lim and D. L. Mobley, *Journal of Chemical Theory and Computation*, 2019, **15**, 1848–1862.

[245] S. Hayward and N. Go, *Annual Review of Physical Chemistry*, 1995, **46**, 223–250.

[246] F. Sittel and G. Stock, *The Journal of Chemical Physics*, 2018, **149**, 150901.

[247] N. Pokhrel and L. Maibaum, *Journal of Chemical Theory and Computation*, 2018, **14**, 1762–1771.

[248] B. A. Berg and T. Neuhaus, *Physics Letters B*, 1991, **267**, 249–253.

[249] D. P. Landau, S.-H. Tsai and M. Exler, *American Journal of Physics*, 2004, **72**, 1294–1302.

[250] A. Barducci, G. Bussi and M. Parrinello, *Physical Review Letters*, 2008, **100**, 020603.

[251] D. Branduardi, G. Bussi and M. Parrinello, *Journal of Chemical Theory and Computation*, 2012, **8**, 2247–2254.

[252] G. M. Torrie and J. P. Valleau, *Journal of Computational Physics*, 1977, **23**, 187–199.

[253] D. Hamelberg, J. Mongan and J. A. McCammon, *The Journal of Chemical Physics*, 2004, **120**, 11919–11929.

[254] Y. Miao, V. A. Feher and J. A. McCammon, *Journal of Chemical Theory and Computation*, 2015, **11**, 3584–3595.

[255] E. Darve and A. Pohorille, *The Journal of Chemical Physics*, 2001, **115**, 9169–9183.

[256] J. Schlitter, M. Engels, P. Krüger, E. Jacoby and A. Wollmer, *Molecular Simulation*, 1993, **10**, 291–308.

[257] B. Isralewitz, M. Gao and K. Schulten, *Current Opinion in Structural Biology*, 2001, **11**, 224–230.

[258] H. Grubmüller, *Physical Review E*, 1995, **52**, 2893–2906.

[259] G. Ciccotti, R. Kapral and E. Vanden-Eijnden, *ChemPhysChem*, 2005, **6**, 1809–1814.

[260] C. Abrams and G. Bussi, *Entropy*, 2014, **16**, 163–199.

[261] R. H. Swendsen and J.-S. Wang, *Physical Review Letters*, 1986, **57**, 2607–2609.

[262] J. D. Chodera and M. R. Shirts, *The Journal of Chemical Physics*, 2011, **135**, 194110.

[263] Y. Sugita and Y. Okamoto, *Chemical Physics Letters*, 2000, **329**, 261–270.

[264] H. Fukunishi, O. Watanabe and S. Takada, *The Journal of Chemical Physics*, 2002, **116**, 9058–9067.

[265] R. M. Neal, *Statistics and Computing*, 1996, **6**, 353–366.

[266] W. Jiang, M. Hodoscek and B. Roux, *Journal of Chemical Theory and Computation*, 2009, **5**, 2583–2588.

[267] S. G. Itoh, A. Damjanović and B. R. Brooks, *Proteins: Structure, Function, and Bioinformatics*, 2011, **79**, 3420–3436.

[268] G. A. Ross, H. E. Bruce Macdonald, C. Cave-Ayland, A. I. Cabedo Martinez and J. W. Essex, *Journal of Chemical Theory and Computation*, 2017, **13**, 6373–6381.

[269] W. Jiang, J. Thirman, S. Jo and B. Roux, *The Journal of Physical Chemistry B*, 2018, **122**, 9435–9442.

[270] A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov and P. N. Vorontsov-Velyaminov, *The Journal of Chemical Physics*, 1992, **96**, 1776–1783.

[271] S. Park and V. S. Pande, *Physical Review E*, 2007, **76**, 016703.

[272] P. H. Nguyen, Y. Okamoto and P. Derreumaux, *The Journal of Chemical Physics*, 2013, **138**, 061102.

[273] A. S. Paluch, J. K. Shah and E. J. Maginn, *Journal of Chemical Theory and Computation*, 2011, **7**, 1394–1403.

[274] A. S. Paluch, D. L. Mobley and E. J. Maginn, *Journal of Chemical Theory and Computation*, 2011, **7**, 2910–2918.

[275] Z. Tan, *Journal of Computational and Graphical Statistics*, 2017, **26**, 54–65.

[276] H. Li, M. Fajer and W. Yang, *The Journal of Chemical Physics*, 2007, **126**, 024106.

[277] J. Kim and J. E. Straub, *The Journal of Chemical Physics*, 2010, **133**, 154101.

[278] A. C. Pan, T. M. Weinreich, S. Piana and D. E. Shaw, *Journal of Chemical Theory and Computation*, 2016, **12**, 1360–1367.

[279] K. M. Åberg, A. P. Lyubartsev, S. P. Jacobsson and A. Laaksonen, *The Journal of Chemical Physics*, 2004, **120**, 3770–3776.

[280] Z. You, L. Li, J. Lu and H. Ge, *The Journal of Chemical Physics*, 2018, **149**, 084114.

[281] L. Yang and Y. Qin Gao, *The Journal of Chemical Physics*, 2009, **131**, 214109.

[282] T. Mori, R. J. Hamers, J. A. Pedersen and Q. Cui, *The Journal of Physical Chemistry B*, 2014, **118**, 8210–8220.

[283] K. A. Armacost, G. B. Goh and C. L. Brooks III, *Journal of Chemical Theory and Computation*, 2015, **11**, 1267–1277.

[284] R. L. Hayes, K. A. Armacost, J. Z. Vilseck and C. L. Brooks III, *The Journal of Physical Chemistry B*, 2017, **121**, 3626–3635.

[285] X. Ding, J. Z. Vilseck, R. L. Hayes and C. L. Brooks III, *Journal of Chemical Theory and Computation*, 2017, **13**, 2501–2510.

[286] J. Z. Vilseck, X. Ding, R. L. Hayes and C. L. Brooks III, *Journal of Chemical Theory and Computation*, 2021, **17**, 3895–3907.

[287] J. L. Knight and C. L. Brooks III, *Journal of Chemical Theory and Computation*, 2011, **7**, 2728–2739.

[288] D. F. Hahn and P. H. Hünenberger, *Journal of Chemical Theory and Computation*, 2019, **15**, 2392–2419.

[289] J. Z. Vilseck, K. A. Armacost, R. L. Hayes, G. B. Goh and C. L. Brooks III, *The Journal of Physical Chemistry Letters*, 2018, **9**, 3328–3332.

[290] K.-K. Han, *Physics Letters A*, 1992, **165**, 28–32.

[291] C. D. Christ and W. F. van Gunsteren, *The Journal of Chemical Physics*, 2008, **128**, 174112.

[292] C. D. Christ and W. F. van Gunsteren, *Journal of Chemical Theory and Computation*, 2009, **5**, 276–286.

[293] D. Sidler, A. Schwaninger and S. Riniker, *The Journal of Chemical Physics*, 2016, **145**, 154114.

[294] A. Kurut, R. Fonseca and W. Boomsma, *The Journal of Physical Chemistry B*, 2018, **122**, 1195–1204.

[295] J. S. Liu and R. Chen, *Journal of the American Statistical Association*, 1998, **93**, 1032–1044.

[296] J. Machta, *Physical Review E*, 2010, **82**, 026704.

[297] W. Wang, J. Machta and H. G. Katzgraber, *Physical Review E*, 2015, **92**, 063307.

[298] M. Rousset and G. Stoltz, *Journal of Statistical Physics*, 2006, **123**, 1251–1272.

[299] E. Lyman and D. M. Zuckerman, *The Journal of Chemical Physics*, 2009, **130**, 081102.

[300] D. A. Rufa, H. E. Bruce Macdonald, J. Fass, M. Wieder, P. B. Grinaway, A. E. Roitberg, O. Isayev and J. D. Chodera, 2020, bioRxiv:2020.07.29.227959.

[301] J. Zhang, R. Chen, C. Tang and J. Liang, *The Journal of Chemical Physics*, 2003, **118**, 6102–6109.

[302] K. Tang, J. Zhang and J. Liang, *PLOS Computational Biology*, 2014, **10**, e1003539.

[303] Samuel W. K. Wong, Jun S. Liu and S. C. Kou, *The Annals of Applied Statistics*, 2018, **12**, 1628–1654.

[304] H. Christiansen, M. Weigel and W. Janke, *Physical Review Letters*, 2019, **122**, 060602.

[305] H. Christiansen, M. Weigel and W. Janke, *Journal of Physics: Conference Series*, 2019, **1163**, 012074.

[306] A. Doucet, S. Godsill and C. Andrieu, *Statistics and Computing*, 2000, **10**, 197–208.

[307] P. J. van Leeuwen, H. R. Künsch, L. Nerger, R. Potthast and S. Reich, *Quarterly Journal of the Royal Meteorological Society*, 2019, **145**, 2335–2365.

[308] M. Amaya, N. Linde and E. Laloy, *Geophysical Journal International*, 2021, **226**, 1220–1238.

[309] R. Weiss, P. Glösekötter, E. Prestes and M. Kolberg, *Journal of Intelligent & Robotic Systems*, 2020, **99**, 335–357.

[310] P. Del Moral, *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics*, 1997, **325**, 653–658.

[311] P. M. Djuric, J. H. Kotecha, J. Zhang, Y. Huang, T. Ghirmai, M. F. Bugallo and J. Miguez, *IEEE Signal Processing Magazine*, 2003, **20**, 19–38.

[312] A. Doucet and A. M. Johansen, in *The Oxford Handbook of Nonlinear Filtering*, Oxford University Press, Oxford, New York, 2011, pp. 656–704.

[313] P. J. Reynolds, D. M. Ceperley, B. J. Alder and W. A. Lester, *The Journal of Chemical Physics*, 1982, **77**, 5593–5603.

[314] M. N. Rosenbluth and A. W. Rosenbluth, *The Journal of Chemical Physics*, 1955, **23**, 356–359.

[315] G. C. A. M. Mooij and D. Frenkel, *Journal of Physics: Condensed Matter*, 1994, **6**, 3879–3888.

[316] Y. Sakai and K. Hukushima, *Journal of the Physical Society of Japan*, 2016, **85**, 104002.

[317] S. Syed, A. Bouchard-Côté, G. Deligiannidis and A. Doucet, 2020, arXiv:1905.02939.

[318] P. Diaconis, S. Holmes and R. M. Neal, *The Annals of Applied Probability*, 2000, **10**, 726–752.

[319] B. K. Radak and B. Roux, *The Journal of Chemical Physics*, 2016, **145**, 134109.

[320] D. J. Sindhikara, D. J. Emerson and A. E. Roitberg, *Journal of Chemical Theory and Computation*, 2010, **6**, 2804–2808.

[321] S. Park, *Physical Review E*, 2008, **77**, 016709.

[322] C. Zhang and J. Ma, *The Journal of Chemical Physics*, 2008, **129**, 134112.

[323] A. Patriksson and D. van der Spoel, *Physical Chemistry Chemical Physics*, 2008, **10**, 2073–2077.

[324] W. D. Vousden, W. M. Farr and I. Mandel, *Monthly Notices of the Royal Astronomical Society*, 2016, **455**, 1919–1937.

[325] J. L. MacCallum, M. I. Muniyat and K. Gaalswyk, *The Journal of Physical Chemistry B*, 2018, **122**, 5448–5457.

[326] D. Sidler, M. Cristòfol-Clough and S. Riniker, *Journal of Chemical Theory and Computation*, 2017, **13**, 3020–3030.

[327] C. Yang, H. Kim and Y. Pak, *Journal of Chemical Theory and Computation*, 2020, **16**, 1827–1833.

[328] S. Syed, V. Romaniello, T. Campbell and A. Bouchard-Côté, 2021, arXiv:2102.07720.

[329] W. Jiang and B. Roux, *Journal of Chemical Theory and Computation*, 2010, **6**, 2559–2565.

[330] V. I. Manousiouthakis and M. W. Deem, *The Journal of Chemical Physics*, 1999, **110**, 2753–2756.

[331] A. Doucet, N. de Freitas and N. Gordon, in *Sequential Monte Carlo Methods in Practice*, ed. A. Doucet, N. de Freitas and N. Gordon, Springer, New York, 2001, pp. 3–14.

[332] P. Del Moral, *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*, Springer, New York, London, 2011.

[333] Y. Zhou, A. M. Johansen and J. A. Aston, *Journal of Computational and Graphical Statistics*, 2016, **25**, 701–726.

[334] L. Y. Barash, M. Weigel, M. Borovský, W. Janke and L. N. Shchur, *Computer Physics Communications*, 2017, **220**, 341–350.

[335] A. Kong, *Technical Reports, Department of Statistics, University of Chicago*, 1992, 348.

[336] L. Martino, V. Elvira and F. Louzada, *Signal Processing*, 2017, **131**, 386–401.

[337] A. Buchholz, N. Chopin and P. E. Jacob, *Bayesian Analysis*, 2021, **16**, 745–771.

[338] C. A. Naesseth, F. Lindsten and T. B. Schön, *Foundations and Trends® in Machine Learning*, 2019, **12**, 307–392.

[339] A. Rizzi, J. Chodera, L. Naden, K. Beauchamp, P. Grinaway, J. Fass, A. D. Wade, B. Rustenburg, G. A. Ross, A. Krämer, H. B. Macdonald, D. W. Swenson, A. Simmonett, J. Rodríguez-Guerra, D. A. Rufa, M. F. Henry, S. Roet and hb0402, *OpenMMTools*, 2019, `https://github.com/choderalab/openmmtools`.

[340] G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni and G. Bussi, *Computer Physics Communications*, 2014, **185**, 604–613.

[341] N. Michaud-Agrawal, E. J. Denning, T. B. Woolf and O. Beckstein, *Journal of Computational Chemistry*, 2011, **32**, 2319–2327.

[342] R. J. Gowers, M. Linke, J. Barnoud, T. J. E. Reddy, M. N. Melo, S. L. Seyler, J. Domański, D. L. Dotson, S. Buchoux, I. M. Kenney and Oliver Beckstein, Proceedings of the 15th Python in Science Conference, Austin, 2016, pp. 98–105.

[343] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, A. Vijaykumar, A. P. Bardelli, A. Rothberg, A. Hilboll, A. Kloeckner, A. Scopatz, A. Lee, A. Rokem, C. N. Woods, C. Fulton, C. Masson, C. Häggström, C. Fitzgerald, D. A. Nicholson, D. R. Hagen, D. V. Pasechnik, E. Olivetti, E. Martin, E. Wieser, F. Silva, F. Lenders, F. Wilhelm, G. Young, G. A. Price, G.-L. Ingold, G. E. Allen, G. R. Lee, H. Audren, I. Probst, J. P. Dietrich, J. Silterra, J. T. Webber, J. Slavič, J. Nothman, J. Buchner, J. Kulick, J. L. Schönberger, J. V. de Miranda Cardoso, J. Reimer, J. Harrington, J. L. C. Rodríguez, J. Nunez-Iglesias, J. Kuczynski, K. Tritz, M. Thoma, M. Newville, M. Kümmerer, M. Bolingbroke, M. Tartre, M. Pak, N. J. Smith, N. Nowaczyk, N. Shebanov, O. Pavlyk, P. A. Brodtkorb, P. Lee, R. T. McGibbon, R. Feldbauer, S. Lewis, S. Tygier, S. Sievert, S. Vigna, S. Peterson, S. More, T. Pudlik, T. Oshima, T. J. Pingel, T. P. Robitaille, T. Spura, T. R. Jones, T. Cera, T. Leslie, T. Zito, T. Krauss, U. Upadhyay, Y. O. Halchenko, Y. Vázquez-Baeza and SciPy 1.0 Contributors, *Nature Methods*, 2020, **17**, 261–272.

[344] A. Morton and B. W. Matthews, *Biochemistry*, 1995, **34**, 8576–8588.

[345] B. Q. Wei, W. A. Baase, L. H. Weaver, B. W. Matthews and B. K. Shoichet, *Journal of Molecular Biology*, 2002, **322**, 339–355.

[346] P. Czodrowski, G. Hölzemann, G. Barnickel, H. Greiner and D. Musil, *Journal of Medicinal Chemistry*, 2015, **58**, 457–465.

[347] Y. Sugita and Y. Okamoto, *Chemical Physics Letters*, 1999, **314**, 141–151.

[348] Y. Mori and H. Okumura, *The Journal of Physical Chemistry Letters*, 2013, **4**, 2079–2083.

[349] J. I. Monroe and M. R. Shirts, *Journal of Computer-Aided Molecular Design*, 2014, **28**, 401–415.

[350] F. Faizi, G. Deligiannidis and E. Rosta, *Journal of Chemical Theory and Computation*, 2020, **16**, 2124–2138.

[351] F. Chen, L. Lovász and I. Pak, Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing, New York, 1999, pp. 275–281.

[352] K. S. Turitsyn, M. Chertkov and M. Vucelja, *Physica D: Nonlinear Phenomena*, 2011, **240**, 410–414.

[353] Y. Sakai and K. Hukushima, *Journal of the Physical Society of Japan*, 2013, **82**, 064003.

[354] N. Rathore, M. Chopra and J. J. de Pablo, *The Journal of Chemical Physics*, 2004, **122**, 024111.

[355] A. Kone and D. A. Kofke, *The Journal of Chemical Physics*, 2005, **122**, 206101.

[356] R. Denschlag, M. Lingenheil and P. Tavan, *Chemical Physics Letters*, 2009, **473**, 193–195.

[357] F. A. Escobedo and F. J. Martínez-Veracoechea, *The Journal of Chemical Physics*, 2007, **127**, 174103.

[358] F. A. Escobedo and F. J. Martinez-Veracoechea, *The Journal of Chemical Physics*, 2008, **129**, 154107.

[359] J. Breen and S. Kirkland, *Linear Algebra and its Applications*, 2017, **520**, 306–334.

[360] N. Hansen, in *Towards a New Evolutionary Computation: Advances in the Estimation of Distribution Algorithms*, ed. J. A. Lozano, P. Larrañaga, I. Inza and E. Bengoetxea, Springer, Berlin, Heidelberg, 2006, pp. 75–102.

[361] N. Hansen, A. Auger, R. Ros, S. Finck and P. Pošík, Proceedings of the 12th Annual Conference Companion on Genetic and Evolutionary Computation, New York, 2010, pp. 1689–1696.

[362] Y. Li and K. Nam, *Journal of Chemical Theory and Computation*, 2020, **16**, 4776–4789.

[363] J. S. T. Wong, J. J. Forster and P. W. F. Smith, *Statistics and Computing*, 2020, **30**, 799–816.

[364] L. Breiman, *Machine Learning*, 1996, **24**, 123–140.

[365] B. W. Zhang, J. Xia, Z. Tan and R. M. Levy, *The Journal of Physical Chemistry Letters*, 2015, **6**, 3834–3840.

[366] Z. Tan, J. Xia, B. W. Zhang and R. M. Levy, *The Journal of Chemical Physics*, 2016, **144**, 034107.

[367] B. W. Zhang, N. Deng, Z. Tan and R. M. Levy, *Journal of Chemical Theory and Computation*, 2017, **13**, 4660–4674.

[368] M. Rocklin, Proceedings of the 14th Python in Science Conference, Austin, 2015, pp. 130–136.

[369] G. O. Roberts and J. S. Rosenthal, *Journal of Applied Probability*, 2007, **44**, 458–475.

[370] J. D. Hol, T. B. Schon and F. Gustafsson, 2006 IEEE Nonlinear Statistical Signal Processing Workshop, Cambridge, 13, pp. 79–82.

[371] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke and T. E. Oliphant, *Nature*, 2020, **585**, 357–362.

[372] S. K. Lam, A. Pitrou and S. Seibert, Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC - LLVM '15, Austin, 2015, pp. 1–6.

[373] N. Hansen, Y. Akimoto and P. Baudis, *Pycma*, 2020, https://github.com/CMA-ES/pycma.

[374] E. Fix and J. L. Hodges, *International Statistical Review / Revue Internationale de Statistique*, 1989, **57**, 238–247.

[375] P. A. M. Dirac, *The Principles of Quantum Mechanics*, Oxford University Press, Oxford, New York, 1981.

[376] P. Grinaway, J. D. Chodera, H. B. Macdonald, J. M. Behr, I. Zhang, D. A. Rufa, J. Rodríguez-Guerra, M. F. Henry, M. Wittman, W. Glass, S. Albanese, A. Silveira and L. N. Naden, *Perses*, 2014, https://github.com/choderalab/perses.

[377] B. Baum, D. Hangauer, A. Heine and G. Klebe, *Exploring Thrombin S1-Pocket*, Protein Data Bank, 2008, https://www.rcsb.org/structure/2ZFF, (accessed September 2021).

[378] M. Amaral, D. B. Kokh, J. Bomke, A. Wegener, H. P. Buchstaller, H. M. Eggenweiler, P. Matias, C. Sirrenberg, R. C. Wade and M. Frech, *Nature Communications*, 2017, **8**, 2276.

[379] H. M. Baumann, V. Gapsys, B. L. de Groot and D. L. Mobley, *The Journal of Physical Chemistry B*, 2021, **125**, 4241–4261.

[380] H. Chen, J. D. C. Maia, B. K. Radak, D. J. Hardy, W. Cai, C. Chipot and E. Tajkhorshid, *Journal of Chemical Information and Modeling*, 2020, **60**, 5301–5307.

[381] M. Bergdorf, A. Robinson-Mosher, X. Guo, K.-H. Law and D. E. Shaw, *DE Shaw Research Technical Report DESRES/TR—2021-01*, 2021.

[382] K. Fukunaga and L. Hostetler, *IEEE Transactions on Information Theory*, 1975, **21**, 32–40.

[383] S. Kullback and R. A. Leibler, *The Annals of Mathematical Statistics*, 1951, **22**, 79–86.

[384] J. R. Hershey and P. A. Olsen, 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, Honolulu, 2007, pp. IV–317–IV–320.