CALIFORNIA STATE UNIVERSITY, NORTHRIDGE

AUTOMATED, EFFICIENT AND RIGOROUS ABSOLUTE BINDING FREE ENERGY CALCULATIONS

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Chemistry

By

Steven Ayoub Jr

May 2023

The thesis of Steven Ayoub Jr is approved:

Ravi Abrol	Date
Maosheng Miao	Date
Tyler Luchko, Chair	Date

California State University, Northridge

Dedication

I would like to dedicate this work to my parents, specifically my mother. My mother has always been my rock. She has been there to love, support and encourage me when things get hard. My father has encouraged me to achieve a higher education and continue my academic career. And finally, I would like to dedicate this work to my older sister. She has never failed to support me and cheer me on.

Acknowledgements

I would like to thank Dr.Maria Elena Zavala, the director of the NIH Bridges for the Doctorate Program at CSUN, for providing mentor support throughout my master's. I'd also like to acknowledge and thank Jesus Hernandez, a fellow chemistry graduate student, who has been a great classmate and friend. Also, like to thank my girlfriend Martha Avila-Zavala who has supported me in many ways but also cheered me on and pushed me to keep going. Above all, I would like to acknowledge and thank my supervisor, Dr. Tyler Luchko, who has been an amazing mentor and teacher throughout my master's journey. He is always approachable, present, genuine and has provided me with an abundance of guidance for my project. I feel honored to have worked under Dr. Luchko. One day, I hope to share my knowledge with future young aspiring research students as he did with me.

Table of Contents

List of Tables

List of Figures

ABSTRACT

AUTOMATED, EFFICIENT AND RIGOROUS ABSOLUTE BINDING FREE ENERGY CALCULATIONS

By

Steven Ayoub Jr

Master of Science in Chemistry

Identifying small drug-like molecules for therapeutic targets of interest is an ongoing effort in drug discovery. Accurate absolute binding free energy calculations (ABFEs) in silico, can reduce the time and cost of exploring a diverse set of potential drug candidates that may have been overlooked experimentally. ABFEs can overcome many challenges faced by relative binding free energy calculations when screening a diverse set of molecules but these calculations can be difficult to implement and perform. Currently ABFE calculations are dominated by expensive explicit solvent models or the use of severe approximated end state methods. For example, explicit solvent models must simulate every atom and, in case of water occupancy of a buried site, large potential energy barrier can lead to slow water exchange and insufficient sampling. Additionally, Lennard-Jones soft-core potentials must be used during the creation or annihilation of particles, but this creates additional complexity when performing these ABFE calculations. To circumvent these drawbacks, we introduced the software ISDDM.py, an automated Python workflow that implements an approach that uses faster implicit solvents, such as generalized Born (GB), which greatly reduces computational costs. We adapted the double decoupling method (DDM), which uses conformational restraints, and paired it with GB solvent to enhance convergence. Our work here attempts to reduce the computational expense of absolute binding free energies

by using implicit solvent models and the ISDDM.py package to handle all the input files, monitor the simulations and perform post-analysis. The software was tested on 83 unique host-guest complex systems and reports encouraging results in estimating binding free energies. We saw overall good correlation with experiments and out outperformed most the SAMPL4 and SAMPL9 challenge submission's RBFE metrics, though large systematic errors were present in the raw results. From our raw results, we saw a significant bias associated with the charge of the guest molecules, where positively charged guest molecules had significantly underestimated binding free energy predictions, and negatively charge guests were overestimated. Looking at all systems as a whole we report an correlation coefficient of 0.76 with an RMSE of 4.1 kcal/mol and a slope of 1.8. By applying a simple linear correction base on guest charge we observed much better correlation with and deviation from experiment, with RMSE values less than 1.06 kcal/mol and a correlation coefficient of 0.87. An MAE of 0.9 kcal/mol and MSE of 0 kcal/mol

Chapter 1

Introduction

From the plagues of history, to influenza epidemics, cancer and the most recent COVID-19 virus outbreak, the world population is constantly facing life-threatening events and diseases. Drug discovery, production, and implementation have been a huge part of preventing or improving outcomes from these events and diseases. Modern drug development requires the screening of a vast and diverse chemical space, with the aim to simultaneously optimize properties such as protein-ligand affinity. However, this process is often risky, costly, and very time-consuming. Currently, the conventional process of a potential drug candidate to reach the market has been estimated to take an average of 14 years, reaching up to 1.8 billion USD and an attrition rate as high as 96% [1]. Therefore, the ability to efficiently and quickly scan a large chemical for potential drug candidates is a major focus in drug discovery.

Attempts to exploring chemical space date back to the early drug discovery period (before the 19th century), during which chemists expanded their knowledge of chemical space at an exponential rate, from a handful of new substances in 1800 up to 11,000 in 1868 [2] to the drug discovery and drug development period (20th century and beyond). It was not until the "post-genome era," during the late 20th century that a huge surge of protein crystal structures were made publicly available and became targets of therapeutic interest for small-molecule drug discovery [3]. This resulted in a shift of drug discovery strategies that focused largely on high-throughput screening (HTS), leading to the identification of multiple therapeutic drug targets and the creation of very large compound libraries [4]. Despite this, HTS operates on the fundamental basis that assumes no prior knowledge of the drug binding site on the target protein. This led to an "irrational" drug design approach in which many pharmaceutical companies saw a decrease in launched drug candidates that were successful in the market. [4]

As an alternative approach, computer-aided drug discovery (CADD) provides a rational structure-based approach that can yield valuable information about intermolecular interactions between protein and ligand systems, as well as the binding affinity. CADD has already been successfully applied in all areas of drug discovery and has brought new drug compounds to the market for diverse diseases, including HIV-1-inhibiting drugs (amprenavir [5] and raltitrexed [6]), as well as topoisomerase II and IV inhibitor, such as norfloxacin[7].The use of molecular simulations in drug discovery has had a major impact in the past decade, as the rapid development of computer hardware and new algorithms has allowed computations to becomes time-affordable. The computational microscope provided by CADD shows how ligands bind to a targeted receptor, proteins interact with each other, and proteins aggregate in the cell membrane. It is not only important for rational drug design but also for the field of biophysics and biochemistry. An example is molecular docking programs which provides a fast and efficient way to screen a large number of potential drug candidates. [8] Molecular docking has been applied to screen ultra large libraries contain-

ing millions of molecules. [9]

In recent years, modern algorithms and rapid development of faster graphics processor units (GPUs) [10, 11, 12] have enabled great strides in estimating protein-ligand molecule binding free energies. Binding free energy calculations are readily used in drug discovery to rank potential candidates by their binding affinity to a host target molecule [13]. Binding free energy calculations paired with molecular dynamics (MD) offers the ability to use full atomic description and have been proposed to improve upon docking scoring methods [14]. MD simulations compute the time evolution of a system and are used as a sampling tool to recover statistical ensembles. A force field is used to describe the potential energy of a system, representing both bonded and non-bonded (van der Waals) and electrostatic interactions. Several force fields commonly used in MD simulations are AMBER [15, 16], CHARMM [17], and GROMOS [18]. Today, classical MD simulations of protein and ligand binding events can last up to a few microseconds and are sufficiently accurate for drug design. Despite this success, classical MD simulations often suffer from insufficient sampling of conformational states due to the high computational demands of simulations greater than a microsecond in length.

The latter of these problems, insufficient sampling, is due to the size of the integration time-step of MD, usually limited to femtoseconds or shorter (dictated by bond vibrations), whereas binding process are in the micro/millisecond time scale. This leads to large computational costs to sample important configurations that have may a major contribution towards binding affinity. Beyond the computational cost, insufficient sampling is a well-recognized problem in molecular simulations due to high energy barriers separating conformations, resulting in biological systems to become trapped in deep energy wells of their potential energy surfaces [19]. Enhanced sampling techniques can accelerate thermodynamics calculations by making rare transition conformations more likely. This can be achieved by, for example, modifying the potential energy surfaces to include a bias potential to the Hamiltonian on the system or increasing the temperature [20]. Commonly used methods include adaptive force bias [21], umbrella sampling [22], steered MD [23], metadynamics [24], and replica exchange molecular dynamics (REMD) [25]. Additionally, endpoint methods, such as molecular mechanics Poisson-Boltzmann Surface Area (MM/PBSA) and molecular mechanics generalized Born Surface Area (MM/GBSA), provide a cheaper way to estimate binding free energy without simulating the binding process [26, 27]. However, these techniques are still computationally intensive, requiring a significant number of ensembles to converge completely.

Rather than following a physical pathway between two states, one can construct a non-physical pathway that involves transforming one molecule to another, avoiding the need of exhaustive sampling. This type of unphysical pathway is referred to as "alchemical transformation" and has emerged as a useful tool for predicting free energy differences, such as the binding of a small molecule to a receptor. The term "alchemical" was coined by the fact that, in silico, a molecule can be transformed into a different, even non-interacting, molecule, where its respective potentials are switched off through a series of intermediate steps. Most alchemical methods utilize free energy perturbation (FEP) [28] or thermody-

namic integration [29] to mathematically calculate the change in free energy. Early efforts demonstrated that MD and Monte Carlo simulations could carry out these calculations for biomolecular systems [30, 31].

Currently, the most popular alchemical method calculates the relative binding free energy (RBFE). RBFE calculations yield the change in binding free energy between two compounds by artificially transforming one compound to the other in the binding site. RBFE requires prior knowledge of the reference compound's binding affinity and that the two compounds are structurally and chemically similar. RBFE play a significant role in the lead optimization in the later states of drug discovery, such as making small changes to a lead candidate compound to create a molecule with higher affinity towards targeted host system [32]. However, RBFE requires structural similarity between molecules, which in turn limits its capability of exploring chemical space.

Conversely, absolute binding free energy (ABFE) calculations are capable of computing the standard free energy of binding between a given compound and the receptor of interest. The standard free energy of binding is computed from the reversible work of decoupling the ligand from the binding site and recouping it with the bulk solvent [33]. Importantly, ABFE can be applied to a diverse chemical space without requiring prior knowledge of any binding affinities or structural similarities between compounds [14, 34]. Therefore, ABFE is more suitable for the virtual screening of diverse ligands but suffers from computational expense and the design (setup/running calculations) can be complicated [35].

Rigorous ABFE calculations require a thermodynamic pathway of overlapping intermediate states to connect the bound and unbound end-states to produce reliable binding affinity calculations. A widely used approach, the double decoupling method (DDM) [34] uses a series of alchemical steps to estimate the absolute binding free energy. The DDM cycle consists of two physically meaningful end states (i.e bound and unbound states) and a series of intermediate states where the ligand is decoupled from the environment. The ligand intermolecular and intramolecular interactions are removed through a series of intermediate windows by scaling its electrostatics and van der Waals (VDW) interactions to zero. In this way, the two molecules (receptor and ligand) are brought together without having to account for complicated interactions during the binding process and then the interactions with the ligand are gradually turned back on, allowing it to complete the binding process. Nevertheless, the DDM method is known to suffer from poor convergence without intramolecular interactions of the ligand. The ligands may collide or overlap in physical space, creating enormous potential energies and numerical instabilities. To prevent this from occurring, artificial restraints are used to keep the ligand from leaving the binding pocket and stop it from interacting with the environment.[36] The use of these restraints is necessary to ensure the bound configuration is sampled during the simulations, aiding in good phase space overlap and faster convergence.[37]

Of course, accurate sampling only helps if the molecular model, including the solvent environment, is accurate. Explicit solvents are full atomic representations of water molecules that surround the solute, providing a physically detailed and generally accurate model; DDM was conceived with explicit solvents in mind. However, the inclusion

of these explicit water molecules causes the simulations to be computationally expensive because the degrees of freedom of each water molecule must be calculated for every time-step [38]. Additionally, explicit solvent can cause "singularity events" when solute-solvent steric overlap occurs. The use of soft-core potentials [39] is used in order to minimize or resolve the instabilities observed during the decoupling of the VDW parameters by shifting the inter-particle distance between the solute and solvent atoms. To avoid the extra complexity and computational costs of using explicit solvents, implicit solvents may be used. Implicit solvents are well suited for the DDM approach as the solvent degrees of freedom (DOF) are integrated out and solute-solvent steric overlap does not occur. Rather than physically representing water molecules, implicit solvent models provide a mean-field representation. Two popular end-state techniques that utilized implicit solvent models that have been widely exploited in free energy calculations are molecular mechanics/Poisson-Boltzmann surface area (MM/PBSA) and molecular mechanics/generalized Born surface area (MM/GBSA) [40, 41]. However, both of these methods have severe thermodynamic approximations [42], for example, ignoring structural changes in receptor and ligand upon binding.

Our work attempts to reduce the complexity and computational cost of ABFE calculations by using implicit solvent models while achieving reliability and accuracy. To do this, we use a modified DDM cycle that uses conformational restraints and the generalized Born (GB) implicit solvent model to enhance convergence. Our DDM cycle provides improved sampling at the end states while reducing the need for sampling in intermediate states. We aim to prove that ABFE calculations can be performed for diverse ligands and can be done in a fast, efficient, and effective manner. To this end, we created the Implicit Solvent DDM tool (`ISDDM.py`), a command line Python interface, to facilitate and automate the use of ABFE calculations for virtual compound screening. The `ISDDM.py` package begins with a parametrized model of the complex and performs all calculations of the alchemical cycle, post-processing, and data analysis. Our implementation utilizes the open-source AmberTools molecular modeling suite to maximize computational throughput. We have tested `ISDDM.py` for accuracy by calculating the binding free energy of a series of host-guest systems based on five hosts: cucurbit[7]uril, pillar[6]arene, octa-acid and α/β-cyclodextrin. ABFEs are calculate for 83 unique host-guest complexes drawn from the literature [43, 44, 45, 46, 47].

Chapter 2 provides underlying theory behind this work, going over binding free energy calculations, use of implicit solvent models, free energy perturbation and introduces our modified double decoupling thermodynamic cycle. In chapter 3 we discuss the chemistry of host-guest systems and our simulation setup parameters. Chapters 4 and 5 present our results and discuss our findings.

Chapter 2

Theory

2.1 Binding Free Energy

Free energy drives a system towards equilibrium and determines the probability that a system will adopt a given state. The Helmholtz free energy, A, of a system in the canonical ensemble (i.e constant number of particles, volume and temperature) is given by

$$F = -\frac{1}{\beta} \ln Q, \tag{2.1}$$

where $\beta = kT$, k Boltzmann's constant and T is the temperature. Q is the partition function of the system,

$$Q = \frac{1}{h^{3N} N!} \int \int e^{-\beta H(p,r)} dp dr,$$

where N is the number of particles in the system and h is Plank's constant. The integral is performed over all $3N$ positions, r, and momenta, p. The Hamiltonian $H(p, r)$ consists of a potential energy term, $U(x)$, that depends on the given configuration, and kinetic energy term, $K(p)$, that depends on particle momenta. In general, evaluating Q is very difficult and not feasible in more complicated cases, such as larger systems with strong interactions between particles. However, since we are interested in the free energy differences, ΔA, between the unbound state (state 0) and bound state (state1), we require only the ratio of partition functions Q_0 and Q_1 :

$$\Delta F_{10} = F_1 - F_0 = -\beta^{-1} \ln \frac{Q_1}{Q_0}. \tag{2.2}$$

As neither the temperature or particles' masses change, the momenta of the two partition functions integrate out and cancel, so only the configurational part of the partition function needed to be considered, and the free free energy of binding can be expressed as

$$\Delta G^{\circ}_{10} = -k_B T ln \frac{\int_{V_1} \int_{\Gamma_1} e^{-\beta[U_1(x)+pV_1]} dV dx}{\int_{V_0} \int_{\Gamma_0} e^{-\beta[U_0(x)+pV_o]} dV dx},$$

where V_0 and V_1 are the box volumes, p is the pressure, and Γ_0 and Γ_1 are the phase space volumes of the bound and unbound states. Due to the compressibility of water, at 1 atm the effect of changes in average of volume on binding is negligible [48], meaning the pV-term can be ignored and the Helmholtz free energy closely approximates the Gibbs free energy

$$\Delta G^{\circ}_{10} \approx \Delta F^{\circ}_{10} = -k_B T ln \frac{\int_{\Gamma_1} e^{-\beta U_1(x)} dx}{\int_{\Gamma_0} e^{-\beta U_0(x)} dx}. \tag{2.3}$$

For complex systems, the partition function has no analytical solution and simulations

need to be used to sample the accessible phase space. Simulating binding events are usually dominated by rate dissociations, which can be on the millisecond timescale or longer for some drugs [35]. In section 2.1.1, we will discuss how alchemical free energy calculations can yield predictions that do not require direct simulations of binding/unbinding events.

2.1.1 Free Energy Perturbation

While simulating the binding process of ligand and receptor provides estimates of binding affinity [49], current common molecular dynamic software packages and high-end GPUs reach a only a few hundred of ns/day [50] for typical proteins. As a results, brute force calculations are unappealing for high throughput drug discovery timescales. Free energy perturbation calculations provide a way to estimate the binding affinity without simulating the binding event. The earliest and most well known method to estimate free energy difference based on the perturbation theory was introduced by Zwanzig[51]. If we then add and subtract $U_0(q)$ from $U_1(x)$ in equation 2.3:

$$\begin{aligned}\Delta F_{10} &= -\frac{1}{\beta}\ln\left[\frac{\int e^{-\beta(U_1(q)-U_0(q)+U_0(q))dq}}{Q_0}\right] \\ &= -\frac{1}{\beta}\ln\left[\frac{\int e^{-\beta(U_1(q)-U_0(q))}e^{-\beta(U_0(q))}dq}{Q_0}\right]\end{aligned}$$

we obtain the definition of exponential averaging (EXP)

$$\Delta F_{10} = -\frac{1}{\beta}\ln\left\langle e^{-\beta(U_1(q)-U_0(q))}\right\rangle_0 = -\frac{1}{\beta}\left\langle e^{-\beta\Delta U_{10}(q)}\right\rangle_0,$$

here $\Delta U_{10} = U_1 - U_0$ and the $\langle\rangle_0$ indicates the equilibrium sampling was calculated in state 0. The above equations applies to a forward transformation from state 0 to state 1. Generally sampling in both directions is used to check for convergence, but in practice their convergence properties may not be the same [52]. In addition, since EXP is only sampled at one state and then perturbed to the other, convergence tends to be slow and is also sensitive to lack of phase space overlap between states [53]

2.1.2 Multistate Bennett Acceptance Ratio (MBAR)

A recurring challenge in ligand-receptor binding is to sample sufficient data to estimate binding affinity to adequate precision. Earlier methods, such as EXP, are one sided and do not make the most efficient use of data when samples from more than one state are available [54]. Multistate Bennett acceptance method (MBAR) is a superior estimator in that it has the lowest possible variance and is asymptotically unbiased [54]. MBAR makes use of all configurations sampled from each state, though sufficient phase space overlap is still required for accurate estimates of binding affinity. In the MBAR formulation, each state i is characterized by a specific potential energy function, $U_i(x)$, inverse temperature, β_i, pressure, p_i and chemical potential, μ_i, depending on the ensemble. The reduced potential energy, $u_i(x)$, of state i is defined as

$$\mathrm{u}_i(x) = \beta_i\left[U_i(x) + p_iV(x) + \mu_i^T N(x)\right],$$

where $N(x)$ is the number of molecules and $U_i(x)$ potential energy function at state i. In order to produce an estimate of for the difference in dimensionless free energies we write

$$\Delta f_{ij} = f_j - f_i = -\ln\frac{c_j}{c_i} = -\ln\frac{\int_\Gamma dx q_j(x)}{\int_\Gamma dx q_i(x)},$$

where f_i are the dimensionless free energies. Once conformations are sampled with Boltzmann statistics, the estimating equations for dimensionless free energies are

$$\hat{f}_i = -ln \sum_{j=1}^{K}\sum_{n=1}^{N_j} \frac{exp[-u_i(x_{jn})]}{\sum_{k=1}^{K} N_k exp[\hat{f}_k - u_k(x_{jn})]},$$

where $\hat{f}_i$ are the MBAR estimates of the free energy of the true free energy, f_i. Note that $\hat{f}_i$ are additive, so only the difference $\Delta\hat{f}_{ij}$ are meaningful. However, when samples from different states have poor phase space overlap, the equation above will fail to converge or $\hat{f}_i$ could have large uncertainties. Lastly, MBAR also provides a direct expression to estimate the statistical uncertainty in free energy differences:

$$\delta^2\Delta\hat{f}_{ij} \equiv \text{cov}\left(-\ln\frac{\hat{c}_j}{\hat{c}_i} - \ln\frac{\hat{c}_j}{\hat{c}_i}\right) = \hat{\Theta}_{ii} - 2\hat{\Theta}_{ij} + \hat{\Theta}_{jj}.$$

Where$\hat{\Theta}_{ij}$ is the covariance matrix, $\hat{\Theta}_{ij} = cov(\Theta_i, \Theta_j)$.

2.1.3 Intermediate states-provide better phase space overlap

In order to compute accurate free energy differences between two states, their phase space integrals must have sufficient overlap, meaning that the two states are similar. Typically, this is almost never the case for bound and unbound states, as the end-states have poor space phase overlap, leading to large uncertainties in the computed the free energy difference. Therefore, appropriately selected intermediate states can make free energy calculations more efficient by bridging the gap between the target end-states [55]. Each intermediate state is defined by a different Hamiltonian, and since free energy is a state function, the path that connects the states is unimportant. Therefore, we can select a cycle that connects any two states through several nonphysical intermediate ones, such as those shown in figure 2.1.

However, there is still a large change between each state in figure 2.1, and these must also be connected by intermediate states. Though other paths may be used, the simplest path to construct between two Hamiltonians is linear:

$$H(\lambda) = \lambda H_1 + (1-\lambda)H_0 \tag{2.4}$$

where λ is a coupling parameter denoting an intermediate state and is given a value between 0 and 1. When $\lambda = 0$, the system is in the H_0 state, and when $\lambda = 1$ it is in the H_1 state. For values between 0 and 1, the system is in a mixed state. For the cycle depicted in figure 2.1,

the additional mixed-states between the major ones shown are simulated with equation 2.4. Each of these intermediate states are simulated independently. From these simulations the free energy difference between state i and its successor $i+1$ can be calculated and binding free energy can be recovered by the sum of all these $\Delta G_{i,i+1}$,

$$\Delta G_{\mathrm{b}} = \sum \Delta G_{i,i+1}.$$

2.1.4 Our Thermodynamic Pathway

Our thermodynamic cycle is based on a modified DDM cycle and employs conformational restraints and GB solvent to enhance convergence. All of the sampling is done in the fast GB implicit solvent model in states 1-8 figure 2.1.ent to enhance convergence. All of the sampling is done in the fast GB implicit solvent model in states 1-8 figure 2.1.

1. States 1 & 8 : The unbound and bound host-guest systems are simulated in GB solvent. In this study, we use replica exchange molecular dynamics (REMD) to efficiently explore the different potential energy surfaces.

2. $1 \rightarrow 2$: Artificial intramolecular conformational restraints are turned on. These restraints are placed on every atom of the ligand and receptor. These restraints restrict the motion of the receptor and ligand to the last frame of state 8. Limiting the range of motion prevents steric clashes and smooths the convergence in the later steps.

3. $2 \rightarrow 3$: The host and guest systems are decoupled from the solvent environment by turning off GB.

4. $3 \rightarrow 4$: The ligand is annihilated by first turning off Lennard-Jones interactions and then scaling the ligand partial charges to zero through several nonphysical λ states.

5. $4 \rightarrow 5$: Orientational restraints are introduced, which prevent the ligand from leaving the binding pocket when it is not interacting with the environment nor the receptor. We select the orientational restraints to bind the ligand relative to the receptor, using the last frame of step 8 as a reference. In this project, we used restraints proposed by Boresch et al. [36], which keep the ligand in a specific orientation relative to the receptor. The method selects a total of six heavy atoms: three from the receptor and three from the ligand. Then harmonic restraints are applied, which are comprised by one distance, two angles and three dihedrals shown in figure 2.2. Once the restraints are applied, an analytical formula can be then used to compute the $\Delta G^{\mathrm{solv}}_{\mathrm{restr}}$ shown below

$$\Delta G_{\mathrm{restraints}} = kTln\left[\frac{8\pi^2 V(K_r K_{\theta 1} K_{\theta 2} K_{\phi 1} K_{\phi 2} K_{\phi 3})^{1/2}}{r_1^2 \sin\theta_1 \sin\theta_2 (2\pi kT)^3}\right],$$

   where K denotes the applied restraints force constant. The r subscript denoted the distant. θ and ϕ denote the angle and torsion restraints.

6. $5 \rightarrow 6$: The change in free energy across this step is 0, since the intra/intermolecular interactions have been turn off. The Boresch restraints have already been applied

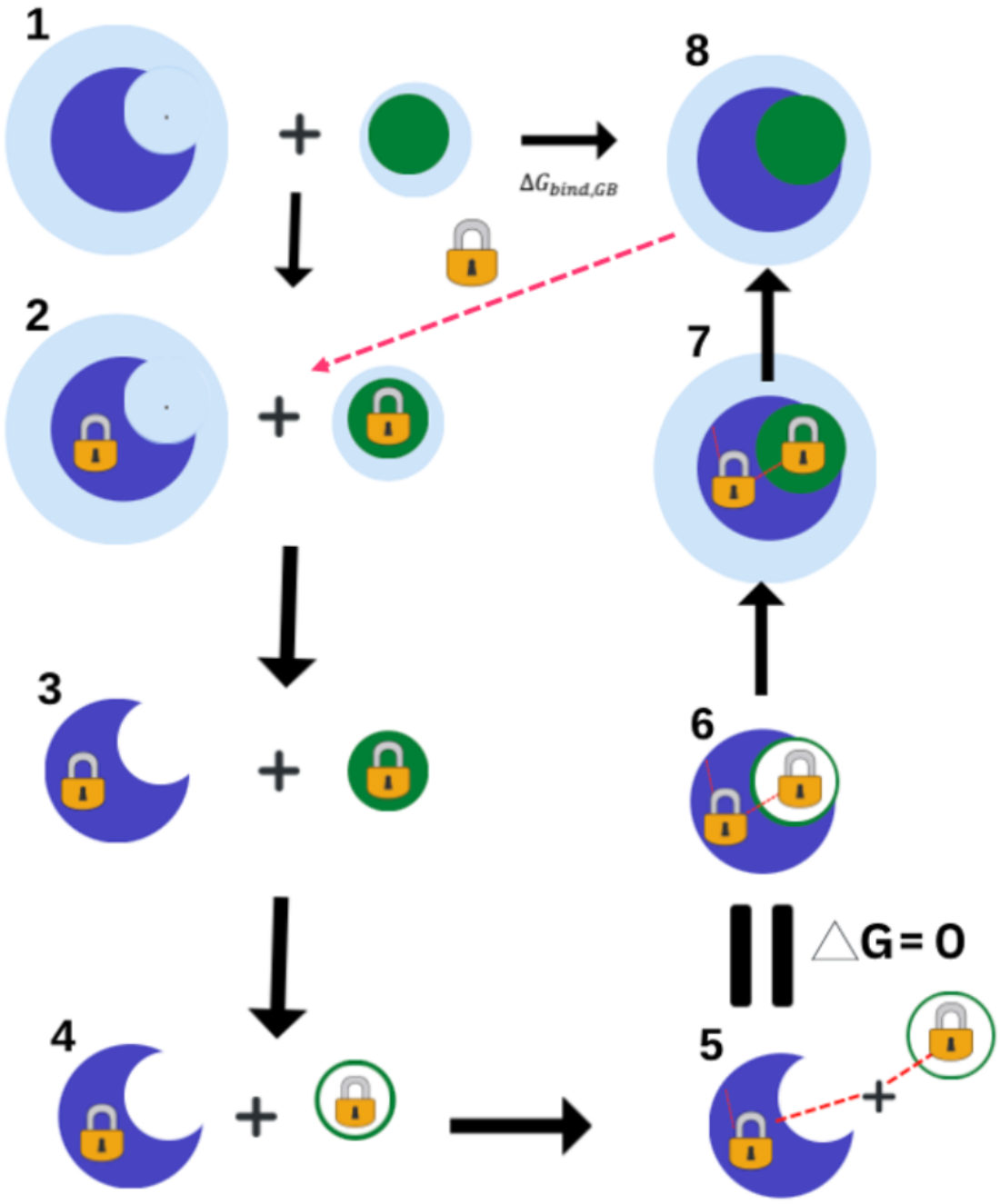

Figure 2.1: Major states in our binding free energy thermodynamic cycle. The free energy change of binding in GB solvent is the free energy difference between states 1 and 8. Conformational restraints (padlock) indicates the presence of harmonic distance restraints between each atom separated by < 6 Å. The filled green circle indicates that ligand charges are active, while the unfilled circle indicates charges have been set to 0. The light blue background is representation of GBSA solvent. Orientational restraints (red dotted line) are applied in states 5 and 6. State 6 is where the ligand is decoupled from the receptor but is orientationally restrained relative to receptor. State 5 is equivalent to state 6 and not simulated.

and the receptor and ligand do not interact. In step 5, the ligand and receptor are in separate simulations and not interacting. In step 6, the ligand and receptor are in the same simulate and not interacting.

7. $6 \rightarrow 7$: Solute-solvent and ligand inter/intramolecular interactions are restored. This is accomplished in a series of steps. First the Lennard-Jones interaction between the host and ligand are turned back on, followed by restoration of solute-solvent interactions. Finally ligand charges are re-introduced in a series of λ-windows.

8. $7 \rightarrow 8$: Conformational and orientational restraints are incrementally removed.

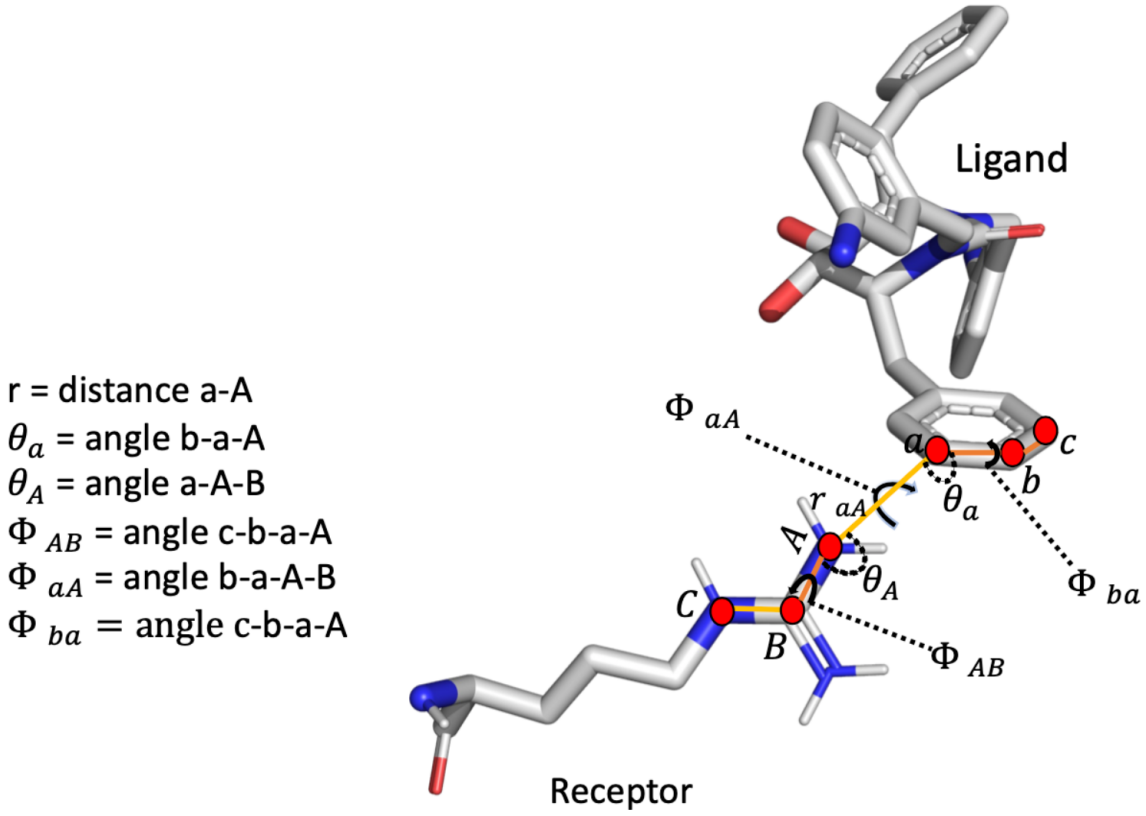

Figure 2.2: Implementation of Boresch restraints for absolute binding free energy calculations for an example receptor-ligand pair. Atoms "a," "b," and "c" belong to the ligand (on the top right), while atoms "A," "B," and "C" belong to the protein (on bottom left).

2.2 Molecular Modeling of physical chemistry of binding

To estimate the free energy differences accurately, a molecular model is required that describes the relevant physical properties of a system correctly. Specifically, the Hamiltonian used to calculate the energy and forces should ensure that all configurations will be sampled with the correct probability. In practice, the choice of the Hamiltonian is often a compromise between accuracy and efficiency. Two major factors that dictate the cost of

the energy and force evaluation are the degrees of freedom and the functional form of the Hamiltonian. A quantum-mechanical description of the system provides the most detailed description, as the electrons' degrees of freedom are modeled explicitly and the interaction energy is calculated by solving electronic structure of the system. In the molecular mechanics (MM) description, molecules are represented by particles, which represent atoms or groups of atoms. The quantum-mechanical description of the system will generally be more computationally demanding than classical molecular models. Here we focus on classical Hamiltonians, commonly referred as to "force fields", that incorporate fixed point charges.

In this thesis, we use the general Amber force field (GAFF) [16, 15] for our MD simulations. The potential energy function that describes the force field of the Amber software package is

$$V_{\text{AMBER}}(r^N) = E_{\text{bonded}} + E_{\text{non-bonded}},$$

where

$$E_{\text{bonded}} = E_{\text{bonds}} + E_{\text{angles}} + E_{\text{torisions}}$$

and

$$E_{\text{non-bonded}} = E_{\text{vdW}} + E_{\text{electrostatics}}.$$

The function is written explicitly as

$$V_{\text{AMBER}}(r^N) = \sum_{i_{\text{bonds}}} k_{bi}(l_i - l_i^0)^2 + \sum_{i_{\text{angles}}} k_{ai}(\theta_i - \theta_i^0)^2 + \sum_{i_{\text{torisions}}} \sum_n \frac{1}{2} V_i^n [1 + \cos(n\omega_i - \gamma_i) +$$

$$\sum_{j=1}^{N-1} \sum_{i=j+1}^{N} f_{ij}^{LJ} \left[\left(\frac{r_{ij}^0}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}^0}{r_{ij}} \right)^{6} \right] + f_{ij}^{\text{elec}} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}.$$

E_{bonded} consists of terms describing bond-stretching, bond-angle bending, and dihedral-angle bending. For bond-stretching, k_{bi} represents the force constant between the ith atom pair, l_i^0 represents the equilibrium bond length, and l_i represents the current bond length. For bond-angle bending, k_{ai} represents the force constant, θ_i^0 is the bond angle at equilibrium and θ_i is the current bond angle between atoms. For dihedral-angle bending, V_i^n represents the dihedral energy barrier, γ is the phase angle for dihedral angle ω. $E_{\text{electrostatics}}$ consists of electrostatics and van der Waals energies. The van der Waals energy is calculated using the 6-12 Lennard-Jones model[56] with equilibrium distance r_{ij}^0 and and energy ϵ_{ij}. Lastly, the electrostatic energy is defined by Coulomb's law where each q is the charge on atom i and j and r_{ij} is the distance from each other. f_{ij}^{LJ} and f_{ij}^{elec} are scaling factors for the Lennard-Jones and electrostatic energy based atoms on the bond separations: $f_{ij} = 0$ for atoms separated by 1-2 bonds, $f_{ij}^{\text{LJ}} = 1/1.2$ and $f_{ij}^{\text{elec}} = 1/2.0$ for atoms separated by 3 bonds and $f_{ij} = 1$ for all others. The factor of 2 ensures the equilibrium distance is r_{ij}^0.

2.2.1 Implicit solvent models

Water plays a fundamental role in molecular association and mediates protein-ligand interactions. Commonly, explicit water models are used, in which a large number of indi-

vidual water molecules are placed around the solute and modeled with approximately the same theoretical level as the solute. While explicit models are generally a realistic representation of an water environment, they are computationally demanding. Moreover, water exchange between bulk and buried cavities can be slow and leads to inefficient sampling of water occupancy states [57, 58]. Alternatively, implicit solvation provides a mean-field model of the solvent, which in turn provides a significant computational advantage by integrating out explicit solvent DOF and reduction of interacting atoms. However, the computational advantage comes at a cost. Some models lack hydrogen bonding between the solute and solvent, or otherwise cause unphysical sampling [59]. Due to its speed, implicit solvent is an attractive approach for sampling the vast chemical space in search of potential drug candidates.

Dielectric continuum implicit solvent models can estimate the free energy of solvation as a function of the conformation of the solute. The solvation free energy is the reversible work of transferring a solute from the gas phase to solution. The ΔG_{solv} is partitioned into two parts: a polar contribution ΔG_{el}, which depends on the electrostatic contributions, a non-polar contribution ΔG_{npol}, which depends on the shape of the solute cavity

$$\Delta G_{\text{solv}} = \Delta G_{\text{el}} + \Delta G_{\text{npol}}.$$

The generalized Born (GB) model treats solvent as a uniform dielectric and is a fast and analytical method to compute the ΔG_{el}. In Amber the "canonical" generalized Born model approximates the Green's function of the Poisson equation as

$$\Delta G_{\text{el}} = -\frac{1}{2}\left(\frac{1}{\epsilon_{in}} - \frac{\exp(-\kappa f_{\text{GB}})}{\epsilon_{out}}\right)\sum_{ij}\frac{q_i q_j}{f_{\text{GB}}(r_{ij}, R_i, R_j)},$$

where r_{ij} is the distance between atoms$j$ and i, q_i and q_j are the partial charges, the R_i and R_j are the so-called Born radii and κ is the Debye-Hückel screening parameter. $f_{GB}()$ is a smooth function that takes several forms. A widely used functional form of $f_{GB}()$ [60] is

$$F_{\text{GB}}(r_{ij}, R_i, R_j) = [r_{ij}^2 + R_i R_j \exp(-r_{ij}^2/4R_i R_j)]^{1/2}$$

where R_i and R_j denote the effective Born radius atoms, i and j, which represent each atom's degree of burial within the solute. Larger effective radii reflect a deeper burial, where the surrounding atoms reduce the extent the atom charge is screened by the solvent.

In dielectric continuum models, the external dielectric medium and the solute will have different dielectric constants. For example, the dielectric of the solvent is typically set to 78.5 to mimic water, while the solute interior will have an dielectric between 1 to 4. Previous work has shown the internal dielectric can be adjusted in order to improved the accuracy of MM/GBSA estimates of protein-ligand affinity [61].

Chapter 3

Methods

3.1 Host-guest systems

Host-guest (HG) systems have emerge as valuable benchmarks for the quantitative assessment of modeling errors for the interactions of ligand-host binding. They tend to have low molecular weight and capture most of the interactions involved in protein-ligand binding, such as hydrogen bonding, van der Waals interactions and conformational changes upon binding. Due to their small molecular size, sampling can be done more efficiently than for protein-ligand systems, and experimental data are available for a wide variety of HG systems. Therefore, HG systems are good candidates to help develop computational modeling scheme sof receptor-ligand binding affinity.

To test our thermodynamic cycle and software implementation of it, we selected host-guest systems from the Statistical Assessment of the Modeling of Proteins and Ligands (SAMPL): SAMPL4 [43], SAMPL7 [62] and SAMPL9 [46] (figure 3.1). α- and β-cyclodextrin hosts and additional guests for octa-acid were obtained from the Taproom repository on Github [47]. All the SAMPL experimental data and initial `mol2` files were downloaded from the GitHub repositories.

3.1.1 Choice of Host-Guest System

In this work, we tested against several macrocyclic host families, cucurbit[7]uril, (CB7), pillar[n]arene, octa-acid (OAH) and α/β-cyclodextrins. These have a diverse binding affinity to variety of small drug-like compounds [62, 43]. From the ligand data available, we tested a total of 83 unique host-ligand complexes.

Cucurbit[n]urils (CB[n]s), including CB[5], CB[6], CB[7], CB[8] and CB[10], are a family of hosts with a molecular structure containing n glycoluril units connected via $2n$ methylene bridges, forming a barrel shape macrocycle with a central hydrophobic cavity. In addition, they have carbonyls protruding out from the hydrophobic cavity. Figure 3.1 shows chemical structures of the CB7 guests, C1–C14, bearing a net charge ranging from e to $2e$.

Pillar[n]arenes are another class of supramolecular compounds having a symmetrical rigid cylindrical structures, with CH_2 linkers connecting the phenylene groups, each of which phenylene group contains two anionic functional group arms. Pillar[n]arene are an emerging class of synthetic macrocycles, having a broad applications from molecular recognition and adsorption to synergistic catalysis and many other aspects [63]. In this study, we tested against Pillar[6]arene (WP6), which bears six carboxylated moieties on the upper and lower rims that can be, in part, protonated at pH = 7.4 [64]. We calculated the binding affinities of 13 ammonium/diammonium cationic ligands (guests) against WP6, depicted in figure 3.1.

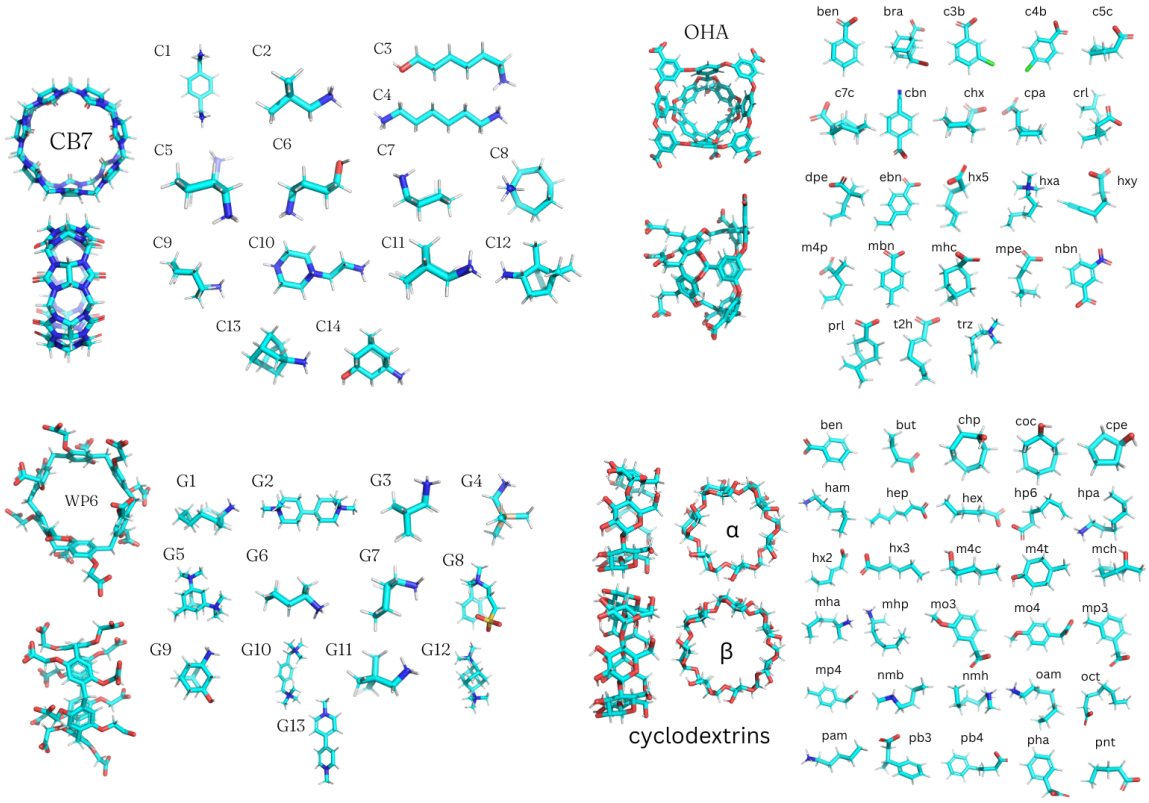

Figure 3.1: Host-guest systems dataset. (Top left.) CB7 receptor and 14 guest ligands from the SAMPL4 challenge dataset. (Bottom left.) WP6 receptor and 13 guest molecules from the SAMPL9 dataset. (Top right.) Octa-acid receptor and 23 guest molecules. (Bottom right.) α/β-cyclodextrins and 30 guest molecules. Octa-acid and α/β-cyclodextrins host-guest complexes were from GitHub repository as benchmark systems. Together, all host systems comprise 83 unique host-guest pairs.

Octa-acid (OAH) is a water soluble a deep-cavity cavitand, which possesses eight water-solubilizing outer surface carboxylate groups, and a deep 10 Å hydrophobic pocket suitable for variety of guest molecules [65]. OAH is known to bind with hydrophobic or amphiphilic guests in aqueous solution, perhaps due to the hydrophobic effect of the desolvation of guests entering the deep hydrophobic binding pocket [43]. Figure 3.1 OAH shows the chemical structures of 23 guests containing a carboxylic acid functional groups that were binding to OAH.

The last benchmark systems were α- and β-cyclodextrins, which are cyclic oligosaccharides, consisting of six and seven glucose monomers respectively. The guest molecules we used with α and β cyclodextrins contain ammonium or carboxylate groups and carry a net charge of e or $-e$ respectively. The two hosts and 41 host-guest complexes shown in figure 3.1.

3.1.2 System preparation

All the ligands were docked into the binding site of an equilibrated host system using Autodock Vina docking software [66] to generate several ligand poses. In this procedure, the host is considered rigid and ligands were centered at the host center of mass. Prior to being used in the docking calculations, we converted `mol2` files of the hosts and ligands to the `pdbqt` format, using OpenBabel [67] with the `-p` option to specify the pH used in the experiments. All ligands were fully flexible during the docking calculations. Autodock Vina carried out a cluster analysis based on all-atom root mean square deviation ranking the dock conformations in order of increasing binding free energies. The lowest energy ligand pose was selected for each host-guest complex for the initial end-state configuration.

The simulations of all host-guest systems were setup by LEaP program in AmberTools software suite [68]. The partial atomic charges were generated using the AM1-BCC model [69] and generalized Amber Force Field v2 (GAFFv2) force field[16, 15] was applied to all guest and host systems using the `antechamber` program in AmberTools21 [68].

3.2 Simulation details

All production simulations were run by our Python script, `ISDDM.py`, (see section §4.1 for implementation details) using `sander` from Amber 21 [68] for the molecular dynamics calculations. For each system, the bound end-state was simulated first to provide a reference structure for restraints. The unbound end-state and all intermediate states were then simulated independently. All systems used hydrogen mass repartitioning, in which mass of hydrogen atoms are increased by a factory of 4 and the mass of the bonded heavy atoms is decreased by an equivalent amount [70]. This has be shown to slow the highest frequency of hydrogen covalent bonds oscillation, which permitted us to use a time-step size of 4 fs.

3.2.1 End-state Simulations

Simulations for states 1 and 8 used replica exchange molecular dynamics simulation (REMD)[71]. The ionic concentration was set to 0.3 M for all host guest systems. We

prepared the initial structures of states 1 and 8 for REMD simulation with 5000 steps of steepest descent energy minimization using the `sander` or `sander.MPI` program . From these starting structures, we ran REMD using `sander.MPI` with eight temperature copies distributed from 298K to 500K for 2,500,000 exchange attempts (`numexchg=2500000`) with exchanges attempted every step (`nstlim=1`). For all host-guest bound end-state simulations, a center-of-mass flat bottom restraint potential was used. Coordinate trajectories at the target temperature of 298 K were extracted by `pytraj`[72]. A total of 10,000 frames were saved for each of the end-states.

3.2.2 Intermediate Simulations

All simulations of states $2-7$ used the last frame of of REMD simulations in state 8 for initial structures. `ISDDM.py` constructed both the orientational and conformational restraints from the last frame of complex configuration from REMD simulation in state 8. The last frame of state 8 was split into the unbound ligand and host systems using `strip` command from `cpptraj`. In all states `gbsa=0` and in states $3-7$ GB was not used (`igb=6`). In state 4 the ligand charges were set to 0 by the `parmed` function `parmed.tools.action.charge`. State 5 is not sampled. The difference in free energy going from state 4 to 5 is computed analytically, as described in section 2.1.4, and the difference in free energy from state 5 to 6 is 0. State 6 is sampled with ligand charges set to 0 and Lennard-Jones interactions between the ligand and host system were removed via the `add_exclusions` function in `parmed.tools.action.charge`. State 7 consisted of several intermediate states, in which the Lennard-Jones potential between the ligand and host was reintroduced in a single step, followed by increasing the GB dielectric constant, using values of of 3.925, 7.85, 15.7, 39.25 and 78.5. This was followed by turning on the ligand charges to 50% and then 100% (fully charge). Between states 7 and 8, we slowly removed the restraint force through a series of intermediate states.

3.2.2.1 Selection and Implementation of Restraint Forces

Between states 1 and 2, conformational restraints were increased over a series of intermediate states and then orientational and conformational restraints were removed between states 7 and 8 in the same fashion. For the benchmark testing, a series of 29 lambda-windows were generated with rotational and orientational force spaced evenly on $\log_2$-scale. The lowest energy conformational restraint force constant was determine to be 2^{-8} kcal/mol/Å and the maximum to be 2^4 kcal/mol/Å . From previous results [73], a minimum orientational force constant of 2^{-4} kcal/mol/Å and maximum 2^8 kcal/mol/Å gave acceptable errors for highly charge guest molecules. Host and guest molecule conformations were restrained by applying a harmonic distance restraint between every atom and each neighbor within 6 Å. To maintain the guest molecule in the correct binding mode, `ISDDM.py` constructed orientational restraints [36], which requires an distance restraint r_{aA} , two angles, θ_A and θ_a, and three torsions, ϕ_{AB}, ϕ_{aA} and ϕ_{ba} (figure 2.2). First the atoms on the ligand, a, and receptor, A, are selected for the distance restraints. Ligand atom a is the closest heavy-atom to the ligand center of mass (COM), where A is the receptor heavy-atom that is the shortest distance from a. Then the script selects the best suited heavy atoms, b and B, to create angles θ_A, θ_a, ϕ_{AB}, ϕ_{aA}, and ϕ_{ba} between 80 and 100 de-

grees. It is important that the angle is close to 90 degrees in order to maintain a well define bound state.

3.3 Free Energy Analysis Using MBAR

The Python package PyMBAR version 3.1 [74] was used to calculate free energy differences along the states 1-4 and 6-8, using the `mbar` module. MBAR requires the potential energies of each coordinate trajectory to be calculated with the potential function of all other states of the same molecule. For example, potential energies for the coordinate trajectory from state 2 need to be calculated for all other states between 1 and 4. These potential energies were calculated using `sander` with `imin=5`.

3.3.1 Time Series Analysis

The number of independent samples in an MD trajectory is limited by time correlations between frames and the time required to relax to the equilibrium average. If this is not accounted for, the final analysis will underestimate the error [75]. To limit our analysis to independent samples, we used the `timeseries` module of PyMBAR to remove correlated frames and the relaxation period of the energy trajectories [76]. These independent samples were then run through the `mbar` function to calculate the final free energy change for all the thermodynamic steps shown in figure 2.1.

3.4 Statistical metrics of absolute binding

To assess the ABFE predictions of GB in this automation workflow protocol, we applied several metrics, several of which are unique to the SAMPL challenges [43]. Standard metrics included the best-fit linear regression slope and Pearson correlation coefficient, R^2. We also calculated the root mean squared error (RMSE),

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\Delta G_i^{\text{calc}} - \Delta G_i^{\text{exp}})}{n}},$$

mean absolute error (MAE),

$$\text{MAE} = \frac{\sum_{i=1}^{n} \left|\Delta G_i^{\text{calc}} - \Delta G_i^{\text{exp}}\right|}{n},$$

and mean signed error (MSE),

$$\text{MSE} = \sum_{i=1}^{n} \frac{\left(\Delta G_i^{\text{calc}} - \Delta G_i^{\text{exp}}\right)}{n},$$

where ΔG^{exp} and ΔG^{calc} are the experimental and GB predicated binding affinity values for guest molecule i and n is the number of measurements. The RMSE_o is the root mean

square error of predicted binding affinities after subtracting the average signed error,

$$\text{RMS}E_o = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\Delta G_i^{\text{exp}} - \Delta G_i^{\text{calc}} - \frac{1}{n}\sum_{j-1}^{n}\left(\Delta G_j^{\text{exp}} - \Delta G_j^{\text{calc}}\right)\right]^2}. \qquad (3.1)$$

and RMSE$_r$, is used to measure the accuracy of relative binding free energies by considering all guest molecules,

$$\text{RMS}E_r = \sqrt{\frac{2}{n\left(n-1\right)}\sum_{i}^{n}\sum_{j=i+1}^{n}\left[\left(\Delta G_j^{\text{calc}} - \Delta G_i^{\text{calc}}\right) - \left(\Delta G_j^{\text{exp}} - \Delta G_i^{\text{exp}}\right)\right]^2}. \qquad (3.2)$$

Chapter 4

Results

Here we use `ISDDM.py` to carry out absolute binding free energy calculations on the host-guest systems describe in the previous section. In section §4.1 we describe the automated workflow for using `ISDDM.py`, including how it prepares all the necessary input files, manages simulations, and and processes the data. In section §4.2, we report on the accuracy of the method compared to experiment, and in section §4.3, we assess the convergence of each step of the cycle. Finally, in section §4.4, we look at the overall scaling performance for the time required for the cycle to finish.

4.1 Implementation Details

`ISDDM.py` is a Python program that automates preparing, running and post-processing all the MD simulations for calculating the free energies. The workflow of `ISDDM.py` is shown in figure 4.1. In the input file (see appendix A for details), users provide parameters for the method, including the location of a force field parameter file and initial coordinates for the complex. When the program is executed, it first carries out REMD GB simulations of states 1 and 8. Optionally, the user may provide their own completed simulations of the end-states instead. From the final frame of the complex in state 8, conformational and orientational restraints are generated. The same frame is used to generate the initial coordinates for states 2-7. For states 2-4, the receptor and ligand are spilt from the complex end-state structure. All the intermediate MD states are run simultaneously and executed in parallel using the maximum number of processors specified in the input file. As MD simulations complete, their trajectories are pushed into the post-analysis queue to gather all the necessary state comparison and energies. The post-processed data is then parsed into a Pandas DataFrame [77], which is then submitted to MBAR where the final change in BFE is calculated for each system. All steps within in the workflow are handled using the Toil package [78], which is a pure Python workflow engine that allows us to create multiple simulations and run them in parallel.

4.2 Calculations of binding free energies for host-guest systems

4.2.1 Cucurbit[7]uril (CB7)

We calculated the binding free energies of 14 cationic guest to CB7, which had a neutral net charge. Figure 4.2b shows the correlation scatter plot of CB7 host-guest binding free energies versus experimental measurements. Ligands are colored by net charge, which ranges from $+1e$ to $+2e$. The computed free energies from the automated workflow model had a significant absolute error (RMSE $= 6.9\,\mathrm{kcal/mol}$) but a good correlation ($R^2 = 0.8$) (table 4.1). Other metrics, slope $= 1.4$, MAE $= 6.7\,\mathrm{kcal/mol}$ and MSE $= -6.7\,\mathrm{kcal/mol}$, agree that while the correlation was reasonable, the absolute binding free energies were underestimated for all the cationic ligands, which may suggest favorable polar interaction with ammonium and carbonyl rich parts of CB7.

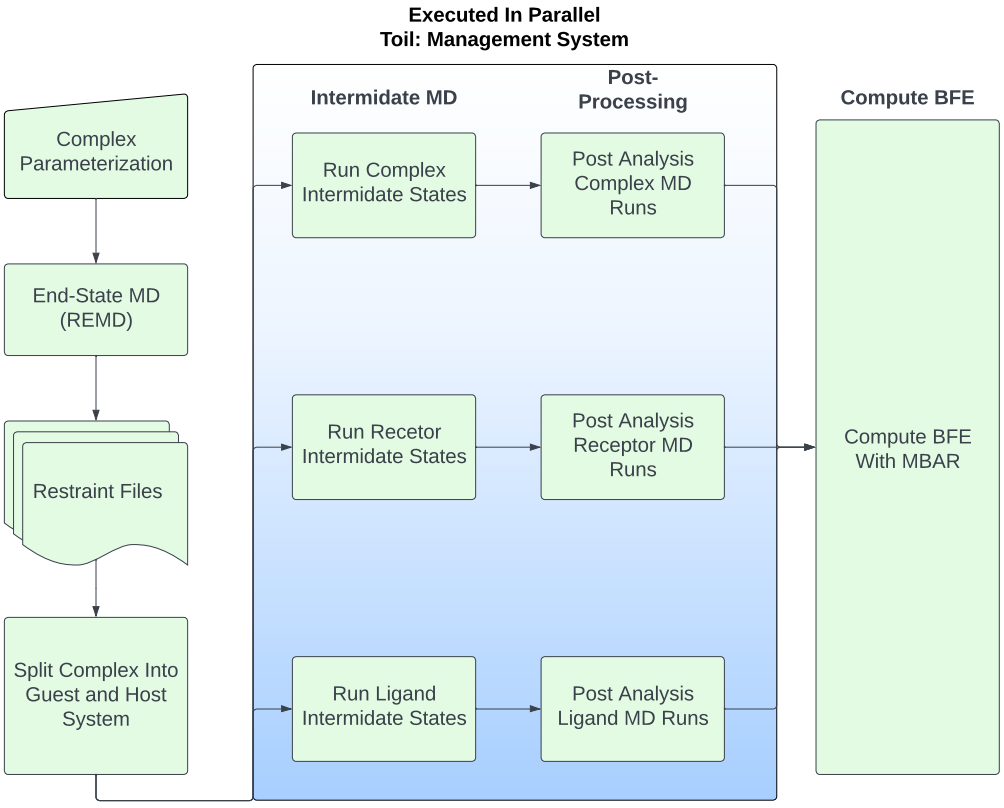

Figure 4.1: Automated workflow pipeline for `ISDDM.py`.

4.2.2 Pillar[6]ene (WP6)

The 13 guest ligands were tested against WP6, which had a net negative charge of $-12e$, as shown in figure 3.1c. Of the 12 guests, 11 had net charges of $+1e$ to $+2e$, due to ammoniums, and one, G8, had a net neutral charge. We see the binding free energy for all positively charged guest were underestimated, similar to what we observed with CB7 (figure 4.2c). However, the binding free energy of G8 was overestimated in contrast to the rest of the ligands. Binding free energies had a significant deviation from experimental values (RMSE=6.64 kcal/mol) and good but not strong decent correlation ($R^2 = 0.52$). On average, the predicated binding affinity was $6.1\,\mathrm{kcal/mol}$ and an MSE of $-6.0\,\mathrm{kcal/mol}$. The magnitude of MAE and MSE indicate poor accuracy of predicated binding free energy values. The large MAE and MSE are probably caused by the lack of charge hydration of the positively charge guest molecules in the GB model, which results in the overestimation of binding.

4.2.3 Octa-acid (OAH)

The host system OAH had a deep 10 Å hydrophobic pocket and eight water-solubilizing carboxylate groups at the outer surface, which gave it a negative net charge of $-8e$ (figure 4.2d). ABFEs of 23 guests were calculated OAH. 18 guests had a net charge of $-1e$ and two (tzh,hxa) had a netcharge of $+1e$. When compared to experiment, calculated binding free energies had a smaller RMSE than for WP6 and CB7 (RMSE $= 2.9\,\mathrm{kcal/mol}$) but poor correlation ($R^2 = 0.07$). The poor correlation is due in large part to the two cationic guest molecules (trz,hxa). If omitted, the correlation improves to $R^2 = 0.54$. Including all data, the mean average error of binding affinity for OAH was $2.2\,\mathrm{kcal/mol}$ and the MSE was $0.6\,\mathrm{kcal/mol}$.

4.2.4 α/β-cyclodextrins

We calculated the binding free energies of 38 guests, with net charges varying from $+1e$ to $-1e$, to α/β–cyclodextrins, which had neutral net charges. Binding free energy metrics for the two molecules were similar (table 4.1). Both β-cyclodextrin (BCD) and α-cyclodextrin (ACD) had RMSE $= 1.6\,\mathrm{kcal/mol}$ compared to experiment. ACD had slope of 0.8 with MAE $= 1.4\,\mathrm{kcal/mol}$ and MSE $= 0.07\,\mathrm{kcal/mol}$; while BCD had a slope of 1.5 and MAE $= 1.4\,\mathrm{kcal/mol}$ and MSE $= -0.7\,\mathrm{kcal/mol}$. Interestingly, we observed a stronger correlation between calculated and experimental for BCD ($R^2 = 0.65$) compared to ACD ($R^2 = 0.17$). This likely be due to the fact BCD has a larger dynamic range of binding free energies than ACD. Figure 4.2e-f shows a clear bias in the error of the ABFE due to ligand charges, which was also seen in the other host guest systems. Here, we see the ABFEs for the negatively charged ligands were overestimated and positively charge ligands were underestimated. These results suggest significant bias due to ligand charge.

4.2.5 Calculations of binding in all host guest systems

Figure 4.2a shows the scatter plot of all the systems we tested. Table 4.1 shows that the statistics for the combined dataset had an RMSE of 4.1 and correlation of 0.76. In all the test cases, we clearly see the ligand net charge strongly affects the calculated ABFEs. This

may be due to the fact that GB does not naturally handle charge hydration asymmetry [79]. However, the host charge has little impact on the ABFE. WP6 and OAH have net charges of -6 and -8 respectively, but seem to follow the same trends as the other molecules.

System	Slope	R^2	RMSE_std	RMSE_o	RMSE_r	MAE	MSE
All-Systems	1.8	0.76	4.1	3.7	5.2	3.1	-1.9
CB7	1.4	0.78	6.9	1.8	2.6	6.7	-6.7
WP6	1.6	0.52	6.6	2.9	4.3	6.1	-6.0
OAH	0.5	0.007	2.9	2.9	4.2	2.2	0.6
ACD	0.8	0.17	1.6	1.6	2.3	1.4	0.07
BCD	1.5	0.65	1.6	1.3	2.0	1.4	-0.7

Table 4.1: Comparison of all host guest systems. Units for RMSE_std, RMSE_o, RMSE_r, MAE and MSE are in kcal/mol.

4.2.6 Generalize Born hydration asymmetry correction

Given that that the GB model used does not correctly handle charge hydration asymmetry [79], and we have observed a bias in predicted ABFEs associated with ligand charge, we applied a linear hydration asymmetry correction,

$$\Delta G_{\text{bind}}^{\text{corr}} = (1 + a)\Delta G_{\text{bind}}^{\text{GB}} + b, \tag{4.1}$$

where a and b are fit parameters.

Charge	a	b
-1	-0.18	-1.89
0	-0.41	-0.97
1	-0.44	-0.23
2	-0.58	-2.30

Table 4.2: Hydration asymmetry linear corrections fitted parameters. Units in kcal/mol.

We applied four separate corrections, where host-guest pairs were grouped by the ligand net charge (table 4.2). We see that the correction improved all metrics for all host system calculations (figure 4.3 and table 4.3). For the full dataset, the correlation coefficient was improved to 0.87 and the RMSE is reduced to 1.06 kcal/mol. Other metrics are similarly improved. The correction depends only on ligand charge and is independent of host charge. One reason why host charge appears irrelevant is that both WP6 and OAH charge carbonyl atoms were solvent exposed when the ligand was bound, so there was no change in hydration for these charge groups.

4.3 Overlap Matrix Exploring the Degree of Space Phase Overlap

When using any perturbation approach, such as FEP or MBAR, a poor phase-space overlap between states can result in an inaccurate free energy estimate [75]. To assess the

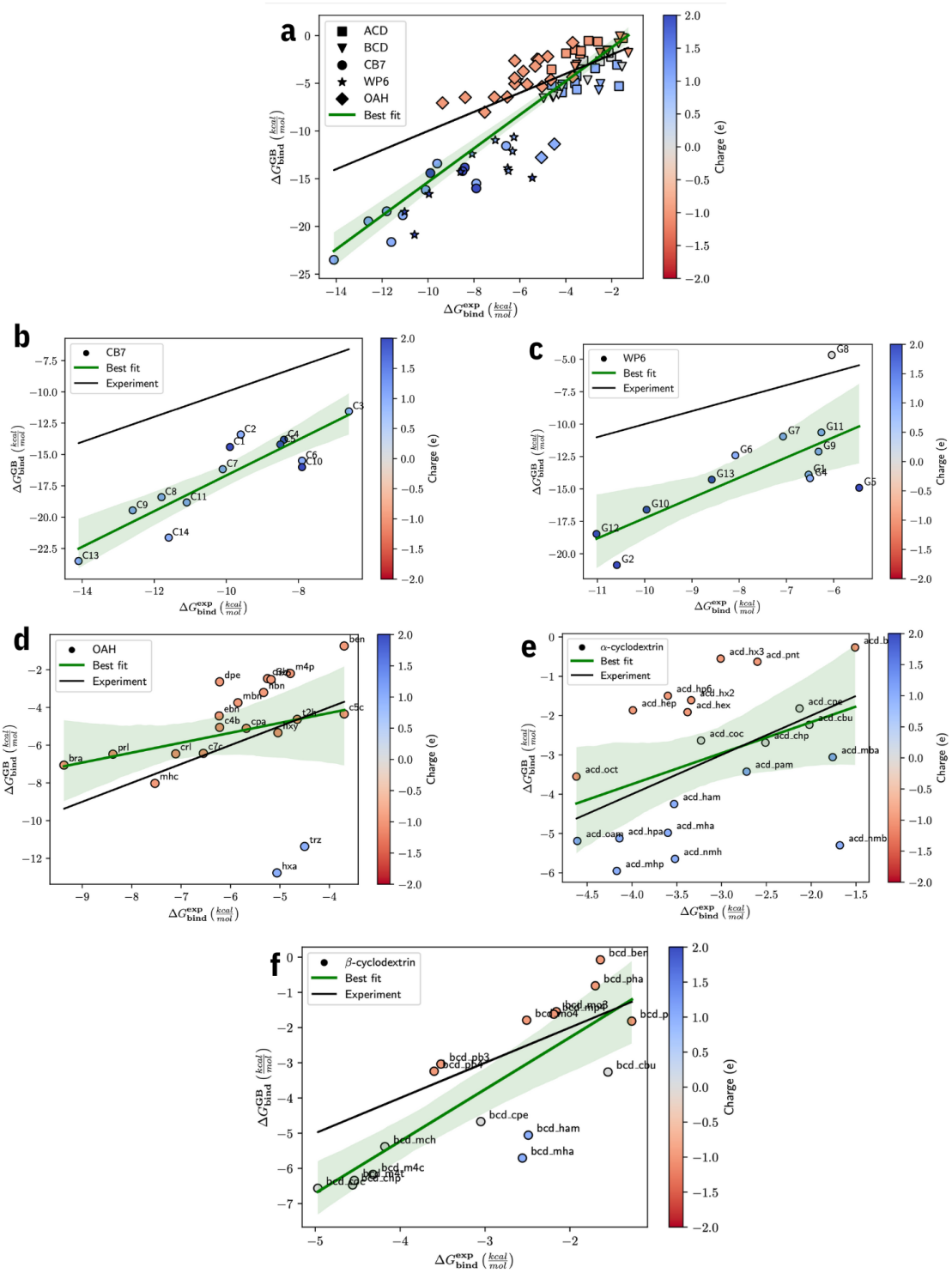

Figure 4.2: Calculated absolute binding free energies ($\Delta G^{\text{bind}}_{\text{GB}}$) with GAFFV2 force field and $GB^{\mathbf{OBC}}$ compared with experiment for (a) all host-guest pairs, (b) CB7, (c) WP6, (d) octa-acid (e) α-cyclodextrin and (f) β-cyclodextrin. The black solid line is the experimental identity. The solid green line indicates the linear regression fitting for host guest systems.

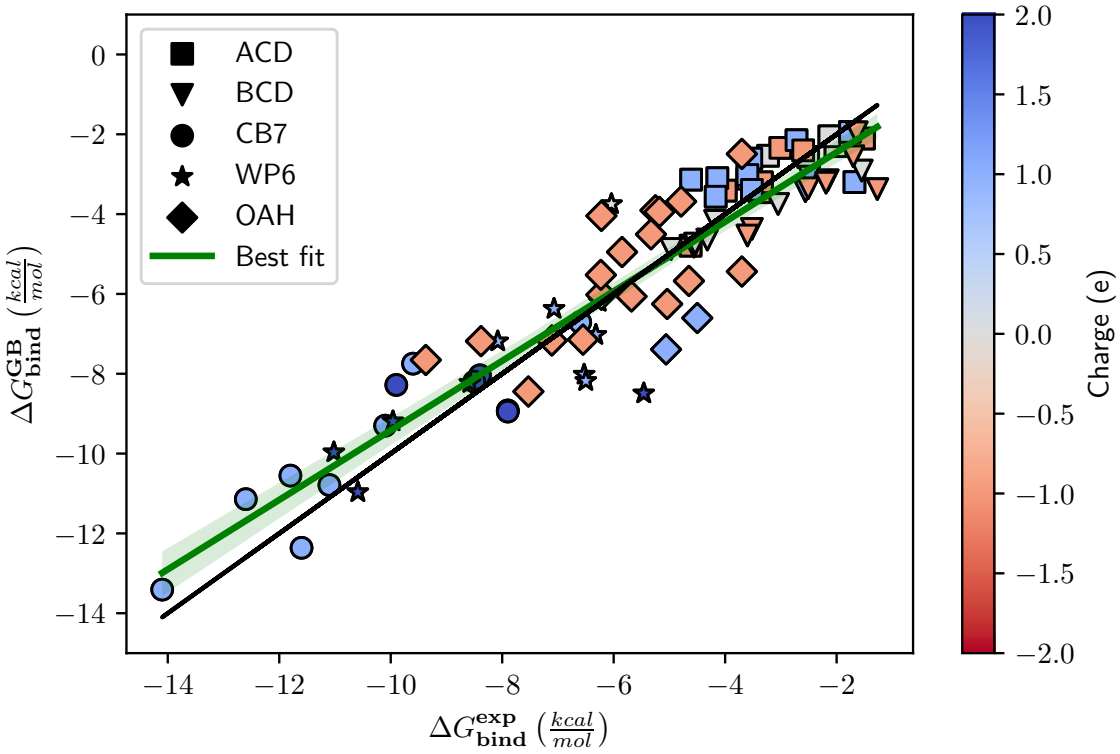

Figure 4.3: Host-guest ABFEs with generalized Born-Hydration Asymmetry Correction. The different shapes in the legend box corresponds to the different host systems.

System	R^2	Slope	RMSE	ME	RMSE$_o$	RMSE$_o$
All	0.87	0.87	1.06	-0.0	1.06	1.51
ACD	0.54	0.53	0.67	-0.25	0.62	0.91
BCD	0.83	0.66	0.88	0.71	0.52	0.76
CB7	0.80	0.81	1.03	-0.44	0.94	1.38
OAH	0.38	0.67	1.30	-0.11	1.30	1.88
WP6	0.51	0.70	1.39	0.10	1.39	2.05

Table 4.3: Comparison of all host guest systems after the application of equation 4.1. Units for RMSE_std, RMSE$_o$, RMSE$_r$, MAE and MSE are in kcal/mol.

convergence of our implementation of our thermodynamic cycle in figure 2.1, we created overlap matrices [75] (figures 4.4 and 4.5). The matrix shows the probability of a sample state i being from state j, indicating the degree of overlap between both states. To obtain reliable free energy estimates, the overlap matrix should at least be tridiagonal. I.e., all $(i, i+1)$ and $(i-1, i)$ elements should be greater than 0.03 [75].

In figure 4.4 highlights an overlap matrix for WP6-G1 complex where every column denotes every simulation in states 6-8 in figure 2.1. (λ_0, λ_0), shows that if sampling is done in λ_0 and analyzed in λ_0 , there is a 0.74 overlap. I.e., 74% of equilibrated configurations are in λ_0. The probability of finding a microstate sampled from state λ_0 in λ_1 is 0.26 and so is finding the microstate sampled from state λ_1 in state λ_0. Overlap for $\lambda_2 - \lambda_9$ in figure 4.4 is also sufficient. In fact, the overlap of states λ_3 and λ_8 is 0.11, which suggests the intermediate windows may be unnecessary. However states λ_9 and λ_{10} show no overlap, and states $\lambda_9 - \lambda_{15}$ show poor overlap (< 0.1), suggesting that additional intermediate states should be introduced. To ensure sufficient overlap for the full cycle, we increased number of restraint windows to 29 resulting in a total of 41 windows shown in figure 4.5, which shows good phase space overlap.

4.4 Performance Scaling

As `ISDDM.py` utilizes the `Toil` workflow management system [78], users are able to employ multiple processors to speedup ABFE calculations. Since all intermediate states (3-7) are independent, all MD simulations, post-analysis calculations and MBAR computations can be parallelized over the available processors identified by the user. Figure 4.6 shows the processor time to complete all intermediate states 2-7 (excluding the end-state simulations). protocol for the 150-atom CB7-C1 complex system. The automation application was able to achieve near perfect scaling as the number of processes increases.

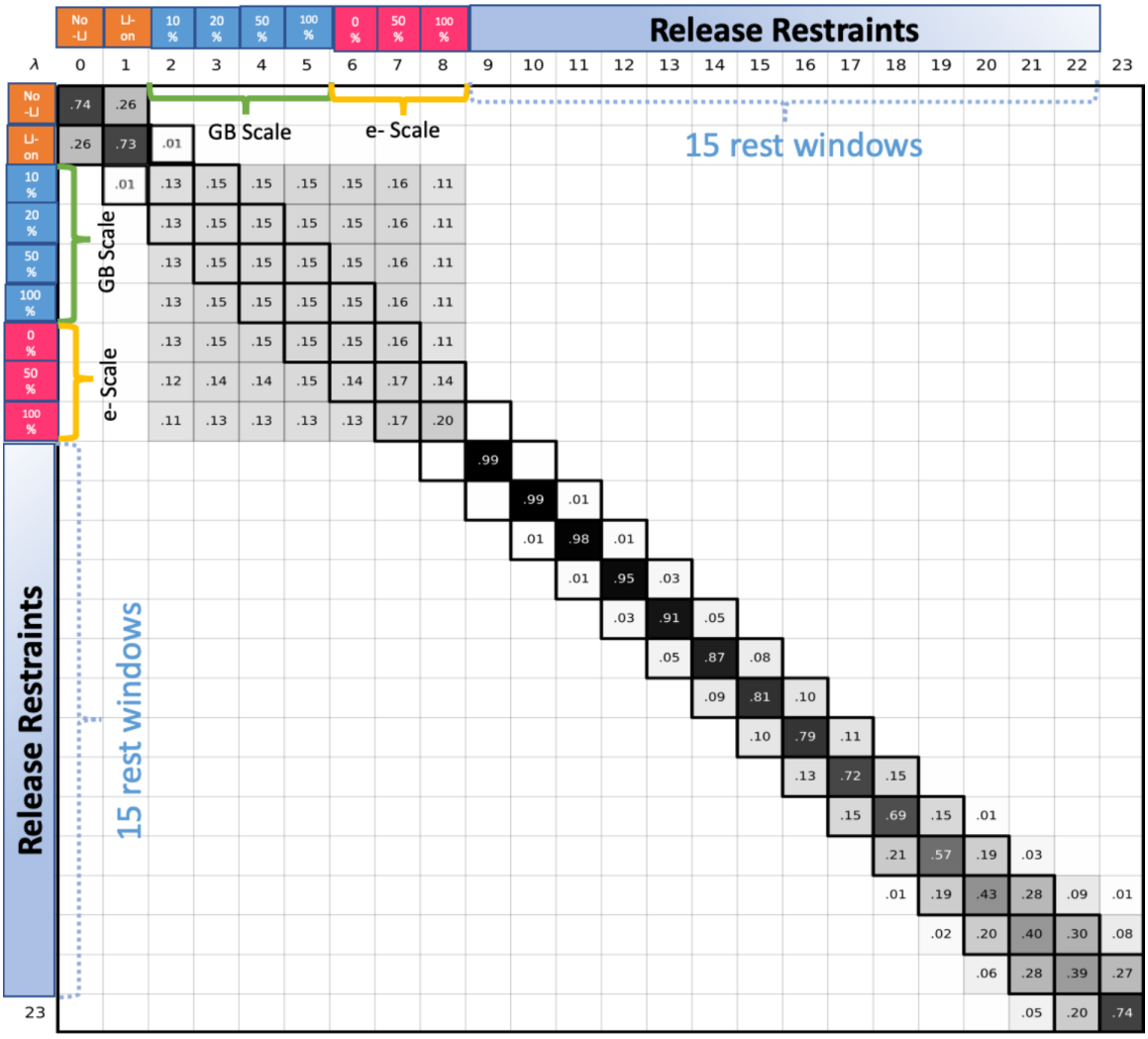

Figure 4.4: Overlap matrix for WP6-G1 with 15 restraint windows. The overlap matrix showcases all the windows for the WP6-G1 complex where every column in denotes denote every simulation performed in states 6-8 in figure 2.1**.** Column λ_0 denotes state 6, where the Lennard-Jones interactions of the ligand were turned off . In state λ_1, ligand Lennard-Jones interactions are reintroduced in the simulation. In λ_{2-5}, the GB external dielectric constant is set to 5%, 10%, 20% 50% and 100% of its final value, 78.5. In λ_{6-8}, the electrostatics of the ligand charges were set to 0%, 50% and 100%. States λ_{9-22} correspond to the release of conformational from $2^4 - 2^{-8}$ kcal/mol and orientational restraints from $2^8 - 2^{-4}$ kcal/mol with uniform logarithmic spacing. λ_{23} is the end-state simulation of the complex. The overlap matrix with 15 restraints windows shows no overlap for state 8 to state 11. The probability of finding a microstate sampled in state 11 in state 12 is 0.01.

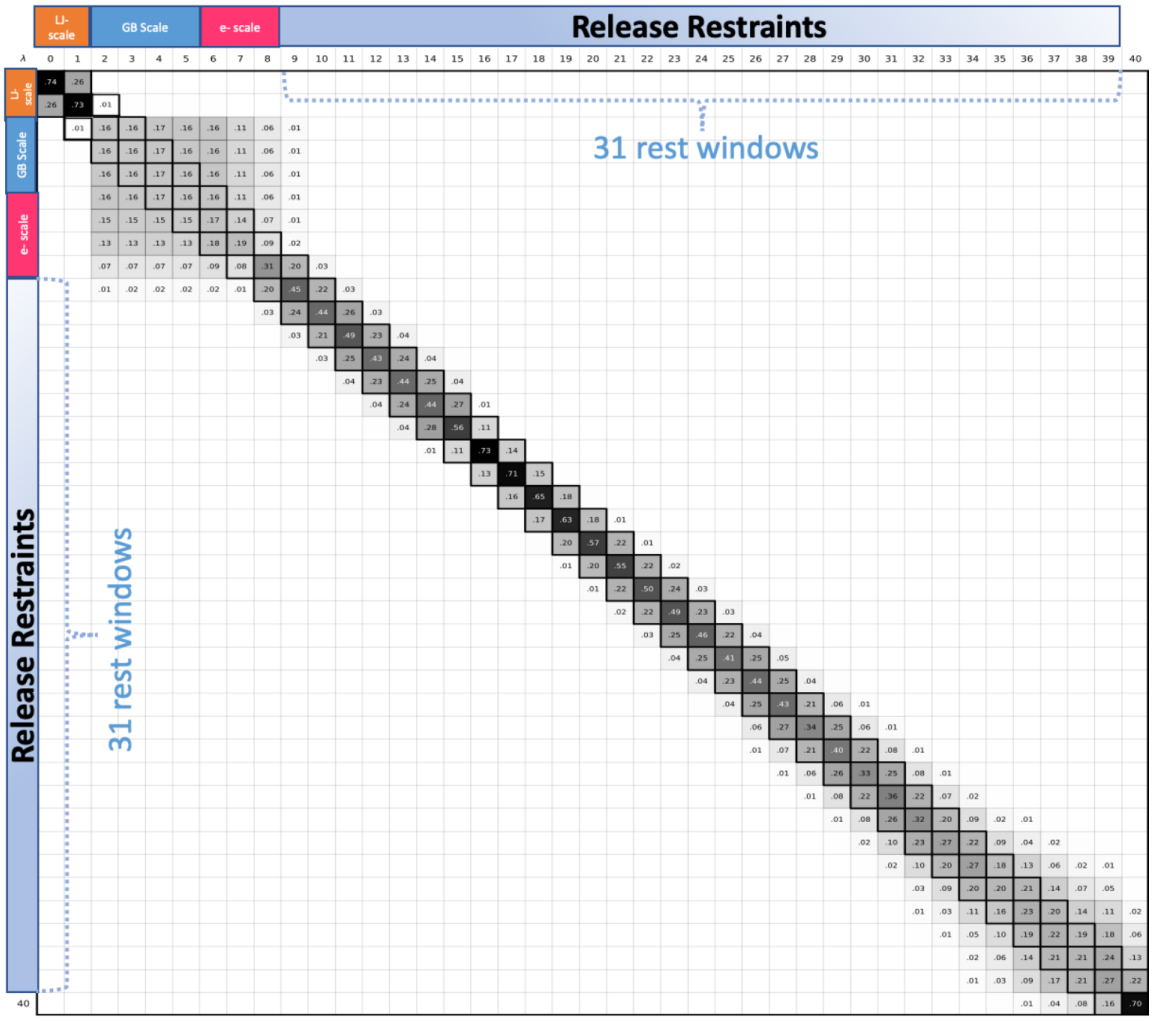

Figure 4.5: Overlap matrix for WP6-G1 with 29 restraint windows for states 6-8 in figure 2.1. There is good overlap throughout all intermediate windows.

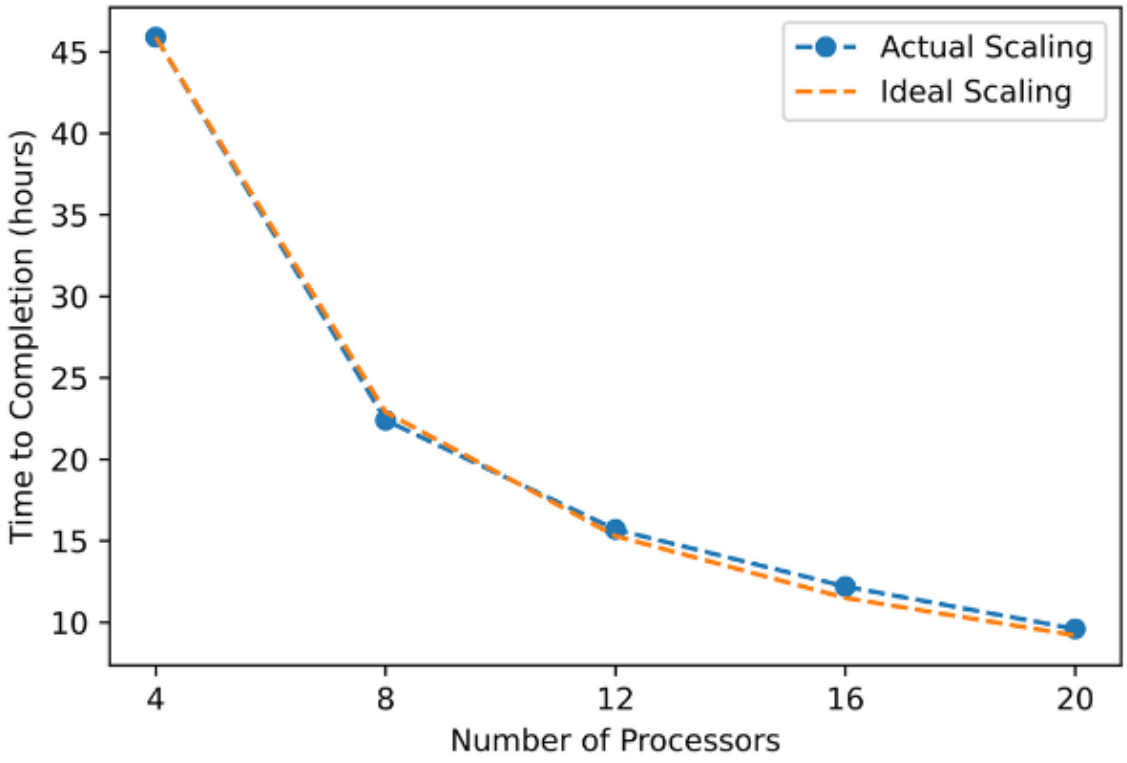

Figure 4.6: `ISDDM.py` processor scaling performance for the 150-atom CB7-C1 complex. The y-axis is the times required for entire workflow to finish and x-axis is the number of processors used.

Chapter 5

Discussion

5.1 Comparison with Other Methods

5.1.1 CB7 SAMPL4 Comparison

Overall, binding affinity prediction for `ISDDM.py` workflow performed well for CB7 system compared to other submission on to the SAMPL4 challenge [43]. If our uncorrected results were submitted, they would have ranked as one of the top performers among the challenge submissions (table 5.1) and our corrected results would have out performed all entries. Only two other methods matched our uncorrected Pearson correlation ($R^2 = 0.78$): the RRHO rigid rotor harmonic oscillator approximation [80], which employed the COSMO-RS (conductor-like screening model for real solvents) solvent model, and the OST (orthogonal space tempering) with TIP3P-MOD (transferable interaction potential three-point). It should be noted that both these methods are significantly more computational demanding, where COSMO-RS utilizes quantum chemical calculations, and the OST method employs an explicit solvent model. However, all the other metrics of the uncorrected method are better than for these two methods.

Two methods used Bennet acceptance ratio (BAR) method, which evaluates free energy differences between two states at a time, rather than all states as done by MBAR. The two submission that used BAR both used atomic multipole optimized energetics for biomolecular simulation (AMOEBA) force field with explicit solvent. Both had a slope of 1.3. However, our uncorrected method had a better slope, RMSE_o and RMSE_r than these methods.

The final three entires in table 5.1 are retrospective studies that used three variations on the QM-VM2 method [81], which utilizes quantum mechanics with implicit solvent. QM-VM2 restricted Hartree-Fock (RHF-D3) and QM-VM2(TPSS-D3) have an improved slope, but similar metrics compared to our uncorrected method. The quantum mechanical method should provide a more accurate model but is computationally intensive. However, we see our faster sampling uncorrected method is achieving same level of accuracy with quantum mechanical method and is promising in virtual screening campaigns.

5.1.2 WP6 SAMPL9 Comparison

There were eight submissions for WP6 dataset from SAMPL9 challenge. As the official results have not been published at this time, the submissions were taken from the GitHub page [82]. Looking first at the RMSE (table 5.2), the DOCKING/SMINA/VINARDO method had the smallest RMSE ($1.8\,\text{kcal/mol}$), followed by the machine learning method with RMSE $= 1.97\,\text{kcal/mol}$ and extended linear interaction energy (ELIE) method with RMSE $= 2.5\,\text{kcal/mol}$. However, all three methods had very low Pearson correlation coefficients of 0.15, 0.40 and 0.09 for machine learning, ELIE and Docking/SMINA method respectively, which suggests a poor correlation between experimental and prediction binding

Solvent Model	Method	Slope	R^2	$RMSE_o$	$RMSE_r$	MAE	MSE
Ours-GBSA	MBAR(DDM/GB)	1.4	0.78	1.8	2.6	6.7	–6.7
Ours-GBSA (ΔG^{GB}_{corr})	MBAR(DDM/GB)	0.81	0.8	0.94	1.38		
BRI BEM	SIE	0.18	0.6	1.8	2.7	-	-
BRI BEM	SIE +HB	0.19	0.6	1.8	2.6	-	-
AMSOL	DOCK 3.7	-0.5	0.1	5.4	7.9	-	-
COSMO-RS	RRHO	1.8	0.8	2.5	3.7	-	-
COSMO	QM-M2	0.7	0.2	3	4.5	-	-
PBSA	M2	2	0.7	3.4	5	-	-
TIP4P	FEP	2.2	0.6	4.9	7.2	-	-
TIP4P	FEP	1.8	0.6	3.9	5.7	-	-
TIP4P-EW	FEP	1.6	0.3	5.6	8.2	-	-
TIP4P-EW	FEP	1.3	0.3	4.7	6.9	-	-
TIP3P/4P-EW	Enthalpy	1.6	0.7	2.7	4	-	-
TIP3P	Enthalpy	2.2	0.7	4	5.8	-	-
TIP3P-EW	Enthalpy	1.4	0.6	2.7	4	-	-
AMOEBA	BAR	1.3	0.6	2.2	3.3	-	-
TIP4P	Water Count	-0.4	0	5.8	8.6	-	-
TIP3P-MOD	OST	1.4	0.8	1.9	2.8	-	-
TIP3P	PMF	-0.4	0.1	4	5.8	-	-
AMOEBA	BAR	1.3	0.6	2.2	3.3	-	-
TIP3P	EES	1.9	0.7	3.4	5	-	-
TIP3P	EES	1.5	0.6	2.9	4.2	-	-
QM-VM2(DFTB3-D3)**	-	0.3	0.0	3.6	5.2	7.2	-7.2
QM-VM2(RHF-D3)**	-	1.0	0.6	1.7	2.5	2.3	-2.3
QM-VM2(TPSS-D3)**	-	0.9	0.7	1.4	2.1	2.4	2.2

Table 5.1: Predictions from SAMPL4 CB7 dataset. All entries were from the SAMPL4 CB7 dataset.Units for RMSE_std, $RMSE_o$, $RMSE_r$, MAE and MSE are in kcal/mol.

affinity. Whereas both our uncorrected and corrected Pearson correlation rank the second highest, only slightly lower than DDM/AMOEBA/BAR method with $R^2$of 0.56.

The other five methods were had an average RMSE of about $\gtrsim 3\,\mathrm{kcal/mol}$. The RMSE for our uncorrected ΔG_{gb}^{bind} ($6.6\,\mathrm{kcal/mol}$) was the largest deviation of the rank submissions. The trained MM/PBSA ELIE method is very similar to end-point calculations using an implicit solvent model GBSA. However this method required some system-specific training step, sharing some similarities with ML based method, and it is not fully physics based nor directly transferable to other systems, especially for targets without known binding affinities [83]. However, our corrected results would have had the lowest RMSE error ($1.39\,\mathrm{kcal/mol}$) compared to all participants.

Method	RMSE	R^2	MAE	MSE
Ours-GBSA	6.6	0.52	6.1	0.10
Ours-GBSA (ΔG_{corr}^{GB})	1.39	0.51	1.12	-0.1
DDM/AMOEBA/BAR	2.8	0.56	2.03	-0.8
APR/GAFF2/TIP3P/MD-US/MBAR	6.7	0.08	6.2	6.1
APR/OFF1.2.0/TIP3P/MD-US/MBAR	3.1	0.09	2.2	1.5
APR/OFF2.0.0/TIP3P/MD-US/MBAR	3.8	0.11	2.9	2.8
MACHINE-LEARNING/NNET/DRAGON-descriptors	1.97	0.15	1.5	0.61
ELIE/GAFF2-ABCG2/TIP3P/MD/MMPBSA	2.5	0.40	1.9	1.9
vDSSB/GAFF2/OPC3/HREM	3.7	0.36	3.3	2.8
DOCKING/SMINA/VINARDO	1.8	0.091	1.5	-0.57

Table 5.2: Predictions from SAMPL9 WP6 dataset. All entries were from the SAML9 dataset. nits for RMSE_std, RMSE$_o$, RMSE$_r$, MAE and MSE are in kcal/mol.

5.1.3 OAH SAMPL4 comparison

To the best of our knowledge, there are no published benchmarks of the entire dataset used here. However, the SAMPL4 challenge used a smaller set of nine guests molecules [43]. One submission of particular interest used the BEDAM method [84] to evaluate binding free energy with the OPLS-AA classical force field [85] and Analytical Generalized Born plus NonPolar (AGBNP2) [86] continuum solvation model. This was one of the most accurate predictions of the SAMPL4 challenge, with RMSE$_o = 1.3\,\mathrm{kcal/mol}$, RMSE$_r = 0.9\,\mathrm{kcal/mol}$ and one of the highest Person correlation coefficient ($R^2 = 0.9$). Restricting our results to the same nine guests, our uncorrected method gave RMSE$_o = 1.3\,\mathrm{kcal/mol}$ and RMSE$_r = 2.0\,\mathrm{kcal/mol}$ and correlation of $R^2 = 0.64$. In contract to the GB method we used, the AGBNP2 method was tuned to properly treat enclosed hydration sites to achieve good accuracy [87]. This suggests that our results could be improved with an more physically realistic solvent model.

As an example of how a better solvent model would improve our results, our uncorrected predictions for octa-acid were adversely affected primarily the much lower binding free energies predicted for the positively charge ligands (trz, hxa), which remain too low, even after correction. Beyond charge hydration asymmetry, this may be the result of the

large negative charge ($-8e$) on the host, coupled with a lack of charge screening from our GB model. Together, these may have caused the conformational reorganization of a benzoate ring for positively charge guests. During our end-state simulations, we also observed the alkylammonium head group of the ligand forming strong short range electrostatic interactions with the benzoate group (figure 5.1). A previous study showed that better incorporation of ionic screening can disrupt this interaction and weaken binding [88]. Where the negative charge guests, performs charge repulsion keeping like charges away from each other as we saw with our results.

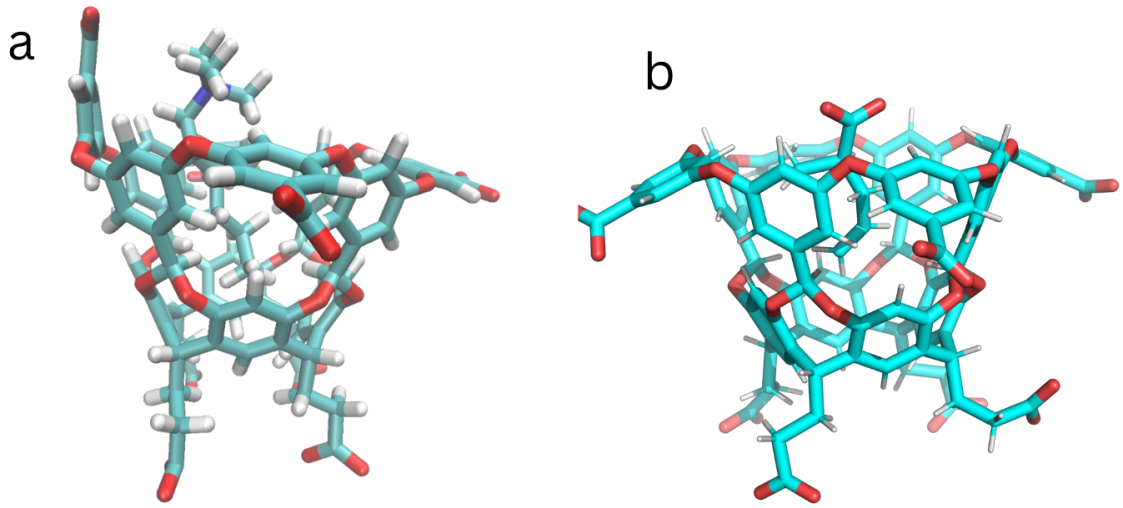

Figure 5.1: OAH-binding models with cationic and negatively charge guest. a. Benzoate ring flip to form a short range ionic interactions with the alkylammonium head group of the guest (hxa). b the negativity charge guest (m4p) without benzoate ring flipping.

5.2 Generalize Born Hydration Asymmetry Correction Considerations

The simplicity of the GB implicit solvent model provides fast estimation of solvation effects, making it very popular for virtual screening. GB is arguably one the faster implicit models, as it approximates the Poisson-Boltzmann equation with a simple analytical formula. However, the version used here is unable to account for several solvation properties that are accounted for in explicit solvent models.

One major approximation is the neglect of charge hydration asymmetry (CHA) [79], which contributed towards solvation free energy ΔG_{solv} on the sign of the charges of the solute. When considering our entire host-guest dataset (figure 4.2a), we clearly see a bias for all charged guest ligands. The absolute binding free energy estimates for positively charge guests were underestimated and negatively charge guests were overestimated. We also observe that the guest molecules have the dominant source of CHA error, as the charge of the host system had little or no effect.

Applying a charge hydration correction based on the guests net charge (table 4.2), resulted in better metrics (table 4.3) across all host-guest systems. The success of this cor-

Method	RMSE	R^2	MAE	MSE	RMSE_o	RMSE_r
Ours-GBSA	1.8	0.64	1.5	1.2	1.3	2.0
Ours-GBSA (ΔG^{GB}_{corr})	1.05	0.64	1.0	0.14	1.0	1.6
DFT-opt/LCCSD	7.4	0.37	5.1	-4.4	6.11	9.2
BEDAM/AGBNP2	1.68	0.9	1.4	1.4	0.86	1.3
NBB/DFT-D3	6.2	0.01	5.2	-4.2	4.3	6.5
Dock/WILMA/SIE	1.3	0.67	1.2	0.90	0.96	1.4
FEP/AM1-BCC	1.2	0.85	0.90	0.6	1.02	1.5
Dock/WILMA/SIE+explicit H-bonding terms	1.3	0.76	1.2	0.96	0.91	1.4
DFT-opt	5.8	0.53	3.6	0.4	5.8	8.7
expanded ensemble/explicit	1.7	0.81	1.3	1.1	1.3	1.9
FEP/RESP	1.0	0.90	0.8	0.5	0.86	1.3
DOCK3.7	5.8	0.7	5.7	5.7	0.9	1.4
DFT-D3/COSMO-RS	6.4	0.35	5.0	5.0	3.9	5.9

Table 5.3: Predictions from SAMPL4 OAH dataset. All entries were from the SAML4 dataset(consist of only 9 guest molecules).The guest molecules that were tested against octa-acid were ben, ebn, c3b, mbn, c4b, mhc, c5c, c7c and chxfigure 3.1. Units for RMSE_std, RMSE_o, RMSE_r, MAE and MSE are in kcal/mol.

rection and the values of the fit parameters suggests that the lack of charge hydration asymmetry in GB is the issue. The Amber molecular modeling suite contains four variants of GB models, which are GB^{HCT}[89], GB^{OBC}[60],GB^{neck2}[90] and GBNSR6 [91]. Of these, GBNSR6 is unique in how the effective Born radii are computed. GBNSR6 ("Numerical Surface R6 GB"), one of newest additions to Amber suite, computes the effective Born radii using an "R6" numerical integration overall the molecular surface of the solute and optionally accounts for CHA. The CHA correction has been shown to improved electrostatic binding free energies for diverse set of biomolecular complexes [92]; however, GBNSR6 does not produce analytic forces and cannot be used for MD simulations, which mean it cannot be used directly in our method. Alternately, the 3-dimensional reference interaction site model (3D-RISM) of molecular solvation [93, 94, 95], which describes the solvent as a classical density distribution of the solvent sites around the solute, yields a more detailed molecular structure than GB models and naturally accounts for properties such as CHA and hydrogen bonding. Unfortunately, 3D-RISM is computational demanding and not suitable for high throughput screening campaigns at this time. However, improved implementation of 3D-RISM may change this.

5.3 Scalable Workflow

Designing a scalable, fully automated absolute binding free energy workflow that can manage the executions of simulations and takes advantage of high-performance computing resources is a desirable tool for drug discovery pipelines. Our automation application was able to achieve near-perfect scaling as the number of processes increases for the 150-atom CB7-C1 complex system as seen in figure 4.6. These results suggest we have strong scaling performance in which the time of completion decreases rapidly as we add more CPU cores to our parallel workflow. A potential bottleneck of the scalability of our workflow is moving toward larger systems, such as proteins, which contain a few thousand heavy atoms. As the system grows, we expect the time of completion to increase as the number of atoms squared which will affect our overall turnaround time. However, our automated workflow is designed to support GPU-accelerated engines such as `pmemd.cuda`. In the future, as we move to small proteins using GPUs will be benchmarked and tested for performance scaling.

Chapter 6

Conclusion

The goal of this research was to create a simpler user interface that can automate a rigorous, modified ABFE cycle, allowing it to be applied over a broad chemical space. Although other automated absolute binding free energies programs exist [96, 97], `ISDDM.py` is unique in that it improves sampling by using a combination of implicit solvents and conformational restraints. To test the automation of `ISDDM.py`, we applied it to a series of host-guest systems, most of which were presented as community challenges by SAMPL group. We saw overall good correlation between experiment and our method (DDM/GB), which outperformed most the SAMPL4 and SAMPL9 challenge submission for RBFE metrics, though large systematic errors were present in the raw results.

One major issue with using the GB solvation method is its inability to account charge hydration asymmetry, leading towards the strong electrostatics interactions from charge guest molecules and bias in the results. From all the host guest systems shown in figure figure 4.2a, we saw that positively charged guest molecules had significantly underestimated binding free energy predictions, and negatively charge guests were overestimated. By applying a simple linear correction based on guest net charge, we observed much better correlation with and deviation from experiment, with RMSE values less than $1.4\,\mathrm{kcal/mol}$. This correction greatly improved our overall accuracy across the entire dataset; however, this correction is likely not transferable. That is, the correction is not a solution but provides support that charge hydration asymmetry is the dominant source of error. Rather, a more accurate solvation model that could account for CHA and other thermodynamic water properties GB lacks will need to be employed.

In future analysis, we plan on both improving the solvation model and testing the method retrospectively on protein-ligand predictions and prospectively on host-guest systems. Protein-ligand binding affinity is public available in databases such as BindingDB [98], which includes curated datasets from the Drug Design Data Resource (D3R) and Community Structure Activity Resource (CSAR). Participation in future blind challenges, such as SAMPL, will test our method's ability to make prospective predictions. In addition, we will attempt to improve the overall accuracy by bookending with more accurate solvation models such as 3D-RISM. 3D-RISM is computational demanding but potential remedies for this could include implementing 3D-RISM on GPUs or using Monte Carlo sampling scheme rather molecular dynamics route. Overall, `ISDDM.py`shows promise to perform fast, accurate absolute binding free energy calculations that combines the sampling efficiency of GB in a rigorous and fully automated application.

Appendix A

Configuration user input

The user controls the entire `ISDDM.py` workflow by supplying a YAML-format user input files, which consists of only 36 lines. Each parameter in the input file is defined below.

A.1 `system_parameters`

The system parameters denote the software and hardware resources there user wishes to use.

`executable:str` The Amber MD engine the user wishes to execute (`sander`, `pmemd`, `pmemd.CUDA`, etc.) for the the end-state and intermediate MD simulations.

`mpi_command:str` The command used to run MPI programs on the computer system; e.g., `mpirun`, `mpiexec`, or `srun`.

`memory:float` Amount of memory needed to run an intermediate MD simulation.

`storage:float` Storage required for an individual MD simulation. Units are in Gigabytes (default=2G).

A.2 `number_of_cores_per_system`

Number of processor to be used for each individual ligand, receptor and complex system.

`complex_ncores:int` Number of processors to be used for a single complex intermediate molecular dynamics simulation.

`receptor_ncores:int` Number of processors to be used for a single receptor intermediate molecular dynamics simulation.

`ligand_ncores:int` Number of processors to be used for a single ligand intermediate molecular dynamics simulation.

A.3 `endstate_parameter_files`

Users need to specify a path to AMBER `parm7` and `rst7` files to designated bound receptor-ligand-complex system. The user must first parameterize the complex with there desired force fields and charge model. Lastly, Amber masks are used to denote ligand and receptor atoms from the complex parameter file.

`complex_parameter_filename:str` The path to an AMBER `parm7` topology file, which defines which atoms are bonded to each other.

complex_coordinate_filename:str The path to an AMBER `rst7` file, which defines where each atom is located on a 3-dimensional coordinate plane.

receptor_mask:str Amber mask syntax to select receptor atoms

ligand_mask:str Amber mask syntax to select ligand atoms

A.4 workflow

endstate_method:str End-state simulation type. The default setting, `remd`, end-state simulation runs replica exchange molecular dynamics. If the user wishes to run a standard MD simulation, will specify `standard_MD`.The user can specify their own pre-calculated end-state simulation by setting value to 0 .

A.4.1 endstate_arguments

Users must define the end-state simulation to use. REMD is the default setting, but users can specify there own end-state simulation as well.

nthread_complex:int. (Optional) Number of processes to run for all complex replicas.

nthread_receptor:int. (Optional) Number of processes to run for all receptor replicas.

nthread_ligand:int. (Optional) Number of processes to run for all ligand replicas.

ngroups:int (Optional) Specifies total number of copies (replicas).

target_temperature:float. (Optional) Target temperature to be extracted from REMD simulation.

remd_template_mdin:str. (Optional) Preference template that specifies general input `mdin` arguments for REMD simulations.

equilibrate_template_mdin:str. (Optional) Preference template that specifies general input `mdin` arguments equilibration simulation.

temperatures: list[float]. (Optional) Specify a list of temperature to be ran for REMD.

A.5 Intermediate States Arguments

The intermediate argument keywords asks for the general input parameters such as GB (GB^{HCT}[89],GB^{OBC}[60],$GB-neck2$[90], scaling restraints and ligand interactions.

igb:int GB version for all calculation: GB^{HCT} [89] (`igb=1`), GB^{OBC} [60] (`igb=2`), GB^{OBC} [99] (`igb=5`), $GB-neck2$ [90] (`igb=8`).

exponent_conformational_forces:list[float] Strength of harmonic conformational restraints are specified by a list of integers to calculate powers of 2, for example -8 gives a restraint coefficient of $2^{-8}\,\mathrm{kcal/mol} = 0.00390625$ kcal/mol.

exponent_orientational_forceslist[float]: Strength of harmonic orientational restraints are specified by a list of integers to calculate powers of 2.

restraint_type: int Orientational restraint type to generate orientational based on center of mass of ligand and receptor. default:2.

temperature: float (Optional.) Temperature to run intermediate simulations if complete end-state simulations are provided.

charges_lambda_window: list[float] Values to use when scaling the GB external dielectric.

$$\Delta G_{\text{bind}}^{\mathbf{GB}} \left(\frac{kcal}{mol} \right)$$

Appendix B

Sample Input File

```
system_parameters:
        working_directory: working/path/directory
        executable: pmemed.MPI
        mpi_command: srun # system dependent
        top_directory_name: output_directory_name

endstate_parameter_files:
        complex_parameter_filename:  cb7-mol01_Hmass.parm7
        complex_coordinate_filename:  cb7-mol01_Hmass.nc

number_of_cores_per_system:
        complex_ncores: 2
        ligand_ncores: 1
        receptor_ncores: 1

AMBER_masks:
        receptor_mask: ':CB7'
        ligand_mask: ':M01'

workflow:
        endstate_method: remd
        endstate_arguments:
                nthreads_complex: 8
                nthreads_receptor: 8
                nthreads_ligand: 8
                ngroups: 4
                target_temperature: 300
                remd_template_mdin: remd.template
                equilibrate_mdin_template: equil.template
                temperatures: [260, 280, 300, 320]
        intermidate_states_arguments:
                mdin_intermidate_config: inter.yaml
                igb_solvent: 2
                temperature: 300
                exponent_conformational_forces: [-4, 2,4]
                exponent_orientational_forces: [-8, 4, 8]
                restraint_type: 2
                charges_lambda_window: [0, 0.5, 1]
                gb_extdiel_windows: [0.5, 1]
```

Bibliography

[1] Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to Improve R&D Productivity: The Pharmaceutical Industry's Grand Challenge. *Nature Reviews Drug Discovery* **2010**, *9*, 203–214.

[2] Leal, W.; Llanos, E. J.; Bernal, A.; Stadler, P. F.; Jost, J.; Restrepo, G. The Expansion of Chemical Space in 1826 and in the 1840s Prompted the Convergence to the Periodic System. *Proceedings of the National Academy of Sciences* **2022**, *119*, e2119083119.

[3] Reymond, J.-L. The Chemical Space Project. *Accounts of Chemical Research* **2015**, *48*, 722–730.

[4] Jhoti, H.; Rees, S.; Solari, R. High-Throughput Screening and Structure-Based Approaches to Hit Discovery: Is There a Clear Winner? *Expert Opinion on Drug Discovery* **2013**, *8*, 1449–1453.

[5] Clark, D. E. What Has Computer-Aided Molecular Design Ever Done for Drug Discovery? *Expert Opinion on Drug Discovery* **2006**, *1*, 103–110.

[6] Anderson, A. C. The Process of Structure-Based Drug Design. *Chemistry & Biology* **2003**, *10*, 787–797.

[7] Batool, M.; Ahmad, B.; Choi, S. A Structure-Based Drug Discovery Paradigm. *International Journal of Molecular Sciences* **2019**, *20*, 2783.

[8] Wang, G.; Zhu, W. Molecular Docking for Drug Discovery and Development: A Widely Used Approach but Far from Perfect. *Future Medicinal Chemistry* **2016**, *8*, 1707–1710.

[9] Yu, Y.; Cai, C.; Zhu, Z.; Zheng, H. *Uni-Dock: A GPU-Accelerated Docking Program Enables Ultra-Large Virtual Screening*; Preprint, 2022.

[10] Friedrichs, M. S.; Eastman, P.; Vaidyanathan, V.; Houston, M.; Legrand, S.; Beberg, A. L.; Ensign, D. L.; Bruns, C. M.; Pande, V. S. Accelerating Molecular Dynamic Simulation on Graphics Processing Units. *Journal of Computational Chemistry* **2009**, *30*, 864–872.

[11] Stone, J. E.; Hardy, D. J.; Ufimtsev, I. S.; Schulten, K. GPU-accelerated Molecular Modeling Coming of Age. *Journal of Molecular Graphics and Modelling* **2010**, *29*, 116–125.

[12] Harvey, M. J.; Giupponi, G.; Fabritiis, G. D. ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. *Journal of Chemical Theory and Computation* **2009**, *5*, 1632–1639.

[13] Deng, Y.; Roux, B. Computations of Standard Binding Free Energies with Molecular Dynamics Simulations. *The Journal of Physical Chemistry B* **2009**, *113*, 2234–2246.

[14] Cournia, Z.; Allen, B. K.; Beuming, T.; Pearlman, D. A.; Radak, B. K.; Sherman, W. Rigorous Free Energy Simulations in Virtual Screening. *Journal of Chemical Information and Modeling* **2020**, *60*, 4153–4169.

[15] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* **1995**, *117*, 5179–5197.

[16] Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *Journal of Computational Chemistry* **2004**, *25*, 1157–1174.

[17] Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *Journal of Computational Chemistry* **1983**, *4*, 187–217.

[18] Christen, M.; Hünenberger, P. H.; Bakowies, D.; Baron, R.; Bürgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Kräutler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; van Gunsteren, W. F. The GROMOS Software for Biomolecular Simulation: GROMOS05. *Journal of Computational Chemistry* **2005**, *26*, 1719–1751.

[19] Onuchic, J. N.; Wolynes, P. G. Theory of Protein Folding. *Current Opinion in Structural Biology* **2004**, *14*, 70–75.

[20] Yang, Y. I.; Shao, Q.; Zhang, J.; Yang, L.; Gao, Y. Q. Enhanced Sampling in Molecular Dynamics. *The Journal of Chemical Physics* **2019**, *151*, 070902.

[21] Darve, E.; Rodríguez-Gómez, D.; Pohorille, A. Adaptive Biasing Force Method for Scalar and Vector Free Energy Calculations. *The Journal of Chemical Physics* **2008**, *128*, 144120.

[22] Patey, G. N.; Valleau, J. P. A Monte Carlo Method for Obtaining the Interionic Potential of Mean Force in Ionic Solution. *The Journal of Chemical Physics* **1975**, *63*, 2334–2339.

[23] Grubmüller, H.; Heymann, B.; Tavan, P. Ligand Binding: Molecular Mechanics Calculation of the Streptavidin-Biotin Rupture Force. *Science* **1996**, *271*, 997–999.

[24] Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proceedings of the National Academy of Sciences* **2002**, *99*, 12562–12566.

[25] Hamelberg, D.; Mongan, J.; McCammon, J. A. Accelerated Molecular Dynamics: A Promising and Efficient Simulation Method for Biomolecules. *The Journal of Chemical Physics* **2004**, *120*, 11919–11929.

[26] Hou, T.; Wang, J.; Li, Y.; Wang, W. Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 1. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations. *Journal of Chemical Information and Modeling* **2011**, *51*, 69–82.

[27] Miller, B. R.; McGee, T. D.; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E. *MMPBSA.Py* : An Efficient Program for End-State Free Energy Calculations. *Journal of Chemical Theory and Computation* **2012**, *8*, 3314–3321.

[28] Bennett, C. H. Efficient Estimation of Free Energy Differences from Monte Carlo Data. *Journal of Computational Physics* **1976**, *22*, 245–268.

[29] Kirkwood, J. G. Quantum Statistics of Almost Classical Assemblies. *Physical Review* **1933**, *44*, 31–37.

[30] Jorgensen, W. L.; Ravimohan, C. Monte Carlo Simulation of Differences in Free Energies of Hydration. *The Journal of Chemical Physics* **1985**, *83*, 3050–3054.

[31] Mobley, D. L.; Klimovich, P. V. Perspective: Alchemical Free Energy Calculations for Drug Discovery. *The Journal of Chemical Physics* **2012**, *137*, 230901.

[32] Schindler, C. E. M. et al. Large-Scale Assessment of Binding Free Energy Calculations in Active Drug Discovery Projects. *Journal of Chemical Information and Modeling* **2020**, *60*, 5457–5474.

[33] Feng, M.; Heinzelmann, G.; Gilson, M. K. Absolute Binding Free Energy Calculations Improve Enrichment of Actives in Virtual Compound Screening. *Scientific Reports* **2022**, *12*, 13640.

[34] Gilson, M.; Given, J.; Bush, B.; McCammon, J. The Statistical-Thermodynamic Basis for Computation of Binding Affinities: A Critical Review. *Biophysical Journal* **1997**, *72*, 1047–1069.

[35] 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]. *Living Journal of Computational Molecular Science* **2020**, *2*.

[36] Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. Absolute Binding Free Energies: A Quantitative Approach for Their Calculation. *The Journal of Physical Chemistry B* **2003**, *107*, 9535–9551.

[37] Mobley, D. L.; Chodera, J. D.; Dill, K. A. On the Use of Orientational Restraints and Symmetry Corrections in Alchemical Free Energy Calculations. *The Journal of Chemical Physics* **2006**, *125*, 084902.

[38] Durrant, J. D.; McCammon, J. A. Molecular Dynamics Simulations and Drug Discovery. *BMC Biology* **2011**, *9*, 71.

[39] 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. *Journal of Chemical Information and Modeling* **2020**, *60*, 5595–5623.

[40] Wang, J.; Hou, T.; Xu, X. Recent Advances in Free Energy Calculations with a Combination of Molecular Mechanics and Continuum Models. *Current Computer Aided-Drug Design* **2006**, *2*, 287–306.

[41] Wang, W.; Donini, O.; Reyes, C. M.; Kollman, P. A. Biomolecular Simulations: Recent Developments in Force Fields, Simulations of Enzyme Catalysis, Protein-Ligand, Protein-Protein, and Protein-Nucleic Acid Noncovalent Interactions. *Annual Review of Biophysics and Biomolecular Structure* **2001**, *30*, 211–243.

[42] Swanson, J. M.; Henchman, R. H.; McCammon, J. A. Revisiting Free Energy Calculations: A Theoretical Connection to MM/PBSA and Direct Calculation of the Association Free Energy. *Biophysical Journal* **2004**, *86*, 67–74.

[43] Muddana, H. S.; Fenley, A. T.; Mobley, D. L.; Gilson, M. K. The SAMPL4 Host–Guest Blind Prediction Challenge: An Overview. *Journal of Computer-Aided Molecular Design* **2014**, *28*, 305–317.

[44] Rizzi, A.; Murkli, S.; McNeill, J. N.; Yao, W.; Sullivan, M.; Gilson, M. K.; Chiu, M. W.; Isaacs, L.; Gibb, B. C.; Mobley, D. L.; Chodera, J. D. Overview of the SAMPL6 Host–Guest Binding Affinity Prediction Challenge. *Journal of Computer-Aided Molecular Design* **2018**, *32*, 937–963.

[45] Yin, J.; Henriksen, N. M.; Slochower, D. R.; Shirts, M. R.; Chiu, M. W.; Mobley, D. L.; Gilson, M. K. Overview of the SAMPL5 Host–Guest Challenge: Are We Doing Better? *Journal of Computer-Aided Molecular Design* **2017**, *31*, 1–19.

[46] Amezcua, M.; Mobley, D. L. Samplchallenges/SAMPL9: Version 0.1: Initial Files for SAMPL9 Host-Guest. Zenodo, 2021.

[47] Slochower, D. Taproom. 2022.

[48] Shirts, M. R. In *Computational Drug Discovery and Design*; Baron, R., Ed.; Springer New York: New York, NY, 2012; Vol. 819; pp 425–467.

[49] De Jong, D. H.; Schäfer, L. V.; De Vries, A. H.; Marrink, S. J.; Berendsen, H. J. C.; Grubmüller, H. Determining Equilibrium Constants for Dimerization Reactions from Molecular Dynamics Simulations. *Journal of Computational Chemistry* **2011**, *32*, 1919–1928.

[50] Kutzner, C.; Páll, S.; Fechner, M.; Esztermann, A.; Groot, B. L.; Grubmüller, H. More Bang for Your Buck: Improved Use of GPU Nodes for GROMACS 2018. *Journal of Computational Chemistry* **2019**, *40*, 2418–2431.

[51] Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *The Journal of Chemical Physics* **1954**, *22*, 1420–1426.

[52] Pohorille, A.; Jarzynski, C.; Chipot, C. Good Practices in Free-Energy Calculations. *The Journal of Physical Chemistry B* **2010**, *114*, 10235–10253.

[53] 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. *The Journal of Chemical Physics* **2005**, *122*, 144107.

[54] Shirts, M. R.; Chodera, J. D. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *The Journal of Chemical Physics* **2008**, *129*, 124105.

[55] Wu, D.; Kofke, D. A. Phase-Space Overlap Measures. I. Fail-safe Bias Detection in Free Energies Calculated by Molecular Simulation. *The Journal of Chemical Physics* **2005**, *123*, 054103.

[56] D.A. Case, H.M. Aktulga, K. Belfon, I.Y. Ben-Shalom, J.T. Berryman, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, G.A. Cisneros, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, K. Kasavajhala, M.C. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K.A. O'Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, A. Shajan, J. Shen, C.L. Simmerling, N.R. Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, J. Wang, H. Wei, R.M. Wolf, X. Wu, Y. Xiong, Y. Xue, D.M. York, S. Zhao, and P.A. Kollman (2021), Amber 2021, University of California, San Francisco.

[57] Bergazin, T. D.; Ben-Shalom, I. Y.; Lim, N. M.; Gill, S. C.; Gilson, M. K.; Mobley, D. L. Enhancing Water Sampling of Buried Binding Sites Using Nonequilibrium Candidate Monte Carlo. *Journal of Computer-Aided Molecular Design* **2021**, *35*, 167–177.

[58] Cournia, Z.; Allen, B.; Sherman, W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *Journal of Chemical Information and Modeling* **2017**, *57*, 2911–2937.

[59] Zhang, J.; Zhang, H.; Wu, T.; Wang, Q.; van der Spoel, D. Comparison of Implicit and Explicit Solvent Models for the Calculation of Solvation Free Energy in Organic Solvents. *Journal of Chemical Theory and Computation* **2017**, *13*, 1034–1043.

[60] Onufriev, A.; Bashford, D.; Case, D. A. Modification of the Generalized Born Model Suitable for Macromolecules. *The Journal of Physical Chemistry B* **2000**, *104*, 3712–3720.

[61] Ravindranathan, K.; Tirado-Rives, J.; Jorgensen, W. L.; Guimarães, C. R. W. Improving MM-GB/SA Scoring through the Application of the Variable Dielectric Model. *Journal of Chemical Theory and Computation* **2011**, *7*, 3859–3865.

[62] Amezcua, M.; El Khoury, L.; Mobley, D. L. SAMPL7 Host–Guest Challenge Overview: Assessing the Reliability of Polarizable and Non-Polarizable Methods for Binding Free Energy Calculations. *Journal of Computer-Aided Molecular Design* **2021**, *35*, 1–35.

[63] Guo, L.; Du, J.; Wang, Y.; Shi, K.; Ma, E. Advances in Diversified Application of Pillar[n]Arenes. *Journal of Inclusion Phenomena and Macrocyclic Chemistry* **2020**, *97*, 1–17.

[64] Deng, C.-L.; Cheng, M.; Zavalij, P. Y.; Isaacs, L. Thermodynamics of Pillararene·guest Complexation: Blinded Dataset for the SAMPL9 Challenge. *New Journal of Chemistry* **2022**, *46*, 995–1002.

[65] Gibb, C. L. D.; Gibb, B. C. Binding of Cyclic Carboxylates to Octa-Acid Deep-Cavity Cavitand. *Journal of Computer-Aided Molecular Design* **2014**, *28*, 319–325.

[66] Forli, S.; Huey, R.; Pique, M. E.; Sanner, M. F.; Goodsell, D. S.; Olson, A. J. Computational Protein–Ligand Docking and Virtual Drug Screening with the AutoDock Suite. *Nature Protocols* **2016**, *11*, 905–919.

[67] O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *Journal of Cheminformatics* **2011**, *3*, 33.

[68] Ambertools21 D.A. Case, H.M. Aktulga, K. Belfon, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, G.A. Cisneros, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, C. Jin, K. Kasavajhala, M.C. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K.A. O'Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, N.R. Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, Y. Xue, D.M. York, S. Zhao, and P.A. Kollman (2021), Amber 2021, University of California, San Francisco.

[69] Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *Journal of Computational Chemistry* **2002**, *23*, 1623–1641.

[70] Hopkins, C. W.; Le Grand, S.; Walker, R. C.; Roitberg, A. E. Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *Journal of Chemical Theory and Computation* **2015**, *11*, 1864–1874.

[71] Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chemical Physics Letters* **1999**, *314*, 141–151.

[72] Welcome — Pytraj 2.0.2.Dev0 Documentation. https://amber-md.github.io/pytraj/latest/index.html.

[73] Barton, Michael. Fast Absolute Binding Free Energy Calculations with Alchemical Pathways in Implicit Solvent Models. Diss. California State University, Northridge, 2021.

[74] Shirts, M.; Beauchamp, K.; Naden, L.; Chodera, J.; Rodríguez-Guerra, J.; Martiniani, S.; Stern, C.; Henry, M.; Fass, J.; Gowers, R.; McGibbon, R. T.; Dice, B.; Jones, C.; Dotson, D.; Burgin, T. Choderalab/Pymbar: 3.1.1. Zenodo, 2022.

[75] Klimovich, P. V.; Shirts, M. R.; Mobley, D. L. Guidelines for the Analysis of Free Energy Calculations. *Journal of Computer-Aided Molecular Design* **2015**, *29*, 397–411.

[76] Chodera, J. D. A Simple Method for Automated Equilibration Detection in Molecular Simulations. *Journal of Chemical Theory and Computation* **2016**, *12*, 1799–1805.

[77] Pandas Documentation — Pandas 2.0.0 Documentation. https://pandas.pydata.org/docs/.

[78] Vivian, J. et al. Toil Enables Reproducible, Open Source, Big Biomedical Data Analyses. *Nature Biotechnology* **2017**, *35*, 314–316.

[79] Mukhopadhyay, A.; Fenley, A. T.; Tolokh, I. S.; Onufriev, A. V. Charge Hydration Asymmetry: The Basic Principle and How to Use It to Test and Improve Water Models. *The Journal of Physical Chemistry B* **2012**, *116*, 9776–9783.

[80] Grimme, S. Supramolecular Binding Thermodynamics by Dispersion-Corrected Density Functional Theory. *Chemistry - A European Journal* **2012**, *18*, 9955–9964.

[81] Xu, P.; Sattasathuchana, T.; Guidez, E.; Webb, S. P.; Montgomery, K.; Yasini, H.; Pedreira, I. F. M.; Gordon, M. S. Computation of Host–Guest Binding Free Energies with a New Quantum Mechanics Based Mining Minima Algorithm. *The Journal of Chemical Physics* **2021**, *154*, 104122.

[82] The SAMPL9 Blind Prediction Challenges for Computational Chemistry. The SAMPL Challenges, 2023.

[83] Liu, X.; Zheng, L.; Qin, C.; Zhang, J. Z. H.; Sun, Z. Comprehensive Evaluation of End-Point Free Energy Techniques in Carboxylated-Pillar[6]Arene Host–Guest Binding: I. Standard Procedure. *Journal of Computer-Aided Molecular Design* **2022**, *36*, 735–752.

[84] Gallicchio, E.; Lapelosa, M.; Levy, R. M. Binding Energy Distribution Analysis Method (BEDAM) for Estimation of Protein-Ligand Binding Affinities. *Journal of Chemical Theory and Computation* **2010**, *6*, 2961–2977.

[85] Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society* **1996**, *118*, 11225–11236.

[86] Gallicchio, E.; Paris, K.; Levy, R. M. The AGBNP2 Implicit Solvation Model. *Journal of Chemical Theory and Computation* **2009**, *5*, 2544–2564.

[87] Gallicchio, E.; Chen, H.; Chen, H.; Fitzgerald, M.; Gao, Y.; He, P.; Kalyanikar, M.; Kao, C.; Lu, B.; Niu, Y.; Pethe, M.; Zhu, J.; Levy, R. M. BEDAM Binding Free Energy Predictions for the SAMPL4 Octa-Acid Host Challenge. *Journal of Computer-Aided Molecular Design* **2015**, *29*, 315–325.

[88] Pal, R. K.; Haider, K.; Kaur, D.; Flynn, W.; Xia, J.; Levy, R. M.; Taran, T.; Wickstrom, L.; Kurtzman, T.; Gallicchio, E. A Combined Treatment of Hydration and Dynamical Effects for the Modeling of Host–Guest Binding Thermodynamics: The SAMPL5 Blinded Challenge. *Journal of Computer-Aided Molecular Design* **2017**, *31*, 29–44.

[89] Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Pairwise Solute Descreening of Solute Charges from a Dielectric Medium. *Chemical Physics Letters* **1995**, *246*, 122–129.

[90] Mongan, J.; Simmerling, C.; McCammon, J. A.; Case, D. A.; Onufriev, A. Generalized Born Model with a Simple, Robust Molecular Volume Correction. *Journal of Chemical Theory and Computation* **2007**, *3*, 156–169.

[91] Aguilar, B.; Onufriev, A. V. Efficient Computation of the Total Solvation Energy of Small Molecules via the R6 Generalized Born Model. *Journal of Chemical Theory and Computation* **2012**, *8*, 2404–2411.

[92] Izadi, S.; Harris, R. C.; Fenley, M. O.; Onufriev, A. V. Accuracy Comparison of Generalized Born Models in the Calculation of Electrostatic Binding Free Energies. *Journal of Chemical Theory and Computation* **2018**, *14*, 1656–1670.

[93] Beglov, D.; Roux, B. An Integral Equation To Describe the Solvation of Polar Molecules in Liquid Water. *The Journal of Physical Chemistry B* **1997**, *101*, 7821–7826.

[94] Kovalenko, A.; Hirata, F. Self-Consistent Description of a Metal–Water Interface by the Kohn–Sham Density Functional Theory and the Three-Dimensional Reference Interaction Site Model. *The Journal of Chemical Physics* **1999**, *110*, 10095–10112.

[95] Luchko, T.; Gusarov, S.; Roe, D. R.; Simmerling, C.; Case, D. A.; Tuszynski, J.; Kovalenko, A. Three-Dimensional Molecular Theory of Solvation Coupled with Molecular Dynamics in Amber. *Journal of Chemical Theory and Computation* **2010**, *6*, 607–624.

[96] Fu, H.; Chen, H.; Cai, W.; Shao, X.; Chipot, C. BFEE2: Automated, Streamlined, and Accurate Absolute Binding Free-Energy Calculations. *Journal of Chemical Information and Modeling* **2021**, *61*, 2116–2123.

[97] Heinzelmann, G.; Gilson, M. K. *Automated Docking Refinement and Virtual Compound Screening with Absolute Binding Free Energy Calculations*; Preprint, 2020.

[98] Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: A Public Database for Medicinal Chemistry, Computational Chemistry and Systems Pharmacology. *Nucleic Acids Research* **2016**, *44*, D1045–D1053.

[99] Onufriev, A.; Bashford, D.; Case, D. A. Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model. *Proteins: Structure, Function, and Bioinformatics* **2004**, *55*, 383–394.