University of Southampton Research Repository

Copyright © and Moral Rights for this thesis and, where applicable, any accompanying data are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis and the accompanying data cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content of the thesis and accompanying research data (where applicable) must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holder/s.

When referring to this thesis and any accompanying data, full bibliographic details must be given, e.g.

Thesis: Author (Year of Submission) "Full thesis title", University of Southampton, name of the University Faculty or School or Department, PhD Thesis, pagination.

Data: Author (Year) Title. URI [dataset]

UNIVERSITY OF SOUTHAMPTON

Faculty of Engineering and Physical Sciences
School of Chemistry

Protein-Ligand Binding Free Energies From Ab-Initio Quantum Mechanical Calculations

DOI: 10.1002/0470841559.ch1

by

Lennart Gundelach

MSc, MSci

ORCiD: 0000-0002-7888-8818

A thesis for the degree of
Doctor of Philosophy

May 2023

University of Southampton

Abstract

Faculty of Engineering and Physical Sciences
School of Chemistry

Doctor of Philosophy

Protein-Ligand Binding Free Energies From Ab-Initio Quantum Mechanical Calculations

by Lennart Gundelach

The accurate prediction of protein-ligand binding free energies with tractable computational methods has the potential to revolutionize drug discovery. Modeling the protein-ligand interaction at a quantum mechanical level, instead of relying on empirical classical mechanical methods, is an essential step toward this goal. In this body of research, we explore the QM-PBSA method to calculate quantum mechanical free energies of binding based on full-protein linear-scaling density functional theory calculations. We apply the QM-PBSA method to the T4-lysozyme L99A/M102Q protein and investigate the convergence, precision, and reproducibility of the QM-PBSA method. Additionally, we compare three different exchange-correlation functionals and different empirical dispersion corrections. Building on our findings in the well-characterized T4-lysozyme we calculate quantum mechanical protein-ligand free energies of binding for a set of ligands binding to the pharmaceutically highly relevant bromodomain containing protein 4 (BRD4) after an extensive investigation of the protein system at the classical mechanical level. BRD4 plays a key role in many cancers. The inhibition of BRD4 can suppress the cancer growth of acute myeloid leukemia, diffuse large B cell lymphoma, prostate cancer, and breast cancer. We demonstrate the predictive power of QM-PBSA in BRD4 as compared to its classical mechanical analog MM-PBSA and show, beyond doubt, that full-protein quantum mechanical calculations are both viable and tractable on modern supercomputers and in an academic context.

Contents

List of Figures

List of Tables

Declaration of Authorship

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

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as: [1, 2]

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

Acknowledgements

Thank you to Chris Skylaris, Thomas Fox, and Christofer Tautermann for their guidance and supervision. I am grateful to the EPSRC, TMCS-CDT, and Boehringer Ingelheim for the funding that made this research possible. Additionally, I would like to acknowledge the supercomputer facilities and their support teams, IRIDIS5, at the University of Southampton, the national supercomputers ARCHER and ARCHER2, and Young at UCL. Thank you also to Davide Sarpa for his feedback on the manuscript. Lastly, I would like to thank my parents for watching Katherine many an afternoon while I finished this thesis.

Chapter 1

Introduction

Proteins are the fundamental building blocks of life and are involved in almost every physiological process in the human body. Consequently, proteins play a crucial role in virtually all diseases. By changing the function of specific proteins, the processes involved in disease pathways can be altered or disrupted, leading to a reduction in symptoms or complete recovery. This is achieved by administering pharmaceutical drugs which circulate through the body and bind to the target protein, changing or disrupting its function. The discovery of new molecules which can bind to target proteins is one of the fundamental challenges of medicinal chemistry. Traditionally, the strength and selectivity of the binding of a small drug molecule to a target protein were determined experimentally through a variety of experimental protocols of varying accuracy. However, this approach requires the synthesis of the candidate molecule and expensive and time-consuming experimental procedures.

For more than 60 years, chemists have been pursuing computational methods for predicting the binding of drug molecules, also referred to as ligands, to proteins as an alternative approach. While today computational methods play a key role in pharmaceutical drug discovery and design, our ability to predict the binding of diverse small molecules accurately and reliably to the thousands of proteins in the human genome is still limited [3–6]. One contributing factor is that most computational methods developed to predict protein-ligand binding solve only the equations of classical mechanics, ignoring, for the most part, the more complex and subtle role that quantum mechanical interactions play in protein-ligand binding. Ignoring quantum mechanical effects greatly simplifies the mathematics and computational cost of simulations but omits fundamental physical interactions like polarization, charge transfer, and many-body quantum effects [7–9]. While there are classical mechanical simulations techniques that aim to include some of the quantum effects via severe approximations or empirical corrections, they fundamentally fail to capture the full range of chemistry and physics that occurs at the interface of a protein and a drug molecule[7, 8].

In recent years, the pursuit of quantum mechanical protein-ligand binding simulations has become tractable due to methodological advancements and the availability of ever-increasing computational resources. This body of research explores the application and development of one such quantum mechanical protein-ligand binding free energy method, QM-PBSA.

In the following two chapters, the theoretical background and mathematical underpinning of protein-ligand binding are introduced, as well as the computational techniques utilized. The QM-PBSA quantum mechanical binding energy method is motivated and described in detail. In Chapter 4, the QM-PBSA study of 7 ligands binding to the protein T4-lysozyme is presented in which the optimal method parameters, like the choice of exchange-correlation functional and convergence of the quantum mechanical binding energies, are explored. Chapter 5 introduces the protein system BRD4 and our thorough classical mechanical investigation, which lays the foundation for the QM-PBSA quantum mechanical study of BRD4. The conclusion in Chapter 6 summarizes the overall findings and outlines potential future extensions to the research presented in this thesis.

Chapter 2

Background

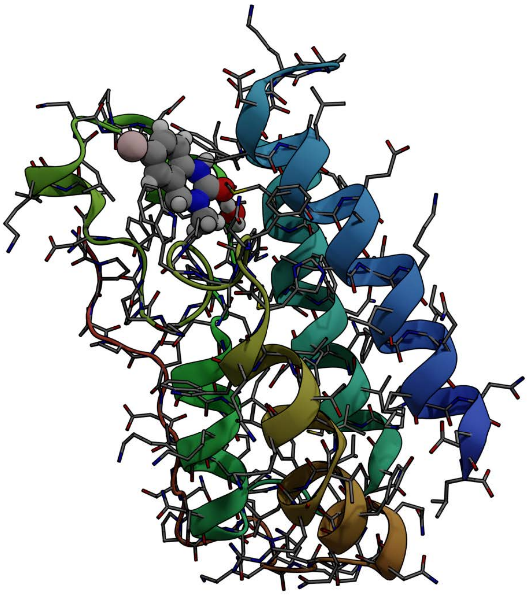


FIGURE 2.1: Stick and ribbon representation of the BRD4(1) protein structure. A small ligand, represented as a space-filling model,is also shown in the binding site.

2.1 Protein-ligand binding

Protein-ligand binding refers to the reversible non-covalent interaction in which a protein and second molecule form a complex. This process is shown diagrammatically in Figure 2.2. The term ligand refers to a molecule that binds to the protein with high affinity and specificity and could be a small drug-like molecule, second protein, or diverse range of bio-molecules. In the context of this thesis, protein-ligand binding refers to a small drug-like molecule binding to a target protein. The interaction is often described as molecular recognition, where an active site on the protein ”recognizes”

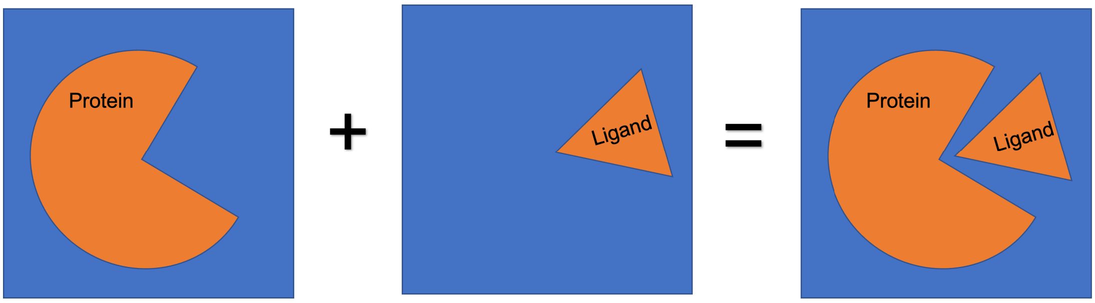


FIGURE 2.2: Simple schematic of the binding of a drug molecule (ligand) to the binding site of a target protein to form a protein-ligand complex.

specific ligands. The simplified metaphor of a lock, the protein binding site, and key, the ligand, conveys the basic idea of protein-ligand binding and is schematically visualized in Figure 2.2.

A more nuanced view of protein-ligand binding acknowledges the vital role of conformational changes in the protein-ligand binding site. The induced fit theory [10] describes the possibility of ligand-induced conformational changes in the binding site in which protein residues are re-arranged into the necessary alignment for binding. The theory of conformational selection [11] posits that the unbound protein can adopt different conformations and that the ligand may preferentially bind to some of these conformations. In reality, both the conformational motion of the protein and ligand flexibility play a crucial role in protein-ligand binding. Complexation is a fundamentally dynamic process. The binding is mediated through the complex interplay of non-covalent interactions like van der Waals, hydrogen bonding, hydrophobic forces, halogen bonding, and $\pi - \pi$ interactions. Additionally, interactions with the surrounding solvent and local conformational changes in the binding site play an important role. The interplay of these varied factors makes protein-ligand binding incredibly complex and difficult to predict accurately. As a result, no single physical or chemical theory can explain which ligands will bind how strongly to a given protein binding site and why.

Experimentally, protein-ligand binding is described by the binding constant and the binding affinity (Gibbs free energy of binding). Consider the binding of protein A with ligand B as a chemical reaction $A + B \rightarrow AB$ and the unbinding as $AB \rightarrow A + B$. At equilibrium, the binding transition and unbinding transitions occur at the same rate, which is expressed in terms of on-rate, k_{on}, and off-rate constants, k_{off}, as follows,

$$k_{on}[A][B] = k_{off}[AB], \tag{2.1}$$

where $[A]$, $[B]$, and $[AB]$ are the concentrations of the unbound protein, unbound ligand, and bound complex, respectively. The binding constant, K_{AB}, which is simply the equilibrium constant of the above reactions, is then defined as,

$$K_{AB} = \frac{k_{on}}{k_{off}} = \frac{[AB]}{[A][B]}. \tag{2.2}$$

The experimentally observable binding constant can be related to the theoretical Gibbs free energy change upon binding (Equation 2.6) to compare experimental results with theoretical or computational predictions. The Gibbs free energy of binding, or binding affinity, describes the strength of binding, or in other words, the energetic benefit of complexation. If the total energy of the complex is lower than that of the individual constituents, binding is energetically favorable. The greater this energetic difference, i.e., the more negative the binding free energy, the more favorable the binding of the ligand to the protein.

The following section outlines the mathematical description of protein-ligand binding and the formulation of protein-ligand binding energies, the prediction of which lies at the heart of this research.

2.2 The statistical mechanical basis of protein-ligand binding free energies

The binding free energy is a measure of the affinity of the process by which two molecules form a complex by non-covalent association. What follows is a derivation of the key equations that facilitate a mathematical description of protein-ligand binding and form the foundation for understanding the computational approaches used to predict protein-ligand binding energies. First, the chemical potential of a solute molecule in a solvent is derived (2.2.1.1), followed by the chemical potential of a complex of solutes in solvent (2.2.1.2). From this, an expression for the relative binding free energy is obtained (2.2.2). This expression is simplified by the introduction of an implicit solvent model (2.2.3). Finally, the entropic contributions to binding are considered (2.2.4).

2.2.1 Protein ligand binding in equilibrium at standard state

The derivations presented here closely follow those in [12]. The reaction of interest may be defined as,

$$A + B \rightarrow AB(Complex), \tag{2.3}$$

Where A is a protein, B a ligand and AB represents the complex formed upon binding. The equilibrium state of this reaction in terms of the chemical potential, μ is given by,

$$\mu_{sol,A} + \mu_{sol,B} = \mu_{sol,AB}. \tag{2.4}$$

To compare properties calculated for different systems and to simplify the mathematical form, a hypothetical standard state is defined [12]. Each species (A,B,AB) is at the standard concentration of $C^0 = 1$ M in the solvent but does not interact with other molecules of A, B or AB. In this limit of low concentration, the chemical potential $\mu_{sol,i}$ of each species is equivalent to the standard state chemical potential $\mu^0_{sol,i}$. The standard free energy of binding is then given by,

$$\Delta G^0_{AB} = \mu^0_{sol,AB} - \mu^0_{sol,A} - \mu^0_{sol,B}, \tag{2.5}$$

where μ^0 are standard state chemical potentials. This can be linked to experiment via the binding constant K_{AB}, using the expression,

$$\Delta G^0_{AB} = -RT \ln K_{AB}. \tag{2.6}$$

The binding constant K_{AB} is dimensionless because concentrations are given in M and the standard concentration is defined as $C^0 = 1$ M in the hypothetical standard state defined above.

2.2.1.1 Solute molecule in solvent

At standard state, the expression for the chemical potential of molecule A in solution, in terms of the canonical partition functions of molecule A in solvent, $Q_{N,A}(V_{N,A})$, and the solvent without solute, $Q_{N,0}(V_{N,0})$, is given by,

$$\mu^0_{sol,A} = -RT \ln \left[\frac{1}{V_{N,A} C^0} \frac{Q_{N,A}(V_{N,A})}{Q_{N,0}(V_{N,0})} \right] + P^0 \bar{V}_A, \tag{2.7}$$

where $\bar{V}_A = V_{N,A} - V_{N,0}$ is the change in volume due to the addition of one solute molecule to the N molecule solvent. $\mu^0_{sol,A}$ is the standard chemical potential of the solute in the gas phase plus the work (isobaric) needed to transfer the solute into the solvent. The work term is usually negligible.

Consider the general classical Hamiltonian,

$$H(\bar{p}_A, \bar{p}_S, \bar{r}'_A, \bar{r}_S) = \sum_i^{M_A+M_S} \frac{p_i^2}{2m_i} + U(\bar{r}'_A \bar{r}_S), \tag{2.8}$$

where M_A, M_S are the number of atoms in solute A and the solvent, respectively, $\bar{r}'_A$ are the lab frame coordinates of the solute atoms and $\bar{r}_s$ those of the solvent atoms. $\bar{p}_A$ and $\bar{p}_S$ are the momenta of the solute and solvent atoms, respectively. $U(\bar{r}'_A, \bar{r}_B)$ is the potential energy of the system at solute configuration $\bar{r}'_A$ and solvent configuration $\bar{r}_S$.

The ratio of canonical partition functions in equation 2.7 is then given by,

$$\frac{Q_{N,A}}{Q_{N,O}} = \frac{\int d\bar{p}_A d\bar{p}_S \int d\bar{r}'_A d\bar{r}_S \exp\left\{-\beta\left[\sum_{i=1}^{M_A+M_S} \frac{p_i^2}{2m_i} + U(\bar{r}'_A, \bar{r}_S)\right]\right\}}{\sigma_A \int d\bar{p}_S \int d\bar{r}_S \exp\left\{-\beta\left[\sum_{i=M_A+1}^{M_A+M_S} \frac{p_i^2}{2m_i} + U(\bar{r}_S)\right]\right\}}. \tag{2.9}$$

The momentum integrals are over all momentum space, i.e. from $-\infty$ to ∞. The positional integrals are overall atomic configurations in which the molecule A is intact and contained inside the container/simulation cell. σ_A is symmetry factor and equal to 1 for non-symmetric molecules. A constant prefactor, which will cancel in binding energy calculations has been omitted. The prefactor consists of a term h^{-3}, where h is the Planck constant, which makes the partition function dimensionless, and a factorial term for counting the indistinguishable solvent particles. In the summation in the denominator, the sum begins at $i = M_A + 1$ to sum only over solvent molecule momenta.

To simplify this integral, the lab frame coordinates of the solute $\bar{r}'_A$ can be split into internal and external coordinates by defining a molecular reference frame for the internal motion of the molecule. The choice of molecular reference frame is arbitrary and does not effect the final free energies. Three atoms are used to define the molecular frame. Atom one (a1) defines the origin and is at position $(0,0,0)$ in the molecular reference frame. The straight line connecting a1 and atom 2 (a2) defines the direction of the x-axis. Thus, in the molecular reference frame, a2 is always on the x-axis, i.e. $(x,0,0)$. The direction of the y-axis is define as that of the vector connecting a2 and atom three (a3) minus the x-component. The z-axis is defined as the vector cross product of the x- and y-axis. Thus a3 always lies in the plane $z = 0$, i.e. $(x,y,0)$. The position and orientation of all other atoms of the molecule are define with respect to this internal molecular reference frame. The 6 degrees of freedom describing the positions and orientations of a1, a2 ,and a3 with respect to the lab frame are termed the external coordinates of the molecules and denoted by $\vec{\zeta}_A = (R_{a1}, R_{a2}, R_{a3}, \zeta_{a1}, \zeta_{a2}, \zeta_{a3})$. The remaining $3M_A - 6$ degrees of freedom (DOF) are termed the internal coordinates of the molecular frame and denoted $\bar{r}_A$.

Note now, that the potential energy U only depends on the relative position and orientation of solute and solvent molecules and not on the position of the solute with respect to the lab frame, i.e. U is not a function of $\vec{\zeta}_A$. Hence the integrals in 2.9 over the external coordinates can be evaluated directly beginning with the three positional external DOF,

$$\int_{V_{NA}} d\bar{R}_A = V_{NA}, \tag{2.10}$$

where V_{NA} is the system volume at equilibrium at standard pressure of 1 atmosphere. Similarly, using spherical polar coordinates, the three orientational DOF can be integrated out,

$$\int_{\theta=0}^{\pi} \sin\theta d\theta \int_{\psi=0}^{2\pi} d\psi \int_{\phi=0}^{2\pi} d\phi = 4\pi^2[-\cos\theta]_0^{\pi} = 8\pi^2. \tag{2.11}$$

Thus, splitting the positional coordinates into internal and external coordinates, the external coordinates can be integrated out of the partition functions leaving a factor of $8\pi^2 V_{N,A}$.

Under approximation of classical mechanics, the momentum terms can likewise be integrated out yielding a factor of $(2\pi m_i RT)^{\frac{3}{2}}$ for each atom, as shown below.

$$I = \int_{-\infty}^{\infty} d\bar{p}_A \exp\left\{-\beta\left[\sum_{i=1}^{M_A+M_S} \frac{p_i^2}{2m_i}\right]\right\} \tag{2.12}$$

$$= \int_{-\infty}^{\infty} d\bar{p}_A \exp\left\{-\beta\left[\frac{p_1^2}{2m_1}\right]\right\} \exp\left\{-\beta\left[\frac{p_2^2}{2m_2}\right]\right\}... \tag{2.13}$$

$$= \prod_{i=1}^{M_A+M_S} \int_{-\infty}^{\infty} d\bar{p}_A \exp\left\{-\beta\left[\frac{p_i^2}{2m_i}\right]\right\}, \tag{2.14}$$

applying the standard integral,

$$\int_{-\infty}^{\infty} dx \exp\{-ax^2\} = \sqrt{\frac{\pi}{a}}, \tag{2.15}$$

$$I = \prod_{i=1}^{M_A+M_S} (2\pi m_i RT)^{\frac{3}{2}} \tag{2.16}$$

In the ratio of canonical partition functions, the momentum contributions from solvent atoms cancel, as they are present in both $Q_{N,A}$ and $Q_{N,0}$. The final expression for the standard chemical potential of molecule A in solvent, in terms of the internal Cartesian coordinates of the solute and the solvent coordinates, $\bar{r}_A, \bar{r}_S$, is,

$$\mu^0_{sol,A} = -RT\ln\left[\frac{8\pi^2}{C^0\sigma_A}\prod_i^{M_A}(2\pi m_i RT)^{\frac{3}{2}}\frac{Z_{N,A}}{Z_{N,0}}\right] + P^0\bar{V}_A, \tag{2.17}$$

where σ_A is the symmetry number of A, and Z are the configurational integrals,

$$Z_{N,A} = \int \exp\{-\beta U(\bar{r}_A, \bar{r}_S)\} d\bar{r}_A d\bar{r}_S, \tag{2.18}$$

$$Z_{N,0} = \int \exp\{-\beta U(\bar{r}_S)\} d\bar{r}_S. \tag{2.19}$$

2.2.1.2 Complex of solutes in solvent

The expression in 2.17 for the standard chemical potential of molecule A in solvent cannot be immediately extended to the standard chemical potential of the complex. First, the internal and external coordinates of the complex must be defined. By convention, define the external coordinates of molecule A, the host, as the external coordinates of the complex. Thus, $\vec{\zeta}_A$ describes the position and orientation of the complex with respect to the lab frame. Upon binding, the external coordinates of molecule B , $\vec{\zeta}_B$, become internal coordinates of the complex. They describe the relative position and orientation of molecule B, usually a ligand, with respect to the host molecule A. Additionally, define a step function $I(\vec{\zeta}_b)$ such that $I(\vec{\zeta}_B) = 1$ for complexed configurations and $I(\vec{\zeta}_B) = 0$ for non-complexed configurations. The standard chemical potential of the complex is then,

$$\mu^0_{sol,AB} = -RT \ln \left[\frac{8\pi^2}{C^0 \sigma_{AB}} \prod_i^{M_A+M_B} (2\pi m_i RT)^{\frac{3}{2}} \frac{Z_{N,AB}}{Z_{N,0}} \right] + P^0 \bar{V}_{AB}, \tag{2.20}$$

where,

$$Z_{N,AB} = \int I(\zeta_B) J_{\zeta_B} \exp\{-\beta U(\bar{r}_A, \bar{r}_B, \bar{r}_S, \zeta_B)\} d\bar{r}_A d\bar{r}_S d\bar{r}_B d\zeta_B, \tag{2.21}$$

and $J(\zeta_B)$ is the Jacobian determinant of the Eulerian rotation of molecule B relative to A. σ_{AB} is the symmetry number of the complex solute. It can be shown that as long as I satisfies two conditions, the chemical potential is independent of the choice of I. The first condition is that the region in which $I(\zeta_B) = 1$ should include all configurations that significantly contribute to the chemical potential of the complex. The second condition states that the region should not include so large a volume of uncomplexed configurations that these contribute to the chemical potential.

By substitution of equations 2.21 and 2.18 into equation 2.5, the free energy of binding at standard state of a host-ligand system is given by,

$$\Delta G^0_{AB} = -RT \ln \left[\frac{C^0}{8\pi^2} \frac{\sigma_A \sigma_B}{\sigma_{AB}} \frac{Z_{N,AB} Z_{N,0}}{Z_{N,A} Z_{N,B}} \right] + P^0 \Delta \bar{V}_{AB}, \tag{2.22}$$

where $\Delta \bar{V}_{AB} = \bar{V}_{AB} - \bar{V}_A - \bar{V}_B$. All mass dependent terms have cancelled in this expression. The final pressure term can be ignored at low pressures.

2.2.2 Relative binding free energies

It is often sufficient to calculate relative free energy differences. It follows from equation 2.22 that the difference in binding free energies at standard state between two different ligands B and C in complex with protein A is given by,

$$\Delta G^0_{AC} - \Delta G^0_{AB} = -RT\ln\left(\frac{\sigma_A B}{\sigma_{AC}}\frac{Z_{N,AC}}{Z_{N,AB}}\right) + P^0(\bar{V}_{AC} - \bar{V}_{AB}) \tag{2.23}$$

$$- \left[-RT\ln\left(\frac{\sigma_B}{\sigma_A}\frac{Z_{N,C}}{Z_{N,B}}\right) + P^0(\bar{V}_C - \bar{V}_B)\right], \tag{2.24}$$

where the first term gives the work of alchemical transformation of ligand B to ligand C inside protein receptor A. The second term gives the work of the same transformation in solution. This expression forms the basis for the rigorous but expensive Thermodynamic Integration (TI) and Free Energy Perturbation (FEP).

2.2.3 Implicit treatment of solvent

The configurational integrals introduced above each involve both the solute and the solvent coordinates. Consider the interaction of the solute and solvent for a given configuration of the system,

$$\Delta U(\bar{r}_A, \bar{r}_S) = U(\bar{r}_A, \bar{r}_S) - U(\bar{r}_A) - U(\bar{r}_S). \tag{2.25}$$

Following the same derivation as previously, the standard state chemical potential of solute A in solution can be re-written in terms of a modified configurational integral, Z_A, that explicitly separates the contributions from the solute and solvent as follows,

$$Z_A = \frac{Z_{N,A}}{Z_{N,0}} = \int \exp\{-\beta[U(\bar{r}_A) + W(\bar{r}_A)]\}d\bar{r}_A, \tag{2.26}$$

where $W(\bar{r}_A)$ is the work of transferring the solute A in configuration $\bar{r}_A$ from the gas phase to the solvent. This term implicitly depends on temperature and pressure and encapsulates the most important effects of the solvent on the chemical potential of the solute in solution. The function W is given by,

$$W(\bar{r}_A) = -RT\ln\left(\frac{\int \exp\{-\beta\Delta U(\bar{r}_A, \bar{r}_S)\}\exp\{-\beta U(\bar{r}_S)\}d\bar{r}_S}{\int \exp\{-\beta U(\bar{r}_S)\}d\bar{r}_S}\right). \tag{2.27}$$

To show this, start from equations 2.18 and 2.19,

$$Z_A = \frac{Z_{N,A}}{Z_{N,0}} = \frac{\int \exp\{-\beta U(\bar{r}_A, \bar{r}_S)\} d\bar{r}_A d\bar{r}_S}{\int \exp\{-\beta U(\bar{r}_S)\} d\bar{r}_S} \tag{2.28}$$

$$= \frac{\int \exp\{-\beta U(\bar{r}_A)\} \exp\{-\beta \left[U(\bar{r}_A, \bar{r}_S) - U(\bar{r}_A) + U(\bar{r}_S) - U(\bar{r}_S)\right]\} d\bar{r}_A d\bar{r}_S}{\int \exp\{-\beta U(\bar{r}_S)\} d\bar{r}_S} \tag{2.29}$$

$$= \frac{\int \exp\{-\beta U(\bar{r}_A)\} \int \exp\{-\beta \Delta U(\bar{r}_A, \bar{r}_S)\} \exp\{-\beta U(\bar{r}_S)\} d\bar{r}_A d\bar{r}_S}{\int \exp\{-\beta U(\bar{r}_S)\} d\bar{r}_S} \tag{2.30}$$

$$= \int \exp\{-\beta U(\bar{r}_A)\} \exp\left\{ -RT \ln \left[\frac{\int \exp\{-\beta \Delta U(\bar{r}_A, \bar{r}_S)\} \exp\{-\beta U(\bar{r}_S)\} d\bar{r}_S}{\int \exp\{-\beta U(\bar{r}_S)\} d\bar{r}_S} \right] \right\}^{-\beta} d\bar{r}_A \tag{2.31}$$

$$= \int \exp\{-\beta U(\bar{r}_A)\} \exp\{-\beta W(\bar{r}_A)\} d\bar{r}_A \tag{2.32}$$

$$= \int \exp\{-\beta \left[U(\bar{r}_A) + W(\bar{r}_A)\right]\} d\bar{r}_A, \tag{2.33}$$

and thus arrived at equation 2.26, the configurational integral of the solute A in solvent with the solvent contribution W separated from the gas phase potential U. Similar expressions to the above can be derived for Z_B and the complex, Z_{AB}.

$W(\bar{r}_A)$ is usually approximated by an implicit solvent model. These models do not treat the solvent coordinates $\bar{r}_S$ explicitly but instead assume the solvent to be continuous and often homogeneous. As a result, all integrals over the solvent coordinates vanish from the binding free energy (equation 2.22), leaving only integrals over the coordinates of the protein, ligand and complex. The above treatment can be extended to include solvent pH in the solvation model.

2.2.4 Entropy changes in binding

Recall the expression for the standard state binding free energy of a host guest system,

$$\Delta G^0_{AB} = -RT \ln \left[\frac{C^0}{8\pi^2} \frac{\sigma_A \sigma_B}{\sigma_{AB}} \frac{Z_{N,AB} Z_{N,0}}{Z_{N,A} Z_{N,B}} \right] + P^0 \Delta \bar{V}_{AB}. \tag{2.34}$$

This expression contains both the enthalpic and entropic contributions to the binding free energy. The change in entropy upon binding can be investigated using the standard relation,

$$\Delta S^0 = -\left(\frac{\partial \Delta G^0}{\partial T} \right)_P, \tag{2.35}$$

which is the partial derivative of the change in free energy with respect to temperature at constant pressure. Recall, that the external DOF of the ligand B become internal degrees of the complex upon binding. Complexation leads to the restriction of these DOF and hence to a change in entropy. This change in entropy due to only the external coordinates of ligand B is investigated first and termed the external entropy.

Recall also, that due to the cancellation of all momentum terms in the binding free energy, binding does not restrain the momentum space but only the spatial degrees of freedom.

To determine the external entropy change upon binding, the binding free energy ΔG^0_{AB} must be re-written in terms of only the exernal DOF of the ligand B. Again, this is achieved via the definition of a potential of mean force by first defining,

$$\Delta U(\bar{r}_A, \bar{r}_B, \zeta_B) \equiv U(\bar{r}_A, \bar{r}_B, \zeta_B) - U(\bar{r}_A) - U(\bar{r}_B), \tag{2.36}$$

$$\Delta W(\bar{r}_A, \bar{r}_B, \zeta_B) \equiv W(\bar{r}_A, \bar{r}_B, \zeta_B) - W(\bar{r}_A) - W(\bar{r}_B), \tag{2.37}$$

and then the potential of mean force,

$$w(\zeta_B) \equiv -RT \times \ln \tag{2.38}$$

$$\left[\frac{\int \exp\{-\beta\,[\Delta U(\bar{r}_A, \bar{r}_B, \zeta_B) + \Delta W(\bar{r}_A, \bar{r}_B, \zeta_B)]\}\exp\{-\beta\,[U(\bar{r}_A) + U(\bar{r}_B) + W(\bar{r}_A) + W(\bar{r}_B)]\}d\bar{r}_A d\bar{r}_B}{\exp\{-\beta\,[U(\bar{r}_A) + U(\bar{r}_B) + W(\bar{r}_A) + W(\bar{r}_B)]\}d\bar{r}_A d\bar{r}_B}\right] \tag{2.39}$$

$$= -RT \ln\langle\exp\{-\beta\,[\Delta U(\bar{r}_A, \bar{r}_B, \zeta_B) + \Delta W(\bar{r}_A, \bar{r}_B, \zeta_B)]\}\rangle_{\bar{r}_A, \bar{r}_B}, \tag{2.40}$$

where $\langle\rangle_{\bar{r}_A, \bar{r}_B}$ is an ensemble average over the configurations $\bar{r}_A$ and $\bar{r}_B$ when the two solutes are far apart, i.e. un-complexed. Note that both $\Delta U(\bar{r}_A, \bar{r}_B, \zeta_B)$ and $\Delta W(\bar{r}_A, \bar{r}_B, \zeta_B) \to 0$ when solutes are widely separated since the inter-molecular energies will be zero and $U(\bar{r}_A, \bar{r}_B, \zeta_B) = U(\bar{r}_A) + U(\bar{r}_B)$ for $\zeta \to \infty$. Hence, $w(\zeta_B) \to 0$ as $\zeta \to \infty$.

The potential of mean force is used to re-write the binding free energy as an integral over only the external DOF of ligand B,

$$\Delta G^0_{AB} = -RT \ln\left[\frac{C^0}{8\pi^2}\frac{\sigma_A \sigma_B}{\sigma_{AB}} \int I(\zeta_B) J(\zeta_B) \exp\{-\beta w(\zeta_B)\} d\zeta_B\right] + P^0 \Delta \bar{V}_{AB} \tag{2.41}$$

$$= -RT \ln X + P^0 \Delta \bar{V}_{AB}, \tag{2.42}$$

where $\exp\{-\beta w(\zeta_B)\}$ has replaced $\frac{Z_{AB}}{Z_A Z_B}$ and the other symbols are as defined previously.

To find the external entropy change apply equation 2.35 to equation 2.42,

$$\Delta S^0 = -\left(\frac{\partial \Delta G^0(\zeta_B)}{\partial T}\right)_P = -R \ln X - RT \frac{\partial \ln X}{\partial T}_P, \tag{2.43}$$

where,

$$RT\frac{\partial \ln X}{\partial T}_P = RT\frac{\partial}{\partial T_P}\int I(\zeta_B)J(\zeta_B)\exp\left\{-\frac{1}{RT}w(\zeta_B)\right\}d\zeta_B \tag{2.44}$$

$$= RT\int I(\zeta_B)J(\zeta_B)\frac{\exp\left\{-\frac{w(\zeta_B)}{RT}\right\}[w(\zeta_B) - Tw'(\zeta_B)]}{RT^2} \tag{2.45}$$

$$= -\frac{1}{T}\left[\int\frac{I(\zeta_B)J(\zeta_B)w(\zeta_B)d\zeta_B}{\exp\left\{\frac{w(\zeta_B)}{RT}\right\}} - T\int\frac{I(\zeta_B)J(\zeta_B)\frac{\partial w(\zeta_B)}{\partial T}d\zeta_B}{\exp\left\{\frac{w(\zeta_B)}{RT}\right\}}\right] \tag{2.46}$$

$$= -\frac{1}{T}\langle w(\zeta_B)\rangle_{AB} + \langle\frac{\partial w(\zeta_B)}{\partial T}\rangle_{AB}. \tag{2.47}$$

Hence,

$$\Delta S^0_{AB} = -\frac{1}{T}(\Delta G^0_{AB} - P^0\Delta\bar{V}_{AB}) + \frac{1}{T}\langle w(\zeta_B)\rangle_{AB} - \langle\frac{\partial w(\zeta_B)}{\partial T}\rangle_{AB} + \Delta S^0_{PdV}, \tag{2.48}$$

where $S^0_{PdV} \approx 0$ unless the reaction occurs at constant volume. The external entropy is the change in entropy under the potential of mean force $w(\zeta_B)$ if only the external DOF of ligand B, i.e. ζ_B, were considered. If there are no other DOF, $\frac{\partial w(\zeta_B)}{\partial T} = 0$ and $S^0_{PdV} = 0$ hence,

$$\Delta S^0_{ext} \equiv -\frac{1}{T}(\Delta G^0_{AB} - P^0\Delta\bar{V}_{AB}) + \frac{1}{T}\langle w(\zeta_B)\rangle_{AB}. \tag{2.49}$$

If we assume that at standard pressure $P^0\Delta\bar{V}_{AB}$ is negligible, then the binding free energy can be written as,

$$\Delta G^0_{AB} = \langle w(\zeta_B)\rangle_{AB} - T\Delta S^0_{ext}, \tag{2.50}$$

where $\langle w(\zeta_B)\rangle_{AB}$ is the expectation value over all complexed configurations AB of the potential of mean force, i.e. the ensemble average of the gas phase and solvation interaction energies where the ensemble average is taken over un-complexed configurations of A and B.

Returning to equation 2.48 for the total change in entropy upon binding, the change in internal entropy is defined next. The internal entropy corresponds to the change in conformational freedom of the solutes upon binding. According to Gilson [12], it is the entropy over and above the external entropy that is found by treating the internal degrees of freedom explicitly. It is however still assumed, that the solvation term W has no entropic component. Because the internal entropy is the entropy over and above the external entropy, it must be equal to $\langle\frac{\partial w(\zeta_B)}{\partial T}\rangle_{AB}$, the only term not included in the external entropy.

First recall,

$$w(\zeta_B) = -RT\ln\langle\exp\left\{-\beta\left[\Delta U(\bar{r}_A,\bar{r}_B,\bar{\zeta}_B)+\Delta W(\bar{r}_A,\bar{r}_B,\bar{\zeta}_B)\right]\right\}\rangle_{r_a,r_b,apart} \tag{2.51}$$

$$= -RT\ln\left[\frac{\int\exp\{-\beta\,[\Delta U(\zeta_B)+\Delta W(\zeta_B)]]\}\exp\{-\beta\,[U(\bar{r}_A)+W(\bar{r}_A)+U(\bar{r}_B)+W(\bar{r}_B]\}d\bar{r}_A d\bar{r}_B}{\int\exp\{-\beta\,[U(\bar{r}_A)+W(\bar{r}_A)+U(\bar{r}_B)+W(\bar{r}_B]\}d\bar{r}_A d\bar{r}_B}\right] \tag{2.52}$$

$$= -RT\ln X = -RT\ln\frac{A}{B}, \tag{2.53}$$

then,

$$\frac{\partial w}{\partial T} = -R\ln X - RT\frac{X'}{X} = -\frac{R}{T}w(\zeta_B) - RT\frac{X'}{X}, \tag{2.54}$$

where,

$$X' = \frac{\partial}{\partial T}\left(\frac{A}{B}\right) = \frac{A'}{B} - \frac{AB'}{B^2}. \tag{2.55}$$

Consider now each term in the above equation separately. Both terms are divided by X in 2.54. The first term is,

$$\frac{\frac{A'}{B}}{\frac{A}{B}} = \frac{A'}{A} \tag{2.56}$$

$$= \frac{1}{T^2}\frac{\int\left[U(\bar{r}_A,\bar{r}_B,\bar{\zeta}_B)+W(\bar{r}_A,\bar{r}_B,\bar{\zeta}_B)\right]\exp\left\{-\beta\left[U(\bar{r}_A,\bar{r}_B,\bar{\zeta}_B)+W(\bar{r}_A,\bar{r}_B,\bar{\zeta}_B)\right]\right\}d\bar{r}_A d\bar{r}_B}{\int\exp\left\{-\beta\left[U(\bar{r}_A,\bar{r}_B,\bar{\zeta}_B)+W(\bar{r}_A,\bar{r}_B,\bar{\zeta}_B)\right]\right\}d\bar{r}_A d\bar{r}_B} \tag{2.57}$$

$$+ \frac{1}{T}\frac{\int\left[U'(\bar{r}_A,\bar{r}_B,\bar{\zeta}_B)+W'(\bar{r}_A,\bar{r}_B,\bar{\zeta}_B)\right]\exp\left\{-\beta\left[U(\bar{r}_A,\bar{r}_B,\bar{\zeta}_B)+W(\bar{r}_A,\bar{r}_B,\bar{\zeta}_B)\right]\right\}d\bar{r}_A d\bar{r}_B}{\int\exp\left\{-\beta\left[U(\bar{r}_A,\bar{r}_B,\bar{\zeta}_B)+W(\bar{r}_A,\bar{r}_B,\bar{\zeta}_B)\right]\right\}d\bar{r}_A d\bar{r}_B} \tag{2.58}$$

$$= -\frac{1}{T^2}\langle U(\bar{r}_A,\bar{r}_B,\bar{\zeta}_B)+W(\bar{r}_A,\bar{r}_B,\bar{\zeta}_B)\rangle_{\bar{r}_A,\bar{r}_B,comp} - \frac{1}{T}\langle U'(\bar{r}_A,\bar{r}_B,\bar{\zeta}_B)+W'(\bar{r}_A,\bar{r}_B,\bar{\zeta}_B)\rangle_{\bar{r}_A,\bar{r}_B,comp}, \tag{2.59}$$

where the expectation value is over the ensemble of the complex AB. The average is over the complex coordinates only as $U(\bar{r}_A,\bar{r}_B,\bar{\zeta}_B)+W(\bar{r}_A,\bar{r}_B,\bar{\zeta}_B)=0$ when the solutes are far apart.

The second term is,

$$\frac{\frac{AB'}{B^2}}{\frac{A}{B}} = \frac{B'}{B} \tag{2.60}$$

$$= \frac{1}{T^2} \frac{\int [U(\bar{r}_A) + W(\bar{r}_A) + U(\bar{r}_B) + W(\bar{r}_B))] \exp\{\beta [U(\bar{r}_A) + W(\bar{r}_A) + U(\bar{r}_B) + W(\bar{r}_B)]\} d\bar{r}_A d\bar{r}_B}{\int \exp\{\beta [U(\bar{r}_A) + W(\bar{r}_A) + U(\bar{r}_B) + W(\bar{r}_B)]\} d\bar{r}_A d\bar{r}_B} \tag{2.61}$$

$$- \frac{1}{T} \frac{\int [U'(\bar{r}_A) + W'(\bar{r}_A) + U'(\bar{r}_B) + W'(\bar{r}_B)] \exp\{\beta [U(\bar{r}_A) + W(\bar{r}_A) + U(\bar{r}_B) + W(\bar{r}_B)]\} d\bar{r}_A d\bar{r}_B}{\int \exp\{\beta [U(\bar{r}_A) + W(\bar{r}_A) + U(\bar{r}_B) + W(\bar{r}_B)]\} d\bar{r}_A d\bar{r}_B} \tag{2.62}$$

$$= \frac{1}{T^2} \left[\langle U(\bar{r}_A) + W(\bar{r}_A) \rangle_{\bar{r}_A} + \langle U(\bar{r}_B) + W(\bar{r}_B) \rangle_{\bar{r}_B} - T \langle U'(\bar{r}_A) + W'(\bar{r}_A) \rangle_{\bar{r}_A} + \langle U'(\bar{r}_B) + W'(\bar{r}_B) \rangle_{\bar{r}_B} \right]. \tag{2.63}$$

Next, the temperature derivatives of W are assumed to be zero. This corresponds to neglecting the change in entropy of the solvent. Recall also, that the internal entropy is given by $\langle \frac{\partial w(\zeta_B)}{\partial T} \rangle_{AB}$. The two terms calculated above are expectation values, i.e. just numbers, and are not effected by $\langle \rangle_{AB}$. Only the term $-\frac{R}{T} w(\zeta_B)$ is effected. Thus, the internal entropy of binding is given by,

$$\Delta S^{int}_{AB} = -\frac{1}{T} \langle w(\zeta_B) \rangle_{AB} \tag{2.64}$$

$$+ \frac{1}{T} \left[\langle U(\bar{r}_A, \bar{r}_B, \bar{\zeta}_B) + W(\bar{r}_A, \bar{r}_B, \bar{\zeta}_B) \rangle_{AB} - \langle U(\bar{r}_A) + W(\bar{r}_A) \rangle_{\bar{r}_A} - \langle U(\bar{r}_B) + W(\bar{r}_B) \rangle_{\bar{r}_B} \right]. \tag{2.65}$$

By combining the internal and external entropy, the binding free energy from equation 2.50 can be re-written as,

$$\Delta G^0_{AB} = \langle w(\zeta_B) \rangle_{AB} - T \Delta S^{ext}_A B \tag{2.66}$$

$$= \langle U(\bar{r}_A, \bar{r}_B, \bar{\zeta}_B) + W(\bar{r}_A, \bar{r}_B, \bar{\zeta}_B) \rangle_{AB} - \langle U(\bar{r}_A) + W(\bar{r}_A) \rangle_{\bar{r}_A} \tag{2.67}$$

$$- \langle U(\bar{r}_B) + W(\bar{r}_B) \rangle_{\bar{r}_B} - T \Delta S^{int}_{AB} - T \Delta S^{ext}_{AB} - \left[P^0 \Delta \bar{V}_{AB} + T \Delta S^0_{PdV} + T \Delta S^0_{solv} \right].$$

The first term is the gas phase plus solvation energy of the complex averaged over the complex ensemble. Terms two and three are the gas phase plus solvation energies of the protein A and ligand B over their respective ensembles. Term 4 is the change in internal entropy, i.e. the change in conformation entropy of the protein and ligand upon binding. Term 5 is the external entropy which gives the change in entropy due to the confinement of the ligand external coordinates inside the binding site. The final three terms in the square brackets are usually assumed to be zero. To some extent, the change in entropy of the solvent, the final term, is included via the parameterization of the implicit solvent model. Written like this, the U and W terms capture all the enthalpic contribution to binding.

The above expression for the binding free energy of a protein and ligand forming a non-covalent complex is the basis for many approximate approaches to calculating binding free energies. The strong similarity of the terms above to those in the popular MM-PBSA method will be highlighted in later sections and the various approximation made outlined in detail.

Chapter 3

Methods

3.1 Classical mechanical protein-ligand free energies of binding

Because the binding of small molecules to proteins and other large biomolecules is of central importance in biology and medicine, much effort has been dedicated to computationally estimating the binding affinities associated with these processes. A zoo of methods has been proposed since the 60s. First, a brief overview of the different classical mechanical approaches is given, followed by sections introducing the methods used in this body of research in greater detail.

3.1.1 An overview of methods

Arguably the most utilized and essential computational method in the context of protein-ligand binding is molecular dynamics (MD) [13], an atomistic simulation of the dynamics of interacting atoms. In molecular dynamics, each atom is described by a set of properties like its mass and charge and a set of parameters that determine its interatomic interaction with other atoms. So-called molecular mechanics force fields are used to compute all pairwise interactions between the atoms in the simulated system, which generally consists of solute molecules in a solvent. From the interaction energies, forces are derived, and the dynamical motion of the system is simulated at an atomistic level of detail. MD forms the foundation for most binding free energy protocols as they rely on the ability to sample multiple different physical configurations of the protein-ligand complex, its unbound constituents, or the binding pathway of the ligand.

Generally speaking, the different approaches to estimating protein-ligand binding may be divided into two categories. The first group is more theoretically rigorous and usually derived directly from the underlying statistical mechanical theory but is expensive and often challenging to implement. The two most notable methods in this category are free energy perturbation (FEP) and thermodynamic integration (TI). Both methods rely on chemically transforming molecules through fictitious, yet theoretically rigorous thermodynamic processes [14].

The second group contains more approximate, often empirically motivated methods that follow a ″quick and dirty″ philosophy, aiming to provide acceptable accuracy at relatively low cost/complexity. A hallmark of these methods is severe assumptions that allow the splitting of the energy function into separable, individually tractable energy components. This group's most widely used method is the molecular mechanics Poisson-Boltzmann surface area method (MM-PBSA) [15].

3.1.2 Molecular dynamics

The basic idea of molecular dynamics is that given the knowledge of the type and position of a group of atoms, the forces exerted on each atom by all other atoms can be calculated. Using Newton's classical mechanical equations of motion, the position of each atom can then be propagated forward in time. This step is repeated iteratively for tiny time steps, generating a ″movie″ of the movement of each atom in the system over time.

The core applications of MD were classified into three groups by Karplus et al. [16]. The first is to sample configuration space, for example, when refining experimental crystallographic structures. The second aims to calculate the thermodynamic properties of the system at equilibrium, like atomic mean-square fluctuation amplitudes. Examining the dynamics of a process over time is the final application of MD. All of the above are utilized in the context of binding free energy estimation.

The most basic discretized numerical algorithm for the forward propagation of the particle positions is the Verlet algorithm. Consider an ensemble of N particles with mass m_i and position vectors $\bar{x}_i(t)$ at time t. From Newton's equations of motion, the acceleration of the particle $m_i\ddot{\bar{x}} = \bar{f}_i(\bar{x}(t)) = -\nabla_{x,i}V(\bar{x}(t))$ where $\bar{f}_i(\bar{x}(t))$ is the force and and $V(\bar{x}(t))$ a scalar potential function called force field in the context of molecular dynamics. Using the central difference approximation for second order derivatives, we write,

$$\frac{\mathrm{d}^2\bar{x}_n}{\mathrm{d}t^2} \approx \frac{\frac{\bar{x}_{n+1}-\bar{x}_n}{\delta t} - \frac{\bar{x}_n-\bar{x}_{n-1}}{\delta t}}{\delta t} \tag{3.1}$$

$$= \frac{\bar{x}_{n+1} - 2\bar{x}_n + \bar{x}_{n-1}}{\delta t^2} = \ddot{\bar{x}}(t), \tag{3.2}$$

where the time step $\delta t > 0$ and $t_n = n\delta t$. Re-arranging the above equation in terms of $\bar{x}_{n+1}$ and substituting for $\ddot{\bar{x}}(t)$ from Newton's equations of motions yields the expression,

$$\bar{x}_{n+1} = 2\bar{x}_n - \bar{x}_{n-1} + \frac{\bar{f}(t)}{m}\delta t^2. \tag{3.3}$$

Re-writing this in terms of δt and for each particle i results in the Verlet expression for the next position,

$$\bar{x}_i(t + \delta t) = 2\bar{x}_i(t) - \bar{x}_i(t - \delta t) + \frac{\bar{f}_i(t)}{m_i}\delta t^2. \tag{3.4}$$

This discretization error of this integration scheme has order δt^4 because the first and third order terms cancel due to the central difference approach. The initial conditions at $t_0 = 0$ are known ($\bar{x}_0$ and $\bar{v}_0 = \dot{\bar{x}}_0$) but for the calculation of $\bar{x}_2$ the position at $\bar{x}_1$ must already be known. Using a Taylor expansion of Newton's equations of motion $\bar{x}_1$ can be calculated from the known initial position and velocity using,

$$\bar{x}_1 \approx \bar{x}_0 + \bar{v}_0\delta t + \frac{1}{2}\dot{\bar{v}}_0\delta t^2. \tag{3.5}$$

In the context of molecular dynamics simulations, knowledge of the velocity of each particle in the system is important to compute thermodynamic properties like kinetic energy and for the temperature control of the simulation. In the basic Verlet scheme, the velocity is not explicitly given. There are two approaches to resolving this limitation. First, the velocity can be estimated at time t from the position terms using the mean value theorem,

$$\bar{v}(t) \approx \frac{\bar{x}(t + \delta t) - \bar{x}(t - \delta t)}{2\delta t}, \tag{3.6}$$

with order δt^2 error. The velocity step is always one step behind the position. The larger error in the approximation of the velocity, δt^2, as compared to the position, δt^4, is avoided by the second approach to explicitly incorporating the velocity. The Velocity-Verlet algorithm is a modification of the simple Verlet algorithm in which the current position and current velocity are used to compute the next step rather than the current and previous position in traditional Verlet. Thus the position and velocity are known simultaneously at no increased memory cost. The equations for the Velocity-Verlet, which has the same order accuracy as the simple Verlet, are,

$$\bar{x}(t + \delta t) = \bar{x}(t) + \bar{v}(t)\delta t + \frac{1}{2}\ddot{\bar{x}}(t)\delta t^2 \tag{3.7}$$

$$\bar{v}(t + \delta t) = \bar{x}(t) + \frac{\ddot{\bar{x}}(t) + \ddot{\bar{x}}(t + \delta t)}{2}\delta t. \tag{3.8}$$

A basic algorithmic implementation of the Velocity-Verlet, written in pseudo-code, for a single particle would go as follows:

```
pos = (0.0,0.0,0.0)
vel = (1.0,2.0,3.0)
acc = (0.0,0.0,0.0)
```

```
mass = 1.0
dt = 0.001
for i in range(start,end):
    new_pos = pos + vel * dt + acc*dt*dt*0.5
    new_acc = apply_force_field(new_pos,mass)
    new_vel = vel + (acc+new_acc)*dt*0.5

    pos = new_pos
    vel = new_vel
    acc = new_acc
```

where apply_force_field() is a function call of the molecular dynamics force field, which calculates the force, or acceleration, on each particle based on its current position. The following section introduces the concept of force fields in more detail.

3.1.2.1 Molecular mechanics force fields

The interaction energies and, subsequently, forces between the atoms are calculated using molecular mechanics force fields. These consist of energy terms summed, pairwise, over each atom in the system. A basic force field commonly used in the context of biomolecular systems will include terms for both bonded (covalent) interactions and non-bonded (non-covalent) interactions [17] as shown in Equations 3.9, 3.10 and 3.11.

$$E_{total} = E_{bonded} + E_{nonbonded} \tag{3.9}$$

$$E_{total} = E_{bonds} + E_{angles} + E_{dihedrals} + E_{electrostatic} + E_{VdW} \tag{3.10}$$

$$\begin{aligned} E_{total} = & \sum_{bonds} K_b(b - b_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} \sum_{n=1}^{N} K_{\phi,n}(1 + \cos(n\phi - \delta_n)) \\ & + \sum_{ij} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} + \sum_{ij} \epsilon_{ij} \left[\left(\frac{\sigma_{min,ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma_{min,ij}}{r_{ij}} \right)^{6} \right] \end{aligned} \tag{3.11}$$

The bonded interaction is split into bond stretching, angle bending, and dihedral torsion as shown in Equation 3.11. The stretching and bending terms are formulated as harmonic oscillators and have two parameters each, a force constant (K_b, K_θ) and the equilibrium value (b_0, θ_0). The torsion term is more complicated and has more parameters, and the implementations vary significantly between force fields [18].

The non-bonded term in Equation 3.11 usually comprises an electrostatic and van der Waals term. The electrostatic term is computed additively with Coulomb's law from the fixed atom-centered charges q_i and q_j, the distance between the atom pair, and the electric constant ϵ_0. The van der Waals term uses Lennard-Jones-type potentials characterized by a well-depth ϵ_{ij} and $\sigma_{min,ij}$, the distance at which the potential energy is zero. The typical choice of exponents 6 and 12 is borne of computational convenience, as the 12th power can be obtained cheaply by squaring the sixth power. However, close to the atom centers, the short-range repulsion rises exponentially, and these terms do not describe this behavior well.

Empirical parameters introduced above are fit against high-accuracy quantum mechanical simulations or experimental data. This process is referred to as parameterization. Different force fields may include additional cross terms, like approximate polarization, or are parameterized for a specific class of systems like proteins, small molecules, or solvents. Modern force fields with well-defined domains of applicability go far beyond the simple formulation described here [18–21]. The limitations of classical mechanical force fields are discussed in detail in Section 3.2.1.

What follows is a brief introduction to the molecular mechanics force field utilized in this body of research.

ff14SB and ff19SB This family of Amber force fields for proteins has been developed since the early 90s. The popular ff94 force field was constructed from the terms in Equations 3.9, 3.10 and 3.11 and described by its authors as ”minimalist” in its functional form [22]. Amber ff99SB improved secondary protein structure balance and dynamics. ff14SB re-fit all amino-acid side-chain dihedral parameters from an improved training set [23]. Additionally, parameters for protonation states of ionizable side chains were implemented. ff19SB [18] is the most recent force field of the SB family of protein force fields and improves amino-acid-depending properties and backbone dihedral parameters for every amino acid. A backbone side-chain coupling correction is also implemented for the first time. In the development of ff19SB, emphasis was placed on the interchangeability of water models, but usage is recommended with modern OPC and OPC3 water force field models.

ff15ipq ff15ipq [24] is a continuation of the ff14ipq force field and is built on a self-consistent physical model. Its flagship feature is that it can derive implicitly polarized atomic charges in the presence of an explicit solvent. Unlike other force field fitting procedures, the parameterization of ff15ipq explicitly considers the influence of the water model on the solute’s charge distribution. Additionally, two sets of force field parameters are fit, one in the presence of solvent and the other in the gas phase.

GAFF1 and GAFF2 The general Amber force field, or GAFF, is a force field for small organic molecules developed by Wang et al. [25]. It is compatible with all Amber force fields for proteins and features parameters for most pharmaceutical molecules. Its functional form is equivalent to Equations 3.9, 3.10 and 3.11. GAFF can uses molecule-specific partial charges generated by semi-empirical quantum mechanical calculations. GAFF2 is the second iteration of the GAFF force field.

Water force fields Various water-specific models representing the bulk solvent water molecules in molecular dynamics simulations exist. The simplest water models are rigid and describe only non-bonded interactions. Their potentials contain an electrostatic (Coulombic) and dispersion-repulsion term (Lennard-Jones potential). In 3-point water models, each site corresponds to an atom and has a point charge, but only the oxygen atoms are assigned Lennard-Jones parameters. In rigid water models, the geometry of the atoms is held fixed. This dramatically reduces the computational cost and allows the explicit inclusion of millions of solvent molecules in large-scale molecular dynamics simulations. TIP3P is a popular example of a 3-point rigid water model [26]. SPCE is a rigid 3-point water model that additionally includes an average polarization correction term [27]. Fully polarizable water models exist but are not commonly used and not explored further in this thesis.

OPC is newer, rigid, and also non-polarizable. However, it employs a 4-point approach with a single van der Waals center on the oxygen nucleus [28]. The fourth point in OPC is a dummy point located near the oxygen and between the two hydrogens. The dummy point is given a negative charge which can improve the electrostatic distribution of the explicit water molecule. Furthermore, OPC differs from other water models because it places fewer constraints on the relative positions and charges during parameter fitting. Improved accuracy in the thermodynamics of ligand binding over older water models like TIP3P has been demonstrated with OPC [28]. OPC3 is a 3-point version of the OPC water model, which maintains significant improvements in accuracy over TIP3P and SPCE [29].

3.1.2.2 Molecular dynamics ensembles

In setting up a molecular dynamics simulation, we constrain a collection of atoms in a simulation cell and aim to study the dynamics of this system in isolation. In reality, however, such a system exists in the context of a wider environment with which it interacts. We must define to what extent the simulated system may interact with its surroundings. The different types of thermodynamic ensembles describe the degree of separation from the surroundings. In the context of molecular dynamics simulations, the microcanonical, canonical, and isothermal-isobaric ensembles are the most relevant.

In the microcanonical ensemble, the simulation cell is fully isolated from the surroundings. This ensemble is abbreviated as NVE because the number of particles (N), the volume of the simulation cell (V), and total energy in the system (E) are constant throughout the simulation. Thus throughout the MD simulations, potential and kinetic energy are exchanged while the total energy is conserved.

For the simulation of proteins, the canonical and isothermal-isobaric ensembles are more commonly used. In the canonical ensemble, the system is immersed in a heat bath of constant temperature T. A heat bath is considered to be sufficiently large such that heat moving from the system to the bath cannot change the temperature of the bath. The system's temperature is held constant through heat transfer with the bath. In addition to the constant temperature, the number of molecules (N) and volume of the simulation cell (V) are held constant, leading to the nomenclature NVT. In this ensemble, the total energy of the system changes. Computationally, a thermostat algorithm is required to add and remove energy from the boundary of the simulation cell. The Berendsen [30] and Anderson [31] thermostats are popular choices.

Finally, in the isothermal-isobaric ensemble, the system's volume can change, but the system's pressure must match the pressure exerted on the system by the surroundings. Additionally, the system is submerged in a heat bath of temperature T, as in the canonical ensemble. The number of particles also remains fixed, leading to the nomenclature NPT. In addition to the thermostat regulating heat transfer, a barostat is needed to control the pressure.

3.1.2.3 Periodic boundary conditions and particle mesh Ewald summation

To perform molecular dynamics on a computational unit cell of finite size, the boundary conditions along the edges of the simulation cell must be defined. Consider the molecular dynamics simulation of a protein in an explicit solvent. If rigid walls bound the simulation cell, the interaction of the solute and solvent molecules with the boundary would impact the bulk properties of the system. To circumvent this interaction with the boundary, periodic boundary conditions are employed. An infinite number of exact replicas or images of the simulation cell are constructed around it. Only the atoms in the original simulation cell are explicitly considered. If a particle leaves the simulation cell on one side, its replica from the adjacent cell enters on the opposite side of the simulation cell. When studying non-periodic systems like a protein in solvent, the size of the simulation cell must be sufficient to prevent the solute from interacting strongly with its images in the adjacent cells. In practice, this is achieved by adding a sufficiently deep layer of solvent molecules around the solute, unusually about 10 Å, to screen any self-interaction of the solute.

To avoid calculating the interaction of the simulated system with infinite replicas of itself at increased distances, and to reduce the computational cost, the non-bonded interactions, as described by the molecular mechanics force field, are truncated. Because of the rapid decay of the Lennard-Jones interactions with distance, they are usually truncated beyond 8 Å. A tail correction is computed for the contribution from beyond the cutoff distance by assuming the atoms are isotropically distributed beyond the cutoff. A further correction is required to ensure proper energy conservation despite the discontinuity in the potential introduced by the truncation of the Lennard-Jones interaction.

The electrostatic interactions are much more long-ranged and require more sophisticated treatment. Ewald summation [32] provides a computationally efficient solution. The electrostatic interaction is divided into one short-range and one long-range term. While the short-range term is calculated normally in real-space, the long-range term is calculated in reciprocal-space using a Fourier transform. Because the short-range interactions converge quickly in real-space and the long-range interactions converge quickly in Fourier-space, they can be truncated with little loss in accuracy. The computational implementation of this approach is called particle-mesh-Ewald (PME) and uses the highly computationally efficient fast Fourier transform [33]. The PME requires periodic boundary conditions and a charge-neutral simulation cell.

3.1.2.4 Molecular dynamics in Amber20: a simulation from start to finish

A brief overview is given of the steps involved in setting up and running an MD simulation of a protein-ligand complex in Amber [34]. The general procedure translates to any other MD simulation program.

First, knowledge of the atomic positions of each atom in the protein-ligand complex is required. Crystallographic structures are generally taken as a starting point and distributed in the PDB file format. Any unwanted molecules, ions, or solvent molecules are stripped from the PDB file. Hydrogen atoms, which are not always resolved in crystallographic structures, are added to the structure using Amber's LEaP utility. For amino acids with non-standard protonation states, the residues names in the PDB are changed accordingly to ensure correct placements of hydrogens. The utility pdb2amber can be used to ensure atom names and atom types correspond to the conventions used within Amber.

Suppose a ligand is present in the structure. In that case, the partial atomic charges for each atom in the ligand can be calculated using Antechamber with the semi-empirical quantum mechanical AM1-BCC method [35]. Next, LEaP is used to load the desired force field parameters for the protein and ligand, along with the ligand partial charges. A water force field is also loaded if the system is to be solvated. Ions to neutralize the total charge of the system are added where necessary. Next, the solutes are solvated in pre-equilibrated boxes of explicit waters corresponding to the choice of water force field. Finally, the necessary input files for MD in Amber are generated.

Before a production simulation can be started, however, the system must be thoroughly equilibrated. Equilibration occurs in several steps that involve minimization, heating the system to the temperature of the heat bath using the NVT ensemble, and adjusting the system volume and pressure in an NPT ensemble. Many philosophies for ”good” equilibration exist.

The production simulation can be set up when the system is deemed sufficiently equilibrated. Key considerations are the choice of ensemble, usually NVT or NPT, the choice of a thermostat and (if necessary) barostat, and the length of the simulation. The SHAKE algorithm [36] is commonly used to constrain the bonds involving hydrogen. This is done because due to the small mass of hydrogens, their bond vibrations have a very high frequency and thus limit the smallest possible time-step. By freezing these hydrogen involving bonds, longer time steps (usually 0.002 ps) can be used. This allows for more sampling at a lower computational cost. As introduced in the previous section, a cutoff distance beyond which non-bonded interactions are no longer calculated is also set, usually to 8.0 Å. The length of the simulation will depend on the system size, availability of computational resources, and specific goals of the simulation.

After completion of the simulation, the generated trajectories can be visualized in molecular viewing tools like VMD [37]. Amber’s Cpptraj [34] can be used to analyze many aspects of the trajectory, like the positional fluctuation of individual protein residues or the formation of hydrogen bonds between the protein and ligand. An ensemble of structures taken from the MD trajectory can serve as input for various binding free energy methods, like MM-PBSA and MM-GBSA, which are introduced in the next section.

3.1.3 The MM-PBSA/GBSA method

The MM-PBSA method, introduced by Kollman et al. in 2000 [15], is a popular end-point method for estimating relative free energies of binding. The method reduces the computational cost by making two key simplifications: 1) sampling only at the end-points of the binding process and 2) treating the solvent implicitly.

By sampling only from the binding end-points, the number of configurations required to converge the free energies is reduced dramatically. The binding pathway or intermediate states are not sampled like in other, more thermodynamically rigorous methods. Molecular dynamics simulations are used to generate a representative ensemble of configuration that serves as input to MM-PBSA, which is essentially a post-processing step performed on the output of dynamics simulations. While explicit waters are used in the dynamics simulation, the water molecules are stripped from the ensemble structures and replaced by an implicit solvent model in the MM-PBSA binding energy calculation. What constitutes ”adequate sampling” and a ”representative ensemble” is a topic of ongoing research and debate.

In MM-PBSA, the free energy of binding of a ligand, B, to a protein, A, is defined as the difference of the average free energy, $\langle G \rangle$, of the complex and its constituents,

$$\Delta G_{bind} = \langle G^{AB} \rangle - \langle G^{A} \rangle - \langle G^{B} \rangle. \tag{3.12}$$

The free energies of the complex, ligand, and protein are deconstructed into the following terms,

$$\langle G \rangle = \langle E_{MM} \rangle + \langle G_{solvation} \rangle - T\langle S \rangle, \tag{3.13}$$

where $\langle E_{MM} \rangle$ is the gas phase molecular mechanics energy, $\langle G_{solvation} \rangle$ is the free energy of solvation, and $-T\langle S \rangle$, is an entropy correction term. The gas phase energy is calculated using a molecular mechanics force field and the solvation energy using the PBSA implicit solvation model. Hence the name MM-PBSA.

To express the calculated binding free energy in terms of the individual energy and entropy terms, substitute equation 3.13 into equation 3.12 to yield,

$$\Delta G_{bind} = \langle \Delta E_{MM} \rangle + \langle \Delta G_{solvation} \rangle - T\langle \Delta S \rangle = \langle \Delta H_{bind} \rangle - T\langle \Delta S \rangle, \tag{3.14}$$

where $\Delta H = H^{AB} - H^{A} - H^{B}$ is the net change in enthalpy upon binding, and ΔE_{MM}, $\Delta G_{solvation}$ and ΔS are defined analogously.

3.1.3.1 The one-trajectory and three-trajectory approaches

There are two approaches to generating an ensemble of configurations for MM-PBSA. The first approach, which follows naturally from Equation 3.12 above, is to sample the dynamics of the protein-ligand complex, the unbound protein, and the unbound ligand separately using three MD simulations. From these, the free energies $\langle G^{AB} \rangle$, $\langle G^{A} \rangle$ and$\langle G^{B} \rangle$ are obtained. The phase-space of the unbound protein and unbound ligand are explicitly sampled. Thus, small changes in the conformational freedom of the ligand and protein upon binding can, in principle, be assessed. This has implications for the entropic contributions to binding discussed further in Section 3.1.3.5.

More commonly, however, only a single MD simulation of the protein-ligand complex is run. Ensembles of configurations for the unbound ligand and unbound protein are generated by deleting, in turn, one of the molecules from the trajectory of the complex. The free energies for the unbound ligand, $\langle G^B \rangle$, and protein, $\langle G^A \rangle$, are then computed from these ensembles. Thus, only a single dynamics simulation of the protein-ligand complex is needed. Furthermore, all intramolecular energy contributions cancel in Equation 3.12 for the binding free energy. This reduces noise in the calculation and improves the convergence of the energy change upon binding.

Some studies report that the single-trajectory approach is more accurate [38, 39]. However, by simulating only the complex, conformational changes of the protein and ligand upon binding are ignored. Consequently, the snapshots generated cannot be used to estimate the entropic contribution from the free protein and ligand. It follows that the single-trajectory approach is only appropriate when no significant structural changes to the protein or ligand are expected upon binding or if their energetic effects cancel in the calculation of relative binding free energies. A two-trajectory approach has also been proposed, in which the free ligand is simulated in addition to the complex. This would allow the inclusion of the ligand reorganization energy [40].

3.1.3.2 The gas phase term

As described in Section 3.1.2, the gas phase energy, $\langle E_{MM} \rangle$, computed from classical mechanical force fields will include terms corresponding to bonds, angles, torsions, van der Waals, and electrostatic interactions.

Molecular mechanical force fields do not allow for the breaking or forming of covalent bonds. When the single-trajectory approach is used, the bonded terms therefore cancel. In the three-trajectory approach, different conformations of the unbound protein and ligand could lead to minor bonded contributions to the net gas phase energy, $\langle \Delta E_{MM} \rangle$.

3.1.3.3 The solvation term: PBSA and GBSA

While molecular dynamics simulations are usually performed with explicit water molecules, the MM-PBSA binding energy method uses an implicit solvent model that replaces the water molecules with a continuum solvent. This is one of the two key simplifying assumptions of the MM-PBSA approach and significantly reduces the computational cost. In Section 2.2.3 an expression for the work function of transferring a solute molecule from the gas phase to the solvent was derived. By treating the solvent implicitly, the integrals over the solvent coordinates vanish from the expression. This section describes the details of the PBSA and GBSA solvation models.

In Equation 3.13 $\langle G_{solvation} \rangle$ is the mean free energy of solvation of the system. This term further split into a polar and non-polar term, $G_{solvation} = G_{pol} + G_{non-pol}$.

The polar solvation term The polar solvation term gives the change in free energy upon the transfer of a charged molecule from the gas phase to the solvent. It is calculated by replacing the explicit solvent of the snapshots generated by MD simulations with an implicit continuum solvent and numerically evaluating the Poisson-Boltzmann (PB) equation,

$$\nabla\epsilon(\bar{r})\nabla\phi(\bar{r}) = -4\pi(\rho(\bar{r}) + \rho_m(\bar{r})), \tag{3.15}$$

where $\epsilon(\bar{r})$ is the position-dependent dielectric constant, ϕ is the electrostatic potential, and ρ and ρ_m are the charge densities of the solute and mobile charge carriers, respectively. The electrostatic potential is given by the Coulombic term in the molecular mechanics force field. Usually, the implicit solvent is treated as a homogeneous medium with a constant dielectric, i.e. $\epsilon(\bar{r}) = \epsilon$. Similarly, mobile charge carriers are not generally considered in PBSA solvation and $\rho_m = 0$. After a cavity is formed and the solute is placed inside the solvent, the solute volume is assigned a dielectric constant of 1. Some studies propose using higher solvent dielectric constants of 2 or 4. In 2011 Hou et al. [41] showed that the choice of the dielectric constant should depend on the nature of the protein-ligand interface, with higher charged interfaces benefiting from a higher interior dielectric constant. However, a follow-up study in 2014 found that an interior dielectric of 4 yielded the best overall results [42]. For water, the continuum dielectric constant is about 80. The simplified Poisson equation is then given by,

$$\nabla^2\phi(\bar{r}) = -\frac{4\pi}{\epsilon}\rho(\bar{r}), \tag{3.16}$$

where ϵ has two values, one inside the solute and one in the solvent. More advanced solvent models treat the dielectric constant as in-homogeneous and implement smooth transition between the solvent ($\epsilon = 80$ for water) and solute ($\epsilon = 1, 2, 4$) regions [43]. The Poisson equation is a three-dimensional partial differential equation and is solved numerically using a finite-difference method on a discretized grid.

An alternative approach for the polar solvation term is the Generalized Born (GB) method. It is based on the analytic expression known as the Born expression,

$$\Delta G_{solv} = -\left(1 - \frac{1}{\epsilon_{out}}\right)\frac{q^2}{2A}, \tag{3.17}$$

for the free energy required to transfer a spherical ion of radius A with a charge q at its center from the gas phase to a solvent, characterized by a continuum dielectric of ϵ_{out}. In this expression, both the solvent-solvent and ion-solvent interactions are encapsulated and depend on the ion radius A, ion charge q, and the solvent continuum dielectric ϵ_{out}. Equation 3.17 is an exact solution to the Poisson equation for a spherical solute with a single charge at its center. It can be shown that in the limit where $\epsilon_{out} \to \infty$,

$$\Delta G_{solv} = -\frac{1}{2}\left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{out}}\right)\sum_{i,j}\frac{q_i q_j}{\sqrt{r_{ij}^2 + R_i R_j}}, \tag{3.18}$$

for an arbitrary distribution of charges i inside a spherical solute with internal dielectric ϵ_{in}. R_i are called effective radii and depend on the radius of the solute sphere A and the position r_i of the charges i inside the spherical solute,

$$R_i = A - \frac{r_i^2}{A}. \tag{3.19}$$

Real molecules are, however, not spherical. Suppose we approximate the molecule as a collection of charges i embedded in spheres of radii a_i and assume that the separation of the spheres r_{ij} is much larger than their radii a_i. In that case, the solvation energy may be written as a summation over individual Born terms (Equation 3.17).

$$\Delta G_{solv} \approx \sum_{i}^{N} -\left(1 - \frac{1}{\epsilon_{out}}\right)\frac{q_i^2}{2a_i} - \frac{1}{2}\sum_{i}^{N}\sum_{j\neq i}^{N}\frac{q_i q_j}{r_{ij}}\left(\frac{1}{\epsilon_{out}} - 1\right) \tag{3.20}$$

Generalized Born solvation models seek to find simple analytic formulas similar to Equation 3.20. A very successful Generalized Born model [44] is given by,

$$\Delta G_{solv} \approx \left(1 - \frac{1}{\epsilon_{out}}\right)\frac{1}{2}\sum_{i,j}^{N}\frac{q_i q_j}{f_{ij}^{GB}}, \tag{3.21}$$

where f_{ij}^{GB} is the GB-kernel [45] and given by,

$$f_{ij}^{GB} = \left[r_{ij}^2 + R_i R_j \exp\left\{\frac{-r_{ij}^2}{4R_i R_j}\right\}\right]^{\frac{1}{2}}, \tag{3.22}$$

where R_i are the effective Born radii of atoms in the solute. These Born radii depend on the true radius of the atoms but also on the radii and positions of all other atoms in the solute. There are a variety of approaches to estimating the Born radii, but for the purpose of this thesis, they are treated as empirical inputs to the Generalized Born solvation model. While the PB solvation term requires the numerical solution of a partial differential equation, the GB model is a simple summation over atoms in the solute and thus significantly computationally cheaper.

Studies comparing the MM-PBSA and MM-GBSA methods showed that the accuracy of the methods is comparable, but system-dependent [39, 41, 42].

The non-polar solvation term The non-polar term gives the energy required to form a cavity for the solute in the solvent. Physically it is composed of the cost of creating the cavity (cavitation) and the dispersion (attractive) and repulsion (repulsive) due to van der Waals interactions between the solvent and solute. This second contribution is referred to as the dispersive term.

In its most basic form, the non-polar term is modeled via the Solvent Accessible Surface Area (SASA) using the linear relationship,

$$G_{non-pol} = \gamma SASA + b, \tag{3.23}$$

where γ is related to the surface tension and b is a correction parameter. The most empirical approach is to fit γ, and b to a training set of experimental results [46]. These terms lose their physical significance in such approaches and are only fitted parameters. Various algorithms exist to define the SASA, but a common implementation rolls a fictitious sphere of some pre-defined radius, usually 1.7 Å, across the solute.

Extensions to this approach attempt to explicitly include a van der Waals dispersion term derived from the solvent accessible volume/area or otherwise [39, 47]. Often, the dispersive effects are included in an approximate manner by either re-scaling the surface-tension or parameterizing γ to account for dispersive effects (GBSA). The MD code Amber20 computes a separate dispersion term via a surface-integration approach [48]. More complicated approaches like the polarized continuum model (PCM), in which the charge distribution across a closed surface around the solute is calculated, can also be used to estimate solvation energies[49][50].

In theory, the cavitation energy should also contain an entropic contribution from the re-organization of the solvent around the solute. However, this is usually ignored and is partially captured by the parameterization of the solvent. Continuum solvent models ignore specific water interactions. The inclusion of some explicit waters to circumvent this issue has been proposed [15, 51].

One problem with the non-polar term is that it incorrectly handles the SASA of buried cavities [46]. A post-hoc correction term can be applied for the cavity filled with a non-interacting dummy ligand prior to binding. This issue does not arise for binding sites directly exposed to the solvent.

3.1.3.4 MM-PBSA/GBSA with explicit waters

In traditional MM-PBSA and -GBSA all explicit water molecules are stripped from the MD trajectory before calculating binding energies. However, it has been found that in some systems, that including some explicit water molecules as part of the protein can improve the resulting energies. In 2003 Spackova et al. [52] included the 20 closest water molecules to the ligand in a DNA duplex system but noticed no drastic impact on binding energies obtained. In 2009, Wong et al. [53] found that including all interfacial water in a protein-protein system degraded results and increased statistical errors while including only two specific bridging waters produced the correct energy trends. In a more systematic approach, Maffucci et al. [54] tested the Nwat-GBSA method in which the N closest waters to the ligand are kept in MM-GBSA protein-ligand binding free energy calculations. Significant improvements in correlation to experiment were observed in two of four protein systems and minor improvements, probably within uncertainties, in the other two systems. The same group tested the Nwat-GBSA method in protein-protein systems in 2016, finding improvement in the correlation of up to 30% across 20 protein-protein systems [55]. In 2014, Mikulskis et al. [51] used MM-GBSA to predict the binding energies of 9 ligands in ferritin and tested the straightforward Nwat method as well as a more involved approach, aiming to include a contribution from the displacements of water from the binding site due to ligand binding. Both methods lead to moderate improvements in the estimated energies, but normal mode entropies were found not to work when explicit waters are included. Also, in 2014, Zhu et al. [56] included waters within a cutoff distance of the protein and ligand heavy atoms in four protein-ligand systems finding significant improvements in correlation in three of four systems. Improvements were especially significant where water molecules mediated protein-ligand hydrogen bonds. In 2017, Aldeghi et al. [57] used the Nwat method on a set of 11 ligands in BRD4 and found a moderate improvement in the correlation of predicted against experimental binding energies. Furthermore, Guest et al. [58] showed in 2021 that 5-6 structural waters in the BRD4 binding site are highly conserved for the vast majority of available crystal structures of BRD4-ligand complexes and that docked binding poses were improved by the explicit inclusion of the structural waters as part of the receptor.

3.1.3.5 Entropy in MM-PBSA/GBSA

The change in entropy upon binding is the most challenging term in MM-PBSA to estimate and usually has the largest statistical uncertainty [39]. Intuitively, ΔS describes the change in conformational freedom of the protein and ligand upon binding. The binding process will change both the positional and orientational degrees of freedom. Gohlke et al. [59] express the change in entropy as the difference in the conformational entropy due to motions of the ligand and protein of the bound complex and the conformational entropy of the free ligand and protein.

In most applications of MM-PBSA, as well as its original formulation by Kollman et al. [15], the change in entropy upon binding is estimated as the difference in conformational entropy of the complex and its constituents, as calculated by normal mode analysis (NMA) [15]. NMA is based on the idea that a system at equilibrium that is displaced from equilibrium by a small perturbation will experience a restoring force that returns it to its equilibrium conformation [60]. At equilibrium, the potential energy of the systems is considered to be at a global minimum. The global minimum assumption in normal mode entropy calculations is the assumption that the protein-ligand complex exists in a unique, low-energy conformation that represents the global minimum of the potential energy surface. This assumption is used to simplify the calculation of the normal mode entropy by assuming that the vibrational modes of the complex are harmonic and that the potential energy surface around the minimum is parabolic. However, the global minimum assumption may not always hold true, especially for flexible systems or for systems with multiple low-energy conformations.

Consider the power series expansion of the potential energy V as a function of the coordinates q describing a protein conformation around the equilibrium conformation q^0,

$$V(q) \approx V(q^0) + \left(\frac{\partial V}{\partial q_i}\right)^0 \eta_i + \frac{1}{2}\left(\frac{\partial^2 V}{\partial q_i \partial q_j}\right)^0 \eta_i \eta_j, \tag{3.24}$$

where $\eta_i = q_i - q_i^0$ and q_i are the instantaneous configurations of the protein components. At the equilibrium conformation q^0, the first term in Equation 3.24 becomes a constant, i.e. the energy minimum, and can be set to zero. The second term is also zero because the derivative at a local minimum is always zero. This leaves the second order partial differential equation,

$$V(q) = \frac{1}{2}\left(\frac{\partial^2 V}{\partial q_i \partial q_j}\right)^0 \eta_i \eta_j = \frac{1}{2}\eta_i H_{ij} \eta_j, \tag{3.25}$$

where H_{ij} is the Hessian matrix of second-order partial derivatives of the potential with respect to the protein components. The molecular mechanics force field gives the potential energy V. Combining Equation 3.25 with the kinetic energy term,

$$T(q) = \frac{1}{2}M\frac{\mathrm{d}^2\eta_i}{\mathrm{d}t^2}, \tag{3.26}$$

where M is the diagonal mass matrix, yields the differential equation,

$$\frac{1}{2}M\frac{\mathrm{d}^2\eta_i}{\mathrm{d}t^2} + \frac{1}{2}\eta_i H_{ij} \eta_j = 0. \tag{3.27}$$

One solution to this differential equation is of the form,

$$\eta_i = a_{ik}\cos(\omega_k t + \delta_k), \tag{3.28}$$

which describes an oscillation of amplitude a_{ik} and frequency ω_k around the equilibrium conformation. Substituting this solution into the equation of motion yields the eigenvalue problem $HA = \lambda A$ where H is the Hessian matrix, and A contains the normal mode vectors (eigenvectors). The eigenvalues in λ are the squares of the frequencies of the normal modes. Using an analytic expression, the normal mode frequencies are used to calculate the vibrational entropy of the protein system. Larger vibrational amplitudes correspond to larger entropic contributions of the normal modes [60]. Because the normal modes are orthogonal, i.e. independent, the total entropy is a summation over all normal modes. In practice, the protein system is stripped of explicit solvent molecules and minimized before the normal mode frequencies are calculated by. To approximate the impact of the solvent, the calculation is generally performed in an implicit solvent.

The harmonic oscillator approximation assumes that the potential energy surface around the minimum of the potential is parabolic and that the motion of the system is described by the harmonic oscillator equation. In diffusive systems, the potential energy surface is not necessarily parabolic, and the motion of the system is not well described by the harmonic oscillator equation. The diffusion process involves a random walk of the particles, and the energy landscape can be complex and multi-modal. The use of the harmonic oscillator approximation in this context can result in inaccurate estimates of the vibrational frequencies and associated entropy contributions. One consequence of the harmonic oscillator approximation for diffusive systems is that it may underestimate the contribution of low-frequency modes to the entropy. Diffusive systems typically exhibit slow, collective motions that can contribute significantly to the entropy, but may not be well-described by the harmonic oscillator approximation. This can lead to an underestimation of the entropy contribution, and potentially inaccurate predictions of thermodynamic properties. Another consequence of the harmonic oscillator approximation for diffusive systems is that it may not capture the anharmonicity of the potential energy surface, which can be important for systems with strong intermolecular interactions. The anharmonicity of the potential can lead to the coupling of vibrational modes and the appearance of overtones, which can have a significant impact on the thermodynamics of the system.

Quasi-harmonic analysis can take into account anharmonic effects in the potential energy surface [61]. It is an extension of the harmonic oscillator approximation. In quasi-harmonic analysis, the potential energy surface is approximated as a sum of harmonic and anharmonic terms. The harmonic part is calculated is in normal mode analysis from the Hessian matrix, which describes the second derivatives of the potential energy with respect to the atomic coordinates. The anharmonic part is obtained from higher-order derivatives of the potential energy. The anharmonic terms account for deviations from the harmonic approximation and can be important for systems with strong intermolecular interactions, such as proteins.

As described in the previous section, often, only a single MD trajectory of the entire complex is used in the MM-PBSA method, and the structures of the unbound ligand and protein are extracted by simple deletion of the other solute as well as the solvent molecules from the simulation of the complex. If these structures are used for the entropy calculation, the free protein and ligand contributions are ignored because the unbound protein and ligand were not sampled. One standard approximation to address this problem is calculating the relative binding free energies of similar ligands in the same protein binding site. Given that the ligands are sufficiently similar, the entropy contributions from the free protein and ligand are assumed to cancel. Often, this is taken one step further, and it is assumed that all entropy terms cancel. This implies that the total change in entropy upon binding is the same for two different ligands. The change in entropy of the solvent is usually completely ignored in MM-PBSA calculations. To some extent, it is implicitly included via the parameterization of the implicit solvent model.

Recently, the interaction entropy (IE) method by Duan et al. [62] has gained popularity because it is a post-processing method that does not require additional simulations beyond the original MD. The Interaction Entropy method (IE) is based on exponential averaging of the fluctuation of the protein-ligand interaction energy for electrostatic and van der Waals energy contributions. The protein-ligand interaction energy, E_{pl}^{int}, is defined as the difference in the gas phase energy of the protein-ligand complex and its constituents, i.e., the separate host and ligand. The fluctuation of the interaction energy of each snapshot around the ensemble-averaged interaction energy, $\langle E_{pl}^{int} \rangle$, is calculated, $\Delta E_{pl}^{int} = E_{pl}^{int} - \langle E_{pl}^{int} \rangle$. The interaction entropy is an exponential average of the fluctuation in interaction energy around the ensemble average,

$$-T\Delta S = KT \ln \langle e^{\beta \Delta E_{pl}^{int}} \rangle, \tag{3.29}$$

where,

$$\langle e^{\beta \Delta E_{pl}^{int}} \rangle = \frac{1}{N} \sum_{i=1}^{N} e^{\beta \Delta E_{pl}^{int}(t_i)}, \tag{3.30}$$

where N is the number of snapshots and is $\Delta E_{pl}^{int}(t_i)$ calculated for each snapshot i.

Link to statistical mechanics In section 2.2.4, a derivation of the entropy change upon binding was presented in some detail. This section shows which of the theoretically present entropy contributions are included in the MM-PBSA + NMA framework and highlights some potential issues.

Formally, the change in entropy upon complexation, at standard state, relative to the change in free energy is defined as,

$$\Delta S^0 = -\left(\frac{\partial \Delta G^0}{\partial T} \right)_P. \tag{3.31}$$

Recall the separation of the ligand and protein DOF in section 2.2 into internal and external DOF. Upon binding, the ligand's external DOF are converted to the complex's internal DOF. Intuitively, the external DOF of the ligand are restricted to the binding site by the complexation with the protein. This loss of freedom in the external DOF of the ligand is the fist component of the ”external entropy” change due to the external coordinates of the ligand. The second component is due to the 6 new internal degrees of freedom from converting the ligand’s external DOF to internal DOF of the complex.

To illustrate this, consider a single-trajectory simulation of the ligand phenol binding to the protein T4-lysozyme LA99/M102Q, which is the next chapter’s subject. NMA of the ligand will yield 39 vibrational modes, 6 of which are zero and correspond to the 6 external rotational and translational DOF. Similarly, the host protein has 7806 vibrational modes with 6 zero-energy modes. The complex has 7845 modes, also with 6 zero-energy modes. Thus, the complex has 6 more non-zero vibrational modes, or internal DOF, than the host + ligand. This results in the vibrational entropy of the complex being higher than that of the host and ligand combined, in this case $\Delta S^{vib} = S^{vib}_{comp} - S^{vib}_{host} - S^{vib}_{lig} = 2\,\mathrm{cal} * \mathrm{mol}^{-1}\mathrm{K}^{-1}$. However, the change in translational and rotational entropy is negative because the complex only has 6 external DOF while the host and ligand each have 6. Under the gas phase assumption of NMA, the translational and rotational entropy are calculated from analytic formulas [63]. The translational entropy is logarithmically proportional to the total mass. The rotational entropy is logarithmically proportional to the principle moments of inertia. The change in translational entropy is constant over the MD trajectory as there are no changes in the mass of the constituents, taking a value of $-39\,\mathrm{cal} * \mathrm{mol}^{-1}\mathrm{K}^{-1}$ in the example system. The total translational entropy of the ligand is $39\,\mathrm{cal} * \mathrm{mol}^{-1}\mathrm{K}^{-1}$ indicating that the entirety of the ligand’s translational entropy is lost upon binding. This is true by construction in the single-trajectory approach. The change in rotational entropy varies slightly over the MD simulation due to the spatial re-organization of the atoms. Its value fluctuates around $-28\,\mathrm{cal} * \mathrm{mol}^{-1}\mathrm{K}^{-1}$, which is very close in magnitude to the total rotational entropy of the ligand. The total change in external entropy upon binding is thus negative.

It is important to emphasize that the translational, vibrational, and rotational entropy terms in the single-trajectory approach are equivalent to the ”external entropy”, as derived in section 2.2.4, because they arise due to the external coordinates of the ligand becoming internal coordinates of the complex. The ”internal entropy” due to only the internal coordinates of the complex cannot be estimated in the single-trajectory approach because there is no conformational change upon binding. In the three-trajectory approach, the change in vibrational entropy would include a new contribution from the internal conformational changes of the protein and ligand upon binding. This ”internal entropy” change could not be separated from the external entropy change due to the 6 new vibrational modes in the complex. The change in translational entropy would remain the same. However, more significant variations in the rotational entropy could be expected due to the full sampling of the free protein and ligand.

The entropy term corresponding to the solvent coordinates gives the change in entropy of the solvent upon binding of the solutes. This term is neglected in the gas phase approximation of NMA. This is equivalent to the assumption that the solvation energies are independent of temperature at the standard pressure [12].

3.1.3.6 Connecting statistical mechanics to the MM-PBSA method

In 2004, Swanson et al. [40] made an attempt to link the approximate MM-PBSA/GBSA method to statistical mechanics of binding. Building on the configurational integrals derived by Gilson [12], Swanson applied a series of approximations, which are outlined and applied below.

Analogous to equation 2.26, the configurational integral for the complex AB in an implicit solvent, using the same symbols as above, is given by,

$$Z_{AB} = \frac{Z_{N,AB}}{Z_{N,0}} = \int \exp\{-\beta[U(\bar{r}_A, \bar{r}_B, \zeta_B) + W(\bar{r}_A, \bar{r}_B, \zeta_B)]\} d\bar{r}_A d\bar{r}_B d\zeta, \tag{3.32}$$

where ζ are the external coordinates of the ligand that have become internal coordinates of the complex AB. The first assumption made by Swanson is that fluctuations in ζ are small for a bound ligand and hence the higher order coupling of ζ to the internal degrees of freedom of the protein and ligand are small. Thus both the potential and solvation term are separable as follows,

$$U(\bar{r}_A, \bar{r}_B, \zeta_B) + W(\bar{r}_A, \bar{r}_B, \zeta_B) = U_1(\zeta_B) + W_1(\zeta_B) + U_2(\bar{r}_A, \bar{r}_B) + W_2(\bar{r}_A, \bar{r}_B). \tag{3.33}$$

This assumption is supported by the observation, that displacement of ζ during an MD simulation of the complex is small. Now define the potential of mean force,

$$w(\zeta_B) = -RT \ln \left[\int \exp\{-\beta\, [U(\bar{r}_A, \bar{r}_B, \zeta_B) + W(\bar{r}_A, \bar{r}_B, \zeta_B)]\} d\bar{r}_A d\bar{r}_B \right], \tag{3.34}$$

and apply the approximation in 3.33,

$$w(\zeta_B) = -RT \ln \left[\int \exp\{-\beta\, [U_1(\zeta_B) + W_1(\zeta_B) + U_2(\bar{r}_A, \bar{r}_B) + W_2(\bar{r}_A, \bar{r}_B)]\} d\bar{r}_A d\bar{r}_B \right] \tag{3.35}$$

$$= U_1(\zeta_B) + W_1(\zeta_B) - RT \ln \left[\int \exp\{-\beta\, [U_2(\bar{r}_A, \bar{r}_B) + W_2(\bar{r}_A, \bar{r}_B)]\} d\bar{r}_A d\bar{r}_B \right]. \tag{3.36}$$

The next approximation made is that the translational and rotational DOF of the ligand external coordinates ζ_B can be decoupled such that,

$$U(\zeta_B) = U(x_1, x_2, x_3) + U(\zeta_1, \zeta_2, \zeta_3), \tag{3.37}$$

and likewise for $W(\zeta_B)$. Inserting the potential of mean force $w(\zeta_B)$ into the complex configurational integral (3.32),

$$Z_{AB} = \int \exp\{-\beta\,[U(\zeta_B) + W(\zeta_B)]\}d\zeta_B \int \exp\{-\beta\,[U_2(\bar{r}_A, \bar{r}_B) + W_2(\bar{r}_A, \bar{r}_B)]\}d\bar{r}_A d\bar{r}_B \tag{3.38}$$

and decoupling rotation and translation in ζ_B,

$$Z_{AB} = z_B^{trans} z_B^{rot} \int \exp\{-\beta\,[U_2(\bar{r}_A, \bar{r}_B) + W_2(\bar{r}_A, \bar{r}_B)]\}d\bar{r}_A d\bar{r}_B = z_B^{trans} z_B^{rot} Z'_{AB}, \tag{3.39}$$

where,

$$z_B^{trans} = \int \exp\{-\beta\,[U(x_1, x_2, x_3) + W(x_1, x_2, x_3)]\}dx_1 dx_2 dx_3, \tag{3.40}$$

$$z_B^{rot} = \int \exp\{-\beta\,[U(\zeta_1, \zeta_2, \zeta_3) + W(\zeta_1, \zeta_2, \zeta_3)]\}d\zeta_1 d\zeta_2 d\zeta_3, \tag{3.41}$$

and

$$Z'_{AB} = \int \exp\{-\beta\,[U_2(\bar{r}_A, \bar{r}_B) + W_2(\bar{r}_A, \bar{r}_B)]\}d\bar{r}_A d\bar{r}_B. \tag{3.42}$$

The standard state binding free energy of the complex is then given by,

$$\Delta G^0_{AB} = -RT \ln \left[\frac{C^0}{8\pi^2} z_B^{trans} z_B^{rot} \frac{Z'_{AB}}{Z_A Z_B} \right], \tag{3.43}$$

where the symmetry numbers σ_{AB}, σ_A and σ_B have been set to 1 (non-symmetric system).

To practically evaluate the configurational integrals, it is assumed that the energy landscape can be explored by a sufficiently long MD simulation. Swanson calls this the 1st order approximation,

$$Z_A = \int \exp\{-\beta\,[U(\bar{r}_A) + W(\bar{r}_A)]\}d\bar{r}_A \approx z_A^{int} \exp\{-\beta \langle E_A \rangle\}, \tag{3.44}$$

where $\langle E_A \rangle = \langle U(\bar{r}_A) + W(\bar{r}_A) \rangle$ and z_A^{int} is the internal configurational integral over all internal degrees of freedom. Similar expressions can be written for Z_B and Z_{AB}. The final assumption made to arrive at the MMPBSA/GBSA method is that the volume of configurational space occupied by the ligand and protein change negligibly upon binding. In other words, the internal energies, i.e. those due to intramolecular interactions, of the protein and ligand remain unchanged upon binding. Mathematically, $z_A^{int} z_B^{int} \approx z_{AB}^{int}$, and hence the internal configurational integrals cancel in the expression for the binding free energy resulting in,

$$\Delta G^0_{AB} = -RT \ln \left[\frac{C^0}{8\pi^2} z_B^{trans} z_B^{rot} \frac{\exp\{-\beta \langle E_{AB} \rangle\}}{\exp\{-\beta \langle E_A \rangle\} \exp\{-\beta \langle E_B \rangle\}} \right] \tag{3.45}$$

$$= \langle E_{AB} \rangle - \langle E_A \rangle - \langle E_B \rangle - RT \ln \left[\frac{C^0}{8\pi^2} z_B^{trans} z_B^{rot} \right]. \tag{3.46}$$

This is equivalent to the single-trajectory approach, in which the internal energies and entropies of the protein and ligand cancel those of the complex exactly. If the solvation energy W is represented by an implicit solvent model, then the above expression for the binding free energy may be written as,

$$\Delta G^0_{AB} = \langle \Delta E_{gas} \rangle - \langle \Delta G_{solv} \rangle - T\Delta S, \tag{3.47}$$

where $\langle \Delta E_{gas} \rangle$ is the ensemble average change in gas phase energy upon binding, $\langle \Delta G_{solv} \rangle$ is the ensemble average change in solvation energy upon binding, and $-T\Delta S$ is the change in entropy due to the restricted translation and rotation of the ligand in the protein binding site. This is precisely the formulation of the MMPBSA/GBSA approach. Furthermore, the assumption that z^{int} cancel is equivalent to the single-trajectory MMPBSA in which only the complex ensemble is sampled. The entropy term,

$$-T\delta S = -RT \ln \left[\frac{C^0}{8\pi^2} z_B^{trans} z_B^{rot} \right], \tag{3.48}$$

is however not necessarily equivalent to the traditional $-T\Delta S$ term included in MM-PBSA via Normal Mode Analysis. It is not clear from this formulation if the gain of 6 new internal DOF in the complex are captured by the configurational integrals z_B^{trans} and z_B^{rot}. Because z_B^{trans} and z_B^{rot} contain integrals over the entire phase space of the external coordinates of the ligand, this formulation of the entropy seems to be equivalent to a complete loss of the rotational and translational entropy of the ligand upon binding, as seen in single-trajectory MM-PBSA.

3.1.3.7 Comparison with alchemical free energy methods

Alchemical free energy methods involve gradually transforming the ligand from the initial to the final state, while continuously monitoring the changes in free energy along the transformation path. The transformation can be done using a variety of protocols, including linear or exponential scaling of the coupling parameter. However, alchemical free energy methods can require significant simulation time to achieve convergence, especially for larger systems [64]. The convergence rate depends on the choice of protocol, which can affect both the accuracy and computational cost of the method. Alchemical free energy methods can provide valuable insights into the energetics of protein-ligand binding, including the contributions of specific residues and water molecules to the binding affinity.

End-point methods, as discussed in this chapter, involve directly computing the free energy difference between the bound and unbound states of the protein-ligand complex. This approach is faster and more straightforward compared to alchemical free energy methods, but it can suffer from insufficient sampling of the configurational space, particularly for larger systems.

In MM-PBSA and MM-GBSA, as well as other end-point binding free energy methods, the ensemble averages, as written in equation 3.47, are taken over the entire system, i.e the whole protein-ligand complex. In contrast, in alchemical free energy methods, the ensemble averages are taken only over the changed regions of the complex, which is generally small compared to the whole system. This leads to more attractive convergence properties of methods like free energy perturbation and thermodynamic integration. Furthermore, in MM-PBSA and MM-GBSA, the total energies of the host and complex are compared with the total energy of the ligand, which is much smaller. Thus, the precision in the total energies has a significant effect on the accuracy of the calculated binding energy. The one-trajectory approach addresses this to some extent, as discussed in section 3.1.3.1.

3.1.3.8 Summary

While MM-PBSA is based on severe approximations and the single-trajectory approach suffers from several conceptual problems, the low computational cost and relative simplicity have popularised the method. The MM-PBSA method is used actively in prospective drug design and lead identification. Recent examples include efforts to identify potential treatments of Covid-19 [65, 66] and Alzheimer's [67, 68] and to better understand Down Syndrome [69]. However, as reviews of this method emphasize, the accuracy of MM-PBSA/GBSA is highly dependent on the system under investigation and depends strongly on the choice of force-field, implicit solvent parameters, and, to a lesser extent, simulation length [39, 42, 47, 70, 71]. While some observers have stated that the severity of its assumptions makes MM-PBSA/GBSA difficult to systematically improve, other see promise in improving the individual terms of the method with new approaches to implicit solvation [48, 50], entropy estimates [62, 72], and higher-accuracy energy methods [1, 38, 73].

3.2 Quantum mechanical protein-ligand free energies of binding

A fundamental limitation common to MM-PBSA/GBSA and all other classical mechanical computational methods of estimating binding free energies is the assumption of the validity of classical mechanics. The atoms and electrons that constitute biological molecules are however governed by the laws of quantum mechanics. Thus, a true description of protein-ligand binding requires a quantum mechanical (QM) treatment. In theory, a full, ab-initio QM approach would be system-independent, parameter-free, and could describe the full spectrum of physical phenomena involved in protein-ligand binding.

Unfortunately, high-level QM methods like coupled-cluster (CC-SD) are prohibitively expensive due to unfavorable scaling of their computational cost with increasing system sizes. Even small-molecule drugs are often too large for routine calculations with these methods. Instead, more approximate or empirical methods can be employed to find a balance between computational cost and theoretical rigor.

In this section, we first outline the motivation for exploring protein-ligand binding with quantum mechanical simulations. We then introduce density functional theory, the electronic structure method at the core of this research. Finally, we provide an overview of literature surrounding quantum mechanical protein-ligand binding energies and highlight recent advances in the field.

3.2.1 What classical mechanics misses

The limitations of classical mechanics force fields can broadly be split into two categories: theoretical and practical.

Beginning with the theoretical, a classical mechanical description of inter- and intra-molecular interactions cannot explicitly describe the behavior of electrons, as these are governed by the laws of quantum mechanics. Instead, a coarse-grained approach that treats atoms as holding charge is required. Any explicit description of electron-mediated interactions is lost. Examples include charge transfer, polarization, halogen bonding, and quantum mechanical many-body effects [3, 8]. In the context of protein-ligand binding prediction, metalloproteins containing a metal ion and highly charged ligands are two examples in which classical mechanical force fields break down because of electron-dominated interactions. Furthermore, traditional molecular mechanics force fields cannot represent chemical reactions and the breaking of covalent bonds.

On the other hand, molecular mechanics force fields also suffer from many practical problems. Because of the central role of some QM interactions in many biological processes, these interactions must be accounted for somehow. This is achieved by using fitted empirical parameters to reproduce experimental or calculated QM results. The performance of traditional force fields becomes heavily dependent on the parameterization quality with a trade-off between accuracy and domain of applicability. Using different force fields for the protein and the ligand in MD simulations of a protein-ligand complex is common. Fundamentally, the application of a force field is limited to those functional groups included in its parametrization. This also extends to the conditions under which the system was parameterized like the presence of solvent, the temperature, or pressure. In the context of protein-ligand binding energy prediction, an example of this are new molecules whose functional groups are not well represented in the training set of the force field.

Few or no fitted parameters are necessary if the electron-mediated quantum mechanical interactions between the protein and ligand can be fully described using quantum mechanical simulations. This not only improves the ease of use but also extends the domain of applicability and transferability of the binding energy protocol.

In addition to the transferability, the quantum mechanical simulations should be able to determine more realistic energies for the protein-ligand system as additional physical phenomena are explicitly accounted for. This has the potential to improve the predictive power of the binding energy protocol.

In this body of research, two classes of QM-based methods are utilized, the fully ab-initio density functional theory (DFT) and a semi-empirical general-purpose tight-binding method. These methods are introduced in the following section.

3.2.2 Density functional theory

Density functional theory is an ab-initio electronic structure method primarily used for calculating ground state energies of a diverse range of physical, chemical, and biological systems. Since the 1970s, it has remained one of the most popular electronic structure methods due to its small number of well-defined approximations, tractable computational cost, and extensive domain of applicability.

In traditional quantum theory, the state of a system is described by the spatial electron wavefunction $\Psi(\bar{r}_n)$ and its evolution in time, as described by the Schrödinger equation where $\bar{r}_n$ are the position vectors of n electrons. QM observables, like the total energy, are given by expectation values of operators acting on the wavefunction Ψ. Consider the total energy Hamiltonian operator for our system of interacting electrons and nuclei, $\hat{H} = \hat{T} + \hat{V}$, where $\hat{T}$ is the kinetic energy operator and $\hat{V}$ the potential energy operator. When applied to the wavefunction, it returns a set of energy eigenvalues, the lowest of which gives the total ground state energy of the system. This eigenvalue problem is generally expressed as,

$$\hat{H} \left| \Psi(\bar{r}_n) \right\rangle = E \left| \Psi(\bar{r}_n) \right\rangle , \tag{3.49}$$

where the energy expectation value E is given by,

$$E = \left\langle \Psi(\bar{r}_n) \right| \hat{H} \left| \Psi(\bar{r}_n) \right\rangle . \tag{3.50}$$

The Hamiltonian can be further split into terms for the nuclear kinetic energy, $\hat{T}_N(\bar{R}_K)$ where $\bar{R}_K$ is the position vector of nucleus K, the electron kinetic energy, $\hat{T}_e(\bar{r}_i)$ where $\bar{r}_i$ is the position vector of electron i, the nuclear-nuclear repulsion, $\hat{V}_{NN}(\bar{R}_K)$, the nuclear-electron attraction, $\hat{V}_{Ne}(\bar{R}_K,\bar{r}_i)$, and the electron-electron repulsion, $\hat{V}_{ee}(\bar{r}_i)$,

$$\hat{H} = \hat{T}_N(\bar{R}_K) + \hat{T}_e(\bar{r}_i) + \hat{V}_{NN}(\bar{R}_K) + \hat{V}_{Ne}(\bar{R}_K,\bar{r}_i) + \hat{V}_{ee}(\bar{r}_i) \tag{3.51}$$

$$\hat{H} = \frac{1}{2}\sum_{K=1}^{N}\frac{\nabla^2_{\bar{R}_K}}{M_K} - \frac{1}{2}\sum_{i=1}^{n}\nabla^2_{\bar{r}_i} + \frac{1}{2}\sum_{K=1}^{N}\sum_{L\neq 1}^{N}\frac{Z_K Z_L}{\bar{R}_{KL}} \tag{3.52}$$

$$-\frac{1}{2}\sum_{K=1}^{N}\sum_{i=1}^{n}\frac{Z_K}{\bar{R}_{Ki}} + \frac{1}{2}\sum_{i=1}^{n}\sum_{j\neq 1}^{n}\frac{1}{\bar{r}_{ij}}, \tag{3.53}$$

for a system of N nuclei with mass M_K and atomic charge Z_K, n electrons and using atomic units $\hbar = m_e = 4\pi\epsilon_0 = 1$ where $\bar{R}_{KL} = |\bar{R}_K - \bar{R}_L|$, $\bar{r}_{ij} = |\bar{r}_i - \bar{r}_j|$, and $\bar{R}_{Ki} = |\bar{R}_K - \bar{r}_i|$. This expression is incredibly complex, with three positional degrees of freedom for each nucleus and electron as well as cross-interaction terms between each electron-electron, nucleus-nucleus, and electron-nucleus pair.

To approach an analytic or numerical solution to this complex equation, the Born-Oppenheimer approximation is made [74]. It observes that because the mass of the nuclei, M_N, is much larger than that of the electrons, m_e, the electrons can be considered to instantaneously re-arrange upon nuclear motion. In other words, the time scale of electronic motion is so much faster than that of nuclear motion that the electrons may be treated as reacting to the potential of a set of nuclei of fixed position. Under this approximation, the nuclear kinetic energy term in 3.51 becomes zero, as the positions of the nuclei are considered fixed. The nuclear-nuclear repulsion term becomes a constant whose action inside the operator does not affect the wavefunction. What remains is the electron Hamiltonian with terms,

$$\hat{H}_{elec} = \hat{T}_e(\bar{r}_i) + \hat{V}_{Ne}(\bar{R}_K,\bar{r}_i) + \hat{V}_{ee}(\bar{r}_i). \tag{3.54}$$

The electron Hamiltonian acts on the electron wavefunction, and the resulting eigenvalues are the total electron energies. The system's total energy is then simply the product of the electron energy and the constant nuclear-nuclear repulsion at fixed nuclear coordinates. Thus the total energy depends only on $\bar{r}_i$ and parametrically on the fixed nuclear positions. By varying the nuclear positions and calculating the resultant total energy, an energy landscape or potential energy surface for the system can be constructed. The nuclear-electron interaction term in the electron Hamiltonian, $\hat{V}_{Ne}(\bar{R}_K,\bar{r}_i)$, is re-written in terms of an external potential $V_{ext}(\bar{r}_i)$,

$$\hat{V}_{Ne}(\bar{R}_K,\bar{r}_i) = \frac{1}{2}\sum_{K=1}^{N}\sum_{i=1}^{n}\frac{Z_K}{\bar{R}_{Ki}} = \sum_{i=1}^{n}V_{ext}(\bar{r}_i). \tag{3.55}$$

Because the electron Hamiltonian is hermitian, it can be shown that any approximate wavefunction, ψ, obeying the same boundary conditions as the exact wavefunction, ϕ, will result in eigenvalues, ϵ, that are larger or equal to the ground state energy, ϵ_0, of the exact wavefunction ϕ. This is known as the variational principle of quantum mechanics.

While the representation of quantum mechanics in terms of wavefunctions lends itself to pen-and-paper theory work, an alternative formulation in terms of electron density forms the basis of density functional theory. The electron density is given by:

$$\rho(\bar{r}) = n \int \Psi(\bar{r}_1, w_1, \bar{r}_2, w_2, \ldots, \bar{r}_n, w_n)\Psi^*(\bar{r}_1, w_1, \bar{r}_2, w_2, \ldots, \bar{r}_n, w_n) d\bar{r}_1 dw_1 d\bar{r}_2 dw_2 \ldots d\bar{r}_n dw_n, \tag{3.56}$$

where $\bar{r}$ are the electron position vectors and w the electron spins of n electrons. Reformulating the electronic structure problem in terms of the electron density is attractive for the development of computational approaches as the density is a function of only three variables, i.e. the three positional coordinates of $\bar{r}$. Furthermore, the electron density can be observed experimentally, unlike the electron wavefunction. Additionally, the total number of electrons is contained in the density and obtained by simple integration of the density over all space. The nuclear charges, or atomic numbers Z, can also be obtained from the slope of the local maxima in the density, which involves taking the partial derivative of the density.

Two theorems by Hohenberg and Kohn [75] lay the foundation for the formulation of density functional theory. In their first theorem, they proved that there is a one-to-one correspondence between the densities and the external potential $V_{ext}(\bar{r})$. In other words, different external potentials will always produce different electron densities. This implies that the electron density completely describes the ground state electronic properties. It follows that the electron kinetic energy, $\hat{T}_e$, and the electron-electron repulsion, $\hat{V}_{ee}$, can be written as functionals of the electron density, $\rho(\bar{r})$. Together these terms are combined into the Hohenberg-Kohn (HK) functional $F_{HK}[\rho(\bar{r})] = \langle\psi| \hat{T}_e + \hat{V}_{ee} |\psi\rangle$. A functional is a map of a function to a scalar field. The only remaining term is the nuclear-electron interaction, described by the external potential $V_{ext}(\bar{r})$. The total energy can be written entirely in terms of the electron density,

$$E = F_{HW}[\rho(\bar{r})] + \int V_{ext}(\bar{r})\rho(\bar{r})d\bar{r} = E[\rho(\bar{r})]. \tag{3.57}$$

Inside the HK-functional, which includes the electron kinetic energy and electron-electron interaction terms, the electron-electron term is further split into a classical Coulombic electron-electron repulsion term, an exchange term arising from particle spin, and an electron correlation term. The expression for the Coulombic term is known, while the electron correlation, exchange, and electron kinetic energy terms do not have known functional forms. Approximate functionals must be constructed to describe these terms.

Hohenberg and Kohn's second theorem [75] uses the variational theorem of quantum mechanics for the wavefunction to derive a variational principle in terms of the electronic density. It states that the energy functional $E[\rho(\bar{r})]$ has as its minimum value the exact ground state energy corresponding to the external potential V_{ext} under the condition that the number of particles remains constant. So given an approximate functional for $F_{HK}[\rho(\bar{r})]$, the variational principle for densities can be used to search for a density $\rho(\bar{r})$ that minimizes the energy $E[\rho(\bar{r})]$.

While Hohenberg and Kohn successfully reformulated the electronic structure problem in terms of the electronic density. However, their approach does not yet offer an advantage over traditional wavefunction approaches. The breakthrough that made density functional theory the most widely applied electronic structure method was made by Kohn and Sham [76], who realized that solving the electronic Hamiltonian for a non-interacting system of electrons is a much easier computational task. The Hamiltonian for a system of non-interacting electrons can be expressed as a sum of single electron operators whose eigenfunctions are Slater determinants of individual electron configurations, and the eigenvalues are the sum of one-electron eigenvalues.

A fictitious system of non-interacting electrons is constructed whose ground-state electron density is the same as that of some real system of interacting electrons. Because of the one-to-one correspondence of densities and external potentials (HK theorem 1), the position and the atomic number of the nuclei are identical in the fictitious and real system. For the non-interacting system, the electron kinetic energy, $T_{e,non-int}$, is just the sum of the individual kinetic energies of each electron. To account for the difference between the non-interacting and real system, a correction to the kinetic energy due to the interacting nature of the electrons, ΔT_e, and a non-classical electron-electron correlation term, ΔV_{ee}, for the electron-electron repulsion are introduced to the energy functional,

$$E[\rho(\bar{r})] = T_{e,non-int}[\rho(\bar{r})] + V_{Ne}[\rho(\bar{r})] + V_{ee}[\rho(\bar{r})] + \Delta T_e[\rho(\bar{r})] + \Delta V_{ee}[\rho(\bar{r})], \tag{3.58}$$

where $V_{Ne}[\rho(\bar{r})]$ is the nuclear-electron interaction, as described by the external potential, and $V_{ee}[\rho(\bar{r})]$ is the classical coulombic electron-electron repulsion. Because the fictitious system is made up of non-interacting electrons, the above energy functional can be expressed in terms of single electron orbitals $|\chi_i\rangle$ using the Slater determinant wavefunctions and $\rho = \sum_{i=1}^{n} \langle\chi_i|\chi_i\rangle$,

$$\begin{aligned} E[\rho(\bar{r})] = \sum_{i=1}^{n} \left(\langle\chi_i| -\frac{1}{2}\nabla_i^2 |\chi_i\rangle - \langle\chi_i| \sum_{K=1}^{N} \frac{Z_K}{|\bar{r}_i - \bar{R}_k|} |\chi_i\rangle \right) \qquad (3.59) \\ + \sum_{i=1}^{n} \langle\chi_i| \frac{1}{2} \int \frac{\rho(\bar{r}')}{|\bar{r}_i - \bar{r}'|} |\chi_i\rangle + E_{xc}[\rho(\bar{r})], \qquad (3.60) \end{aligned}$$

where the first term is the sum of individual kinetic energy expectation values for individual non-interacting electrons, the second term is the expectation value of the nuclear-electron interaction, the third term is the classical electron-electron repulsion, and the final term contains the corrections to the kinetic energy due to interacting electrons as well as a correction to the classical self-interaction. The term is called the exchange-correlation functional and represents the only unknown term in Equation 3.59. The goal now is to find orbitals $|\chi_i\rangle$ that minimize $E[\rho(\bar{r})]$ according to the eigenvalue problem,

$$\hat{h}_i^{KS} |\chi_i\rangle = \epsilon_i |\chi_i\rangle , \tag{3.61}$$

where $\hat{h}_i^{KS}$ is the Kohn-Sham one electron operator,

$$\hat{h}_i^{KS} = -\frac{1}{2}\nabla_i^2 - \sum_{K=1}^{N} \frac{Z_K}{|\bar{r}_i - \bar{R}_k|} + \frac{1}{2}\int \frac{\rho(\bar{r}')}{|\bar{r}_i - \bar{r}'|} + V_{xc}, \tag{3.62}$$

where $V_{xc} = \frac{\partial E_{xc}[\rho(\bar{r})]}{\partial \rho(\bar{r})}$ is the functional derivative with respect to the density. It functions as a one-electron operator for which the expectation value of the Slater determinant is E_{xc}. Because the electrons in the fictitious system do not interact, the system can be described by a summation over the Kohn-Sham operators,

$$\sum_{i=1}^{n} \hat{h}_i^{KS} |\chi_1, \chi_2 \dots \chi_n\rangle = \sum_{i=1}^{n} \epsilon_i |\chi_1, \chi_2 \dots \chi_n\rangle . \tag{3.63}$$

The core benefit of this approach over the initial formulation by Hohenberg and Kohn is that the energy contribution of the non-interacting electron density, which is known, is much larger than the correction terms in the exchange-correlation functional E_{xc}. In Hohenberg and Kohn, the electron kinetic energy was unknown and required an approximate functional form leading to much larger errors in predicted energies. In implementing Kohn-Sham (KS) DFT, the KS orbitals are usually expanded in some functional basis set $\{\phi\}$. The total energy in KS-DFT is a sum of functionals over KS orbitals expressed in a particular basis set for the kinetic energy of the fictitious non-interacting system and functionals of the electron density for the coulomb term, external potential, and the approximate exchange-correlation correction term. By minimizing the total energy with respect to the KS orbitals, under the constraint of maintaining the orthonormality of the orbitals, the calculated energy approaches the ground state energy of the true system of interacting electrons.

3.2.2.1 Exchange-correlation functionals

While some parts of the total energy functional used in Kohn-Sham DFT are known analytically, the so-called exchange-correlation functional must be approximated. The exact form of the exchange-correlation is not known. It must be constructed from physical considerations, limiting cases that can be solved exactly and by benchmarking against higher accuracy quantum methods or experimental results.

As described in the previous section, the exchange-correlation functional captures the correction to the kinetic energy of the non-interacting system due to electron interaction and corrections arising from the interaction of electron spins. In practice, however, the exchange-correlation functional is usually split into two terms, one for electron exchange and one for electron correlation.

The simplest functionals are derived from the exact results of a uniform electron gas (jellium) with a constant electron density. This approximation is known as the local density approximation (LDA), and the exchange functional is given by,

$$E_x^{LDA} = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{\frac{1}{3}} \int \rho^{\frac{4}{3}}(\bar{r})d\bar{r}. \tag{3.64}$$

Various correlation functionals have been proposed to combine with the LDA exchange functional. However, the approximation of a uniform electron density is not very good in the context of biomolecules.

LDA functionals can be improved by incorporating the gradient of the density, $\nabla\rho$, to better approximate non-uniform electron densities (as observed in real molecules). Functionals of this form are known as generalized gradient approximation (GGA) functionals, and PBE [77] was one of the first successful examples and is still extensively used. The PBE exchange functional is given by,

$$E_x^{PBA} = \int \rho^{\frac{4}{3}}(\bar{r}) f(\nabla\rho(\bar{r}))d\bar{r}, \tag{3.65}$$

where f is given by,

$$f = C_\alpha + \beta\left(\frac{x^2}{1+\gamma x^2}\right), \tag{3.66}$$

and $x(\bar{r}) = \frac{\nabla\rho(r)}{\rho^{\frac{4}{3}}(\bar{r})}$ is a dimensionless entity. PBE is the progenitor of functional design based on exact constraints with limited empiricism. Another widely used GGA functional is the more empirically motivated BLYP, which combines the exchange functional B88 with the correlation functional LYP [78]. BLYP relies heavily on the fitting of a range of adjustable parameters.

Beyond GGA functionals, meta-GGA (mGGA) functionals incorporate the electron kinetic energy density and the density gradient. This allows for more flexibility in the functional form, and in general, mGGAs outperform GGAs, but are more computationally intensive. The inclusion of the mGGA exchange roughly doubles the computational cost compared to the GGA functional PBE. In a benchmarking study of Head-Gordon [78], the best performing mGGA tested was a relatively new empirically based functional, B97M-rV [79], which incorporates a non-local dispersion term. This functional performs comparably to hybrid-mGGA functionals [78], which incorporate a portion of exact exchange from Hartree–Fock theory and represent the most complex and computationally demanding category of functionals.

3.2.2.2 Pseudopotentials

The computational cost of traditional density functional theory scales cubically with the total number of electrons in the system. In large systems with heavy atoms, the majority of all electrons making up the system are core electrons. However, binding and other chemical interactions are mediated mainly through the valence electrons. Thus, the explicit consideration of all core electrons significantly increases computational cost while having little impact on calculated energies. Furthermore, close to the atomic core, the potential has a steep gradient and would require a finer grid when discretized, further increasing the computational load. Pseudopotentials solve these problems by approximating the core states with a smooth analytic potential. This analytic function represents the nuclear and electronic core to the valence electrons, decreasing the total number of electrons explicitly considered and circumventing the need for finer grid spacing near the atomic core. The pseudopotentials are constructed according to several conditions for each element in the system.

3.2.2.3 Empirical dispersion corrections in density functional theory

Traditional density functionals like PBE and BLYP do not describe long-range dispersion interactions well [80]. Recall the expression for the dispersion term in classical mechanical force fields from Equation 3.9. At short inter-atomic distances, the R^{-6} dependence of dispersion interactions can be described relatively well due to overlapping electron densities. At longer ranges, however, standard XC functionals fail to describe dispersive effects [80]. To overcome this, a variety of methods have been proposed. The most popular method, DFT+D, applies a post-hoc correction term to the total DFT energy and is calculated in a pairwise additive manner from an empirical potential, similar to a molecular mechanics force field. To avoid double counting of dispersive interactions at short and medium ranges, damping functions are used to damp the dispersion correction at short inter-atomic separations. A variety of empirical parameterizations and damping functions have been developed. The D2, D3, and the newest D4 dispersion corrections by Grimme et al. [81, 82] are among the most widely used. In a benchmarking study on large host-guest systems in 2013 [83], D2/D3 performed on-par with functionals containing explicit dispersion terms and outperformed alternative approaches like one-electron correlation methods.

In addition to the pairwise dispersion corrections, a three-body dispersion correction may be applied. While negligible in small systems, Risthaus et al. [83] found that in systems of hundreds of atoms, the three body term can become significant (1-5% contribution). D3 and D4 feature an approximate three-body dispersion term, E_{abc}, which can be added on top of any pairwise dispersion correction.

While Grimme's DFT+D uses a highly-parameterized empirical force-field-like correction term, non-local correlation functionals can, in principle, describe non-local correlation effects without needing additional correction terms. This approach requires fewer parameters, is more physically motivated, and may produce more accurate results at the price of additional computational cost. One of the first successful attempts at non-local correlation functionals was VdW-DF2 in 2010 [84]. When combined with existing exchange-correlation functionals, significant improvements over traditional dispersion corrected calculations were seen. The inclusion of the non-local correlation term increases the computational cost only slightly. Also, in 2010, the non-local dispersion correlation functional VV10 was developed by Vydrov and Van Voorhis [85]. They proposed a GGA functional combining rPW86 exchange, PBE correlation, and VV10 dispersion correlation. Excellent results for non-covalently bonded systems were obtained. The numerically more efficient rVV10 has two optimization parameters that can be used to incorporate it into existing exchange-correlation functionals [86].

3.2.3 Linear-scaling DFT in ONETEP

While standard density functional theory scales cubically with the number of electrons in the system, linearly-scaling versions of DFT have been developed. The ONETEP code [87] is one such linear-scaling DFT implementation. The one-particle electron density matrix, which decays exponentially with distance, is truncated to achieve linear-scaling. Effectively, the electronic density is localized in space. In ONETEP this is achieved by representing localized orbitals by non-orthogonal generalized Wannier functions (NGWFs). Strict localization is enforced for the NGWFs. As part of the self-consistent minimization, both the orbitals and the electron density are optimized, reducing the number of required NGWFs. The parallelization strategy of ONETEP exploits this locality and can spread individual atoms over many simultaneous processes. A hybrid MPI-OMP parallelization allows for efficient and scaleable distribution of workload. MPI processes split the system into individual atoms or groups of atoms, and each spawns multiple OMP threads. The unique characteristic of ONETEP is that even though it is linear-scaling, it can retain large basis set accuracy as in conventional cubic-scaling DFT calculations [87].

To demonstrate the linear-scaling capabilities of ONETEP, single-point energy calculations at moderate settings were run for lipid bi-layers of different sizes. The lipid layers were constructed from PE DLPE lipids only. The structures were generated using the CHARM-GUI tool [88]. Each structure was equilibrated in Amber16 in explicit solvent. The single-point energy calculations in ONETEP were run with an energy cutoff of 600eV, using the PBE functional. Each calculation converged after 11 outer loop iterations. The calculations used 320 MPI processes, with 4 OMP threads each, for a total of 1280 cores on the Intel Skylake processors of the Iridis5 supercomputer. Figure 3.1 shows the total wall-time against the number of atoms in the system and a visualization of the 1000 and 50000 atom lipid bi-layers. A clear linear relationship is evident. A linear regression yields a slope of 0.62 and an intercept of 1026 seconds with a p-value of 0.0002.

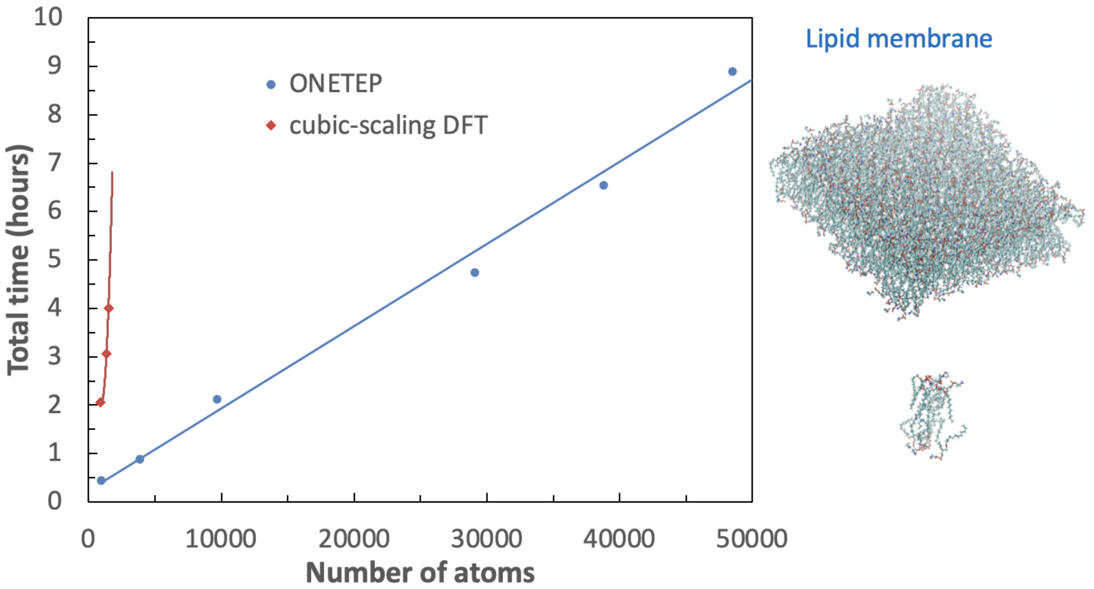


FIGURE 3.1: Linear-scaling of total computing time with the number of atoms in a lipid bi-layer system at an energy cutoff of 600eV using the PBE functional and 1280 cores (320 MPI, 4 OMP). Figure provided by C.K. Skylaris.

3.2.3.1 The solvent model in ONETEP

The ONETEP implicit solvation model is a minimal-parameter Poisson-Boltzmann (PB) based model (Section 3.1.3) which is implemented self-consistently as part of the DFT calculation [43]. The solute cavity is constructed from iso-surfaces of the electronic density, which offers a superior definition of the cavity compared to the rolling sphere approach in traditional PBSA. Using the electronic density to construct the cavity additionally reduces the number of empirical parameters. Because the PB equation is solved self-consistently as part of the DFT calculation, the implicit solvent can alter the electron density. The shape of the cavity can remain fixed to reduce the computational cost or can be updated with each iteration of the self-consistent calculation to increase accuracy. Ionic point charges, which cause singularities in finite-elements Poisson-Boltzmann solvers, are replaced by Gaussian smeared charges that improve the efficiency of the numerical solvers. A single parameter controls this so-called smeared-ion formalism.

The non-polar term of the solvation energy is calculated from the surface area of the iso-surfaces used to construct the cavity. The so-defined solvent accessible surface area

(SASA) is multiplied by the physical surface tension of the solvent, γ, and scaled by an empirical factor of 0.28 to approximately include the dispersion-repulsion between the solvent and solute. Thus the non-polar term is analogous in form to that of traditional PBSA solvation but benefits from the cavity definition based on the electron density of the solute.

3.2.4 Semi-empirical QM methods

While, in principle, free of empirical parameters, ab-initio DFT is computationally expensive. Semi-empirical QM (SEQM) methods have been developed to provide QM accuracy at a significantly lower computational cost. Usually, cost reduction is achieved by incorporating empirical approximations and fitting parameters to avoid expensive computations. Three popular SEQM methods are AM1, GFN2-XTB, and PM6. One common application of these methods is calculating ligand partial charges for use in classical mechanical MD simulations. The GFN2-XTB SEQM method used in this research is briefly introduced in the following section.

3.2.4.1 GFN2-XTB

GFN2-XTB is a semi-empirical tight-binding method developed by Grimme et al. [89]. Fundamentally, it is a more empirical and approximate version of DFT. The ground state energy is Taylor expanded around the self-consistent change in electron density and truncated after the third term. Three-electron and four-electron integrals are set to zero. The remaining three terms are further decomposed into separable energies, some of which are physically based, while others are empirical. Element-specific parameter sets were developed by optimizing the method against large benchmarking data sets. GFN2-XTB is unique because it is general purpose and does not require user-generated parameters.

3.2.5 Choice of QM method in protein-ligand binding energy prediction

There is a zoo of QM methods, varying in the degree of accuracy, theoretical rigor, and computational cost. In the context of protein-ligand binding, only relatively low-cost methods can be used as the systems under investigation have thousands of atoms. Hence, most attempts thus far have focused either on semi-empirical QM or hybrid QM/MM methods. Hybrid QM/MM is a multi-level method in which a small QM region is embedded in a classical MD simulation. In the studies outlined below, the sampling of dynamics, if performed at all, is often done at MM level due to the prohibitive cost of ab-initio QM dynamics sampling for systems of biological interest.

The most commonly applied method is the semi-empirical AM1. Fischer [90], Jamet [91] and Ibrahim [92] all applied AM1 to the ligand only. Ojha [93] used several SEQM methods, including the self-consistent-charge density-functional tight-binding method SCC-DFTB, on the ligand only and found that the different methods performed indistinguishably within the statistical error. Wang and Chen [94] also used SCC-DFTB but included the ligand and 8 active protein residues. Barboult [95] and Sippl [96] used AM1 and included the ligand and a 6 and 5 Å region around the ligand, respectively. Diaz [97] used AM1 and PM3 in a QM/MM-PBSA approach on the TEM-1 enzyme with benzylpenicillin (BP) and cephalothin (CEF). Soederhjelm [98], Zhan [99, 100], Wang [101] and Leonidas [102, 103] used DFT calculations, mostly on the ligands and surrounding sites. These studies featured either no sampling or minimal sampling from MD. Merz et al. [104] used linear-scaling SEQM methods on the whole protein and Ryde [105] used the SEQM methods AM1, RM1, and PM6 to calculate binding energies in three protein-ligand systems using a SEQM-GBSA approach. Anisimov [106] used a hybrid QM/MM method based on the COSMO solvent model. Fragmentation based approaches like PMISP [7, 107] or the fragment molecular orbital (FMO) [108] method using QM calculations have been applied to various protein-ligand systems [109–112]. In 2010, Cole [113] used linear-scaling DFT with MM sampling on an entire protein-protein complex. Fox and Skylaris applied the same method to a protein-ligand complex in 2012 and 2014 [38, 114].

In 2017, Nacimento et al. [115] used FEP together with QM/MM MD at AM1 level to study classical inhibitors (DMT, DNP, GNT, HUP, THA) with acetylcholinesterase (AChE). The study focused on understanding the binding mechanism and contributions to the binding energies. Also, in 2017, Olsson and Ryde [116] compared different approaches at a QM/MM level. They compared the effectiveness of a reference potential FEP approach with a full QM/MM FEP simulation in which the ligand was treated at the SEQM level. Grimme and Ehrlich (2017) [117] proposed a full QM approach based on the interaction energy from DFT, a solvation term from COSMO RS, and an entropy term from semi-empirical DFTB3-D3 hessian. Frush (2017) [118] used a linear interaction energy (LIE) approach with QM/MM energy evaluations on an MD trajectory. The QM region was treated with AM1. In 2016, Ryde et al. [8] tested 4 SEQM methods as part of the D3R challenge. In 2019, Giese and York [119] published a study of small molecule solvation free energies and T4-lysozyme binding free energies using QM/MM MD and force-matched reference potentials. Their approach was to correct the end state of an MM alchemical free energy calculation using a correction term obtained from QM/MM dynamics. To smooth the transition from MM to QM/MM, a MM' reference potential was created and used as an intermediate step in an extended thermodynamic cycle. The resulting free energies were well converged and exhibited low statistical errors (mean unsigned error of 0.04 ± 0.14kcal/mol).

In a review of these recent developments, Cavasotto et al. [3] conclude that QM approaches are still system dependent, too slow for industry, and ignore conformational sampling of protein flexibility. In charged systems, the solvation contributions are not sufficiently accurate and the entropy problems of MM-PBSA have not been addressed. It was found repeatedly that dispersion-correction is mandatory for any method [39]. Despite inconsistent improvements over MM results, a variety of QM and QM/MM studies have shown that QM terms not present in MM force fields do contribute significantly to the gas phase energies [39]. Some methods employ a QM/MM minimization step before evaluating snapshot energies, but no clear improvement could be determined.

3.3 The QM-PBSA method

Methodologically, QM-PBSA [1, 38, 113] is a straightforward modification of MM-PBSA. The gas phase energy $\langle E_{MM} \rangle$ and solvation energy $\langle G_{solvation} \rangle$, calculated using the molecular mechanics force field and PBSA implicit solvation scheme, are replaced by quantum mechanical gas phase and solvation energies from full-protein linear-scaling ab-initio density functional theory calculations in ONETEP [87]. The solvation energy is also estimated from an implicit solvent model, which is implemented self-consistently with minimal parameters in the DFT single-point energy evaluation and contains dispersion-repulsion effects in addition to the cavitation energy [43].

The implicit solvation term uses the QM electron density to define the solute cavity, and the solvation, in turn, changes the electron density during the self-consistent iterative DFT calculation. By extension, QM-GBSA using the simpler Generalized Born model cannot be formulated in the same manner.

The optional entropy correction term in MM-PBSA can be added to the QM-PBSA energy using either normal mode analysis or the Interaction Entropy method [62].

3.3.1 Motivating QM-PBSA

MM-PBSA, and by extension QM-PBSA, are not thermodynamically rigorous and have significant shortcomings in their formulation [12]. The single-trajectory approach, in which molecular dynamics simulations are only performed for the protein-ligand complex, ignores all potential configurational changes in the protein and ligand upon binding. By sampling only the endpoints of the binding process, the binding path of the ligand and the movement of the protein binding site to accommodate the ligand are ignored. The implicit solvent ignores the importance of structural waters and bridging waters between the protein and ligand and gives only a rough estimation of the energetic cost of displacing water molecules from the solvent-accessible binding site during ligand binding. In the context of classical mechanical binding free energy estimation, more rigorous, reliable, and accurate methods like thermodynamic integration (TI) or free energy perturbation (FEP) exist. However, because QM-PBSA and MM-PBSA sample only from the binding endpoints, significantly less sampling and expensive DFT calculations are required to obtain binding energy estimates. In contrast, alchemical free energy methods require sampling at many intermediary points, making them computationally intractable for a DFT-based approach.

In previous works, our group has attempted to extend TI, a rigorous classical mechanical method, to quantum mechanics. While the alchemical transformation was performed classically, the two endpoints of the transformation were subjected to a one-step transition from classical mechanics to quantum mechanics. Using a Monte-Carlo-like acceptance criteria [120–123], the configurations generated by classical mechanics are tested for acceptance into the true QM ensemble. Due to significant differences in the MM and QM phase space, many of the classical configurations are rejected from the QM ensemble. While this approach shows promise for small molecules, the acceptance rate of larger biomolecules becomes vanishingly small. As a result, this more rigorous approach was deemed non-feasible for large-scale biomolecular interactions like protein-ligand binding.

In contrast, the extension from MM-PBSA to QM-PBSA is methodologically trivial. The gas phase and solvation energy can be replaced one-to-one with single-point DFT calculations in implicit solvent. The availability of a PBSA-type implicit solvent in the linear-scaling DFT code ONETEP allows for a fair and direct comparison of the MM and QM binding energies. The snapshots generated for MM-PBSA and QM-PBSA are taken from unconstrained, plain MD simulations of the protein-ligand complex and require no further preparation. Thus, no constraints or biasing force field potentials need to be accounted for in the QM energy evaluations. MM-PBSA and QM-PBSA calculations are also trivially parallel, meaning that every DFT energy evaluation can be run concurrently, leading to attainable time-to-solutions of less than 5 hours on a state-of-the-art HPC system.

A central concern with the QM-PBSA approach is that snapshots generated from classical mechanical MD are evaluated using a quantum mechanical energy Hamiltonian. While ideally, sampling would be performed at a QM level of theory, this is currently not achievable due to the exorbitant computational cost of ab-initio MD. It is not clear to what extent the potential energy surface sampled at the classical mechanical level coincides with the true potential energy surface of the QM Hamiltonian, and the degree of overlap is likely system-dependent. In our formulation of QM-PBSA, we must assume that the location of the minima of the QM and MM potential energy landscape coincide. The hope is to more accurately determine the magnitude (depth) of these energy minima by evaluating the QM energies instead of the MM energies. We must also assume that the shape of the potential minima at MM and QM levels is comparable. One may expect the MM ensemble to be so different from the true QM ensemble that QM energy evaluations on MM snapshots lead to non-relevant energies. The QM landscape would have to be known to diagnose these differences explicitly. In the absence of perfect knowledge of the true QM energy landscape, the rate of convergence of calculated QM energy terms compared to their MM counterparts may serve as an interesting proxy. This question of the convergence of QM binding energies compared to MM binding energies will be addressed in our benchmarking studies of QM-PBSA in Chapters 4 and 5.

In summary, we see QM-PBSA as a stepping stone method that is simple, tractable, and enables the exploration of binding energies based on fully quantum mechanical energy evaluations today. In the following chapter, QM-PBSA is applied to the T4-lysozyme protein-ligand benchmarking system, and key aspects of the QM-PBSA method, like the convergence and choice of exchange-correlation functional, are explored.

Chapter 4

T4-Lysozyme: QM-PBSA Reproducibility, Convergence, and Exchange-Correlation Functional

4.1 Introduction

Having introduced and motivated the QM-PBSA quantum mechanical binding free energy method, this chapter focuses on validating the method in a previously studied benchmarking system and characterizing its convergence and dependence on the choice of exchange-correlation functional.

In 2014, Skylaris et al. [38] first applied the QM-PBSA method to a set of ligands binding to the T4-lysozyme protein. In this chapter, we continue the exploration of T4-lysozyme using the QM-PBSA method and present our study [1] of the reproducibility, convergence characteristics, and comparison of three exchange-correlation functionals and dispersion corrections. We first introduce the target protein system, a T4-lysozyme double mutant, in Section 4.2 followed by a description of the 2014 study by Skylaris et al. [38] upon which this study builds. Section 4.4 introduces our research aims. The design of this computational study and computational details of the methods employed are presented in Section 4.5 followed by the results and discussion in Sections 4.6 and 4.7, respectively.

4.2 The T4-lysozyme double mutant L99A/M102Q

The protein under investigation is the lysozyme of the bacteriophage T4. Bacteriophages are viruses that infect and replicate within bacteria. Two heavily studied genetically engineered mutants are L99A and L99A/M102Q, which have a non-polar and polar buried cavity, respectively [124]. The L99A/M102Q double mutant [125] with a buried polar binding site is used in this study. A visualization of the protein in complex with phenol is shown in Figure 4.1. Due to the moderate size of the protein (2600 atoms) and the relative simplicity of the buried binding site, T4-lysozyme has been used in multiple studies [46, 126] to develop and validate docking methods [125, 127], alchemical free energy methods [128, 129], and other statistical thermodynamic approaches [130]. In 2017, Villarreal et al. [131] tested their hybrid steered molecular dynamics approach on benzylacetate, 2-nitrothiophene, and benzene inside the T4-lysozyme L99A/M102Q double mutant. More recently, Cabeza de Vaca et al. [132] used the L99A mutant and seven benzene derivatives to validate an enhanced Monte-Carlo method for absolute binding free energies achieving a mean unsigned error of 1.2 kcal/mol. Cole et al. [73] derived a quantum mechanical bespoke force-field for the L99A mutant and used it to calculate binding free energies for six benzene derivatives obtaining a mean unsigned error of 0.85 kcal/mol. Niitsu et al. [133] used a replica-exchange method to distinguish binders and non-binders to the L99A mutant. In 2020, Sakae et al. [134] used the T4L99A-phenol system to validate an absolute free-energy method for systems with multiple binding poses. Mobley et al. [135] also propose T4-lysozyme as a ’simple’ benchmarking system for protein-ligand binding free energy prediction because of the large volume of experimental and computational research data available, the conformational stability of the protein and binding sites as well as the small, neutral and rigid ligands. However, they also highlight the limitation that the binding affinity range of known ligands is extremely small.

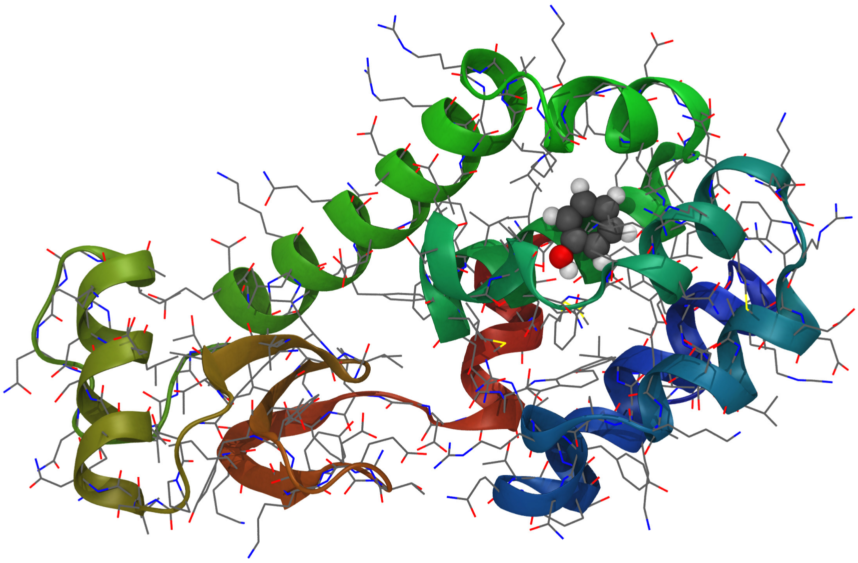


FIGURE 4.1: Phenol bound in the buried binding site of the T4-lysozyme double mutant (L99A/M102Q). Visualization made in VMD [37]. PDB Code: 1LI2

4.3 The original T4-lysozyme QM-PBSA paper

In 2014, Skylaris et al. [38] tested the QM-PBSA method using ab-initio DFT calculations on an entire protein-ligand complex. The relative binding free energies of 8 ligands to the T4-lysozyme double mutant L99A/M102Q were computed. Sampling was performed at the MM-level, using the Amber10 package [136]. A single-trajectory approach was employed, and the ff99SB [137] and GAFF force fields were used for the protein and ligand, respectively. The AM1-BCC method in Amber was used to calculate ligand charges, and the TIP3P water model [26] and neutralizing Cl^- ions were used for the explicit solvent. Only the bound catechol-protein complex was equilibrated using an extensive equilibration protocol. The other ligands were mutated from the catechol afterward. Production calculations of 20 ns were run at 300 K in the NVT ensemble using a 2 fs timestep with a Langevin thermostat and the SHAKE algorithm to constrain bonds to hydrogens. Due to its large size, the 1-phenylsemicarbazide complex structure was prepared and equilibrated separately.

MM-PBSA post-processing was employed on 1000 equally spaced snapshots, using an infinite non-bonded cutoff with a dielectric constant of 80 for the implicit water solvent and an internal dielectric of 1. Entropies were also estimated at MM level using normal mode analysis on 50 snapshots using NAB in Amber10 in the presence of an implicit solvent using the Generalized Born model.

The convergence of the MM-PBSA binding energies with the number of snapshots was used to estimate the error in the QM-PBSA approach. At 50 snapshots, the maximum fluctuations observed were below 0.5 kcal/mol. 50 snapshots were evaluated at a QM level and compared with MM-PBSA on the same snapshots for each ligand.

All QM single-point energy calculations were performed in ONETEP [87] using its minimal parameter implicit solvent model introduced in section 3.2.3.1. The DFT calculations were run at a kinetic energy cutoff of 827 eV using the PBE functional with a DFT+D style dispersion-correction. Implicit solvent parameters from a previous validation study were used. A post-hoc correction term was applied to the protein (host) to correct for the erroneous addition of the surface area of the buried binding pocket to the host cavitation energy. A similar correction was applied in the MM-PBSA approach.

Good agreement was observed between the binding free energies in vacuum at QM and MM level, indicating that the force-fields employed are well parameterized for the system. However, significant differences in the free energies of binding in solvent were observed. The QM-PBSA appeared more accurate, with an overall root mean square error (RMSE) of 2.7 kcal/mol as compared to 4.0 kcal/mol for MM-PBSA. Furthermore, the QM approach correctly predicted one of the two non-binders, whereas the MM approach did not.

4.4 Research goals

Our motivation is to improve the accuracy, transferability, and reproducibility of protein-ligand binding free energy calculations using high-accuracy ab-initio DFT. In this study of 7 ligands binding to the T4-lysozyme double mutant (L99A/M102Q), we push the boundary of DFT-based binding free energy calculations by drastically increasing the number of full-protein DFT calculations to over 2900 with the intent of:

1. Demonstrating the reproducibility of DFT-based binding free energies,
2. Studying the convergence of the QM-PBSA method,
3. Determining statistical errors at different levels of sampling,
4. Comparing the performance of dispersion-including non-local and meta-GGA-nonlocal exchange-correlation functionals VV10[85] and B97M-rV [79, 138] with the popular GGA functional PBE [77],
5. Comparing different empirical dispersion-corrections to the PBE functional and assessing the significance of the three-body dispersion term.

The choice of exchange-correlation functional may have a significant impact on accuracy. A recent benchmarking study by Head-Gordon et al. [78] showed that for a test set of binding energies of small ligands interacting with protein receptors (HSG set) the RMSDs of different functionals varied from 2.68 to 0.11 kcal/mol. Furthermore, the choice of empirical dispersion-correction, or functionals that explicitly include dispersion, was significant.

Improved accuracy in the QM-PBSA vacuum and solvation energies could be combined with advances in the entropy terms and sampling to achieve high-accuracy relative binding free energies for novel drug candidates. Furthermore, exchange-correlation (XC) functional benchmarking studies usually focus on small systems. Thus, establishing the relative accuracy of different XC functionals and dispersion-corrections in the prediction of binding energies in large-scale biological systems is necessary to inform future research utilizing DFT in this context.

An improved understanding of the behavior of the energy/entropy distributions and the rates of convergence of mean energy/entropy terms in both QM-PBSA and MM-PBSA will inform appropriate sampling and allow estimates of statistical errors. Furthermore, an in-depth comparison of the convergence of QM- v.s MM-PBSA methods will elucidate the impact of sampling with classical mechanical while evaluating energies at the quantum mechanical level of theory. To this end, the semi-empirical quantum mechanical (SEQM) method GFN2-XTB [89] is also investigated. Lastly, this study provides the opportunity to test the reproducibility of

DFT calculations within ONETEP by comparing the new results obtained to those reported by Skylaris et al. [38] in 2014.

Overall, our goal is to lay the foundation for large-scale applications of the QM-PBSA method, to comment on best practices and demonstrate that with modern computing capabilities, DFT binding free energy calculations are viable and are a promising avenue of research and industry application.

4.5 Methods

4.5.1 Design of computational study

4.5.1.1 Sampling

In this study, we re-use the conformations generated from molecular dynamics simulations in 2014 by Skylaris et al. [38]. MD sampling was performed using the one-trajectory approach and the force-field ff99SB [137] for the protein and GAFF1 for the ligand in Amber10 [34, 136, 139, 140]. 20 ns of MD were generated for each ligand from which 1000 configurations were extracted. We re-use the identical 1000 MD snapshots for 7 ligands (shown in Figure 4.2) in T4-lysozyme in this study. Skylaris et al. applied the QM-PBSA method to a subset of 50 snapshots, equally spaced within the 1000 extracted from the MD trajectories. Ligand 8, a non-binder, from the 2014 ligand set was excluded from this study. Due to its larger size, it is more prone to inducing side-chain motions in the protein upon binding [135], which are likely not captured in 20 ns of MD.

4.5.1.2 Relative binding free energies and treatment of the non-binder

MM-PBSA and related approximate methods are designed to calculate relative rather than absolute binding free energies. Normalization to the experimental binding energy of a reference ligand is needed to compare our calculated results to absolute experimental binding energies. In 2014, only phenol was considered by Skylaris et al. [38] as the reference ligand. Instead, we chose a different approach based on reviewer feedback obtained during the publication of this research. The root mean squared deviation after the removal of the systematic error (mean signed error, MSE), called RMSDtr, is used to quantify the closeness of predicted to experimental binding energies. The RMSDtr incorporates all choices of reference ligand and yields a single RMSD value instead of a separate RMSD for each choice or reference ligand, simplifying the comparison of methods.

Hydroxyaniline is a non-binder and thus does not have a well-defined experimental binding free energy. The experimental assay used by Boyce et al. [128] could identify measurable binders up to a disassociation constant of 10 milli-mol. This corresponds to a free energy of binding of -2.7 kcal/mol at 300 K, giving the lower limit of the non-binder's free energy of binding. The theoretical upper limit is 0 kcal/mol. All metrics applied to relative binding energies are calculated for the lower limits, the upper limits, and excluding the non-binder.

4.5.1.3 Exchange-correlation functionals

We selected the exchange-correlation functionals PBE, VV10, and B97M-rV for comparison in this QM-PBSA study.

PBE is a generalized gradient approximation (GGA) functional based on exact constraints and minimal empiricism. Because PBE cannot describe long-range correlation effects, an empirical force-field-like dispersion-correction is added. In this study, we test ONETEP's default dispersion-correction [87] and variants of Grimme's D2 [141] and D3 [81, 142, 143] empirical dispersion-corrections. We chose PBE because it is extremely popular and is based on physical considerations with only moderate empiricism.

VV10 was selected because it is a non-local dispersion-including GGA functional. It combines rPW86 exchange, PBE correlation, and VV10 dispersion-correlation [85]. In 2016, Womack et al. [138] implemented a more numerically efficient version, rVV10, into ONETEP.

Beyond GGA functionals, meta-GGAs (mGGA) incorporate the electron kinetic energy density and the density gradient. In a benchmarking study by Head-Gordon et al. [78] the most accurate mGGA was the relatively new empirically-parameterized functional B97M-rV [79], which incorporates rVV10 non-local dispersion.

4.5.1.4 Ligand set A, 50 snapshots: comparison of DFT functionals

To compare DFT functionals, a random subset of 5 ligands, named ligand set A (blue in Figure 4.2), was chosen due to the increased computational cost of evaluating multiple DFT exchange-correlation functionals and dispersion-corrections. The binding energies of methylphenol, fluoroaniline, catechol, hydroxyaniline, and phenol were calculated (set A). The energies of 50 equally spaced snapshots, the identical structures as in 2014 [38], were evaluated at DFT level using functionals PBE [77], VV10 [85] and B97M-rV [79].

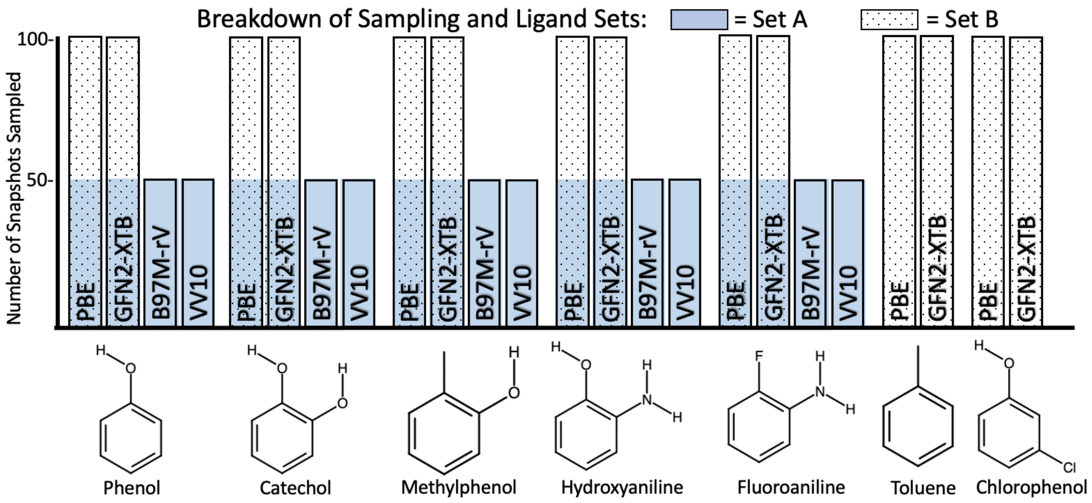


FIGURE 4.2: An overview of sampling and methods for each ligand. Ligand set A consists of the first 5 ligands and 50 snapshots of sampling with DFT functionals PBE, VV10 and B97M-rV. Ligand set B consists of all 7 ligands and 100 snapshots for PBE and GFN2-XTB. Structures drawn using Marvin JS on chem-space.com.

4.5.1.5 Ligand set B, 100 snapshots: convergence, errors and comparison of MM-, SEQM- and QM-PBSA

To investigate the convergence of the QM-PBSA method and to compare it to MM-PBSA, the ligands toluene and chlorophenol were added to ligand set A to form ligand set B (dotted area in Figure 4.2) with 7 ligands. Sampling was increased to 100 equally spaced snapshots. These 100 configurations are equally spaced within the 1000 snapshots generated by Skylaris et al. [38] and include the 50 snapshots of ligand set A. Only the PBE functional, with dispersion-corrections, was evaluated over the 100 snapshots of ligand set B. The semi-empirical tight-binding method GFN2-XTB by Grimme et al. [89] was also tested on the 100 snapshots and 7 ligands of set B.

4.5.2 Computational details

4.5.2.1 MM-PBSA

MM-PBSA post-processing was performed in Amber10 using the force field ff99SB [137] for the protein and GAFF1 for the ligand [136, 139, 140] with an infinite non-bonded cutoff. Because the choice of implicit solvent model can significantly impact the results [42], Poisson-Boltzmann solvation, available in both Amber and ONETEP, was used for consistency. A dielectric constant of 80 was used to represent the solvent water, and a dielectric constant of 1 inside the protein [42].

4.5.2.2 DFT

The linear scaling DFT code ONETEP [87] was used for energy evaluation both in this and the 2014 study by Skylaris et al. [38]. A kinetic energy cutoff of 800 eV was used for all functionals. 4 non-orthogonal generalized Wannier functions (NGWF) were used for carbon, nitrogen, and oxygen, and 1 NGWF was used for hydrogen. For sulfur and fluorine, 9 NGWFS were used. An NGWF radius of 8 atomic units was used throughout. The pseudoatomic solver was used in ONETEP to generate the initial NGWFs. ONETEP default parameters for water at room temperature were used. The default solvent surface tension is 4.7624×10^{-5} Ha/Bohr^2 with an apolar scaling factor of 0.281075 and a solvation β of 1.3. The bulk permittivity was 78.54 and an interior dielectric of 1 was used.

To speed up the solution of the Poisson-Boltzmann equation, the charge at the boundary of the simulation cell was coarse-grained. The default charge coarse-graining factor at the boundary is 5. By increasing this to 10, the energy evaluation of the complex is sped up by 20%. In a series of test calculations, increasing the coarse-graining to 10 resulted in a change of only 0.005 kcal/mol in the total energy in solvent and 0.01 kcal/mol in the binding energies. ONETEP version 5.3.2.6 compiled with the Intel 2019 compilers and Intel MPI 2019 was used. A link to the full set of DFT input and output files is provided in Appendix C.

4.5.2.3 Cavity-correction

The T4-lysozyme (LA99/M102Q) binding site is a buried cavity [125]. The ONETEP and Amber implicit-solvent models incorrectly describe the solvent-accessible surface area (SASA) of buried cavities. Cavity-correction terms appropriate to QM and MM are applied to alleviate this issue. All QM- and MM-PBSA results, therefore, are cavity-corrected.

Both the minimal-parameter PBSA solvent model implemented in ONETEP and the PBSA solvent used at MM-level incorrectly handle the buried cavity in T4-lysozyme (L99A/M102Q). This is a known issue for solvent models based on the solvent accessible surface area described in detail in 2010 by Genheden et al. [46] and in 2014 by Skylaris et al. [38]. In the un-complexed protein calculation, i.e., the host, the surface area of the interior of the buried binding site is counted towards the solvent accessible surface area (SASA) used to calculate the non-polar solvation term. Thus, the non-polar term of the protein is larger than that of the complex. This implies the formation of a larger cavity in the solvent. Conceptually, the SASA model creates an additional, fictitious cavity in the solvent with the SASA of the buried binding site. Because the non-polar terms of both the protein and complex are known, a post-hoc cavity-correction may be applied to remove the additional contribution of the buried cavity to the non-polar solvation energy. The spurious cavitation energy is removed by subtracting the difference in the non-polar terms of the host and complex from the host solvation energy. If the cavity should be filled with water, as is the case for this T4-lysozyme mutant (L99A/M102Q) [144], this correction is applied a second time to remove the cavity SASA from the host SASA. This is necessary because a water-filled cavity reduces the volume of the cavity formed for the solute in the solvent.

This simple approach works immediately for the MM-PBSA calculations, which use a simplistic SASA model with no estimate of the dispersion and repulsion between the solvent and solute (npopt=1 in Amber10). In the QM calculations, dispersion/repulsion is included in the non-polar term. Because the cavity SASA does represent a physical boundary between solvent and solute inside the buried cavity, the dispersion-repulsion due to the cavity SASA term is physically correct. Since the non-polar dispersion in ONETEP is just a scaling factor applied to the non-polar term, it can be separated out and only the excess cavitation energy removed from the host non-polar solvation term. A full derivation is provided in [38].

4.5.2.4 Dispersion

The 2014 binding free energy calculations of Skylaris et al. [38] utilized PBE with a DFT+D style dispersion-correction, based on a damping function by Elstner [145] from 2001 and ONETEP's own parameterization. In this study, Grimme's D2 [141] and variants of the newer D3 dispersion-correction, including a three-body dispersion term (E_{ABC}), are applied to the PBE functional [81, 142, 143]. The dispersion energies for D2 and D3 variants are obtained from Grimme's standalone dftd3 program and manually applied to the DFT(PBE) energies. A link to the full set of dftd3 input and output files is provided in Appendix C.

4.5.2.5 GFN2-XTB

The semi-empirical tight-binding method GFN2-XTB [89] features atom-specific parameterization for most of the periodic table, a self-consistent implicit-solvent model, and the D4 dispersion-correction [82]. It is implemented in the XTB package developed by Grimme et al. [146, 147]. The default settings for GFN2-XTB single-point implicit solvent (water) energy evaluations were used. GFN2-XTB uses a GB solvent model [45] in which the cavitation and dispersion energy is treated with a single parameter, multiplied by the SASA. Thus, the straightforward application of a QM-style cavity-correction is not possible. Results with and without an MM-style cavity-correction are considered. A link to the full set of GFN2-XTB input and output files is provided in Appendix C.

4.5.2.6 Entropy

The entropic contribution to binding was calculated using normal mode analysis (NMA) [148] in the NAB program in Amber16 [136, 149]. The complex, host, and ligand's vibrational, translational, and rotational entropies were evaluated. Before NMA, a two-part energy minimization comprised of a conjugate gradient method, followed by the Newton-Raphson method, was performed on each snapshot with tight convergence criteria using the ff99SB and GAFF1 force fields. A Hawkins, Cramer, Truhlar (HCT) Generalized Born implicit solvent, with an internal dielectric of 1, was used for the frequency calculations and the energy minimizations with infinite non-bonded cutoff. All 1000 available structures for each ligand were evaluated.

4.6 Results

4.6.1 Reproducibility

TABLE 4.1: Total change in enthalpy upon binding, $\langle \Delta H_{bind} \rangle$, in kcal/mol relative to phenol now and in 2014 [38] over the same 50 snapshots. No entropy correction included.

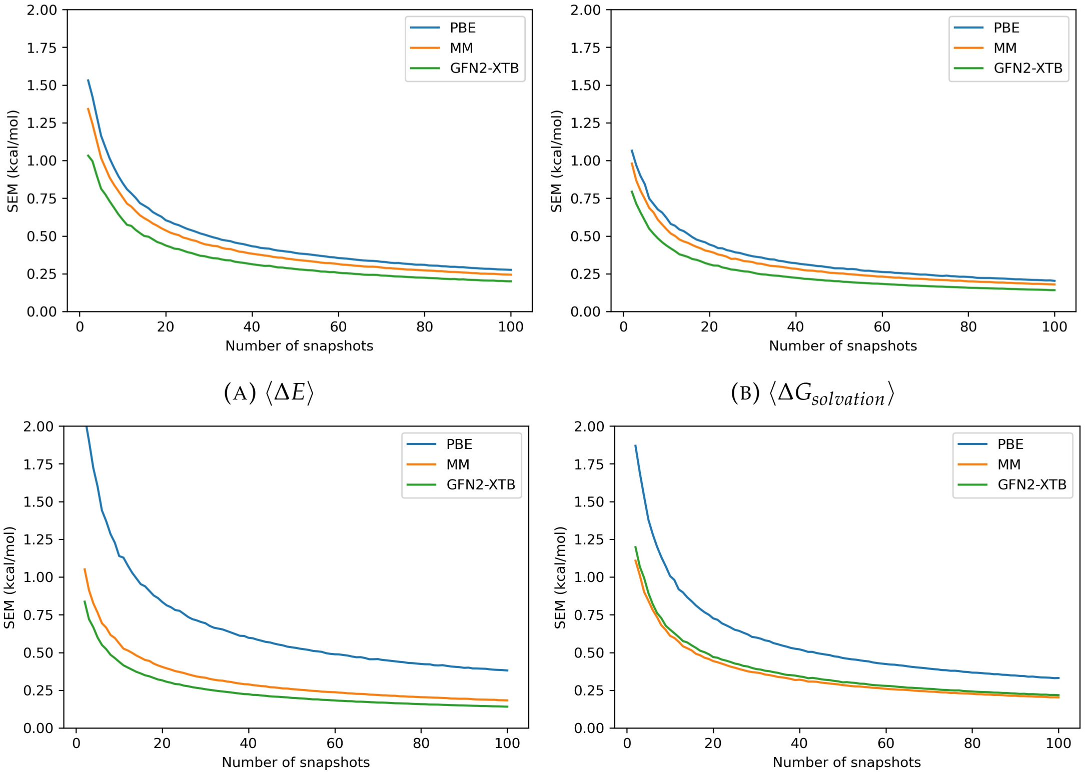
Ligands	PBE	PBE - 2014	Delta
catechol	-8.9	-9.0	0.14
fluoroaniline	-6.0	-5.9	-0.12
hydroxyaniline	-6.2	-6.2	0.00
methylphenol	-8.7	-8.5	-0.16
toluene	-5.0	-4.8	-0.18
chlorophenol	-6.9	-6.9	0.01
		Absolute Mean	0.10

This study provides a valuable opportunity to demonstrate the reproducibility of DFT-based binding free energy calculations. The calculations by Skylaris et al.[38] were performed with a 2012 version (3.1.15.2) of the ONETEP[87] code, while this study uses version 5.3.2.6 from late 2019. As shown in Table 4.1, despite 7 years of active code development, the average absolute difference between the new and old results using the same structures and functional is 0.1 kcal/mol. This underlines the robustness and precision of the DFT methodology and the ONETEP code in particular.

4.6.2 Convergence

4.6.2.1 Standard error of the mean

The standard error of the mean (SEM) measures how far a sample's mean is likely to deviate from the true population mean. The QM-PBSA method averages energy terms



(A) $\langle \Delta E \rangle$

(B) $\langle \Delta G_{solvation} \rangle$

(C) $\langle \Delta G_{solvation-cav-cor} \rangle$

(D) $\langle \Delta H_{bind} \rangle$

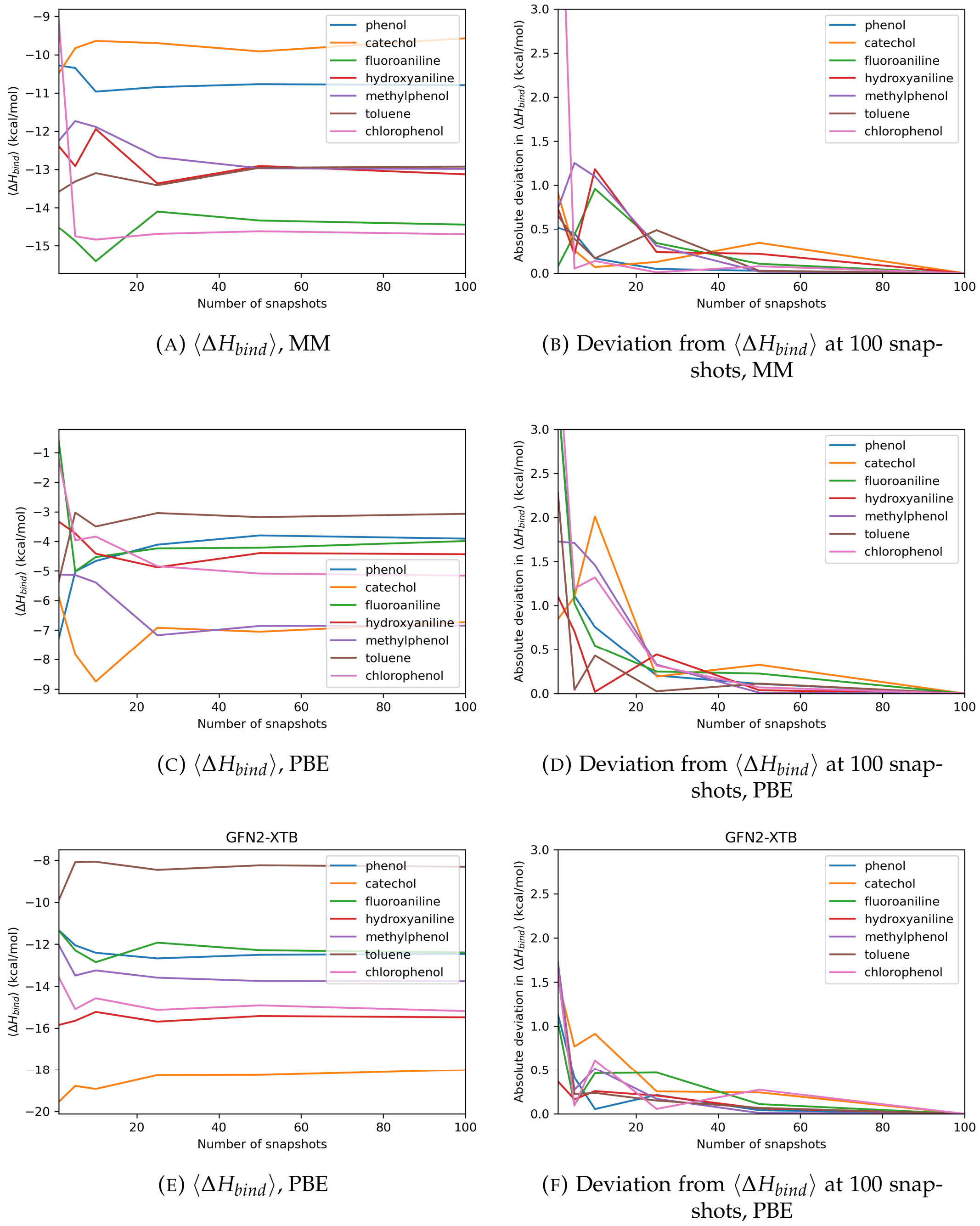
FIGURE 4.3: The standard error of the mean (SEM) calculated by bootstrapping (1000 re-samples) of the change upon binding in the gas-phase energy, $\langle \Delta E \rangle$, solvation energy, $\langle \Delta G_{solvation} \rangle$, cavity-corrected solvation energy, $\langle \Delta G_{solvation-cav-cor} \rangle$, and total enthalpy, $\langle \Delta H_{bind} \rangle = \langle \Delta E \rangle + \langle \Delta G_{solvation} \rangle$, for catechol up to 100 snapshots for MM, the DFT functional PBE and the SEQM method GFN2-XTB .

over an ensemble of snapshots (population sample). Thus, the SEM estimates how the calculated energies differ from the true, fully sampled energies (i.e. the population mean). The SEM assumes normality of the energy distributions. Using the Shapiro-Wilks test [150] we concluded that overall, the use of the SEM is appropriate.

Figure 4.3 shows the SEM convergence, calculated by bootstrapping, of each enthalpy component and the total enthalpy for catechol in T4-lysozyme. The enthalpies shown are net enthalpies. The SEM of the net gas-phase enthalpy, $\langle \Delta E \rangle$, and the solvation energy, $\langle \Delta G_{solvation} \rangle$, in catechol is only slightly higher for PBE than for the other methods. However, the SEM of PBE in the cavity-corrected solvation energy, $\langle \Delta G_{solvation-cav-cor} \rangle$, is significantly higher. At 100 snapshots, the SEM of the cavity-corrected solvation energy is 0.39 kcal/mol, while the other methods have SEMs below 0.18 kcal/mol. This leads to the overall higher SEM in the total enthalpy change upon binding, $\langle \Delta H_{bind} \rangle$, for the PBE method. The higher SEM of PBE is also reflected in the SEM of functionals VV10 and B97M-rV over ligand set A.

Unlike for PBE, the MM-style cavity-correction term applied to the MM-PBSA results (labeled MM) only minimally increases the solvation energy SEM. GFN2-XTB is shown without cavity-correction and has a similar SEM to MM. The above observations are consistent for all ligands (Figures A.1 to A.7 in Appendix A).

4.6.2.2 Absolute deviations



(A) $\langle \Delta H_{bind} \rangle$, MM

(B) Deviation from $\langle \Delta H_{bind} \rangle$ at 100 snapshots, MM

(C) $\langle \Delta H_{bind} \rangle$, PBE

(D) Deviation from $\langle \Delta H_{bind} \rangle$ at 100 snapshots, PBE

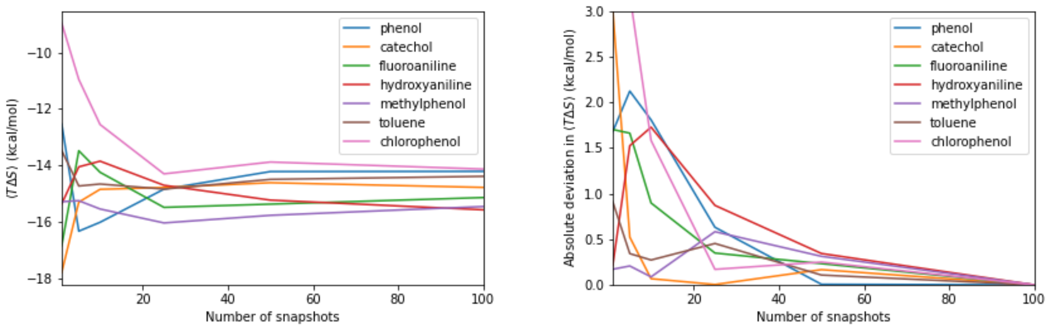
(E) $\langle \Delta H_{bind} \rangle$, PBE

(F) Deviation from $\langle \Delta H_{bind} \rangle$ at 100 snapshots, PBE

FIGURE 4.4: Left: Mean change in total enthalpy upon binding,$\langle \Delta H_{bind} \rangle$, of each ligand at different numbers of equally spaced snapshots. Right: Absolute deviation of $\langle \Delta H_{bind} \rangle$ at different numbers of equally spaced snapshots from the 'converged' mean over 100 snapshots. Methods: MM (a,b), DFT(PBE) (c,d), and GFN2-XTB(e,f).

Figure 4.4 shows the convergence of the mean change in enthalpy upon binding and absolute deviation from $\langle \Delta H_{bind} \rangle$ at 100 snapshots for MM, PBE and GFN2-XTB. Considering first the mean binding energies, all methods appear surprisingly stable between 25, 50, and 100 snapshots. Especially PBE and GFN2-XTB show only small changes in $\langle \Delta H_{bind} \rangle$ from 50 to 100 snapshots. The absolute deviation plots show that $\langle \Delta H_{bind} \rangle$ fluctuates considerably ($\approx$ 1kcal/mol) for all methods below 25 snapshots. This is most pronounced in PBE and is in line with the observation of a larger SEM in PBE. MM and GFN2-XTB show comparable levels of fluctuation. Interestingly, for 25 snapshots and beyond, the absolute deviations from the 'converged' results at 100 snapshots vary very little and are indistinguishable for PBE, MM, and GFN2-XTB. No deviations above 0.5 kcal/mol are observed beyond 25 snapshots. Additional analysis using sets of randomly selected snapshots confirmed that beyond 25 snapshots the convergence, with respect to $\langle \Delta H_{bind} \rangle$ at 100 snapshots, of MM and PBE is indistinguishable (Figures A.10 to A.16 in Appendix A).

4.6.3 Entropy correction



(A) (B)

FIGURE 4.5: (a) Convergence of mean entropy change upon binding, $T\Delta\langle S \rangle$, over 100 equally spaced snapshots. (b) Absolute deviation of $T\Delta\langle S \rangle$ from "converged" value at 100 snapshots at different numbers of equally spaced snapshots.

The entropy term in QM- and MM-PBSA is calculated by normal mode analysis, as detailed in the methods section. The maximum SEM at low numbers of snapshots is lower than for the enthalpic components, especially the DFT cavity-corrected solvation energies. However, the rate of convergence is also slower. The entropy SEM at 100 snapshots is larger than that of the total enthalpy change upon binding calculated with MM and GFN2-XTB, and is comparable to that of PBE.

Figure 4.5 shows a similar analysis for entropy as done for the enthalpic terms. Panels 4.5a and 4.5b show the convergence of the mean net-entropy term and absolute deviation from the mean net-entropy term at 100 snapshots. There are significant fluctuations below 50 snapshots ($>$ 1 kcal/mol). Fluctuations of about 0.5 kcal/mol remain even beyond 50 snapshots, and, compared to the enthalpic terms, the entropy term appears qualitatively slower in convergence.

The degree of entropy sampling significantly changes the RMSDtr of calculated against experimental relative binding free energies. Figure 4.6 shows the RMSDtr over ligand set B at 100 enthalpy snapshots and increasing levels of entropy sampling. Including a small number of entropy snapshots (5,10,25) increases the RMSDtr by up to 1.3 kcal/mol for MM-PBSA and 0.4 kcal/mol for QM-PBSA. At 50 entropy snapshots and beyond, the RMSDtr decreases compared to no entropy correction. The lowest RMSDtr is reached at 100 snapshots of entropy, i.e., the same level of sampling as for the enthalpy terms. Beyond this, the sampling of snapshots not included for calculating the enthalpy terms does not further improve the accuracy against experiment. All three treatments of the non-binder exhibit the increased RMSDtr at low levels of entropy sampling (Figures A.17 to A.19 in Appendix A).

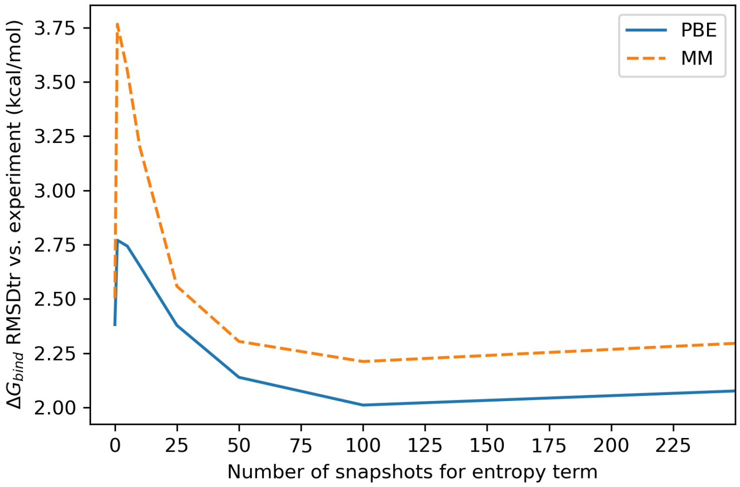


FIGURE 4.6: Root mean square deviation from experiment after removal of mean signed error (kcal/mol) of calculated binding free energies for ligand set B, at different levels of entropy sampling. Enthalpy sampled over 100 snapshots. RMSDtr calculated with upper limit for non-binder (lower limit and binders only in Figures A.17 to A.19 in Appendix A).

4.6.4 Statistical error due to incomplete sampling

The calculated absolute binding free energies are the sum of two separate means, the mean enthalpy and mean entropy, sampled over a selection of protein-ligand conformations, i.e., snapshots. By propagation of errors, the SEM of each ligand's entropy and enthalpy terms and the chosen reference ligand are combined to estimate the total statistical error due to imperfect sampling in the relative binding free energies.

Table 4.2 shows the maximum statistical error due to incomplete sampling across all choices of reference ligands for each method at different levels of sampling. Enthalpy and entropy terms are evaluated over the same 10, 25, 50, and 100 equally spaced snapshots. The maximum statistical errors for PBE are almost identical to those for VV10 and B97M-rV; hence only PBE is shown. We use these values to estimate the uncertainty in our calculated binding free energies going forward. This statistical error or uncertainty, due to imperfect sampling, should not be confused with the RMSDtr of the calculated against experimental binding free energies, used to quantify the closeness of predicted to experimental results.

TABLE 4.2: Maximum statistical errors due to imperfect sampling (SEM) in entropy-corrected relative binding free energies with different methods sampled over 10, 25, 50, and 100 equally spaced snapshots. Enthalpy and entropy terms sampled over same snapshots. Cavity-correction applied for PBE and MM.

	Max Statistical Errors (kcal/mol)			
Snapshots/Methods	10	25	50	100
PBE	1.88	1.22	0.87	0.62
MM	1.63	1.05	0.76	0.54
GFN2-XTB	1.57	1.02	0.73	0.52

4.6.5 Ligand set A: comparing DFT functionals

TABLE 4.3: Root mean square deviation from experimental binding free energies after removal of the mean signed error (kcal/mol) for ligand set A with energies and entropies sampled over the same 50 snapshots. RMSDtr values shown with the non-binder's energy set to 0 kcal/mol (upper limit,[⋆]), -2.7 kcal/mol (lower limit,[‡]) and without the non-binder. The average standard error (SE) in RMSDtr calculated using bootstrapping (10000 re-samples).

	RMSD after removal of systematic error (MSE)		
Method	all ligands[⋆]	all ligands[‡]	binders only
B97M-V	2.49	1.86	1.93
VV10	2.68	2.11	2.21
PBE+ONETEP Disp	2.26	1.76	1.93
PBE+D2	2.25	1.71	1.84
PBE+D3(BJ)	2.25	1.79	1.97
PBE+D3(BJ)+ABC	2.17	1.73	1.91
PBE+D3(BJM)	2.22	1.77	1.95
PBE+D3(BJM)+ABC	2.14	1.71	1.90
Average SE	0.26	0.27	0.30

Table 4.3 shows the root mean square deviation after removal of the mean signed error of the calculated relative binding free energies with respect to experimental binding free energies [125, 128] of ligand set A using 50 enthalpy and 50 entropy snapshots. The RMSDtr is shown for all treatments of the non-binder, and the average standard error (SE) is estimated by bootstrapping with 10000 re-samples.

Overall, VV10 is the worst-performing exchange-correlation functional and has a consistently higher RMSDtr, irrespective of the treatment of the non-binder. The PBE+dispersion methods have slightly lower RMSDtr than B97M-rV when the non-binder is included via the upper or lower bound. However, given the estimated standard error, this difference is likely not significant. For the subset of binders only, B97M-rV and PBE+dispersion methods achieve the same RMSDtr. All the empirical dispersion corrections to PBE perform well but are indistinguishable given the standard error. Including the three-body dispersion term (ABC) always slightly reduces RMSDtr, but the change is much smaller than the standard error.

4.6.6 Ligand set B: PBE, GFN2-XTB and MM

TABLE 4.4: Root mean square deviation from experimental binding free energies after removal of the mean signed error (kcal/mol) for ligand set B with energies and entropies sampled over the same 100 snapshots. RMSDtr values shown with the non-binder's energy set to 0 kcal/mol (upper limit,[⋆]), -2.7 kcal/mol (lower limit,[‡]) and without the non-binder. Average standard error (SE) in RMSDtr calculated using bootstrapping (10000 resamples).

	RMSD after removal of systematic error (MSE)		
Method	all ligands[⋆]	all ligands[‡]	binders only
MM	2.21	1.61	1.55
PBE+ONETEP Disp	2.01	1.57	1.65
PBE+D2	2.09	1.63	1.69
PBE+D3(BJ)	2.11	1.73	1.84
PBE+D3(BJ)+ABC	2.03	1.66	1.77
PBE+D3(BJM)	2.11	1.74	1.85
PBE+D3(BJM)+ABC	2.02	1.67	1.79
GFN2-XTB	3.65	3.16	3.12
Average SE	0.15	0.16	0.17

We now compare the accuracy vs. experiment of MM, PBE+dispersion, and GFN2-XTB on ligand set B at 100 snapshots of enthalpy and entropy. Table 4.4 shows the RMSDtr of the calculated relative binding free energies against experiment[125, 128]. The RMSDtr is shown for all treatments of the non-binder, and the average standard error (SE) is estimated by bootstrapping with 10000 resamples.

Overall, the accuracy against experiment as described by the RMSDtr is comparable for the MM- and QM-PBSA approach. Only the SEQM-PBSA approach using the GFN2-XTB energy method performs significantly worse. The different empirical dispersion corrections are indistinguishable given the standard error. As in ligand set A, the three-body dispersion term slightly reduces the RMSDtr of both the PBE+D3(BJ) and PBE+D3(BJM) methods, but this change is within the estimated standard errors.

4.6.7 Comment on correlation

Correlation to experiment is not included as a quality metric for two main reasons. First, the ligand set is very small. Second, the experimental binding free energies range is only 1.4 kcal/mol, and multiple ligands have identical or near identical experimental energies. The estimated statistical error in the computed relative binding free energies due to incomplete sampling is 0.87 kcal/mol at 50 snapshots and 0.62 kcal/mol at 100 snapshots for PBE. Hence, the correlation values vary greatly depending on the choice of reference ligand and the treatment of the non-binder. Furthermore, the 90% confidence intervals calculated by bootstrapping for Pearson r-values exhibit huge ranges of r-value, often above 0.5. Thus, no meaningful comparison between methods is possible. We conclude that a larger number of ligands and a greater range of experimental binding free energies are critical requirements for future protein-ligand system selection.

4.7 Discussion

4.7.1 Computational cost

Gathering the results for this study posed a serious computational challenge. Excluding initial testing and exploratory work, 3600 ab-initio DFT calculations were completed, 2900 of which were on the entire 2600-atom T4-lysozyme. This was made possible by 1) the linear-scaling of the ONETEP DFT code, 2) the efficient hybrid MPI-OMP parallelization of the ONETEP code, and 3) access to three different HPC centers. Running calculations concurrently on three HPC facilities for 6 months, the DFT calculations alone required more than 1 million core-hours. Taking into account the 21000 normal mode calculations, 15000 empirical dispersion calculations, and 2100 SEQM calculations we estimate a total wall-time of about 30000 hours or 1250 days. With full access to the 5,632 compute nodes of the tier-0 EU HPC facility HAWK, the calculations for this study could have been completed in less than 24 hours.

4.7.2 Convergence and errors

One criticism of QM-PBSA and related methods is that sampling and energy evaluations are performed using different energy functions [8]. We expected this to lead to poor convergence of the QM energy terms compared to the MM energies. While the higher SEM for DFT methods, as compared to MM, initially indicated this to be true, the source of the higher SEM is predominantly the QM cavity-correction.

Further investigation showed that the QM non-polar solvation terms calculated in ONETEP have a larger variance than the MM non-polar terms. This is exacerbated by the functional form of the QM cavity-correction, which combines the host and complex non-polar terms and then scales the result by a factor of 7.116 [38]. The larger variance in the QM non-polar terms on the total binding free energies is thus magnified, leading to a larger overall SEM. The cavity-corrected solvation energy standard deviation for PBE ranges from 2.7 to 3.9 kcal/mol in ligand set B, while the range for MM is 1.3 to 1.9 kcal/mol. One possible reason for the larger variance in the DFT non-polar term is the more complex definition of the binding cavity via electron-density iso-surfaces. GN2-XTB has similar or lower standard deviations than MM.

Detailed analysis of the convergence of the total enthalpy change upon binding of each ligand showed that beyond 25 snapshots, the QM results appear equally converged as the MM results. Analysis of the absolute deviations from the ″converged″ enthalpy change upon binding at 100 snapshots using different numbers of equally spaced and randomly selected snapshots also confirmed this. At low numbers of snapshots (< 25), the DFT energies fluctuated significantly more than those from MM, as reflected by the QM methods' higher SEM. The SEQM GFN2-XTB showed similar convergence to MM, even below 25 snapshots. This is likely because the method does not suffer the increased SEM due to the QM cavity-correction.

In terms of precision, the maximum estimated statistical error for PBE in Table 4.2 is only 0.08 kcal/mol higher than for MM at 100 snapshots and 0.11 kcal/mol higher at 50 snapshots. This further suggests that at and beyond 50 snapshots, the convergence and precision of MM and DFT methods are comparable. The key finding is that in this system, the QM-PBSA method (irrespective of the choice of functional) does not suffer from poorer convergence compared to MM-PBSA.

In this study, snapshots generated from a single MD simulation were used. Sampling snapshots from independent MD simulations may result in larger standard errors of the mean and may impact the rate of convergence of both the MM and QM methods.

These results indicate that the MM force fields (GAFF and ff99SB) used for sampling are well parameterized for this system and produce configurations that overlap well with the true QM ensemble. As a result, the QM calculations converge quickly because no high QM-energy configurations are present in the MM ensemble. In terms of the potential energy landscape, this would mean that the position of energy minima in the MM and QM representation are very similar. The difference in calculated binding free energies results from the different depths and shapes of these energy minima for the different energy functions.

4.7.3 QM-PBSA: improvements and recommendations

4.7.3.1 Choice of DFT functional

Between the three DFT exchange-correlation functionals tested over 50 snapshots on ligand set A, PBE+dispersion is the most promising choice, as it outperforms VV10 in terms of RMSDtr and has very similar RMSDtr to B97M-rV, which is computationally twice as expensive. All the empirical dispersion corrections to PBE perform well but, given the estimated standard error, have indistinguishable RMSDtr. The similar performance of the D2 and D3 empirical dispersion-corrections in this large dispersion-dominated system supports the findings by Risthaus et al. [83] in their 2013 DFT + dispersion benchmarking study.

In the same study, Risthaus et al. [83] found the D3 three-body dispersion term to contribute 2.3% to 14.6% in large, dispersion-dominated systems. We have confirmed that the three-body dispersion term is significant (about 10% of total dispersion) in protein-ligand systems of this size and tends to improve RMSDtr slightly, however, well within the estimated standard error. Given the size of the three-body dispersion term and its tendency to reduce the RMSDtr, we recommend using the D3 empirical dispersion correction due to its ability to include the three-body dispersion term.

Why do the newer and more computationally expensive VV10 and B97M-rV, which explicitly account for dispersion, not improve upon the PBE functional with empirical dispersion in this QM-PBSA study of a large dispersion-dominated system? Application of the non-parametric Kolmogorov-Smirnov test for equality of two one-dimensional distributions showed that the distributions of non-cavity-corrected absolute solvation energies of PBE, VV10, and B97M-rV are very similar. This echoes our past experience using the ONETEP solvent model that showed the solvation energies to be independent of the choice of DFT functional. Based on the Kolmogorov-Smirnov test, the gas-phase energy distributions, which include the dispersion energy, are dissimilar. Both VV10 and B97M-rV use the non-local rVV10 dispersion term. While the dispersion term rVV10 results in different gas-phase energies than the empirical dispersion corrections to PBE, this does not translate to improved accuracy against experiment compared to PBE+dispersion in this QM-PBSA study.

Furthermore, the rVV10 non-local dispersion term cannot describe three-body effects, which we found to be significant using the empirical D3-ABC method. Risthaus et al. [83] suggested that the three-body dispersion term from D3 could be added to dispersion including functionals in a post-hoc fashion; however, this was not tested here.

B97M-rV is a meta-GGA functional and thus almost twice as computationally intensive as the GGA functionals PBE and VV10. The data-set used to design and test B97M-rV consisted almost entirely of small molecules [79]. The only protein-ligand system was a 1686 atom HIV-protease/indinavir complex split into 21 interaction fragment pair structures. In the 2017 benchmarking study by Head-Gordon et al. [78], B97M-rV was the most accurate meta-GGA. However, of the almost 5000 data points tested, there were only 21 protein-ligand fragments (same as above) and 12 protein-DNA complexes with a maximum size of 58 atoms. While we can only comment on the suitability of the functionals to the QM-PBSA method and not their overall accuracy, it is interesting that B97M-rV produced worse or comparable results to PBE+dispersion. A potential explanation is that the accuracy of exchange-correlation functionals on small molecule test sets is not indicative of their applicability to larger systems. This study demonstrates that moving up "Jacob's ladder" of functional complexity does not guarantee improved results.

4.7.3.2 Inclusion of entropy term

Including an entropy correction term appears to decrease the quality of results when insufficient entropy sampling is performed. For this system, sampling the normal mode entropy term over less than 25 snapshots was inadequate, and no entropy sampling should be preferred over poor entropy sampling. This is intuitive as insufficient sampled entropy terms introduce a large statistical error ($>$ 1 kcal/mol) into the binding energies. On the other hand, when the entropy is sampled with more than 25 snapshots, the inclusion of an entropy correction reduces RMSDtr. We found that the best results were obtained when sampling entropy over the same 50 or 100 snapshots used for enthalpy sampling. Sampling beyond this slightly increased errors, possibly due to the sampling of conformations not included in the enthalpy terms. Based on these findings, we are concerned about the use of less than 50 NMA calculations in some applications of MM-PBSA [62, 151, 152].

4.7.3.3 Sampling and statistical error

Based on this study, we recommend QM energy sampling at 50 snapshots. Sampling at 100 snapshots of enthalpy and entropy reduces the maximum estimated statistical error due to imperfect sampling from 0.87 kcal/mol to 0.62 kcal/mol. However, it does not, in this system, significantly reduce RMSDtr. Very stable total enthalpies are observed between 50 and 100 snapshots, and the absolute deviation of the change in total enthalpy upon binding at 50 and 100 snapshots is lower than 0.5 kcal/mol for all ligands.

4.7.3.4 QM- vs MM-PBSA

The extent to which PBE+dispersion can consistently improve MM results can not be clearly stated. However, the results indicate that in this system, the QM-PBSA approach produces relative binding free energies with RMSDtr comparable to MM. None of the methods tested were able to identify hydroxyaniline as a non-binder.

Given the small range of binding energies in the ligand set, the null hypothesis of assigning each ligand the same binding energy yields relatively low errors against experiment. The null hypothesis has an RMSDtr of 0.57 kcal/mol if the non-binder is excluded and 1.00 kcal/mol and 1.87 kcal/mol when the non-binder is included via the lower and upper bound, respectively. Fundamentally, more ligands with a broader range of binding free energies are needed to compare MM- and QM-PBSA.

In this study on the T4-lysozyme double mutant (L99A/M102Q), our linear-scaling DFT-based QM-PBSA method achieves an RMSDtr of about 1.7 kcal/mol across the 6 binders, and MM-PBSA achieves an RMSDtr of 1.6 kcal/mol. We briefly outline the results of some other QM- and SEQM-PB(GB)SA studies to place our results into context. For a more in-depth review of QM based binding free energy calculations, we recommend the review by Ryde and Söderhjelm [8]. In 2011, Anisimov et al. [106, 153] applied a SEQM-PBSA style method using the PM3 Hamiltonian and a COSMO solvation model to 5 ligands binding to the LcK SH2 domain and 4 binders to BACE1. They achieved MAD of 0.7 kcal/mol and 1.7 kcal/mol, respectively. In both cases, the SEQM approach was more accurate than MM-PBSA. In 2012, Mikulskis et al. [105] tested a SEQM-GBSA approach with the AMI, RM1, and PM6 Hamiltonians on three protein-ligand systems. The overall best-performing energy function, AM1, achieved a MADtr of 1.8-12.0 kcal/mol in avidin, 1.1-1.3 kcal/mol in fXa, and 0.3-4.9 kcal/mol in ferritin, depending on the details of the hydrogen bond correction and choice of dispersion correction. Only in the ferritin system was the best SEQM-GBSA method able to outperform MM-GBSA and MM-PBSA convincingly. In 2010, Söderhjelm et al. [107] used the PMISP (polarizable multipole interactions with supermolecular pairs) approach on 7 biotin analogs binding to avidin. They achieved a MADtr of 4.5 kcal/mol, and the QM approach performed worse than MM-PBSA (3.3 kcal/mol).

One of the key motivations to extend binding free energy calculations to the quantum mechanical level is that the QM energy evaluations can, in principle, describe a broader range of physics. In this ligand set and binding site, however, the MM force-field is not challenged by high charges, large polarization, charge transfer, or similar phenomena that are not well described in traditional empirical force fields. This may, in part, explain the similar accuracy of MM- and QM-PBSA in this system. Going forward, efforts should focus on protein-ligand systems that explicitly challenge traditional force fields and where the more involved, quantum mechanical description may be necessary.

4.7.3.5 GFN2-XTB

Both in ligands set A and B, and irrespective of the treatment of the non-binder, GFN2-XTB has the highest RMSDtr. This may not be surprising as GFN2-XTB is relatively new, semi-empirical, general-purpose, and more than 100 times faster than DFT. As for the B97M-rV functional, the GFN2-XTB method was developed based on small molecule data sets and aimed at systems of roughly 1000 atoms [89]. Lastly, the differences in PBSA solvation in the DFT and MM approaches and GBSA solvation in GFN-XTB may have also contributed to the gap in performance [41].

4.8 Conclusions

In this study, we have shown that thousands of ab-initio DFT calculations of full protein-ligand systems are computationally feasible in the context of protein-ligand binding studies for drug design applications. In testing the exchange-correlation functionals PBE, VV10, and B97M-rV we find that the computationally cheapest functional, PBE, is the most promising candidate for the application of the QM-PBSA method. Our findings highlight that benchmarking studies focused almost entirely on small systems may not be representative of the performance of the functionals in a QM-PBSA approach applied to much larger systems (2600 atoms in our case). Different empirical dispersion-corrections to PBE all perform well, but their accuracy against experiment are all within the estimated standard error. The D3 three-body dispersion term is significant in size ($\approx$ 10%) and tends to improve results slightly.

By expanding the QM calculations to 100 snapshots for the PBE functional, we can show that sampling at 50 snapshots is likely sufficient for convergence. While going beyond 50 snapshots reduces statistical error, no improvement in predicted against experimental binding energies is observed. Furthermore, the QM-PBSA and MM-PBSA methods exhibit near indistinguishable convergence beyond 25 snapshots of sampling. This is shown by the similar statistical errors and the convergence of the mean binding energies. In this system, including an entropy correction term is only beneficial when sampled over at least 25 snapshots. Entropy terms with less sampling increase RMSE and reduce correlation. Sampling entropy beyond 100 snapshots does not improve results.

Our study demonstrates that QM-PBSA with full-protein calculations is viable in an academic context and is a useful addition to the toolbox of free energy calculations, especially in cases where force field parametrization may not be sufficiently able to capture effects such as charge transfer and polarisation, which are included by default in quantum descriptions. Looking to the future, we believe that extending more rigorous classical mechanical binding free energy methods to full-QM, using linear-scaling density functional theory, has significant potential. QM-PBSA method is an important stepping stone.

Chapter 5

Protein-Ligand Binding Energies in BRD4

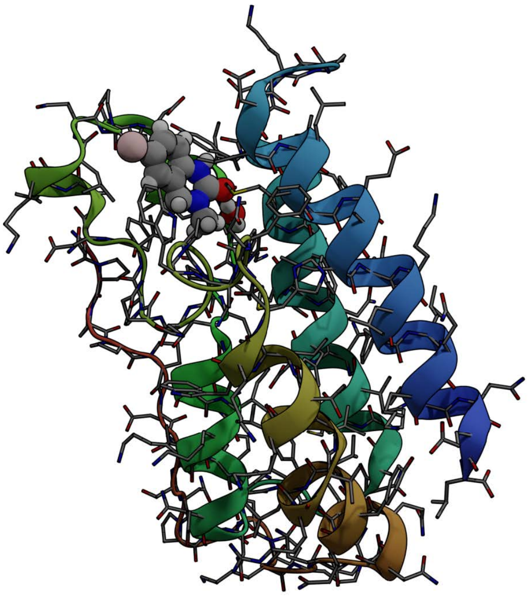


FIGURE 5.1: Stick and ribbon representation of the BRD4(1) protein structure. Ligand 4, represented as a space-filling model, is also shown in the binding site.

5.1 Introduction

While the QM-PBSA study of T4-lysozyme presented in the previous chapter was valuable in comparing method parameters and studying the convergence of QM-PBSA binding energies, T4-lysozyme is a relatively simple and well-established benchmarking system. The next challenge is to apply the QM-PBSA method, with the lessons learned in T4-lysozyme, to a ″real″ protein-ligand system. By ″real,″ we mean a pharmaceutically relevant protein and ligand set with more challenging properties. In addition to the choice of the protein target, a major shortcoming of the T4-lysozyme study was the very short MD sampling with outdated force fields, the low number of ligands, and their small binding affinity range.

This chapter presents our QM-PBSA study of the pharmaceutically highly relevant bromodomain containing protein 4 (BRD4). We aim to address some of the shortcomings of our previous T4-lysozyme study. BRD4 plays a key role in many cancers and is under active investigation as a cancer drug target [154]. The inhibition of BRD4 can suppress the cancer growth of acute myeloid leukemia, diffuse large B cell lymphoma, prostate cancer, and breast cancer [155].

Before calculating expensive QM binding energies, an extensive MD study is conducted in Section 5.3 in which the conformational motion of the protein and its dependence on the choice of MM force field are explored. In Section 5.4 the MM-PBSA and MM-GBSA binding energies are calculated based on trajectories from different force field combinations. Additionally, MM-PBSA and MM-GBSA, with the inclusion of explicit waters, are tested. Building on the solid foundational MM study of BRD4 presented in Sections 5.3 and 5.4, QM-PBSA binding energies are calculated and compared to MM-PBSA/GBSA in Section 5.5.

5.2 The BRD4(1) system

Bromodomains are small protein domains of about 110 amino acids responsible for recognizing acetyl-lysine residues on proteins. There are a wide variety of bromodomain-containing proteins. BRD4 is a member of the well-studied BET family of bromodomain-containing proteins [156]. In this study, we focus on the first bromodomain of the BRD4 protein, commonly referred to as BRD4(1) and pictured in Figure 5.1. For convenience, the ’(1)’ nomenclature is dropped in the remainder of the text. This bromodomain contains 121 amino acids and 2035 atoms. Its binding site has clear access to the solvent and is located between two hydrophobic loop structures called the ZA- and BC-loops, which connect the four anti-parallel alpha helices that make up the structure’s body. The ZA- and BC- loops are highlighted in blue and red in Figure 5.2.

The 121 protein residues of the BRD4(1) structure under investigation are only a single protein domain of a larger macromolecule. The residue numbering employed in this chapter’s analysis is defined by the 121 residue protein domain rather than the entire biological entity. To find the corresponding residue number in the whole genome, add 44 to the residue number reported here.

The BRD4 bromodomain and a set of 10 ligands, labeled 1-10 in order of ascending binding energy with a potency range of 5 kcal/mol, have been proposed by Mobley et al. [135] as a potential binding free energy benchmarking system. The benchmarking set was recently used in an absolute binding energy study by Huggins et al. [157]. High-quality co-crystal structures are available for most ligands, along with experimental binding energies. The ligands are uncharged, chemically diverse, and span a binding range of about 5 kcal/mol. The binding site is solvent accessible, and many of the ligands form hydrogen bonds with the Asp137 residue inside the binding site [58]. A full table including structures, SMILES, experimental energies, and PDB codes is presented in [135] and an abbreviated version in Figure B.1 of Appendix B.

The experimental binding energies of 4/10 ligands were obtained using the Alphascreen technology with reported average estimated experimental errors of 0.20 kcal/mol and a maximum estimated experimental error of 0.23 kcal/mol [158–160]. Among the Alphascreen ligands are a confirmed non-binder, ligand 1, which was found to be inactive at a concentration of 250 μM, and a weak binder, ligand 2, with 32% inhibition at 250 μM. The experimental binding energies of the other 6 ligands, all binders, were measured using isothermal titration calorimetry. The reported mean experimental errors were 0.05 kcal/mol and the maximum experimental error was 0.07 kcal/mol [161–164]. The reported experimental errors are likely significantly underestimated.

Due to the pharmaceutical relevance of bromodomain-containing proteins, BRD4 has been the subject of extensive computational research with a noticeable increase in interest in the past 4 years. A variety of binding energy studies on BRD4 have been performed using MD [165], QM/MM [166, 167], absolute binding free energy (ABFE) methods [57, 157, 168], MM-PBSA [57, 169, 170] and MM-GBSA [66, 167, 171–173].

5.2.1 Previous computational studies

We introduce some studies to which we will refer in the discussion but cannot include all computational research into BRD4. In 2013 Steiner et al. [174] studied the flexibility and accessibility of the binding site in 20 bromodomains, including BRD4, reporting alternate side chain arrangements observed in molecular dynamics (MD) simulations using the CHARMM PARAM22 + TIP3P force fields. Kuang et al. [166] used classical MD with umbrella sampling and QM/MM MD to study two inhibitors to BRD4, $(+/-)$-JQ1. One of these inhibitors, $(+)$-JQ1, is ligand 9 in our ligand set. In 2017, Aldeghi et al. [57] compared binding free energy predictions of 11 ligands in BRD4 from an absolute binding energy method with MM-PBSA energies. Also, in 2017, Heinzelmann et al. [165] used the attach-pull-release absolute binding energy method on a set of 7 ligands in BRD4. Cheng et al. [167] compared the binding affinity of RVX208 and RVX297 to the first and second bromodomain of BRD4 using QM/MM

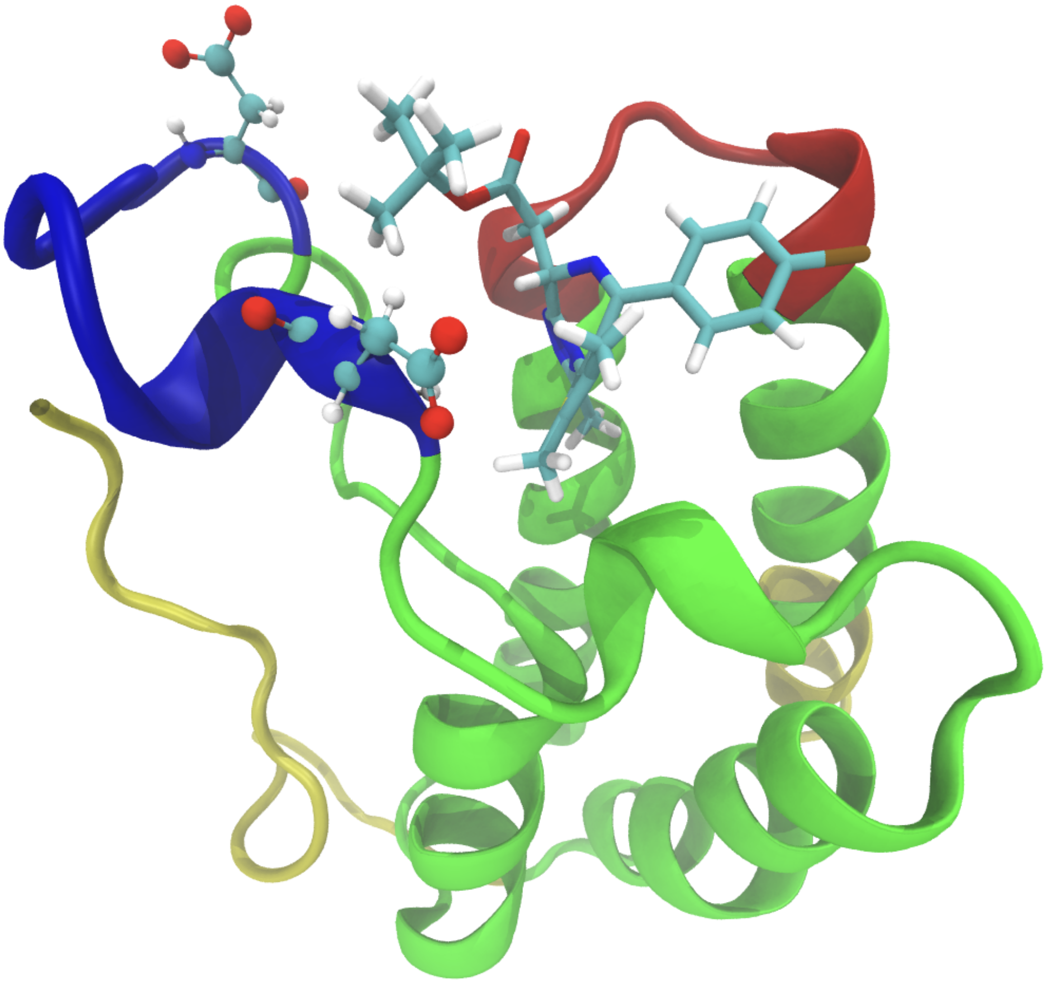


FIGURE 5.2: ZA-loop and BC-loop of BRD4(1) highlighted in blue and red, respectively. Ligand 9 bound in binding site.

MD as well as MM-GBSA. RVX208 is ligand 6 in our ligand set. In 2018 Su et al. published two papers studying first 3, then 5 ligands binding to BRD4 using MM-PBSA [169, 170]. In 2019, 2020, and 2021 Wang et al. [66, 171, 172] studied 9 inhibitors in BRD4 using MM-GBSA and principal component analysis. Rodriguez et al. [173] examined binding energies and binding poses of olonine in BRD4 using MM-GBSA with explicit waters in 2020. In 2022, Guest et al. [168] applied two alchemical free energy methods to BRD4, and Huggins et al. [157] performed absolute binding free energy calculations on the same ligand set used in this study. A 2019 review article by Myrianthopoulos et al. [175] summarizes the vast body of in-silico studies in BRDs.

5.3 Molecular mechanics force fields and BRD4 conformational motion

Due to the high computational cost of QM-PBSA binding energy calculations, only a limited number of snapshots can be considered per ligand. As a result, the method can be sensitive to the choice of snapshots sampled from an unstable trajectory with, for example, multiple different binding modes or even ligand dissociation events. Thus, before embarking on expensive QM-PBSA calculations in BRD4, we thoroughly explore the dynamics of the protein-ligand complex with various force field combinations. In this section, we perform 900 ns of MD for the set of 10 ligands binding to BRD4 using 6 different force field combinations. We study the trajectory stability and occurrence of ligand dissociation and conformational changes in the protein. The stability and convergence of the MD trajectories are evaluated using the root mean square deviation (RMSD) of the complex structure from the original structure. Based on an initial residue-wise root mean square fluctuation (RMSF) analysis, the RMSD of the protein backbone, the ZA- and BC-loop, the ZA-loop only, and the ligand only are considered. Furthermore, instances of ligand dissociation are compared across trajectories from different force field combinations. Finally, a conformational change in the ZA-loop is identified, characterized, and its dependence on the choice of force field is established.

5.3.1 Methods

5.3.1.1 Molecular dynamics and force fields

This study focuses on the various choices of molecular mechanics force fields for proteins and how they may influence the protein-ligand dynamics of BRD4. In modeling a protein-ligand system, three sets of force field parameters are utilized. One for the ligand, one for the protein, and a third for the explicit solvent (usually water). We introduce the force fields used in this study below.

For the ligand parameters, the generalized Amber force fields GAFF2 and its predecessor GAFF1 for organic molecules are used [25]. GAFF1 and GAFF2 are compatible with all protein force fields and explicit water representations in Amber and contain parameters for most organic and pharmaceutical functional groups.

For the protein, the force fields ff14SB [23], ff19SB [18] and ff15ipq [24] are used. For the representation of the explicit water molecules in the dynamics simulations the water models TIP3P [26], OPC [28], OPC3 [29] and SPCE [27] were used. These force field are introduced in Section 3.1.2.1.

The water model used in the initial parameterization of the protein force field can lead to beneficial error canceling between the chosen water model and force field. Because of this, ff14SB usually is combined with the TIP3P water model [23] and ff15ipq with SPCE [24]. In the development of ff19SB, emphasis was placed on the interchangeability of water models but usage is recommended with OPC or OPC3. The combination of ff19SB with TIP3P, which is explicitly warned against in the Amber20 manual, has also been tentatively tested. The protein force field ff14SB is paired with the general force field GAFF for the ligand, while the newer ff19SB and ff15ipq are paired with the updated GAFF2.

In an attempt to test a ligand force field other than GAFF1 and GAFF2, the force field combination ff19SB with OpenFF(2.0) [176] for the ligand was created using the ParmEd [177] utility. However, the combination of OpenFF with ff19SB has only recently been introduced and has not been fully characterized.

5.3.1.2 Computational details

System preparation The initial structures for BRD4 in complex with the 10 ligands were taken from the published benchmarking data set of Mobley et al. [135]. Co-crystal structures are available for 9/10 of the ligands. Hydrogen atoms were added to the ligands using Open Babel [178], and the AM1-BCC charge model, as implemented in Amber's antechamber, was used by Mobley et al. to assign ligand atom partial charges. Mol2 and sdf files are provided in the benchmarking dataset for each ligand. The ligand for which no experimental structure is available, named ligand 1 in the Mobley set, was docked, by Mobley et al., into BRD4 using AutoDock Vina [179].

Parameter files were generated by the authors of this study for the protein using ff19SB, ff14SB, and ff15ipq. For the ligands, the force fields GAFF1 and GAFF2 are used. In the solvation step, 11000 water molecules and 32 Na+ and 35 Cl- ions (0.15 M NaCl) are added. The excess of 3 Cl- ions was added to neutralize the protein's net charge. Joung and Cheatham's ion parameters are used [180]. Our extensive equilibration protocol has been applied and is described below.

Equilibration protocol

- Minimization 1: 1000 steps conjugate gradient and 1000 steps steepest descent, nonbonded cut-off of 8 Å, restraint of 1000 $\text{kcal} \cdot \text{mol}^{-1} \cdot Å^{-2}$ on all heavy atoms.
- Minimization 2: 1000 steps conjugate gradient and 1000 steps steepest descent, nonbonded cut-off of 8 Å, restraint of 1000 $\text{kcal} \cdot \text{mol}^{-1} \cdot Å^{-2}$ on complex heavy atoms to relax the water residues.

- Heating: 100 ps of heating of the solvent with NVT ensemble from 100 K to 300. Heavy atoms of the complex restrained with 1000 kcal $\cdot$ mol^{-1} $\cdot$ $\AA^{-2}$.
- Density: Adjust the size of the box via 500 ps of NPT and heavy atoms of the complex restrained with 1000 kcal $\cdot$ mol^{-1} $\cdot$ $\AA^{-2}$.
- Cooling: 100 ps of cooling of the solvent with NVT ensemble from 300 K to 100. Heavy atoms of the complex restrained with 1000 kcal $\cdot$ mol^{-1} $\cdot$ $\AA^{-2}$.
- Minimization: 1000 steps conjugate gradient and 1000 steps steepest descent, nonbonded cut off is 8 A, for each decreasing constraint of 1000, 500, 100, 50, 20, 10, 5, 2, 0 kcal $\cdot$ mol^{-1} $\cdot$ $\AA^{-2}$.
- Heating: 100 ps of heating of the system with NVT ensemble from 100 K to 300 K with no constraints.
- Equilibration: 100 ps simulation at 300 K in NPT ensemble with no constraints.

Molecular dynamics All production calculations are performed with the same settings in an NPT ensemble. The SHAKE algorithm is used to constrain hydrogen bonds, and the time step is 0.002 ps. Temperature regulation is achieved using Langevin dynamics with a collision frequency of 2.0 ps^{-1} at 300 K. The Berendsen barostat with isotropic position scaling at atmospheric pressure and a pressure relaxation time of 2 ps regulates the pressure. A cutoff of 8.0 Å is used for non-bonded interactions. Production calculations are performed with the force field and water model combinations shown in Table 5.1. For each force field combination, three replicas of 300 ns each are run for a total of 900 ns per ligand.

TABLE 5.1: Summary of force field combinations employed. * Suk Joung and Thomas E. Cheatham as implemented in Amber20.

Force Field Combination	Protein	Ligand	Water	Ions *
ff14SB-TIP3P	ff14SB [23]	GAFF1	TIP3P [26]	J. & C. TIP3P [180, 181]
ff15ipq-SPCE	ff15ipq [24]	GAFF2	SPCE [27]	J. & C. SPCE
ff19SB-TIP3P	ff19SB [18]	GAFF2	TIP3P	J. & C. TIP3P
ff19SB-OPC	ff19SB	GAFF2	OPC [28]	J. & C. TIP4PEW
ff19SB-OPC3	ff19SB	GAFF2	OPC3 [29]	J. & C. TIP3P
ff19SB-OpenFF-OPC3	ff19SB	OpenFF 2.0 [176]	OPC3	J. & C. TIP3P

5.3.2 Results

5.3.2.1 RMSF analysis

The protein's residue-wise root mean squared fluctuation (RMSF) was calculated for all MD trajectories to identify critical areas of protein motion throughout the molecular dynamics (MD) simulations. Figure 5.3 shows the residue-wise RMSF

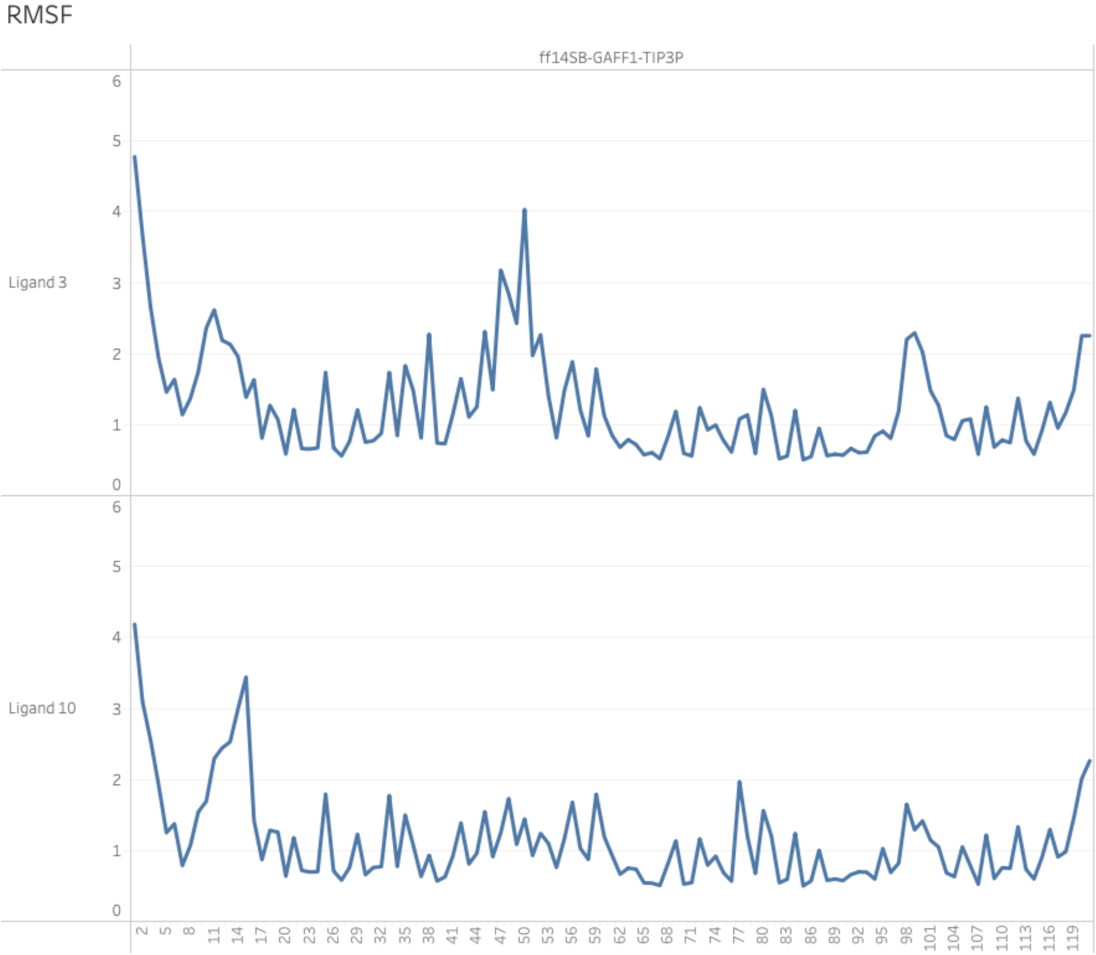


FIGURE 5.3: Residue-wise root mean squared fluctuation (RMSF, Å) of the protein BRD4(1) bound with ligands 3 and 10 over a single 300 ns molecular dynamics simulation using the ff14SB-TIP3P force field combination.

values for two 300 ns simulations of BRD4 bound to ligands 3 and 10 using the ff14SB-TIP3P force field combination.

As illustrated in Figure 5.3, two areas of motion are the protein's tail ends, which we define as residues 1-16 and residues 115-121. This definition of the protein "ends" is somewhat arbitrary and based on observations of RMSF and RMSD values and visual inspection of the trajectories. It is not based on protein structure or function. The tail ends are free to move in the solvent throughout the simulation and are located far away from the binding site. Comparable movement of the ends is observed across all 10 protein-ligand complexes and different choices of force fields and explicit solvent models. In the full biological entity of BRD4, the tail ends would be attached to the second bromodomain of BRD4 and would exhibit significantly less motion.

The two other regions of protein motion are the ZA- and BC-loop structures which encase the solvent exposed binding site. For the following analysis, we define the ZA-loop as residues 45 to 54 and the BC-loop as residues 95 to 103. The RMSF plot for ligand 3 in Figure 5.3 has a clear peak in the ZA-loop region and a secondary, smaller peak in the BC-loop region. In contrast, minimal fluctuation in the position of the protein loops is seen for BRD4 bound with ligand 10. Figures 5.2 and 5.5 give a closer view of the two loop structures.

Besides the tail ends and the two loop structures, BRD4 undergoes very little motion during the MD trajectories for all ligands and force fields tested.

5.3.2.2 Trajectory stability

TABLE 5.2: Average and maximum of the root mean squared deviation (RMSD, Å) from the initial pose of molecular dynamics trajectories with different force fields and water models across a set of 10 ligands binding to BRD4. The RMSD of the protein backbone (excluding tail ends), protein backbone of only the ZA- and BC-loops, backbone of only the ZA-loop, and the RMSD of the ligand are shown. Each replica is 300 ns of MD and RMSD values are sampled at 1500 snapshots per replica using Amber20's PyTraj.

		Replica 0		Replica 1		Replica 2		Total	
RMSD Mask	Force Field / RMSD	avg	max	avg	max	avg	max	avg	max
Backbone-NoEnds	ff15ipq-SPCE	1.2	2.7	1.3	4.5	1.4	5.7	1.3	5.7
	ff14SB-TIP3P	1.1	2.3	1.4	3.9	1.2	2.3	1.2	3.9
	ff19SB-TIP3P	1.0	1.7	1.0	2.2	0.9	1.6	1.0	2.2
	ff19SB-OPC	0.9	1.7	1.0	2.1	1.1	5.6	1.0	5.6
	ff19SB-OPC3	1.0	2.2	1.0	2.1	0.9	1.8	1.0	2.2
Backbone-ZA- and BC-loops	ff15ipq-SPCE	1.6	5.4	1.7	5.3	1.9	6.9	1.7	6.9
	ff14SB-TIP3P	1.5	4.4	2.1	5.7	1.6	4.6	1.7	5.7
	ff19SB-TIP3P	1.2	3.5	1.3	4.2	1.2	2.6	1.2	4.2
	ff19SB-OPC	1.1	2.9	1.2	4.1	1.4	7.0	1.3	7.0
	ff19SB-OPC3	1.3	4.0	1.3	3.9	1.2	3.2	1.2	4.0
Backbone-ZA-loop	ff15ipq-SPCE	1.8	7.1	1.8	6.9	2.1	8.3	1.9	8.3
	ff14SB-TIP3P	1.6	5.8	2.5	7.6	1.9	6.3	2.0	7.6
	ff19SB-TIP3P	1.3	3.8	1.5	5.7	1.3	3.5	1.4	5.7
	ff19SB-OPC	1.2	3.7	1.4	5.4	1.5	8.3	1.4	8.3
	ff19SB-OPC3	1.5	5.2	1.4	5.1	1.3	4.3	1.4	5.2
Ligand	ff15ipq-SPCE	2.6	13.1	2.4	10.2	3.5	56.1	2.8	56.1
	ff14SB-TIP3P	2.3	9.6	4.5	66.7	2.6	9.2	3.1	66.7
	ff19SB-TIP3P	2.1	6.8	2.2	9.6	2.1	12.8	2.1	12.8
	ff19SB-OPC	1.9	9.4	2.1	7.8	2.2	6.6	2.1	9.4
	ff19SB-OPC3	2.0	5.9	1.9	5.9	1.9	4.8	2.0	5.9

The root mean squared deviation (RMSD) from the initial configuration was calculated for all trajectories to assess the stability of the molecular dynamics trajectories for BRD4 in complex with a set of 10 ligands with a variety of force fields and explicit water models. For each of the three 300 ns replicas, the RMSD for 1500 snapshots was evaluated. The RMSD gives insight into the degree of change of different regions of the protein and ligand conformations throughout the simulation compared to the initial structure. Based on the residue-wise RMSDF analysis in Section 5.3.2.1, the RMSD was measured for the following regions of the protein-ligand complexes:

- Backbone NoEnds : protein backbone excluding tail ends (residues 17-114),
- Backbone ZA- and BC-loop: backbone of ZA- and BC-loops (residues 45-54 and 95-103),
- Backbone ZA-loop: backbone of ZA-loop only (residues 45-54),
- Ligand: movement of the ligand only.

The protein tail ends were excluded as they move freely in the solvent, do not contribute to ligand binding, and their movement obfuscates details of the RMSD calculation.

Due to the large number of trajectories, only the average and maximum RMSD values across the entire ligand set are shown in Table 5.2 for each force field combination. Three replicas of 300 ns and results across all three replicas are shown.

Not shown in Table 5.2 is the experimental combination of Amber's ff19SB for the protein with the OpenFF ligand force field and the three-point OPC3 water model. With maximum backbone RMSD values of 8.3 and 7.7 Å in replicas 0 and 1, respectively, there are snapshots where the protein has undergone considerable conformational changes. All 10 ligands immediately dissociate from the protein for the experimental ff19SB-OpenFF-OPC3 force field combination. Because of this, the force field combination ff19SB-OpenFF-OPC3 is excluded from the study and not discussed further.

The remaining force fields can be grouped into two categories: ff19SB-based and non-ff19SB-based. Consider first the backbone RSMD of the protein with the tail ends removed, labeled backbone NoEnds in Table 5.2. The non-ff19SB-based ff14SB-TIP3P and ff15ipq-SPCE have similar average RMSD values of 1.25 to 1.50 Å across the three replicas with total average RMSDs of 1.33 and 1.42 Å, respectively. The three ff19SB-based force field combinations all have total average RMSDs of about 1 Å, indicating that the protein undergoes less motion and remains closer to the original pose throughout the trajectories than in the non-ff19SB simulations. Across the three replicas, the ff19SB-based force field combinations are very similar in average RMSD. For 8 out of 9 ff19SB-based replicas, the maximum protein backbone RMSD is about 2 Å. Lastly, considering the standard deviation (std) of the backbone RMSD distributions, the std of the ff19SB-based force field combinations is generally less than half that of the non-ff19SB-based force field combination.

As identified in the RMSF analysis, the ZA- and BC-loop structures are the two main regions of protein motion in BRD4. This is reflected in the larger backbone RMSD values of the loop regions compared to the whole protein. The ff19SB-based force field combinations have lower RMSD values than ff14SB and ff15ipq, with total average RMSD values of about 1.25 Å and 1.70 Å, respectively. While the maximum RMSD values are closer, the RMSD standard deviations highlight that ff19SB-based simulations spend significantly less time in protein loop conformations that differ significantly from the original pose.

The higher average and max RMSD values for the ZA-loop region reveal that most of the protein motion is in the ZA- rather than the BC-loop. Especially for ff14SB-TIP3P and ff15ipq-SPCE, max RMSD values of 6-8 Å are observed, indicating significant conformational changes. These conformational changes in the ZA-loop are discussed in more detail in Section 5.3.2.3.

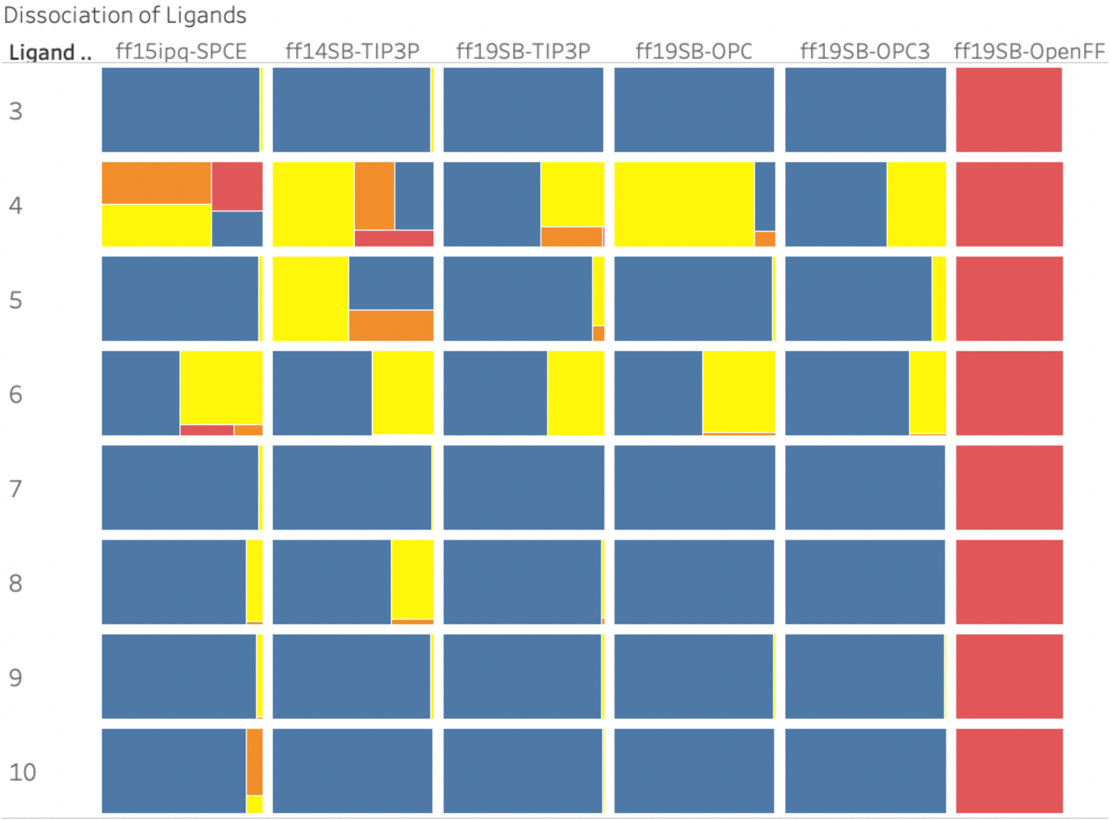


FIGURE 5.4: Proportions of ligand RMSD values across three replicas of 300 ns each for 8 binders with BRD4. Blue : Bound (RMSD < 3 Å), Yellow : Bound but significant motion of ligand in binding site (3 Å < RSMD < 5 Å), Orange : Different binding pose or in process of unbinding (5 Å < RMSD < 10 Å), Red : Ligand fully dissociated (10 Å < RMSD). RMSD value of ligand calculated relative to first snapshot in trajectory and for 1500 snapshots per 300 ns replica.

Finally, consider the motion of the ligand only in Table 5.2. Figure 5.4 visualizes the proportions of snapshots in which the ligand is dissociated for the 8 binders in the ligand set. Overall, the ff19SB-based combinations have lower average RMSD values and almost no ligand dissociation. In contrast, ff14SB-TIP3P and ff15ipq-SPCE have higher average RMSD values and multiple clear instances of ligands dissociating, as evident in Figure 5.4. While ligands 4 and 6 have regions of ligand motion between 3 and 5 Å for all force fields, ff19SB-based combinations have smaller regions with RMSD beyond 5 Å. For ligands 3, 5, 7, 8, 9, and 10, ff19SB-OPC and ff19SB-OPC3 ligand RMSD values are almost all < 3 Å, i.e., the ligand is strongly bound. At the same time, ff14SB-TIP3P has significant unstable ligand configurations in ligands 5 and 8. Likewise, ff15ipq-SPCE in ligands 8, 9, and 10. These differences show that the average ligand RMSD is higher across the board for ff14SB-TIP3P and ff15ipq-SPCE compared to ff19SB-based force field combinations.

5.3.2.3 ZA-loop motion

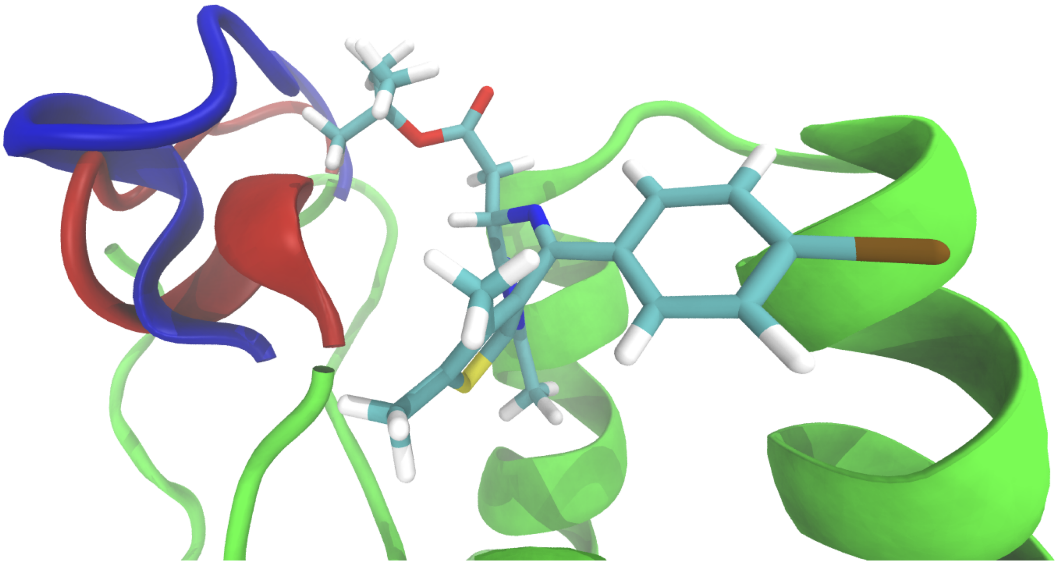


FIGURE 5.5: Opened (red) and closed (blue) ZA-loop conformation in BRD4 bound to ligand 9 using ff14SB+TIP3P force field combination.

The RMSF plot in Figure 5.3 and ZA-loop RMSD values in Table 5.2 indicate that the majority of protein motion observed, especially for force fields ff14SB-TIP3P and ff15ipq-SPCE, is concentrated in the ZA-loop. The ZA-loop is attached to the main body of the protein by two Asp residues shown in Figure 5.2 (residues 45 and 53). Figure 5.5 is a closer view of the ZA-loop structure in BRD4 in complex with ligand 9. The blue ZA-loop configuration, which we will refer to as the closed configuration, is the conformation observed in the co-crystal structures of BRD4 and the initial pose for our simulations. The red conformation of the ZA-loop, which we call the open conformation, is adopted by the protein in several simulations for multiple ligands. In this section, we characterize the open loop conformation and determine for which ligands and force field combinations this alternate ZA-loop conformation is observed.

Visually, the ZA-loop opening moves the tip of the ZA-loop outward, away from the binding site. The Asp residues on either side of the ZA-loop act like hinges for the outward loop rotation [165]. The transition takes a few nanoseconds but once adopted, the open-loop conformation is maintained for the rest of the trajectory. Visual inspection of the trajectories indicates two separate instances of the ZA-loop opening. First, ZA-loop opening after ligand dissociation, and second, ZA-loop opening while the ligand is bound and remains bound.

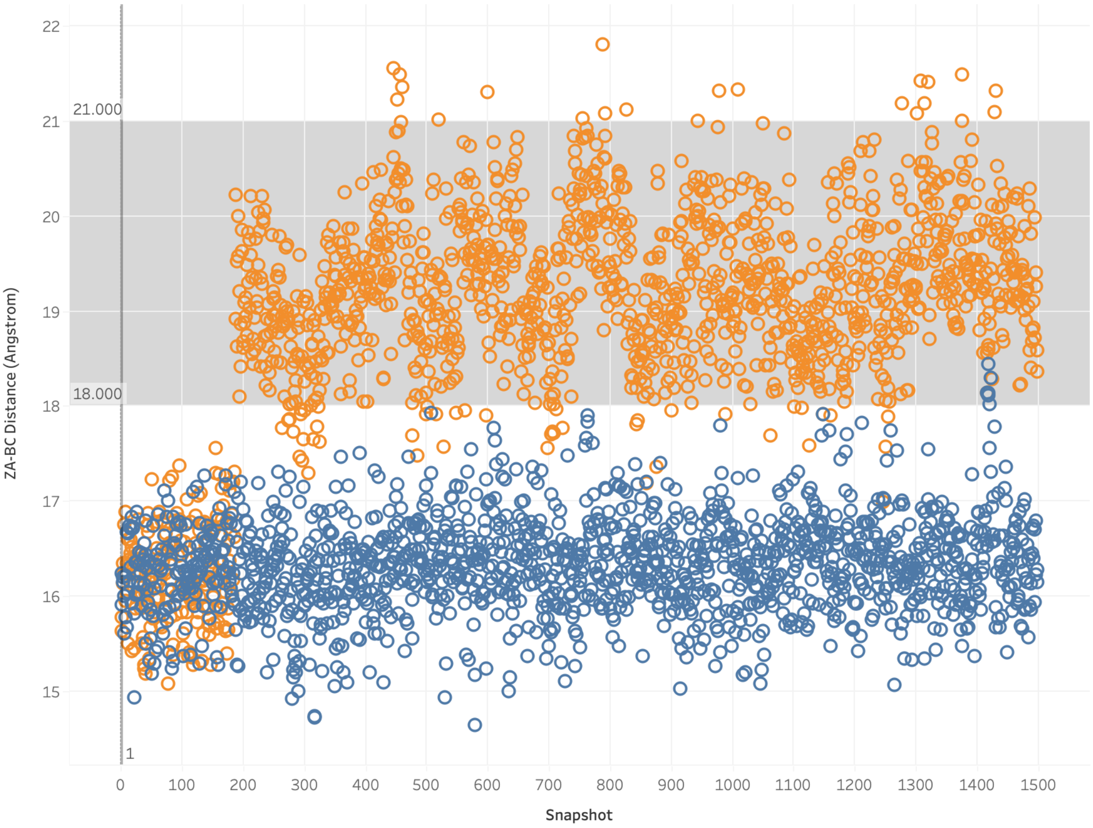


FIGURE 5.6: Distance in Å between the ZA- and BC-loops of BRD4 bound to ligand 9. Replica 0 in blue and replica 1 in orange. Grey band shows area indicative of open ZA-loop conformation.

The ZA-loop opening can be consistently identified by measuring the distance between the top of the ZA-loop and the BC-loop and by calculating the dihedral angles of the Asp hinges, which mediate the ZA-loop opening. Figures 5.6 and 5.7 show the distance between the ZA- and BC-loops (Å) and Asp-hinge dihedral angles for each snapshot in replicas 0 and 1 of BRD4 bound with ligand 9 using ff14SB-TIP3P.

Replica 0, shown in blue, does not undergo ZA-loop opening, and the ZA- to BC-loop distance remains in the range of 15-18 Å, and Asp dihedrals do not change systematically. Replica 1, in orange, undergoes ZA-loop opening at snapshot 200. This is evidenced by the increase in ZA- to BC-loop distance (18-21 Å) and abrupt change in dihedral angles. The increased loop distance and changed dihedral remain constant until the end of the simulation.

Using the ZA- to BC-loop distance and dihedral angles, ZA-loop opening is easily and reliably identified. For trajectories generated using the force field combination ff14SB for the protein, GAFF1 for the ligand, and the TIP3P water model, ZA-loop opening is observed for ligands 1, 3, 4, 5, 6, 7, and 9. For ligands 1 and 4, the ZA-loop opening is a

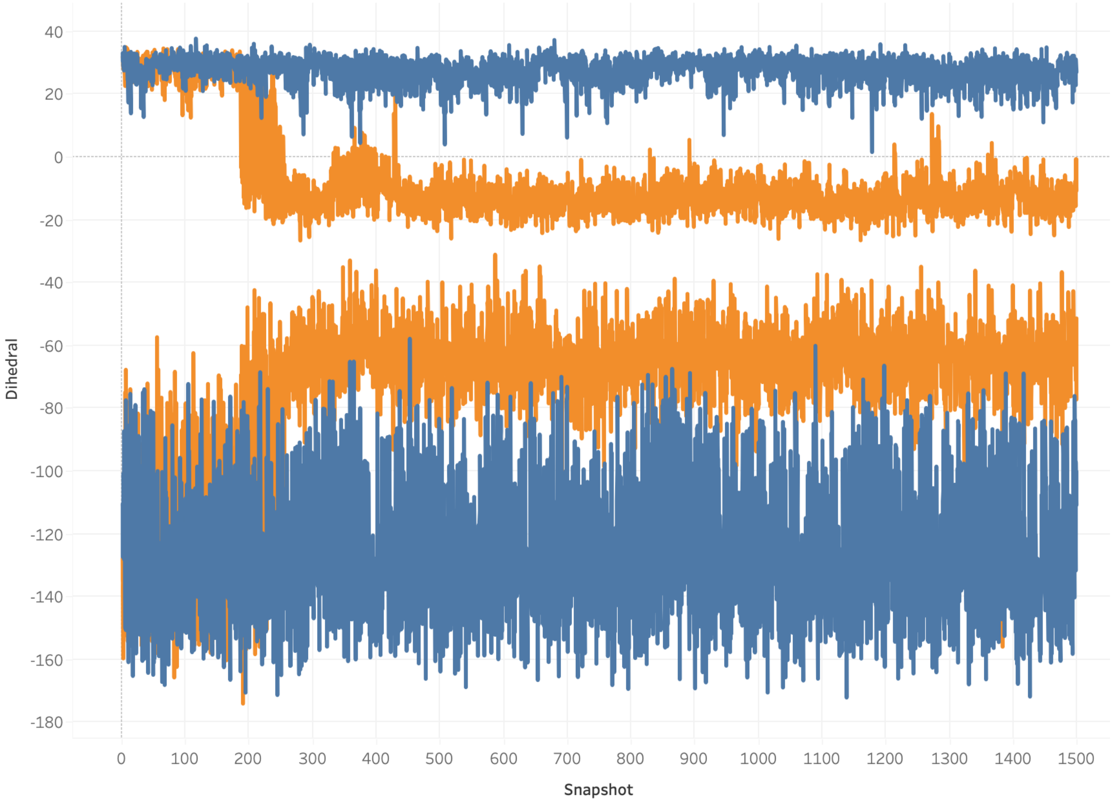


FIGURE 5.7: Asp45 and Asp53 dihedral angles (degrees) for replica 0 (blue) and replica 1 (orange) for BRD4 bound to ligand 9 using ff14SB-TIP3P force field combination.

result of the dissociation of the ligand from the protein. For the force field combination ff15ipq-SPCE, the ZA-loop opening is observed for ligands 4 and 6 but occurs after the ligand dissociates from the protein. No instances of ZA-loop opening are observed in trajectories generated using the ff19SB-TIP3P, ff19SB-OPC, and ff19SB-OPC3 force field combinations. For ff19SB-OpenFF-OPC3, all ligands dissociate, and ZA-loop motions resembling the ZA-loop opening occur in the unbound protein.

In summary, ZA-loop opening occurs when the ligand dissociates for force field combinations ff14SB-TIP3P, ff15ipq-SPCE, and ff19SB-OpenFF-OPC3. Only in ff14SB-TIP3P trajectories is the ZA-loop opening observed with ligands 3, 5, 6, 7, and 9 bound to the protein and is irreversible on the timescale of the trajectories.

5.3.3 Discussion

5.3.3.1 ff19SB produces more stable trajectories

In terms of trajectory stability, the ff19SB protein force field combined with GAFF2 for the ligand and the OPC, OPC3, or TIP3P water models have significantly lower protein backbone RMSD values ($\approx$ 1 Å) compared to ff14SB+GAFF+TIP3P and ff15ipq+GAFF2+SPCE ($\approx$ 1.3 Å). The spread of backbone RMSD values in ff19SB-based combination is also less than half of that in non-ff19SB-based force fields. Specifically, in the ZA- and BC-loops, the ff14SB and ff15ipq force field combinations lead to significantly more protein motion. This is partly due to the open ZA-loop conformation observed only in ff14SB and ff15ipq, which leads to maximum ZA-loop RMSD values of 6 to 8 Å. Overall, the 8 binders remain more firmly bound in ff19SB-based trajectories, while ligands in the ff14SB trajectories take on multiple, often significantly different, binding modes throughout the simulation. There are also no instances of complete ligand dissociation in the ff19SB-based force field combinations, whereas multiple ligands dissociate completely in ff14SB and, to a lesser extent, ff15ipq.

5.3.3.2 Protein conformational change in ff14SB

The ZA-loop opening described in Section 5.3.2.3 and illustrated in Figure 5.5, can be identified by measuring the distance between the ZA- and BC-loops and the dihedral angles of the Asp protein residues at the base of the ZA-loop. This ZA-loop transition was previously observed and characterized in simulations of the unbound BRD4 protein by Heinzelmann et al. in 2017 [165]. According to Heinzelmann, the transition is irreversible on the timescale of hundreds of ns, occurs after 20 to 60 ns in simulations of the unbound protein, and is about -2.5 kcal/mol more favorable than the closed-loop conformation in the crystal structure of the unbound protein. Heinzelmann did not observe the ZA-loop opening for bound protein-ligand complexes, but it appears no extensive unconstrained sampling of the protein in complex with the ligand was performed. The trajectories were produced using ff14SB with GAFF2 and the TIP3P and SPCE water models. Recently, Huggins et al. [157] published an absolute binding free energy study in BRD4 and confirmed the ZA-loop opening in the unbound protein, which they observed in 4 out of 5 100 ns MD simulations using the protein force field ff14SB and SPC/E water model. Long timescale MD simulations were not performed on the bound protein-ligand complex, and ZA-loop opening was not observed in the 2.4 ns λ-windows of the free energy calculations. In 2021 Guest et al. [58] studied 297 co-crystal structures of BRD4 bound to small molecules and showed that BRD4 maintains high structural similarity regardless of the ligand choice. Although the ZA-loop showed the most flexibility in the active site, the differences observed are tiny compared to the ZA-loop transformation produced by our ff14SB-based MD simulations.

In our ff14SB trajectories, the ZA-loop opening is observed with the ligand bound for 5 out of 10 ligands. As observed by Heinzelmann et al. [165], the ZA-loop opening appears to be mediated by two Asp hinges whose dihedral angles change characteristically (Figure 5.7) during the loop opening. In our simulations of the bound protein-ligand complexes, the ZA-loop opening occurs, for some ligands, in only one or two of the three 300 ns replicas, indicating that the transition may, on average, take hundreds of ns to occur. Other research has shown that ligand binding stabilizes the ZA-loop motion in BRD4 [170], potentially explaining the longer time to transition in the complex than in the unbound protein.

The ZA-loop opening is not observed for any ligands in trajectories generated using ff19SB-based force field combinations. The difference we observe between trajectories using ff19SB and ff14SB may be explained by the improved treatment of protein backbone dihedral parameters in ff19SB compared to its predecessor ff14SB (or ff99SB) [18]. In ff19SB, explicit coupling between the ψ and ϕ dihedral parameters is enabled by training against two-dimensional quantum mechanical energy surfaces rather than only local minima in ff14SB. Furthermore, the treatment of polarization was made more consistent in mapping ff19SB dihedral parameters, and amino acid specific correction maps were added to account for the dependence of atomic dihedral parameters on the protein residue. Finally, a backbone side-chain coupling correction is also considered for the first time in ff19SB. These improvements in ff19SB over ff14SB (or ff99SB) lead to better reproduction of amino acid specific differences and closer agreement with experimental properties [18].

Fundamentally, it is not possible to assess which trajectories are closer to the true dynamics of the protein-ligand complex. However, the above results should be carefully considered in choosing a trajectory as input for the expensive QM-PBSA binding energy study of BRD4. To our assessment, the dissociation of multiple binders, adoption of non-crystallographic binding poses, and the ZA-loop conformation not observed in any crystallographic structure of BRD4 indicate that the ff19SB trajectories should be chosen over those generated with ff14SB. The more stable trajectories of ff19SB-based force field combinations, which feature more strongly conserved binding modes similar to the ligand's crystallographic poses, thus lend themselves more easily as inputs for quantum mechanical binding free energy prediction.

5.3.3.3 Pairing of water model with ff19SB

In the development of ff19SB, particular emphasis was placed on disentangling the interdependence of the protein force field and the choice of an explicit water model. In contrast, the ff14SB protein force field relies on fortuitous error cancellation with the TIP3P water model [18], which was used during the parameterization of ff14SB. Regarding trajectory stability, the pairing of ff19SB+GAFF2 with OPC, OPC3, and TIP3P leads to indistinguishable backbone, ligand, and ZA-loop RMSD values. Furthermore, no ligands dissociate, and no ZA-loop motion is seen, irrespective of the choice of water model for ff19SB. This indicates that the significant differences between ff19SB-based combinations and ff14SB+TIP3P arise primarily from the change in the protein force field.

5.3.4 Conclusion

Based on the findings of this preliminary MD study of BRD4 with a series of force field combinations, we selected ff19SB as our protein force field of choice and the 900ns of MD generated as the basis for our QM-PBSA study presented in Section 5.5.

It is important to stress that the above exploration is insufficient to make general claims about the relative accuracy and reliability of the force fields explored. Due to the limited sampling of 3x300 ns, long-time-scale protein motions are not sufficiently captured to formulate statistical hypotheses about their importance nor any thermodynamic properties associated with these motions. Additionally, we have only explored these force field combinations in a single protein bound with 10 ligands. However, the ZA-loop opening in the BRD4 bound protein-ligand complex for ff14SB illustrates that the choice of molecular mechanics force field can substantially impact the conclusions drawn and binding energies derived from molecular dynamics studies of protein-ligand binding.

While our findings suggest that the observed differences are due to changes made between the protein force field ff14SB and ff19SB, the inconsistent use of GAFF1 and GAFF2 between these two force field combinations in our study may have also had an unforeseen impact. To thoroughly explore the ZA-loop transition and its force field dependence, we would recommend running the following additional simulations and analyses:

- replica all ff14SB simulations with GAFF2 instead of GAFF1,
- multiple long-time-scale MD simulations of the unbound protein with ff14SB and ff19SB would show for which force fields the unbound protein adopts the open-loop conformation,
- multiple ff19SB simulations starting from the open-loop configuration produced by the ff14SB simulations to see if the switch to the ff19SB force field would lead to the closing of the ZA-loop from the open-loop confirmation,
- at least 3 additional re-runs for each force field and ligand to gather more significant statistics on the occurrence and time-scale of ZA-loop opening if ff14SB,
- the conformational energy of the protein-ligand complex in the open and closed loop could be explored. Heinzelmann et al. [165] performed such an analysis for the loop-opening in the unbound protein.

For our QM-PBSA study, we believe that this preliminary study demonstrated that the 900 ns of MD generated by the force field combination ff19SB+GAFF2+OPC3 should be used as input for end-point binding free energy calculations. The following section explores MM-PBSA and MM-GBSA binding energy predictions in detail before investigating QM-PBSA in BRD4 in Sections 5.5.

5.4 MM-PBSA and MM-GSA binding energies

Having analyzed the MD trajectories of BRD4 bound to 10 ligands generated by a selection of force field combinations in the previous chapter, we now apply the end-point classical mechanical binding free energy methods MM-PBSA and MM-GBSA to the generated trajectories. We explore the inclusion of explicit water molecules as part of the protein in the binding energy calculations and the Interaction Entropy correction term. This section serves as a basis against which the later QM-PBSA results may be compared and additionally explores the impact of the force field and input trajectories on the calculated binding energies in BRD4.

5.4.1 Methods

5.4.1.1 Design of computational study

The 900ns of MD trajectories per ligand generated in the previous chapter are used as input for MM-PBSA and MM-GBSA. Binding energies calculated from the ff14SB and ff19SB trajectories, together with different explicit water models, are compared. The Interaction Entropy (IE) correction term from Duan et al. [62] is tested. Lastly, different numbers of water molecules are included explicitly as part of the protein in the MM-PBSA and MM-GBSA calculation. The predictive power as compared to experiment of the computed relative binding free energies is evaluated using the RMSDtr and Spearman's correlation coefficient r_s.

The MM-PBSA and MM-GBSA methods are suitable only for predicting relative, rather than absolute, binding free energies. The root mean squared deviation after removal of the systematic error (mean signed error), called RMSDtr, introduced in Section 4.5.1.2, is used to describe the error against experiment. While the RMSDtr metric assesses the closeness of predicted and experimental relative binding energies, Spearman's rank correlation factor r_s quantifies the predictive power of the relative binding energy rank ordering. A value of 1 corresponds to a calculated rank order equivalent to the experimental binding energies. Because Spearman's r_s describes the rank ordering of binding energies, it can be applied directly to the calculated binding energies.

The chosen ligand set of 10 ligands contains one confirmed non-binder, ligand 1, and one ligand of very low binding affinity, ligand 2 (32% inhibition at 250 μM). As these ligands' actual binding energies are unknown, they are excluded from the RMSDtr calculations. The non-binders are, however, included in the rank orderings as quantified by Spearman's r_s with "experimental" binding energies of 0 kcal/mol.

Estimates of the statistical error due to finite sampling in the computed metrics are obtained using the block averaged standard error (BASE) method[182]. The trajectory is split into M segments of length n. Then, the metric of interest, for example, Spearman's r_s, is calculated for each of the M segments. The standard deviation in r_s among the block averages, σ_n, is then used to estimate the overall standard error at block size n using $BASE(n) = \frac{\sigma_n}{\sqrt{M}}$. At $n = 1$ the BASE is equal to the analytic standard error (SE) of the metric across the whole trajectory. As n increases to up to one-fifth of the total trajectory length, the estimated BASE increases until it reaches a plateau. The value at which the BASE plateaus is a reliable estimator for the true standard error [182]. This study uses the maximum BASE value obtained as the upper bound on the uncertainty in the RMSDtr and r_s.

5.4.1.2 Computational details

MM-PBSA MM-PBSA is performed using Amber20's MMPBSA.py utility. The two-term non-polar solvation term is used in which the cavity and dispersion terms are treated separately. The recommended values are used for the cavity offset (-0.5692) and cavity surface tension (0.0378). An ionic strength of 0, fill ration of 4.0, solute dielectric of 1, and Amber's pre-calculated atomic radii are used.

MM-GBSA MM-GBSA is performed using Amber20's MMPBSA.py utility. Default settings in Amber20 with a modified GB model (igb=5) developed by A. Onufriev, D. Bashford, and D.A. Case were used, and parameters α, β, and γ set to 1.0, 0.8, and 4.85, respectively [34]. A solute dielectric constant of 1 is used. A link to the MM-PBSA and MM-GBSA inputs and outputs is provided in Appendix C.

Explicit waters The inclusion of explicit water molecules is not specifically supported in Amber's MMPBSA.py framework. The 'closest' method in cpptraj was used to create trajectories containing the N closest water molecules to the ligand in each snapshot. Thus the total number of waters remains constant throughout the trajectory, although the individual explicit waters are not conserved. Including 1-10, 15, 25, 50, and 100 waters was tested, and binding energies were calculated using MMPBSA.py with the water molecules defined as part of the host.

Interaction Entropy The Interaction Entropy method by Duan et al. [62] was implemented in Python and applied to the MM-PBSA/GBSA results as a post-processing step. The Interaction Entropy is outlined in more detail in Section 3.1.3.5.

5.4.2 Results

5.4.2.1 Block averaged standard errors

To estimate the statistical errors in the metrics calculated from the relative binding free energy estimates, i.e., the RMSDtr and Spearman's r_s, the standard error (SE) of these metrics is calculated using the block averaged standard error (BASE). Figure 5.8 shows the block averaged standard error (BASE) of the RMSDtr of predicted against experimental binding free energies using MM-GBSA and the force field combination ff19SB-OPC. The series labeled 0N represents the BASE when no explicit waters are considered in the MM-GBSA calculation. At a block size of 1 snapshot, the BASE is equivalent to the metric's analytic SE, which represents a lower bound on the statistical error in the calculated metric due to imperfect sampling. It corresponds roughly to the SE for the RMSDtr calculated by bootstrapping. The bootstrapped SE for the RMSDtr are 0.01-0.02 kcal/mol for MM-PBSA and -GBSA binding energies without explicit waters for the combined 4500 snapshot ensemble and between 0.02-0.04 kcal/mol for individual replicas with 1500 snapshots each. A more realistic estimate of the statistical error due to finite sampling in the RMSDtr is given by the value at which the BASE, as shown in Figure 5.8, plateaus with increasing block size (n). We use the maximum BASE value observed for $1 < \mathrm{n} < 900$ snapshots as the estimated upper limit of the statistical error in the calculated RMSDtr and correlation coefficients. The example BASE plot of figure 5.8 corresponds to an RMSDtr BASE of 0.07 kcal/mol for the MM-GBSA binding energies without explicit waters using the ff19SB-OPC3 force field combination. The full set of maximum BASE values are given in Tables B.1, B.2, B.3, and B.4 in Appendix B.

5.4.2.2 Binding energies without explicit waters

The RMSDtr and Spearman's rank correlation coefficient (r_s) are calculated for the MM-PBSA and -GBSA binding energies for force field combinations ff14SB-TIP3P, ff19SB-OPC, ff19SB-OPC3, and ff19SB-TIP3P to assess the accuracy and predictive power. Binding energies are calculated with and without the Interaction Entropy correction term. Table 5.3 shows the RMSDtr for the 8 binders and Spearman's r_s for all 10 ligands in BRD4. The maximum BASE are given in Appendix B in Tables B.1 to B.4.

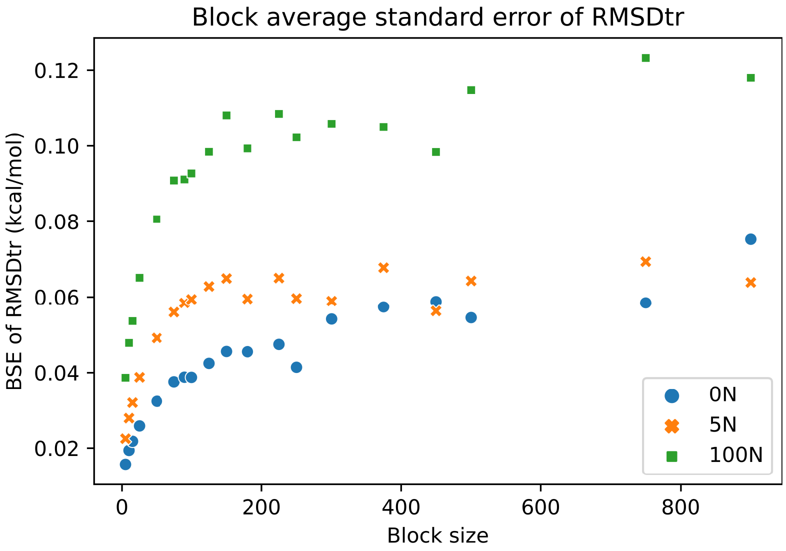


FIGURE 5.8: Block averaged standard error (BASE, kcal/mol) at different block sizes (n) of the root mean square deviation (RMSDtr) of the predicted MM-GBSA against experimental binding free energies using ff19SB-OPC3 across three replicas of 300 ns each. BASE is shown for binding energies without explicit waters (0N), 5 explicit waters (5N), and 100 explicit waters (100N).

Consider first the null-hypothesis that all ligands have the same absolute binding energy. Over the set of only binders, the null-hypothesis yields an RMSDtr of 1.45 kcal/mol due to the limited binding affinity range of 4.46 kcal/mol. Including the non-binders yields a null-hypothesis RMSDtr of 3.50 kcal/mol. The correlation coefficients of the null-hypothesis are always zero.

MM-PBSA RMSDtr values range from 2.35 kcal/mol for ff14SB-TIP3P to 3.28 kcal/mol for ff19SB-OPC when no entropy correction is included. Maximum BASE range from 0.07 to 0.16 kcal/mol across methods. The force field combination ff14SB-TIP3P achieves the best rank ordering for MM-PBSA with and without the IE correction term. However the calculated ligand rank order does not correspond well to the experimental results ($r_s = 0.32$ and $r_s = 0.53$). All three ff19SB-based force field combinations produce r_s of about 0, i.e., do not show predictive power in ordering the ligand binding energies. Maximum BASE in r_s range from 0.01 to 0.04.

When the Interaction Entropy correction term is added to the MM-PBSA results, RMSDtr values become significantly larger across all force field combinations. The r_s of ff14SB-TIP3P and ff19SB-OPC are slightly improved but suffer significant deterioration in RMSDtr by the inclusion of the IE term. BASE for the RMSDtr also increase, with maximum BASE between 0.17 and 0.7 kcal/mol.

For MM-GBSA, RMSDtr values range from 2.09 kcal/mol for ff19SB-OPC3 to 3.05 kcal/mol for ff14SB-TIP3P when no entropy correction is included. Maximum BASE values are, on average, 0.3 kcal/mol lower than for MM-PBSA. MM-GBSA rank

TABLE 5.3: Root mean squared deviation after the removal of systematic error (RMSDtr, kcal/mol) of the predicted against experimental binding free energies for the 8 binders in complex with BRD4 using MM-PBSA and -GBSA with different force field combinations. Spearman's r_s of rank ordering of calculated against experimental binding energies for all 10 ligands (binder and non-binder) in BRD4. Results with and without the Interaction Entropy correction term are shown. Binding energies were estimated from an ensemble of 4500 snapshots comprising 900 ns of MD (three replicas combined). No explicit waters are included in MM-PBSA and -GBSA. Corresponding maximum block averaged standard error (BASE) is shown in Tables B.1 to B.4.

Method	Force Field	Entropy	RMSDtr (kcal/mol)	Spearman's r
MM-PBSA	ff14SB-TIP3P	No	2.35	0.32
		IE	7.14	0.53
	ff19SB-TIP3P	No	2.65	-0.07
		IE	4.72	0.04
	ff19SB-OPC	No	3.28	0.00
		IE	5.05	0.26
	ff19SB-OPC3	No	2.96	-0.01
		IE	6.11	-0.06
MM-GBSA	ff14SB-TIP3P	No	3.05	0.81
		IE	8.74	0.74
	ff19SB-TIP3P	No	2.21	0.76
		IE	6.43	0.56
	ff19SB-OPC	No	2.19	0.78
		IE	5.14	0.67
	ff19SB-OPC3	No	2.09	0.81
		IE	5.18	0.53
Null Hypothesis			1.45	0.00

orderings are surprisingly good, with all force field combinations achieving $r_s > 0.7$. As for MM-PBSA, the inclusion of the IE correction term deteriorates the RMSDtr significantly and reduces r_s across all force field combinations.

Overall, the best performing method is MM-GBSA using ff19SB-OPC3 with RMSDtr of 2.09 kcal/mol and $r_s = 0.81$ and BASEs of 0.07 kcal/mol and 0.03, respectively. Interaction Entropy correction terms increase error and reduce correlation while significantly increasing BASE for both MM-PBSA and MM-GBSA.

5.4.2.3 Binding energies with explicit waters

When ligands are bound, the BRD4 binding site features several structural waters within the binding site. We explore the effect of including the N closest waters to the ligand as explicit waters in the MM-PBSA and MM-GBSA binding free energy calculations on the error against experiment (RMSDtr) and rank ordering of ligands (Spearman's r_s). Figures 5.9 and 5.10 illustrate the change in RMSDtr, for the set of 8 binders, with increasing numbers of explicit water included in the binding energy calculation. Figures 5.11 and 5.12 give the r_s ranking correlations with increasing

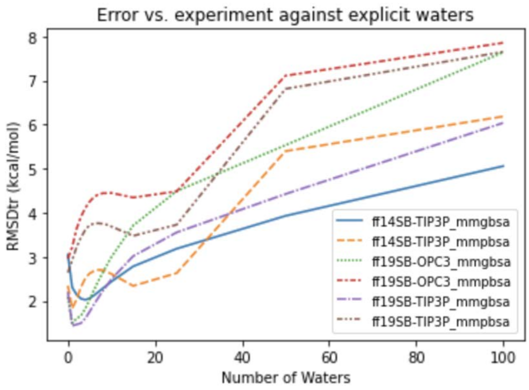


FIGURE 5.9: Error of predicted against experimental binding energies as RMSDtr (kcal/mol) at different numbers of explicit waters using different force field combinations with MM-PBSA and MM-GBSA.

number of explicit waters. Full tables of all RMSDtr and r_s values are provided in Tables B.6 to B.9 of Appendix B.

The ff19SB-OPC results have been intentionally omitted due to the severe effect of including explicit waters on the calculated binding energies for this force field combination. The inclusion of any explicit waters increases the RMSDtr values dramatically. For MM-GBSA the RMSDtr increases from 2.2 kcal/mol without water to 5.0 kcal/mol with just one explicit water and increases steadily to an RMSDtr of 110.0 kcal/mol when the 100 closest waters to the ligand are included. These errors indicate an underlying problem in the calculated energies, as further discussed in section 5.4.3.

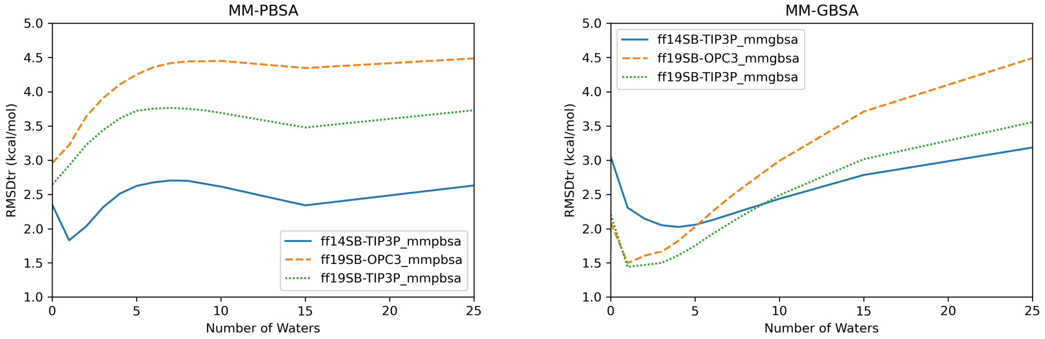


FIGURE 5.10: Error of predicted against experimental binding energies (RMSDtr, kcal/mol) at different numbers of explicit waters using different force field combinations with MM-GBSA (left) and MM-PBSA (right) without entropy correction.

Figure 5.9 clearly illustrates that the inclusion of 25, 50, or 100 closest waters in the binding energy calculation universally deteriorates the quality of predictions in terms of RMSDtr. RMSDtr values increase compared to the results without explicit waters for MM-GBSA and MM-PBSA for all force field combinations. At 100 explicit waters, all RMSDtr values are above 5 kcal/mol. For r_s, Figures 5.11 and 5.12 show that beyond 10 explicit waters, r_s become significantly worse as compared to those when no explicit waters are considered.

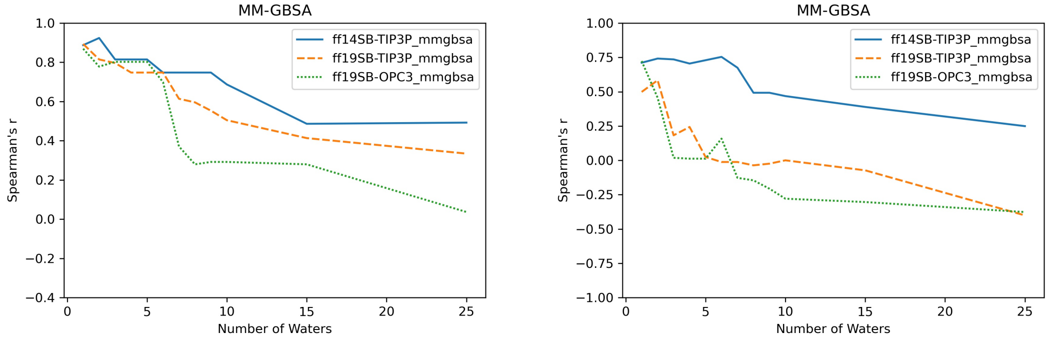


FIGURE 5.11: Spearman's r_s against number of explicit water molecules in MM-GBSA binding free energies without entropy correction (left) and with Interaction Entropy (right).

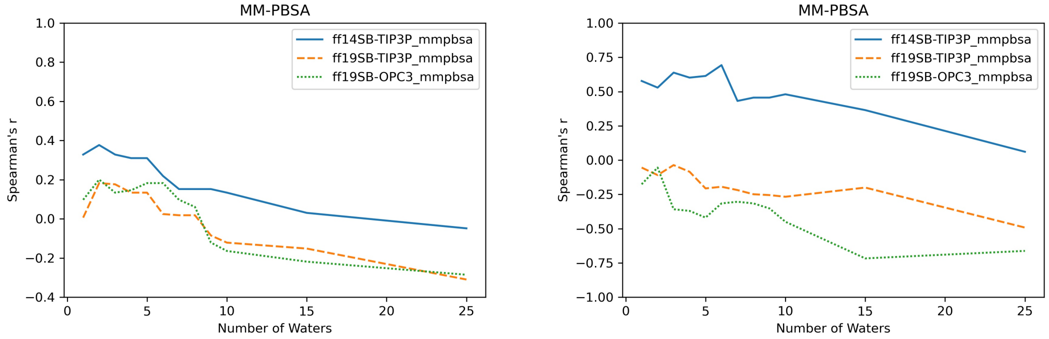


FIGURE 5.12: Spearman's r_s against number of explicit water molecules in MM-PBSA binding free energies without entropy correction (left) and with Interaction Entropy (right).

Figure 5.10 shows the RMSDtr of MM-PBSA and MM-GBSA with 0 to 25 explicit waters. Overall, MM-GBSA RMSDtr values are improved by including 1 to 5 explicit waters. This holds for ff19SB- and ff14SB-based force field combinations but with ff19SB-based combinations attaining significantly lower RMSDtr by $\approx$ 1kcal/mol. Beyond 8 explicit waters, RMSDtr values increase above the values without explicit waters. Considering the maximum BASE in the RMSDtr of 0.14 kcal/mol (Table B.1), these observed changes are significant. Including explicit waters does not systematically improve RMSDtr values for MM-PBSA energies. Only ff14SB+TIP3P shows a slight reduction in RMSDtr with 1 to 3 explicit waters.

Figure 5.11 and 5.12 show r_S with increasing numbers of explicit waters. When no entropy term is included, the inclusion of 1 to 5 explicit waters does not significantly alter the observed r_s. Beyond 5 explicit waters, rank ordering worsens across the board for both MM-PBSA and MM-GBSA. The inclusion of the IE correction term does not alter this behavior.

5.4.3 Discussion

We observe a significant dependence of the RMSDtr on the choice of force field combination in this MD study of BRD4. This observation is in line with the substantial differences in protein conformational motion between ff19SB and ff14SB, as discussed in Section 5.3. Overall, MM-GBSA produces significantly better ligand rank orderings than MM-PBSA, irrespective of the choice of force field. MM-GBSA obtains the most consistent results from ff19SB-based force field combinations yielding RMSDtr values of 2.21, 2.19, and 2.09 kcal/mol and r_s of 0.76, 0.78, and 0.81. MM-PBSA, on the other hand, appears incapable of ranking this ligand set in BRD4 with the computational protocol employed in this study.

Overall, the quality of binding energy estimates is not impacted significantly by the pairing of OPC, OPC3, or TIP3P with ff19SB. Correlation metrics between the choices of water model are indistinguishable for both MM-PBSA and MM-GBSA with or without the inclusion of explicit waters in the binding free energy estimates. This is in contrast to results by Heinzelmann et al. [165], whose protein-ligand binding energy predictions in BRD4 depended strongly on the choice of water model when using the ff14SB protein force field. Working on the same ligand set in BRD4 as our research, Huggins et al. [157] found that the RMSE of absolute binding energy predictions using ff14SB depended heavily on the choice of water model (TIP3P, SPCE, TIP4P-Ewald, and AM1-BCC ligand charge model).

An outlier from the above description is ff19SB+OPC, for which the inclusion of explicit waters significantly deteriorates the RMSDtr and correlation coefficients obtained by both MM-PBSA and -GBSA. We hypothesize this is because OPC is the only four-point water model tested and is currently incompatible with the implementation of MMPBSA.py in Amber20. We have reported this observation to the developers.

Including the IE term increases the RMSDtr values of all tested methods and force field combinations with and without the inclusion of explicit waters. Regarding the rank ordering of ligands, including the IE term does not lead to an overall improvement. Instead, the r_s values for almost all force field combinations with and without explicit waters worsen.

In terms of the closeness of predicted against experimental relative binding energies in this ligand set in BRD4, MM-GBSA benefits clearly from the inclusion of 1-5 waters and maintains low RMSDtr for 10 or less explicit waters. For MM-PBSA only the ff14SB+TIP3P force field has reduced RMSDtr upon inclusion of one or two explicit waters. The near-perfect rank ordering of MM-GBSA is maintained or slightly improved when 1-5 waters are included but generally deteriorates beyond 7 explicit waters. MM-PBSA rank ordering is slightly improved by including 1 to 5 waters but deteriorates significantly beyond 8 explicit waters. Including 25 or more explicit waters increases RMSDtr and reduces r_s in all instances. At 25 or more explicit waters the BASE in the RMSDtr also increases substantially (figure 5.8).

TABLE 5.4: Percentage occupation of water bridged interactions between ligand and protein residues 54, 97, 49, and 39. "None" is the percentage of snapshots in which no water bridge was detected. Based on 1500 snapshots extracted from 300ns of MD using ff19SB+GAFF2+OPC3.

Residues / Ligands	Tyr54	Asn97	none	Leu49	Pro39
1	54.27	0.20	45.47	0.00	0.00
2	38.93	18.73	40.40	3.47	0.00
3	25.93	2.80	54.67	14.13	0.00
4	71.40	7.20	23.80	0.00	0.00
5	34.27	4.53	59.40	4.60	0.00
6	33.13	1.00	53.93	2.60	1.00
7	48.87	0.13	51.00	0.00	0.00
8	55.67	0.93	18.93	0.13	33.47
9	26.67	6.20	41.60	32.07	0.00
10	7.33	11.73	55.73	22.60	0.00

Given these findings, the inclusion of 1 to 5 explicit waters in MM-GBSA binding energy calculations in BRD4 can be justified. The BRD4 binding site features 5-6 structural waters when ligands are bound [58], which may explain the improvement in predicted binding energies when including some or all of these waters as part of the protein. In 2021, Guest et al. [58] showed that 5-6 structural waters are conserved in most BRD4-ligand co-crystal structures and docking poses. Additionally, they showed that free energy perturbation (FEP) binding energy predictions are improved with the explicit inclusion of these waters. They observed a water bridged hydrogen bond forming between the ligands and the Asn97 residue of the protein. Based on this observation by Guest et al. we used cpptraj in Amber20 to find and track the occurrence and up-time of bridging water interactions between the ligand and protein in our input trajectories. Table 5.4 shows the percentage of snapshots for which a water bridge was found between the ligand and protein. Only residues with water bridge occupation of more than 10% are included. This analysis confirms that water-mediated interactions play an important role in our ligand set in BRD4. Consider, for example, ligands 8 and 4, in which water bridged hydrogen bonds are observed in about 80% of snapshots. The most prominent water-bridged interaction in our ligand set is between residue Tyr54 and the ligand. While ligand 1 maximally exhibits one concurrent bridging interaction, the other ligands can have 2-4 concurrent bridging interactions between the protein and ligand.

Our findings add to a growing volume of research that supports the use of some explicit waters in MM-GBSA and MM-PBSA protein-ligand binding free energy calculations. In 2013, Maffucci et al. [54] demonstrated significant improvement in ligand ranking by including 20 explicit waters in MM-GBSA calculations for topoisomerase and penicillopepsin. Mikulskis et al. [51] showed an improvement when including explicit waters in MM-GBSA for 9 ligands binding to ferritin. Also, in 2014, Zhu et al. [56] improved MM-PBSA calculations by including 1-7 explicit waters in a series of JNK3 kinase inhibitors. In 2016, Maffucci et al. [55] demonstrated improvements in protein-protein binding calculations by including explicit waters in MM-GBSA. Most recently, Rodriguez et al. [173] used MM-GBSA with 8 explicit waters to investigate the binding of olonine in BRD4.

In 2017, Aldeghi et al. [57] performed explicit water MM-PBSA on 11 ligands in BRD4 using the Nwat method by Maffucci et al. [54, 55]. Their Interaction Entropy correction term did not improve ligand rank ordering, and they also observed a significant increase in statistical error. In contrast to our findings, rank orderings were slightly improved by including 20 to 50 explicit waters. Aldeghi's MM-PBSA produced far better ligand ranking than we observed for MM-PBSA in our ligand set, despite 5 common ligands between the sets. Snapshots for MM-PBSA were generated from Hamiltonian-exchange Langevin dynamics with a simulation length of only 10 ns and three replicas. Simulations were performed in Gromacs using ff99SB-ILDN force field for the protein, GAFF1.5 for the ligand, and the TIP3P water model. MM-PBSA was performed in GMXPBSA on 501 snapshots. A solute dielectric constant of 2 was used compared to a value of 1 in our study. This choice was made as previous studies on other protein-ligand systems had shown improved correlations at higher dielectric constants for MM-PBSA [70]. Furthermore, a non-zero salt concentration was used by Aldeghi et al., as well as a different surface tension for the nonpolar term in the solvent accessible surface area term of the PBSA solvation model. It is well established that MM-PBSA and -GBSA are sensitive to even small changes in the simulation protocol in general [70], and in BRD4 [157]. The comparison of Aldeghi's MM-PBSA results and our own clearly illustrates this.

5.4.4 Conclusion

In this section, the end-point binding free energy methods MM-PBSA and MM-GBSA were applied to the MD trajectories generated for ff14SB- and ff19SB-based force field combinations. The effect of the Interaction Entropy correction term was tested and the inclusion of increasing numbers of explicit waters. These classical mechanical protein-ligand binding free energy predictions in BRD4 serve as the benchmark against which the quantum mechanical binding energies computed in the rest of this chapter are compared.

As in Section 5.3, the conclusions of this MM-PBSA/GBSA study of BRD4 cannot be naively extended to other protein-ligand systems. We are not advocating MM-GBSA with 1-5 explicit waters as a ″go-to″ method in protein-ligand binding energy prediction. The results from this section serve as a point of comparison for the following quantum mechanical study and highlight some important concepts in end-point binding free energy prediction. The choice of force field can, as for the protein motion, significantly impact estimated binding energies. Additionally, including some explicit waters appears to be beneficial in BRD4 and corroborates other studies investigating explicit water MM-PBSA/GBSA. The Interaction Entropy correction term falls short and does not improve estimated energies. A more detailed discussion of the IE term is provided as part of the QM-PBSA study below.

The method of choosing and including explicit waters employed in this study is very basic and leads to several problems. While the number of water molecules is the same in every snapshot, the individual waters present may be different molecules. Selecting waters based on their proximity to the ligand also leads to the inclusions of waters outside the binding site, which are unlikely to play an important role in ligand binding. A more sophisticated approach should select all bridging waters and only those waters inside the binding site. Mikulskis et al. [51] proposed a more general approach of including different numbers of explicit waters in each snapshot and using weighted averages when calculating the final binding energies. They also attempt to capture the effect of the displacement of waters during binding with the MM-GBSA approach.

Lastly, it is worth noting that it has been suggested that using a dielectric constant higher than 1 inside the solute can improve MM-PBSA/GBSA binding energy predictions [70]. We chose a value of 1 in this study because there has been no research exploring the dielectric constant in QM-PBSA binding energies. Thus, the default value of 1 was chosen for both QM-PBSA and MM-PBSA/GBSA to allow direct comparison of results.

5.5 QM-PBSA binding energies

Having thoroughly explored our ligand set in BRD4 at the classical mechanical level of theory, in this chapter, the quantum mechanical QM-PBSA binding free energies are calculated and compared to their classical mechanical counterparts. Based on Section 5.3, the force field combination ff19SB-GAFF2-OPC3 was selected. This force field combination led to the most stable MD trajectories with no ligand dissociation and did not lead to the ZA-loop conformational change often observed with ff14SB. We compare quantum mechanical QM-PBSA with traditional classical mechanical MM-PBSA and MM-GBSA discussed in Section 5.4. Our comparison focuses on the direct comparison of QM-PBSA and its classical analog MM-PBSA. We compute entropy correction terms using normal mode analysis (NMA) and Interaction Entropy (IE) [62]. Finally, we compare the results in BRD4 with our previous QM-PBSA study of the much simpler T4-lysozyme [1] protein binding with 7 ligands from Chapter 4.

5.5.1 Methods

5.5.1.1 Design of computational study

900ns of MD for the force field combination ff19SB-GAFF2-OPC3 generated in Section 5.3 are used as the basis for the binding energy calculations. From the combined 900 ns trajectory for each ligand, 4500 snapshots are extracted and binding energies evaluated using MM-PBSA and MM-GBSA as in Section 5.4. A subset of 50 snapshots, equally spaced in time, is taken from the 4500, and QM-PBSA binding energies are calculated for all ligands. Normal mode entropy estimates are calculated for the subset of 50 snapshots only. Interaction Entropy terms are calculated for MM-PBSA and MM-GBSA across all 4500 snapshots. For the QM-PBSA, MM-PBSA, and MM-GBSA results over 50 snapshots, the Interaction Entropy from only those 50 snapshots is considered.

As in Section 5.4, the root mean squared deviation after removal of the systematic error (mean signed error), called RMSDtr, calculated over the set of binders and Spearman's rank correlation factor r_s over the whole ligand set are used to assess the quality of the computed binding energies. The treatment of the non-binder, ligand 1, and weak binder, ligand 2, is consistent with the MM study in Section 5.4.

Estimates of the statistical error due to finite sampling in the computed metrics are obtained using the block averaged standard error (BASE) method [182] introduced in Section 5.4.1.1. In addition to the BASE, the standard error in the computed metrics is calculated using bootstrapping (with replacement) over 1000 iterations to provide a second estimate of statistical error.

5.5.1.2 Computational details

QM-PBSA, DFT QM energy evaluations are carried out by the linear-scaling DFT code ONETEP [87]. The general purpose exchange-correlation functional PBE [77] and D3 dispersion correction [142] were found to be the most promising in our previous study on the T4-lysozyme [1]. A kinetic energy cutoff of 800eV is used. 4 non-orthogonal generalized Wannier functions (NGWFs) are used for carbon, nitrogen, and oxygen, and 1 NGWF is used for hydrogen. For sulfur, fluorine, and bromine 9 NGWFs are used. An NGWF radius of 8.0 atomic units is used throughout. ONETEP default parameters for water at room temperature are used. The default solvent surface tension is $4.7624 \times 10^{-5}\mathrm{Ha/Bohr}^2$ with an apolar scaling factor of 0.281075 and a solvation β of 1.3. The bulk permittivity is 78.54, and an interior dielectric of 1 is used. To speed up the solution of the Poisson-Boltzmann equation, the charge at the boundary of the simulation cell is coarse-grained. The default charge coarse-graining factor at the boundary is 5, but by increasing this to 10, the energy evaluation of the complex is sped up by 20%. A link to QM-PBSA input and output files is included in Appendix C.

Normal mode analysis Normal mode analysis is performed in Amber20 using MMPBSA.py and default settings. The complex, host, and ligand's vibrational, translational, and rotational entropies are evaluated. Before NMA, MMPBSA.py performs a two-part energy minimization comprised of a conjugate gradient method, followed by the Newton-Raphson method on each snapshot with tight convergence criteria. A Hawkins, Cramer, Truhlar (HCT) Generalized Born implicit solvent with an internal dielectric of 1 is used for the frequency calculations and the energy minimizations with infinite non-bonded cutoff.

5.5.2 Results

5.5.2.1 Convergence of energy terms

TABLE 5.5: Standard error of the mean (SEM) in kcal/mol for the absolute binding energy (G_bind), net gas phase energy (Gas-phase) and net solvation energy (G_solv) averaged across the ligand set for QM-PBSA and MM-PBSA over 50 snapshots.

Method / SEM (kcal/mol)	QM-PBSA	MM-PBSA
G_bind	0.44	0.47
Gas-phase	0.64	0.65
G_solv	0.40	0.63

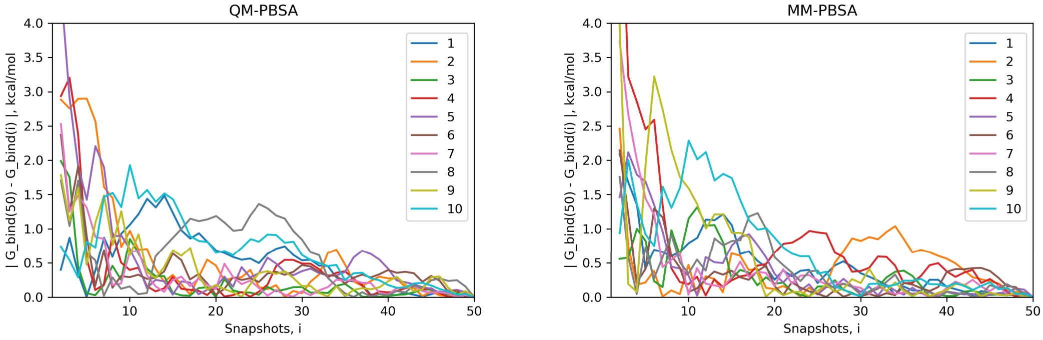


FIGURE 5.13: Absolute deviation of computed absolute binding energies at different numbers of snapshots from the computed energies at 50 snapshots for QM-PBSA (left) and MM-PBSA (right) in kcal/mol for all ligands.

The standard error of the mean (SEM) measures how far a sample's mean is likely to deviate from the true population mean. The MM-PBSA and QM-PBSA methods average each energy term across the ensemble of snapshots, i.e., the population sample. Table 5.5 shows the SEM, averaged over the ligand set, for the absolute binding energies, net gas phase energies, and net solvation energies for QM-PBSA and

MM-PBSA. Net energies are the difference between the complex and its constituents. The SEM of the absolute binding energies and net gas phase energies over 50 snapshots of QM-PBSA and MM-PBSA are almost identical. For the net solvation energy, QM-PBSA has a slightly larger SEM by 0.2 kcal/mol. Figure 5.13 shows the absolute deviation of the computed absolute binding energies at increasing numbers of snapshots from the "converged" value at 50 snapshots with QM-PBSA and MM-PBSA. The convergence behavior is comparable between MM and QM, and beyond 30 snapshots, the absolute deviations are below 0.5 kcal/mol for most ligands. The mean absolute difference in absolute binding energies across the ligand set between MM-PBSA at 50 and 4500 snapshots is 0.36 kcal/mol. This means that on average, if all 4500 snapshots are included, the calculated binding energies differ by 0.36 kcal/mol from those calculated over only 50 snapshots.

5.5.2.2 Absolute binding free energies

TABLE 5.6: QM-PBSA absolute gas phase energies and binding energies without entropy and with Interaction Entropy and normal mode correction terms in kcal/mol. The non-binder (ligand 1) and weak binder (ligand 2) are allocated experimental binding energies of 0 kcal/mol.

Ligand	Gas-Phase	G_bind	G_bind(IE)	G_bind(NMA)	Experiment
1	-36.1	-34.3	-29.4	-19.4	0
2	-45.2	-37.3	-27.0	-20.9	0
3	-45.8	-44.1	-37.6	-26.2	-5.95
4	-32.6	-26.2	-15.0	-10.7	-6.36
5	-44.5	-40.4	-33.8	-23.3	-7.40
6	-55.3	-42.7	-36.2	-23.8	-7.84
7	-48.4	-45.2	-40.0	-26.7	-8.16
8	-42.9	-42.5	-34.7	-24.8	-8.99
9	-47.2	-46.4	-40.2	-28.4	-9.64
10	-47.7	-46.9	-38.5	-27.8	-10.4

While end-point binding free energy methods like QM-PBSA and MM-PBSA are only suitable for calculating relative free energies of binding, an explicit look at the absolute binding energies can prove informative. Table 5.6 shows the net gas-phase energy and absolute binding energies for each ligand without entropy correction and with normal mode and Interaction Entropy corrections. Ligand 4, pictured in Figure 5.15, is a clear outlier and is excluded from the RMSDtr and Spearman's r_s analysis. The absolute binding energy of ligand 4 without entropy is 16 kcal/mol more positive (weaker) than the average across the rest of the ligand set and 8 kcal/mol weaker than the known non-binder, ligand 1. A full discussion of ligand 4 is presented in Section 5.5.3.2 of the discussion. Additionally, Table 5.6 illustrates that QM-PBSA should not be used to calculate absolute binding energies.

5.5.2.3 Predictive power

The RMSDtr and Spearman's r_s are used to quantify the closeness of calculated to experimental relative binding energies and rank ordering of ligands by binding energy. Table 5.7 shows the RMSDtr and Spearman's r_s values for QM-PBSA, MM-PBSA, and MM-GBSA. Results are shown for the ensemble subset of 50 snapshots evaluated using QM-PBSA, and MM results are additionally shown across the entire ensemble of 4500 snapshots. Metrics are calculated without entropy, normal mode entropy, and Interaction Entropy. As normal mode calculations for only 50 snapshots were performed, no results for MM-PBSA/GBSA over 4500 snapshots with normal mode correction are shown. BASE and bootstrapped SE for each metric are shown in Table B.5 in Appendix B. Ligand 4, the outlier, is excluded from the analysis in this section and further discussed in Section 5.5.3.2.

Consider first the results without entropy correction. QM-PBSA has an RMSDtr of 1.77 kcal/mol and r_s of 0.83. This is 1.22 kcal/mol lower than MM-PBSA and 0.89 kcal/mol lower than MM-GBSA, over the same 50 snapshots, and thus a statistically significant improvement with respect to the estimated statistical error of less than 0.2 kcal/mol. MM-PBSA has no predictive power regarding rank ordering the ligands ($r_s = 0.01$). The ligand rank order produced by MM-GBSA is comparable with that of QM-PBSA. Considering the full 4500 snapshots, MM-PBSA does not change significantly in terms of RMSDtr or r_s. MM-GBSA, on the other hand, shows an improved RMSDtr of 2.07 kcal/mol, which is 0.3 kcal/mol larger than that of QM-PBSA over 50 snapshots.

Overall, the inclusion of the normal mode entropy correction term improves the performance of QM-PBSA, reducing the RMSDtr by 0.17 kcal/mol and increasing r_s by 0.02. However, these changes are within the estimated uncertainties of the two metrics. This is in part due to the fact that the inclusion of normal mode entropy increases the BASE and bootstrapped SE for both RMSDtr and r_s. The inclusion of normal mode entropy further deteriorates MM-PBSA. MM-GBSA shows a significant drop in r_s of 0.26 while improving RMSDtr by 0.49 kcal/mol, which is, however, of comparable magnitude as the BASE of 0.47 kcal/mol.

Including the Interaction Entropy correction increases RMSDtr significantly and reduces r_s across the board. The increase in error against experiment caused by the Interaction Entropy term is much larger in the set of 4500 snapshots than in the set of only 50 snapshots. The BASE in RMSDtr increases significantly from 50 to 4500 snapshots when the IE term is included.

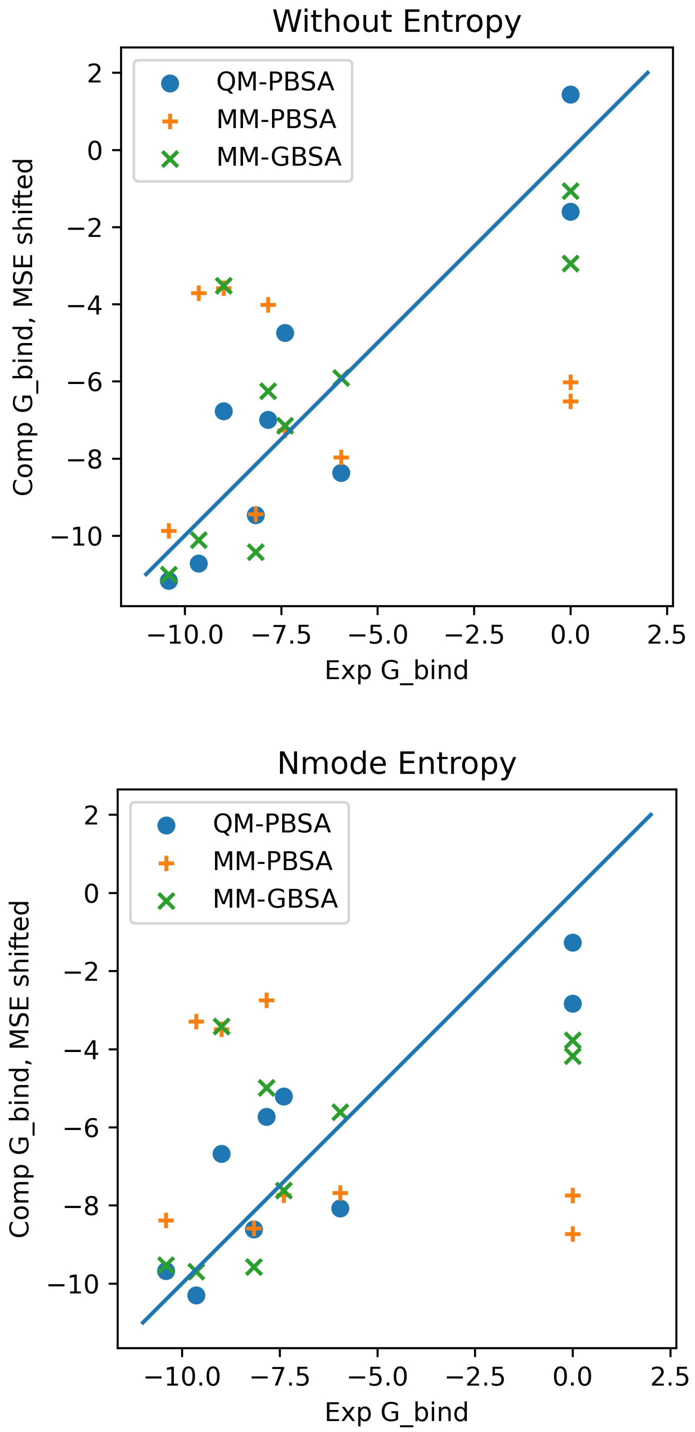


FIGURE 5.14: Plots of computed binding energies (shifted by mean signed error) against experimental binding energies for 9 ligands in BRD4 for QM-PBSA, MM-PBSA and MM-GBSA.

TABLE 5.7: Root mean squared deviation of calculated binding energies shifted by the mean signed error (RMSDtr) against the experimental binding energies of 7 binders in BRD4 in kcal/mol. Spearman's rank order coefficient, r_s, of calculated binding energies against the experimental binding energies of 9 ligands (binders and non-binders) in BRD4. Metrics are shown for QM-PBSA, MM-PBSA, and MM-GBSA without entropy and with normal mode and Interaction Entropy term. Metrics for MM are shown over 50 and over the complete set of 4500 snapshots. Block average standard errors and bootstrapped standard errors are provided in Table B.5 in Appendix B. Normal mode entropies are only calculated for 50 snapshots.

Snapshots	Method	Entropy	RMSDtr (kcal/mol)	Spearman's r_s
50	QM-PBSA	No Entropy	1.77	0.83
		Normal Mode	1.60	0.85
		Interaction Entropy	2.30	0.76
	MM-PBSA	No Entropy	2.99	0.01
		Normal Mode	3.05	-0.30
		Interaction Entropy	3.84	-0.13
	MM-GBSA	No Entropy	2.66	0.79
		Normal Mode	2.17	0.53
		Interaction Entropy	3.22	0.48
4500	MM-PBSA	No Entropy	2.89	-0.03
		Normal Mode	NA	NA
		Interaction Entropy	6.18	0.03
	MM-GBSA	No Entropy	2.07	0.78
		Normal Mode	NA	NA
		Interaction Entropy	5.54	0.53

5.5.3 Discussion

5.5.3.1 QM-PBSA

In the context of this set of ligands in BRD4, QM-PBSA produces relative binding free energies that are closer to experiment and have better rank ordering than MM-PBSA binding energies. MM-GBSA performs better than MM-PBSA but is still inferior to QM-PBSA in terms of RMSDtr and rank ordering. Normal mode entropies improve QM-PBSA results slightly but deteriorate MM-PBSA results significantly. In our hands, Interaction Entropy corrections reduce the quality of estimated binding energies across the board and are difficult to converge, mirroring observations by other authors [183–185].

The ligand rank orderings produced by QM-PBSA are very good, with $r_s = 0.85$ when normal mode entropy is included and BASE and bootstrapped SE of 0.08 and 0.06, respectively. Figure 5.14 shows plots of the experimental against computed binding energies where the computed energies have been shifted by the mean signed error as in the calculation of the RMSDtr. Plots are shown with and without normal mode entropy and for the whole ligand set (excluding ligand 4). Visual comparison of QM-PBSA with MM-PBSA also clearly shows an improvement in predictive power.

With an RMSDtr of 1.60 kcal/mol and r_s of 0.85, QM-PBSA with normal mode entropy has comparable accuracy against experiment and predictive power as MM-GBSA with the inclusion of 1-3 explicit waters and no entropy correction, which was the best performing method in Section 5.4.

We see similar results in comparing the results of this BRD4 study to our previous QM-PBSA study in T4-lysozyme. In T4-lysozyme RMSDtr values of 1.84 kcal/mol were obtained using the PBE exchange-correlation functional and D3 dispersion and normal mode entropy compared to an RMSDtr of 1.60 kcal/mol in BRD4. Interestingly, the standard error (SE) values for the absolute binding energies and net solvation energies are slightly lower, by about 0.3 kcal/mol, in BRD4 than in T4-lysozyme at 50 snapshots of sampling. This is because BRD4 has a solvent-exposed binding site and does not require a cavity-correction term to account for the buried cavity binding site in the T4-lysozyme [1, 38]. The cavity-correction term was the largest source of standard error in the T4-lysozyme system. Both in BRD4 and T4-lysozyme, the QM energies converge at the same rate as the MM energies, indicating that, for these systems, evaluating structures generated using classical mechanical force fields with a quantum mechanical energy Hamiltonian is viable. If this observation is shown to hold for more protein-ligand systems, the QM-PBSA or similar approaches can provide a realistic and accessible method of exploring protein-ligand binding at a quantum mechanical level of theory.

In 2017, Aldeghi et al. [57] studied 11 ligands in BRD4 using both MM-PBSA and an absolute alchemical binding free energy method. The single-trajectory MM-PBSA approach achieved Spearman's $r_s = 0.72$ without entropy and 0.61 with an entropy correction term. Absolute alchemical energies improved upon this with $r_s = 0.85$. Also, in 2017, Heinzelmann et al. [165] applied the attach-pull-release method to 7 ligands binding to BRD4, achieving Kendall τ between 0.33 and 0.50 and RMSE of 1.14-3.21 kcal/mol depending on the details of the method. In 2022 Guest et al. [168] applied FEP and multi-site lambda dynamics (MSλD) in BRD4 and reported Spearman's r_s of 0.7 and 0.8, respectively, for 14 compounds. Average accuracy against experiment for FEP was reported as 1.0 ± 1.3 kcal/mol and 0.7 ± 0.5 kcal/mol after the exclusion of one outlier. Also, in 2022, Huggins et al. [157] applied an alchemical binding free energy method to the 8 binders from the Mobley BRD4 benchmark set also used in this study. Using the ff14SB protein force field, the previous iteration of the ff19SB force field used in this study, they reported RMSE of 1.44 to 3.36 kcal/mol and Kendall τ of 0.31 to 0.56 depending on the choice of water model (TIP3P, TIP4P-Ewald, and SPCE) and ligand charge model (AM1-BCC and RESP). Notably, ligand 4, which was excluded from our ligand set as an outlier for the QM-PBSA method, was not an outlier for the alchemical method employed by Huggins et al. [157]

In comparison, our QM-PBSA method achieves RMSDtr of 1.6 kcal/mol across 7 binders and an r_s of 0.85 across the 9 ligands (binders and non-binders) in BRD4. These results align with those of more thermodynamically rigorous classical mechanical methods; however, as summarized in Section 5.5.3.4, the computational cost of QM-PBSA is orders of magnitude larger.

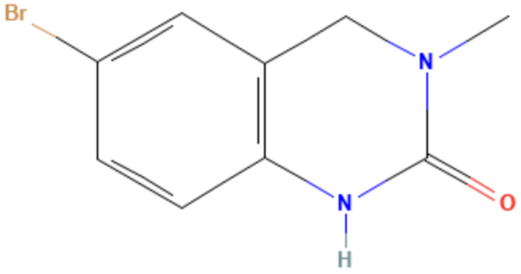


FIGURE 5.15: Ligand 4, the outlier.

5.5.3.2 The outlier

The computed QM-PBSA binding energies of ligand 4, the outlier, are 16 kcal/mol more positive than the average binding energy across the ligand set, while the experimental binding energy for ligand 4 is essentially in the middle of the binding set. We discuss here the attempts made to explain and correct this outlier.

Ligand 4, pictured in Figure 5.15, is the smallest ligand in the set and undergoes the most motion during the MD simulation as measured by the ligand RMSD. During the 900 ns of MD, several similar binding poses are sampled by ligand 4. We re-ran the 900 ns MD for ligand 4 with a 2 kcal mol^{-1} Å^{-2} harmonic restraint imposed on the ligand in an attempt to sample only the binding mode closest to the initial configuration. 10 restrained snapshots were evaluated by QM-PBSA but the resulting absolute binding energies for ligand 4 differed only by 0.5 kcal/mol from those of the un-restrained trajectory. Thus, we concluded that sampling multiple related binding modes in itself is not the cause of ligand 4's significantly underestimated QM binding energies.

Since ligand 4 has the only bromine atom in the test set, we examined whether the DFT calculations might treat the bromine incorrectly. We replaced the bromine in ligand 4 with a hydrogen and re-equilibrated and re-ran 900ns of MD. This altered ligand, ligand 4H, is a real binder with similar binding energy as the bromine variant [158]. After evaluating three snapshots of ligand 4H with QM-PBSA, we found that the net difference in absolute binding energies by changing the bromine to a hydrogen was only -0.52 kcal/mol. Thus we abandoned this hypothesis.

It is well established that the BRD4 binding site features a number of structural waters, which are observed both in crystal structures and simulation [58]. Using Amber20's cpptraj we analyzed the bridging waters between the ligands and BRD4 protein. The occupations percentages for each water bridge are shown in Table 5.4 of Section 5.4.3. We observed that for ligand 4 the bridging water between the ligand and the protein residue Tyr54 is highly conserved and present in more than 70% of snapshots with the same water molecule being conserved over many nanoseconds of MD. While bridging waters are present in the trajectories for the other ligands, the average occupation is less than 30 % with no single ligand, except ligand 4, having an occupation of more than 55% for the Tyr54 water bridge. Additionally, the mean number of water bridges per snapshot is 1.2 for ligand 4, while all other ligands, except ligand 6, have mean values $<<$ 1. We hypothesized that the small and loosely bound ligand 4 requires the bridging water interactions with the protein to bind. The explicit inclusion of these bridging waters in the QM-PBSA calculation may 'fix' the underestimated QM binding energies. We performed QM-PBSA over 50 snapshots for ligand 4 in which any bridging waters between the ligand and protein were explicitly included as part of the protein in the DFT calculations. This approach is analogous to explicit water MM-PBSA, which has successfully been applied in several studies [51, 53–58].

Our own MM-PBSA/GBSA calculations with explicit waters in the Section 5.4 showed that including a few waters can improve the quality of predicted binding energies. While the net gas-phase energy of our explicit water QM-PBSA for ligand 4 did become 10 kcal/mol more favorable to binding, the solvation and normal mode entropy became less favorable for binding. As a result, the entropy corrected calculated binding energy of ligand 4 with explicit bridging waters was only about 1 kcal/mol stronger than the original estimate. Thus the inclusion of the highly conserved bridging water in ligand 4 did not bring its outlier energy into line with the rest of the ligand set. Two reported X-ray structures of ligand 4 in BRD4 (PDB:4HBV) and the highly conserved CREBBP bromodomain (PDB:4NYV) show subtle differences in the binding modes of the ligand and changes in the structure and number of waters in the binding site. A potential issue is that our dynamics simulation does not sufficiently sample the different arrangements of structural waters in the binding site.

5.5.3.3 Interaction Entropy

Including an IE correction term increases RMSDtr and worsens rank ordering in all cases. The negative impact of the IE term is especially pronounced when all 4500 snapshots are included in the MM-PBSA and MM-GBSA calculations. Overall, the IE correction also increases the BASE and bootstrapped SE significantly. Figure 5.16 shows the calculated IE correction for MM-PBSA as increasing numbers of snapshots are included, up to 4500. The sharp increases in IE indicate that individual snapshots with peak interaction energies dominate the IE results. We also observe that the IE

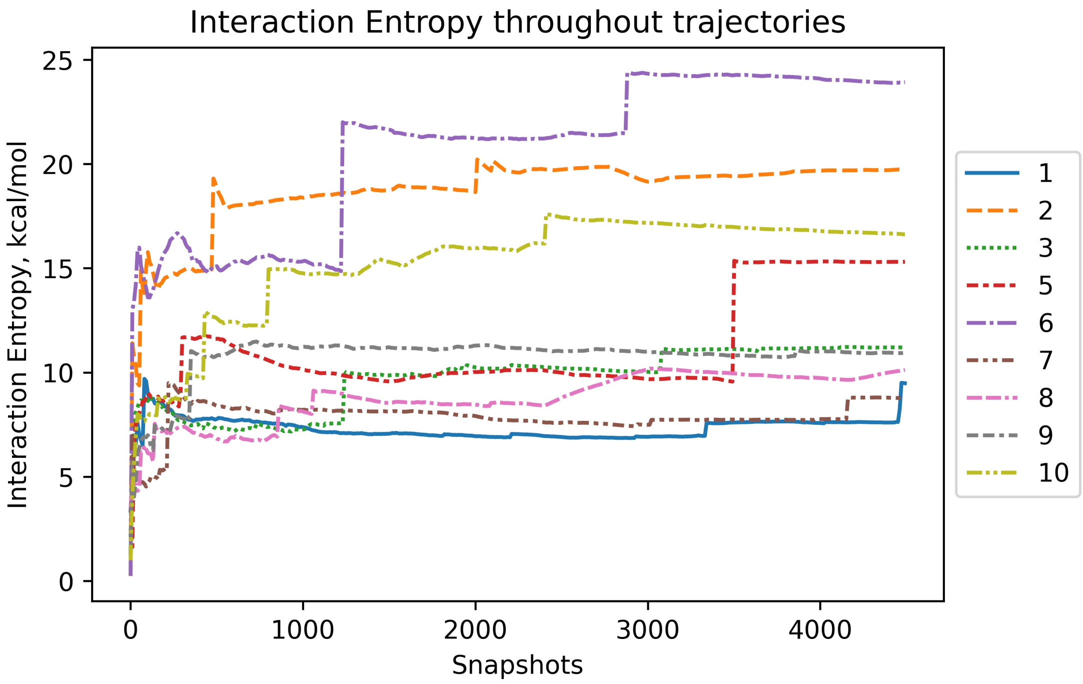


FIGURE 5.16: Interaction Entropy correction term in kcal/mol for MM-PBSA as increasing numbers of snapshots are included for each ligand binding to BRD4.

increases as more snapshots are included for almost all ligands. This observation is in line with the findings of Ekberg et al. [185], who showed that the IE increases with the number of snapshots sampled and that the exponential average is poorly conditioned. They suggest that when the standard deviation of the interaction energy is larger than 3.5 kcal/mol, the Interaction Entropy method becomes almost impossible to converge. In our test set, the standard error of the interaction energy is 4.8 kcal/mol when averaged across the ligand set. Thus, according to Ekberg et al., the IE method is not applicable. Similarly, Menzer et al. [184] observed that rare events dominate the IE term. Also, in 2018, Kohut et al. [183] showed that IE depends strongly on the highest peak interaction energies and does not significantly change after peak energy has distorted the IE term. This behavior is mirrored in our results in Figure 5.16. Kohut et al. concluded that the IE method is not applicable, especially when conformational changes occur.

5.5.3.4 Computational cost

To calculate quantum mechanical protein-ligand binding energies using the QM-PBSA method on a total of 10 ligands binding to BRD4 using 50 snapshots per ligand required about 7.3 million core-hours on AMD EPYC 7742 62-core processors on the ARCHER2 supercomputer. A single solvated DFT single-point energy evaluation on the BRD4 protein (2035 atoms) has a wall-time of about 5 hours on 8 nodes with 128 cores each (dual socket AMD EPYC). Given access to the whole ARCHER2 supercomputer, the entire QM-PBSA BRD4 study could have been completed in less than 10 hours due to the trivially parallel nature of QM-PBSA. Including additional calculations and initial testing, we estimate a total usage of 10 million core-hours on ARCHER2 and IRIDIS5 (University of Southampton supercomputer).

5.5.4 Conclusion

This study is, to our knowledge, the first application of full-protein DFT binding energy calculations on a real-world, pharmaceutically relevant protein and ligand set. Building on our QM-PBSA validation study in T4-lysozyme [1], we demonstrate the application of QM-PBSA in BRD4. We find a significant improvement in accuracy against experiment and ligand rank ordering over classical mechanical MM-PBSA. The best results are obtained by including a normal mode entropy correction term. As in the T4-lysozyme study, the QM binding energies appear equally converged as the MM binding energies with SEM < 0.5 kcal/mol at 50 snapshots of sampling. Exploring whole protein-ligand complexes at a quantum mechanical level of theory is both computationally and methodologically viable and opens a variety of opportunities for further investigation like the potential applications of extracting further information from the full-QM electronic densities for protein-ligand systems.

Chapter 6

Conclusions

In this body of research, we have investigated the use of full-protein density functional theory calculations to predict protein-ligand free energies of binding. The accurate prediction of ligand binding energies and their relative rank ordering is of central importance in the growing field of computational drug discovery and design. Based on our explorative study of the T4-lysozyme protein system [1], we selected optimal parameters for applying the quantum mechanical QM-PBSA binding free energy method. In the first application of a DFT-based full-protein QM binding energy method on a pharmaceutically relevant protein, BRD4, we demonstrated near-perfect ligand rank ordering and errors relative to experiment of 1.6 kcal/mol when the normal mode entropy term is included. The quantum mechanical binding energies appear similarly converged as their classical mechanical counterparts. The QM-PBSA calculations in BRD4 used molecular dynamics trajectories from our in-depth molecular dynamics study of BRD4, in which the trajectory stability, protein motion, and ligand dissociation were investigated for several force field combinations. Additionally, classical MM-PBSA and MM-GBSA binding energies for multiple force field combinations were calculated. Including explicit water molecules as part of the end-point binding free energy calculations was considered along with an entropy correction term.

In this final chapter, we outline the state of the QM-PBSA method after three years of active research, illustrate avenues of continued research and showcase some current advances in the wider field of quantum mechanical protein-ligand binding energy prediction. We conclude with some closing remarks.

6.1 The state of QM-PBSA

Having applied the QM-PBSA quantum mechanical binding energy method to two distinct proteins and ligand sets, we summarize the key findings, outline some outstanding issues, and highlight avenues for continued research.

The biggest challenge of studying protein-ligand interaction at a quantum mechanical level of theory is the large size of proteins. Even when working with only the active regions of a protein, like in the first bromodomain of the bromodomain containing protein 4, the systems under investigation still have thousands of atoms and tens of thousands of electrons. By choosing linear-scaling DFT as our quantum method, we can perform thousands of full-protein DFT calculations on our protein systems of interest. Our T4-lysozyme and BRD4 QM-PBSA studies have shown beyond doubt that full-protein DFT calculations are viable and can be performed at scale in an academic research environment. One of our core missions is to communicate this capability with the wider computational chemistry community to promote more research in this field.

The second major challenge in quantum mechanical protein-ligand binding energy prediction is the convergence of the calculated binding energies. Because sampling is performed at the classical mechanical level, the QM energy evaluation in QM-PBSA is performed on a classical ensemble of snapshots. While this is undeniably a theoretical weak point of our methodology, we can show in T4-lysozyme and BRD4 that the convergence of the QM-PBSA binding energies mirrors that of the MM-PBSA and MM-GBSA binding energies. Because it is not possible to evaluate the true convergence of the predicted binding energies in an absolute sense, it is promising that our QM-PBSA method achieves the same convergence characteristics as the popular and much-applied classical mechanical MM-PBSA. Additionally, our extended sampling of 100 QM-PBSA snapshots in T4-lysozyme indicates that 50 snapshots of sampling per ligand are sufficient to obtain reliable binding energy estimates. The BRD4 study confirmed this. It is, however, unclear if this observation is transferable to other protein-ligand systems. Fundamentally, the QM-PBSA methodology needs to be applied to more protein-ligand systems to fully characterize its convergence across different types of proteins and ligand sets.

We have obtained mixed results regarding the predictive power of QM-PBSA. In T4-lysozyme, the accuracy against experiment, measured by the RMSDtr, is comparable between QM-PBSA and MM-PBSA. In BRD4, QM-PBSA outperforms its classical analog MM-PBSA significantly. However, MM-GBSA, with the inclusion of a few explicit waters, results in comparable RMSDtr and r_s at a fraction of the computational cost. Again, QM-PBSA must be benchmarked against more systems to form a full image of its potential to predict more accurate and reliable protein-ligand free energies of binding.

A core piece of peer-review feedback obtained in the publication of our research was that it is difficult to tell to what extent the methodological improvement of MM force field to DFT in the gas-phase energies can improve overall binding energies when the approximate implicit solvation scheme and normal mode estimation remain unaltered. While the minimal parameter implicit solvation model in ONETEP is more involved and accurate than traditional PBSA solvation [43], the simplification of implicit solvation remains severe, as illustrated by our findings regarding the role of explicit waters in the BRD4 binding site. Given the improvements in MM-GBSA binding energies by including some explicit waters in BRD4, including explicit waters in QM-PBSA is a potential avenue to improve the methodology. Due to the linear-scaling, adding a few explicit water molecules would have no noticeable impact on the computational cost of QM-PBSA.

Nonetheless, the approximations inherent in the formulation of MM-PBSA and QM-PBSA may prove too severe to systematically improve the accuracy of predicted binding energies by treating the individual terms of the method with more accurate methodologies. In two reviews of hybrid QM/MM-PBSA methods, Ryde et al. [39] and Söderhjelm et al. [8] both express the sentiment that MM-PBSA is not accurate enough for predictive drug design, and that its approximations are too severe to be used as the basis of more accurate methods like QM-PBSA. On the other hand, the methodological ease of modifying MM-PBSA to QM-PBSA, the low sampling requirements, and the trivially-parallel nature of this end-point method makes it an attractive starting point for quantum mechanical approaches.

The research volume in quantum mechanical protein-ligand free energy of binding prediction is still minimal. In 2022, excluding the author's work, we found only four publications on QM protein-ligand binding energies. Vennelakanti et al. [9] published an opinion piece about the future of large-scale QM and QM/MMM for predictive modeling in enzymes and proteins. Maier et al. [186] benchmarked a QM molecule-in-molecules (MIM) approach, which partitions the protein system into small, overlapping fragments on which independent QM calculations are performed. The energies of the fragments are then recombined to recover the total energy. They reported improvements over traditional MM-PBSA/GBSA and Pearson correlation coefficients between 0.81 and 0.97. However, no errors against experiment were given, and the method only applies to sets of structurally similar ligands. Additionally, no sampling is performed in the MIM approach; instead, single energy-minimized crystal structures are used to compute binding energy estimates. This eliminates the possibility of capturing multiple binding modes, binding site flexibility, and ligand flexibility. Chen et al. [187] applied the GFN-FF force field and the family of GFN-xTB SEQM methods to 90 protein-ligand complexes. They truncated the protein for the SEQM calculations and tested different truncation radii. GFN2-xTB was the best performing SEQM method tested with mean absolute errors after removing the systematic error of 7 kcal/mol for charged systems and 5 kcal/mol for neutral systems. The correlations of the SEQM approach and traditional MM-PBSA were comparable. Lastly, Kirsopp et al. [188] computed quantum mechanical protein-ligand interactions natively on a quantum computer. They studied 12 inhibitors to the BACE1 protein on different quantum computers and a simulated quantum computer. They obtained coefficients of determinants, R^2, of 0.55, 0.77, and 0.56, depending on the QM-hardware. In comparison, a DFT-based approach achieved $R^2 = 0.65$.

6.2 Future work

In the absence of resource limitations, the QM-PBSA method should be tested on various other protein-ligand benchmark systems to confirm its convergence characteristics and assess its predictive power. The protein systems BACE-1, FXR, CDK2 Kinase, SYK Kinase, CathepsinS, and Stromelysin-1 all have publicly available ligand sets with co-crystal structures and experimental binding energies and are between 1700 and 3300 atoms large. A full QM-PBSA benchmarking study of the roughly 100 ligands across the protein systems listed in the previous sentence would require just under 100 million core-hours on the current generation AMD EPYC CPUs. This estimation was made in the context of an application to compute time on Europe's largest supercomputer. At the time of application, we had only completed the QM-PBSA study on T4-lysozyme. In the final decision, we were asked to investigate an additional, more realistic protein system before embarking on a large-scale benchmarking study. We expect the promising results obtained in our subsequent QM-PBSA study of BRD4 will strengthen the case for a large HPC grant, which would be required to perform QM-PBSA on multiple protein systems. We have developed the technological infrastructure to deploy and manage such volume calculations.

Thus far, the binding energies have been considered the only output of the expensive QM-PBSA calculations. While the accurate prediction of protein-ligand binding affinities is the ultimate goal, the binding energy as a single number gives little insight into the binding mechanism. In every solvated DFT single-point energy evaluation performed in QM-PBSA, the full-protein electron densities and electrostatic potentials are calculated and written to the disk. These electron densities are, in principle, a full QM representation of the protein-ligand system. It encodes all the physical interactions described by DFT. From our BRD4 study presented in Chapter 5 we have access to full-protein QM electron densities for 50 configurations each of BRD4 in complex with 10 ligands. These electron densities represent a novel and rich data source to understand the mechanisms of protein-ligand binding at the quantum mechanical level of theory. By systematically identifying regions of change between the electron densities of the protein in complex with different ligands or competing binding poses of the same ligand, the concrete changes in electron behavior could be studied and visualized. However, in our experience, the electron densities remain very similar, and changes are often tiny and difficult to detect in the huge three-dimensional density matrix. We propose that machine learning algorithms for image recognition could be applied to detect similarities and differences between electron densities. These algorithms take two or three-dimensional matrices with intensities, or RGB values, at each point as input parameters and may identify patterns that a difficult to discover in such an extensive three-dimensional data set manually.

In addition to their direct study, the electron densities and potentials could be used as input for the training of classical mechanical force fields like the bespoke-QM force field QUBE [189]. The QUBE force field uses quantum mechanical calculations to derive system-specific bespoke force field parameters. While the approach is currently still limited to small molecules, the electron densities and potentials of full-protein calculations could be incorporated into the training data in the future to improve, for example, protein residue-specific dihedral parameters. While the field of machine-learned molecular mechanics force fields is still mostly considering small molecules, as the field advances to protein force fields, the QM electron densities and potentials produced by QM-PBSA binding energy estimations may become a valuable source of training data [190].

6.3 Closing remarks

While the scientific output of our research efforts has been described in detail in Chapters 4 and 5, we would like to highlight some additional aspects. There is, to our knowledge, no precedent for the volume of full-protein DFT calculations performed for this body of research. Including exploratory calculations, we estimate that more than 10000 full-protein DFT calculations were performed. To achieve this, we developed scalable solutions for managing and running many concurrent calculations and utilities to collect, store, and analyze the simulation outputs. Many of these have been made available through the supporting information of our publications.

A key objective of our research was to demonstrate to the wider computational chemistry community that treating full proteins at a quantum mechanical level of theory is viable and tractable for individual research groups. Through conferences, our publications, and collaborations with other researchers and industry, we have aimed to motivate and empower others to invest their time into the prospects of quantum mechanical protein-ligand simulations.

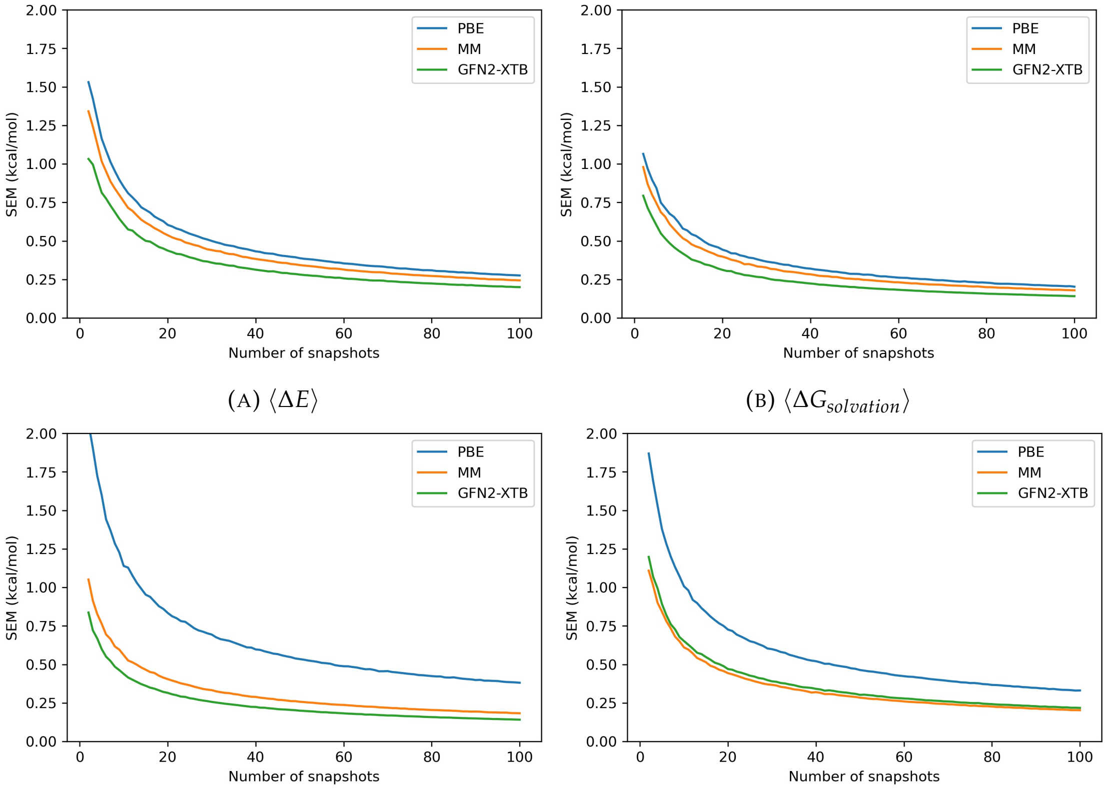
It is our firm belief that if a fraction of the scientific effort that has been invested into developing highly advanced and complicated classical mechanical methods for estimating protein-ligand binding over the past 60 years were to be invested in quantum-based approaches, significant improvements in accuracy, transferability, and domain of applicability could be achieved given the availability of modern high-performance computing resources.

Appendix A

T4-Lysozyme

Figures A.1-A.7: The standard error of the mean (SEM) calculated by bootstrapping (1000 re-samples) of the change upon binding in the gas-phase energy, $\langle \Delta E \rangle$, solvation energy, $\langle \Delta G_{solvation} \rangle$, cavity-corrected solvation energy, $\langle \Delta G_{solvation-cav-cor} \rangle$, and total enthalpy, $\langle \Delta H_{bind} \rangle = \langle \Delta E \rangle + \langle \Delta G_{solvation} \rangle$, for each ligand up to 100 snapshots for MM, the DFT functional PBE and the SEQM method GFN2-XTB.

Figures A.10-A.16: Absolute deviation of $\langle \Delta H_{bind} \rangle$ at different numbers of randomly selected snapshots from the 'converged' mean over 100 snapshots. Two different sets of random snapshots shown in blue and orange.



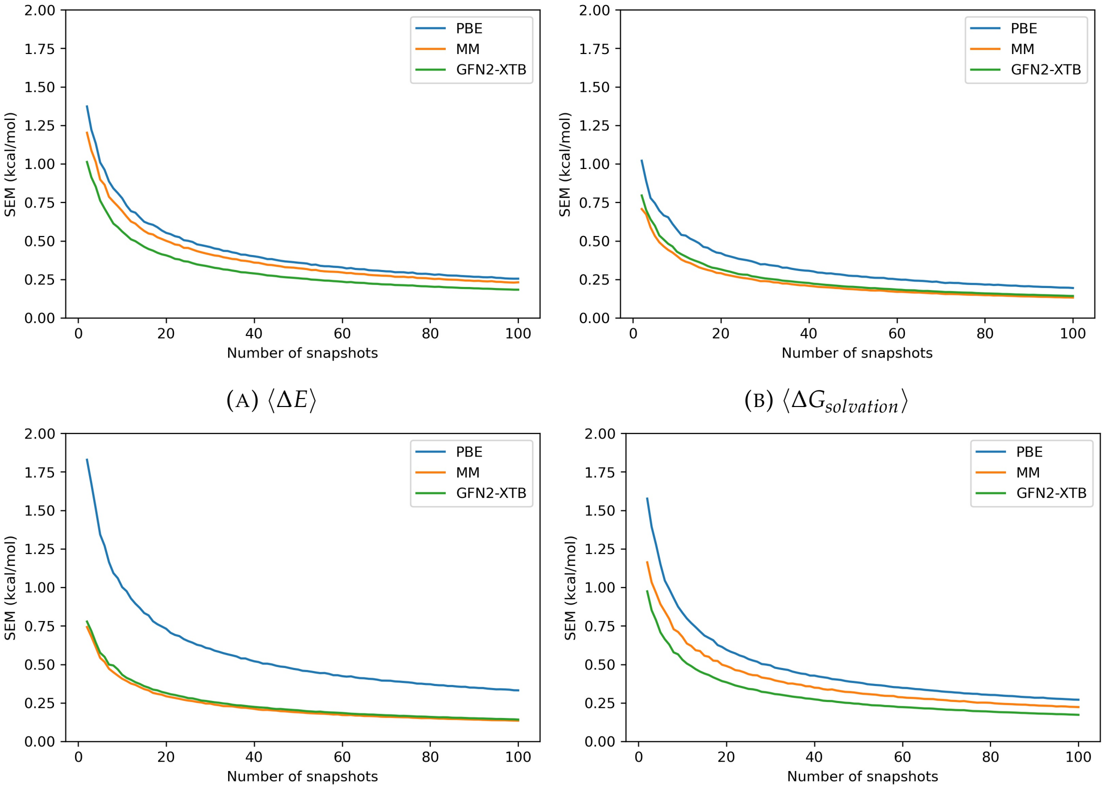
(A) $\langle \Delta E \rangle$

(B) $\langle \Delta G_{solvation} \rangle$

(C) $\langle \Delta G_{solvation-cav-cor} \rangle$

(D) $\langle \Delta H_{bind} \rangle$

FIGURE A.1: catechol



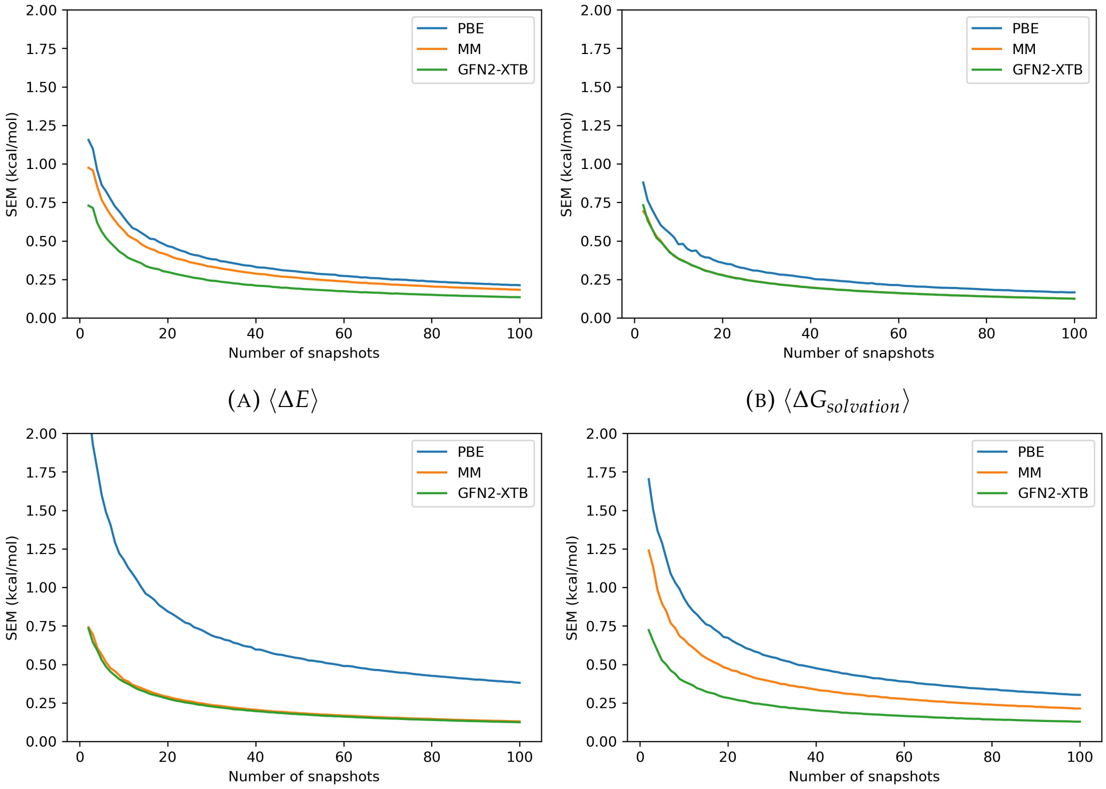
(A) $\langle \Delta E \rangle$

(B) $\langle \Delta G_{solvation} \rangle$

(C) $\langle \Delta G_{solvation-cav-cor} \rangle$

(D) $\langle \Delta H_{bind} \rangle$

FIGURE A.2: phenol



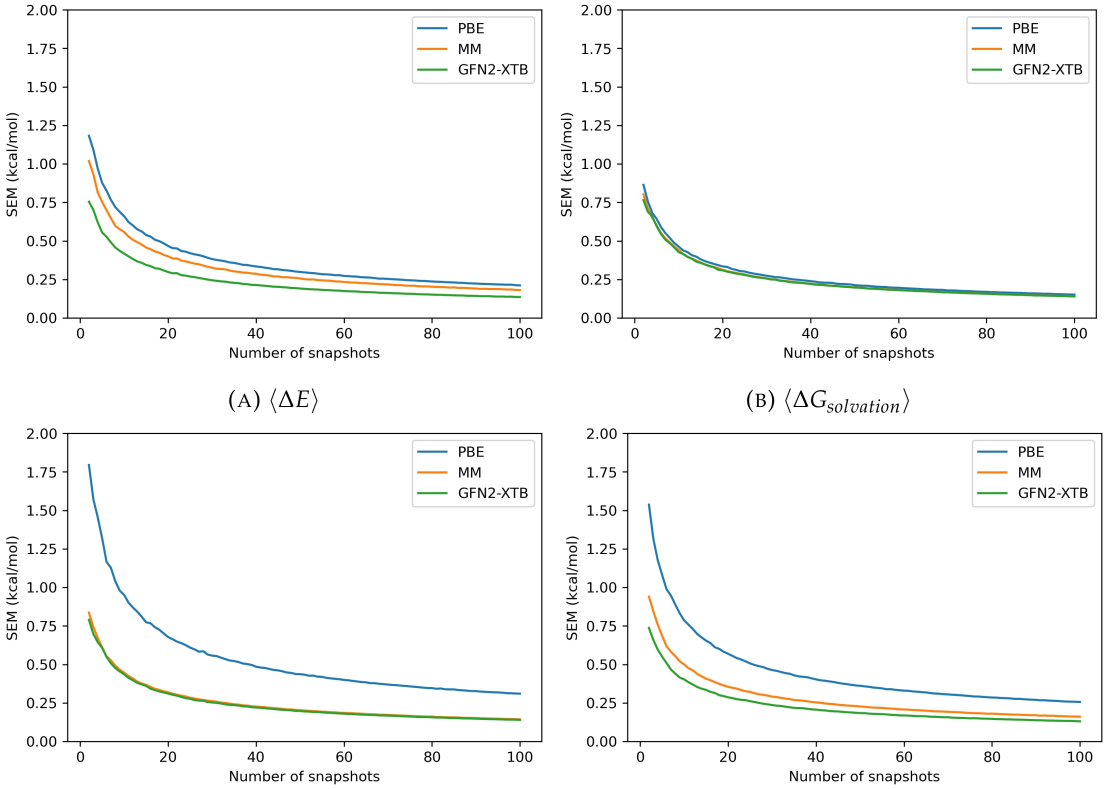
(A) $\langle \Delta E \rangle$

(B) $\langle \Delta G_{solvation} \rangle$

(C) $\langle \Delta G_{solvation-cav-cor} \rangle$

(D) $\langle \Delta H_{bind} \rangle$

FIGURE A.3: fluoroaniline



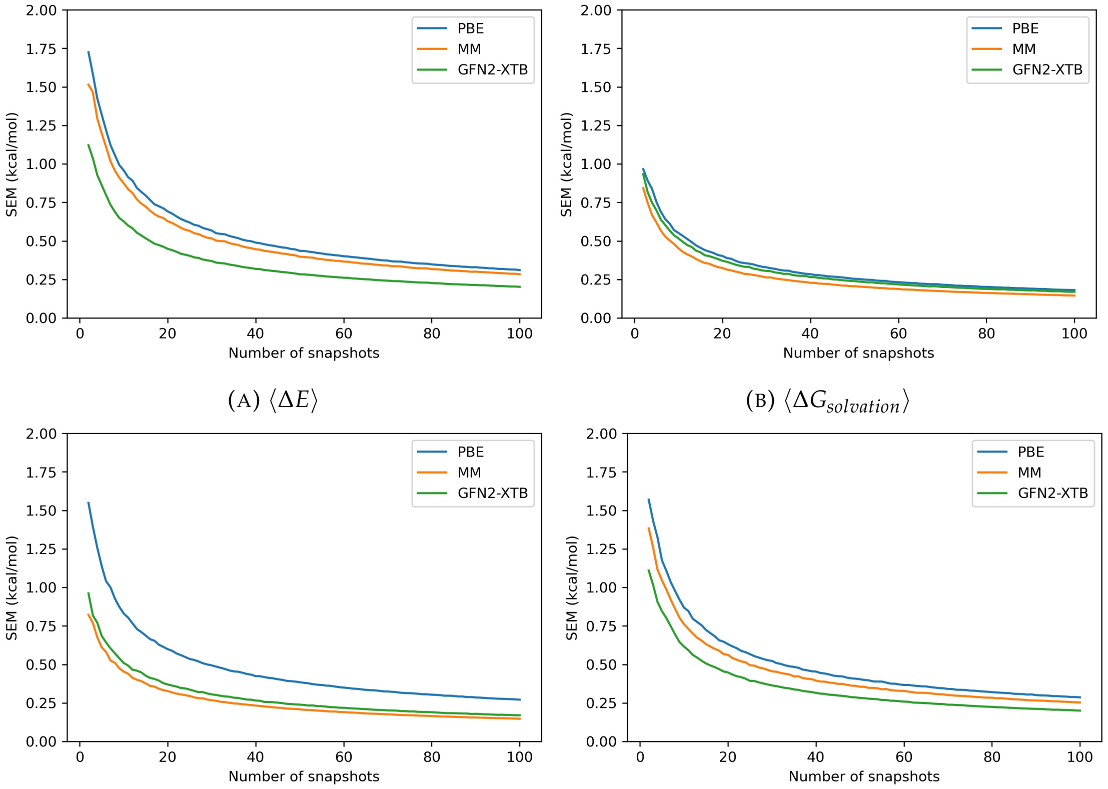
(A) $\langle \Delta E \rangle$

(B) $\langle \Delta G_{solvation} \rangle$

(C) $\langle \Delta G_{solvation-cav-cor} \rangle$

(D) $\langle \Delta H_{bind} \rangle$

FIGURE A.4: methylphenol



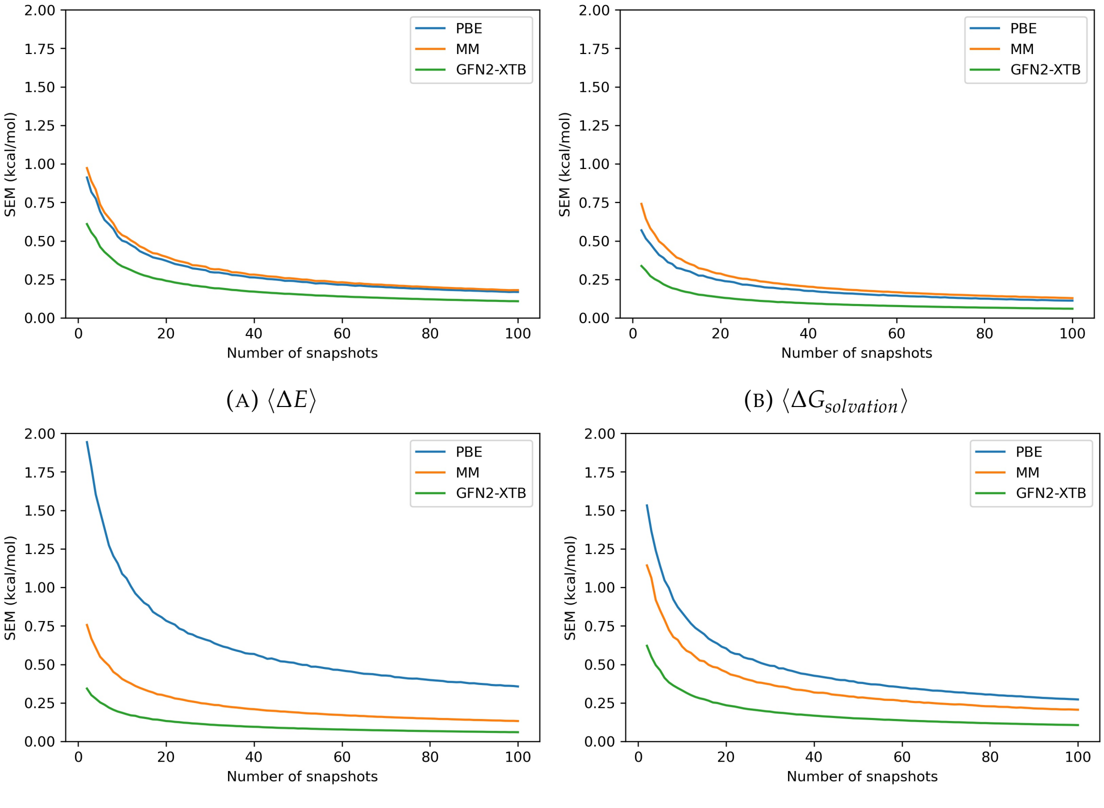
(A) $\langle \Delta E \rangle$

(B) $\langle \Delta G_{solvation} \rangle$

(C) $\langle \Delta G_{solvation-cav-cor} \rangle$

(D) $\langle \Delta H_{bind} \rangle$

FIGURE A.5: hydroxyaniline



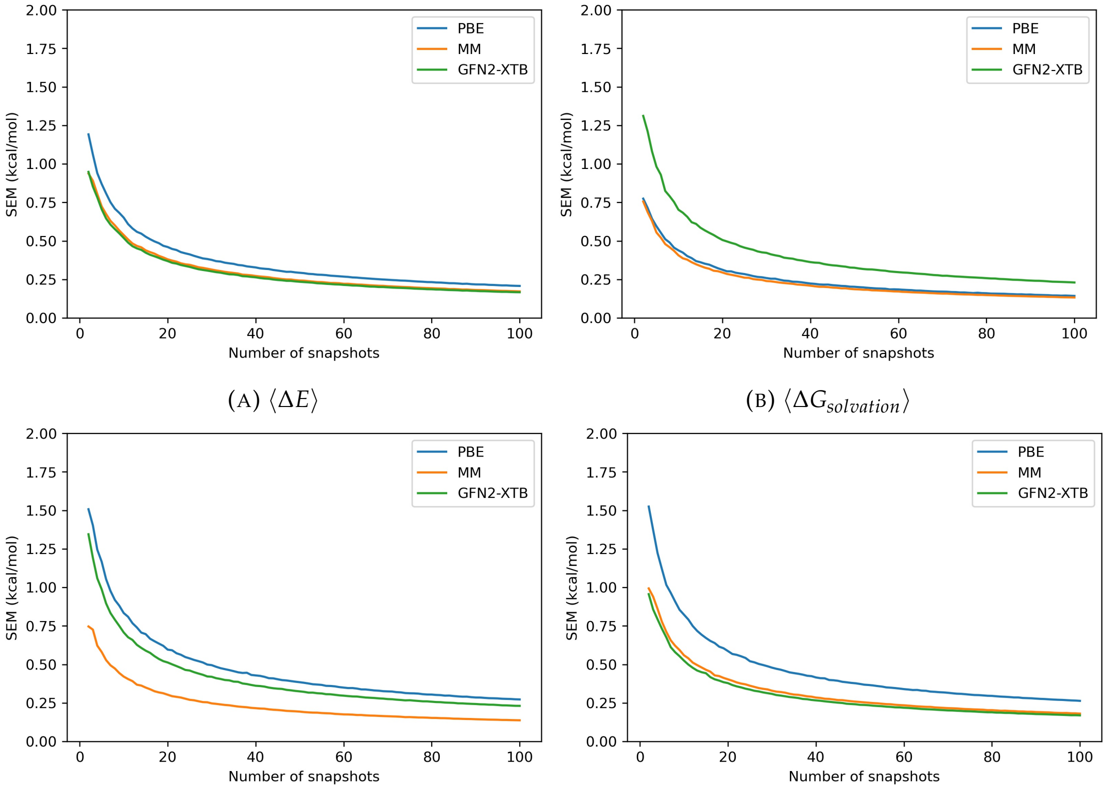
(A) $\langle \Delta E \rangle$

(B) $\langle \Delta G_{solvation} \rangle$

(C) $\langle \Delta G_{solvation-cav-cor} \rangle$

(D) $\langle \Delta H_{bind} \rangle$

FIGURE A.6: toluene



(A) $\langle \Delta E \rangle$ (B) $\langle \Delta G_{solvation} \rangle$

(C) $\langle \Delta G_{solvation-cav-cor} \rangle$ (D) $\langle \Delta H_{bind} \rangle$

FIGURE A.7: chlorophenol

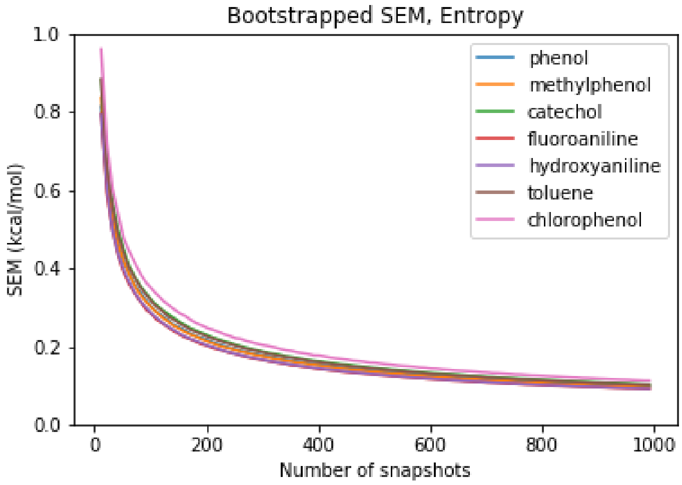
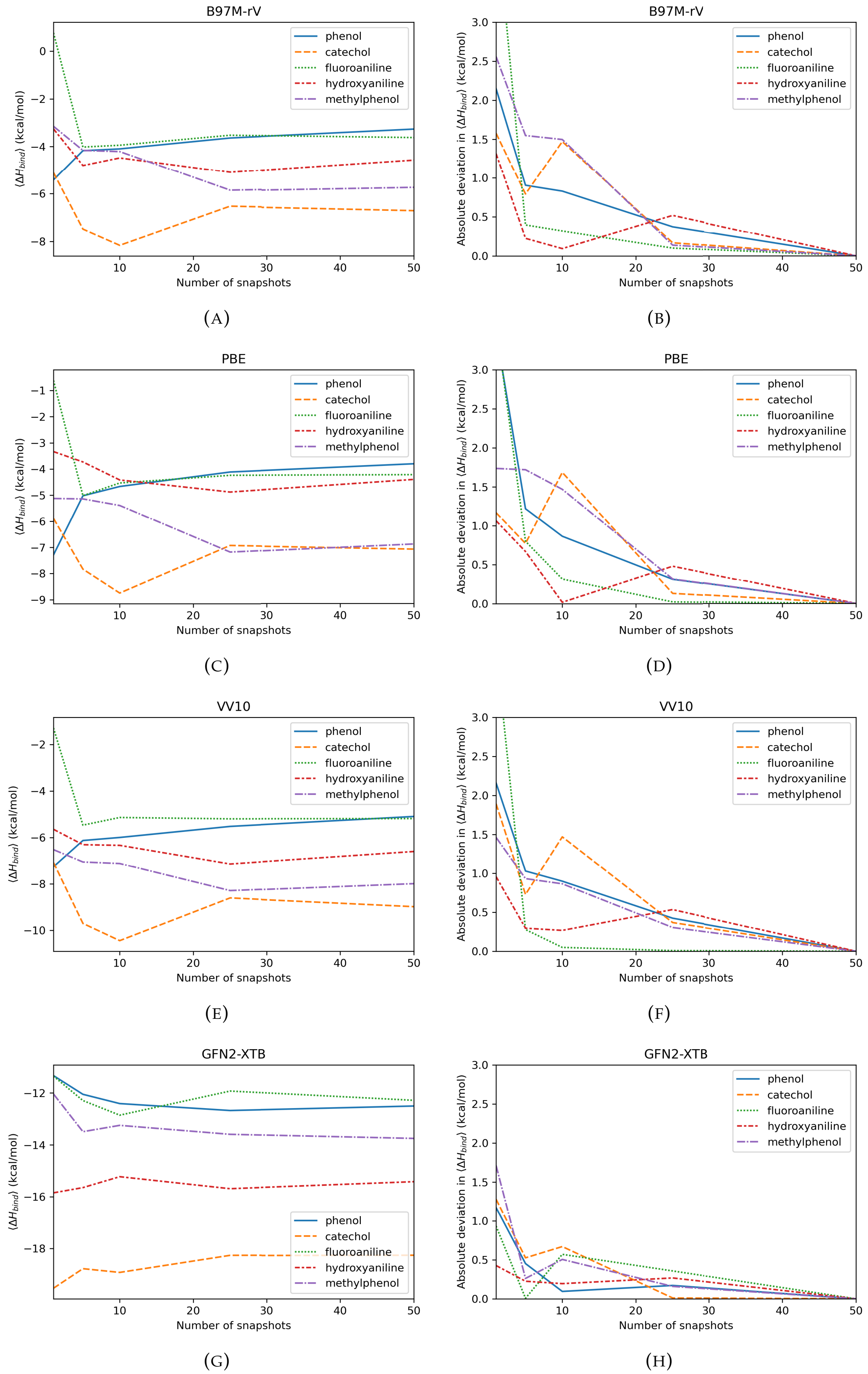


FIGURE A.8: Bootstrapped (1000 re-samples) SEM of entropy correction term, $T \times \langle \Delta S \rangle$, up to 100 snapshots for each ligand.



(A)

(B)

(C)

(D)

(E)

(F)

(G)

(H)

FIGURE A.9: Left: Mean change in total enthalpy upon binding,$\langle \Delta H_{bind} \rangle$, of each ligand at different numbers of equally spaced snapshots. Right: Absolute deviation of $\langle \Delta H_{bind} \rangle$ at different numbers of equally spaced snapshots from the 'converged' mean over 50 snapshots. Methods: B97M-rV (a,b),PBE (c,d), VV10 (e,f), GFN2-XTB (g,h).

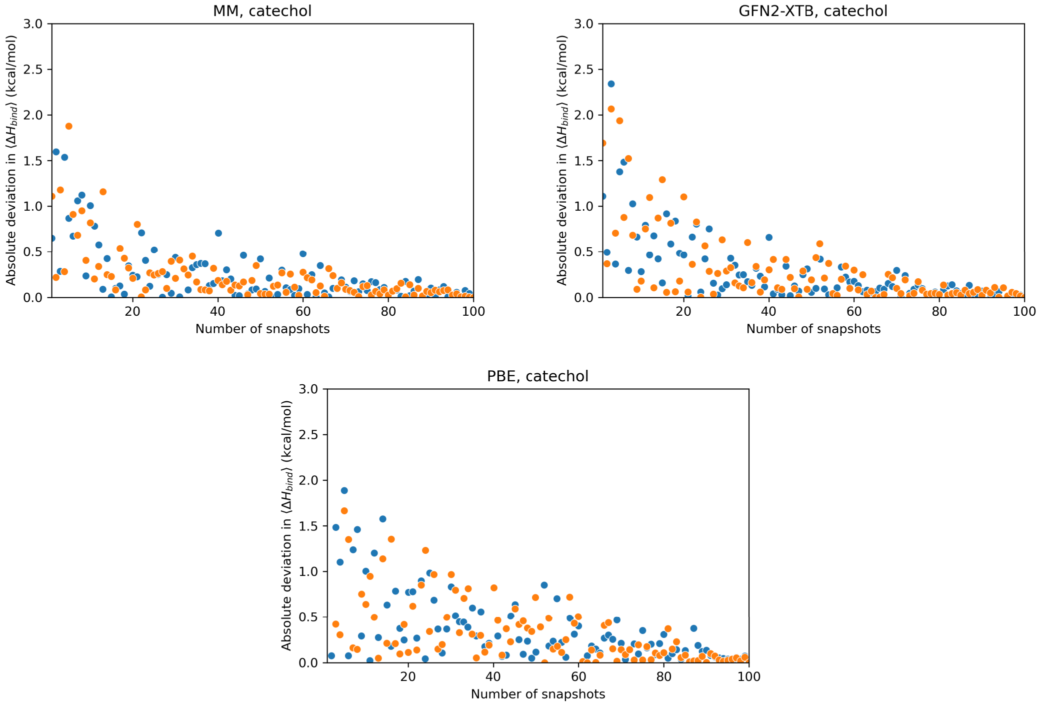


FIGURE A.10: catechol

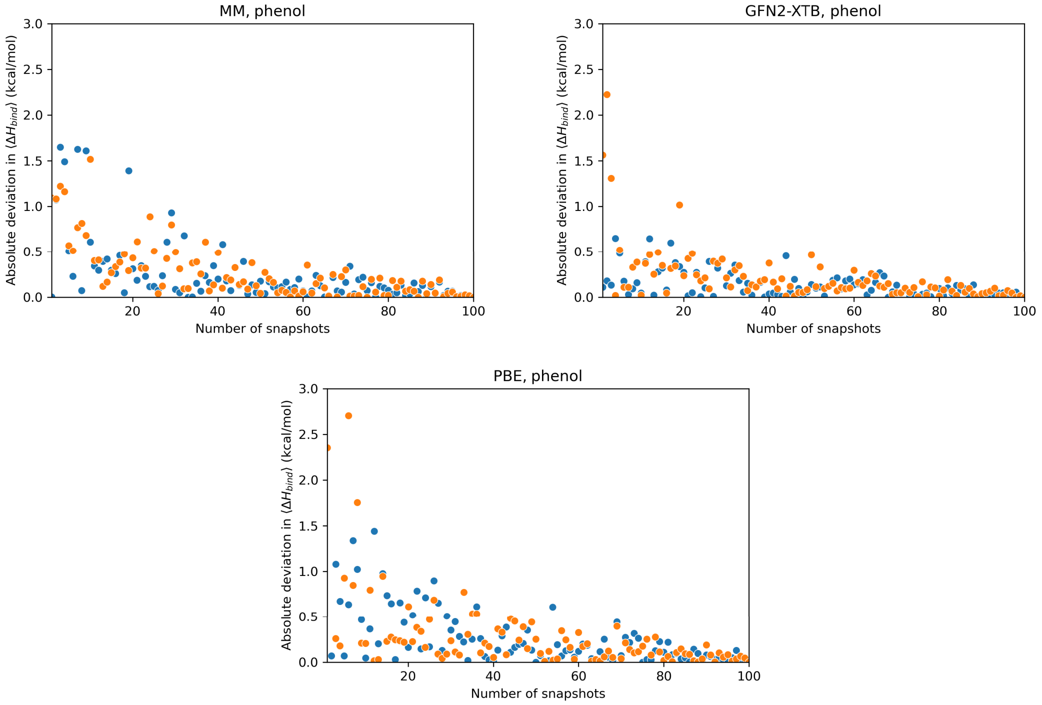


FIGURE A.11: phenol

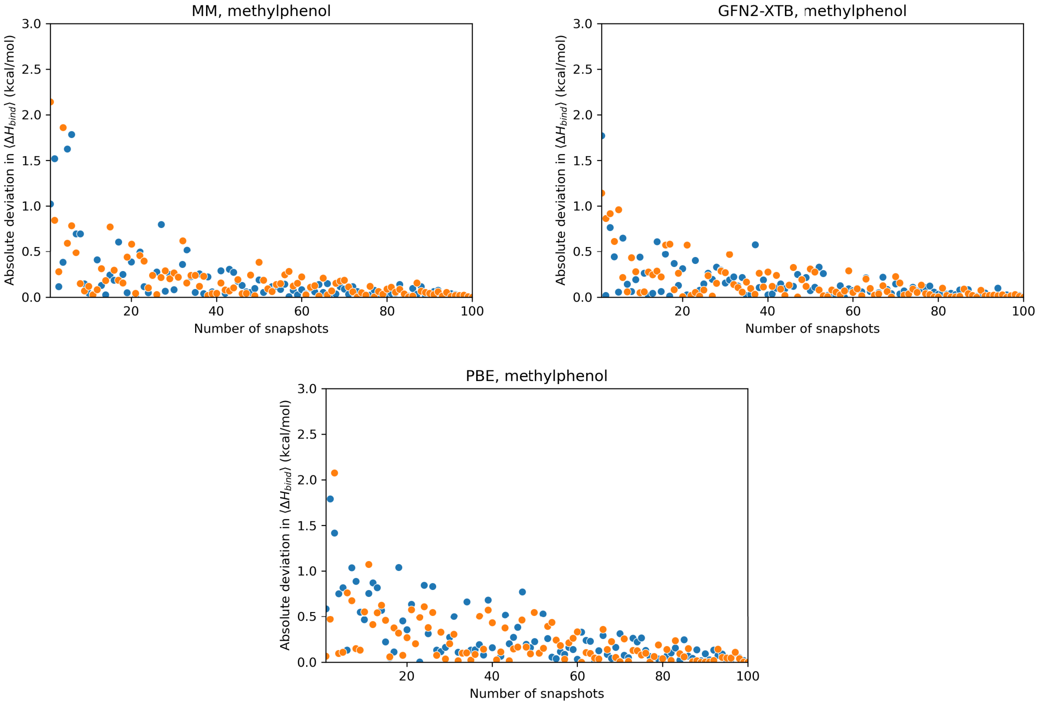


FIGURE A.12: methylphenol

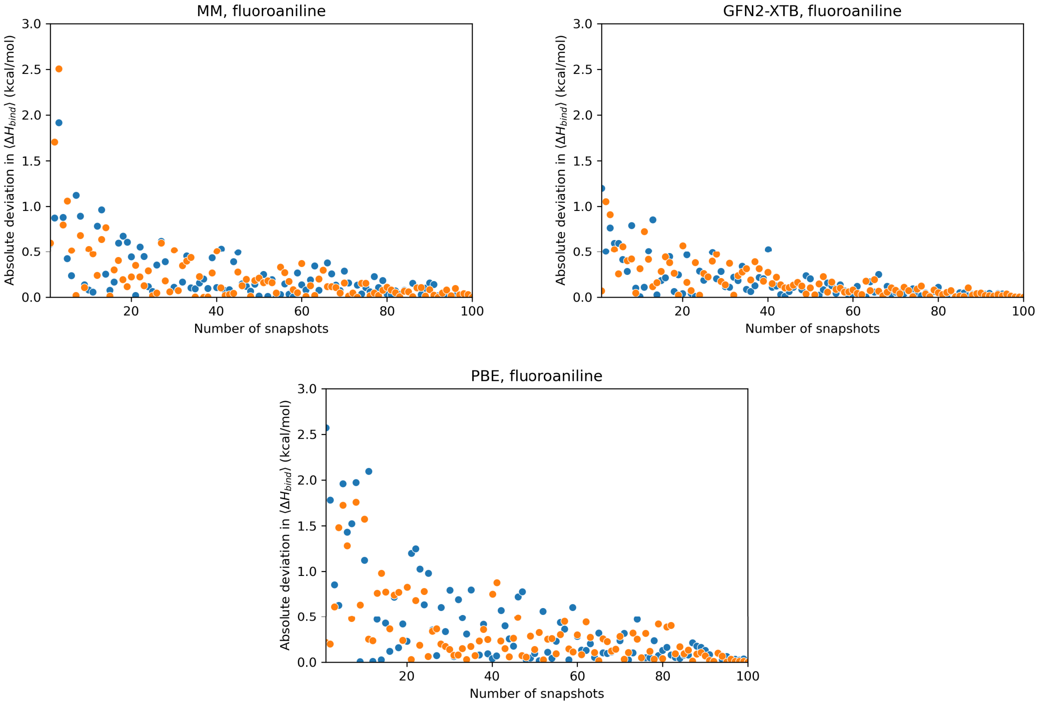


FIGURE A.13: fluoroaniline

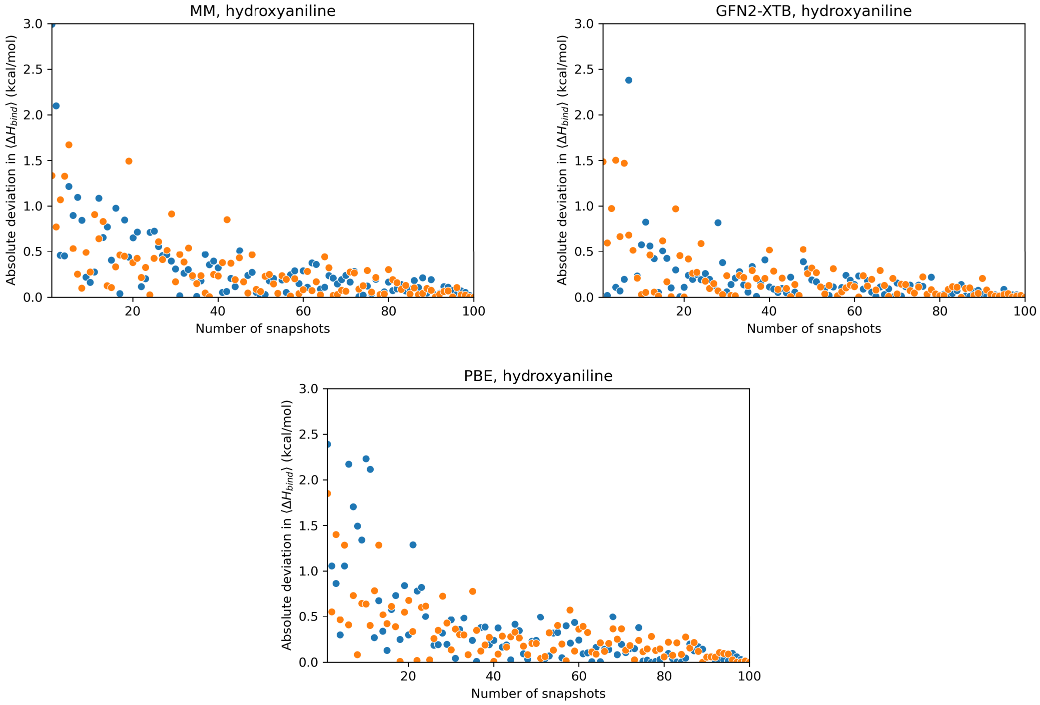


FIGURE A.14: hydroxyaniline

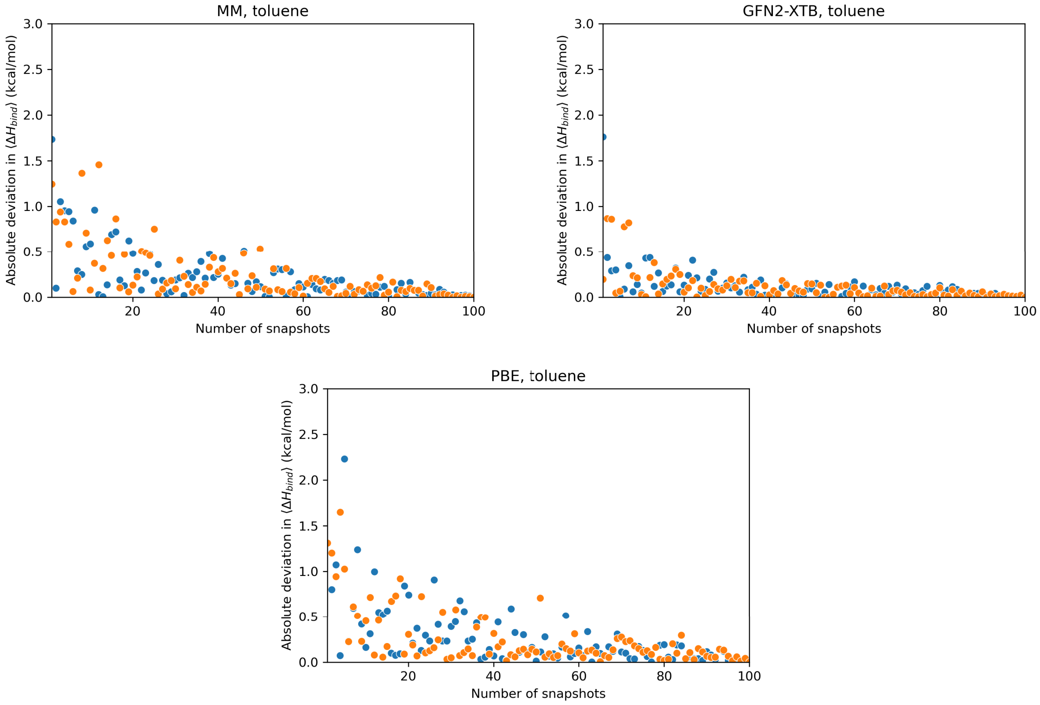


FIGURE A.15: toluene

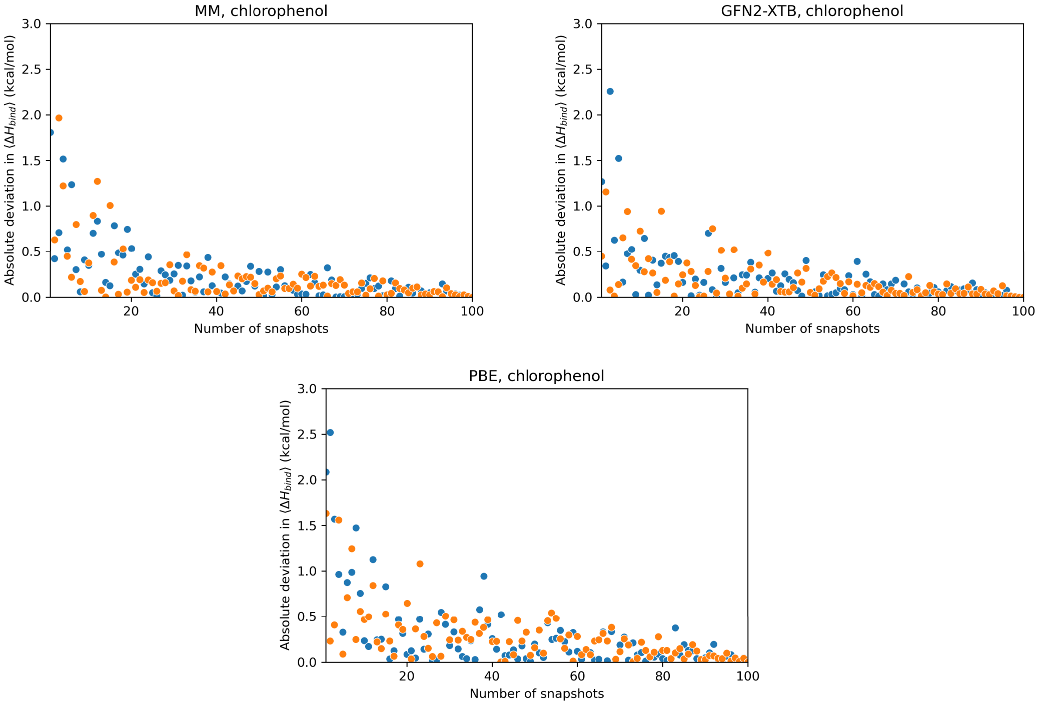


FIGURE A.16: chlorophenol

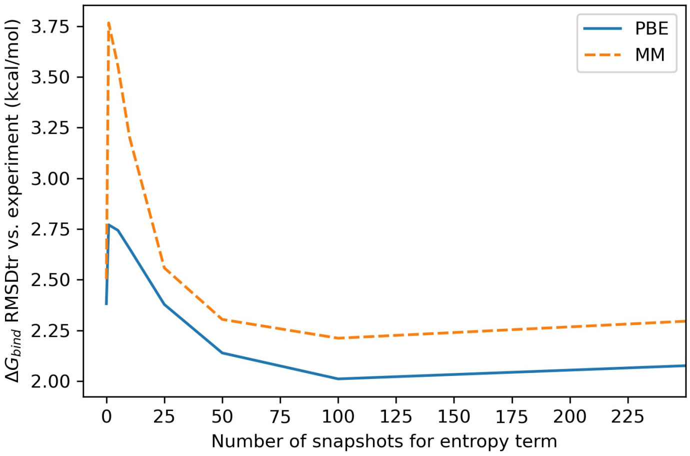


FIGURE A.17: Root mean square error deviation from experiment after removal of mean signed error (kcal/mol) of calculated binding free energies for ligand set B, at different levels of entropy sampling. Enthalpy sampled over 100 snapshots. RMSDtr calculated with upper limit for non-binder.

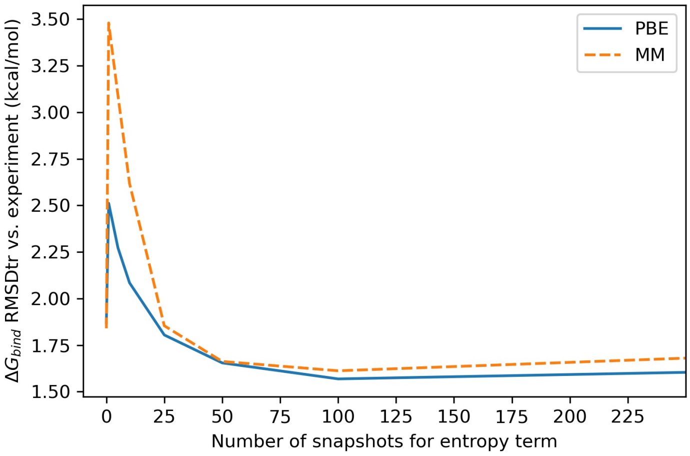


FIGURE A.18: Root mean square error deviation from experiment after removal of mean signed error (kcal/mol) of calculated binding free energies for ligand set B, at different levels of entropy sampling. Enthalpy sampled over 100 snapshots. RMSDtr calculated with lower limit for non-binder.

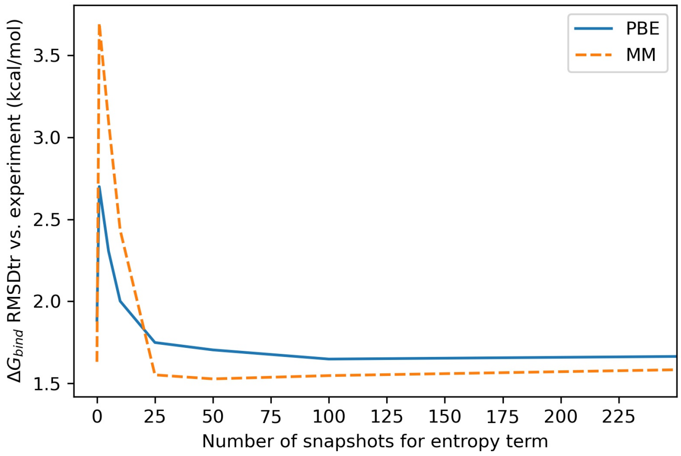


FIGURE A.19: Root mean square error deviation from experiment after removal of mean signed error (kcal/mol) of calculated binding free energies for ligand set B, at different levels of entropy sampling. Enthalpy sampled over 100 snapshots. RMSDtr calculated with binders only.

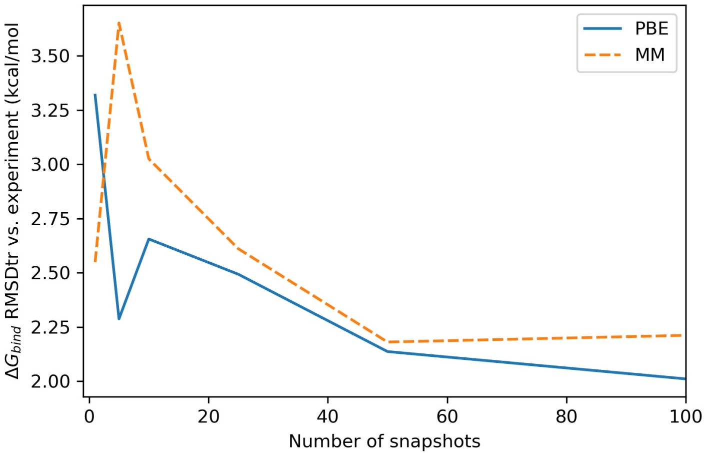


FIGURE A.20: Root mean square error deviation from experiment after removal of mean signed error (kcal/mol) of calculated binding free energies for ligand set B, at different levels of entropy and energy sampling. Enthalpy and energy sampled over same snapshots. RMSDtr calculated with upper limit for non-binder.

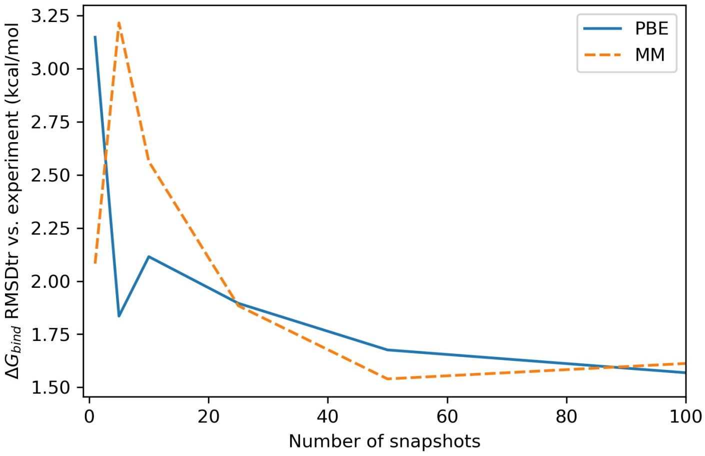


FIGURE A.21: Root mean square error deviation from experiment after removal of mean signed error (kcal/mol) of calculated binding free energies for ligand set B, at different levels of entropy and energy sampling. Enthalpy and energy sampled over same snapshots. RMSDtr calculated with lower limit for non-binder.

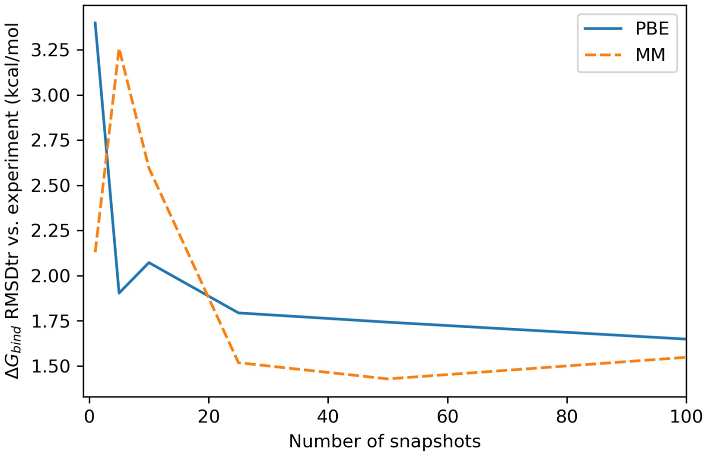


FIGURE A.22: Root mean square error deviation from experiment after removal of mean signed error (kcal/mol) of calculated binding free energies for ligand set B, at different levels of entropy and energy sampling. Enthalpy and energy sampled over same snapshots. RMSDtr calculated with binders only.

Appendix B

BRD4

TABLE B.1: Maximum block averaged standard errors in RMSDtr (kcal/mol) of MM-PBSA and MM-GBSA without entropy correction for 10 ligand binding to BRD4 with different numbers of explicit waters (XN).

	0N	3N	5N	10N	25N
ff14SB-TIP3P_mmgbsa	0.113887	0.143778	0.121237	0.109799	0.127694
ff14SB-TIP3P_mmpbsa	0.161999	0.126941	0.131458	0.150861	0.100190
ff19SB-OPC3_mmgbsa	0.065149	0.047955	0.047753	0.063251	0.085598
ff19SB-OPC3_mmpbsa	0.111130	0.093385	0.088731	0.096996	0.085509
ff19SB-OPC_mmgbsa	0.093136	0.206653	0.225344	0.337697	0.912879
ff19SB-OPC_mmpbsa	0.129076	0.905175	1.374065	4.578537	4.749408
ff19SB-TIP3P_mmgbsa	0.070586	0.084165	0.105516	0.119304	0.123853
ff19SB-TIP3P_mmpbsa	0.096735	0.061421	0.062102	0.060126	0.109208

TABLE B.2: Maximum block averaged standard errors in RMSDtr (kcal/mol) of MM-PBSA and MM-GBSA with Interaction Entropy correction for 10 ligand binding to BRD4 with different numbers of explicit waters (XN).

	0N	3N	5N	10N	25N
ff14SB-TIP3P_mmgbsa	0.354684	0.444692	0.420979	0.743306	0.863566
ff14SB-TIP3P_mmpbsa	0.437248	0.402654	0.370999	0.831603	0.850804
ff19SB-OPC3_mmgbsa	0.169021	0.236866	0.284536	0.200134	0.556649
ff19SB-OPC3_mmpbsa	0.704806	0.476319	0.914648	0.469965	0.857187
ff19SB-OPC_mmgbsa	0.343108	2.439163	1.011918	1.706925	2.631498
ff19SB-OPC_mmpbsa	0.242921	9.091368	12.769034	15.579301	10.734350
ff19SB-TIP3P_mmgbsa	0.247113	0.450745	0.643758	0.534131	0.452680
ff19SB-TIP3P_mmpbsa	0.215740	0.292108	0.519423	0.773108	1.082458

TABLE B.3: Maximum block averaged standard errors in Spearman's r of MM-PBSA and MM-GBSA without entropy correction for 10 ligand binding to BRD4 with different numbers of explicit waters (XN).

	0N	3N	5N	10N	25N
ff14SB-TIP3P_mmgbsa	0.039464	0.029677	0.024196	0.032561	0.043276
ff14SB-TIP3P_mmpbsa	0.030009	0.037180	0.049081	0.053779	0.070333
ff19SB-OPC3_mmgbsa	0.032882	0.011037	0.023905	0.019419	0.021269
ff19SB-OPC3_mmpbsa	0.041054	0.018614	0.015115	0.042225	0.018981
ff19SB-OPC_mmgbsa	0.016873	0.046490	0.037105	0.021954	0.034867
ff19SB-OPC_mmpbsa	0.025289	0.041166	0.025647	0.027044	0.041475
ff19SB-TIP3P_mmgbsa	0.011623	0.012483	0.040791	0.043208	0.042862
ff19SB-TIP3P_mmpbsa	0.038651	0.038696	0.042142	0.020286	0.029816

TABLE B.4: Maximum block averaged standard errors in Spearman's r of MM-PBSA and MM-GBSA with Interaction Entropy correction for 10 ligand binding to BRD4 with different numbers of explicit waters (XN).

	0N	3N	5N	10N	25N
ff14SB-TIP3P_mmgbsa	0.023287	0.068015	0.065219	0.095281	0.101481
ff14SB-TIP3P_mmpbsa	0.164355	0.123980	0.149120	0.162340	0.095670
ff19SB-OPC3_mmgbsa	0.043273	0.044291	0.046246	0.096725	0.055534
ff19SB-OPC3_mmpbsa	0.050621	0.066010	0.061014	0.109621	0.113346
ff19SB-OPC_mmgbsa	0.038676	0.056860	0.057404	0.080917	0.079079
ff19SB-OPC_mmpbsa	0.069964	0.090772	0.051550	0.139993	0.121452
ff19SB-TIP3P_mmgbsa	0.041556	0.061794	0.073116	0.053676	0.063685
ff19SB-TIP3P_mmpbsa	0.088453	0.112944	0.079008	0.086278	0.114220

TABLE B.5: Block averaged standard erros (BASE) and bootstrapped standard errors (SE) for RMSDtr and Spearman's r_s for QM-PBSA, MM-PBSA and MM-GBSA without entropy correction, normal moden entropy and Interaction Entropy correction at different numbers of snapshots.

Snapshots	Method	Entropy	RMSDtr BASE	RMSDtr Bootsrapped SE	r_s BASE	r_s Bootsrapped SE
50	QM-PBSA	No Entropy	0.19	0.15	0.03	0.02
		Normal Mode	0.23	0.26	0.08	0.06
		Interaction Entropy	0.21	0.20	0.06	0.08
	MM-PBSA	No Entropy	0.18	0.18	0.07	0.04
		Normal Mode	0.24	0.25	0.15	0.19
		Interaction Entropy	0.45	0.41	0.09	0.11
	MM-GBSA	No Entropy	0.20	0.15	0.02	0.05
		Normal Mode	0.47	0.27	0.06	0.09
		Interaction Entropy	0.21	0.29	0.06	0.11
4500	MM-PBSA	No Entropy	0.12	0.18	0.04	0.04
		Normal Mode	NA	NA	NA	NA
		Interaction Entropy	0.71	1.15	0.12	0.21
	MM-GBSA	No Entropy	0.05	0.13	0.03	0.04
		Normal Mode	NA	Na	NA	NA
		Interaction Entropy	0.73	1.10	0.09	0.14

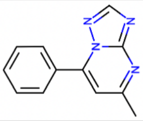
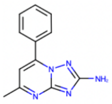
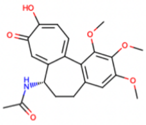
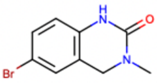
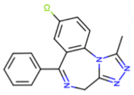
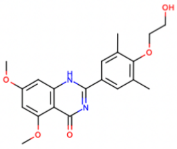
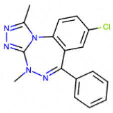
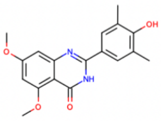
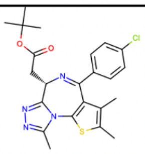
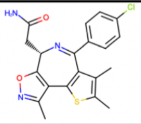
Ligand Number	Ligand Structure	G_bind (kcal/mol)	PDB Code
1		Non-Binder	
2		Weak-Binder	4MEQ
3		-5.95	4LYS
4		-6.36	4HBV
5		-7.40	3U5J
6		-7.84	4MR4
7		-8.16	3U5L
8		-8.99	4MR3
9		-9.64	RMXF
10		-10.41	4LRG

FIGURE B.1: Benchmarking ligand set for BRD4(1) with ligand structure, experimental binding energies and PDB codes [158–164] as compiled by Mobley et al.[135]

TABLE B.6: Spearman's r_s for MM-PBSA and MM-GBSA with different force field combinations at different numbers of explicit waters without entropy correction.

	0N	1N	2N	3N	4N	5N	6N	7N	8N	9N	10N	15N	25N	50N	100N
ff14SB-TIP3P mmgbsa	0.809	0.888	0.924	0.815	0.815	0.815	0.748	0.748	0.748	0.748	0.687	0.486	0.492	0.389	0.231
ff14SB-TIP3P mmpbsa	0.316	0.328	0.377	0.328	0.310	0.310	0.219	0.152	0.152	0.152	0.134	0.030	-0.049	-0.468	-0.571
ff19SB-OPC3 mmgbsa	0.815	0.869	0.778	0.802	0.802	0.802	0.699	0.371	0.280	0.292	0.292	0.280	0.036	-0.055	-0.383
ff19SB-OPC3 mmpbsa	-0.012	0.097	0.201	0.134	0.146	0.182	0.182	0.097	0.061	-0.122	-0.164	-0.219	-0.286	-0.626	-0.638
ff19SB-OPC mmgbsa	0.778	0.444	0.426	0.061	-0.146	-0.146	-0.146	-0.280	-0.371	-0.371	-0.371	-0.286	-0.304	-0.450	-0.492
ff19SB-OPC mmpbsa	0.000	-0.140	0.091	0.292	0.456	0.590	0.657	0.705	0.778	0.778	0.729	0.644	-0.669	-0.644	-0.644
ff19SB-TIP3P mmgbsa	0.760	0.894	0.815	0.796	0.748	0.748	0.748	0.614	0.596	0.553	0.505	0.413	0.334	0.249	-0.188
ff19SB-TIP3P mmpbsa	-0.073	0.006	0.182	0.176	0.134	0.134	0.024	0.018	0.018	-0.085	-0.122	-0.152	-0.310	-0.626	-0.638

TABLE B.7: Root mean squared deviation after removal of the systematic error (RMS-Dtr, kcal/mol) for MM-PBSA and MM-GBSA with different force field combinations at different numbers of explicit waters without entropy correction.

	0N	1N	2N	3N	4N	5N	6N	7N	8N	9N	10N	15N	25N	50N	100N
ff14SB-TIP3P mmgbsa	3.049	2.307	2.147	2.052	2.026	2.058	2.126	2.204	2.284	2.360	2.439	2.785	3.185	3.930	5.059
ff14SB-TIP3P mmpbsa	2.349	1.832	2.033	2.312	2.514	2.626	2.677	2.705	2.700	2.657	2.614	2.342	2.631	5.399	6.185
ff19SB-OPC3 mmgbsa	2.091	1.495	1.601	1.663	1.824	2.032	2.249	2.452	2.638	2.817	2.996	3.711	4.492	5.534	7.641
ff19SB-OPC3 mmpbsa	2.961	3.222	3.636	3.907	4.108	4.252	4.359	4.417	4.443	4.444	4.450	4.346	4.487	7.110	7.856
ff19SB-OPC mmgbsa	2.192	5.020	6.942	7.884	9.071	10.469	12.039	13.720	15.476	17.384	19.330	29.934	49.973	81.515	110.114
ff19SB-OPC mmpbsa	3.284	10.567	13.915	15.465	17.281	19.985	24.061	28.363	33.834	39.112	44.393	64.139	105.615	619.528	712.993
ff19SB-TIP3P mmgbsa	2.207	1.439	1.466	1.496	1.605	1.757	1.922	2.074	2.223	2.359	2.493	3.017	3.556	4.425	6.039
ff19SB-TIP3P mmpbsa	2.645	2.924	3.221	3.438	3.611	3.723	3.754	3.763	3.752	3.728	3.690	3.478	3.730	6.811	7.652

TABLE B.8: Spearman's r_s for MM-PBSA and MM-GBSA with different force field combinations at different numbers of explicit waters with Interaction entorpy correction.

	0N	1N	2N	3N	4N	5N	6N	7N	8N	9N	10N	15N	25N	50N	100N
ff14SB-TIP3P mmgbsa	0.736	0.711	0.742	0.736	0.705	0.729	0.754	0.675	0.492	0.492	0.468	0.389	0.249	0.261	0.195
ff14SB-TIP3P mmpbsa	0.529	0.578	0.529	0.638	0.602	0.614	0.693	0.432	0.456	0.456	0.480	0.365	0.061	-0.043	-0.243
ff19SB-OPC3 mmgbsa	0.535	0.723	0.456	0.018	0.012	0.012	0.158	-0.128	-0.146	-0.207	-0.280	-0.304	-0.377	-0.340	-0.407
ff19SB-OPC3 mmpbsa	-0.061	-0.176	-0.055	-0.359	-0.371	-0.419	-0.316	-0.304	-0.316	-0.353	-0.450	-0.717	-0.663	-0.717	-0.717
ff19SB-OPC mmgbsa	0.675	0.164	0.030	-0.164	-0.249	-0.353	-0.310	-0.274	-0.310	-0.377	-0.274	-0.261	-0.377	-0.383	-0.468
ff19SB-OPC mmpbsa	0.261	-0.365	-0.049	0.109	0.492	0.535	0.705	0.693	0.693	0.693	0.766	0.742	-0.571	-0.584	-0.614
ff19SB-TIP3P mmgbsa	0.559	0.498	0.584	0.182	0.243	0.024	-0.012	-0.012	-0.036	-0.024	0.000	-0.073	-0.401	-0.401	-0.432
ff19SB-TIP3P mmpbsa	0.043	-0.055	-0.109	-0.036	-0.085	-0.207	-0.195	-0.219	-0.249	-0.255	-0.267	-0.201	-0.492	-0.480	-0.468

TABLE B.9: Root mean squared deviation after removal of the systematic error (RMS-Dtr, kcal/mol) for MM-PBSA and MM-GBSA with different force field combinations at different numbers of explicit waters with Interaction Entropy correciton.

	0N	1N	2N	3N	4N	5N	6N	7N	8N	9N	10N	15N	25N	50N	100N
ff14SB-TIP3P mmgbsa	8.745	8.094	7.680	8.135	8.238	8.267	8.235	8.012	8.518	8.263	8.021	8.919	10.176	11.452	12.612
ff14SB-TIP3P mmpbsa	7.140	6.476	5.782	5.940	5.921	5.755	5.592	5.666	6.397	6.322	5.933	6.658	8.092	11.166	12.389
ff19SB-OPC3 mmgbsa	5.179	4.546	4.420	5.021	5.829	5.961	5.586	5.676	6.815	7.534	7.252	9.606	10.797	12.356	15.036
ff19SB-OPC3 mmpbsa	6.107	5.684	5.917	6.794	7.366	7.288	7.147	7.073	8.216	8.598	8.488	9.711	10.727	14.825	16.173
ff19SB-OPC mmgbsa	5.144	8.275	10.883	14.337	19.182	22.724	27.847	33.687	33.557	38.844	46.529	58.268	83.544	129.849	179.597
ff19SB-OPC mmpbsa	5.049	10.443	15.585	17.796	17.851	19.477	22.519	24.221	24.656	25.145	24.041	38.224	131.993	647.913	752.060
ff19SB-TIP3P mmgbsa	6.431	6.134	6.167	6.244	5.965	5.236	6.089	5.534	5.389	5.923	6.346	7.707	9.651	11.016	13.492
ff19SB-TIP3P mmpbsa	4.719	5.200	4.884	4.899	5.005	5.127	5.490	5.475	5.193	5.719	6.134	6.739	9.575	12.032	13.828

Appendix C

Links to Data Repositories

T4-Lysozyme Link to input and output files:
https://github.com/gundelach/ESI-PCCP

BRD4: MD, MM-PBSA, and MM-GBSA Link to input and output files:
https://github.com/gundelach/ESI-JCIM

BRD4: QM-PBSA Link to input and output files:
https://doi.org/10.5258/SOTON/D2384

Bibliography

[1] L. Gundelach, T. Fox, C. Tautermann, and C.-K. Skylaris, "Protein–ligand free energies of binding from full-protein DFT calculations: convergence and choice of exchange–correlation functional", Physical Chemistry Chemical Physics **23**, 9381 (2021).

[2] L. Gundelach, T. Fox, C. Tautermann, and C.-K. Skylaris, "BRD4: Quantum mechanical protein-ligand binding free energies using the full-protein DFT-based QM-PBSA method", Physical Chemistry Chemical Physics **In-Review** (2022).

[3] C. N. Cavasotto, N. S. Adler, and M. G. Aucar, "Quantum chemical approaches in structure-based virtual screening and lead optimization", Frontiers in Chemistry **6**, 188 (2018).

[4] C. N. Cavasotto, M. G. Aucar, and N. S. Adler, "Computational chemistry in drug lead discovery and design", International Journal of Quantum Chemistry **119**, `10.1002/qua.25678` (2019).

[5] V. B. Luzhkov, "Molecular modelling and free-energy calculations of protein–ligand binding", Russian Chemical Reviews **86**, 211–230 (2017).

[6] B. J. Williams-Noonan, E. Yuriev, and D. K. Chalmers, "Free Energy Methods in Drug Design: Prospects of "Alchemical Perturbation" in Medicinal Chemistry", Journal of Medicinal Chemistry **61**, 638–649 (2018).

[7] P. Söderhjelm, J. Kongsted, S. Genheden, and U. Ryde, "Estimates of ligand-binding affinities supported by quantum mechanical methods", Interdisciplinary Sciences: Computational Life Sciences **2**, 21–37 (2010).

[8] U. Ryde and P. Söderhjelm, "Ligand-Binding Affinity Estimates Supported by Quantum-Mechanical Methods", Chemical Reviews **116**, 5520–5566 (2016).

[9] V. Vennelakanti, A. Nazemi, R. Mehmood, A. H. Steeves, and H. J. Kulik, "Harder, better, faster, stronger: Large-scale QM and QM/MM for predictive modeling in enzymes and proteins", Current Opinion in Structural Biology **72**, 9–17 (2022).

[10] D. E. Koshland, "Application of a Theory of Enzyme Specificity to Protein Synthesis", Proceedings of the National Academy of Sciences **44**, 98–104 (1958).

[11] B. Ma, S. Kumar, C. J. Tsai, and R. Nussinov, "Folding funnels and binding mechanisms", Protein Engineering **12**, 713–720 (1999).

[12] M. K. Gilson, J. A. Given, B. L. Bush, and J. A. McCammon, "The statistical-thermodynamic basis for computation of binding affinities: A critical review", Biophysical Journal **72**, 1047–1069 (1997).

[13] S. A. Hollingsworth and R. O. Dror, "Molecular Dynamics Simulation for All", Neuron **99**, 1129–1143 (2018).

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

[15] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case, and T. E. Cheatham, "Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models", Accounts of Chemical Research **33**, 889–897 (2000).

[16] M. Karplus and J. A. McCammon, "Molecular dynamics simulations of biomolecules", Nature Structural Biology **9**, 646–652 (2002).

[17] K. Vanommeslaeghe, O. Guvench, and A. D. MacKerell, "Molecular Mechanics", Current Pharmaceutical Design **20**, 3281–3292 (2014).

[18] C. Tian, K. Kasavajhala, K. A. Belfon, L. Raguette, H. Huang, A. N. Migues, J. Bickel, Y. Wang, J. Pincay, Q. Wu, and C. Simmerling, "Ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution", Journal of Chemical Theory and Computation **16**, 528–552 (2020).

[19] A. Warshel, M. Kato, and A. V. Pisliakov, "Polarizable force fields: History, test cases, and prospects", Journal of Chemical Theory and Computation **3**, 2034–2045 (2007).

[20] C. Zhang, C. Lu, Z. Jing, C. Wu, J. P. Piquemal, J. W. Ponder, and P. Ren, "AMOEBA Polarizable Atomic Multipole Force Field for Nucleic Acids", Journal of Chemical Theory and Computation **14**, 2084–2108 (2018).

[21] A. Albaugh, H. A. Boateng, R. T. Bradshaw, O. N. Demerdash, J. Dziedzic, Y. Mao, D. T. Margul, J. Swails, Q. Zeng, D. A. Case, P. Eastman, L. P. Wang, J. W. Essex, M. Head-Gordon, V. S. Pande, J. W. Ponder, Y. Shao, C. K. Skylaris, I. T. Todorov, M. E. Tuckerman, and T. Head-Gordon, "Advanced Potential Energy Surfaces for Molecular Simulation", Journal of Physical Chemistry B **120**, 9811–9832 (2016).

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

[23] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, "ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB", Journal of Chemical Theory and Computation **11**, 3696–3713 (2015).

[24]K. T. Debiec, D. S. Cerutti, L. R. Baker, A. M. Gronenborn, D. A. Case, and L. T. Chong, "Further along the Road Less Traveled: AMBER ff15ipq, an Original Protein Force Field Built on a Self-Consistent Physical Model", Journal of Chemical Theory and Computation **12**, 3926–3947 (2016).

[25]J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general amber force field", Journal of Computational Chemistry **25**, 1157–1174 (2004).

[26]W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water", The Journal of Chemical Physics **79**, 926–935 (1983).

[27]H. J. Berendsen, J. R. Grigera, and T. P. Straatsma, "The missing term in effective pair potentials", Journal of Physical Chemistry **91**, 6269–6271 (1987).

[28]S. Izadi, R. Anandakrishnan, and A. V. Onufriev, "Building water models: A different approach", Journal of Physical Chemistry Letters **5**, 3863–3871 (2014).

[29]S. Izadi and A. V. Onufriev, "Accuracy limit of rigid 3-point water models", Journal of Chemical Physics **145**, 074501 (2016).

[30]H. J. Berendsen, J. P. Postma, W. F. Van Gunsteren, A. Dinola, and J. R. Haak, "Molecular dynamics with coupling to an external bath", The Journal of Chemical Physics **81**, 3684–3690 (1984).

[31]H. C. Andersen, "Molecular dynamics simulations at constant pressure and/or temperature", The Journal of Chemical Physics **72**, 2384–2393 (1980).

[32]P. P. Ewald, "Die Berechnung optischer und elektrostatischer Gitterpotentiale", Annalen der Physik **369**, 253–287 (1921).

[33]T. Darden, D. York, and L. Pedersen, "Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems", The Journal of Chemical Physics **98**, 10089–10092 (1993).

[34]D. Case, H. Aktulga, K. Belfon, I. Ben-Shalom, J. Berryman, S. Brozell, D. Cerutti, T. Cheatham, G. Cisneros, V. Cruzeiro, T. Darden, R. Duke, G. Giambasu, M. Gilson, H. Gohlke, A. Goetz, R. Harris, S. Izadi, S. Izmailov, K. Kasavajhala, M. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K. O'Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, A. Shajan, J. Shen, C. Simmerling, N. Skrynnikov, J. Smith, J. Swails, R. Walker, J. Wang, J. Wang, H. Wei, R. Wolf, X. Wu, Y. Xiong, Y. Xue, D. York, S. Zhao, and P. Kollman, "Amber20", University of California, San Francisco (2022).

[35]A. Jakalian, D. B. Jack, and C. I. Bayly, "Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation", Journal of Computational Chemistry **23**, 1623–1641 (2002).

[36] C. L. Brooks, “Computer simulation of liquids”, Journal of Solution Chemistry **18**, 99–99 (1989).

[37] W. Humphrey, A. Dalke, and K. Schulten, “VMD – Visual Molecular Dynamics”, Journal of Molecular Graphics **14**, 33–38 (1996).

[38] S. J. Fox, J. Dziedzic, T. Fox, C. S. Tautermann, and C. K. Skylaris, “Density functional theory calculations on entire proteins for free energies of binding: Application to a model polar binding site”, Proteins: Structure, Function and Bioinformatics **82**, 3335–3346 (2014).

[39] S. Genheden and U. Ryde, “The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities”, Expert Opinion on Drug Discovery **10**, 449–461 (2015).

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

[41] T. Hou, J. Wang, Y. Li, and W. Wang, “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 **51**, 69–82 (2011).

[42] H. Sun, Y. Li, S. Tian, L. Xu, and T. Hou, “Assessing the performance of MM/PBSA and MM/GBSA methods. 4. Accuracies of MM/PBSA and MM/GBSA methodologies evaluated by various simulation protocols using PDBbind data set”, Physical Chemistry Chemical Physics **16**, 16719–16729 (2014).

[43] J. Dziedzic, H. H. Helal, C. K. Skylaris, A. A. Mostofi, and M. C. Payne, “Minimal parameter implicit solvent model for ab initio electronic-structure calculations”, EPL **95**, 43001 (2011).

[44] A. V. Onufriev and D. A. Case, “Generalized Born Implicit Solvent Models for Biomolecules”, Annual Review of Biophysics **48**, 275–296 (2019).

[45] W. Clark Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, “Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics”, Journal of the American Chemical Society **112**, 6127–6129 (1990).

[46] S. Genheden, J. Kongsted, P. Söderhjelm, and U. Ryde, “Nonpolar solvation free energies of protein-ligand complexes”, Journal of Chemical Theory and Computation **6**, 3558–3568 (2010).

[47] C. Wang, D. Greene, L. Xiao, R. Qi, and R. Luo, “Recent developments and applications of the MMPBSA method”, Frontiers in Molecular Biosciences **4**, 87 (2018).

[48] C. Tan, Y. H. Tan, and R. Luo, “Implicit nonpolar solvent models”, Journal of Physical Chemistry B **111**, 12263–12274 (2007).

[49]V. Barone, M. Cossi, and J. Tomasi, “A new definition of cavities for the computation of solvation free energies by the polarizable continuum model”, Journal of Chemical Physics **107**, 3210–3221 (1997).

[50]S. Genheden, P. Mikulskis, L. Hu, J. Kongsted, P. Söderhjelm, and U. Ryde, “Accurate predictions of nonpolar solvation free energies require explicit consideration of binding-site hydration”, Journal of the American Chemical Society **133**, 13081–13092 (2011).

[51]P. Mikulskis, S. Genheden, and U. Ryde, “Effect of explicit water molecules on ligand-binding affinities calculated with the MM/GBSA approach”, Journal of Molecular Modeling **20**, 1–11 (2014).

[52]N. Špačková, T. E. Cheatham, F. Ryjáček, F. Lankaš, L. Van Meervelt, P. Hobza, and J. Šponer, “Molecular dynamics simulations and thermodynamics analysis of DNA-drug complexes. Minor groove binding between 4,6-diamidino-2-phenylindole and DNA duplexes in solution”, Journal of the American Chemical Society **125**, 1759–1769 (2003).

[53]S. Wong, R. E. Amaro, and J. Andrew McCammon, “MM-PBSA captures key role of intercalating water molecules at a protein-Protein interface”, Journal of Chemical Theory and Computation **5**, 422–429 (2009).

[54]I. Maffucci and A. Contini, “Explicit Ligand Hydration Shells Improve the Correlation between MM-PB/GBSA Binding Energies and Experimental Activities”, Journal of Chemical Theory and Computation **9**, 2706–2717 (2013).

[55]I. Maffucci and A. Contini, “Improved Computation of Protein-Protein Relative Binding Energies with the Nwat-MMGBSA Method”, Journal of Chemical Information and Modeling **56**, 1692–1704 (2016).

[56]Y. L. Zhu, P. Beroza, and D. R. Artis, “Including explicit water molecules as part of the protein structure in MM/PBSA calculations”, Journal of Chemical Information and Modeling **54**, 462–469 (2014).

[57]M. Aldeghi, M. J. Bodkin, S. Knapp, and P. C. Biggin, “Statistical Analysis on the Performance of Molecular Mechanics Poisson-Boltzmann Surface Area versus Absolute Binding Free Energy Calculations: Bromodomains as a Case Study”, Journal of Chemical Information and Modeling **57**, 2203–2221 (2017).

[58]E. E. Guest, S. D. Pickett, and J. D. Hirst, “Structural variation of protein-ligand complexes of the first bromodomain of BRD4”, Organic and Biomolecular Chemistry **19**, 5632–5641 (2021).

[59]H. Gohlke, ed., *Protein-Ligand Interactions* (Wiley-VCH Verlag GmbH Co. KGaA, Weinheim, Germany, Apr. 2012).

[60] J. A. Bauer and V. Bauerová-Hlinková, "Normal Mode Analysis: A Tool for Better Understanding Protein Flexibility and Dynamics with Application to Homology Models", in *Homology molecular modeling - perspectives and applications* (IntechOpen, Oct. 2021).

[61] G. P. Pereira and M. Cecchini, "Multibasin Quasi-Harmonic Approach for the Calculation of the Configurational Entropy of Small Molecules in Solution", Journal of Chemical Theory and Computation **17**, 1133–1142 (2021).

[62] L. Duan, X. Liu, and J. Z. Zhang, "Interaction Entropy: A New Paradigm for Highly Efficient and Reliable Computation of Protein–Ligand Binding Free Energy", Journal of the American Chemical Society **138**, 5722–5728 (2016).

[63] B. Tidor and M. Karplus, "The contribution of vibrational entropy to molecular association: The dimerization of insulin", Journal of Molecular Biology **238**, 405–414 (1994).

[64] S. Genheden and U. Ryde, "Comparison of end-point continuum-solvation methods for the calculation of protein-ligand binding free energies", Proteins: Structure, Function and Bioinformatics **80**, 1326–1342 (2012).

[65] V. K. Bhardwaj, R. Singh, J. Sharma, V. Rajendran, R. Purohit, and S. Kumar, "Identification of bioactive molecules from tea plant as SARS-CoV-2 main protease inhibitors", Journal of Biomolecular Structure and Dynamics, 1–10 (2020).

[66] J. Wang, "Fast Identification of Possible Drug Treatment of Coronavirus Disease-19 (COVID-19) through Computational Drug Repurposing Study", Journal of Chemical Information and Modeling **60**, 3277–3286 (2020).

[67] R. Singh, V. Bhardwaj, P. Das, and R. Purohit, "Natural analogues inhibiting selective cyclin-dependent kinase protein isoforms: a computational perspective", Journal of Biomolecular Structure and Dynamics **38**, 5126–5135 (2020).

[68] A. Kaur, S. Shuaib, D. Goyal, and B. Goyal, "Interactions of a multifunctional di-triazole derivative with Alzheimer's Aβ42 monomer and Aβ42 protofibril: A systematic molecular dynamics study", Physical Chemistry Chemical Physics **22**, 1543–1556 (2020).

[69] V. K. Bhardwaj, R. Singh, J. Sharma, P. Das, and R. Purohit, "Structural based study to identify new potential inhibitors for dual specificity tyrosine-phosphorylation-regulated kinase", Computer Methods and Programs in Biomedicine **194**, 105494 (2020).

[70] H. Sun, L. Duan, F. Chen, H. Liu, Z. Wang, P. Pan, F. Zhu, J. Z. Zhang, and T. Hou, "Assessing the performance of MM/PBSA and MM/GBSA methods. 7. Entropy effects on the performance of end-point binding free energy calculation approaches", Physical Chemistry Chemical Physics **20**, 14450–14460 (2018).

[71] L. Xu, H. Sun, Y. Li, J. Wang, and T. Hou, "Assessing the performance of MM/PBSA and MM/GBSA methods. 3. the impact of force fields and ligand charge models", Journal of Physical Chemistry B **117**, 8408–8421 (2013).

[72] I. Y. Ben-Shalom, S. Pfeiffer-Marek, K. H. Baringhaus, and H. Gohlke, "Efficient Approximation of Ligand Rotational and Translational Entropy Changes upon Binding for Use in MM-PBSA Calculations", Journal of Chemical Information and Modeling **57**, 170–189 (2017).

[73] D. J. Cole, I. Cabeza De Vaca, and W. L. Jorgensen, "Computation of protein-ligand binding free energies using quantum mechanical bespoke force fields", MedChemComm **10**, 1116–1120 (2019).

[74] M. Born and W. Heisenberg, "Zur Quantentheorie der Molekeln", Annalen der Physik **379**, 1–31 (1924).

[75] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas", Physical Review **136**, B864 (1964).

[76] W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects", Physical Review **140**, A1133 (1965).

[77] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple", Physical Review Letters **77**, 3865–3868 (1996).

[78] N. Mardirossian and M. Head-Gordon, "Thirty years of density functional theory in computational chemistry: An overview and extensive assessment of 200 density functionals", Molecular Physics **115**, 2315–2372 (2017).

[79] N. Mardirossian and M. Head-Gordon, "Mapping the genome of meta-generalized gradient approximation density functionals: The search for B97M-V", The Journal of Chemical Physics **142**, 074111 (2015).

[80] S. Grimme, "Density functional theory with London dispersion corrections", Wiley Interdisciplinary Reviews: Computational Molecular Science **1**, 211–228 (2011).

[81] S. Grimme, S. Ehrlich, and L. Goerigk, "Effect of the damping function in dispersion corrected density functional theory", Journal of Computational Chemistry **32**, 1456–1465 (2011).

[82] E. Caldeweyher, C. Bannwarth, and S. Grimme, "Extension of the D3 dispersion coefficient model", Journal of Chemical Physics **147**, 034112 (2017).

[83] T. Risthaus and S. Grimme, "Benchmarking of London dispersion-accounting density functional theory methods on very large molecular complexes", Journal of Chemical Theory and Computation **9**, 1580–1591 (2013).

[84] K. Lee, É. D. Murray, L. Kong, B. I. Lundqvist, and D. C. Langreth, "Higher-accuracy van der Waals density functional", Physical Review B - Condensed Matter and Materials Physics **82**, `10.1103/PhysRevB.82.081101` (2010).

[85]O. A. Vydrov and T. Van Voorhis, “Nonlocal van der Waals density functional: The simpler the better”, Journal of Chemical Physics **133**, 244103 (2010).

[86]N. Mardirossian, L. Ruiz Pestana, J. C. Womack, C. K. Skylaris, T. Head-Gordon, and M. Head-Gordon, “Use of the rVV10 Nonlocal Correlation Functional in the B97M-V Density Functional: Defining B97M-rV and Related Functionals”, Journal of Physical Chemistry Letters **8**, 35–40 (2017).

[87]J. C. Prentice, J. Aarons, J. C. Womack, A. E. Allen, L. Andrinopoulos, L. Anton, R. A. Bell, A. Bhandari, G. A. Bramley, R. J. Charlton, R. J. Clements, D. J. Cole, G. Constantinescu, F. Corsetti, S. M. Dubois, K. K. Duff, J. M. Escartín, A. Greco, Q. Hill, L. P. Lee, E. Linscott, D. D. O’Regan, M. J. Phipps, L. E. Ratcliff, Á. R. Serrano, E. W. Tait, G. Teobaldi, V. Vitale, N. Yeung, T. J. Zuehlsdorff, J. Dziedzic, P. D. Haynes, N. D. Hine, A. A. Mostofi, M. C. Payne, and C. K. Skylaris, “The ONETEP linear-scaling density functional theory program”, Journal of Chemical Physics **152**, 174111 (2020).

[88]J. Lee, D. S. Patel, J. Ståhle, S. J. Park, N. R. Kern, S. Kim, J. Lee, X. Cheng, M. A. Valvano, O. Holst, Y. A. Knirel, Y. Qi, S. Jo, J. B. Klauda, G. Widmalm, and W. Im, “CHARMM-GUI Membrane Builder for Complex Biological Membrane Simulations with Glycolipids and Lipoglycans”, Journal of Chemical Theory and Computation **15**, 775–786 (2019).

[89]C. Bannwarth, S. Ehlert, and S. Grimme, “GFN2-xTB - An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions”, Journal of Chemical Theory and Computation **15**, 1652–1671 (2019).

[90]F. Gräter, S. M. Schwarzl, A. Dejaegere, S. Fischer, and J. C. Smith, “Protein/ligand binding free energies calculated with quantum mechanics/molecular mechanics”, Journal of Physical Chemistry B **109**, 10474–10483 (2005).

[91]M. Retegan, A. Milet, and H. Jamet, “Exploring the binding of inhibitors derived from tetrabromobenzimidazole to the CK2 protein using a QM/MM-PB/SA approach”, Journal of Chemical Information and Modeling **49**, 963–971 (2009).

[92]M. A. Ibrahim, “Performance assessment of semiempirical molecular orbital methods in describing halogen bonding: Quantum mechanical and quantum mechanical/molecular mechanical-molecular dynamics study”, Journal of Chemical Information and Modeling **51**, 2549–2559 (2011).

[93]K. D. Dubey and R. P. Ojha, “Binding free energy calculation with QM/MM hybrid methods for Abl-Kinase inhibitor”, Journal of Biological Physics **37**, 69–78 (2011).

[94]Y. T. Wang and Y. C. Chen, “Insights from QM/MM modeling the 3D structure of the 2009 H1N1 influenza A virus neuraminidase and its binding interactions with antiviral drugs”, Molecular Informatics **33**, 240–249 (2014).

[95] F. Barbault and F. Maurel, “Is inhibition process better described with MD(QM/MM) simulations? the case of urokinase type plasminogen activator inhibitors”, Journal of Computational Chemistry **33**, 607–616 (2012).

[96] K. Wichapong, A. Rohe, C. Platzer, I. Slynko, F. Erdmann, M. Schmidt, and W. Sippl, “Application of docking and QM/MM-GBSA rescoring to screen for novel Myt1 kinase inhibitors”, Journal of Chemical Information and Modeling **54**, 881–893 (2014).

[97] N. Díaz, D. Suárez, K. M. Merz, and T. L. Sordo, “Molecular dynamics simulations of the TEM-1 β-lactamase complexed with cephalothin”, Journal of Medicinal Chemistry **48**, 780–791 (2005).

[98] M. Kaukonen, P. Söderhjelm, J. Heimdal, and U. Ryde, “QM/MM-PBSA method to estimate free energies for reactions in proteins”, Journal of Physical Chemistry B **112**, 12537–12548 (2008).

[99] X. Chen, X. Zhao, Y. Xiong, J. Liu, and C. G. Zhan, “Fundamental reaction pathway and free energy profile for hydrolysis of intracellular second messenger adenosine 3,5-cyclic monophosphate (cAMP) catalyzed by phosphodiesterase-4”, Journal of Physical Chemistry B **115**, 12208–12219 (2011).

[100] H. Lu, X. Huang, M. D. M. Abdulhameed, and C. G. Zhan, “Binding free energies for nicotine analogs inhibiting cytochrome P450 2A6 by a combined use of molecular dynamics simulations and QM/MM-PBSA calculations”, Bioorganic and Medicinal Chemistry **22**, 2149–2156 (2014).

[101] M. Wang and C. F. Wong, “Rank-ordering protein-ligand binding affinity by a quantum mechanics/molecular mechanics/Poisson-Boltzmann-surface area model”, Journal of Chemical Physics **126**, 026101 (2007).

[102] S. Manta, A. Xipnitou, C. Kiritsis, A. L. Kantsadi, J. M. Hayes, V. T. Skamnaki, C. Lamprakis, M. Kontou, P. Zoumpoulakis, S. E. Zographos, D. D. Leonidas, and D. Komiotis, “3-Axial CH 2OH Substitution on Glucopyranose does not Increase Glycogen Phosphorylase Inhibitory Potency. QM/MM-PBSA Calculations Suggest Why”, Chemical Biology and Drug Design **79**, 663–673 (2012).

[103] K. E. Tsitsanou, J. M. Hayes, M. Keramioti, M. Mamais, N. G. Oikonomakos, A. Kato, D. D. Leonidas, and S. E. Zographos, “Sourcing the affinity of flavonoids for the glycogen phosphorylase inhibitor site via crystallography, kinetics and QM/MM-PBSA binding studies: Comparison of chrysin and flavopiridol”, Food and Chemical Toxicology **61**, 14–27 (2013).

[104] K. M. Merz, “Limits of free energy computation for protein-ligand interactions”, Journal of Chemical Theory and Computation **6**, 1769–1776 (2010).

[105] P. Mikulskis, S. Genheden, K. Wichmann, and U. Ryde, “A semiempirical approach to ligand-binding affinities: Dependence on the hamiltonian and corrections”, Journal of Computational Chemistry **33**, 1179–1189 (2012).

[106]V. M. Anisimov and C. N. Cavasotto, “Quantum mechanical binding free energy calculation for phosphopeptide inhibitors of the Lck SH2 domain”, Journal of Computational Chemistry **32**, 2254–2263 (2011).

[107]P. Söderhjelm, J. Kongsted, and U. Ryde, “Ligand affinities estimated by quantum chemical calculations”, Journal of Chemical Theory and Computation **6**, 1726–1737 (2010).

[108]D. G. Fedorov, “The fragment molecular orbital method: theoretical development, implementation in GAMESS, and applications”, Wiley Interdisciplinary Reviews: Computational Molecular Science **7**, e1322 (2017).

[109]T. Sawada, D. G. Fedorov, and K. Kitaura, “Role of the key mutation in the selective binding of avian and human influenza hemagglutinin to sialosides revealed by quantum-mechanical calculations”, Journal of the American Chemical Society **132**, 16862–16872 (2010).

[110]R. Kurauchi, C. Watanabe, K. Fukuzawa, and S. Tanaka, “Novel type of virtual ligand screening on the basis of quantum-chemical calculations for protein-ligand complexes and extended clustering techniques”, Computational and Theoretical Chemistry **1061**, 12–22 (2015).

[111]U. Tagami, K. Takahashi, S. Igarashi, C. Ejima, T. Yoshida, S. Takeshita, W. Miyanaga, M. Sugiki, M. Tokumasu, T. Hatanaka, T. Kashiwagi, K. Ishikawa, H. Miyano, and T. Mizukoshi, “Interaction Analysis of FABP4 Inhibitors by X-ray Crystallography and Fragment Molecular Orbital Analysis”, ACS Medicinal Chemistry Letters **7**, 435–439 (2016).

[112]A. Heifetz, E. I. Chudyk, L. Gleave, M. Aldeghi, V. Cherezov, D. G. Fedorov, P. C. Biggin, and M. J. Bodkin, “The Fragment Molecular Orbital Method Reveals New Insight into the Chemical Nature of GPCR-Ligand Interactions”, Journal of Chemical Information and Modeling **56**, 159–172 (2016).

[113]D. J. Cole, C. K. Skylaris, E. Rajendra, A. R. Venkitaraman, and M. C. Payne, “Protein-protein interactions from linear-scaling first-principles quantum-mechanical calculations”, EPL **91**, 37004 (2010).

[114]S. J. Fox, “Protein-ligand binding affinities from large-scale quantum mechanical simulations”, PhD thesis (University of Southampton, 2012).

[115]É. C. Nascimento, M. Oliva, K. Świderek, J. B. Martins, and J. Andrés, “Binding Analysis of Some Classical Acetylcholinesterase Inhibitors: Insights for a Rational Design Using Free Energy Perturbation Method Calculations with QM/MM MD Simulations”, Journal of Chemical Information and Modeling **57**, 958–976 (2017).

[116]M. A. Olsson and U. Ryde, “Comparison of QM/MM Methods To Obtain Ligand-Binding Free Energies”, Journal of Chemical Theory and Computation **13**, 2245–2253 (2017).

[117] S. Ehrlich, A. H. Göller, and S. Grimme, “Towards full Quantum-Mechanics-based Protein–Ligand Binding Affinities”, ChemPhysChem **18**, 898–905 (2017).

[118] E. H. Frush, S. Sekharan, and S. Keinan, “In Silico Prediction of Ligand Binding Energies in Multiple Therapeutic Targets and Diverse Ligand Sets - A Case Study on BACE1, TYK2, HSP90, and PERK Proteins”, Journal of Physical Chemistry B **121**, 8142–8148 (2017).

[119] T. J. Giese and D. M. York, “Development of a Robust Indirect Approach for MM → QM Free Energy Calculations That Combines Force-Matched Reference Potential and Bennett’s Acceptance Ratio Methods”, Journal of Chemical Theory and Computation **15**, 5543–5562 (2019).

[120] C. Cave-Ayland, C. K. Skylaris, and J. W. Essex, “A Monte Carlo Resampling Approach for the Calculation of Hybrid Classical and Quantum Free Energies”, Journal of Chemical Theory and Computation **13**, 415–424 (2017).

[121] C. Sampson, T. Fox, C. S. Tautermann, C. Woods, and C. K. Skylaris, “A ”Stepping Stone” Approach for Obtaining Quantum Free Energies of Hydration”, Journal of Physical Chemistry B **119**, 7030–7040 (2015).

[122] J. Morado, P. N. Mortenson, J. W. M. Nissink, M. L. Verdonk, R. A. Ward, J. W. Essex, and C. K. Skylaris, “Generation of Quantum Configurational Ensembles Using Approximate Potentials”, Journal of Chemical Theory and Computation **17**, 7021–7042 (2021).

[123] J. Morado, P. N. Mortenson, M. L. Verdonk, R. A. Ward, J. W. Essex, and C. K. Skylaris, “ParaMol: A Package for Automatic Parameterization of Molecular Mechanics Force Fields”, Journal of Chemical Information and Modeling **61**, 2026–2047 (2021).

[124] B. S. Hudson and D. Harris, “T4 phage lysozyme: a protein designed for understanding tryptophan photophysics”, Time-Resolved Laser Spectroscopy in Biochemistry II **1204**, 80 (1990).

[125] B. Q. Wei, W. A. Baase, L. H. Weaver, B. W. Matthews, and B. K. Shoichet, “A model binding site for testing scoring functions in molecular docking: This work is dedicated to the memory of Andy Morton (1964-1997)”, Journal of Molecular Biology **322**, 339–355 (2002).

[126] N. Ando, B. Barstow, W. A. Baase, A. Fields, B. W. Matthews, and S. M. Gruner, “Structural and thermodynamic characterization of T4 lysozyme mutants and the contribution of internal cavities to pressure denaturation”, Biochemistry **47**, 11097–11109 (2008).

[127] A. P. Graves, R. Brenk, and B. K. Shoichet, “Decoys for docking”, Journal of Medicinal Chemistry **48**, 3714–3728 (2005).

[128] S. E. Boyce, D. L. Mobley, G. J. Rocklin, A. P. Graves, K. A. Dill, and B. K. Shoichet, "Predicting Ligand Binding Affinity with Alchemical Free Energy Methods in a Polar Model Binding Site", Journal of Molecular Biology **394**, 747–763 (2009).

[129] Y. Deng and B. Roux, "Calculation of standard binding free energies: Aromatic molecules in the T4 lysozyme L99A mutant", Journal of Chemical Theory and Computation **2**, 1255–1273 (2006).

[130] E. Gallicchio, M. Lapelosa, and R. M. Levy, "Binding energy distribution analysis method (BEDAM) for estimation of protein-ligand binding affinities", Journal of Chemical Theory and Computation **6**, 2961–2977 (2010).

[131] O. D. Villarreal, L. Yu, R. A. Rodriguez, and L. Y. Chen, "Computing the binding affinity of a ligand buried deep inside a protein with the hybrid steered molecular dynamics", Biochemical and Biophysical Research Communications **483**, 203–208 (2017).

[132] I. Cabeza De Vaca, Y. Qian, J. Z. Vilseck, J. Tirado-Rives, and W. L. Jorgensen, "Enhanced Monte Carlo Methods for Modeling Proteins Including Computation of Absolute Free Energies of Binding", Journal of Chemical Theory and Computation **14**, 3279–3288 (2018).

[133] A. Niitsu, S. Re, H. Oshima, M. Kamiya, and Y. Sugita, "De Novo Prediction of Binders and Nonbinders for T4 Lysozyme by gREST Simulations", Journal of Chemical Information and Modeling **59**, 3879–3888 (2019).

[134] Y. Sakae, B. W. Zhang, R. M. Levy, and N. Deng, "Absolute Protein Binding Free Energy Simulations for Ligands with Multiple Poses, a Thermodynamic Path That Avoids Exhaustive Enumeration of the Poses", Journal of Computational Chemistry **41**, 56–68 (2020).

[135] D. L. Mobley and M. K. Gilson, "Predicting Binding Free Energies: Frontiers and Benchmarks", Annual Review of Biophysics **46**, 531–558 (2017).

[136] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods, "The Amber biomolecular simulation programs", Journal of Computational Chemistry **26**, 1668–1688 (2005).

[137] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, "Comparison of multiple amber force fields and development of improved protein backbone parameters", Proteins: Structure, Function and Genetics **65**, 712–725 (2006).

[138] J. C. Womack, N. Mardirossian, M. Head-Gordon, and C. K. Skylaris, "Self-consistent implementation of meta-GGA functionals for the ONETEP linear-scaling electronic structure package", Journal of Chemical Physics **145**, 204114 (2016).

[139] D. Case, T. Darden, T. Cheatham, C. Simmerling, J. Wang, R. Duke, R. Luo, M. Crowley, R. Walker, W. Zhang, K. Merz, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossváry, K. Wong, F. Paesani, J. Vanicek, X. Wu, S. Brozell, T. Steinbrecher, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, D. Mathews, M. Seetin, C. Sagui, V. Babin, and P. Kollman, "Amber10", University of California, San Francisco (2008).

[140] J. W. Ponder and D. A. Case, "Force fields for protein simulations", Advances in Protein Chemistry **66**, 27–85 (2003).

[141] S. Grimme, "Semiempirical GGA-type density functional constructed with a long-range dispersion correction", Journal of Computational Chemistry **27**, 1787–1799 (2006).

[142] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, "A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu", Journal of Chemical Physics **132**, 154104 (2010).

[143] D. G. Smith, L. A. Burns, K. Patkowski, and C. D. Sherrill, "Revised Damping Parameters for the D3 Dispersion Correction to Density Functional Theory", Journal of Physical Chemistry Letters **7**, 2197–2203 (2016).

[144] W. A. Baase, L. Liu, D. E. Tronrud, and B. W. Matthews, "Lessons from the lysozyme of phage T4", Protein Science **19**, 631–641 (2010).

[145] M. Elstner, P. Hobza, T. Frauenheim, S. Suhai, and E. Kaxiras, "Hydrogen bonding and stacking interactions of nucleic acid base pairs: A density functional-theory based treatment", Journal of Chemical Physics **114**, 5149–5155 (2001).

[146] P. Pracht, E. Caldeweyher, S. Ehlert, and S. Grimme, "A Robust Non-Self-Consistent Tight-Binding Quantum Chemistry Method for large Molecules", ChemRxiv, 1–19 (2019).

[147] S. Grimme, C. Bannwarth, and P. Shushkov, "A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1-86)", Journal of Chemical Theory and Computation **13**, 1989–2009 (2017).

[148] J. Srinivasan, T. E. Cheatham, P. Cieplak, P. A. Kollman, and D. A. Case, "Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices", Journal of the American Chemical Society **120**, 9401–9409 (1998).

[149] D. Case, R. Betz, D. Cerutti, T. Cheatham, T. Darden, R. Duke, T. Giese, H. Gohlke, A. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K. Merz, G. Monard, H. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, D. Roe, A. Roitberg, C. Sagui, C. Simmerling, W. Botello-Smith, J. Swails, R. Walker, J. Wang, R. Wolf,

X. Wu, L. Xiao, and P. Kollman, “Amber16”, University of California, San Francisco (2016).

[150] S. S. Shapiro and M. B. Wilk, “An analysis of variance test for normality (complete samples)†”, Biometrika **52**, 591–611 (1965).

[151] S. Zhong, K. Huang, Z. Xiao, X. Sheng, Y. Li, and L. Duan, “Binding Mechanism of Thrombin-Ligand Systems Investigated by a Polarized Protein-Specific Charge Force Field and Interaction Entropy Method”, Journal of Physical Chemistry B **123**, 8704–8716 (2019).

[152] P.-C. Su, C.-C. Tsai, S. Mehboob, K. E. Hevener, and M. E. Johnson, “Comparison of radii sets, entropy, QM methods, and sampling on MM-PBSA, MM-GBSA, and QM/MM-GBSA ligand binding energies of F . tularensis enoyl-ACP reductase (FabI)”, Journal of Computational Chemistry **36**, 1859–1873 (2015).

[153] V. M. Anisimov, A. Ziemys, S. Kizhake, Z. Yuan, A. Natarajan, and C. N. Cavasotto, “Computational and experimental studies of the interaction between phospho-peptides and the C-terminal domain of BRCA1”, Journal of Computer-Aided Molecular Design **25**, 1071–1084 (2011).

[154] B. Donati, E. Lorenzini, and A. Ciarrocchi, “BRD4 and Cancer: Going beyond transcriptional regulation”, Molecular Cancer **17**, 1–13 (2018).

[155] L. Lu, Z. Chen, X. Lin, L. Tian, Q. Su, P. An, W. Li, Y. Wu, J. Du, H. Shan, C. M. Chiang, and H. Wang, “Inhibition of BRD4 suppresses the malignancy of breast cancer cells via regulation of Snail”, Cell Death and Differentiation **27**, 255–268 (2020).

[156] Y. Taniguchi, “The bromodomain and extra-terminal domain (BET) family: Functional anatomy of BET paralogous proteins”, International Journal of Molecular Sciences **17**, 27827996 (2016).

[157] D. J. Huggins, “Comparing the Performance of Different AMBER Protein Forcefields, Partial Charge Assignments, and Water Models for Absolute Binding Free Energy Calculations”, Journal of Chemical Theory and Computation, acs.jctc.1c01208 (2022).

[158] P. V. Fish, P. Filippakopoulos, G. Bish, P. E. Brennan, M. E. Bunnage, A. S. Cook, O. Federov, B. S. Gerstenberger, H. Jones, S. Knapp, B. Marsden, K. Nocka, D. R. Owen, M. Philpott, S. Picaud, M. J. Primiano, M. J. Ralph, N. Sciammetta, and J. D. Trzupek, “Identification of a chemical probe for bromo and extra C-terminal bromodomain inhibition through optimization of a fragment-derived hit”, Journal of Medicinal Chemistry **55**, 9831–9837 (2012).

[159] L. R. Vidler, P. Filippakopoulos, O. Fedorov, S. Picaud, S. Martin, M. Tomsett, H. Woodward, N. Brown, S. Knapp, and S. Hoelder, “Discovery of novel small-molecule inhibitors of BRD4 using structure-based virtual screening”, Journal of Medicinal Chemistry **56**, 8073–8088 (2013).

[160]V. S. Gehling, M. C. Hewitt, R. G. Vaswani, Y. Leblanc, A. Coîté, C. G. Nasveschuk, A. M. Taylor, J. C. Harmange, J. E. Audia, E. Pardo, S. Joshi, P. Sandy, J. A. Mertz, R. J. Sims, L. Bergeron, B. M. Bryant, S. Bellon, F. Poy, H. Jayaram, R. Sankaranarayanan, S. Yellapantula, N. Bangalore Srinivasamurthy, S. Birudukota, and B. K. Albrecht, "Discovery, design, and optimization of isoxazole azepine BET inhibitors", ACS Medicinal Chemistry Letters **4**, 835–840 (2013).

[161]X. Lucas, D. Wohlwend, M. Hügle, K. Schmidtkunz, S. Gerhardt, R. Schüle, M. Jung, O. Einsle, and S. Günther, "4-Acyl pyrroles: Mimicking acetylated lysines in histone code reading", Angewandte Chemie - International Edition **52**, 14055–14059 (2013).

[162]P. Filippakopoulos, S. Picaud, O. Fedorov, M. Keller, M. Wrobel, O. Morgenstern, F. Bracher, and S. Knapp, "Benzodiazepines and benzotriazepines as protein interaction inhibitors targeting bromodomains of the BET family", Bioorganic and Medicinal Chemistry **20**, 1878–1886 (2012).

[163]P. Filippakopoulos, J. Qi, S. Picaud, Y. Shen, W. B. Smith, O. Fedorov, E. M. Morse, T. Keates, T. T. Hickman, I. Felletar, M. Philpott, S. Munro, M. R. McKeown, Y. Wang, A. L. Christie, N. West, M. J. Cameron, B. Schwartz, T. D. Heightman, N. La Thangue, C. A. French, O. Wiest, A. L. Kung, S. Knapp, and J. E. Bradner, "Selective inhibition of BET bromodomains", Nature **468**, 1067–1073 (2010).

[164]S. Picaud, C. Wells, I. Felletar, D. Brotherton, S. Martin, P. Savitsky, B. Diez-Dacal, M. Philpott, C. Bountra, H. Lingard, O. Fedorov, S. Müller, P. E. Brennan, S. Knapp, and P. Filippakopoulos, "RVX-208, an inhibitor of BET transcriptional regulators with selectivity for the second bromodomain", Proceedings of the National Academy of Sciences of the United States of America **110**, 19754–19759 (2013).

[165]G. Heinzelmann, N. M. Henriksen, and M. K. Gilson, "Attach-Pull-Release Calculations of Ligand Binding and Conformational Changes on the First BRD4 Bromodomain", Journal of Chemical Theory and Computation **13**, 3260–3275 (2017).

[166]M. Kuang, J. Zhou, L. Wang, Z. Liu, J. Guo, and R. Wu, "Binding Kinetics versus Affinities in BRD4 Inhibition", Journal of Chemical Information and Modeling **55**, 1926–1935 (2015).

[167]C. Cheng, H. Diao, F. Zhang, Y. Wang, K. Wang, and R. Wu, "Deciphering the mechanisms of selective inhibition for the tandem BD1/BD2 in the BET-bromodomain family", Physical Chemistry Chemical Physics **19**, 23934–23941 (2017).

[168]E. E. Guest, L. F. Cervantes, S. D. Pickett, C. L. Brooks, and J. D. Hirst, "Alchemical Free Energy Methods Applied to Complexes of the First Bromodomain of BRD4", Journal of Chemical Information and Modeling, acs.jcim.1c01229 (2022).

[169] J. Su, X. Liu, S. Zhang, F. Yan, Q. Zhang, and J. Chen, "A computational insight into binding modes of inhibitors XD29, XD35, and XD28 to bromodomain-containing protein 4 based on molecular dynamics simulations", Journal of Biomolecular Structure and Dynamics **36**, 1212–1224 (2018).

[170] J. Su, X. Liu, S. Zhang, F. Yan, Q. Zhang, and J. Chen, "A theoretical insight into selectivity of inhibitors toward two domains of bromodomain-containing protein 4 using molecular dynamics simulations", Chemical Biology and Drug Design **91**, 828–840 (2018).

[171] E. Wang, G. Weng, H. Sun, H. Du, F. Zhu, F. Chen, Z. Wang, and T. Hou, "Assessing the performance of the MM/PBSA and MM/GBSA methods. 10. Impacts of enhanced sampling and variable dielectric model on protein-protein Interactions", Physical Chemistry Chemical Physics **21**, 18958–18969 (2019).

[172] Y. Wang, S. Murlidaran, and D. A. Pearlman, "Quantum simulations of SARS-CoV-2 main protease Mpro enable high-quality scoring of diverse ligands", Journal of Computer-Aided Molecular Design **35**, 963–971 (2021).

[173] Y. Rodríguez, G. Gerona-Navarro, R. Osman, and M. M. Zhou, "In silico design and molecular basis for the selectivity of Olinone toward the first over the second bromodomain of BRD4", Proteins: Structure, Function and Bioinformatics **88**, 414–430 (2020).

[174] S. Steiner, A. Magno, D. Huang, and A. Caflisch, "Does bromodomain flexibility influence histone recognition?", FEBS Letters **587**, 2158–2163 (2013).

[175] V. Myrianthopoulos and E. Mikros, "From bench to bedside, via desktop. Recent advances in the application of cutting-edge in silico tools in the research of drugs targeting bromodomain modules", Biochemical Pharmacology **159**, 40–51 (2019).

[176] D. L. Mobley, C. C. Bannan, A. Rizzi, C. I. Bayly, J. D. Chodera, V. T. Lim, N. M. Lim, K. A. Beauchamp, D. R. Slochower, M. R. Shirts, M. K. Gilson, and P. K. Eastman, "Escaping Atom Types in Force Fields Using Direct Chemical Perception", Journal of Chemical Theory and Computation **14**, 6076–6092 (2018).

[177] M. R. Shirts, C. Klein, J. M. Swails, J. Yin, M. K. Gilson, D. L. Mobley, D. A. Case, and E. D. Zhong, "Lessons learned from comparing molecular dynamics engines on the SAMPL5 dataset", Journal of Computer-Aided Molecular Design **31**, 147–161 (2017).

[178] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open Babel: An Open chemical toolbox", Journal of Cheminformatics **3**, 1–14 (2011).

[179] O. Trott and A. J. Olson, "AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading", Journal of Computational Chemistry **31**, NA–NA (2009).

[180] I. S. Joung and T. E. Cheatham, "Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations", Journal of Physical Chemistry B **112**, 9020–9041 (2008).

[181] S. Joung and T. E. Cheatham, "Molecular dynamics simulations of the dynamic and energetic properties of alkali and halide ions using water-model-specific ion parameters", Journal of Physical Chemistry B **113**, 13279–13290 (2009).

[182] A. Grossfield and D. M. Zuckerman, *Chapter 2 Quantifying Uncertainty and Sampling Quality in Biomolecular Simulations*, 2009.

[183] G. Kohut, A. Liwo, S. Bösze, T. Beke-Somfai, and S. A. Samsonov, "Protein-Ligand Interaction Energy-Based Entropy Calculations: Fundamental Challenges for Flexible Systems", Journal of Physical Chemistry B **122**, 7821–7827 (2018).

[184] W. M. Menzer, C. Li, W. Sun, B. Xie, and D. D. Minh, "Simple Entropy Terms for End-Point Binding Free Energy Calculations", Journal of Chemical Theory and Computation **14**, 6035–6049 (2018).

[185] V. Ekberg and U. Ryde, "On the Use of Interaction Entropy and Related Methods to Estimate Binding Entropies", Journal of Chemical Theory and Computation **17**, 5379–5391 (2021).

[186] S. Maier, B. Thapa, J. Erickson, and K. Raghavachari, "Comparative assessment of QM-based and MM-based models for prediction of protein–ligand binding affinity trends", Physical Chemistry Chemical Physics **24**, 14525–14537 (2022).

[187] Y. Q. Chen, Y. J. Sheng, Y. Q. Ma, and H. M. Ding, "Efficient calculation of protein–ligand binding free energy using GFN methods: the power of the cluster model", Physical Chemistry Chemical Physics **24**, 14339–14347 (2022).

[188] J. J. Kirsopp, C. Di Paola, D. Z. Manrique, M. Krompiec, G. Greene-Diniz, W. Guba, A. Meyder, D. Wolf, M. Strahm, and D. Muñoz Ramo, "Quantum computational quantification of protein–ligand interactions", International Journal of Quantum Chemistry, 10.1002/QUA.26975 (2022).

[189] J. T. Horton, A. E. Allen, L. S. Dodda, and D. J. Cole, "QUBEKit: Automating the Derivation of Force Field Parameters from Quantum Mechanics", Journal of Chemical Information and Modeling **59**, 1366–1381 (2019).

[190] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K. R. Müller, "Machine Learning Force Fields", Chemical Reviews **121**, 10142–10186 (2021).