THE UNIVERSITY *of* EDINBURGH

New Methods for Alchemical Absolute Binding Free Energy Calculations

Finlay Clark

Submitted in partial fulfilment
of the requirements for the
degree of Doctor of Philosophy

The University of Edinburgh

2025

Abstract

Alchemical absolute binding free energy (ABFE) calculations are of increasing interest in drug discovery. They can predict the binding affinities of structurally dissimilar ligands to their targets and offer higher accuracy than alternative methods. However, their widespread application is still limited by high computational cost, lack of automation, and inaccuracy. This work investigates methods to improve the speed, accuracy, and automation of ABFE calculations.

Receptor-ligand restraint schemes were compared. A new scheme based on multiple distance restraints was proposed which avoids inherent instabilities and may provide convergence benefits. This produced results comparable to the common Boresch scheme, while omitting orientational restraints led to large errors.

A fully automated ABFE workflow was developed, including automated λ-window selection, the ensemble-based detection of equilibration, and the adaptive allocation of sampling time based on inter-replicate statistics. The workflow produced equivalent results to a nonadaptive scheme over several test systems, while often accelerating equilibration.

White's marginal standard error rule was reformulated to provide a spectrum of equilibration detection heuristics applicable to single simulations. These were tested on ensembles of synthetic time series modelled on free energy change estimates from long ABFE calculations. Methods that more thoroughly accounted for autocorrelation often showed late and variable truncation times, while methods that less thoroughly accounted for autocorrelation often showed early truncation, relative to the optimal truncation point. A method was identified which achieved robust performance across test sets by balancing these extremes.

The performance of extremely fast ABFE calculations was investigated over a range of test systems and compared to "standard-length" calculations. Short ABFE calculations showed large absolute errors, but often retained similar ranking performance to "standard" calculations.

Lay Summary

Drug discovery is expensive and slow. Drugs normally work by sticking ("binding") to a target protein, which triggers a beneficial biological effect. If the strength of binding could be predicted with computers, scientists could avoid making molecules which are bad drugs because they do not bind to the target protein. This could make drug discovery faster and cheaper, accelerating the creation of new treatments. This work focuses on a method to predict the binding strengths of varied molecules, called alchemical absolute binding free energy (ABFE) calculations.

Unfortunately, ABFE calculations are currently too slow to be routinely used in drug discovery. Also, users often have to make many decisions about how the calculations are run, which wastes human time or results in the use of bad defaults.

This work investigates methods to make ABFE calculations more efficient and automated. These include methods to automatically decide how much time to spend on a prediction, and what data to use. The methods are made available in free software. The automated methods are shown to be reliable, saving human time, and sometimes improve efficiency, saving computer time. It is also shown that very fast ABFE calculations can sometimes provide useful predictions. Hopefully, these methods will help ABFE calculations become a useful tool in drug discovery, and contribute to the creation of new drugs.

Acknowledgements

As an early undergraduate, I had no interest in graduate research. I am glad I changed my mind. I feel lucky to have had this opportunity, and I am grateful to the many people who made my job so enjoyable.

One of the best parts of my role has been the freedom allowed by my academic supervisors, Julien Michel and Danny Cole. Thank you Julien for keeping me on track while trusting me with independence, and for creating an excellent research environment. Being in your group has been a pleasure. Danny, thank you for your enthusiastic guidance. I am excited to work closely with you soon.

Completing industrial placements was hugely valuable and helped me understand the broader context of my work. Thank you Graeme Robb for organising such useful placements, and thank you to the whole AstraZeneca computational chemistry team for making them so enjoyable. Thank you Martin Packer for suggesting a suitable benchmark system and providing input files.

Almost all of the work in this thesis is based on code developed and maintained by Lester Hedges and Christopher Woods (BioSimSpace and Sire). Thank you for being so responsive and helpful when I had issues, and patient and encouraging when I made (minor) contributions.

Also, thank you to Mathew Topper, whose PhD thesis template I stole. Thank you Toni Mey and Gerhard König, for conducting my viva.

Too many people deserve thanks in room 234 (both permanent and lunch-time residents). It has been a pleasure learning from and having lunch with you all. In particular, thank you Audrius for the conversations and computer help. If I had listened more carefully I could have avoided several disasters. Anna, sorry for getting you into bikes. Eva, thanks for the chats about music, books, and restraints. Chenfeng, thanks for sharing your rap at last. Remember you promised me an recording on graduation. João, thanks for the chats about science, music, and everything else. Shivani, thanks for making me sit up straight and eat lunch on time. One day I will upgrade from bread sandwiches. Roy, you are a quick learner and I am sure you will do some great work. Adele, thanks for getting me through my first SLURM installation.

My enjoyment of work has benefitted from plenty of time spent away from the lab. Thanks to my biking friends - you are all insane. It's hard to say who is the craziest but on balance I think Adam is the biggest danger to himself and Dave is the biggest danger to society. Sam comes a close second in both categories. Thank you to my Cambridge flatmates Henry and Dani, who made my second placement in Cambridge a lot of fun - especially the late-night Rage Against the Machine sessions.

My family has been very supportive of my work, as with everything else. Thank you. Granny, I am afraid I cannot say when I will earn "proper money".

Lastly, thank you Emily. Thanks for your patience when I work too much, and when maths abducts my brain. Thanks for your love, always.

Contents

Figures and Tables

Figures

Tables

Nomenclature

Roman Symbols

c^o	Standard State Concentration, $\frac{1}{1660}$ Å^{-2}
F	Helmholtz Free Energy
G	Gibbs Free Energy
g	Statistical Inefficiency
H	Enthalpy
k_B	Boltzmann Constant
Q	Partition Function
S	Entropy
T	Temperature
U	Potential Energy
V^o	Standard State Volume, $\frac{1}{c^o}$, 1660 Å^2
Z	Configurational Partition Function

Greek Symbols

β	Inverse temperature, $1/k_\mathrm{B}T$
Δ	Denotes a finite change in a quantity
λ	Alchemical progress parameter which interpolates Hamiltonians
μ	Chemical potential
σ	Standard deviation

Other Symbols

$\hat{x}$	Denotes that x is estimated
$\langle ... \rangle$	Denotes the average of the contained quantity
^{o}	Denotes a *standard* free energy or other quantity

Acronyms / Abbreviations

ABFE Absolute Binding Free Energy

ADME Adsoption, Distribution, Metabolism, and Excretion

AMBER Assisted Model Building and Energy Refinement

BAR Bennett Acceptance Ratio

CADD Computer-Aided Drug Design
CI Confidence Interval
CPU Central Processing Unit
DNA Deoxyribonucleic Acid
ESS Effective Sample Size
FEP Free Energy Perturbation
GPU Graphics Processing Unit
GROMACS GROningen MAchine for Chemical Simulation
HREX Hamiltonian Replica Exchange
IC_{50} Half Maximal Inhibitory Concentration
ITC Isothermal Titration Calorimetry
LJ Lennard-Jones
MAE Mean Absolute Error
MBAR Multistate Bennett Acceptance Ratio
MC Monte Carlo
MCMC Markov Chain Monte Carlo
MD Molecular Dynamics
MM Molecular Mechanics
MM/GBSA Molecular Mechanics/ Poisson-Bolzmann Surface Area
MM/PBSA Molecular Mechanics/ Generalised Born Surface Area
MSER Marginal Standard Error Rule
MUE Mean Unsigned Error
NPT Number Pressure Temperature
NVE Number Volume Energy
NVT Number Volume Temperature
OFF Open Force Field
PDB Protein Data Bank
PME Particle-Mesh Ewald
QM Quantum Mechanics
QSAR Quantitative Structure-Activity Relationship
RBFE Relative Binding Free Energy
REST Replica Exchange with Solute Tempering
RF Reaction Field
RMSE Root Mean Square Error
SAR Structure-Activity Relationship
SEM Standard Error of the Mean
SOMD Sire/OpenMM Molecular Dynamics

SPR Surface Plasmon Resonance

TI Thermodynamic Integration

Chapter 1

Introduction

Drug discovery is expensive and slow. It is a complex inter-disciplinary research effort which begins with the identification of an unmet medical need and proceeds to large-scale clinical trials. It has a failure rate of over 90 % in the clinical stages alone, and is estimated to take 15 years and $2 billion USD.[2,3] To accelerate the development of new treatments, this process must be made more efficient.

Drugs are commonly small molecules which exert their biological effect by binding to a disease-modulating target. If the binding affinity could be predicted quickly and accurately, this would prevent the wasteful synthesis and testing of inactive molecules, substantially accelerate early-stage small-molecule drug discovery. Alchemical absolute binding free energy (ABFE) calculations are promising because they predict the binding affinities of structurally dissimilar molecules more accurately than alternative methods. However, their practical application remains limited by high computational cost, lack of automation, and inaccuracy. Here, computer-aided drug discovery is briefly reviewed, the theoretical basis of ABFE calculations is outlined, and outstanding challenges are highlighted.

1.1 Computer-Aided Drug Design

1.1.1 The Drug Discovery Pipeline

This work addresses a challenge in the development of small molecules as drugs. Most drugs approved by the United States Food and Drug Administration in 2023 were small molecules.[4] These generally exert their desired biological effect by acting as ligands which bind to a target (usually a protein), which may alter chemical signalling or inhibit enzyme activity, for example.[5]

The dominant approach in small-molecule drug discovery is target-based discovery.[6] It begins with identifying an unmet medical need and a drug target. Once the target is confirmed to offer a valid route to treating the disease, small molecules showing promising activity against this target, referred to as "hits", are identified. Selected hits are then optimised to generate "lead" compounds with more favourable property profiles, which are further optimised to produce a preclinical drug candidate. These stages are the "hit discovery", "hit-to-lead", and "lead optimisation" stages of early drug discovery. Clinical trials then determine the safety (phase I), efficacy (phase II), and large-population efficacy (phase III) of the drug candidate, providing the evidence required for regulatory approval.[7]

This process is estimated to take 15 years and $2 billion USD.[3] However, this is an average which masks a failure rate of over 90 % in the clinical stages alone.[2] Clearly, innovations which reduce the financial and time investment required for drug discovery would accelerate the development of new therapeutics. While the clinical stages are the most expensive, most opportunities to expedite drug discovery exist in the earlier stages.[8] Additionally, improvements to early-stage drug discovery may reduce failures in the clinical stages by improving ligand properties.

1.1.2 Early-Stage Drug-Discovery: The Hit Discovery, Hit-to-lead, and Lead Optimisation Stages

Collectively, hit discovery, hit-to-lead, and lead optimisation are intended to yield a preclinical drug candidate likely to prove effective and safe in clinical trials. This is challenging because because the effectiveness of a drug is decided by many interrelated properties which may be difficult to predict and measure.[9] The drug candidate must not only bind to the validated target but also show favourable absorption, distribution, metabolism, and excretion (ADME), toxicity, and specificity profiles.

Binding affinity is often one of the first properties to be effectively optimised. The purpose of hit discovery is to yield diverse molecules that show promising activity in a screening assay, which requires sufficient binding affinity.[7] High-throughput screening of large compound libraries (10^5 - 10^6 molecules) remains a popular, if resource-intensive, approach to hit discovery.[10]

The hit-to-lead stage involves the optimisation of promising hit molecules. Generally, three to five chemical series are selected to reduce the chance of becoming stuck in unfavourable property space.[7] The molecular property space is explored through iterative "design make test analyse" cycles. During the deductive design and make stages, models of property space are used to predict improved molecules; during the inductive test and analyse stages, the results of tests are used to evaluate and improve these models.[11] These models include chemist's "chemical intuition", machine learning, molecular modelling, and (quantitative) structure-activity relationship ((Q)SAR) models. Heuristics like Lipinski's Rule of 5 are often used to roughly map favourable regions in complex property space to simple physical and chemical properties.[12,13] Lipinsiki's Rule of 5 predicts that molecules which violate no more than one of a set of four rules (no more than 5 hydrogen bond donors, no more than 10 hydrogen bond acceptors, molecular mass less than 500 Dalton, and calculated $\log P$ (n-octanol-water partition coefficient) no more than 5) are more likely to be orally bioavailable.[12] The tests cover a broad range of important properties, including activity (screening against target), selectivity (e.g. kinase panel screening), toxicity (e.g. screening against off-targets such as hERG), and ADME (e.g. Caco-2 permeability).[14–16] Promising compounds are selected for further improvement during lead optimisation. Animal models may also be used at this stage, although they are most heavily used during pre-clinical testing, where their high cost is justified by the rich information they yield on likely in-human performance.[17]

Throughout these stages, binding affinity plays a key role. The molecule is unlikely to be an effective drug if the binding affinity is low. Increasing affinity reduces the required dose and may increase specificity, reducing side effects due to interactions with off-targets.[18] However, it should be noted that over-reliance on binding affinity optimisation to a single target is likely a highly inefficient approach to drug discovery.[6]

A multitude of methods exist to measure binding affinity.[19] For competitive antagonists (for example, a molecule which blocks an enzyme's active site), it is often measured indirectly as the half-maximal inhibitory concentration (IC_{50}) from a biological assay.[20] Isothermal titration calorimetry (ITC), regarded as the gold standard of affinity measurements, involves titration of the target and ligand at constant temperature while the heat released or absorbed is measured. This allows direct estimation of binding affinity and its enthalpic and entropic components.[19] Another more direct measurement of binding affinity is surface plasmon resonance (SPR), where one binding partner is immobilised on a thin

metal sheet while a solution of the other is passed over. Changes in the refractive index of the sheet are proportional to changes in the bound mass resulting from binding events. These changes are monitored and used to determine binding affinity.[19]

1.1.3 Computational Binding Affinity Prediction

Unfortunately, accurate and direct affinity measurement methods such as ITC and SPR are relatively expensive and slow. Additionally, performing any affinity measurement requires either the presence of the compound in an accessible library or its potentially costly synthesis. Rapid and accurate computational estimates of binding affinity would lift these restrictions, allowing the screening of vast virtual libraries and avoiding the wasteful synthesis of low-affinity compounds. Even noisy affinity estimates may provide substantial efficiency gains; if the goal is to improve a molecule's affinity 10-fold, then affinity estimates with 1 kcal mol^{-1} of noise may reduce the required number of compounds to be synthesised 5-fold, under some assumptions.[21]

Computer-aided drug design methods can be classified as either ligand-based, when only the structure of the ligand is available, or structure-based, when the three-dimensional structure of the target is known. Recent advances in protein structure prediction (e.g. AlphaFold) and structural biology (for example in cryo electron microscopy) have driven the increasing use of the more information-rich structure-based methods,[22–27] and the discussion here is restricted to these. A more complete discussion of the concepts underlying these methods is given in the remainder of this chapter, but here they are briefly put in the context of drug discovery.

The first stage in structure-based affinity prediction is obtaining the ligand-target complex structure. If a co-crystal structure is unavailable, molecular docking is commonly used to generate the complex. Docking algorithms generally search for favourable binding poses of a fully flexible ligand within a rigid receptor and rank the poses generated using physical, empirical, knowledge-based, or machine-learning scoring functions.[28] The performance of later affinity predictions is strongly dependent on the accuracy of the input poses.[29,30]

The second stage is selection of the affinity prediction method. These span a wide range of costs and accuracies. Scoring functions from docking offer fast affinity estimates suitable for screening up to $\approx 10^9$ molecules, but docking scores generally correlate poorly with experimental affinities.[31] Recently, there has been enormous interest in machine learning scoring functions, but these often learn data set biases instead of physical interactions, making them poorly generalisable to complexes dissimilar to the training data.[32–35]

End-state methods are a slower but more accurate class of techniques. These include linear interaction energy (LIE) and molecular mechanics Poisson-Boltzmann surface area (MM/PBSA).[36–40] In the LIE method, simulations of the solvated ligand-target complex and the solvated ligand are performed and differences in average ligand interaction energies are calculated using a molecular mechanics force field (an approximate energy function).[37,41,42] An expression containing these terms is fit to available experimental binding affinities and used to predict the affinities of new molecules. MM/PBSA calculations are often based on a single simulation of the ligand-target complex.[42] They approximate the binding affinity as a sum of terms which include molecular mechanics energies and polar and non-polar hydration free energies which are calculated using the Poisson-Boltzmann method and solvent accessible surface area method, respectively.[38,39] MM/GBSA is a cheaper variant which calculates the polar hydration free energy term using the generalised Born method.[39] LIE, MM/PBSA, and MM/GBSA generally perform similarly.[42,43] Another end-state approach is mining minima, in which the configurational partition functions of the complex, free ligand, and free target are approximated as sums of local configurational integrals from low energy local minima.[44,45] Unfortunately, the accuracy of end-state methods is limited by approximations such as the implicit treatment of solvation and the neglect or approximation of entropic contributions to binding.[46]

In principle, methods based on quantum mechanics (QM) should offer the most rigorous estimates of binding affinity.[47] However, the high cost and poor scaling of QM methods make thorough conformational sampling impossible. Hence, improvements from more accurate energies are offset by errors from incomplete conformational sampling.

In practice, the slowest but most accurate affinity prediction methods use simulations with molecular mechanics force fields to sample end-states of interest and spanning intermediate states. These methods are often referred to as "rigorous" and "physics-based" because they are rigorously derived from the prin-

ciples of statistical mechanics and make less physical approximations than other methods.[48–51] Alchemical relative binding free energy (RBFE) calculations have become the gold standard for affinity prediction in pharmaceutical research and are routinely employed in hit-to-lead and lead optimisation efforts.[52–54] These use unphysical, "alchemical", intermediate states which interconvert structurally related ligands. However, the requirement for similar ligands bound to the same target with the same binding pose excludes important problems such as calculating the RBFEs of structurally dissimilar ligands to a common target, ranking different binding poses, optimising selectivity or promiscuity by calculating the RBFEs of the same ligand to various (off) targets, and predicting the functional response of ligand binding from the change in binding affinity between different conformational states of the same target.[55]

These limitations are overcome by alchemical absolute binding free energy (ABFE) calculations, which completely remove a ligand's intermolecular interactions.[49–51] These have shown promise for computing the binding affinities of diverse fragments and merged molecules,[1] evaluating binding selectivity,[56] and predicting the functional response of ligand binding.[55] ABFE calculations are also promising for applications in the hit discovery stage. The ability to perform high-throughput screening computationally would lift the restrictions imposed by limited physical compound libraries, allowing the efficient screening of larger and more diverse virtual libraries. Current approaches to virtual screening are varied but usually employ a hierarchy of methods, where fast methods are used to rapidly eliminate most molecules before slower, higher-accuracy methods are applied.[57] However, the performance of virtual screening is often limited by the poor accuracy of fast affinity prediction methods. ABFE calculations are promising as a final stage in virtual screening because they provide relatively accurate affinity estimates for structurally diverse molecules.[58] They outperform end-state calculations and substantially increase hit rates in virtual screening over docking alone.[59–61] Their real-world utility in virtual screening is highlighted by examples including a winning entry to the Critical Assessment of Computational Hit-Finding Experiments Challenge #1.[62,63]

Clearly, ABFE calculations are promising for many applications in early-stage drug discovery. However, their widespread use is limited by their high computational cost compared to relative calculations, lack of automation, and sometimes limited accuracy. The remainder of this chapter examines the theoretical basis of alchemical ABFE calculations and highlights areas for improvement.

1.2 Statistical Thermodynamics

"Rigorous" methods for binding affinity estimation are derived from the principles of statistical mechanics. Statistical mechanics connects microscopic properties to macroscopic experimental observables, such as binding affinity measured with SPR.

The central problem in statistical mechanics is determining the probability of a system being in a particular "microstate", uniquely defined by microscopic properties such as the momenta and relative positions of all particles. These probabilities are affected by macroscopic constraints on the system, such as constant temperature, pressure, and number of particles in a lab experiment. From these probabilities, average values of observables can be calculated. For macroscopic systems containing $\approx 10^{23}$ particles, fluctuations around these averages are usually negligible, making them effectively exact. The solution to this problem is given by the (generalised) Boltzmann distribution.

1.2.1 The Boltzmann Distribution

As put by Schrödinger:[64]

"There is, essentially, only one problem in statistical thermodynamics: the distribution of a given amount of energy E over N identical systems. Or perhaps better: to determine the distribution of an assembly of N identical systems over the possible states in which this assembly can find itself, given that the energy of the assembly is a constant E."

This assembly is referred to as an ensemble, and it contains a very large number of systems, N. The systems are coupled in the sense the total energy of the ensemble is constant, but the energy of interaction between systems is negligible. In an experiment or simulation, we can imagine that we are observing a single system coupled to a "bath" of all the other ensemble members.

The solution is the Boltzmann distribution,[65,66]

$$p_i \propto e^{-\beta E_i} = e^{-\frac{E_i}{k_\mathrm{B}T}}, \tag{1.1}$$

where p_i is the probability of finding a system in the microstate i with energy E_i. $\beta = \frac{1}{k_\mathrm{B}T}$, where k_B is the Boltzmann constant and T is the absolute temperature.

The law of equal a priori probabilities states that each way of distributing the N systems over the allowed microstates (satisfying the constraint of constant total energy) is equally probable. From this starting point, the probabilities of each microstate can be derived by maximising the number of ways of distributing the N systems, which is equivalent to maximising the entropy. This is the Boltzmann distribution, and it applies when N is large.[67] The normalised probabilities are

$$p_i = \frac{e^{-\beta E_i}}{\sum_j e^{-\beta E_j}} = \frac{e^{-\beta E_i}}{Q}, \tag{1.2}$$

where the normalising constant, $Q = \sum_j e^{-\beta E_j}$, is called the partition function. All of the thermodynamic properties of the system can be calculated from it.[67]

Ensembles are often referred to by the constraints imposed on each system. For example, the NVE (microcannonical) ensemble fixes the number of particles, the volume, and the energy of the systems. These are all extensive variables (proportional to the size of the system). By fixing the E_is in Equation 1.2, it can be seen that microstate probabilities in the NVE ensemble are given by a flat distribution

$$p_{i,NVE} = \frac{1}{\Omega}, \tag{1.3}$$

where Ω is the number of microstates.

The NVT (canonical) ensemble constrains the intensive variable (independent of the system size) temperature instead of the extensive variable energy. Fixing this intensive variable for every system corresponds to fixing an extensive variable of the ensemble. This can be shown using the first law of thermodynamics

$$dE = \left(\frac{\partial E}{\partial S}\right)_{V,N} dS - \left(-\frac{\partial E}{\partial V}\right)_{S,N} dV + \sum_{k=1}^{K} \left(-\frac{\partial E}{\partial N_k}\right)_{S,V,N_{i\neq k}} dN_k \tag{1.4}$$

$$= TdS - PdV + \sum_{k=1}^{K} \mu_k dN_k, \tag{1.5}$$

where S is the entropy, P is the pressure, and there are N_k molecules of each of the K different types of molecule with chemical potential μ_k. As the ensemble is at thermodynamic equilibrium, the entropy is maximised and $dS = 0$. V and all N_k are fixed, and therefore $dV = dN_k = 0, \forall k$. Hence, $dE = 0$ and the energy of the ensemble is constant. These are the assumptions made during the derivation of the Boltzmann equation, and therefore the probabilities of microstates in the NVT ensemble are given by the Boltzmann distribution (Equation 1.2).[68]

Laboratory experiments are usually conducted at constant pressure, rather than constant volume, so the NPT (isothermal isobaric) ensemble is appropriate. In this case, applying the first law (Equation 1.5) to the ensemble yields

$$dE + PdV = 0, \tag{1.6}$$

which implies that $E + PV$ is constant for the ensemble. $H = E + PV$ is the enthalpy. Re-deriving the Boltzmann distribution with the constraint of fixed total enthalpy, rather than energy, yields the probabilities of microstates in the NPT ensemble

$$p_{i,NPT} = \frac{e^{-\beta H_i}}{\sum_j e^{-\beta H_j}}. \tag{1.7}$$

Equations 1.7 and 1.2 are both special cases of the generalised Boltzmann distribution.[69]

Under the ergodic hypothesis, the infinite time average of a quantity of interest, A, is equivalent to the ensemble average

$$\langle A \rangle_{\text{Time}} \overset{\text{Time}\to\infty}{=} \sum_j A_j p_j, \tag{1.8}$$

where A_j is the value of A for the microstate j. This assumption is sometimes used to justify the estimation of ensemble averages using time averages. Alternatively, statistical mechanics can be viewed as a form of statistical inference, rather than a physical theory, which is not reliant on assumptions such as ergodicity and equal a priori microstate probabilities to be correct. Maximising the entropy is justified as a way of making maximally unbiased predictions, rather than as a law of physics.[70] This provides a more general conceptual framework but produces the same relationships as above.

While this discussion has assumed discrete microstates, all results can be trivially generalised to the continuous case by replacing sums with integrals.

1.2.2 Energy, Entropy, and Free Energy

We are usually interested in macrostates related to experimental observables. Macrostates contain many microstates. For a system containing a ligand and its protein target, two interesting macrostates are the bound (protein-ligand complex) and free (separated protein and ligand) states (Figure 1.1). Considering a simple model of this system provides intuition for energy, entropy, and free

energy. Let the bound and free macrostates consist of Ω_{Bound} and Ω_{Free} microstates, all with energy E_{Bound} or E_{Free}. The relative probabilities of observing each macrostate in the NVT ensemble are

$$\frac{p_{\text{Bound}}}{p_{\text{Free}}} = \frac{\sum_{\text{Bound}} e^{-\frac{E_{\text{Bound}}}{k_{\text{B}}T}}}{\sum_{\text{Free}} e^{-\frac{E_{\text{Free}}}{k_{\text{B}}T}}} \tag{1.9}$$

$$= \frac{\Omega_{\text{Bound}} e^{-\frac{E_{\text{Bound}}}{k_{\text{B}}T}}}{\Omega_{\text{Free}} e^{-\frac{E_{\text{Free}}}{k_{\text{B}}T}}} \tag{1.10}$$

$$= \frac{\Omega_{\text{Bound}}}{\Omega_{\text{Free}}} e^{-\frac{\Delta E_{\text{Bind}}}{k_{\text{B}}T}}, \tag{1.11}$$

where $\sum_{\text{Free}}$ and $\sum_{\text{Bound}}$ are sums over the microstates in the free and bound macrostates and $\Delta E_{\text{Bind}} = E_{\text{Bound}} - E_{\text{Free}}$.

A increase in $\frac{p_{\text{Bound}}}{p_{\text{Free}}}$ corresponds to greater binding affinity. For a potent ligand in a system much larger than the binding site, a good model is $\Omega_{\text{Free}} > \Omega_{\text{Bound}}$ (there are more ways for a ligand to be unbound than bound) and $\Delta E_{\text{Bind}} > 0$ (the ligand forms strong interactions with the protein). These effects oppose each other: the reduced number of microstates disfavours the bound macrostate, while the decrease in energy favours it. An increase in temperature disfavours binding by reducing the effect of the energy decrease and $\frac{p_{\text{Bound}}}{p_{\text{Free}}} \to \frac{\Omega_{\text{Bound}}}{\Omega_{\text{Free}}}$ as $T \to \infty$. As the importance of the energy term is affected by temperature, but the microstate number ratio is not, these terms can be obtained separately using measurements of binding affinities at different temperatures.

Equation 1.11 can be written as

$$\frac{p_{\text{Bound}}}{p_{\text{Free}}} = \frac{e^{\frac{k_{\text{B}}T \ln \Omega_{\text{Bound}}}{k_{\text{B}}T}}}{e^{\frac{k_{\text{B}}T \ln \Omega_{\text{Free}}}{k_{\text{B}}T}}} e^{-\frac{\Delta E_{\text{Bind}}}{k_{\text{B}}T}} \tag{1.12}$$

$$= \frac{e^{\frac{TS_{\text{Bound}}}{k_{\text{B}}T}}}{e^{\frac{TS_{\text{Free}}}{k_{\text{B}}T}}} e^{-\frac{\Delta E_{\text{Bind}}}{k_{\text{B}}T}} \tag{1.13}$$

$$= e^{-\frac{\Delta E_{\text{Bind}} - T\Delta S_{\text{Bind}}}{k_{\text{B}}T}}, \tag{1.14}$$

where we have used Boltzmann's entropy formula $S = k_{\text{B}} \ln \Omega$. This clarifies that entropy specifies the number of microstates in a macrostate in a form which can be easily added to energies. This is especially useful since entropies and energies often cannot be measured separately. Instead, only the relative probabilities of macrostates at a given temperature may be obtained, meaning that only the "free

energy" of binding is obtained

$$\frac{p_{\text{Bound}}}{p_{\text{Free}}} = e^{-\frac{\Delta F_{\text{Bind}}}{k_{\text{B}}T}}, \tag{1.15}$$

where $\Delta F_{\text{Bind}} = \Delta E_{\text{Bind}} - T\Delta S_{\text{Bind}}$ is the Helmholtz free energy change upon binding (and $F = E - TS$). Free energies are additive, like energies, but include the effects of changing numbers of microstates through the entropy term. At a given temperature, $\frac{p_{\text{Bound}}}{p_{\text{Free}}}$, and therefore the binding affinity, is completely specified by the free energy of binding.

Comparing Equations 1.15 and 1.9 shows that

$$\Delta F_{\text{Bind}} = -k_{\text{B}}T \ln \frac{Q_{\text{Bound}}}{Q_{\text{Free}}}. \tag{1.16}$$

The probabilities of the microstates, and therefore the binding affinity, may depend on the constraints imposed on the ensemble. Laboratory experiments normally correspond to the NPT, not the NVT ensemble. For the NPT ensemble, all energies in the preceding discussion should be replaced by enthalpies, and $\frac{p_{\text{Bound}}}{p_{\text{Free}}}$ is determined by the Gibbs free energy of binding, $\Delta G_{\text{Bind}} = \Delta H_{\text{Bind}} - T\Delta S_{\text{Bind}}$.

1.2.3 The Statistical Mechanics of Ligand Binding

The discussion in Section 1.2.2 provides intuition but is imprecise. Here, the equations in Section 1.2.2 are generalised and connected to precise definitions of binding affinity. Their correspondence to experimental affinity measurements is discussed.

The equations given in Section 1.2.2 apply to discrete microstates, but microstates are approximated as continuous in classical mechanics. Writing the energies as a function of the phase-space coordinates (the positions $\boldsymbol{r}$ and momenta $\boldsymbol{p}$ of all particles), $\boldsymbol{\Gamma}$, Equation 1.9 becomes

$$\frac{p_{\text{Bound}}}{p_{\text{Free}}} = -k_{\text{B}}T \ln \frac{Q_{\text{Bound}}}{Q_{\text{Free}}} \tag{1.17}$$

$$= \frac{\int_{\text{Bound}} e^{-\frac{E(\boldsymbol{\Gamma})}{k_{\text{B}}T}} d\boldsymbol{\Gamma}}{\int_{\text{Free}} e^{-\frac{E(\boldsymbol{\Gamma})}{k_{\text{B}}T}} d\boldsymbol{\Gamma}}, \tag{1.18}$$

where $\int_{\text{Bound}}$ and $\int_{\text{Free}}$ indicate integration over the phase space coordinates of the bound and free states. In the general case where microstates have different probabilities, S is given by the Gibbs entropy formula

$$S = -k_{\text{B}} \sum_i p_i \ln p_i \tag{1.19}$$

which reduces to the Boltzmann formula when microstates when all microstates have the same probability. This differs from Shannon's entropy formula only by the factor of k_{B},[71] hinting at the deep connection between statistical mechanics and information theory.[70]

The energy is given by the Hamiltonian, $\mathcal{H}$

$$E(\mathbf{\Gamma}) = \mathcal{H}(\boldsymbol{p}, \boldsymbol{r}) = \sum_{i=1}^{N_{\text{Atoms}}} \frac{|\boldsymbol{p_i}|^2}{2m_i} + U(\boldsymbol{r}), \tag{1.20}$$

where N_{Atoms} is the number of atoms in the system, $\boldsymbol{p_i}$ and m_i are the momentum and mass of atom i, and $U(\boldsymbol{r})$ is the potential energy. Because $\mathcal{H}$ consists of parts which depend only on the momenta and only on the configuration of the system (positions), partition functions can be factored into momenta and configuration-dependent components. As the masses and temperatures are the same between the bound and free states, the momenta-dependent components of Equation 1.18 cancel, leaving only the configurational partition functions, Z_{Bound} and Z_{Free},

$$\frac{p_{\text{Bound}}}{p_{\text{Free}}} = -k_{\text{B}}T \ln \frac{Z_{\text{Bound}}}{Z_{\text{Free}}} \tag{1.21}$$

$$= \frac{\int_{\text{Bound}} e^{-\frac{U(\boldsymbol{r})}{k_{\text{B}}T}} d\boldsymbol{r}}{\int_{\text{Free}} e^{-\frac{U(\boldsymbol{r})}{k_{\text{B}}T}} d\boldsymbol{r}}. \tag{1.22}$$

If the entire system is translated or rotated in the laboratory frame of reference, the potential energies do not change. In other words, the potential energies are invariant to the laboratory frame of reference. Therefore integration need only be performed over the $3N_{\text{Atoms}} - 6$ internal degrees of freedom when calculating the configurational integrals. Therefore we will swap $\boldsymbol{r}$ for $\boldsymbol{r'}$, which denotes integration over only the internal degrees of freedom.

The system's potential energy in the free state is not invariant to changes in the coordinates of the ligand relative to the protein; if all other atomic positions are fixed and the ligand is translated, the system energy will change. However, the free state is usually defined so that there is no correlation between the degrees

of freedom of the protein and the ligand. Therefore, the energies are invariant to the relative coordinates of the protein and ligand *on average* and Equation 1.22 can be written

$$\frac{p_{\text{Bound}}}{p_{\text{Free}}} = \frac{\int_{\text{Bound}} e^{-\frac{U(\boldsymbol{r'})}{k_{\text{B}}T}} d\boldsymbol{r'}}{\int_{\text{Free}} d\boldsymbol{r_{PL}} \int_{\text{Free}} e^{-\frac{U(\boldsymbol{r'})}{k_{\text{B}}T}} d(\boldsymbol{r'} - \boldsymbol{r_{PL}})} \tag{1.23}$$

$$= \frac{\int_{\text{Bound}} e^{-\frac{U(\boldsymbol{r'})}{k_{\text{B}}T}} d\boldsymbol{r'}}{8\pi^2 V_{\text{PL,Free}} \int_{\text{Free}} e^{-\frac{U(\boldsymbol{r'})}{k_{\text{B}}T}} d(\boldsymbol{r'} - \boldsymbol{r_{PL}})}, \tag{1.24}$$

where $\boldsymbol{r_{PL}}$ denotes the ligand's 6 external degrees of freedom relative to the protein.[72] $V_{\text{PL,Free}}$ is the volume accessible to the ligand in the free state when the position of the protein is fixed, and comes from integrating over the positional part of $\boldsymbol{r_{PL}}$. $8\pi^2$ comes from integrating the orientational part of $\boldsymbol{r_{PL}}$ (the z-axis of a Cartesian coordinate system can be pointed towards any point on the surface of a sphere, giving 4π steradians, and for every z-axis position the x and y-axes can be rotated through 2π radians).

Equation 1.24 shows that $\frac{p_{\text{Bound}}}{p_{\text{Free}}}$ depends on the definitions of the free and bound states, both explicitly through the definition of bound and free configurations and implicitly through the system size, which affects $V_{\text{PL,Free}}$. To remove the system size dependence, standard state bound and free states are used, where the solutes (complex, ligand, and protein) are present at the standard state concentration ($C^o = 1$ mol l^{-1}) but only interact with the solvent. As a result, $V_{\text{PL,Free}}$ is replaced by the standard state volume $V^o = \frac{1}{C^o} = 1661$ Å^3. So long as protein-ligand interactions are negligible in the free state, this makes $\frac{p_{\text{Bound}^\circ}}{p_{\text{Free}^\circ}}$ independent of the definition of the free configurations and the size of the box. Similarly, $\frac{p_{\text{Bound}^\circ}}{p_{\text{Free}^\circ}}$ is insensitive to the exact definition of the bound state (as long as all important low-energy configurations are included), because Boltzmann factors with large negative energies dominate the integral over bound configurations.[49] Finally, we have

$$\frac{p_{\text{Bound}^\circ}}{p_{\text{Free}^\circ}} = \frac{\int_{\text{Bound}} e^{-\frac{U(\boldsymbol{r'})}{k_{\text{B}}T}} d\boldsymbol{r'}}{8\pi^2 V^o \int_{\text{Free}} e^{-\frac{U(\boldsymbol{r'})}{k_{\text{B}}T}} d(\boldsymbol{r'} - \boldsymbol{r_{PL}})}, \tag{1.25}$$

or equivalently

$$\Delta F^o_{\text{Bind}} = -k_{\text{B}}T \ln \frac{\int_{\text{Bound}} e^{-\frac{U(\boldsymbol{r'})}{k_{\text{B}}T}} d\boldsymbol{r'}}{8\pi^2 V^o \int_{\text{Free}} e^{-\frac{U(\boldsymbol{r'})}{k_{\text{B}}T}} d(\boldsymbol{r'} - \boldsymbol{r_{PL}})}. \tag{1.26}$$

Since ΔG^o_{Bind} and ΔF^o_{Bind} only differ by $P^o \Delta V_{\text{Bind}}$ which is negligible at atmospheric pressure, they are often used interchangeably.

In their landmark work, Gilson et al. derived an expression similar to 1.26 beginning from the chemical potentials of the protein, ligand, and complex in solution (Equation 13).[49] Their expression includes symmetry numbers, but these can generally be neglected.[73] Woo and Roux also derived a similar expression (Equation 5) from a different perspective, relying heavily on potentials of mean force (PMFs).[51] A PMF is a free energy surface, $W(x)$, along some coordinate x, for example, the intermolecular distance between the ligand and receptor. Its absolute value is arbitrary, but differences in the PMF yield free energy differences between different values of the coordinate. For example, the relative probability of finding the system at the value of x_2, p_{x_2}, compared to at x_1, p_{x_1} is

$$\frac{p_{x_2}}{p_{x_1}} = e^{-\frac{W(x_2)-W(x_1)}{k_{\text{B}}T}}. \tag{1.27}$$

This work focuses on computational methods to evaluate Equation 1.26. Ideally, these predictions should be compared with experiments which more directly measure $\frac{p_{\text{Bound}}}{p_{\text{Free}}}$, such as ITC and SPR (Section 1.1.2). The dissociation constant, K^o_{D}, can be calculated from the plot of heat absorbed against molar ratio of receptor and ligand using the slope at the midpoint (for ITC), or the ratio of dissociation and association rates (for SPR).

$$K^o_{\text{D}} = \frac{[L][R]}{c^o[LR]} = -k_{\text{B}}T \ln \Delta G^o_{\text{Bind}}, \tag{1.28}$$

where $[L]$, $[R]$, and $[LR]$ are the concentrations of the ligand, receptor, and receptor-ligand complex measured under constant temperature and pressure, and concentrations approximate activities.[49]

Unfortunately, often only IC_{50} measurements (Section 1.1.2) are available as they are easier to obtain. These can be related to the inhibitory constant, K_{i}, using the Cheng-Prusoff equation,[74]

$$K_{\text{i}} = \frac{\text{IC}_{50}}{1 + \frac{[S]}{K_{\text{m}}}}, \tag{1.29}$$

where K_m is the Michaelis constant and $[S]$ is the concentration of the substrate with which the ligand is competing. If all binding events result in inhibition, K_D can be approximated as K_i. However, IC_{50}s are can be effected by a multitude of factors, such as difficulty reaching the drug target in a tissue culture. To reduce issues with direct conversion of IC_{50} values to free energies of binding, predictions of changes in IC_{50}s may be made.[20]

1.3 Describing the Potential Energy Landscape

Evaluating Equation 1.26 requires a potential energy function, $U(\boldsymbol{r'})$. To provide an accurate affinity estimate, this function should be accurate for the low-energy configurations which dominate the integrals at room temperature. Among many other considerations, this means that the interactions of the ligand with the polar solvent (water) and less polar protein should be accurately balanced, that the dominant conformations of the protein and ligand should have accurate relative energies, and that this should be the case for a wide variety of drug-like molecules.

Quantum mechanical methods provide the most accurate potential energies, but are so expensive that $U(\boldsymbol{r'})$ can only be evaluated for a few sets of $\boldsymbol{r'}$.[47] This introduces errors into integrals in Equation 1.26 which offset any improvements due to accurate energies. Instead, Equation 1.26 is usually evaluated using fixed-charge classical potentials (also known as molecular mechanics force fields), which are orders of magnitude faster to compute.[75]

1.3.1 Fixed-Charge Classical Potentials

Fixed-charge classical potentials intuitively decompose $U(\boldsymbol{r'})$ into components arising from bonds, angles, torsions, improper torsions, electrostatic interactions (Ele) and van der Waals (vdW) interactions

$$U(\boldsymbol{r'}) = U_\mathrm{Bond}(\boldsymbol{r'}) + U_\mathrm{Angle}(\boldsymbol{r'}) + U_\mathrm{Torsion}(\boldsymbol{r'}) + U_\mathrm{Improper}(\boldsymbol{r'}) + U_\mathrm{Ele}(\boldsymbol{r'}) + U_\mathrm{vdW}(\boldsymbol{r'}). \tag{1.30}$$

The bond and angle terms sum over all bonds (d) and angles (θ) between bonded atoms. For each bond or angle, i, the energy is given by a harmonic potential

$$U_\mathrm{Bond,i}(d_i) = K_{\mathrm{Bond},i}(d_i - d_{0,i})^2 \tag{1.31}$$

$$U_\mathrm{Angle,i}(\theta_i) = K_{\mathrm{Angle},i}(\theta_i - \theta_{0,i})^2, \tag{1.32}$$

where $_0$ denotes an equilibrium value. A harmonic potential is also often used for the improper dihedral terms, which model the out-of-plane distortion of planar groups and maintain the correct chirality of tetrahedral centres.[75]

The torsional functional form is

$$U_{\text{Torsion,i}}(\theta_i) = K_{\text{Torsion},i}[1 + \cos(m\theta_i - \gamma)], \tag{1.33}$$

where m is the multiplicity and γ is the phase angle (typically 0 or π). Typically, multiple torsional terms are applied per torsion.[75]

The electrostatic energies are calculated using the Coulomb potential

$$U_{\text{Ele,i,j}}(d_{ij}) = \frac{q_i q_j}{4\pi\epsilon_0 d_{ij}}, \tag{1.34}$$

where d_{ij} is the distance between atoms i and j, ϵ_0 is the permittivity of vacuum, and q_i and q_j are point charges assigned to each atom (or virtual site). Van der Waals interactions are usually described using the Lennard-Jones potential

$$U_{\text{vdW,i,j}}(d_{ij}) = 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{d_{ij}}\right)^{6}\right], \tag{1.35}$$

where ϵ_{ij} is the Lennard-Jones well depth and σ_{ij} is the collision diameter (where $U_{\text{vdW}} = 0$).

The non-bonded terms ($U_{\text{Ele}}(\boldsymbol{r'})$ and $U_{\text{vdW}}(\boldsymbol{r'})$) are summed over all pairs of atoms in different molecules or separated by at least 4 bonds (non-bonded interactions separated by 4 bonds are typically scaled down).[75] As there are many pairs, these are the most expensive terms to evaluate. To reduce this expense, Lennard-Jones interactions are usually ignored when d_{ij} exceeds a cut-off. While methods to correct for the missing interactions are available,[76] the corrections are often negligible for binding free energy calculations.[77] More importantly, the truncation of the potential must be smoothed to avoid discontinuities in the energies and forces, and the choice of smoothing method may affect results.[78]

Unlike the attractive Lennard-Jones interactions, electrostatic interactions decrease more slowly than the volume of a sphere increases ($\frac{1}{d_{ij}^3}$) and neglecting interactions beyond the cut-off introduces substantial errors.[75] Instead, the reaction field (RF) method assumes a homogeneous dielectric continuum beyond

the cut-off.[79] An alternative is lattice summation methods such as particle mesh Ewald, where the Coulomb interactions are split into a short-range sum which is calculated in real space, and a sum of smooth long-range terms which are evaluated in reciprocal space using Fourier transforms.[80,81]

Given the functional form shown above, a key question is how to obtain and assign parameters (force constants, equilibrium values, etc.). For a general force field, these parameters should work well for a broad range of molecules in a particular category (e.g. proteins or drug-like small molecules). The parameters are typically obtained by fitting to a combination of quantum-mechanical data (e.g. optimised geometries, and torsional potentials) and experimental data (e.g. densities and heats of vaporisation).[82,83] Traditionally, parameters are assigned using "atom types", which are themselves assigned based on the environment of each atom in a molecule. This results in many redundant parameters and makes force fields difficult to extend. "Direct chemical perception" avoids these issues by assigning parameters directly based on the unmodified chemical graph of the molecule.[84] Both methods assign all parameters other than small molecule partial charges. These are often obtained with the AM1-BCC method, which corrects the charges produced by a fast semi-empirical quantum method (Austin Model 1) with bond-charge corrections to reproduce the HF/6-31G* electrostatic potentials.[85,86] A promising recent approach is training graph neural networks to assign parameters.[87]

The molecular mechanics energy functions presented above represent many severe approximations. For example, the dynamic electron distributions in real molecules are reduced to point charges of fixed magnitude. This leads to inaccuracies when atoms are close and the electron distributions should overlap, and when changes in the environment should change the electron distribution. Force fields can be modified to include the effects of charge-penetration and explicit polarisation, but it is often not clear whether improvements in accuracy will be observed for a specific application and whether they will be worth the increased cost.[88,89] In general, improving molecular mechanics force fields is challenging because they are not systematically improvable and it is often unclear which parameters and which aspects of functional form are responsible for inaccuracies.

1.3.2 Machine Learning Potentials

Machine learning offers a framework for training highly flexible functions which are extremely promising as atomistic potentials. Appealingly, they are systematically improvable. Recent machine learning potentials show an accuracy comparable to the quantum chemical methods they were trained on at orders of magnitude lower cost, and transferable potentials are now available for small organic molecules.[90–93] However, they are still prohibitively expensive compared to molecular mechanics force fields, and are yet to yield substantial improvements in binding or hydration free energy calculations in affordable protocols.[94–96] Hybrid classical/machine learning potentials which better balance accuracy and speed are promising.[97]

1.4 Sampling the Potential Energy Landscape

Selecting a suitable force field does not allow the immediate evaluation of Equation 1.26 because the space of all possible $\boldsymbol{r'}$ is vast. This makes uniform sampling of the configurational space highly inefficient, as very few configurations have low enough energy to contribute significantly to the integrals.

Two complementary approaches to evaluating these integrals are reducing the dimensionality of $\boldsymbol{r'}$, and focusing sampling on samples with non-negligible Boltzmann factors. The dimensionality of $\boldsymbol{r'}$ can be minimised by solvating the solutes in the smallest possible box. Periodic boundary conditions are used to avoid artefacts from vacuum interfaces.[98] The box must be sufficiently large to avoid interactions of the solutes with periodic copies of themselves, and increasingly spherical box shapes reduce the required number of solvent molecules needed for the same minimum solute-periodic copy distance.

Two common methods for sampling configurations by their Boltzmann weights are Markov Chain Monte Carlo (MCMC) and molecular dynamics (MD).[99] Free energy calculations usually use MD for most sampling.

1.4.1 Molecular Dynamics

The definition of momentum and Newton's second law are

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{r} = \boldsymbol{M^{-1}p}, \tag{1.36}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{p} = -\nabla_{\boldsymbol{r}} U(\boldsymbol{r}), \tag{1.37}$$

respectively, where $\boldsymbol{M}$ is a diagonal matrix of the atom masses and t is the time. Beginning from an initial point in phase space, the forces (potential energy gradients) on each atom are used to update their positions and momenta as a function of time, generating a trajectory.

Unfortunately, these equations cannot generally be solved analytically for systems of interest. Instead, finite-difference methods are used, where positions and momenta are updated at small time steps. These methods aim to satisfy time-reversibility, closely preserve the true trajectory, satisfy the conservation of energy and momentum, and allow the use of a large time step. They should be symplectic, meaning that they exactly conserve a "shadow Hamiltonian" which is close to the true Hamiltonian and becomes equal as the time step tends to 0.[99] A popular choice is the velocity Verlet algorithm.[100]

Because applying Newton's equations conserves the total energy of the system, samples from the NVE ensemble are obtained. To enforce constant temperature rather than constant energy, thermostat algorithms are applied to simulations. For example, the Andersen thermostat randomly updates velocities to match those from the Maxwell-Boltzmann distribution. This mimics "collisions" with a heat bath.[101] Another approach is to use Langevin dynamics, where a frictional term ($\gamma\boldsymbol{p}$) and a random force ($\boldsymbol{\eta}(t)$) are added to Equation 1.37

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{p} = -\nabla_{\boldsymbol{r}} U(\boldsymbol{r}) - \gamma\boldsymbol{p} + \boldsymbol{\eta}(t), \tag{1.38}$$

where γ and $\boldsymbol{\eta}(t)$ are related by the fluctuation-dissipation theorem.[102]

Sampling from the NPT ensemble also requires a barostat to enforce constant pressure. A simple approach is to update the volume using Metropolis Monte Carlo moves.[41]

The computational cost of a fixed-length MD simulation is inversely proportional to the time step. The fastest frequency motions, which are bonds to hydrogen, determine the maximum stable time step. Commonly, bonds involving hydrogen are made rigid using constraint algorithms, which allows the use of longer time steps without losing the conservation of energy and creating instabilities.[103] Since Equation 1.26 is independent of masses, mass from carbons can be moved to bonded hydrogens without affecting free energy predictions, slowing down bond vibrations. This hydrogen mass repartitioning allows constraint algorithms to work at larger time steps and has enabled the routine use of 4 fs time steps in free energy calculations.[104]

1.4.2 Enhanced Sampling

Important configurations are often separated by large energy barriers which MD trajectories cross slowly. A vast number of algorithms are available to enhance configurational sampling.[105] An algorithm frequently used in free energy calculations is Hamiltonian replica exchange (HREX) and particularly replica exchange with solute scaling (REST2). HREX involves swapping coordinates between parallel simulations of the same system with modified Hamiltonians, and is useful when sampling of a slow degree of freedom is accelerated with some of the Hamiltonians.[106] REST2 is an example of HREX which introduces Hamiltonians with reduced solvent-solute interactions, mimicking the effect of increased temperature and generating varied conformations which are reintroduced to simulations with Hamiltonians of interest.[107]

1.4.3 Equilibration, Convergence, and Uncertainty

The assumption of ergodicity states that ensemble averages are equal to infinite sampling time averages.[99] However, simulations cannot be run for an infinite time, and expected errors compared to the infinite time results should be quantified.

Often, simulations are started from initial configurations which have vanishingly small probabilities under the Boltzmann distribution. This can produce an initial bias in calculated quantities, which decreases with sampling time as the system evolves towards more probable regions of phase space. This is often called equilibration, and initial unequilibrated samples are usually discarded to reduce bias. When there are many simulations, as in an alchemical free energy calculation, it

is desirable to select the truncation point in an automated manner rather than manually inspecting all outputs. However, the two automated truncation point selection methods popular in the molecular simulation community have not been thoroughly assessed.[108,109]

Convergence describes the approach of a quantity to its infinite sampling time value with increasing sampling. Errors reported for molecular simulations are normally estimates of convergence, in that they predict the expected difference to the infinite sampling time value. MD trajectories slowly explore configurational space, meaning that sampled configurations are highly correlated. This increases the expected difference to the infinite sampling value compared to independently sampled configurations. As a result, the autocorrelation function for the quantity of interest must be calculated when estimating uncertainties from a single molecular dynamics run.[110] However, the accurate estimation of autocorrelation is notoriously difficult. Additionally, single runs often become stuck in local energy minima, meaning that uncertainties are severely underestimated.

A better alternative is to estimate sampling uncertainty from the means of (equilibrated) repeat runs. These tend to explore a larger area of conformational space.[111] The most actionable estimate of uncertainty is a confidence interval - a bound which is expected to cover the infinite time value with a specified probability. Confidence intervals can be obtained from the means of replicate runs by relying on the central limit theorem and assuming that the run means are normally distributed

$$\mathrm{CI}_{100(1-\alpha)} = \bar{x} \pm t\sqrt{\frac{\hat{\sigma}^2}{n}}, \tag{1.39}$$

where $\bar{x}$ is the estimated mean over replicate runs, $\hat{\sigma}$ is the estimated standard deviation, n is the number of replicate runs, and t is the critical value from the Student's t-distribution corresponding to a two-tailed confidence interval at a given confidence for $n-1$ degrees of freedom. However, this is only valid when each run contains many independent samples, in which case there would be no advantage over using a single run.

Bootstrapping is an alternative approach which assumes that the sample distribution is representative of the underlying population distribution.[112] Instead of rerunning the entire set of runs many times to find the true uncertainty, this is simulated by resampling from the initial set of runs. Uncertainties for any metric can then trivially be calculated based on the sets of resampled runs.

For example, the two-sided 95% CIs for the mean would be given by the lower 2.5th and upper 97.5th percentiles for the means over the sets of resampled runs. However, the assumption of representative sample distributions is invalid unless many independent simulations have been performed. Using t-based confidence intervals based on multiple replicate runs appears to be the least bad option.

Even if the sampling error is quantified precisely, many other sources of error can contribute to differences between experimental and calculated quantities.[113] Firstly, experimental errors may be substantial. Secondly, the simulated system may differ significantly from the experimental system, for example in the ionisation state of protein side-chains. Thirdly, force fields may contribute large errors to the energies. Fourthly, the simulation and experiment may measure different things - there may be multiple binding sites measured in an affinity experiment, while a simulation may only treat one. While these are often the largest sources of error, there are many more which are usually presumed to be minor, such as constraining bond lengths.

1.5 Rigorous Free Energy Calculations

"Rigorous" binding free energy calculations evaluate ratios of partition functions such as those in Equation 1.26. Given a suitable force field and sampling strategy, it may be tempting to evaluate $\frac{p_{\text{Bound}}}{p_{\text{Free}}}$ by running a long simulation of a protein and ligand and counting the relative number of bound and free configurations obtained. Comparing Equations 1.22, 1.24 and 1.26 shows that

$$\Delta F^{o}_{\text{Bind}} = -k_{\text{B}}T \ln \frac{V_{\text{PL,Free}} p_{\text{Bound}}}{V^{o} p_{\text{Free}}}. \tag{1.40}$$

However, this is not often useful for computing $\Delta F^{o}_{\text{Bind}}$ because the binding, and especially unbinding processes usually take much longer than MD simulations can be run for, producing poor estimates of $\frac{p_{\text{Bound}}}{p_{\text{Free}}}$.

Instead, these calculations exploit the path independence of the free energy to construct alternative pathways between the free and bound states along which converged sampling is more achievable. Usually, sampling is performed at discrete points along these pathways.[114] Therefore, we require two final ingredients to evaluate the free energy of binding: a method to evaluate free energy differences between discrete states and a choice of pathway.

1.5.1 Calculating Free Energy Changes Between States

We will assume that the potential energy function of the system depends on a parameter called λ, which is 0 and 1 at initial and final states of interest, respectively. Simulations at a single value of λ are sometimes called λ windows.

Samples from a Single State

The free energy change between states i and j can be written

$$\Delta F_{ij} = -k_{\mathrm{B}}T \ln \frac{Q_j}{Q_i} \tag{1.41}$$

$$= -k_{\mathrm{B}}T \ln \frac{\int e^{-\frac{U(\boldsymbol{r},\lambda_j)}{k_{\mathrm{B}}T}} \mathrm{d}\boldsymbol{r}}{Q_i} \tag{1.42}$$

$$= -k_{\mathrm{B}}T \ln \frac{\int e^{-\frac{U(\boldsymbol{r},\lambda_j)+U(\boldsymbol{r},\lambda_i)-U(\boldsymbol{r},\lambda_i)}{k_{\mathrm{B}}T}} \mathrm{d}\boldsymbol{r}}{Q_i} \tag{1.43}$$

$$= -k_{\mathrm{B}}T \ln \frac{\int e^{-\frac{\Delta U_{ij}(\boldsymbol{r})}{k_{\mathrm{B}}T}} e^{-\frac{U(\boldsymbol{r},\lambda_i)}{k_{\mathrm{B}}T}} \mathrm{d}\boldsymbol{r}}{Q_i} \tag{1.44}$$

$$= -k_{\mathrm{B}}T \ln \left\langle e^{-\frac{\Delta U_{ij}(\boldsymbol{r})}{k_{\mathrm{B}}T}} \right\rangle_i, \tag{1.45}$$

where $\Delta U_{ij}(\boldsymbol{r}) = U(\boldsymbol{r},\lambda_j) - U(\boldsymbol{r},\lambda_i)$ and $\langle ... \rangle_i$ indicates an average over $\boldsymbol{r}$ from state i. This is the Zwanzig equation.[115]

As the states become infinitely close in λ, the energy differences tend to 0 and the logarithm and exponential in Equation 1.45 can be replaced with their first-order Taylor expansions

$$\mathrm{d}F = \left\langle \frac{\partial U(\boldsymbol{r},\lambda)}{\partial \lambda} \right\rangle_\lambda \mathrm{d}\lambda. \tag{1.46}$$

Integrating this over λ to obtain free energy changes is called thermodynamic integration (TI).[116] When all samples are from state i, the TI estimate of ΔF_{ij} is

$$\Delta F_{ij} = \left\langle \frac{\partial U(\boldsymbol{r},\lambda)}{\partial \lambda} \right\rangle_i \Delta\lambda_{ij}, \tag{1.47}$$

which is an approximation of the Zwanzig equation. The Zwanzig equation is exact in the limit of infinite sampling of state i, but the TI estimate additionally requires that the states are infinitely close in λ. Hence, the Zwanzig equation is preferred over TI when samples are from a single state.

The Zwanzig equation provides reliable free energy estimates when the thermally accessible configurations of state j are a subset of those for state i, but this is not normally the case. When the energy differences are large, the exponential average in Equation 1.45 becomes dominated by few samples and the estimate is biased and noisy.

Samples from Two States

Substantially more reliable estimates are usually obtained by sampling both states i and j. The TI free energy change estimate is obtained from the average average gradients, which corresponds to the trapezoid rule

$$\Delta F_{ij} = \left(\frac{1}{2} \left\langle \frac{\partial U(\boldsymbol{r}, \lambda)}{\partial \lambda} \right\rangle_i + \frac{1}{2} \left\langle \frac{\partial U(\boldsymbol{r}, \lambda)}{\partial \lambda} \right\rangle_j \right) \Delta \lambda_{ij}. \tag{1.48}$$

It may be tempting to apply the Zwanzig estimator similarly, averaging the estimates from the forwards and reverse directions

$$\Delta F_{ij} = -\frac{1}{2} k_\mathrm{B} T \ln \left\langle e^{-\frac{\Delta U_{ij}(\boldsymbol{r})}{k_\mathrm{B} T}} \right\rangle_i + \frac{1}{2} k_\mathrm{B} T \ln \left\langle e^{-\frac{\Delta U(\boldsymbol{r})_{ji}}{k_\mathrm{B} T}} \right\rangle_j. \tag{1.49}$$

This may somewhat cancel errors from the forward and reverse perturbations, but usually the systematic errors are asymmetric and do not completely cancel.[117,118] A better approach is to combine the samples from both states to create a mixture distribution, forgetting which state they came from (i.e. discarding information about which state they were sampled from).[119,120] The free energy differences from this virtual mixture state to states i and j can then be estimated using the Zwanzig equation

$$\Delta F_{\mathrm{Mix},j} = -k_\mathrm{B} T \ln \left\langle e^{-\frac{U(\boldsymbol{r},\lambda_j) - U_\mathrm{Mix}(\boldsymbol{r})}{k_\mathrm{B} T}} \right\rangle_\mathrm{Mix} \tag{1.50}$$

$$= -k_\mathrm{B} T \ln \left\langle e^{-\frac{\Delta U_{\mathrm{Mix},j}(\boldsymbol{r})}{k_\mathrm{B} T}} \right\rangle_\mathrm{Mix}, \tag{1.51}$$

and

$$\Delta F_{\mathrm{Mix},i} = -k_\mathrm{B} T \ln \left\langle e^{-\frac{\Delta U_{\mathrm{Mix},i}(\boldsymbol{r})}{k_\mathrm{B} T}} \right\rangle_\mathrm{Mix}, \tag{1.52}$$

where $U_{\mathrm{Mix}}(\boldsymbol{r})$ is the effective energy function which would produce the mixture distribution were it sampled directly

$$U_{\mathrm{Mix}}(\boldsymbol{r}) = -k_{\mathrm{B}}T \ln \left(\frac{N_i}{N_{\mathrm{Tot}}} e^{-\frac{U(\boldsymbol{r},\lambda_i)-\Delta F_{\mathrm{Mix},i}}{k_{\mathrm{B}}T}} + \frac{N_j}{N_{\mathrm{Tot}}} e^{-\frac{U(\boldsymbol{r},\lambda_j)-\Delta F_{\mathrm{Mix},j}}{k_{\mathrm{B}}T}} \right). \tag{1.53}$$

$N_{\mathrm{Tot}} = N_i + N_j$ is the total number of samples from states i and j. Importantly, the free energy differences to be estimated are required to construct $U_{\mathrm{Mix}}(\boldsymbol{r})$. Hence Equation 1.51 must be solved iteratively for $\Delta F_{\mathrm{Mix},i}$ and $\Delta F_{\mathrm{Mix},j}$, or equivalently, ΔF_{ij}. Shifting $U_{\mathrm{Mix}}(\boldsymbol{r})$ by an arbitrary offset does not change its likelihood given the mixture distribution, so F_{Mix} is arbitrary and only ΔF_{ij} is meaningful.

Bennett first derived this method in 1976 by minimising the error in ΔF_{ij} in the large-sample limit, and it is often referred to as the Bennett acceptance ratio (BAR).[121] Later, BAR (and TI) were rigorously examined by the statistics community as general methods for estimating ratios of normalising constants by MCMC,[122,123] and they are now applied to statistical inference in diverse fields including psychology.[124] BAR can also be derived by finding the maximum likelihood estimator of the free energy change using (reverse) logistic regression.[119,125] This perspective is equivalent to optimising $U_{\mathrm{Mix}}(\boldsymbol{r})$ to match the obtained mixture distribution.

Theoretically, when two states are sampled, BAR provides the lowest variance free energy estimate of any asymptotically unbiased estimator in the large-sample regime.[125] Practically, BAR is superior to the Zwanzig equation and TI.[126,127] Unfortunately, BAR becomes unreliable when the potential energy landscapes of i and j are too dissimilar because the energy differences to the mixture state become large and the exponential average in Equation 1.51 is dominated by few configurations. The phase-space overlap quantifies this similarity.[128] To increase overlap and obtain reliable free energy estimates, many intermediate λ windows are usually required.

Samples from Many Intermediate States

When there are many intermediate states, the efficiency of TI can be improved using more sophisticated numerical quadrature methods. However, summing the free energy differences obtained with BAR is still preferable.[129]

When there is overlap between non-adjacent states, BAR can be improved by constructing the mixture distribution with all samples from all states, in which case

$$U_{\mathrm{Mix}}(\boldsymbol{r}) = -k_{\mathrm{B}}T \ln \sum_k \frac{N_k}{N_{\mathrm{Tot}}} e^{-\frac{U(\boldsymbol{r},\lambda_k)-\Delta F_{\mathrm{Mix},k}}{k_{\mathrm{B}}T}}. \tag{1.54}$$

This is a generalisation of BAR which is the lowest variance asymptotically unbiased estimator when multiple states are sampled. It is equivalent to BAR when there is no overlap between non-adjacent states. This approach was developed in the statistics community and popularised in the molecular simulation community as the Multistate Bennett Acceptance Ratio (MBAR).[119,130,131] A clear and simple derivation with notation familiar to chemists is given by Ding.[132]

MBAR is only the minimum variance estimator when the samples from each state are equally autocorrelated. However, this is no justification for discarding samples to reduce correlation, because it remains a robust estimator as long as the autocorrelations are not drastically unequal.[122] Similarly, good uncertainty estimates can be obtained when samples are correlated,[133,134] and samples should not be discarded for this purpose.

An alternative to using many intermediate states is to run many short non-equilibrium switching simulations. Initial configurations are generated by extracting samples from equilibrium simulations of the end states. λ is then scaled to the value at the opposite end-state. If these simulations were infinitely long, then the work done in "pushing" λ would be the reversible, equilibrium work, equal to the free energy change; the averages in all free energy change estimators would be replaced with a single value and the estimates would simplify to the work done. However, short simulations are run, which require larger, non-equilibrium, work to switch λ. The Zwanzig, BAR, and MBAR estimators traditionally use ΔU estimates for single conformations, which can be treated as the work done in infinitely fast non-equilibrium switches. Hence, the ΔU values in the Zwanzig and BAR equations can be the work done during infinitely fast or infinitely slow switching. This makes it less surprising that non-equilibrium work values from finite time switching can also be used.[135,136]

1.5.2 Selecting States

The final ingredient required to estimate the free energy of binding is a choice of pathway.

Physical-Path-Based Binding Free Energy Calculations

The most obvious pathway is a physical one, where the ligand is gradually separated from the protein. The statistical mechanical framework was given by Woo and Roux.[51] Usually, a series of equilibrium MD simulations are performed with the ligand restrained at increasing distances from the protein. The initial state, where the ligand and protein are unrestrained, corresponds to the bound integral in Equation 1.26. However, since the relative protein-ligand degrees of freedom, $\boldsymbol{r_{PL}}$, are restrained rather than completely frozen with constraints in the final separated state, Equation 1.26 effectively becomes

$$\Delta F^{o}_{\text{Bind}} = -k_{\text{B}}T \ln \frac{\int_{\text{Bound}} e^{-\frac{U(\boldsymbol{r'})}{k_{\text{B}}T}} d\boldsymbol{r'} \int_{\text{Free,Restr}} e^{-\frac{U(\boldsymbol{r_{PL}})}{k_{\text{B}}T}} d\boldsymbol{r_{PL}}}{8\pi^2 V^{o} \int_{\text{Free,Restr}} e^{-\frac{U(\boldsymbol{r'})}{k_{\text{B}}T}} d\boldsymbol{r'}}, \tag{1.55}$$

where "Restr" indicates the presence of protein-ligand restraints and $\int_{\text{Free,Restr}} e^{-\frac{U(\boldsymbol{r_{PL}})}{k_{\text{B}}T}} d\boldsymbol{r_{PL}}$ corrects for the difference between the restrained free state integral in the denominator and the constrained free state integral in Equation 1.26. The restraints are chosen so the correction can be evaluated by analytical or numerical integration without simulation. Isolating the terms which require sampling yields

$$\Delta F^{o}_{\text{Bind}} = -k_{\text{B}}T \ln \frac{\int_{\text{Bound}} e^{-\frac{U(\boldsymbol{r'})}{k_{\text{B}}T}} d\boldsymbol{r'}}{\int_{\text{Free,Restr}} e^{-\frac{U(\boldsymbol{r'})}{k_{\text{B}}T}} d\boldsymbol{r'}} - \Delta F^{o}_{\text{Release}}, \tag{1.56}$$

where

$$\Delta F^{o}_{\text{Release}} = -k_{\text{B}}T \ln \frac{8\pi^2 V^{o}}{\int_{\text{Free,Restr}} e^{-\frac{U(\boldsymbol{r_{PL}})}{k_{\text{B}}T}} d\boldsymbol{r_{PL}}} \tag{1.57}$$

is the correction for releasing the restrained ligand to the standard state volume. Additional restraints on the ligand's conformation are sometimes applied before separation and released afterwards to reduce sampling issues.

The geometric route is the best choice for computing absolute binding free energies for two large binding partners, such as two proteins.[137] However, it is problematic when there is no unhindered route to extract a ligand from a buried binding site.

Alchemical Absolute Binding Free Energy Calculations

The alchemical route overcomes this limitation by using unphysical, "alchemical" intermediate states. Gilson laid out the theoretical basis of alchemical ABFE calculations,[49] and Boresch provided an easily implemented practical method.[50] Rather than physically extracting the ligand, these calculations entirely remove the intermolecular interactions of the ligand restrained in the binding site, then switch the interactions of the ligand back on in solvent.

In detail, the free and restrained state in Equation 1.56 is replaced with a decoupled and restrained state, where the non-interacting ligand is restrained in the binding site. This is effectively a free state because there are no correlations between the ligand and protein degrees of freedom, other than through the restrained relative external degrees of freedom. Equation 1.56 becomes

$$\Delta F^{o}_{\mathrm{Bind}} = -k_{\mathrm{B}}T \ln \frac{\int_{\mathrm{Bound}} e^{-\frac{U(\boldsymbol{r'})}{k_{\mathrm{B}}T}} d\boldsymbol{r'}}{\int_{\mathrm{Decoup,Restr}} e^{-\frac{U(\boldsymbol{r'})}{k_{\mathrm{B}}T}} d\boldsymbol{r'}} - k_{\mathrm{B}}T \ln \frac{\int_{\mathrm{Decoup,Free}} e^{-\frac{U(\boldsymbol{r'})}{k_{\mathrm{B}}T}} d\boldsymbol{r'}}{\int_{\mathrm{Free}} e^{-\frac{U(\boldsymbol{r'})}{k_{\mathrm{B}}T}} d\boldsymbol{r'}} - \Delta F^{o}_{\mathrm{Release}}, \tag{1.58}$$

where "Decoup" denotes decoupled in the sense of Gilson et al., meaning that the ligand has no intermolecular interactions (but may also have perturbed intramolecular interactions).[49] The second term is the free energy of hydration (plus contributions from any changes in the intramolecular interactions), which corrects for the lack of ligand-solvent interactions in the decoupled, restrained state. The contributions from the protein degrees of freedom cancel between the integrals in this term, so only the ligand in water needs to be sampled. The sequence of transformations used to compute the first and second terms are called the bound and free legs, respectively.

Often, the restraints, charges, and LJ terms are introduced or removed in separate stages of each leg. In this case, $\Delta F^{o}_{\mathrm{Bind}}$ can be expressed as a sum of contributions from each stage

$$\Delta F^{o}_{\mathrm{Bind}} = -\Delta F_{\mathrm{Bound,Restrain}} - \Delta F_{\mathrm{Bound,Discharge}} - \Delta F_{\mathrm{Bound,Vanish}} - \Delta F^{o}_{\mathrm{Release}} + \Delta F_{\mathrm{Free,Discharge}} + \Delta F_{\mathrm{Free,Vanish}} + \Delta F_{\mathrm{SymCorr}}, \tag{1.59}$$

where "Bound" and "Free" denote the bound and free legs, "Restrain" denotes the introduction of the restraints (beginning from the unrestrained ligand in the binding site), "Discharge" indicates removal of ligand partial charges, "Vanish" denotes removal of ligand LJ terms, and $\Delta F_{\text{SymCorr}}$ is a correction for symmetrical regions of the potential energy landscape which are not sampled as the restraints are introduced, but which are subject to a very high energy penalty from the restraint.[73] An alchemical ABFE cycle is shown in Figure 1.1.

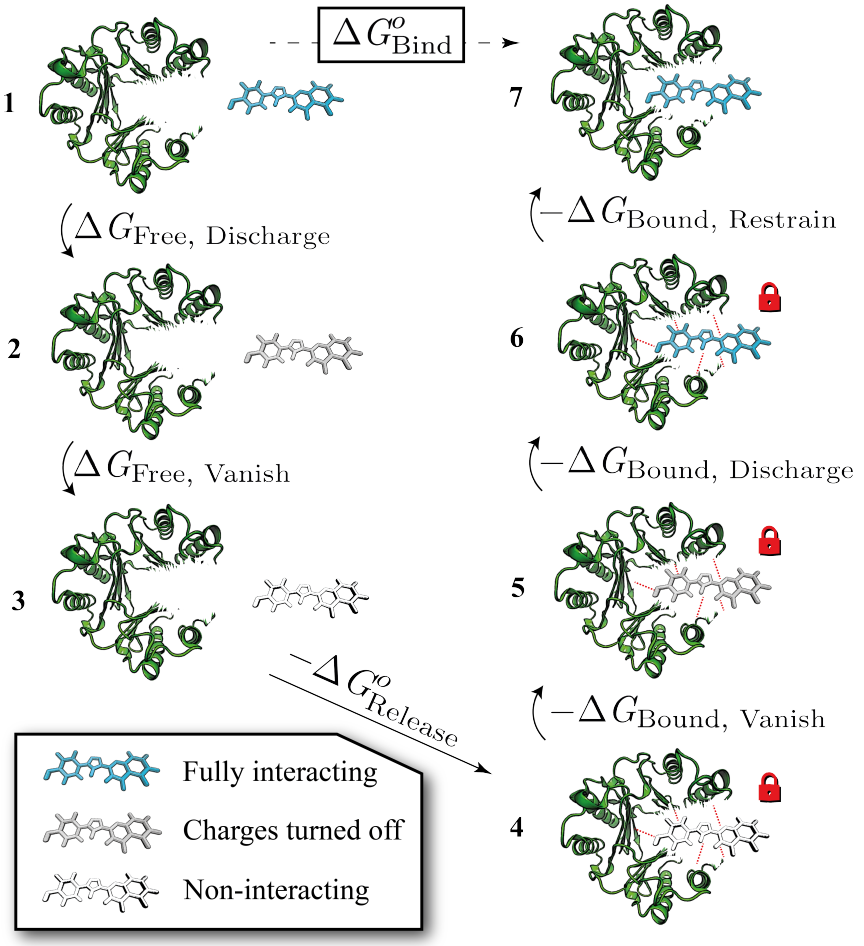


Figure 1.1: An alchemical thermodynamic cycle for the calculation of ABFEs, which proceeds from the unbound state (1) to the bound state (7) through a series of unphysical "alchemical" intermediates. The red dashed lines indicate protein-ligand restraints and the locks indicate that receptor-ligand restraints are active.

The choice of intermediate states for each stage is critical. In particular, the LJ energy is infinite at $r = 0$. If LJ interactions are removed by linearly scaling with $1 - \lambda$, this results in the "end-point catastrophe", where very large energy differences are observed between the decoupled end state and nearby λ states. Instead, it is necessary to use a soft-core potential which smoothly removes these interactions as λ is increased.[138]

In addition, the choice of protein-ligand restraints are critical to the success of an ABFE calculation. The most common restraints are those proposed by Boresch,[50] which define coordinate systems on the ligand and protein using three anchor points (often atomic positions) in each. The six relative external degrees of freedom, $\boldsymbol{r_{PL}}$, are then restrained. However, these can create instabilities and may not provide optimal convergence.

Alchemical relative binding free energy (RBFE) calculations are closely related to alchemical ABFE calculations, but the ligand is transformed into a structurally similar ligand rather than a non-interacting version of itself. They yield $\Delta\Delta F$ estimates and do not require restraints. The smaller perturbation produces fewer sampling challenges and makes RBFE calculations easier to converge, but they are limited to similar ligands bound to the same target with the same binding pose, at least within the single or hybrid topology frameworks.[114]

1.6 Outstanding Challenges and Outline of the Thesis

In summary, the ability to rapidly and accurately predict binding affinities would substantially increase efficiency in early-state drug development. "Rigorous" affinity prediction methods based on statistical mechanics use molecular dynamics or Monte Carlo methods to sample approximate energy functions and are more accurate than alternatives. Alchemical methods avoid the computationally intractable direct simulation of ligand binding and unbinding by using unphysical, "alchemical" intermediate states which provide an more efficient route to calculating free energy changes. Alchemical RBFE calculations have become a routine tool in the pharmaceutical industry, but they are limited to similar ligands bound to the same target with the same binding pose. This excludes important problems, such as ranking the affinities of structurally diverse molecules, ranking different binding poses, optimising ligand selectivity or promiscuity, and predicting the functional response of ligand binding.

These problems can be tackled by alchemical ABFE calculations. In particular, their applicability to structurally diverse compounds makes them attractive as a final, accurate filter in virtual screening. However, their use is limited by their higher cost and reduced accuracy compared to relative calculations. In addition, many protocol decisions often have to be made manually, and robust automation is required to facilitate high-throughput applications. This thesis investigates methods to improve the accuracy, efficiency, and automation of alchemical ABFE calculations.

In particular, the choice of receptor-ligand restraints is critical to the success of alchemical ABFE calculations. Boresch restraints are easily implemented, but can produce instabilities and are limited to restraining only the relative external degrees of freedom of the receptor and ligand. This may not provide optimal convergence. In Chapter 2, different restraint schemes are compared and automated methods for selecting restraint parameters are suggested. A restraints scheme based on multiple distance restraints between anchor points in the receptor and ligand is proposed, which avoids the inherent instabilities of Boresch restraints and may provide convergence benefits by more strongly restricting the relative movements of the receptor and ligand.

Building on an automated restraint selection scheme from Chapter 2, Chapter 3 presents a fully-automated open-source workflow to facilitate the high-throughput application of ABFE calculations. It includes algorithms to automate the selection of λ windows, the allocation of simulation time, and the truncation of unequilibrated data. Chapter 4 presents further automated truncation point selection algorithms, and quantitatively compares them using large amounts of synthetic data.

Finally, virtual screening requires diverse ligands to be ranked by their binding affinities. A quantitative understanding of the trade-off between simulation time and ranking performance is important to conduct efficient virtual screening campaigns. Chapter 3 investigates the effect of changing simulation time on the ranking performance of ABFE, using the automated workflow developed in Chapters 2 and 3. The ranking performance of very short ABFE calculations, which are especially appealing for virtual screening, is investigated.

Chapter 2

Comparison of Receptor–Ligand Restraint Schemes

Alchemical absolute binding free energy calculations require restraints between the receptor and ligand to restrict their relative positions and, optionally, orientations. The restraints proposed by Boresch are commonly used, but they must be carefully selected in order to sufficiently restrain the ligand and to avoid inherent instabilities. Applying multiple distance restraints between anchor points in the receptor and ligand provides an alternative framework without inherent instabilities which may provide convergence benefits by more strongly restricting the relative movements of the receptor and ligand. However, there is no simple method to calculate the free energy of releasing these restraints due to the coupling of the internal and external degrees of freedom of the receptor and ligand. Here, a method to rigorously calculate free energies of binding with multiple distance restraints by imposing intramolecular restraints on the anchor points is proposed. Absolute binding free energies for the human macrophage migration inhibitory factor/MIF180, system obtained using a variety of Boresch restraints and rigorous and nonrigorous implementations of multiple distance restraints are compared. It is shown that several multiple distance restraint schemes produce estimates in good agreement with Boresch restraints. In contrast, calculations without orientational restraints produce erroneously favorable free energies of binding by up to approximately 4 kcal mol$^{\check{}}$1.

These approaches offer new options for the deployment of alchemical absolute binding free energy calculations, which may enhance stability and accelerate convergence in some cases. In addition, this work illustrates that orientational restraints can dramatically improve the convergence of ABFE calculations, and provides a robust algorithm to derive restraint parameters for Boresch restraints, reducing the requirement for user input and facilitating large-scale deployment. This remainder of this chapter is included unmodified as published:

Clark, F.; Robb, G.; Cole, D. J.; Michel, J. Comparison of Receptor–Ligand Restraint Schemes for Alchemical Absolute Binding Free Energy Calculations. *J. Chem. Theory Comput.* **2023**, *19*, 3686–3704.

2.1 Introduction

The *in silico* prediction of protein-ligand binding affinities is an important problem in drug discovery. The ability to rapidly and accurately calculate affinities for arbitrary protein-ligand systems would allow the efficient prioritisation of compounds for synthesis and testing, accelerating the hit-to-lead and lead optimisation stages of drug discovery.[54]

Recent improvements in computing power and automation have brought this vision closer to realisation.[139–143] In particular, alchemical methods are ideally suited for application during the hit-to-lead and lead optimisation stages of drug discovery,[54,144] as well as in the later stages of virtual screening.[58] Along with path-based methods,[137] alchemical simulations form a class of exact (in the limit of complete sampling and a perfect description of the potential energy) methods based on molecular dynamics or Monte Carlo sampling which provide binding affinity predictions of greater accuracy than alternatives.[145] Modern computing resources now support routine use of alchemical relative binding free energy (RBFE) calculations to add value during drug discovery campaigns.[54].

RBFE calculations avoid the computationally intractable challenge of converging unbiased simulations of ligand binding and unbinding by instead gradually interconverting two structurally similar ligands. Interconversion proceeds through unphysical "alchemical" intermediates and is done in both the bound and unbound states.[114]. Based on a thermodynamic cycle, the free energy differences for each step are summed to yield the difference in the free energy of binding between

the ligands. RBFE calculations are used routinely and protocols for their robust deployment have been researched in detail.[146–149] However, as a result of the requirement for a common ligand core, binding pose, and binding site, the following valuable problems typically lie outside the scope of RBFE calculations:[58,150]

1. Calculating the RBFEs of structurally dissimilar ligands to a common target
2. Calculating RBFEs of the same ligand to the same protein with different binding poses
3. Calculating the RBFEs of the same ligand to different targets
4. Calculating the absolute binding free energy of a given ligand to a given target

Alchemical absolute binding free energy (ABFE) calculations escape these limitations by following a more general thermodynamic cycle in which the ligand's intermolecular interactions are completely turned off.[58] In principle, these calculations can be used to calculate the binding free energies of structurally diverse molecules to varied targets, making them attractive for drug discovery. However, the alchemical ABFE framework presents challenges not encountered during an RBFE calculation. Restraints must be applied between the protein and ligand to avoid convergence issues such as those associated with the ligand "wandering" out of the binding site as its intermolecular interactions are removed.[50] Restraints may also be required to avoid errors in the calculated binding free energies when the bound state is implicitly defined to include configurations where the ligand is anywhere in the entire simulation box relative to the receptor, as is the case when restraints are not used. However, these errors only affect weak binders (Section A.1).

In general, it is non-trivial to select the optimum receptor-ligand restraints. Furthermore, ABFE calculations can be challenging to converge and therefore computationally costly because the ligand is completely removed.[61,151] As a result, application studies still combine RBFE and ABFE, with ABFE applied more successfully to low molecular-weight compounds.[152] Thus, there are barriers to the routine application of ABFE calculations.

The performance and accessibility of ABFE calculations would be improved if it was trivial to select receptor-ligand restraints which resulted in stable simulations and produced optimal convergence. While progress has been made in this direction with tools for automated or partially-automated restraint selection,[61,140–142,153] there is still no restraint type or selection method which completely solves this issue.

Receptor-ligand restraints of a variety of forms have been proposed. Following early work utilising restraints on a single ligand atom[154,155], the first theoretically rigorous approach involving restraints on all of the external degrees of freedom (DoF) of the ligand was the Body Restraint Algorithm of Hermans and Wang.[156] Later, the Virtual Bond Algorithm (VBA) of Boresch et al. was introduced,[50] which involves restraining one distance, two bond angles, and three dihedral angles between six anchor points defined by the receptor and ligand. This provided a more convenient method to restrain the relative external DoF of the receptor and ligand, along with a simple analytical correction for releasing the restraints. The VBA has found widespread use and is often referred to as "Boresch restraints".

However, despite their popularity, Boresch restraints suffer from a number of limitations and must be carefully applied to avoid numerical instabilities and sampling issues.[157] For instance, if the anchor points are tied to the positions of highly flexible portions of the ligand or protein which do not strongly interact, then the restraints will be unable to maintain a binding pose similar to the restrained and interacting system, potentially leading to slow convergence of free energy estimates. Since only six relative external degrees of freedom can be restrained within this framework there are limits on the extent to which ligand motions can be restricted. Thus, additional restraints on the intramolecular degrees of freedom of the ligand may be required to improve convergence for flexible ligands.[158] Furthermore, if the restraints are poorly chosen, small changes in the Cartesian coordinates of the anchor points can result in large jumps in the six DoF defined in the VBA framework, resulting in the application of large forces, which can cause simulations to crash. This frequently occurs when sets of three contiguous anchor points approach collinearity,[140] which can result in the application of large forces through the dihedral restraints.

Alternative restraint schemes have recently been proposed to address these issues; Fu et al. proposed a method in which the restrained six external DoF are derived by finding the optimal rotation of the ligand which minimises its root-mean-square deviation (RMSD) with respect to a protein-ligand complex reference structure (after correcting for rotation and translation of the protein).[159] By moving away from the six anchor points of the VBA, this was intended to simplify the selection of stable and efficient restraints. The "distance to bound configuration" (DBC) restraint is also intended to simplify restraint selection and to minimise the variance of free energy estimates for removing the ligand's intermolecular interactions.[160–162] This is achieved by directly restraining the RMSD of a subset of the ligand coordinates within the frame of reference of the binding site in

order to optimally restrict the accessible configurational volume as the ligand intermolecular interactions are removed. However, because this scheme couples the internal and external degrees of freedom of the protein and ligand, there is no simple way to calculate the free energy of releasing the non-interacting ligand to the standard state. This necessitates a final stage to release the DBC restraints to a single harmonic restraint, for which the free energy of release is simple to calculate.

Another alternative to Boresch restraints is to restrain the distance between multiple receptor-ligand atom pairs. These restraints offer several advantages; for example, they can be intuitively selected to match native receptor-ligand interactions such as hydrogen bonds, thus closely mimicking the interacting state. This may accelerate convergence by tightly restricting ligand motion while intermolecular interactions are removed. Furthermore, these restraints do not suffer from the numerical instabilities inherent to the Boresch restraints scheme. Indeed, multiple distance restraints were used in an early study of the binding of biotin to streptavidin.[163] However, the naïve application of multiple distance restraints is theoretically incorrect, because they introduce coupling between the internal and external degrees of freedom of the protein and ligand, preventing the rigorous calculation of the free energy of releasing the non-interacting ligand.[164] Despite this, a recent implementation has been described which relied on the assumption that the restraints were sufficiently weak that such coupling was negligible, and that the free energy of turning on the restraints was close to zero.[165] However, this was not verified, and the scheme has not been systematically compared to Boresch restraints.

To address this, this study compares the absolute binding free energies obtained using Boresch restraints and different implementations of multiple distance restraints for a single ligand (MIF-180) binding to a single protein (human macrophage migration inhibitory factor, or MIF). This was suggested as a good model system by Qian et al.[166] because MIF is a pharmaceutically relevant protein, but of moderate size (342 residues), and no major conformational changes occur in the protein upon ligand binding (Figure 2.1). In this work, the the standard binding free energy of MIF-180 was calculated using multiple sets of Boresch restraint parameters. The results were compared to those produced by non-rigorous implementations of multiple distance restraints similar to that of Mendoza-Martinez et al.[165], and two rigorous multiple distance restraint schemes: one inspired by Salari et al.[160], and one newly developed.

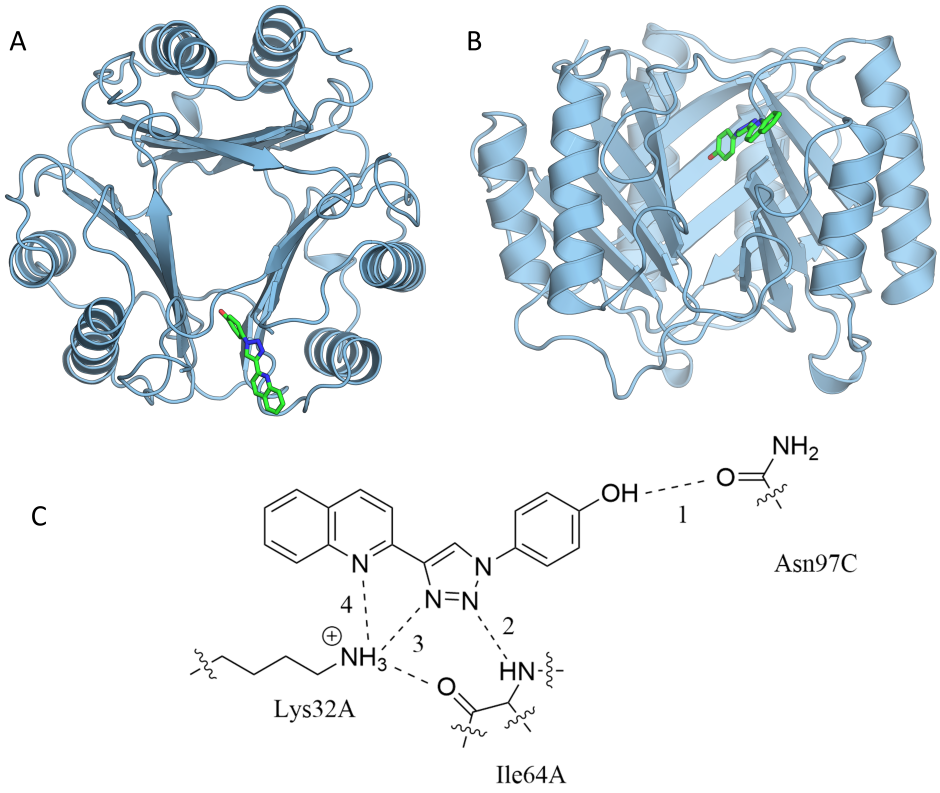


Figure 2.1: (A) and (B): MIF with MIF180 bound, rendered with PyMOL 2.5.[167] (C) Hydrogen bonding interactions between MIF and MIF180 in the tautomerase active site, redrawn from Qian et al..[166]

2.2 Theory

2.2.1 Alchemical Absolute Binding Free Energy Calculations

ABFEs can be computed using an alchemical cycle (Figure 2.2). The standard free energy of binding is calculated by adding up the terms around the cycle:

$$\begin{aligned}\Delta G^{o}_{\text{Bind}} =& \Delta G_{\text{Free, Discharge}} + \Delta G_{\text{Free, Vanish}} - \Delta G^{o}_{\text{Release}} - \Delta G_{\text{Bound, Vanish}} \\ & - \Delta G_{\text{Bound, Discharge}} - \Delta G_{\text{Bound, Restrain}} + \Delta G_{\text{Sym. Corr.}} \\ =& \Delta G_{\text{Free}} + \Delta G^{o}_{\text{Bound}},\end{aligned} \tag{2.1}$$

where “Free” and “Bound” indicate that the ligand is in solution or in the receptor binding site, “Discharge” means removal of ligand Coulombic interactions, “Vanish” indicates removal of the ligand Lennard-Jones (LJ) interactions, and “Restrain” means introduction of intermolecular restraints between the receptor

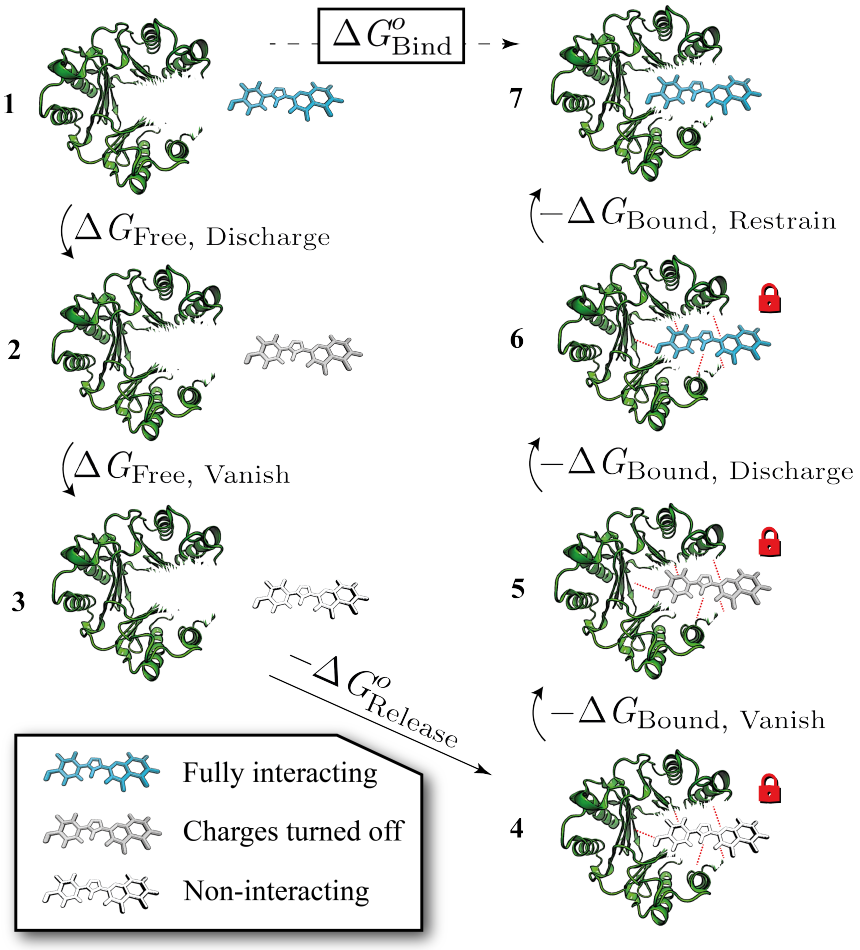


Figure 2.2: An alchemical thermodynamic cycle for the calculation of ABFEs. The red dashed lines indicate protein-ligand restraints. The locks indicate that receptor-ligand restraints are active. It is generally computationally intractable to obtain $\Delta G^{o}_{\mathrm{Bind}}$ by direct simulation, but the alchemical cycle allows $\Delta G^{o}_{\mathrm{Bind}}$ to be obtained through a series of states which are less challenging to sample at equilibrium. $\Delta G^{o}_{\mathrm{Release}}$ is calculated without simulation.

and ligand. $\Delta G^{o}_{\mathrm{Release}}$ is the correction for releasing the receptor-ligand restraints when the ligand has no intermolecular interactions, and $\Delta G_{\mathrm{Sym.\ Corr.}}$ accounts for symmetries broken by the introduction of restraints.[73] The bound leg refers to all calculations where the ligand is in the binding site and includes the symmetry correction, and the free leg describes all calculations where the ligand is in solution.

When only the intermolecular components of the ligand non-bonded interactions (Coulombic or Lennard-Jones (LJ)) are removed, this is often termed "decoupling", while "annihilation" may refer to removal of both the inter- and intramolecular components.[168] However, the terminology of Gilson et al. is used here:[49] "decoupling" denotes removal of of the ligand intermolecular interactions while enforcing receptor-ligand restraints, irrespective of how the intramolecular interactions are treated.

2.2.2 Receptor-Ligand Restraints

The free energy of releasing the restraint on the decoupled ligand is given by the ratio of configurational integrals

$$\Delta G_{\text{Release}} = -k_{\text{B}}T \ln \frac{Z_{\text{State 3}}}{Z_{\text{State 4}}}, \tag{2.2}$$

where k_{B} is the Boltzmann constant, T is the temperature, and $Z_{\text{State 3}}$ and $Z_{\text{State 4}}$ are the configurational integrals for states 3 and 4, as defined in Figure 2.2. However, this result is dependent on the size of the water box in State 3, V_{Box}. A more useful quantity is the standard free energy of binding

$$\Delta G^{o}_{\text{Release}} = -k_{\text{B}}T \ln \frac{Z_{\text{State 3}}V^{o}}{Z_{\text{State 4}}V_{\text{Box}}}, \tag{2.3}$$

which is independent of V_{Box}. $V^{o} = 1660\text{Å}^{3}$ is the standard state volume. This ratio could be evaluated using a simulation in which the restrained decoupled ligand is completely released into the simulation box, but this would be slow to converge. Instead, this ratio must be simplified so that it can be evaluated without simulation. It can be shown (Section A.2) that Equation 2.3 can be written

$$\begin{aligned}\Delta G^{o}_{\text{Release}} = &- k_{\text{B}}T \ln V^{o} 8\pi^{2} + k_{\text{B}}T \ln \int e^{-\frac{W_{\text{r}}(\mathbf{x}_{\text{Ext}})}{k_{\text{B}}T}} |\mathbf{J}| \text{d}\mathbf{x}_{\text{Ext}} \\ &- \Delta G_{\text{Preorg.}} - \Delta G_{\text{Distort.}},\end{aligned} \tag{2.4}$$

where $W_{\text{r}}(\mathbf{x}_{\text{Ext}})$ is the potential of mean force (PMF) of the receptor-ligand restraint energy with respect to the six relative receptor-ligand external DoF, $\mathbf{x}_{\text{Ext}}$. The form of the Jacobian determinant, $|\mathbf{J}|$, depends on the coordinate transformation used to extract the relative external degrees of freedom from the

internal degrees of freedom of the complex. $\Delta G_{\text{Preorg.}}$ accounts for straining of the receptor and decoupled ligand when $W_{\text{r}}(\mathbf{x}_{\text{Ext}})$ is at its minimum, while $\Delta G_{\text{Distort.}}$ accounts for further distortion of the receptor and ligand when $W_{\text{r}}(\mathbf{x}_{\text{Ext}})$ is not at its minimum.

For an arbitrary set of receptor-ligand restraints, there is no straightforward way to evaluate this expression. The standard solution to obtain $\Delta G^{o}_{\text{Release}}$ is to select a set of receptor-ligand restraints for which the restraint energy, $U_{\text{r}}(\mathbf{x}_{\text{Ext}})$, is a function of only the receptor-ligand relative external degrees of freedom. In this case, $\Delta G_{\text{Preorg.}} = \Delta G_{\text{Distort.}} = 0$, because changing the relative coordinates of the decoupled ligand and receptor has no effect on the internal DoF and $W_{\text{r}}(\mathbf{x}_{\text{Ext}}) = U_{\text{r}}(\mathbf{x}_{\text{Ext}})$. Restraints of this form are described as not coupling the internal and external degrees of freedom of the receptor and ligand. Intuitively, such restraints do no "squeeze" or "stretch" the receptor or decoupled ligand. The free energy of releasing these restraints is

$$\Delta G^{o}_{\text{Release}} = -k_{\text{B}}T \ln V^{o} 8\pi^{2} + k_{\text{B}}T \ln \int e^{-\frac{U_{\text{r}}(\mathbf{x}_{\text{Ext}})}{k_{\text{B}}T}} |\mathbf{J}| \text{d}\mathbf{x}_{\text{Ext}}, \tag{2.5}$$

which can be integrated directly. Thus, the ideal receptor-ligand restraints would not couple the internal and external degrees of freedom of the receptor and ligand, so that Equation 2.5 is valid. The ideal restraints would also ensure optimal convergence of the bound stages. This might be achieved by mimicking the native receptor-ligand interactions as closely as possible,[150] thus minimally perturbing the fully-interacting complex while maximally restricting the accessible configurational volume during decoupling.[161] However, the extent to which this can be achieved is limited when restraining only six DoF. Hence, it may be desirable to use restraints which do couple these internal and relative external degrees of freedom in order to accelerate convergence of the decoupling calculations. In this case, the restraints which do couple these degrees of freedom should be released through simulation to those which do not (and the associated free energy change accounted for), or some degree of error must be tolerated in the calculation of $\Delta G^{o}_{\text{Release}}$. Finally, the ideal restraints would lack instabilities, and would be simple to select in an easily automatable manner.[141,159,169]

2.2.3 Boresch Restraints

Boresch restraints (Figure 2.3) only affect the six relative external degrees of freedom of the ligand with respect to the receptor, and do not couple the receptor and ligand internal and external degrees of freedom.[50] As a result, Equation 2.5

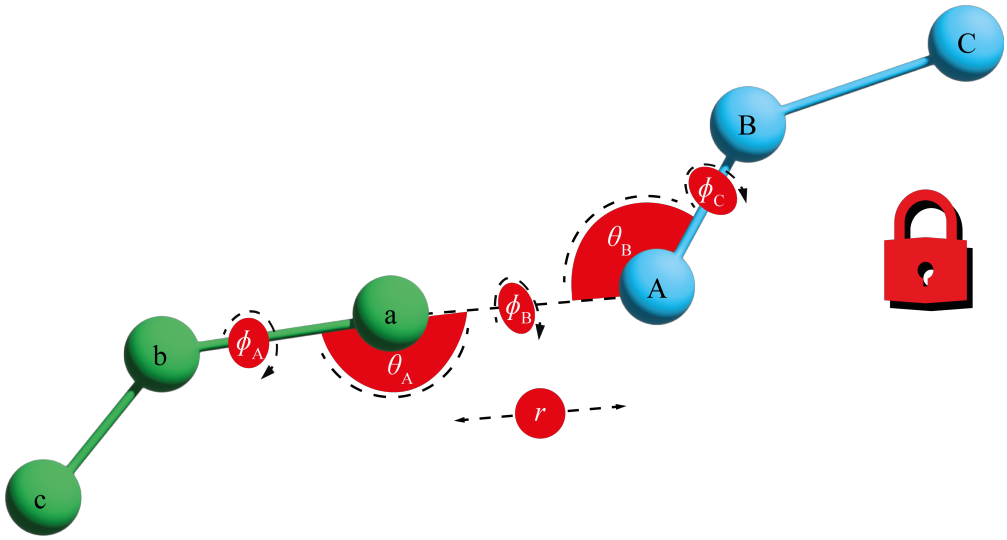


Figure 2.3: The general form of Boresch restraints.[50] Three anchor points (a, b, and c) are selected based on the receptor (green) coordinates and three (A, B, and C) are selected based on the ligand (blue) coordinates. The restrained six external degrees of freedom are defined as one distance (r), two bond angles (θ_A and θ_B), and three dihedral angles (ϕ_A, ϕ_B, and ϕ_C). In this diagram, the contiguous anchor points b, a, and A are close to collinear and therefore these restraints would be expected to be unstable.

can be easily evaluated by numerical integration of the expression

$$\begin{aligned}\Delta G^{o}_{\mathrm{Release}} = &- k_\mathrm{B}T \ln V^{o} 8\pi^2 \\ &+ k_\mathrm{B}T \ln \int_0^{\infty}\int_0^{\pi}\int_0^{2\pi} e^{-\frac{u(r)+u(\theta_\mathrm{A})+u(\phi_\mathrm{A})}{k_\mathrm{B}T}} r^2 \sin\theta_\mathrm{A}\,\mathrm{d}\theta_\mathrm{A}\mathrm{d}\phi_\mathrm{A}\mathrm{d}r \\ &+ k_B T \ln \int_0^{\pi}\int_0^{2\pi}\int_0^{2\pi} e^{-\frac{u(\theta_\mathrm{B})+u(\phi_\mathrm{B})+u(\phi_\mathrm{C})}{k_\mathrm{B}T}} \sin\theta_\mathrm{B}\mathrm{d}\theta_\mathrm{B}\mathrm{d}\phi_\mathrm{B}\mathrm{d}\phi_\mathrm{C}, \quad (2.6)\end{aligned}$$

where $u(x)$ is the restraining potential applied to the degree of freedom x. These degrees of freedom are defined in Figure 2.3. The second term in Equation 2.6 integrates over the position of anchor atom A with respect to the receptor coordinates, and does not depend on anchor atoms B and C. The third term integrates over the orientation of the ligand with respect to the receptor. Hence, Boresch restraints can be used to restrain only the position of the anchor point A with respect to the receptor, by setting $u(\theta_\mathrm{B})$, $u(\phi_\mathrm{B})$, and $u(\phi_\mathrm{C})$ to 0.

Although it is common to use harmonic restraining potentials and to select the anchor points as atomic positions, these are not constraints of the framework. For example, periodic dihedral restraints can be used,[142] and anchor points may be derived from multiple atomic positions or centres of mass.[166] Harmonic restraints are used in this work. In this case, Equation 2.6 can be evaluated analytically as

$$\Delta G^{o}_{\mathrm{Release}} = -k_{\mathrm{B}}T \ln \left[\frac{V^{o} 8\pi^{2}}{r_{0}^{2} \sin\theta_{\mathrm{A},0} \sin\theta_{\mathrm{B},0}} \frac{\left(K_{\mathrm{r}} K_{\theta_{\mathrm{A}}} K_{\theta_{\mathrm{B}}} K_{\phi_{\mathrm{A}}} K_{\phi_{\mathrm{B}}} K_{\phi_{\mathrm{C}}}\right)^{1/2}}{(2\pi k_{\mathrm{B}}T)^{3}} \right], \quad (2.7)$$

where K denotes a force constant and 0 denotes an equilibrium value.[50] This assumes that r, $\sin\theta_{\mathrm{A}}$, and $\sin\theta_{\mathrm{B}}$ can be taken out of the integrals in Equation 2.6 and replaced by their equilibrium values.

As mentioned previously, when the anchor points are arranged so that large changes in the six DoF defined in the VBA framework result from small changes to the Cartesian coordinates of atoms, this can result in large forces and simulation crashes. To avoid such instabilities, anchor points must be carefully selected to avoid the collinearity of any 3 contiguous anchor points.

Algorithms have been proposed for the selection of Boresch restraints.[1,61,140,150,170,171] At a minimum, these aim to select stable restraints based on the geometry of the complex, while more sophisticated methods aim to enhance convergence by directly (e.g. based on H-bonds) or indirectly (based on minimum total variance of the distance, angles, and dihedrals) mimicking strong receptor-ligand interactions based on a short unrestrained simulation. However, there is no obviously superior method which has been shown to guarantee selection of numerically stable restraints with optimal convergence properties.

2.2.4 Multiple Distance Restraints

Restraints schemes based on multiple distance restraints do not suffer from the inherent instabilities of Boresch restraints, and allow the ligand to be restrained to a greater extent (Figure 2.4). In this work, harmonic or flat-bottomed distance restraints were used. Using the former, U_{r} is given by

$$U_{\mathrm{r}}(r_1, \ldots, r_N) = \sum_{n=1}^{N} \frac{1}{2} K_n (r_n - r_{n,0})^2, \quad (2.8)$$

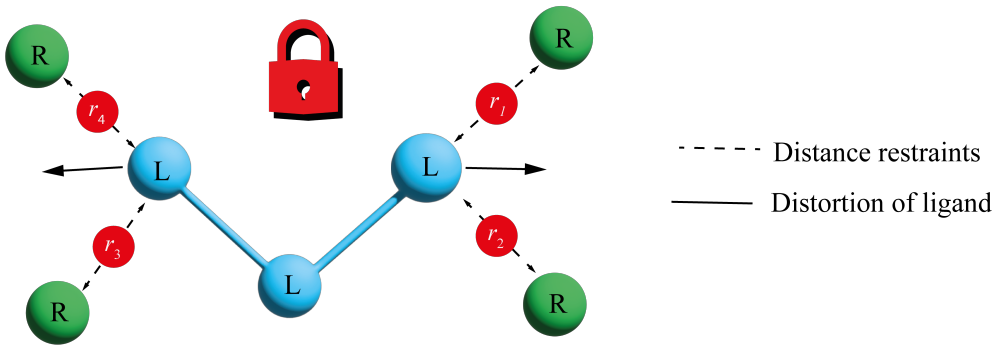


Figure 2.4: A ligand restrained using multiple distance restraints. Four distance restraints (r_1 - r_4) are applied between pairs of atoms in the ligand (blue) and the receptor (green). Applying distance restraints between several receptor-ligand atom pairs can allow greater restriction of ligand movement than a typical six DoF restraint scheme. However, multiple distance restraints can couple the internal and external degrees of freedom of the receptor and ligand, as shown by the distortion of the ligand, preventing rigorous calculation of the free energy of releasing the decoupled ligand.

and with the latter the total restraint energy is given by

$$U_{\mathrm{r}}(r_1,\ldots,r_N) = \sum_{n=1}^{N} \begin{cases} 0 & \text{if } |r_n - r_{n,0}| \leq r_{n,\mathrm{fb}} \\ \frac{1}{2}K_n(|r_n - r_{n,0}| - r_{n,\mathrm{fb}})^2 & \text{if } |r_n - r_{n,0}| > r_{n,\mathrm{fb}}, \end{cases} \tag{2.9}$$

where r_n is the distance between atoms in a restrained pair, $r_{n,0}$ is the equilibrium distance, $r_{n,\mathrm{fb}}$ is the flat-bottomed radius, and K_n is the force constant for restrained pair n. N is the total number of restrained pairs.

The main flaw of multiple distance restraints is that, applied naïvely, they may couple the internal and external degrees of freedom of the receptor and ligand, preventing the simplification of Equation 2.4 to 2.5 and hence the calculation of $\Delta G^o_{\mathrm{Release}}$. Here we investigate three approaches to circumvent this difficulty (Figure 2.5). The first is to implement multiple distance restraints in such a way that the error introduced is negligible (the "naïve" approach). This is the basis for recent implementations in the Michel group[165,172,173], in which sets of relatively permissive flat-bottomed restraints were used. An approximate value for $\Delta G^o_{\mathrm{Release}}$ is calculated by numerical integration of

$$\begin{aligned} \Delta G^o_{\mathrm{Release}} \approx & -k_B T \ln\left(8\pi^2 V^o\right) + \\ & k_B T \ln \int_0^{x_{\mathrm{Box}}} \int_0^{y_{\mathrm{Box}}} \int_0^{z_{\mathrm{Box}}} \int_0^{2\pi} \int_0^{\pi} \int_0^{2\pi} \sin\theta e^{-\frac{U_{\mathrm{r}}(x,y,z,\psi,\theta,\phi)}{k_{\mathrm{B}}T}} \mathrm{d}x\mathrm{d}y\mathrm{d}z\mathrm{d}\psi\mathrm{d}\theta\mathrm{d}\phi, \end{aligned} \tag{2.10}$$

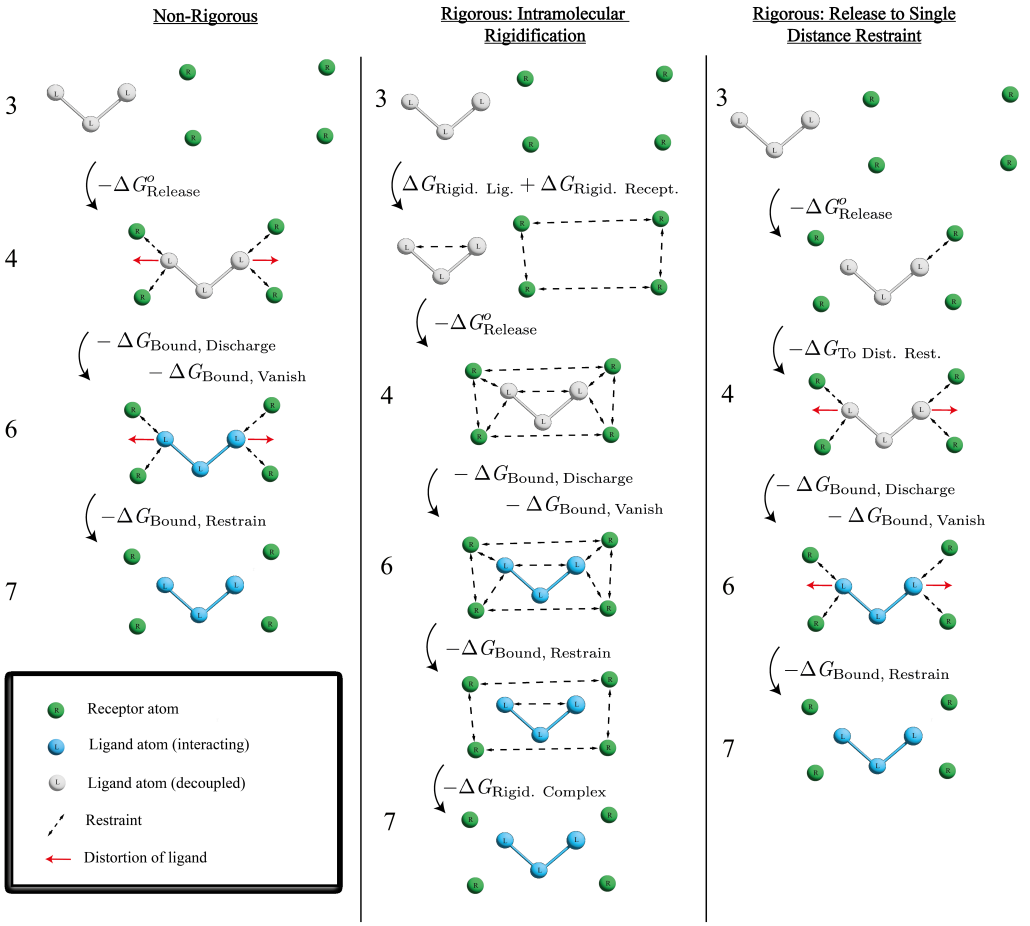


Figure 2.5: Multiple distance restraints schemes. The numbers indicate the most similar states from Figure 2.2. From top to bottom, all schemes start in State 3 and progress to State 7 (Figure 2.2). For the example shown, the "naïve" (non-rigorous) multiple distance restraint scheme may suffer from large systematic error due to distortion of the flexible ligand, which is not accounted for in the $-\Delta G^{o}_{\text{Release}}$ stage. This is avoided in the other schemes by either applying strong intramolecular restraints, or releasing all but one distance restraint.

where x_{Box}, y_{Box}, and z_{Box} are the side lengths of the simulation box, and x, y, z, ψ, θ, and ϕ are the Cartesian coordinates of the centre of mass and the Euler angles of the ligand in the frame of reference of the receptor. This is evaluated by taking the average intramolecular coordinates of the anchor points from a simulation of State 4. The receptor and ligand are then assumed to be rigid, allowing $U_{\text{r}}(x, y, z, \psi, \theta, \phi)$ to be calculated by translating and rotating the ligand anchor points with respect to the receptor anchor points and evaluating $U_{\text{r, Rigid}}(r_1, \ldots, r_N) = U_{\text{r, Rigid}}(x, y, z, \psi, \theta, \phi)$ at each point, where "Rigid" shows that the anchor points have been fixed to their average intramolecular positions.

This assumes that $\Delta G_{\text{Preorg.}} = 0$ (Equation 2.4), and produces a bias towards more negative free energies of binding, because $\Delta G_{\text{Preorg.}} \geq 0$. The error will be substantial when the average intramolecular positions of the anchor points for the decoupled complex are very different to those in the free ligand and receptor, or the restraints substantially restrict the conformational freedom of the receptor or ligand, thus enforcing substantial "preorganisation". The magnitude of the error would be expected to increase with the number and strength of restraints, and decreasing volume of the flat-bottom region. This is likely to be particularly problematic for systems where there are substantial changes in the conformations of the ligand and the binding site upon binding. In addition, the use of average positions assumes that $\Delta G_{\text{Distort.}} = 0$ and $U_{\text{r, Rigid}}(x, y, z, \psi, \theta, \phi) = W_{\text{r}}(x, y, z, \psi, \theta, \phi)$. This produces a slight bias towards more positive free energies of binding, because a flexible system will distort to minimise the sum $W_{\text{r}}(x, y, z, \psi, \theta, \phi) + \Delta G_{\text{Distort., Point}}(x, y, z, \psi, \theta, \phi) \leq U_{\text{r, Rigid}}(x, y, z, \psi, \theta, \phi)$ for a given relative position and orientation. $\Delta G_{\text{Distort., Point}}(x, y, z, \psi, \theta, \phi)$ is related to $\Delta G_{\text{Distort.}}$ by integration over the six external DoF. Overall, neglecting flexibility is expected to result in an erroneously negative $\Delta G^{o}_{\text{Release}}$ due to neglect of $\Delta G_{\text{Preorg.}}$. Obtaining an accurate free energy of binding under these approximations depends upon the restraints being sufficiently permissive and there being no large changes in the average intramolecular coordinates of the anchor points during binding.

Here we investigate an alternative scheme to eliminate these sources of error: applying intramolecular restraints to rigidify the anchor points. The free energy of applying and releasing these in the free and bound stages can be explicitly calculated with simulation. The restraints should be sufficiently strong to "preorganise" the system, and substantially stronger than the intermolecular restraints. This guarantees that $\Delta G_{\text{Preorg.}} = \Delta G_{\text{Distort.}} = 0$ and that $W_{\text{r}}(x, y, z, \psi, \theta, \phi) = U_{\text{r, Rigid}}(x, y, z, \psi, \theta, \phi)$, rendering Equation 2.10 rigorous.

An alternative rigorous approach to the implementation of multiple distance restraints is to evaluate the free energy change, $\Delta G_{\text{To Dist. Rest.}}$, for releasing them to a single harmonic or flat-bottomed restraint after decoupling. Once only a single distance restraint is active, the free energy of release can be calculated exactly, as is done for the DBC restraint.[160,161] The free energy of releasing a single distance restraint can be calculated using

$$\Delta G^{o}_{\text{Release}} = -k_{\text{B}}T \ln V^{o} + k_{\text{B}}T \ln 4\pi \int_{0}^{\infty} r^{2} e^{-\frac{U_{\text{r}}(r)}{k_{\text{B}}T}} \, \text{d}r, \tag{2.11}$$

where r is the distance between two anchor atoms and $U_{\mathrm{r}}(r)$ takes the form of Equation 2.8 or 2.9.

2.3 Methods

2.3.1 System Preparation

The MIF/MIF180 systems were set up approximately following the methodology of Qian et al.[166] The structures of MIF and MIF180 were obtained from the crystal structures of complexes of MIF with both MIF180 (relatively poor resolution - 2.6 Å, PDB ID 4WR8) and the structurally similar MIF190 (relatively high resolution - 1.8 Å, PDB ID 4WRB).[174] Five extra copies of the biological assembly were removed from 4WR8 by deleting all chains from D onwards. The higher resolution structure was superimposed on the lower resolution structure by aligning chain B from 4WR8 with chain A from 4WRB using PyMOL.[167] Crystallographic waters (from 4WRB) were retained and all other non-protein and non-ligand atoms were discarded. For atoms with alternative locations, the location with the greatest occupancy was selected. Thirteen missing atoms were added using pdb4amber.[175] H++ (version 3.2) and PROPKA (version 3.4.0) were used to suggest the protonation state of all residues.[176–178] Consistent with the literature for the apo protein,[179] a large pKa shift was predicted in both cases for the N-terminal proline residue which is present in the binding site, which produces the neutral form at a pH of 7. The suggested protonation sites for all histidines were taken from H++. Hydrogens were added to MIF using H++ and a hydrogen ion was removed from the N of each of the protonated terminal prolines so as to maintain the hydrogen bond to Tyr36. Based on the results of Qian et al., the neutrality of the terminal proline was maintained in the complex.

There are no parameters available for neutral N-terminal proline in the AMBER ff14sb forcefield.[180] Therefore, antechamber 21.0 was used to parameterise the neutral proline with an NME-capped C-terminus, using AM1-BCC partial charges and AMBER atom types.[85,86] The remainder of the MIF protein was parameterised using the AMBER ff14sb force field. Hydrogens were added to MIF180 using Open Babel (version 3.0.0) and BioSimSpace (version 2020.1.0) was used to parameterise the ligand in the *syn* conformation (Figure A.2) with the GAFF2.11 force field and AM1-BCC charges using antechamber 21.0.[181,182]

Absolute Binding Free Energy Calculations

Alchemical ABFE calculations were performed using the double decoupling method,[49] according to the cycle shown in Figure 2.2 and the multiple distance restraint schemes shown in Figure 2.5:

$$\begin{aligned}\Delta G^{o}_{\text{Bind}} =& \Delta G_{\text{Free, Discharge}} + \Delta G_{\text{Free, Vanish}} - \Delta G^{o}_{\text{Release}} - \Delta G_{\text{Bound, Vanish}} \\ & - \Delta G_{\text{Bound, Discharge}} - \Delta G_{\text{Bound, Restrain}} + \Delta G_{\text{Sym. Corr.}} \\ & + \Delta G_{\text{Rest. Scheme}} \qquad (2.12)\\ =& \Delta G_{\text{Free}} + \Delta G^{o}_{\text{Bound}} \qquad (2.13)\end{aligned}$$

$$\Delta G_{\text{Rest. Scheme}} = \begin{cases} 0 & \text{if Boresch} \\ \Delta G_{\text{Rigid. Recept.}} + \Delta G_{\text{Rigid. Lig.}} & \\ -\Delta G_{\text{Rigid. Complex}} & \text{if intramol. rigid.} \\ -\Delta G_{\text{To Dist. Rest.}} & \text{if release,} \end{cases} \qquad (2.14)$$

where $\Delta G_{\text{Rest. Scheme}}$ collects any additional terms which are specific to the restraints scheme used. The relevant restraints scheme for each form of $\Delta G_{\text{Rest. Scheme}}$ is shown on the right of Equation 2.14: "Boresch" includes both Boresch restraints and the "non-rigorous" implementation of multiple distance restraints shown in Figure 2.5, "intramol. rigid." refers to multiple distance restraints with intramolecular rigidification, and "release" denotes multiple distance restraints with release to a single distance restraint. $\Delta G_{\text{Rigid. Recept.}}$ and $\Delta G_{\text{Rigid. Lig.}}$ are the free energy changes for intramolecular rigidification of receptor and ligand anchor points, $-\Delta G_{\text{Rigid. Complex}}$ accounts for the release of intermolecular restraints in the receptor- interacting ligand complex, and $\Delta G_{\text{To Dist. Rest.}}$ is the free energy of releasing several distance restraints to a single distance restraint. For the intramolecular rigidification scheme, the intramolecular restraints are applied in State 3, and released in state 7. $\Delta G_{\text{Sym. Corr.}}$ is at least $-k_{\text{B}}T \ln 3$ for this system, because MIF is a homotrimer with three equivalent binding sites, and all restraints restrict the ligand to a single binding site. $\Delta G_{\text{Sym. Corr.}}$ may be larger if the restraints also break symmetries arising from the structure of the ligand (Section A.6). ΔG_{Free} and $\Delta G^{o}_{\text{Bound}}$ are the overall free and bound leg contributions to $\Delta G^{o}_{\text{Bind}}$, where $\Delta G^{o}_{\text{Bound}}$ includes all terms other than $\Delta G_{\text{Free, Discharge}}$ and $\Delta G_{\text{Free, Vanish}}$.

No correction is included to account for the truncation of the tails of the LJ potentials because initial simulations yielded negligible corrections ($\approx$ 0.1 kcal mol^{-1}).[76] Protein-ligand restraints were introduced, and charges and LJ interactions were incrementally removed by scaling the coupling parameter, λ, from

0 to 1. For calculations with Boresch restraints and multiple distance restraints without intramolecular rigidification or release to a single distance restraint, the force constants of the protein-ligand restraints and the magnitude of the charges were scaled linearly with λ. The soft-core potential implemented in Sire (with a LJ soft-core parameter set to 2.0) which is based on the potentials of Zacharias and McCammon,[138] and Michel et al.,[183] was used to scale the LJ interactions. Eight evenly-spaced windows and 18 non-evenly spaced λ-windows (0.000, 0.028, 0.056, 0.111, 0.167, 0.222, 0.278, 0.333, 0.389, 0.444, 0.500, 0.556, 0.611, 0.667, 0.722, 0.778, 0.889, and 1.000) were used for the free discharging and vanishing stages, respectively. Six non-evenly spaced windows (0.000, 0.125, 0.250, 0.375, 0.500, and 1.000), 8 evenly spaced windows, and 36 non-evenly spaced windows (31 evenly spaced windows from λ=0 to 0.750, then 5 evenly spaced windows from λ=0.750 to 1.000) were used for the bound restraining, discharging, and vanishing stages, respectively. The window spacings were selected to yield sufficient overlap without excessive numbers of windows based on initial test simulations. The intramolecular components of both the Coulombic and LJ interactions between ligand atoms were completely removed.

For calculations where multiple distance restraints were released to a single distance restraint, $\Delta G^{o}_{\text{Release}}$ was calculated by numerical integration of Equation 2.11, and the force constants were scaled with λ^5 over 21 evenly-spaced λ windows.[161] The same protocol was used to introduce and remove all restraints in the multiple distance restraints simulations with intramolecular restraints.

Restraints were selected to optimally mimic native protein-ligand interactions by post-processing a 6 ns simulation of the fully interacting complex.[155,158] From the first frame, all heavy atoms in the protein within 10 Å of the ligand, and all heavy atoms in the ligand were selected. To avoid anchor points with poor correlation in position, the distances between all possible protein-ligand atom pairs from this selection were tracked over the trajectory, and the 200 pairs with the lowest standard deviation were selected. For multiple distance restraints, only the lowest variance pair for any anchor point was retained, provided that neither of the anchor points had already been selected for use in another restraint. For Boresch restraints, all pairs were taken as candidate anchor points a and A (Figure 2.3). For each pair, adjacent heavy atoms were selected to complete the sets of Boresch anchor points. These sets were ordered by increasing total variance of the Boresch DoF, as done by Alibay,[1,170] and sets of anchor points were discarded if the average values of θ_{A} or θ_{B} were below 30 or above 150 degrees. The equilibrium values for all restraints were taken to be their average values

during the unrestrained simulation. Force constants were selected so that in the decoupled state, the harmonic restraints would generate the same distributions as observed in the coupled state.[155] Gaussian distributions in the coupled state were assumed and the variances of the Boresch DoF were used to calculate the force constants (Section A.7).

For the Boresch restraints, $\Delta G^{o}_{\text{Release}}$ was calculated by numerical integration of Equation 2.6. This was used in preference to the analytical correction to avoid potential errors introduced by the approximations required to derive Equation 2.7. For multiple distance restraints, numerical integration of Equation 2.10 was performed using the "standardstatecorrection" utility available within Sire,[184] using all frames of the trajectory, a buffer of 5 Å, a translational volume element of 0.25 Å, and 30 orientations per $[0, 2\pi]$ Euler angle interval.

2.3.2 Molecular Dynamics Protocol

Solvation and equilibration were performed using BioSimSpace.[182] The protein-ligand complex was placed in a periodic cube of side 84 Å (determined by the longest edge of the axis-aligned bounding box plus 15 Å padding on each side) and solvated with TIP3P water molecules.[185] 150 mM NaCl was added. The system was energy-minimised using PMEMD (50000 steps).[175] Equilibration in the NVT ensemble was performed using PMEMD (5 ps with all non-solvent atoms restrained and heating from 0 to 298 K, followed by 50 ps with only backbone atoms restrained, then 50 ps with no restraints), followed by equilibration in the NPT ensemble at 1 atm and 298 K using PMEMD.CUDA (400 ps with all non-solvent heavy atoms restrained, followed by 2 ns with no restraints).

The free ligand was solvated with TIP3P water and 150 mM NaCl in a periodic box of side length 40 Å. 50000 steps of minimisation were performed using PMEMD. Equilibration in the NVT ensemble was performed using PMEMD (5 ps with all non-solvent atoms restrained and heating from 0 to 298 K, followed by 50 ps with no restraints), followed by equilibration in the NPT ensemble with PMEMD.CUDA (1 atm and 298 K with restraints on non-solvent heavy atoms for 200 ps followed by 2 ns with no restraints). A Langevin thermostat and Berendsen barostat were used for the relevant equilibration steps.[186]

All alchemical simulations were performed using the software SOMD,[187] available within Sire (version 2022.2.0).[184] SOMD was modified to allow the use of Boresch restraints, the scaling of restraints with λ^5, and the simultaneous use of different restraints. The code implementing Boresch restraints has been integrated into the

main branch of Sire. An Andersen thermostat (collision frequency 10 ps^{-1}) and Monte Carlo barostat (25 time steps between isotropic box scaling attempts) were used to maintain a temperature and pressure of 298 K and 1 atm.[41,101] A timestep of 4 fs was used in combination with the leap-frog Verlet integrator and hydrogen mass repartitioning (using a repartitioning factor of 4).[188] All bond lengths were constrained. The reaction field method was used with a dielectric constant of 78.3,[189] and a cut-off of 12 Å was used for all non-bonded interactions. Energy minimisation was performed prior to each simulation with a maximum of 1000 iterations. The bound stage vanish λ windows were run for 8 ns, and all others for 6 ns. Free energy differences for each stage were estimated using the Multistate Bennett-Acceptance Ratio (MBAR) for the final 5 ns of all simulations.[121,131] Coordinates were saved every 20 ps.

All simulations were repeated five times with independent starting velocities. Because the MBAR uncertainties estimated from single runs were small compared to the variation between repeat runs, errors are reported as 95% confidence intervals based on the deviation between independent replicates, assuming Gaussian distributions and using t-values for 4 degrees of freedom. Student's t-test was used to assess evidence for a significant difference at 95% confidence.

2.4 Results and Discussion

2.4.1 Boresch Restraints

Restraint Selection

From an initial set of restraining simulations, two binding poses were identified (Figure 2.6) which interconverted slowly on the timescale of the simulations (6 ns). To allow comparison of different restraints for a single binding pose, all restraints were fit to binding pose A, other than a single set of Boresch restraints which was fit to pose B. The calculation for pose B was carried out to allow comparison to the experimental free energy of binding. The overall free energy of binding obtained with only pose A alone is not generally shown, because neglect of binding pose B would make direct comparison to experiment misleading.

Three sets of Boresch restraints were selected initially for pose A (Figure 2.7). The first was the best-scoring set (B1) based on the minimum-variance algorithm, and mimics the phenol-Asn97C hydrogen bond (Figure 2.7). To test varied anchor point positions, the second set (B2) was selected as the top-scoring restraints

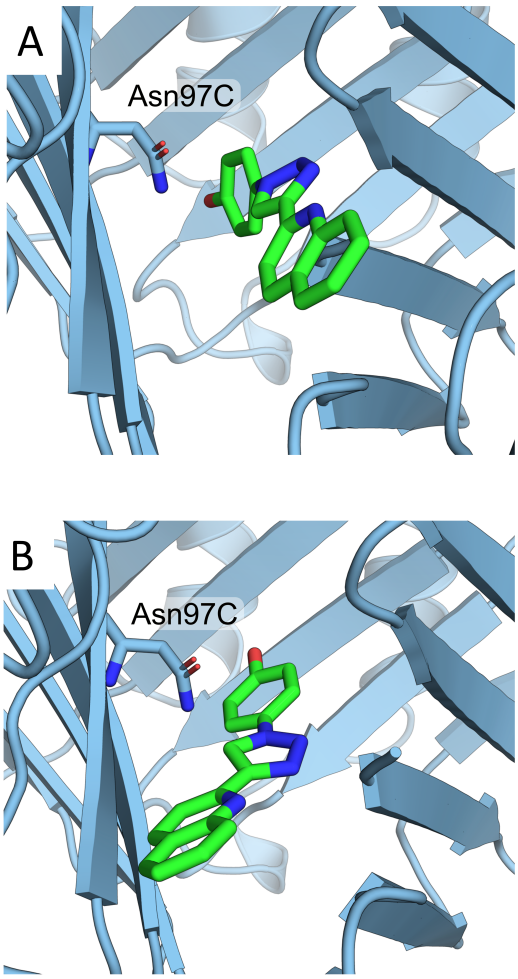


Figure 2.6: Alternative binding poses A (panel A) and B (panel B). Interconversion occurred rarely on the timescale of the simulation (6 ns). The Asn on which restraints B1 and B1-poseB are based is shown. Rendered with PyMOL.[167]

with anchor points outwith the phenol moiety - these were based on the triazole ring and were 7th best-scoring overall. Finally, selection was constrained to the quinoline moiety (B3). The parameters of restraint sets B1, B2, and B3 are given in Table A.1. The protein anchor points were located in residues with low root-mean-square fluctuations (RMSFs) of the α-carbon positions although this was not directly targeted by the algorithm (Figure A.3). B1-poseB was the highest scoring set of anchor points based on a simulation including only pose B. Similar to the B1 restraints, these were based on the phenol group of MIF-180 and Asn97C.

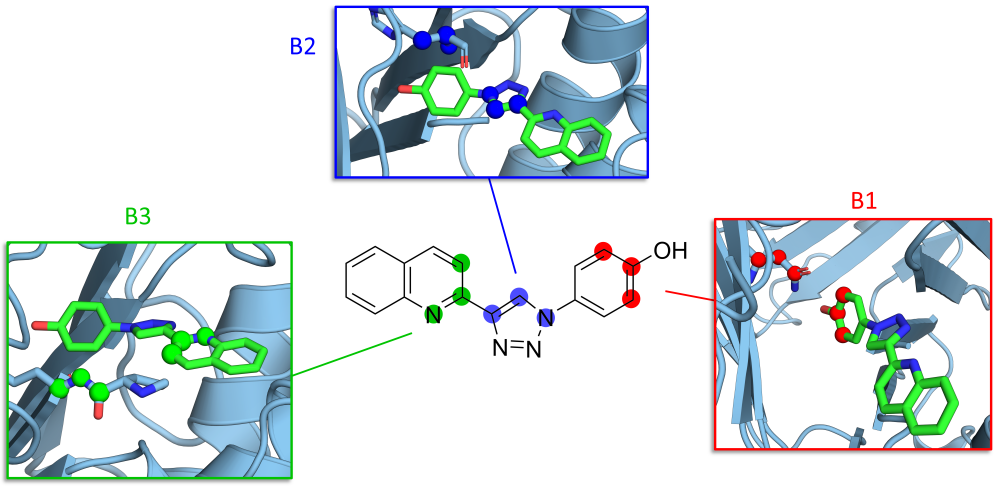


Figure 2.7: Sets of Boresch anchor points B1 (red), B2 (blue), and B3 (green). Anchor points are circled or shown as spheres. Windows rendered with PyMOL.[167]

A discussion of the challenges faced during restraint selection, symmetry corrections, and the strengths and limitations of the restraint selection algorithm is given in Section A.6. Improvements to the restraints selection algorithm are also proposed - in particular, we recommend scoring possible restraints using a metric calculated from the variances of the DoF of the prospective restraints, rather than using the total variances directly.

Results with Force Constants Fit to Simulation

Calculations were performed with Boresch restraints with force constants fit to simulation. Force constants were fit based on the variance of the Boresch DoF in State 7, as discussed in Section A.7. Several calculations were performed for pose A (B1, B2, and B3) to allow comparison between restraints (Table 2.1). A single calculation was performed for pose B (B1-poseB) to allow overall comparison with experiment by combining the results for both poses.

The results shown for $-\Delta G^{o}_{\text{Release}}$ were calculated by numerical integration of Equation 2.6. While use of the Boresch analytical correction introduces large errors in certain regimes (force constants very weak, r very short, θ_{A} or θ_{B} close to 0 or π rad), only small deviations between the analytical and numerical corrections, no greater than 0.04 kcal mol^{-1}, were found for any of the Boresch restraints. This is in accordance with previous studies.[190].

Table 2.1: Bound Leg Contributions to $\Delta G^{o}_{\mathrm{Bind}}$ with Boresch Restraints[a]

	Restraints								
Contribution	**B1**	**B2**	**B3**	**B1-P**	**B3-P***	**B1-10**	**B2-10**	**B3-10****	**B1-20**
$-\Delta G^{o}_{\mathrm{Release}}$	9.76	9.90	9.10	10.02	8.91	6.69	6.39	6.53	7.90
$-\Delta G_{\mathrm{Bound,\ Vanish}}$	-2.06 ± 1.08	-2.20 ± 0.94	-0.73 ± 0.25	-2.22 ± 0.19	-0.68 ± 0.06	-1.04 ± 1.21	0.09 ± 1.03	0.58 ± 1.96	-1.98 ± 1.18
$-\Delta G_{\mathrm{Bound,\ Discharge}}$	-12.99 ± 0.63	-11.90 ± 0.48	-11.86 ± 0.22	-11.76 ± 0.19	-11.71 ± 0.77	-12.46 ± 0.59	-11.62 ± 0.47	-12.04 ± 0.70	-12.64 ± 0.19
$-\Delta G_{\mathrm{Bound,\ Restrain}}$	-1.48 ± 0.04	-1.74 ± 0.03	-1.70 ± 0.11	-1.66 ± 0.08	-1.77 ± 0.11	-0.40 ± 0.01	-0.59 ± 0.08	-0.94 ± 0.21	-0.77 ± 0.20
$\Delta G_{\mathrm{Sym.\ Corr.}}$	-1.06	-0.65	-0.65	-1.06	-0.65	-1.06	-0.65	-0.65	-1.06
$\Delta G^{o}_{\mathrm{Bound}}$	-7.83 ± 1.25	-6.60 ± 1.05	-5.84 ± 0.36	-6.69 ± 0.28	-5.91 ± 0.78	-8.27 ± 1.35	-6.37 ± 1.13	-6.53 ± 2.09	-8.55 ± 1.21

[a] Results for binding pose A, in kcal mol^{-1}. Uncertainties stated as 95 % confidence intervals based on the variance of 5 replicate runs, assuming Gaussian distributions. Restraint parameters are given in Table A.1. -10 and -20 denote that all force constants were set to 10 or 20 kcal mol^{-1} Å^{-2} [rad^{-2}], respectively. * Run 1 excluded from average due to under-sampling of water in binding site during the vanishing stage. ** Due to simulations crashing, results based on 4 independent replicates with partial completion of many lambda windows.

The average of repeat runs for B1 - 3 and B1-poseB generally showed good convergence as assessed by lack of drift with increasing sampling time (Section A.8). However, there were substantial differences between replicate runs, most notably during the vanishing stage, which generally contributed the greatest uncertainty to $\Delta G^{o}_{\mathrm{Bound}}$. In all cases, this uncertainty could be traced back to a few windows around $\lambda = 0.4$ (Figure 2.8).

The majority of this uncertainty can be attributed to the entry of water to the binding site. Binding site water was defined as being simultaneously within 8 Å of the N atom in Pro1A, and CG2 in Val106A, which are on opposite sides of the binding site. At the start of the vanishing stage there were no waters in the binding site, which increased to an average of approximately 4.5 after decoupling, in good agreement with the crystal structure of free MIF (PDB ID 1gd0).[191] A sudden jump in water occupancy of the binding site occurred around $\lambda = 0.4$, the region of divergence of the PMFs for B1. Here, it was found that water may only enter the binding site by "forcing" the partially vanished ligand to the side of the binding site (Figure 2.9).

In cases where the binding site showed high water occupancy (runs 1, 2, and 4), the resultant strain favoured vanishing of the ligand, producing divergence of the vanish stage PMF towards more favourable free energies of vanishing. The opposite was true for runs 3 and 5, which had very low average water occupancies at $\lambda = 0.4$. This is in accord with the results of Rogers et al.,[192] who found negative correlation between water occupancy and the gradient of the potential energy with respect to λ at intermediate stages of the vanishing leg. Although dramatic fluctuations in the number of water molecules in the binding sites were observed at higher values of λ, this did not translate into additional uncertainty,

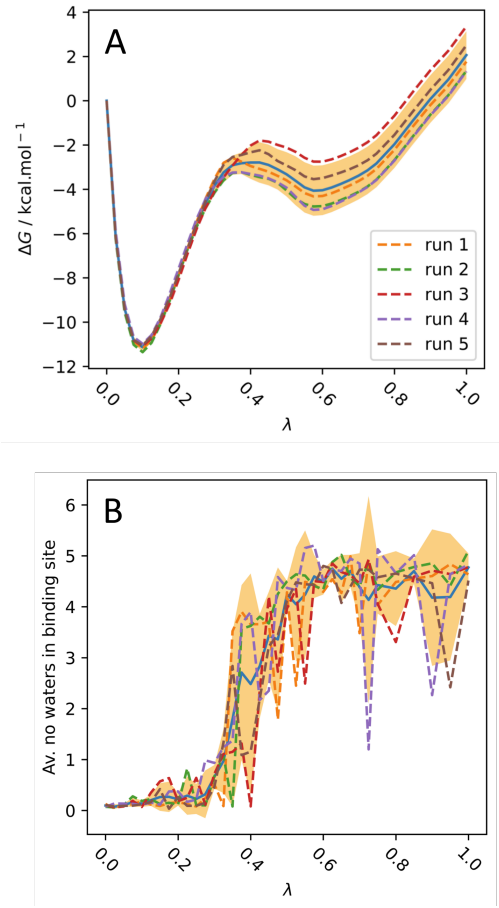


Figure 2.8: Panel A: PMF along λ for the bound vanish stage for B1. Panel B: Average number of waters in the binding site (defined as the overlap of two spheres of radius 8 Å centred on the N atom in Pro1A, and CG2 in Val106A) against λ during the vanish stage for B1. There is a strong correlation between the water occupancy of the binding sites and the divergence of the PMFs around $\lambda = 0.4$. The shaded area shows the 95% confidence interval and the solid blue line shows the mean.

because at this stage the LJ terms are mostly removed and the ligand is able to pass through atoms relatively smoothly. Some dips in the water occupancy were found to be due to the rotation of the side chain of Met2A to obstruct the end of the binding site.

Based on preliminary simulations, the equilibration time for the vanish stage simulations was increased to 3 ns per window to remove systematic error from slow movement of water into the binding site. However, as shown by Figure 2.8, occasionally water failed to enter the binding site during the entire simulation at values of λ where several waters entered the binding site during replicate runs. Therefore, some systematic error likely persisted in some cases. This appeared to be true for the B1 simulations. Irrespective of whether the force constants were

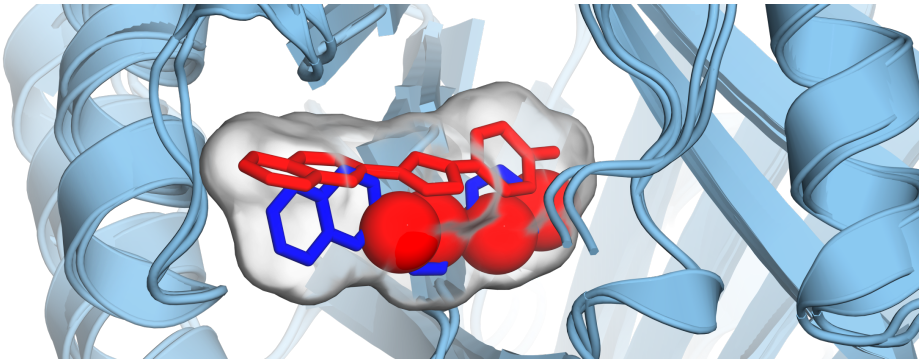


Figure 2.9: Waters (shown as spheres) in the binding site at $\lambda = 0.4$ for B1 at 6.48 ns. Only binding site waters (as defined as the overlap of two spheres of radius 8 Å centred on the N atom in Pro1A and CG2 in Val106A) are shown. The run 1 (red ligand/ waters) trajectory was superimposed on that from run 3 (blue ligand/ no waters present) by aligning MIF. Surface generated based on the ligands alone to approximately show the binding pocket.The ligand must be pushed to the side of the binding pocket to provide space for the water. Rendered with PyMOL.[167]

fit to simulation (B1), or set to 10 (B1-10) or 20 (B1-20) kcal mol^{-1}Å^{-2} or kcal mol^{-1} rad^{-2}, the results were more negative than for B2 and B3. In most cases the differences were significant at 95 % confidence. In addition, these simulations generally showed the greatest uncertainties between replicates, demonstrating greater random error. This was despite the B1 restraints being selected as optimum based on the minimum variance algorithm, and the protein anchor points being based on a stable Asn forming part of a β-sheet, highlighting the difficulty of selecting optimal restraints.

It seemed unlikely that the offset was due to the restraint of the phenol group, because multiple distance restraint schemes based on this group did not produce such negative free energies of vanishing (see Section 2.2.4). To ensure that this was not due to flexibility of the Asn side-chain, the calculation was repeated using ligand anchors in the phenol but protein anchors only in the backbone and Cα of this Asn. The result was very similar (-7.88 ± 0.95 kcal mol^{-1}), showing that the issue was not side-chain flexibility. Finally, the calculation was repeated using ligand atoms in the phenol but protein anchors in a different residue (Ile64A) - see B1-P. This resulted in a $\Delta G^{o}_{\mathrm{Bound}}$ of -6.69 ± 0.28 kcal mol^{-1} which had a much smaller uncertainty and was closer to the results for B2 and B3. The difference between B1 and B1-P was significantly different based on a Student's t-test, although it is acknowledged that the assumption of normally distributed free energy differences is likely to be incorrect.[193] In this system especially, where the dominant source of error appears to be the slow hopping of water between

energy minima, the central limit theorem is unlikely to be applicable. Regardless, the B1-P results were substantially less negative than the B1-10 and B1-20 results, suggesting that basing the protein anchors on the Asn for the B1 restraints may have made the simulations more susceptible to water sampling issues. It is possible that the restraints between the phenol and this Asn, which quickly forms a H-bond to water upon its entry to the binding site, created particularly high barriers to the entry of water, giving rise to systematic error. Regardless, the difference observed suggests that the approach of running independent replicate simulations with different restraints, as taken by Alibay et al.,[1] is sensible.

The B3 simulations were also rerun with protein anchor atoms on a different residue (Ala38A) - see B3-P. There was very good agreement between runs 2 - 5 for the bound vanish stage, but $\Delta G^{o}_{\text{Bound}}$ was around 3 kcal mol^{-1} more negative for run 1 due to comparatively low number of waters in the binding site over just 3 λ windows (see Section A.9). This highlights the importance of correctly sampling rehydration of the binding site upon decoupling.[194] It has been demonstrated that hybrid sampling approaches combining molecular dynamics with Monte Carlo water moves, in both the μVT and NPT ensembles, can improve the performance of relative binding free energy calculations.[195–197] In this system, it is possible that these approaches may perform poorly as a result of low acceptance probabilities, because water has to strain the partially decoupled ligand to enter the binding site around $\lambda = 0.4$, where proper sampling of rehydration is most critical. Non-equilibirum candidate Monte Carlo may overcome this by allowing for relaxation of the ligand position.[194] The strong dependence of the free energy on the correct sampling of binding-site water may make this a good system for testing methods for enhancing sampling of rehydration. Regardless, the $\Delta G^{o}_{\text{Bound}}$ obtained after discarding the data from run 1 (-5.91 $\pm$ 0.78 kcal mol^{-1}) was very similar to that obtained for B3 (-5.84 $\pm$ 0.36 kcal mol^{-1}).

Ignoring the results for B1, which seemed to be especially affected by water sampling issues, the remaining simulations with force constants fit to simulation showed generally good agreement (within 1 kcal mol^{-1}), demonstrating reasonable reproducibility with this restraints scheme. Averaging B2, B3, B1-P, and B3-P yielded a $\Delta G^{o}_{\text{Bound}}$ of -6.26 $\pm$ 0.72 kcal mol^{-1} for binding pose A.

Although the aim of this study was to compare the values of $\Delta G^{o}_{\text{Bound}}$ obtained with different restraints, a single set of five replicate calculations were carried out for the free leg to allow comparison with experiment. Comparison with experiment cannot be used for the comparison of restraint schemes as there are likely to be

systematic errors from the force field, but it is used here as a crude check on the overall results. There was excellent agreement between replicates of the free leg simulations and convergence was achieved quickly (Section A.10), yielding $\Delta G_{\text{Free}} = -3.08 \pm 0.14$ kcal mol^{-1} ($\Delta G_{\text{Free, Discharge}} = 9.03 \pm 0.07$ kcal mol^{-1}, and $\Delta G_{\text{Free, Vanish}} = -12.11 \pm 0.11$ kcal mol^{-1}). This was in spite of *syn-anti* interconversion, which was observed during at least one run for every lambda window, but which occurred slowly on the timescale of the simulations. $\Delta G^{o}_{\text{Bound}}$ for binding pose B (with the B1-poseB restraints) was -6.63 ± 1.11 kcal mol^{-1} ($-\Delta G^{o}_{\text{Release}} = 9.76$ kcal mol^{-1}, $-\Delta G_{\text{Bound, Discharge}} = -2.66 \pm 1.04$ kcal mol^{-1}, $-\Delta G_{\text{Bound, Vanish}} = -10.77 \pm 0.39$ kcal mol^{-1}, $-\Delta G_{\text{Bound, Restrain}} = -1.90 \pm 0.1$ kcal mol^{-1}, and $\Delta G_{\text{Sym. Corr.}} = -1.06$ kcal mol^{-1}). Combining the average result for pose A (-6.28 ± 0.49 kcal mol^{-1}, using all results from Table 2.1 excluding all B1 calculations and B3-10 due to the issues discussed) with the the result for pose B according to $\Delta G^{o}_{\text{Bound}} = -k_B T \ln(\exp(-\beta \Delta G^{o}_{\text{Bound, 1}}) + \exp(-\beta \Delta G^{o}_{\text{Bound, 2}}))$ yielded an overall $\Delta G^{o}_{\text{Bound}}$ of -6.89 ± 0.74 kcal mol^{-1} and $\Delta G^{o}_{\text{Bind}} = -9.97 \pm 0.76$ kcal mol^{-1}.[168]. The overall result was in good agreement with the experimental binding free energy of -8.98 ± 0.28 kcal mol^{-1},[198] but more negative than the value calculated by Qian et al. using molecular dynamics and the AMBER ff14sb and GAFF force fields (-7.47 ± 0.99 kcal mol^{-1}).[166] However, a clear comparison with the results of Qian et al. is prevented by a number of methodological differences.[166] For example, they only observe the *anti* and *syn* conformers of MIF180 during the free and bound legs, respectively, and they apply a penalty of 1.60 kcal mol^{-1} to account for this. Here, interconversion was observed in both the free and bound states and no correction was applied. Furthermore, Qian et al. do not apply a symmetry correction to account for the threefold symmetry of MIF, and do not perform calculations for an alternative binding pose.[166]

Results without Orientational Restraints

It was found that orientational restraints were essential for achieving reliable free energy estimates. The requirement for orientational restraints was investigated by repeating the B1 and B2 calculations without the orientational component of the restraint (K_{θ_B}, K_{ϕ_B}, and K_{ϕ_C} were set to 0) - see B1-o and B2-o in Table 2.2. B1 was also repeated setting all force constants other than K_r to 0, thus retaining only a single distance restraint - see B1-d. This resulted in large and significant shifts to more negative free energies of binding by 1.72, 3.84, and 2.05 kcal mol^{-1}, for B1-o, B2-o, and B1-d, respectively. Despite this, there was no obvious drift of the free energies with increasing sampling time (Section A.11).

Table 2.2: Bound Leg Contributions to ΔG^o_{Bind} without Orientational Restraints[a]

	Restraints		
	B1-o	**B2-o**	**B1-d**
$-\Delta G^o_{\text{Release}}$	4.94	4.51	1.08
$-\Delta G_{\text{Bound, Vanish}}$	-0.60 ± 0.71	-1.47 ± 0.99	1.67 ± 1.11
$-\Delta G_{\text{Bound, Discharge}}$	-12.50 ± 0.40	-11.93 ± 0.20	-11.89 ± 0.44
$-\Delta G_{\text{Bound, Restrain}}$	-0.74 ± 0.01	-0.91 ± 0.14	-0.09 ± 0.00
$\Delta G_{\text{Sym. Corr.}}$	-0.65	-0.65	-0.65
$\Delta G^o_{\text{Bound}}$	-9.55 ± 0.82	-10.44 ± 1.02	-9.88 ± 1.19

[a] All values in kcal mol^{-1}. Uncertainties stated as 95 % confidence intervals based on the variance of 5 replicate runs, assuming Gaussian distributions. -o and -d mean that all force constants other than k_r, k_{θ_A}, and k_{ϕ_A}, or k_r, were set to 0, respectively.

The lack of orientational restraints allows the mixing of binding poses A and B, but this is also not a plausible source of the error introduced. In the limit of perfect sampling, the free energy of binding can be no more negative than that calculated by combining the free energies of binding of the two poses as was done previously. The exception to this would be if there were other binding poses which were numerous or more favourable, which seems unlikely.

Instead, the negative offset is very likely due to the failure to sample all relevant orientations at intermediate values of λ during vanishing. The close agreement between B1-o and B1-d suggests that the offset is due to the removal of the orientational component of the restraint. As the LJ interactions are removed, the sampling of orientations different to that of the binding pose will become favourable. The gradient of the free energy change as the LJ interactions are removed will likely be less positive in these alternative orientations, because they were high in energy when the LJ interactions were at full strength. Therefore, failure to sample these orientations due to large barriers should give excessively positive free energies of vanishing, resulting in erroneously negative free energies of binding, as is observed.

This is illustrated by the divergence of the PMF for the bound vanish stage for B2-o (Figure 2.10). The divergence shows a strong correlation with the orientational sampling at $\lambda = 0.325$, and not with the presence of water in the binding site (Figure A.16). For run 4, the plots of ϕ_{C} show that sampling was largely restricted to the orientation of the original binding mode. This resulted in the most positive gradient of the PMF and the most favourable free energy of binding. During runs 1, 2, and 5, the ligand rotated length-ways in the binding site by around 90 degrees with respect to the initial pose, resulting in a less positive PMF gradient.

During run 3, the ligand rotated around 180 degrees length-ways in the binding site with respect to the initial pose. The gradient here was evidently even less positive, resulting in a substantially less negative free energy of binding. The slow interconversion between orientations explains the lack of drift of the results with increasing simulation time - there are large barriers between orientations which prevent equilibrium sampling on the timescale of the simulations. To support

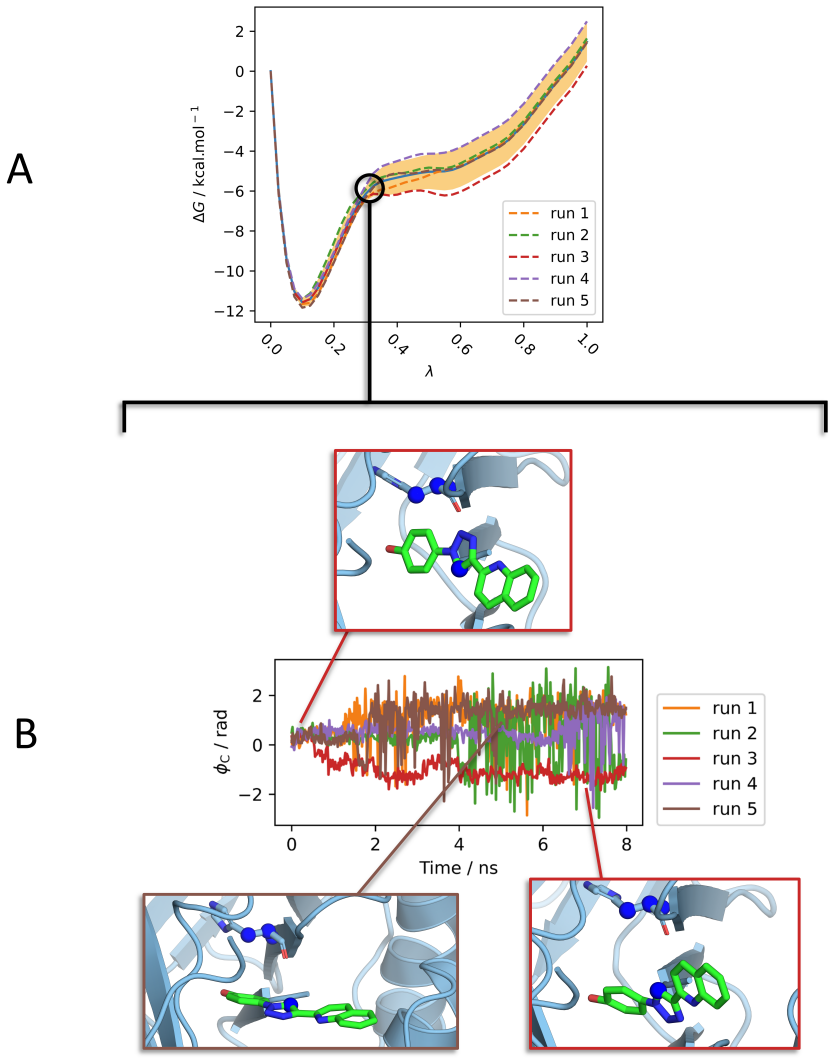


Figure 2.10: (A) The PMFs for the bound vanish stage for B2-o. These diverge around $\lambda = 0.325$. (B) The unrestrained Boresch DoF ϕ_C for B2-o at $\lambda = 0.325$, showing the presence of multiple slowly inter-converting orientations. Snapshots taken from run 3 (0.06 ns), run 5 (5.06 ns), and run 3 (7.22 ns), from left to right. The anchor points used to define ϕ_C are shown as spheres. Windows rendered with PyMOL.[167]

this explanation, the free energy of releasing B1 to B1-o in the decoupled state was calculated by scaling the strength of the orientational force constants with

λ, using the same set of λ windows as for the bound vanish stage. The free energy difference for releasing the restraints calculated by simulation was 4.92 $\pm$ 0.02 kcal mol^{-1} (MBAR 95 % C.I. estimate for a single simulation), in close agreement with the difference calculated by numerical integration by subtracting the two $\Delta G^{o}_{\text{Release}}$ terms (4.82 kcal mol^{-1}). This confirms that sampling a large increase in configurational space is not problematic when there are no barriers, at least when the growth is sufficiently slow - the issue is sampling the "rugged" configurational space at intermediate stages of decoupling. Comparison to prior work on the use of orientational restraints is given in Section A.13.

The offset introduced by the removal of orientational restraints may be reduced or removed through the use of the Hamiltonian-replica exchange (HREX) method. This is because in the fully decoupled state, there are no barriers to orientational rearrangements, and HREX has been found to improve sampling when the barriers in configurational space are low in at least one state.[199]. This allows free sampling of varied orientations, and mixing of these into the intermediate λ states using HREX may improve orientational sampling. However, it may not remove the bias; Lapelosa et al. found that convergence of their HREX ABFE calculations for a large and flexible ligand could only be achieved with orientational restraints.[200]

Performance of Common Default Force Constants

The performance of the Boresch restraints with the force constants fit to simulation were compared to those with all force constants set to the common defaults of 10 or 20 kcal mol^{-1} Å^{-2} [rad^{-2}], denoted by -10 and -20 (Table 2.1).[168,201,202] No significant differences in $\Delta G^{o}_{\text{Bound}}$ were found compared to when the force constants were fit to simulation, which was unsurprising given that the theoretical independence of the binding free energy with respect to the strength of restraints has been previously confirmed.[50,190] If there were any improvements in precision or increases in the rate of convergence with the force constants fit to simulation, these were not observed above the noise generated by other sources of error.

The only difference when default force constants were used was that several simulations crashed, very likely due to the collinearity of contiguous anchor points. While all simulations completed successfully when the force constants were fit to simulation, 1 λ window failed for B2-10 (this was rerun) and 12 failed for B3-10.

For B3-10, the minimum energy penalty arising from the restraints for setting a θ_A or θ_B to 0 or 180 degrees was approximately 5 k_BT, meaning that collinearity was relatively likely and crashes may have been anticipated. However, for B2-10, the minimum penalty for collinearity was approximately 10 k_BT. This makes crashes in the decoupled state highly unlikely, but when the ligand is still interacting with the protein it may become trapped in unusual orientations which distort the anchor points towards collinearity. This was the cause of the crash for B3-10 run 1: during the vanish $\lambda = 0.475$ window the simulation failed after θ_A approaches 0 (Figure 2.11). This occurred because the ligand became trapped underneath the terminal proline residue, resulting in θ_A approaching collinearity. A better restraint selection algorithm would have discounted restraints if the

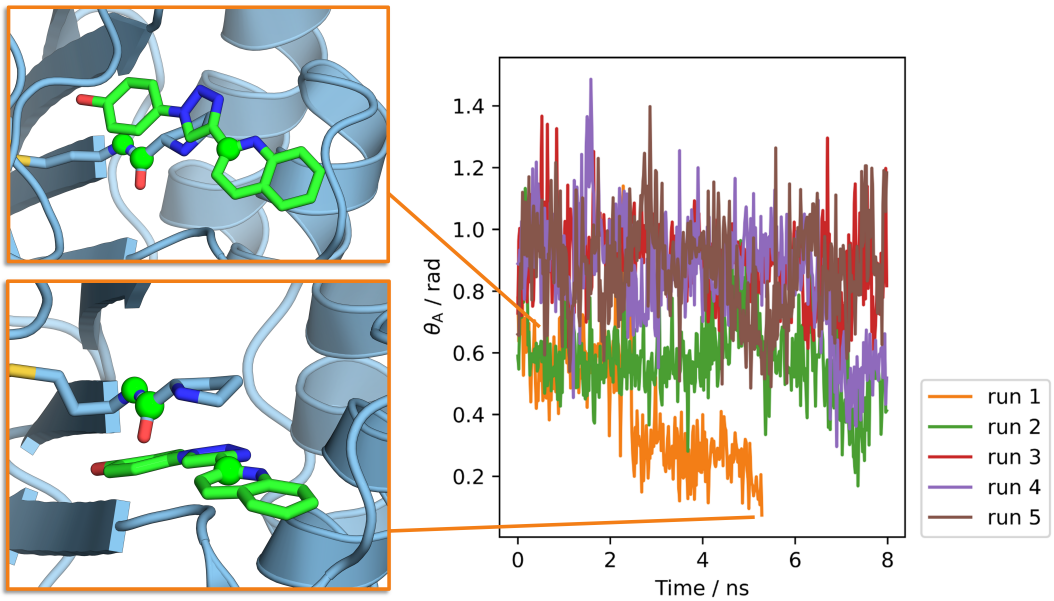


Figure 2.11: The restrained angle θ_A for B3-10 at $\lambda = 0.475$ during the vanish stage. For run 1, θ_A tended towards 0 as the ligand became trapped under the terminal proline and the simulation crashed. Anchor points used in the definition of θ_A are shown as spheres. Snapshots taken from run 1 at 1.12 ns (upper image) and 5.30 ns (lower image). Windows rendered with PyMOL.[167]

energy penalty from the restraints for collinearity was below some threshold, rather than only checking the equilibrium angles. Regardless, this highlights that instabilities with Boresch restraints can be an issue even when sensible restraint selections are made. It also illustrates that fitting the force constants to simulation (or at least using higher force constants) can produce more stable restraints.

Table 2.3: Bound Leg Contributions to $\Delta G^{o}_{\text{Bind}}$ with Multiple Distance Restraints[a]

	Restraints						
Contribution	**M-All**	**M-Rig**	**M-Rig-N***	**M-All-R**	**M-Hand-R**	**M-Hand**	**M-Hand-1**
$\Delta G_{\text{Rigid. Lig.}}$	-	0.50 ± 0.00	-	-	-	-	-
$\Delta G_{\text{Rigid. Recept.}}$	-	10.36 ± 0.09	-	-	-	-	-
$-\Delta G^{o}_{\text{Release}}$	15.68 ± 0.37	10.05 ± 0.17	9.97 ± 0.03	2.40	1.44	4.35 ± 0.27	5.82 ± 0.26
$-\Delta G_{\text{To Dist. Rest.}}$	-	-	-	16.66 ± 0.29	3.07 ± 0.24	-	-
$-\Delta G_{\text{Bound, Vanish}}$	-7.71 ± 1.05	-2.48 ± 1.25	-2.83 ± 0.90	-7.71 ± 1.05	1.58 ± 1.05	1.58 ± 1.05	1.81 ± 1.06
$-\Delta G_{\text{Bound, Discharge}}$	-13.54 ± 0.45	-13.23 ± 0.46	-12.97 ± 0.10	-13.54 ± 0.45	-11.89 ± 0.43	-11.89 ± 0.43	-12.99 ± 0.64
$-\Delta G_{\text{Bound, Restrain}}$	-3.67 ± 0.10	-1.33 ± 0.03	-1.35 ± 0.03	-3.67 ± 0.10	-0.02 ± 0.02	-0.02 ± 0.02	-0.28 ± 0.12
$-\Delta G_{\text{Rigid. Complex}}$	-	-8.31 ± 0.08	-	-	-	-	-
$\Delta G_{\text{Sym. Corr.}}$	-1.06	-1.06	-1.06	-0.65	-0.65	-0.65	-0.65
$\Delta G^{o}_{\text{Bound}}$	-10.31 ± 1.20	-5.50 ± 1.35	-8.24 ± 0.91	-6.51 ± 1.18	-6.47 ± 1.16	-6.64 ± 1.16	-6.30 ± 1.27

[a] Results for binding pose A, in kcal mol^{-1}. Uncertainties stated as 95 % confidence intervals based on the variance of 5 replicate runs assuming Gaussian distributions. The distance restraint dictionaries used for all protocols are given in Section A.14. -N indicates that the protocol was repeated with no intramolecular rigidification, -R indicates repetition with release to the single strongest distance restraint, and -1 indicates repetition with the flat-bottomed diameter set to 1 Å for all restraints. $\Delta G_{\text{To Dist. Rest.}}$ is the free energy of releasing the multiple distance restraints to a single distance restraint.* 10 replicate runs were used. The convergence of $-\Delta G^{o}_{\text{Release}}$ with respect to the standardstatecorrection parameters was confirmed (Section A.15).

2.4.2 Multiple Distance Restraints

With Intramolecular Rigidification

Multiple distance restraints provide a framework which is free from the inherent instabilities of Boresch restraints, and which allows more complete restraint of the ligand than Boresch restraints alone. However, their naïve application renders the ABFE framework theoretically inexact. To illustrate this, harmonic distance restraints were applied to every heavy atom in the ligand and their lowest variance unique partner heavy atom in the protein (protocol M-All, 22 receptor-ligand distance restraints). As expected, this produced an excessively negative $\Delta G^{o}_{\text{Bound}}$ estimate, in excess of 3 kcal mol^{-1} more negative than most of the Boresch restraints, indicating that $\Delta G_{\text{Preorg.}}$ was large (Table 2.3). The distance restraint dictionaries used for all protocols are given in Section A.14.

This was contrasted with the rigorous multiple distance restraints scheme with intramolecular rigidification of anchor points (Figure 2.12), referred to as M-Rig. Anchor points were selected as for M-All, except that only anchor points in the phenol moiety of the ligand were used in order to avoid the tight restraints restricting rotatable bonds in the ligand; this would likely require an enhanced sampling approach, such as umbrella sampling,[51] to restrain bond rotation before applying the restrictive intermolecular restraints. This resulted in 7 receptor-ligand distance restraints. The intramolecular restraints were implemented as

harmonic distance restraints between all pairs of anchor atoms within the given molecule, with force constants of 75 kcal mol^{-1} Å^{-2}. For simulations where $-\Delta G^{o}_{\text{Release}}$ was calculated using Equation 2.10, it was confirmed that the estimate had converged with respect to the number of points used for numerical integration (Section A.15). This scheme gave $\Delta G^{o}_{\text{Bound}} = -5.50 \pm 1.35$ kcal mol $^{-1}$, in good agreement with the previous Boresch calculations, providing proof-of-concept of a rigorous implementation of multiple distance restraints.

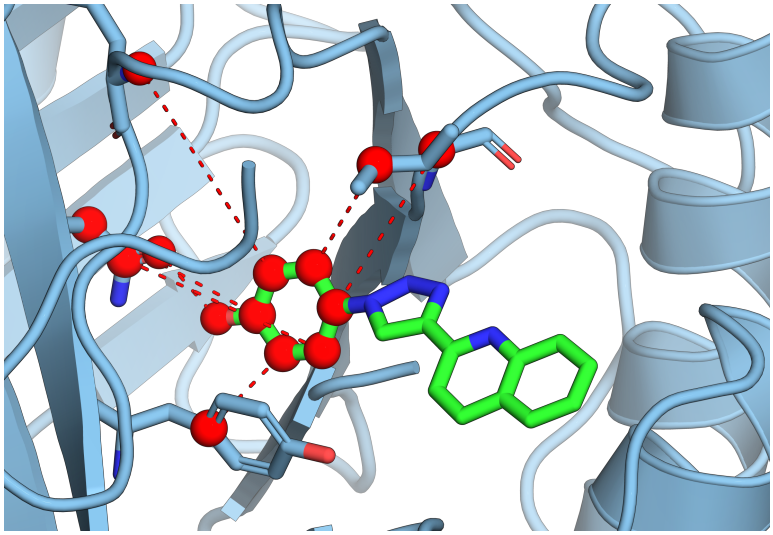


Figure 2.12: Anchor points (red) used for the M-Rig restraints. Intermolecular restraints are shown as red dashed lines, while intramolecular restraints not shown. Rendered with PyMOL.[167]

The magnitude of $\Delta G_{\text{Preorg.}}$ can be roughly estimated as $\Delta G_{\text{Rigid. Recept.}} + \Delta G_{\text{Rigid. Lig.}} - \Delta G_{\text{Rigid. Complex}} = 2.55 \pm 0.17$ kcal mol^{-1} (see Section A.18). This is in excellent agreement with the difference observed when the procedure was repeated without intramolecular restraints (M-Rig-N), which was 2.74 ± 1.63 kcal mol^{-1}, providing evidence that this is the main source of the error observed when multiple distance restraints are applied in a non-rigorous manner.

The intramolecular restraints are required to prevent distortion of the intramolecular degrees of freedom by the intermolecular restraints, which would otherwise introduce error into the $\Delta G^{o}_{\text{Release}}$ calculated using Equation 2.10. The strength of the intramolecular restraints required to eliminate this error is dependent on how how permissive the intermolecular restraints are; highly restrictive intermolecular restraints necessitate aggressive rigidification.

For the M-Rig calculations, strong intermolecular restraints were used and therefore strong intramolecular restraints were also required. The large value of $\Delta G_{\text{Preorg.}}$ which resulted from the strong intermolecular restraints allowed us to show that that intramolecular rigidification eliminates this error, by comparison of M-Rig and M-Rig-N. However, the strong intermolecular restraints resulted in a significant additional computational cost during the rigidification stage. The λ protocols for the rigidification stages could likely have been optimised to reduce the total number of windows to approximately 25 (Section A.17), but this would still result in an additional computational cost comparable to the vanishing stage. Hence, we would not recommend multiple distance restraints with intramolecular rigidification when using strong intermolecular restraints. Instead, this scheme is expected to perform best with more permissive intermolecular restraints. An efficient implementation may be as follows:

1. Select flat-bottomed, rather than harmonic, intermolecular restraints and select the flat-bottomed regions so to include almost all distances sampled during the unrestrained simulation of the fully-interacting complex
2. Perform the rigidification simulations simultaneously starting from States 3 and 4. Monitor the convergence of $\Delta G_{\text{Preorg.}}$ as the intramolecular restraint strength is increased and stop the calculations once convergence is achieved
3. Calculate $\Delta G^{o}_{\text{Release}}$ with Equation 2.10 based on the trajectory for the most strongly rigidified version of State 4

In cases where such flat-bottomed restraints are used, and there is little rearrangement of the ligand or binding site upon decoupling, $\Delta G_{\text{Preorg.}}$ would be expected to converge immediately and the above scheme would reduce to the naïve multiple distance restraints scheme. When this is not the case, the naïve scheme would be expected to yield incorrect results, while the above scheme should remain correct.

With Release to Single Distance Restraint

Based on the approach taken with DBC restraints,[160,161] the free energy of releasing all but the strongest distance restraint after decoupling was calculated for M-All (M-All-R). Because a single distance restraint does not couple the internal and external degrees of freedom of the protein and ligand, this allows the rigorous calculation of $-\Delta G^{o}_{\text{Release}}$ using Equation 2.11, removing the requirement for intramolecular restraints.

In contrast to M-All, M-All-R produced a $\Delta G^{o}_{\text{Bound}}$ of -6.51 ± 1.18 kcal mol^{-1}, in good agreement with the Boresch calculations. There appeared to be a slight drift in $\Delta G_{\text{To Dist. Rest.}}$ with time (Figure A.20) towards more negative values, which may be due to the requirement for the ligand to sample all points on the surface of a sphere surrounding the protein anchor point upon decoupling. This may be improved by releasing to a centre-of-mass restraint centred on the binding site, reducing the volume which must be sampled.[161] In addition, scaling the restraint potential differently between the end points may improve convergence; instead of the λ^5 scaling used here, a soft bond stretch potential would likely perform well for the removal of these harmonic restraints.[203]

This process was repeated using a significantly more permissive multiple distance restraint scheme, where 4 flat-bottomed distance restraints were selected to mimic the 4 protein-ligand hydrogen bonds shown in Figure 2.1. This scheme is denoted M-Hand-R as the anchor points were selected by hand, although automated selection to match hydrogen bonds would be straightforward. The radius of the flat-bottomed region was selected to be as small as possible without the restraints engaging at any point during the restraint fitting simulation, ensuring that $\Delta G_{\text{Bound, Restrain}} \approx 0$. The force constants for the half-harmonic potentials were 40 kcal mol^{-1}. This again produced a $\Delta G^{o}_{\text{Bound}}$ in good agreement with the Boresch results and M-Rig (-6.47 ± 1.16 kcal mol^{-1}), indicating that the relatively permissive restraints sufficiently restricted orientational sampling. Furthermore, no substantial drift in the estimate with simulation time was observed (Figure A.21).

Compared to the naïve and Boresch schemes, this scheme required a single additional release stage, which was relatively computationally affordable. Furthermore, computational cost could be further reduced with optimisation of the λ schedule for M-Hand-R, because excellent overlap was achieved between many non-consecutive λ windows (Section A.17). The overlap matrices for the vanishing stages were very similar regardless of the restraint scheme, and the number of windows required for the vanishing stages could not have been reduced below approximately 30 in any case (see Figure A.24 as a representative example). In contrast, for both M-Hand-R and M-Rig-R, it appears that 10 windows would be sufficient, or even fewer in the case of M-Hand-R. For M-Hand-R. Convergence appears to be achieved after around 1 ns sampling per window (not including the 1 ns equilibration), which seems broadly similar to the bound vanish stage

results (not including the 3 ns equilibration). Therefore, when the intermolecular restraints are not extremely numerous and strong, the additional cost associated with the release stage with an optimised λ schedule is expected to be substantially less than a third of the vanish stage.

Although there was no evidence for improved convergence of the decoupling simulations withM-All-R over the Boresch schemes, this might be observed in other systems with highly flexible ligands, where ligand conformational sampling is the dominant source of uncertainty. However, the use of Boresch restraints in combination with RMSD-based restraints on the conformation of the ligand may prove similarly effective.[51]

This scheme is in some ways similar to the DBC scheme, in that a complex restraint (set of restraints) involving many degrees of freedom is released to a single harmonic restraint to allow the calculation of $\Delta G^{o}_{\mathrm{Release}}$. While the DBC restraint is attractively simple - it consists of a single flat-bottomed restraint on the RMSD of a subset of ligand coordinates in the frame of reference of the binding site - multiple distance restraints schemes offer finer control over the strength of restraints applied to different subsections of the system. This may be beneficial, for example in the case of a large and flexible ligand where only part of the ligand interacts strongly with the receptor. If the coordinates of all ligand heavy atoms were included in the DBC restraint, then the DBC coordinate would show very wide fluctuations during simulations of the bound state. Fitting the DBC restraint to encompass the 95th percentile of sampled DBC coordinates would result in a very weak restraint on all sections of the ligand, which may result in sampling issues. Flat-bottomed multiple distance restraints could be fit in a similar way, such that the flat-bottomed regions encompassed almost all distances measured during a simulation of the fully-interacting complex. In contrast to the DBC restraints fit to all ligand heavy atoms, the multiple distance restraints fit to all heavy atoms would closely restrict the portion of the ligand which interacts strongly with the receptor, while allowing large fluctuations in the flexible portion which does not. While multiple distance restraints require many more parameters than the DBC restraint, they can be automatically selected from a simulation of the fully interacting complex using algorithms described in this work. Furthermore, if the user opted not to run such a simulation for restraint selection, distance restraints could be intuitively selected to match receptor-ligand interactions, and reasonable default parameters could be chosen. It may be challenging to select a reasonable flat-bottomed region for a DBC restraint without a trajectory.

Averaging over the rigorous multiple distance restraint schemes (M-Rig, M-All-R, and M-Hand-R), the mean $\Delta G^{o}_{\text{Bound}}$ result for pose A was -6.16 ± 1.43 kcal mol^{-1}. This was in good agreement with the mean result for pose A calculated with Boresch restraints (-6.28 ± 0.49 kcal mol^{-1}, average of B2, B3, B1-P, and B3-P, and B2-10). Ignoring the contribution to the free energy of binding from pose B, for which no calculations were performed with multiple distance restraints, the free energies of binding would have been -9.24 ± 1.44 kcal mol^{-1} and -9.36 ± 0.37 kcal mol^{-1} for multiple distance restraints and Boresch restraints, respectively.

With a Large Flat-Bottomed Region

Finally, a non-rigorous implementation of multiple distance restraints was tested, based on the assumption of no coupling between internal and external degrees of freedom in the limit of weak restraints. The schemes tested were M-Hand, a repetition of M-Hand-R without releasing to a single distance restraint, and M-Hand-1, a repetition of M-Hand with all flat-bottomed diameters set to 1 Å, a reduction from the average diameter of 2.7 Å for M-Hand-R.

A $\Delta G^{o}_{\text{Bound}}$ of -6.64 ± 1.16 kcal mol $^{-1}$ was obtained for M-Hand. This was not significantly different to M-Hand-R or M-Rig, suggesting that the approximations made by the scheme were minor in this case. The results for M-Hand and M-Hand-1 (-6.30 $\pm$ 1.27 kcal mol^{-1}) were very similar, despite the slightly more restrictive restraints. This shows that it is possible to obtain equivalent free energies using the naïve distance restraints scheme and rigorous schemes, so long as the coupling between the internal and relative external DoF of the protein and ligand is weak when the ligand is decoupled. However, increasingly negative binding free energies would be expected as more restrictive restraints are used and with increasing differences between the *apo* and *holo* conformations of the binding site, and the magnitude of the error may be be difficult to predict.

2.5 Conclusion

The free energies of binding for MIF-180 to MIF calculated with varied sets of Boresch restraints were fairly self-consistent and in good agreement with experiment. However, removal of the orientational restraints produced estimates which were up to approximately 4 kcal mol^{-1} more negative, likely because the ligand failed to sample all relevant orientations as the LJ interactions were removed. It was found that the calculations were highly sensitive to the sampling of water

in the binding site at intermediate stages of vanishing, and that under-hydration of the binding site during these stages over as few as 3 λ windows could shift the binding free energy by over 3 kcal mol^{-1} towards more favourable binding. This illustrates the importance of thorough water sampling in alchemical free energy calculations. It would be interesting to test whether HREX, enhancing water sampling with Monte Carlo water moves in either the the μVT and NPT ensembles, or longer simulation times around problematic λ-windows can improve the convergence of binding free energies for this system.[195–197]

Instabilities inherent to the Boresch restraint scheme were highlighted by the failure of several simulations, even with sensible restraint parameters which imposed a minimum energy penalty of around 10 $k_{\mathrm{B}}T$ for collinear anchor points. The use of multiple distance restraints offers an alternative restraint scheme which lacks the instabilities of Boresch restraints and may improve convergence during decoupling by allowing greater restriction of ligand movements. The theory of multiple distance restraints was discussed and a rigorous implementation of the multiple distance restraints scheme was proposed. This utilised intramolecular rigidification of anchor points to prevent coupling between the internal and external DoF. This was shown to produce free energy estimates in good agreement with the Boresch restraints, at least within the large uncertainties encountered with this system (Figure 2.13). This scheme incurred a substantial additional computational cost over Boresch restraints because aggressive rigidification was required to counter the strong intermolecular restraints, but the scheme may offer benefits where more permissive intermolecular restraints are used.

Another rigorous implementation of the multiple distance restraints scheme was tested, which involved releasing the multiple restraints to a single distance restraint after decoupling. In contrast to calculations performed entirely without orientational restraints, this scheme produced free energy estimates in good agreement with the Boresch restraints scheme, at a reduced computational cost compared to the scheme employing intramolecular restraints. The additional computational cost compared to Boresch restraints is expected to be less than a third of the vanish stage unless very many strong intermolecular restraints are used. Additional costs associated with rigorous multiple distance restraints schemes may be compensated for by convergence benefits in some systems, although that was not demonstrated in this work. The mean $\Delta G^{o}_{\mathrm{Bound}}$ calculated based on pose A with rigorous implementations of multiple distance restraints (-6.16 ± 1.43 kcal mol^{-1}) was in close agreement with the mean result calculated with Boresch restraints (-6.28 ± 0.49 kcal mol^{-1}).

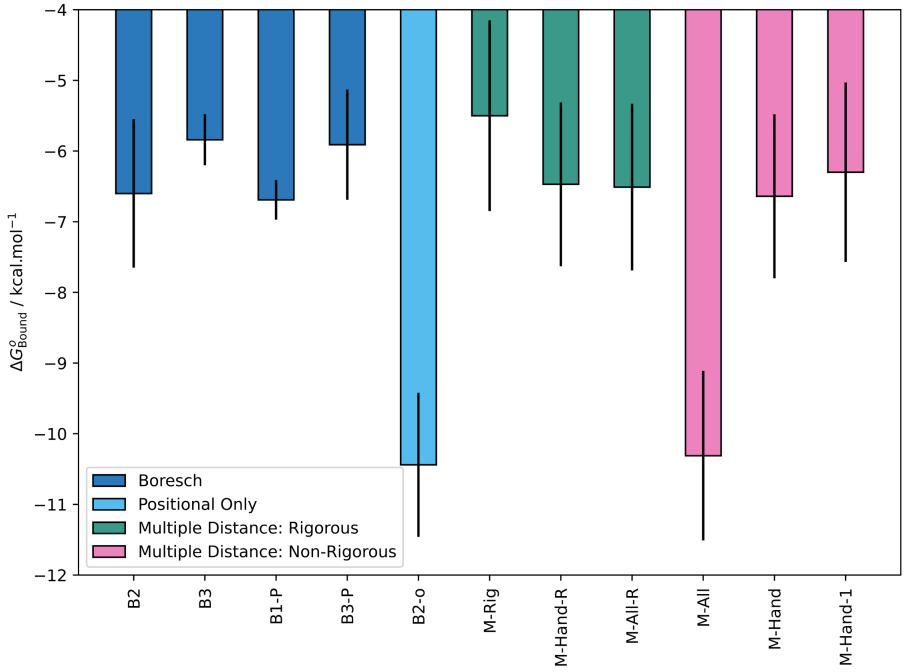


Figure 2.13: Summary of results for $\Delta G^o_{\text{Bound}}$ obtained for binding pose A using a variety of restraints schemes. Uncertainties are 95 % confidence intervals based on the variance of 5 replicate runs, assuming Gaussian distributions. Results for B1 have been omitted as they appeared to be more susceptible to water sampling issues. For B3-P, run 1 was excluded from the average due to under-sampling of water in binding site during the vanishing stage.

Finally, a non-rigorous implementation of the multiple distance restraints scheme was tested, which relied on the assumption of negligible coupling between the internal and external DoF. With strong restraints, this assumption was violated and excessively negative free energies of binding were calculated, but quantities close to the rigorous estimates were obtained with sufficiently permissive restraints. However, it may be difficult to predict the magnitude of the error introduced by this implementation.

The dominant source of uncertainty in these calculations appeared to be the sampling of water, and no convergence benefits were demonstrated with multiple distance restraints over Boresch restraints. Future work may investigate whether convergence benefits are observed in systems where ligand conformational sampling is the dominant source of uncertainty. Further comparison against a wider range of restraint schemes over a variety of systems is also the subject of future work.

In summary, this work demonstrates that absolute binding free energies equivalent to those obtained with Boresch restraints can be calculated using multiple distance restraints. This framework is in principle more stable and may offer convergence benefits during decoupling, although this must be balanced against the additional computational cost incurred by the extra stages required for the rigorous schemes. However, a multiple distance restraints scheme utilising many flat-bottomed potentials to closely restrain the ligand with minimal disturbance of the interacting system may allow the restraining stage to be neglected, as with the DBC restraint,[161] while improving the convergence of the decoupling stages when the ligand is flexible. This work discussed the theory and demonstrated proof-of-principle of rigorous multiple distance restraint schemes; future work may investigate whether the scheme can offer performance benefits.

Chapter 3

Automated Adaptive Absolute Binding Free Energy Calculations

In Chapter 2, various receptor-ligand restraint schemes were investigated and a robust algorithm for the selection of Boresch restraint parameters was proposed. These were implemented in the BioSimSpace simulation package, and tutorials were provided to allow practitioners to easily incorporate them in their workflows.[182,204] While the recommendations given and methods proposed may aid the robust deployment of ABFE calculations at scale, many other aspects of ABFE calculations were not addressed and a fully-automated workflow was not developed.

To lower the often prohibitive computational cost of ABFE calculations, efficient automated workflows are required to reduce both computational cost and human intervention. In this chapter, a fully automated ABFE workflow based on the automated selection of λ windows, the ensemble-based detection of equilibration, and the adaptive allocation of sampling time based on inter-replicate statistics, is presented. It is found that the automated selection of intermediate states with consistent overlap is rapid, robust, and simple to implement. Robust detection of equilibration is achieved with a paired t-test between the free energy estimates at initial and final portions of an ensemble of runs. Reasonable default parameters are determined for all algorithms and show that the full workflow produces equivalent results to a nonadaptive scheme over a variety of test systems, while often accelerating equilibration. The complete workflow is implemented in the open-source package A3FE (github.com/michellab/a3fe).

The remainder of the chapter is included unmodified as published:

Clark, F.; Robb, G. R.; Cole, D. J.; Michel, J. Automated Adaptive Absolute Binding Free Energy Calculations. *J. Chem. Theory Comput.* **2024**, *20*, 7806–7828.

3.1 Introduction

The binding affinity of a drug to its target is an important quantity in early-stage drug discovery. Its quick and accurate prediction would allow the wasteful synthesis of weakly-binding molecules to be avoided. Substantial progress has been made towards this goal, with rigorous "alchemical" free energy calculations based on molecular dynamics or Monte Carlo sampling recommended as the optimal method.[53,54,58] Alchemical relative binding free energy (RBFE) calculations, which evaluate the relative difference in binding affinity between structurally related molecules, are now routinely applied in the hit-to-lead and lead optimisation stages of drug discovery.[205] However, standard RBFE calculations are limited to structurally similar molecules binding to the same target with the same binding mode.[114]

Alchemical absolute binding free energy (ABFE) calculations escape these constraints through their more general formulation.[49,50] RBFE calculations are based on a thermodynamic cycle where two ligands are inter-converted in solution and while bound to their target; ABFE calculations involve removing a single ligand's intermolecular interactions in both environments.[206] This makes ABFE calculations applicable to a wider class of problems,[55,56] including accurately ranking diverse molecules in the final stages of high-throughput virtual screening.[60,61]

Unfortunately, the application of ABFE calculations is limited by their high computational cost. ABFE calculations require a relatively large modification of the system (complete removal of ligand intermolecular interactions) which increases the thermodynamic length around an ABFE cycle compared to an RBFE cycle.[207,208] This necessitates more sampling for a result of equal precision. ABFE calculations are also particularly susceptible to sampling issues such as slow rehydration of the binding site and side-chain rearrangement.[209] Calculating the relative binding free energy between a pair of similar ligands using a default ABFE protocol would be around an order of magnitude more expensive than with a default RBFE protocol,[52,61] and would likely be less precise.

Clearly, ABFE calculations must become more efficient to realise their potential. A promising class of methods yield relative binding free energies between structurally dissimilar ligands while aiming to avoid the sampling issues associated with "emptying"' the binding site.[153,183,210–212] Approaches combining active learning of affinity predictions with free energy calculations have demonstrated substantial efficiency improvements by reducing the number of simulations required to identify potent binders.[213–217] Non-equilibrium calculations involving "swarms" of short switching trajectories may decrease wall-clock time in some cases.[209,218–220] A plethora of enhanced sampling techniques have been applied to alchemical free energy calculations, sometimes producing dramatic improvements in accuracy and efficiency.[106,221–235] However, many of these techniques require system-specific tuning, yield highly system-specific benefits,[209] may require frequent communication between otherwise independent replicas, and may even degrade performance.[193]

There are also areas in the standard ABFE cycle where substantial human and computer time is often wasted through manual set-up or poor defaults, and where automation and adaptive algorithms may offer substantial efficiency improvements.[236–239] This avenue for efficiency improvement is much less well-explored than the approaches mentioned above and may be particularly beneficial for ABFE calculations, which vary widely in terms of sampling requirements and equilibration and convergence behaviour. Here, we investigate three of these areas.

The first area is the spacing of intermediate states. It is usually impossible to obtain an accurate estimate of a free energy difference with alchemical methods by sampling only at the end-states of interest. For overlap-based free energy estimators such as the Bennett Acceptance Ratio (BAR) or the Zwanzig equation,[121] this is because sampling at one end-state almost never produces samples with non-negligible probability at the opposite end state. For thermodynamic integration (TI), errors are introduced because the potential of mean force (PMF) with respect to an alchemical variable (which smoothly interpolates the end state Hamiltonians) is poorly approximated. The problem is solved in both cases by sampling at intermediate states. The variable controlling the interpolation of intermediate states is usually called λ, and intermediate states may be referred to as λ windows.

Optimal intermediate states would minimise the standard error of the overall free energy estimate for a given total sampling time. However, selecting these would require knowledge of the statistical inefficiencies and divergences of the probability distributions along all possible paths between end states.[208,240] It should be noted that the optimal path is influenced by the relative efficiencies of sampling at intermediate states, and is therefore dependent on the sampling methodology as well as the Hamiltonians. It appears challenging to obtain this information without costly simulations which would reduce overall efficiency. Nevertheless, a number of approximate approaches have been developed.

Blondel,[241] and later Pham and Shirts,[242] derived schemes to select minimum-variance pathways between end-state Hamiltonians. Their derivations assumed infinitely close spacing of intermediate states along the pathway and ignored statistical inefficiencies, which reduced the optimised schedules' efficiencies. Reinhardt and Grubmüller also ignored correlation times but accounted for the finite spacing of intermediate states.[243] There have been attempts to reduce correlation times by smoothing energy barriers,[232,244] and by constraining the selected path to avoid problematic regions of parameter space.[245,246]

A particularly simple approach is to optimise only the number and spacing of states along a pre-determined path. We limit ourselves to this approach here. Most approaches to this problem involve short initial simulations, although König et al. developed a simple metric to estimate the required number of intermediate states based on the energy difference between minimised end-state structures.[247] Several methods have been proposed that space intermediate states by equal thermodynamic distances. This has been shown to minimise total variance of the free energy estimate for any unbiased estimator, assuming infinitely close spacing of states.[248] Methods which do,[249,250] and do not,[251,252] account for the autocorrelation of samples have been suggested. Methods based on heuristics have also been proposed.[253,254] Recently, Zhang et al. built models to predict non-local phase-space overlap, which were used to select λ-schedules with equal phase space overlap between adjacent windows.[255] However, the use of any protocol for intermediate state selection is still rare in alchemical free energy calculations, and extensive investigation of optimal selection protocols and potential benefits is lacking.

The second area is the allocation of sampling time. For a free energy calculation using information from two states, Bennett noted that the equal allocation of sampling time between states is at worst half-optimal.[121] This is because the allocation of 1 hour of sampling to each state (2 hours total sampling) cannot yield a worse estimate than the optimal allocation of a total of 1 hour of sampling between both states, given that the variance of the estimate decreases monotonically with the number of samples from each state. However, an absolute binding free energy calculation uses samples from many (N) states. The lower bound on the estimation efficiency with equal sampling time then falls to $1/N$ of the optimal allocation. Therefore large efficiency improvements may be made by the optimal allocation of sampling time when few states have a much higher computational cost per independent sample.

Sun et al. adaptively allocated simulation time during absolute hydration free energy calculations and an RBFE calculation.[256] Time allocation was based on the time derivatives of the variance of the free energy estimates between adjacent λ windows obtained with the Bennett Acceptance Ratio for a single run.[121] However, this did not substantially improve the efficiency of the RBFE calculation due to relatively short and similar autocorrelation times of the perturbed energy differences between windows. A similar approach is more likely to benefit ABFE calculations, which are more susceptible to dramatic variations in correlation time between windows as a result of sampling issues. Indeed, Mendoza-Martinez et al. demonstrated that the equilibration of ABFE calculations could be accelerated by allocating sampling time based on the uncertainties of the inter-window free energy changes.[165] In contrast to Sun et al., errors were calculated from the difference between replicate runs, which is a robust method for uncertainty quantification.[111,151,257–260] However, this protocol lacks a rigorous derivation, and no automated implementation is available. While this work focuses on the adaptive allocation of sampling time within single ABFE calculations, Li et al. showed how adaptively allocating simulation time to individual calculations within a network could substantially improve statistical precision.[239]

The final area is the detection of equilibration and convergence. Intermediate state simulations in free energy calculations are usually started from coordinates sampled with the ligand fully interacting. These coordinates may have a very low probability under the equilibrium distributions of other states, producing initial transients in the estimated free energy changes as the system relaxes. It is standard practice to discard samples from these equilibration (burn-in) periods to reduce bias in the final free energy estimate.[114] Convergence refers to the

approach of the free energy estimate to its asymptotic (infinite sampling) value. Methods to assess equilibration and convergence are essential to increase confidence that finite-time simulations are representative of the underlying stationary distributions. However, it is never possible to state this with certainty.[261]

In this work, free energy estimates are termed "unequilibrated" if they show clear bias compared to results from longer simulations, or if they change significantly with additional sampling time at late times. Otherwise, they are termed "equilibrated". This definition is consistent with its use in the field, but does not guarantee the removal of bias compared to the true infinite sampling value. We refer to estimates as "unconverged" if there is evidence that they are unstable or inaccurate compared to their infinite sampling values. Hence, "unequilibrated" estimates are always "unconverged", and "equilibrated" estimates may still be "unconverged". This contrasts with frequent practice in the literature, where "equilibrated" estimates are always described as "converged".

In free energy calculations, the most well-known equilibration detection protocols are those of Chodera and Yang et al..[109,262] Yang et al. monitored the behaviour of the reverse cumulative average of the free energy gradient. Contamination by non-equilibrated data was detected by deviation from normality. Chodera proposed a non-parametric method which selects the unequilibrated region by maximising the effective sample size of the production region. Both methods are based on the data from a single run. As a result, both are susceptible to erroneously discarding the majority of simulation data when the system becomes trapped in a local minimum. Convergence is typically assessed using uncertainties obtained from a single run (after block-averaging or sub-sampling to remove correlation) or, more rigorously, from differences between repeated runs.[258]

Despite similarity of the underlying problem, different diagnostics are used in the Markov Chain Monte Carlo (MCMC) literature.[263] The most common method for assessing sampling seems to be the Gelman-Rubin diagnostic,[264] which effectively compares the variances estimated within and between MCMC chains. If individual chains (simulations) are initially overdispersed and become trapped in local minima, the inter-chain variances will be greater than the intra-chain variances. This manifests as a Gelman-Rubin $\hat{R} > 1$, indicating that individual chains have failed to converge to the same stationary distribution. MCMC equilibration detection methods include the Geweke test,[265] which checks for a significant difference

between the means of the initial and final portions of the data. These diagnostics have occasionally been used in molecular simulation (for example by García and Hasse) but,[266] to our knowledge, they are yet to be exploited in free energy calculations.

Here, we attempt to improve the efficiency of ABFE calculations by developing and evaluating automated protocols in each of the areas mentioned: the spacing of intermediate states, the allocation of sampling time, and the detection of equilibration. Where possible, we use ensemble-based metrics to increase robustness. We have sought to minimise user time, in addition to computational time, by implementing our protocols in an easy-to-use open-source Python package, A3FE, which is itself based on BioSimSpace and SOMD.[182,267]

3.2 Theory

3.2.1 The Optimum Allocation of Resources Along a Path

In this work, we assume that our path through alchemical space is already defined (by our choice of soft-core potentials, etc.), and that progress along this path is controlled by a single interpolating variable, λ. We want to determine the optimal allocation of sampling time along this path which minimises the uncertainty in the final free energy estimate for a constant total sampling time.

It is convenient to start by assuming infinitely close spacing of intermediate states, which makes the uncertainties of estimators such as TI, Zwanzig, and BAR equivalent (see section B.1). Under this assumption, the estimated free energy change between $\lambda = 0$ and 1 is,[116]

$$\Delta\widehat{F}_{0,1} = \int_0^1 \left\langle \frac{\partial H}{\partial \lambda} \right\rangle_\lambda \mathrm{d}\lambda. \tag{3.1}$$

The Hamitonian, $H(\lambda, \mathbf{x})$, is dependent on λ and on the positions and momenta of all particles in the system, $\mathbf{x}$, but is written H for brevity. The free energy change calculated is only an estimate (denoted by the hat operator) because we cannot fully enumerate the desired ensemble. Instead, $\langle ... \rangle_\lambda$ denotes an average obtained from statistical sampling (molecular dynamics) at λ and taken to be equivalent to the ensemble average (the ergodic hypothesis). The variance of the

estimated free energy change is

$$\sigma^2(\Delta\widehat{F}_{0,1}) = \int_0^1 \sigma^2\left(\left\langle\frac{\partial H}{\partial\lambda}\right\rangle_\lambda\right)\mathrm{d}\lambda \tag{3.2}$$

$$= \int_0^1 \frac{1}{n_\lambda}\sigma^2\left(\frac{\partial H}{\partial\lambda}\right)\mathrm{d}\lambda \tag{3.3}$$

$$= \int_0^1 \frac{g_\lambda}{t_\lambda}\sigma^2\left(\frac{\partial H}{\partial\lambda}\right)\mathrm{d}\lambda, \tag{3.4}$$

where n_λ is the number of uncorrelated samples obtained at λ, $\sigma^2\left(\left\langle\frac{\partial H}{\partial\lambda}\right\rangle_\lambda\right)$ is the variance of the mean gradient of the Hamiltonian with respect to λ, $\sigma^2\left(\frac{\partial H}{\partial\lambda}\right)$ is the variance of the gradient of the Hamiltonian with respect to λ, $g_\lambda = \frac{t_\lambda}{n_\lambda}$ is the statistical inefficiency at λ, t_λ is the sampling time at λ, and we have assumed independent sampling at different values of λ. Defining the fractional sampling time as $\pi_\lambda = \frac{t_\lambda}{t_{\text{Tot.}}}$, where $t_{\text{Tot.}}$ is the total sampling time for the transformation, yields,

$$\widehat{\sigma^2(\Delta F_{0,1})} = \frac{1}{t_{\text{Tot.}}}\int_0^1 \frac{g_\lambda}{\pi_\lambda}\sigma^2\left(\frac{\partial H}{\partial\lambda}\right)\mathrm{d}\lambda \tag{3.5}$$

The total fractional sampling time must equal 1. $\sigma^2(\Delta F_{0,1})$ can be minimised subject to this constraint by creating the Lagrangian function

$$\mathcal{L} = \frac{1}{t_{\text{Tot.}}}\int_0^1 \frac{g_\lambda}{\pi_\lambda}\sigma^2\left(\frac{\partial H}{\partial\lambda}\right)\mathrm{d}\lambda - \gamma\left(1 - \int_0^1 \pi_\lambda \mathrm{d}\lambda\right) \tag{3.6}$$

and setting its derivative with respect to π_λ to 0 for all λ

$$\frac{\partial\mathcal{L}}{\partial\pi_\lambda} = 0 = -\frac{g_\lambda}{t_{\text{Tot.}}\pi_\lambda^2}\sigma^2\left(\frac{\partial H}{\partial\lambda}\right) + \gamma \tag{3.7}$$

$$\pi_\lambda^2 = \frac{g_\lambda}{t_{\text{Tot.}}\gamma}\sigma^2\left(\frac{\partial H}{\partial\lambda}\right) \tag{3.8}$$

$$\pi_\lambda = \sqrt{\frac{g_\lambda}{t_{\text{Tot.}}\gamma}}\sigma\left(\frac{\partial H}{\partial\lambda}\right). \tag{3.9}$$

The integrals over λ have disappeared because this condition must be true for all λ. This approach is similar to that of Sun et al. but is based on TI rather than BAR.[256] Equation 3.8 rearranges to $-\frac{\gamma}{t_{\text{Tot.}}} = \text{Const.} = -\frac{g_\lambda}{t_\lambda^2}\sigma^2\left(\frac{\partial H}{\partial\lambda}\right)$, where $-\frac{g_\lambda}{t_\lambda^2}\sigma^2\left(\frac{\partial H}{\partial\lambda}\right)$ is the derivative of the variance of the estimated free energy change (Equation 3.4) with respect to t_λ at a fixed value of λ. This means that the time derivative of the variance of the mean is a constant for all λ when the allocation

of sampling time is optimum, as discussed by Sun et al..[256] The intuition is that if there are states with higher time derivatives of the variance, then sampling time would have been better allocated to these than states where the additional sampling time reduced uncertainty less.

The Lagrange multiplier, γ, can be determined by substituting Equation 3.9 into the condition $\int_0^1 \pi_\lambda \mathrm{d}\lambda = 1$:

$$\sqrt{\gamma} = \frac{1}{\sqrt{t_{\text{Tot.}}}} \int_0^1 \sqrt{g_\lambda}\sigma \left(\frac{\partial H}{\partial \lambda} \right) \mathrm{d}\lambda. \tag{3.10}$$

Hence the choice of fractional sampling time allocation which minimises the overall uncertainty is

$$\pi_\lambda = \frac{\sqrt{g_\lambda}\sigma \left(\frac{\partial H}{\partial \lambda} \right)}{\int_0^1 \sqrt{g_\lambda}\sigma \left(\frac{\partial H}{\partial \lambda} \right) \mathrm{d}\lambda}. \tag{3.11}$$

Substituting this into Equation 3.5 to obtain the minimum variance of the free energy change estimates provides a limit for the minimum variance of the overall free energy change,

$$\sigma^2(\Delta \widehat{F}_{0,1}) \geq \sigma^2_{\text{Min.}}(\Delta \widehat{F}_{0,1}) \tag{3.12}$$

$$\sigma^2(\Delta \widehat{F}_{0,1}) \geq \frac{1}{t_{\text{Tot.}}} \left(\int_0^1 \sqrt{g_\lambda}\sigma \left(\frac{\partial H}{\partial \lambda} \right) \mathrm{d}\lambda \right)^2 \tag{3.13}$$

$$\sigma^2(\Delta \widehat{F}_{0,1}) \geq \frac{1}{t_{\text{Tot.}}} \left(\int_0^1 \sqrt{t_\lambda}\sigma \left(\left\langle \frac{\partial H}{\partial \lambda} \right\rangle_\lambda \right) \mathrm{d}\lambda \right)^2 \tag{3.14}$$

The expected fractional reduction in simulation time required for equivalent uncertainty with optimal spacing can be quantified with

$$\text{IF} = \frac{\sigma^2_{\text{Min.}}(\Delta \widehat{F}_{0,1}))}{\sigma^2(\Delta \widehat{F}_{0,1})}, \tag{3.15}$$

which is equivalent to the improvement factor (IF) of Lundborg et al. when $\sigma^2(\Delta \widehat{F}_{0,1})$ is obtained with a uniform distribution of sampling time.[250]

We could have presented an equivalent discussion in terms of thermodynamic length.[207,208,240,248,250] Thermodynamic length provides a measure of the distance between thermodynamic states and quantifies the length of paths between them. Minimising the thermodynamic length of a transformation minimises the dissipated work for slow but finite-time transformations (when correlation is accounted for or samples are uncorrelated). $\int_0^1 \sqrt{t_\lambda}\sigma \left(\left\langle \frac{\partial H}{\partial \lambda} \right\rangle_\lambda \right) \mathrm{d}\lambda$, or equivalently

$\int_0^1 \sqrt{g_\lambda}\sigma\left(\frac{\partial H}{\partial \lambda}\right) \mathrm{d}\lambda$, can be regarded as a measure of thermodynamic length which accounts for autocorrelation. If autocorrelation is ignored, $g = 1$ for all values of λ and the appropriate measure of thermodynamic length is $\int_0^1 \sigma\left(\frac{\partial H}{\partial \lambda}\right) \mathrm{d}\lambda$. However, the assumption of $g = 1$ means that minimising this length does not minimise overall uncertainty.

In equilibrium free energy calculations, simulations are usually carried out at discrete values of λ. Two methods to modify π to minimise $\sigma^2(\Delta\widehat{F}_{0,1})$ are changing the density of λ widows, and changing the allocation of sampling time to each window.

Changing the Density of Windows

π can be increased around a given value of λ by increasing the density of λ windows. For a transformation with K windows where there is an equal allocation of sampling time to each window, each of the $K-1$ $\Delta\lambda$ "steps" is effectively allocated a fraction of $\frac{1}{K-1}$ of the total simulation time. The approximate (due to the earlier assumption of infinitely close states) optimum spacing between windows k and $k+1$, $\Delta\lambda_{k,k+1}$, can be found from Equation 3.11:

$$\frac{t_{\Delta\lambda_{k,k+1}}}{t_{\mathrm{Tot.}}} = \frac{1}{K-1} = \frac{\int_{\lambda_k}^{\lambda_{k+1}} \sqrt{g_\lambda}\sigma\left(\frac{\partial H}{\partial \lambda}\right) \mathrm{d}\lambda}{\int_0^1 \sqrt{g_\lambda}\sigma\left(\frac{\partial H}{\partial \lambda}\right) \mathrm{d}\lambda} \tag{3.16}$$

$$\int_{\lambda_k}^{\lambda_{k+1}} \sqrt{g_\lambda}\sigma\left(\frac{\partial H}{\partial \lambda}\right) \mathrm{d}\lambda = \frac{\int_0^1 \sqrt{g_\lambda}\sigma\left(\frac{\partial H}{\partial \lambda}\right) \mathrm{d}\lambda}{K-1} \tag{3.17}$$

$$\int_{\lambda_k}^{\lambda_{k+1}} \sqrt{t_\lambda}\sigma\left(\left\langle\frac{\partial H}{\partial \lambda}\right\rangle_\lambda\right) \mathrm{d}\lambda = \frac{\int_0^1 \sqrt{t_\lambda}\sigma\left(\left\langle\frac{\partial H}{\partial \lambda}\right\rangle_\lambda\right) \mathrm{d}\lambda}{K-1}, \tag{3.18}$$

Alternatively, assuming $g = 1$ for all λ,

$$\int_{\lambda_k}^{\lambda_{k+1}} \sigma\left(\frac{\partial H}{\partial \lambda}\right) \mathrm{d}\lambda = \frac{\int_0^1 \sigma\left(\frac{\partial H}{\partial \lambda}\right) \mathrm{d}\lambda}{K-1}. \tag{3.19}$$

Hence, to determine the optimal spacing, we need an estimate of $\sqrt{t_\lambda}\sigma\left(\left\langle\frac{\partial H}{\partial \lambda}\right\rangle_\lambda\right)$ or $\sigma\left(\frac{\partial H}{\partial \lambda}\right)$ as a function of λ. One way to obtain this is from an initial short set of simulations with non-optimal spacing. The windows can then be spaced according to Equation 3.19 or 3.18 by ensuring that the areas under the $\sigma\left(\frac{\partial H}{\partial \lambda}\right)$ or $\sqrt{t_\lambda}\sigma\left(\left\langle\frac{\partial H}{\partial \lambda}\right\rangle_\lambda\right)$ curves between each value of λ are equal. We can either specify $N_{\mathrm{Wind.}}$, or the entire right-hand sides of Equations 3.19 or 3.18. We term these "thermodynamic speeds" following Minh, who discussed the spacing of windows according to Equation 3.19.[251] While the units of kcal mol^{-1} for a "speed" may

seem unnatural, it follows from the widely accepted definition of thermodynamic length, which has units of energy. Because the speed measures the step size taken between each λ-window, it also has units of length and therefore energy. We specify a constant thermodynamic speed, ensuring that equivalent spacing is obtained between different transformations of varying thermodynamic length. This algorithm is illustrated in Figure 3.1 for the case where $\sigma\left(\frac{\partial H}{\partial \lambda}\right)$ is used.

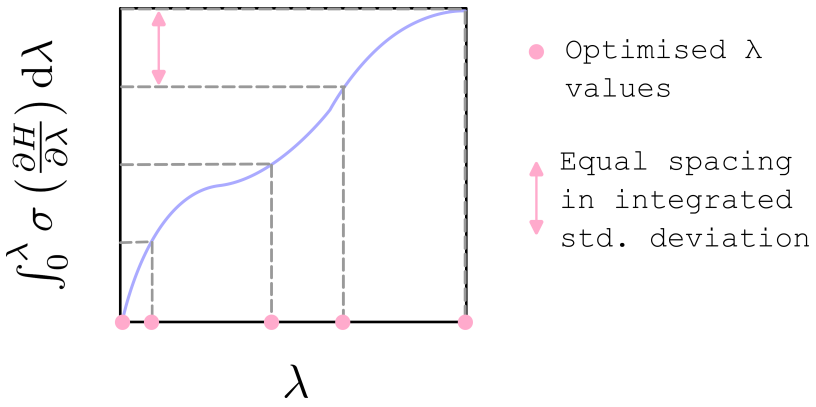


Figure 3.1: An illustration of the λ-spacing algorithm, where $\sigma\left(\frac{\partial H}{\partial \lambda}\right)$ is used. The variation of $\sigma\left(\frac{\partial H}{\partial \lambda}\right)$ with λ is estimated from short initial simulations with a fixed default λ schedule and used to estimate $\int_0^\lambda \sigma\left(\frac{\partial H}{\partial \lambda}\right) \mathrm{d}\lambda$ as a function of λ. The selected "thermodynamic speed" determines the constant step size in $\int_0^\lambda \sigma\left(\frac{\partial H}{\partial \lambda}\right) \mathrm{d}\lambda$, which is used to select the optimsed λ-schedule. Regions where $\int_0^\lambda \sigma\left(\frac{\partial H}{\partial \lambda}\right) \mathrm{d}\lambda$ increases quickly as a function of λ suggest quickly changing probability distributions in configuration space, which produce more densely-spaced λ windows.

A common approach to spacing λ windows is to aim for consistent and sufficient overlap between adjacent windows.[128] The uncertainty of the BAR estimator can be expressed in terms of overlap,[121] but also in terms of the variance of the gradient (under our assumption of infinitely close window spacing, see Section B.1). Hence, these spacing approaches are closely related and the variance of the gradient must predict the overlap in the limit of infinitely closely spaced windows. Intuitively, this is because the dimensionless variance of the gradient is the average squared relative change in probability with λ

$$\beta^2\sigma^2\left(\frac{\partial H}{\partial \lambda}\right) = \left\langle \left(\frac{1}{p(\mathbf{x},\lambda)}\frac{\partial p(\mathbf{x},\lambda)}{\partial \lambda}\right)^2 \right\rangle_\lambda , \tag{3.20}$$

where $p(\mathbf{x},\lambda)$ is the probability of observing the phase-space coordinates $\mathbf{x}$ at λ and $\beta = \frac{1}{k_\mathrm{B}T}$ (Section B.2).

Changing the Allocation of Sampling Time

Varying window spacing with a constant time allocation per window is one strategy to achieve optimal sampling times at each λ, minimising overall uncertainty. A complementary strategy is to allocate different sampling times to each window.

For K windows, we can approximate Equation 3.5 as

$$\sigma^2(\Delta\widehat{F}_{0,1}) = \sum_{k=1}^{K} \sigma^2(\Delta\widehat{F}_k) = \sum_{k=1}^{K} w_k^2 \sigma^2 \left(\left\langle \frac{\partial H}{\partial \lambda} \right\rangle_\lambda \right) = \sum_{k=1}^{K} \frac{w_k^2}{n_k} \sigma^2 \left(\frac{\partial H}{\partial \lambda} \right), \quad (3.21)$$

where there are K states with state number k, w_k is the state weight given by the numerical integration method, and ΔF_k is the contribution of a single state to the overall free energy change, such that $\Delta F_{0,1} = \sum_{k=1}^{K} \Delta F_k$.

From Equation 3.11, we can write

$$\frac{t_{\text{Optimal},k}}{t_{\text{Tot.}}} = \frac{w_k \sqrt{g_k} \sigma \left(\frac{\partial H}{\partial \lambda} \right)}{\sum_{k=0}^{K} w_k \sqrt{g_k} \sigma \left(\frac{\partial H}{\partial \lambda} \right)}, \quad (3.22)$$

where $t_{\text{Optimal},k}$ is the estimate of the optimal sampling time at state k . The data currently available from state k can be used to effectively estimate the statistical inefficiency according to $g_k = \frac{t_{\text{Current},k}}{n_{\text{Current},k}}$, where $t_{\text{Current},k}$ is the current sampling time at state k and $n_{\text{Current},k}$ is the current number of uncorrelated samples. Additionally, the denominator of Equation 3.11 can be written in terms of the minimum achievable variance of the free energy change estimate (Equation 3.14)

$$\frac{t_{\text{Optimal},k}}{t_{\text{Tot.}}} = \frac{w_k \sqrt{\frac{t_{\text{Current},k}}{n_{\text{Current},k}}} \sigma \left(\frac{\partial H}{\partial \lambda} \right)}{\sqrt{t_{\text{Tot.}}} \sigma_{\text{Min.}}(\Delta\widehat{F}_{0,1})} \quad (3.23)$$

$$t_{\text{Optimal},k} = \sqrt{\frac{t_{\text{Current},k} t_{\text{Tot.}}}{\sigma^2_{\text{Min.}}(\Delta\widehat{F}_{0,1})}} \sigma_{\text{Current}}(\Delta\widehat{F}_k). \quad (3.24)$$

Equation 3.24 shows how to estimate the optimal sampling time at state k from the current state sampling time and standard error of the the state free energy change contribution, $\sigma_{\text{Current}}(\Delta\widehat{F}_k)$. However, we avoid allocating sampling time with a specified $t_{\text{Tot.}}$ for each calculation because this reduces the efficiency over multiple calculations compared to specifying a shared "runtime constant", $C = \frac{\sigma^2_{\text{Min.}}(\Delta\widehat{F}_{0,1})}{t_{\text{Tot.}}}$

$$t_{\text{Optimal,k}} = \sqrt{\frac{t_{\text{Current},k}}{C}} \sigma_{\text{Current}}(\Delta\widehat{F}_k). \quad (3.25)$$

This allows the total simulation time per calculation to vary to reduce the overall uncertainty for the set of calculations. Repeated cycles of simulation and re-estimation of the optimum sampling time with Equation 3.25 are performed until the estimated optimum sampling time is equal to, or less than, the current sampling time.

Sun et al. derived an adaptive time allocation protocol based on the intra-run BAR errors.[256] Our approach differs in that we calculate the uncertainties ($\sigma_{\text{Current}}(\Delta\widehat{F}_k)$) from the differences between independently-equilibrated replicate runs, rather than from intra-run fluctuations. To allow easy decomposition of the uncertainties into contributions from single states, increasing independence, we base our algorithm on the errors from TI rather than BAR. Theoretically, each set of repeat simulations at a given state can be run independently. During the preparation of this work, Yu et al. proposed a similar algorithm.[268]

3.2.2 Detection of equilibration

Our equilibration detection heuristic takes inspiration from the Gewke test,[265] which checks the equality of means of initial and final portions of the data. Typically, the first 10 % and last 50 % of the data are used. Significantly different means are taken to indicate a systematic drift and therefore lack of equilibration.

We aim to detect systematic drift of the free energy estimate from an ensemble of independent replicates while ignoring systematic differences between the replicate estimates. Therefore, we perform a paired t-test between the free energy estimates from the first 10 % and final 50 % of the data for each replicate. This is repeated while sequentially discarding more data from the start of the simulation up to some threshold. The data is accepted as equilibrated when there is no evidence for a significant difference between the paired samples at the 95 % confidence level.

3.3 Methods

3.3.1 System Preparation

In general, the initial test systems were prepared from crystal structures, retaining crystal waters and discarding all other non-protein and non-ligand atoms. Details of initial structure preparation for individual systems are given in Section B.3 and the rationale for system selection is given in Section 3.4.1. Alternate atom

locations with the greatest occupancy were selected (or the A state, if occupancies were equal). Missing atoms were added using pdb4amber.[175] The protonation states of all residues were assigned using H++ (version 3.2).[178] Proteins and crystal waters were parameterised with the AMBER ff14SB and TIP3P force fields using antechamber 22.0.[180,185,269] Ligands were protonated using Open Babel (version 3.0.0) and parameterised with OpenForceField 2.0.0 and AM1-BCC partial charges through BioSimSpace (version 2023.0.0).[83,85,86,181,182] The protein-ligand complexes and free ligands were solvated with TIP3P water and 0.15 M of NaCl in rhombic dodecahedral boxes, ensuring a minimum distance of 15 Å between the solute and the box edge. The Cyclophilin D systems were taken from the "abfe_2fs" input supplied by Alibay et al. (GAFF2 with AM1-BCC partial charges, ff99SB-ILDN, and TIP3P force fields).[1,270] These were used with two minor modifications - the ligands were extracted and reparameterised with GAFF2.11 and AM1-BCC partial charges using antechamber 22.0 to avoid issues with non-integer charges, and the boxes were expanded to rhombic dodecahedral boxes with a minimum solute-box edge distance of 15 Å to allow the use of a 12 Å cutoff for reaction-field electrostatics.

3.3.2 Molecular Dynamics Protocols

All solvated systems (both "initial" and obtained from Alibay et al.) were subject to the standard A3FE minimisation and equilibration workflow, in which all simulations were performed using GROMACS 2021.3 through BioSimSpace (version 2023.0.0).[271] Systems were energy minimised for 1000 steps, followed by equilibration in the NVT ensemble (5 ps with all non-solvent atoms restrained and heating from 0 to 298 K, followed by 50 ps with restraints on all backbone atoms for the complexes only, then 50 ps with no restraints). NPT equilibration was then performed at 1 atm and 298 K (400 ps with restraints on non-solvent heavy atoms, followed by 1 ns with no restraints). Finally, independent 5 ns NPT equilibration runs were carried out for each independent replicate run for all systems to provide varied starting conformations. All restraints used a force constant of 10 kcal mol^{-1} Å^{-2}. The V-rescale and C-rescale algorithms were used for the relevant equilibration steps, with time constants of 1 and 4 ps, respectively.[272,273] All equilibration molecular dynamics simulations used a timestep of 2 fs and a cut-off of 8 Å for short-range interactions, with electrostatic interactions computed with the PME algorithm.[80]

3.3.3 Absolute Binding Free Energy Calculations

Individual Calculations

ABFE calculations were carried out using the double decoupling method of Gilson et al.,[49] according to the thermodynamic cycle shown in Figure 2 of Clark et al..[274] The simulations with the ligand in solvent collectively make up the free leg, while those with the receptor-ligand complex make up the bound leg. Sets of calculations where interactions of a given type are introduced or removed are termed stages: receptor-ligand restraints were introduced, charges were scaled, and Lennard-Jones (LJ) terms were scaled in the restrain, discharge, and vanish stages, respectively. Standard free energies of binding were calculated according to

$$\begin{aligned}\Delta G^{o}_{\text{Bind}} =& \Delta G_{\text{Free, Discharge}} + \Delta G_{\text{Free, Vanish}} - \Delta G^{o}_{\text{Release}} - \Delta G_{\text{Bound, Vanish}} \\ & - \Delta G_{\text{Bound, Discharge}} - \Delta G_{\text{Bound, Restrain}} + \Delta G_{\text{Sym. Corr.}} \\ =& \Delta G_{\text{Free}} + \Delta G^{o}_{\text{Bound}},\end{aligned} \tag{3.26}$$

where $\Delta G_{\text{Sym. Corr.}}$ includes any required symmetry corrections and $\Delta G^{o}_{\text{Release}}$ is the free energy of releasing the non-interacting ligand to the standard state volume.[73] ΔG_{Free} and $\Delta G^{o}_{\text{Bound}}$ are the overall free and bound leg contributions to $\Delta G^{o}_{\text{Bind}}$, where $\Delta G^{o}_{\text{Bound}}$ includes all terms other than $\Delta G_{\text{Free, Discharge}}$ and $\Delta G_{\text{Free, Vanish}}$. The relative receptor-ligand external degrees of freedom were restrained using Boresch restraints and $\Delta G^{o}_{\text{Release}}$ was calculated using an analytical correction term.[50,275] The coupling parameter, λ, was scaled from 0 to 1 to gradually introduce Boresch restraints, and to remove charges and LJ terms in the relevant stages. λ linearly scaled the magnitude of restraint force constants and charges, while a soft core potential based on that of Zacharias et al. was used to scale LJ terms (soft-core parameter set to 2.0).[138]

Individual simulations were carried out using SOMD (Sire/ OpenMM Molecular Dynamics),[187,276] which is available within Sire (version 2023.1.3).[267] OpenMM's LangevinMiddleIntegrator was used (friction coefficient of 1 ps^{-1}, coupled to a heat bath at 298 K).[277] Pressure was maintained at 1 atm using a Monte Carlo barostat (isotropic box scaling attempts every 25 time steps). Hydrogen mass repartitioning (with a repartitioning factor of 3) was used to allow a timestep of 4 fs, as done in earlier ABFE studies with this simulation engine.[172,278] Bonds to hydrogen were constrained. A cut-off of 12 Å was used for all non-bonded interactions and electrostatics were treated with the reaction field method with a

dielectric constant of 78.3.[189] We did not correct for the truncation of the tails of the LJ potentials because we found this produced negligible corrections. Energy minimisation with a maximum of 1000 iterations was performed prior to every simulation. The gradient of the free energy with respect to λ was calculated every 200 timesteps.

Overall Workflow

The complete workflow (Figure 3.2) was carried out using the Python package A3FE (version 0.1.0), which is available on Github at github.com/michellab/a3fe and is built on BioSimSpace and SOMD.[182,267] The final 5 ns equilibration run of the fully interacting complex for the first replicate was analysed to select the Boresch restraints using BioSimSpace. This algorithm fits the force constants to the fluctuations observed for prospective sets of anchor points, then selects the stable anchor set which maximally restricts the configurational volume available to the fully decoupled ligand. This effectively mimics strong native receptor-ligand interactions and ensures that $\Delta G_{\text{Bound, Restrain}} \approx 1.23$ kcal mol^{-1} (see section S6 in Clark et al. for a discussion and Hedges et al. for a tutorial).[204,274] The same restraints were used for all replicate runs to ensure that the free energy gradients at a given λ value would converge to the same values with infinite sampling. Different initial configurations were used for each leg of each repeat run, as extracted from the end of the 5 ns equilibration step. Within each leg, the same initial configuration was used for each λ-window.

The λ-schedule from Clark et al.,[274] which was manually optimised for the MIF/MIF180 system, was used as a default. This used 8 windows for the free discharge stage, 18 for the free vanish stage, 6 for the bound restrain stage, 8 for the bound discharge stage, and 36 for the bound vanish stage. This schedule was used to run all non-adaptive simulations for the initial test systems. For adaptive simulations, the λ-schedule at each stage was generated by running very short simulations (0.1 ns, no replicates) at all default λ values. These were used to estimate $\sigma\left(\frac{\partial H}{\partial \lambda}\right)$ as a function of λ, and the new λ-schedule was generated so that the area under the $\sigma\left(\frac{\partial H}{\partial \lambda}\right)$ vs λ curve was equal between adjacent λ values. The size of these areas, and hence the spacing of the windows, were determined by the user-specified "thermodynamic speed" parameter.[251]

The adaptive runs began with 0.2 ns simulations for all replicates. Per-window free energy changes were calculated from the product of $\left\langle \frac{\partial H}{\partial \lambda} \right\rangle_\lambda$ and weights from the trapezoidal rule. From the inter-run deviations, the standard errors of the mean free energy changes were calculated and used to predict the optimal run

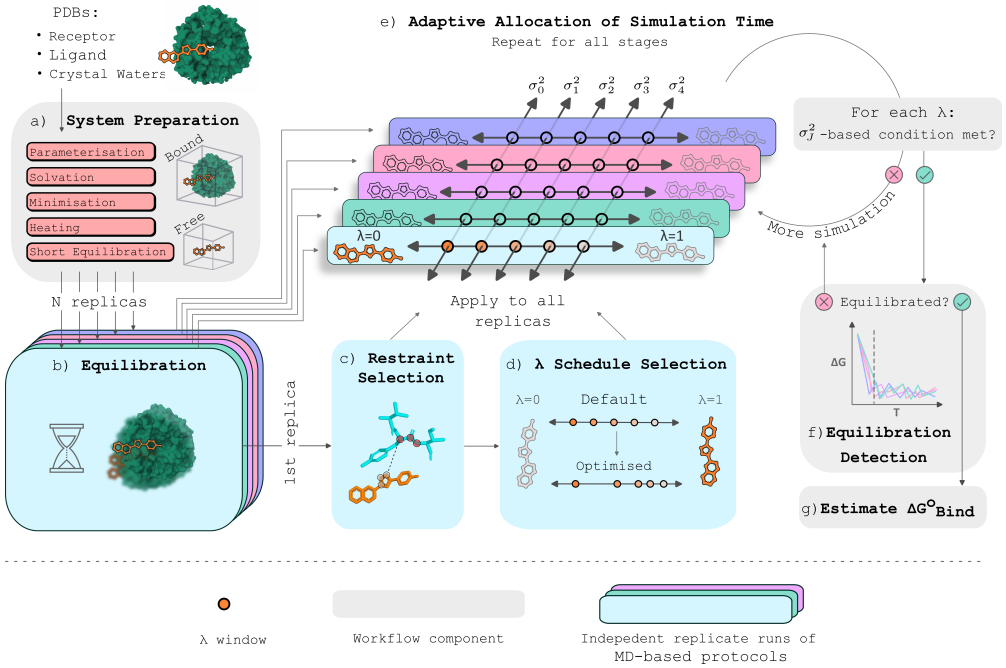


Figure 3.2: Scheme of the automated absolute binding free energy workflow implemented in the A3FE Python package. Following a) initial system preparation, b) individual 5 ns equilibration runs were carried out for each of the N replicate runs for each leg to provide diverse input conformations for the free energy calculations. c) The first of these for the bound leg was analysed to select Boresch restraints. d) A single short replicate was run for each of the default λ values to provide an estimate of the variance of the gradient, from which the final λ values were calculated. The same Boresch restraints and λ scheme were used for all replicate production runs. e) Simulation time was allocated by iteratively calculating the inter-run standard errors of the per-window ΔG estimates, estimating the optimal run time for each window, and allocating sampling time accordingly. Once the estimated optimal sampling times were achieved, f) the equilibration time was determined for each stage using an ensemble-based equilibration detection heuristic, g) and the overall free energy changes were calculated using MBAR.

times according to Equation 3.25. Additional simulation time was then allocated accordingly, and iterations of sampling time allocation and analysis were performed until predicted optimum sampling times agreed with the allocated sampling times. Equal sampling times were allocated to all replicate windows at the same value of λ. To promote smooth convergence to the "optimal" time, the total sampling time for each window was no more than doubled during each cycle. To smooth the allocation of sampling time between windows, it was enforced that the total sampling time for a given window could be no less than half that of the adjacent windows. Total sampling time was controlled with the user-input runtime constant (C in Equation 3.25). To obtain true optimum efficiency, times and statistical inefficiencies should be measured in terms of computing time (or

power consumption/ financial cost), rather than in nanoseconds of simulation time. This accounts for the differences in computational costs of different systems. To account for the different computational costs of different legs for different systems, $t_{\mathrm{Current},k}$ and $t_{\mathrm{Optimal,k}}$ were multiplied by the relative computational cost of the current calculation leg, resulting in an overall factor of $\frac{1}{\sqrt{\mathrm{Rel.Cost.}}}$ on the right hand side of Equation 3.25. Costs were calculated for all systems from the λ selection simulations, which were all run on the same machine with four GeForce RTX 2080 SUPER GPUs. These were expressed relative to the MIF180 bound leg, which had an absolute computational cost of 0.21 hours ns^{-1}. Hence, Equation 3.25 was used as shown with simulation times for the MIF180 bound legs, but where the computational cost was greater, the predicted optimal time was reduced. The approach of using relative costs was intended to improve the repeatability of the workflow - once users have calculated the absolute cost of the MIF180 bound leg, for which input files are supplied, this can be supplied to A3FE and the resultant simulation times (in ns) should be comparable to those shown here. The effect of varying hardware was removed by expressing computational time in GPU hours on GeForce RTX 2080 SUPER GPUs, obtained by multiplying simulation time in ns by the absolute cost calculated during the λ selection simulations.

Equilibration detection was performed after the adaptive allocation of sampling time was complete. This was done by splitting all time-ordered data for a stage into 100 blocks, so that each contained the same fraction of data (but not necessarily sampling time) from each λ window. The free energy change at each percent of total stage simulation time was computed using the Multistate Bennett Acceptance Ratio (MBAR) as implemented in pymbar 3.1.1.[121,131] The mean ΔG estimates for the first 10 % and last 50 % of the data were calculated for each run and a paired t-test was performed on the per-run differences. This was repeated after discarding 17 %, 33 %, and 50 % of total simulation time. Equilibration was equated with the first p-value greater than 0.05. Unequilibrated data were discarded and $\Delta G^{o}_{\mathrm{Bind}}$ was calculated with pymbar. The result was expressed as a mean over all N replicates and uncertainties were calculated as 95 % t-based confidence intervals (using $N-1$ degrees of freedom). Where not provided, uncertainties in experimental affinity data were assumed to be 0.5 kcal mol^{-1}.[20,279]

3.4 Results and Discussion

3.4.1 The Behaviour of Non-Adaptive Runs

To inform the development of our adaptive protocols, we first performed non-adaptive ABFE calculations on five protein-ligand systems (Figure 3.3). These were intended to span a range of ligand sizes, rates of equilibration, and sampling challenges. T4L/Benzene is a common test system for ABFE methodologies;[50,190] MIF/MIF180 is subject to slow rehydration of the binding site upon ligand decoupling;[166,274] the MDM2 complexes were used to develop the group's previous adaptive protocol;[165] and the slow equilibration of the PDE2a complex was highlighted by Huggins.[171] T4L, MIF, MDM2, and PDE2a denote the L99A mutant of T4 lysozyme, human macrophage migration inhibition factor, mouse double minute 2 homolog with a truncated lid, and phosphodiesterase 2a, respectively (Section B.3). We ran 5 independent replicate runs for each system, all using the manually-optimised λ-schedule from our previous work.[274]. Firstly, we investigated short-timescale behaviour by running all windows for a total of 0.2 ns and using all data for analysis. We then carried out a more realistic protocol optimised for MIF/MIF180 in our previous work - all windows were run for 6 ns with the initial 1 ns discarded to equilibration, other than the bound vanish stages, which were run for 8 ns with the initial 3 ns discarded to equilibration. Finally, we investigated longer timescale behaviour by running all windows for 30 ns and discarding the first 10 ns of data to equilibration. The protocols are referred to as 0.2 ns, 6 ns, and 30 ns. As we aimed to compare between protocols rather than with experiment, we tolerated two potential sources of error which we may not have otherwise; the *syn-anti* isomerisation in MIF180 was rarely sampled, in contrast to our previous work using GAFF2.11,[166,274] and we performed calculations for the charged ligand Pip2 using a reaction field treatment of electrostatics without maintaining neutrality of the box or applying corrections.[280]

None of the overall free energy changes are strictly converged

Our main interest was in the equilibration and convergence behaviour of the non-adaptive runs, and their relative performance compared to adaptive protocols. However, we include experimental binding free energies in Table 3.1 to check that our results are reasonable. The 30 ns computed free energies are broadly similar to the experimental values, giving some assurance that our default protocol is reasonable. We note that the calculated values appear excessively negative for ligands

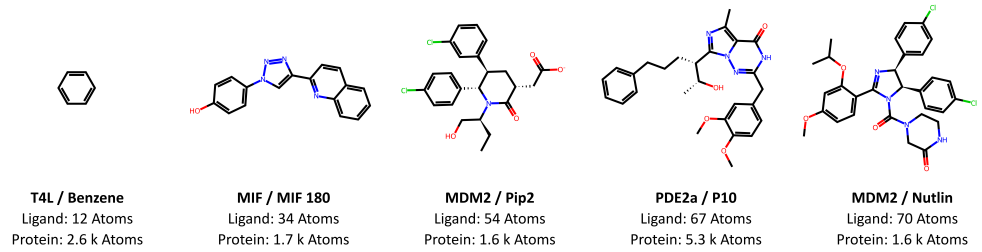


Figure 3.3: Summary of initial test complexes. Drawn with RDKit.[281]

Table 3.1: Non-Adaptive $\Delta G^{o}_{\mathrm{Bind}}$ for Initial Test Systems[a]

	0.2 ns	6 ns	30 ns	Exp. $\Delta G^{o}_{\mathrm{Bind}}$	Exp. Source
T4L	-4.74 ± 0.79	-3.80 ± 0.87	-4.03 ± 0.80	-5.19 ± 0.16	ITC, Morton et al.[282]
MIF	-16.56 ± 2.49	-10.16 ± 1.27	-10.48 ± 1.14	-8.98 ± 0.28	FPA, Cisneros et al.[198]
MDM2-Pip2	-17.32 ± 1.68	-13.49 ± 1.27	-13.94 ± 0.86	-9.11 ± 0.01	ITC, Michelsen et al.[283]
PDE2a	-29.40 ± 2.89	-17.26 ± 1.49	-16.96 ± 1.97	-14.35 ± 0.50*	SPA, Obach et al.[284]
MDM2-Nutlin	-17.88 ± 4.07	-15.56 ± 2.25	-16.92 ± 1.98	-11.14 ± 0.27	ITC, Mendoza-Martinez et al.[165]

[a] All quantities in kcal mol^{-1}. Calculation uncertainties stated as 95 % confidence intervals based on the variance of 5 replicate runs, assuming Gaussian distributions, while experimental uncertainties are given as standard deviations. Symmetry corrections are included. A detailed breakdown of the components of the calculated binding free energies is given in Table B.1 and all Boresch restraint parameters are given in Table B.2. * Uncertainty estimated based on Hahn et al..[20,279] ITC, FPA, and SPA denote isothermal titration calorimetry, fluorescence polarisation assay, and scintillation proximity assay.

larger than benzene, in particular the MDM2 ligands. This is to be anticipated for a wide variety of sampling challenges in equilibrium ABFE calculations, because the bound leg simulations are started from the holo structure. Any failure to relax towards the apo structure and free ligand conformations during the simulation can be regarded as erroneous "preorganisation" of the system for binding, producing erroneously favourable estimates of binding. This effect can generally be expected to increase for larger ligands with more substantial sampling challenges and more favourable binding affinities, and may be avoided using bidirectional non-equilibrium switching protocols which account for the apo state.[219] While we make broad comparisons with experimental affinities here, we note that these are from different sources, including biochemical assays.

More pertinently, the initial test systems show a range of equilibration behaviours. While the results obtained for the brief 0.2 ns runs are not significantly different to those from the costly 30 ns runs for T4L, there are significant and substantial differences for MDM2-Pip2, MIF, and especially PDE2a. As shown by the breakdown of binding energy components (Table B.4) and Figure B.1,

the dominant contribution to slow equilibration comes from the bound vanish legs, where sampling is impeded by the protein and where a second-order phase-change like transition occurs as the Van der Waals interactions are removed.[285] The diversity of equilibration behaviour suggests that equal-time equilibration rules are likely to be sub-optimal. However, all cases of slow equilibration are similar in that the overall free energies become less favourable with increasing sampling time, which is to be expected given the discussion of sampling issues above.

The uncertainties in Table 3.1 do not generally decrease proportionally to the inverse square root of sampling time. This suggests that the increased sampling time does not produce a proportional number of independent samples; rather independent runs are becoming trapped in separate regions of conformational space.[111,286] Again, Table B.4 shows that sampling in the bound leg, and in particular the bound vanish leg is to blame. In contrast to the bound leg, the uncertainties for the free leg generally decrease with sampling time.

While the 6 ns and 30 ns runs appeared "equilibrated", we performed further checks to assess convergence. For strictly converged simulations, we would expect replicate runs to sample the same distributions of our parameter of interest. We tested whether the gradients obtained for independent runs were sampled from the same distributions using the Kruskal-Wallis H-test (after subsampling according to the statistical inefficiency determined individually for each run).[262] The Kruskal-Wallis H-test is a non-parametric equivalent of the one-way analysis of variance (ANOVA), and is an extension of the Mann-Whitney U-test for more than two groups.[287] It tests the null hypothesis of equal medians for all underlying (population) distributions; the alternative hypothesis is that the population median of one group is different to at least one other group and that all samples do not originate from the same distribution. Figure B.2 shows that the proportion of windows showing significant differences at the 95 % confidence level is always above 70 % for the bound vanish and bound discharge stages. However, for the free legs, the fraction increases with ligand size from around 0 for T4L to similar to the bound fractions for MDM2-Nutlin. This indicates increasing sampling issues with increasing ligand size. All calculations contain many windows where replicate runs produce significantly different median gradients, meaning that none of our calculations are strictly converged, even after 30 ns sampling time per window. We emphasise that our results would be judged as "converged" by the standards of most of the literature; the reported lack of convergence is due to our strict criterion. In Section B.7, we reanalyse results from Alibay et al.[1,288]

to demonstrate that most literature ABFE results are likely also unconverged by this criterion. This confirms that replicate runs are essential for reliable free energy estimates and uncertainty quantification for the bound leg runs, and many free leg runs, at least without improved sampling.[193,257,258] In accordance with Wan et al., this suggests that the optimal number of replicates for a given sampling time is likely to be the maximum number of replicates which remain long enough to achieve sufficient equilibration.[286] We note that averaging results from multiple replicate runs which become trapped in separate regions of configuration space is not theoretically rigorous,[289] because there is no guarantee that the distribution among these conformational regions is not biased by the starting conformations. However, this approach appears robust in practice.[257]

The Gellman-Rubin diagnostic is a popular metric used to diagnose convergence in MCMC simulations based on the difference between intra-run and inter-run variances.[264] However, we found that the tendency of the uncertainties not to decrease with increasing sampling time (the tendency of systems to remain trapped in local minima) meant that this diagnostic was not useful for determining the termination point of our simulations. Effectively, the convergence of our simulations often did not improve with increasing sampling time, and therefore the degree of convergence was not a useful stopping criterion. We note that a more recent version of the Gelman-Rubin diagnostic may be more useful for this purpose.[290] Still, we found this diagnostic to be very useful for highlighting windows where substantial sampling problems occurred (Section B.13).

The standard deviations of the gradients equilibrate quickly, while the standard errors equilibrate slowly

The standard deviation of the gradient ($\sigma\left(\frac{\partial H}{\partial \lambda}\right)$) and the time-normalised standard error of the mean gradient ($\sqrt{t_\lambda}\sigma\left(\left\langle\frac{\partial H}{\partial \lambda}\right\rangle_\lambda\right)$) are plotted against λ for PDE2a in Figure 3.4. This illustrates the general trends for all test systems (the plots for all remaining systems are given in Section B.8).

The standard deviations are in excellent agreement between the 0.2, 6, and 30 ns runs, illustrating rapid equilibration (Figure 3A). In general, the curves of $\sigma\left(\frac{\partial H}{\partial \lambda}\right)$ are extremely similar between the vanish stages, but $\sigma\left(\frac{\partial H}{\partial \lambda}\right)$ is slightly higher for the discharge legs. This may stem from greater rearrangement of the polar solvent environment than of the less polar binding pocket upon discharging. The shapes of the standard deviation curves are very similar between systems for the

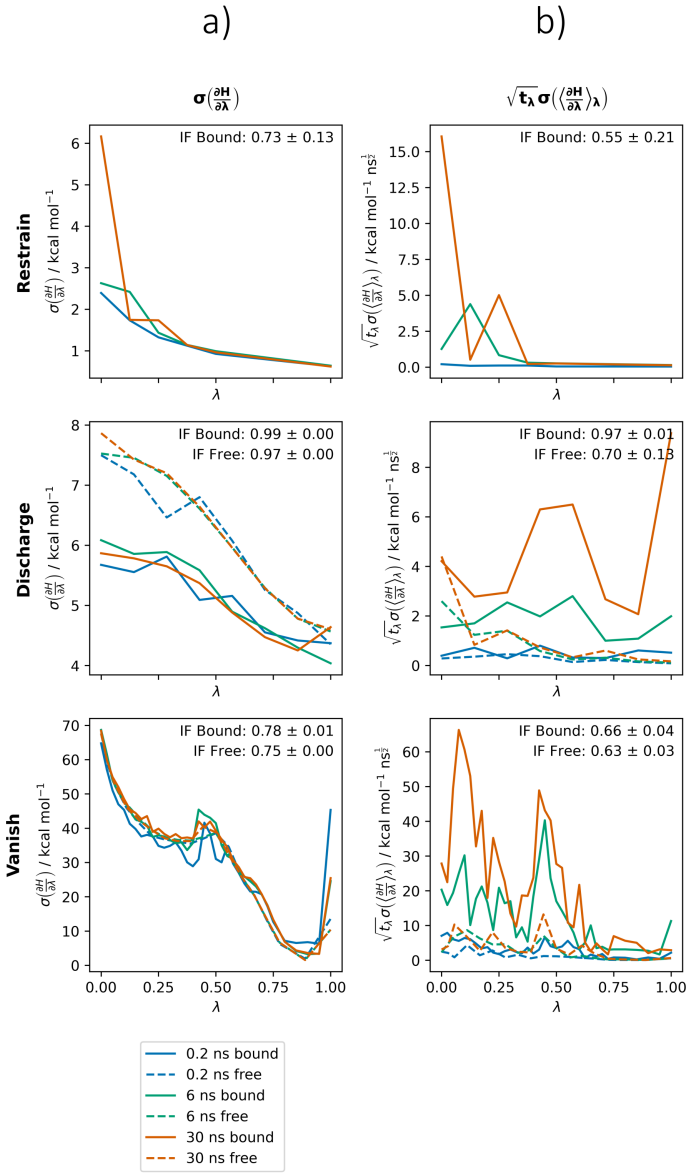


Figure 3.4: a) Standard deviation of the gradient and b) time-normalised standard error of the mean gradient against λ for all calculation stages for PDE2a. Improvement factors (IF, Equation 3.15) are computed with respect to equally-spaced λ-windows.

vanish stages, with the exception of the bound vanish stage for T4L, which shows a substantially lower standard deviation around $\lambda = 0.5$ (Figure B.4). This is likely because the volume vacated by the decoupled ligand is not filled by water in the hydrophobic T4L L99A binding site.

In contrast to the standard deviations, the time-normalised standard errors of the means of the gradients appear unequilibrated (Figure 3.4b). Increasing sampling time produces higher $\sqrt{t_\lambda}\sigma\left(\left\langle\frac{\partial H}{\partial\lambda}\right\rangle_\lambda\right)$ values which again reflects the fact that the uncertainties do not generally decrease with sampling time. This increase can happen for two reasons; firstly, because the replicates remain trapped in local minima in configuration space , producing constant $\sigma\left(\left\langle\frac{\partial H}{\partial\lambda}\right\rangle_\lambda\right)$ at increased time and hence increased $\sqrt{t_\lambda}\sigma\left(\left\langle\frac{\partial H}{\partial\lambda}\right\rangle_\lambda\right)$. In other words, our estimate of the statistical inefficiency has failed to converge because no transitions between the relevant conformational minima have been observed. Alternatively, this can be due to replicates exploring new local minima in configuration space, increasing both $\sigma\left(\left\langle\frac{\partial H}{\partial\lambda}\right\rangle_\lambda\right)$ and $\sqrt{t_\lambda}\sigma\left(\left\langle\frac{\partial H}{\partial\lambda}\right\rangle_\lambda\right)$. As expected from the uncertainties in Table B.4, $\sqrt{t_\lambda}\sigma\left(\left\langle\frac{\partial H}{\partial\lambda}\right\rangle_\lambda\right)$ is generally much smaller and more consistent with increasing simulation times for the free legs over the bound legs, reflecting the fact that the protein environment poses substantial sampling challenges which dramatically increase the statistical inefficiency. The exception to this is the free vanish stage for Nutlin, which shows a substantial peak in $\sqrt{t_\lambda}\sigma\left(\left\langle\frac{\partial H}{\partial\lambda}\right\rangle_\lambda\right)$ around $\lambda = 0.5$ which is similar in magnitude between the bound and free legs. This is due to the overlap of intramolecular groups (Section B.9). $\sqrt{t_\lambda}\sigma\left(\left\langle\frac{\partial H}{\partial\lambda}\right\rangle_\lambda\right)$ is dependent on the statistical inefficiency, which is determined by the slowest timescales of exchange between conformational minima which produce different gradients. As a result, it is noisy. Despite the noise, some trends can be observed: $\sqrt{t_\lambda}\sigma\left(\left\langle\frac{\partial H}{\partial\lambda}\right\rangle_\lambda\right)$ is generally highest between $\lambda = 0$ and 0.6 for the bound vanish stages where the ligand is more strongly interacting with the protein environment, but subsequently drops as the ligand becomes weakly interacting and sampling is improved. For the bound restrain stage, there are often peaks at $\lambda = 0$ where the restraint is absent and the ligand can explore poses which were not observed during the restraint-fitting simulations.

These observations have consequences for the design of adaptive protocols. Because $\sigma\left(\frac{\partial H}{\partial\lambda}\right)$ equilibrates quickly, minimum variance protocols can be found from short test simulations. These protocols are equivalent to minimum thermodynamic length protocols where the metric does not account for statistical inefficiency. Because $\sqrt{t_\lambda}\sigma\left(\left\langle\frac{\partial H}{\partial\lambda}\right\rangle_\lambda\right)$ equilibrates slowly, minimum standard error of the mean protocols cannot easily be found from short simulations, and require adaptive protocols. These protocols are equivalent to minimum thermodynamic length protocols where the metric accounts for statistical inefficiency.

If we wish to use the MBAR free energy estimator, we must calculate the energies of samples obtained at a given value of λ at all other values of λ. Protocols which adaptively change the number of λ-windows according to Equation 3.18 are then inconvenient because of the need to compute energies at all values of λ at which may be used. This can result in a substantial overhead. Therefore, we select λ values to achieve minimum variance using Equation 3.19 based on a short set of initial simulations of negligible cost. $\sigma\left(\frac{\partial H}{\partial \lambda}\right)$ was determined as a function of λ for a default set of windows using a single replicate run with 0.1 ns sampling per window. This is effectively equivalent to selecting states to achieve equal and sufficient overlap when statistical inefficiency is ignored. We then adaptively allocate simulation time to achieve minimum standard error of the mean according to Equation 3.25.

3.4.2 Performance of Window Spacing, Time Allocation, and Equilibration Detection Algorithms

The window spacing protocol is reliable and spacing impacts equilibration

To independently assess the performance of the automated λ-spacing protocol, we carried out tests on the MIF complex where equal sampling time was allocated to each window (i.e. the window spacing algorithm was used but the adaptive sampling algorithm was not). We generated λ-schedules using thermodynamic speeds of 0.5, 1.0, 2.0, and 4.0 kcal mol^{-1} (Equation 3.19). The cost of the protocol was negligible, but it generated more consistent off-diagonal overlap values than the manually-optimised schedule (Figure 3.5). This was especially true for the vanish stages, where the thermodynamic length varied least linearly as a function of λ. Figure 3.5 a) shows λ against the the normalised λ index ($\frac{\text{Index}}{\text{No.windows}}$), where a steeper curve of of λ against the normalised index indicates a lower density of windows. Here, the adaptive protocols reduces the density of states close to $\lambda = 1$, avoiding the increase in overlap observed for the non-adaptive protocol in this region (Figure 3.5a). As expected based on Section 3.4.1, the same number of windows were normally selected between the bound and free legs, and the number of windows approximately halved when the thermodynamic speed was doubled (Figure 3.6). A speed of 2.0 kcal mol^{-1} appeared close to the maximum, because low off-diagonal overlaps were observed (according to the rule-of-thumb that off-diagonal values should not be lower than 0.03).[128] This is in agreement with the results of Rizzi.[252]

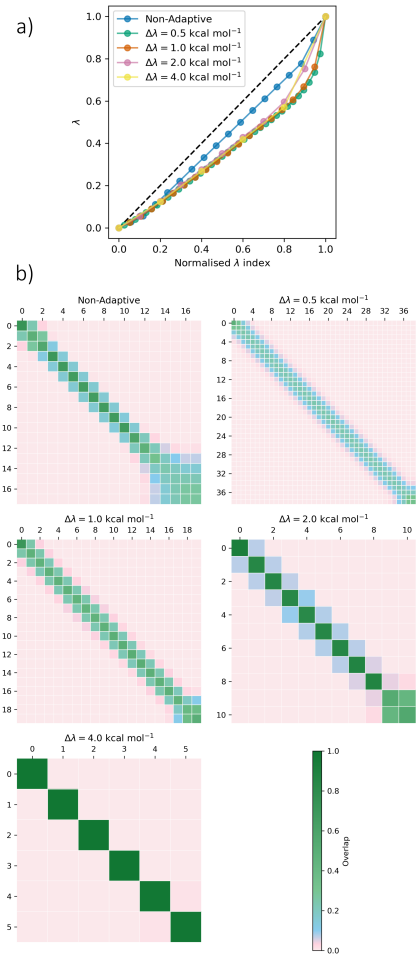


Figure 3.5: Automated selection of λ windows for the free vanish stage for MIF. a) λ against normalised λ index ($\frac{\text{Index}}{\text{No.windows}}$) and b) overlap matrices using a manually-optimised λ-schedule, and using the automated method to space windows with thermodynamic speeds of 0.5, 1.0, 2.0, and 4.0 kcal mol^{-1}. The automated method produces more consistent off-diagonal overlap than the manually-optimised schedule. Plots for all other stages are shown in Section B.10

Simulations with the generated λ-schedules showed no visible reduction in the overall standard error of the mean of the free energy change for equivalent sampling time compared to the "manually-optimised" protocol (Figure 3.7); rather, all schedules other than that with 4.0 kcal mol^{-1} speed showed similar inter-run uncertainties. This is as expected given that the protocol did not account for differences in statistical inefficiency between stages. This may have also been anticipated from the results of Nguyen and Minh, who found that given sufficient overlap, the uncertainty collapsed purely as a function of total sampling time.[291] However, the 4.0 kcal mol^{-1} thermodynamic speed protocol often showed substantially greater uncertainty, as anticipated from the negligible overlap.

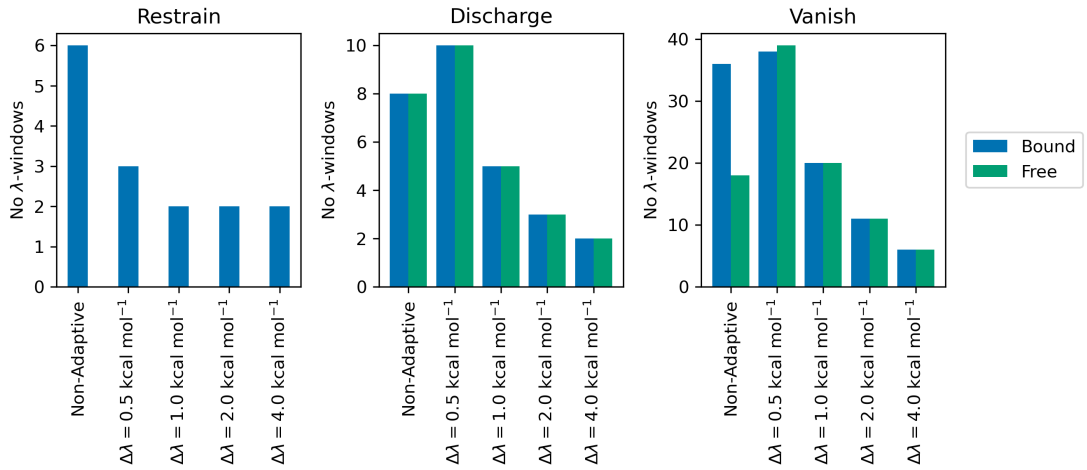


Figure 3.6: Number of windows selected for each stage for MIF. Non-adaptive shows the default number of windows, which were previously manually optimised for MIF/MIF180.[274] All other schedules were generated with the automated procedure using thermodynamic speeds of 0.5, 1.0, 2.0, and 4.0 kcal mol^{-1}.

The most substantial effect of increasing the spacing up to 2.0 kcal mol^{-1} was that equilibration was accelerated (Figure 3.8) for a given total simulation time. This was discussed by Rizzi.[252] The effect is intuitive - reducing the number of windows increases the total sampling time per window for a given total simulation time. This produces faster relaxation for a given computational cost. This suggests that for slowly relaxing systems where replica exchange is not used, the thermodynamic speed should be made as large as possible without negatively impacting the MBAR estimate through insufficient overlap. A speed of 2.0 kcal mol^{-1} appears to work well for this system. If this method were used with replica exchange, benefits would still be expected at relatively high thermodynamic speeds,[252] but the need for a sufficiently high exchange rate may favour lower speeds.[291]

This protocol is effectively a simplified version of those of Minh and Rizzi.[251,252] We believe that it has a more rigorous basis than more empirical approaches which minimise free energy differences between windows,[253] because it effectively targets equal divergences of the probability distributions between adjacent states. Adding a constant energy offset to a Hamiltonian would produce a free energy difference while maintaining perfect overlap, and therefore spacing states according to free energy differences has a weak theoretical foundation. This protocol requires that the λ-schedule for the trial simulations is sufficiently dense to faithfully capture variations in $\sigma\left(\frac{\partial H}{\partial \lambda}\right)$ with λ. However, for a reasonably chosen alchemical path (e.g. a reasonable choice of soft-core) there should be no sudden changes in $\sigma\left(\frac{\partial H}{\partial \lambda}\right)$ with λ (Figure 3.4) meaning that the initial λ-schedule does not have

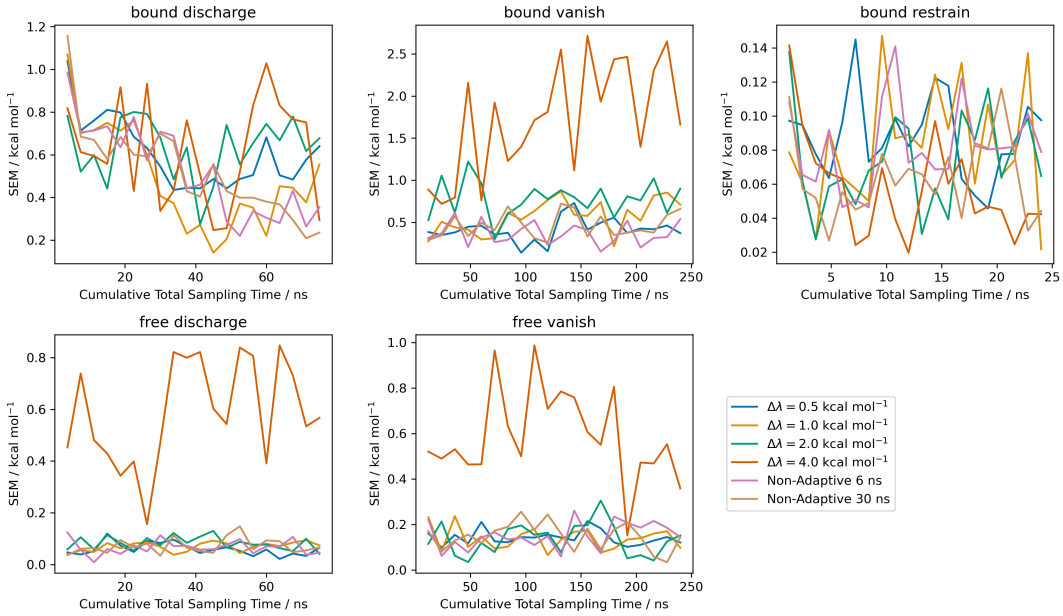


Figure 3.7: Non- cumulative inter-run standard errors of the mean of estimated ΔG against sampling time for all stages with different λ schedules. Equal simulation times were allocated to each window for within each schedule. Data were split into 20 equal blocks and MBAR was run on each block. By non-cumulative, we mean that the data for each block were analysed independently, so that the overall error is not expected to decrease with $\frac{1}{\sqrt{\text{SamplingTime}}}$. The data for truncated before analysis so that the per-stage computational costs were equal to that of the cheapest stage.

to be particularly dense. The advantage of using a fixed set of initial states is that initial simulations can be run in parallel, which may reduce wall-clock time compared to iterative schemes.[254] This protocol is also simpler than others based on similar principles.[255] Given that the automated protocol is simple to implement, appears robust, and produces superior λ-schedules compared to more costly manual selection, we recommend its use in alchemical free energy workflows.

The adaptive sampling scheme concentrates sampling where there are sampling issues

To isolate the effect of the adaptive time allocation algorithm, it was tested while λ-schedules were kept constant. This algorithm should provide the greatest advantage when two conditions are met: there is a sampling issue which is restricted to a few windows, and the timescale of the sampling issue is not dramatically longer than a typical λ simulation. This would mean that the default equal-time allocation would be a poor choice, and that the uncertainty could be affordably reduced with additional sampling of problematic windows.

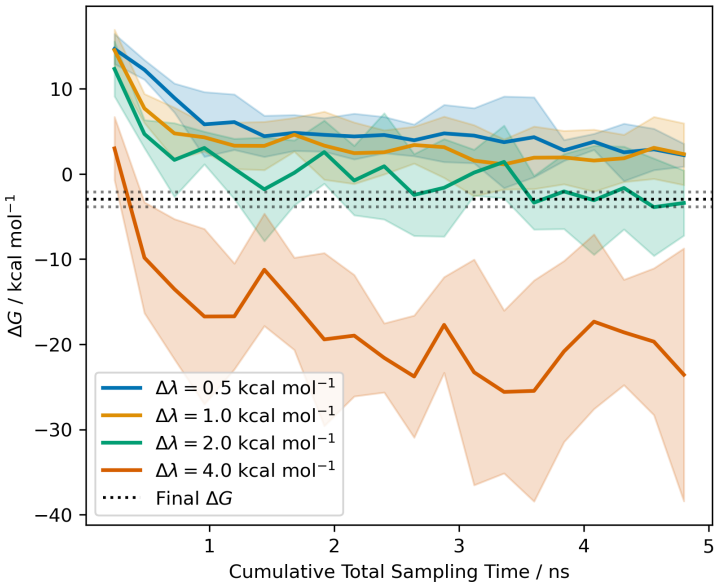


Figure 3.8: Equilibration of the MIF bound vanish stage against total sampling time with varying λ-window spacing. The mean is shown as a solid line, and shaded regions indicate 95 % t-based confidence intervals based on inter-replicate differences. Wider spacing leads to accelerated equilibration towards the 30 ns result (black-dotted line with 95 % CI shown as sparse dotted lines). The longer-time equilibration and results for the free vanish stage are shown in Section B.10.

To test whether the adaptive scheme produced benefits in a favourable case, we performed adaptive and non-adaptive runs for the free vanish stage of MDM2-Pip2. This was selected because it had the most favourable SEM-based improvement factor of any stage (0.37 ± 0.13, Figure B.6), indicating a sampling issue localised to a few windows. The adaptive and non-adaptive time-allocation runs used the same optimised λ-schedule with speed 1.0 kcal mol^{-1}. To provide improved statistics, 20 replicates were used for each protocol. A runtime constant of 1×10^{-10} kcal^2 mol^{-2} ns^{-1} was used for the adaptive runs and the non-adaptive data were truncated slightly above 2 ns per simulation per window to give an equal total computational cost.

Figure 3.9 shows the allocation of sampling time for the adaptive protocol. At $\lambda = 0$, the intramolecular Lennard-Jones forces are still fully active and the two chlorophenyl rings of Pip2 remain parallel throughout the simulation. Above $\lambda = 0.2$, these forces are substantially weakened, and the ligand relaxes so that the rings point in opposite directions. There is an intermediate regime around $\lambda = 0.1$ where the ligand slowly exchanges between both conformations, substantially increasing the uncertainty of the free energy differences between runs. This produces a spike in the allocated simulation time. The main effect

of the differing allocation of sampling time is to accelerate equilibration (Figure 3.9). At longer sampling times, the adaptive protocol produces estimates which are significantly closer to the final 30 ns estimate than the non-adaptive protocol. At very short timescales this trend is reversed, which suggests that fast relaxations are better sampled by the equal-time protocol, while the adaptive protocol more effectively samples the slow exchange between different ring orientations.

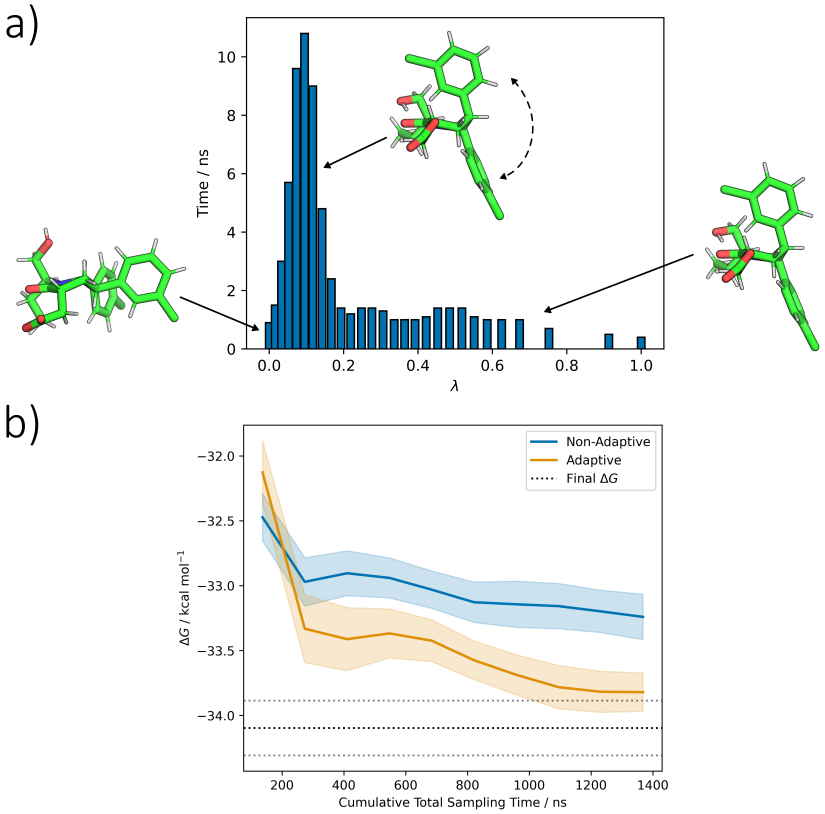


Figure 3.9: a) Allocation of sampling time against λ for the Pip2 free vanish leg. Sampling times shown are per-simulation, and 20 replicate simulations were used for each λ-window. The fluctuation between conformations around $\lambda = 0.1$ causes divergence of the free energy esimates between runs, and produces a peak in the sampling times. b) Equilibration of the free vanish stage towards the 30 ns result (black dotted line with sparser dotted lines showing 95 % CI). Data was split into 10 blocks before analysis with MBAR. Shaded areas show 95 % t-based CIs. The adaptive and non-adaptive runs use the same λ-schedule, and therefore the accelerated equilibration is due to the concentration of sampling time around the problematic $\lambda = 0.1$ region by the adaptive sampling algorithm.

We note that while the adaptive algorithm produces desirable improvements in the rate of equilibration, it was intended to reduce uncertainty. It fails to reduce inter-run uncertainty for Pip2, at least at earlier total sampling times (Figure B.16). This is explained by the large bias in initial conformations, in which

the chlorophenyl rings are always parallel. Concentrating sampling around the $\lambda = 0.1$ region in fact initially increased the inter-run uncertainty, because it resulted in a wider, but less biased, spread of free energies; however, uncertainty compared to the long-time result was reduced. If each window were initialised by a conformation which had been thoroughly equilibrated at the given value of λ, this algorithm should also reduce inter-run uncertainty.

We further tested the adaptive protocol on the bound vanish stage of the T4L complex, again with 20 runs per protocol and with a shared λ-schedule selected with a speed of 1.0 kcal mol^{-1} (Section B.13). Here, the sampling issue resulted from the occasional water molecule moving into the empty binding site around $\lambda = 1.0$. This issue was also noted by Lagardère et al.[225] The adaptive algorithm allocated substantially more sampling time to the affected windows, but in contrast to the Pip2 example, this did not produce any observable benefit. This is because the timescale of water entry to the binding site is far greater than what was allocated by our algorithm. Hence, we expect that the adaptive algorithm will rarely offer dramatic within-stage improvements, because the timescales of most sampling issues are expected to be much longer than the simulation times that can reasonably be allocated. This is in agreement with the recent work of Yu et al.,[268] who investigated the performance of a similar algorithm for computing the hydration free energies of trivalent rare earth elements in ionic liquids. They found that the algorithm offered substantial reductions in uncertainty at equivalent computational cost for the Coulombic contribution at 600 K. However, at 300 K the timescale of the sampling issue was dramatically increased and the adaptive protocol offered little improvement over equal-time allocation.

One advantage of our time allocation protocol compared to replica-exchange methods is that information does not have to be shared frequently between simulations. This makes the workflow substantially easier to deploy to high-performance computing clusters, where individual simulations can be parallelised across different nodes. If the only metric used to decide simulation time was the inter-replicate ΔG uncertainty for a given window, then only the N replicates at each window would have to finish before further simulation time could be allocated to that window. However, we also share some information between λ windows, as we require that the total simulation time allocated is no less than half that of the adjacent windows. This ensures that information about problematic regions is effectively shared between nearby windows, but it also increases the coupling between simulations, which now require information

from adjacent windows before they can be resubmitted. Despite the reduction in coupling, we note that our protocol is likely to be less efficient than replica-exchange-based methodologies in terms of absolute computing time, especially highly flexible schemes such as ensemble of expanded ensemble.[292] In principle, there is no reason why this algorithm should not be applied in combination with replica exchange methods; for example, windows with low statistical inefficiencies could be run for less time between exchange attempts than windows with high statistical inefficiencies. However, this is more complicated, since the statistical inefficiency at a given λ window becomes a function of the sampling at all other λ windows and the inter-window exchange probabilities, making it less clear where additional sampling time would be best allocated. Alternatively, windows showing large inter-run ΔG deviations could be targeted by enhanced sampling algorithms. For example, if many replicates were used, the low dimensional descriptor which best discriminates between the run mean ΔG estimates may be sought and used to guide enhanced sampling.

The adaptive sampling algorithm behaves predictably for the free legs, and less predictably for the bound legs

We sought to understand the reproducibility of the allocated sampling times between repeat runs of the entire protocol, and the effect of changing simulation parameters on the time allocated. In this section, all algorithms (adaptive simulation time allocation, automated window spacing, and equilibration detection) were used. Table 3.2 shows the allocation of sampling time between protocols labelled according to their runtime constant (r, kcal^2 mol^{-2} ns^{-1}), the thermodynamic speed used for window spacing (s, kcal mol^{-1}), and the number of independently-equilibrated repeat runs per protocol (n). The overall free energies are generally not significantly different to the non-adaptive 30 ns result. There is reasonable consistency between the times allocated for repeats of the r0.005-s1-n5 protocol, showing that similar sampling issues and related uncertainties are detected between overall repeats. Sampling time is mostly concentrated around $\lambda = 0.4$ during the bound vanish stage, where water begins to enter the binding site.[274] In the ideal case, where the uncertainty decreases proportionally to the square root of sampling time, all protocols with the same runtime constant should be allocated equal sampling times, regardless of the number of λ-windows or number of repeats. This appears to be the case for the free legs, where doubling the number of replicates, or halving or doubling the number of λ-windows does not appear to affect the allocated sampling time. In contrast, the bound vanish leg

Table 3.2: Variation of Allocated Simulation Time with Simulation Parameters for MIF[a]

Protocol	Parameters			Simulation Times / ns						ΔG^o_{bind} / kcal mol^{-1}
	Run Time Constant / kcal^2 mol^{-2} ns^{-1}	Thermodynamic Speed / kcal mol^{-1}	No. Replicates	Bound Restrain	Bound Discharge	Bound Vanish	Free Discharge	Free Vanish	Total	
r0.005, s1, n5, repeat 1	0.005	1.0	5	14	111	282	19	62	488	-11.80 ± 1.87
r0.005, s1, n5, repeat 2	0.005	1.0	5	21	112	306	16	62	517	-11.37 ± 0.74
r0.005, s1, n5, repeat 3	0.005	1.0	5	14	126	308	12	46	506	-11.28 ± 1.66
r0.005, sOrig., n5	0.005	Default spacing	5	6	126	196	12	44	384	-12.12 ± 1.07
r0.001, s1, n5	0.001	1.0	5	98	311	1286	32	114	1840	-11.37 ± 0.59
r0.005, s1, n10	0.005	1.0	10	14	112	185	17	66	394	-12.27 ± 0.88
r0.005, s0.5, n5	0.005	0.5	5	6	100	190	15	54	366	-12.05 ± 1.56
r0.005, s2, n5, repeat 1	0.005	2.0	5	8	324	406	14	52	803	-9.52 ± 0.67
r0.005, s2, n5, repeat 2	0.005	2.0	5	17	430	257	12	56	772	-11.15 ± 1.56
r0.005, s4, n5	0.005	4.0	5	12	79	551	16	62	720	-9.47 ± 1.91

[a] ΔG^o_{bind} shown with with 95 % t-based confidence intervals). In the protocol names, r denotes the run time constant (kcal^2 mol^{-2} ns^{-1}), s denotes the thermodynamic speed (kcal mol^{-1}) used to determine the λ-window spacing, and n denotes the number of replicates in each ensemble run. sOrig denotes the use of the default λ-spacing. A detailed breakdown of the calculated free energies is given in Table B.3.

simulations are prone to becoming stuck in local minima and the uncertainties do not generally decrease as expected with additional sampling time (Section 3.4.1). Here, increasing the number of λ-windows (r0.0005-sOrig-n5, r0.0005-s0.5-n5) or the number of replicates (r0.0005-s1-n10) appears to decrease the total allocated sampling time, likely indicating that reduced uncertainty is achieved with equivalent sampling time when the number of independent simulations is greater, in agreement with Wan et al..[286] However, this trend is not replicated for the bound discharge leg. Increasing the λ-spacing with speeds of 2 and 4 kcal mol^{-1} generally increased the total sampling time, but did not have a reliable effect on the sampling time for the bound stages, possibly indicating lower reliability of sampling issue detection with fewer independent simulations. Surprisingly, the r0.0005-s4-n5 result remained reasonable despite extremely poor overlap (for the bound vanish leg, the highest off-diagonal overlap was 0.02). When the runtime constant was reduced by a factor of 5, the time allocations for the free leg increased by a factor of approximately $\sqrt{5}$ as expected, although the proportional increase was greater for the bound stages. Overall, this indicates that the algorithm is robust and behaves as expected when uncertainties decrease in proportion to the square root of the sampling time, as assumed. However, when simulations become trapped in local minima, the overall sampling time allocation can be affected by parameters other than just the runtime constant.

We note that the ability of the adaptive algorithm to detect sampling issues is highly dependent on using diverse starting conformations. We performed two adaptive runs of the T4L system with 5 replicate runs, a runtime constant of 0.0005 kcal^2 mol^{-2} ns^{-1}, and λ-windows spaced with a speed of 1 kcal mol^{-1}. The first protocol used the default of independently equilibrated (in the fully interacting state) structures for each repeat, while the second protocol used the input structure generated for the first repeat for all repeats. A dramatically greater allocation of sampling time for the bound vanish stage of the first protocol (458 ns) was observed compared to the second protocol (16 ns). This suggests that using different starting velocities alone may be insufficient to sample diverse local minima in the initial stages of the adaptive runs. However, the shared starting structure protocol produced an overall result (-5.26 ± 0.35 kcal mol^{-1}) in better agreement with experiment (-5.19 ± 0.16 kcal mol^{-1}) than the diverse starting structure protocol (-4.33 ± 0.67 kcal mol^{-1}), perhaps suggesting that entry of any waters to this binding site is an artefact of simulation.

Detecting equilibration based on multiple replicates improves reliability

We assessed our ensemble-based equilibration detection method based on its performance on the bound vanish stage for all initial test systems (Figure 3.10). These were run adaptively with the parameters given in Section 3.4.3, meaning that the adaptive window spacing and time allocation algorithms were used in combination with the equilibration detection heuristic. This stage was chosen as it displayed the most pronounced equilibration behaviour. In particular, the MIF and PDE2a complexes showed pronounced initial transients, which were successfully removed by the paired t-test method. For all other systems, any initial transients were substantially smaller and this method discarded no time to equilibration. Inspection of the equilibration times for all stages showed that the paired t-test method rarely discards any data when there is no visible initial transient (Figure B.22). This is in contrast to the popular method of Chodera which often discards a substantial amount of data even when there is no obvious initial transient, even when it is applied to the mean trace (see MDM2-Nutlin bound vanish in Figure 3.10, T4L bound restrain in Figure B.22). We compared the absolute differences between the ΔG estimates obtained with both equilibration detection methods to the final 30 ns non-adaptive bound vanish results (Figure B.23). In every case, the difference is smaller for the paired t-test, which could

be due to reduced variance arising from retaining more of the data. However, we have a small sample size of 5, and there is no evidence for a significant difference between the two methods at 95 % confidence based on the Wilcoxon signed-rank test (p=0.06).

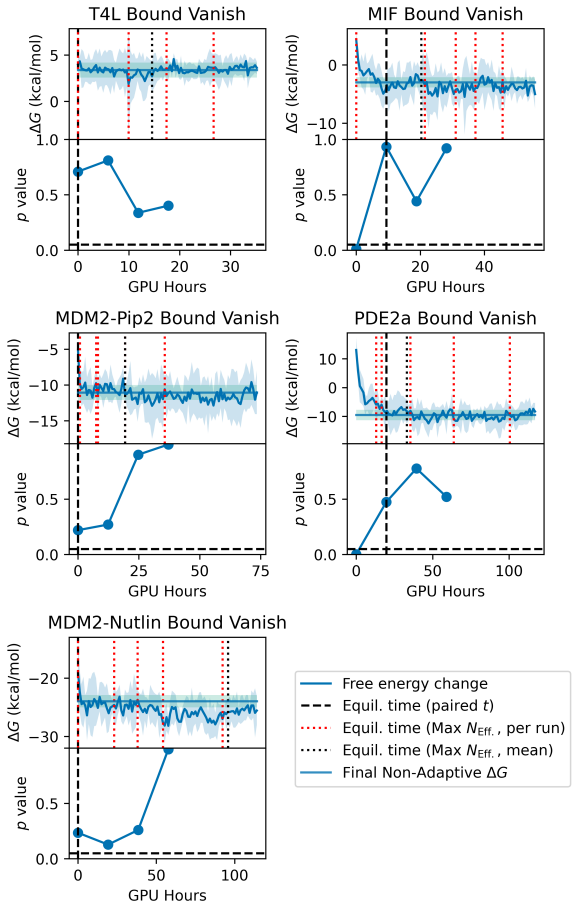


Figure 3.10: Selection of equilibration times for the bound vanish stage by the paired t-test method and Chodera's method (applied to both the mean trace and individual replicates).[262] The upper windows show traces of ΔG obtained by dividing the data up into 100 equal blocks and running MBAR on each. Shaded areas indicate 95 % inter-run t-based confidence intervals. Final Non-Adaptive ΔG are taken from the non-adaptive 30 ns runs. Lower windows show the p-values obtained by truncating the data up to the time shown and performing paired t-tests on the first 10 % and last 50 % of the remaining data. The first p-value > 0.05 is used as a heuristic to indicate equilibration.

The paired t-test method is only applicable to an ensemble of simulations, but Chodera's method has the advantage that it can be applied to single runs as well as the mean trace. However, applying Chodera's method to the individual runs for the bound vanish stages reveals substantial variability of the times discarded, generally ranging from no data discarded to nearly all data removed. This is

concerning because we care about removing systematic bias in our ensemble result, where the bias results from a relaxation process shared by all runs. We expect this to have a comparable timescale between runs, and therefore replicates should have similar equilibration times. It appears that explicitly using information about the inter-run reproducibility of the transient in the paired t method increases robustness. During initial testing, we found the paired t-test method to be substantially more sensitive to the presence of initial transients than an unpaired t-test on the initial and final portions of the data.

While the performance of the paired t-test method appeared promising, we note several potential limitations. Firstly, it is not a rigorous statistical test but a heuristic; $p > 0.05$ shows no significant evidence for drift of the free energy estimate, but we take it to mean that there is no drift of the free energy estimate ("The Fallacy of the Transposed Conditional").[293] The use of a parametric test may not be justified. Furthermore, like all equilibration detection methods we are aware of, this method requires that initial simulations have been run on a timescale comparable to that of the initial transient; very slow equilibration will not be apparent from very short simulations. Our workflow likely depends on the correlation between slow initial relaxation and large inter-run uncertainty, which means that sufficient sampling time is allocated to the bound vanish legs to allow the initial transients to be detected and removed. In addition, the method is sensitive to parameters such as the number of t-tests performed and the number of replicate runs. With increasing numbers of tests, the chances of falsely "detecting equilibration" early increase. This could be mitigated by taking the equilibration point to be the first test with $p < 0.05$ starting from the test with most truncation. With more replicate runs, the power of the test increases and a lack of equilibration is more likely to be detected.

One deficiency of Chodera's method is that it never explicitly signals lack of equilibration, but rather selects an optimal truncation point given the data that has already been collected. The paired t-test method could in principle detect a lack of equilibration, triggering the allocation of additional simulation time, although this was not observed for this test set.

3.4.3 Overall Performance of an Optimised Adaptive Protocol

"Optimal" parameters were selected for the adaptive algorithms based on the tests described above. We selected a runtime constant of 0.0005 $\text{kcal}^2\ \text{mol}^{-2}\ \text{ns}^{-1}$ as this generally produced runs of sufficient length to match the non-adaptive 30 ns results. We retained our default of 5 replicate runs as this seemed to represent a reasonable compromise between fast equilibration and reliable detection of sampling issues, although we did not investigate this in detail. Because we did not use replica exchange and hence did not have to consider replica mixing, we spaced λ-windows with a large thermodynamic speed of 2 kcal mol^{-1} to accelerate equilibration. Finally, we retained the default equilibration detection parameters as these appeared to deliver robust performance.

As an initial test of overall performance, we applied the adaptive protocol to the set of initial test systems. This is effectively a "training" set as we used it to select the adaptive parameters, and it is therefore not representative of prospective performance. Hence, we then assessed the "optimised" protocol on an unseen "test" set.

The adaptive protocol accelerates the equilibration of the initial test set

For all systems, only two λ windows were selected for turning on the restraints, but high overlap (> 0.2) was observed for this stage (Figure B.24). This reflects the fact that the restraint selection algorithm selects restraints which minimally perturb the ligand's native interactions. As expected from the $\sigma\left(\frac{\partial H}{\partial \lambda}\right)$ plot (Figure B.4), similar λ-schedules were usually generated for the same transformation in the free and bound legs, meaning that reasonable schedules for the bound leg may be determined from the cheaper free leg initial test simulations. However, there were exceptions - for example, more windows were used for the PDE2a free discharge leg than the bound discharge leg due to the greater variance of the gradient observed in the free leg. A detailed breakdown of the sampling time allocations is provided in Section B.18. Generally, the compute time was concentrated in the bound discharge and especially the bound vanish stages, and increased with increasing ligand size.

The "optimised" adaptive protocol produced extremely similar results to the long time 30 ns and manually-optimised 6 ns non-adaptive protocols (Table 3.3). This was despite using over 3 times less GPU time than the 6 ns protocol, over 14 times less than the 30 ns protocol (Table B.5), and requiring no manual tweaking of the equilibration times of λ-schedules. We note that this is not an entirely fair test,

Table 3.3: Predicted $\Delta G^{o}_{\mathrm{Bind}}$ for Initial Test Systems with Adaptive and Long-Run Non-Adaptive Protocols[a]

	"Optimised" Adaptive	30 ns Non-Adaptive
T4L	-4.28 ± 0.73	-4.03 ± 0.80
MIF	-10.54 ± 0.78	-10.48 ± 1.14
MDM2-Pip2	-13.98 ± 1.67	-13.94 ± 0.86
PDE2a	-18.36 ± 1.49	-16.96 ± 1.97
MDM2-Nutlin	-16.14 ± 1.92	-16.92 ± 1.98

[a] All quantities in kcal mol^{-1}. Calculation uncertainties stated as 95 % confidence intervals based on the variance of 5 replicate runs, assuming Gaussian distributions. Symmetry corrections are included (Section B.5). A detailed breakdown of the components of the calculated binding free energies is given in Table B.4.

as we used some of these systems to optimise the adaptive algorithm parameters in the previous sections. However, this provides further reassurance that the parameters and overall protocol are robust, and shows that the adaptive protocol produces equivalent results to a manually-optimised non-adaptive protocol at lower computational cost.

We may also have tried to save compute simply by shortening the non-adaptive runs. To investigate the feasibility of this, we compared the free energy estimates obtained from the adaptive and non-adaptive protocols at equivalent computational time, ignoring equilibration and retaining all data. Figure 3.11 shows that the adaptive protocol produces substantial increases in the rate of equilibration for the bound vanish stages for MIF and PDE2a, the two systems with the most pronounced initial transients. As these stages make the dominant contributions to overall calculation cost, this produces accelerated equilibration of the entire dataset (Figure B.31). Because the allocated sampling times are not sharply peaked for these stages (Figure B.26), we attribute this improvement to the wider λ spacing. Manually selecting such widely-spaced windows while retaining sufficient overlap would be extremely labour-intensive, illustrating the potential of the adaptive protocol to minimise human, as well as computer time. Even if the non-adaptive protocols were run for the same total compute time as the adaptive protocol for the bound vanish stages, Figure 3.11 shows very long equilibration times would be required for MIF and PDE2a. Identifying the requirement for such long equilibration times would be time-consuming; alternatively, uniformly

applying very long equilibration times would be wasteful. In contrast, the adaptive procedure only discards data to equilibration for the problematic MIF and PDE2a systems (Figure 3.10). This suggests inherent advantages to the automated, adaptive protocol even when the total computational cost is equivalent.

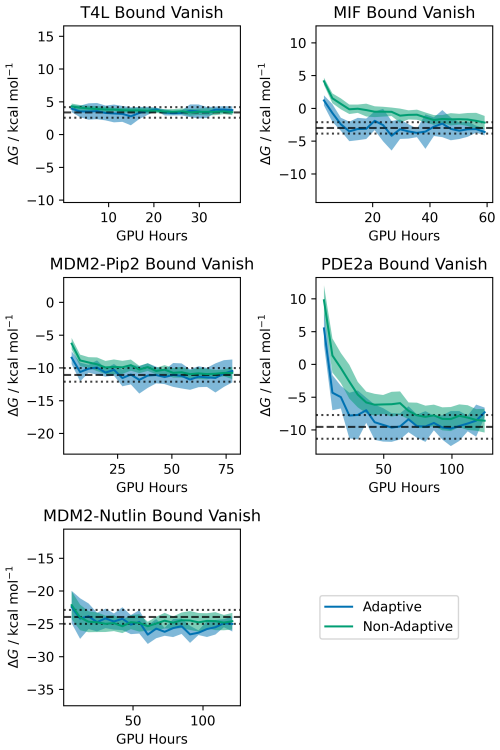


Figure 3.11: Estimated ΔG against sampling time for the bound vanish stages of the initial test systems. Data were split into 20 equal blocks and MBAR was run on each block. "adaptive" refers to the "optimised" adaptive protocol, and the non-adaptive results were obtained by averaging the non-adaptive 6 ns and 30 ns runs to more clearly show the differences in equilibration behaviour. The data for the non-adaptive runs was truncated before analysis so that the per-stage computational costs were equal to the adaptive calculation. For this reason the 95 % t-based confidence intervals (shaded areas) are expected to be a factor of $\sqrt{2}$ smaller for the non-adaptive protocol. The final non-adaptive ΔG is taken from the 30 ns non-adaptive result (black dotted line with 95 % CIs shown by sparser dotted lines). The y-axis spacings are the same for all plots. Plots for all stages are shown in Section B.20.

Although the adaptive algorithm produced improvements in the rate of equilibration, we found no clear differences between the uncertainty in the free energy estimates at equivalent computational effort either within stages (Figure B.30), or over entire calculations (Figure B.32). This likely reflects two facts: that the $\sqrt{t_\lambda}\sigma\left(\left\langle\frac{\partial H}{\partial\lambda}\right\rangle_\lambda\right)$-based improvement factors are fairly close to 1 (Figure B.5), indicating little scope for improvement with adaptive time allocation; and

that allocating time to problematic windows often does not reduce uncertainties because individual runs become stuck in local minima, indicating that the timescales of sampling issues are too long. An alternative approach would be to allocate more replicates to problematic λ windows, but this would reduce the rate of equilibration.

Despite the lack of uncertainty reduction, the algorithm successfully allocated sufficient time to achieve equilibration in the bound vanish legs, while allocating dramatically less time to most free stages. This again suggests a correlation between slow equilibration and large inter-run uncertainty.

The adaptive protocol shows robust performance on an "unseen" test set

We sought to test the robustness of our workflow using a set of new systems which were not used to select the "optimal" workflow parameters. We selected the Cyclophilin-D set from Alibay et al.'s study of ABFE in the fragment optimisation process.[1] Excluding ligand 4 (because its affinity fell outside the SPR detection limits) produced a set of 9 fragments and merged molecules with a wide dynamic range.[294] Alibay et al.'s results were re-analysed and plotted excluding this ligand in Section B.21.

First, we performed non-adaptive runs for all ligands using the same number of λ-windows for each stage as Alibay et al.. We allocated 5 ns per window with 1 ns discarded to equilibration, which reflects common practice in the literature. Reassuringly, the non-adaptive SOMD runs produced a very similar correlation with experiment compared to the GROMACS runs of Alibay et al. (R^2 values of $0.81^{0.91}_{0.46}$ and $0.79^{0.95}_{0.38}$ (upper and lower 95 % CIs), respectively - see Table B.6), despite the allocation of around 4 times less simulation time. As found by Alibay et al., our predictions were systematically more negative than experiment (Table 3.4). A more detailed analysis and comparison to the results of Alibay et al. is given in Section B.21.

Recalculating all free energies with the "optimised" adaptive protocol produced extremely similar results to the non-adaptive protocol (Figure 3.12 and Table 3.4) with good correlation with experiment, albeit for a small set of ligands. There was no significant evidence at 95 % confidence for differences in either the uncertainty between the adaptive and non-adaptive runs, nor the unsigned error compared to experiment ($p = 0.36$ in both cases using the Wilcoxon ranked-sign test). A relatively large amount of simulation time was allocated to these calculations using the same parameters as for the initial test systems, because the bound systems

Table 3.4: Predicted $\Delta G^{o}_{\text{Bind}}$ for Cyclophilin D [a]

	Adaptive	Non-adaptive 5 ns	Alibay	Exp. $\Delta G^{o}_{\text{Bind}}$
Ligand 2	-10.56 ± 1.06	-10.52 ± 1.21	-8.18 ± 0.76	-9.06 ± 0.50
Ligand 3	-5.06 ± 2.78	-4.73 ± 1.21	-4.71 ± 0.27	-2.93 ± 0.50
Ligand 4	-4.92 ± 1.92	-6.93 ± 1.79	-4.14 ± 1.00	-2.90 ± 0.50
Ligand 8	-7.30 ± 0.73	-7.34 ± 1.06	-7.24 ± 0.73	-4.04 ± 0.50
Ligand 14	-14.38 ± 2.17	-15.44 ± 1.39	-12.92 ± 0.54	-11.22 ± 0.50
Ligand 16	-9.27 ± 1.79	-10.49 ± 2.04	-10.54 ± 0.60	-8.42 ± 0.50
Ligand 27	-9.14 ± 1.67	-10.81 ± 1.42	-10.21 ± 1.36	-7.57 ± 0.50
Ligand 39	-13.99 ± 1.97	-13.58 ± 1.09	-12.62 ± 0.59	-8.43 ± 0.50
Ligand 40	-14.72 ± 0.88	-14.51 ± 0.81	-11.78 ± 0.65	-8.08 ± 0.50

[a] All quantities in kcal mol^{-1}. Calculation uncertainties stated as 95 % t-based confidence intervals based on the variance of 5 replicate runs, assuming Gaussian distributions. "Alibay" results were taken from Alibay et al.,[1] and the experimental results were taken from Grädler et al., who used surface plasmon resonance.[294] The experimental uncertainties were assumed to be 0.5 kcal mol^{-1}. A detailed breakdown of the components of the calculated binding free energies is given in Table B.7.

were relatively small and inexpensive (around half the cost of the MIF bound system). However, the adaptive protocol was still around 1.5 times cheaper than the non-adaptive protocol, and around 6 times cheaper than the protocol of Alibay et al.. Unlike the MIF and PDE2a systems, these complexes equilibrated quickly, meaning that the accelerated equilibration afforded by the adaptive protocol did not provide a substantial advantage (Figure B.40). However, the fact that the adaptive protocol produced equivalent results to the non-adaptive protocol on an "unseen" test set suggests that it is robust and that our "optimal" parameters are likely to perform well for other systems. Furthermore, the adaptive protocol performed sensible optimisations without any human intervention, for example reducing the number of λ windows for the restrain stage from 12 (which produced very high overlap) to 2. A more detailed analysis is given in Section B.22.

3.5 Conclusion

We have presented an automated workflow for ABFE calculations based on the automated selection of λ windows, the ensemble-based detection of equilibration, and the adaptive allocation of sampling time based on inter-replicate uncertainties in the per-window free energy estimates. Our central conclusions are:

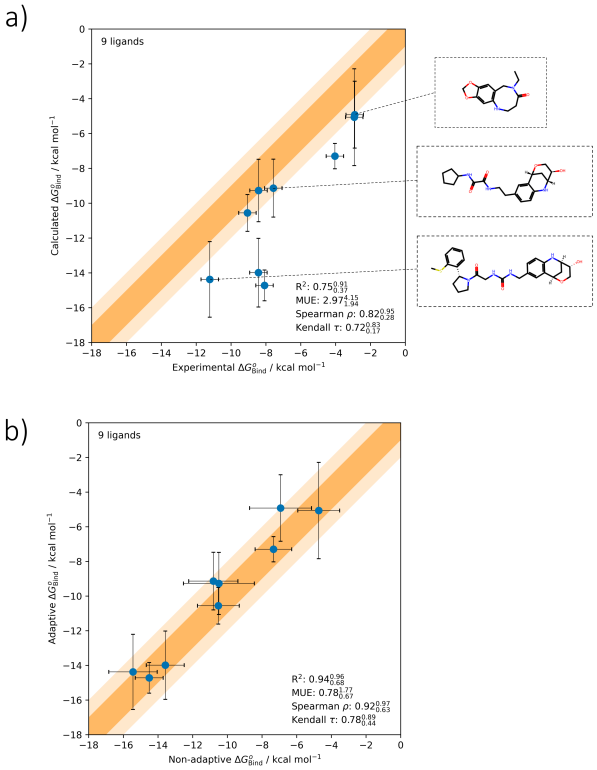


Figure 3.12: a) Experimental free energies of binding for Cyclophilin D against predicted free energies of binding obtained using the "optimised" adaptive protocol and b) comparison of results obtained using the "optimised" adaptive protocol and the non-adaptive protocol. Selected ligands are shown in a) to illustrate the structural diversity. Experimental free energies were obtained from Grädler et al..[294] The darker and lighter shaded areas show 1 and 2 kcal mol^{-1} deviations from exact agreement, respectively. Error bars show 95 % confidence intervals, which were assumed to be 0.5 kcal mol^{-1} for experiment and calculated from the deviation between 5 replicate runs for the predictions. 95 % confidence intervals on statistics were calculated by bootstrapping with 10000 iterations of resampling.

(1) The automated selection of λ windows to give consistent overlap can be achieved using very short initial test simulations with non-optimal spacing. This is cheap, robust, and simple to implement. Increasing the spacing of windows can accelerate equilibration by reducing the concentration of total sampling time at the start of simulations.

(2) An ensemble-based equilibration detection heuristic based on a paired t-test between the free energy estimates at initial and final portions of a run appears robust.

(3) Adaptively allocating sampling time based on the uncertainty in free energy changes between repeat runs increased sampling times where there were sampling issues. The bound vanish leg was usually allocated the most computing time, because it usually involved the most severe sampling issues. In a rare case where the sampling issue had a timescale amenable to simulation, this algorithm accelerated equilibration. However, we generally found no evidence that the adaptive time allocation algorithm reduced uncertainty for equivalent computational cost, likely because the timescales of sampling issues could not be matched by our simulations.

(4) We found reasonable default parameters for all algorithms and tested the performance of our overall workflow on a range of systems. In all cases, the adaptive protocol produced free energy estimates equivalent to the non-adaptive protocols. For systems which showed slow equilibration, the adaptive protocol substantially increased the rate of equilibration, likely due to the wide spacing of λ windows.

(5) We have provided an open-source implementation of all algorithms in the Python package A3FE, which is available on GitHub (github.com/michellab/a3fe).

We hope this work helps to facilitate the more efficient, automated calculations which will be required to unlock the potential of ABFE calculations for drug discovery.

Here, we have taken the narrow view that the optimal ABFE workflow would produce minimal overall uncertainty in all free energy estimates for a fixed computational cost. However, practitioners in industrial drug discovery may not need precise estimates of affinities for weak binders, so long as they are correctly identified as weak. It is also likely more useful to accurately predict the affinities of dissimilar ligands with distinct receptor interactions than more similar ligands for which ΔG^o_{Bind} may be predicted with more confidence using a machine learning model trained on previous ABFE results. Future work may take a more realistic view of the "optimal" workflow, for example by quickly terminating simulations which show weak binding. This is likely to be robust because, as we have discussed, unequilibrated free energy estimates tend to overestimate affinity. Tighter coupling between free energy protocols and active learning protocols may also be useful - for example, the number of replicas and simulation time may be increased in proportion to the uncertainty of a machine-learned predictor of affinity.

Chapter 4

Robust Automated Truncation Point Selection

Chapters 2, and especially 3, presented algorithms to automate various decisions in ABFE calculations. These were implemented in the open-source software A3FE, which was shown to provide comparable or better performance to non-adaptive workflows while saving user time. These algorithms included a replicate run-based heuristic for equilibration detection, which was less prone to over-discarding data than Chodera's popular method.[108] However, it was not possible to draw quantitative conclusions on the performance of the heuristics because there was limited ABFE data for testing, and it would have been prohibitively expensive to generate much more. In addition, the heuristic proposed was not applicable to single runs, which are often used in practice.

In this chapter, White's marginal standard error rule is reformulated to provide a spectrum of truncation point selection heuristics that differ in their treatment of autocorrelation.[295] These include a method effectively equivalent to Chodera's, and are all applicable to single runs. To allow quantitative conclusions to be drawn, these methods are tested on ensembles of synthetic time series modelled on free energy change estimates from the long absolute binding free energy calculations in Chapter 3. It is shown that methods that more thoroughly account for autocorrelation often show late and variable truncation times, while methods that less thoroughly account for autocorrelation often show early truncation, relative to the optimal truncation point. This increases variance and bias, respectively. A method that achieves robust performance across our test sets by balancing

these two extremes is recommended. Since none of the methods reliably detect insufficient sampling, we refer to truncation point selection, rather than equilibration detection. All heuristics tested are implemented in the open-source Python package RED (github.com/fjclark/red).

The remainder of the chapter is included unmodified as published:

Clark, F.; Cole, D. J.; Michel, J. Robust Automated Truncation Point Selection for Molecular Simulations. *J. Chem. Theory Comput.* **2025**, *21*, 88-101. doi.org/10.1021/acs.jctc.4c01359.

4.1 Introduction

Quantities calculated from molecular simulations are often subject to initial bias due to unrepresentative starting configurations. The resultant systematic error can be reduced by truncating data from the start of the simulation and calculating the quantity of interest using the remaining "equilibrated" data. However, discarding too many data unnecessarily increases random error. It is common practice to select a fixed truncation point, which is unlikely to be optimal across many simulations, or to select by visually inspecting the data, which is only feasible for a few simulations. Robust automated truncation point selection methods are required to minimise errors in calculated quantities while facilitating automation. Here we address the selection of an optimal truncation point given some data, and not the much harder problem of detecting when sufficient data have been collected (so that enough representative configurations of the equilibrium distribution have been sampled). Therefore, we refer to "truncation point selection" and not "equilibration detection", which might be expected to refer to the latter problem. We define the optimal truncation point to be that minimising the root mean square error (RMSE) compared to the value which would be obtained with infinite sampling.

Two common heuristics for automated truncation point selection in molecular simulations are those of Yang et al.[109] and Chodera.[108] Yang et al. proposed a method based on reverse cumulative averaging, where a final "equilibrated" portion of the data are assumed to be Gaussian. The "equilibrated" region is extended in the reverse direction until significant deviation from this distribution is detected, indicating "unequilibrated" data. However, the assumption of normality is not always justified,[108,193] and the robust selection of an intial "equilibrated" region may be challenging. Chodera suggested selecting the truncation point

which maximises the effective sample size, demonstrating the method on a simple system with rapid convergence and short correlation times. However, this method is prone to selecting late truncation points for more correlated data from more realistic applications.[296,297]

There is a substantial literature on truncation point selection outside the field of molecular simulation.[265,295,298,299] In particular, the marginal standard error rule (MSER) family of methods has been found to be simple, effective, and easily automatable.[295,298,300] These methods select the truncation point by minimising the marginal standard error, and are very similar to Chodera's method of maximising the effective sample size; the main difference is that the MSER methods popular in operational research account for correlation less thoroughly, if at all, and their applicability to correlated molecular simulation data is uncertain.

Here, we compare the performance of a spectrum of MSER methods which apply increasingly rigorous techniques to account for correlation. These include the original MSER method and a method which is effectively identical to Chodera's heuristic.[108,295] To rigorously assess their performance, we test these methods on synthetic data modelled on long absolute binding free energy calculations for a variety of protein-ligand complexes. This work complements the recent study of Oliveira et al. by testing a wide range of MSER methods and quantitatively assessing their performance.[301] We limit ourselves to the analysis of single runs, but note that truncation point selection and uncertainty quantification based on an ensemble of repeat runs are likely to be more robust, if more expensive.[296] All methods discussed are implemented in the open-source Python package RED (Robust Equilibration Detection, github.com/fjclark/red, where equilibration is used in the sense of finding the optimal truncation point). This provides alternatives for the PyMBAR timeseries "detect_equilibration" function.[108,302]

4.2 Theory

4.2.1 Bias and Standard Deviation

As discussed by Chodera,[108] we are usually interested in the ensemble average of a quantity $A(\boldsymbol{x})$

$$\langle A \rangle_\pi = \int A(\boldsymbol{x})\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \tag{4.1}$$

where $\boldsymbol{x}$ is the vector of the system's phase-space coordinates and $\pi(\boldsymbol{x})$ is the probability of observing $\boldsymbol{x}$ in the ensemble of interest. $\langle A \rangle_\pi$ is often estimated with $\langle A \rangle_{[n_0,N]}$, the average of a series of samples from a simulation

$$\langle A \rangle_{[n_0,N]} = \frac{1}{N_{n_0}} \sum_{n=n_0}^{N} A(\boldsymbol{x}_n), \tag{4.2}$$

where N is the total number of samples, n_0 is the number of the first sample used in the average (earlier samples are discarded), $N_{n_0} = N - n_0 + 1$, and $\boldsymbol{x}_n$ is the vector of phase-space coordinates for the nth sample of the system. Assuming ergodicity, in the limit of an infinite number of samples, N_{n_0}, $\langle A \rangle_{[n_0,N]} = \langle A \rangle_{[n_0,\infty]} = \langle A \rangle_\pi$.[99] With finite sampling, $\langle A \rangle_{[n_0,N]}$ is an approximation of $\langle A \rangle_\pi$ with an associated error. The expected root-mean-squared error (RMSE) of the estimate can be separated into contributions from bias and variance

$$\begin{aligned} \mathrm{RMSE}_{\mathrm{Trajs}}(\langle A \rangle_{[n_0,N]}) &= \sqrt{\left\langle (\langle A \rangle_{[n_0,N]} - \langle A \rangle_\pi)^2 \right\rangle_{\mathrm{Trajs}}} && (4.3) \\ &= \sqrt{\mathrm{Var}_{\mathrm{Trajs}}(\langle A \rangle_{[n_0,N]}) + \mathrm{Bias}^2_{\mathrm{Trajs}}(\langle A \rangle_{[n_0,N]})} && (4.4) \\ &= \sqrt{\mathrm{SD}^2_{\mathrm{Trajs}}(\langle A \rangle_{[n_0,N]}) + \mathrm{Bias}^2_{\mathrm{Trajs}}(\langle A \rangle_{[n_0,N]})}, && (4.5) \end{aligned}$$

where $\langle ... \rangle_{\mathrm{Trajs}}$ indicates an average over the ensemble of all possible simulation trajectories with initial phase-space coordinates $\boldsymbol{x_0}$ taken from some allowed set (often with the same positions). $\mathrm{Var}_{\mathrm{Trajs}}(\langle A \rangle_{[n_0,N]})$, $\mathrm{Bias}_{\mathrm{Trajs}}(\langle A \rangle_{[n_0,N]})$, and $\mathrm{SD}_{\mathrm{Trajs}}(\langle A \rangle_{[n_0,N]})$ denote the variance, bias, and standard deviation of $\langle A \rangle_{[n_0,N]}$ over the ensemble of trajectories. The (well known) full derivation of Equation 4.5 is given in Section C.1.[108] Explicitly, the expected bias and variance over the ensemble of trajectories are

$$\begin{aligned} \mathrm{SD}_{\mathrm{Trajs}}(\langle A \rangle_{[n_0,N]}) &= \sqrt{\left\langle (\langle A \rangle_{[n_0,N]} - \langle \langle A \rangle_{[n_0,N]} \rangle_{\mathrm{Trajs.}})^2 \right\rangle_{\mathrm{Trajs}}} && (4.6) \\ \mathrm{Bias}_{\mathrm{Trajs}}(\langle A \rangle_{[n_0,N]}) &= \langle \langle A \rangle_{[n_0,N]} \rangle_{\mathrm{Trajs}} - \langle A \rangle_\pi. && (4.7) \end{aligned}$$

These terms oppose each other - choosing an early truncation point (low n_0) will produce a larger $\mathrm{Bias}_{\mathrm{Trajs}}$ due to the inclusion of early unrepresentative samples, and a smaller $\mathrm{SD}_{\mathrm{Trajs}}$ due to the inclusion of more samples (assuming similar starting configurations). Conversely, late truncation points reduce $\mathrm{Bias}_{\mathrm{Trajs}}$ but increase $\mathrm{SD}_{\mathrm{Trajs}}$. When truncating simulation data, the objective is to minimise $\mathrm{RMSE}_{\mathrm{Trajs}}$ by balancing $\mathrm{Bias}_{\mathrm{Trajs}}$ and $\mathrm{SD}_{\mathrm{Trajs}}$.

4.2.2 Heuristics for Truncation Point Selection

We take "optimal" to mean "minimising $\mathrm{RMSE}_{\mathrm{Trajs}}(\langle A\rangle_{[n_0,N]})$", and we aim to select the optimal truncation point. This is impossible in practice because $\mathrm{RMSE}_{\mathrm{Trajs}}(\langle A\rangle_{[n_0,N]})$ as a function of n_0 is unknown. However, many heuristics have been proposed. White's marginal standard error rule (MSER) selects an n_0 by minimising the marginal standard error of the mean of A,[295] and is usually presented without explicitly accounting for autocorrelation

$$n_0* = \underset{N>n_0>0}{\operatorname{argmin}} \left[\frac{1}{N_{n_0}^2} \sum_{n=n_0}^{N} (A(\boldsymbol{x_n}) - \langle A\rangle_{[n_0,N]})^2 \right], \tag{4.8}$$

where n_0* is the selected n_0. We present a more general formulation of MSER where autocorrelation may be explicitly accounted for

$$n_0* = \underset{N>n_0>0}{\operatorname{argmin}} \left[\widehat{\mathrm{SD}}_{\mathrm{Trajs}}(\langle A\rangle_{[n_0,N]}) \right] \tag{4.9}$$

$$= \underset{N>n_0>0}{\operatorname{argmin}} \left[\widehat{\mathrm{Var}}_{\mathrm{Trajs}}(\langle A\rangle_{[n_0,N]}) \right], \tag{4.10}$$

where the hat operator indicates an estimate made from the available trajectory (or trajectories). Equations 4.9 and 4.10 are equivalent because the same truncation point minimises the variance and the standard deviation. A naïve estimator of $\mathrm{Var}_{\mathrm{Trajs}}$ is

$$\widehat{\mathrm{Var}}_{\mathrm{Trajs,\ Naïve}}(\langle A\rangle_{[n_0,N]}) = \frac{1}{N_{n_0}} \sum_{t=-(N_{n_0}-1)}^{N_{n_0}-1} \hat{\gamma}_{t,[n_0,N]} \tag{4.11}$$

$$= \frac{1}{N_{n_0}} \left(\hat{\gamma}_{0,[n_0,N]} + 2 \sum_{t=1}^{N_{n_0}-1} \hat{\gamma}_{t,[n_0,N]} \right), \tag{4.12}$$

where $\hat{\gamma}_{t,[n_0,N]}$ is the estimate of the autocovariance at lag t. This measures the degree of linear dependence between values in a time series separated by t samples

$$\gamma_t = \gamma_{-t} = \mathrm{Cov}(A(\boldsymbol{x}_n), A(\boldsymbol{x}_{n+t})) \tag{4.13}$$

$$\hat{\gamma}_{t,[n_0,N]} = \frac{1}{N_{n_0}} \sum_{n=n_0}^{N-t} (A(\boldsymbol{x}_n) - \langle A\rangle_{[n_0,N]})(A(\boldsymbol{x}_{n+t}) - \langle A\rangle_{[n_0,N]}). \tag{4.14}$$

Cov denotes covariance. We have assumed time-reversibility, hence $\gamma_t = \gamma_{-t}$ and Equation 4.11 simplifies to Equation 4.12.

It is helpful to compare Equation 4.12 to the familiar variance of the mean formula which assumes that samples are uncorrelated

$$\widehat{\mathrm{Var}}_{\mathrm{Trajs,Uncor}}(\langle A\rangle_{[n_0,N]}) = \frac{1}{N_{n_0}^2}\sum_{n=n_0}^{N}(A(\boldsymbol{x_n}) - \langle A\rangle_{[n_0,N]})^2 \tag{4.15}$$

$$= \frac{1}{N_{n_0}}\widehat{\mathrm{Var}}_{[n_0,N]}(A(\boldsymbol{x})) \tag{4.16}$$

$$= \frac{1}{N_{n_0}}\hat{\gamma}_{0,[n_0,N]}. \tag{4.17}$$

This shows that the correlated variance estimate should always be larger than the uncorrelated estimate (when the correlations are positive). This is due to the autocovariance terms for lags greater than 0, which are missing from Equation 4.17. When the uncorrelated variance estimate is used in the general MSER equation (Equation 4.10), the traditional MSER equation (Equation 4.8) is recovered. However, samples from molecular simulations are often strongly positively autocorrelated, meaning that Equation 4.15 underestimates the true variance of the mean.

For simplicity, we do not correct for bias in our autocovariance estimates arising because the true mean is unknown.[303,304] Additionally, we do not correct autocovariance terms for the finite size of the time series used to estimate them (we divide by N_{n_0} although there are only $N_{n_0} - t$ terms in the sum). While this means that $\langle\hat{\gamma}_{t,[n_0,N_a]}\rangle_{\text{Trajs.}} \neq \langle\hat{\gamma}_{t,[n_0,N_b]}\rangle_{\text{Trajs.}} \neq \langle\hat{\gamma}_{t,[n_0,\infty]}\rangle_{\text{Trajs.}}$, where $N_a \neq N_b$, they will be very close to equal for the dominant autocovariance terms where $t \ll N_{n_0}$, and the lack of correction allows the variance of $\langle A\rangle_{[n_0,N]}$ to be estimated by summing the uncorrected $\hat{\gamma}_{t,[n_0,N]}$ terms (Section 4.2.3).

$\widehat{\mathrm{Var}}_{\mathrm{Trajs}}(\langle A\rangle_{[n_0,N]})$ can be thought of as a surrogate for $\mathrm{RMSE}_{\mathrm{Trajs}}(\langle A\rangle_{[n_0,N]})$ which can be calculated with data from a single simulation; by picking the truncation point which minimises $\widehat{\mathrm{Var}}_{\mathrm{Trajs}}(\langle A\rangle_{[n_0,N]})$, we hope to minimise $\mathrm{RMSE}_{\mathrm{Trajs}}(\langle A\rangle_{[n_0,N]})$. This is reasonable because at sufficiently late truncation points when the bias is negligible, often $\widehat{\mathrm{Var}}_{\mathrm{Trajs}}(\langle A\rangle_{[n_0,N]})$ becomes a good estimator of the $\mathrm{Var}_{\mathrm{Trajs}}(\langle A\rangle_{[n_0,N]})$ component of $\mathrm{RMSE}_{\mathrm{Trajs}}(\langle A\rangle_{[n_0,N]})$, which penalises late truncation. For early truncation points, increasing bias increases $\widehat{\mathrm{Var}}_{\mathrm{Trajs}}(\langle A\rangle_{[n_0,N]})$, mimicking the $\mathrm{Bias}_{\mathrm{Trajs}}(\langle A\rangle_{[n_0,N]})$ contribution to $\mathrm{RMSE}_{\mathrm{Trajs}}(\langle A\rangle_{[n_0,N]})$ which favours late truncation. This relies on the violation of the time-reversibility assumption by the bias at early times to inflate $\widehat{\mathrm{Var}}_{\mathrm{Trajs}}(\langle A\rangle_{[n_0,N]})$, whereas $\mathrm{Var}_{\mathrm{Trajs}}(\langle A\rangle_{[n_0,N]})$ may not increase with increasing bias.

Chodera proposed choosing the truncation point which maximises the effective sample size, which can be expressed as

$$n_0* = \underset{N>n_0>0}{\operatorname{argmin}} \left[\frac{\widehat{\operatorname{Var}}_{\text{Trajs}}(\langle A \rangle_{[n_0,N]})}{\widehat{\operatorname{Var}}_{[n_0,N]}(A(\boldsymbol{x}))} \right], \tag{4.18}$$

which is very similar to the general MSER (Equation 4.10). It differs only by the factor of $\frac{1}{\widehat{\operatorname{Var}}_{[n_0,N]}(A(\boldsymbol{x}))}$. This will decrease sensitivity to bias compared to MSER, because an initial transient will increase $\widehat{\operatorname{Var}}_{[n_0,N]}(A(\boldsymbol{x}))$, but this effect is expected to be small when autocorrelation is thoroughly accounted for (by considering all important terms in the autocorrelation function), as in Chodera's work.[108] Hence, the main difference between MSER and Chodera's method lies in the calculation of $\widehat{\operatorname{Var}}_{\text{Trajs}}(\langle A \rangle_{[n_0,N]})$. Therefore, we use Equation 4.10 to select truncation points throughout this work, and compare different methods to calculate $\widehat{\operatorname{Var}}_{\text{Trajs}}(\langle A \rangle_{[n_0,N]})$.

4.2.3 Calculation of the Variance of the Mean

All truncation point selection heuristics tested in this work are special cases of the general MSER equation (Equation 4.10). They differ only in the way the variance of the mean is estimated. Calculating $\widehat{\operatorname{Var}}_{\text{Trajs}}(\langle A \rangle_{[n_0,N]})$ by adding all autocovariance terms (Equation 4.12) is problematic because at long lag times, these become dominated by noise. This makes $\widehat{\operatorname{Var}}_{\text{Trajs}}(\langle A \rangle_{[n_0,N]})$ noisy.[305] Instead, we consider a spectrum of methods which differ in the number and weighting of terms in the autocovariance sum they consider.

As discussed previously, the simplest method is simply to ignore all $\hat{\gamma}_{t,[n_0,N]}$ terms when $t > 0$, as in the original MSER (Equation 4.15). This will substantially underestimate the variance of the mean when A is highly correlated.

Lag window estimators of the variance of the mean take the form

$$\widehat{\operatorname{Var}}_{\text{Trajs,Window}}(\langle A \rangle_{[n_0,N]}) = \frac{1}{N_{n_0}} \left(\hat{\gamma}_{0,[n_0,N]} + 2 \sum_{t=1}^{N_{n_0}-1} w(t, N_{n_0}) \hat{\gamma}_{t,[n_0,N]} \right), \tag{4.19}$$

where $w(t, N_{n_0})$ is a weight function ("lag window"), which may depend on N_{n_0}. $w(t, N_{n_0})$ is greater than or equal to 0 and usually no larger than 1. This reduces the noise from later autocovariance terms by down-weighting them.[306] Larger lag windows will generally underestimate the autocovariance less, at the expense of more noise.

A closely related method is batch means.[307] The data are split into blocks of size b, and the variance of the overall mean is estimated from the variance of the block means

$$\widehat{\mathrm{Var}}_{\mathrm{Trajs,Batch}}(\langle A\rangle_{[n_0,N]}) = \frac{1}{N_b^2}\sum_{j=1}^{N_b}(\langle A\rangle_{[n_0+(j-1)b,n_0+jb-1]} - \langle A\rangle_{[n_0,N]})^2, \qquad (4.20)$$

where $N_b = N_{n_0}/b$ is the number of blocks, assuming b is a factor of N_{n_0}. This relies on the fact that the batch means become uncorrelated as $b \to \infty$. This method implicitly accounts for autocorrelation in the MSER-m family of methods, where the data are block-averaged using blocks of size m before MSER is applied.[300,308] However, using overlapping batches, rather than the non-overlapping batches in Equation 4.20, yields a less noisy variance estimate.[309,310] The overlapping batch means variance estimate is effectively equivalent to a lag window estimate (Equation 4.19) using a Bartlett lag window (a kind of triangular window) of the appropriate size,[311] and therefore Equation 4.20 can be thought of as a noisy (if cheaper) special case of Equation 4.19. Hence, we do not consider the batch means method. The circular bootstrap and moving blocks bootstrap methods are also asymptotically equivalent to the overlapping batch means method.[134,312–314]

A final class of methods terminates the sum of autocovariance terms according to Markov chain-specific criteria. These are Geyer's initial sequence estimators.[305] Geyer noted that for stationary, reversible, and irreducible Markov chains, the sequence of sums of pairs of covariance terms

$$\Gamma_m = \gamma_{2m} + \gamma_{2m+1} \qquad (4.21)$$

is positive, decreasing, and convex. Geyer's initial positive, monotone, and convex sequence estimators impose each of these properties (and all preceding properties), respectively. For example, the initial positive sequence (IPS) estimator of the variance of the mean is

$$\widehat{\mathrm{Var}}_{\mathrm{Trajs,IPS}}(\langle A\rangle_{[n_0,N]}) = \frac{1}{N_{n_0}}\left(-\hat{\gamma}_{0,[n_0,N]} + 2\sum_{m=0}^{M}\hat{\Gamma}_{m,[n_0,N]}\right) \tag{4.22}$$

$$\hat{\Gamma}_{m,[n_0,N]} = \hat{\gamma}_{2m,[n_0,N]} + \hat{\gamma}_{2m+1,[n_0,N]}, \tag{4.23}$$

where M is the largest value of m for which $\hat{\Gamma}_{m,[n_0,N]} > 0, m \leq M$. Chodera used a similar method where the autocovariance sum is truncated at the first negative autocovariance term.[108,302] Of the methods discussed, these estimators tend to produce the largest estimates of autocorrelation and hence the largest and most realistic estimates of the variance of the mean. While many other schemes exist to estimate the autocovariance function and variance of the mean,[110,134,315–323] a detailed comparison is beyond the scope of this study.

Finally, we note that all single-run estimates of the variance of the mean are likely to be severe underestimations, due to the tendency of molecular simulation trajectories to become trapped in local minima.[111]

4.2.4 Assessing Truncation Point Selection Algorithms

A variety of metrics have been used to assess truncation point selection algorithms in the operations research literature, including percentage bias removed by truncation, the closeness of the truncation point to the first unbiased point, the coverage of the true mean by the calculated confidence intervals, and the variability of the truncation point.[298,308,324] However, we believe that the most natural metric is the RMSE of the estimated means to the true mean over an ensemble of test trajectories.[108] This most closely reflects our intentions when using these heuristics - to minimise error to the true value.

For these RMSEs to be meaningful, we need to test with ensembles of trajectories which reflect real use cases, and we must know, or have good estimates of, $\langle A\rangle_\pi$ for observables of interest. This is not straightforward with real simulation data; systems complex enough to be interesting often require long simulations to remove bias and accurately estimate $\langle A\rangle_\pi$, and are often expensive to simulate, making the generation of an ensemble of trajectories unfeasible. Cheaper trajectory ensembles can be generated using simpler test systems, such as Chodera's liquid

argon,[108] but this does not reflect real use-cases such as protein-ligand binding free energy calculations.[114] Simple models used in the operations research literature have the same problem here.[298,308,324] We attempt to address this issue by fitting models to data from compute-intensive absolute binding free energy calculations. From these models, we can cheaply generate ensembles of synthetic trajectories with exactly known properties.

4.3 Methods

We began with data from absolute binding free energy calculations. We used long calculations to allow more time for the initial transient to decay and to reduce errors in the estimated autocorrelation functions. To make the data more representative of absolute binding free energy calculations in general, we used data from a diverse range of protein-ligand complexes. The real data are computationally expensive to generate and unsuitable for testing since the infinite-sampling results are unknown. Therefore, the initial transients and autocorrelation functions of the real data were estimated and used to generate large numbers of correlated synthetic time series from uncorrelated Gaussian noise. The properties of the synthetic time series did not need to exactly match the real data they were modelled on, only to be reasonably representative of absolute binding free energy data. Because the infinite sampling time results were known for the synthetic time series, truncation point selection heuristics could be evaluated by their RMSE to the infinite sampling time results.

4.3.1 Generation of Simulation Data

Data were taken from our recent absolute binding free energy study - specifically, for the long (30 ns per simulation) non-adaptive runs.[296] Simulation details are given in Clark et al..[296] In free energy calculations, the Hamiltonian, H, is parameterised by λ. λ scales the strength of intermolecular interactions of the ligand in absolute binding free energy calculations. The real data consist of simulations run at $\lambda = 0$, where (a subset of) interactions are fully active, at $\lambda = 1$, where these interactions are removed, and at several intermediate values.[296] An important quantity is the gradient $\frac{\partial H}{\partial \lambda}$, as the free energy change for a transformation can be estimated by integrating the mean gradient over λ. Therefore, we base our testing on the time series of (integrated) $\frac{\partial H}{\partial \lambda}$. All sampled $\frac{\partial H}{\partial \lambda}$ are provided on Zenodo.[296,325] To provide challenging test data, we mainly use

data from the "bound vanish" stage of the thermodynamic cycle, where the ligand intermolecular Lennard-Jones terms are progressively removed while the ligand is restrained in the protein binding site. The bound vanish stage generally shows the most pronounced initial transient and longest correlation times. The data include 5 different test systems (see section S3 of Clark et al. for more details) with ligands ranging from small and rigid benzene up to a 70-atom ligand, ensuring diverse initial transient and autocorrelation behaviour.[296] The complexes were the L99A mutant of T4 lysozyme (T4L) with benzene,[50,190] human macrophage migration inihibition factor (MIF) with the ligand MIF180,[166,274] mouse double minute 2 homologue (with a truncated lid) with the ligands Pip2 and Nutlin,[165] and phosphodiesterase 2a with the ligand "P10".[171] Some analysis was also performed for the free vanish stage, where the ligand intermolecular Lennard-Jones interactions are removed in water. All simulations were performed with a 4 fs time step and $\frac{\partial H}{\partial \lambda}$ was saved every 200 steps. For each system, a time series of estimated free energy change against simulation time was produced by integrating $\frac{\partial H}{\partial \lambda}$ over λ using the trapezoidal rule.[116] Hence, the A of previous sections is replaced with ΔG. These time series were averaged over the 5 replicate runs performed for each system to provide less noisy estimates of model parameters discussed in Section 4.3.2.

4.3.2 Generation of Synthetic Data

The general strategy for modelling the data is illustrated in Figure 4.1: initial transients were fitted to the first 10 ns of simulation data, while autocovariance functions were fit to the final 20 ns of approximately stationary data. These were used to produce large ensembles of synthetic time series, where each synthetic time series was generated from different uncorrelated Gaussian noise, but shared the same initial transient and autocovariance functions.

In detail, the bound vanish stage time series generally showed no substantial drift over the final 20 ns (from 10 - 30 ns, Figures C.1 and C.2). Therefore, the "true" infinite sampling time free energy changes were calculated as the mean over the last 20 ns, and were subtracted from the time series. The initial transients were generally well-fit by a single exponential decay (Figure C.3), although some quickly-decaying bias was evident for most systems at short timescales. To account for this, a second exponential with short half-life was fit after subtracting the first exponential, and retained only if the pre-exponential term was positive (Figure C.4).

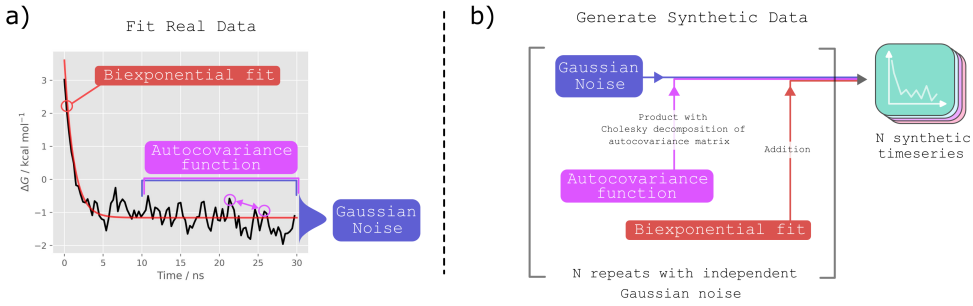


Figure 4.1: The creation of synthetic datasets modelled on real data. a) A 30 ns time series of free energy change estimaters against simulation time was obtained from an absolute binding free energy calculation (here we averaged over 5 repeats for the Lennard-Jones term removal stage when the ligand is bound to the protein). The initial transient was captured with a biexponential fit. The final 20 ns generally showed no substantial overall trend and were approximately Gaussian. These data were used to fit an autocovariance function. b) To generate synthetic time series, we started with a vector of uncorrelated Gaussian noise. The desired autocorrelation structure was reintroduced by taking its product with the Cholesky decomposition of the autocovariance matrix, and the initial transient was reintroduced by adding the biexponential fit. This was repeated N times with different Gaussian noise (but the same autocovariance function and biexponential trend) to generate N synthetic time series. We used $N = 1000$.

Autocovariance functions were calculated for each system using the final 20 ns portions of the time series. The first two points were calculated directly from the averaged time series, while the remaining points were obtained by interpolating the convex gamma function computed following Geyer (Figures C.6 and C.7).[305] Figure C.5 shows that the final 20 ns portions of the time series were well-modelled by Gaussian distributions (although some systems showed small but significant deviations from normality). Therefore, synthetic ensembles of trajectories were generated for each system starting from vectors of uncorrelated standard normal noise. 1000 noise vectors were generated for each system. The Cholesky decompositions of the autocovariance matrices were obtained and the resulting lower-triangular matrices were used to reintroduce the desired correlation structure to the uncorrelated noise vectors. The exponentials were then added to introduce the initial transients. Examples of synthetic time series are shown in Figures C.9 and C.10.

Synthetic ensembles of trajectories were also created for the free vanish stage for the T4L and PDE2a systems, which were chosen to represent highly uncorrelated and relatively correlated time series, respectively (Section C.3). Trajectory ensembles were generated as for the bound vanish stage, except no exponential trends were fit.

The synthetic data will not be perfect models of the true data, because the initial transients, autocovariance functions, and infinite sampling time "true" free energy changes will not be perfectly estimated. This does not matter, so long as the properties of the synthetic time series are reasonably representative of absolute binding free energy data in general. Hence, we are not concerned with precisely modelling specific features of the real data, for example, the large drop observed for MDM2-Nutlin Repeat 1 around 11 ns (Figure C.2). Relatedly, the mean time series for T4L drifts from ≈ 3.5 to 4.0 kcal mol^{-1} between 10 and 30 ns, in the opposite direction to the apparent initial transient. This appeared to violate our assumption of no drift in this region and produced a long-tailed autocovariance function (Figure C.6 and Table 4.1). However, it does not matter if this does not match the "true", infinite sampling autocovariance function for T4L, because the autocovariance function appears realistic:[110,326] it shows a fast initial decay followed by long, approximately exponential tail (Figure C.8).

4.3.3 Testing of Heuristics

Three types of method were tested for calculating the variance: an uncorrelated estimate which used Equation 4.17 ("Uncorrelated Estimate", equivalent to White's original MSER), window methods using Equation 4.19 with window sizes 5, 50, and $\sqrt{N_{n_0}}$, and initial sequence methods based on Geyer's initial positive, monotone, and convex sequence rules.[305] We tested the $\sqrt{N_{n_0}}$ window size due to its use in selecting batch and window sizes for variance estimation.[307,327] We use triangular windows of the form

$$w(t) = \begin{cases} 0 & \text{if } |t| > s, \\ 1 - \frac{|t|}{s} & \text{if } |t| \leq s, \end{cases} \tag{4.24}$$

where s is the window size. Chodera's method for terminating the autocovariance function is not one of Geyer's initial sequence methods. However, we included it with the initial sequence rules (as "Initial Sequence: Chodera") due to its similarity. We also tested a variant of Geyer's initial convex sequence method where each successive variance calculation (with increasing truncation) was restricted to

a gamma series (Equation 4.23) no longer than that from the previous truncation point (named "Initial Sequence: Smoothed Lag Convex"). In all cases, we did not calculate the marginal standard error for the last 10 % of the data, to avoid noisy estimates producing very late truncation.[328]

Each method was applied to every member of the synthetic ensembles after truncating the synthetic data after 8 ns, which is fairly typical of an absolute binding free energy calculation.[274] In addition to the "Standard" datasets (8 ns truncation and generated as described above), we tested all methods on a "Short" dataset where the truncation point was 0.2 ns, a "Subsampled" dataset where 99 of every 100 datapoints were discarded, a "Noisy" dataset where the autocovariance terms were scaled up by $\sqrt{5}$, and a "Block-Averaged" dataset, where successive blocks of 100 data points were replaced by their averages. Reported uncertainties in $\mathrm{RMSE}_{\mathrm{Trajs}}(\langle \Delta G \rangle_{[n_0,N]})$, $\mathrm{SD}_{\mathrm{Trajs}}(\langle \Delta G \rangle_{[n_0,N]})$, and $\mathrm{SD}_{\mathrm{Trajs}}(\langle \Delta G \rangle_{[n_0,N]})$ for each of the heuristics are 95 % confidence intervals which were calculated by bootstrapping synthetic trajectories 10000 times with replacement.

All tested methods were implemented in the open-source Python package RED (Robust Equilibration Detection, where "equilibration" is used in the sense of finding the optimal truncation point) available at github.com/fjclark/red. A complete workflow to reproduce the study beginning from the absolute binding free energy gradient data is provided at github.com/michellab/Robust-Equilibration-Detection-Paper and all data are provided on Zenodo.[329]

4.4 Results

4.4.1 Initial Sequence Methods are Prone to Over-Discarding

To understand how the methods performed on time series with no initial transient, we applied them to the free vanish stage synthetic trajectory ensembles. We used the T4L system (benzene in water) and the PDE2a system (ligand P10 from Huggins in water) as examples of relatively low variance and correlation, and relatively high variance and correlation systems, respectively.[171] Details of the fitting procedure and parameters are given in Section C.3. As the synthetic data contained no biases, the optimum truncation time was 0 ns and late truncation indicated problems with the heuristics.

For the relatively low variance and correlation T4L system, all methods consistently selected discard times very close to 0 ns, producing RMSEs within uncertainty of the optimal fixed truncation time limit (at 0 ns - Section C.4). It was reassuring to observe this expected behaviour. However, for the relatively high variance and correlation PDE2a system, methods which more fully accounted for autocorrelation were increasingly prone to over-discarding data (Figure 4.2). This was particularly true for the initial sequence methods, which occasionally discarded over half of the data despite the lack of any bias. This was reflected by increases in the RMSEs. The "Initial Sequence: Positive" method should produce the largest variance of the mean estimates for a given time series because it chooses the latest truncation point for the autocovariance series and does not reduce the autocovariance sum by enforcing monotonicity or convexity of the gamma series. This method was the most prone to over-discarding data, corroborating the trend that generalised MSER methods which produce larger variance of the mean estimates are more prone to erroneously discarding data. Within the initial sequence methods, applying Geyer's initial monotone and convex rules slightly reduced erroneous late truncation compared to the initial positive sequence method.

Using the first 100 PDE2a synthetic trajectories, we verified that the maximum effective sample size heuristic implemented in PyMBAR's timeseries module gave identical results to the implementation in RED.[108,302] We then compared the truncation points selected using this method and the equivalent minimum marginal standard error heuristic (see Equations 4.18 and 4.10), where we did not employ the adaptive integration scheme described by Chodera et al..[302] As expected, the truncation points selected were generally very similar: identical truncation times were selected in 81 % of cases and the difference was less than 0.1 % (of the total time series length) for 95 % of the trajectories. The minimum standard error method showed a slight bias for later truncation times over the maximum effective sample size method - the mean (ESS - MSE) difference was -0.34 % of the total time series length. However, we found that using the marginal standard error method fixed several problem cases where time series (not the synthetic data discussed above) were contaminated by one or a few very different initial samples, which the effective sample size method failed to remove.

We also examined the effect of the adaptive integration scheme described by Chodera et al. on the same trajectories.[302] This scheme reduces the computational cost by calculating the autocorrelation less frequently at greater lag times, but we found that it produced more erroneously late truncation times (Section C.5).

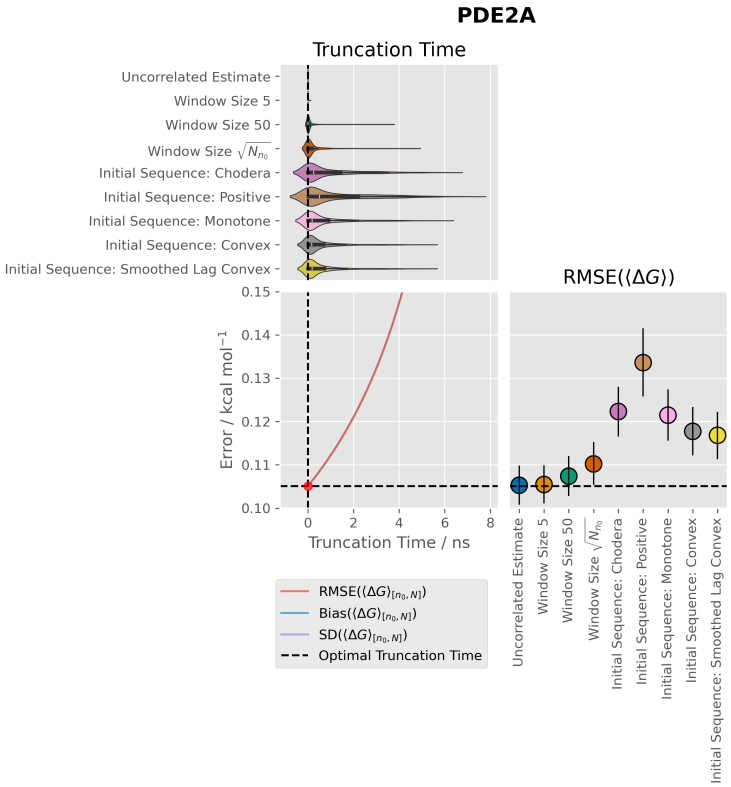


Figure 4.2: Discard times, RMSEs, and underlying time series properties for the free vanish stage for PDE2A. The top panel shows kernel density estimates of the distributions of times discarded with each method. The bottom left panel shows the RMSEs which would be obtained with an infinitely large ensemble of synthetic time series with fixed truncation points. The red dot indicates the optimum fixed-time truncation point. This is at time 0 ns as there is no bias. The bottom right panel shows the RMSEs obtained with each generalised MSER method. Uncertainties are 95 % confidence intervals obtained with 10000 iterations of bootstrapping with replacement. Note that the y-axis has been shifted from 0 to make differences between the methods clearer.

4.4.2 The $\sqrt{N_{n_0}}$ Window Method Balances Bias Sensitivity and Variability

Next, we tested the methods on the "Standard" synthetic ensembles generated from the bound vanish stages. The fitted parameters used to generate the synthetic ensembles were fairly diverse, with the variance of the mean estimates, max lag index (for the initial convex sequence estimate of the autocovariance series), and pre-exponential factors for the "slow" exponential fits varying over an order of magnitude (Table 4.1). The half-lives of the "slow" exponential fits varied from 0.3 to 1.6 ns. Further details of the parameter fitting and synthetic data generation are given in Section C.2.

Table 4.1: Model Parameters Fitted to Bound Vanish Stages of Absolute Binding Free Energy Calculations[a]

	Half-life (ns)	a (kcal mol^{-1})	Fast Half-life (ns)	Fast a (kcal mol^{-1})	Total Variance (kcal^2 mol^{-2})	Max Lag Index
T4L	0.33	0.86	∞	0	110	7175
MIF	0.88	4.7	0.0040	14	55	749
MDM2-Nutlin	1.6	2.6	0.0052	21	81	577
MDM2-PIP2	0.80	3.6	0.0057	12	41	335
PDE2A	0.33	13	0.019	13	520	1699

[a] Total variance refers to the total variance of the mean, obtained by summing the autocovariance series from - max lag index to + max lag index, where the series and maximum lag indices were estimated according to Geyer's initial convex sequence rules. "a" refers to the pre-exponential factors.

We compared the ensemble RMSEs obtained by each generalised MSER method to the minimum possible RMSE obtained by applying the optimal fixed-time truncation to all time series. In general, methods which produced larger variance of the mean estimates were prone to choosing late truncation points, increasing $\mathrm{SD}_{\mathrm{Trajs}}(\langle \Delta G \rangle_{[n_0,N]})$, while methods which underestimated the variance of the mean were prone to choosing early truncation points, increasing $\mathrm{Bias}_{\mathrm{Trajs}}(\langle \Delta G \rangle_{[n_0,N]})$. This is illustrated for one of the synthetic MIF time series in Figure 4.3 (we show $\frac{1}{\widehat{\mathrm{Var}}_{\mathrm{Trajs}}(\langle \Delta G \rangle_{[n_0,N]})}$ in place of $\widehat{\mathrm{Var}}_{\mathrm{Trajs}}(\langle \Delta G \rangle_{[n_0,N]})$ to make the behaviour in the region of minimum $\widehat{\mathrm{Var}}_{\mathrm{Trajs}}(\langle \Delta G \rangle_{[n_0,N]})$ clearer). The "Uncorrelated Estimate" (original MSER) is only sensitive to offset of the time series mean through the first term of the autocovariance function, $\hat{\gamma}_{0,[n_0,N]}$ (Equation 4.17). As a result, it produces underestimates of $\widehat{\mathrm{Var}}_{\mathrm{Trajs}}(\langle \Delta G \rangle_{[n_0,N]})$ which vary smoothly with increasing truncation but are insensitive to bias. This results in early truncation. The estimates of $\hat{\gamma}_{0,[n_0,N]}$ quickly stabilise with increasing truncation and the variance of the mean estimates become proportional to $\frac{1}{N_{n_0}}$. In contrast, the initial sequence methods are sensitive to offset of the time series mean through many more terms of the autocovariance function. As a result, offsetting the mean more effectively counter-balances the $\frac{1}{N_{n_0}}$ term to increase the variance of the mean estimate. This prevents early truncation. However, the $\widehat{\mathrm{Var}}_{\mathrm{Trajs}}(\langle \Delta G \rangle_{[n_0,N]})$ estimates become much noisier, especially at late truncation times, which can produce spurious minima in the variance of the mean estimation, leading to late truncation. Sudden dips in $\widehat{\mathrm{Var}}_{\mathrm{Trajs}}(\langle \Delta G \rangle_{[n_0,N]})$ were often caused by sudden decreases in the maximum lag index used to calculate the autocovariance sum. For the Chodera method in particular, troughs at single values of n_0 were often observed (Figure 4.3). However, issues with troughs in the maximum lag index were generally removed or reduced with increasingly stringent initial sequence methods. In general, Geyer's initial monotone and convex sequence methods produced the smoothest traces of $\widehat{\mathrm{Var}}_{\mathrm{Trajs}}(\langle \Delta G \rangle_{[n_0,N]})$ of the "initial sequence"

methods (including Chodera's). The window method with window size $\sqrt{N_{n_0}}$ compromises the extremes of the uncorrelated estimate and the initial sequence methods by including enough treatment of correlation to avoid excessively early truncation, while avoiding noisy variance of the mean estimates which can produce late truncation.

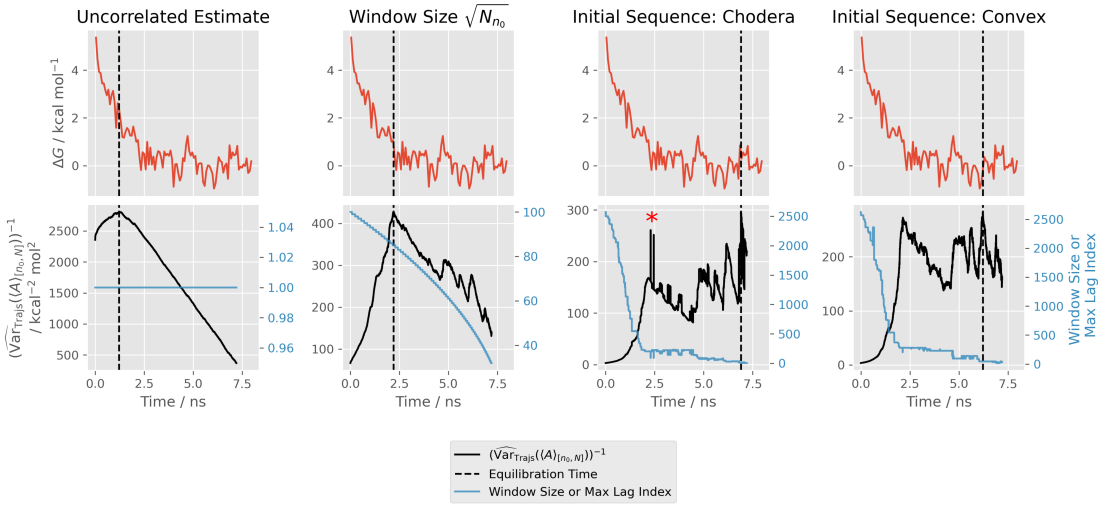


Figure 4.3: Performance of several generalised MSER heuristics on a single bound vanish synthetic times series for for MIF. To show features close to the selected truncation points, we show the inverse of the squared (marginal) standard error $(\widehat{\mathrm{Var}}_{\mathrm{Trajs}}(\langle \Delta G \rangle_{[n_0,N]}))$. With the Chodera method, sudden dips in the maximum lag index cause spikes in the inverse $\widehat{\mathrm{Var}}_{\mathrm{Trajs}}(\langle \Delta G \rangle_{[n_0,N]})$ (two spikes are marked by the red asterisk). The time series have been block-averaged with 100 blocks to make the trends clearer.

More quantitatively, the performance of all methods across the "standard" synthetic ensembles is compared in Figure 4.4 and Table 4.2. The test systems can be divided into two groups: those where the optimal fixed-time truncation point is around or above 50 % of the simulation time (MIF, MDM2-Nutlin, and MDM2-Pip2), and those where it is substantially less (T4L, PDE2a). Unsurprisingly, methods prone to late truncation (the initial sequence methods), performed best in the first set, while methods prone to early truncation performed best in the latter set. For example, MDM2-Pip2 had a relatively late optimal fixed truncation time (around 4 ns). The uncorrelated estimate method generally truncated at less than 1 ns for this system, producing an ensemble RMSE around 4 times greater than the fixed truncation point minimum. For the window estimators, the mean truncation times increased and the ensemble RMSEs decreased with increasing window size up to the $\sqrt{N_{n_0}}$ window estimator - this produced an ensemble RMSD close to the optimal fixed truncation time limit and to the initial sequence

estimator methods, which performed the best. In contrast, the PDE2a system had a relatively early optimal fixed truncation time (around 1.5 ns). Here, the best-performing methods were the window estimators with window sizes of 5 and 50, both with an ensemble RMSE of $0.28^{0.29}_{0.27}$ kcal mol^{-1}. The RMSE for the uncorrelated estimate and window size $\sqrt{N_{n_0}}$ methods were marginally worse ($0.31^{0.32}_{0.30}$ and $0.32^{0.34}_{0.31}$ kcal mol^{-1}, respectively), while the RMSEs for the initial sequence methods were substantially higher ($\approx$ 0.46 kcal mol^{-1}) due to late truncation. Worryingly, the initial sequence methods showed strongly bimodal distributions of discard times, centred around 1.5 and 7 ns. Hence, the strong bias sensitivity of the initial sequence methods comes with an increased propensity to truncate all data but the final local trough or plateau.

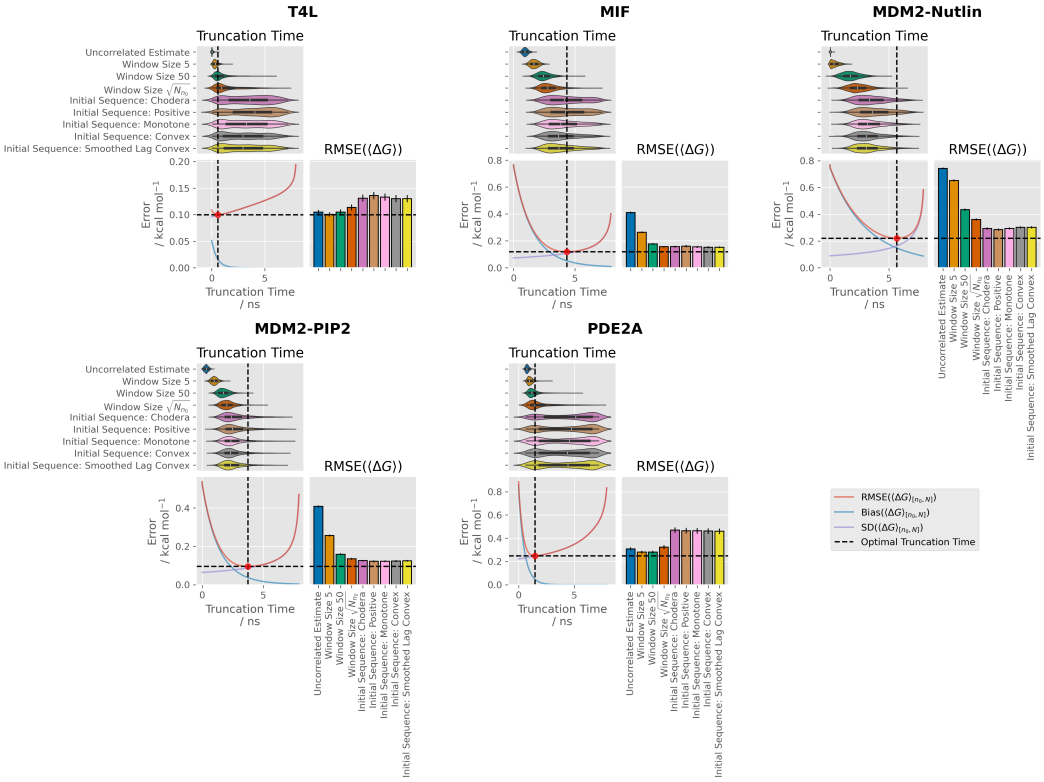


Figure 4.4: Discard times, RMSEs, and underlying time series properties for the bound vanish stage time series. Within each system block, the top panels show kernel density estimates of the distributions of times discarded with each method. The bottom left panels show the RMSEs which would be obtained with an infinitely large ensemble of synthetic time series with fixed truncation points. The red dot indicates the optimum fixed-time truncation point. Bottom right panels show the RMSEs obtained over the full synthetic ensemble for each method. Uncertainties are 95 % confidence intervals obtained with 10000 iterations of bootstrapping with replacement.

Table 4.2: Ensemble RMSEs for all Generalised MSER Heuristics for the "Standard" Bound Vanish Data[a]

Method	T4L	MIF	MDM2-Nutlin	MDM2-PIP2	PDE2A
Uncorrelated Estimate	$0.105^{0.109}_{0.101}$	$0.411^{0.417}_{0.404}$	$0.743^{0.749}_{0.737}$	$0.408^{0.413}_{0.403}$	$0.308^{0.321}_{0.295}$
Window Size 5	$0.100^{0.105}_{0.096}$	$0.264^{0.270}_{0.257}$	$0.653^{0.660}_{0.646}$	$0.256^{0.262}_{0.251}$	$0.280^{0.292}_{0.268}$
Window Size 50	$0.105^{0.110}_{0.100}$	$0.178^{0.184}_{0.172}$	$0.434^{0.443}_{0.425}$	$0.159^{0.164}_{0.154}$	$0.280^{0.293}_{0.267}$
Window Size $\sqrt{N_{n_0}}$	$0.114^{0.119}_{0.108}$	$0.157^{0.163}_{0.151}$	$0.361^{0.369}_{0.352}$	$0.135^{0.140}_{0.130}$	$0.323^{0.341}_{0.306}$
Initial Sequence: Chodera	$0.131^{0.138}_{0.125}$	$0.157^{0.164}_{0.149}$	$0.293^{0.302}_{0.285}$	$0.125^{0.130}_{0.121}$	$0.469^{0.493}_{0.445}$
Initial Sequence: Positive	$0.136^{0.143}_{0.129}$	$0.161^{0.169}_{0.153}$	$0.285^{0.294}_{0.276}$	$0.122^{0.127}_{0.117}$	$0.465^{0.489}_{0.440}$
Initial Sequence: Monotone	$0.133^{0.140}_{0.126}$	$0.155^{0.163}_{0.148}$	$0.294^{0.303}_{0.285}$	$0.122^{0.127}_{0.118}$	$0.463^{0.488}_{0.439}$
Initial Sequence: Convex	$0.130^{0.137}_{0.123}$	$0.153^{0.160}_{0.146}$	$0.302^{0.311}_{0.293}$	$0.124^{0.128}_{0.119}$	$0.461^{0.485}_{0.437}$
Initial Sequence: Smoothed Lag Convex	$0.130^{0.137}_{0.124}$	$0.152^{0.159}_{0.145}$	$0.302^{0.310}_{0.293}$	$0.125^{0.129}_{0.120}$	$0.459^{0.483}_{0.435}$

[a] All values in kcal mol^{-1}. Uncertainties are 95 % confidence intervals obtained by bootstrapping over 10000 iterations with replacement.

We found that the $\sqrt{N_{n_0}}$ window method was the most robust truncation selection heuristic on these data. Its intermediate treatment of autocorrelation balanced the propensities for early and late truncation shown by methods ignoring correlation, and methods more completely accounting for correlation, respectively. Methods at the extremes of the correlation spectrum performed well on the early truncation point set and poorly on the late truncation point set or vice versa. In contrast, the $\sqrt{N_{n_0}}$ window method never performed more than 29 % worse (by RMSE) than the top-performing automated method. It also did not produce the wide and sometimes bimodal distributions of truncation times seen for the initial sequence estimators, which we view as an advantage in itself.

There were systems where all methods systematically underestimated the truncation point (particularly MDM2-Nutlin and MDM2-Pip2). However, the best-performing methods generally showed an ensemble RMSE close to the optimal fixed truncation point minimum, and the $\sqrt{N_{n_0}}$ window method never performed much worse. Figure 4.4 shows that the $\sqrt{N_{n_0}}$ window method would never perform much worse, and would often perform substantially better compared to a common default fixed truncation time of 20% (1.6 ns). Further analyses of the "standard" ensembles are presented in Section C.6.

The left-most local minimum (LLM) method has been proposed to avoid the over-truncation of data with MSER.[299] The left-most local minimum is selected as the truncation point, rather than the usual global minimum. However, we found that this produced very early truncation points when applied to the initial sequence generalised MSER methods, because the curves of MSE against truncation point were noisy and had early minima. For our data, it would be unhelpful to apply this to the other truncation heuristics, because they already showed a tendency towards early truncation.

The observation that sudden troughs in the maximum lag time led to instabilities with the Chodera method prompted us to try the "Initial Sequence: Smoothed Lag Convex" method, in which the maximum lag times at subsequent truncation times were never allowed to increase. Geyer's initial positive, monotone, and convex sequence rules were also applied. However, the performance was very similar to the "Initial Sequence: Convex" method.

Oliveira et al. qualitatively tested MSER (original, batch, and LLM variants) in grand cannonical Monte Carlo simulations and concluded that it out-performed methods including Chodera's maximum effective sample size heuristic, which they found prone to late truncation.[301] However, we found that the original MSER method ("uncorrelated estimate") often severely underestimated the optimal truncation time, producing very large RMSEs (for example, see MDM2-Pip2 in Table 4.2). A key difference between Oliveira et al.'s data and ours is that the initial transient is fairly hidden by the high variance of our data (see Figure C.1 for the non block-averaged time series). It is likely important to account for autocorrelation to effectively recover the trend from the noise. This is not the case for Oliveira et al.'s data, where reasonable truncation times appear to be estimated without accounting for autocorrelation, avoiding the associated risk of late truncation. Hence, the optimal truncation point selection heuristic will likely depend on the properties of the data. However, as discussed below, the $\sqrt{N_{n_0}}$ window method remains a good choice for several variants of our synthetic dataset.

4.4.3 Similar Results are Obtained with Noisier Data

To increase confidence that our results were generalisable, and to better understand the methods, we retested all heuristics on modified synthetic ensembles. Detailed results are given in Section C.6.

"Noisy" ensembles were created in the same way as the "standard" data, but all autocovariance terms were increased by a factor of $\sqrt{5}$. This was intended to reverse the effect of averaging over the 5 repeat runs before the synthetic data parameters were extracted. The effect of the added noise was to shorten the optimum truncation time (Figure C.17), which was reflected by earlier truncation time distributions for all heuristics. However, the relative performance of all methods on all systems remained very similar to the standard data (Figure C.17 and Table C.3).

To investigate whether these heuristics were applicable to simulations run at a single value of λ, rather than the free energy change integrated over λ, we modelled synthetic data on a a single λ state simulation. Data were modelled on a particularly noisy bound vanish λ state, which generally produced higher ratios of total variance to the slow exponential prefactor. However, the comparative performances of the heuristics were similar to the "standard" synthetic ensemble (Section C.7).

4.4.4 Subsampling and Block Averaging Reduce Differences Between the Methods Due to Reduced Autocorrelation

To check whether our results were affected by the frequency of data collection, we subsampled 1 out of every 100 data points. Reducing the number of data points increased the ratio of $\mathrm{SD}_{\mathrm{Trajs}}(\langle \Delta G \rangle_{[n_0,N]})$ to $\mathrm{Bias}_{\mathrm{Trajs}}(\langle \Delta G \rangle_{[n_0,N]})$, shifting the optimal truncation points to slightly earlier times (Figure C.17). As the subsampling interval was a large fraction of the length of the autocorrelation functions used to generate the data (Table 4.1), this also dramatically reduced the autocorrelation of the time series. It was therefore unsurprising that all generalised MSER methods, which differ only in their treatment of autocorrelation, performed similarly on all subsampled trajectory ensembles (Table C.4). The truncation time distributions for all methods became similar: those from the "uncorrelated estimate" generally shifted later, and the truncation time distributions from the initial sequence methods became narrower and shifted earlier. Unlike with the "standard" data, no methods produced RMSEs dramatically greater than the fixed truncation time minimum. For these reasons, it may seem tempting to subsample all time series before applying the heuristics. However, this is problematic - if the mean is calculated from the truncated subsampled data, it will have an substantially higher RMSE than for the original data, due to the loss of information in the discarded samples. Applying the truncation time from the subsampled data is also ill-advised because the subsampled data have earlier optimal truncation times than original data; the truncation times for most methods will be biased towards erroneously early times.

Block averaging is another method which reduces autocorrelation between consecutive data points. However, unlike subsampling, it does not affect the underlying bias-standard deviation trade-off of the time series, and the optimal fixed-time truncation points remain the same. As discussed in Section 4.2.3, estimation of the variance of the mean by block averaging is closely related to estimating

the variance of the mean using window estimators with a Bartlett window.[309] Hence, block averaging the data before analysis effectively turns the "uncorrelated estimate" into a window estimator, and increases the effective window sizes of the window size estimators. However, we repeated the tests after block averaging all "standard" time series (block size 100) as a way of simulating less correlated and less noisy data. Block averaging dramatically improved the performance of the "uncorrelated estimate" on the late truncation point systems (Figure C.17 and Table C.5), which was unsurprising given that this effectively became a window method with a large window. Generally, the distributions of discard times became wider and shifted to later times for all methods. For the early truncation point systems, this generally worsened the performance of the window estimators so that they became comparable with the initial sequence estimators. In terms of RMSE, the initial sequence methods usually did not perform significantly worse than without block-averaging. The bimodal discard time distributions observed for the initial sequence methods on the "standard" dataset were observed for the window estimators for some systems.

4.4.5 All Methods Usually Fail to Detect Insufficient Sampling

This work compares heuristics for truncation point selection once sampling is complete. However, it would be useful if these methods could also detect when all samples are highly biased, indicating that few representative samples from the equilibrium distribution have been obtained and more sampling is required. In other words, it would be useful if truncation point selection heuristics could be used for equilibration detection. In the context of MSER, truncation times over 50 % of the simulation time have been taken to indicate a lack of equilibration.[328] We tested this heuristic by applying all generalised MSER methods to the first 0.2 ns of the "standard" synthetic datasets (Figure C.17 and Table C.6). For all time series, the optimal fixed-time truncation points were at over 50 % of the total time, and usually close to 100 %. PDE2a had large and quickly decaying bias at these short timescales (Table 4.1), which produced a steep decrease of RMSE with increasing truncation time compared to all other systems. For all other systems, the distributions of discard times were relatively narrow and centred at less than 0.05 ns for all methods. Hence, all methods failed to detect insufficient sampling according to the 50 % time rule. However, for PDE2a, the size 50 window method and all methods with fuller treatment of correlation showed median discard times at greater than 50 % of the total time. This indicates that when large and rapidly decaying initial transients dominate slowly increasing random errors, the 50 %

time rule may be effective as a simulation-stopping criterion (especially when used with generalised MSER methods which more fully account for autocorrelation). This appears to be the case for Oliveira et al.'s data. However, our results suggest that this is unlikely to be the case for most free energy calculations.

4.4.6 Practical Use of Generalised MSER Methods

General Applicability

We have attempted to create synthetic data sets with varied autocovariance functions and biases. Across our data sets, the $\sqrt{N_{n0}}$ window method favourably compromised bias sensitivity and discard time variability, and appeared reasonably robust. By increasing the variability of our data, we increased confidence that this applies to higher variance data (Section 4.4.3); by decreasing the autocorrelation (Section 4.4.4) we increased confidence that this applies to less correlated data. However, we acknowledge our data's coverage of possible time series generated from molecular simulations remains limited.

The initial transients in our synthetic data decay quickly enough that most bias is eliminated by the end of a "standard" length binding free energy simulation. This is a sweet spot for testing truncation heuristics because the choice of truncation point is especially critical. We expect the data to be reasonably representative of absolute binding free energy calculations, but time series from other molecular simulations may not occupy this sweet spot. In the case where a rapidly-decaying initial transient only biases the first few samples of a long time series, it is still important not to erroneously discard large fractions of the data (Section 4.4.1). Also, large but short-lived initial transients may have a fairly long-lived effect on cumulative averages if no truncation is performed. Therefore, the use of these heuristics is still justified. In the case where insufficient sampling has been performed and all data remain highly biased, these heuristics can still help to reduce bias. A key strength of these automated methods is that they can be applied without prior knowledge of the data. When few time series are generated or there is substantial prior knowledge of likely initial transient behaviour, manual selection or fixed truncation points, respectively, may produce better results.

When applying generalised MSER methods to free energy calculations, a practical question is whether they should be applied to the overall time series (e.g. after combining all λ state gradients using thermodynamic integration), or applied individually to the gradients/perturbed energies from each window. As the optimal truncation points are expected to vary substantially between states, the former

approach may reduce the quantity of data unnecessarily discarded, while the latter should reduce noise in the time series. We have not fully answered this question, but we have shown that the $\sqrt{N_{n_0}}$ window method appears to offer reasonable performance on synthetic data fit both to the overall stage free energy change (Table 4.2) and to a single noisy λ state (Section C.7). These heuristics appear suitable for both approaches.

Calculating Uncertainty

While inter-run uncertainty estimates are more robust,[258] single-run uncertainty estimates are required in the absence of replicates. Flyvbjerg and Petersen argue for non-overlapping block-averaging rather than directly using the autocovariance function to calculate uncertainty,[303] while a current best practices guide for molecular simulation does not recommend either method over the other.[330] We make a case for the use of the autocovariance function. Flyvbjerg and Petersen argue that the block-averaging method is cheaper than calculating the autocovariance function and avoids having to select the truncation point of the sum of autocovariance terms. However, using overlapping block averages provides a reduced variance variance estimate compared to non-overlapping block averages, and is equivalent to applying an appropriately sized Bartlett window to the autocovariance function.[309,310] This clarifies that the problem of truncation point selection is not avoided by using a block averaging method; it is closely related to the issue of block size selection. Furthermore, the additional computational cost (which is likely negligible with modern hardware) can bring a reduced variance variance estimate. As discussed by Gowers et al., use of the autocovariance function is also easily automatable.[331] In addition, methods we are aware of for estimating the optimal block size require calculating the autocovariance function as an intermediate step,[134,332–334] making more direct use of the autocovariance function simpler. Therefore, calculating variance of the mean estimates using autocovariance function-based methods may be preferable. Using uncorrelated or window estimators (without determining a suitable window size based on the data) may substantially underestimate the variance of the mean due to the incomplete treatment of correlation. Of the initial sequence methods, Geyer's initial positive sequence method will often provide the smallest underestimate because it produces the latest truncation of the autocorrelation function. However, this will still systematically underestimate the variance of the mean, as noise in

the autocorrelation function will result in early truncation (as negative Γ terms will appear earlier in the series with increasing noise). Therefore, approaches that fit the tail of the autocorrelation function to an exponential will likely provide superior uncertainty estimates.[315,326]

To illustrate the underestimation of uncertainty, we calculated the percentage of synthetic time series where the estimated 95 % confidence intervals covered the true mean (the coverage) after truncation (Table C.9). We did this for the "standard" free vanish stage synthetic ensembles, which did not contain initial transients. For the relatively uncorrelated T4L system, the coverage of the confidence intervals were very similar for all methods and close to the expected 95 %. In contrast, for the more correlated PDE2a system, the coverages for all systems were well below 95 %, even though there were still almost 1000 effectively uncorrelated samples. They were worst for the uncorrelated estimate (44 %), and improved up to the $\sqrt{N_{n_0}}$ window method (71 %) as autocorrelation was increasingly accounted for. Coverage then decreased slightly for the initial sequence methods ($\approx$ 64%). We note that further downward bias can be introduced by truncating time series with MSER methods prior to calculating confidence intervals, as done here; by definition, MSER picks the truncation point which minimises the estimated standard error. This bias will be greater for methods which are more prone to over-discarding. As shown by Equation 4.18, Chodera's method of maximising the effective sample size is closely related to MSER, so the issue of downwards-biased uncertainty estimates also applies. When using generalised MSER or maximum effective sample size truncation metrics, users should be aware that they may be introducing additional downwards bias to subsequent uncertainty estimates. This is without even considering that single runs often become stuck in local minima in configuration space, an issue which multiple runs may help to diagnose.

Subsampling is Not Recommended

Producing better estimates with less information is impossible, so subsampling is not recommended when the cost of saving and using samples is negligible.[305] Samples which are not fully correlated each contain new information (albeit less than uncorrelated samples), and discarding them adds noise to the variance and mean estimates. Subsampling may substantially increase the variance of the mean estimate and the variance estimate. Concretely, we show in Section C.9 that

when a sufficiently long stationary time series has a purely exponential correlation function with a half-life much greater than the sampling interval, subsampling at the interval of the statistical inefficiency increases the variance of the mean by 31%.

For this reason, we recommend against subsampling input data for the Bennett Acceptance Ratio (BAR) and the Multistate Bennett Acceptance Ratio (MBAR) free energy estimators.[121,131] These are no longer the maximum likelihood free energy estimators when applied to correlated samples. Still, they remain robust estimators, and correcting for the correlation is unlikely to yield much improvement unless there are large differences in correlation between states.[119,335] However, subsampling input for MBAR is often recommended because the asymptotic variance estimate given by Kong et al. and used in the original MBAR paper is derived for uncorrelated samples, meaning it is a large underestimate when there is substantial correlation.[130,131] While subsampling decreases bias in this uncertainty estimate, it adds noise to the free energy and uncertainty estimates. A better approach may be to use an uncertainty estimator which accounts for correlation, for example, block bootstrapping,[134,334,336] or a corrected asymptotic estimator.[119,133]

Speed

The computational cost of the generalised MSER methods is likely negligible compared to the cost of generating the data. However, computational cost may become a consideration when the algorithm is run repeatedly, for example if applied to individual λ states in alchemical free energy calculations. In general, the computational cost increases with increasing treatment of correlation, as more terms in the autocovariance sum must be evaluated. Hence, the $\sqrt{N_{n_0}}$ window method also has the advantage of speed over the initial sequence methods. In our current implementation (RED 0.1.1),[337–339] we generally found the ratio of speeds to be $\approx 1 : 2 : 8$ for the Uncorrelated, Window Size $\sqrt{N_{n_0}}$, and all "Initial Sequence" methods, respectively on the "standard" time series of 10^4 points. However, this is highly dependent on the implementation.

4.5 Conclusion

We have reformulated White's MSER to provide a spectrum of truncation point selection heuristics which differ in their treatment of autocorrelation. These include a method effectively equivalent to Chodera's maximum effective sample size heuristic.[108] We tested these methods by generating ensembles of synthetic time series modelled on free energy change estimates from long absolute binding free energy calculations. Heuristics were assessed by their RMSE to the infinite-time ensemble average of the synthetic data.

We observed a general trade-off: methods which more thoroughly accounted for autocorrelation often showed late and variable truncation times; methods which less thoroughly accounted for autocorrelation often showed early truncation, relative to the optimal fixed-time truncation point. This increased variance and bias, respectively. We found that using a window estimator of the variance with size $\sqrt{N_{n_0}}$ provided a good compromise between these extremes, generally providing robust truncation point estimates. Rerunning our analyses on noisier and less correlated data produced similar conclusions.

All methods tested in this work are implemented in the open-source Python package RED (Robust Equilibration Detection, where equilibration is used in the sense of finding the optimal truncation point), available from the PyPI, conda-forge, and at github.com/fjclark/red. This provides alternatives for the PyMBAR timeseries "detect_equilibration" function. While these methods were useful for selecting truncation points given data, they were not generally useful for detecting insufficient sampling (lack of equilibration) in our data.

Future work may involve testing these heuristics on a broader range of synthetic data. In addition, further studies may focus on adapting these heuristics to multiple repeat runs (using globally centered autocovariances).[340,341] This is likely a more robust approach and may reduce some of the issues we encountered with late truncation.

Chapter 5

Rapid Absolute Binding Free Energy Calculations for Virtual Screening

5.1 Introduction

An attractive application for ABFE calculations is virtual screening,[58] in which a hierarchy of computational methods is employed to identify potential hit molecules for a target of interest. Often, virtual screens begin from a vast virtual library, such as the ENAMINE REAL database, which contains over 9 billion molecules.[342] Initially, filtering tools and empirical rules may be applied to dramatically reduce the search space, often followed by docking and scoring.[343] Unfortunately, the docking scores often correlate poorly with experimental binding affinity.[344] Due to their relatively high accuracy and applicability to diverse chemical scaffolds, ABFE calculations are well-suited to accurately rescoring the molecules with the most favourable docking scores before molecules are selected for synthesis.[58,60,61,345] Virtual screening pipelines incorporating ABFE calculations have shown good performance in a prospective setting - for example, see a winning entry to the first Critical Assessment of Computational Hit-Finding Experiments (CACHE) Challenge.[62]

However, virtual screening requires affinity prediction for many molecules, making the high cost of ABFE (and RBFE) calculations additionally prohibitive. Part of the solution is undoubtedly the more efficient use of the expensive affinity information generated; this has motivated substantial industrial interest in active-learning protocols incorporating free energy calculations.[214–217,346–348] However, a complementary and under-explored solution may be the use of extremely fast alchemical free energy calculations, using orders of magnitude less simulation

time than is "standard". Several studies have noted that very short ABFE calculations may produce estimates with similar precision and accuracy compared to "standard" simulation times,[349] or at least retain useful ranking performance.[348] Despite this, there have been no systematic studies on the influence of simulation time on the performance of ABFE calculations.

The A3FE software presented in Chapter 3, which implements algorithms from Chapters 2 and 3, allows ABFE calculations to be run and analysed with no system-specific input other than the prepared structures. This facilitates high-throughput ABFE studies with minimal user intervention. Here, A3FE is used to assess the performance of ABFE calculations as the total simulation time is varied by two orders of magnitude. Emphasis is placed on ranking performance, which is critical to the success of virtual screening. In addition, the performance of fast ABFE calculations on a variety of systems is investigated.

5.2 Methods

5.2.1 System Preparation

Five proteins were studied: ephrin type-B receptor 4 (EphB4) in complex with 38 ligands,[350,351] cyclophilin D (CycloD) with nine ligands,[1,294] heat shock protein 90 (HSP90) with 18 ligands,[1,352] tyrosine kinase 2 (TYK2) with 13 ligands,[353–355] and p38 mitogen-activated protein kinase (p38) with 29 ligands.[355,356] EphB4 was prepared from PDB ID 2vwz using Schrödinger's Preparation Wizard with default settings (crystal waters were retained, capping groups were added, the orientations of asparagine, glutamine, and histidine residues were evaluated, and protonation states were defined). The protein and crystal waters were parameterised with the AMBER ff14SB and TIP3P force fields using antechamber 22.0.[180,185,269] EphB4 ligands were prepared with standard LigPrep settings, docked with reference to 2vwz using Glide with maximal common substructure core constraints, and parameterised with GAFF2.11 and AM1-BCC partial charges using antechamber 22.0.[85,86,357] Martin Packer is thanked for providing the prepared protein structure and ligand SDFs. The CycloD complexes from Chapter 3 were used (see Section 3.3.1). HSP90 and ligands were also prepared from Alibay et al. as described in Section 3.3.1 for CycloD. The TYK2 and p38 systems were obtained from the protein-ligand benchmark version 0.2.1;[355] the proteins and

crystal waters were parameterised with the AMBER ff14SB and TIP3P force fields,[180,185,269] and the ligands with GAFF2.11 and AM1-BCC partial charges using antechamber 22.0.[85,86,357] Otherwise, all set up was performed as described in Section 3.3.

5.2.2 Molecular Dynamics Protocols

All calculations were run using A3FE. Where not specified, the protocols given in Section 3.3 were used. Three non-adaptive protocols spanning three orders of magnitude of simulation time were considered: "0.01 ns", "0.1 ns", and "1 ns", which employed 0.01, 0.1, and 1 ns of sampling per window, respectively (Figure 5.1). The only other protocol used was "5 ns" for Cyclophilin D, for which ABFE data were already available from Chapter 3. For all protocols, the first 20 % of simulation time was discarded prior to analysis. The restrained and unrestrained NPT equilibration stages were shortened to 50 ps, and the five independent NPT equilibration stages used to fit the restraints were shortened to 100 ps, but otherwise the default A3FE equilibration procedure was used (Section 3.3). A λ schedule optimised for MIF/MIF180 (Section 3.3.1) with the automated window spacing algorithm with a thermodynamic speed of 1 kcal mol^{-1} (Chapter 3) was used for all ligands. Specifically, windows at λ = 0.0 and 0.1 were used for the bound restraining stage; at 0.0, 0.291, 0.54, 0.776, and 1.0 for the bound discharge stage; at 0.0, 0.026, 0.054, 0.083, 0.111, 0.14, 0.173, 0.208, 0.247, 0.286, 0.329, 0.373, 0.417, 0.467, 0.514, 0.564, 0.623, 0.696, 0.833, and 1.0 for the bound vanish stage; at 0.0, 0.222, 0.447, 0.713, and 1.0 for the free discharge stage; and at 0.0, 0.026, 0.055, 0.09, 0.126, 0.164, 0.202, 0.239, 0.276, 0.314, 0.354, 0.396, 0.437, 0.478, 0.518, 0.559, 0.606, 0.668, 0.762, and 1.0, for the free vanish stage. The only exception was CycloD, for which the non-adaptive results presented in Chapter 3 were re-analysed for fractions of the total simulation time (hence the λ schedule and equilibration protocol described in Section 3.3 were used).

EphB4 experimental free energies were calculated from reported IC50 values assuming that IC50 $\approx K_{\mathrm{D}}$.[350,351] Experimental binding affinities for other systems were obtained from Alibay et al. or the protein-ligand benchmark.[1,355] Uncertainties in free energy results for individual ligands are reported as 95 % t-based confidence intervals based on the deviations between 5 replicate runs. Uncertainties in statistics for a set of ligands are reported as 95 % confidence intervals generated by bootstrapping: results for each ligand were resampled with replacement 10000

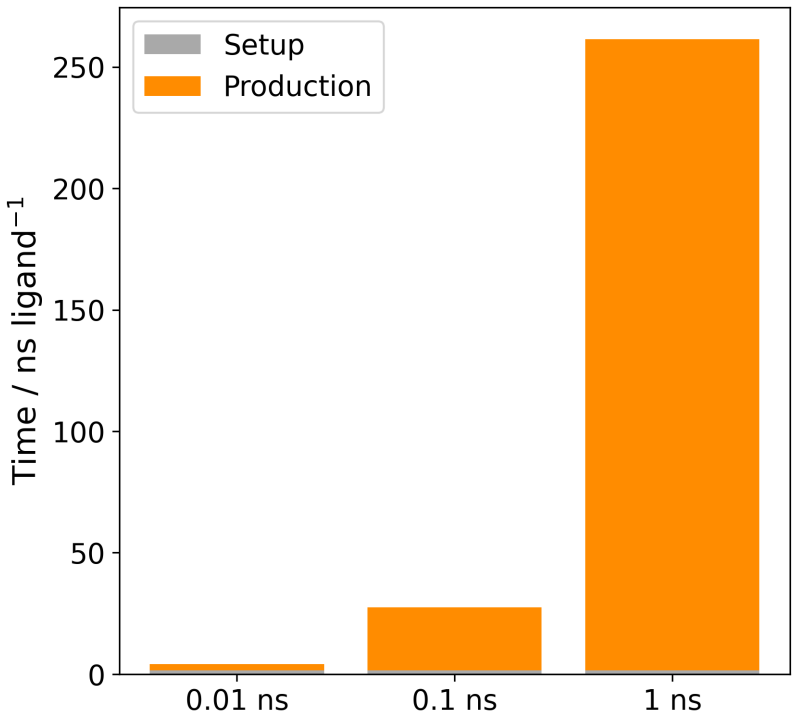


Figure 5.1: Total molecular dynamics time per ligand for ABFE protocols.

times. The uncertainties associated with experiment and per-ligand free energy estimates were included by resampling from Gaussian distributions associated with each. Where not provided, uncertainties in experimental affinity data were assumed to be 0.5 kcal mol^{-1}.[20,279]

5.3 Results and Discussion

First, the effect of total simulation time on ranking performance was investigated using EphB4.[350,351] The ABFE results were analysed for the 1, 0.1, and 0.01 ns protocols, in which 1, 0.1, and 0.01 ns of simulation time were allocated per λ window. This system is not a realistic test-case for ABFE, because only small R-group changes occur, meaning RBFE would be preferable. However, the many ligands (38) and wide dynamic range of experimental binding affinities allow robust performance assessment. The effect of changing total simulation time was also investigated using the CyloD system of Alibay et al..[1] This set includes diverse fragments and merged ligands, which provides a more realistic use case for ABFE. The performance of the 0.1 ns protocol was then evaluated using the HSP90 system from Alibay et al. (which is also a realistic ABFE use case) and two common RBFE benchmark systems: TYK2 and p38.[1,355]

5.3.1 The Impact of Simulation Time on Ranking Performance

The 1 ns protocol showed reasonable ranking performance on the 38-ligand EphB4 set (Figure 5.2 and Table 5.1); Kendall's τ was $0.56^{0.61}_{0.34}$. The results showed a substantial negative offset (MUE $5.76^{6.29}_{5.21}$ kcal mol^{-1}) which was likely due to insufficient sampling during the bound leg, which could entail failure of the protein to relax to the apo structure and failure of the binding site to rehydrate. Overestimation of the true K_{D} from the IC50 values may also contribute. However, obtaining accurate ranking and relative affinity estimates is the primary concern during virtual screening. For relative estimates, the above factors may cancel out between ligands, and absolute offsets are irrelevant. However, the non-offset plots are included to illustrate the effect of increasing sampling time on the offset, and because these offsets may be relevant when using substantially different receptor structures for different ligands.

Reducing the simulation time over two orders of magnitude substantially increased the overall offsets (Figure 5.2 and Table 5.1). The MUE increased by around 5 kcal mol^{-1} with each order of magnitude reduction in simulation time. There was also a clear increase in the variance of predictions for single ligands: the mean variances were $10.83^{13.29}_{8.55}$, $3.45^{4.18}_{2.78}$, and $2.45^{2.96}_{1.95}$ kcal mol^{-1} for the 0.01, 0.1, and 1 ns protocols, respectively. The ratios of variances between the 0.01/0.1 ns protocols and 0.1/1 ns protocols were 3.1 and 1.4, respectively, much less than the 10 expected assuming uncorrelated sampling. As discussed in Chapter 3, this suggests that windows in individual ABFE calculations become trapped in local regions of configurational space, producing correlated results. As a result, a given increase in computational cost yields a smaller increase in precision. However, despite the increased offset and greater variances, the ranking performances across all protocols were extremely similar; the 0.01 ns protocol produced a Kendall's τ of $0.56^{0.57}_{0.27}$, equal to the 1 ns result despite using two orders of magnitude less simulation time.

The retention of good ranking performance despite greater variance for individual calculations may result from an increased tendency to over-predict the binding affinity of stronger binders. This is shown by the lines of best fit in Figure 5.2, which are steeper for shorter protocols. A correlation between larger ligands and stronger binding affinities may explain this - larger, more flexible ligands tend to have greater affinities but often pose greater sampling challenges. Undersampling in the bound leg produces overestimates of binding affinity, possibly producing the relative overprediction of larger ligands. If this is the case, the short ABFE

protocols may be partially relying on this correlation for their ranking performance, which may obscure a loss of ability to differentiate molecules of similar size and flexibility. However, the retention of ranking performance with dramatic reductions in simulation time is promising and suggests the potential utility of fast ABFE in virtual screening.

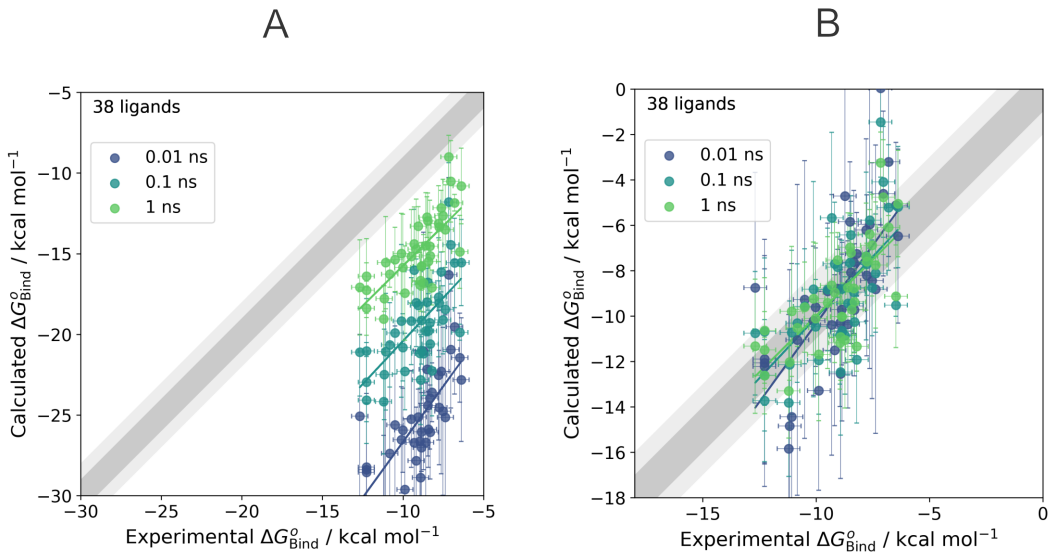


Figure 5.2: Correlation of calculated affinities with experiment for EphB4 using varying sampling times. A) Calculated affinities against experimental affinities and B) offset calculated affinities against experimental affinities. Offsets were calculated so that the mean of the calculated affinities equalled the mean of the experimental affinities. Uncertainties in the calculated values are 95 % t-based confidence intervals calculated from the deviations between 5 replicate runs. The darker and lighter shaded regions indicate the 1 and 2 kcal mol^{-1} deviations from exact agreement, respectively.

In virtual screening, the relative affinities estimated by ABFE calculations are more important than the absolute predictions themselves. Important questions are: "Given a calculated affinity difference, how confident am I in correctly selecting the stronger binder?" or conversely, "Given an experimental affinity difference, how confident am I that ABFE will correctly predict which binder is stronger?". These questions provide alternative viewpoints for evaluating binding affinity prediction performance (Figure 5.3). Panels A and B of Figure 5.3 answer these questions for the 1 ns protocol. Panel A shows that when the predicted absolute value of $\Delta\Delta G^o_{\mathrm{Bind,Calc.}}$ is greater than 2 kcal mol^{-1}, the 1 ns protocol correctly predicts the stronger binder with high confidence. When the predicted absolute value of $\Delta\Delta G^o_{\mathrm{Bind,Calc.}}$ is less than 2 kcal mol^{-1}, the 1 ns protocol mis-predicts the stronger binder almost as often as it predicts correctly. Panel B shows that when the measured absolute value of $\Delta\Delta G^o_{\mathrm{Bind,Exp.}}$ between two ligands is between 0

Table 5.1: Performance metrics for EphB4 ABFE calculations[a]

Metric	Protocol		
	0.01 ns	0.1 ns	1 ns
MUE / kcal mol^{-1}	$16.33^{17.28}_{15.44}$	$10.35^{11.04}_{9.70}$	$5.76^{6.29}_{5.21}$
RMSE / kcal^2 mol^{-2}	$16.50^{17.60}_{15.78}$	$10.54^{11.28}_{10.00}$	$5.95^{6.54}_{5.53}$
R^2	$0.50^{0.53}_{0.14}$	$0.45^{0.55}_{0.17}$	$0.55^{0.61}_{0.25}$
Spearman ρ	$0.77^{0.77}_{0.38}$	$0.67^{0.76}_{0.36}$	$0.75^{0.80}_{0.48}$
Kendall τ	$0.56^{0.57}_{0.27}$	$0.49^{0.56}_{0.25}$	$0.56^{0.61}_{0.34}$
Mean per-Ligand σ^2 / kcal^2 mol^{-2}	$10.83^{13.29}_{8.55}$	$3.45^{4.18}_{2.78}$	$2.45^{2.96}_{1.95}$

[a] Uncertainties are 95 % confidence intervals generated by bootstrapping (10000 iterations of resampling with replacement).

and 1 kcal mol^{-1}, the probability of predicting the stronger binder is close to the random value of 0.5. In contrast, this probability rises to around 0.8 when the absolute value of $\Delta\Delta G^o_{\text{Bind,Exp.}}$ rises to between 2 and 3 kcal mol^{-1}, and continues to rise as the measured free energy differences continue to increase. Dropping the number of replicate runs in the 1 ns protocol from 5 to 1 produces a consistent reduction in the probabilities of predicting the stronger binder. Surprisingly, panel C suggests that reducing the simulation time over 2 orders of magnitude does not dramatically reduce the probability of predicting the stronger binder of a pair, consistent with the retention of good ranking performance.

To investigate whether short-timescale performance was retained for a more realistic ABFE use-case, the non-adaptive results for CycloD from Chapter 3 were re-analysed using small fractions of the total simulation time (Figure 5.4 and Table 5.2). As observed for EphB4, reducing the simulation time produced increasingly negative offsets in the computed free energy differences, which were relatively greater for stronger binders. In addition, identical ranking performance was obtained with the 0.01 ns protocol compared to the 5 ns protocol (Kendall's τ $0.61^{0.83}_{0.11}$). This suggests that extremely short ABFE protocols may usefully rank structurally diverse ligands in some cases. This is especially promising as the protocol used here involves independent sampling for each λ window. Hence, the windows are trivially parallelisable and extremely short wall-clock times may be obtained when substantial compute is available.

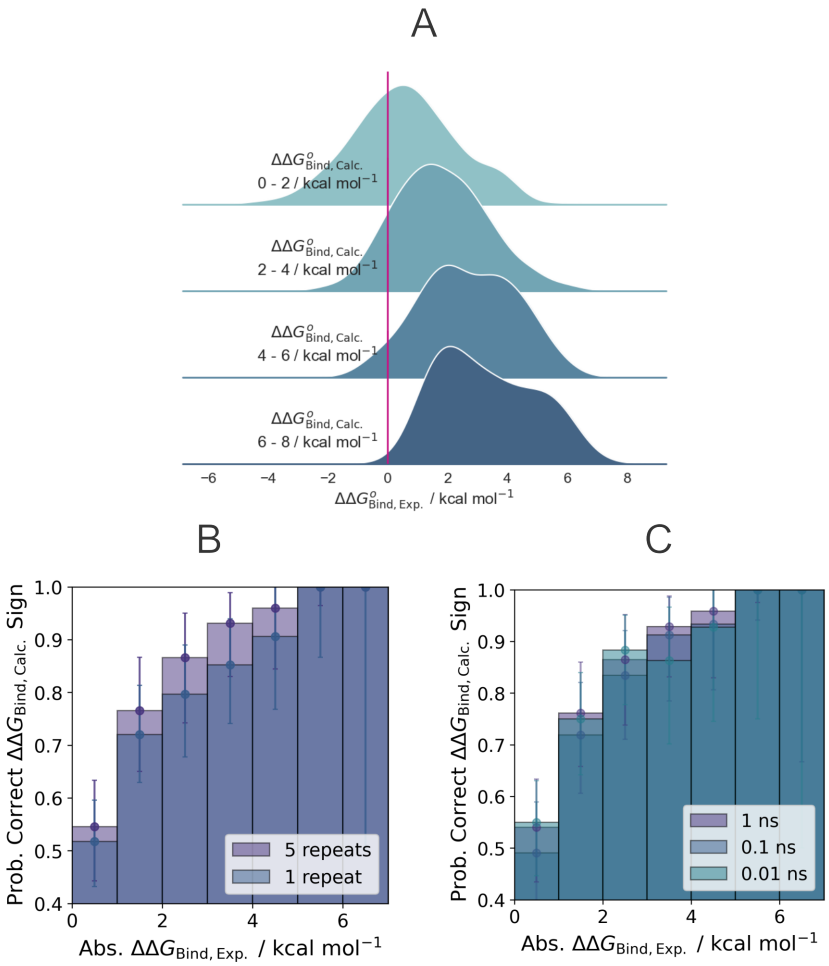


Figure 5.3: Ranking ability of ABFE protocols by experimental $\Delta\Delta G^o_{\text{Bind}}$. A) Kernel density estimates of distributions of $\Delta\Delta G^o_{\text{Bind,Exp.}}$ for given $\Delta\Delta G^o_{\text{Bind,Calc.}}$ intervals. B) Probability of $\Delta\Delta G^o_{\text{Bind,Calc.}}$ having the correct sign given the absolute value of $\Delta\Delta G^o_{\text{Bind,Calc.}}$ using the 1 ns protocol and all 5 replicate results per ligand, or a single replicate run. Absolute $\Delta\Delta G^o_{\text{Bind,Calc.}}$ binned with bin width 1 kcal mol^{-1}. C) Probability of $\Delta\Delta G^o_{\text{Bind,Calc.}}$ having the correct sign given the absolute value of $\Delta\Delta G^o_{\text{Bind,Calc.}}$ using different ABFE protocols. All protocols use 5 replicates per ligand, and the "1 ns" results are identical to the "5 repeats" results in B). Uncertainties are 95 % confidence intervals generated by bootstrapping (10000 iterations of resampling with replacement).

5.3.2 The Performance of Fast ABFE on Further Test Systems

To further assess the performance of fast ABFE calculations, the 0.1 ns protocol was applied to the 18-ligand HSP90 system from Alibay et al. (a realistic ABFE use case with relatively diverse ligands) and two common RBFE benchmark systems: TYK2 and p38 (Figure 5.5 and Table 5.3).[1,355]

Table 5.2: Performance metrics for CycloD ABFE calculations[a]

Metric	Protocol		
	0.01 ns	0.1 ns	5 ns
MUE / kcal mol^{-1}	$16.71^{19.22}_{14.45}$	$6.98^{8.63}_{5.46}$	$3.52^{4.49}_{2.65}$
RMSE / kcal^2 mol^{-2}	$17.08^{19.71}_{14.89}$	$7.43^{9.12}_{5.82}$	$3.85^{4.79}_{2.98}$
R^2	$0.86^{0.95}_{0.40}$	$0.78^{0.93}_{0.40}$	$0.81^{0.93}_{00.47}$
Spearman ρ	$0.77^{0.93}_{0.20}$	$0.73^{0.93}_{0.27}$	$0.77^{0.93}_{0.23}$
Kendall τ	$0.61^{0.83}_{0.11}$	$0.55^{0.83}_{0.17}$	$0.61^{0.83}_{0.11}$

[a] Uncertainties are 95 % confidence intervals generated by bootstrapping (10000 iterations of resampling with replacement).

Table 5.3: Performance metrics for HSP90, TYK2, and P38 0.1 ns ABFE calculations[a]

Metric	Target		
	HSP90	TYK2	P38
MUE / kcal mol^{-1}	$8.58^{9.47}_{7.74}$	$5.83^{6.73}_{5.01}$	$11.10^{12.19}_{9.96}$
RMSE / kcal^2 mol^{-2}	$8.78^{9.76}_{7.94}$	$5.94^{6.97}_{5.32}$	$11.57^{12.64}_{10.68}$
R^2	$0.59^{0.80}_{0.11}$	$0.07^{0.30}_{0.00}$	$0.18^{0.36}_{0.03}$
Spearman ρ	$0.63^{0.84}_{0.19}$	$0.24^{0.58}_{-0.30}$	$0.57^{0.70}_{0.24}$
Kendall τ	$0.47^{0.67}_{0.12}$	$0.13^{0.41}_{-0.21}$	$0.38^{0.50}_{0.16}$

[a] Uncertainties are 95 % confidence intervals generated by bootstrapping (10000 iterations of resampling with replacement).

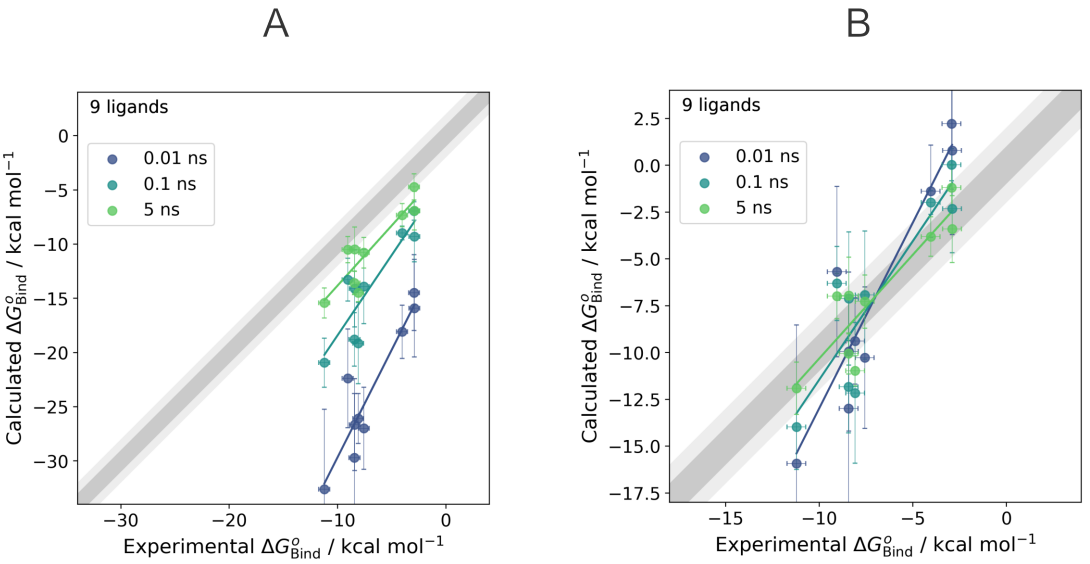


Figure 5.4: Correlation of calculated affinities with experiment for CycloD using varying sampling times. A) Calculated affinities against experimental affinities and B) offset calculated affinities against experimental affinities. Offsets were calculated so that the mean of the calculated affinities mean equalled the mean of the experimental affinities. Uncertainties in the calculated values are 95 % t-based confidence intervals calculated from the deviations between 5 replicate runs. The darker and lighter shaded regions indicate the 1 and 2 kcal mol^{-1} deviations from exact agreement, respectively.

The initial HSP90 calculations revealed a single obvious outlier (Figure D.1): the predicted $\Delta G^o_{\mathrm{Bind}}$ for ligand 28 (see Alibay et al.) was around 8 kcal mol^{-1} more positive than expected based on the other ligands. The solvent-exposed binding site of HSP90 is known to pose water sampling challenges in ABFE calculations.[1,209] Alibay et al. employed an MD/MC water-sampling methodology during equilibration of the fully-interacting complex[196] and noted that only 1 binding-site water was added to the ligand 24 complex. In contrast, 3 waters were added to all other complexes. However, they found that ABFE calculations for ligand 24 with 3 or 1 waters in the binding site had no significant effect on the predicted affinity. Examining the input files for ligand 24 and 31 obtained from Alibay et al. revealed that a buried water was missing from the ligand 24 complex where there was missing electron density in the crystal structure (Figure D.1). In contrast to the results of Alibay et al., inserting this water into the ligand 24 complex produced a dramatic change in the calculated $\Delta G^o_{\mathrm{Bind}}$, which dropped from -7.78 ± 1.70 kcal mol^{-1} to -15.05 ± 2.82 kcal mol^{-1}, in line with the other results. This suggests that this water is incorrectly omitted from the crystal structure, and the ability of the short ABFE protocol to distinguish

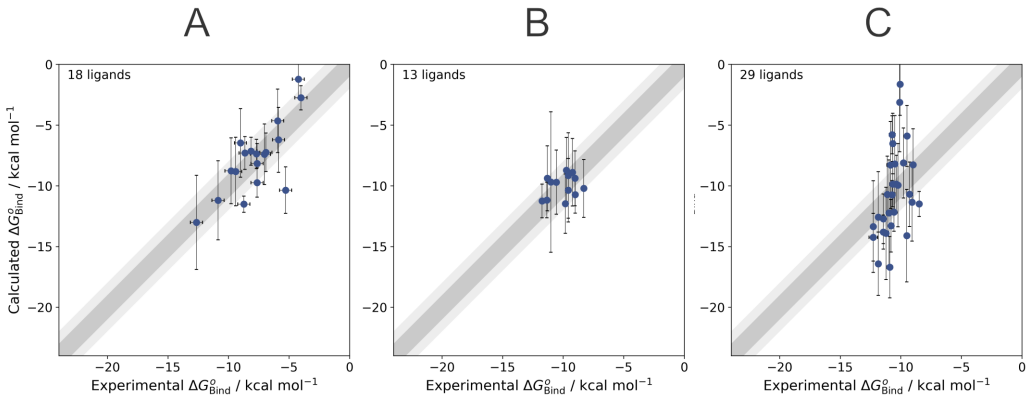


Figure 5.5: Correlation of calculated affinities with experiment for A) HSP90, B) TYK2, and C) P38 using 0.1 ns protocol. Affinities are offset for each target so that the mean of the calculated affinities equals the mean of the experimental affinities. Uncertainties in the calculated values are 95 % t-based confidence intervals calculated from the deviations between 5 replicate runs. The darker and lighter shaded regions indicate the 1 and 2 kcal mol^{-1} deviations from exact agreement, respectively.

likely incorrect input structures is reassuring. While this result contrasts with the findings of Alibay et al., it is consistent with the results of Baumann et al.,[209] who found that starting HSP90 ABFE calculations without waters in the binding site produced less favourable binding affinities by several kcal mol^{-1}.

With the missing water molecule inserted into the ligand 24 complex binding site, the HSP90 0.1 ns ABFE results showed reasonable ranking performance (Kendall's τ $0.47^{0.67}_{0.12}$, Table 5.3) providing further confidence that fast ABFE calculations may usefully rank for structurally diverse ligands. While the results obtained by Alibay et al. appear slightly better (for example, Pearson r = 0.96 ± 0.03, compared to $0.77^{0.89}_{0.34}$ here), it is unclear whether this is due to extensive sampling (20 ns per window compared to 0.1 ns per window here) or other methodological differences (Alibay et al. use GROMACS 2021, while SOMD is used here).[1]

In contrast to EphB4, CycloD, and HSP90, the ranking performance was close to random for TYK2 with the 0.1 ns protocol (Kendall's τ $0.13^{0.41}_{-0.21}$). However, the range of experimental binding affinities was small (around 3 kcal mol^{-1}) and the RMSE of the mean-offset results ($1.18^{1.20}_{2.52}$ kcal mol^{-1}) was not excessive. Therefore, better ranking may be observed with a wider range of experimental affinities. However, the results were poor for P38. While there was some correlation (Kendall's τ $0.38^{0.50}_{0.16}$), the spread of the mean-offset results was large (RMSE $3.28^{4.36}_{2.76}$ kcal mol^{-1}). The two largest outliers (excessively positive$\Delta G^o_{\text{Bind,Calc.}}$)

were found to have a substituted benzene group pointing in a different direction to all other ligands. It is possible that the input poses were not correct, and that it was not possible to recover the correct poses with the limited 0.1 ns sampling. Alternatively, the wide spread of affinities may suggest that sufficient conformational sampling is essential to obtain accurate relative affinities for this series and that 0.1 ns is insufficient.

Schrödinger's FEP+ program has been used to calculate ABFEs for TYK2 and P38 with much lower error and greater correlation than shown here (R^2 values of 0.72 and 0.55 for TYK2 and P38, respectively).[61] While they used more ligands per target, resulting in a ≈ 1 kcal mol^{-1} wider range of experimental binding affinities for TYK2, their results are unambiguously superior. It is unclear how much the increased performance is due to increased sampling time, improved sampling algorithms such replica exchange with solute scaling (REST2), or a combination of both.[107] To isolate the effect of sampling time, future work should compare the current TYK2 and P38 results to those obtained with longer protocols.

5.4 Conclusion

This work demonstrates that reducing sampling time by orders of magnitude in ABFE calculations does not necessarily harm ranking performance. This is promising for virtual screening applications, as it suggests that fast ABFE calculations may be used to rapidly re-rank many top-scoring ligands from docking, or to perform fast active learning cycles. This is likely to substantially improve the enrichment of active compounds.[60] The study was facilitated by an automated workflow implemented in the A3FE software, which is based on the methods presented in Chapters 2 and 3.

However, this study is incomplete, and substantially more validation should be carried out to determine whether fast ABFE protocols are useful for virtual screening in general. Few systems were studied in this work. While an effort was made to include sets of structurally diverse ligands, short ABFE calculations should be tested with many varied systems to determine when these protocols do, and do not perform well. In addition, "standard length" and shorter ABFE protocols should be compared for many more systems to isolate the effect of sampling time. It would be interesting to see if increased sampling improved the poor performances we observed for TYK2 and P38. The effect of reducing sampling time may be more pronounced when using enhanced sampling methods

such as Hamiltonian replica exchange (HREX) and should be investigated. It would also be interesting to determine if fast ABFE protocols can retain performance for systems where molecular weight does not correlate well with the binding affinities.

Another limitation of the study is that the fast ABFE calculations were not compared to inherently "fast" end-state methods such as molecular mechanics Poisson-Boltzmann surface area (MM/PBSA) calculations.[38,40,358] An MM/PBSA-based active-learning virtual screening workflow has recently been described by Loeffler et al..[347] This method typically performs less well than "standard-length" ABFE calculations,[59] but this has not been determined for short ABFE calculations which have a comparable computational cost.

Finally, fast ABFE calculations should be validated using data which more closely reflects the intended use case of virtual screening. That is, their ability to improve enrichment in a virtual screening pipeline should be assessed similarly to Feng et al. for "standard-length" ABFE calculations.[60]

Chapter 6

Conclusion

This work investigated methods to improve the accuracy, efficiency, and automation of alchemical ABFE calculations.

In Chapter 2, receptor-ligand restraint schemes were compared for computing the ABFE of the MIF/MIF180 complex. A restraint scheme based on multiple distance restraints between receptor and ligand atoms was proposed, which avoids the inherent instabilities of the popular Boresch restraints, and may improve convergence by more strongly restricting the relative movements of the receptor and ligand. This was shown to produce estimates in good agreement with Boresch restraints. In contrast, calculations without orientational restraints produced erroneously favourable free energies of binding by up to ≈ 4 kcal mol^{-1}. Methods for automatically selecting restraint parameters, in particular by fitting force constants to fluctuations, were shown to be reliable. These approaches offer new options for the deployment of ABFE calculations.

Building on these automated restraint selection algorithms, a fully-automated open-source ABFE workflow was developed in Chapter 3 to facilitate high-throughput applications. Algorithms to automate the selection of λ windows, the allocation of simulation time, and the truncation of unequilibrated data were presented. The simple window selection procedure was shown to be inexpensive and robust, as was the truncation point selection heuristic. Allocating sampling time based on inter-run uncertainties was shown to concentrate sampling time where there were sampling issues, but this only improved convergence in rare cases when the timescales of the sampling issues were comparable to the maximum achievable sampling time. Overall, the workflow provided equivalent results to non-adaptive schemes for a variety of systems, while often accelerating equilibration.

While an equilibration detection heuristic was proposed in Chapter 3, it was only applicable to multiple replicates and quantitative testing was challenging due to the high computational cost of the ABFE data. In Chapter 4, White's marginal standard error rule was reformulated to provide a spectrum of truncation point selection heuristics differing in their treatment of autocorrelation.[295] These included a method effectively equivalent to Chodera's,[108] and were all applicable to single runs. To allow quantitative conclusions to be drawn, these methods were tested on ensembles of synthetic time series modelled on free energy change estimates from the long absolute binding free energy calculations in Chapter 3. Methods that more thoroughly accounted for autocorrelation often showed late and variable truncation times, while methods that less thoroughly accounted for autocorrelation often showed early truncation, relative to the optimal truncation point. A method was recommended that achieved robust performance across the test sets by balancing these two extremes. None of the methods reliably detected insufficient sampling. All methods were implemented in the open-source Python package RED (robust equilibration detection), which was designed to replace the equivalent functions from the widely used software PyMBAR. It is hoped that this will improve the reliability of computed free energy changes.

Finally, the automated workflow developed in Chapter 3 was used to assess the ranking performance of very short ABFE calculations in Chapter 5. For series of EphB4 and Cyclophillin-D binders, it was found that reducing the simulation time over two orders of magnitude introduced large negative offsets to the computed free energies, but preserved equivalent ranking performance. Fast calculations using only 0.1 ns of simluation time per window were shown to produce good ranking for HSP90 ligands, but poorer performance for TYK2 and P38. However, it was unclear if these poor performances were due to the short simulation times, as the full-length simulations were not performed due to time constraints. Regardless, these results warrant further investigation of short ABFE calculations; the availability of reliably accurate ranking predictions an order of magnitude faster than with "standard" ABFE would be extremely useful for virtual screening calculations.

The methods and software presented in this thesis provide robust automation of ABFE calculations and may improve efficiency by reducing user and computer time. However, the relative merits and recommended use cases of some of these methods should be determined through more comprehensive studies. Specifically, the multiple distance restraints scheme proposed was shown to generate equivalent results to the Boresch scheme, but large-scale testing on a wide variety of com-

plexes will be required to determine whether they deliver convergence benefits. The promising retention of ranking ability at drastically reduced simulations time was only demonstrated for a few systems, and large-scale testing will be required to better understand the general performance and domain of applicability of fast ABFE calculations.

More work is required to develop efficient ABFE protocols for problems relevant to the pharmaceutical industry. For example, the methods in this thesis were tested on receptor-ligand complexes which had crystal structures available, or at least were prepared from the crystal structure of similar ligands. In virtual screening, the uncertainty of the binding pose is likely to be much higher, which may require multiple calculations of different binding poses. Streamlined integrated workflows, for example using induced-fit docking on AlphaFold structures followed by scoring with ABFE, are required to provide value in drug discovery.[359] This work has implicitly taken the narrow view that all ABFE predictions are equally important. This is patently untrue in real drug discovery. Different predictions may be more or less informative depending on prior knowledge of affinities for similar molecules. Equivalently, ABFE predictions are less useful for a molecule when a machine learning affinity prediction has high certainty. Active learning of ABFE affinity predictions is promising method to make intelligent use of expensive information.[348] Additionally, if a molecule is predicted to bind very poorly, the prediction need not be precise. However, if similar and high affinities are predicted for two molecules, higher precision is desirable. More useful ABFE workflows would minimise the cost of answering questions asked in drug discovery, rather than simpler targets such as minimising the root mean squared error of all predictions.

Similarly, absolute calculations may not provide the most efficient alchemical pathways to free energy differences of interest. Alternatives should be compared. For example, if a relative free energy difference between structurally dissimilar ligands is required, it may be more efficient to simultaneously decouple one ligand while recoupling the other.[153,183] This might prevent sampling issues arising from slow rehydration of the empty binding site, for instance. The alchemical transfer method is also promising for relative binding free energy calculations with dissimilar ligands, as well as binding selectivity calculations.[212,360]

Finally, there has been immense interest in predicting binding affinities with machine learning. However, general machine learning affinity models are plagued by failure to generalise beyond their training sets.[32–35] Until much more high-quality training data is available, the complementary use of ABFE calculations and problem-specific machine-learned models through active learning may be more fruitful.

Appendix A

Comparison of Receptor-Ligand Restraint Schemes

A.1 Restraints are not Necessary to Prevent Errors Arising from an Incorrect Definition of the Bound State for Strong Binders

When no receptor-ligand restraints are used in an ABFE calculation, the implicit definition of the bound state includes configurations where the ligand is anywhere in the entire simulation box relative to the receptor. This section shows that even in the limit of perfect sampling, this introduces errors in computed binding free energies for weak binders, but not for sufficiently strong binders in a sufficiently small simulation box.

In a hypothetical alchemical calculation with perfect sampling, receptor-ligand restraints would only be required for weak binders. This is because for strong binders, the free energy of the macrostate where the interacting ligand is confined to the binding site is almost identical to the free energy of the macrostate where the ligand is allowed to explore the entire simulation box. This argument is equivalent to that of Gilson et al.,[49] who showed that the standard chemical potential is insensitive to the cutoff function (e.g. restraints) for strong binders. Intuitively, the probability of finding the ligand outside the binding site when it is fully interacting is negligible, and therefore including the unbound configurations makes only a negligible difference to the definition of the state when the box is not extremely large. In such a hypothetical simulation with a strong binder, perfect sampling, and no restraints, the ligand would be decoupled while it was

allowed to explore the entire simulation box in the presence of the receptor. The standard state dependence would then be included by correcting the binding free energy by $-k_{\mathrm{B}}T\ln\frac{V_{\mathrm{Box}}}{V^o}$, where V_{Box} is the volume of the simulation box, and $V^o = 1660$ Å^3 is the standard state volume. The free energy obtained would be the same as when restricting ligand sampling to the binding site (assuming that this definition included all low energy configurations,[49] or that restraints were introduced starting with the ligand kinetically trapped in the binding site and the free energy cost of turning them on was accounted for), as done in this work.

In contrast, for a weak binder, the free energy of the macrostate where the interacting ligand can explore the entire simulation box is substantially different to the macrostate where the ligand only samples the binding site. Therefore, starting from the macrostate where the ligand may sample the entire simulation box introduces substantial error, and restraints are required to ensure that the ligand only samples the binding site. In this case, restraints are required for the correct binding site definition.

To illustrate this, the free energies obtained from these hypothetical simulations can be calculated A) restricting the ligand sampling to the binding site ($\Delta G^o_{\mathrm{Bind,\ Site}}$) and B) allowing the ligand to sample the entire simulation box ($\Delta G^o_{\mathrm{Bind,\ Box}}$). $\Delta G^o_{\mathrm{Bind,\ Site}}$ gives the difference in free energy between the fully-interacting ligand in the binding site and the fully interacting ligand in a water box of volume 1660 Å^3. Therefore, assuming negligible interaction of the ligand and receptor outside of the binding site, and negligible volume of the receptor compared to the box volume, the two free energies of binding are related by

$$\Delta G^o_{\mathrm{Bind,\ Box}} = -k_{\mathrm{B}}T\ln\left(\exp^{-\frac{\Delta G^o_{\mathrm{Bind,\ Site}}}{k_{\mathrm{B}}T}} + \frac{V_{\mathrm{Box}}}{V^o}\right).$$

Taking $V_{\mathrm{Box}} = 512000$ Å^3 (side length of 80 Å) and $T = 298$ K, the two binding free energies can be compared (Figure A.1):

When $\Delta G^o_{\mathrm{Bind,\ Site}} > -k_{\mathrm{B}}T\ln\frac{V_{\mathrm{Box}}}{V^o}$, $\Delta G^o_{\mathrm{Bind,\ Box}}$ tends to $-k_{\mathrm{B}}T\ln\frac{V_{\mathrm{Box}}}{V^o} = -3.4$ kcal mol^{-1}, the entropy of releasing a water from a box of volume 1660 to 51200 Å^3. For $\Delta G^o_{\mathrm{Bind,\ Site}} < -k_{\mathrm{B}}T\ln\frac{V_{\mathrm{Box}}}{V^o}$, the difference between the two free energies, $\Delta G^o_{\mathrm{Bind,\ Site}}$ and $\Delta G^o_{\mathrm{Bind,\ Box}}$, is negligible and thus restraints are not required to define the binding site. Thus, in the case of strong binders, receptor-ligand restraints are only required to prevent sampling issues (for example the ligand would have to sample outside the binding site as soon as the unbound state began to contribute significantly to the configurational integral).

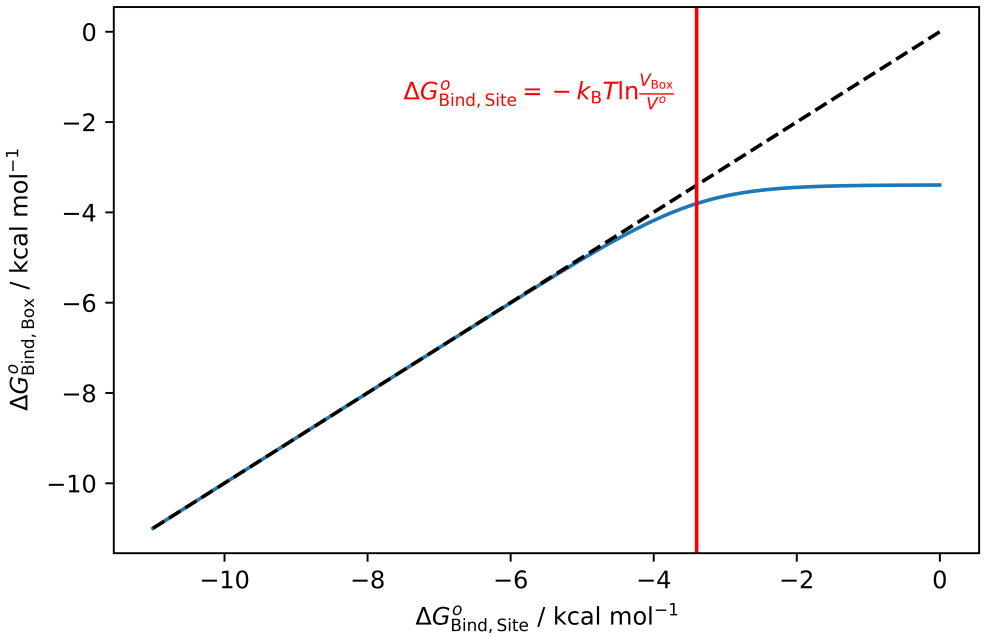


Figure A.1: $\Delta G^o_{\text{Bind, Site}}$ against $\Delta G^o_{\text{Bind, Box}}$ for a box size of 51200 Å^3 at 298 K. The dashed line shows $\Delta G^o_{\text{Bind, Site}} = \Delta G^o_{\text{Bind, Box}}$. For $\Delta G^o_{\text{Bind, Site}} < -k_B T \ln \frac{V_{\text{Box}}}{V^o}$, there is a negligible difference between the two free energies, and thus restraints are not required to define the binding site.

A.2 Derivation of a General Expression for The Standard Free Energy of Releasing Receptor-Ligand Restraints

The free energy of releasing the restraint on the decoupled ligand is given by the ratio of configurational integrals of states 3 and 4 as defined in the main text (Figure 2.2)

$$\Delta G_{\text{Release}} = -k_B T \ln \frac{Z_{\text{State 3}}}{Z_{\text{State 4}}}, \tag{A.1}$$

where k_B is the Boltzmann constant, T is the temperature, and $Z_{\text{State 3}}$ and $Z_{\text{State 4}}$ are the configurational integrals for states 3 and 4. It can be shown (Equation 10 of Boresch et al.[50]) that this can be written as

$$\Delta G_{\text{Release}} = -k_B T \ln \frac{Z_{\text{R, Free}} Z_{\text{L, Free}}}{Z_C}, \tag{A.2}$$

where "R, Free", "L, Free", and C denote the configurational integrals for the free receptor, the free decoupled ligand, and the complex where there are restraints between the decoupled ligand and the receptor, respectively. Note that $Z_{\mathrm{R,\ Free}}$ and Z_{C} include contributions from the interaction of the receptor with the solvent and involve integration over the solvent degrees of freedom (DoF), while $Z_{\mathrm{L,\ Free}}$ does not include solvent interactions or integration over the solvent degrees of freedom. By integrating out the six external degrees of freedom, equation A.2 can be rewritten as

$$\Delta G_{\mathrm{Release}} = -k_{\mathrm{B}}T \ln \frac{\tilde{Z}_{\mathrm{R,\ Free}}\tilde{Z}_{\mathrm{L,\ Free}}V_{\mathrm{Box}}8\pi^2}{\tilde{Z}_{\mathrm{C}}}, \tag{A.3}$$

where V_{Box} is the volume of the simulation box and $\tilde{Z}$ denotes integration over the $3N-6$ internal degrees of freedom, where N is the number of atoms in the ligand for $\tilde{Z}_{\mathrm{L,\ Free}}$, the number of atoms in the receptor and water box for $\tilde{Z}_{\mathrm{R,\ Free}}$, or the number of atoms in the restrained receptor-decoupled ligand complex and water box for $\tilde{Z}_{\mathrm{C}}$. Equation A.3 stresses that $\Delta G_{\mathrm{Release}}$ is dependent on the box size, which is undesirable. As discussed by Gilson et al.[49], a correction must be applied in order to yield the standard absolute binding free energies. This standard state dependence was missing from early "double annihilation" calculations,[361] but is correctly accounted for in the modern "double decoupling" alchemical approach. This correction is applied by replacing V_{Box} with $V^o = 1660\text{Å}^3$, the standard state volume, to yield the standard free energy of releasing the restraints

$$\Delta G^o_{\mathrm{Release}} = -k_{\mathrm{B}}T \ln \frac{\tilde{Z}_{\mathrm{R,\ Free}}\tilde{Z}_{\mathrm{L,\ Free}}V^o8\pi^2}{\tilde{Z}_{\mathrm{C}}}, \tag{A.4}$$

which is independent of the box size. $\tilde{Z}_{\mathrm{C}}$ can be expanded, giving

$$\begin{aligned}\Delta G^o_{\mathrm{Release}} = &- k_{\mathrm{B}}T \ln \tilde{Z}_{\mathrm{R\ Free}}\tilde{Z}_{\mathrm{L\ Free}}V^o8\pi^2 \\ &+ k_{\mathrm{B}}T \ln \int e^{-\frac{W_{\mathrm{r}}(\mathbf{x}_{\mathrm{Ext}})+W_{\mathrm{R,\ Comp.}}(\mathbf{x}_{\mathrm{Ext}})+W_{\mathrm{L,\ Comp.}}(\mathbf{x}_{\mathrm{Ext}})}{k_{\mathrm{B}}T}}|\mathbf{J}|\mathrm{d}\mathbf{x}_{\mathrm{Ext}}\end{aligned} \tag{A.5}$$

where $W_{\mathrm{r}}(\mathbf{x}_{\mathrm{Ext}})$ and $W_{\mathrm{L,\ Comp.}}(\mathbf{x}_{\mathrm{Ext}})$ are the potentials of mean force (PMFs) associated with the receptor-ligand restraint energy and the decoupled ligand internal energy (as part of the complex), respectively. $W_{\mathrm{R,\ Comp.}}(\mathbf{x}_{\mathrm{Ext}})$ is the PMF of the receptor internal energy (as part of the complex) and the remaining contributions to the total system energy. These are with respect to the six relative external degrees of freedom of the receptor and ligand, which are contained in

the vector $\mathbf{x}_{\text{Ext}}$. These PMFs result from integration over the $3N-12$ remaining internal degrees of freedom of the complex. The form of the Jacobian determinant, $|\mathbf{J}|$, depends on the coordinate transformation used to extract the relative external degrees of freedom from the internal degrees of freedom of the system.

$W_{R,\text{Comp.}}(\mathbf{x}_{\text{Ext}})$ and $W_{L,\text{Comp.}}(\mathbf{x}_{\text{Ext}})$ can be expanded to yield

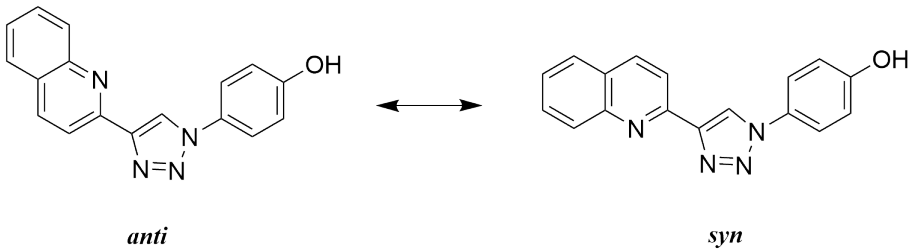
$$\begin{aligned}\Delta G^o_{\text{Release}} = &- k_{\text{B}}T \ln \tilde{Z}_{\text{R Free}} \tilde{Z}_{\text{L Free}} V^o 8\pi^2 \\ &+ k_{\text{B}}T \ln \int e^{-\frac{W_{\text{r}}(\mathbf{x}_{\text{Ext}})}{k_{\text{B}}T}} \times \\ &\quad e^{-\frac{\Delta_1 W_{\text{R, Comp.}}(\mathbf{x}_{\text{Ext},0}) + \Delta_2 W_{\text{R, Comp.}}(\mathbf{x}_{\text{Ext}}) + \tilde{G}_{\text{R, Free}}}{k_{\text{B}}T}} \times \\ &\quad\quad e^{-\frac{\Delta_1 W_{\text{L, Comp.}}(\mathbf{x}_{\text{Ext},0}) + \Delta_2 W_{\text{L, Comp.}}(\mathbf{x}_{\text{Ext}}) + \tilde{G}_{\text{L, Free}}}{k_{\text{B}}T}} |\mathbf{J}| \mathrm{d}\mathbf{x}_{\text{Ext}}, \end{aligned} \tag{A.6}$$

where the vector $\mathbf{x}_{\text{Ext},0}$ contains the values of the relative external degrees of freedom at which the PMF $W_{\text{r}}(\mathbf{x}_{\text{Ext}})$ has its minimum. $\tilde{G}_{\text{L, Free}}$ is the free energy of the internal degrees of freedom of the decoupled ligand in the absence of receptor-ligand restraints, $\Delta_1 W_{L,\text{Comp.}}(\mathbf{x}_{\text{Ext},0})$ is the free energy of "preorganising" the ligand intramolecular degrees of freedom when the restraint energy is at its minimum, and $\Delta_2 W_{\text{L, Comp.}}(\mathbf{x}_{\text{Ext}}) = W_{\text{L, Comp.}}(\mathbf{x}_{\text{Ext}}) - \Delta_1 W_{L,\text{Comp.}}(\mathbf{x}_{\text{Ext},0}) - \tilde{G}_{\text{L, Free}}$ is the additional free energy penalty from further distortion of the ligand internal degrees of freedom. It is 0 when $\mathbf{x}_{\text{Ext}} = \mathbf{x}_{\text{Ext},0}$. These terms give rise to $\tilde{Z}_{\text{L, Free}}$, $\Delta G_{\text{L, Preorg.}}$, and $\Delta G_{\text{L, Distort.}}$, respectively. There are equivalent terms for the receptor, which also include contributions from solvent interactions. Rewriting Equation A.6 gives

$$\begin{aligned}\Delta G^o_{\text{Release}} &= - k_{\text{B}}T \ln \tilde{Z}_{\text{R Free}} \tilde{Z}_{\text{L Free}} V^o 8\pi^2 \\ &\quad + k_{\text{B}}T \ln \int e^{-\frac{W_{\text{r}}(\mathbf{x}_{\text{Ext}})}{k_{\text{B}}T}} |\mathbf{J}| \mathrm{d}\mathbf{x}_{\text{Ext}} + k_{\text{B}}T \ln \tilde{Z}_{\text{R, Free}} \tilde{Z}_{\text{L, Free}} \\ &\quad - \Delta G_{\text{R, Preorg.}} - \Delta G_{\text{L, Preorg.}} - \Delta G_{\text{R, Distort.}} - \Delta G_{\text{L, Distort.}} \\ &= - k_{\text{B}}T \ln V^o 8\pi^2 + k_{\text{B}}T \ln \int e^{-\frac{W_{\text{r}}(\mathbf{x}_{\text{Ext}})}{k_{\text{B}}T}} |\mathbf{J}| \mathrm{d}\mathbf{x}_{\text{Ext}} - \Delta G_{\text{Preorg.}} - \Delta G_{\text{Distort.}}, \end{aligned} \tag{A.7}$$

where $\Delta G_{\text{Preorg.}}$ and $\Delta G_{\text{Distort.}}$ are the sums of the preorganisation and distortion terms for the receptor and the ligand.

A.3 The *syn* and *anti* conformations of MIF180



anti ***syn***

Figure A.2: The *syn* and *anti* conformations of MIF180.

A.4 Parameters for Boresch Restraints

Table A.1: Parameters for Boresch restraints, as labelled in Figure 2.3. K refers to a force constant and 0 denotes and equilibrium value.

	Restraint Set					
Parameter	**B1**	**B2**	**B3**	**B1-poseB**	**B1-P**	**B3-P**
c (index)	4949	937	51	4946	965	605
b (index)	4944	933	47	4942	961	601
a (index)	4946	935	49	4944	963	603
A (index)	11	10	12	14	5	6
B (index)	2	13	6	4	3	9
C (index)	3	20	19	5	14	12
r_0 (Å)	5.92	8.14	7.84	8.69	6.94	8.48
$\theta_{A,0}$ (rad)	1.85	2.06	0.81	1.54	0.66	1.24
$\theta_{B,0}$ (rad)	1.59	1.89	1.74	1.52	2.10	2.09
$\phi_{A,0}$ (rad)	-0.30	1.68	2.59	-1.22	-0.14	2.55
$\phi_{B,0}$ (rad)	-1.55	1.52	-1.20	2.80	-1.32	-0.20
$\phi_{C,0}$ (rad)	2.90	0.20	2.63	-0.29	-0.11	2.15
K_r (kcal mol^{-1} Å^{-2})	25.49	10.92	10.25	12.25	14.27	6.32
K_{θ_A} (kcal mol^{-1} rad^{-2})	66.74	126.83	49.44	90.22	91.57	73.31
K_{θ_B} (kcal mol^{-1} rad^{-2})	38.39	98.43	99.26	115.70	100.27	63.90
K_{ϕ_A} (kcal mol^{-1} rad^{-2})	215.36	189.35	51.25	189.91	71.33	68.44
K_{ϕ_B} (kcal mol^{-1} rad^{-2})	49.23	58.81	25.98	42.35	72.02	69.07
K_{ϕ_C} (kcal mol^{-1} rad^{-2})	49.79	100.72	95.41	174.58	71.25	44.54

A.5 RMSF of Residues Containing Anchor Points

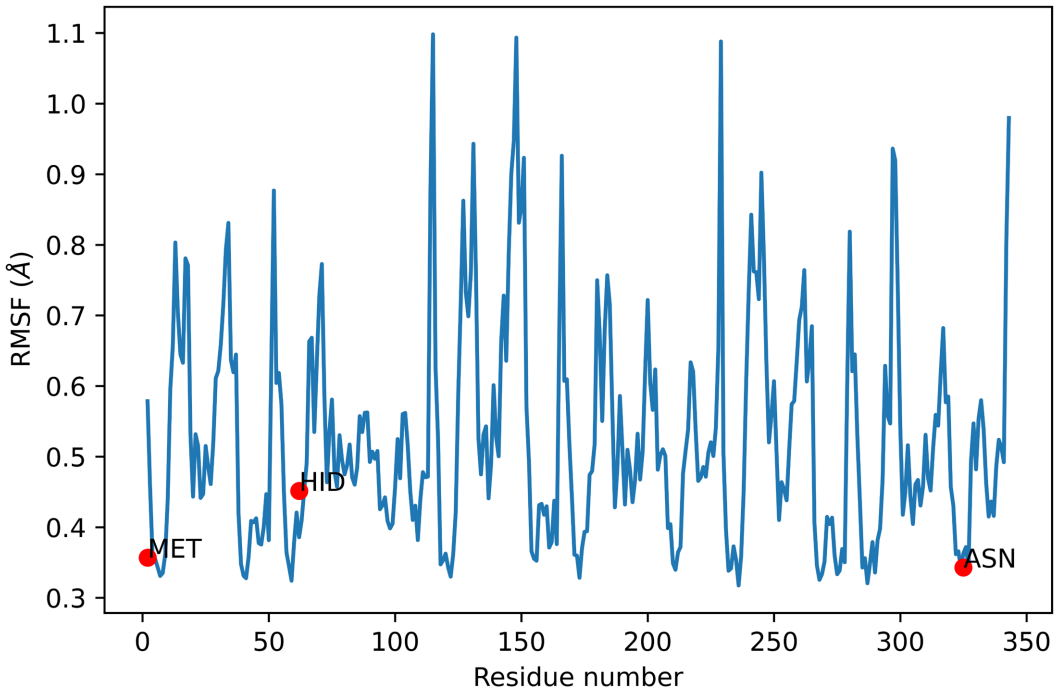


Figure A.3: The restraint selection algorithm produces protein anchor points on stable residues. The RMSF of the Cα atom of each residue was calculated over the unrestrained simulation used for restraint selection, after all frames were aligned to the average structure. MET, HID, and ASN residues highlighted in red contain the anchor points used for the B3, B2, and B1 restraints, respectively.

A.6 Challenges in Restraint Selection

The selection of Boresch restraints highlighted challenges related to the presence of multiple binding poses and the treatment of symmetry. Fitting the force constants to simulation provides a crude estimate of the PMFs with respect to the Boresch degrees of freedom. Thus, the force constants for the Boresch DoF are effectively doubled during the restraining stage. Under the approximations of Gaussian distributions and no correlation between the Boresch DoF, the free energy of turning on the restraints can be estimated from Equation 2.7; this should be approximately $-k_B T \ln((2^6)^{1/2}) = -1.23$ kcal mol^{-1} at 298 K in all cases.

This observation allows the identification of issues during the restraining stages. $\Delta G_{\text{Bound, Restrain}} > 5$ kcal mol^{-1} was calculated from initial simulations, which indicated an issue with binding poses which slowly interconverted on the timescale of simulation (Figure 2.6). This resulted from an almost immediate change to an alternate binding pose from the input structure during the simulation used to fit the restraints. Because this pose persisted throughout the unrestrained simulation, relatively tight force constants were selected, resulting in the calculation of large free energies of restraining to the alternative pose starting from the original input structure.

Subsequent calculations were run by fitting restraints for a single binding pose, and rerunning simulations from where the alternate pose was sampled. Reruns were only required at low values of λ during the restraining calculations as sampling of the alternate pose was prevented by the engaged restraints. These switches were easily identified by checking for large jumps in $\mathrm{d}H/\mathrm{d}\lambda$, as suggested by Baumann et al..[209]

A completely automated restraint selection algorithm would account for symmetry, which the method used here does not. The flip of the phenol group in MIF180 occurs infrequently on the timescale of the simulations, and does not occur at all during the early stages of decoupling in the bound leg. When transitions between symmetrical energy minima (e.g. flip of the phenol) are not sampled, and the symmetry of the minima is not broken by the introduction of restraints, some authors have applied symmetry corrections.[168] However, this is unnecessary,[73] because free energy changes are obtained from distributions of energies,[131] or their gradients with respect to λ.[116] These distributions are identical regardless

of whether one or all of the symmetrical minima are sampled, and therefore the free energy estimates do not change. Thus, no symmetry corrections are required in general here (beyond those accounting for the restraint of the ligand to a single binding site when there are three identical binding sites present per protein).

However, when the receptor-ligand restraints break this symmetry, this must be accounted for. The symmetric wells must either be sampled in equilibrium as the symmetry is broken (the phenol must sample both orientations in equilibrium as the restraints are turned on), or in the case where only one minimum is sampled (the phenol does not flip), a correction must be applied.[73] The latter occurs for the B1 restraints, which impose a large energy penalty for the flip of the phenol. The phenol does not flip during the restraining simulations and hence the associated correction is $-k_BT\ln 2 = -0.41$ kcal mol^{-1}. However, placing restraints on the phenol necessitates equilibrium sampling about the triazole-phenol dihedral during decoupling, which is not the case for the other restraints (because for the other restraints, rotation around this bond results in transitions between symmetrical energy minima). The requirement for both a symmetry correction and equilibrium sampling about the triazole-phenol dihedral could be removed by introducing a flat-bottomed restraint to prevent the phenol flip, as suggested by Wang et al..[158] This would only be required during the bound leg simulations. Alternatively, symmetry-adapted restraints could be used.[161]

It was observed that several of the restraint sets selected had large equilibrium r distances (Table A.1) of around 8 Å. It is possible that this resulted from a bias of the restraint selection algorithm towards large values of r as a result of scoring by minimum variance of the Boresch DoF alone: for the same absolute movement of anchor points a or A normal to the vector between anchors A and a, the variation in θ_A and θ_B will be smaller when r is greater. A more natural choice of "score" for prospective restraints may be the total configurational volume accessible in the decoupled state, calculated using Equation 2.7 for the force constants fitted to simulation. A smaller accessible configurational volume would score higher. The objective is the same as with scoring by total variance: to select restraints which mimic the strongest receptor-ligand interactions. The two metrics are closely related, because the force constants calculated from the variances of the DoF are used to calculate the accessible configurational volume. However, as a result of the Jacobian terms in Equation 2.7, this metric is more consistent between restraint sets with different equilibrium distances and angles, avoiding issues with bias as

discussed above. It would also be sensible to exclude groups which may be stable while interacting with the ligand, but which may be substantially more mobile when the ligand is decoupled, for example by restricting the protein anchor atoms to the backbone and Cα atoms.[170]

A.7 Fitting of Restraint Force Constants

Restraint force constants were fitted by measuring the distributions of the DoF to be restrained in the absence of any restraints, for the fully interacting complex. At equilibrium, a harmonic oscillator produces a Gaussian distribution about its equilibrium value

$$P(x) = \sqrt{\frac{K}{2k_{\mathrm{B}}T\pi}} e^{-\frac{K(x-x_0)^2}{2k_{\mathrm{B}}T}}, \tag{A.8}$$

where $P(x)$ is the probability of the restrained DoF taking the value x, x_0 is the mean value of the DoF where no force is applied by the restraint, k_{B} is the Boltzmann constant, T is the temperature in Kelvin, and K is the force constant. Comparing this to the standard expression for a Normal distribution

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{A.9}$$

where $\mu = x_0$ and σ^2 is the variance, it can be seen that setting $K = \frac{k_{\mathrm{B}}T}{\sigma^2}$ makes Equations A.8 and A.9 equivalent.

If the distribution of the DoF about its mean in the fully interacting, non-restrained state is Gaussian, then measuring the variance of this distribution and setting $K = \frac{k_{\mathrm{B}}T}{\sigma^2}$ will mean that when the ligand is decoupled, the distribution of the DoF about its mean will be the same as in the fully interacting, non-restrained state. The force constants were fit in this way to minimise changes to the distributions of the restrained DoF during decoupling, which was intended to enhance convergence. Where only a limited number of DoF can be restrained, as with Boresch restraints, selecting the DoF with the minimum variance in the full-interacting, non-restrained state meant that K was as large as possible, hence the restraints were as restrictive as possible, without having to restrict the movement of an otherwise high-variance DoF, potentially causing convergence issues.

A.8 Convergence of Boresch Simulations with Force Constants Fit to Simulation

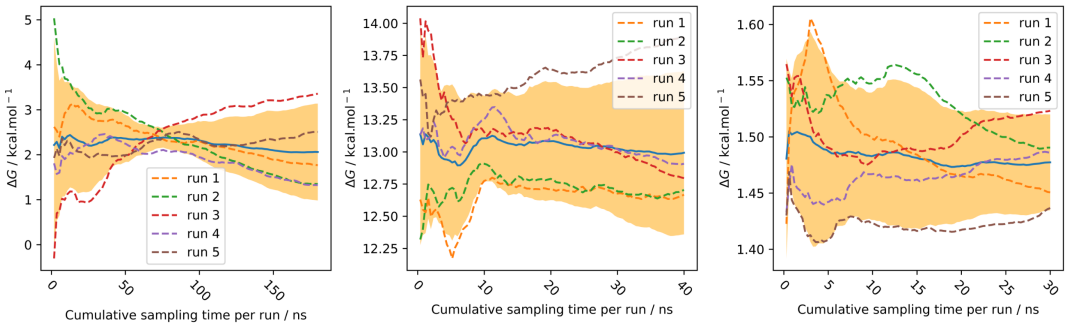


Figure A.4: Convergence of the bound leg simulations with B1 with cumulative sampling time per window. From left to right: the vanish, discharge, and restrain stages. Shaded area shows the 95% confidence interval and the solid blue line shows the mean.

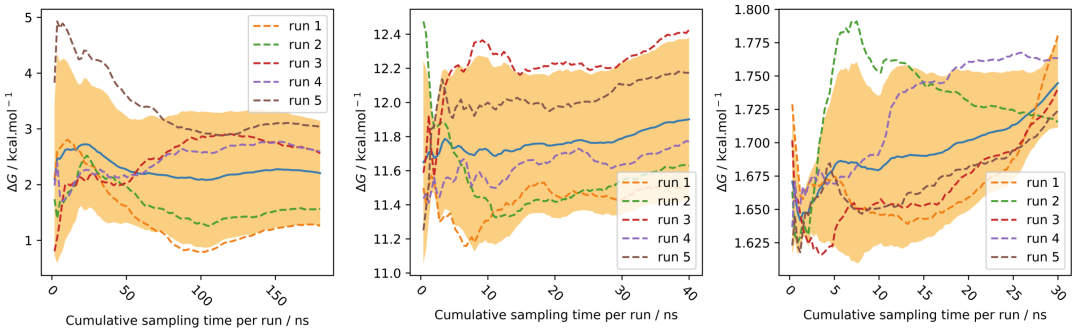


Figure A.5: Convergence of the bound leg simulations for B2 with cumulative sampling time per window. From left to right: the vanish, discharge, and restrain stages. Shaded area shows the 95% confidence interval and the solid blue line shows the mean.

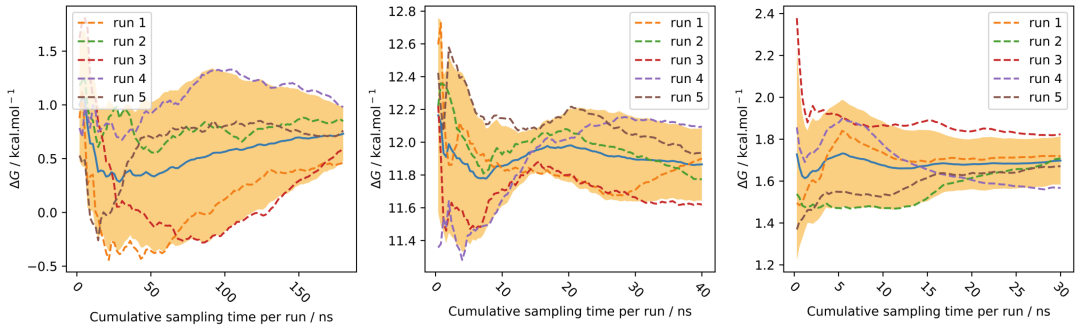


Figure A.6: Convergence of the bound leg simulations for B3 with cumulative sampling time per window. From left to right: the vanish, discharge, and restrain stages. Shaded area shows the 95% confidence interval and the solid blue line shows the mean.

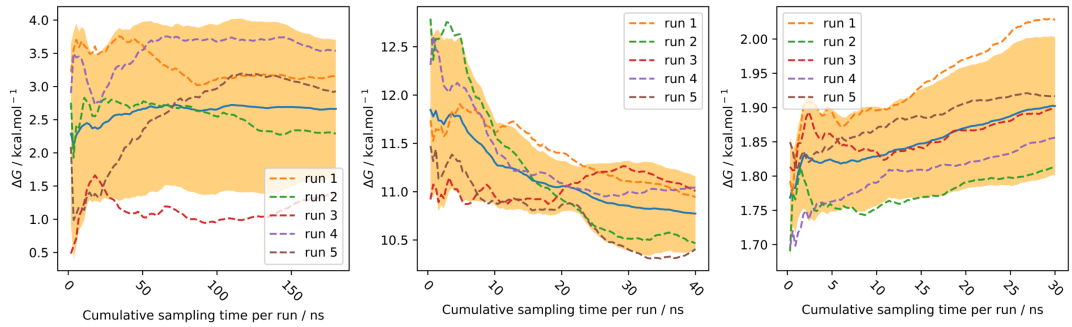


Figure A.7: Convergence of the bound leg simulations for B1-poseB with cumulative sampling time per window. From left to right: the vanish, discharge, and restrain stages. Shaded area shows the 95% confidence interval and the solid blue line shows the mean.

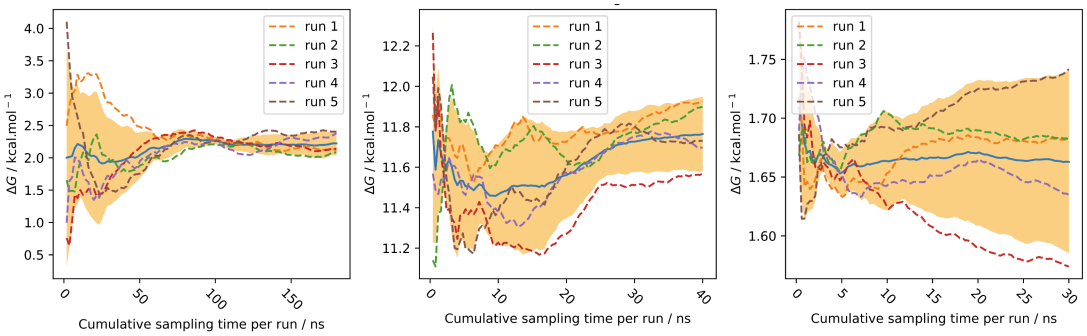


Figure A.8: Convergence of the bound leg simulations for B1-P with cumulative sampling time per window. From left to right: the vanish, discharge, and restrain stages. Shaded area shows the 95% confidence interval and the solid blue line shows the mean.

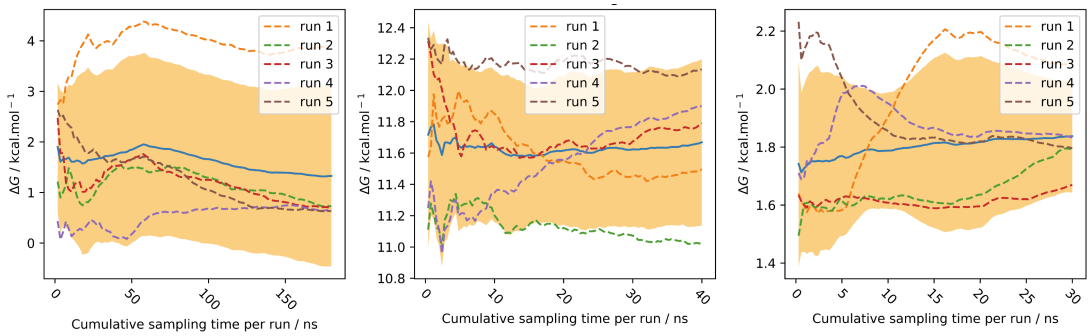


Figure A.9: Convergence of the bound leg simulations for B3-P with cumulative sampling time per window. From left to right: the vanish, discharge, and restrain stages. Shaded area shows the 95% confidence interval and the solid blue line shows the mean.

A.9 Bound Vanish PMF and Number of Waters in the Binding Site for B3-P

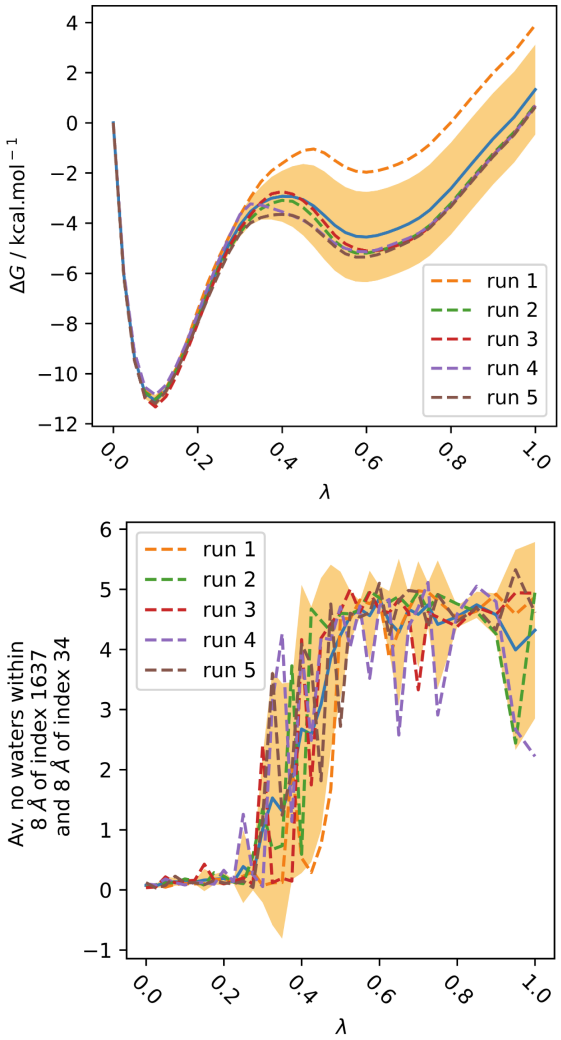


Figure A.10: Left - PMF along λ for bound vanish stage for B3-P. Right - Average number of waters in the binding site (as defined as the overlap of two spheres of radius 8 Å centred on the N atom in Pro1A, and CG2 in Val106A - with indices 1637 and 34) against λ during the vanish stage for B3-P. The shaded area shows the 95% confidence interval, and the solid blue line shows the mean. The under-sampling of waters in the binding site over three λ windows near $\lambda = 0.4$ results in the bound vanish PMF for run 1 divering by over 3 kcal mol^{-1}.

A.10 Convergence of Free Leg Simulations

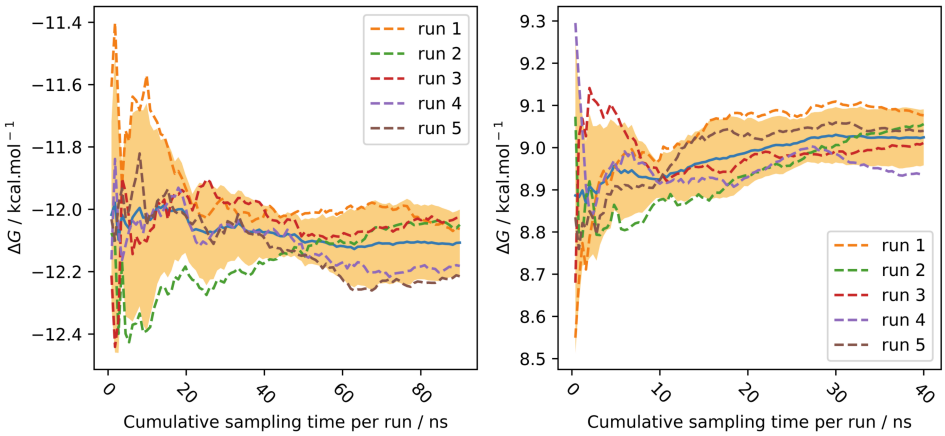


Figure A.11: Convergence of the free leg simulations with cumulative sampling time per window. Left: the vanish stage, right: the discharge stage. Shaded area shows the 95% confidence interval.

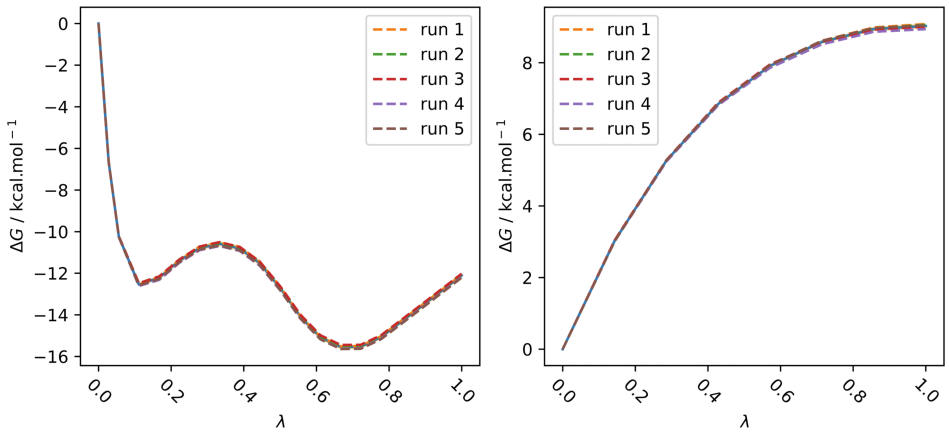


Figure A.12: Potentials of mean force along lambda for the free leg simulations.

A.11 Convergence of Boresch Simulations with No Orientational Component

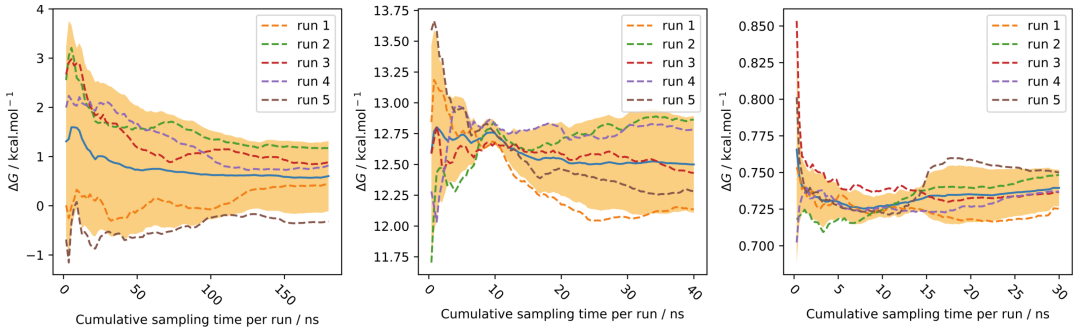


Figure A.13: Convergence of the bound leg simulations with B1-o with cumulative sampling time per window. From left to right: the vanish, discharge, and restrain stages. Shaded area shows the 95% confidence interval and the solid blue line shows the mean.

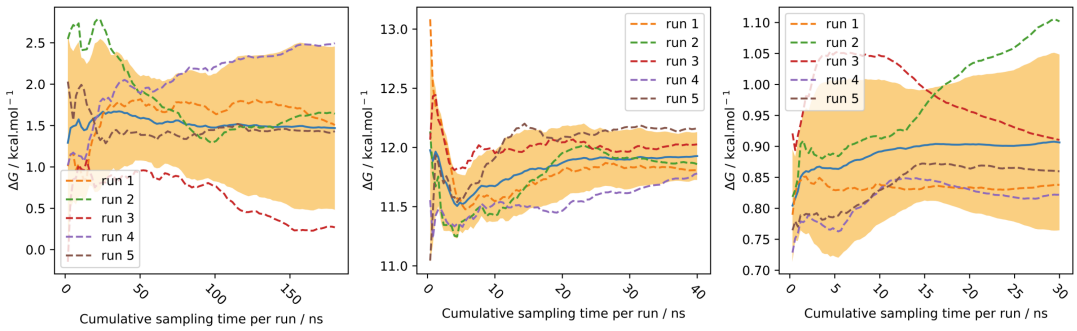


Figure A.14: Convergence of the bound leg simulations with B2-o with cumulative sampling time per window. From left to right: the vanish, discharge, and restrain stages. Shaded area shows the 95% confidence interval and the solid blue line shows the mean.

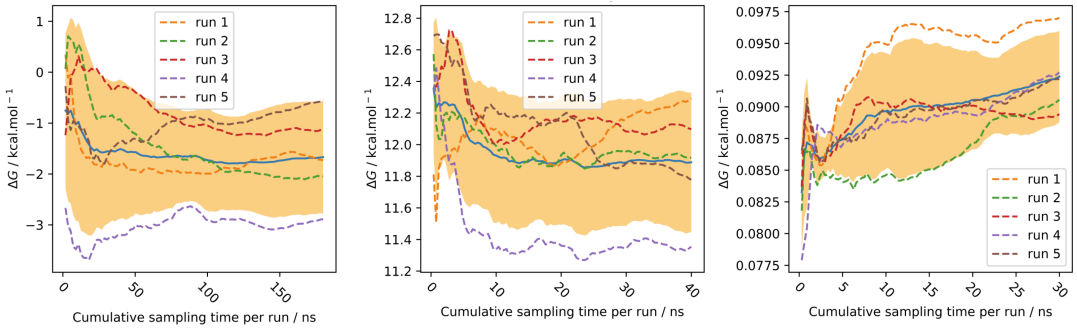


Figure A.15: Convergence of the bound leg simulations with B1-d with cumulative sampling time per window. From left to right: the vanish, discharge, and restrain stages. Shaded area shows the 95% confidence interval and the solid blue line shows the mean.

A.12 Number of Waters in the Binding Site for B2-o

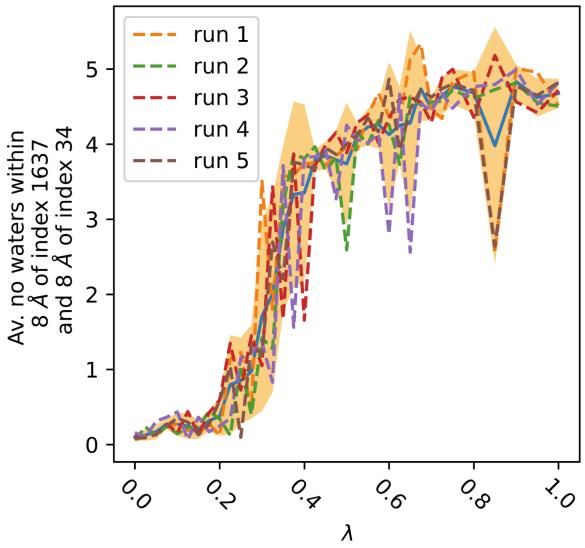


Figure A.16: Right - Average number of waters in the binding site (as defined as the overlap of two spheres of radius 8 Å centred on atoms of index 1637 and 34) against λ during the vanish stage for B2-o.

A.13 Comparison of Results without Orientational Restraints to the Literature

In this work, it was found that a lack of orientational restraints resulted in an erroneously negative free energy of binding due to failure to sample all relevant orientations during decoupling. Mobley et al. proposed a similar explanation,[168] but found that the offset was towards erroneously positive free energies of binding, which is surprising given the results observed here. However, this may have been due to the incorrect calculation of $-\Delta G^{o}_{\mathrm{Release}}$ in this preceding study.

A value of -5.26 kcal mol^{-1} was calculated for $\Delta G^{o}_{\mathrm{Release}}$ for phenol binding to a T4 lysozyme mutant when using a single distance restraint with a force constant of 10 kcal mol^{-1} Å^{-2} (see $\Delta G^{\mathrm{H_2O}}_{\mathrm{restr}}$ for "No orientational restraints" in Table 1 of Mobley et al.). The the free energy of releasing the restraint is calculated using

$$\Delta G^{o}_{\mathrm{Release}} = k_{\mathrm{B}}T \ln\left(\frac{1}{V^{o}} 4\pi \int_{0}^{r_{\mathrm{Box}}} r^{2} e^{-\frac{k(r-r_0)^2}{2RT}} \mathrm{d}r\right)$$

where the terms are as defined for Equation 2.11. If $r_0 = 0$, a value of -5.27 kcal mol^{-1} is obtained at 298 K, in excellent agreement with the result of Mobley et al.. However, this is incorrect, because $r_0 \neq 0$. r_0 is not stated by Mobley et al., so $r_0 = 3.48$ Å was taken from Table 1 of Boresch et al. for "CL1 FULL", where the same anchors in the protein are used (although the correction does not change dramatically if this distance is changed slightly). This value of r_0 yields a correction of - 1.71 kcal mol^{-1}, giving an overall free energy for the run without orientational restraints of 9.09±0.09 kcal mol^{-1} (subtracting the 0.41 kcal mol^{-1} symmetry correction which is not required when there are no orientational restraints), which is less than 1 kcal mol^{-1} different to the answer obtained with orientation decomposition (10.03 ± 0.05 kcal mol^{-1}). This value would be in even better agreement with a greater r_0. Therefore, it seems likely that in this case there were no large sampling issues caused by the lack of orientational restraints, and the observed deviation may have resulted from the erroneous use of $r_0 = 0$.

A.14 Restraint Dictionaries used for Multiple Distance Restraints Simulations

The format of the restraint dictionaries is {(anchor index 1, anchor index 2): (equilibrium distance, (force constant)/2, flat-bottomed radius), ...}.

A.14.1 M-Rig

Intermolecular restraints dictionary:

{(21, 4950): (2.72, 19.55, 0), (11, 4946): (5.92, 12.75, 0), (3, 4909): (3.63, 9.64, 0), (5, 4949): (6.87, 9.28, 0), (4, 971): (4.17, 7.42, 0), (2, 1613): (8.67, 7.07, 0), (14, 963): (5.56, 6.81, 0)}

Intramolecular restraints dictionary for the ligand:

{(21, 11): (1.364, 100, 0), (21, 3): (2.38159, 100, 0), (21, 5): (3.65882, 100, 0), (21, 4): (3.66517, 100, 0), (21, 2): (2.40022, 100, 0), (21, 14): (4.15453, 100, 0), (11, 3): (1.398, 100, 0), (11, 5): (2.42473, 100, 0), (11, 4): (2.41599, 100, 0), (11, 2): (1.398, 100, 0), (11, 14): (2.79524, 100, 0), (3, 5): (1.398, 100, 0), (3, 4): (2.78413, 100, 0), (3, 2): (2.41718, 100, 0), (3, 14): (2.41514, 100, 0), (5, 4): (2.41787, 100, 0), (5, 2): (2.79877, 100, 0), (5, 14): (1.398, 100, 0), (4, 2): (1.398, 100, 0), (4, 14): (1.398, 100, 0), (2, 14): (2.42338, 100, 0)}

Intramolecular restraints dictionary for the protein:

{(4950, 4946): (2.40799, 100, 0), (4950, 4909): (6.23833, 100, 0), (4950, 4949): (1.229, 100, 0), (4950, 971): (7.85489, 100, 0), (4950, 1613): (8.07314, 100, 0), (4950, 963): (9.03817, 100, 0), (4946, 4909): (7.67178, 100, 0), (4946, 4949): (1.522, 100, 0), (4946, 971): (9.71347, 100, 0), (4946, 1613): (8.0947, 100, 0), (4946, 963): (11.01378, 100, 0), (4909, 4949): (6.40042, 100, 0), (4909, 971): (9.2935, 100, 0), (4909, 1613): (11.23339, 100, 0), (4909, 963): (10.82613, 100, 0), (4949, 971): (8.9755, 100, 0), (4949, 1613): (8.49461, 100, 0), (4949, 963): (10.22414, 100, 0), (971, 1613): (7.35912, 100, 0), (971, 963): (2.53581, 100, 0), (1613, 963): (9.33324, 100, 0)}

M-All

Intermolecular restraints dictionary:

{(21, 4950): (2.72, 19.55, 0), (11, 4946): (5.92, 12.75, 0), (3, 4909): (3.63, 9.64, 0), (5, 4949): (6.87, 9.28, 0), (4, 971): (4.17, 7.42, 0), (2, 1613): (8.67, 7.07, 0), (14, 963): (5.56, 6.81, 0), (20, 959): (4.73, 6.07, 0), (10, 950): (6.68, 5.77, 0), (12, 51): (9.15, 5.64, 0), (18, 4951): (9.52, 5.31, 0), (13, 47): (6.0, 5.02, 0), (19, 49): (8.75, 4.94, 0), (6, 53): (9.8, 3.92, 0), (17, 34): (4.24, 3.64, 0), (16, 45): (7.8, 3.59, 0), (9, 548): (8.14, 2.81, 0), (15, 48): (9.26, 2.74, 0), (7, 584): (8.25, 1.63, 0), (8, 1633): (10.93, 1.6, 0), (0, 1737): (6.64, 1.28, 0), (1, 4914): (10.4, 1.13, 0)}

M-hand

Intermolecular restraints dictionary was:

{(21, 4950): (2.72, 20, 0.61), (18, 961): (3.09, 20, 0.61), (17, 512): (3.25, 20, 2.06), (19, 512): (3.69, 20, 2.13)}

A.15 Convergence of Restraint Correction for Multiple Distance Restraints Schemes

It was confirmed that estimates of $\Delta G^o_{Release}$ had converged with respect to the number of integration points. It was found that evaluating the restraint energy at a relatively large number of orientations was essential to obtain an accurate estimate. Examples are shown below. 50 frames were saved per ns of simulation.

Table A.2: Convergence of $\Delta G^o_{\mathrm{Release}}$ for M-Rig with increasing number of grid points used for numerical integration. s is the number of frames to skip between two snapshot evaluations, b is the amount by which the bounding rectangle of the restrained host atoms coordinates is extended in each dimension, d is the edge length of a translational volume element, and o is the number of orientations at which the restraint energy is to evaluated per Euler angle interval ([0,2π] for ϕ and ψ, [0, π] for θ). Insensitivity to b and d was confirmed in preliminary simulations. Uncertainties are the 95 % C.I.s obtained from the variance between 5 replicate runs by assuming Gaussian distributions.

s	**b** / Å	**d** / Å	**o**	$\Delta G^o_{\mathbf{Release}}$
1	4	0.25	6	-13.28 ± 4.16
1	4	0.25	12	-10.19 ± 0.28
1	4	0.10	12	-10.19 ± 0.28
1	4	0.25	18	-10.08 ± 0.05
1	4	0.25	24	-10.03 ± 0.07
1	4	0.25	30	-10.02 ± 0.04

Table A.3: Convergence of $\Delta G^o_{\text{Release}}$ for M-Hand-1 with increasing number of grid points used for numerical integration. s is the number of frames to skip between two snapshot evaluations, b is the amount by which the bounding rectangle of the restrained host atoms coordinates is extended in each dimension, d is the edge length of a translational volume element, and o is the number of orientations at which the restraint energy is to evaluated per Euler angle interval ([0,2π] for ϕ and ψ, [0, π] for θ). Insensitivity to b and d was confirmed in preliminary simulations. Uncertainties are the 95 % C.I.s obtained from the variance between 5 replicate runs by assuming Gaussian distributions. Note that the final result shown below is slightly different to that shown in the main text, because these checks were performed before a bug affecting flat-bottomed restraints was fixed and all other affected simulations were repeated.

s	**b** / Å	**d** / Å	**o**	$\Delta G^o_{\text{Release}}$
1	4	0.25	6	-11.41 ± 8.89
1	4	0.25	12	-5.83 ± 0.24
1	4	0.10	12	-5.83 ± 0.24
1	4	0.25	18	-5.74 ± 0.14
1	4	0.25	24	-5.73 ± 0.15
1	4	0.25	30	-5.72 ± 0.15

Table A.4: Convergence of $\Delta G^o_{\text{Release}}$ for M-All with increasing number of grid points used for numerical integration. s is the number of frames to skip between two snapshot evaluations, b is the amount by which the bounding rectangle of the restrained host atoms coordinates is extended in each dimension, d is the edge length of a translational volume element, and o is the number of orientations at which the restraint energy is to evaluated per Euler angle interval ([0,2π) for ϕ and ψ, [0, π] for θ). Insensitivity to b and d was confirmed in preliminary simulations. Uncertainties are the 95 % C.I.s obtained from the variance between 5 replicate runs by assuming Gaussian distributions.

s	**b** / Å	**d** / Å	**o**	$\Delta G^o_{\text{Release}}$
1	4	0.25	6	-61.55 ± 12.12
1	4	0.25	18	-13.68 ± 9.60
1	4	0.25	24	-16.32 ± 0.92
1	4	0.25	30	-15.68 ± 0.37

A.16 Convergence of Multiple Distance Restraint Simulations

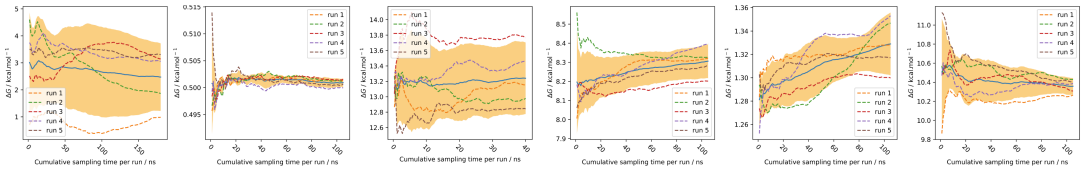


Figure A.17: Convergence of the bound leg simulations for M-Rig with cumulative sampling time per window. From left to right: the vanish, rigidify lig, discharge, rigidify complex, restrain, and rigidify recept. stages. Shaded area shows the 95% confidence interval and the solid blue line shows the mean.

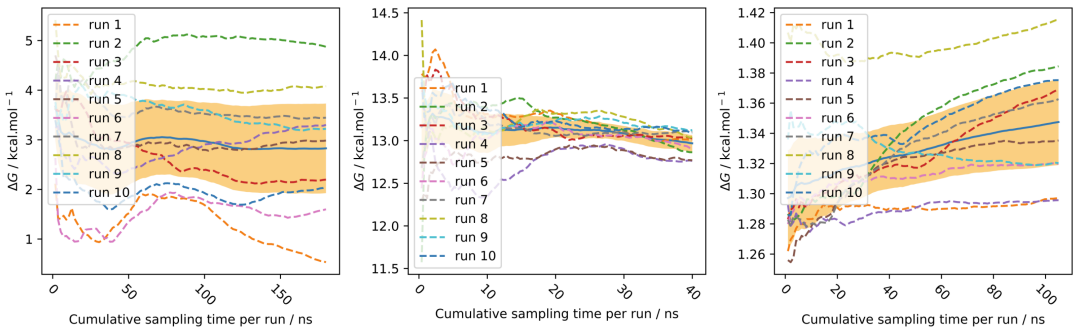


Figure A.18: Convergence of the bound leg simulations for M-Rig-N with cumulative sampling time per window. From left to right: the vanish, discharge, and restrain stages. Shaded area shows the 95% confidence interval and the solid blue line shows the mean.

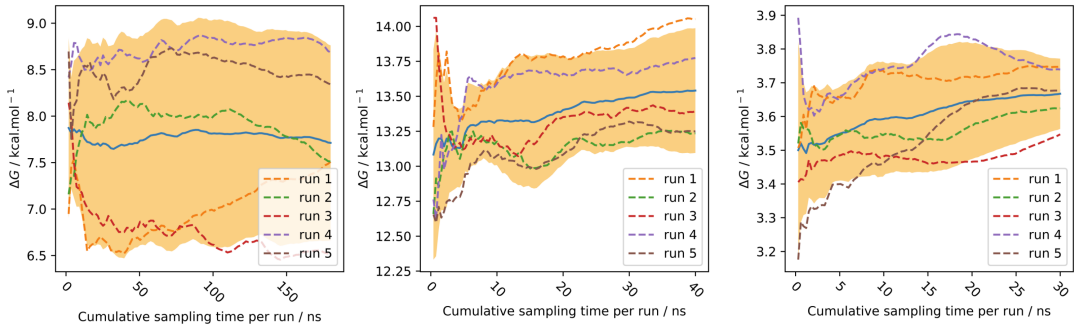


Figure A.19: Convergence of the bound leg simulations for M-All with cumulative sampling time per window. From left to right: the vanish, discharge, and restrain stages. Shaded area shows the 95% confidence interval and the solid blue line shows the mean.

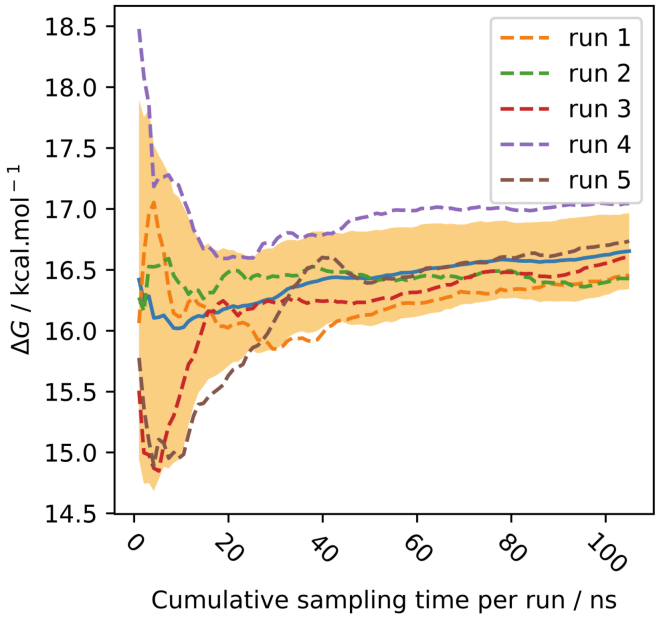


Figure A.20: Convergence of the free energy of releasing all distance restraints other than the single strongest instance for M-All-R, with cumulative sampling time per window. Shaded area shows the 95% confidence interval and the solid blue line shows the mean.

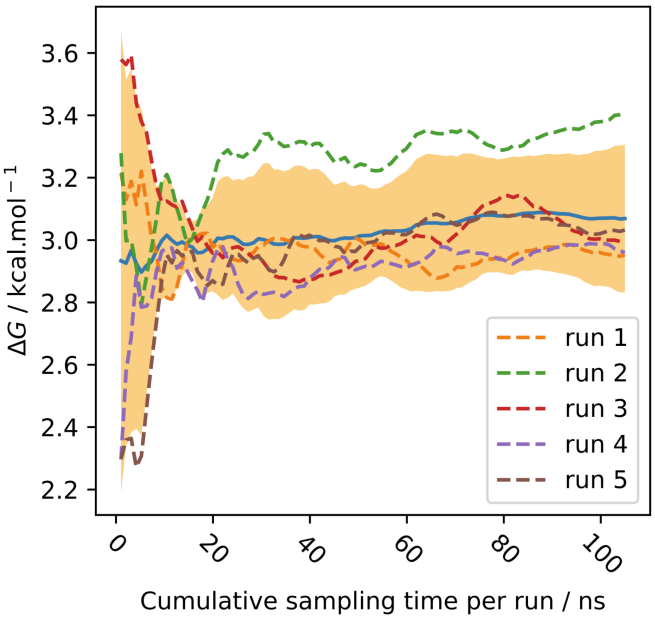


Figure A.21: Convergence of the free energy of releasing all distance restraints other than the single strongest instance for M-Hand-R, with cumulative sampling time per window. Shaded area shows the 95% confidence interval and the solid blue line shows the mean.

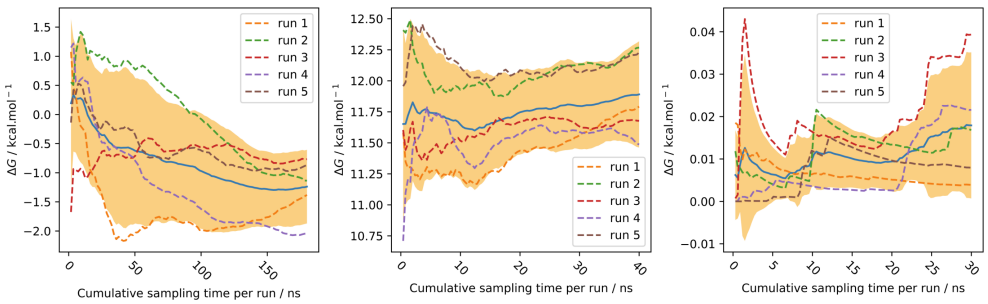


Figure A.22: Convergence of the bound leg simulations for M-Hand with cumulative sampling time per window. From left to right: the vanish, discharge, and restrain stages. Shaded area shows the 95% confidence interval and the solid blue line shows the mean.

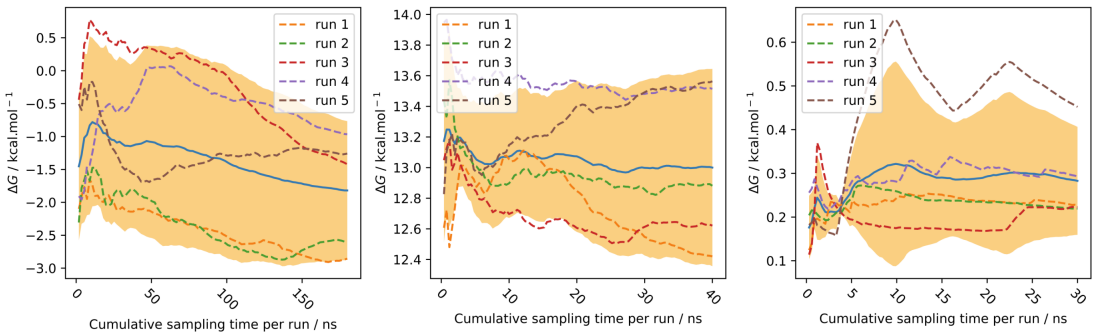


Figure A.23: Convergence of the bound leg simulations for M-Hand-1 with cumulative sampling time per window. From left to right: the vanish, discharge, and restrain stages. Shaded area shows the 95% confidence interval and the solid blue line shows the mean.

A.17 Overlap Matrices for Selected Multiple Distance Restraint Protocols

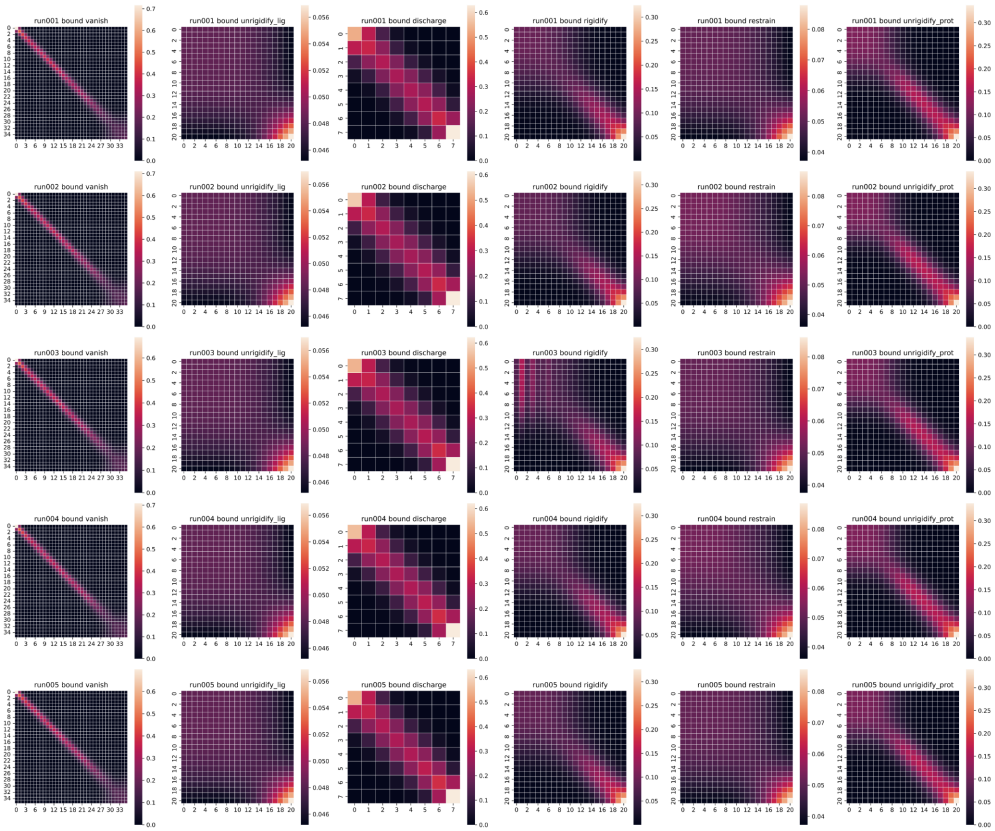


Figure A.24: Overlap matrices for M-Rig

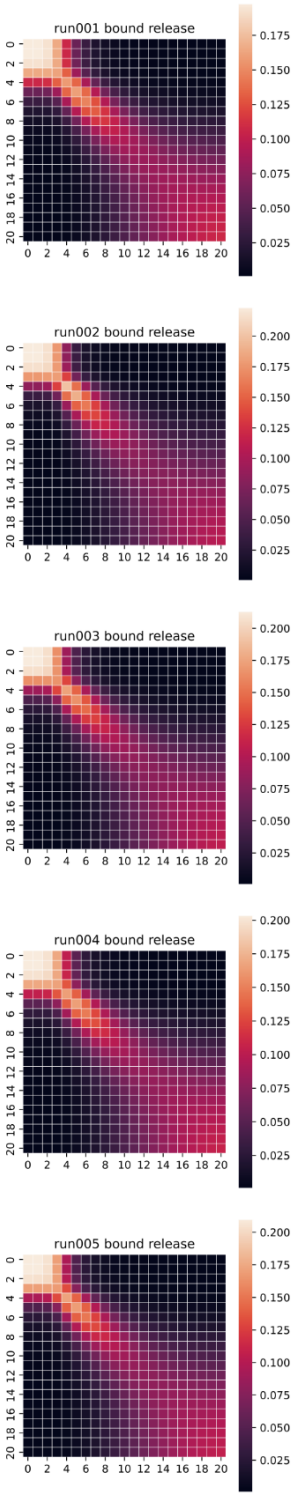


Figure A.25: Overlap matrices for M-Hand-R

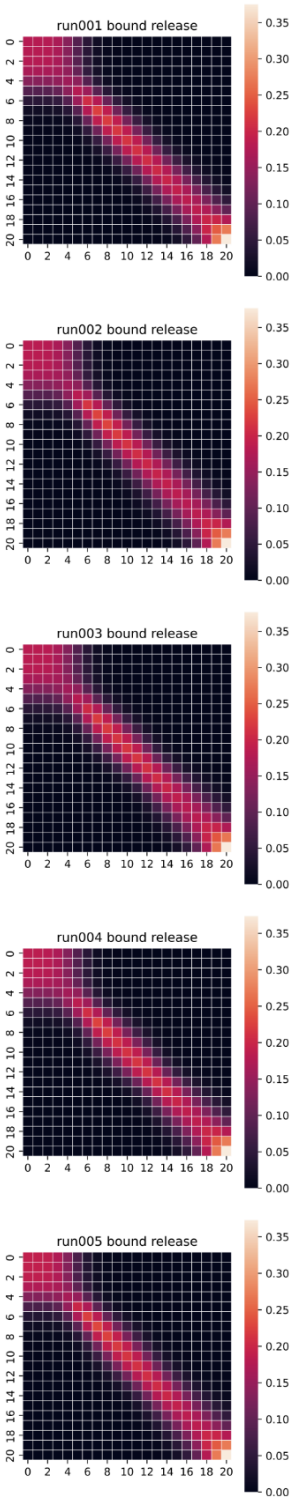


Figure A.26: Overlap matrices for M-All-R

A.18 Convergence of Free Energy of Preorganisation with Increasing Restraint Strength

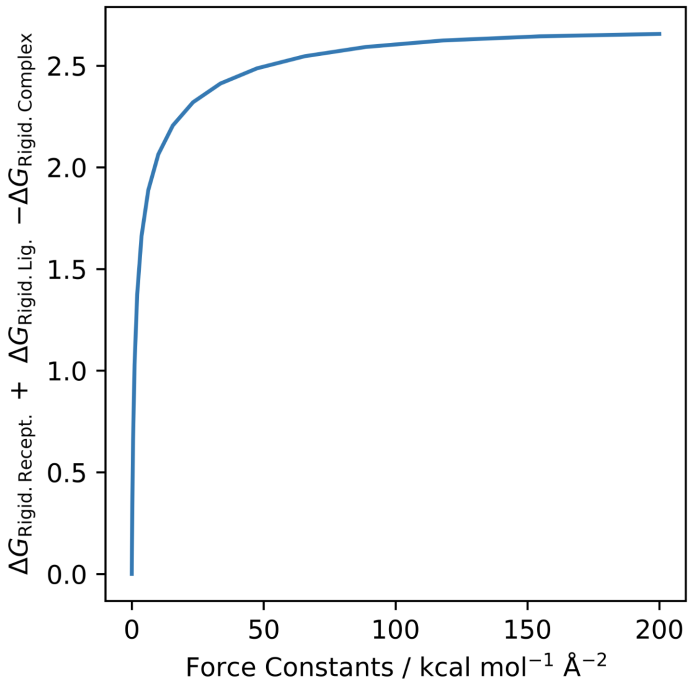


Figure A.27: Convergence of $\Delta G_{\text{Rigid. Recept.}} + \Delta G_{\text{Rigid. Lig.}} - \Delta G_{\text{Rigid. Complex}} \approx \Delta G_{\text{Preorg.}}$ for M-Rig with respect to increasing strength of the intramolecular restraints, where the maximum strength of restraints was 200 kcal mol^{-1}. This is approximate because there are intermolecular interactions between the ligand and receptor when the restraints are introduced - this error could be eliminated by introduced the intramolecular restraints when the ligand is decoupled. A restraint strength of $>$ 50 kcal mol^{-1} Å^{-2} would have been required to obtain result which was approximately converged with respect to increasing strength of the intramolecular restraints. For M-Rig, the intramolecular restraints of strength 75 kcal mol^{-1} produced $\Delta G_{\text{Rigid. Recept.}} + \Delta G_{\text{Rigid. Lig.}} - \Delta G_{\text{Rigid. Complex}} = 2.55$ kcal mol^{-1}, which was very similar to the result with 200 kcal mol^{-1} restraints (2.66 kcal mol^{-1}).

Appendix B

Automated Adaptive Absolute Binding Free Energy Calculations

B.1 Uncertainties of TI, Zwanzig, and BAR in the Limit of Infinitely Close States

Under the assumption of infinitely closely-spaced intermediate states, the uncertainties of the TI, Zwazig, and BAR estimates become equivalent. This discussion is similar to that given by Nguyen and Minh.[291]

B.1.1 TI

In the theoretical limit of infinitely many states, the free energy change can be calculated with thermodynamic integration according to

$$\widehat{\Delta F} = \int_0^1 \left\langle \frac{\partial H}{\partial \lambda} \right\rangle_\lambda \mathrm{d}\lambda, \tag{B.1}$$

where $\widehat{\Delta F}$ denotes the estimated free energy change and $\langle ... \rangle_\lambda$ denotes an average obtained at a fixed value of λ.

Assuming independent sampling at different values of λ, the total variance can be obtained by integrating the uncertainties at each value of λ

$$\sigma^2(\widehat{\Delta F}) = \int_0^1 \sigma^2 \left(\left\langle \frac{\partial H}{\partial \lambda} \right\rangle_\lambda \right) \mathrm{d}\lambda. \tag{B.2}$$

B.1.2 Zwanzig

The free energy difference between states 0 and 1 can be calculated using samples from state 0 using the Zwanzig equation

$$\widehat{\Delta F} = -\beta^{-1} \ln \left\langle e^{-\beta \Delta u(\boldsymbol{x})} \right\rangle_0 , \tag{B.3}$$

where $\langle ... \rangle_0$ denotes an average over samples obtained at state 0, $\beta = \frac{1}{k_\mathrm{B} T}$, and $\boldsymbol{x}$ is a point in configuration space. When the states are close enough in λ such that $\beta \Delta u(\boldsymbol{x}) << 1$ can be assumed, the exponential term in Equation B.3 can be approximated using a first-order Taylor expansion

$$\widehat{\Delta F} \approx -\beta^{-1} \ln \langle 1 - \beta \Delta u(\boldsymbol{x}) \rangle_0 \tag{B.4}$$

$$\approx -\beta^{-1} \ln(1 - \langle \beta \Delta u(\boldsymbol{x}) \rangle_0) . \tag{B.5}$$

Taking another Taylor expansion yields

$$\widehat{\Delta F} \approx \langle \Delta u(\boldsymbol{x}) \rangle_0 . \tag{B.6}$$

In the case of several intermediate states, Equation B.6 could be applied repeatedly to obtain the overall free energy change

$$\widehat{\Delta F} \approx \sum_{k=1}^{K-1} \langle \Delta u_{k+1,k}(\boldsymbol{x}) \rangle_k , \tag{B.7}$$

where k is the state number and there are K states. In the limit of infinitely close states this becomes equivalent to Equation B.1 and hence the uncertainty is also given by Equation B.2.

B.1.3 BAR

The Bennett Acceptance Ratio (BAR) can be understood as reweighting from the mixture distribution, π_Mix, constructed with samples from two states.[121,362]

$$\widehat{F_1} = -\beta^{-1} \ln \left\langle \frac{e^{-\beta u_1(\boldsymbol{x})}}{\widehat{\pi}_\mathrm{Mix}(\boldsymbol{x})} \right\rangle_{\pi_\mathrm{Mix}} , \tag{B.8}$$

$$\widehat{\pi}_\mathrm{Mix}(\boldsymbol{x}) = \frac{N_0}{N_0 + N_1} e^{-\beta u_0(\boldsymbol{x}) + \beta \widehat{F_0}} + \frac{N_1}{N_0 + N_1} e^{-\beta u_1(\boldsymbol{x}) + \beta \widehat{F_1}}, \tag{B.9}$$

where N_k is the total number of samples from state k. Setting $\widehat{F_0} = 0$ and assuming $N_0 = N_1 = N$

$$\widehat{\Delta F} = -\beta^{-1} \ln \left\langle \frac{e^{-\beta u_1(\boldsymbol{x})}}{\frac{1}{2}(e^{-\beta u_0(\boldsymbol{x})} + e^{-\beta u_1(\boldsymbol{x}) + \beta \widehat{\Delta F}})} \right\rangle_{\pi_{\text{Mix}}} . \tag{B.10}$$

Dividing the top and bottom of the fraction by $e^{-\beta u_0(\boldsymbol{x})}$ yields

$$\widehat{\Delta F} = -\beta^{-1} \ln \left\langle \frac{e^{-\beta \Delta u(\boldsymbol{x})}}{\frac{1}{2}(1 + e^{-\beta \Delta u(\boldsymbol{x}) + \beta \widehat{\Delta F}})} \right\rangle_{\pi_{\text{Mix}}} . \tag{B.11}$$

As the states become infinitely close, $\beta \Delta u(\boldsymbol{x})$ and $\beta \Delta \widehat{F}$ become much less than 1. This justifies the use of first-order Taylor expansions of the exponentials

$$\widehat{\Delta F} = -\beta^{-1} \ln \left\langle \frac{1 - \beta \Delta u(\boldsymbol{x})}{1 - \frac{1}{2}(\beta \Delta u(\boldsymbol{x}) - \beta \widehat{\Delta F})} \right\rangle_{\pi_{\text{Mix}}} . \tag{B.12}$$

Using the fact that $\frac{1}{1-x} = 1 + x$ for $x << 1$ (where x is $\frac{1}{2}(\beta \Delta u(\boldsymbol{x}) - \beta \widehat{\Delta F})$)

$$\widehat{\Delta F} = -\beta^{-1} \ln \left\langle (1 - \beta \Delta u(\boldsymbol{x})) \left(1 + \frac{1}{2}(\beta \Delta u(\boldsymbol{x}) - \beta \widehat{\Delta F} \right) \right\rangle_{\pi_{\text{Mix}}} \tag{B.13}$$

$$= -\beta^{-1} \ln \left\langle 1 - \frac{1}{2} \beta \Delta u(\boldsymbol{x}) - \frac{1}{2} \beta \widehat{\Delta F} \right\rangle_{\pi_{\text{Mix}}} , \tag{B.14}$$

where we have ignored the cross terms (such as $(\beta \Delta u(\boldsymbol{x}))^2$), which are negligible. As the two states become infinitely close in λ, the weights of each state in π_{Mix} become equal, and $\langle ... \rangle_{\pi_{\text{Mix}}}$ becomes the arithmetic mean over the two states. Furthermore, the configurations sampled in each state become indistinguishable, meaning that $\langle ... \rangle_{\pi_{\text{Mix}}} \approx \langle ... \rangle_0 \approx \langle ... \rangle_1$. Arbitrarily choosing state 0, and taking the first-order Taylor expansion of the logarithm yields

$$\widehat{\Delta F} \approx \langle \Delta u(\boldsymbol{x}) \rangle_0 , \tag{B.15}$$

which is the same as Equation B.6., Hence, the uncertainty is given by Equation B.2, as before.

B.2 The Relationship Between the Variance of the Gradient and Overlap

The variance of the gradient of the free energy with respect to λ, sampled at some value of λ, is

$$\mathrm{Var}\left(\frac{\partial H}{\partial \lambda}\right) = \left\langle \left(\frac{\partial H}{\partial \lambda}\right)^2 \right\rangle_\lambda - \left\langle \frac{\partial H}{\partial \lambda} \right\rangle_\lambda^2 \tag{B.16}$$

where $H(\boldsymbol{x};\lambda)$ is the Hamiltonian which depends on the sampled point in configuration space, $\boldsymbol{x}$, and the coupling parameter, λ. $\langle ... \rangle_\lambda$ denotes an average at a given value of λ.

As discussed by Blondel,[241] offsetting all energies at a given value of λ by a constant does not affect the variance. For simplicity, the offset is chosen to be $F(\lambda)$, the free energy at a given value of λ, and the new Hamiltonians are denoted by H'. Hence

$$\mathrm{Var}\left(\frac{\partial H}{\partial \lambda}\right) = \left\langle \left(\frac{\partial H'}{\partial \lambda}\right)^2 \right\rangle_\lambda - \left\langle \frac{\partial H'}{\partial \lambda} \right\rangle_\lambda^2 \tag{B.17}$$

We also note that

$$e^{-\beta H'(\boldsymbol{x},\lambda)} = e^{-\beta H(\boldsymbol{x},\lambda)+\beta F(\lambda)} = p(\boldsymbol{x},\lambda) \tag{B.18}$$

where $p(\boldsymbol{x},\lambda)$ is the normalised probability of sampling phase-space point $\boldsymbol{x}$ at a given value of λ.

Because all λ states have been offset by their free energies, the free energy changes between the new states is 0, and therefore $\left\langle \frac{\mathrm{d}H'}{\mathrm{d}\lambda} \right\rangle_\lambda = 0$, simplifying Equation B.18

$$\mathrm{Var}\left(\frac{\partial H}{\partial \lambda}\right) = \left\langle \left(\frac{\partial H'}{\partial \lambda}\right)^2 \right\rangle_\lambda \tag{B.19}$$

Noting that

$$\frac{\partial H'}{\partial \lambda} = \frac{\frac{\partial H'}{\partial \lambda} e^{-\beta H'}}{e^{-\beta H'}} = -\frac{1}{\beta}\frac{\partial \log p(\mathbf{x},\lambda)}{\partial \lambda} \tag{B.20}$$

Equation B.19 can be rewritten:

$$\beta^2 \mathrm{Var}\left(\frac{\partial H}{\partial \lambda}\right) = \int \left(\frac{\partial \ln p(\mathbf{x},\lambda)}{\partial \lambda}\right)^2 p(\mathbf{x},\lambda)\mathrm{d}\mathbf{x} \tag{B.21}$$

This tells us that the dimensionless variance of the gradient is the Fischer information. Rewriting this equation

$$\beta^2 \mathrm{Var}\left(\frac{\partial H}{\partial \lambda}\right) = \left\langle \left(\frac{1}{p(\mathbf{x},\lambda)}\frac{\partial p(\mathbf{x},\lambda)}{\partial \lambda}\right)^2 \right\rangle_\lambda \tag{B.22}$$

shows that the dimensionless variance of the gradient is the expected value of the squared relative probability change when λ is varied.

B.3 Details of Structure Preparation for Initial Test Systems

Where details of structure preparation are omitted below, the steps described in Section 3.1 were followed.

B.3.1 T4L

The complex of benzene with the L99A mutant of T4 lysozyme was prepared from PDB ID 4W52 as described in Section 3.1.[363]

B.3.2 MIF

The complex of human macrophage migration inhibitory factor (MIF) with the ligand MIF180 was prepared as described by Clark et al..[274] The only differences were the use of Open Force Field 2.0.0 for the small molecule, rather than GAFF2.11, and the use of a rhombic dodecahedral box, rather than a cubic box.

B.3.3 MDM2-Pip2

For the complex of MDM2 (with truncated lid - "Short") with Pip2, AMBER prm7 and rst7 files were obtained from Mendoza-Martinez et al..[165] The only modifications made were the reparameterisation of the ligand with Open Force Field 2.0.0 with AM1-BCC partial charges (rather than the original GAFF), and the resolvation using a rhombic dodecahedral box, as described in Section 3.1. The experimental binding affinity was taken from the "17-125" column for the Pip-2 ITC data in Table 1 of Michelsen et al..[283]

B.3.4 PDE2a

A PDB structure for Phosphodiesterase 2a in complex with ligand "P10" was obtained from Huggins[171], specifically from https://github.com/djhuggins/Holoware-TestCases/blob/main/PDE2/ff14_tip3p_am1bcc/P10/input/complex.pdb. Due to all water molecules having incorrect angles for TIP3P, water hydrogens were removed and reintroduced using tleap from ambertools 22.0. Otherwise, preparation was performed as described in Section 3.1.

B.3.5 MDM2-Nutlin

The complex of mouse double minute 2 homolog (MDM2) (with truncated lid - "Short") with Pip2 was prepared by aligning PDB ID 4WT2 to PDB ID 4HG7. The protein structure was taken from 4WT2 from residue 17 (S) to the end (110, V), as this most closely matched the experimental construct. The ligand and crystallographic waters were taken from 4HG7, and subsequent preparation was performed as described in Section 3.1.

B.4 Detailed Results for Non-Adaptive ABFE Calculations on Initial Test Systems

Table B.1: Components of Non-Adaptive $\Delta G^{o}_{\mathrm{Bind}}$ for Initial Test Systems[a]

	Bound Restrain	Bound Discharge	Bound Vanish	Free Discharge	Free Vanish	Restraint Correction	Symmetry Correction	Exp. $\Delta G^{o}_{\mathrm{Bind}}$
T4L 0.2 ns	2.03 ± 0.15	0.27 ± 0.13	4.04 ± 0.73	2.07 ± 0.03	-7.14 ± 0.19	-7.08	-0.41	-5.19 ± 0.16
T4L 6 ns	1.97 ± 0.06	0.21 ± 0.05	3.10 ± 0.96	2.05 ± 0.01	-7.25 ± 0.05	-7.08	-0.41	-5.19 ± 0.16
T4L 30 ns	1.96 ± 0.02	0.18 ± 0.02	3.38 ± 0.81	2.05 ± 0.00	-7.24 ± 0.02	-7.08	-0.41	-5.19 ± 0.16
MIF 0.2 ns	1.70 ± 0.21	19.12 ± 2.56	2.21 ± 1.43	14.35 ± 0.13	-17.58 ± 0.27	-10.35	-0.65	-8.98 ± 0.28
MIF 6 ns	1.71 ± 0.07	18.04 ± 1.28	-3.51 ± 0.45	14.23 ± 0.02	-17.85 ± 0.05	-10.35	-0.65	-8.98 ± 0.28
MIF 30 ns	1.81 ± 0.10	17.68 ± 1.03	-2.98 ± 0.87	14.17 ± 0.11	-17.84 ± 0.03	-10.35	-0.65	-8.98 ± 0.28
MDM2-Nutlin 0.2 ns	2.64 ± 1.18	139.34 ± 1.25	-21.78 ± 4.02	140.59 ± 1.23	-48.12 ± 1.57	-9.85	0.00	-11.14 ± 0.27
MDM2-Nutlin 6 ns	2.33 ± 0.79	138.82 ± 1.16	-25.06 ± 1.88	140.17 ± 0.34	-49.49 ± 0.51	-9.85	0.00	-11.14 ± 0.27
MDM2-Nutlin 30 ns	2.33 ± 0.34	138.17 ± 0.89	-23.94 ± 1.06	140.16 ± 0.23	-50.39 ± 0.67	-9.85	0.00	-11.14 ± 0.27
MDM2-Pip2 0.2 ns	1.65 ± 0.09	54.28 ± 0.75	-6.88 ± 1.38	53.93 ± 0.36	-32.34 ± 1.26	-10.14	0.00	-9.11 ± 0.01
MDM2-Pip2 6 ns	1.67 ± 0.03	52.82 ± 0.38	-11.12 ± 0.71	53.54 ± 0.12	-33.81 ± 0.69	-10.14	0.00	-9.11 ± 0.01
MDM2-Pip2 30 ns	1.76 ± 0.17	52.74 ± 0.39	-11.06 ± 1.05	53.46 ± 0.11	-34.10 ± 0.21	-10.14	0.00	-9.11 ± 0.01
PDE2A 0.2 ns	1.74 ± 0.17	214.78 ± 0.85	4.20 ± 2.70	214.23 ± 0.35	-33.15 ± 1.53	-10.24	0.00	-14.35 ± 0.50
PDE2A 6 ns	1.84 ± 0.17	214.54 ± 0.54	-9.55 ± 1.70	213.49 ± 0.18	-34.16 ± 0.56	-10.24	0.00	-14.35 ± 0.50
PDE2A 30 ns	1.94 ± 0.14	214.22 ± 0.62	-9.53 ± 1.81	213.40 ± 0.12	-33.97 ± 0.34	-10.24	0.00	-14.35 ± 0.50

[a] All quantities in kcal mol^{-1}. Uncertainties stated as 95 % confidence intervals based on the variance of 5 replicate runs, assuming Gaussian distributions.

B.5 Restraints Parameters and Discussion of Symmetry Corrections

Table B.2: Parameters for Boresch restraints for initial test systems, as labelled in Figure 3 of Clark et al..[274] K refers to a force constant and 0 denotes an equilibrium value.

	T4L	MIF	MDM2-Pip2	PDE2A	MDM2-Nutlin
r1	1550	952	1307	4395	1330
r2	1530	950	1295	4393	1318
r3	1552	959	1309	4408	1332
l1	4	10	19	11	28
l2	3	13	14	10	6
l3	5	20	20	12	21
r_0 / Å	7.69	5.66	5.92	6.59	7.05
θ_{A0} / Å	1.30	2.14	1.80	1.71	1.32
θ_{B0} / Rad	1.48	1.48	1.28	1.27	1.43
ϕ_{A0} / Rad	2.56	1.84	-2.57	2.35	-2.78
ϕ_{B0} / Rad	2.94	3.09	1.05	0.00	1.21
ϕ_{C0} / Rad	1.41	0.22	-0.96	0.21	0.91
k_r / kcal mol^{-1} Å^{-2}	6.20	18.00	8.06	12.42	7.82
$k_{\theta A}$ / kcal mol^{-1} Rad^{-2}	28.76	87.34	92.70	76.28	62.12
$k_{\theta B}$ / kcal mol^{-1} Rad^{-2}	24.82	89.72	88.14	85.28	141.90
$k_{\phi A}$ / kcal mol^{-1} Rad^{-2}	59.86	101.06	76.44	228.30	53.84
$k_{\phi B}$ / kcal mol^{-1} Rad^{-2}	0.80	115.14	161.32	139.72	132.66
$k_{\phi C}$ / kcal mol^{-1} Rad^{-2}	55.18	99.32	149.54	101.36	199.86

The symmetry corrections shown in Table B.1 were selected as follows: for T4L, benzene was observed to rotate freely about its 6-fold axis of symmetry at $\lambda = 0$ during the restraining simulations, but was not observed to flip over. Therefore, we included a correction of $-k_B T \ln 2$. For MIF, we included a correction of $-k_B T \ln 3$ to account for the three-fold symmetry of the protein. No symmetry corrections were required for MDM2-Pip2, MDM2-Nutlin, or PDE2a.

B.6 Estimated Free Energy Changes Against Sampling Times for Initial Non-Adaptive Runs

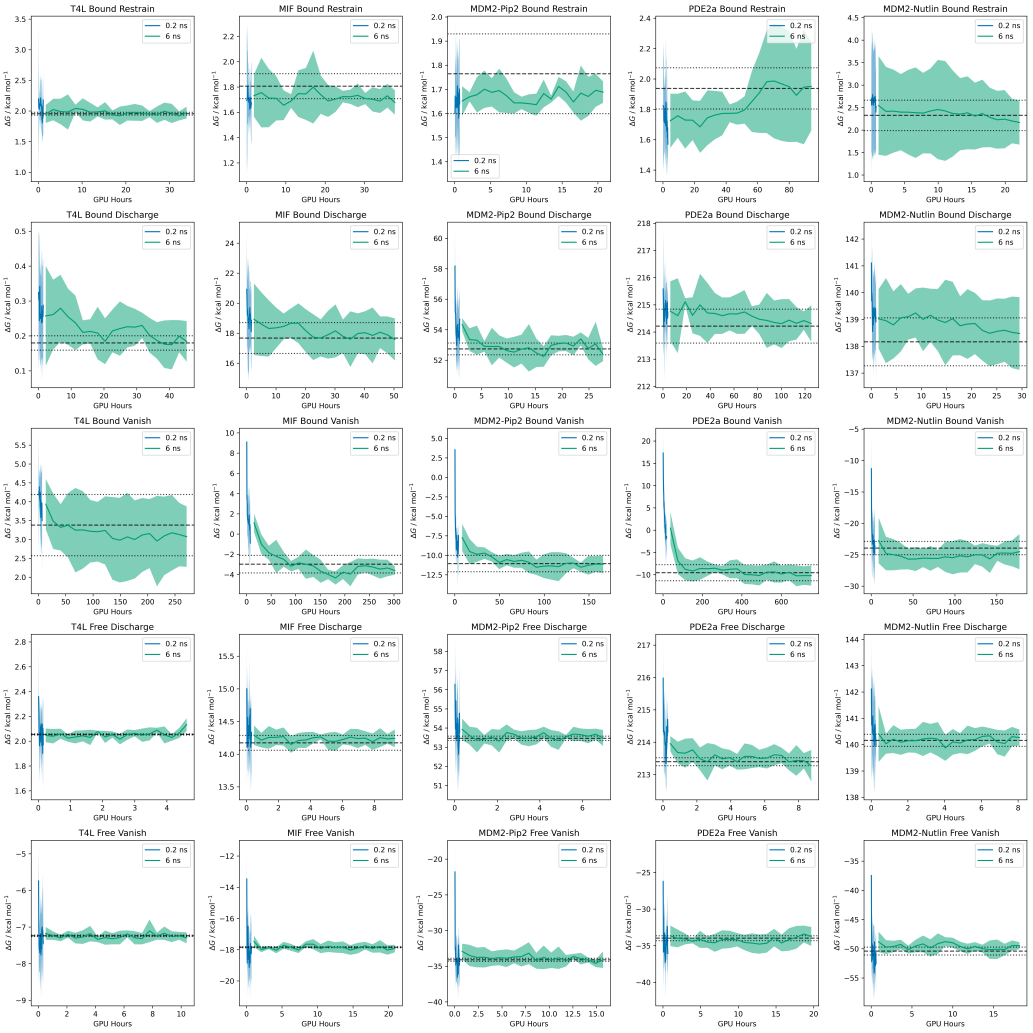


Figure B.1: Estimated free energy changes against sampling times for all stages of all initial non-adaptive runs. Traces were calculated by dividing the data for each run into 20 blocks and performing MBAR individually on each block. No data were discarded to equilibration. Shaded areas show the 95 % t-based confidence intervals based on deviations between replicate runs. Results are shown for the 0.2 and 6 ns runs, and the final 30 ns result (discarding the initial 10 ns of each window to equilibration) is shown as a dotted line, with 95 % CI boundaries shown as sparser dotted lines.

B.7 Summary of Kruskal-Wallis H-tests on Gradient Distributions

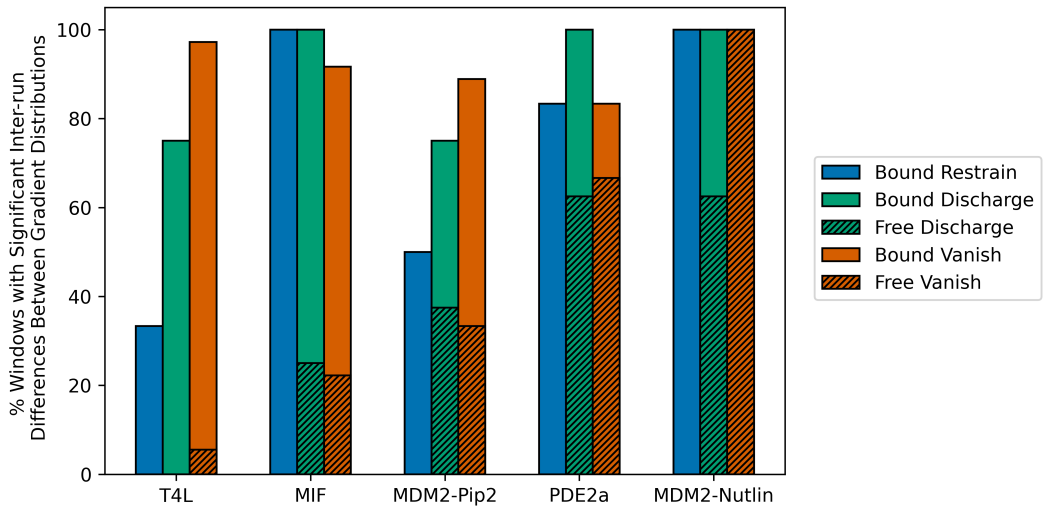


Figure B.2: Percentage of λ-windows for each stage for which there is a significant difference between gradient distributions, as indicated by $p< 0.05$ from the Kruskal-Wallis H-test. This is always high for the bound discharge and bound vanish stages, and increases from around 0 to > 60 with ligand size for the free stages.

Figure B.2 shows significant differences between gradient distributions between replicate runs for most λ-windows for all bound leg simulations. A strict criterion for convergence is that all repeat simulations should sample from the same distribution. By this criterion, none of the calculations are strictly converged, as discussed in the main text. To illustrate that this is likely also true for most literature ABFE studies, we reanalysed the results of Alibay et al..[1,288] The analysis (Figure B.3 yielded similar results, confirming that most literature ABFE results are also likely unconverged by this strict criterion.

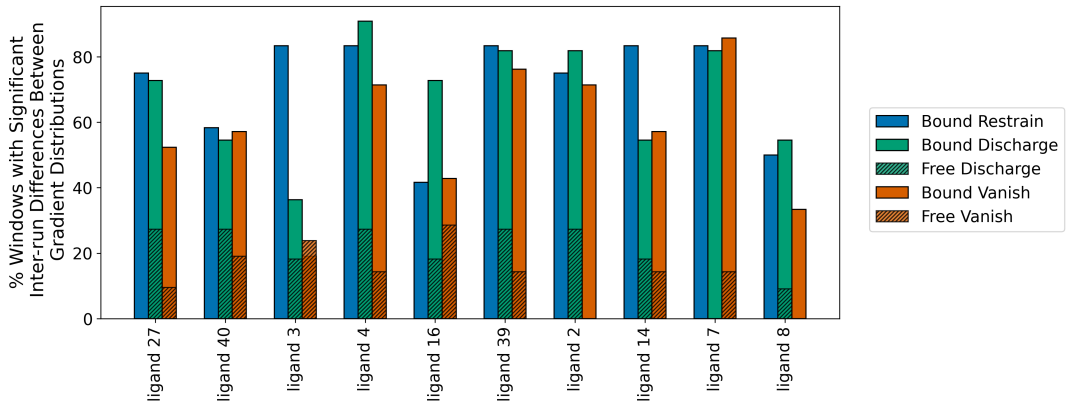


Figure B.3: Percentage of λ-windows for each stage for which there is a significant difference between gradient distributions for Alibay et al.'s Cyclophilin D ABFE calculations.[1,288] Significant differences were indicated by $p< 0.05$ from the Kruskal-Wallis H-test.

B.8 Standard Deviations of the Gradients and Time-Normalised Standard Errors of the Mean Gradients

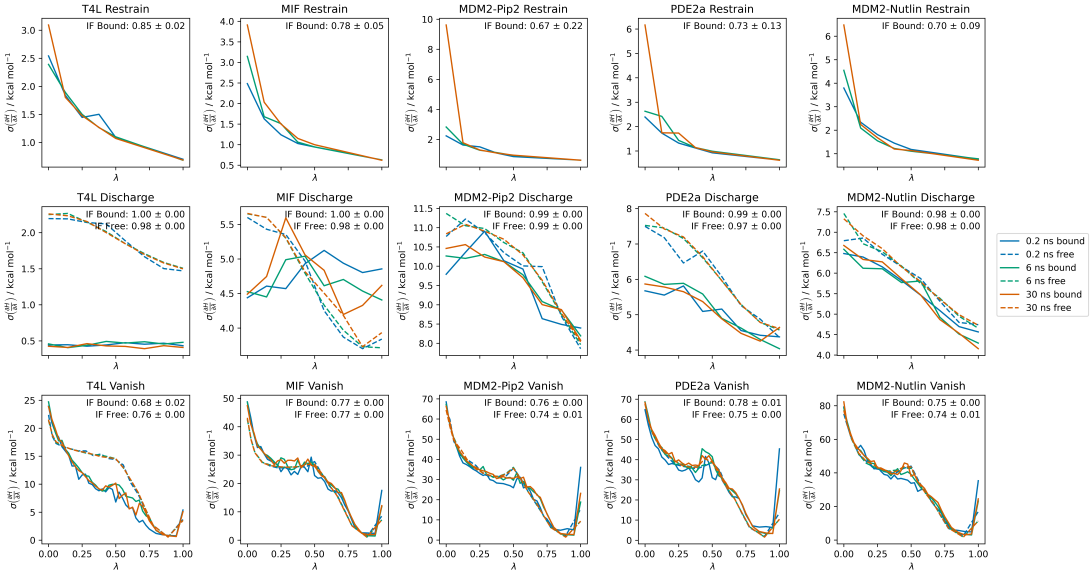


Figure B.4: Standard deviation of the gradient against λ for initial non-adaptive runs. Improvement factor (IF) is calculated with respect to equally-spaced λ-windows.

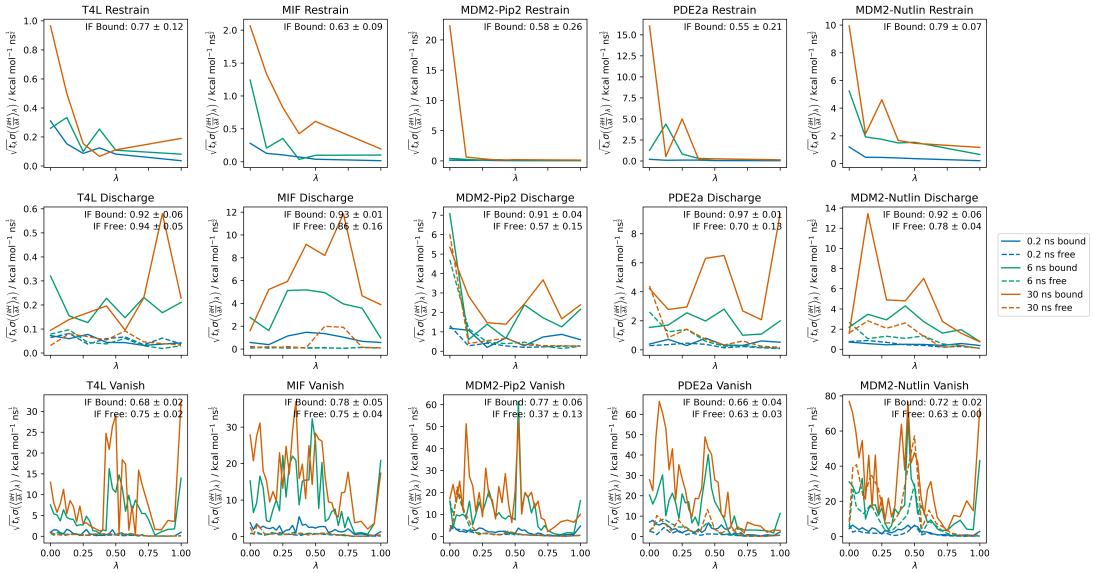


Figure B.5: Time-normalised standard error of the mean gradient against λ for initial non-adaptive runs. Improvement factor (IF) is calculated with respect to equally-spaced λ-windows.

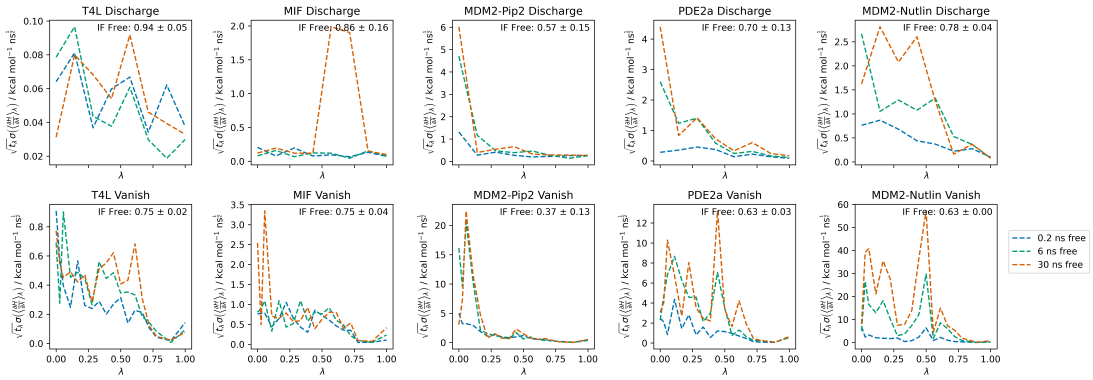


Figure B.6: Time-normalised standard error of the mean gradient against λ for initial non-adaptive runs, showing only results for the free leg. Improvement factor (IF) is calculated with respect to equally-spaced λ-windows.

B.9 Nutlin Conformations Observed During the Free Vanish Leg At and Above $\lambda = 0.5$

The time-normalised standard error of the mean gradients for the free vanish stage were high relative to the other ligands, and comparable to those for the bound legs. The peak around $\lambda = 0.5$ visible in Figure B.6 is very likely due to different dominant conformations of Nutlin resulting in different amounts of overlap and differing gradients between runs. Below $\lambda = 0.5$, the intramolecular repulsive interactions are strong enough to avoid direct overlap as seen in b) of Figure B.4. The window at $\lambda = 0.5$ is the first where this occurs, producing more negative gradients for runs 3 and 5 where conformation b) is dominant, and less negative gradients for runs 1, 2, and 4, where conformation a) is dominant.

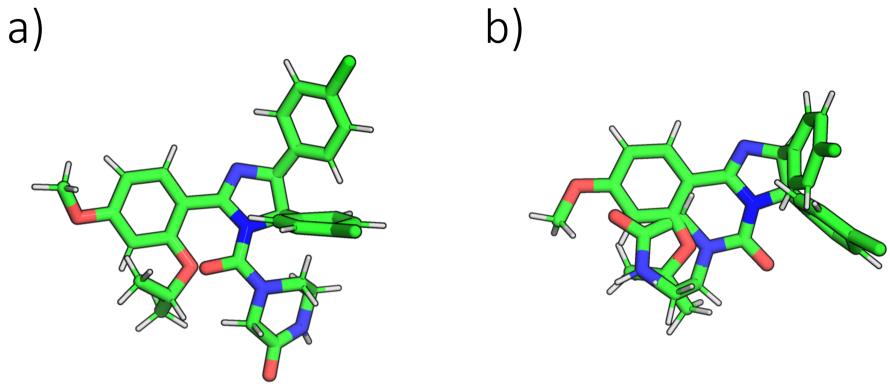


Figure B.7: Examples of typical conformations observed for Nutlin during the free vanish stages a) with the piperazinone ring pointing away from the isopropoxy group, reducing atomic overlap and b) with the piperazinone ring pointing towards the isopropoxy group, resulting in substantial atomic overlap.

B.10 Testing Automated Window Spacing with MIF

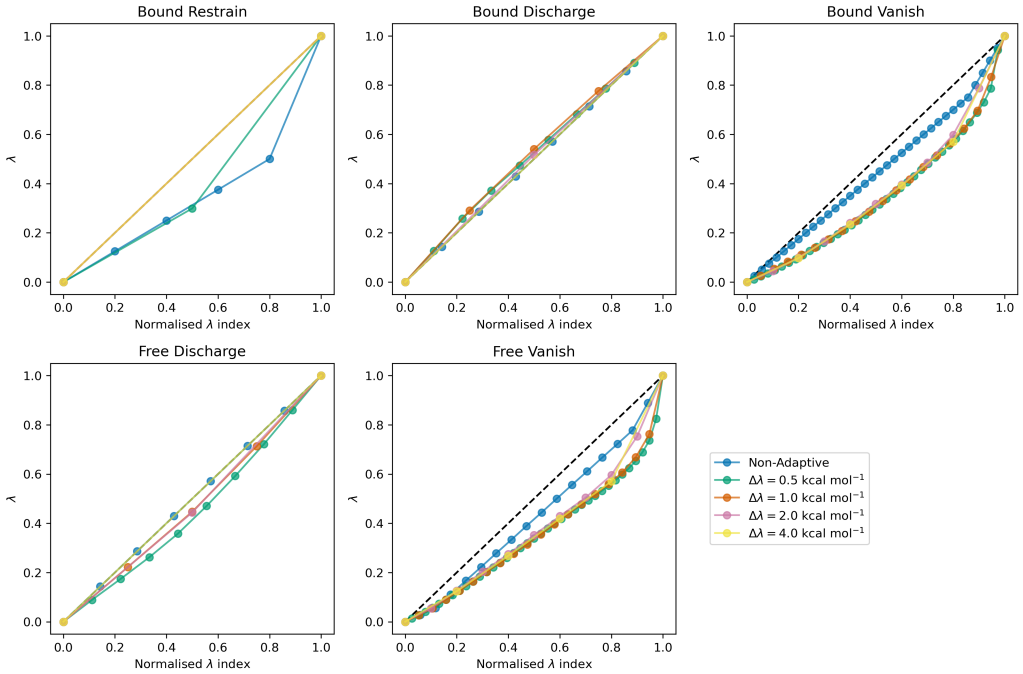


Figure B.8: λ against normalised λ index ($\frac{\text{Index}}{\text{No.windows}}$) for all legs and stages for MIF using a manually-optimised λ schedule (non-adaptive), and using the automated method to space windows with thermodynamic speeds of 0.5, 1.0, 2.0, and 4.0 kcal mol^{-1}.

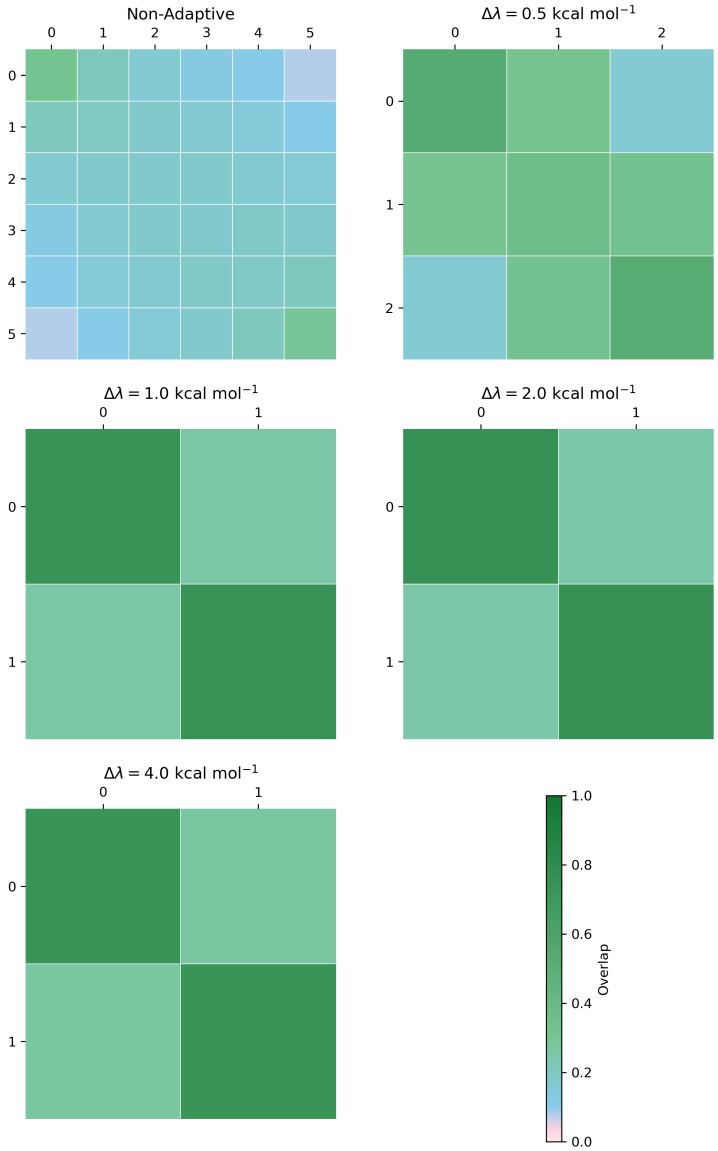


Figure B.9: Overlap matrices for the bound restrain stage for MIF, using a manually-optimised λ schedule, and using the automated method to space windows with thermodynamic speeds of 0.5, 1.0, 2.0, and 4.0 kcal mol^{-1}.

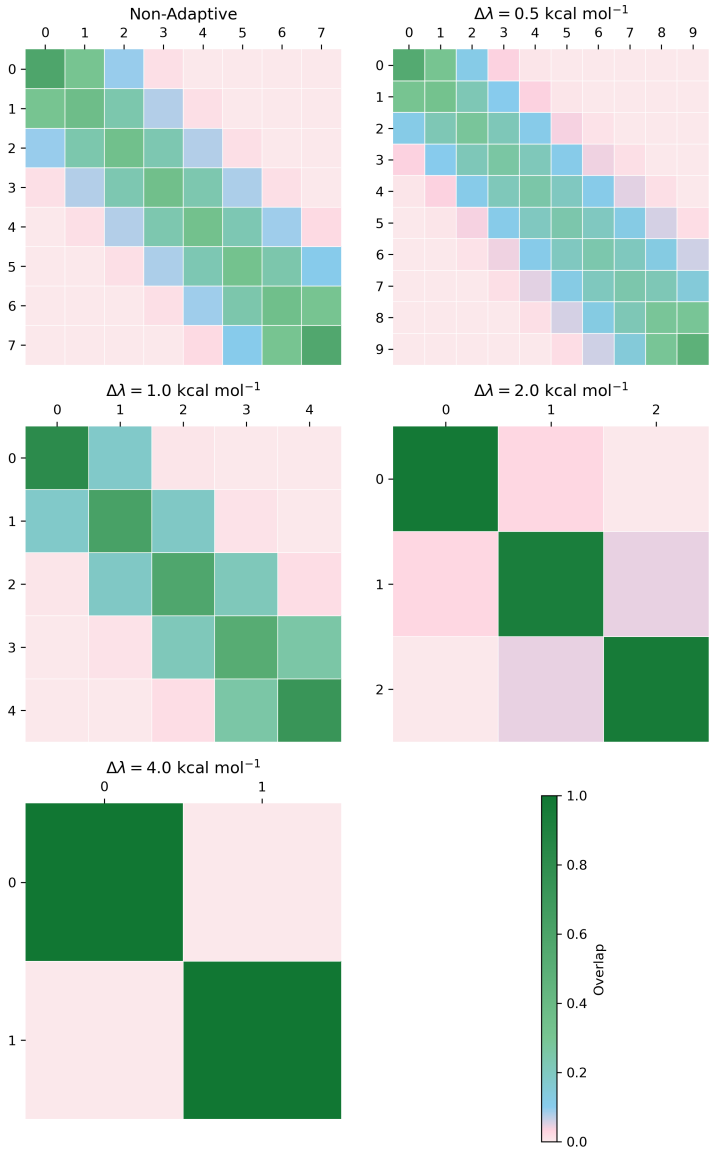


Figure B.10: Overlap matrices for the bound discharge stage for MIF, using a manually-optimised λ schedule, and using the automated method to space windows with thermodynamic speeds of 0.5, 1.0, 2.0, and 4.0 kcal mol^{-1}.

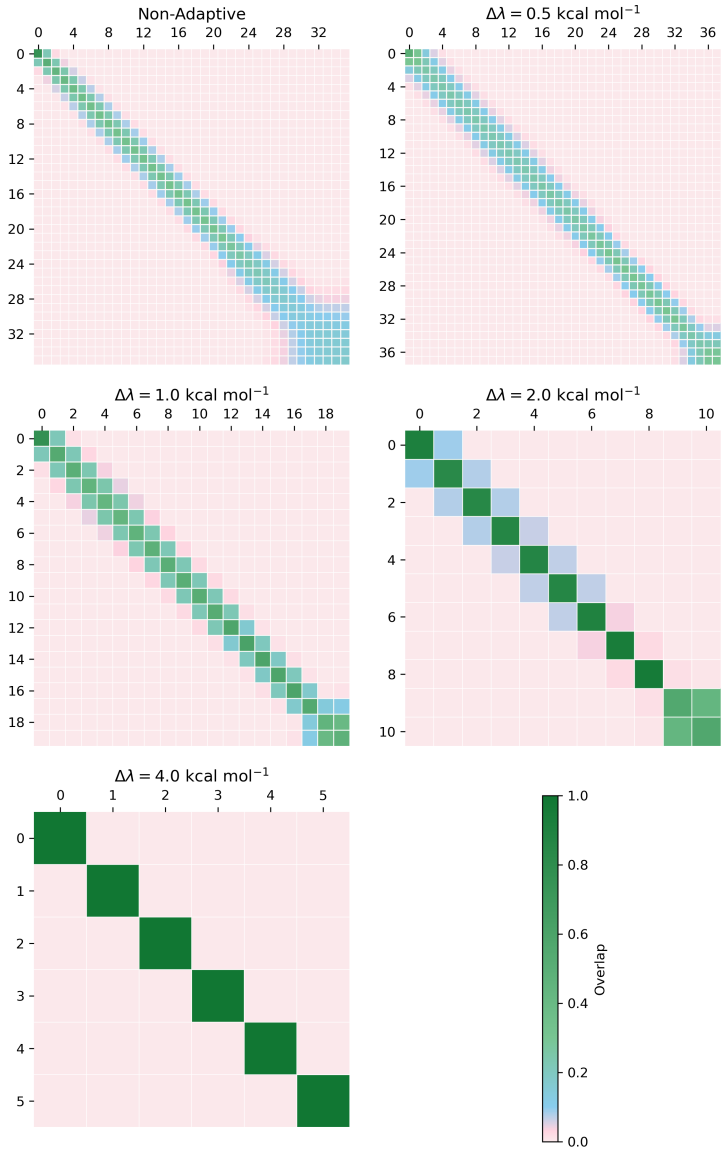


Figure B.11: Overlap matrices for the bound vanish stage for MIF, using a manually-optimised λ schedule, and using the automated method to space windows with thermodynamic speeds of 0.5, 1.0, 2.0, and 4.0 kcal mol^{-1}.

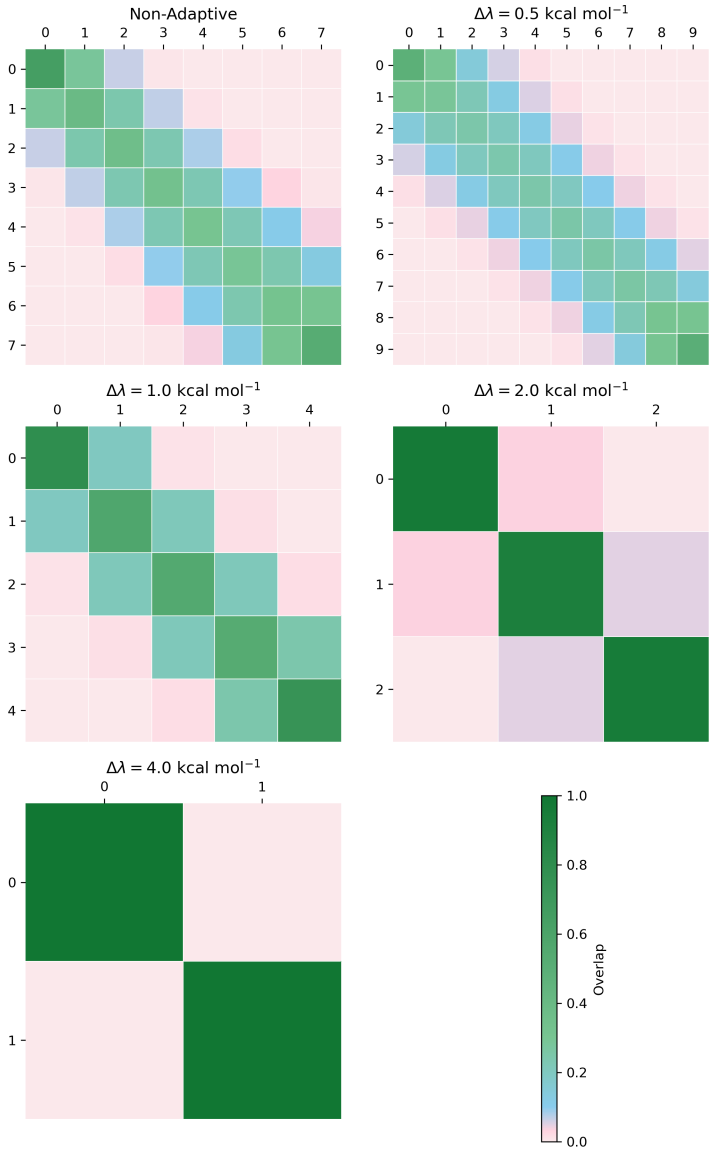


Figure B.12: Overlap matrices for the free discharge stage for MIF, using a manually-optimised λ schedule, and using the automated method to space windows with thermodynamic speeds of 0.5, 1.0, 2.0, and 4.0 kcal mol^{-1}.

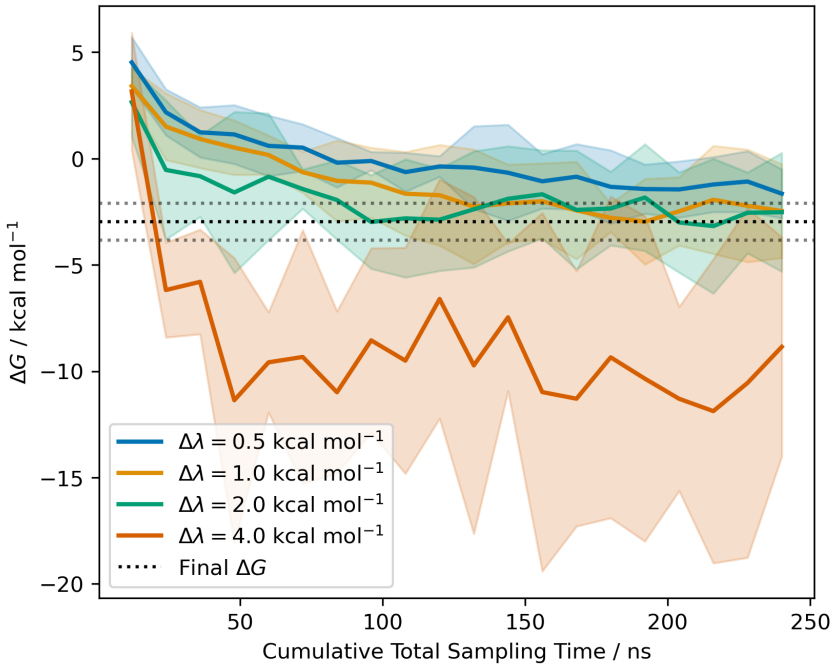


Figure B.13: Longer-time equilibration of the MIF bound vanish stage against total sampling time with varying λ-window spacing. Individual replicates are shown with dashed lines, the mean is shown as a solid line, and shaded regions indicate 95 % t-based confidence intervals. Wider spacing leads to accelerated equilibration towards the 30 ns result (black-dotted line with 95 % CI shown as sparse dotted lines).

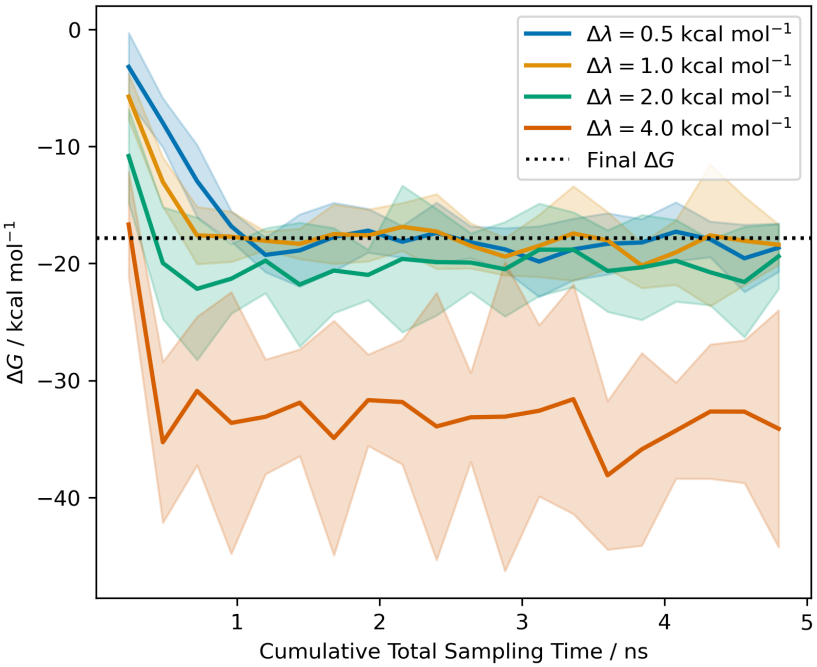


Figure B.14: Short-time equilibration of the MIF free vanish stage against total sampling time with varying λ-window spacing. Individual replicates are shown with dashed lines, the mean is shown as a solid line, and shaded regions indicate 95 % t-based confidence intervals. Wider spacing leads to accelerated equilibration towards the 30 ns result (black-dotted line with 95 % CI shown as sparse dotted lines).

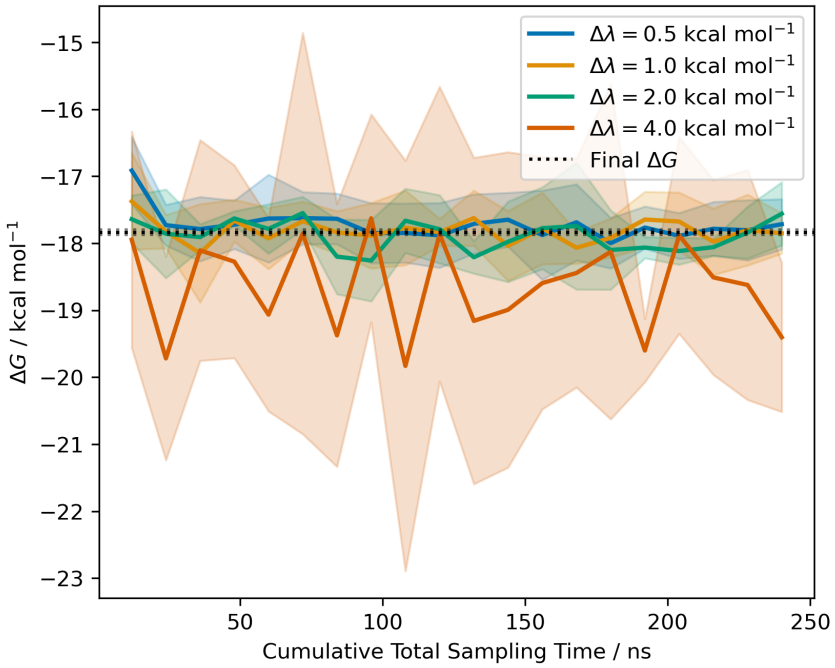


Figure B.15: Longer-time equilibration of the MIF free vanish stage against total sampling time with varying λ-window spacing. Individual replicates are shown with dashed lines, the mean is shown as a solid line, and shaded regions indicate 95 % t-based confidence intervals. Wider spacing leads to accelerated equilibration towards the 30 ns result (black-dotted line with 95 % CI shown as sparse dotted lines).

B.11 Free Energies for Variation of Allocated Simulation Time with Adaptive Parameters for MIF

Table B.3: Components of Free Energies for Variation of Allocated Simulation Time with Simulation Parameters for MIF Experiments [a]

	Bound Restrain	Bound Discharge	Bound Vanish	Free Discharge	Free Vanish	Restraint Correction	Symmetry Correction	Exp. $\Delta G^{o}_{\mathrm{Bind}}$
r0.005, s1, n5, repeat 1	1.73 ± 0.15	18.17 ± 0.82	-2.13 ± 1.25	14.25 ± 0.11	-17.97 ± 0.20	-10.35	0.65	-8.98 ± 0.28
r0.005, s1, n5, repeat 2	1.72 ± 0.13	17.98 ± 1.04	-2.19 ± 0.83	14.21 ± 0.06	-17.77 ± 0.05	-10.35	0.65	-8.98 ± 0.28
r0.005, s1, n5, repeat 3	1.73 ± 0.15	18.61 ± 1.14	-2.79 ± 1.43	14.25 ± 0.06	-17.68 ± 0.37	-10.35	0.65	-8.98 ± 0.28
r0.005, sOrig., n5	1.71 ± 0.16	17.70 ± 1.06	-1.06 ± 0.64	14.16 ± 0.11	-17.62 ± 0.10	-10.35	0.65	-8.98 ± 0.28
r0.001, s1, n5	1.79 ± 0.15	18.28 ± 0.74	-2.56 ± 0.42	14.24 ± 0.04	-17.79 ± 0.08	-10.35	0.65	-8.98 ± 0.28
r0.005, s1, n10	1.73 ± 0.11	18.07 ± 0.70	-1.40 ± 0.89	14.24 ± 0.11	-17.80 ± 0.19	-10.35	0.65	-8.98 ± 0.28
r0.005, s0.5, n5	1.69 ± 0.15	18.41 ± 1.40	-1.91 ± 0.43	14.20 ± 0.12	-17.75 ± 0.20	-10.35	0.65	-8.98 ± 0.28
r0.005, s2, n5, repeat 1	1.70 ± 0.10	17.74 ± 1.35	-3.53 ± 0.94	14.35 ± 0.13	-17.66 ± 0.08	-10.35	0.65	-8.98 ± 0.28
r0.005, s2, n5, repeat 2	1.80 ± 0.21	17.74 ± 1.20	-2.08 ± 0.91	14.31 ± 0.15	-17.69 ± 0.44	-10.35	0.65	-8.98 ± 0.28
r0.005, s4, n5	1.71 ± 0.15	18.14 ± 1.02	-3.61 ± 1.59	14.78 ± 0.59	-17.70 ± 0.44	-10.35	0.65	-8.98 ± 0.28

[a] All quantities in kcal mol^{-1}. Uncertainties stated as 95 % confidence intervals based on the variance of 5 replicate runs, assuming Gaussian distributions. In the protocol names, r denotes the run time constant ($\mathrm{kcal}^2\ \mathrm{mol}^{-2}\ \mathrm{ns}^{-1}$), s denotes the thermodynamic speed (kcal mol^{-1}) used to determine the λ-window spacing, and n denotes the number of replicates in each ensemble run. sOrig denotes the use of the default λ-spacing.

B.12 Standard Error of the Mean Free Energy Change for the Pip2 Free Stage with Adaptive and Non-Adaptive Protocols

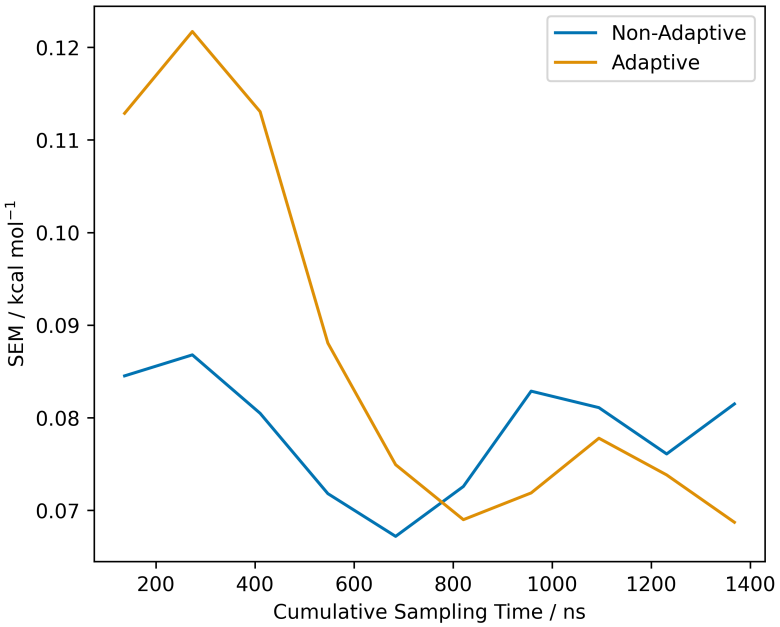


Figure B.16: Standard error of the mean free energy change for the free vanish stage of Pip2. Data were split into 10 blocks prior to analysis with MBAR, and the uncertainties only account for the data in the current block (and are therefore not expected to decrease with $\sqrt{N_{\text{Blocks}}}$).

B.13 Effect of the Adaptive Allocation of Sampling Time to the Bound Vanish Stage of T4L/Benzene

During the bound vanish stage of the non-adaptive 30 ns T4L run, the Gelman-Rubin $\hat{R}$ clearly identified convergence issues above $\lambda = 0.4$, and in particular close to $\lambda = 1$ (Figure B.17). This was also reflected by the divergence of the potential of mean force with respect to λ in this region (Figure B.18).

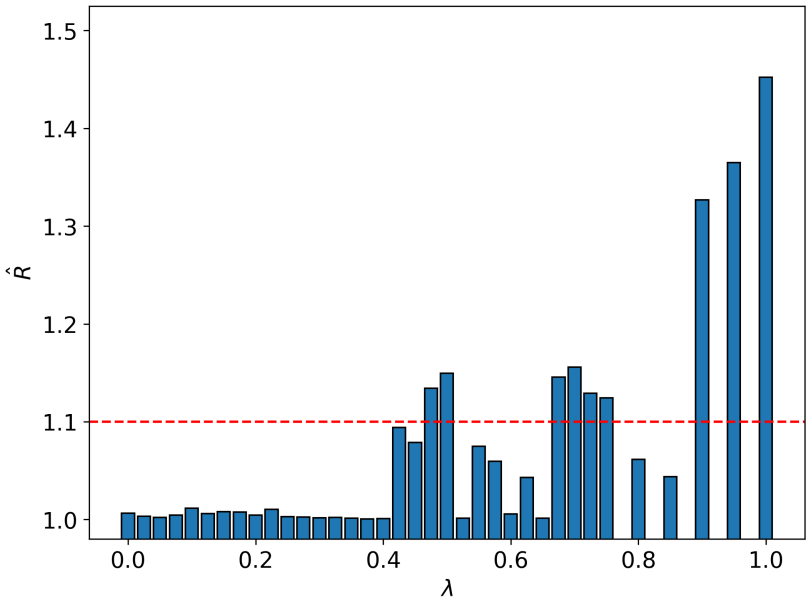


Figure B.17: Gelman-Rubin $\hat{R}$ against λ for the bound vanish stage of the 30 ns T4L non-adapitve run. $\hat{R} > 1.1$ is taken to indicate convergence issues. $\hat{R}$ was computed with ArviZ 0.15.1.[364]

This was found to be due to the occasional entry of water to the binding site for some replicates above $\lambda = 0.4$. To check if we could remedy this using the adaptive sampling time algorithm, we performed adaptive and non-adaptive runs using 20 replicates and the same λ schedules (speed = 1.0 kcal mol^{-1}). The non-adaptive run was run for 20 ns per window, and the runtime constant of the adaptive run was tweaked until comparable total simulation time was achieved. As expected, the adaptive algorithm concentrated sampling time around $\lambda = 1$ (Figure B.19). However, the plots of free energy estimates against simulation time for the adaptive and non-adaptive protocols appear very similar (Figure B.20) , and there was no evidence for a significant difference in variance at 95 % confidence (Levene test, p=0.92). In addition, Figure B.21 shows that the water occupancy of the binding site remained very variable between replicates and appeared similar between the adaptive and non-adaptive protocols (water

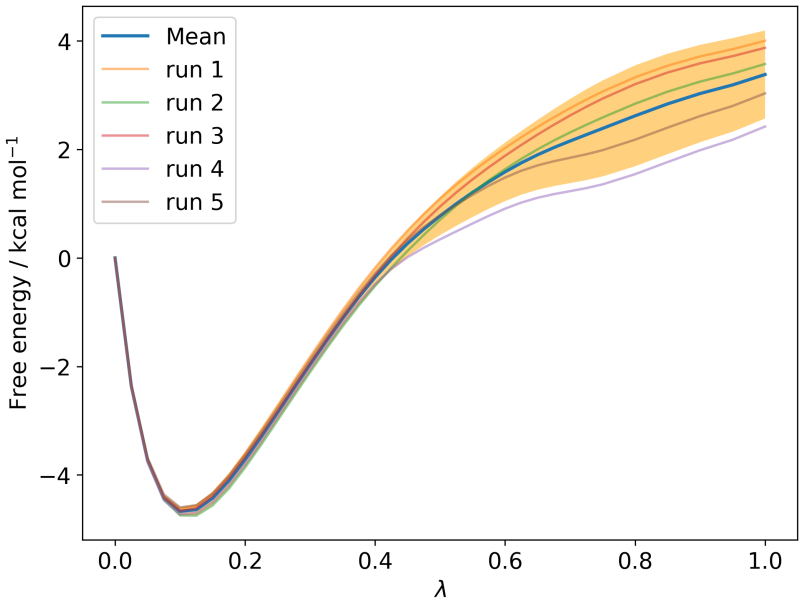


Figure B.18: MBAR-derived potential of mean force against λ for the bound vanish stage of the 30 ns T4L non-adapitve run.

occupancy was assessed by the mean number of waters with 6 Å of either of two atoms on opposite sides of the benzene rings). This suggests that the adaptive sampling algorithm is ineffective here because the timescales of water entry and exit from the binding site are beyond what can reasonably be allocated by the adaptive algorithm.

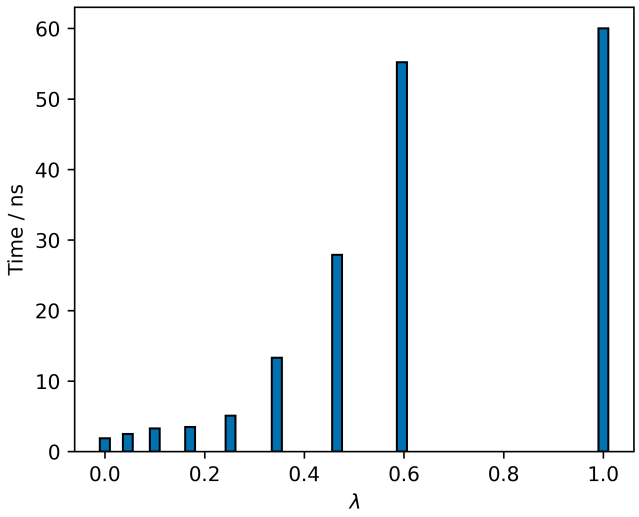


Figure B.19: Allocation of sampling time against λ for the T4L bound vanish leg.

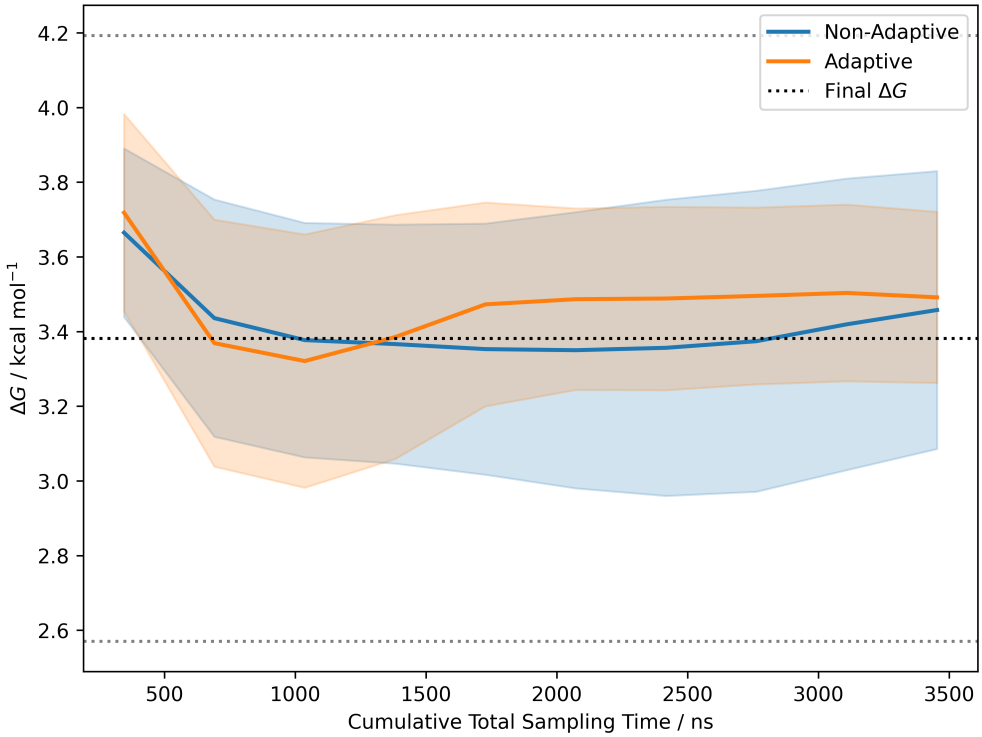


Figure B.20: Estimated free energy change against simulation time for the T4L bound vanish stage with adaptive and non-adaptive protocols. The 30 ns result is shown by a black dotted line with sparser dotted lines showing 95 % CI. Data was split into 10 blocks before analysis with MBAR. Shaded areas show 95 % t-based CIs. The adaptive and non-adaptive runs use the same λ-schedule.

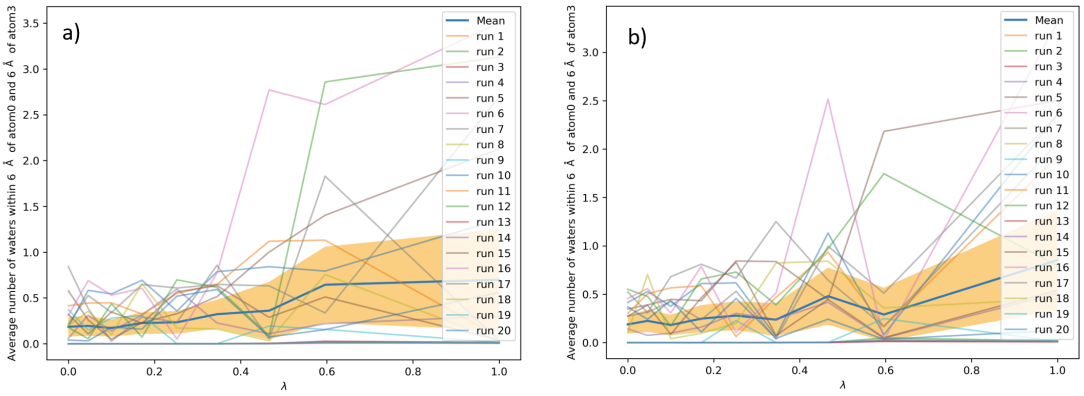


Figure B.21: Average number of waters in the binding site for a) the non-adaptive and b) the adaptive protocol. Shaded area shows the 95 % t-based CI, although the assumption of Gaussian distributions is clearly unreasonable here. Atoms 0 and 3 are on opposite sides of the benzene ring.

B.14 Selection of Equilibration Times for All Stages of Initial Test Systems

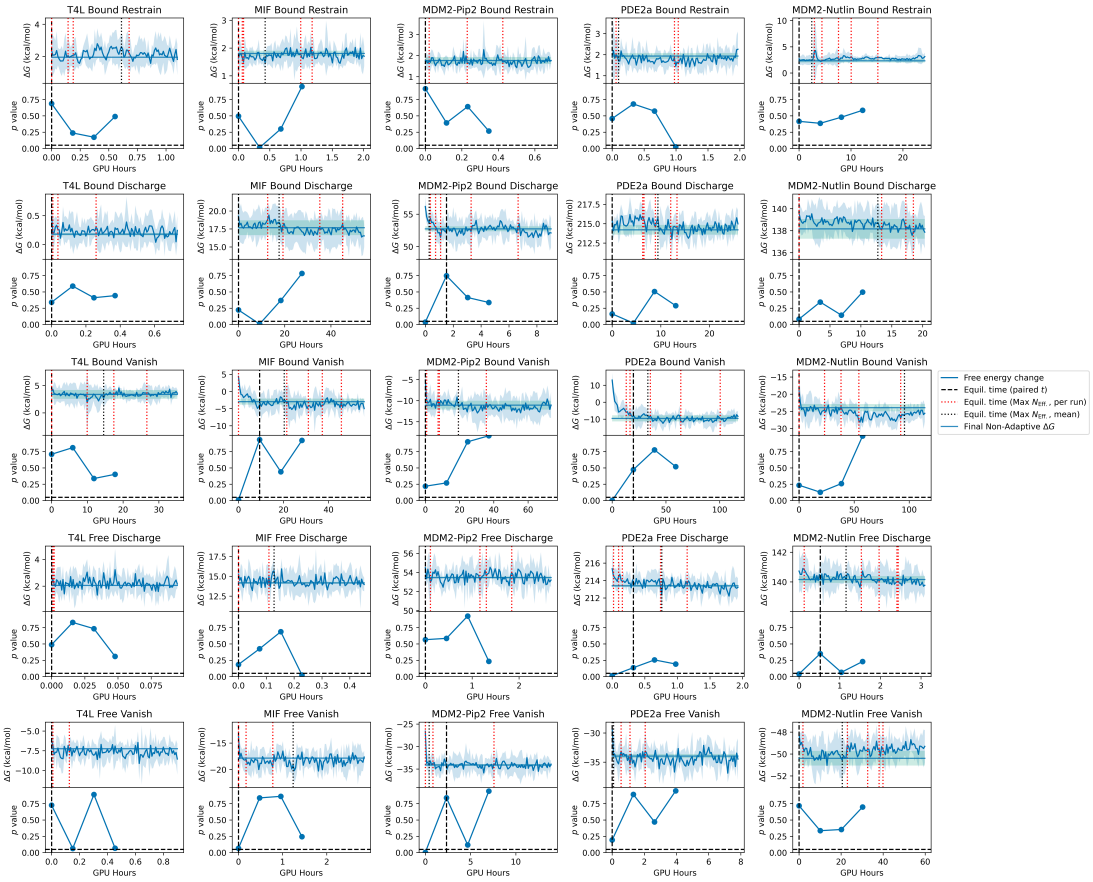


Figure B.22: Selection of equilibration times for all stages for initial test systems by the paired t-test method and Chodera's method (applied to both the mean trace and individual replicates).[262] The upper windows show traces of ΔG obtained by dividing the data up into 100 equal blocks and running MBAR on each. Shaded areas indicate 95 % inter-run t-based confidence intervals. Final Non-Adaptive ΔG taken from the non-adaptive 30 ns runs. Lower windows show the p-values obtained by truncating the data up to the time shown and performing paired t-tests on the first 10 % and last 50 % of the remaining data. The first p-value > 0.05 is used as a heuristic to indicate equilibration.

B.15 Comparison of Absolute Differences in Final Free Energy Estimates Compared to Long-Time Result Using Different Equilibration Methods

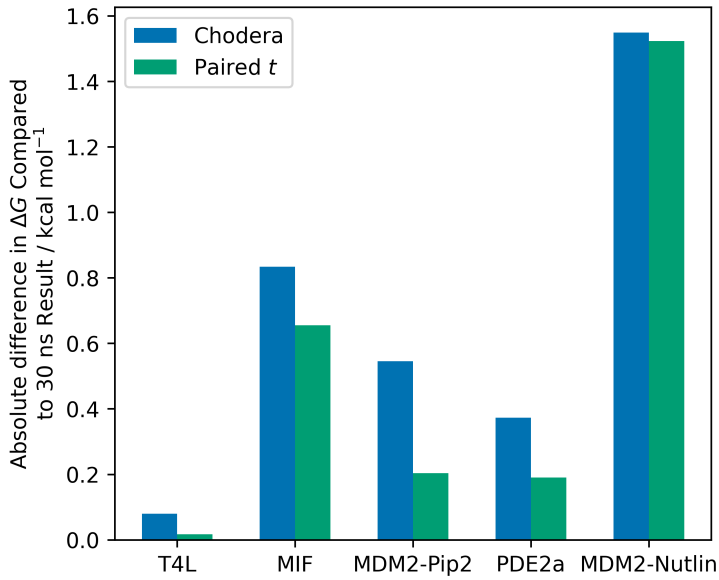


Figure B.23: Comparison of the absolute error between bound vanish ΔG estimates using different equilibration methods, and the long-time non-adaptive results for these stages (from the 30 ns runs). The absolute errors are always smaller for the paired t method compared to Chodera's method (run on the mean trace). However, there is no evidence for a significant difference at 95 % based on the Wilcoxon signed-rank test ($p = 0.06$).

B.16 Detailed Results for "Optimised" Adaptive ABFE Calculations on Initial Test Systems

Table B.4: Components of Non-Adaptive $\Delta G^{o}_{\mathrm{Bind}}$ for Initial Test Systems[a]

	Bound Restrain	Bound Discharge	Bound Vanish	Free Discharge	Free Vanish	Restraint Correction	Symmetry Correction	Exp. $\Delta G^{o}_{\mathrm{Bind}}$
T4L adaptive	1.97 ± 0.06	0.23 ± 0.14	3.64 ± 0.71	2.11 ± 0.08	-7.22 ± 0.15	-7.08	-0.41	-5.19 ± 0.16
T4L non-adaptive	1.96 ± 0.02	0.18 ± 0.02	3.38 ± 0.81	2.05 ± 0.00	-7.24 ± 0.02	-7.08	-0.41	-5.19 ± 0.16
MIF adaptive	1.71 ± 0.15	17.71 ± 1.61	-2.70 ± 1.16	14.21 ± 0.14	-17.72 ± 0.13	-10.35	-0.65	-8.98 ± 0.28
MIF non-adaptive	1.81 ± 0.10	17.68 ± 1.03	-2.98 ± 0.87	14.17 ± 0.11	-17.84 ± 0.03	-10.35	-0.65	-8.98 ± 0.28
MDM2-Pip2 adaptive	1.67 ± 0.00	52.73 ± 0.68	-10.55 ± 1.18	53.50 ± 0.29	-33.78 ± 0.74	-10.14	0.00	-9.11 ± 0.01
MDM2-Pip2 non-adaptive	1.76 ± 0.17	52.74 ± 0.39	-11.06 ± 1.05	53.46 ± 0.11	-34.10 ± 0.21	-10.14	0.00	-9.11 ± 0.01
PDE2A adaptive	1.73 ± 0.15	214.69 ± 0.70	-8.22 ± 1.82	213.47 ± 0.23	-33.87 ± 0.27	-10.24	0.00	-14.35 ± 0.82
PDE2A non-adaptive	1.94 ± 0.14	214.22 ± 0.62	-9.53 ± 1.81	213.40 ± 0.12	-33.97 ± 0.34	-10.24	0.00	-14.35 ± 0.82
MDM2-Nutlin adaptive	2.61 ± 0.52	138.62 ± 1.17	-24.63 ± 0.84	140.16 ± 0.35	-49.55 ± 0.51	-9.85	0.00	-11.14 ± 0.27
MDM2-Nutlin non-adaptive	2.33 ± 0.34	138.17 ± 0.89	-23.94 ± 1.06	140.16 ± 0.23	-50.39 ± 0.67	-9.85	0.00	-11.14 ± 0.27

[a] All quantities in kcal mol^{-1}. Uncertainties stated as 95 % confidence intervals based on the variance of 5 replicate runs, assuming Gaussian distributions. "adaptive" refers to the "optimised" adaptive protocol, while "non-adaptive" refers to the the 30 ns non-adaptive protocol.

B.17 Selection of Windows for Initial Test Systems with "Optimised" Adaptive Protocol

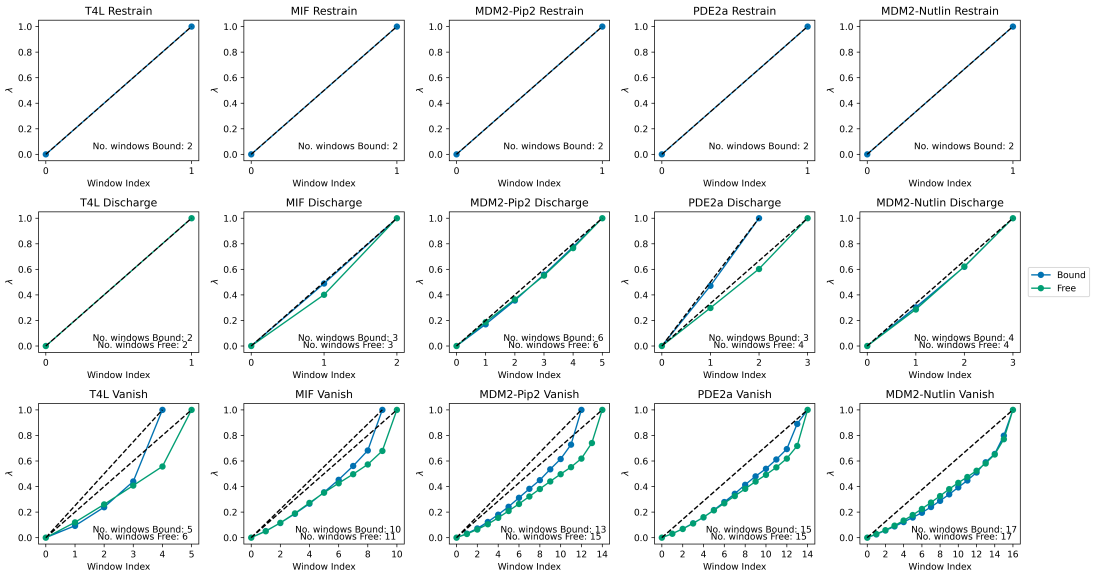


Figure B.24: λ values selected for each of the initial test systems using a thermodynamic speed of 2 kcal mol^{-1}.

B.18 Detailed Breakdown of Sampling Time Allocation for Initial Test Systems with "Optimised" Non-Adaptive Protocol

Note that the total allocation of simulation time depends not only on the inter-run error, but also the relative computational cost of the leg for the given system (Figure B.25). If the statistical inefficiencies were the same for all windows, the allocated sampling times (Figures B.26 and B.27) should be a function only of the number of windows allocated. In this case, the sampling time per window would be expected to be the same for windows of the same computational cost, and hence large differences in sampling time per window (Figure B.28) can highlight stages containing windows with particularly high statistical inefficiencies.

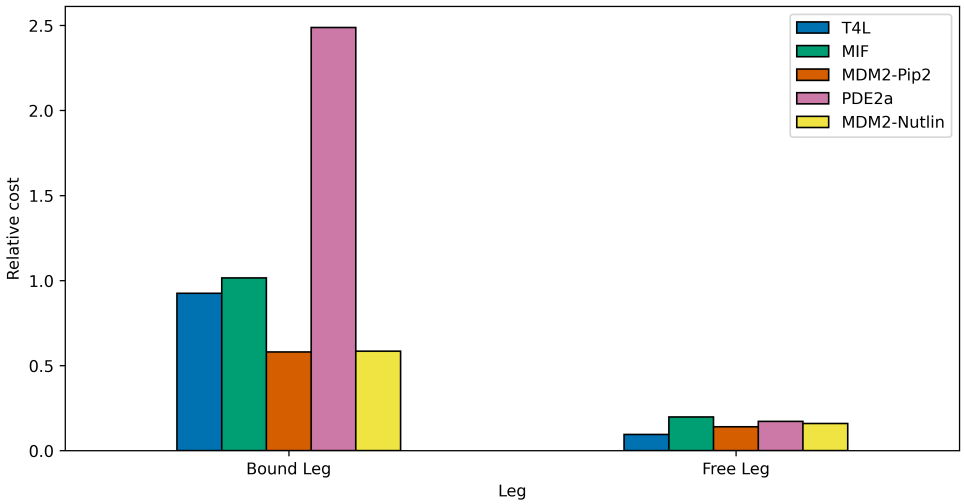


Figure B.25: Relative computational costs of each leg for the initial test systems, calculated relative to the MIF180 bound leg. Calculated based on the average time taken to run a ns of simulation on a single GPU. The absolute cost of the MIF180 bound leg was 0.21 GPU hours / ns. All computational costs were assessed running on NVIDIA GeForce RTX2080 SUPER GPUs.

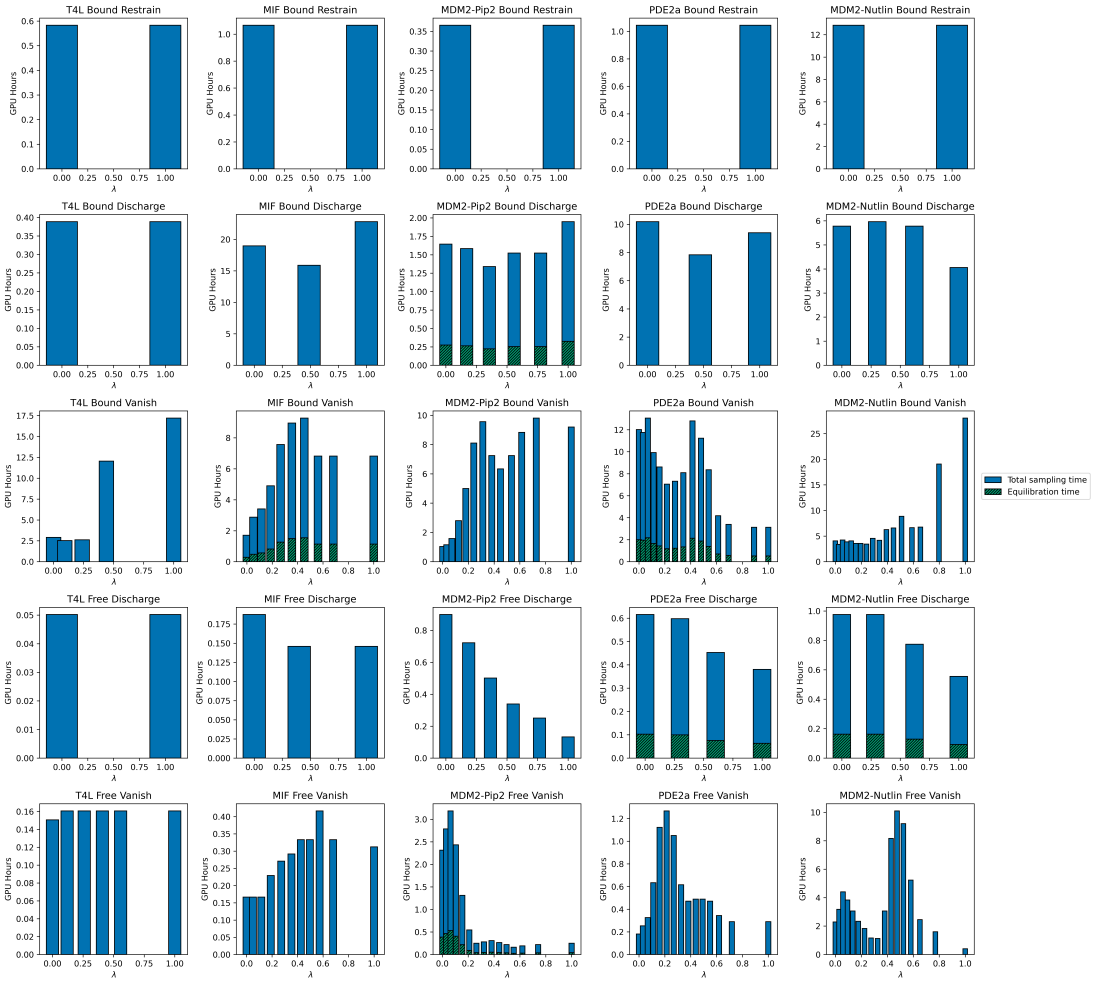


Figure B.26: Per-window breakdown of sampling times allocated to the initial test systems with the "Optimised" adaptive protocol.

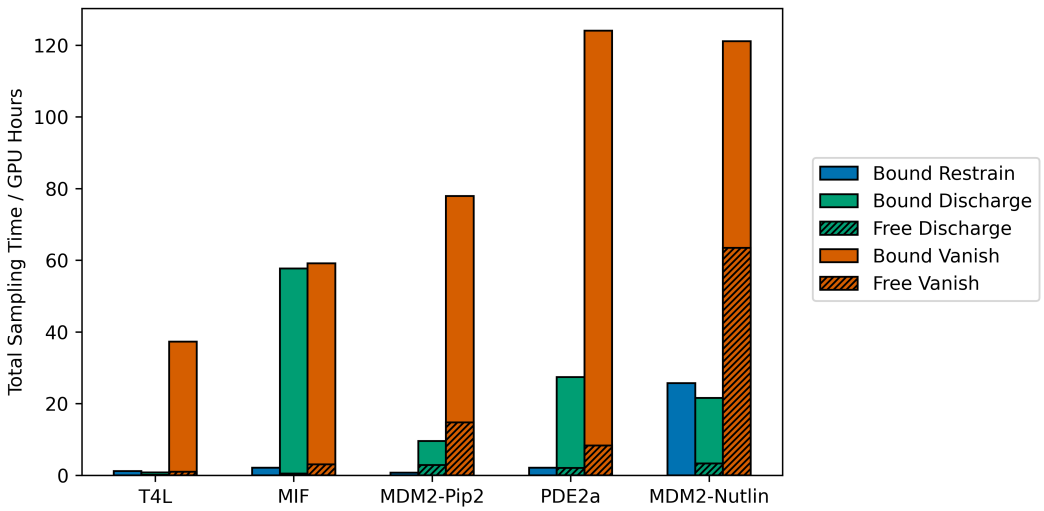


Figure B.27: Per-stage breakdown of sampling times allocated to the initial test systems with the "Optimised" adaptive protocol.

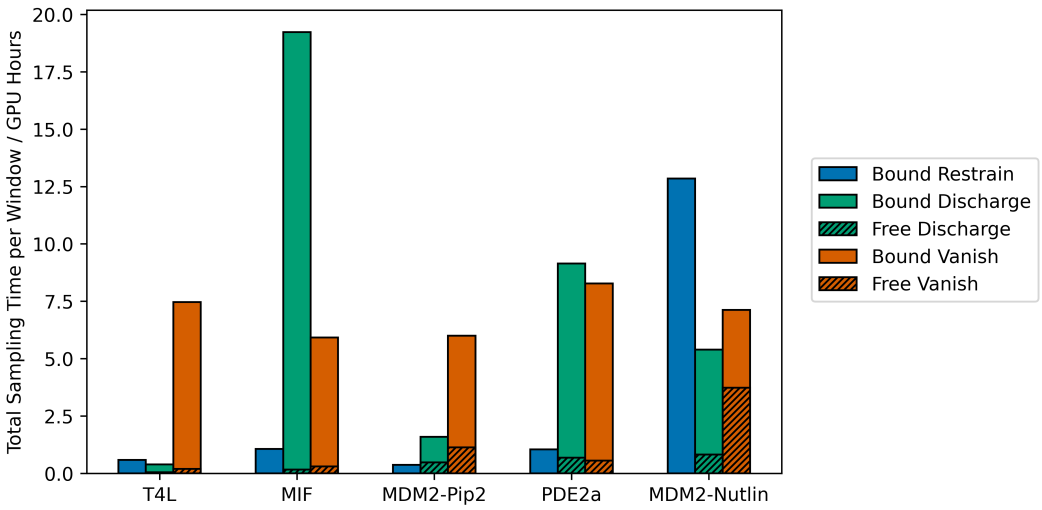


Figure B.28: Per-stage breakdown of the per-window sampling times allocated to the initial test systems with the "Optimised" adaptive protocol.

B.19 Total Allocated GPU Times for Adaptive and Non-Adaptive Protocols on Initial Systems

Table B.5: Total Allocated GPU Times for Adaptive and Non-Adaptive Protocols on Initial Systems[a]

	T4L	MIF	MDM2-Pip2	PDE2a	MDM2-Nutlin	Total
Adaptive	40	122	106	164	235	668
0.2 ns	10	12	7	27	7	63
6 ns	377	429	250	1000	255	2311
30 ns	1536	1762	1029	4059	1053	9438

[a] All sampling times in GPU hours.

B.20 Free Energy Estimates Against Sampling Time for Initial Test Systems with Adaptive and Non-Adaptive Protocols

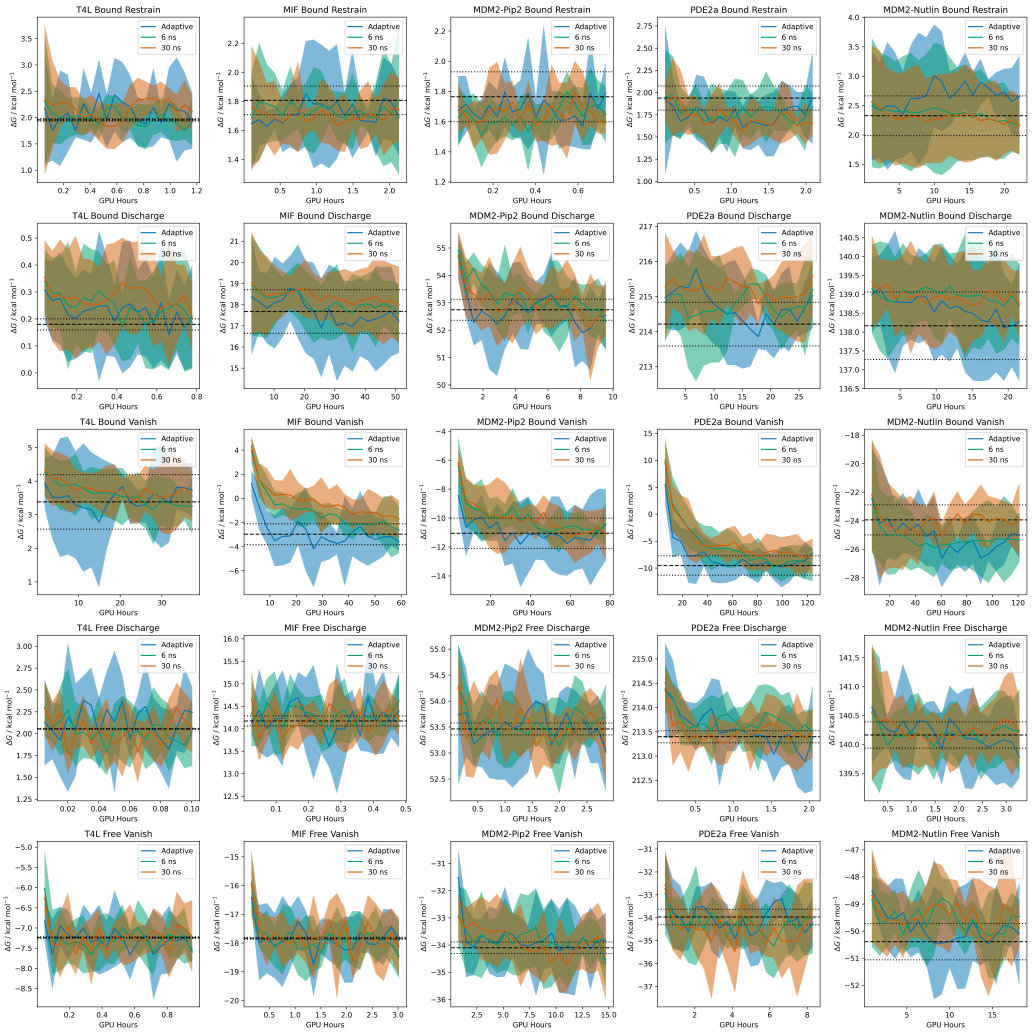


Figure B.29: Estimated ΔG against sampling time for all stages of the initial test systems. Data were split into 20 equal blocks and MBAR was run on each block. "adaptive" refers to the "optimised" adaptive protocol and "6 ns" and "30 ns" refer to the respective non-adaptive protocols. The data for truncated before analysis so that the per-stage computational costs were equal to that of the cheapest stage. Shaded areas show 95 % t-based confidence intervals. The final non-adaptive ΔG is taken from the 30 ns non-adaptive result (black dotted line with 95 % CIs shown by sparser dotted lines).

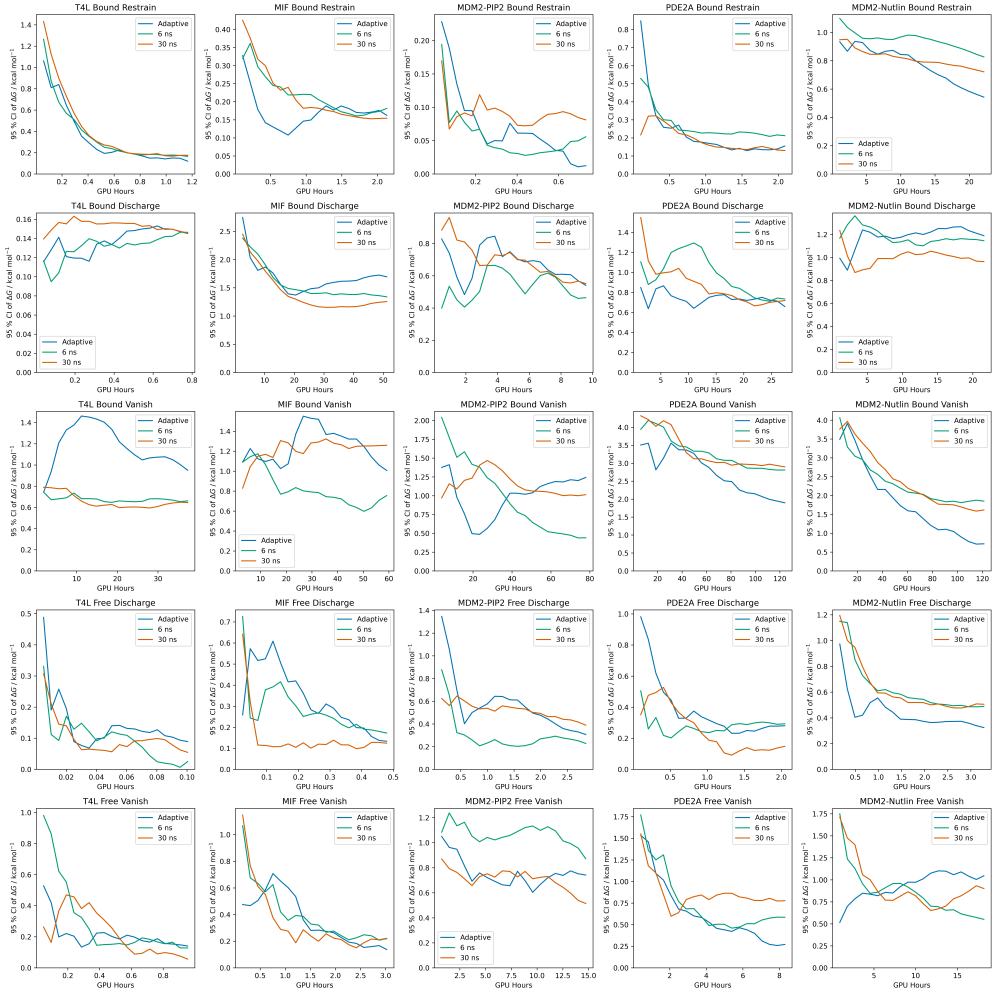


Figure B.30: 95 % t-based confidence intervals of the estimated ΔG against sampling time for all stages of the initial test systems. Data were split into 20 equal blocks and MBAR was run on each block. Analysis was performed cumulatively, meaning that all prior data were included in the error calculation for each time point, and the overall error is expected to decrease with $\frac{1}{\sqrt{\text{SamplingTime}}}$. "adaptive" refers to the "optimised" adaptive protocol and "6 ns" and "30 ns" refer to the respective non-adaptive protocols. Final portions of the data were truncated before analysis so that the per-stage computational costs were equal to that of the cheapest stage.

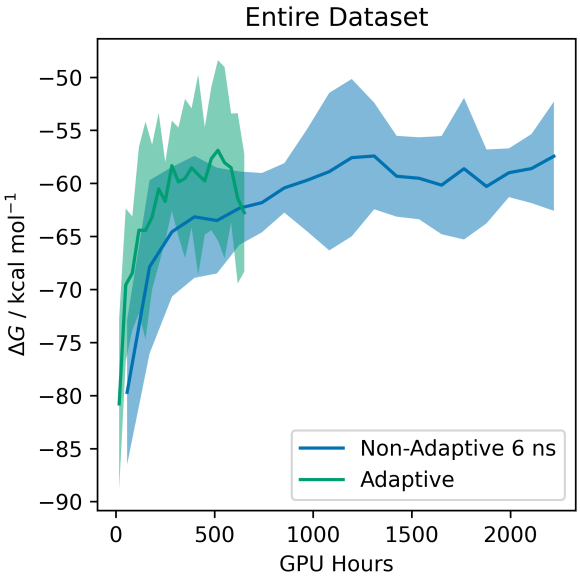


Figure B.31: Estimated ΔG against sampling time for entire dataset of initial test systems. Data were split into 20 equal blocks and MBAR was run on each block. The results are shown at the centre of each block. "Adaptive" refers to the "optimised" adaptive protocol. Shaded areas show 95 % t-based confidence intervals.

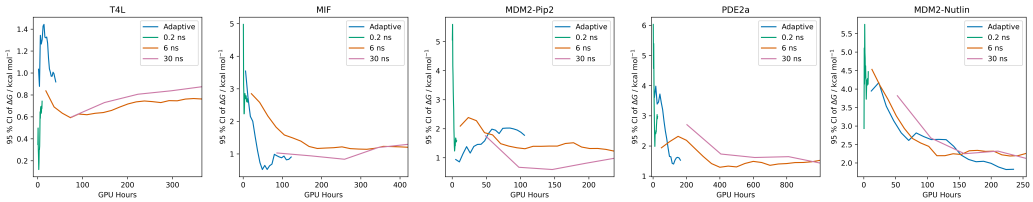


Figure B.32: 95 % t-based confidence intervals of the estimated ΔG against sampling time for all stages of the initial test systems. Data were split into 20 equal blocks and MBAR was run on each block. Analysis was performed cumulatively, meaning that all prior data were included in the error calculation for each time point, and the overall error is expected to decrease with $\frac{1}{\sqrt{\text{SamplingTime}}}$. "adaptive" refers to the "optimised" adaptive protocol and "6 ns" and "30 ns" refer to the respective non-adaptive protocols. The results are shown at the centre of each block.

B.21 Detailed Cyclophilin-D Results

Having excluded ligand 4, Alibay's results were re-analysed using our analysis protocol to produce Figure B.33 and the corresponding metrics in Table B.6.

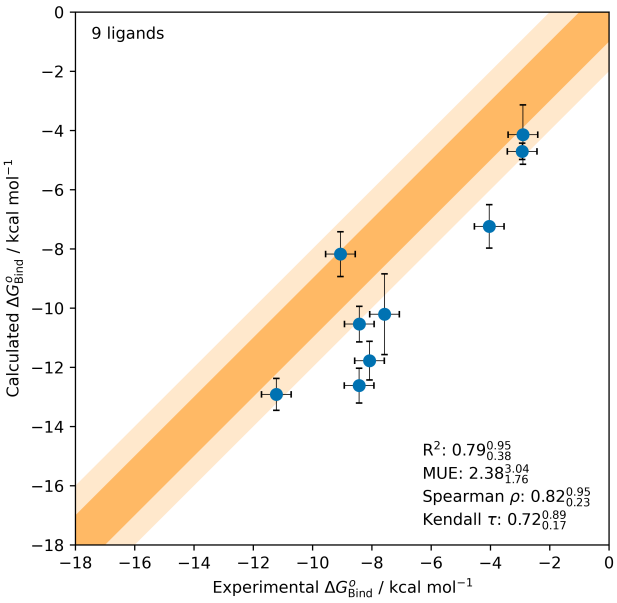


Figure B.33: Experimental free energies of binding for Cyclophilin D against predicted free energies of binding as calculated by Alibay et al..[1] Experimental free energies were obtained from Grädler et al..[294] The darker and lighter shaded areas show 1 and 2 kcal mol^{-1} deviations from exact agreement, respectively. Error bars show 95 % confidence intervals, which were assumed to be 0.5 kcal mol^{-1} for experiment and calculated from the deviation between 5 replicate runs for the predictions. 95 % confidence intervals on statistics were calculated by bootstrapping with 10000 iterations of resampling.

Detailed results are summarised for our adaptive and non-adaptive runs in Table B.7, while the Boresch restraint parameters are shown in Table B.8. There were significant differences between the unsigned deviation from experiment and inter-replicate deviations between the results from Alibay et al. and our non-adaptive protocol (p=0.01 and 0.004, respectively, from Wilcoxon signed-rank tests). Specifically, our results showed a larger offset towards more negative free energies of binding, and larger inter-run deviations. The larger inter-run deviations may be partially explained by the much shorter run times of our non-adaptive protocol (5 ns per window) compared to those used by Alibay et al. (20 ns per window). However, given the tendency of inter-run uncertainties not to decrease with increased sampling time (Section 4.1.1), combined with the larger offset to experiment, this

Table B.6: Performance Metrics for Cyclophilin-D Free Energy Prediction Methods[a]

	Alibay	Non-adaptive	Adaptive	Adaptive vs Non-adaptive
r	0.89 (0.61, 0.98)	0.90 (0.68, 0.97)	0.86 (0.61, 0.95)	0.97 (0.82, 0.98)
r^2	0.79 (0.38, 0.95)	0.81 (0.46, 0.94)	0.75 (0.37, 0.91)	0.94 (0.68, 0.96)
MUE	2.38 (1.76, 3.04)	3.52 (2.61, 4.49)	2.97 (1.94, 4.15)	0.78 (0.67, 1.77)
RMSE	2.61 (1.98, 3.24)	3.84 (2.95, 4.79)	3.49 (2.31, 4.71)	1.04 (0.85, 2.19)
Spearman ρ	0.82 (0.23, 0.95)	0.77 (0.27, 0.93)	0.82 (0.28, 0.95)	0.92 (0.63, 0.97)
Kendall τ	0.72 (0.17, 0.89)	0.61 (0.17, 0.83)	0.72 (0.17, 0.83)	0.78 (0.44, 0.89)

[a] Performance metrics for the prediction of Cyclophilin-D experimental free energies using "Alibay" - the ABFE methodology of Alibay et al.;[1] "Non-adaptive" - our non-adaptive workflow; and "Adaptive" - our "optimal" adaptive workflow. "Adaptive vs Non-adaptive" shows metrics for the correlation between the adaptive and non-adaptive results. Uncertainties are given as 95 % confidence intervals, obtained by bootstrapping with 10000 iterations of resampling.

may also suggest systematically poorer sampling in our SOMD-based protocols compared to the GROMACS-based workflow of Alibay et al.. This could be due to many factors, such as the choice of soft-core potential, or the exact definition of the decoupled state.

Finally, Figures B.34 and 10 show the correlation of the non-adaptive results with experiment, and the non-adaptive results with the adaptive results, respectively.

Table B.7: Detailed Breakdown of Predicted $\Delta G^o_{\mathrm{Bind}}$ for Cyclophilin D[a]

	Bound Restrain	Bound Discharge	Bound Vanish	Free Discharge	Free Vanish	Restraint Correction	Symmetry Correction	Exp. $\Delta G^o_{\mathrm{Bind}}$
2 adaptive	1.63 ± 0.05	111.38 ± 1.43	2.16 ± 1.17	106.38 ± 0.19	-12.22 ± 0.19	-10.46	0.00	-9.06 ± 0.50
2 non-adaptive	1.67 ± 0.05	111.63 ± 0.76	1.76 ± 0.70	106.26 ± 0.08	-12.20 ± 0.34	-10.46	0.00	-9.06 ± 0.50
3 adaptive	2.89 ± 2.18	12.76 ± 0.30	-0.31 ± 0.91	10.04 ± 0.09	-9.64 ± 0.22	-9.87	0.00	-2.93 ± 0.50
3 non-adaptive	1.72 ± 0.11	12.39 ± 0.52	0.80 ± 0.99	9.99 ± 0.01	-9.68 ± 0.06	-9.87	0.00	-2.93 ± 0.50
4 adaptive	1.69 ± 0.08	109.43 ± 1.28	-3.48 ± 1.58	107.30 ± 0.16	-13.98 ± 0.12	-9.39	0.00	-2.90 ± 0.50
4 non-adaptive	2.14 ± 0.77	109.98 ± 1.32	-2.52 ± 0.78	107.25 ± 0.17	-13.98 ± 0.09	-9.39	0.00	-2.90 ± 0.50
8 adaptive	1.71 ± 0.40	-54.56 ± 0.38	4.45 ± 1.07	-57.24 ± 0.24	-7.90 ± 0.18	-9.43	0.00	-4.04 ± 0.50
8 non-adaptive	1.68 ± 0.29	-54.40 ± 0.72	4.19 ± 1.02	-57.36 ± 0.15	-7.94 ± 0.10	-9.43	0.00	-4.04 ± 0.50
14 adaptive	1.72 ± 0.08	92.41 ± 0.79	1.81 ± 2.14	88.20 ± 0.23	-17.50 ± 0.35	-10.86	0.00	-11.22 ± 0.50
14 non-adaptive	1.71 ± 0.02	92.34 ± 0.62	2.46 ± 1.86	88.15 ± 0.11	-17.95 ± 0.14	-10.86	0.00	-11.22 ± 0.50
16 adaptive	1.55 ± 0.07	87.80 ± 0.37	-0.62 ± 1.66	86.61 ± 0.64	-17.35 ± 0.25	-10.19	0.00	-8.42 ± 0.50
16 non-adaptive	1.58 ± 0.03	87.95 ± 0.36	0.51 ± 2.02	86.72 ± 0.53	-17.35 ± 0.31	-10.19	0.00	-8.42 ± 0.50
27 adaptive	1.71 ± 0.07	-22.53 ± 1.14	1.52 ± 1.70	-26.52 ± 0.14	-12.84 ± 0.36	-10.92	0.00	-7.57 ± 0.50
27 non-adaptive	1.75 ± 0.05	-22.33 ± 1.27	2.59 ± 0.47	-26.74 ± 0.15	-12.96 ± 0.22	-10.92	0.00	-7.57 ± 0.50
39 adaptive	1.81 ± 0.24	73.58 ± 1.33	2.29 ± 1.31	67.45 ± 0.13	-14.22 ± 0.28	-10.46	0.00	-8.43 ± 0.50
39 non-adaptive	1.96 ± 0.70	73.54 ± 1.24	1.58 ± 1.24	67.31 ± 0.02	-14.27 ± 0.17	-10.46	0.00	-8.43 ± 0.50
40 adaptive	1.69 ± 0.25	73.82 ± 0.82	4.58 ± 0.60	67.94 ± 0.14	-12.82 ± 0.19	-10.24	0.00	-8.08 ± 0.50
40 non-adaptive	1.74 ± 0.25	73.51 ± 0.38	4.27 ± 0.78	67.78 ± 0.11	-13.02 ± 0.29	-10.24	0.00	-8.08 ± 0.50

[a] All quantities in kcal mol^{-1}. Calculation uncertainties stated as 95 % t-based confidence intervals based on the variance of 5 replicate runs, assuming Gaussian distributions. "adaptive" refers to our "optimal" adaptive protocol, and "non-adaptive" to our non-adaptive procotol. Experimental results were taken from Grädler et al., who used surface plasmon resonance.[294] The experimental uncertainties were assumed to be 0.5 kcal mol^{-1}.

Table B.8: Parameters for Boresch restraints for the Cyclophilin D ligands, as labelled in Figure 3 of Clark et al..[274] K refers to a force constant and 0 denotes an equilibrium value.

	2	3	4	8	14	16	27	39	40
r1	1554	1632	1534	1535	1567	1567	1554	1558	1558
r2	1552	1630	1532	1533	1565	1565	1552	1556	1556
r3	1564	1645	1544	1545	1577	1577	1564	1568	1568
l1	13	13	11	15	20	8	4	10	7
l2	12	3	14	1	19	13	3	7	10
l3	14	6	16	4	21	25	7	12	19
r_0 /	5.19	6.12	4.22	4.96	4.63	6.37	3.68	5.89	5.77
θ_{A0} /	0.69	1.20	1.72	0.74	1.58	0.60	1.65	0.83	0.87
θ_{B0} / Rad	2.51	1.71	1.31	0.64	1.80	1.62	1.30	1.11	1.13
ϕ_{A0} / Rad	-0.24	-1.87	1.65	0.03	1.11	-0.39	1.56	0.00	-0.00
ϕ_{B0} / Rad	-0.85	0.28	-1.06	-0.93	-1.02	-0.54	-1.16	2.72	2.38
ϕ_{C0} / Rad	0.22	0.94	-2.61	-3.08	-2.51	0.15	-2.30	-3.05	3.12
k_r / kcal mol^{-1} $^{-2}$	19.86	9.46	8.38	17.86	18.68	16.80	21.18	18.32	19.32
$k_{\theta\mathrm{A}}$ / kcal mol^{-1} Rad^{-2}	214.50	109.66	34.20	128.18	92.20	190.46	73.02	174.74	130.04
$k_{\theta\mathrm{B}}$ / kcal mol^{-1} Rad^{-2}	103.74	34.98	30.84	61.74	128.28	110.76	64.32	81.08	90.62
$k_{\phi\mathrm{A}}$ / kcal mol^{-1} Rad^{-2}	91.72	244.78	114.02	52.02	193.68	69.42	176.98	119.12	85.94
$k_{\phi\mathrm{B}}$ / kcal mol^{-1} Rad^{-2}	25.70	102.60	69.22	13.84	144.82	33.40	216.74	77.78	91.02
$k_{\phi\mathrm{C}}$ / kcal mol^{-1} Rad^{-2}	33.22	60.16	37.46	10.04	90.32	84.82	70.06	73.10	47.98

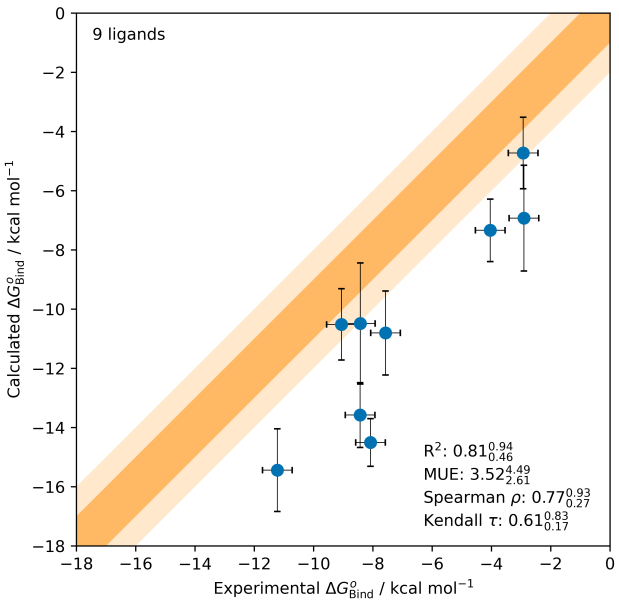


Figure B.34: Experimental free energies of binding for Cyclophilin D against predicted free energies of binding obtained using the non-adaptive protocol. Experimental free energies were obtained from Grädler et al..[294] The darker and lighter shaded areas show 1 and 2 kcal mol^{-1} deviations from exact agreement, respectively. Error bars show 95 % confidence intervals, which were assumed to be 0.5 kcal mol^{-1} for experiment and calculated from the deviation between 5 replicate runs for the predictions. 95 % confidence intervals on statistics were calculated by bootstrapping with 10000 iterations of resampling.

B.22 Details of Cyclophilin-D Adaptive Runs

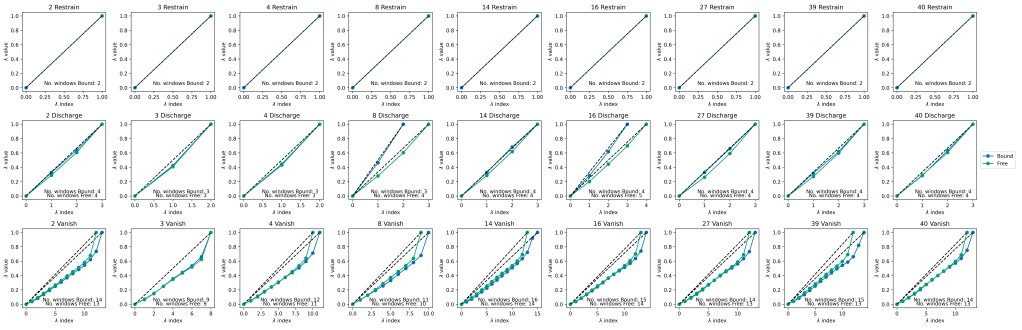


Figure B.35: λ values selected for each of the Cyclophilin D ligands using a thermodynamic speed of 2 kcal mol^{-1}.

Table B.9: Allocation of Sampling Time Between Cyclophilin D Ligands using Adaptive and Non-Adaptive Protocols[a]

	adaptive	non-adaptive
Ligand 2	104	161
Ligand 3	146	137
Ligand 4	216	154
Ligand 8	65	174
Ligand 14	124	164
Ligand 16	77	189
Ligand 27	123	180
Ligand 39	161	170
Ligand 40	83	177
Total	1098	1505

[a] All times given in GPU-hours.

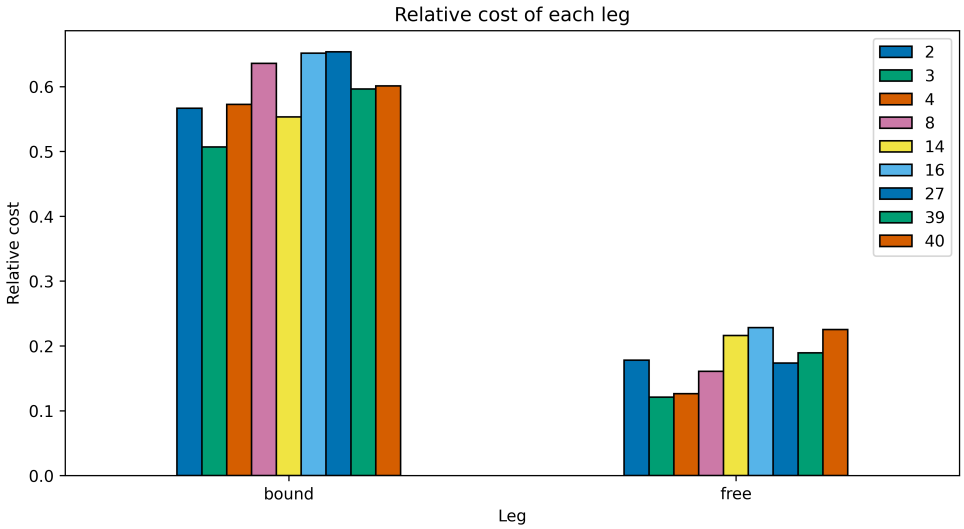


Figure B.36: Relative computational costs of each leg for the Cyclophilin D systems, calculated relative to the MIF180 bound leg. Calculated based on the average time taken to run a ns of simulation on a single GPU. The absolute cost of the MIF180 bound leg was 0.21 GPU hours / ns. All computational costs were assessed running on NVIDIA GeForce RTX2080 SUPER GPUs.

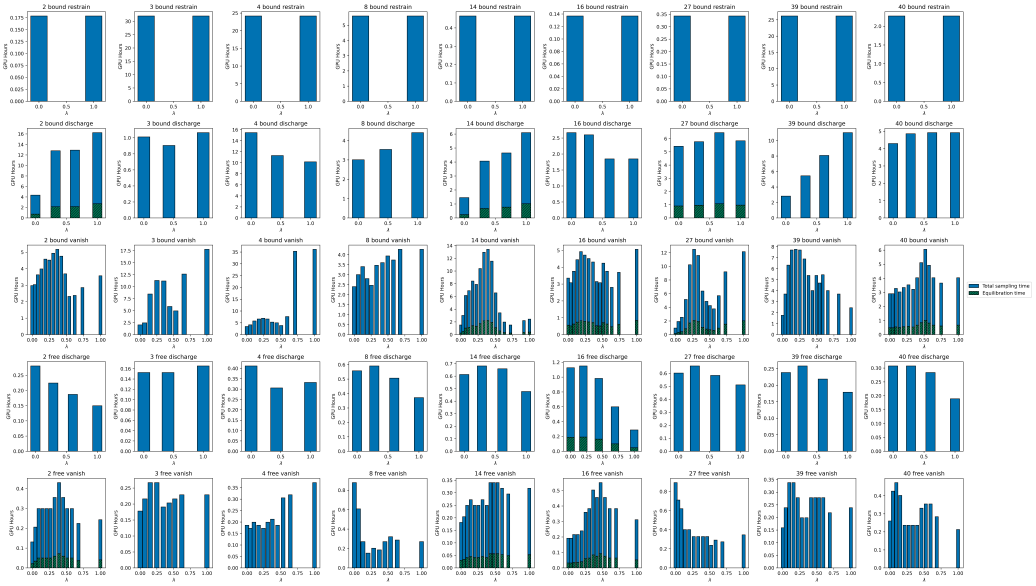


Figure B.37: Per-window breakdown of sampling times allocated to the Cyclophilin D systems with the "Optimised" adaptive protocol.

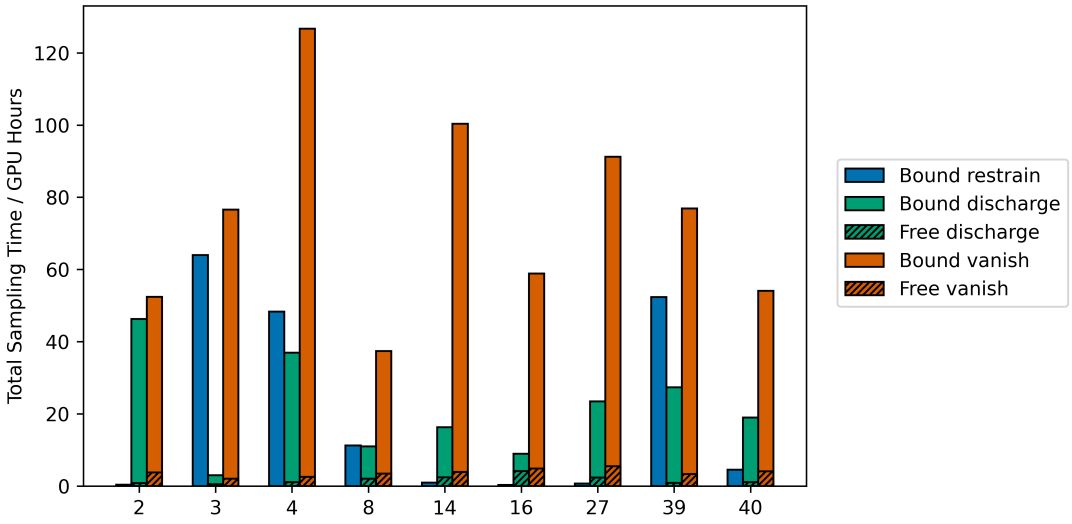


Figure B.38: Per-stage breakdown of sampling times allocated to the Cyclophilin D systems with the "Optimised" adaptive protocol.

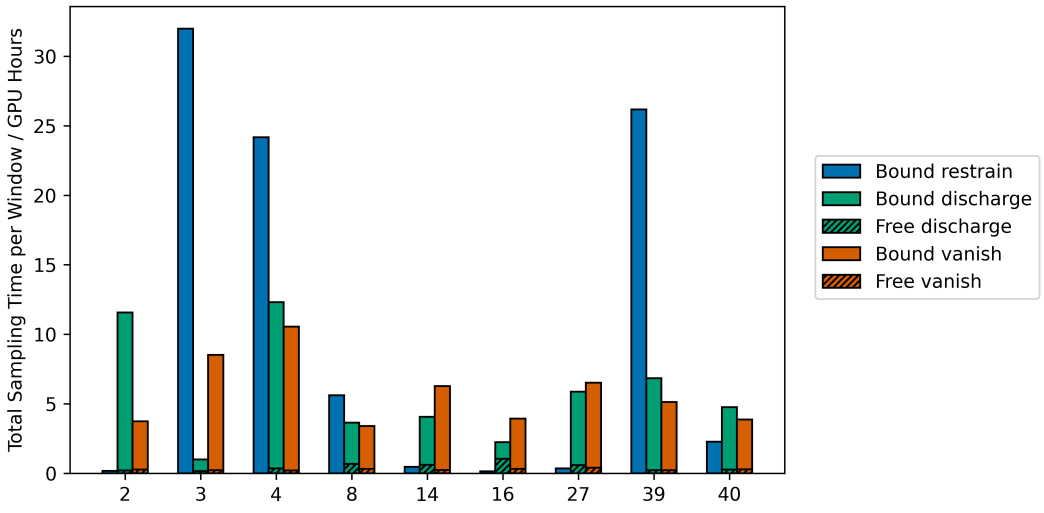


Figure B.39: Per-stage breakdown of the per-window sampling times allocated to the Cyclophilin D systems with the "Optimised" adaptive protocol.

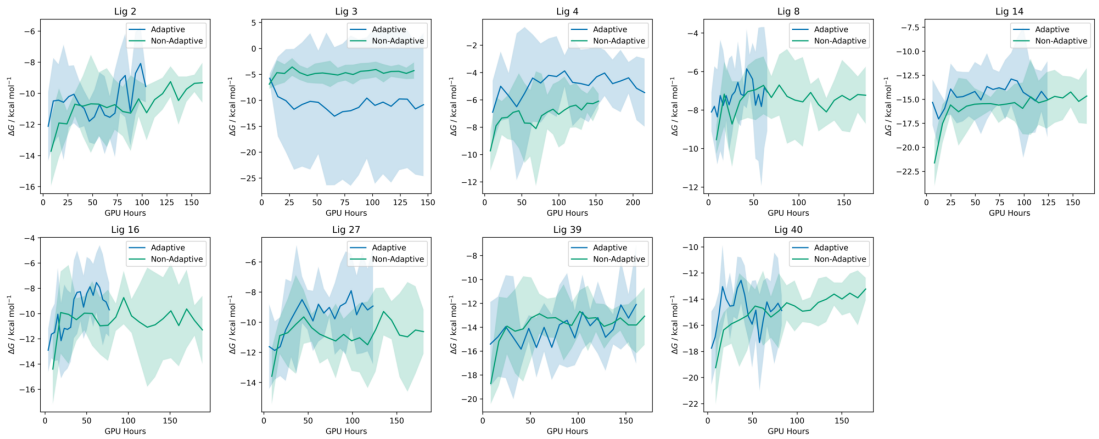


Figure B.40: Estimated ΔG against sampling time for the Cyclophilin D ligands. All data were split into 20 equal blocks and MBAR was run on each block. Shaded areas show 95 % t-based confidence intervals. The apparent large deviation for ligand 3 adaptive runs occurred due to this weak binder flipping in the binding site during the $\lambda = 0$ restrain-stage window for some runs. This could equally have occurred during the non-adaptive runs. Despite the large deviation shown above, running MBAR on all the data produced a reasonable result much closer to the expected ≈ 1.23 kcal mol^{-1}, as shown in Table B.7.

Appendix C

Robust Automated Truncation Point Selection

C.1 Decomposition of RMSE into Bias and Variance Terms

It is well known that the overall RMSE can be decomposed into bias and variance terms.[108] For clarity, the full derivation is given here in our notation:

$$\mathrm{RMSE}_{\mathrm{Trajs}} = \sqrt{\left\langle (\langle A\rangle_{[n_0,N]} - \langle A\rangle_\pi)^2 \right\rangle_{\mathrm{Trajs}}} \tag{C.1}$$

$$= \sqrt{\left\langle \langle A\rangle^2_{[n_0,N]} - 2\langle A\rangle_{[n_0,N]}\langle A\rangle_\pi + \langle A\rangle^2_\pi \right\rangle_{\mathrm{Trajs}}} \tag{C.2}$$

$$= \sqrt{\left\langle \langle A\rangle^2_{[n_0,N]} \right\rangle_{\mathrm{Trajs}} - 2\left\langle \langle A\rangle_{[n_0,N]} \right\rangle_{\mathrm{Trajs}} \langle A\rangle_\pi + \langle A\rangle^2_\pi} \tag{C.3}$$

$$= \sqrt{\left\langle \langle A\rangle^2_{[n_0,N]} \right\rangle_{\mathrm{Trajs}} - \left\langle \langle A\rangle_{[n_0,N]} \right\rangle^2_{\mathrm{Trajs}} + \left\langle \langle A\rangle_{[n_0,N]} \right\rangle^2_{\mathrm{Trajs}} - 2\left\langle \langle A\rangle_{[n_0,N]} \right\rangle_{\mathrm{Trajs}} \langle A\rangle_\pi + \langle A\rangle^2_\pi} \tag{C.4}$$

$$= \sqrt{\left\langle \langle A\rangle^2_{[n_0,N]} - 2\langle\langle A\rangle_{[n_0,N]}\rangle_{\mathrm{Trajs}}\langle A\rangle_{[n_0,N]} + \langle\langle A\rangle_{[n_0,N]}\rangle^2_{\mathrm{Trajs.}} \right\rangle_{\mathrm{Trajs}} + (\langle\langle A\rangle_{[n_0,N]}\rangle_{\mathrm{Trajs}} - \langle A\rangle_\pi)^2} \tag{C.5}$$

$$= \sqrt{\left\langle (\langle A\rangle_{[n_0,N]} - \langle\langle A\rangle_{[n_0,N]}\rangle_{\mathrm{Trajs.}})^2 \right\rangle_{\mathrm{Trajs}} + (\langle\langle A\rangle_{[n_0,N]}\rangle_{\mathrm{Trajs}} - \langle A\rangle_\pi)^2} \tag{C.6}$$

$$= \sqrt{\mathrm{Var}_{\mathrm{Trajs}} + \mathrm{Bias}^2_{\mathrm{Trajs}}}. \tag{C.7}$$

C.2 Modelling the Bound Vanish Stages of the Absolute Binding Free Energy Data

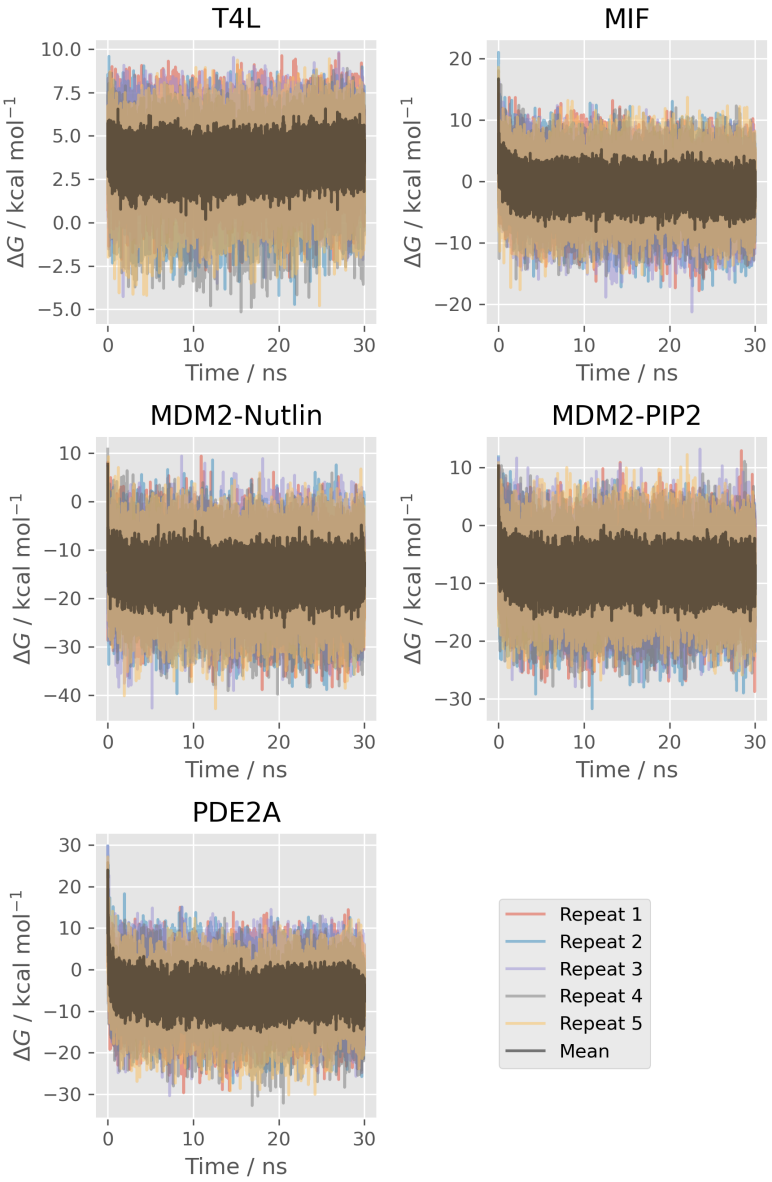


Figure C.1: Time series of sampled free energy changes for all systems for bound vanish stages of the absolute binding free energy calculations. ΔG estimates obtained by integrating the gradients of the free energy over all windows using the trapezoidal rule.

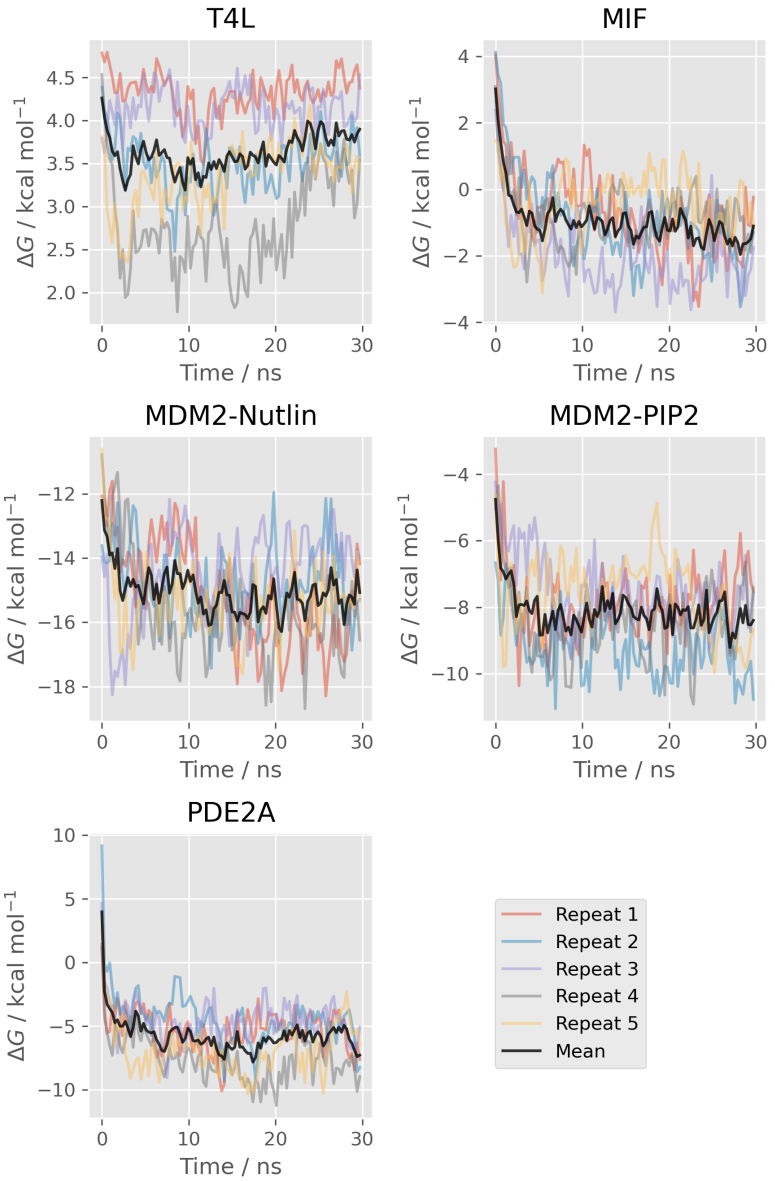


Figure C.2: Block-averaged time series of sampled free energy changes for all systems for bound vanish stages of the absolute binding free energy calculations. Block averaging was performed using 100 blocks to more clearly show trends. ΔG estimates obtained by integrating the gradients of the free energy over all windows using the trapezoidal rule.

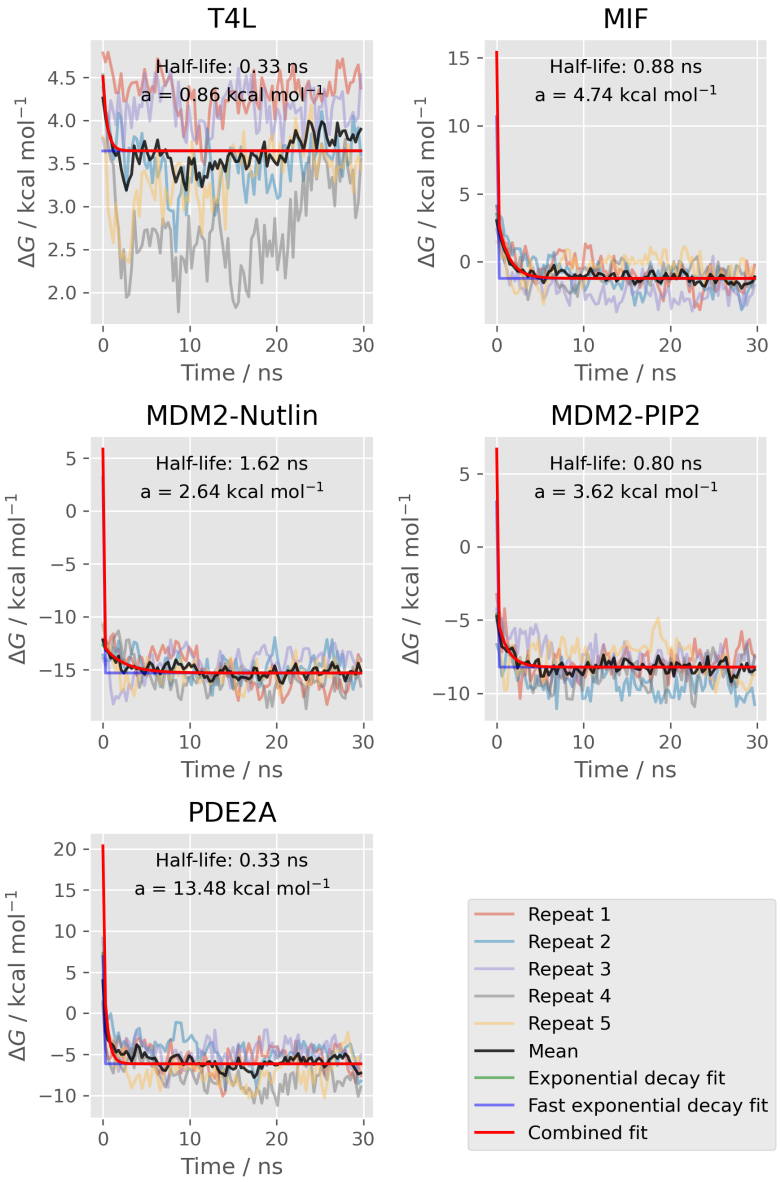


Figure C.3: Exponential fits to the bound vanish leg time series of the absolute binding free energy calculations. The overall fit is composed of a "fast" and an initial exponential fit. Time series block averaged with 100 blocks to more clearly show trends.

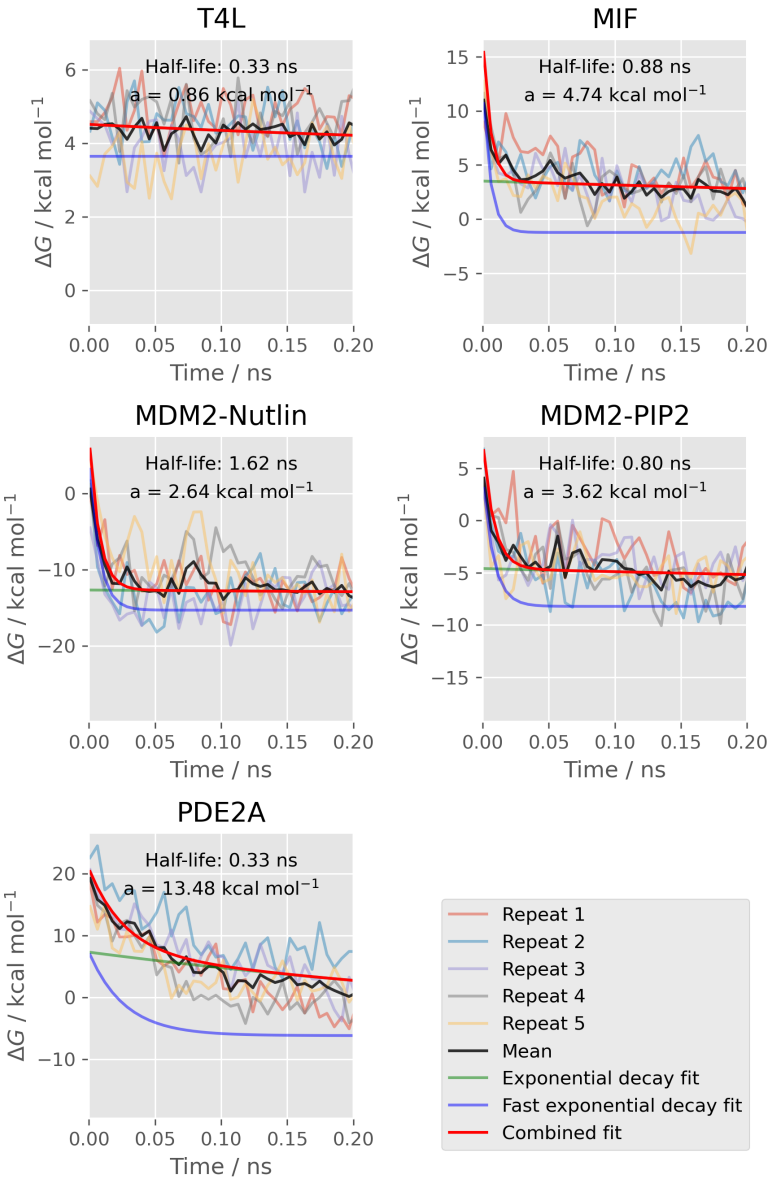


Figure C.4: Exponential fits to the bound vanish leg time series of the absolute binding free energy calculations, zoomed in to show only the first 0.2 ns. In this region, the additional "fast" exponential fit is required to model the trend. Time series block averaged with 100 blocks to more clearly show trends.

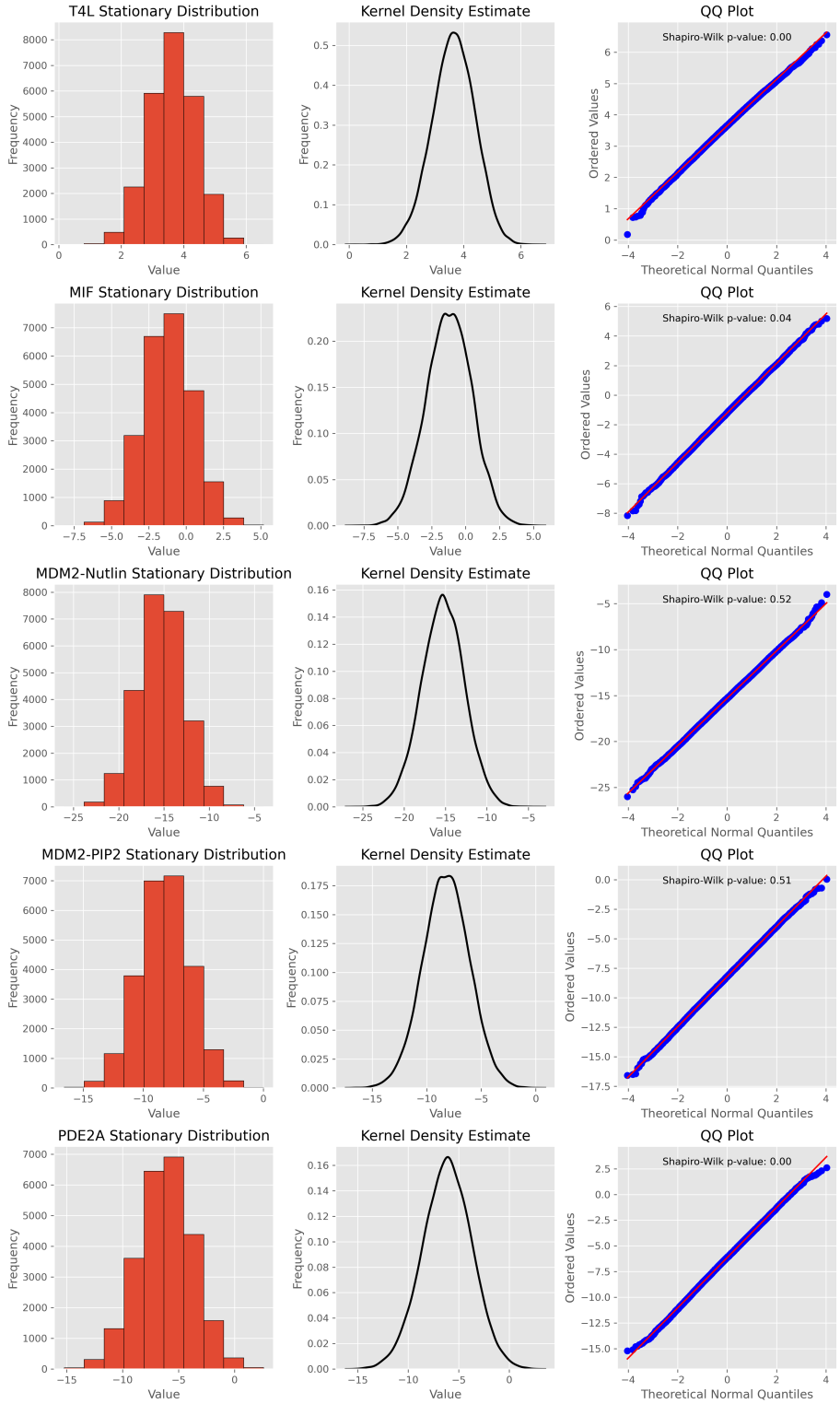


Figure C.5: Histograms, kernel density estimates, and QQ plots of the distributions of ΔG estimates obtained over the last 20 ns of the bound vanish stages of the absolute binding free energy calculations. For some systems, the Shapiro-Wilk test showed significant evidence for non-normality of the distributions. However, the deviations were small, meaning that modelling the distributions as normal was reasonable.

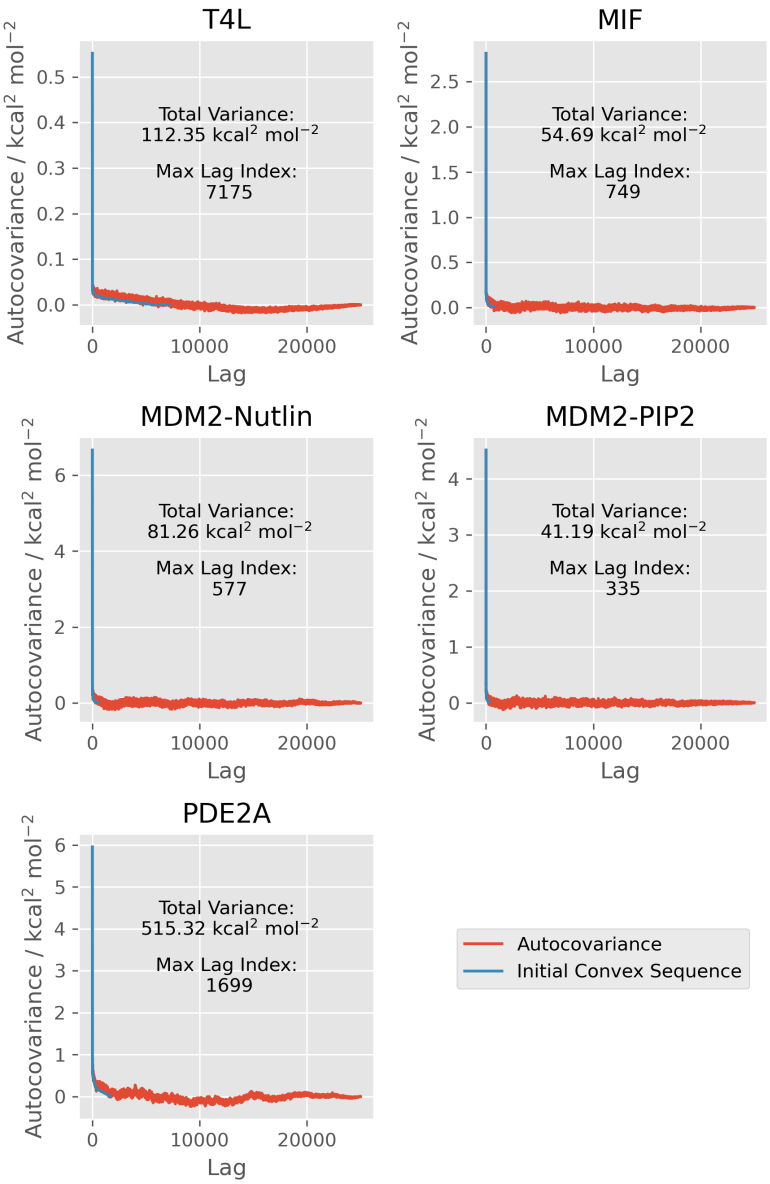


Figure C.6: Estimated ("Autocovariance") and fitted autocovariance ("Initial Convex Sequence") functions obtained from the final 20 ns of the bound vanish stages of the absolute binding free energy calculations.

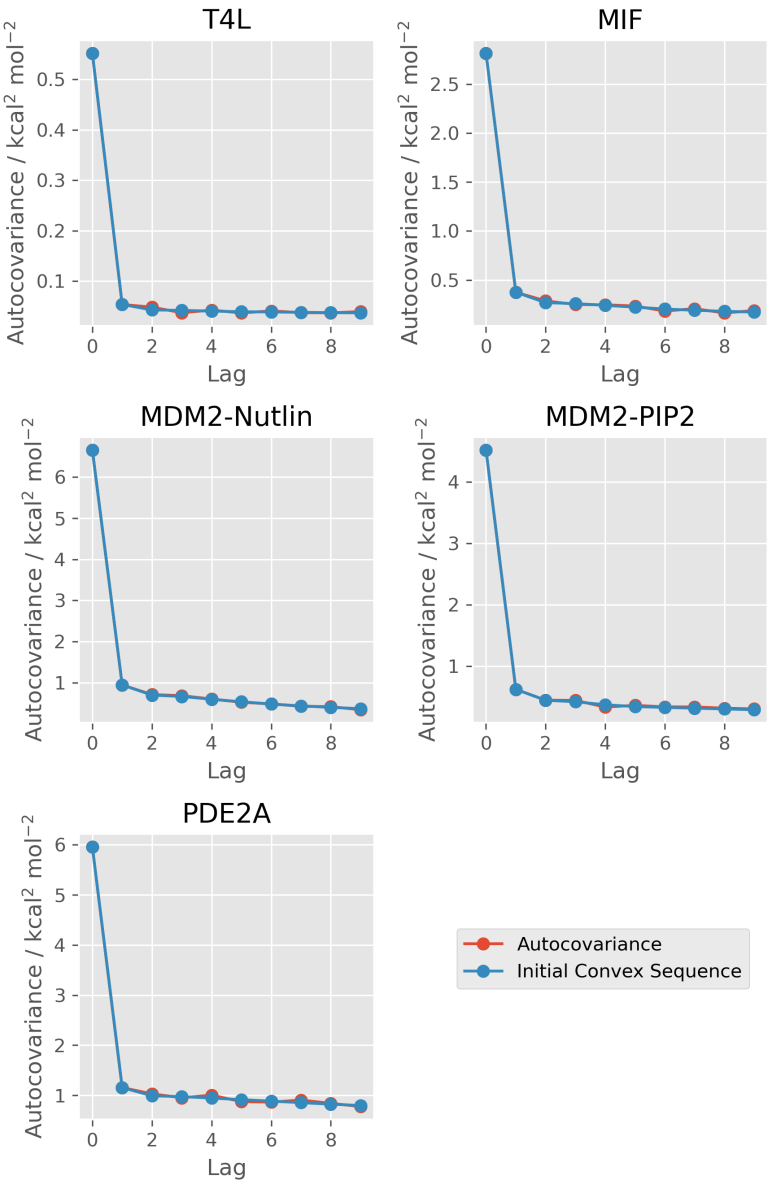


Figure C.7: Early lag time estimated ("Autocovariance") and fitted autocovariance ("Initial Convex Sequence") functions obtained from the final 20 ns of the bound vanish stages of the absolute binding free energy calculations.

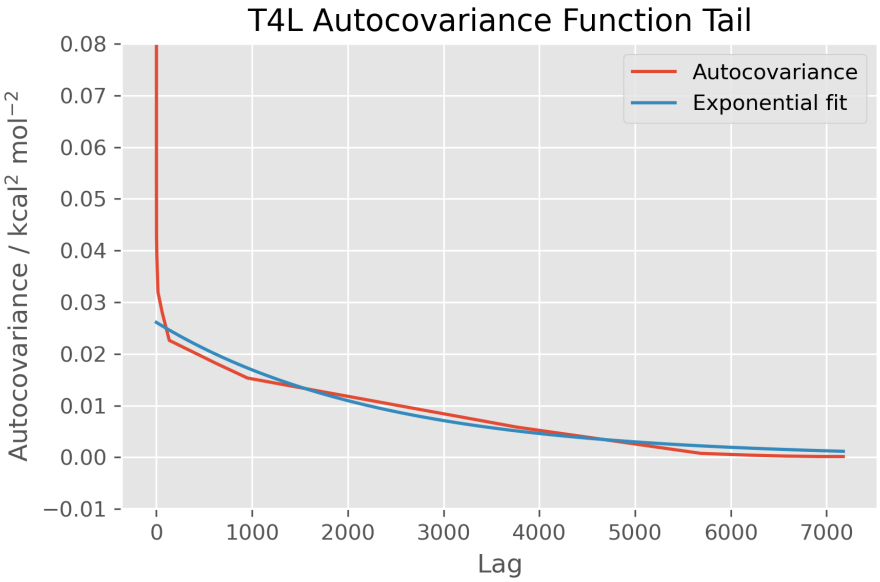


Figure C.8: The long tail of the estimated T4L autocovariance function is reasonably well described by a relatively slow exponential decay. This is preceded by relatively fast decay at early lag times.

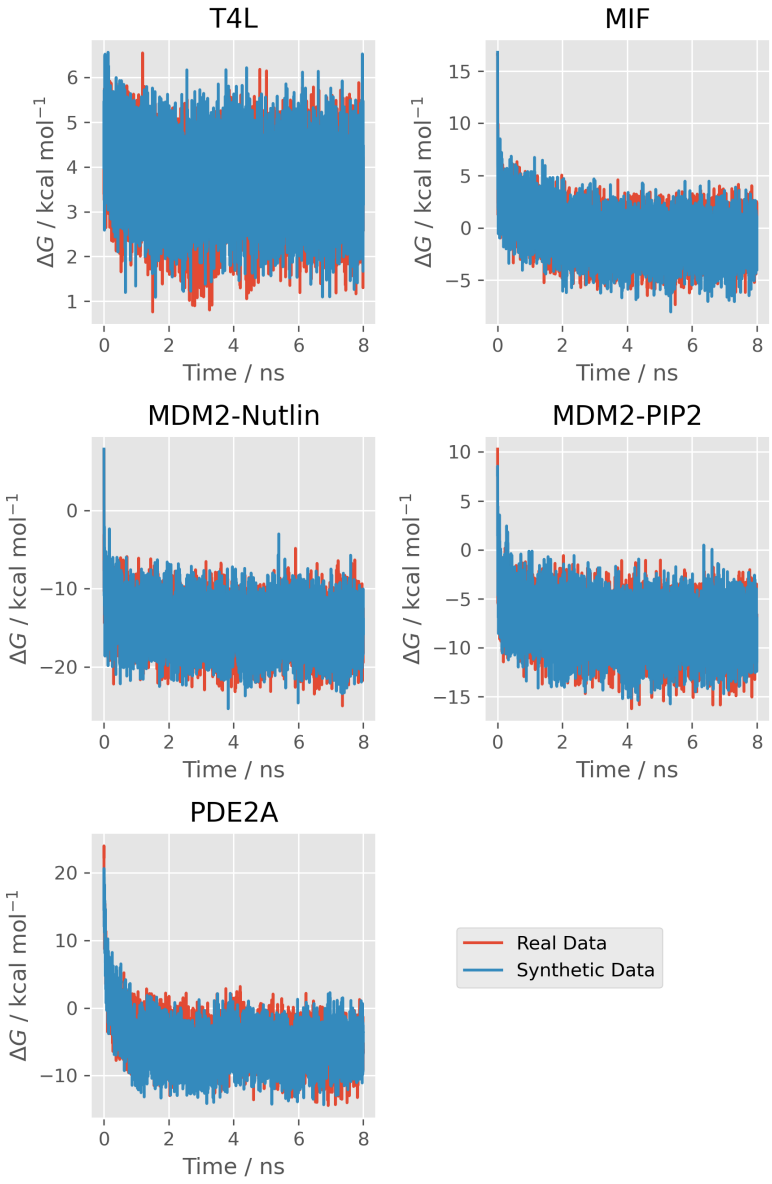


Figure C.9: Random examples of synthetic ΔG time series against the time series from simulation they were fit to.

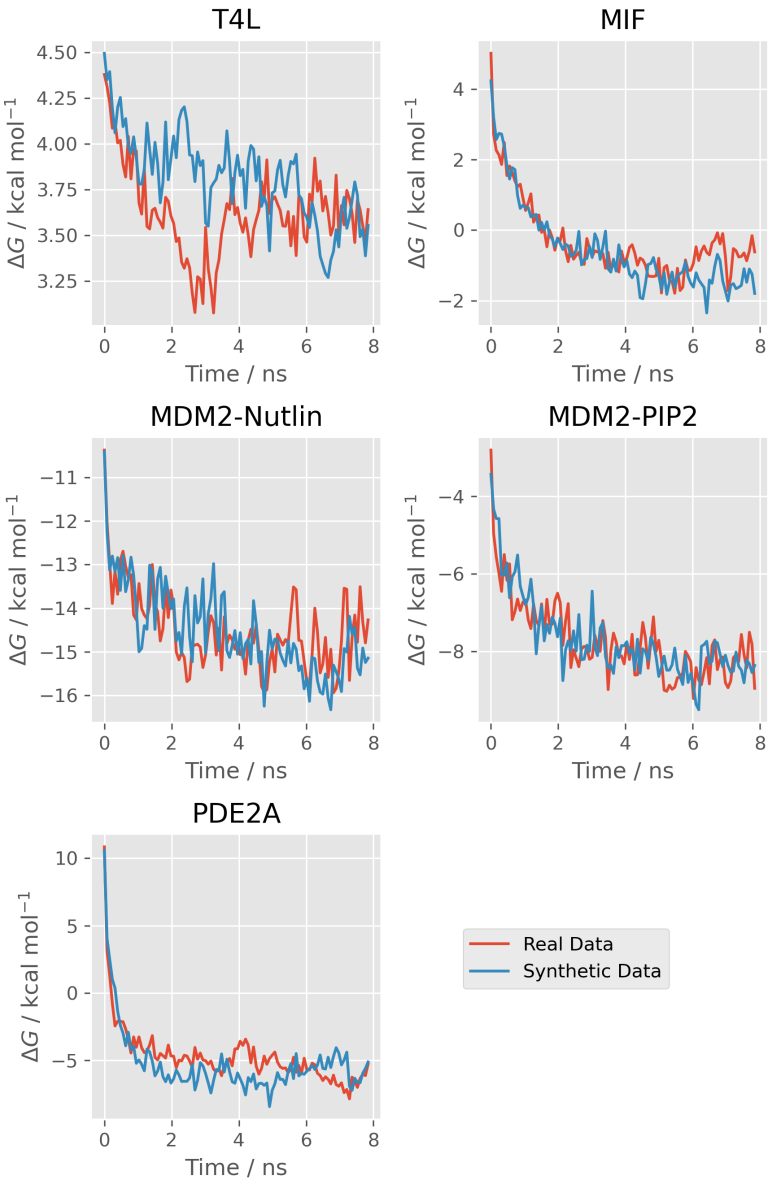


Figure C.10: Random examples of synthetic ΔG time series against the time series from simulation they were fit to. Block averaged with 100 blocks to more clearly show trends and features.

Table C.1: Model Parameters Fitted to Free Vanish Stages of Absolute Binding Free Energy Calculations[a]

	Half-life (ns)	a (kcal mol^{-1})	Fast Half-life (ns)	Fast a (kcal mol^{-1})	Total Variance (kcal^2 mol^{-2})	Max Lag Index
T4L	∞	0.00	∞	0.00	1.5	13
PDE2A	∞	0.00	∞	0.00	110	865

[a] Total variance refers to the total variance of the mean, obtained by summing the autocovariance series from - max lag index to + max lag index, where the series and maximum lag indices were estimated according to Geyer's initial convex sequence rules. "a" refers to the pre-exponential factors.

C.3 Modelling the Free Vanish Stages of the Absolute Binding Free Energy Data

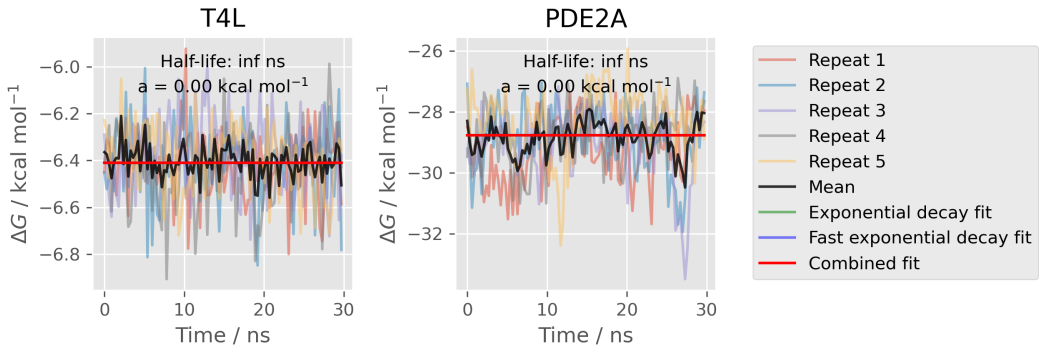


Figure C.11: Exponential fits to the free vanish leg time series of the absolute binding free energy calculations. Time series block averaged with 100 blocks to more clearly show trends.

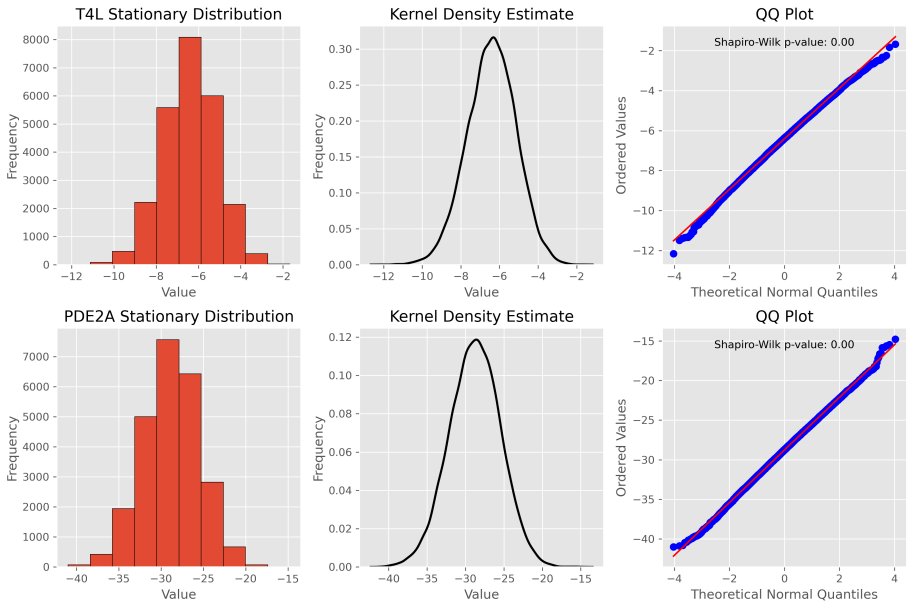


Figure C.12: Histograms, kernel density estimates, and QQ plots of the distributions of ΔG estimates obtained over the last 20 ns of the free vanish stages of the absolute binding free energy calculations. For both systems, the Shapiro-Wilk test showed significant evidence for non-normality of the distributions. However, the deviations were small, meaning that modelling the distributions as normal was reasonable.

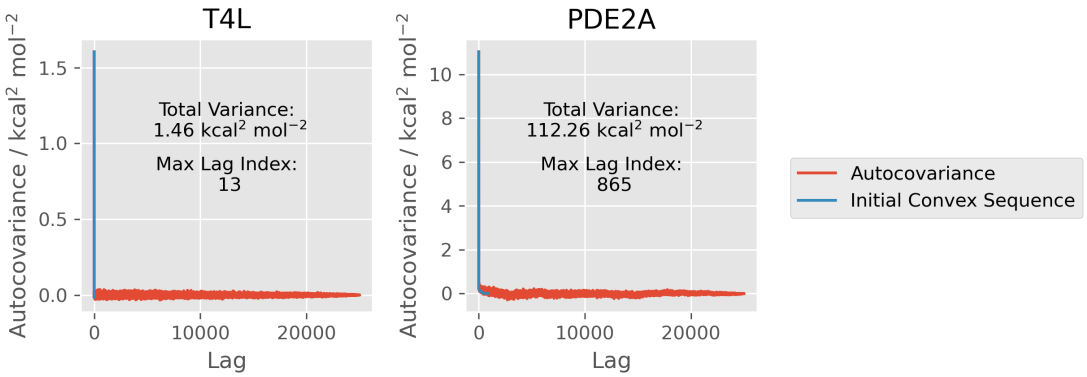


Figure C.13: Estimated ("Autocovariance") and fitted autocovariance ("Initial Convex Sequence") functions obtained from the final 20 ns of the free vanish stages of the absolute binding free energy calculations.

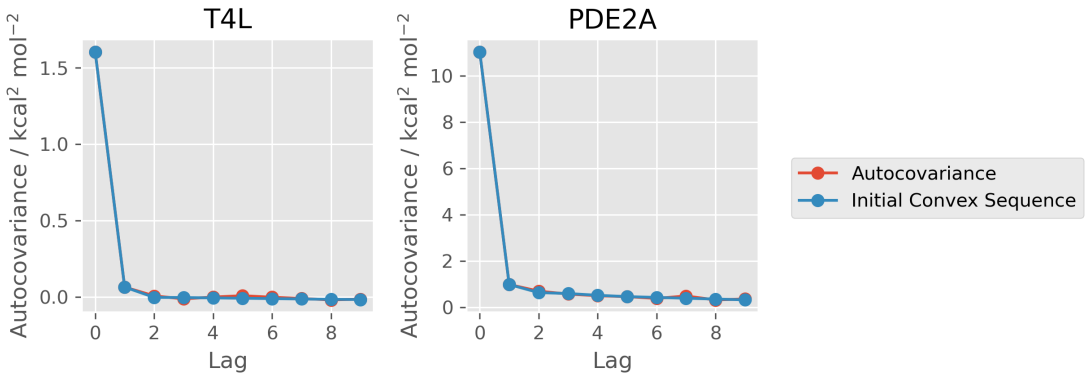


Figure C.14: Early lag time estimated ("Autocovariance") and fitted autocovariance ("Initial Convex Sequence") functions obtained from the final 20 ns of the free vanish stages of the absolute binding free energy calculations.

C.4 Performance of Truncation Heuristics on Free Vanish Leg Time Series

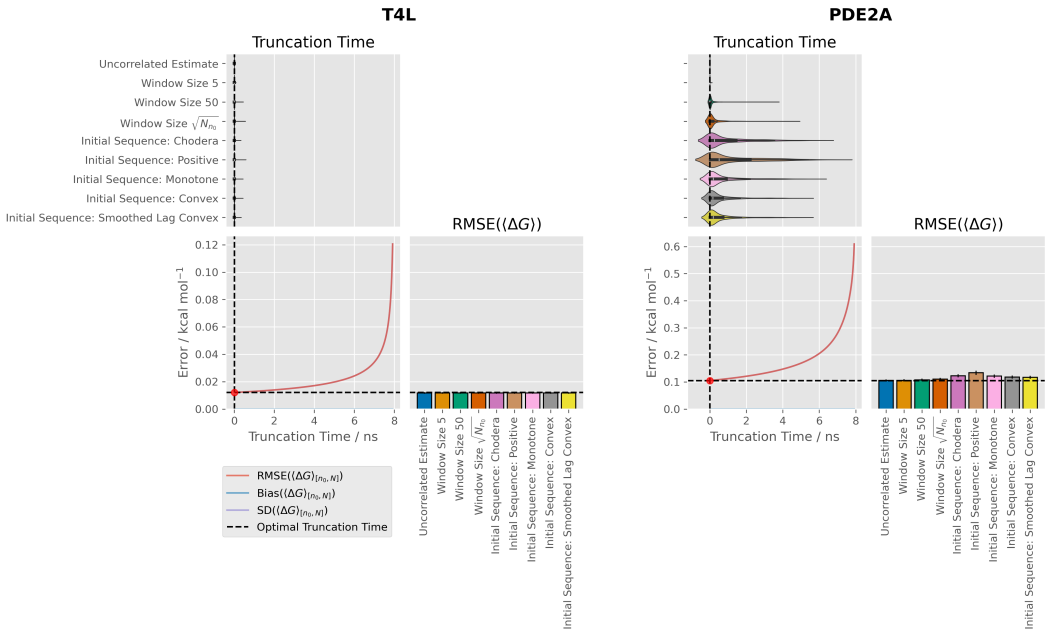


Figure C.15: Discard times, RMSEs, and underlying time series properties for the free vanish stage time series for T4L and PDE2A. The top panels show kernel density estimates of the distributions of times discarded with each method. The bottom left panels show the RMSEs which would be obtained with an infinitely large ensemble of synthetic time series with fixed truncation points. The red dot indicates the optimum fixed-time truncation point. This is at 0 as there are no biases. The right panels show the RMSEs obtained with each generalised MSER method.

Table C.2: Ensemble RMSEs for all Generalised MSER Heuristics for the Free Vanish Data[a]

Method	T4L	PDE2A
Uncorrelated Estimate	$0.0117^{0.0123}_{0.0112}$	$0.105^{0.110}_{0.101}$
Window Size 5	$0.0117^{0.0123}_{0.0112}$	$0.105^{0.110}_{0.101}$
Window Size 50	$0.0118^{0.0123}_{0.0112}$	$0.107^{0.112}_{0.103}$
Window Size $\sqrt{N_{n_0}}$	$0.0118^{0.0123}_{0.0112}$	$0.110^{0.115}_{0.105}$
Initial Sequence: Chodera	$0.0118^{0.0123}_{0.0112}$	$0.122^{0.128}_{0.117}$
Initial Sequence: Positive	$0.0118^{0.0123}_{0.0112}$	$0.134^{0.142}_{0.126}$
Initial Sequence: Monotone	$0.0118^{0.0123}_{0.0112}$	$0.121^{0.128}_{0.116}$
Initial Sequence: Convex	$0.0118^{0.0123}_{0.0112}$	$0.118^{0.123}_{0.112}$
Initial Sequence: Smoothed Lag Convex	$0.0118^{0.0123}_{0.0112}$	$0.117^{0.122}_{0.111}$

[a] All values in kcal mol^{-1}. Uncertainties are 95 % confidence intervals obtained by bootstrapping over 10000 iterations with replacement.

C.5 Investigation of Adaptive Integration Scheme with the Max ESS Heuristic on PDE2a Free Vanish Leg Time Series

Figure C.16: Kernel density estimate plot of truncation times selected with Chodera's maximum effective sample size heuristic (as implemented in PyMBAR's timeseries module) with and without the adaptive integration scheme described by Chodera et al.
.[108,131,302] These synthetic trajectories have no bias and hence the optimum truncation time is at index 0. The heuristics were tested on the first 100 synthetic trajectories. The use of the adaptive integration scheme produced more erroneously late truncation times. The final index shown corresponds to 8 ns of simulation data.

Table C.3: Ensemble RMSEs for all Generalised MSER Heuristics for the "Noisy" Bound Vanish Data[a]

Method	T4L	MIF	MDM2-Nutlin	MDM2-PIP2	PDE2A
Uncorrelated Estimate	$0.218^{0.228}_{0.207}$	$0.715^{0.726}_{0.703}$	$0.771^{0.783}_{0.759}$	$0.536^{0.544}_{0.527}$	$0.634^{0.659}_{0.609}$
Window Size 5	$0.219^{0.229}_{0.209}$	$0.495^{0.508}_{0.483}$	$0.757^{0.769}_{0.745}$	$0.464^{0.473}_{0.454}$	$0.573^{0.597}_{0.550}$
Window Size 50	$0.236^{0.249}_{0.224}$	$0.333^{0.345}_{0.321}$	$0.663^{0.677}_{0.649}$	$0.307^{0.317}_{0.297}$	$0.584^{0.608}_{0.559}$
Window Size $\sqrt{N_{n_0}}$	$0.260^{0.273}_{0.248}$	$0.293^{0.305}_{0.281}$	$0.578^{0.593}_{0.564}$	$0.262^{0.271}_{0.252}$	$0.639^{0.671}_{0.608}$
Initial Sequence: Chodera	$0.293^{0.306}_{0.281}$	$0.318^{0.335}_{0.302}$	$0.491^{0.507}_{0.475}$	$0.251^{0.261}_{0.242}$	$0.980^{1.028}_{0.932}$
Initial Sequence: Positive	$0.297^{0.310}_{0.284}$	$0.326^{0.343}_{0.309}$	$0.481^{0.497}_{0.465}$	$0.248^{0.262}_{0.235}$	$0.948^{0.995}_{0.902}$
Initial Sequence: Monotone	$0.292^{0.305}_{0.279}$	$0.317^{0.334}_{0.300}$	$0.492^{0.508}_{0.477}$	$0.243^{0.253}_{0.234}$	$0.956^{1.004}_{0.908}$
Initial Sequence: Convex	$0.287^{0.300}_{0.275}$	$0.308^{0.324}_{0.292}$	$0.504^{0.519}_{0.489}$	$0.244^{0.253}_{0.234}$	$0.949^{0.998}_{0.902}$
Initial Sequence: Smoothed Lag Convex	$0.286^{0.299}_{0.273}$	$0.308^{0.324}_{0.293}$	$0.502^{0.517}_{0.487}$	$0.244^{0.254}_{0.235}$	$0.951^{1.000}_{0.904}$

[a] All values in kcal mol^{-1}. Uncertainties are 95 % confidence intervals obtained by bootstrapping over 10000 iterations with replacement.

Table C.4: Ensemble RMSEs for all Generalised MSER Heuristics for the "Subsampled" Bound Vanish Data[a]

Method	T4L	MIF	MDM2-Nutlin	MDM2-PIP2	PDE2A
Uncorrelated Estimate	$0.128^{0.134}_{0.123}$	$0.454^{0.468}_{0.439}$	$0.732^{0.750}_{0.714}$	$0.428^{0.442}_{0.415}$	$0.388^{0.405}_{0.371}$
Window Size 5	$0.132^{0.137}_{0.126}$	$0.346^{0.359}_{0.333}$	$0.657^{0.677}_{0.637}$	$0.363^{0.377}_{0.348}$	$0.386^{0.403}_{0.370}$
Window Size 50	$0.133^{0.139}_{0.128}$	$0.325^{0.337}_{0.313}$	$0.600^{0.619}_{0.581}$	$0.353^{0.367}_{0.339}$	$0.388^{0.406}_{0.370}$
Window Size $\sqrt{N_{n_0}}$	$0.137^{0.144}_{0.131}$	$0.328^{0.340}_{0.315}$	$0.626^{0.646}_{0.605}$	$0.351^{0.366}_{0.337}$	$0.405^{0.434}_{0.380}$
Initial Sequence: Chodera	$0.136^{0.142}_{0.129}$	$0.356^{0.369}_{0.342}$	$0.662^{0.683}_{0.642}$	$0.371^{0.386}_{0.357}$	$0.411^{0.440}_{0.387}$
Initial Sequence: Positive	$0.149^{0.157}_{0.142}$	$0.317^{0.329}_{0.304}$	$0.589^{0.610}_{0.568}$	$0.349^{0.365}_{0.334}$	$0.445^{0.476}_{0.416}$
Initial Sequence: Monotone	$0.138^{0.145}_{0.132}$	$0.357^{0.370}_{0.343}$	$0.659^{0.679}_{0.639}$	$0.381^{0.395}_{0.367}$	$0.417^{0.446}_{0.391}$
Initial Sequence: Convex	$0.138^{0.144}_{0.131}$	$0.358^{0.371}_{0.345}$	$0.662^{0.682}_{0.642}$	$0.379^{0.393}_{0.365}$	$0.414^{0.444}_{0.388}$
Initial Sequence: Smoothed Lag Convex	$0.139^{0.145}_{0.132}$	$0.355^{0.368}_{0.342}$	$0.658^{0.678}_{0.638}$	$0.378^{0.392}_{0.364}$	$0.416^{0.445}_{0.390}$

[a] All values in kcal mol^{-1}. Uncertainties are 95 % confidence intervals obtained by bootstrapping over 10000 iterations with replacement.

C.6 Performance of Truncation Heuristics on All Bound Vanish Leg Time Series Ensembles

Note that in some cases, the relatively low RMSE of the "Window Size 50" method on the block averaged data results from the large block size preventing late truncation when there are few data points (e.g. a window size of 50 requires more than 50 points, which is half of the time series length).

Table C.5: Ensemble RMSEs for all Generalised MSER Heuristics for the "Block Averaged" Bound Vanish Data[a]

Method	T4L	MIF	MDM2-Nutlin	MDM2-PIP2	PDE2A
Uncorrelated Estimate	$0.116^{0.121}_{0.110}$	$0.162^{0.168}_{0.156}$	$0.395^{0.403}_{0.386}$	$0.144^{0.149}_{0.140}$	$0.280^{0.295}_{0.266}$
Window Size 5	$0.138^{0.145}_{0.132}$	$0.162^{0.169}_{0.155}$	$0.306^{0.316}_{0.296}$	$0.128^{0.134}_{0.122}$	$0.391^{0.412}_{0.370}$
Window Size 50	$0.116^{0.121}_{0.111}$	$0.133^{0.138}_{0.128}$	$0.297^{0.304}_{0.289}$	$0.113^{0.118}_{0.108}$	$0.289^{0.301}_{0.276}$
Window Size $\sqrt{N_{n_0}}$	$0.142^{0.148}_{0.136}$	$0.161^{0.168}_{0.154}$	$0.292^{0.302}_{0.283}$	$0.125^{0.132}_{0.118}$	$0.417^{0.438}_{0.395}$
Initial Sequence: Chodera	$0.143^{0.149}_{0.137}$	$0.167^{0.174}_{0.159}$	$0.298^{0.308}_{0.288}$	$0.128^{0.135}_{0.122}$	$0.430^{0.451}_{0.409}$
Initial Sequence: Positive	$0.146^{0.152}_{0.140}$	$0.172^{0.180}_{0.164}$	$0.282^{0.293}_{0.272}$	$0.131^{0.139}_{0.124}$	$0.454^{0.475}_{0.433}$
Initial Sequence: Monotone	$0.145^{0.151}_{0.139}$	$0.167^{0.175}_{0.160}$	$0.298^{0.308}_{0.288}$	$0.131^{0.138}_{0.125}$	$0.444^{0.465}_{0.423}$
Initial Sequence: Convex	$0.144^{0.151}_{0.138}$	$0.167^{0.175}_{0.159}$	$0.299^{0.309}_{0.289}$	$0.131^{0.137}_{0.124}$	$0.443^{0.464}_{0.422}$
Initial Sequence: Smoothed Lag Convex	$0.145^{0.151}_{0.138}$	$0.167^{0.175}_{0.159}$	$0.294^{0.304}_{0.284}$	$0.130^{0.137}_{0.123}$	$0.443^{0.464}_{0.422}$

[a] All values in kcal mol^{-1}. Uncertainties are 95 % confidence intervals obtained by bootstrapping over 10000 iterations with replacement.

Table C.6: Ensemble RMSEs for all Generalised MSER Heuristics for the "Short" Bound Vanish Data[a]

Method	T4L	MIF	MDM2-Nutlin	MDM2-PIP2	PDE2A
Uncorrelated Estimate	$0.727^{0.738}_{0.716}$	$4.397^{4.418}_{4.377}$	$2.637^{2.664}_{2.609}$	$3.393^{3.416}_{3.370}$	$10.594^{10.650}_{10.538}$
Window Size 5	$0.719^{0.730}_{0.708}$	$4.347^{4.368}_{4.326}$	$2.578^{2.606}_{2.549}$	$3.328^{3.352}_{3.304}$	$9.970^{10.033}_{9.910}$
Window Size 50	$0.686^{0.698}_{0.674}$	$4.251^{4.276}_{4.225}$	$2.554^{2.589}_{2.521}$	$3.233^{3.262}_{3.205}$	$9.586^{9.645}_{9.526}$
Window Size $\sqrt{N_{n_0}}$	$0.703^{0.714}_{0.691}$	$4.290^{4.312}_{4.267}$	$2.557^{2.588}_{2.527}$	$3.269^{3.295}_{3.243}$	$9.669^{9.731}_{9.606}$
Initial Sequence: Chodera	$0.710^{0.722}_{0.699}$	$4.301^{4.324}_{4.278}$	$2.566^{2.598}_{2.534}$	$3.283^{3.310}_{3.257}$	$9.671^{9.734}_{9.609}$
Initial Sequence: Positive	$0.694^{0.706}_{0.683}$	$4.267^{4.292}_{4.242}$	$2.559^{2.592}_{2.525}$	$3.245^{3.274}_{3.216}$	$9.534^{9.597}_{9.472}$
Initial Sequence: Monotone	$0.705^{0.717}_{0.694}$	$4.286^{4.310}_{4.262}$	$2.567^{2.599}_{2.535}$	$3.271^{3.299}_{3.244}$	$9.622^{9.683}_{9.560}$
Initial Sequence: Convex	$0.707^{0.718}_{0.695}$	$4.293^{4.317}_{4.269}$	$2.572^{2.604}_{2.541}$	$3.274^{3.302}_{3.248}$	$9.645^{9.708}_{9.582}$
Initial Sequence: Smoothed Lag Convex	$0.706^{0.717}_{0.694}$	$4.292^{4.315}_{4.269}$	$2.568^{2.600}_{2.536}$	$3.273^{3.300}_{3.246}$	$9.645^{9.708}_{9.584}$

[a] All values in kcal mol^{-1}. Uncertainties are 95 % confidence intervals obtained by bootstrapping over 10000 iterations with replacement.

Figure C.17: Discard times, RMSEs, and underlying time series properties for the bound vanish stage time series for all data sets. The top panels show kernel density estimates of the distributions of times discarded with each method. The bottom left panels show the RMSEs which would be obtained with an infinitely large ensemble of synthetic time series with fixed truncation points. The red dot indicates the optimum fixed-time truncation point.The right panels show the ensemble average RMSEs obtained with each of the truncation heuristics. Error bars are 95 % confidence intervals obtained by bootstrapping over 1000 iterations with replacement.

Figure C.18: Discard times, unsigned errors, and underlying time series properties for the bound vanish stage time series for all data sets. The top panels show kernel density estimates of the distributions of times discarded with each method. The bottom left panels show the RMSEs which would be obtained with an infinitely large ensemble of synthetic time series with fixed truncation points. The red dot indicates the optimum fixed-time truncation point. The right panels show the distributions of unsigned errors obtained over the synthetic ensembles.

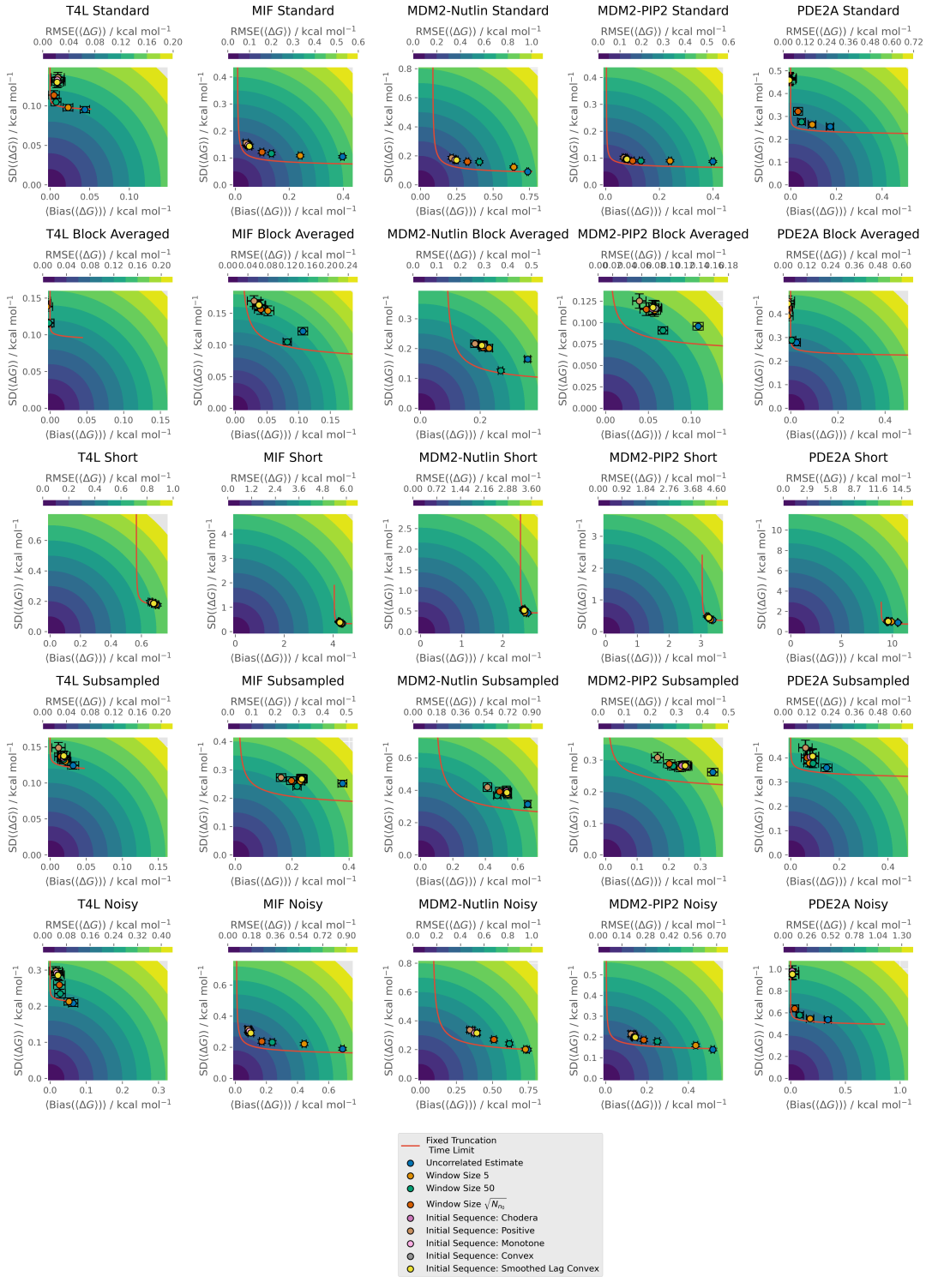


Figure C.19: Decomposition of errors made by generalised MSER methods over ensembles of synthetic trajectories. Red lines show the fixed truncation point limits. Error bars are 95 % confidence intervals obtained by bootstrapping over 1000 iterations with replacement.

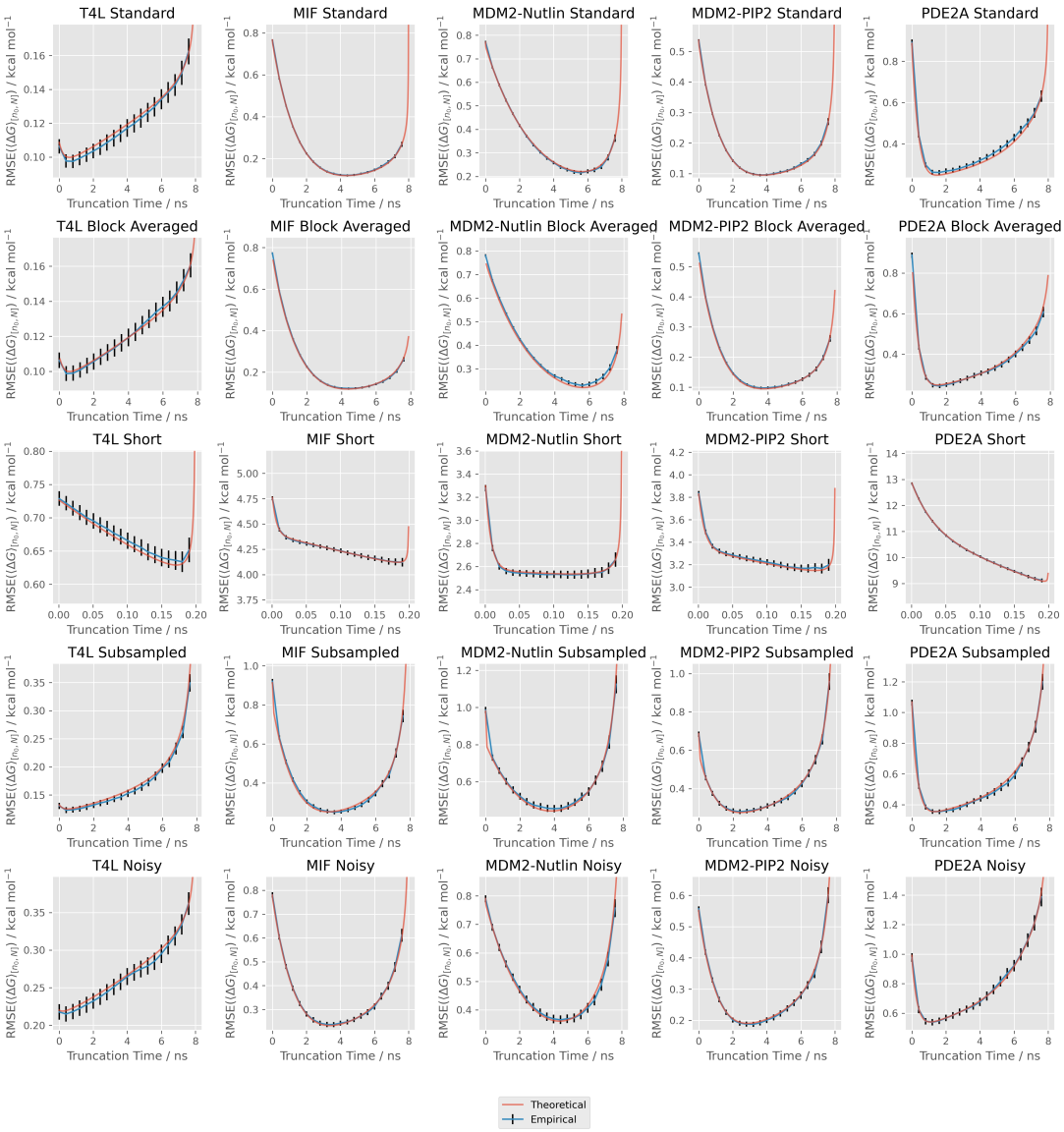


Figure C.20: Theoretical versus empirical RMSEs obtained for all synthetic time series ensembles for all systems. Error bars are 95 % confidence intervals obtained by bootstrapping over 1000 iterations with replacement.

Table C.7: Model Parameters Fitted to $\lambda = 0.45$ Bound Vanish Window of Absolute Binding Free Energy Calculations[a]

	Half-life (ns)	a (kcal mol^{-1})	Fast Half-life (ns)	Fast a (kcal mol^{-1})	Total Variance (kcal^2 mol^{-2})	Max Lag Index
T4L	0.16	2.5	∞	0	5000	4995
MIF	3.2	19	0.70	0.45	7100	859
MDM2-Nutlin	3.1	12	∞	0	29000	1475
MDM2-PIP2	3.0	11	∞	0	6500	1023
PDE2A	0.27	43	∞	0	49000	2231

[a] Total variance refers to the total variance of the mean, obtained by summing the autocovariance series from - max lag index to + max lag index, where the series and maximum lag indices were estimated according to Geyer's initial convex sequence rules. "a" refers to the pre-exponential factors.

C.7 Performance of Heuristics on Synthetic Data Modelled on Single λ State

To test the applicability of the methods to single λ states, rather than a free energy change integrated over all states, we modelled synthetic data on a single λ state. We chose the bound vanish $\lambda = 0.45$ state, which tends to show high noise and a substantial initial transient. Rather than integrating over all states using the trapezoidal rule, this window was given a weight of 1. Otherwise, the synthetic data were fit using the same procedure as for the "standard" synthetic ensembles. This generally produced greater ratios of total variance to the slow preexponential factor (Table C.7) compared to the "standard" synthetic data (Table 1).

Similarly to the "standard" synthetic ensembles, the methods which less thoroughly account for autocorrelation select earlier truncation times and perform better on systems with earlier optimal truncation times, and vice versa (Table C.8 and Figures C.21 and C.22). As with the "standard" ensembles, the $\sqrt{N_{n_0}}$ window method appeared to strike a reasonable compromise between bias sensitivity and truncation time variability.

Table C.8: Ensemble RMSEs for all Generalised MSER Heuristics for the Single Window Bound Vanish Data[a]

Method	T4L	MIF	MDM2-Nutlin	MDM2-PIP2	PDE2A
Uncorrelated Estimate	$0.68^{0.71}_{0.65}$	$9.00^{9.06}_{8.95}$	$5.79^{5.90}_{5.69}$	$5.21^{5.26}_{5.16}$	$2.41^{2.50}_{2.31}$
Window Size 5	$0.69^{0.72}_{0.66}$	$8.15^{8.22}_{8.07}$	$5.65^{5.76}_{5.54}$	$5.07^{5.13}_{5.02}$	$2.33^{2.42}_{2.23}$
Window Size 50	$0.81^{0.85}_{0.77}$	$6.50^{6.59}_{6.40}$	$5.30^{5.42}_{5.18}$	$4.46^{4.53}_{4.39}$	$2.40^{2.52}_{2.30}$
Window Size $\sqrt{N_{n0}}$	$0.91^{0.96}_{0.87}$	$5.57^{5.68}_{5.47}$	$4.82^{4.95}_{4.69}$	$3.90^{3.98}_{3.83}$	$2.94^{3.11}_{2.77}$
Initial Sequence: Chodera	$0.96^{1.00}_{0.92}$	$4.62^{4.74}_{4.51}$	$4.65^{4.83}_{4.48}$	$3.15^{3.24}_{3.06}$	$4.38^{4.60}_{4.15}$
Initial Sequence: Positive	$1.00^{1.04}_{0.95}$	$4.67^{4.78}_{4.55}$	$4.70^{4.89}_{4.52}$	$3.15^{3.24}_{3.06}$	$4.29^{4.51}_{4.06}$
Initial Sequence: Monotone	$0.99^{1.04}_{0.95}$	$4.66^{4.77}_{4.55}$	$4.63^{4.81}_{4.46}$	$3.18^{3.26}_{3.09}$	$4.32^{4.54}_{4.09}$
Initial Sequence: Convex	$0.98^{1.02}_{0.93}$	$4.68^{4.79}_{4.57}$	$4.63^{4.81}_{4.45}$	$3.22^{3.31}_{3.14}$	$4.32^{4.55}_{4.10}$
Initial Sequence: Smoothed Lag Convex	$0.98^{1.02}_{0.94}$	$4.67^{4.78}_{4.56}$	$4.65^{4.83}_{4.48}$	$3.23^{3.31}_{3.15}$	$4.29^{4.52}_{4.07}$

[a] All values in kcal mol^{-1}. Uncertainties are 95 % confidence intervals obtained by bootstrapping over 10000 iterations with replacement.

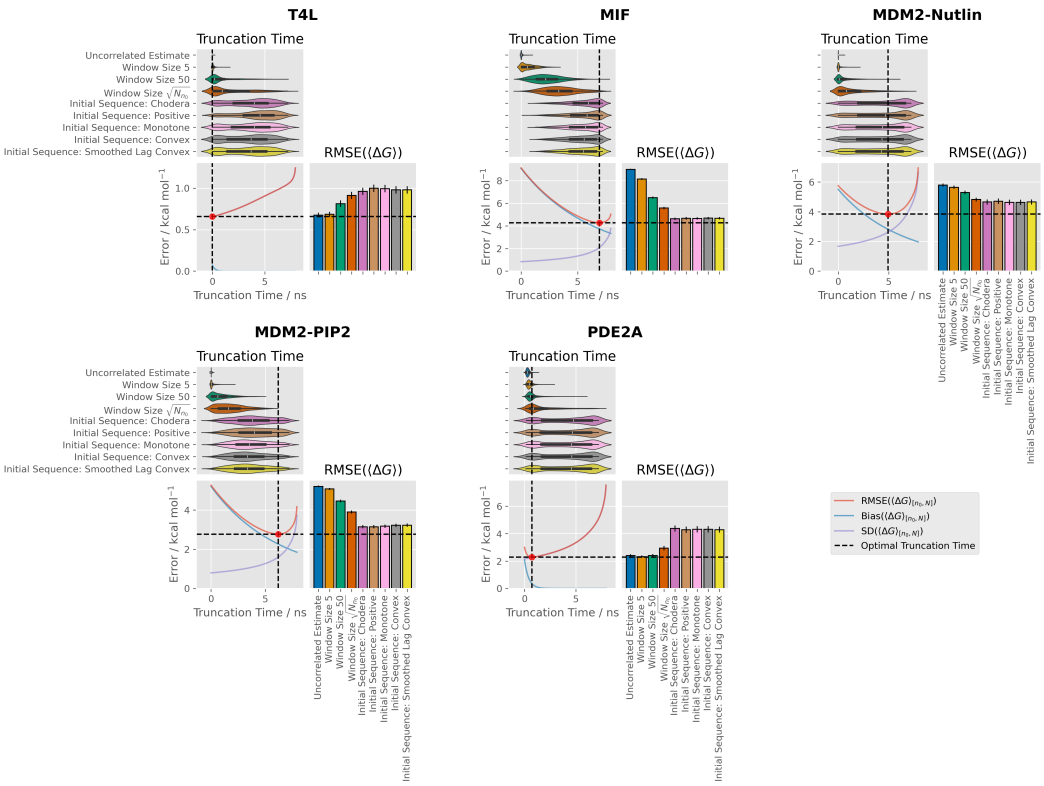


Figure C.21: Discard times, RMSEs, and underlying time series properties for the bound vanish stage single λ window time series. The top panels show kernel density estimates of the distributions of times discarded with each method. The bottom left panels show the RMSEs which would be obtained with an infinitely large ensemble of synthetic time series with fixed truncation points. The red dot indicates the optimum fixed-time truncation point. Bottom right pannels show the RMSEs obtained over the full synthetic ensemble for each method. Uncertainties are 95 % confidence intervals which were obtained by 10000 iterations of boostrapping with replacement.

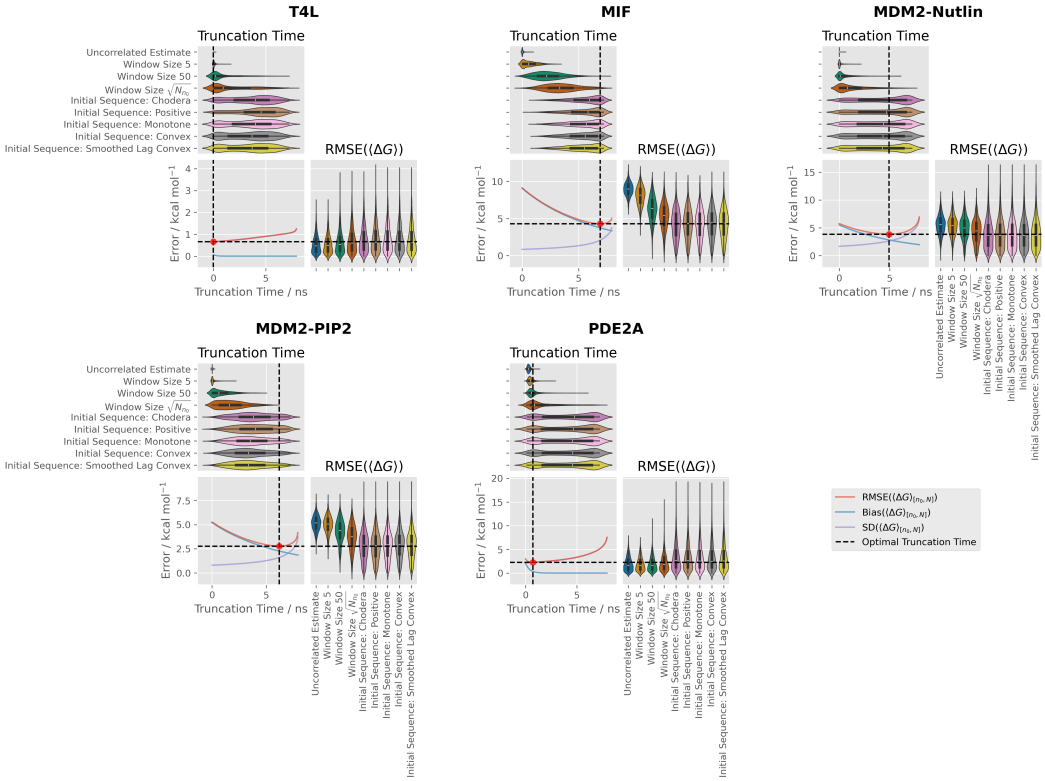


Figure C.22: Discard times, unsigned errors, and underlying time series properties for the bound vanish stage single λ window time series. The top panels show kernel density estimates of the distributions of times discarded with each method. The bottom left panels show the RMSEs which would be obtained with an infinitely large ensemble of synthetic time series with fixed truncation points. The red dot indicates the optimum fixed-time truncation point. The bottom right pannels show the kernel density estimates of the distributions of unsigned errors obtained for each method.

C.8 Coverage of Confidence Intervals for Truncated Free Vanish Data

Table C.9: Coverage of 95% Confidence Intervals for Truncated Free Vanish Data[a]

System	**T4L**	**PDE2A**
Dataset	standard	standard
Uncorrelated Estimate	97	44
Window Size 5	97	50
Window Size 50	97	65
Window Size $\sqrt{N_{n_0}}$	97	71
Initial Sequence: Chodera	97	63
Initial Sequence: Positive	97	67
Initial Sequence: Monotone	97	64
Initial Sequence: Convex	97	63
Initial Sequence: Smoothed Lag Convex	97	63

[a] All coverages given as a % of ensemble members for which the 95% confidence interval included the true ensemble value (hence ideally all coverages should be $\approx 95\%$). Confidence intervals were calculated from each individual truncated time series using the same method used to select the truncation point by minimising the marginal standard error.

C.9 The Variance Increase Resulting from Subsampling

A time series of length N is subsampled at intervals of S (so that the new series has length ($N_{\text{Sub}} = \frac{N}{S}$). There is no initial bias and no truncation is performed. The variance of the mean of the subsampled time series, $\text{Var}_{\text{Trajs,Sub}}(\langle A \rangle_{[0,N_{\text{Sub}}]})$, will be larger than that of the original time series, $\text{Var}_{\text{Trajs}}(\langle A \rangle_{[0,N]})$, by the factor

$$\frac{\text{Var}_{\text{Trajs,Sub}}(\langle A \rangle_{[0,N_{\text{Sub}}]})}{\text{Var}_{\text{Trajs}}(\langle A \rangle_{[0,N]})} = \frac{\frac{1}{N_{\text{Sub}}}\left(\gamma_0 + 2\sum_{t'=1}^{N_{\text{Sub}}-1}\gamma'_{t'}\right)}{\frac{1}{N}\left(\gamma_0 + 2\sum_{t=1}^{N_{\text{Sub}}-1}\gamma_t\right)} \tag{C.8}$$

$$= S\frac{\gamma_0 + 2\sum_{t'=1}^{N_{\text{Sub}}-1}\gamma'_{t'}}{\gamma_0 + 2\sum_{t=1}^{N-1}\gamma_t}, \tag{C.9}$$

where $'$ denotes the lag times and autocovariances of the subsampled time series, and all other terms are defined in Section 2 of the main text. Taking out the common factor of γ_0, this can be rewritten as

$$\frac{\text{Var}_{\text{Trajs,Sub}}(\langle A \rangle_{[0,N_{\text{Sub}}]})}{\text{Var}_{\text{Trajs}}(\langle A \rangle_{[0,N]})} = S\frac{1 + 2\sum_{t'=1}^{N_{\text{Sub}}-1}\frac{\gamma'_{t'}}{\gamma_0}}{1 + 2\sum_{t=1}^{N-1}\frac{\gamma_t}{\gamma_0}} \tag{C.10}$$

$$= S\frac{g_{\text{Sub}}}{g}, \tag{C.11}$$

where $g = 1 + 2\sum_{t=1}^{N-1}\frac{\gamma_t}{\gamma_0}$ is the statistical inefficiency of the original time series, and g_{Sub} is the statistical inefficiency of the subsampled time series. When there is no autocorrelation, $g = g_{\text{Sub}}$ and the increase in variance is proportional to the subsampling interval. This is the worst case scenario. When the time series is autocorrelated, subsampling reduces the autocorrelation, meaning that $g > g_{\text{Sub}}$ and $\frac{g_{\text{Sub}}}{g} < 1$, hence proportional increase in error is less than if the samples were uncorrelated (when S is constant).

Often, data are subsampled according to their statistical inefficiency.[20] Assuming a perfect estimate of the statistical inefficiency, $S = g$ and

$$\frac{\text{Var}_{\text{Trajs,Sub}}(\langle A \rangle_{[0,N_{\text{Sub}}]})}{\text{Var}_{\text{Trajs}}(\langle A \rangle_{[0,N]})} = g_{\text{Sub}}. \tag{C.12}$$

Following Janke,[110] we explore the idealised case when an autocorrelation function is a single exponential with decay constant τ. Then, the statistical inefficiency is

$$g = 1 + 2\sum_{t=1}^{N-1}\frac{\gamma_t}{\gamma_0} \tag{C.13}$$

$$= 1 + 2\sum_{t=1}^{N-1} e^{-\frac{t}{\tau}}. \tag{C.14}$$

Assuming that $N \gg \tau$,

$$g \approx 1 + 2\sum_{t=1}^{\infty} e^{-\frac{t}{\tau}} = -1 + 2\sum_{t=0}^{\infty} e^{-\frac{1}{\tau}} \tag{C.15}$$

$$= -1 + \frac{2}{1 - e^{-\frac{1}{\tau}}} \tag{C.16}$$

$$= \frac{1 + e^{-\frac{1}{\tau}}}{1 - e^{-\frac{1}{\tau}}} \tag{C.17}$$

$$= \coth\left(\frac{1}{2\tau}\right) \tag{C.18}$$

The subsampled decay constant, τ_{Sub}, is related to τ by

$$e^{-\frac{t'}{\tau_{\text{Sub}}}} = e^{-\frac{t}{S\tau_{\text{Sub}}}} = e^{-\frac{t}{\tau}}, \tag{C.19}$$

hence

$$\tau_{\text{Sub}} = \frac{\tau}{S}. \tag{C.20}$$

When subsampling is performed at the interval of the perfectly-estimated statistical inefficiency, then

$$\tau_{\text{Sub}} = \frac{\tau}{\coth\left(\frac{1}{2\tau}\right)} \tag{C.21}$$

$$= \tau \tanh\left(\frac{1}{2\tau}\right). \tag{C.22}$$

If $\tau \gg \frac{1}{2}$, then $\frac{1}{2\tau} \ll 1$ and

$$\tau_{\text{Sub}} \approx \frac{\tau}{2\tau} = \frac{1}{2}. \tag{C.23}$$

Then

$$\frac{\mathrm{Var}_{\mathrm{Trajs,Sub}}(\langle A\rangle_{[0,N_{\mathrm{Sub}}]})}{\mathrm{Var}_{\mathrm{Trajs}}(\langle A\rangle_{[0,N]})} \approx \coth(1) = 1.31. \tag{C.24}$$

Hence, for a stationary time series with a single exponential autocorrelation function, when the decay constant is much greater than $\frac{1}{2}$ (e.g. the half-life is much greater than the sampling interval) and the number of samples is much greater than the decay constant, subsampling increases the variance of the mean by 31%.

Appendix D

Rapid ABFE calculations for Virtual Screening

D.1 Initial HSP90 Results with Outlier

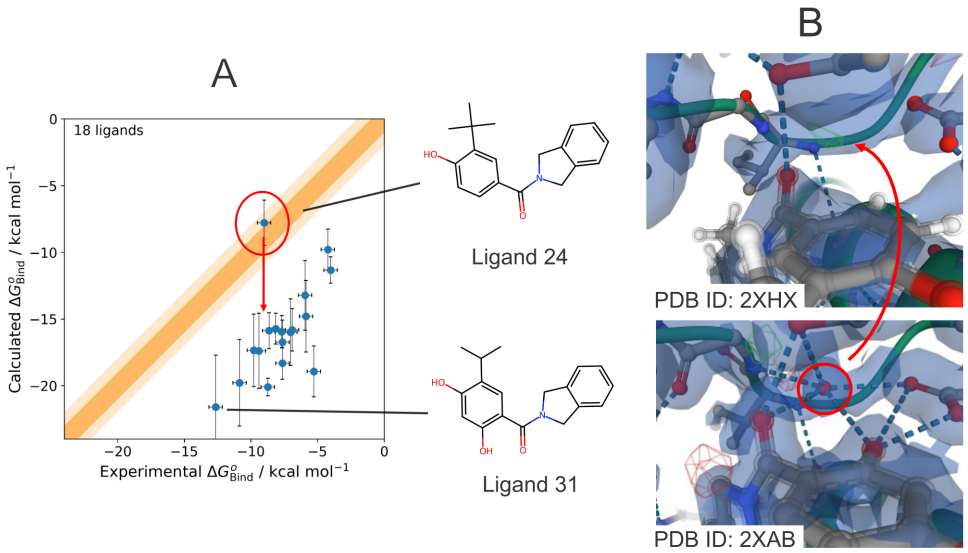


Figure D.1: Initial HSP90 Results with Outlier. A) Calculated against experimental free energies of binding for HSP90 ligands, highlighting the outlier (ligand 28 from Alibay et al.) and a non-outlier (ligand 31 from Alibay et al.).[1] B) The crystal structures used to prepare the ligand 28 complex (PDB ID 2XHX) and the ligand 31 complex (PDB ID 2XAB) by Alibay et al., highlighting the absence of a water molecule at the site of missing electron density in the ligand 28 complex. When the complexes were aligned and this water was transplanted from the ligand 31 to ligand 28 complex, the calculated ΔG^o_{Bind} for ligand 28 dropped from -7.78 ± 1.70 kcal mol^{-1} to -15.05 ± 2.82 kcal mol^{-1}, in line with the other results. Ligands drawn with RDKit.[281]

Bibliography

[1] Alibay, I.; Magarkar, A.; Seeliger, D.; Biggin, P. C. Evaluating the Use of Absolute Binding Free Energy in the Fragment Optimisation Process. *Commun. Chem.* **2022**, *5*, 105.

[2] Dowden, H.; Munro, J. Trends in Clinical Success Rates and Therapeutic Focus. *Nat. Rev. Drug Discov.* **2019**, *18*, 495–496.

[3] Research and Development in the Pharmaceutical Industry | Congressional Budget Office. https://www.cbo.gov/publication/57126, Thu, 04/08/2021 - 12:00.

[4] Mullard, A. 2023 FDA Approvals. *Nat. Rev. Drug Discov.* **2024**, *23*, 88–95.

[5] Rang, H. P. The Receptor Concept: Pharmacology's Big Idea. *Br. J. Pharmacol.* **2006**, *147*, S9–S16.

[6] Sadri, A. Is Target-Based Drug Discovery Efficient? Discovery and "Off-Target" Mechanisms of All Drugs. *J. Med. Chem.* **2023**, *66*, 12651–12677.

[7] Sinha, S.; Vohora, D. In *Pharmaceutical Medicine and Translational Clinical Research*; Vohora, D., Singh, G., Eds.; Academic Press: Boston, 2018; pp 19–32.

[8] Sadybekov, A. V.; Katritch, V. Computational Approaches Streamlining Drug Discovery. *Nat.* **2023**, *616*, 673–685.

[9] M. Hann, M. Molecular Obesity, Potency and Other Addictions in Drug Discovery. *Med. Chem. Commun.* **2011**, *2*, 349–355.

[10] Brown, D. G.; Boström, J. Where Do Recent Small Molecule Clinical Development Candidates Come From? *J. Med. Chem.* **2018**, *61*, 9442–9468.

[11] Schneider, G. Automating Drug Discovery. *Nat. Rev. Drug Discov.* **2018**, *17*, 97–113.

[12] Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings1. *Adv. Drug Deliv. Rev.* **2001**, *46*, 3–26.

[13] Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the Chemical Beauty of Drugs. *Nat. Chem.* **2012**, *4*, 90–98.

[14] Goldstein, D. M.; Gray, N. S.; Zarrinkar, P. P. High-Throughput Kinase Profiling as a Platform for Drug Discovery. *Nat. Rev. Drug Discov.* **2008**, *7*, 391–397.

[15] Pollard, C. E.; Valentin, J.-P.; Hammond, T. G. Strategies to Reduce the Risk of Drug-Induced QT Interval Prolongation: A Pharmaceutical Company Perspective. *Br. J. Pharmacol.* **2008**, *154*, 1538–1543.

[16] van Breemen, R. B.; Li, Y. Caco-2 Cell Permeability Assays to Measure Drug Absorption. *Expert Opin. Drug Metab. Toxicol.* **2005**, *1*, 175–185.

[17] Singh, V. K.; and Seed, T. M. How Necessary Are Animal Models for Modern Drug Discovery? *Expert Opin. Drug Discov.* **2021**, *16*, 1391–1397.

[18] Hopkins, A. L.; Mason, J. S.; Overington, J. P. Can We Rationally Design Promiscuous Drugs? *Curr. Opin. Struct. Biol.* **2006**, *16*, 127–136.

[19] Kairys, V.; Baranauskiene, L.; Kazlauskiene, M.; Matulis, D.; Kazlauskas, E. Binding Affinity in Drug Design: Experimental and Computational Techniques. *Expert Opin. Drug. Discov.* **2019**, *14*, 755–768.

[20] Hahn, D.; Bayly, C.; Boby, M. L.; Macdonald, H. B.; Chodera, J.; Gapsys, V.; Mey, A.; Mobley, D.; Benito, L. P.; Schindler, C.; Tresadern, G.; Warren, G. Best Practices for Constructing, Preparing, and Evaluating Protein-Ligand Binding Affinity Benchmarks [Article v1.0]. *Living J. Comput. Mol. Sci.* **2022**, *4*, 1497–1497.

[21] Mobley, D. L.; Klimovich, P. V. Perspective: Alchemical Free Energy Calculations for Drug Discovery. *J. Chem. Phys.* **2012**, *137*, 230901.

[22] Jumper, J. et al. Highly Accurate Protein Structure Prediction with AlphaFold. *Nat.* **2021**, *596*, 583–589.

[23] Renaud, J.-P.; Chari, A.; Ciferri, C.; Liu, W.-t.; Rémigy, H.-W.; Stark, H.; Wiesmann, C. Cryo-EM in Drug Discovery: Achievements, Limitations and Prospects. *Nat. Rev. Drug Discov.* **2018**, *17*, 471–492.

[24] Fernandez-Leiro, R.; Scheres, S. H. W. Unravelling Biological Macromolecules with Cryo-Electron Microscopy. *Nat.* **2016**, *537*, 339–346.

[25] Jones, N. Crystallography: Atomic Secrets. *Nat.* **2014**, *505*, 602–603.

[26] Liu, W. et al. Serial Femtosecond Crystallography of G Protein–Coupled Receptors. *Science* **2013**, *342*, 1521–1524.

[27] Wei, H.; McCammon, J. A. Structure and Dynamics in Drug Discovery. *NPJ Drug Discov.* **2024**, *1*, 1–8.

[28] Stanzione, F.; Giangreco, I.; Cole, J. C. In *Progress in Medicinal Chemistry*; Witty, D. R., Cox, B., Eds.; Elsevier, 2021; Vol. 60; pp 273–343.

[29] Schaller, D. A.; Christ, C. D.; Chodera, J. D.; Volkamer, A. Benchmarking Cross-Docking Strategies in Kinase Drug Discovery. *J. Chem. Inf. Model* **2024**,

[30] Ohadi, D.; Kumar, K.; Ravula, S.; DesJarlais, R. L.; Seierstad, M. J.; Shih, A. Y.; Hack, M. D.; Schiffer, J. M. Input Pose Is Key to Performance of Free Energy Perturbation: Benchmarking with Monoacylglycerol Lipase. *J. Chem. Inf. Model* **2024**,

[31] Gorgulla, C.; Boeszoermenyi, A.; Wang, Z.-F.; Fischer, P. D.; Coote, P. W.; Padmanabha Das, K. M.; Malets, Y. S.; Radchenko, D. S.; Moroz, Y. S.; Scott, D. A.; Fackeldey, K.; Hoffmann, M.; Iavniuk, I.; Wagner, G.; Arthanari, H. An Open-Source Drug Discovery Platform Enables Ultra-Large Virtual Screens. *Nat.* **2020**, *580*, 663–668.

[32] Wallach, I.; Heifets, A. Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization. *J. Chem. Inf. Model.* **2018**, *58*, 916–932.

[33] Sieg, J.; Flachsenberg, F.; Rarey, M. In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2019**, *59*, 947–961.

[34] Chen, L.; Cruz, A.; Ramsey, S.; Dickson, C. J.; Duca, J. S.; Hornak, V.; Koes, D. R.; Kurtzman, T. Hidden Bias in the DUD-E Dataset Leads to Misleading Performance of Deep Learning in Structure-Based Virtual Screening. *PLOS ONE* **2019**, *14*, e0220113.

[35] Boyles, F.; Deane, C. M.; Morris, G. M. Learning from Docked Ligands: Ligand-Based Features Rescue Structure-Based Scoring Functions When Trained on Docked Poses. *J. Chem. Inf. Model.* **2022**, *62*, 5329–5341.

[36] Åqvist, J.; Medina, C.; Samuelsson, J.-E. A New Method for Predicting Binding Affinity in Computer-Aided Drug Design. *Protein Eng. Des. Sel.* **1994**, *7*, 385–391.

[37] Hansson, T.; Marelius, J.; Åqvist, J. Ligand Binding Affinity Prediction by Linear Interaction Energy Methods. *J. Comput. Aided Mol. Des.* **1998**, *12*, 27–35.

[38] Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate-DNA Helices. *J. Am. Chem. Soc.* **1998**, *120*, 9401–9409.

[39] Massova, I.; Kollman, P. A. Combined Molecular Mechanical and Continuum Solvent Approach (MM-PBSA/GBSA) to Predict Ligand Binding. *Perspect. Drug Discov. Des.* **2000**, *18*, 113–135.

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

[41] Åqvist, J.; Wennerström, P.; Nervall, M.; Bjelic, S.; Brandsdal, B. O. Molecular Dynamics Simulations of Water and Biomolecules with a Monte Carlo Constant Pressure Algorithm. *Chem. Phys. Lett.* **2004**, *384*, 288–294.

[42] Genheden, S.; Ryde, U. Comparison of the Efficiency of the LIE and MM/GBSA Methods to Calculate Ligand-Binding Energies. *J. Chem. Theory Comput.* **2011**, *7*, 3768–3778.

[43] Rifai, E. A.; van Dijk, M.; Vermeulen, N. P. E.; Yanuar, A.; Geerke, D. P. A Comparative Linear Interaction Energy and MM/PBSA Study on SIRT1–Ligand Binding Free Energy Calculation. *J. Chem. Inf. Model.* **2019**, *59*, 4018–4033.

[44] Head, M. S.; Given, J. A.; Gilson, M. K. "Mining Minima": Direct Computation of Conformational Free Energy. *J. Phys. Chem. A* **1997**, *101*, 1609–1618.

[45] Gilson, M. K.; Stewart, L. E.; Potter, M. J.; Webb, S. P. Rapid, Accurate, Ranking of Protein–Ligand Binding Affinities with VM2, the Second-Generation Mining Minima Method. *J. Chem. Theory Comput.* **2024**, *20*, 6328–6340.

[46] Roux, B.; Chipot, C. Editorial Guidelines for Computational Studies of Ligand Binding Using MM/PBSA and MM/GBSA Approximations Wisely. *J. Phys. Chem. B* **2024**,

[47] Gundelach, L.; Fox, T.; S. Tautermann, C.; Skylaris, C.-K. Protein–Ligand Free Energies of Binding from Full-Protein DFT Calculations: Convergence and Choice of Exchange–Correlation Functional. *Phys. Chem. Chem. Phys.* **2021**, *23*, 9381–9393.

[48] Jorgensen, W. L.; Ravimohan, C. Monte Carlo Simulation of Differences in Free Energies of Hydration. *J. Chem. Phys.* **1985**, *83*, 3050–3054.

[49] Gilson, M.; Given, J.; Bush, B.; McCammon, J. The Statistical-Thermodynamic Basis for Computation of Binding Affinities: A Critical Review. *Biophys. J.* **1997**, *72*, 1047–1069.

[50] Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. Absolute Binding Free Energies: A Quantitative Approach for Their Calculation. *J. Phys. Chem. B* **2003**, *107*, 9535–9551.

[51] Woo, H.-J.; Roux, B. Calculation of Absolute Protein–Ligand Binding Free Energy from Computer Simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6825–6830.

[52] Wang, L. et al. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J. Am. Chem. Soc.* **2015**, *137*, 2695–2703.

[53] Breznik, M.; Ge, Y.; Bluck, J. P.; Briem, H.; Hahn, D. F.; Christ, C. D.; Mortier, J.; Mobley, D. L.; Meier, K. Prioritizing Small Sets of Molecules for Synthesis through In-Silico Tools: A Comparison of Common Ranking Methods. *ChemMedChem* **2023**, *18*, e202200425.

[54] Cournia, Z.; Allen, B.; Sherman, W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *J. Chem. Inf. Model.* **2017**, *57*, 2911–2937.

[55] Vögele, M.; Zhang, B. W.; Kaindl, J.; Wang, L. Is the Functional Response of a Receptor Determined by the Thermodynamics of Ligand Binding? *J. Chem. Theory Comput.* **2023**, *19*, 8414–8422.

[56] Aldeghi, M.; Heifetz, A.; Bodkin, M. J.; Knapp, S.; Biggin, P. C. Predictions of Ligand Selectivity from Absolute Binding Free Energy Calculations. *J. Am. Chem. Soc.* **2017**, *139*, 946–957.

[57] Li, F. et al. CACHE Challenge #1: Targeting the WDR Domain of LRRK2, A Parkinson's Disease Associated Protein. *J. Chem. Inf. Model.* **2024**, *64*, 8521–8536.

[58] Cournia, Z.; Allen, B. K.; Beuming, T.; Pearlman, D. A.; Radak, B. K.; Sherman, W. Rigorous Free Energy Simulations in Virtual Screening. *J. Chem. Inf. Model.* **2020**, *60*, 4153–4169.

[59] Aldeghi, M.; Bodkin, M. J.; Knapp, S.; Biggin, P. C. Statistical Analysis on the Performance of Molecular Mechanics Poisson–Boltzmann Surface Area versus Absolute Binding Free Energy Calculations: Bromodomains as a Case Study. *J. Chem. Inf. Model.* **2017**, *57*, 2203–2221.

[60] Feng, M.; Heinzelmann, G.; Gilson, M. K. Absolute Binding Free Energy Calculations Improve Enrichment of Actives in Virtual Compound Screening. *Sci. Rep.* **2022**, *12*, 13640.

[61] Chen, W.; Cui, D.; Jerome, S. V.; Michino, M.; Lenselink, E. B.; Huggins, D. J.; Beautrait, A.; Vendome, J.; Abel, R.; Friesner, R. A.; Wang, L. Enhancing Hit Discovery in Virtual Screening through Absolute Protein–Ligand Binding Free-Energy Calculations. *J. Chem. Inf. Model.* **2023**, *63*, 3171–3185.

[62] Gutkin, E.; Gusev, F.; Gentile, F.; Ban, F.; Benjamin Koby, S.; Narangoda, C.; Isayev, O.; Cherkasov, A.; G. Kurnikova, M. In Silico Screening of LRRK2 WDR Domain Inhibitors Using Deep Docking and Free Energy Simulations. *Chem. Sci.* **2024**, *15*, 8800–8812.

[63] Li, Z. et al. Identify Potent SARS-CoV-2 Main Protease Inhibitors via Accelerated Free Energy Perturbation-Based Virtual Screening of Existing Drugs. *Proc. Natl. Acad. Sci.* **2020**, *117*, 27381–27387.

[64] Schrödinger, E. *Statistical Thermodynamics*, revised edition ed.; Dover Publications: Mineola, NY, 1989; Originally published in 1946 by Cambridge University Press.

[65] Boltzmann, L. *Wissenschaftliche Abhandlungen: Bd. 1865-1874*; Chelsea Publishing Company, Incorporated, 1909; Vol. 215.

[66] Gibbs, J. W. *Elementary principles in statistical mechanics: developed with especial reference to the rational foundations of thermodynamics*; C. Scribner's sons, 1902.

[67] Dill, K.; Bromberg, S. *Molecular Driving Forces: Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience*, 2nd ed.; Garland Science: New York, 2010.

[68] Jaynes, E. T. Gibbs vs Boltzmann Entropies. *Am. J. Phys.* **1965**, *33*, 391–398.

[69] Gao, X.; Gallicchio, E.; Roitberg, A. E. The Generalized Boltzmann Distribution Is the Only Distribution in Which the Gibbs-Shannon Entropy Equals the Thermodynamic Entropy. *J. Chem. Phys.* **2019**, *151*, 034113.

[70] Jaynes, E. T. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106*, 620–630.

[71] Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.

[72] Steinberg, I. Z.; Scheraga, H. A. Entropy Changes Accompanying Association Reactions of Proteins. *J. Biol. Chem.* **1963**, *238*, 172–181.

[73] Duboué-Dijon, E.; Hénin, J. Building Intuition for Binding Free Energy Calculations: Bound State Definition, Restraints, and Symmetry. *J. Chem. Phys.* **2021**, *154*, 204101.

[74] Cheng, Y.-C.; Prusoff, W. H. Relationship between the Inhibition Constant (KI) and the Concentration of Inhibitor Which Causes 50 per Cent Inhibition (I50) of an Enzymatic Reaction. *Biochem. Pharmacol.* **1973**, *22*, 3099–3108.

[75] Riniker, S. Fixed-Charge Atomistic Force Fields for Molecular Dynamics Simulations in the Condensed Phase: An Overview. *J. Chem. Inf. Model.* **2018**, *58*, 565–578.

[76] Shirts, M. R.; Mobley, D. L.; Chodera, J. D.; Pande, V. S. Accurate and efficient corrections for missing dispersion interactions in molecular simulations. *J. Phys. Chem. B.* **2007**, *111*, 13052–13063.

[77] Aldeghi, M.; Bluck, J. P.; Biggin, P. C. In *Computational Drug Discovery and Design*; Gore, M., Jagtap, U. B., Eds.; Springer: New York, NY, 2018; pp 199–232.

[78] Whitmore, L. M.; Shirts, M. R. Force Switching and Potential Shifting Lead to Errors in Free Energies of Alchemical Transformations. *arXiv* **2024**, *2410.14187*, DOI: 10.48550/arXiv.2410.14187.

[79] Barker, J.; Watts, R. Monte Carlo Studies of the Dielectric Properties of Water-like Models. *Mol. Phys.* **1973**, *26*, 789–792.

[80] Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N·log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

[81] Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

[82] Qiu, Y. et al. Development and Benchmarking of Open Force Field v1.0.0—the Parsley Small-Molecule Force Field. *J. Chem. Theory Comput.* **2021**, *17*, 6262–6280.

[83] Boothroyd, S. et al. Development and Benchmarking of Open Force Field 2.0.0: The Sage Small Molecule Force Field. *J. Chem. Theory Comput.* **2023**, *19*, 3251–3275.

[84] Mobley, D. L.; Bannan, C. C.; Rizzi, A.; Bayly, C. I.; Chodera, J. D.; Lim, V. T.; Lim, N. M.; Beauchamp, K. A.; Slochower, D. R.; Shirts, M. R.; Gilson, M. K.; Eastman, P. K. Escaping Atom Types in Force Fields Using Direct Chemical Perception. *J. Chem. Theory. Comput.* **2018**, *14*, 6076–6092.

[85] Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132–146.

[86] Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comput. Chem.* **2002**, *23*, 1623–1641.

[87] Wang, Y.; Fass, J.; Kaminow, B.; E. Herr, J.; Rufa, D.; Zhang, I.; Pulido, I.; Henry, M.; Macdonald, H. E. B.; Takaba, K.; D. Chodera, J. End-to-End Differentiable Construction of Molecular Mechanics Force Fields. *Chem. Sci.* **2022**, *13*, 12016–12033.

[88] Rackers, J. A.; Wang, Q.; Liu, C.; Piquemal, J.-P.; Ren, P.; Ponder, J. W. An Optimized Charge Penetration Model for Use with the AMOEBA Force Field. *Phys. Chem. Chem. Phys.* **2017**, *19*, 276–291.

[89] Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A. J.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B* **2010**, *114*, 2549–2564.

[90] Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B. E(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials. *Nat. Commun.* **2022**, *13*, 2453.

[91] Batatia, I.; Kovács, D. P.; Simm, G. N. C.; Ortner, C.; Csányi, G. MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields. *arXiv* **2022**, DOI: 10.48550/ARXIV.2206.07697.

[92] Kovács, D. P.; Moore, J. H.; Browning, N. J.; Batatia, I.; Horton, J. T.; Kapil, V.; Witt, W. C.; Magdău, I.-B.; Cole, D. J.; Csányi, G. MACE-OFF23: Transferable Machine Learning Force Fields for Organic Molecules. *arXiv* **2023**, DOI: 10.48550/arXiv.2312.15211.

[93] Anstine, D.; Zubatyuk, R.; Isayev, O. AIMNet2: A Neural Network Potential to Meet Your Neutral, Charged, Organic, and Elemental-Organic Needs. *ChemRxiv* **2024**, DOI: 10.26434/chemrxiv-2023-296ch-v2.

[94] Sabanés Zariquiey, F.; Galvelis, R.; Gallicchio, E.; Chodera, J. D.; Markland, T. E.; De Fabritiis, G. Enhancing Protein–Ligand Binding Affinity Predictions Using Neural Network Potentials. *J. Chem. Inf. Model.* **2024**, *64*, 1481–1485.

[95] Karwounopoulos, J.; Wu, Z.; Tkaczyk, S.; Wang, S.; Baskerville, A.; Ranasinghe, K.; Langer, T.; Wood, G. P. F.; Wieder, M.; Boresch, S. Insights and Challenges in Correcting Force Field Based Solvation Free Energies Using a Neural Network Potential. *J. Phys. Chem. B.* **2024**, *128*, 6693–6703.

[96] Karwounopoulos, J.; Bieniek, M.; Wu, Z.; Baskerville, A. L.; König, G.; Cossins, B. P.; Wood, G. P. F. Evaluation of Machine Learning/Molecular Mechanics End-State Corrections with Mechanical Embedding to Calculate Relative Protein–Ligand Binding Free Energies. *J. Chem. Theory Comput.* **2025**, *21*, 967–977.

[97] Wang, Y. et al. On the Design Space between Molecular Mechanics and Machine Learning Force Fields. *arXiv* **2024**, DOI: 10.48550/arXiv.2409.01931.

[98] Braun, E.; Gilmer, J.; Mayes, H. B.; Mobley, D. L.; Monroe, J. I.; Prasad, S.; Zuckerman, D. M. Best Practices for Foundations in Molecular Simulations [Article v1.0]. *Living J. Comp. Mol. Sci.* **2019**, *1*, 5957–5957.

[99] Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press, 2017.

[100] Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. A Computer Simulation Method for the Calculation of Equilibrium Constants for the Formation of Physical Clusters of Molecules: Application to Small Water Clusters. *J. Chem. Phys.* **1982**, *76*, 637–649.

[101] Andersen, H. C. Molecular Dynamics Simulations at Constant Pressure and/or Temperature. *J. Chem. Phys.* **1980**, *72*, 2384–2393.

[102] Schneider, T.; Stoll, E. Molecular-Dynamics Study of a Three-Dimensional One-Component Model for Distortive Phase Transitions. *Physs. Rev. B* **1978**, *17*, 1302–1322.

[103] van Gunsteren, W.; Berendsen, H. Algorithms for Macromolecular Dynamics and Constraint Dynamics. *Mol. Phys.* **1977**, *34*, 1311–1327.

[104] Hopkins, C. W.; Le Grand, S.; Walker, R. C.; Roitberg, A. E. Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *J. Chem. Theory Comput.* **2015**, *11*, 1864–1874.

[105] Hénin, J.; Lelièvre, T.; Shirts, M. R.; Valsson, O.; Delemotte, L. Enhanced Sampling Methods for Molecular Dynamics Simulations [Article v1.0]. *Living J. Comp. Mol. Sci.* **2022**, *4*, 1583–1583.

[106] Chodera, J. D.; Shirts, M. R. Replica Exchange and Expanded Ensemble Simulations as Gibbs Sampling: Simple Improvements for Enhanced Mixing. *J. Chem. Phys.* **2011**, *135*, 194110.

[107] Wang, L.; Friesner, R. A.; Berne, B. J. Replica Exchange with Solute Scaling: A More Efficient Version of Replica Exchange with Solute Tempering (REST2). *J. Phys. Chem. B* **2011**, *115*, 9431–9438.

[108] Chodera, J. D. A Simple Method for Automated Equilibration Detection in Molecular Simulations. *J. Chem. Theory Comput.* **2016**, *12*, 1799–1805.

[109] Yang, W.; Bitetti-Putzer, R.; Karplus, M. Free Energy Simulations: Use of Reverse Cumulative Averaging to Determine the Equilibrated Region and the Time Required for Convergence. *J. Chem. Phys.* **2004**, *120*, 2618–2628.

[110] Janke, W. In *Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms*; Grotendorst, J., Marx, D., Murmatsu, A., Eds.; John von Neumann Institute for Computing, 2002; Vol. 10.

[111] Caves, L. S. D.; Evanseck, J. D.; Karplus, M. Locally Accessible Conformations of Proteins: Multiple Molecular Dynamics Simulations of Crambin. *Protein Sci.* **1998**, *7*, 649–666.

[112] Horowitz, J. L. Bootstrap Methods in Econometrics. *Annu. Rev. Econ.* **2019**, *11*, 193–224.

[113] Ross, G. A.; Lu, C.; Scarabelli, G.; Albanese, S. K.; Houang, E.; Abel, R.; Harder, E. D.; Wang, L. The Maximal and Current Accuracy of Rigorous Protein-Ligand Binding Free Energy Calculations. *Commun. Chem.* **2023**, *6*, 1–12.

[114] Mey, A. S.; Allen, B. K.; Bruce Macdonald, H. E.; Chodera, J. D.; Hahn, D. F.; Kuhn, M.; Michel, J.; Mobley, D. L.; Naden, L. N.; Prasad, S.; Rizzi, A.; Scheen, J.; Shirts, M. R.; Tresadern, G.; Xu, H. Best Practices for Alchemical Free Energy Calculations [Article v1.0]. *LiveCoMS* **2020**, *2*.

[115] Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22*, 1420–1426.

[116] Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* **1935**, *3*, 300–313.

[117] Kofke, D. A.; Cummings, P. T. Quantitative Comparison and Optimization of Methods for Evaluating the Chemical Potential by Molecular Simulation. *Mol. Phys.* **1997**, *92*, 973–996.

[118] Lu, N.; Kofke, D. A. Accuracy of Free-Energy Perturbation Calculations in Molecular Simulation. I. Modeling. *J. Chem. Phys.* **2001**, *114*, 7303–7311.

[119] Geyer, C. J. *Estimating Normalizing Constants and Reweighting Mixtures*; Technical Report 568, 1994.

[120] Shirts, M. R. Reweighting from the Mixture Distribution as a Better Way to Describe the Multistate Bennett Acceptance Ratio. 2017; `10.48550/arXiv.1704.00891`.

[121] Bennett, C. H. Efficient Estimation of Free Energy Differences from Monte Carlo Data. *J. Comput. Phys.* **1976**, *22*, 245–268.

[122] Meng, X.-L.; Wong, W. H. Simulating Ratios of Normalizing Constants Via a Simple Identity: A Theoretical Exploration. *Stat. Sin.* **1996**, *6*, 831–860.

[123] Gelman, A.; Meng, X.-L. Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling. *Stat. Sci.* **1998**, *13*, 163–185.

[124] Gronau, Q. F.; Sarafoglou, A.; Matzke, D.; Ly, A.; Boehm, U.; Marsman, M.; Leslie, D. S.; Forster, J. J.; Wagenmakers, E.-J.; Steingroever, H. A Tutorial on Bridge Sampling. *J. Math. Psychol.* **2017**, *81*, 80–97.

[125] Shirts, M. R.; Bair, E.; Hooker, G.; Pande, V. S. Equilibrium Free Energies from Nonequilibrium Measurements Using Maximum-Likelihood Methods. *Phys. Rev. Lett.* **2003**, *91*, 140601.

[126] Shirts, M. R.; Pande, V. S. Comparison of Efficiency and Bias of Free Energies Computed by Exponential Averaging, the Bennett Acceptance Ratio, and Thermodynamic Integration. *J. Chem. Phys.* **2005**, *122*, 144107.

[127] de Ruiter, A.; Boresch, S.; Oostenbrink, C. Comparison of Thermodynamic Integration and Bennett Acceptance Ratio for Calculating Relative Protein-Ligand Binding Free Energies. *J. Comp. Chem.* **2013**, *34*, 1024–1034.

[128] Klimovich, P. V.; Shirts, M. R.; Mobley, D. L. Guidelines for the Analysis of Free Energy Calculations. *J. Comput.-Aided Mol. Des.* **2015**, *29*, 397–411.

[129] Bruckner, S.; Boresch, S. Efficiency of Alchemical Free Energy Simulations. II. Improvements for Thermodynamic Integration. *J. Comp. Chem.* **2011**, *32*, 1320–1333.

[130] Kong, A.; McCullagh, P.; Meng, X.-L.; Nicolae, D.; Tan, Z. A Theory of Statistical Models for Monte Carlo Integration. *J. R. Stat. Soc. Ser. B Stat. Method.* **2003**, *65*, 585–604.

[131] Shirts, M. R.; Chodera, J. D. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J. Chem. Phys.* **2008**, *129*, 124105.

[132] Ding, X. Bayesian Multistate Bennett Acceptance Ratio Methods. *J. Chem. Theory Comput.* **2024**, *20*, 1878–1888.

[133] Li, X. S.; Van Koten, B.; Dinner, A. R.; Thiede, E. H. Understanding the Sources of Error in MBAR through Asymptotic Analysis. *J. Chem. Phys.* **2023**, *158*, 214107.

[134] Politis, D. N.; White, H. Automatic Block-Length Selection for the Dependent Bootstrap. *Econom. Rev.* **2004**, *23*, 53–70.

[135] Jarzynski, C. Nonequilibrium Equality for Free Energy Differences. *Phys. Rev. Lett.* **1997**, *78*, 2690–2693.

[136] Crooks, G. E. Entropy Production Fluctuation Theorem and the Nonequilibrium Work Relation for Free Energy Differences. *Phys. Rev. E* **1999**, *60*, 2721–2726.

[137] Gumbart, J. C.; Roux, B.; Chipot, C. Standard Binding Free Energies from Computer Simulations: What Is the Best Strategy? *J. Chem. Theory Comput.* **2013**, *9*, 794–802.

[138] Zacharias, M.; Straatsma, T. P.; McCammon, J. A. Separation-shifted Scaling, a New Scaling Method for Lennard-Jones Interactions in Thermodynamic Integration. *J. Chem. Phys.* **1994**, *100*, 9025–9031.

[139] Salomon-Ferrer, R.; Götz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theory Comput.* **2013**, *9*, 3878–3888.

[140] Heinzelmann, G.; Gilson, M. K. Automation of Absolute Protein-Ligand Binding Free Energy Calculations for Docking Refinement and Compound Evaluation. *Sci Rep* **2021**, *11*, 1116.

[141] Fu, H.; Chen, H.; Cai, W.; Shao, X.; Chipot, C. BFEE2: Automated, Streamlined, and Accurate Absolute Binding Free-Energy Calculations. *J. Chem. Inf. Model.* **2021**, *61*, 2116–2123.

[142] Rizzi, A.; Chodera, J.; Naden, L.; Beauchamp, K.; Albanese, S.; Grinaway, P.; Prada-Gracia, D.; Rustenburg, B.; Ajsilveira; Saladi, S.; Boehm, K.; Gmach, J.; Rodríguez-Guerra, J. Choderalab/Yank: 0.25.2 - Bugfix Release. Zenodo, 2019.

[143] Huggins, D. J.; Biggin, P. C.; Dämgen, M. A.; Essex, J. W.; Harris, S. A.; Henchman, R. H.; Khalid, S.; Kuzmanic, A.; Laughton, C. A.; Michel, J.; others Biomolecular simulations: From dynamics and mechanisms to computational assays of biological activity. *WIREs Comput Mol Sci.* **2019**, *9*, e1393.

[144] De Simone, A.; Georgiou, C.; Ioannidis, H.; Gupta, A. A.; Juárez-Jiménez, J.; Doughty-Shenton, D.; Blackburn, E. A.; Wear, M. A.; Richards, J. P.; Barlow, P. N.; others A computationally designed binding mode flip leads to a novel class of potent tri-vector cyclophilin inhibitors. *Chem. Sci.* **2019**, *10*, 542–547.

[145] Breznik, M.; Ge, Y.; Bluck, J. P.; Briem, H.; Hahn, D. F.; Christ, C. D.; Mortier, J.; Mobley, D. L.; Meier, K. Prioritizing Small Sets of Molecules for Synthesis through *In-silico* Tools: A Comparison of Common Ranking Methods. *ChemMedChem* **2022**,

[146] Mey, A. S.; Juárez-Jiménez, J.; Hennessy, A.; Michel, J. Blinded predictions of binding modes and energies of HSP90-α ligands for the 2015 D3R grand challenge. *Bioorg. Med. Chem.* **2016**, *24*, 4890–4899.

[147] Mey, A. S.; Jiménez, J. J.; Michel, J. Impact of domain knowledge on blinded predictions of binding energies by alchemical free energy calculations. *J. Comput. Aided Mol. Des.* **2018**, *32*, 199–210.

[148] Loeffler, H. H.; Bosisio, S.; Duarte Ramos Matos, G.; Suh, D.; Roux, B.; Mobley, D. L.; Michel, J. Reproducibility of free energy calculations across different molecular simulation software packages. *J. Chem. Theory Comput.* **2018**, *14*, 5567–5582.

[149] Granadino-Roldan, J. M.; Mey, A. S.; Pérez González, J. J.; Bosisio, S.; Rubio-Martinez, J.; Michel, J. Effect of set up protocols on the accuracy of alchemical free energy calculation over a set of ACK1 inhibitors. *PLoS One* **2019**, *14*, e0213217.

[150] Lee, T.-S.; Allen, B. K.; Giese, T. J.; Guo, Z.; Li, P.; Lin, C.; McGee, T. D.; Pearlman, D. A.; Radak, B. K.; Tao, Y.; Tsai, H.-C.; Xu, H.; Sherman, W.; York, D. M. Alchemical Binding Free Energy Calculations in AMBER20: Advances and Best Practices for Drug Discovery. *J. Chem. Inf. Model.* **2020**, *60*, 5595–5623.

[151] Rizzi, A. et al. The SAMPL6 SAMPLing Challenge: Assessing the Reliability and Efficiency of Binding Free Energy Calculations. *J. Comput.-Aided Mol. Des.* **2020**, *34*, 601–633.

[152] Georgiou, C.; McNae, I.; Wear, M.; Ioannidis, H.; Michel, J.; Walkinshaw, M. Pushing the Limits of Detection of Weak Binding Using Fragment-Based Drug Discovery: Identification of New Cyclophilin Binders. *J. Mol. Biol.* **2017**, *429*, 2556–2570.

[153] Baumann, H. M.; Dybeck, E.; McClendon, C. L.; Pickard, F. C.; Gapsys, V.; Pérez-Benito, L.; Hahn, D. F.; Tresadern, G.; Mathiowetz, A. M.; Mobley, D. L. Broadening the Scope of Binding Free Energy Calculations Using a Separated Topologies Approach. *J. Chem. Theory Comput.* **2023**, *19*, 5058–5076.

[154] Hermans, J.; Shankar, S. The Free Energy of Xenon Binding to Myoglobin from Molecular Dynamics Simulation. *Isr. J. Chem.* **1986**, *27*, 225–227.

[155] Roux, B.; Nina, M.; Pomès, R.; Smith, J. Thermodynamic Stability of Water Molecules in the Bacteriorhodopsin Proton Channel: A Molecular Dynamics Free Energy Perturbation Study. *Biophys. J.* **1996**, *71*, 670–681.

[156] Hermans, J.; Wang, L. Inclusion of Loss of Translational and Rotational Freedom in Theoretical Estimates of Free Energies of Binding. Application to a Complex of Benzene and Mutant T4 Lysozyme. *J. Am. Chem. Soc.* **1997**, *119*, 2707–2714.

[157] Procacci, P.; Macchiagodena, M. On the NS-DSSB Unidirectional Estimates in the SAMPL6 SAMPLing Challenge. *J. Comput.-Aided Mol. Des.* **2021**, *35*, 1055–1065.

[158] Wang, J.; Deng, Y.; Roux, B. Absolute Binding Free Energy Calculations Using Molecular Dynamics Simulations with Restraining Potentials. *Biophys. J.* **2006**, *91*, 2798–2814.

[159] Fu, H.; Cai, W.; Hénin, J.; Roux, B.; Chipot, C. New Coarse Variables for the Accurate Determination of Standard Binding Free Energies. *J. Chem. Theory Comput.* **2017**, *13*, 5173–5178.

[160] Salari, R.; Joseph, T.; Lohia, R.; Hénin, J.; Brannigan, G. A Streamlined, General Approach for Computing Ligand Binding Free Energies and Its Application to GPCR-Bound Cholesterol. *J. Chem. Theory Comput.* **2018**, *14*, 6560–6573.

[161] Ebrahimi, M.; Hénin, J. Symmetry-Adapted Restraints for Binding Free Energy Calculations. *J. Chem. Theory Comput.* **2022**, *18*, 2494–2502.

[162] Santiago-McRae, E.; Ebrahimi, M.; Sandberg, J. W.; Brannigan, G.; Hénin, J. Computing Absolute Binding Affinities by Streamlined Alchemical Free Energy Perturbation (SAFEP) [Article v1.0]. *Living J. Comput. Mol. Sci.* **2023**, *5*, 2067–2067.

[163] Miyamoto, S.; Kollman, P. A. Absolute and Relative Binding Free Energy Calculations of the Interaction of Biotin and Its Analogs with Streptavidin Using Molecular Dynamics/Free Energy Perturbation Approaches. *Proteins* **1993**, *16*, 226–245.

[164] Gallicchio, E.; Levy, R. M. *Advances in Protein Chemistry and Structural Biology*; Elsevier, 2011; Vol. 85; pp 27–80.

[165] Mendoza-Martinez, C.; Papadourakis, M.; Llabrés, S.; Gupta, A. A.; Barlow, P. N.; Michel, J. Energetics of a Protein Disorder–Order Transition in Small Molecule Recognition. *Chem. Sci.* **2022**, *13*, 5220–5229.

[166] Qian, Y.; Cabeza de Vaca, I.; Vilseck, J. Z.; Cole, D. J.; Tirado-Rives, J.; Jorgensen, W. L. Absolute Free Energy of Binding Calculations for Macrophage Migration Inhibitory Factor in Complex with a Druglike Inhibitor. *J. Phys. Chem. B* **2019**, *123*, 8675–8685.

[167] *The PyMOL Molecular Graphics System Version 2.5.4*; Schrödinger, LLC, 2022.

[168] Mobley, D. L.; Chodera, J. D.; Dill, K. A. On the Use of Orientational Restraints and Symmetry Corrections in Alchemical Free Energy Calculations. *J. Chem. Phys.* **2006**, *125*, 084902.

[169] Fu, H.; Chen, H.; Blazhynska, M.; Goulard Coderc de Lacam, E.; Szczepaniak, F.; Pavlova, A.; Shao, X.; Gumbart, J. C.; Dehez, F.; Roux, B.; Cai, W.; Chipot, C. Accurate Determination of Protein:Ligand Standard Binding Free Energies from Molecular Dynamics Simulations. *Nat. Protoc.* **2022**, *17*, 1114–1141.

[170] Alibay, I. IAlibay/MDRestraintsGenerator: MDRestraintsGenerator 0.1.0. Zenodo, 2021.

[171] Huggins, D. J. Comparing the Performance of Different AMBER Protein Forcefields, Partial Charge Assignments, and Water Models for Absolute Binding Free Energy Calculations. *J. Chem. Theory Comput.* **2022**, *18*, 2616–2630.

[172] Bosisio, S.; Mey, A. S. J. S.; Michel, J. Blinded Predictions of Host-Guest Standard Free Energies of Binding in the SAMPL5 Challenge. *J. Comput.-Aided Mol. Des.* **2017**, *31*, 61–70.

[173] Papadourakis, M.; Bosisio, S.; Michel, J. Blinded predictions of standard binding free energies: lessons learned from the SAMPL6 challenge. *J. Comput. Aided Mol. Des.* **2018**, *32*, 1047–1058.

[174] Dziedzic, P.; Cisneros, J. A.; Robertson, M. J.; Hare, A. A.; Danford, N. E.; Baxter, R. H. G.; Jorgensen, W. L. Design, Synthesis, and Protein Crystallography of Biaryltriazoles as Potent Tautomerase Inhibitors of Macrophage Migration Inhibitory Factor. *J. Am. Chem. Soc.* **2015**, *137*, 2996–3003.

[175] Case, D. et al. Amber 2021. 2021.

[176] Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical p K_{a} Predictions. *J. Chem. Theory Comput.* **2011**, *7*, 525–537.

[177] Søndergaard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of p K_{a} Values. *J. Chem. Theory Comput.* **2011**, *7*, 2284–2295.

[178] Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V. H++ 3.0: Automating pK Prediction and the Preparation of Biomolecular Structures for Atomistic Molecular Modeling and Simulations. *Nucleic Acids Res.* **2012**, *40*, W537–W541.

[179] Swope, M. Direct Link between Cytokine Activity and a Catalytic Site for Macrophage Migration Inhibitory Factor. *EMBO J.* **1998**, *17*, 3534–3541.

[180] Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.

[181] O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminf.* **2011**, *3*, 33.

[182] Hedges, L.; Mey, A.; Laughton, C.; Gervasio, F.; Mulholland, A.; Woods, C.; Michel, J. BioSimSpace: An Interoperable Python Framework for Biomolecular Simulation. *JOSS* **2019**, *4*, 1831.

[183] Michel, J.; Verdonk, M. L.; Essex, J. W. Protein-Ligand Complexes: Computation of the Relative Free Energy of Different Scaffolds and Binding Modes. *J. Chem. Theory Comput.* **2007**, *3*, 1645–1655.

[184] Woods, C. J.; Mey, A. S. J. S.; Calabrò, G.; Michel, J. Sire Molecular Simulation Framework. 2019.

[185] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926–935.

[186] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

[187] Calabrò, G.; Woods, C. J.; Powlesland, F.; Mey, A. S. J. S.; Mulholland, A. J.; Michel, J. Elucidation of Nonadditive Effects in Protein–Ligand Binding Energies: Thrombin as a Case Study. *J. Phys. Chem. B* **2016**, *120*, 5340–5350.

[188] Hockney, R.; Goel, S.; Eastwood, J. Quiet High-Resolution Computer Models of a Plasma. *J. Comput. Phys.* **1974**, *14*, 148–158.

[189] Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. A Generalized Reaction Field Method for Molecular Dynamics Simulations. *J. Chem. Phys.* **1995**, *102*, 5451–5459.

[190] Deng, Y.; Roux, B. Calculation of Standard Binding Free Energies: Aromatic Molecules in the T4 Lysozyme L99A Mutant. *J. Chem. Theory Comput.* **2006**, *2*, 1255–1273.

[191] Orita, M.; Yamamoto, S.; Katayama, N.; Aoki, M.; Takayama, K.; Yamagiwa, Y.; Seki, N.; Suzuki, H.; Kurihara, H.; Sakashita, H.; Takeuchi, M.; Fujita, S.; Yamada, T.; Tanaka, A. Coumarin and Chromen-4-one Analogues as Tautomerase Inhibitors of Macrophage Migration Inhibitory Factor: Discovery and X-ray Crystallography. *J. Med. Chem.* **2001**, *44*, 540–547.

[192] Rogers, K. E.; Ortiz-Sánchez, J. M.; Baron, R.; Fajer, M.; de Oliveira, C. A. F.; McCammon, J. A. On the Role of Dewetting Transitions in Host–Guest Binding Free Energy Calculations. *J. Chem. Theory Comput.* **2013**, *9*, 46–53.

[193] Bhati, A. P.; Coveney, P. V. Large Scale Study of Ligand–Protein Relative Binding Free Energy Calculations: Actionable Predictions from Statistically Robust Protocols. *J. Chem. Theory Comput.* **2022**, *18*, 2687–2702.

[194] Ge, Y.; Wych, D. C.; Samways, M. L.; Wall, M. E.; Essex, J. W.; Mobley, D. L. Enhancing Sampling of Water Rehydration on Ligand Binding: A Comparison of Techniques. *J. Chem. Theory Comput.* **2022**, *18*, 1359–1381.

[195] Ben-Shalom, I. Y.; Lin, Z.; Radak, B. K.; Lin, C.; Sherman, W.; Gilson, M. K. Accounting for the central role of interfacial water in protein–ligand binding free energy calculations. *J. Chem. Theory Comput.* **2020**, *16*, 7883–7894.

[196] Ben-Shalom, I. Y.; Lin, C.; Kurtzman, T.; Walker, R. C.; Gilson, M. K. Simulating water exchange to buried binding sites. *J. Chem. Theory Comput.* **2019**, *15*, 2684–2691.

[197] Ross, G. A.; Russell, E.; Deng, Y.; Lu, C.; Harder, E. D.; Abel, R.; Wang, L. Enhancing water sampling in free energy calculations with grand canonical Monte Carlo. *J. Chem. Theory Comput.* **2020**, *16*, 6061–6076.

[198] Cisneros, J. A.; Robertson, M. J.; Valhondo, M.; Jorgensen, W. L. A Fluorescence Polarization Assay for Binding to Macrophage Migration Inhibitory Factor and Crystal Structures for Complexes of Two Potent Inhibitors. *J. Am. Chem. Soc.* **2016**, *138*, 8630–8638.

[199] Hahn, D. F.; König, G.; Hünenberger, P. H. Overcoming Orthogonal Barriers in Alchemical Free Energy Calculations: On the Relative Merits of λ-Variations, λ-Extrapolations, and Biasing. *J. Chem. Theory Comput.* **2020**, *16*, 1630–1645.

[200] Lapelosa, M.; Gallicchio, E.; Levy, R. M. Conformational transitions and convergence of absolute binding free energy calculations. *J. Chem. Theory Comput.* **2012**, *8*, 47–60.

[201] Khalak, Y.; Tresadern, G.; Aldeghi, M.; Baumann, H. M.; Mobley, D. L.; de Groot, B. L.; Gapsys, V. Alchemical Absolute Protein–Ligand Binding Free Energies for Drug Design. *Chem. Sci.* **2021**, *12*, 13958–13971.

[202] Aguayo-Ortiz, R.; Dominguez, L. Unveiling the Possible Oryzalin-Binding Site in the α-Tubulin of *Toxoplasma Gondii*. *ACS Omega* **2022**, *7*, 18434–18442.

[203] Wang, L.; Deng, Y.; Wu, Y.; Kim, B.; LeBard, D. N.; Wandschneider, D.; Beachy, M.; Friesner, R. A.; Abel, R. Accurate Modeling of Scaffold Hopping Transformations in Drug Discovery. *J. Chem. Theory Comput.* **2017**, *13*, 42–54.

[204] Hedges, L. O.; Bariami, S.; Burman, M.; Clark, F.; Cossins, B. P.; Hardie, A.; Herz, A. M.; Lukauskis, D.; Mey, A. S. J. S.; Michel, J.; Scheen, J.; Suruzhon, M.; Woods, C. J.; Wu, Z. A Suite of Tutorials for the BioSimSpace Framework for Interoperable Biomolecular Simulation [Article v1.0]. *Living J. Comput. Mol. Sci.* **2023**, *5*, 2375–2375.

[205] Schindler, C. E. M. et al. Large-Scale Assessment of Binding Free Energy Calculations in Active Drug Discovery Projects. *J. Chem. Inf. Model.* **2020**, *60*, 5457–5474.

[206] York, D. M. Modern Alchemical Free Energy Methods for Drug Discovery Explained. *ACS Phys. Chem Au* **2023**, *3*, 478–491.

[207] Weinhold, F. Metric Geometry of Equilibrium Thermodynamics. *J. Chem. Phys.* **1975**, *63*, 2479–2483.

[208] Crooks, G. E. Measuring Thermodynamic Length. *Phys. Rev. Lett.* **2007**, *99*, 100602.

[209] Baumann, H. M.; Gapsys, V.; de Groot, B. L.; Mobley, D. L. Challenges Encountered Applying Equilibrium and Nonequilibrium Binding Free Energy Calculations. *J. Phys. Chem. B* **2021**, *125*, 4241–4261.

[210] Michel, J.; Essex, J. W. Hit identification and binding mode predictions by rigorous free energy simulations. *J. Med. Chem.* **2008**, *51*, 6654–6664.

[211] Rocklin, G. J.; Mobley, D. L.; Dill, K. A. Separated Topologies—A Method for Relative Binding Free Energy Calculations Using Orientational Restraints. *J. Chem. Phys.* **2013**, *138*, 085104.

[212] Azimi, S.; Gallicchio, E. Binding Selectivity Analysis from Alchemical Receptor Hopping and Swapping Free Energy Calculations. *J. Phys. Chem. B* **2024**, *128*, 10841–10852.

[213] Konze, K. D.; Bos, P. H.; Dahlgren, M. K.; Leswing, K.; Tubert-Brohman, I.; Bortolato, A.; Robbason, B.; Abel, R.; Bhat, S. Reaction-Based Enumeration, Active Learning, and Free Energy Calculations To Rapidly Explore Synthetically Tractable Chemical Space and Optimize Potency of Cyclin-Dependent Kinase 2 Inhibitors. *J. Chem. Inf. Model.* **2019**, *59*, 3782–3793.

[214] Khalak, Y.; Tresadern, G.; Hahn, D. F.; De Groot, B. L.; Gapsys, V. Chemical Space Exploration with Active Learning and Alchemical Free Energies. *J. Chem. Theory Comput.* **2022**, *18*, 6259–6270.

[215] Gusev, F.; Gutkin, E.; Kurnikova, M. G.; Isayev, O. Active Learning Guided Drug Design Lead Optimization Based on Relative Binding Free Energy Modeling. *J. Chem. Inf. Model.* **2023**, *63*, 583–594.

[216] Thompson, J.; Walters, W. P.; Feng, J. A.; Pabon, N. A.; Xu, H.; Maser, M.; Goldman, B. B.; Moustakas, D.; Schmidt, M.; York, F. Optimizing Active Learning for Free Energy Calculations. *Artif. Intell. Life Sci.* **2022**, *2*, 100050.

[217] Eckmann, P.; Wu, D.; Heinzelmann, G.; Gilson, M. K.; Yu, R. MFBind: A Multi-Fidelity Approach for Evaluating Drug Compounds in Practical Generative Modeling. *arXiv* **2024**, DOI: 10.48550/arXiv.2402.10387.

[218] Gapsys, V.; Pérez-Benito, L.; Aldeghi, M.; Seeliger, D.; van Vlijmen, H.; Tresadern, G.; de Groot, B. L. Large Scale Relative Protein Ligand Binding Affinities Using Non-Equilibrium Alchemy. *Chem. Sci.* **2020**, *11*, 1140–1152.

[219] Gapsys, V.; Yildirim, A.; Aldeghi, M.; Khalak, Y.; van der Spoel, D.; de Groot, B. L. Accurate Absolute Free Energies for Ligand–Protein Binding Based on Non-Equilibrium Approaches. *Commun. Chem.* **2021**, *4*, 61.

[220] Wan, S.; Bhati, A. P.; Coveney, P. V. Comparison of Equilibrium and Nonequilibrium Approaches for Relative Binding Free Energy Predictions. *J. Chem. Theory Comput.* **2023**, acs.jctc.3c00842.

[221] Hsu, W.-T.; Piomponi, V.; Merz, P. T.; Bussi, G.; Shirts, M. R. Alchemical Metadynamics: Adding Alchemical Variables to Metadynamics to Enhance Sampling in Free Energy Calculations. *J. Chem. Theory Comput.* **2023**, *19*, 1805–1817.

[222] Wang, L.; Deng, Y.; Knight, J. L.; Wu, Y.; Kim, B.; Sherman, W.; Shelley, J. C.; Lin, T.; Abel, R. Modeling Local Structural Rearrangements Using FEP/REST: Application to Relative Binding Affinity Predictions of CDK2 Inhibitors. *J. Chem. Theory Comput.* **2013**, *9*, 1282–1293.

[223] Woods, C. J.; Essex, J. W.; King, M. A. The Development of Replica-Exchange-Based Free-Energy Methods. *J. Phys. Chem. B* **2003**, *107*, 13703–13710.

[224] Jiang, W.; Roux, B. Free Energy Perturbation Hamiltonian Replica-Exchange Molecular Dynamics (FEP/H-REMD) for Absolute Ligand Binding Free Energy Calculations. *J. Chem. Theory Comput.* **2010**, *6*, 2559–2565.

[225] Lagardère, L.; Maurin, L.; Adjoua, O.; El Hage, K.; Monmarché, P.; Piquemal, J.-P.; Hénin, J. Lambda-ABF: Simplified, Portable, Accurate, and Cost-Effective Alchemical Free-Energy Computation. *J. Chem. Theory Comput.* **2024**,

[226] Deng, Y.; Roux, B. Computation of Binding Free Energy with Molecular Dynamics and Grand Canonical Monte Carlo Simulations. *J. Chem. Phys.* **2008**, *128*, 115103.

[227] Ross, G. A.; Russell, E.; Deng, Y.; Lu, C.; Harder, E. D.; Abel, R.; Wang, L. Enhancing Water Sampling in Free Energy Calculations with Grand Canonical Monte Carlo. *J. Chem. Theory Comput.* **2020**, *16*, 6061–6076.

[228] Ross, G. A.; Bruce Macdonald, H. E.; Cave-Ayland, C.; Cabedo Martinez, A. I.; Essex, J. W. Replica-Exchange and Standard State Binding Free Energies with Grand Canonical Monte Carlo. *J. Chem. Theory Comput.* **2017**, *13*, 6373–6381.

[229] Ben-Shalom, I. Y.; Lin, Z.; Radak, B. K.; Lin, C.; Sherman, W.; Gilson, M. K. Accounting for the Central Role of Interfacial Water in Protein–Ligand Binding Free Energy Calculations. *J. Chem. Theory Comput.* **2020**, *16*, 7883–7894.

[230] Lee, T.-S.; Tsai, H.-C.; Ganguly, A.; York, D. M. ACES: Optimized Alchemically Enhanced Sampling. *J. Chem. Theory Comput.* **2023**, *19*, 472–487.

[231] Zheng, L.; Chen, M.; Yang, W. Random Walk in Orthogonal Space to Achieve Efficient Free-Energy Simulation of Complex Systems. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 20227–20232.

[232] König, G.; Glaser, N.; Schroeder, B.; Kubincová, A.; Hünenberger, P. H.; Riniker, S. An Alternative to Conventional λ-Intermediate States in Alchemical Free Energy Calculations: λ-Enveloping Distribution Sampling. *J. Chem. Inf. Model.* **2020**, *60*, 5407–5423.

[233] Li, H.; Fajer, M.; Yang, W. Simulated Scaling Method for Localized Enhanced Sampling and Simultaneous "Alchemical" Free Energy Simulations: A General Method for Molecular Mechanical, Quantum Mechanical, and Quantum Mechanical/Molecular Mechanical Simulations. *J. Chem. Phys.* **2007**, *126*, 024106.

[234] Sugita, Y.; Kitao, A.; Okamoto, Y. Multidimensional Replica-Exchange Method for Free-Energy Calculations. *J. Chem. Phys.* **2000**, *113*, 6042–6051.

[235] Kong, X.; Brooks, C. L. λ-Dynamics: A New Approach to Free Energy Calculations. *J. Chem. Phys.* **1996**, *105*, 2414–2423.

[236] Fu, H.; Chipot, C.; Shao, X.; Cai, W. Standard Binding Free-Energy Calculations: How Far Are We from Automation? *J. Phys. Chem. B* **2023**, *127*, 10459–10468.

[237] De Oliveira, C.; Leswing, K.; Feng, S.; Kanters, R.; Abel, R.; Bhat, S. FEP Protocol Builder: Optimization of Free Energy Perturbation Protocols Using Active Learning. *J. Chem. Inf. Model.* **2023**, *63*, 5592–5603.

[238] Koby, S. B.; Gutkin, E.; Patel, S.; Kurnikova, M. An Automated On-The-Fly Optimization of Resource Allocation for High-Throughput Protein-Ligand Binding Free Energy Simulations. 2023.

[239] Li, P.; Li, Z.; Wang, Y.; Dou, H.; Radak, B. K.; Allen, B. K.; Sherman, W.; Xu, H. Precise Binding Free Energy Calculations for Multiple Molecules Using an Optimal Measurement Network of Pairwise Differences. *J. Chem. Theory Comput.* **2022**, *18*, 650–663.

[240] Sivak, D. A.; Crooks, G. E. Thermodynamic Metrics and Optimal Paths. *Phys. Rev. Lett.* **2012**, *108*, 190602.

[241] Blondel, A. Ensemble variance in free energy calculations by thermodynamic integration: Theory, optimal "Alchemical" path, and practical solutions. *J. Comput. Chem.* **2004**, *25*, 985–993.

[242] Pham, T. T.; Shirts, M. R. Identifying Low Variance Pathways for Free Energy Calculations of Molecular Transformations in Solution Phase. *J. Chem. Phys.* **2011**, *135*, 034114.

[243] Reinhardt, M.; Grubmüller, H. Determining Free-Energy Differences Through Variationally Derived Intermediates. *J. Chem. Theory Comput.* **2020**, *16*, 3504–3512.

[244] König, G.; Ries, B.; Hünenberger, P. H.; Riniker, S. Efficient Alchemical Intermediate States in Free Energy Calculations Using λ-Enveloping Distribution Sampling. *J. Chem. Theory Comput.* **2021**, *17*, 5805–5815.

[245] Naden, L. N.; Pham, T. T.; Shirts, M. R. Linear Basis Function Approach to Efficient Alchemical Free Energy Calculations. 1. Removal of Uncharged Atomic Sites. *J. Chem. Theory Comput.* **2014**, *10*, 1128–1149.

[246] Naden, L. N.; Shirts, M. R. Linear Basis Function Approach to Efficient Alchemical Free Energy Calculations. 2. Inserting and Deleting Particles with Coulombic Interactions. *J. Chem. Theory Comput.* **2015**, *11*, 2536–2549.

[247] König, G.; Brooks, B. R.; Thiel, W.; York, D. M. On the Convergence of Multi-Scale Free Energy Simulations. *Mol. Simul.* **2018**, *44*, 1062–1081.

[248] Shenfeld, D. K.; Xu, H.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. Minimizing Thermodynamic Length to Select Intermediate States for Free-Energy Calculations and Replica-Exchange Simulations. *Phys. Rev. E* **2009**, *80*, 046705.

[249] Lindahl, V.; Lidmar, J.; Hess, B. Riemann Metric Approach to Optimal Sampling of Multidimensional Free-Energy Landscapes. *Phys. Rev. E* **2018**, *98*, 023312.

[250] Lundborg, M.; Lidmar, J.; Hess, B. On the Path to Optimal Alchemistry. *Protein J.* **2023**, *42*, 477–489.

[251] Minh, D. D. L. Alchemical Grid Dock (AlGDock): Binding Free Energy Calculations between Flexible Ligands and Rigid Receptors. *J. Comput. Chem.* **2020**, *41*, 715–730.

[252] Rizzi, A. Improving Efficiency and Scalability of Free Energy Calculations through Automatic Protocol Optimization. Ph.D. thesis, Weill Medical College of Cornell University, United States – New York, 2020.

[253] Zeng, J.; Qian, Y. Adaptive Lambda Schemes for Efficient Relative Binding Free Energy Calculation. *J. Comput. Chem.* **2023**, jcc.27287.

[254] Midgley, S.; Bariami, S.; Habgood, M.; Mackey, M. Adaptive Lambda Scheduling: A method for computational efficiency in Free Energy Perturbation simulations. *ChemRxiv* **2024**, DOI: 10.26434/chemrxiv-2024-3fxcj.

[255] Zhang, S.; Giese, T. J.; Lee, T.-S.; York, D. M. Alchemical Enhanced Sampling with Optimized Phase Space Overlap. *J. Chem. Theory Comput.* **2024**,

[256] Sun, Z. X.; Wang, X. H.; Zhang, J. Z. H. BAR-based Optimum Adaptive Sampling Regime for Variance Minimization in Alchemical Transformation. *Phys. Chem. Chem. Phys.* **2017**, *19*, 15005–15020.

[257] Bhati, A. P.; Wan, S.; Wright, D. W.; Coveney, P. V. Rapid, Accurate, Precise, and Reliable Relative Free Energy Prediction Using Ensemble Based Thermodynamic Integration. *J. Chem. Theory Comput.* **2017**, *13*, 210–222.

[258] Bhati, A. P.; Wan, S.; Hu, Y.; Sherborne, B.; Coveney, P. V. Uncertainty Quantification in Alchemical Free Energy Methods. *J. Chem. Theory Comput.* **2018**, *14*, 2867–2880.

[259] Knapp, B.; Ospina, L.; Deane, C. M. Avoiding False Positive Conclusions in Molecular Simulation: The Importance of Replicas. *J. Chem. Theory Comput.* **2018**, *14*, 6127–6138.

[260] Adler, M.; Beroza, P. Improved Ligand Binding Energies Derived from Molecular Dynamics: Replicate Sampling Enhances the Search of Conformational Space. *J. Chem. Inf. Model.* **2013**, *53*, 2065–2072.

[261] Cowles, M. K.; Carlin, B. P. Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *J. Am. Stat. Assoc.* **1996**, *91*, 883–904.

[262] Chodera, J. D. A Simple Method for Automated Equilibration Detection in Molecular Simulations. *J. Chem. Theory Comput.* **2016**, *12*, 1799–1805.

[263] Roy, V. Convergence Diagnostics for Markov Chain Monte Carlo. *Annu. Rev. Stat. Appl.* **2020**, *7*, 387–412.

[264] Gelman, A.; Rubin, D. B. Inference from Iterative Simulation Using Multiple Sequences. *Statist. Sci.* **1992**, *7*.

[265] Geweke, J. Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. 1991; Staff Report 148, Federal Reserve Bank of Minneapolis.

[266] García, E. J.; Hasse, H. Studying Equilibria of Polymers in Solution by Direct Molecular Dynamics Simulations: Poly(N-isopropylacrylamide) in Water as a Test Case. *Eur. Phys. J. Spec. Top.* **2019**, *227*, 1547–1558.

[267] Woods, C. J.; Hedges, L. O.; Mulholland, A. J.; Malaisree, M.; Tosco, P.; Loeffler, H. H.; Suruzhon, M.; Burman, M.; Bariami, S.; Bosisio, S.; Calabro, G.; Clark, F.; Mey, A. S. J. S.; Michel, J. Sire: An Interoperability Engine for Prototyping Algorithms and Exchanging Information between Molecular Simulation Programs. *J. Chem. Phys.* **2024**, *160*, 202503.

[268] Yu, Z.; Batista, E. R.; Yang, P.; Perez, D. Acceleration of Solvation Free Energy Calculation via Thermodynamic Integration Coupled with Gaussian Process Regression and Improved Gelman–Rubin Convergence Diagnostics. *J. Chem. Theory Comput.* **2024**, *20*, 2570–2581.

[269] Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J. Mol. Graphics Modell.* **2006**, *25*, 247–260.

[270] Alibay, I.; Magarkar, A.; Seeliger, D.; Philip, B. C. Benchmark Set Inputs for Absolute Binding Free Energy Calculations of Fragment Optimisations. 2022; `https://doi.org/10.5281/zenodo.5913469`.

[271] Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25.

[272] Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

[273] Bernetti, M.; Bussi, G. Pressure Control Using Stochastic Cell Rescaling. *J. Chem. Phys.* **2020**, *153*, 114107.

[274] Clark, F.; Robb, G.; Cole, D. J.; Michel, J. Comparison of Receptor–Ligand Restraint Schemes for Alchemical Absolute Binding Free Energy Calculations. *J. Chem. Theory Comput.* **2023**, *19*, 3686–3704.

[275] Boresch, S. On Analytical Corrections for Restraints in Absolute Binding Free Energy Calculations. *J. Chem. Inf. Model.* **2024**, *64*, 3605–3609.

[276] Eastman, P. et al. OpenMM 8: Molecular Dynamics Simulation with Machine Learning Potentials. *J. Phys. Chem. B* **2024**, *128*, 109–116.

[277] Zhang, Z.; Liu, X.; Yan, K.; Tuckerman, M. E.; Liu, J. Unified Efficient Thermostat Scheme for the Canonical Ensemble with Holonomic or Isokinetic Constraints via Molecular Dynamics. *J. Phys. Chem. A* **2019**, *123*, 6056–6079.

[278] Papadourakis, M.; Bosisio, S.; Michel, J. Blinded Predictions of Standard Binding Free Energies: Lessons Learned from the SAMPL6 Challenge. *J. Comput.-Aided Mol. Des.* **2018**, *32*, 1047–1058.

[279] Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A. The Experimental Uncertainty of Heterogeneous Public Ki Data. *J. Med. Chem.* **2012**, *55*, 5165–5173.

[280] Petrov, D.; Perthold, J. W.; Oostenbrink, C.; de Groot, B. L.; Gapsys, V. Guidelines for Free-Energy Calculations Involving Charge Changes. *J. Chem. Theory Comput.* **2024**, *20*, 914–925.

[281] Landrum, G. RDKit: Open-source Cheminformatics. 2006.

[282] Morton, A.; Baase, W. A.; Matthews, B. W. Energetic Origins of Specificity of Ligand Binding in an Interior Nonpolar Cavity of T4 Lysozyme. *Biochemistry* **1995**, *34*, 8564–8575.

[283] Michelsen, K.; Jordan, J. B.; Lewis, J.; Long, A. M.; Yang, E.; Rew, Y.; Zhou, J.; Yakowec, P.; Schnier, P. D.; Huang, X.; Poppe, L. Ordering of the N-Terminus of Human MDM2 by Small Molecule Inhibitors. *J. Am. Chem. Soc.* **2012**, *134*, 17059–17067.

[284] Obach, R. S.; Walker, G. S.; Sharma, R.; Jenkinson, S.; Tran, T. P.; Stepan, A. F. Lead Diversification at the Nanomole Scale Using Liver Microsomes and Quantitative Nuclear Magnetic Resonance Spectroscopy: Application to Phosphodiesterase 2 Inhibitors. *J. Med. Chem.* **2018**, *61*, 3626–3640.

[285] Pal, R. K.; Gallicchio, E. Perturbation Potentials to Overcome Order/Disorder Transitions in Alchemical Binding Free Energy Calculations. *J. Chem. Phys.* **2019**, *151*, 124116.

[286] Wan, S.; Bhati, A. P.; Wade, A. D.; Coveney, P. V. Ensemble-Based Approaches Ensure Reliability and Reproducibility. *J. Chem. Inf. Model.* **2023**,

[287] Kruskal, W. H.; and Wallis, W. A. Use of Ranks in One-Criterion Variance Analysis. *J. Am. Stat. Assoc.* **1952**, *47*, 583–621.

[288] Alibay, I.; Magarkar, A.; Seeliger, D.; Philip, B. C. Sampled $\Delta H/\Delta\lambda$ and ΔH Data from ABFE Calculations of 10 Ligands Bound to Cyclophilin D. 2022; `https://doi.org/10.5281/zenodo.5904110`.

[289] Cooke, B.; Schmidler, S. C. Statistical Prediction and Molecular Dynamics Simulation. *Biophys. J.* **2008**, *95*, 4497–4511.

[290] Vats, D.; Knudson, C. Revisiting the Gelman–Rubin Diagnostic. *Stat. Sci.* **2021**, *36*, 518–529.

[291] Nguyen, T. H.; Minh, D. D. L. Intermediate Thermodynamic States Contribute Equally to Free Energy Convergence: A Demonstration with Replica Exchange. *J. Chem. Theory Comput.* **2016**, *12*, 2154–2161.

[292] Hsu, W.-T.; Shirts, M. R. Ensemble of Expanded Ensembles: A Generalized Ensemble Approach with Enhanced Flexibility and Parallelizability. *arXiv* **2023**, *arXiv:2308.06938*, DOI: 10.48550/arXiv.2308.06938.

[293] Nicholls, A. Confidence Limits, Error Bars and Method Comparison in Molecular Modeling. Part 2: Comparing Methods. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 103–126.

[294] Grädler, U.; Schwarz, D.; Blaesse, M.; Leuthner, B.; Johnson, T. L.; Bernard, F.; Jiang, X.; Marx, A.; Gilardone, M.; Lemoine, H.; Roche, D.; Jorand-Lebrun, C. Discovery of Novel Cyclophilin D Inhibitors Starting from Three Dimensional Fragments with Millimolar Potencies. *Bioorg. Med. Chem. Lett.* **2019**, *29*, 126717.

[295] White, K. P. An Effective Truncation Heuristic for Bias Reduction in Simulation Output. *Simulation* **1997**, *69*, 323–334.

[296] Clark, F.; Robb, G. R.; Cole, D. J.; Michel, J. Automated Adaptive Absolute Binding Free Energy Calculations. *J. Chem. Theory Comput.* **2024**, *20*, 7806–7828.

[297] Fan, S.; Iorga, B. I.; Beckstein, O. Prediction of Octanol-Water Partition Coefficients for the SAMPL6-logP Molecules Using Molecular Dynamics Simulations with OPLS-AA, AMBER and CHARMM Force Fields. *J. Comput. Aided Mol. Des.* **2020**, *34*, 543–560.

[298] Hoad, K.; Robinson, S.; Davies, R. Automating Warm-up Length Estimation. *J. Oper. Res. Soc.* **2010**, *61*, 1389–1403.

[299] Pasupathy, R.; Schmeiser, B. The Initial Transient in Steady-State Point Estimation: Contexts, a Bibliography, the MSE Criterion, and the MSER Statistic. Proc. 2010 Winter Simul. Conf. 2010; pp 184–197.

[300] Spratt, S. Heuristics for the Startup Problem. M.Sc. thesis, Department of Systems Engineering, University of Virginia, 1998.

[301] Oliveira, F. L.; Luan, B.; Esteves, P. M.; Steiner, M.; Neumann Barros Ferreira, R. pyMSER-An Open-Source Library for Automatic Equilibration Detection in Molecular Simulations. *J. Chem. Theory Comput.* **2024**, *20*, 8559–8568.

[302] Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Seok, C.; Dill, K. A. Use of the Weighted Histogram Analysis Method for the Analysis of Simulated and Parallel Tempering Simulations. *J. Chem. Theory Comput.* **2007**, *3*, 26–41.

[303] Flyvbjerg, H.; Petersen, H. G. Error Estimates on Averages of Correlated Data. *J. Chem. Phys.* **1989**, *91*, 461–466.

[304] Vogelsang, T. J.; Yang, J. Exactly/Nearly Unbiased Estimation of Autocovariances of a Univariate Time Series With Unknown Mean. *J. Time Ser. Anal.* **2016**, *37*, 723–740.

[305] Geyer, C. J. Practical Markov Chain Monte Carlo. *Stat. Sci.* **1992**, *7*, 473–483.

[306] Geyer, C. J. Markov Chain Monte Carlo Maximum Likelihood. Computing Science and Statistics, Proceedings of the 23rd Symposium on the Interface. 1991; pp 156–163.

[307] Alexopoulos, C.; Seila, A. F. Implementing the Batch Means Method in Simulation Experiments. Proc. 28th Winter Simul. Conf. USA, 1996; pp 214–221.

[308] White, K.; Cobb, M.; Spratt, S. A Comparison of Five Steady-State Truncation Heuristics for Simulation. Proc. 32nd Winter Simul. Conf. USA, 2000; pp 755–760.

[309] Meketon, M. S.; Schmeiser, B. Overlapping Batch Means: Something for Nothing? Proc. 16th Conf. Winter Simul. Dallas, TX, 1984; pp 226–230.

[310] Song, W. T.; Schmeiser, B. W. Variance of the Sample Mean: Properties and Graphs of Quadratic-Form Estimators. *Oper. Res.* **1993**, *41*, 501–517.

[311] Bartlett, M. S. Periodogram Analysis and Continuous Spectra. *Biometrika* **1950**, *37*, 1–16.

[312] Politis, D. N.; Romano, J. P. In *Exploring the Limits of Bootstrap*; LePage, R., Billard, L., Eds.; John Wiley: New York, 1992; pp 263–270.

[313] Künsch, H. R. The Jackknife and the Bootstrap for General Stationary Observations. *Ann. Stat.* **1989**, *17*, 1217–1241.

[314] Liu, R. Y.; Singh, K. In *Exploring the Limits of Bootstrap*; LePage, R., Billard, L., Eds.; Wiley: New York, 1992; pp 225–248.

[315] Straatsma, T.; Berendsen, H.; Stam, A. Estimation of Statistical Errors in Molecular Simulation Calculations. *Mol. Phys.* **1986**, *57*, 89–95.

[316] Wolff, U. Monte Carlo Errors with Less Errors. *Comput. Phys. Commun.* **2004**, *156*, 143–153.

[317] Liu, Y.; Vats, D.; Flegal, J. M. Batch Size Selection for Variance Estimators in MCMC. *Methodol. Comput. Appl.* **2022**, *24*, 65–93.

[318] Evertz, H. G. The Loop Algorithm. *Adv. Phys.* **2003**, *52*, 1–66.

[319] Thompson, M. B. A Comparison of Methods for Computing Autocorrelation Time. *arXiv* **2010**, *arXiv:1011.0175*, DOI: 10.48550/arXiv.1011.0175.

[320] Jonsson, M. Standard Error Estimation by an Automated Blocking Method. *Phys. Rev. E* **2018**, *98*, 043304.

[321] Politis, D. N. Adaptive Bandwidth Choice. *J. NONPARAMETR. STAT.* **2003**, *15*, 517–533.

[322] Goodman, J.; Weare, J. Ensemble Samplers with Affine Invariance. *Commun. Appl. Math. Comput. Sci.* **2010**, *5*, 65–80.

[323] Hess, B. Determining the Shear Viscosity of Model Liquids from Molecular Dynamics Simulations. *J. Chem. Phys.* **2002**, *116*, 209–217.

[324] Franklin, W. W.; White, K. P. Stationarity Tests and MSER-5: Exploring the Intuition behind Mean-Squared-Error-Reduction in Detecting and Correcting Initialization Bias. 2008 Winter Simul. Conf. Miami, FL, USA, 2008; pp 541–546.

[325] Clark, F. Sampled $\Delta H/\Delta\lambda$ from Non-Adaptive ABFE Calculations. 2024; `https://doi.org/10.5281/zenodo.11520013`.

[326] Janke, W.; Sauer, T. Application of the Multicanonical Multigrid Monte Carlo Method to the Two-Dimensional Φ^4-Problem: Autocorrelations and Interface Tension. *J. Stat. Phys.* **1995**, *78*, 759–798.

[327] Flegal, J. M.; Jones, G. L. Batch Means and Spectral Variance Estimators in Markov Chain Monte Carlo. *Ann. Stat.* **2010**, *38*, 1034–1070.

[328] Hoad, K.; Robinson, S. Implementing MSER-5 in Commercial Simulation Software and Its Wider Implications. Proc. 2011 Winter Simul. Conf. WSC. Phoenix, AZ, USA, 2011; pp 495–503.

[329] Clark, F. Data to Reproduce "Robust Automated Equilibration Detection for Molecular Simulations"'. 2024; `https://doi.org/10.5281/zenodo.13902735`.

[330] Grossfield, A.; Patrone, P. N.; Roe, D. R.; Schultz, A. J.; Siderius, D.; Zuckerman, D. M. Best Practices for Quantification of Uncertainty and Sampling Quality in Molecular Simulations [Article v1.0]. *Living J. Comput. Mol. Sci.* **2019**, *1*, 5067–5067.

[331] Gowers, R. J.; Farmahini, A. H.; Friedrich, D.; Sarkisov, L. Automated Analysis and Benchmarking of GCMC Simulation Programs in Application to Gas Adsorption. *Mol. Simul.* **2018**, *44*, 309–321.

[332] Song, W. T.; Schmeiser, B. W. Optimal Mean-Squared-Error Batch Sizes. *Manag. Sci.* **1995**, *41*, 110–123.

[333] Song, W. T. On the Estimation of Optimal Batch Sizes in the Analysis of Simulation Output. *Eur. J. Oper. Res.* **1996**, *88*, 304–319.

[334] Patton, A.; Politis, D. N.; White, H. Correction to "Automatic Block-Length Selection for the Dependent Bootstrap" by D. Politis and H. White. *Econom. Rev.* **2009**, *28*, 372–375.

[335] Meng, X.-L.; Schilling, S. Warp Bridge Sampling. *J. Compt. Graph. Stat.* **2002**, *11*, 552–586.

[336] Tan, Z.; Gallicchio, E.; Lapelosa, M.; Levy, R. M. Theory of Binless Multi-State Free Energy Estimation with Applications to Protein-Ligand Binding. *J. Chem. Phys.* **2012**, *136*, 144102.

[337] Harris, C. R. et al. Array Programming with NumPy. *Nature* **2020**, *585*, 357–362.

[338] Lam, S. K.; Pitrou, A.; Seibert, S. Numba: A LLVM-based Python JIT Compiler. Proc. Second Workshop LLVM Compil. Infrastruct. HPC. New York, NY, USA, 2015; pp 1–6.

[339] Seabold, S.; Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. 9th Python in Science Conference. 2010.

[340] Banerjee, A.; Vats, D. Efficient Multivariate Initial Sequence Estimators for MCMC. *arXiv.org* **2024**, *arXiv:2406.15874v1*, DOI: 10.48550/arXiv.2406.15874v1.

[341] Agarwal, M.; Vats, D. Globally Centered Autocovariances in MCMC. *J. Comput. Graph. Stat.* **2022**, *31*, 629–638.

[342] Shivanyuk, A. N.; Ryabukhin, S. V.; Tolmachev, A.; Bogolyubsky, A.; Mykytenko, D.; Chupryna, A.; Heilman, W.; Kostyuk, A. Enamine real database: Making chemical diversity real. *Chem. Today* **2007**, *25*, 58–59.

[343] Lionta, E.; Spyrou, G.; K. Vassilatis, D.; Cournia, Z. Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances. *Curr. Top. Med. Chem.* **2014**, *14*, 1923–1938.

[344] Cutrona, K. J.; Newton, A. S.; Krimmer, S. G.; Tirado-Rives, J.; Jorgensen, W. L. Metadynamics as a Postprocessing Method for Virtual Screening with Application to the Pseudokinase Domain of JAK2. *J. Chem. Inf. Model.* **2020**, *60*, 4403–4415.

[345] Malmstrom, R. D.; Watowich, S. J. Using Free Energy of Binding Calculations To Improve the Accuracy of Virtual Screening Predictions. *J. Chem. Inf. Model.* **2011**, *51*, 1648–1655.

[346] Gorantla, R.; Kubincová, A.; Suutari, B.; Cossins, B. P.; Mey, A. S. J. S. Benchmarking Active Learning Protocols for Ligand-Binding Affinity Prediction. *J. of Chem. Inf. Model.* **2024**, *64*, 1955–1965.

[347] Loeffler, H. H.; Wan, S.; Klähn, M.; Bhati, A. P.; Coveney, P. V. Optimal Molecular Design: Generative Active Learning Combining REINVENT with Precise Binding Free Energy Ranking Simulations. *J. Chem. Theory Comput.* **2024**, *20*, 8308–8328.

[348] Crivelli-Decker, J. E.; Beckwith, Z.; Tom, G.; Le, L.; Khuttan, S.; Salomon-Ferrer, R.; Beall, J.; Gómez-Bombarelli, R.; Bortolato, A. Machine Learning Guided AQFEP: A Fast and Efficient Absolute Free Energy Perturbation Solution for Virtual Screening. *J. Chem. Theory Comput.* **2024**,

[349] Heinzelmann, G.; Huggins, D. J.; Gilson, M. K. BAT2: An Open-Source Tool for Flexible, Automated, and Low Cost Absolute Binding Free Energy Calculations. *J. Chem. Theory Comput.* **2024**, *20*, 6518–6530.

[350] Bardelle, C.; Cross, D.; Davenport, S.; Kettle, J. G.; Ko, E. J.; Leach, A. G.; Mortlock, A.; Read, J.; Roberts, N. J.; Robins, P.; Williams, E. J. Inhibitors of the Tyrosine Kinase EphB4. Part 1: Structure-based Design and Optimization of a Series of 2,4-Bis-Anilinopyrimidines. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 2776–2780.

[351] Bardelle, C.; Coleman, T.; Cross, D.; Davenport, S.; Kettle, J. G.; Ko, E. J.; Leach, A. G.; Mortlock, A.; Read, J.; Roberts, N. J.; Robins, P.; Williams, E. J. Inhibitors of the Tyrosine Kinase EphB4. Part 2: Structure-based Discovery and Optimisation of 3,5-Bis Substituted Anilinopyrimidines. *Bioorganic & Medicinal Chemistry Letters* **2008**, *18*, 5717–5721.

[352] Murray, C. W. et al. Fragment-Based Drug Discovery Applied to Hsp90. Discovery of Two Lead Series with High Ligand Efficiency. *J. of Med. Chem.* **2010**, *53*, 5942–5955.

[353] Liang, J. et al. Lead Identification of Novel and Selective TYK2 Inhibitors. *Eur. J. Med. Chem.* **2013**, *67*, 175–187.

[354] Liang, J. et al. Lead Optimization of a 4-Aminopyridine Benzamide Scaffold To Identify Potent, Selective, and Orally Bioavailable TYK2 Inhibitors. *J. Med. Chem.* **2013**, *56*, 4521–4536.

[355] Hahn, D. F.; Wagner, J. R. Protein-Ligand Benchmark Dataset for Free Energy Calculations. 2022; `https://doi.org/10.5281/zenodo.6600875`.

[356] Goldstein, D. M. et al. Discovery of 6-(2,4-Difluorophenoxy)-2-[3-Hydroxy-1-(2-Hydroxyethyl)Propylamino]-8-Methyl-8 *H* -Pyrido[2,3- *d*]Pyrimidin-7-One (Pamapimod) and 6-(2,4-Difluorophenoxy)-8-Methyl-2-(Tetrahydro-2 *H* -Pyran-4-Ylamino)Pyrido[2,3- *d*]Pyrimidin-7(8 *H*)-One (R1487) as Orally Bioavailable and Highly Selective Inhibitors of P38α Mitogen-Activated Protein Kinase. *J. Med. Chem.* **2011**, *54*, 2255–2265.

[357] Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

[358] Genheden, S.; Ryde, U. The MM/PBSA and MM/GBSA Methods to Estimate Ligand-Binding Affinities. *Expert Opin. Drug Discov.* **2015**, *10*, 449–461.

[359] Miller, E. B. et al. Reliable and Accurate Solution to the Induced Fit Docking Problem for Protein–Ligand Binding. *J. Cheem. Theory Comput.* **2021**, *17*, 2630–2639.

[360] Sabanés Zariquiey, F.; Pérez, A.; Majewski, M.; Gallicchio, E.; De Fabritiis, G. Validation of the Alchemical Transfer Method for the Estimation of Relative Binding Affinities of Molecular Series. *J. Chem. Inf. Model.* **2023**, *63*, 2438–2444.

[361] Jorgensen, W. L.; Buckner, J. K.; Boudon, S.; Tirado-Rives, J. Efficient Computation of Absolute Free Energies of Binding by Computer Simulations. Application to the Methane Dimer in Water. *J. Chem. Phys.* **1988**, *89*, 3742–3746.

[362] Shirts, M. R. Reweighting from the mixture distribution as a better way to describe the multistate Bennett acceptance ratio. *arXiv preprint arXiv:1704.00891* **2017**, *arXiv:1704.00891*, DOI: 10.48550/arXiv.1704.00891.

[363] Merski, M.; Fischer, M.; Balius, T. E.; Eidam, O.; Shoichet, B. K. Homologous Ligands Accommodated by Discrete Conformations of a Buried Cavity. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 5039–5044.

[364] Kumar, R.; Carroll, C.; Hartikainen, A.; Martin, O. ArviZ a Unified Library for Exploratory Analysis of Bayesian Models in Python. *JOSS* **2019**, *4*, 1143.