View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by OTHES

universität
wien

Dissertation

Titel der Dissertation

"Methodological Studies concerning Free Energy Simulations"

Verfasser

Gerhard König

angestrebter akademischer Grad

Doktor der Naturwissenschaften (Dr. rer. nat.)

Wien, 2010

Studienkennzahl laut Studienblatt:	A 091 490
Dissertationsgebiet laut Studienblatt:	Molekulare Biologie
Betreuer:	ao. Univ-Prof. Dr. Stefan Boresch

Meinem Vater

Vorwort

In dieser Dissertation werde ich vier miteinander verwandte methodologische Studien betreffend Freien Energierechnungen präsentieren. Diese basieren auf vier Manuskripten, die entweder schon in internationalen, peer-reviewten Zeitschriften publiziert wurden, oder zur Veröffentlichung vorgesehen sind:

1.) Unorthodox uses of Bennett's acceptance ratio method

Gerhard König, Stefan Bruckner, Stefan Boresch
Journal of Computational Chemistry, **30**(11), 1712-18 (2009)
(entspricht Kapitel 3)

2.) Non-Boltzmann Sampling and Bennett's Acceptance Ratio Method: How to profit from bending the rules

Gerhard König, Stefan Boresch
Akzeptiert vom Journal of Computational Chemistry (2010)
(entspricht Kapitel 4)

3.) Hydration Free Energies of Amino Acids: Why Side Chain Analog Data Are Not Enough

Gerhard König, Stefan Boresch
Journal of Physical Chemistry B, **113**(26), 8967-8974 (2009)
(entspricht Kapitel 5)

4.) Absolute hydration free energies of blocked amino acids: Are current estimates of protein solvation overvalued ?

Gerhard König, Stefan Bruckner, Stefan Boresch
Wird eingereicht
(entspricht Kapitel 6)

Bis auf die letzte Arbeit, zu der mein Kollege Stefan Bruckner und ich in gleichen Teilen beigetragen haben, handelt es sich bei den hier aufgeschlüsselten Publikationen um Erstautorenschaften meinerseits. Die darin beschriebenen Computerexperimente wurden gemeinsam mit meinem Betreuer, Prof. Stefan Boresch, konzipiert und (ausgenommen die Ethan-Methanol und Phosphotyrosin-Rechnungen in der ersten Publikation) von mir persönlich durchgeführt.

Danksagung

An dieser Stelle möchte ich mich bei allen bedanken, die beim Entstehen dieser Arbeit mitgeholfen haben.

Am meisten Dank gebührt ohne jeden Zweifel meinem Betreuer, Prof. Stefan Boresch. Ich kenne kaum jemanden, der behaupten kann einen Betreuer mit derart offenen Türen zu haben (im wahrsten Sinne des Wortes). Immer, wenn ich Ideen oder (was öfter der Fall war) Fragen hatte, konnte ich einfach mal schnell in den Nachbarraum gehen und fand dort eine immer sprudelnde Quelle des Wissens vor. Auf der anderen Seite hat er mir aber auch immer den Spielraum gelassen in Form von meinen kleinen Nebenprojekten (die teilweise auch hier enthalten sind) eigene Wege zu gehen. Am meisten bewundere ich die unendliche Geduld, die er mit mir gehabt hat und ich schätze mich deshalb glücklich, in ihm ein derart großartiges Vorbild für die akademische Welt zu haben.

Auch bei meinen Kollegen vom Institut für computergestützte biologische Chemie möchte ich mich nicht nur für das tolle Arbeitsklima, sondern auch dafür, dass sie mir stets mit Rat und Tat zur Seite standen, ganz herzlich bedanken: Christian Schröder, Sonja Maurer und Michael Haberler, die sich als Versuchskaninchen und Korrekturleser zur Verfügung gestellt haben, Stefan Bruckner, mit dem ich viele Computerexperimente gemeinsam durchgeführt habe, sowie Gregor Neumayr und Thomas Taylor, die immer für eine interessante Diskussion gut waren. Nicht zu vergessen ist auch Prof. Othmar Steinhauser, der mich in die wunderbare Welt der Simulation eingeführt hat und immer ein offenes Ohr für meine Anliegen hatte.

Schließlich möchte ich mich noch ganz herzlich bei meinen Freunden und meiner Familie bedanken, die mich stets unterstützt und auch als ein ausgleichender Gegenpol zu meiner manchmal überbordenden Aktivität gewirkt haben. Danke für die vielfältige Welt ausserhalb des Elfenbeinturms.

Zusammenfassung

Die Bestimmung von freien Energieunterschieden ist essentiell für die Untersuchung von zahlreichen Prozessen, wie Wirkmechanismen von Medikamenten, den Verlauf von enzymatischen Reaktionen oder die Löslichkeit von Chemikalien. Mittels Molekulardynamiksimulationen sind freie Energierechnungen in der Lage freie Energieunterschiede mit hoher Genauigkeit zu bestimmen. Diese Genauigkeit ist jedoch mit einem gewaltigen Rechenaufwand verbunden, sodass eine Bestimmung oft Tage oder Wochen dauert. Aus diesem Grunde werden erhebliche Anstrengungen zur Optimierung dieser Techniken unternommen.

Der erste Teil dieser Dissertation beschreibt die Anwendung der Bennett's Acceptance Ratio Methode (BAR) auf Probleme, wo konventionelle freie Energierechnungen nicht anwendbar sind. Dies illustriert die Vielseitigkeit dieser Methode. Desweiteren demonstrieren wir eine Erweiterung von BAR zur Behandlung von gebiasten Simulationen, die nicht der klassischen Boltzmann-Verteilung gehorchen. Wir bezeichnen diese Methode ergo als Non-Boltzmann Bennett (NBB). Im Rahmen von einigen praktischen Anwendungen wird anschliessend gezeigt, wie eine kreative Wahl des gebiasten Zustands die Effizienz von freien Energierechnungen erhöhen kann.

Im zweiten Teil werden BAR und NBB zur Bestimmung von Hydrationsenergien verwendet. Speziell in der Proteinfaltung oder beim Binden von Liganden spielen die Energiekosten für die (De-)Solvatation eine erhebliche Rolle. Unglücklicherweise können die Hydrationsenergien von Aminosäuren nicht experimentell bestimmt werden, weshalb oft auf Schätzungen mittels Seitenkettenanaloga zurückgegriffen wird. Die Annahme, dass Seitenketten repräsentativ für volle Aminosäuren sind, ist jedoch bisher nicht wissenschaftlich getestet worden. Daher wurden die Hydrationsenergien sowohl von Aminosäuren, als auch von Seitenkettenanaloga bestimmt. Es zeigt sich dabei eine erhebliche Diskrepanz, was sich auf zwei Effekte zurückführen läßt: Solventexklusion und Selbstsolvatation. Während beim Erstem der Zugang zum Lösungsmittel sterisch versperrt wird, ensteht zweiteres durch Wechselwirkungen des Aminosäurerückgrats mit polaren Gruppen der Seitenkette. Da viele Techniken in der computergestützten Chemie Selbstlösung nicht berücksichtigen, hat dies schwere Auswirkungen auf die Genauigkeit. Wir illustrieren dies anhand von impliziten Solventmethoden und diskutieren den Einfluss unserer Ergebnisse auf Proteinstudien.

Abstract

The determination of free energy differences is fundamental to the study of several processes such as the binding of drugs to proteins, the paths of enzymatic reactions or the solubility of chemical compounds. By employing molecular dynamics simulations, free energy calculations are capable to compute such free energy differences with high accuracy. However, this accuracy comes at excessive computational costs, often requiring days or weeks to obtain exact results. Thus, considerable effort still has to be invested in the optimization of such techniques.

The first half of this dissertation focuses on the application of Bennett's Acceptance Ratio method (BAR) to problems where standard methods to compute free energy differences are not feasible. This highlights the unique versatility of BAR. Furthermore, we demonstrate how to extend BAR in order to make use of non-Boltzmann probability distributions in biased simulations. We refer to this method as Non-Boltzmann Bennett (NBB). The NBB method is illustrated by several examples that demonstrate how a creative choice of the biased state can also improve the efficiency of free energy simulations.

The second half is concerned with the application of BAR and NBB to the study of hydration free energies. Especially in protein folding or ligand binding (de)-solvation penalties can contribute considerably to the free energy difference. Unfortunately, hydration free energies of amino acids cannot be measured experimentally. Thus, approximations based on side chain analog data are used instead. However, the assumption that side chain analogs are representative for full amino acids has never been thoroughly tested. We, therefore, computed both relative and absolute solvation free energies of amino acids and side chain analogs, showing that the results can deviate considerably due to two effects: Solvent exclusion and self-solvation. While the former accounts for the reduction of solute–solvent interactions due to sterical occlusions, the latter arises from interactions between the backbone and the polar functional groups of the side chains. Since several techniques in computational chemistry do not account for self-solvation, this finding has severe consequences. We illustrate this for several implicit solvent models and briefly discuss the implication of our results for the field of protein science.

Contents

1 Introduction 1

2 Basic concepts of free energy simulations 9

2.1 Statistical mechanics and the free energy 9
2.2 Force Fields 11
2.3 Molecular simulations 12
2.4 Calculating free energy differences 14
2.5 Practical considerations concerning relative free energy differences . . 19

3 Unorthodox uses of Bennett's acceptance ratio method 24

3.1 Introduction 24
3.2 Theory 27
3.3 Methods 28
3.4 Results 31
    3.4.1 Ethane to methanol 31
    3.4.2 Change of cut-off radius 31
    3.4.3 Change of force field 34
    3.4.4 Solvation free energies from implicit solvent models 36
3.5 Concluding Discussion 37

4 Non-Boltzmann Sampling and Bennett's Acceptance Ratio Method: How free energy simulations can profit from bending the rules. 40

4.1 Introduction 41
4.2 Theory 44
    4.2.1 The Non-Boltzmann Bennett Method 44
    4.2.2 Applications of NBB 46

4.3 Methods 47
  4.3.1 Leucine–Asparagine 47
  4.3.2 Five-atomic systems 48
  4.3.3 Pseudoglycine–Glycine 49
  4.3.4 Implicit solvent 50
4.4 Results 51
  4.4.1 Leucine-Asparagine 51
  4.4.2 Five-atomic systems 53
  4.4.3 Pseudoglycine – glycine 55
  4.4.4 Implicit solvent 57
4.5 Conclusions 57

5 Hydration free energies of amino acids: Why side chain analog data are not enough 62
5.1 Introduction 63
5.2 Methods 66
5.3 Results 68
  5.3.1 Side chain analog and amino acid results 68
  5.3.2 Estimating contributions from SE and SS to $\Delta\Delta\Delta A$ 71
  5.3.3 Molecular origin of self-solvation and its conformation-dependence 75
  5.3.4 Some comments on specific results 78
  5.3.5 Implicit solvent results 79
5.4 Concluding Discussion 83
5.5 Supporting Information 85

6 Absolute hydration free energies of blocked amino acids: Are current estimates of protein solvation overvalued ? 91
6.1 Introduction 92
6.2 Methods 95
6.3 Results and Discussion 98
6.4 Conclusions 105
6.5 Some comments concerning implicit solvent models 109

Final discussion 113

Bibliography **115**

Declaration by the author **133**

Curriculum Vitae **135**

Chapter 1

Introduction

Predictions not only form the basis of scientific discovery, but also constitute a crucial factor in planning and optimizing system designs in engineering. According to the philosopher Karl Popper the capability to develop clear and testable predictions is actually the one central feature that characterizes a mature scientific discipline [1]. However, in case of complex systems, such as encountered in the biological domain, making useful quantitative predictions is far from trivial, since their properties are characterized by a phenomenon that is called *emergence* [2]. The term emergence conveys the idea that multiple basic entities which form relatively simple interactions can combine to a very intricate collective whose properties cannot be directly explained by the properties of its individual constituents. This concept was already described by Aristotle and became proverbial with the words "The whole is greater than the sum of its parts" [3].

Humans, for example, are not simple bags filled with water and some secret ingredients. Our bodies are composed of different forms of cells, and all components of these cells, such as enzymes or the DNA, are subject to a very intricate self-regulatory system, whose complex responses to itself and the environment make out the very essence of the emergent quality that we casually call "life". Emergence is also part of everyone's life, as illustrated by human society itself, which sometimes appears to be quite independent of the will of its individuals (e.g., in social processes such as decision making in a state or, on a more local level, in committees). The same concept can also be transferred to animal societies, whose interactions are less complex (e.g. the behavior of ant colonies or the formation of ecosystems based on simple predator-prey relationships). Similar considerations also apply to the shape of weather phenomena in meteorology, the behavior of stock markets in economics, or the interactions of neurons in the brain that lead to human thought [4].

Biochemistry is mainly characterized by the reactions and interactions between biomolecules such as proteins. On the molecular level an average protein consists of several thousand atoms. In principle one has to compute the interactions of each atom with all other atoms in proximity, a task that becomes increasingly difficult with growing system size (e.g., in a system of 10 atoms 100 interactions have to be considered, while for 1000 atoms the number of interactions grows to $1,000,000$). In addition, proteins are not isolated entities, but interact with other proteins and substrates in an aqueous environment. Thus, one has to account for the influence of water molecules in the surrounding, as well as for other freely diffusing compounds that are essential for the function of the protein. Due to this high level of complexity, processes taking place in living matter can appear to behave quite cryptic or even erratic at times (as every biology student is painfully aware of).

Although experienced molecular biologists have been able to achieve a good understanding for the sometimes chaotic twists and turns of the biochemical pathways, there is a continuing interest to find generally applicable theoretical means that can be employed in this field. Thus, the holy grail of every biochemist is to fully describe biological phenomena in terms of physicochemical processes (Richard Feynman articulated this view particularly clearly with the words ”Everything that living things do can be understood in terms of the jiggling and wiggling of atoms” [5]). With the advance of computer technology, a full atomistic representation of the dynamics of macromolecules on a mesoscopic level is gradually becoming feasible. Thus, the importance of accurate simulations of the properties and dynamics of proteins (and other biomolecules) will increase even further, especially in areas where experimental methods are not applicable (due to methodological restrictions) or exceedingly expensive. In the long term, computer simulations might even pave the way to the rational design of enzymes or, in combination with systems biology, even whole biochemical pathways [6]. This could turn biotechnology to a fully developed engineering science with a large emphasis on in silico approaches, making the creation of novel biological systems possible (a concept also known as *synthetic biology*).

To be of practical use for such purposes, simulations must be able to calculate the chemical properties of a system to high accuracy. One of the most fundamental thermodynamic properties of a chemical system is its free energy (A). The concept of free energy is mostly encountered in form of free energy differences (ΔA) associated with chemical reactions, i.e. processes that lead from one state to another state (e.g. $2\ H_2 + O_2 \rightarrow 2\ H_2O$). The free energy difference between the two states involved determines the probability and direction of chemical reactions (such as the synthesis of biomolecules), conformational changes (such as protein folding) or transfer processes (such as solvation). Thus, the development of theoretical means to determine free energy differences has been the focus of generations of researchers [7–11].

One half of this dissertation focuses on the study of solvation free energies of amino acids and their corresponding side chain analogs. This is motivated by the great impact of the so-called hydrophobic effect for correctly predicting properties of biomolecules. However, the importance of determing solvation free energies is not restricted to biomolecules. Various properties of potential drugs also depend on hydrophobicity. The affinity of a drug for its target is affected by the polarity of the drug itself as well as of the protein binding site. An equally important issue, however, is bioavailability: Since most drugs are administered orally, they have to absorbed in the gastrointestinal tract [12]. In this context, the percentage of the dose reaching the circulation is called the bioavailability. Since too hydrophilic drugs are unable to pass through cell membranes, most of the drug would be lost and higher doses would be necessary to reach the necessary drug concentration in the blood. Hydrophobic drugs are, therefore, more economical. However, too hydrophobic drugs will accumulate in the cell membranes and, thus, can lead to toxic effects [13]. Therefore, data on the hydrophobicity of a compound is also relevant for absorption, distribution, metabolism, excretion and toxicity (ADMET) tests in drug development.

Similar considerations also apply to the study of environmental effects of hydrophobic compounds. Research in this area has been sparked with the discovery of the hazardous properties of the well-known synthetic pesticide DDT in the second half of the 20^{th} century [14]. Strongly hydrophobic compounds such as DDT get easily absorbed by soil and are highly persistent, with half lifes ranging from days to years. Especially in aquatic ecosystems, hydrophobic compounds are quickly absorbed by organisms, thus getting into the global food chain, where they accumulate in top-level predators. Due to its reproductive toxicity, DDT almost lead to the extinction of various birds of prey, including the national symbol of the USA, the bald eagle. Therefore, the evaluation of the hydrophobicity of potentially hazardous chemicals has caught considerable attention. In Europe the necessity of testing chemicals for (eco-)toxicity led to the European Community Regulation on chemicals and their safe use (REACH). This law entered into force in 2007, and requires the registration and evaluation of about 143,000 chemical substances marketed within the union. Considering such extraordinary efforts, the employment of computer simulations to characterize at least some aspects of these chemicals could pose a fast, cost-effective and ecological alternative to normal laboratory work in analytical chemistry. However, such delicate applications require methods that are both accurate and reliable.

Today, so-called "free energy simulations" are the most accurate and general methodology in the field of computational chemistry. In the biological domain, they have been successfully applied to the calculation of binding affinities of ligands [15, 16], the study of enzymatic reactions [17], of molecular solvation [18, 19], and of protein stability as a function of point mutations [20]. However, free energy simulations are currently subject to a number of limitations. First, if the two end states of the free energy difference of interest (denoted as 0 and 1, respectively) are too different, unphysical intermediate states have to be introduced in order to achieve convergence. Since such intermediate states are commonly realized by mixing the potential energy functions (U) of both end states according to a mixing factor λ (i.e., $U_\lambda = (1 - \lambda)\ U_0 + \lambda\ U_1$), they are also referred to as λ-states or λ-points. By adding λ-states, the total simulation length is of course correspondingly multiplied, leading to considerable computational costs. Up to 21 λ-points (or even more) are required when using conventional thermodynamic integration free energy simulations.

Second, the exact computation of free energy differences requires adequate sampling of all relevant low energy conformations of a state. Especially in biomolecular systems, the energy landscape is characterized by local energy minima, which are frequently separated by very large energy barriers that are hardly ever crossed during a normal simulation. Thus, molecular dynamics simulations often get trapped in local minima, often requiring days or weeks of computing time even on modern computer clusters to collect the necessary data, and it is never clear whether some important information is still missing (a situation similar to traveling in a foreign country with an ordinary car in very steep mountainous terrain —without a map— trying to find the deepest river).

In this thesis, we try to address some of the issues mentioned above. In particular, two chapters will focus on the advantages of the Bennett's acceptance ratio method (BAR) [9]. Although BAR was originally conceived in the mid-seventies, it was not until the recent rise of non-equilibrium versions of free energy calculations [10, 11] that its efficiency was systematically compared to the two traditional workhorses of free energy calculations, i.e., thermodynamic integration (TI) [7] and the exponential formula (EF, also known as thermodynamic perturbation) [8]. In a detailed study, Shirts and Pande showed that BAR is more efficient than both TI and EF [21], requiring significantly fewer λ-points to obtain correct results. This finding and the fact that BAR is a minimum variance, maximum likelihood estimator of the free energy difference resulted in an increased popularity of BAR.

In Chapter 3, we start by illustrating some applications of BAR to problems where both TI and EF are not practical anymore. This study is based on the observation that BAR can compute free energy differences with fewer λ-states than TI and EF. For simple systems free energy differences can be calculated without any intermediate states at all. This is demonstrated for several standard benchmark systems (e.g., the free energy difference between ethane and methanol in aqueous solution). Then, we show how BAR can be used to compute quite unorthodox free energy differences directly, such as the free energy difference resulting from changing the treatment of electrostatic interactions, from switching the force field, or from using an implicit solvent model. Such calculations could prove advantageous for force field development or the validation of implicit solvent methods.

The problem of insufficient sampling is addressed in Chapter 4. Our starting point is the observation that simulations that do not adhere to the classical Boltzmann rule (so-called non-Boltzmann sampling) are able to enhance the exploration process (i.e., able to cross energy barriers faster), thus obtaining correct results with simulation times that can be several times shorter than in normal simulations [22–24]. This is usually achieved by adding a so-called biasing potential to the normal simulation setup. However, to obtain correct free energy differences from such biased simulations, it is necessary to account for the effects of the bias in the (post-production) analysis. We demonstrate that this can be accomplished quite simply with a slight modification of Bennett's Acceptance Ratio method. Due to its similarity to Non-Boltzmann Thermodynamic Integration (NBTI) [24] and in honour of the Austrian physicist Ludwig Boltzmann, we refer to this technique as Non-Boltzmann Bennett (NBB). We illustrate the method by several examples and show how a creative choice of the biased state can also improve the efficiency of free energy simulations.

However, the methodological aspects of free energy simulations are not restricted to reducing the number of λ-states or improving the sampling during simulations. A more fundamental problem is the employment of the additivity principle in macromolecular chemistry and biology [25]. This principle assumes that the components of a molecule contribute independently to some process and, therefore, the total change of the free energy of the molecule is given by the sum of its components. Thus, on a system-theoretical level, "additivity" is the direct opposite of the aforementioned "emergence" principle that dominates in biological phenomena.

In the context of solvation free energies, the use of additive methods is relatively widespread since the solvation free energies of complex molecules, such as proteins or even amino acids, cannot be measured experimentally [26]. Therefore, estimates of these solvation free energies were obtained from small molecules by adding contributions of model compounds. E.g., full amino acids were separated into a model compound representing the backbone (e.g., N-methylacetamide) [27] and the amino acid side chains (side chain analogs, e.g., methanol for Ser etc.) [28]. Fragment based methods to determine the solvation free energy [29–31], as well as some hydrophobicity scales [32] can thus be regarded as special extensions of the additivity principle. In particular, the side chain solvation free energies reported by Wolfenden and co-workers [28] are widely used as model systems for amino acids and proteins.

In Chapter 5, we employ free energy simulations to compute relative solvation free energies for several pairs of amino acids with N-acetyl-methylamide blocking groups and compare them with the corresponding results of side chain analogs. This serves to test the assumption whether the solvation free energies of side chain and protein backbone are additive. In particular, we focus on two effects. First the reduction of solute–solvent interactions due to sterical occlusions, which is called solvent exclusion. The second effect is the so-called self-solvation, which arises from interactions between the backbone and the polar functional groups of the side chains. Our approach is driven by the hypothesis that those two effects are most likely the major causes of possible non-additivities of solvation free energies. Thus, the accuracy of additive approaches will depend on the magnitude of these effects. If the changes of the solvation free energy due to self-solvation and solvent exclusion are significant, the correct prediction of solvation effects will depend on the ability of a method to account for both. For this purpose, we complement the free energy simulations of amino acids and side chain analogs by simulations in which we compute relative solvation free energy differences between unphysical systems, e.g., amino acids with all backbone and/or side chain charges set to zero. These data make it possible to estimate the respective contributions from solvent exclusion and self-solvation to the solvation free energy of amino acids of blocked amino acids. In addition, we analyze interactions between side chain and backbone of polar amino acids. Using Ser as a representative example of small, polar amino acids, we explore the influence of backbone conformation on solvent affinity. This test is also employed for several implicit solvent models to explore their conformance with our explicit solvent results.

Finally, in Chapter 6 we present absolute solvation free energies for blocked N-acetyl-methylamide amino acids and compare them with results for non-zwitterionic amino acids and side chain analogs. These calculations were motivated by a recent study by Chang et al. [33]. On the one hand, Chang et al. found clear deviations from the additivity principle for zwitterionic amino acids; on the other hand, they postulated additive behavior for the non-zwitterionic amino acids. In this part of the thesis we investigate the cause of this discrepancy and continue to assess errors that may arise due to the employment of additivity principles for the determination of solvation free energies. We close with a discussion of our findings and their potential impact on applications in the biological domain.

The remainder of this thesis is organized as follows. First, we briefly outline the basic concepts of free energy simulations (Chapter 2). We then present a study concerning the use of BAR in the context of rather unusual free energy differences (Chapter 3), followed by a demonstration how a combination of BAR with the employment of biased states can improve the efficiency of free energy simulations (Chapter 4). In Chapter 5, we present results for relative solvation free energy differences of blocked amino acids and their corresponding side chain analogs, determining the effect of solvent exclusion and self-solvation. Finally (Chapter 6), we compute absolute solvation free energies of blocked amino acids and compare our results to other studies.

Chapter 2

Basic concepts of free energy simulations

For readers not familiar with the theoretical background of free energy simulations, we outline the basic principles on the following pages. Since a full explanation of the statistical mechanical foundation clearly lies outside the scope of this thesis, we focus on a short (and, therefore, necessarily incomplete) introduction of most of the technical terms and concepts that will be encountered during the rest of this work. We restrict ourselves to classical mechanics since no quantum-mechanical calculations are included in this thesis. More thorough treatments of the subject can be found elsewhere [34–37].

2.1 Statistical mechanics and the free energy

The theoretical foundation of computer simulations of (bio)molecular systems lies in statistical mechanics [34, 38]. Statistical mechanics describe how macroscopic properties (such as pressure, viscosity or free energies) can be explained in terms of molecular configurations. In this context, a central concept is the so-called *phase space*. To illustrate what is meant by phase space, consider that for a system consisting of N atoms, 6N values are required to determine the three components of the coordinates and of the momenta of each atom. Each combination of the 3N positions and 3N momenta defines a particular point, a so-called *microstate*, in the 6N-dimensional space, which is referred to as phase space. A microstate is a particular combination of all degrees of freedom of a state of a thermodynamic system (every knowable aspect of every part of a specific configuration), e.g. a certain conformation of a molecule. To describe macroscopic phenomena, data from multiple microstates that populate the macroscopic system has to be combined. In this context, the number of useful microstates is restricted by the external constraints on the system, e.g. whether the system is thermally isolated, kept at a fixed temperature with a large heat reservoir, or open. *Ensembles* are sets of points in phase space that fulfill such criteria. E.g., points in phase space whose momenta correspond to a certain temperature belong to the canonical ensemble, while systems that are fully isolated are restricted to parts of phase space that have exactly the same energy, which corresponds to the microcanonical ensemble. To calculate the properties of interest, molecular simulations generate a sequence of microstates from an ensemble, which can be analyzed in detail.

The property of interest to us is the Helmholtz free energy (A) of a molecular system. In the canonical ensemble A is given by

$$A = -k_B T \ln Z \tag{2.1}$$

where k_B is the Boltzmann constant, T is the absolute temperature in Kelvin and Z is the so-called partition function. In the following, we will concentrate on the configurational partition function, since in classical mechanics the kinetic energy contributions to the partition function and, hence the the free energy can be taken care of analytically. The partition function is a function of all microstates of a system that fulfill the constraints of the ensemble (i.e., all imaginable combinations of atomic coordinates for a given number of atoms N) and encodes the thermodynamic properties of a system.

In the canonical ensemble, the partition function is given by a sum[1] over all microstates of the system

$$Z = \sum_i \exp\left(-\frac{U_i}{k_B T}\right) \tag{2.2}$$

where i denotes a particular microstate of the system, and U_i is the potential energy of microstate i. The partition function thus represents a special summation (or, in classical physics, integration) over parts of the phase space.

2.2 Force Fields

In the case of molecular simulations, U_i is usually calculated from all atomic coordinates of the system with a so-called force field. For the development of a force field the quantum mechanical interactions between all atoms in the system have been reduced to classical terms. The quantum-physical basis for this approximation is provided by the Hellmann-Feynman-Electrostatic-Theorem [38–41], which follows from the Born-Oppenheimer-Approximation[2]. In classical force fields, the potential energy consists of two groups of terms:

$$U = U^{bonded} + U^{nonbonded} \tag{2.3}$$

U_{bonded} is a sum of special terms that mimic the chemical bonds between atoms. A minimum set for the description of a molecule is given by bonds (which directly link

---

[1] We note that in a system that obeys the laws of classical physics (such as encountered in molecular dynamics simulations) the states are not quantized, and, therefore, the sum in Equation 2.2 should be replaced by an integral over all degrees of freedom, i.e.

$$Z = \int \dots \int \exp\left(-\frac{U(\vec{r},\vec{q})}{k_B T}\right) d\vec{r}\, d\vec{q}$$

where $\vec{r}$ denotes the atomic coordinates and $\vec{q}$ the associated momenta.
However, we think that the notation used above improves the readability and consistency of this chapter. Besides, due to the finite precision of floating point numbers in computers, the states in computer simulations are *de facto* quantized.

[2] The Born-Oppenheimer-Approximation [42] is based upon the fact that electrons move much faster than the nuclei because of the difference in weight. Thus the electronic system can always respond quickly to changes of the nuclei. This allows the decoupling of the motions of nuclei and electrons.

two atoms), bond angles (the angle between three atoms that are connected by two consecutive bonds) and dihedrals (the relative rotational angle between two bonds, which are separated by a third bond, which serves as axis of rotation):

$$U_{bonded} = \sum_{bonds} K_b \left(l - l_0\right)^2 + \sum_{angles} K_\theta \left(\theta - \theta_0\right)^2 + \sum_{dihedrals} K_\phi \left[1 + \cos\left(n\phi - \gamma\right)\right] \quad (2.4)$$

In Equation 2.4, K denotes the force constant for the respective harmonic potential, l the bond length, l_0 the equilibrium bond length , θ the bond angle, θ_0 the equilibrium bond angle, ϕ the dihedral angle, γ the phase shift, and n is the periodicity of the dihedral term. These terms are necessary for the basic description of the geometry. However, additional terms and crossterms can be introduced for a better description.

For larger distances the electron densities can be approximated by point charges at the positions of the nuclei. Thus, the summation of the nucleic charges and the electron charges leads to partial charges, which are used for the electrostatic Coulomb interactions (Coul). However, two other terms are necessary for a proper description of interactions between atoms that are not covalently bound: The repulsive interactions and the attractive van der Waals dispersion forces between atoms at closer range. This is typically done with the Lennard Jones (LJ) term, where the repulsion is described by a r^{-12} term, while the attraction is given by a r^{-6} term. The non-bonded interactions are calculated over all pairs of atoms (j and k with $1 \leq j < k \leq N$ will in the following denote the indices of the atoms involved)

$$U_{nonb} = U_{Coul} + \ U_{LJ} = \sum_{j,k} \frac{q_j q_k}{r_{jk}} + \sum_{j,k} \left(\frac{A_{jk}}{r_{jk}^{12}} - \frac{B_{jk}}{r_{jk}^{6}} \right) \quad (2.5)$$

In the Coulomb term, q is the atomic partial charge, while in the Lennard-Jones term the parameters A and B determine the point of minimal energy ($A = \epsilon\ r_m^{12},\ B = 2\epsilon\ r_m^6$, with ϵ being the well depth and r_m as the distance of minimum energy)

2.3 Molecular simulations

To determine the different microstates i in Equation 2.2, so-called molecular dynamics simulations [43–46], or, alternatively, Monte Carlo simulations [47] are employed. While molecular dynamics simulations solve Newton's equations of motion for molecular structures, Monte Carlo simulations employ random moves, which are accepted or rejected based on the Metropolis rule [47]. During such simulations, the phase space of a molecular system is sampled. Molecular dynamics and Monte Carlo simulations are devised to produce conformations according to their probability (as given by their Boltzmann weight). Thus, these methods are sometimes referred to as "Boltzmann sampling". The basis of this approach is the fact that low energy conformations are preferred in nature; i.e., the potential energy U_i is directly linked to the probability ρ_i of the corresponding state i according to

$$\rho_i = \frac{n_i}{n_{tot}} = \frac{\exp\left(-\frac{U_i}{k_B T}\right)}{Z} \tag{2.6}$$

where n_i is the number of members of the ensemble residing in state i and n_{tot} is the total number of members.

Equation 2.6 makes clear that low energy regions are usually more important than high energy regions, since they have higher probability. This is illustrated when calculating the expectation value (or ensemble average) of any property Θ of a system. Its expectation value is given by:

$$\langle \Theta \rangle = \sum_i \Theta_i \; \rho_i \tag{2.7}$$

Since molecular dynamics simulations already produce conformations according to their probability ρ_i, it is quite easy to obtain estimates of the ensemble averages from averages over the corresponding time series[3], i.e.

$$\langle \Theta \rangle_{sim.} = \frac{\sum_t \Theta_t}{n_t} \tag{2.8}$$

where $\langle \Theta \rangle_{sim.}$ denotes the estimate of an ensemble average calculated in a molecular dynamics or Monte Carlo simulation, t is the index of a conformation sampled in the simulation and n_t is the total number of conformations generated in the corresponding simulation.

If the data collected in a molecular simulation does not reflect the correct probability distributions of all contributing states (e.g., if the simulations were too short),

---

[3] We would like to point out that in Monte Carlo simulations a time series means a sequence of conformations generated by the computer based on random moves. Thus this sequence does not reflect any real time behavior.

the expectation values computed from it will most likely be erroneous. This is especially true if some of the dominant low energy regions of phase space were omitted. Since the free energy of a system is also a property that depends on the whole partition function (see Equation 2.1), correct sampling is crucial to obtain accurate results.

2.4 Calculating free energy differences

Trying to calculate the free energy directly according to Equation 2.7 is quite impractical for most purposes since computing the partition function requires sampling of *all* microstates of the system (see Equation 2.2). The dilemma of calculating the partition function with molecular dynamics or Monte Carlo simulations becomes clear when one recalls that such simulations generate conformations according to their probability ρ_i, as given in Equation 2.6; i.e., they do not produce conformations randomly, but were designed to yield ensemble averages without weighting each frame. Thus, simple averaging over the frames resulting from a simulation is enough to obtain an ensemble average. However, if we express the free energy in terms of such an ensemble average (cf. 2.7), the following equation results:

$$A = k_B T \ln \left\langle + \frac{U_i}{k_B T} \right\rangle = k_B T \ln \sum_i \exp\left(+ \frac{U_i}{k_B T} \right) \rho_i \qquad (2.9)$$

While the probability ρ_i is large in case of low energy regions, the term to its left, $\left(+\frac{U_i}{k_B T}\right)$, is large in case of high energy regions. Thus, there is a conflict of the two terms in the sum in Equation 2.9, leading to poor convergence. Consequently, attempts to compute absolute free energies from molecular dynamics or Monte Carlo simulations are usually inaccurate.

Fortunately, the absolute free energy of a system is not required for most applications in chemistry. More important is the free energy change associated with chemical reactions or transfer processes. All such calculations have in common that the initial state (0) is transformed to another (final) state (1). Based on Equation 2.1, the Helmholtz free energy difference (ΔA) between two states 0 and 1 is given by

$$\Delta A = -k_B T \ln \frac{Z(1)}{Z(0)} \qquad (2.10)$$

By itself, Equation 2.10 appears to be without merit, since we now have to compute two partition functions instead of just one. However, it is possible to reformulate the equation above in terms of ensemble averages of either state 0 or state 1 (the indices 0 and 1 indicate that the ensemble averages are calculated over all coordinate frames generated for state 0 or 1, respectively) [8]:

$$\Delta A = -k_B T \ln \left\langle \exp\left[-\frac{U(1) - U(0)}{k_B T}\right]\right\rangle_0$$

$$\text{or, equivalently} \tag{2.11}$$

$$\Delta A = +k_B T \ln \left\langle \exp\left[-\frac{U(0) - U(1)}{k_B T}\right]\right\rangle_1$$

We now rewrite the formulation for the ensemble average of state 0 in Equation 2.11 to the form of Equation 2.9, thus obtaining

$$\Delta A = -k_B T \ln \sum_i \exp\left[-\frac{U_i(1) - U_i(0)}{k_B T}\right] \rho_i(0) \tag{2.12}$$

In contrast to Equation 2.9, the summation here is over the *difference* $(U(1) - U(0))$ of the potential energies of the respective end states instead of the potential energy itself. Assuming that the overall shape of the potential energy surfaces are similar, high energy conformations of one state will also be high energy conformations of the other state (for completely nonsensical structures this will always be true), the two large values $U(1)$ and $U(0)$ will more or less cancel each other out[4]. Thus, in most cases, the $\rho_i(0)$ term will dominate in the summation, which greatly increases the convergence (and accuracy) relative to calculating the two associated free energies individually according to Equation 2.9.

If the two end states involved are very dissimilar, the assumption made above is not valid and the terms $U(1)$ and $U(0)$ will not cancel each other out, leading to poor convergence. Thus, the larger the differences between the shapes of the energy landscapes of states 0 and 1, the more difficult it becomes to compute free energy

---

[4] Similar considerations will also apply to low energy regions

differences by molecular dynamics simulation. However, if the two end states are too dissimilar, the free energy difference calculation can be broken down into multiple smaller substeps. This approach improves the similarity (or phase space overlap) of the two end states involved in each substep of the calculation, yielding a better convergence of the free energy results. To do so, intermediate states between the two end states 0 and 1 have to be simulated. Usually, such intermediate states are created by mixing the energy parameters of the states 0 and 1. The mixture ratio is given by the so-called coupling parameter λ. Customarily, values of λ range between $0 \leq \lambda \leq 1$, with the initial state corresponding to $\lambda = 0$ and the final state to $\lambda = 1$. The simplest (but not necessarily the best[5]) combination of states 0 and 1 as a function of λ is

$$U(\lambda) = \lambda U(1) + (1 - \lambda) U(0) \tag{2.13}$$

By using such a step-wise approach, the free energy difference of interest is obtained as the sum of n_λ smaller free energy simulations, i.e.

$$\Delta A = \sum_{l=1}^{n_\lambda - 1} \Delta A(\lambda_l \rightarrow \lambda_{l+1}) \tag{2.14}$$

Based on this trick, multiple (formally exact) techniques have been derived from Equation 2.14 to calculate free energy differences. The most prominent examples are the Exponential Formula [8] (EF, also known as Thermodynamic Perturbation), Bennett's Acceptance Ratio method [9] (BAR) and Thermodynamic Integration [7] (TI).

The Exponential Formula [8] (EF) is a direct application of Equation 2.11. Each substep of a free energy difference calculation is usually based on a trajectory from a single state.

$$\Delta A^{EF-FW}(\lambda_l \rightarrow \lambda_{l+1}) = -k_B T \ln \left\langle \exp\left[-\frac{U(\lambda_{l+1}) - U(\lambda_l)}{k_B T}\right]\right\rangle_{\lambda_l} \tag{2.15}$$

Again, the index λ_l indicates that the ensemble average is calculated over all coordinate frames generated for state λ_l. This illustrates a free energy simulation con-

---

[5] Usually, so-called *soft core* potentials [48] are employed to improve the stability of free energy simulations. However, these potentials do not linearly depend on the λ value ($dU(\lambda)/d\lambda$).

ducted in "forward" direction. The opposite (a free energy simulation conducted in "backward" direction) would be the case if the expectation value is calculated from a trajectory of state λ_{l+1}. Here the free energy difference is given by

$$\Delta A^{EF-BW}\left(\lambda_l \rightarrow \lambda_{l+1}\right) = k_B T \ln \left\langle \exp\left[-\frac{U\left(\lambda_l\right) - U\left(\lambda_{l+1}\right)}{k_B T}\right]\right\rangle_{\lambda_{l+1}} \tag{2.16}$$

Bennett's Acceptance Ratio method [9] (BAR) requires two simulations, one generating a trajectory containing n_{λ_l} coordinate frames for the initial state λ_l (potential energy function $U\left(\lambda_l\right)$), the other generating $n_{\lambda_{l+1}}$ coordinate sets for the final state λ_{l+1} (potential energy function $U\left(\lambda_{l+1}\right)$). Bennett showed that the free energy difference of a substep $\Delta A^{BAR}\left(\lambda_l \rightarrow \lambda_{l+1}\right)$ can formally be written as [9]

$$\Delta A^{BAR}\left(\lambda_l \rightarrow \lambda_{l+1}\right) = k_B T \left(\ln \frac{\sum_{\lambda_{l+1}} f(U\left(\lambda_l\right) - U\left(\lambda_{l+1}\right) + C)}{\sum_{\lambda_l} f(U\left(\lambda_{l+1}\right) - U\left(\lambda_l\right) - C)} - \ln \frac{n_{\lambda_{l+1}}}{n_{\lambda_l}}\right) + C \tag{2.17}$$

where f is the Fermi function,

$$f(x) = \frac{1}{1 + \exp(\frac{x}{k_B T})} \tag{2.18}$$

and

$$C = k_B T \ln \frac{Z_{\lambda_l} n_{\lambda_{l+1}}}{Z_{\lambda_{l+1}} n_{\lambda_l}}. \tag{2.19}$$

Equation 2.17 by itself would be without merit since the unknown constant C is essentially the sought after quantity (ratio of the partition functions of state λ_l and λ_{l+1}). However, Bennett showed that C can be found through an iterative procedure based on the condition

$$\sum\nolimits_1 f(U\left(\lambda_l\right) - U\left(\lambda_{l+1}\right) + C) = \sum\nolimits_0 f(U\left(\lambda_{l+1}\right) - U\left(\lambda_l\right) - C), \tag{2.20}$$

Once C has been determined so that Equation 2.20 is satisfied, the free energy difference is given by

$$\Delta A^{BAR}\left(\lambda_l \rightarrow \lambda_{l+1}\right) = -k_B T \ln \frac{n_{\lambda_{l+1}}}{n_{\lambda_l}} + C \tag{2.21}$$

The last method, Thermodynamic Integration [7] (TI), involves numerical quadrature to determine the free energy difference. Contrary to the two methods mentioned before this method employs the coupling parameter λ directly. This is motivated by considering λ as a continuous variable that can be used for differentiation/integration. Using the fundamental theorem of calculus, the free energy difference can be written as

$$\Delta A = A\left(1\right) - A\left(0\right) = \int_0^1 \frac{\partial A\left(\lambda\right)}{\partial \lambda} d\lambda = -k_B T \int_0^1 \left[\frac{\partial \ln Z\left(\lambda\right)}{\partial \lambda}\right] d\lambda \tag{2.22}$$

which immediately leads to the working equation for TI

$$\Delta A^{TI} = \int_0^1 d\lambda \left\langle \frac{\partial U\left(\lambda\right)}{\partial \lambda} \right\rangle_\lambda \tag{2.23}$$

In practice, this integral is evaluated by conducting several simulations at discrete values of λ to evaluate $\left\langle \frac{\partial U(\lambda)}{\partial \lambda} \right\rangle_\lambda$ and then employing numerical quadrature to approximate the integral. Usually, this is done by using the simple trapezoidal rule; however, we want to point out that recent experiments by our coworker Stefan Bruckner indicate that other methods for numerical quadrature (e.g., Gauss-Legendre or Clenshaw-Curtis[6]) are far more efficient than the trapezoidal rule [49, 50], provided the integrand is relatively well-behaved[7].

In closing, we would like to point out that, recently, non-equilibrium techniques to compute free energy differences were discovered by Jarzynski [10] and Crooks [11].

---

[6] Which are equivalent to special polynomial-fitting approaches.

[7] The shape of the integrand depends on the simulation setup (e.g., the soft-core-scheme employed).

2.5 Practical considerations concerning relative free energy differences

Often one is interested in a direct comparison of the free energy differences of two processes. This can be done by considering their relative free energy difference ($\Delta\Delta A$). In such cases, clever use of thermodynamic cycles can help to find more efficient ways to compute the required free energy changes [51]. Since the initial and the final state are identical in a cyclic process, and since the free energy is a state function, the total free energy change of going around a cycle has to be zero according to the laws of thermodynamics. Notably, this feature is independent of the number or kind of intermediate states involved in the cycle. Since free energy differences only depend on the end points, we are at liberty to choose pathways in between in any way it suits our purpose. Thus, thermodynamic cycles can be employed to divide a relatively complex process into a number of substeps. Often, these substeps are easier to compute than the original free energy of interest.

To illustrate this with an example, consider that we want to compare the solvation free energies of two compounds, A and B (as, e.g., in Chapter 5). This comparison normally involves the calculation of two absolute solvation free energies, which is feasible today, but requires a considerable computational effort. However, an alternative approach consists in determing the (relative) free energy difference ($\Delta\Delta A_{solv}$) between the two states *directly.* The corresponding thermodynamic cycle is depicted in Figure 2.1. Now, for the calculation of the total solvation free energy difference ($\Delta\Delta A_{solv}$), we need the solvation free energies of compounds A and B, denoted as ΔA^{A}_{solv} and ΔA^{B}_{solv}, which corresponds to $-\Delta A_1$ and ΔA_3 in Figure 2.1. Thus, the relative solvation free energy difference $\Delta\Delta A_{solv}$ of interest is given by

$$\Delta\Delta A_{solv} = \Delta A^{A}_{solv} - \Delta A^{B}_{solv} = \Delta A_1 + \Delta A_3 \tag{2.24}$$

These free energy differences correspond to the vertical arrows in Figure 2.1.

To form a thermodynamic cycle, we have to add horizontal arrows, which correspond to so-called "alchemical" free energy simulations. Such alchemical free

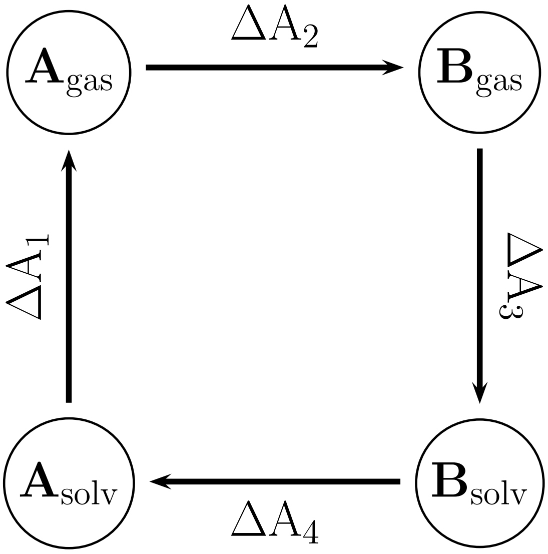

ΔA_2
$\mathbf{A}_{\text{gas}}$
$\mathbf{B}_{\text{gas}}$
ΔA_1
ΔA_3
$\mathbf{A}_{\text{solv}}$
$\mathbf{B}_{\text{solv}}$
ΔA_4

Figure 2.1: Thermodynamic cycle for determing solvation free energy differences

energy simulations transform compound A to compound B (e.g., turning lead to gold). While this is a rather difficult task in the real world (to say the least), such restrictions of physics do not apply to computer simulations. In silico, this corresponds to simply substituting the force field parameters of A by the parameters of B, which is quite trivial and can be done both in gas phase and solution ((ΔA_2) and $-\Delta A_4$ in Figure 2.1). Since the sum of a thermodynamic cycle is always zero,

$$\Delta A_1 + \Delta A_2 + \Delta A_3 + \Delta A_4 = 0 \tag{2.25}$$

the solvation free energy difference $\Delta\Delta A_{solv}$ can also be expressed in terms of the two alchemical free energy differences

$$\Delta\Delta A_{solv} = \Delta A_1 + \Delta A_3 = -\Delta A_2 - \Delta A_4 \tag{2.26}$$

Equation 2.26 means that if the solutes are (somewhat) similar, only a relatively simple mutation has to be carried out once in gas phase and once in solution (ΔA_2 and ΔA_4) to compare of the solvent affinities of two compounds. Since the simulation in gas phase includes just the solute, the computational costs are reduced considerably.

A more complex example of employing thermodynamic cycles is the analysis of ligand binding. In drug development, one is often not interested in the absolute binding free energy, but rather in a comparison of the relative binding affinity of drug candidates. Such a process is illustrated in Figure 2.2. Let the ligands L_1 and L_2 be two putative inhibitors of a receptor R. In principle, the two binding affinities (ΔA_{bind}) of the ligands L_1 and L_2 (horizontal arrows) could be determined in separate free energy calculations and compared afterwards. This would require the transfer of the ligand from a large distance into the binding pocket to form the intermolecular complex LR [52] or a stepwise procedure referred to as double-decoupling [53].

However, the relative binding affinity ($\Delta\Delta A_{bind}$) can be determined more easily by calculating the alchemical free energies indicated by vertical arrows in Figure 2.2. The first free energy difference, $\Delta \mathrm{A}_{\mathrm{aqu.}}(\mathrm{L}_1 \rightarrow \mathrm{L}_2)$, is the alchemical free energy difference between the two ligands in aqueous solution (this step corresponds to $-\Delta A_4$ in

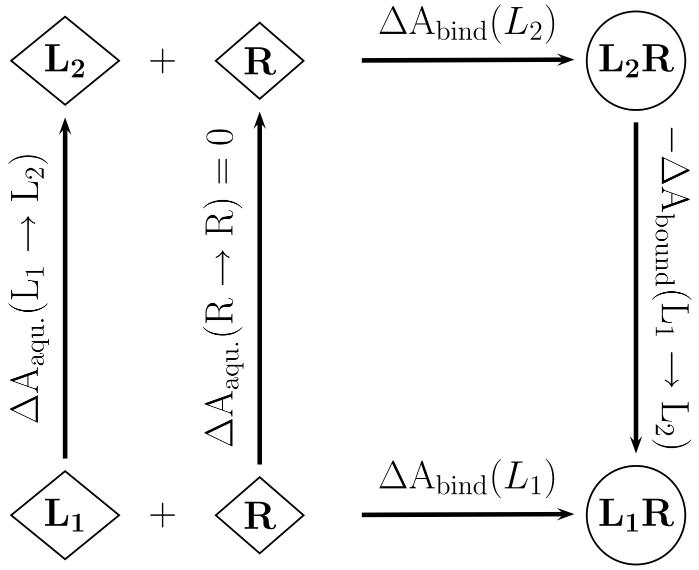


Figure 2.2: Thermodynamic cycle for the determination of the relative free energies of binding of two ligands, L_1 and L_2, to a receptor (R). While receptor and ligands are infinitely far apart on the left side of the figure (and, therefore, can be treated separately as indicated by the +), they form a complex L_xR (x=1,2) on the right side of the cycle by transferring the ligand into the binding pocket ($\Delta A_{bind}(L_x)$). Vertical arrows represent alchemical mutations.

Figure 2.1)[8]. The second free energy difference, $\Delta\mathrm{A}_{\mathrm{bound}}(\mathrm{L}_1 \rightarrow \mathrm{L}_2)$, is the free energy difference between the ligands in the bound state. It involves the transformation of L_1 to L_2 while being located in the binding pocket. For the sake of completeness, also the free energy difference $\Delta A_{aqu.}(R \rightarrow R)$ is included in Figure 2.2, however, the associated change of the free energy is zero. Since

$$\Delta\Delta\mathrm{A}_{\mathrm{bind}} = \Delta\mathrm{A}_{\mathrm{bind}}(\mathrm{L}_2) - \Delta\mathrm{A}_{\mathrm{bind}}(\mathrm{L}_1) = \Delta\Delta\mathrm{A}_{\mathrm{bound}}(\mathrm{L}_1 \rightarrow \mathrm{L}_2) - \Delta\Delta\mathrm{A}_{\mathrm{aqu.}}(\mathrm{L}_1 \rightarrow \mathrm{L}_2)$$

we can calculate the relative free energy difference of binding from non-physical pathways that are more reliable than simulating the physical processes. This is particularly so if the ligands involved have very similar binding modes.

---

[8] This highlights the importance of correct predictions of solvation effects in practical applications, since, in addition to the change of the intramolecular interactions, also the solvation free energy difference plays a critical role in this process.

Chapter 3

Unorthodox uses of Bennett's acceptance ratio method

We illustrate the application of Bennett's acceptance ratio method (BAR) to problems where standard methods to compute free energy differences (thermodynamic integration, exponential formula) are not practical. Our starting point is the observation that BAR can often compute the free energy difference between two states without the need for intermediate states usually employed (and necessary) in alchemical free energy simulations. This is demonstrated first for the free energy difference between ethane and methanol in aqueous solution. We then show how BAR can be used to compute directly rather unusual free energy differences, such as the free energy difference resulting from changing the treatment of electrostatic interactions, from switching the force field, or from using an implicit solvent model. Calculations of this kind should prove useful for force field development and the validation of implicit solvent methods.

3.1 Introduction

Alchemical free energy simulations have become an important tool in the arsenal of the computational chemist. Successful applications, but also continuing challenges are well documented by several reviews, e.g., 54–57 .The vast majority of applications of free energy simulations reported to date used either thermodynamic integration (TI) [7] or the exponential formula (EF), sometimes also referred to as thermodynamic perturbation [8]. Despite the availability of other methods, such as Bennett's acceptance ratio method (BAR) [9], TI and EF have been and still are the workhorses of alchemical free energy simulation. The discovery of non-equilibrium techniques to compute equilibrium free energy differences by Jarzynski [10] and Crooks [11] renewed interest in the comparison of methods that can be used to compute (alchemical) free energy differences; see, e.g., Refs. 58–60. Combined with the demonstration that BAR is the equilibrium equivalent of Crook's theorem [11, 61], the results of these studies led to a rediscovery of BAR. In a detailed comparison to TI and EF, Shirts and Pande showed that BAR was more efficient than either of the two in typical applications of alchemical free energy simulations [21]. As a result, the use of BAR is becoming more and more prevalent.

This paper is concerned with a facet of BAR related to the efficiency of the method, which may open up the possibility of new types of alchemical free energy simulations. Our starting point is the observation that BAR can compute an alchemical free energy difference between two states (systems) in a single step, using only simulations of the end points (states of interest). This contravenes the common wisdom that in most cases alchemical free energy calculations require not only simulations of the two end states, 0 and 1, but also simulations of unphysical intermediate states, formally characterized by the so-called coupling parameter λ. A λ-value of, e.g., 0.6 indicates a hybrid state whose properties are a mixture of approximately 40 % state 0 and 60 % state 1 (the exact properties of this artificial intermediate state depend on the detail of the hybrid potential energy function, see below). This need for intermediate states has several ramifications. First, it is one important factor in making free energy simulations so expensive in terms of computer resources (multiple, long simulations are required to obtain a single quantity). Equally important, it complicates the computer code used to carry out the underlying molecular dynamics (MD) or Monte Carlo (MC) simulations. Each simulation package handles the details of accommodating hybrids differently, i.e., on the code level no two implementations are completely equivalent. The code complexity to handle these intermediate states often entails a (further) performance penalty. Moreover, only a subset of features may be available compared to regular MD (or MC) in a particular program package. Using CHARMM [62] as an example, when computing free energy differences with one of the three available modules (BLOCK, TSM, PERT), most of the otherwise available implicit solvent models are not supported. The additional code to make possible the use of Ewald summation in combination with the BLOCK module was only added last year.

The need for intermediate states can even make certain interesting alchemical free energy simulations impractical. Consider, e.g., the calculation of absolute solvation free energies resulting from the use of an implicit solvent model. [63] Such models typically consist of (at least) an additional energy term, but in most cases they also require the use of a specific cut-off radius and/or specific options for the calculations of the intramolecular electrostatic interactions. E.g., the computation of electrostatic interactions in the EEF1 implicit solvent model [64] requires (i) scaled charges for any charged moieties (e.g., N- and C-terminus, charged side chains), (ii) use of a group based cut-off truncation scheme with a cut-off radius of 9 Å and (iii) a distance dependent dielectric ($\varepsilon(r) = 1/r$). This should be contrasted with gas phase calculations, where one employs the regular charges, no cut-off, and a constant dielectric constant $\varepsilon = 1$. Since electrostatic interactions are computed completely differently in the gas phase and with the EEF1 model, simulations of intermediate states ($0 < \lambda < 1$), as are necessary in TI, are not practical. As we shall show, BAR is capable of computing the free energy difference of adding an implicit solvent model just using simulations in the gas phase and in implicit solvent, without need for intermediate states. The utility of BAR in connection with implicit solvent models was recently pointed out by Mobley et al. [65]

Quite generally, if one can compute a free energy difference of interest from simulations of the physical end points alone, the complications and limitations of the standard implementations of free energy simulations do not apply. In the remainder of this manuscript we give examples of novel applications of alchemical free energy calculations, which exploit exactly this “one-step capability” of BAR. In particular, we studied the following four problems. (1) To demonstrate the “one-step capability” of BAR compared to, e.g., EF, we computed the alchemical free energy difference between ethane and methanol in water using just simulations of the physical end states. (2) We calculated the free energy difference resulting from the use of two different cut-off radii for phosphotyrosine (pTyr) mimetics, illustrating how BAR can be used to account for the free energy cost of changing the treatment of nonbonded interactions. (3) We computed the difference in solvation free energies of capped Ala and Ser (N-acetyl-methylamide amino acids) resulting from switching from the CHARMM22 [66] to the AMBER Cornell et al. force field [67]. Finally, (4) we calculated the free energies of solvation of capped Ala and Ser that one obtains with the EEF1 [64] and FACTS [68] implicit solvent models, illustrating the utility of BAR in connection with implicit solvent models. The remainder of this paper is organized as follows: In Sect. 3.2 we briefly summarize the theory of BAR. Simulation details are provided in Sect. 3.3, followed by the presentation of the results (Sect. 3.4). We conclude with a short discussion concerning the usefulness of these types of calculations.

3.2 Theory

BAR requires two simulations, one generating a trajectory containing n_0 coordinate frames for the initial state 0 (potential energy function U_0), the other generating n_1 coordinate sets for the final state 1 (potential energy function U_1). Bennett showed that the free energy difference $\Delta A^{0\to 1}$ can formally be written as [9]

$$\Delta A^{0\to 1} = k_B T \left(\ln \frac{\sum_1 f(U_0 - U_1 + C)}{\sum_0 f(U_1 - U_0 - C)} - \ln \frac{n_1}{n_0} \right) + C \tag{3.1}$$

where f is the Fermi function,

$$f(x) = \frac{1}{1 + \exp(\beta\, x)} \tag{3.2}$$

and

$$C = k_B T \ln \frac{Q_0 n_1}{Q_1 n_0}. \tag{3.3}$$

The other symbols have their usual meaning; k_B is Boltzmann's constant, T is the temperature, and Q is the (canonical) partition function. The summation indexes 0 and 1 indicate that the sums run over all coordinate frames generated for the initial and final state. Equation 3.1 by itself would be without merit since the unknown constant C is essentially the sought after quantity (ratio of the partition functions of state 0 and 1). However, Bennett showed that C can found through an iterative procedure based on the condition

$$\sum\nolimits_1 f(U_0 - U_1 + C) = \sum\nolimits_0 f(U_1 - U_0 - C), \tag{3.4}$$

Once C has been determined so that Equation 3.4 is satisfied, the free energy difference is given by

$$\Delta A^{0\to 1} = -k_B T \ln \frac{n_1}{n_0} + C \tag{3.5}$$

Starting from Crooks' theorem [11], Shirts et al. rederived BAR using maximum likelihood techniques [61]. The demonstration that BAR can be obtained by a well understood standard technique further increased the attractiveness of the method.

3.3 Methods

All calculations were carried out with CHARMM [62]. In connection with the EEF1 implicit solvent function [64] the CHARMM19 polar hydrogen potential energy function [69] was used as prescribed by the model; in all other calculations we employed the CHARMM22 [66] or the AMBER Cornell et al. all-atom protein force field [67]. The model problems studied were (1) the alchemical free energy difference between ethane and methanol in water, (2) the free energy difference resulting from a change in cut-off radius in simulations of three pTyr mimetics, (3) the free energy difference resulting from using the AMBER [67] rather than the CHARMM [66] force field in calculations studying the solvent affinity of capped Ala and Ser, and (4) the free energy of solvation of Ala and Ser resulting from the use of the EEF1 [64] and the FACTS [68] implicit solvent models.

Free energy differences were computed with BAR, relying solely on simulations of the respective end states. For the ethane to methanol simulations this means that we conducted two MD simulations: one of ethane, one of methanol in aqueous solution. A dual topology hybrid solute mimicking either ethane or methanol was used in both calculations, cf. Ref. 70 for details. In the second test application, each pTyr mimetic was simulated in the gas phase with the cut-off radius of 70 Å used in the original study [71] and with a much longer cut-off radius of 998 Å; free energy differences were obtained from these pairs of simulations. In the AMBER/CHARMM inter-force-field calculations the respective amino acids (Ala, Ser) were simulated both in the gas phase and in aqueous solution using the Cornell et al. [67] and the CHARMM22 [66] force fields (the actual simulations were all carried out with CHARMM using the residue topology and parameter files for the Cornell et al. force field made available

Table 3.1: Overview of free energy calculations and some details of the simulation protocols used in the four model problems studied

BAR				**TI**			
Description	Environment[a]	Reps[b]	ns[c]	Description	Environment[a]	Reps[b]	ns[c]
Ethane to methanol							
short protocol[d]	wat	6	2	Reference calculation	wat	6	21
long protocol[d]	wat	6	20				
Change of cut-off radius							
short to long cut-off[e]	gas	5	168	short cut-off[f]	gas	5	42
				long cut-off[f]	gas	5	42
Change of force field							
CHARMM→AMBER, for Ala and Ser	gas	4	168				
Ala→Ser, CHARMM and AMBER force field	gas	4	168	Ala→Ser, CHARMM and AMBER force field	gas	10	84
CHARMM→AMBER, for Ala and Ser	wat	4	20				
Ala→Ser, CHARMM and AMBER force field	wat	4	20	Ala→Ser, CHARMM and AMBER force field	wat	10	42
Solvation free energies from implicit solvent models							
Ala, Ser	gas→EEF1	5	168				
Ala, Ser	gas→FACTS	5	168				

---

[a] Gas phase (gas), water (wat), or implicit solvent (EEF1 [64] or FACTS [68])

[b] Number of independent free energy simulation carried out with different random seeds for the initial velocities

[c] Total simulation length (in ns) used to obtain the free energy difference of interest. For the "one-step" BAR simulations two simulations at the respective end states of half the length indicated were carried out. For TI the cumulative simulation length of all λ-values simulated, included (re)equilibration is given. A time step of 2 fs was used, with the exception of the ethane to methanol calculations, where the time step was 1 fs

[d] Simulations at the two end states were also used to estimate the free energy difference using EF

[e] for each of the three compounds studied (PP, BP, F_2BP, cf. main text)

[f] Alchemical free energy differences BP→F_2BP, F_2BP→PP, PP→BP;

by Thomas E. Cheatham, III). The solvation free energies resulting from implicit solvent models were calculated by an analogous procedure, i.e., from one trajectory of the respective amino acid in the gas phase and another trajectory with the implicit solvent model applied. All these simulations were plain MD simulations; none of the CHARMM free energy modules were used.

In some cases, we also computed the free energy differences by TI (when this was possible), or used TI to calculate additional free energy differences that could be used to close thermodynamic cycles in order to verify the BAR results. TI calculations were carried out with the PERT free energy module of CHARMM (see the documentation at www.charmm.org), using 21 λ values ($\lambda = 0.00,\ 0.025,\ 0.075, \ldots,\ 0.975,\ 1.00$). An overview of all simulations and some additional details are given in Table 3.1.

The ethane to methanol simulations were set up as described in Ref. 70. Gas phase as well as implicit solvent model simulations were carried out using Langevin dynamics with a friction coefficient of 5 ps^{-1} on all atoms. Random forces were applied according to the target temperature of 300 K. In the solvent simulations of N-acetyl-methylamide Ala and Ser 243 TIP3P water molecules [69,72] were present, and the temperature was maintained at about 300 K by a Nosé-Hoover thermostat [73]. Lennard-Jones interactions were switched off between 9–10 Å, while electrostatic interactions were computed with the Particle Mesh Ewald method [74]. The simulation box was a truncated octahedron, cut out of cube with side length 21.4 Å. In the gas phase (and the implicit solvent simulations) coordinates were saved to disk every 100 steps, whereas trajectories were written every 10 steps in aqueous solution. The standard deviations of the free energy results were determined by repeating each simulation several times, starting from different initial random velocities. The energies required for BAR[1] were extracted from the trajectories using the EAVG command of the BLOCK module of CHARMM and processed by a `Perl` script.

---

[1] for ethane to methanol, we also attempted to compute the free energy difference with EF

3.4 Results

3.4.1 Ethane to methanol

Table 3.2 summarizes the results obtained for the alchemical free energy difference between ethane and methanol in aqueous solution based on just simulations of the two endpoints (ethane, $\lambda = 0$, methanol, $\lambda = 1$). No TI results are given since any numerical integration scheme requires values of $\langle \partial U/\partial \lambda \rangle_\lambda$ at intermediate states $0 < \lambda < 1$. We report the average value obtained from six independent simulations based on shorter and longer trajectories (referred to as "short protocol" and "long protocol in Table 3.2; see also Table 3.1). The results of Table 3.2 should be compared to $\Delta A_{H_2O}^{E \to M} = -3.05 \pm 0.05$ kcal/mole, obtained by TI from six independent series of simulations using 21 λ values each. As one sees immediately, the BAR results agree excellently with this reference value, although the standard deviation for the "short" results is somewhat high (±0.31 kcal/mole). By contrast, the "short" EF results are completely wrong. For the "long" results, the backward EF (M→E) result of +3.30 approaches the correct value, but the standard deviation remains unacceptably high. The forward EF result of +0.08 kcal/mole remains completely wrong, despite underlying simulations of in total 120 ns. The ease of Bennett to arrive at the correct result in one step is all the more remarkable considering that only 0.07 % of energy differences $\langle U_1 - U_0 \rangle_0$ and $-\langle U_0 - U_1 \rangle_1$ overlap, explaining the poor convergence of EF. We stress that the results of Table 3.2 should not be interpreted as BAR being the superior method to compute this particular free energy difference. The correct result for $\Delta A_{H_2O}^{E \to M}$ can be obtained by both TI and EF using relatively short protocols; *however*, only if simulations at intermediate values of λ are used (data not shown). Only BAR is capable of calculating the correct free energy difference from just simulations of the physical endpoints. We note in passing that attempts to use WHAM [75, 76] to calculate the free energy difference based on the same raw data used for BAR did not converge.

3.4.2 Change of cut-off radius

Our second case study was motivated by the correction of a potential inconsistency in the treatment of electrostatic interactions during the computation of alchemical

Table 3.2: Alchemical free energy difference (in kcal/mole) between ethane (E) and methanol (M) in aqueous solution.

	BAR ($\Delta A_{H_2O}^{E\to M}$)	σ[a]	forward EF ($\Delta A_{H_2O}^{E\to M}$)	σ[a]	backward EF ($\Delta A_{H_2O}^{M\to E}$)	σ[a]
short protocol	-3.00	±0.31	+0.95	±0.61	+6.02	±1.04
long protocol	-3.03	±0.04	+0.08	±1.22	+3.30	±0.90

[a] Standard deviation

Table 3.3: Free energy change resulting from the change of a 70 Å (“short c.o.”) to a 998 Å (“long c.o.”) cut-off radius in combination with a shifting function to calculate electrostatic interactions for PP, BP and F_2BP in the gas phase. All free energies are given in kcal/mole; standard deviations of all results were below ±0.04 kcal/mole.

		BAR				**TI**	
	BP	F_2BP	PP		BP→F_2BP	F_2BP→PP	PP→BP
short→long c.o.[a]	−0.29	−1.31	−0.23	short c.o[b]	280.22	−271.12	−9.13
	BP→F_2BP	F_2BP→PP	PP→BP	long c.o.[b]	281.26	−272.21	−9.08
Correction[c]	**-1.03**	**1.09**	**-0.06**	**Difference**[d]	**-1.04**	**1.09**	**-0.05**

[a] Free energy difference resulting from change in cut-off radius

[b] Alchemical free energy difference obtained at indicated cut-off radius

[c] Change in alchemical free energy difference resulting from increase of cut-off radius, obtained from “short→long c.o. results

[d] Change in alchemical free energy difference resulting from increase of cut-off radius, obtained as difference of “long c.o.” and “short c.o.” results

free energy differences between phenol phosphate (PP), benzyl phosphonate (BP) and difluorobenzyl phosphonate (F_2BP) in the gas phase. PP is the side chain analog of pTyr, whereas BP and F_2BP are the side chain analogs of important pTyr mimetics. In particular F_2BP is an important model compound when trying to understand the role of selective fluorination on the potency of potential inhibitors of pTyr binding to protein tyrosine phosphatases and SH2 domains [71].

In Ref. 71 we computed electrostatic interactions using a cut-off of 70 Å in combination with a shifting function (Equation 16 in Ref. 62). Since intramolecular electrostatic interactions in these dianions are extremely strong and vary considerably depending on the system (e.g., the polarity of the CH_2 moiety of BP is the opposite of that of the CF_2 group of F_2BP), this choice of cut-off radius in combination with a shifted potential may have been too short. Using BAR, we computed the effect of extending the short 70 Å radius to the safe value of 998 Å for each of the three compounds; the results are listed in the first row of the left half of Table 3.3. The corresponding corrections for the alchemical free energy differences between respective pairs are given in the last row of the left half of Table 3.3. While the influence of cut-off radius is negligible for $\Delta A_{gas}^{BP\rightarrow PP}$, one sees that both free energy differences involving F_2BP ($\Delta A_{gas}^{BP\rightarrow F_2BP}$ and $\Delta A_{gas}^{F_2BP\rightarrow PP}$) incur an error of $\approx$ 1 kcal/mole as a result of the too short cut-off radius.

The same corrections can of course also be obtained by repeating the alchemical free energy difference calculations with the longer cut-off radius; the influence of cut-off radius is then obtained as the difference between the free energies found with the two cut-off radii. Results of such calculations, using regular TI, are presented in the right half of Table 3.3. As can be seen by comparing the results in the last line of the left and right half of the table, the agreement between the direct calculation using BAR and the indirect calculation using TI is excellent, illustrating the correctness of our use of BAR.

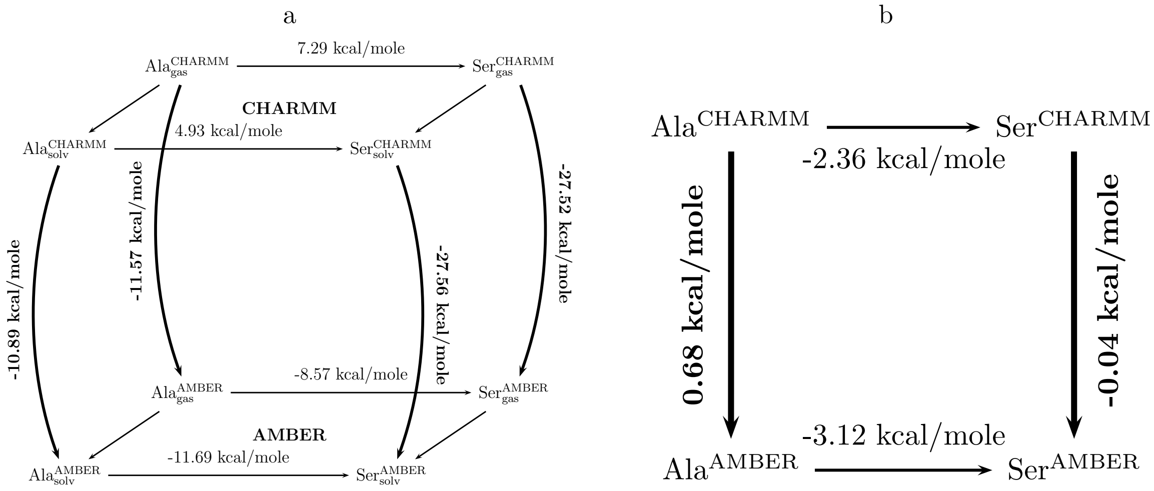
We note that although we found errors of $\approx$ 1 kcal/mole, the conclusions of Ref. 71 are not affected since the solvation free energy differences between the three compounds (the physical relevant quantities) are on the order of 10 kcal/mole and more.

3.4.3 Change of force field

The choice of the remaining case studies was motivated by the observation that the solvation free energy of (capped) amino acids and their corresponding side chain analogs (methane for alanine, methanol for serine etc.) can differ significantly [33,77, 78]. E.g., the relative solvation free energy difference between methane and methanol is -7.39 kcal/mole, whereas that between Ala and Ser is only -2.46 kcal/mole [78]. Using capped Ala and Ser as prototypical examples, we utilized BAR to compute the change in solvation free energy resulting from changing the CHARMM force field [66] to the AMBER Cornell et al. force field [67]. In the next section, we report the solvation free energies of the two amino acids that one finds using two implicit solvent methods.

The diagrams in the top and bottom sides of the "cube" shown in Figure 3.1a represent the standard thermodynamic cycles to compute solvation free energies [51]. Results for the CHARMM force field are shown in the top cycle, the AMBER results are listed in the cycle on the bottom. These free energy differences can be calculated with any of the standard techniques (TI, EF, BAR etc.); the results shown here were obtained with "one-step" BAR, but results of TI calculations led to extremely similar results (differences to BAR results reported here below 0.1 kcal/mole, data not shown). In addition, we used BAR to compute the free energy differences resulting from switching the force fields (thick vertical arrows in Figure 3.1a) for all states involved in the calculation of the solvation free energy difference between Ala and Ser. The most striking aspect of Figure 3.1a is the magnitude of the "vertical" free energy differences between the force fields, which surpass the "horizontal" differences between the apolar Ala and the polar Ser. This finding highlights the arbitrariness of absolute free energies. The picture becomes clearer when one focuses on the differences between relative solvation free energy differences ($\Delta\Delta A^{AA}_{solv} = \Delta A^{AA}_{H_2O} - \Delta A^{AA}_{gas}$) in Figure 3.1b instead of on the absolute effects in Figure 3.1a. Again, the horizontal arrows designate the relative solvation free energy differences ($\Delta\Delta A^{AA}_{solv}$) obtained with the CHARMM (top) and the AMBER force fields (bottom), whereas the vertical arrows denote the inter-force-field free energy differences between the solvation free energies of Ala and Ser obtained when replacing the CHARMM by the AMBER force field, respectively. The two force fields give almost identical

Figure 3.1: Free energy differences resulting from switching from the CHARMM [66] to the AMBER force field [67] for capped Ala and Ser. a) Inter-force field free energy differences, as well as individual free energy differences in gas phase and solution. b) Difference in relative solvation free energies as a consequence of switching force fields, as well as the relative solvation free energy differences ($\Delta\Delta A_{solv}^{AA}$) between the two amino acids.


a
$\mathrm{Ala}_{\mathrm{gas}}^{\mathrm{CHARMM}}$
7.29 kcal/mole
$\mathrm{Ser}_{\mathrm{gas}}^{\mathrm{CHARMM}}$
CHARMM
$\mathrm{Ala}_{\mathrm{solv}}^{\mathrm{CHARMM}}$
4.93 kcal/mole
$\mathrm{Ser}_{\mathrm{solv}}^{\mathrm{CHARMM}}$
-10.89 kcal/mole
-11.57 kcal/mole
-27.56 kcal/mole
-27.52 kcal/mole
$\mathrm{Ala}_{\mathrm{gas}}^{\mathrm{AMBER}}$
-8.57 kcal/mole
$\mathrm{Ser}_{\mathrm{gas}}^{\mathrm{AMBER}}$
AMBER
$\mathrm{Ala}_{\mathrm{solv}}^{\mathrm{AMBER}}$
-11.69 kcal/mole
$\mathrm{Ser}_{\mathrm{solv}}^{\mathrm{AMBER}}$
b
$\mathrm{Ala}^{\mathrm{CHARMM}}$
$\mathrm{Ser}^{\mathrm{CHARMM}}$
-2.36 kcal/mole
0.68 kcal/mole
-0.04 kcal/mole
-3.12 kcal/mole
$\mathrm{Ala}^{\mathrm{AMBER}}$
$\mathrm{Ser}^{\mathrm{AMBER}}$

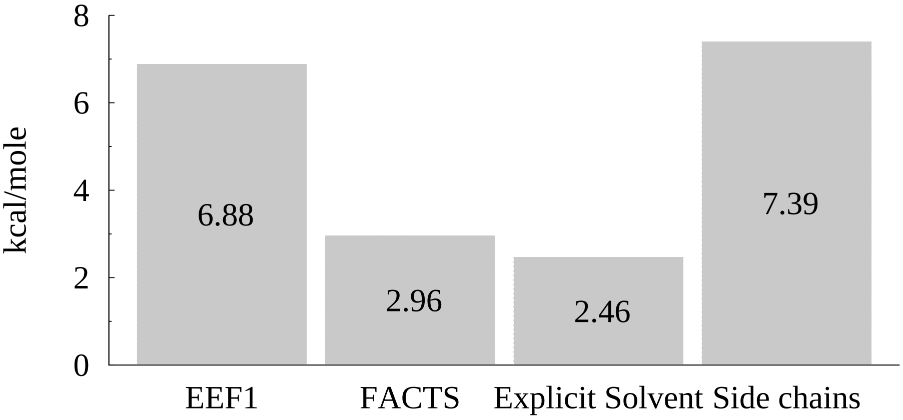


Figure 3.2: Solvation free energies differences (in kcal/mole) between capped Ala and Ser obtained with two implicit solvent methods, EEF1 and FACTS. The explicit solvent results and the side chain analog results (methane–methanol) are also shown.

results for the solvation free energy of Ser; whereas for Ala we find a difference of +0.68 kcal/mole (left vertical arrow in Figure 3.1b). Thus, interestingly, most of the differences in solvation free energies $\Delta\Delta A^{\mathrm{AMBER}}_{solv} - \Delta\Delta A^{\mathrm{CHARMM}}_{solv} = -0.73$ kcal/mole have their origin in the Ala results. Clearly, calculations of this kind can help elucidate the effect which differences in parametrization have on thermodynamic properties, separating intramolecular effects from interactions with solvent.

The results summarized in Figure 3.1b can also be used to gauge the accuracy and precision of our calculations. Summing the four free energy differences (taking into account changes of sign where appropriate!), one finds a cycle closing error of only 0.01 kcal/mole. This negligible error demonstrates the correctness of our CHARMM→AMBER calculations. Also, in Ref. 78 we reported the relative solvation free energy differences between Ala and Ser obtained from standard TI; the values of −2.46 kcal/mole and −3.20 kcal/mole for CHARMM and AMBER, respectively, agree excellently with the present results.

3.4.4 Solvation free energies from implicit solvent models

Using BAR, we directly computed the (absolute) solvation free energies of (capped) Ala and Ser when using the EEF1 and the FACTS implicit solvent models. To compare the results with those obtained with explicit solvent, as well as to the solvent affinity of the respective side chain analogs, we show the relative solvation free energy between Ala and Ser in Figure 3.2. One sees immediately that there are huge differences between the models. While the FACTS result (-2.96 kcal/mole) is relatively close to the explicit solvent value of -2.46 kcal/mole [78], the EEF1 result is off by more than -3.5 kcal/mole and resembles the side chain analog rather than the amino acid result.

As discussed in the Introduction, for most implicit solvent models it would be rather involved to compute the energy and forces at intermediate values of the coupling parameter λ; hence, the possibility of computing the free energy difference in one step is extremely useful. We found that in some cases EF sufficed to compute the solvation free energy of implicit solvent models (using the average of a forward (gas phase → implicit solvent) and backward (implicit solvent → gas phase) calculation), but BAR turned out to be more precise and reliable (data not shown).

The EEF1 implicit solvent model is one of the few implicit solvent methods that is supported by conventional free energy modules in CHARMM; hence, we used the standard thermodynamic cycle [51] and computed $\Delta\Delta A^{AA}_{solv} = \Delta A_{EEF} - \Delta A^{AA}_{gas}$, where ΔA_{EEF} denotes the free energy change between Ala and Ser using EEF1 and ΔA^{AA}_{gas} is the corresponding free energy difference in the gas phase. With $\Delta A_{EEF} = -6.49$ kcal/mole and $\Delta A^{AA}_{gas} = 0.46$ kcal/mole, we obtain $\Delta\Delta A^{AA}_{solv} = -6.95$ kcal/mole, in excellent agreement with the value reported in Figure 3.2 and, thus, verifying the correctness of the direct approach.

3.5 Concluding Discussion

Several studies found that BAR is frequently more efficient than TI and EF in alchemical free energy simulations [21, 58–60]. In this work we presented three examples of calculations that would either have been much more difficult or not possible at all without BAR. E.g., calculating the free energy resulting from the change of cut-off radius (cf. Table 3.3) directly would not be possible in the framework of traditional free energy methods. Admittedly, for the specific system, simply repeating the calculations with a longer cut-off would have been equally quick. However, corrections of this kind may well be advantageous for larger systems; in particular, since already existing trajectories can be reused. The other two examples are of even greater practical relevance. BAR is extremely useful in connection with implicit solvent models, see also Ref. 65. The method makes it possible to calculate and compare solvation free energies resulting from the use of different implicit solvent models, which should prove useful in comparing the quality of such approaches.

The calculation of free energy differences resulting from swapping the underlying force field, is relevant for parametrization in general. As an example, suppose that one has modified an existing force field and that free energy calculations on model compounds were carried out to validate the original parameters. Therefore, upon each modification of the force field, these calculations should be repeated, a tedious and expensive exercise. On the other hand, it is much more likely that regular MD simulations of the systems of interest with the original and the modified force field have been carried out to compare a variety of properties. Using BAR in the manner described here, the trajectories written during these MD simulations can be used to quantify the free energy cost of the force field modification. In other words, BAR permits one to relate relatively cheaply the effect which force field modifications have on thermodynamic properties and to identify the changes which have the largest effect.

At least in connection with CHARMM, "one-step" free energy simulations with BAR have a potential additional benefit not discussed so far. As mentioned in the Introduction, the need for intermediate steps in alchemical free energy calculations adds a layer of complexity to the underlying MD (or MC) code used in free energy simulations, which also impacts performance. E.g., the new fast lookup code of CHARMM [79] cannot be used if either of the three free energy modules is used (BLOCK, TSM, PERT). Further, several usage scenarios of these modules effectively necessitate the use of the generic slow energy routines, as well as the slow nonbonded list routines. The loss in performance is considerable and can be as large as a factor of four. As illustrated by the ethane–methanol example, BAR can compute an alchemical free energy difference in a single step. The required trajectories of ethane and methanol in water (solutes suitably modified by attaching a dummy group representing methanol and ethane, respectively) were obtained with the fastest routines available in CHARMM (lookup table energy routines plus fast nonbonded list generator). While in this simple case (and small system) the shorter simulation times compared to the slow energy routines are rather irrelevant, the potential performance gain is certainly of interest in large(r), real world applications. We are presently exploring further this aspect of using BAR rather than TI or EF.

Even before its recent rediscovery, BAR proved to be a valuable tool. Already some twenty five years ago, Ferguson pointed out the superiority of BAR in situations of poor overlap between states [80]. In a study on the hydration free energy of water, Hummer et al. exploited the efficiency of BAR to illustrate the theoretical considerations with a large number of free energy data, which most likely would have been prohibitively costly to calculate with other methods (such as TI or EF) [81]. Similarly, a more recent work on water conduction through hydrophobic channels is an interesting example of the use of BAR in an unusual context [82]. We hope that the examples presented here show that BAR can facilitate the calculation of free energy differences in non-traditional situations. Quite generally, whenever one has trajectories of two states 0 and 1 and when there is any overlap between $\langle U_1 - U_0 \rangle_0$ and $-\langle U_0 - U_1 \rangle_1$, one can rely on BAR to compute the free energy difference between the two states. Thus, we suspect that there are many more "unorthodox" uses of BAR waiting to be explored.

Chapter 4

Non-Boltzmann Sampling and Bennett's Acceptance Ratio Method: How free energy simulations can profit from bending the rules.

The exact computation of free energy differences requires adequate sampling of all relevant low energy conformations. Especially in systems with rugged energy surfaces, adequate sampling can only be achieved by biasing the exploration process, thus yielding non-Boltzmann probability distributions. To obtain correct free energy differences from such simulations, it is necessary to account for the effects of the bias in the post-production analysis. We demonstrate that this can be accomplished quite simply with a slight modification of Bennett's Acceptance Ratio method, referring to this technique as Non-Boltzmann Bennett. We illustrate the method by several examples and show how a creative choice of the biased state(s) used during sampling can also improve the efficiency of free energy simulations.

4.1 Introduction

The calculation of free energy differences is one of the most promising applications of computational chemistry. It bridges the gap between the microscopic world of molecular simulation and one of the most fundamental macroscopic thermodynamic properties, the free energy. Thus, free energy simulations provide direct means to address a wide range of biologically relevant questions. Successful applications include the calculation of binding affinities of ligands [15,16], the study of enzymatic reactions [17], of molecular solvation [18,19], and of protein stability as a function of point mutations [20].

The vast majority of free energy simulations reported to date were carried out by two families of methods: Thermodynamic Integration (TI) [7] and (several variants of) the Exponential Formula (EF), often also referred to as Thermodynamic Perturbation, Free Energy Perturbation or Exponential Averaging [8]. Recently, however, the use of Bennett's acceptance ratio method (BAR) [9] has become more and more prevalent. Although first described in 1976, it was practically never used in free energy simulations until seven or eight years ago. Several studies have found BAR to be superior in terms of efficiency to TI and EF in alchemical free energy simulations [21, 58–60]. Fewer intermediate states and, hence, shorter total simulation lengths suffice to calculate a free energy difference accurately and precisely compared to other methods, notably TI and EF. In addition, the overlap criterion of BAR [9] provides a rational measure for gauging the quality of a free energy simulation. We note that depending on one's point of view BAR may also be regarded as a sophisticated variant of EF, an equilibrium version of Crook's theorem [11], or even as a non-discretized weighted histogram analysis method (WHAM) conducted on just two states [75, 83].

A major challenge for any computer simulation based method (not just free energy simulations) is the need for adequate sampling. If relevant parts of phase space are not visited during a simulation, any results derived from it are of dubious quality. This problem has been addressed by multiple techniques including, e.g., umbrella sampling [22, 23], Hamiltonian replica exchange [84, 85], accelerated molecular dynamics (AMD) [86], conformational flooding [87] etc. Since all these techniques go beyond the Boltzmann sampling of conventional molecular dynamics (MD) or Monte

Carlo simulations, we subsume them by the term “non-Boltzmann sampling”. Often they can speed up the convergence of simulations significantly. E.g., for the mean field Ising model on N sites theoretical considerations demonstrate that with non-Boltzmann sampling one can obtain estimates in a time which is polynomial in N, whereas time would be exponential in N if conventional Boltzmann sampling were used. [88]. Thus, though the computational cost can still be considerable, an efficient non-Boltzmann sampling technique is by far the better choice for large systems.

Focusing on free energy simulations, it has been found that even the rather trivial case of a side chain being trapped in a conformational minimum can lead to incorrect results or at least to very slow convergence of the calculations. Several groups observed this effect for the case of solvation free energies [24,89–91], and even more so in the context of binding free energy calculations [92,93]. Since in some cases even very long simulations are not enough to escape from such local minima, special techniques have to be applied. Woods et al. suggested replica exchange TI (RETI) in coupling parameter space (λ-space). This special variant of Hamiltonian replica exchange does not involve non-Boltzmann sampling since the system only switches between λ-states at which simulations would be carried out anyways. While RETI enhances configurational sampling (e.g., of the solvent) it is less clear whether this type of exchange moves may help escape from conformational substates. Therefore, other (Hamiltonian) replica exchange schemes have been proposed in the context of TI [94, 95]. A rather different approach is the “Confine-and-release method” by Mobley et al. [93], which relies on the use of constraints; see also Refs. 24, 96, 97. However, the most straightforward approach appears to be applying biasing potentials to the problematic degree(s) of freedom; for TI and EF this has already been investigated [24, 98], and the use of biasing potentials is of course central to WHAM [75, 99].

Here we show that non-Boltzmann sampling can be used equally easily together with BAR. We will refer to this modification of BAR as non-Boltzmann Bennett (NBB), thus reflecting its methodological similarity to non-Boltzmann Thermodynamic Integration (NBTI) [24]. Similarly to NBTI and WHAM, NBB is an extension of umbrella sampling [22]. In addition to demonstrating the utility of NBB for a straightforward task (overcoming hindered rotation of amino acid side chains), we explore additional uses of non-Boltzmann sampling in connection with free energy simulations. We note that a similar application of WHAM in a non-standard context was described recently [100]. In an earlier study [101] we demonstrated "unorthodox" applications of BAR, which are based on the observation that in some situations BAR can compute free energy differences using just the physical end states (which is not possible with, e.g., TI and EF). By exploiting this strength of BAR in combination with "creative" biasing potentials or, rather, biased states, further "unorthodox" uses of BAR/NBB can be devised, which are able to improve the efficiency of free energy simulations.

To emphasize the versatility and usability of NBB, we restrict ourselves to biased states that do not require any specialized computer code. Thus, the calculations described in this work should be repeatable with any simulation package or force field. In particular, we consider the following model tasks/problems: (1) We use a simple biasing potential to overcome hindered rotation of amino acid side chains about the χ_1 angle (Sects. 4.3.1, 4.4.1). (2) For a small toy system we show that the free energy difference between two states with significantly different energy landscapes can be calculated based on simulations of a third state which includes the relevant phase space regions of the systems of interest (Sects. 4.3.2, 4.4.2). (3) We show that NBB can correct small errors in free energy simulations without having to repeat all simulations. Specifically, we use simulations in which a force field term was "forgotten" and show that by appropriate re-weighting the correct free energy difference is obtained as if the full (correct) force field had been used during the underlying MD simulations (Sects. 4.3.3, 4.4.3). Finally, (4) by an analogous approach simulations with a fast implicit solvent model can be utilized to obtain results that correspond to the use of a high quality implicit solvent model (which would be much slower). This last example illustrates in particular how the separation of production and analysis can be exploited to gain efficiency (Sects. 4.3.4, 4.4.4).

We note that there is some potential for confusion concerning nomenclature/terminology. The trajectories obtained during the simulations of the various biased states employed in this work constitute standard Boltzmann sampling for these modified states; yet, sampling is of the non-Boltzmann type for the physical states of interest. Thus, our choice of terminology for the method, *non-Boltzmann*

Bennett, reflects the point of view of the physical system. Further, the list of example applications just given indicates that the utility of NBB goes beyond straightforward biasing of selected degrees of freedom. In fact, there often is no biasing potential in the traditional sense, i.e., a single potential energy term favoring specific regions of phase space. Instead, the full potential energy function is altered more or less subtly to achieve a specific purpose. A related situation is found, e.g., when employing WHAM to analyze data from generalized ensemble simulations [99]. In cases where speaking of a biasing potential might be misleading, we refer to the simulated system as biased state or *sampled state*, and to the the system of interest as *target state*.

The remainder of this paper is organized as follows. First, we outline the theory of NBB (Sect. 4.2.1) and describe the types of biased (or *sampled*) states that we use (Sect. 4.2.2). Methodological details of the simulations are presented in Section 4.3. We then present the results for the four model problems outlined above (Sect. 4.4) and conclude with a short discussion concerning the usefulness of these types of calculations in Section 4.5.

4.2 Theory

4.2.1 The Non-Boltzmann Bennett Method

To compute the free energy difference between two states 0 and 1, BAR utilizes the information obtained from simulations of both states simultaneously [9]. The free energy difference between states 0 (potential energy function U_0) and 1 (potential energy function U_1) is given by

$$\Delta A^{0\to 1} = \beta^{-1}\left(\ln \frac{\langle f(U_0 - U_1 + C)\rangle_1}{\langle f(U_1 - U_0 - C)\rangle_0}\right) + C \tag{4.1}$$

The subscripts 0 and 1 in Equation 4.1 indicate that the ensemble averages $\langle\ \rangle$ are calculated from the trajectories of the initial (0) and final state (1), respectively. The symbol f denotes the Fermi function,

$$f(x) = \frac{1}{1 + \exp(\beta\, x)} \tag{4.2}$$

and

$$C = \beta^{-1} \ln \frac{Q_0 n_1}{Q_1 n_0} \tag{4.3}$$

where Q denotes the respective partition function, β has the usual meaning of $1/k_BT$, and n_0 and n_1 are the number of configurations of state 0 and 1 from which the ensemble averages are evaluated. The unknown constant C, which corresponds essentially to the free energy difference of interest, is found iteratively. Starting from an initial guess, one searches for the value of C so that the argument of the logarithm in Equation 4.1 equals unity since in this case the free energy difference is given by

$$\Delta A^{0\to 1} = -\beta^{-1} \ln \frac{n_1}{n_0} + C \tag{4.4}$$

In NBB, simulations are carried out for a biased (sampled) state (potential energy function U^{biased}) with special properties instead of the target state (the physical system) with the regular potential energy function U. Such states are generated by applying a biasing potential (V^{bias}) to the original system (i.e., $U^{biased} = U + V^{bias}$); however, to remain completely general in the choice of the biased state, we define the biasing potential as

$$V^{bias} = U^{biased} - U \tag{4.5}$$

Torrie and Valleau [22] showed how to obtain an unbiased ensemble average $\langle X \rangle$ of some property X from simulations of a biased state:

$$\langle X \rangle = \frac{\left\langle X \exp\left(\beta V^{bias}\right)\right\rangle_b}{\left\langle \exp\left(\beta V^{bias}\right)\right\rangle_b} \tag{4.6}$$

where we use the notation $\langle\ \rangle_b$ to indicate that the ensemble averages on the right hand side of Equation 4.6 are evaluated from simulations of the biased state. The working equation of NBB is thus easily found by applying Equation 4.6 to the two ensemble averages in Equation 4.1 (with X being $f(U_0 - U_1 + C)$ and $f(U_1 - U_0 - C)$, respectively), i.e.

$$\Delta A^{0\to 1} = \beta^{-1} \ln \left(\frac{\left\langle f(U_0 - U_1 + C) \exp\left(\beta V_1^{bias}\right)\right\rangle_{1,b}}{\left\langle f(U_1 - U_0 - C) \exp\left(\beta V_0^{bias}\right)\right\rangle_{0,b}} \frac{\left\langle \exp\left(\beta V_0^{bias}\right)\right\rangle_{0,b}}{\left\langle \exp\left(\beta V_1^{bias}\right)\right\rangle_{1,b}} \right) + C. \tag{4.7}$$

To use Equation 4.7 one has to evaluate three quantities for each frame of the trajectories: For the biased trajectory of state 0, it is necessary to compute U_0, U_1 and V_0^{bias}, while for state 1, U_0, U_1 and V_1^{bias} are required. Since U_0 and U_1 would have to be also calculated for regular BAR, the computational overhead due to the costs of determining V^{bias} is quite low (except for the case of extremely complicated sampling states, where it is more straightforward to calculate V^{bias} directly according to Equation 4.5).

4.2.2 Applications of NBB

Before turning to specific examples, we briefly discuss some types of sampled (biased) states that should prove useful in free energy simulations. The most straightforward application of classical biasing potentials consists in overcoming a known barrier, such as hindered rotation about a dihedral angle. However, employing specially designed sampled states can enhance sampling in general (instead of just a selected degree of freedom). This can be pivotal in free energy simulations since a free energy difference between two states can only be calculated if their phase spaces overlap; otherwise, intermediate states have to be introduced. By extending the phase spaces of both end states by employing suitable sampled states, the overlap region between them can be enlarged. This can enhance the efficiency of the free energy simulation and decrease the number of necessary intermediate steps [102, 103]. In fact, if the phase space of the sampled state is large enough to envelop both end states, a single trajectory may be sufficient to compute the free energy difference.

As an example of a rather untypical sampled state we consider the case where you detect a (small) error in your simulation setup after you have generated the trajectories which you plan to evaluate with BAR. Normally, one has to rerun all simulations, but depending on the system studied the computational cost may be significant. Instead, provided the relevant regions of phase space were still sampled during the defective simulations, one can regard the sampled state as simulations in the presence of a (admittedly, rather peculiar) biasing potential, $V^{bias} = U^{faulty} - U^{correct}$. The faulty trajectories can be analyzed using NBB instead of regular BAR, leading to the correct free energy difference. Using WHAM, Shirts et al. used an analogous approach to correct for missing dispersion interactions because of (too) short cut-off radii in MD simulations [100].

The idea underlying the approach just described can be utilized to enhance the computational efficiency of free energy simulations by using different levels of accuracy during the production of trajectories and their analysis. A computationally cheap(er), approximate potential energy function is used in the sampled state for the exploration of phase space, followed by an analysis of the trajectories with an exact, but computationally expensive potential function. This corresponds to NBB with a biasing potential $V^{bias} = U^{approximate} - U^{exact}$. Since coordinates are usually saved only at every tenth or even hundredth step of the MD simulation, the expensive energy terms are computed only for a small fraction of the total simulation steps, thus reducing computational cost.

4.3 Methods

All calculations were carried out with CHARMM [62, 104], using the CHARMM22 all-atom force field [66]. The gas phase and implicit solvent model simulations were conducted with Langevin dynamics, using a friction coefficient of 5 ps^{-1} on all atoms and random forces according to a target temperature of 300 K. To justify a time step of 2 fs, hydrogen masses were set to 10 amu. Trajectories were usually written every 100 steps (exceptions will be stated explicitly). Details of explicit solvent simulations are given when describing the respective system.

The standard deviations reported were determined by repeating each free energy simulation four times, starting with different initial random velocities. The energies of the respective states required for BAR and NBB were extracted from the trajectories using the EAVG command of the BLOCK module of CHARMM; the BAR/NBB analysis was carried out by a `Perl` program.

4.3.1 Leucine–Asparagine

The potentials of mean force (PMF) of blocked asparagine (Asp) and leucine (Leu) with respect to χ_1 and χ_2 differ substantially, particularly in the gas phase. For this reason, the calculation of the relative solvation free energy difference between these two amino acids was chosen in Ref. 24 as a model problem to investigate the effect of conformational substates and rotational barriers on the convergence of free energy simulations. Since the effect is particularly pronounced in the gas phase, we calculate the alchemical free energy difference between Asp and Leu in the gas phase using regular BAR and NBB.

The alchemical mutation was set up in the single topology framework [105] as described in Ref. 24. Eleven λ-states ($\lambda = 0.0, 0.1, \ldots, 1.0$) were used. At each λ-state simulations of 10 ns length were carried out. The regular BAR calculations (no biasing potential present) were carried out thrice, starting from different sets of initial χ_1/χ_2 values. In the non-Boltzmann TI (NBTI) calculations described in Ref. 24 an adaptive umbrella potential was used in combination with TI to compute this free energy difference. Here, we used a much simpler, static biasing potential instead: the dihedral energy terms for χ_1 and χ_2 were deleted (which is equivalent to adding a biasing potential that counteracts the dihedral energy terms exactly).

4.3.2 Five-atomic systems

Leitgeb et al. also reported results of several free energy simulations carried out for five-atomic model systems [24]. Because of the smallness of the systems, the free energy differences between them can be calculated by numerical integration of the partition function; i.e., one can obtain reference results independent of free energy simulations. The three model systems (cf. also Ref. 24) are unbranched, nonlinear five-atomic molecules. The equilibrium bond lengths were 1.53 Å, all bond angles were 111°. Two dihedral angle terms (ϕ_1, ϕ_2) were present in each system.

- In system I, the same threefold torsional potential (multiplicity $n_1 = 3$, force constant $k_1 = 2.5$ kcal/mol; $n_2 = 3$, $k_2 = 2.5$ kcal/mol) was applied to both ϕ_1 and ϕ_2, resulting in nine equivalent minima.

- In system II the sum of two potentials was applied simultaneously to each dihedral ($n_{1,1} = 3$, $k_{1,1} = 2$ kcal/mol, $n_{1,2} = 1$, $k_{1,2} = 2$ kcal/mol; $n_{2,1} = 3$, $k_{2,1} = 2$ kcal/mol, $n_{2,2} = 1$, $k_{2,2} = 2$ kcal/mol), resulting in a single global minimum. In addition, there are four local minima that can be potentially reached in a normal MD simulation.

- In system III, the dihedral potentials were the same as in system II, but the dihedral force constants were raised ($n_{1,1} = 3$, $k_{1,1} = 3.5$ kcal/mol, $n_{1,2} = 1$, $k_{1,2} = 3$ kcal/mol; $n_{2,1} = 3$, $k_{2,1} = 3.5$ kcal/mol, $n_{2,2} = 1$, $k_{2,2} = 3$ kcal/mol). In addition, intramolecular electrostatic and Lennard Jones interactions were present. This results in two equivalent global minima and one local minimum that is in principle accessible. However, these three minima are separated by high energy barriers.

We attempted to compute the free energy differences between systems I, II and III in two ways. First, BAR was used based on 84 ns simulations of the respective end states. In addition, since the nine minima of I include all relevant minima of II and III, we used NBB to compute the free energy differences of interest based on the simulations of just state I. To obtain the ensemble averages of systems II and III from system I, we employed the biasing potentials $V_{II}^{bias} = U_I - U_{II}$ and $V_{III}^{bias} = U_I - U_{III}$, respectively.

4.3.3 Pseudoglycine–Glycine

In a recent study of relative solvation free energy differences between blocked amino acids [78], we utilized a glycine-like molecule as intermediate state, to which we refer as pseudoglycine (PG). PG differs from normal glycine (Gly) only by the force field type of the C_α carbon. Obviously, it is useful to know the relative solvation free energy difference between PG and Gly. The exact value of this free energy difference depends on the force field used, e.g., whether the backbone cross term map (CMAP) correction [106] is applied or not.

We computed the free energy differences between between PG and Gly in aqueous solution with and without the CMAP term. The solutes were placed in a truncated octahedron (cut out of a cube with a side length of 37.25 Å); 862 TIP3P water molecules [69,72] were present. The temperature was maintained at about 300 K by a Nosé-Hoover thermostat [73]. Lennard-Jones interactions were switched off between 10 and 12 Å, while electrostatic interactions were computed with the Particle Mesh Ewald method [74]. Free energy differences were computed with BAR based on simulations of just the physical endpoints (no intermediate λ-values were simulated). The total simulation length for each state was 10 ns. Trajectories were saved to disk every tenth step.

We then assume that we only have simulations without CMAP correction available, but are interested in the free energy differences that would be obtained in the presence of the CMAP correction. To do so, we regard the simulations without

CMAP as simulations with CMAP plus a biasing potential that completely counteracts CMAP, i.e., $V^{bias} = -U_{CMAP}$. Thus, using NBB we can compute the free energy difference between PG and Gly as if the CMAP correction had been applied during the simulations.

4.3.4 Implicit solvent

A related strategy can also help increase computational efficiency. In Ref. 78 we compared free energy differences obtained in explicit solvent and implicit solvent simulations. The most accurate implicit solvent model, GBMV [107] gave results in good agreement with explicit solvent. This contrasts, e.g., with the FACTS [68] model, which for some amino acids led to solvation free energies which deviated considerably from the reference results. The higher accuracy of GBMV, however, has a price: GBMV is (at least) ten times slower than FACTS.

Because of this huge disparity in performance, we tried the following approach: Trajectories of the blocked amino acids in implicit solvent were generated with FACTS, but during post-processing (evaluation of the ensemble averages needed for Equation 4.7) GBMV was used. The simulations were set up as described in Ref. 78; in particular, we considered the amino acids alanine (Ala), serine (Ser), valine (Val), threonine (Thr), leucine (Leu), asparagine (Asp), phenylalanine (Phe) and tyrosine (Tyr). Simulation lengths were 200 ns. In addition, the dihedral potential terms for χ_1 and χ_2 were removed during the simulation to facilitate sampling (as described in the first example). This is equivalent to the presence of a second biasing potential. Overall, we therefore employed a biasing potential consisting of several terms, $V^{bias} = V^{bias,1} + V^{bias,2}$, where $V^{bias,1} = U_{FACTS} - U_{GBMV}$ accounts for the faster implicit solvent model and $V^{bias,2} = -U_{dihe}(\chi_1, \chi_2)$ lowers the rotational barriers about the side chain dihedral degrees of freedom.

4.4 Results

4.4.1 Leucine-Asparagine

Our first application of NBB is the computation of the alchemical free energy difference between Leu and Asn in the gas phase ($\Delta A_{gas}^{Leu \rightarrow Asn}$). As reported by Leitgeb et al. [24], TI free energy simulations based on normal MD simulations were very imprecise, with results ranging from -59.7 to -68.8 kcal/mol (which constitutes a difference of ≈ 9 kcal/mol). The underlying sampling problems were overcome by two approaches: Using NBTI, a free energy difference of -61.7 ± 0.47 kcal/mol was found; using an approach originally suggested by Straatsma and McCammon [96], free energy differences ranging from -61.6 to -62.1 kcal/mol were obtained.

In the sampled state of the NBB simulations reported here, we used a much simpler biasing potential compared to Ref. 24; i.e., we simply suppressed the dihedral angle potential energy term for the χ_1 and χ_2 degrees of freedom. In contrast to the adaptive umbrella potential used in earlier work [24], such a static biasing potential does not flatten the PMF since intramolecular nonbonded interactions are not countered. In Figure 4.1 we show the PMF about χ_1 for Leu (left hand side) and Asn (right hand side) before (solid line) and after (dotted line) removing the two dihedral energy terms. One sees that the barriers are lowered by approximately 3 kcal/mol; at the same time the overall shape of the PMF is retained.

To test whether this trivial modification of the potential energy function in combination with NBB suffices to yield correct free energy differences for $\Delta A_{gas}^{Leu \rightarrow Asn}$, we compared results obtained with BAR (based on regular, unbiased trajectories) and NBB (with the biasing potential described above). The results are summarized in Table 4.1. We report results for three sets of simulations, started from different initial combinations of χ_1 and χ_2; the starting conformations are listed in the leftmost column of Table 4.1. The BAR results (second column from the left) suggest problems similar to those encountered in the earlier TI simulations [24]. Depending on the starting conformation of the side chain, results range from -61.3 to -65.2 kcal/mol. The NBB results, on the other hand, vary only between -62.0 and -62.3 kcal/mol, which agrees reasonably with the results of Ref. 24. The example demonstrates that a simple, "conventional" biasing potential can facilitate sufficient

Figure 4.1: Comparison of the potentials of mean force (in kcal/mol) of leucine and asparagine in gas phase before and after deleting the dihedral potentials of χ_1 and χ_2. The arrows indicate the reduction of the respective energy barriers due to the removal of the dihedral potential.

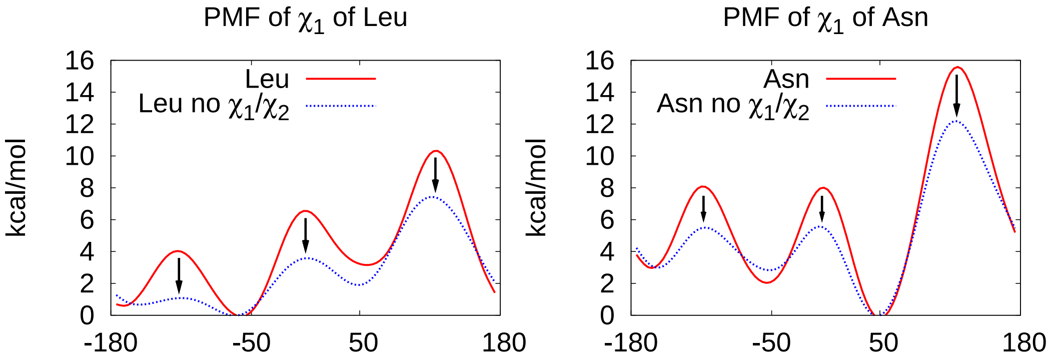


Table 4.1: Comparison of relative free energy calculations (in kcal/mol) between leucine and asparagine in gas phase as calculated with normal BAR and NBB employing biasing potentials on χ_1 and χ_2.

χ_1/χ_2[a]	BAR	NBB
-180/-180	-61.28 ± 0.35	-62.03 ± 0.29
60/-180	-62.32 ± 0.49	-62.33 ± 0.47
60/-80	-65.15 ± 0.44	-62.07 ± 0.16

[a] Initial conformation of side chain dihedrals

sampling of the rotational substates to obtain converged free energy differences.

4.4.2 Five-atomic systems

The PMFs about the two dihedral angles of the three five-atomic model systems I – III are displayed in Figure 4.2. The dark areas mark energy minima; the accessible phase space about the minima is indicated by dotted lines. All regions beyond the dotted lines are so high in energy that they will not be sampled during normal MD simulations. From the comparison of the PMFs of system II and III, it should be clear that computing the free energy difference between these two systems will be somewhat of a challenge. Not only are the minima of the two systems completely different, but the minimum of II is also a high energy region of III and vice versa.

Results of BAR and NBB calculations of the free energy differences between the three systems are listed in Table 4.2. Calculating free energy differences between I and II, or I and III is unproblematic; the BAR results are in good agreement with the quasi-analytical reference results of Ref. 24. Since the respective minima of II and III are also minima of I (cf. Figure 4.2) and since the energy barriers separating the nine minima of I are relatively low (4 kcal/mol), phase space overlap is guaranteed. By contrast, a regular BAR free energy simulation for the free energy difference between II and III did not converge because of lack of phase space overlap. This agrees with failure of earlier attempts to compute this free energy difference with TI [24].

However, the observation that in terms of relevant minima I is a superset of II and III suggests to use simulations of I to compute the free energy difference between II and III. Defining system I as the sampled state contrasts with the conventional conception of biasing potentials. Typically, V^{bias} is given by a specific potential energy term (e.g., in the Asn–Leu example the biasing potential is the sum of two dihedral energy terms). By contrast, here the biasing potential is literally the energy difference between the full potential energy function of II (or III) and I. The result of the NBB calculation for the free energy difference between II and III (together with the unproblematic cases I→II and I→III) is shown in the third column of Table 4.2. The value of -14.38 ± 0.20 kcal/mol is in excellent agreement with the quasi-analytical reference result. One may argue that the free energy difference

Figure 4.2: Potentials of mean force of the five-atomic systems I, II and III as a function of the dihedral angles ϕ_1 and ϕ_2. Dark areas represent important regions of phase space with energy levels below 2 kcal/mol, while dotted lines encircle the accessible phase space with energy levels below 5 kcal/mol.

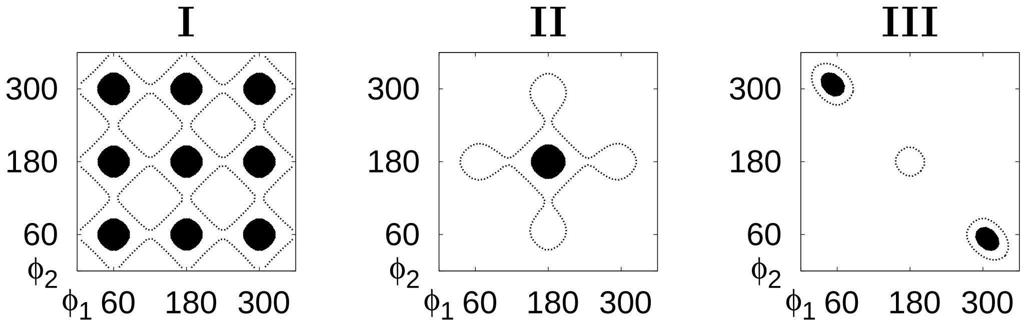


Table 4.2: Free energy differences between the five-atomic systems I, II, and III.

Mutation	BAR	NBB[a]	Reference[b]
I→II	1.23 ± 0.09	1.24 ± 0.03	1.22
I→III	-13.18 ± 0.09	-13.15 ± 0.11	-13.06
II→III	no convergence	-14.38 ± 0.20	-14.28

[a] Based on re-weighted trajectories of system I only

[b] Reference free energy difference obtained by numerical integration of the partition functions [24]

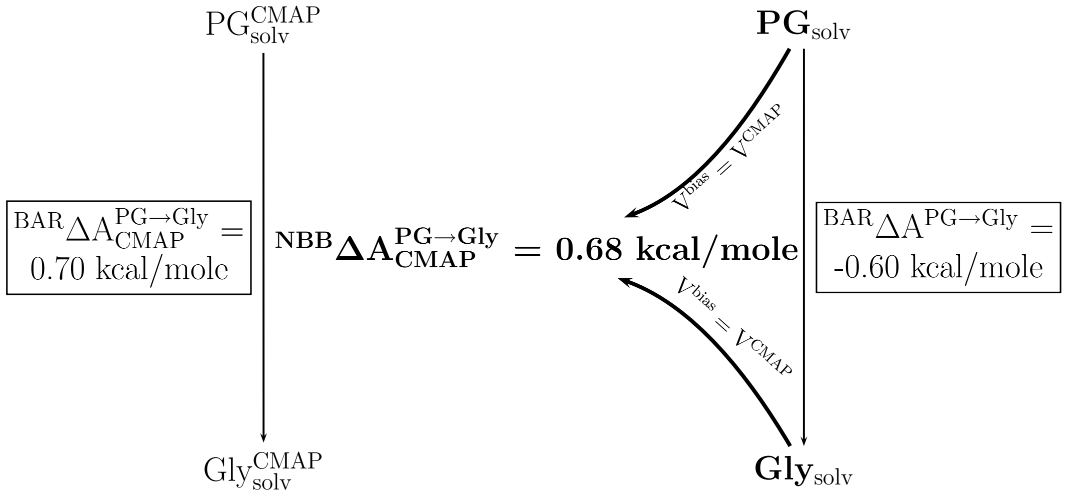
II→III could also have been obtained by intermediate λ-states. Note, however, that this was attempted without success (albeit with TI instead of BAR) in Ref. 24. The large difference between the end states makes finding suitable alchemical intermediates difficult; biasing the system to enhance sampling is more efficient.

4.4.3 Pseudoglycine – glycine

It is, alas, all too common to discover after concluding a long series of simulations that some small error has crept in. We mimic this scenario for the case of the alchemical free energy difference between PG and Gly in aqueous solution by assuming that the CMAP correction [106] was forgotten during the MD simulations of PG and Gly. We consider the trajectories generated without CMAP as biased states obtained with a biasing potential that counteracted CMAP.

The results for this application of NBB are summarized in Figure 4.3. The vertical arrows correspond to the reference calculations, in which the free energy differences between PG and Gly were obtained with regular BAR from simulations with ($^{BAR}\Delta A^{PG\to Gly}_{CMAP}$) and without CMAP ($^{BAR}\Delta A^{PG\to Gly}$). The two results are not devoid of interest since the free energy difference between PG and Gly changes sign when CMAP is applied. The curved arrows in Figure 4.3 indicate the use of NBB to obtain $\Delta A^{PG\to Gly}_{CMAP}$ from the simulations carried out *without* CMAP; we label this free energy difference $^{NBB}\Delta A^{PG\to Gly}_{CMAP}$. As one sees, $^{BAR}\Delta A^{PG\to Gly}_{CMAP} = 0.70$ kcal/mol and $^{NBB}\Delta A^{PG\to Gly}_{CMAP} = 0.68$ kcal/mol agree excellently. The computational overhead of this correction is minimal; all we had to do was to compute the CMAP energy term for each frame of the trajectories of PG and Gly obtained without CMAP. Without NBB, all simulations would have to be repeated. Therefore, NBB may reduce computational cost significantly if free energy differences need to be re-determined under slightly modified simulation conditions or small changes in the force field (e.g., during parametrization).

Figure 4.3: Free energy differences (kcal/mol) in solution between PG and Gly. The two straight arrows indicate the reference results calculated with BAR with (left) and without (right) the CMAP correction. The bold, curved arrows symbolize the use of NBB to compute the free energy difference between PG and Gly *with CMAP* from simulations *without CMAP*.


$\mathrm{PG}_{\mathrm{solv}}^{\mathrm{CMAP}}$
$\mathbf{PG}_{\mathrm{solv}}$
$^{\mathrm{BAR}}\Delta \mathrm{A}_{\mathrm{CMAP}}^{\mathrm{PG}\rightarrow\mathrm{Gly}} =$ 0.70 kcal/mole
$V^{\mathrm{bias}} = V^{\mathrm{CMAP}}$
$^{\mathbf{NBB}}\boldsymbol{\Delta}\mathbf{A}_{\mathbf{CMAP}}^{\mathbf{PG}\rightarrow\mathbf{Gly}} = \mathbf{0.68\ kcal/mole}$
$^{\mathrm{BAR}}\Delta \mathrm{A}^{\mathrm{PG}\rightarrow\mathrm{Gly}} =$ -0.60 kcal/mole
$V^{\mathrm{bias}} = V^{\mathrm{CMAP}}$
$\mathrm{Gly}_{\mathrm{solv}}^{\mathrm{CMAP}}$
$\mathbf{Gly}_{\mathrm{solv}}$

4.4.4 Implicit solvent

Instead of just correcting results, separating the production and the analysis phase of free energy calculations makes possible additional applications. In a recent publication, we compared relative solvation free energies between several blocked amino acids obtained with various implicit solvent models to the free energy differences from explicit solvent simulations. The so-called FACTS implicit solvent model [68] gave good results for some pairs (e.g., Ala–Ser), but failed completely for others (e.g., Asn–Leu). Other implicit solvent models (e.g., GBMV [107]) gave results in much better agreement with explicit solvent simulations; however, at the price of much higher computational cost.

In Table 4.3 we compare the results obtained with explicit solvent (second column), FACTS (third column), GBMV (fourth column) and NBB using FACTS as the sampled state and GBMV as the target state (rightmost column). The results in columns 2–4 were already presented in Ref. 78; the root mean square deviation from the explicit solvent values (last line in Table 4.3) of the free energy differences obtained with FACTS (2.7 kcal/mol) is significantly larger than that of those obtained with GBMV (0.5 kcal/mol). By contrast, the NBB results obtained by post-processing trajectories generated with FACTS by GBMV are of comparable quality (root mean square deviation of 0.5 kcal/mol) to those directly obtained with GBMV. The computational overhead compared with the regular FACTS/BAR calculations is negligible given the increased accuracy, whereas the gain in efficiency compared to GBMV/BAR is dramatic. Since an energy and force calculation using GBMV is at least ten times slower than using FACTS, and since only every 100$^{\text{th}}$ frame was saved during the MD simulations and, hence, needed to be recalculated, the NBB calculations were 11 times faster than the GBMV/BAR simulations, and only 30% slower than the FACTS/BAR simulations.

4.5 Conclusions

We demonstrated the utility of BAR in combination with simulations of suitably biased states (sampled states). Theoretically, the NBB method is based on the reweighting of biased trajectories as first described by Torrie and Valleau [22]. If free

Table 4.3: Comparison of solvation free energy differences between selected amino acids pairs obtained with explicit solvent, FACTS [68], GBMV [107], and NBB based on simulations with FACTS and post-processing with GBMV. All free energy differences are in kcal/mol

	Explicit[a]	FACTS[a]	GBMV[a]	NBB[b]
Ala-Ser	-2.5	-3.0	-2.8	-2.9
Val-Thr	-2.4	-1.4	-2.1	-2.3
Leu-Asn	-6.1	-1.4	-6.5	-7.1
Phe-Tyr	-4.7	-3.1	-3.9	-4.9
Val-Ala	-1.0	-1.8	-0.5	-0.7
Thr-Ser	-1.3	-3.4	-0.5	-1.4
Phe-Ala	0.0	-2.1	0.2	-0.1
Tyr-Ser	2.5	-2.0	2.1	1.9
RMSD[a]		2.7	0.5	0.5

[a] see Ref. 78

[b] FACTS trajectories were reanalyzed with NBB using GBMV

energy differences using BAR/NBB are calculated in a post-processing step as was done here, no special code is required for NBB. In particular; it should be possible to reproduce all examples presented here with any simulation program for biomolecular systems, e.g., CHARMM [104], AMBER [108], GROMACS [109], NAMD [110] etc. Given the theoretical and practical simplicity, it is astonishing that approaches like NBB are not already widely used. It has only recently come to our attention that a related combination of AMD [86] and BAR has been developed independently from our research [111].

When we started to test NBB, we decided to keep biasing potentials extremely simple. This can be seen particularly in the Asn–Leu example. In an earlier study an involved and computationally expensive adaptive umbrella potential was used [24]. However, as demonstrated by the results of Sect. 4.4.1, the same effect can be achieved by the deletion of two dihedral angle terms. The attempts to compute this alchemical free energy difference without the employment of a better sampled state also demonstrate that BAR is as susceptible to insufficient sampling as TI or other methods to compute free energy differences. In all regular BAR results reported in the Table 4.1 there was sufficient overlap ($\geq$ 10%) between forward and backward perturbations; yet, the resulting free energy differences differed by almost 4 kcal/mol depending on the starting conformation. Thus, as useful as it is (since no comparable gauge exists, e.g., in TI) the overlap criterion (or overlap integral) of BAR should be viewed as a necessary, but not a sufficient criterion for the correctness of a free energy simulation. More specifically, the overlap integral is a tremendous help in choosing the necessary number of λ-states, but it cannot prevent errors from insufficient sampling of phase space because of conformational substates, high energy barriers etc.

In contrast to the situation found in the Asn–Leu example, the five-atomic model systems illustrate situations where a straightforward biasing potential is not possible. The special situation that the simulation of a single system sufficed to compute the free energy difference between the two (different) systems of interest bears some resemblance to techniques explored by van Gunsteren and co-workers to obtain multiple free energy differences from the simulation of a single state [15, 112]. However, in contrast to, e.g., the enveloping distribution sampling (EDS) method, we are primarily interested in the free energy difference between two states and not a family of similar states. In the particular situation found in the model problem, sampling (of both end states) needed to be enhanced to achieve overlap in the first place (i.e., without the biased state the free energy results do not converge). One possible generalization of such situations does indeed lead to EDS and related methods. However, another potentially even more troublesome variant is that despite overlap an important region of phase space of the initial and/or the final state is sampled insufficiently. Consider, e.g., alchemical free energy simulations involving (short) peptides. To ensure that the phase space of the peptide(s) is sampled sufficiently at each λ-state one can remove (or counteract) the dihedral energy terms of the peptide backbone [113]. Similarly, in ligand binding calculations biasing potentials could be used to enhance sampling of side chains near the binding site. Such an approach was recently proposed by McCammon and co-workers [111]. In general, different biasing potentials may be required for the two end states (systems of interest).

The PG/Gly and FACTS/GBMV examples reported in Sects. 4.4.3 and 4.4.4 are conceptually quite similar. Simulations are carried out with the potential energy function of the sampled state; the analysis is carried out with the potential energy function appropriate for the target state; the difference between the two is viewed as the biasing potential. Obviously, before adopting such an approach, one has to be reasonably sure that similar regions of phase space would be sampled with either potential energy functions. The model problem of a "forgotten" energy term is definitely no advocacy for sloppy simulations. The approach may, however, be handy during force field development. If free energy simulations with a force field are part of the parametrization / optimization process, then an approach analogous to what was done in the PG/Gly example can be used to obtain free energy differences corresponding to the use of the latest version of a force field based on already available simulations carried out with some earlier version of the force field. Particularly for polarizable force fields, such an incremental approach may save considerable computer time.

The primary motivation for the last example, use of the (fast) FACTS implicit solvent model to obtain free energy differences of similar quality as if the (much) slower GBMV model had been employed, is speeding up the underlying MD simulations, i.e., the generation of the trajectories. Given the computational cost of free energy simulations, several related applications of this idea come to mind: It has been noted for quite some time now that the short cut-off radii (10Å, 8Å or even shorter) made possible by the particle mesh Ewald summation [74] can lead to errors in the calculation of Lennard-Jones potentials since in many programs the same cut-off radius is used for the truncation of Lennard-Jones interactions and the real space part of Ewald summation [100]. This suggests to generate trajectories with a short(er) cut-off radius (e.g., 8 or 10Å), but evaluating the trajectories with BAR/NBB using a more appropriate cut-off radius, such as 14 or 16Å. Provided that at most every tenth simulation step is saved to disk, this would still result in a significant saving of computer time for large systems. In fact, correcting for too short Lennard-Jones cut-off radii as just outlined was described recently by Shirts et al. using WHAM instead of BAR/NBB [100]. Another potential application are free energy simulations in combination with a polarizable force field. E.g., in the recent overview of the AMOEBA polarizable force field, several results of free energy simulations were reported [114]. During the MD simulations in solution, induced dipoles were converged only to 10^{-2} D, but during calculation of energies for use in BAR, induced dipoles were evaluated with a convergence criterion of 10^{-5} D. In principle, one ought to use NBB in this situation, with the change in potential energy resulting from the difference in convergence criterion as the 'biasing' potential. Given the small size of the correction, the use of plain BAR is permissible, but NBB would be the theoretically correct approach.

In our opinion one of the greatest strengths of BAR is its flexibility. In Ref. 101 we presented several unusual applications of BAR. We used the term "unorthodox" since most calculations of free energy differences would not be feasible with other approaches (e.g., TI). Augmenting BAR by employing special sampled states as done in NBB is another step to enhance the flexibility of the method. Thus, we are confident that clever choices of sampled (biased) states will lead to many more "unorthodox" applications of BAR/NBB.

Chapter 5

Hydration free energies of amino acids: Why side chain analog data are not enough

Using molecular dynamics based free energy simulations, we computed relative solvation free energies for pairs of N-acetyl-methylamide amino acids (Ala–Ser, Val–Thr, Phe–Tyr, Val–Ala, Thr–Ser, Phe–Ala, and Tyr–Ser) and compared the results with the relative solvation free energies of the corresponding pairs of side chain analogs. We observed differences in (relative) solvent affinity $\Delta\Delta\Delta A$ between amino acids and side chain analogs of up to sixty six percent, or, in absolute numbers, 4.9 kcal/mole (Ala–Ser). To rationalize these findings, we estimated separately contributions from what we refer to as solvent exclusion and self-solvation. While the former accounts for the reduction in solute–solvent interactions as one part of the solute occludes other parts of the solute, the latter turned out to be the determining contribution for small polar amino acids and could be shown to arise from interactions between the polar backbone and the polar functional group of the respective side chain in the gas phase. Consequently, the solvent affinity of small polar amino acids depends strongly on the backbone conformation. Our results indicate that the still widely used group additivity – solvent exclusion assumption to estimate solvation free energies for large(r) molecules (such as peptides and proteins) from model compound data (such as side chain analogs) is insufficient. To illustrate practical consequences, we compare the explicit solvent results with those of implicit solvent models. While approaches based on the Generalized Born model give results in (mostly) good agreement with explicit solvent, approaches relying (primarily) on the group additivity – solvent exclusion assumption fail to reproduce $\Delta\Delta\Delta A$. Finally, we briefly discuss the implications of our results for hydrophobicity scales.

5.1 Introduction

Since proteins perform their function in aqueous solution, understanding the contribution of solvent to protein stability, protein association and protein–ligand binding is of great theoretical and practical importance. However, while solvation free energies of small molecules can be measured with high accuracy and precision, the same is not the case for macromolecules, such as proteins. One, therefore, estimates these solvation free energies of interest from data obtained for small molecules. In the case of proteins, one typically uses experimentally determined solvation free energies of model compounds representing the peptide bond (e.g., N-methylacetamide) [27] and the amino acid side chains (side chain analogs, e.g., methanol for Ser etc.) [28]. One then assumes that the solvation free energy is additive (*group additivity* (GA) assumption). [115, 116] While there is no theoretical justification, in the case of proteins the GA assumption is considered adequate. In a recent review Wolfenden states: "There appears to be no reason to suppose that such effects are likely to alter the relative solvation properties of the different amino acid side chains significantly, as compared with the relative solvation properties of the corresponding amino acid residues", with "such effects" referring to cooperativity or anticooperativity between functional groups [116]. The GA assumption underlies fragment based methods [29–31], as well as so-called hydrophobicity scales [32]. In particular, the side chain solvation free energies reported by Wolfenden and co-workers [28] are one of the foundations of the widely used scale by Kyte and Doolittle. [117]

One widely used refinement of the GA assumption results from the observation that amino acids in the interior of proteins will obviously contribute very little to its solvent affinity. We refer to such steric effects as *solvent exclusion* (SE). One frequently used approach [118–120] to account for SE consists in scaling the solvation free energy contribution of a fragment by its solvent accessible surface area (SASA) [121]. We shall refer to GA refined by accounting for SE (e.g., by scaling with the SASA) as group additivity – solvent exclusion (GA-SE) assumption. Among its many applications, it formed and forms the basis of many implicit solvent models [64, 118, 119, 122, 123]. For example, the atomic solvation parameter (ASP) model [119] approximates the solvation free energy of a protein by multiplying the SASA of each atom with a solvation parameter for this atom type (which in turn is derived from the side chain analog data by Wolfenden and co-workers [28]).

There are a number of well-documented limitations of the GA-SE assumption. While for apolar groups (atoms) the proportionality to the SASA (as expressed in cal/Å^2) fluctuates within a relatively narrow range, values for polar groups (atoms) vary much more strongly (as pointed out, e.g., by Karplus [124]). Already some twenty years ago Yunger and Cramer [125], as well as Roseman [126], studied limitations of the GA assumption in connection with hydrophobicity scales of amino acids. Such scales are typically based on the relative solubility of model compounds in different phases (usually water and an apolar phase, such as *n*-octanol or even vacuum) [127]. Roseman [126], as well as Yunger and Cramer [125] compared directly measured partition coefficients between water and *n*-octanol for blocked and zwitterionic amino acids, respectively, with estimates obtained from the GA assumption. For polar and charged amino acids they observed large deviations and suggested intramolecular interactions between the backbone and the polar/charged side chains in the apolar phase as the likely cause, fittingly calling this effect *self-solvation* (SS). Data by White and Wimley for short peptides indicate that transfer free energies measured using N-acetyl-methylamide amino acids are not always representative for longer peptides [128]. In an elegant thought experiment Lazaridis and Karplus demonstrated that the GA-SE approximation breaks down for polar and charged groups [129].

Despite these caveats, applications relying at least to some degree on the GA and GA-SE approximations are ubiquitous. The calculation of Kyte and Doolittle hydropathy plots is a routine procedure on the ExPASy Server [130]. GA-SE based estimates were used to estimate the contribution of the solvation free energy to protein folding [120], and the GA-SE approximation is a central element of some widely used implicit solvent models [64, 123]. It is, therefore, of interest to probe the accuracy of estimates of solvation free energies obtained by the GA or GA-SE approximations. Approaches to estimate the solvation free energy of peptides and proteins based on model compound data, could, in principle, be tested by comparing the solvation free energies of increasingly larger systems (e.g., amino acids, di-, tri-, tetra-peptides etc.) with estimates based on model compounds. However, experimental solvation free energies of peptides are not available, and may well be impossible to obtain, given experimental constraints; cf. the discussion in Ref. 26. This leaves computer simulations as the only possibility. There exist several systematic molecular dynamics based free energy simulation (MDFE) studies of hydration free energies of amino acid side chain analogs, e.g., Refs. 33, 89, 131, 132. However, very few computational data are available for (blocked) amino acids, let alone larger peptides [24, 90, 133–135]. Very recently, Chang et al. [33] published a complete comparison of hydration free energies of uncharged and zwitterionic amino acids and their corresponding side chain analogs and found noticeable differences, particularly for charged and polar amino acids.

In this work we apply MDFE to the simplest possible system with biological relevance for which limitations of model compound based estimates are expected. We compare solvation free energy differences of N-acetyl-methylamide amino acids (blocked amino acids) to the solvation free energy differences of the corresponding side chain analogs, which represent a broad range of distinct physicochemical properties (e.g. polarity and size). The goal of the present study is twofold. First, while from a theoretical point of view it is clear that differences must exist, it is not known how large these deviations are. Second, we want to determine the molecular origin of any deviations from the GA and from the GA-SE assumptions. To investigate the influence of SE and SS on the hydration affinity of amino acids, we complement the MDFE of amino acids and side chain analogs by simulations in which we compute relative solvation free energy differences between unphysical systems, e.g., amino acid with all backbone and/or side chain charges set to zero. These data make it possible to estimate the respective contributions from SE and SS to ΔA_{solv} of blocked amino acids. In addition, we analyze interactions between side chain and backbone of polar amino acids. Using Ser as a representative example of small, polar amino acids, we explore the influence of backbone conformation on solvent affinity.

Finally, we report $\Delta\Delta A_{solv}^{AA}$ for the pair Ala–Ser with restrained backbone conformations using a variety of implicit solvent models (ASP [119], EEF1 [64], SASA [123], GBMV [107], GBSW [136] and FACTS [68]), as well as $\Delta\Delta A_{solv}^{AA}$ of all amino acid pairs studied in this work (without backbone restraints).

5.2 Methods

The CHARMM22 all-atom protein force field was used. [137]. $\Delta\Delta A_{solv}^{AA}$ between Ala and Ser was also calculated with the AMBER Cornell et al. force field [67]. All calculations were carried out using CHARMM [138]; free energy differences in explicit solvent were computed by thermodynamic integration (TI) [7] with the PERT module of CHARMM [138]. A brief description of PERT's functionality (originally written by B. Brooks) can be found in Ref. 139; see also the documentation of CHARMM at www.charmm.org. To overcome slow sampling of side chain rotamers, we used Non-Boltzmann Thermodynamic Integration (NBTI) where necessary [24]. Solvation free energy differences with implicit solvent models were calculated using Bennett's acceptance ratio method [9].

All relative solvation free energy differences in explicit solvent were computed using the standard thermodynamic cycle [51], which entails the calculation of alchemical free energy differences between the two physical systems in the gas phase and in solution. For the blocked amino acids, these were calculated in two ways. Following the usual approach, one side chain was transformed into the other, e.g., for the transmutation of Ala→Ser a methyl group was changed into a CH_2OH group. In addition, we also employed a two-step protocol. First, the side chain of the respective first amino acid was mutated into a single hydrogen. In a second step, this intermediate was transmuted into the respective second amino acid. The intermediate state resembles glycine; however, (in the CHARMM force field) the atom type of the C_α carbon differs from that found in glycine; we, therefore, refer to it as pseudoglycine (PG). In the two-step approach, the $\Delta\Delta A_{solv}^{AA}$ of interest is obtained as the combination of two solvation free energy differences relative to the PG intermediate state; i.e., $\Delta\Delta A_{solv}^{Ala\rightarrow Ser} = \Delta\Delta A_{solv}^{PG\rightarrow Ser} - \Delta\Delta A_{solv}^{PG\rightarrow Ala}$.

To understand the physical origin of differences in solvation free energies of amino acids compared to their side chain analogs, we computed various sets of (relative) solvation free energy differences with unphysical endpoints, such as uncharged backbones or uncharged side chains. In addition, we calculated solvation free energy differences between Ala and Ser with the “backbone” of the blocked amino acids restrained to four conformations: An extended conformation (ϕ: 180.0°, ψ: 180.0°), a helical conformation (ϕ: -57.8°, ψ: -47.0°), a β-sheet conformation (ϕ: -119.0°, ψ: +113.0°) and a left-handed helix (ϕ: 57.8°, ψ: 47.0°). For this purpose, harmonic dihedral restraint terms with a force constant of 100 kcal/(mole radian^2) were applied to the ϕ and ψ torsion angles. These calculations allowed us to study the dependence of self-solvation effects on the conformation of the backbone. To differentiate between contributions from Ala and Ser to the relative solvation free energy difference in the presence of backbone restraints, we also conducted MDFE simulations to calculate the solvation free energy difference associated with the change of restraints for the extended conformation to restrains for the sheet conformation. This was accomplished by turning on/off the two sets of restraint potentials as a function of the coupling parameter λ.

Gas phase free energy differences ΔA_{gas}^{AA} were calculated using Langevin dynamics simulations with a friction coefficient of 5 ps^{-1} on all atoms. Random forces were applied according to the target temperature of 300 K. To justify a time step of 2 ps, hydrogen masses were set to 10 amu. The alchemical mutation was split into 21 intermediate steps, using soft core Lennard Jones and electrostatic interactions. The simulation length at each λ value was 4 ns, the first 60 ps of which were discarded as equilibration. Thus, all gas phase simulations had an overall length of 84 ns and were repeated at least five times in the forward and the backward direction (e.g., in the case of Ala–Ser, directing the transformation five times from Ala to Ser and five times from Ser to Ala), starting from different initial random velocities.

In all solvent simulations 243 TIP3P water molecules [140, 141] were present. The simulation box was a truncated octahedron with constant volume. The side length L of the cube from which the octahedron was generated was $L \approx 24.6$ Å, the exact value depending on the solute. Integration of the equations of motion was carried out with the leapfrog algorithm; the time step was 2 fs. The temperature was maintained at about 300 K by a Nosé-Hoover thermostat [73]. SHAKE [142] was used to keep the water geometry rigid. Lennard-Jones interactions were switched off between 9–10 Å, while electrostatic interactions were computed with the Particle Mesh Ewald method [74]. Coordinates obtained after 400 ps of equilibration (which included 200 ps of constant pressure MD, resulting in the final system size for a given solute) served as the starting configuration for the actual MDFE. In addition, each system was equilibrated for 60 ps at every λ-value. The same 21 λ-values as in the gas phase were used. A total simulation length of 54.6 ns was used for the calculation of the solvation free energy differences between capped amino acids $\Delta A^{AA}_{H_2O}$. For the calculations involving unphysical endpoints shorter protocols (ranging from 2.1 to 25.2 ns of total simulation length) were used. Free energy difference calculations were repeated at least three times starting from different initial random velocities both in the forward and the backward direction. In the calculations employing NBTI [24] an adaptive umbrella potential was applied to the χ_1 angle. This potential was updated every 60 ps, discarding 12 ps of equilibration in each iteration. At each of the 21 λ steps 1 ns was used for the buildup of the biasing potential function, followed by 1 ns of data accumulation in the gas phase and 400 ps of data accumulation in solution.

Solvation free energies with implicit solvent models (ASP [119], EEF1 [64], SASA [123], GBMV [107], GBSW [136] and FACTS [68]) were calculated as described in Refs. 65 and 78, based on simulations of 72 to 158 ns length in the gas phase and with the respective implicit solvent model. The results reported are the average of at least five such simulation pairs. Parameters for the EEF1, SASA, GBMV and GBSW implicit solvent models were selected based on scripts generated by the CHARMM-GUI website [143]. The parameters for FACTS [68] were obtained from the example in the CHARMM documentation at www.charmm.org.

5.3 Results

5.3.1 Side chain analog and amino acid results

We calculated the relative solvation free energy differences $\Delta\Delta A_{solv}$ for selected pairs of N-acetyl-X-methylamide amino acids ($\Delta\Delta A^{AA}_{solv}$ of Ala–Ser, Val–Thr, Phe–Tyr, Val–Ala, Thr–Ser, Phe–Ala, and Tyr–Ser) and for pairs of the corresponding side chain analogs ($\Delta\Delta A^{SC}_{solv}$ of methane(Me)–methanol(MeOH), propane(Pr)–ethanol(EtOH), toluene(Tol)–*p*-cresol(p-Cre), Pr–Me, EtOH–MeOH, Tol–Me and p-Cre–MeOH). The results are summarized in 5.1. $\Delta\Delta A^{AA}_{solv}$ is reported in the first column; the corresponding side chain results $\Delta\Delta A^{SC}_{solv}$ in the second column can be compared to the experimental values reported by Wolfenden and co-workers [28], which are shown in the third column. We also include pertinent results for Leu-Asn from Ref. 24. In the rightmost column of 5.1 we list the deviation $\Delta\Delta\Delta A = \Delta\Delta A^{AA}_{solv} - \Delta\Delta A^{SC}_{solv}$ between side chain and amino acid results. One can see immediately that the differences are large in several cases. The amino acids forming the pairs studied differ (primarily) either in polarity (Ala–Ser, Val–Thr, Leu–Asn, Phe–Tyr) or size (Val–Ala, Phe–Ala, Thr–Ser, Tyr–Ser). With the notable exception of Phe–Tyr, the largest deviations $\Delta\Delta\Delta A$ from the respective side chain results (in absolute numbers) are observed for the solvation free energy differences of apolar relative to polar amino acids, i.e., Ala–Ser, Val–Thr and Leu–Asn. In the most extreme case (Ala–Ser), the solvation free energy difference of the blocked amino acids differs by almost 5 kcal/mole from the solvation free energy difference of the corresponding side chain analogs. For amino acid pairs of like polarity and relatively similar size (Val–Ala, Thr–Ser), the differences of approximately 1 kcal/mole between side chain analog and amino acid results are statistically significant, but much smaller than those obtained for the apolar–polar pairs. As the difference in size between two amino acids of similar polarity increases, so does the deviation from the respective side chain analog results (Phe–Ala, Tyr–Ser). The results of 5.1 clearly suggest that the relative solvation free energy differences of amino acid pairs cannot be estimated from the solvation free energy differences of the respective side chain analogs. In other words, the contribution from the backbone to the solvent affinities of amino acids is not uniform, and, thus, side chain data are not sufficient to estimate the solvation free energy of amino acids.

Especially considering the size of the deviations observed in some cases, it is important to validate the correctness of our calculations. For the selected pairs of side chain analogs the root mean square deviation (RMSD) of the computed solvation free energy differences from the experimental data by Wolfenden and co-workers [28] (third column in 5.1) is 0.58 kcal/mole; similarly, the RMSD with respect to the

Table 5.1: Solvation free energy differences of amino acids and their corresponding side chain analogs in kcal/mole

	$\Delta\Delta A^{AA}_{solv}$ [a]	$\Delta\Delta A^{SC}_{solv}$ [b]	$\Delta\Delta A^{Exp}_{solv}$ [c]	$\Delta\Delta\Delta A$ [d]
Ala-Ser	-2.46	-7.39	-7.00	4.93
Val-Thr	-2.44	-7.09	-6.87	4.65
Leu-Asn[e]	-6.10	-11.02	-11.96	4.92
Phe-Tyr	-4.72	-4.64	-5.35	-0.08
Val-Ala	-0.97	-0.04	-0.05	-0.93
Thr-Ser	-1.29	-0.23	-0.18	-1.06
Phe-Ala	0.01	2.05	2.70	-2.06
Tyr-Ser	2.46	-0.15	1.05	2.61

[a] Standard deviations $\leq$ 0.30 kcal/mole

[b] Standard deviations $\leq$ 0.16 kcal/mole

[c] Experimental side chain analog data from Ref. 28

[d] $\Delta\Delta\Delta A = \Delta\Delta A^{AA}_{solv}$-$\Delta\Delta A^{SC}_{solv}$

[e] Data taken from Ref. 24

solvation free energy energy differences of Shirts et al. [89] (the computationally most elaborate study reported to date) is 0.56 kcal/mole. There are no reference data for blocked amino acids; however, Spichty and Karplus recently obtained very similar results for Val–Thr (−2.74 kcal/mole vs. our −2.44 kcal/mole) [144]. Our results agree qualitatively with the values reported by Chang et al. [33] for zwitterionic amino acids and the OPLS-AA force field [145]. Since we used the CHARMM force field [137], it is unlikely that the differences in solvent affinity of side chain analogs and amino acids are an artifact of a particular force field. To investigate this issue further, we computed the solvation free energy difference for the amino acid pair Ala–Ser with the Cornell et al. AMBER force field [67]. We obtained $\Delta\Delta A^{AA}_{solv}$ = −3.12 kcal/mole, a value which also deviates considerably from the side chain analog result of −7.30 kcal/mole [89].

5.3.2 Estimating contributions from SE and SS to $\Delta\Delta\Delta A$

Our approach to estimate SE and SS contributions relies on the computation of solvation free energy differences between blocked amino acids with some or all partial atomic charges set to zero. This allows us to selectively deactivate interactions of the fragment with its surroundings while keeping its steric properties intact. To illustrate the various types of calculations, we adopt the following pictorial notation. A filled square $\blacksquare$ denotes the backbone (with blocking groups), a filled triangle $\blacktriangle$ and diamond $\blacklozenge$ denote two types of side chains, e.g., Ala and Ser. Thus, we represent a blocked amino acid (e.g., Ala) as $\overset{\blacktriangle}{\blacksquare}$. Used by itself, $\blacksquare$ denotes the pseudo-glycine (PG) intermediate state used in the two-step calculations (cf. Methods). Similarly, $\blacktriangle$ or $\blacklozenge$ by themselves refer to the respective side chain analog. Unfilled symbols indicate that the partial charges of the respective part of the system were set to zero, e.g., $\overset{\blacktriangle}{\square}$ indicates an amino acid without partial charges on the backbone.

Solvent exclusion: To estimate contributions from SE ($\Delta\Delta\Delta A^{SE}_{solv}$) to $\Delta\Delta\Delta A$, we computed relative free energy differences of hydration between hypothetical blocked amino acid pairs with (i) the backbone charges set to zero ($\Delta\Delta A^{unch.\,BB}_{solv}$, $\overset{\blacktriangle}{\square} \rightarrow \overset{\blacklozenge}{\square}$), (ii) the side chain charges set to zero ($\Delta\Delta A^{unch.\,SC}_{solv}$, $\overset{\triangle}{\blacksquare} \rightarrow \overset{\lozenge}{\blacksquare}$), and (iii) all charges set to zero ($\Delta\Delta A^{LJ}_{solv}$, $\overset{\vartriangle}{\square} \rightarrow \overset{\lozenge}{\square}$). The difference between $\Delta\Delta A^{unch.\,BB}_{solv}$ and the corresponding side chain analog solvation free energy difference ($\Delta\Delta A^{SC}_{solv}$, $\blacktriangle \rightarrow \blacklozenge$, see 5.1) provides an estimate of the free energy cost resulting from the incomplete solvation of the side chain (compared to the side chain analog case) because of the presence of the backbone (SE of the side chain by the backbone). The apolar interactions of the uncharged backbone with the side chain and the surrounding water are expected to be small and mostly independent of the side chain; i.e., they are expected to cancel from $\Delta\Delta A^{unch.\,BB}_{solv}$. Similarly, $\Delta\Delta A^{unch.\,SC}_{solv}$ provides estimates of SE effects on the backbone resulting from the presence of the side chain. Thus, $\Delta\Delta A^{unch.\,BB}_{solv}$ and $\Delta\Delta A^{unch.\,SC}_{solv}$ contain the two possible contributions from SE to $\Delta\Delta\Delta A$. However, by adding $\Delta\Delta A^{unch.\,BB}_{solv}$ and $\Delta\Delta A^{unch.\,SC}_{solv}$, one would count twice the free energy contribution resulting from the change in apolar interactions of the two side chains with the backbone and the water (i.e., the Lennard-Jones contribution to $\Delta\Delta A^{AA}_{solv}$, which corresponds to $\Delta\Delta A^{LJ}_{solv}$). This suggests to estimate $\Delta\Delta\Delta A^{SE}_{solv}$ as

$$\Delta\Delta\Delta A^{SE}_{solv} = \begin{bmatrix} \overset{\blacklozenge}{\square} & & \overset{\lozenge}{\blacksquare} & & \overset{\lozenge}{\square} \\ \uparrow & + & \uparrow & - & \uparrow \\ \overset{\blacktriangle}{\square} & & \overset{\vartriangle}{\blacksquare} & & \overset{\vartriangle}{\square} \end{bmatrix} - \begin{matrix} \blacklozenge \\ \uparrow \\ \blacktriangle \end{matrix} =$$

$$= \left[\Delta\Delta A^{unch.\,BB}_{solv} + \Delta\Delta A^{unch.\,SC}_{solv} - \Delta\Delta A^{LJ}_{solv}\right] - \Delta\Delta A^{SC}_{solv}. \tag{5.1}$$

The results for $\Delta\Delta A^{unch.\,BB}_{solv}$, $\Delta\Delta A^{unch.\,SC}_{solv}$ and $\Delta\Delta A^{LJ}_{solv}$ can be found in Table 1 of Supporting Information.

Self-solvation: To understand SS in the case of amino acids, one has to quantify the free energy contribution resulting from intramolecular interactions between side chain and backbone. We obtained an estimate of the SS contribution ($\Delta\Delta\Delta A^{SS}_{solv}$) to $\Delta\Delta\Delta A$ based on the calculations of relative solvation free energies with the two-step protocol, in which we computed $\Delta\Delta A^{AA}_{solv}$ between the amino acid of interest and a glycine-like intermediate state (PG, see Methods); i.e., in our pictorial notation we studied $\blacksquare \rightarrow \overset{\blacktriangle}{\blacksquare}$. In addition, we computed solvation free energy differences between the following hypothetical systems: (i) uncharged PG to a completely uncharged amino acid ($\Delta\Delta A^{PG_{unch.} \rightarrow AA_{unch.}}_{solv}$, $\square \rightarrow \overset{\vartriangle}{\square}$), (ii) uncharged PG to an amino acid with uncharged backbone ($\Delta\Delta A^{PG_{unch.} \rightarrow AA_{unch.\,BB}}_{solv}$, $\square \rightarrow \overset{\blacktriangle}{\square}$), and (iii) PG to an amino acid with uncharged side chain ($\Delta\Delta A^{PG \rightarrow AA_{unch.\,SC}}_{solv}$, $\blacksquare \rightarrow \overset{\vartriangle}{\blacksquare}$). Detailed results for all these free energy differences are provided in Table 2 of Supporting Information.

If free energies were additive, the following equation would hold

$$\begin{array}{ccccccc} \blacktriangle & & \vartriangle & & \blacktriangle & & \vartriangle \\ \blacksquare & & \square & & \square & & \blacksquare \\ \uparrow & + & \uparrow & = & \uparrow & + & \uparrow \\ \blacksquare & & \square & & \square & & \blacksquare \end{array}$$

$$\Delta\Delta A_{solv}^{PG\rightarrow AA} + \Delta\Delta A_{solv}^{PG_{unch.}\rightarrow AA_{unch.}} = \Delta\Delta A_{solv}^{PG_{unch.}\rightarrow AA_{unch.\ BB}} + \Delta\Delta A_{solv}^{PG\rightarrow AA_{unch.\ SC}} \tag{5.2}$$

since, as one sees from the pictorial representation, both sides contain the same molecular fragments, as well as the same number of charged and uncharged atoms. Deviations from Equation 5.2 because of SE are likely to be negligible since by definition SE is a result of the occlusion of one part of the system by the presence of neighboring groups, which are identical in all four steps of Equation 5.2. Therefore, the difference between left and right hand side of Equation 5.2 should yield a measure for the contribution of SS to $\Delta\Delta\Delta A$, i.e.,

$$\begin{aligned} \Delta\Delta\Delta A_{solv}^{SS} &= (\Delta\Delta A_{solv}^{PG\rightarrow AA} + \Delta\Delta A_{solv}^{PG_{unch.}\rightarrow AA_{unch.}}) \\ &- (\Delta\Delta A_{solv}^{PG_{unch.}\rightarrow AA_{unch.\ BB}} + \Delta\Delta A_{solv}^{PG\rightarrow AA_{unch.\ SC}}) \end{aligned} \tag{5.3}$$

Summary: The results of our estimates of $\Delta\Delta\Delta A_{solv}^{SE}$ and $\Delta\Delta\Delta A_{solv}^{SS}$ are summarized in 5.1 (detailed data are provided in Table 3 in Supporting Information). For each pair of amino acids studied, 5.1 shows the total deviation between amino acid and side chain analog solvation free energy differences ($\Delta\Delta\Delta A$, see 5.1) as the white background bar, overlayed by bars representing $\Delta\Delta\Delta A_{solv}^{SE}$ (light gray) and $\Delta\Delta\Delta A_{solv}^{SS}$ (dark gray). The RMSD of the side chain analog data from the amino acid data is 2.33 kcal/mole. If we include our estimates of SE and SS, this RMSD drops to 0.08 kcal/mole. If, on the other hand, only SE is taken into account, the RMSD of the side chain analog data from the amino acid remains at 1.99 kcal/mole. This clearly indicates that SE alone is insufficient to explain $\Delta\Delta\Delta A$.

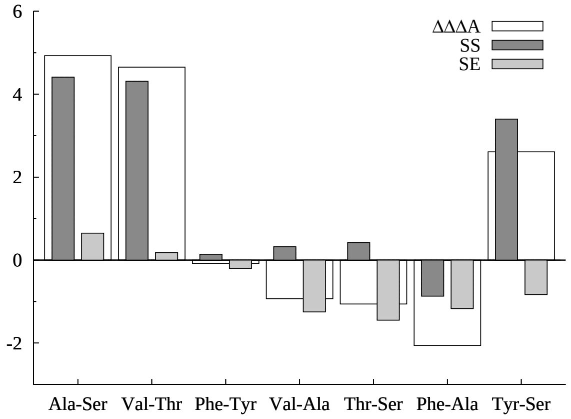


Figure 5.1: Contributions to the solvation free energy differences of amino acids (in kcal/mole). $\Delta\Delta\Delta A$ is the deviation of the $\Delta\Delta A^{AA}_{solv}$ of amino acids relative to the side chain analogs results, while SE and SS refer to the estimates of free energy contributions from solvent exclusion and self-solvation, respectively.

5.3.3 Molecular origin of self-solvation and its conformation-dependence

The analysis summarized in 5.1 clearly demonstrates the quantitative importance of SS. It is of obvious interest to identify the underlying molecular processes. For Ala–Ser, Val–Thr and Tyr-Ser, SS is, in fact, the dominant contribution, accounting for 3.5 — 4.5 kcal/mole of $\Delta\Delta\Delta A$. This value suggests the involvement of a (strong) hydrogen bond [146]. We, therefore, studied side chain – backbone hydrogen bonding patterns in Ser in the gas phase and in solution. Using a rather loose criterion for the presence of a hydrogen bond (donor(D)–acceptor(A) distance < 3.1 Å, $\angle$(D-H-A) $> 100°$), we observed a hydrogen bond between the side chain hydroxyl oxygen and donors in the backbone in 84.3% of the conformations in the gas phase, compared with 32.3% in solution. Also, in water the side chain alternates between two hydrogen bonding partners, whereas in the gas phase interactions between side chain and backbone are mediated by a single (strong) hydrogen bond (see Figure 1 in Supporting Information). Clearly, the presence of water weakens intramolecular backbone – side chain hydrogen bonds.

The above analysis strongly suggests that intramolecular hydrogen bond formation between the side chain functional group and the polar moieties of the backbone is a major contributor to SS in amino acids. If this is the case, then the solvation free energy of amino acids will depend on backbone conformation since it determines the shortest possible distances between backbone and side chain atoms. To explore this conformation dependence, we carried out MDFE for the amino acid pair Ala–Ser during which the solute was restrained to four different backbone conformations, an extended chain, an α-helical, a β-sheet, and a left-handed helix conformation. Since the SE contribution to $\Delta\Delta A^{AA}_{solv}$ for Ala→Ser is quite small (see 5.1), Ala–Ser is an excellent model system to study the conformation dependence of SS.

As can be seen in 5.2, the smallest solvation free energy difference is obtained for the extended conformation. The relative solvation free energy difference between Ala and Ser drops to -1.49 kcal/mole in this case, i.e. just two thirds of the solvation free energy difference obtained for the unrestrained amino acids. A hydrogen bond between side chain and backbone is easily formed in this conformation since the amide hydrogen and the hydroxyl oxygen of the side chain are only ≈ 2 Å

Table 5.2: $\Delta\Delta A^{AA}_{solv}$ for Ala–Ser with restrained backbone conformations in kcal/mole

ϕ/ψ [a]	$\Delta A^{rest.}_{H_2O}$ [b]	$\Delta A^{rest.}_{Gas}$ [c]	$\Delta\Delta A^{rest.}_{solv}$ [d]
extended: 180.0/180.0	3.94	5.43	-1.49
α-helix: -57.8/-47.0	4.44	6.95	-2.51
β-sheet: -119.0/+113.0	3.70	8.73	-5.03
l. h. helix: +57.8/+47.0	3.68	5.79	-2.10

[a] Backbone conformation and target values for restraints on backbone dihedral angles

[b] Free energy change in aqueous solution

[c] Free energy change in the gas phase

[d] Solvation free energy difference

apart. The opposite extreme is found in the β-sheet conformation, where the amide hydrogen points away from the side chain almost at a right angle, resulting in a distance of about 3.6 Å. Clearly, no hydrogen bond formation can take place under these circumstances. Indeed, the relative solvation free energy difference between Ala and Ser in the sheet conformation is -5.03 kcal/mole, a value that approaches the solvation free energy difference between the side chain analogs. The helical conformations, on the other hand, give results which lie in the range of the unrestrained amino acids.

To check the origin of the -3.54 kcal/mole difference in $\Delta\Delta A_{solv}^{AA}$ between Ala and Ser when restrained to the extended and the sheet conformation, we carried out some additional MDFE simulations. Rather than computing the relative solvation free energy difference between Ala and Ser in the presence of backbone restraints (as for the results reported in 5.2), we computed for Ala and Ser separately the relative solvation free energy of the respective amino acid in the extended and in the sheet conformation. These two (relative) free energy differences are -0.66 kcal/mole and -4.18 kcal/mole for Ala and Ser (given for the direction extended chain → sheet conformation). By taking the difference between Ala and Ser, one obtains -3.52 kcal/mole, in excellent agreement with the data of 5.2. Thus, as expected, the major SS contribution is obtained for Ser. These results demonstrate unambiguously and directly (i.e., in terms of the resulting solvation free energy itself) the influence which formation (or non-formation) of intramolecular hydrogen bonds has on self-solvation.

Since we computed relative free energy differences, we have available separately the respective free energy differences in the gas phase (ΔA_{gas}^{rest}) and in solution ($\Delta A_{H_2O}^{rest}$), which are listed in 5.2 as well. While the free energy differences in solution are quite similar (ranging from 3.68 to 4.44 kcal/mole), the gas phase results vary considerably (between 5.43 and 8.73 kcal/mole). Thus, the conformation dependence of the solvation free energy difference, and, hence, the SS contribution originate primarily in the gas phase, and not in solution. Because of the lack of suitable interaction partners in the gas phase, intramolecular stabilization is maximized by forming strong hydrogen bonds between the backbone and the side chain (unless this is prevented by restraining the system to backbone conformations where this is not possible, e.g., a sheet-like conformation). This finding / observation in in disagreement with the conclusions by Chang et al. [33]. While their results (for zwitterionic amino acids) agree qualitatively with the present work, they concluded that the presence of intramolecular backbone – side chain hydrogen bonds in solution led to weaker solvation of the side chains in the amino acids compared to the corresponding side chain analogs; similarly, the backbone was solvated more weakly than glycine. The data reported in 5.2 contradicts this interpretation. Chang et al. [33] compared radial distribution functions and the number of hydrogen bonds between water and the functional groups for both side chain analogs and amino acids. Despite the differences found, this is indirect evidence which is difficult to quantify in terms of a corresponding free energy cost. Since we computed relative solvation free energy differences, we obtained the gas phase and aqueous phase free energy differences directly, and these data suggest unambiguously that most of the SS contribution arises in the gas phase.

5.3.4 Some comments on specific results

As can be seen in 5.1, accounting for SE does suffice in some cases (Val–Ala, Thr–Ser). For apolar side chains (Val–Ala), small SS contributions are expected, but the almost negligible SS result of the polar-polar mutation Thr–Ser is surprising. The apparent lack of SS contributions in this case turns out to result from the cancellation of two large, similar terms. The individual SS estimates for Ser and Thr based on Equation 5.3 are +4.12 kcal/mole and +3.70 kcal/mole, respectively (cf. Table 2 of Supporting Information).

Phe–Tyr is the only apolar–polar pair for which side chain analog and amino acid solvation free energies are virtually identical ($\Delta\Delta\Delta A < 0.1$ kcal/mole). This is a consequence of the size of the side chains. While the oxygen of the side chain hydroxyl groups in, e.g., Ser and Thr is on average < 2.5 Å away from the respective C_α carbon, the mean distance to C_α in Tyr is 6.21 Å. Since no intramolecular hydrogen bonds can be formed, SS contributions are negligible. Further, because of the similar size of the two side chains, SE contributions for the two solutes cancel from $\Delta\Delta A^{AA}_{solv}$.

For the amino acid pairs Ala–Ser, Val–Thr, Tyr–Ser, SS is the dominant contribution. In fact, for the first two pairs the SS contribution from the two small polar amino acid accounts almost completely for the observed $\Delta\Delta\Delta A$. For Tyr–Ser the SS contribution is 3.40 kcal/mole (resulting almost exclusively from Ser (cf. Table 2 in Supporting Information)), but the SE contribution (-1.17 kcal/mole) has the opposite sign and is non-negligible. Finally, while the larger contribution to $\Delta\Delta\Delta A$ for Phe–Ala comes from SE, there is also a SS contribution of ≈ 0.9 kcal/mole. This value shows that even for apolar molecules (groups) intramolecular interactions can lead to deviations from the GA approximation.

5.3.5 Implicit solvent results

Given the large deviations between side chain and blocked amino acid hydration affinities, we decided to test how well a variety of implicit solvent models can reproduce the blocked amino acid results. The implicit solvent models considered, ASP [119], EEF1 [64], SASA [123], GBMV [107], GBSW [136] and FACTS [68], can be loosely grouped into three classes. ASP is a first generation implicit solvent model, purely based on the GA-SE approximation. EEF1 and SASA combine the GA-SE approximation with modified electrostatic interactions (distance dependent dielectric constant, neutralized ionic side chains) to account for dielectric screening. FACTS, GBSW and GBMV, on the other hand, employ the generalized Born model of solvation, supplemented by an additional term for apolar solvation. The following comparison is biased in so far as we compare to a specific water model, TIP3P [140,141], which itself may be subject to error. However, although one of the simplest water models in widespread use, TIP3P was found to give excellent results for side chain analog solvation free energies in a recent comparison of several water models [147].

Our test consisted of two steps. First, we repeated the Ala–Ser calculations with backbone restraints on several backbone conformations (cf. 5.2) using the various implicit solvent models; the respective solvation free energy differences are reported in 5.3. The last two lines of 5.3 give the respective RMSD deviation from the explicit solvent results, as well as the difference between the solvation free energy difference of the amino acids restrained to the β-sheet and the extended chain conformation ($\Delta\Delta\Delta A^{conf}_{ext\rightarrow\beta}$). Extended and β-sheet conformation are particularly

Table 5.3: Comparison of $\Delta\Delta A_{solv}^{AA}$ for Ala–Ser with and without restraints on different backbone conformations using explicit solvent and a variety of implicit solvent models. All free energy differences are in kcal/mole

Conformation	Explicit	ASP	EEF1	SASA	FACTS	GBSW	GBMV
unrestrained	-2.46	-3.00	-7.20	-2.87	-2.99	-2.06	-2.78
extended	-1.49	-3.02	-8.30	-2.92	-2.22	-1.47	-1.82
α-helix	-2.51	-3.47	-6.67	-2.10	-2.92	-1.95	-2.86
β-sheet	-5.03	-3.98	-7.46	-3.82	-5.32	-4.45	-5.11
l. h. helix	-2.10	-4.04	-6.23	-2.06	-2.37	-1.58	-2.24
RMSD[a]	N/A	1.22	4.67	0.88	0.48	0.46	0.27
$\Delta\Delta\Delta A_{ext\rightarrow\beta}^{conf}$[b]	-3.54	-0.96	0.84	-0.90	-3.10	-2.98	-3.29

[a] Root mean square deviation from explicit solvent results

[b] Difference between β-sheet and extended chain result

interesting, since they show the highest variation in terms of solvation free energy differences (-3.54 kcal/mole with explicit solvent, cf. 5.2). The overall worst performance was obtained for EEF1. All other implicit solvent models lead to a $\Delta\Delta A^{AA}_{solv}$ of Ala–Ser without backbone restraints that agrees well with the explicit solvent value of -2.46 kcal/mole (entry "unrestrained" in 5.3). Much larger differences were obtained, however, in the calculations with restrained backbone conformations. Casting aside EEF1, the other two implicit solvent models not based on the generalized Born model, ASP and SASA, are in acceptable agreement as far as RMSD deviation is concerned (1.22 and 0.88 kcal/mole, respectively). The conformation dependence of $\Delta\Delta A^{AA}_{solv}$, as quantified by $\Delta\Delta\Delta A^{conf}_{ext\rightarrow\beta}$, on the other hand, is significantly underestimated (-0.96 and -0.90 kcal/mole for ASP and SASA compared to -3.54 kcal/mole for explicit solvent). By contrast, the generalized Born based models, FACTS, GBSW and GBMV, are all in good agreement with explicit solvent, both in terms of overall RMSD as well as of $\Delta\Delta\Delta A^{conf}_{ext\rightarrow\beta}$. These more complex models clearly give better results than the (primarily) GA-SE based approaches.

In a second step we focused on the three generalized Born based implicit solvent models and used them to calculate the relative solvation free energies for all (unrestrained) amino acid pairs reported in 5.1. The results are summarized in 5.4. Here, noticeable differences between the methods become apparent. While the GBSW and GBMV results are overall in good agreement with explicit solvent simulations (RMSD of 0.51 and 0.41 kcal/mole, respectively), the results obtained with FACTS have a RMSD of 2.67 kcal/mole. The deviations are not uniform and range from acceptable (e.g., Ala–Ser, Val-Ala) to huge ($>$ 5 kcal/mole for Leu–Asn). Also, the large aromatic amino acids (Phe, Tyr) seem to be troublesome, e.g., for the pair Tyr–Ser the FACTS result has the opposite sign compared to explicit solvent calculations. 5.4 suggests that computationally (much) more expensive methods (GBSW, GBMV) lead to significantly better results compared to FACTS (for the capped amino acids studied here the computational cost of FACTS relative to GBSW and GBMV was approximately as 1 : 5 : 15).

Table 5.4: Comparison of $\Delta\Delta A^{AA}_{solv}$ between selected amino acids pairs obtained with explicit solvent and three Generalized Born based implicit solvent models. All free energy differences are in kcal/mole

	Explicit	FACTS	GBSW	GBMV
Ala-Ser	-2.46	-2.99	-2.06	-2.78
Val-Thr	-2.44	-1.41	-2.03	-2.12
Leu-Asn	-6.10	-1.26	-6.99	-6.47
Phe-Tyr	-4.72	-3.10	-4.80	-3.93
Val-Ala	-0.97	-1.80	-0.68	-0.49
Thr-Ser	-1.29	-3.38	-1.34	-0.52
Phe-Ala	0.01	-2.11	-0.08	0.20
Tyr-Ser	2.46	-2.00	1.94	2.07
RMSD[a]		2.67	0.51	0.41

[a] Root mean square deviation from explicit solvent results

5.4 Concluding Discussion

We demonstrated and analyzed the contributions from solvent exclusion and self-solvation to the hydration affinity of blocked amino acids, and showed how these contributions change solvation free energies of amino acids relative to those of their side chain analogs. Our results highlight the importance of self-solvation (being the dominant contribution in polar amino acids). Furthermore, in accord with earlier interpretations by Roseman [126], as well as Yunger and Cramer [125], we showed unambiguously that the self-solvation contribution arises primarily in the gas phase (apolar phase) because of the formation of intramolecular hydrogen bonds between the polar functional group of the side chain and the backbone. This finding entails that the strength of the self-solvation contribution depends on the conformation of the backbone. In the context of larger systems, such as proteins, various intramolecular interactions can lead to self-solvation. A polar side chain in the protein interior could be stabilized by a hydrogen bond (or at least favorable electrostatic interactions) with its own backbone, the backbone or side chain of amino acids in spatial proximity or with other polar compounds nearby (e.g., ligands or co-factors). Furthermore, in a study concerning the partitioning of amino acids and peptides in aqueous two-phase systems Chu and Chen reported marked deviations from the group additivity assumption for peptides that seem to result from intramolecular interactions between hydrophobic residues [148]. Due to the abundance of possible interaction partners, it will be much more difficult to understand self-solvation (and related) effects in larger systems, let alone to quantify them. Furthermore, since self-solvation is a highly local effect, strongly depending on the details of the environment, general rules are difficult to devise.

Our results clearly show that particularly for small to medium sized polar amino acids (Ser, Thr, Asn) side chain data do not suffice to estimate the solvation free energy. Ideally, our results should be verified by experimental solvation free energies. However, solvation free energies can be only measured in the range of +4 kcal/mole to −11 kcal/mole [26]. From our two-step protocols, we know solvation free energies relative to what we refer to as pseudo-glycine. The solvation free energy of this hypothetical reference state with the CHARMM force field is approximately −10.5 kcal/mole (unpublished data); combined with the two-step data our estimate of the (absolute solvation) free energies of, e.g., Ala and Ser is −11.3 and −13.6 kcal/mole. Thus, even blocked Ala appears to be outside the accessible experimental range.

The large deviations between side chain and amino acid solvation free energies, at least for polar amino acids, immediately raise questions concerning the validity and applicability of hydrophobicity scales. Given the multitude of such scales, determined by a variety of methods [127], one cannot give a generic answer. One widely used scale is the so-called hydropathy scale by Kyte and Doolittle [117], which was constructed using Wolfenden's side chain analog solvation free energies as one of the input parameters. Our results suggest that the side chain data are misleading for small to medium sized polar amino acids, whereas they are sufficient for apolar and large, polar amino acids (e.g., Tyr). Kyte and Doolittle stressed and demonstrated the robustness of their scale against variation / choice of raw data (parameters) [117]. Thus, the overall impact of the deviations between side chain and amino acid solvation free energies described here on the Kyte and Doolittle scale may not bee too dramatic, in particular when one is primarily interested in properties such as the GRAVY [117] values of proteins. By contrast, if one were to use the Kyte Doolittle scale (or, for that matter, the actual side chain data by Wolfenden and co-workers) to estimate the solvation contribution of a specific Ser or Thr, then these values would significantly be in error. In addition, use of hydrophobicity scales to estimate solvation contributions of individual amino acids completely neglects the effect of the conformation on the solvent affinity (cf. 5.2).

As reflected by the results reported in 5.3 and 5.4, our model systems also pose challenges for implicit solvent models. Calculating the solvation free energy of, e.g., Ser correctly requires that the implicit solvent model offsets the strong intramolecular interactions between backbone and side chain, i.e., that it weakens the backbone – side chain hydrogen bonds to the same degree as would happen in explicit solvent. As pointed out in a recent review, achieving this delicate balance remains an issue even with the latest generation of continuum electrostatics based implicit solvent models [149]. In our tests, the two methods implementing the Generalized Born model rigorously (GBSW, GBMV) successfully reproduce the explicit solvent results. Since the Generalized Born approach goes beyond the group additivity-solvent exclusion approximation, this is not surprising. Similarly, the failure of models that exclusively or to a large extent rely on the group additivity-solvent exclusion approximation (ASP, EEF1, SASA) was to be expected. The most surprising and mixed result was obtained for a recent model, FACTS, which is based on the Generalized Born model, but attempts to reduce the computational effort by a number of approximations. While FACTS reproduces the conformation dependence of $\Delta\Delta A^{AA}_{solv}$ for the pair Ala-Ser well, it fails for a number of other amino acids. Clearly, capped amino acids (and related solutes) should prove useful as benchmark systems for implicit solvent models.

5.5 Supporting Information

The first three tables present the detailed results of MDFE of (partially) uncharged systems. Table 5.5 summarizes the results used to estimate $\Delta\Delta\Delta A^{SE}_{solv}$ (cf. Equation 5.3.2 of the main manuscript). Similarly, Table 5.6 lists the raw data for individual amino acids relative to the PG intermediate state used to estimate SS contributions (cf. Equation 5.3 of the main manuscript). The resulting values of $\Delta\Delta\Delta A^{SE}_{solv}$ and $\Delta\Delta\Delta A^{SS}_{solv}$ are compiled in Table 5.7, which forms the basis of Figure 5.1 in the main text. In Figure 5.2 histograms of the intramolecular distances of possible hydrogen bonding partners between side chain and backbone are shown for both gas phase and solution.

Table 5.5: **Estimating $\Delta\Delta\Delta A^{SE}_{solv}$:** Relative solvation free energy differences of (partially) uncharged amino acids. All free energies are in kcal/mole. The average standard deviations for $\Delta\Delta A^{unch.\,BB}_{solv}$, $\Delta\Delta A^{unch.\,SC}_{solv}$ and $\Delta\Delta A^{LJ}_{solv}$ are 0.37, 0.41 and 0.33 kcal/mole, respectively. The side chain results $\Delta\Delta A^{SC}_{solv}$ are included for ease of comparison

	$\Delta\Delta A^{SC}_{solv}$ [a]	$\Delta\Delta A^{unch.\,BB}_{solv}$ [b]	$\Delta\Delta A^{unch.\,SC}_{solv}$ [c]	$\Delta\Delta A^{LJ}_{solv}$ [d]	$\Delta\Delta\Delta A^{SE}_{solv}$ [e]
Ala-Ser	−7.39	−6.65	−0.01	0.08	0.65
Val-Thr	−7.09	−6.70	−0.40	−0.19	0.18
Phe-Tyr	−4.64	−4.55	−0.30	−0.01	−0.20
Val-Ala	−0.04	−0.96	−0.89	−0.56	−1.25
Thr-Ser	−0.23	−1.09	−1.36	−0.77	−1.45
Phe-Ala	2.05	0.85	−0.53	−0.56	−1.17
Tyr-Ser	−0.15	−0.92	−0.52	−0.45	−0.83

[a] Side chain analog results from Table 5.1 of the main text

[b] Relative solvation free energy difference between pairs of amino acids with backbone charges set to zero

[c] Relative solvation free energy difference between pairs of amino acids with side chain charges set to zero

[d] Relative solvation free energy difference between pairs of amino acids with all charges set to zero

[e] Estimate of the free energy contribution resulting from SE, cf. Equation 5.3.2 of the main text

Table 5.6: **Estimating ΔA_{solv}^{SS}:** Relative solvation free energy differences of (partially) uncharged amino acids relative to pseudo-glycine (PG). All free energies are in kcal/mole. The average standard deviations for $\Delta\Delta A_{solv}^{PG\rightarrow AA}$, $\Delta\Delta A_{solv}^{PG_{unch.}\rightarrow AA_{unch.\,BB}}$, $\Delta\Delta A_{solv}^{PG\rightarrow AA_{unch.\,SC}}$ and $\Delta\Delta A_{solv}^{PG_{unch.}\rightarrow AA_{unch.}}$ are 0.18, 0.16, 0.23 and 0.25 kcal/mole, respectively.

	$\Delta\Delta A_{solv}^{PG\rightarrow AA}$ [a]	$\Delta\Delta A_{solv}^{PG_{unch.}\rightarrow AA_{unch.\,BB}}$ [b]	$\Delta\Delta A_{solv}^{PG\rightarrow AA_{unch.\,SC}}$ [c]	$\Delta\Delta A_{solv}^{PG_{unch.}\rightarrow AA_{unch.}}$ [d]	$\Delta\Delta\Delta A_{solv}^{SS}$ [e]
Ala	−0.81	0.27	−0.58	0.22	−0.28
Ser	−3.14	−6.31	−0.64	0.31	4.12
Val	0.16	1.23	0.31	0.78	−0.60
Thr	−1.88	−5.22	0.73	1.09	3.70
Phe	−0.83	−0.58	−0.05	0.78	0.58
Tyr	−5.57	−5.40	−0.12	0.77	0.72

[a] Relative solvation free energy difference between PG and the respective amino acid

[b] Relative solvation free energy difference between an uncharged PG and an amino acid with uncharged backbone

[c] Relative solvation free energy difference between PG and the respective amino acid with side chain charges set to zero

[d] Relative solvation free energy difference between an uncharged PG and a completely uncharged amino acid

[e] Estimate of the free energy contribution resulting from self-solvation, cf. Equation 5.3 of the main text

Table 5.7: **SE and SS contributions to the relative solvation free energy differences between pairs of amino acids:** All free energies are in kcal/mole. The results listed in the second ($\Delta\Delta\Delta A_{solv}^{SE}$) and the third ($\Delta\Delta\Delta A_{solv}^{SS}$) column are visualized in Figure 5.1 of the main text.

	$\Delta\Delta A_{solv}^{SC}$ [a]	$\Delta\Delta\Delta A_{solv}^{SE}$ [b]	$\Delta\Delta\Delta A_{solv}^{SS}$ [c]	$\Delta\Delta A_{solv}^{est.}$ [d]	$\Delta\Delta A_{solv}^{AA}$ [e]	Unacct.[f]
Ala-Ser	−7.39	0.65	4.40	−2.34	−2.46	0.12
Val-Thr	−7.09	0.18	4.30	−2.61	−2.44	−0.17
Phe-Tyr	−4.64	−0.20	0.14	−4.70	−4.72	0.02
Val-Ala	−0.04	−1.25	0.32	−0.97	−0.97	0.00
Thr-Ser	−0.23	−1.45	0.42	−1.26	−1.29	0.03
Phe-Ala	2.05	−1.17	−0.86	0.02	0.01	0.03
Tyr-Ser	−0.15	−0.83	3.40	2.42	2.46	−0.04

[a] Side chain analog results from Table 5.1 of the main manuscript

[b] SE contribution according to Table 5.5

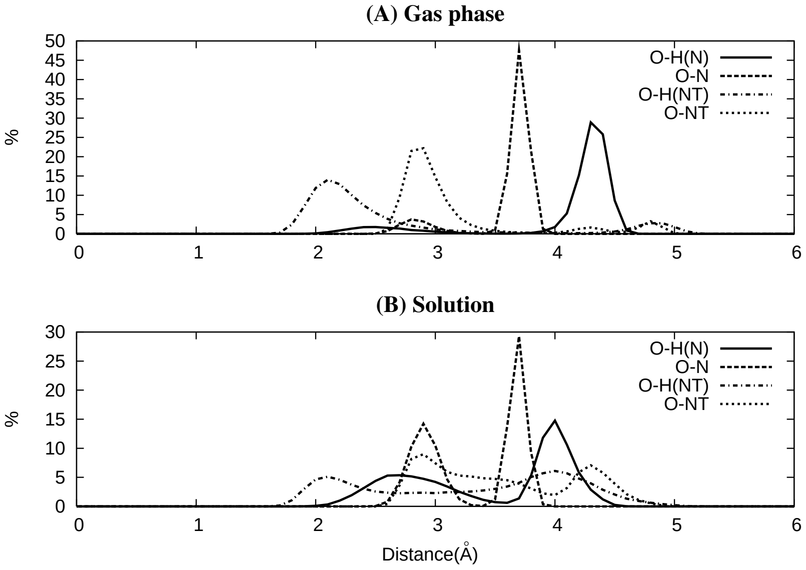
[c] SS contribution according to Table 5.6

[d] Estimate of the relative solvation free energy, $\Delta\Delta A_{solv}^{est.} = \Delta\Delta A_{solv}^{SC} + \Delta\Delta\Delta A_{solv}^{SE} + \Delta\Delta\Delta A_{solv}^{SS}$

[e] Actual solvation free energy difference between the respective pair of amino acids

[f] Deviation $\Delta\Delta A_{solv}^{est.} - \Delta\Delta A_{solv}^{AA}$ between the estimated and the directly calculated relative solvation free energy difference.

Figure 5.2: **Histograms of the intramolecular distances** between the side chain oxygen (O) of Ser and potential hydrogen bonding partners on the backbone (H(N) and H(NT) of the blocking groups, together with their corresponding heavy atoms N and NT) in both gas phase (A) and solution (B). The plots were generated from a 480 ns MD simulation in the gas phase and a 68 ns simulation in solution.



Chapter 6

Absolute hydration free energies of blocked amino acids: Are current estimates of protein solvation overvalued ?

Due to experimental restrictions there are no experimental solvation free energies available for amino acids. Therefore, side chain analog data are often used as model systems in biomolecular studies. This approach basically relies on the assumption that the solvation free energies of the side chain and the backbone are additive. However, in a recent study significant nonadditivities were found for relative solvation free energies of side chain analogs and blocked amino acids, thus casting doubt on this assumption [J. Phys. Chem. B, 113, 8967 (2009)]. To evaluate the additivity of side chain and backbone contributions, we present absolute solvation free energies for blocked N-acetyl-methylamide amino acids which were calculated with molecular dynamics based free energy simulations. By comparing our results for blocked amino acids with solvation free energies for non-zwitterionic amino acids and side chain analogs, we demonstrate that side chain analog data are clearly insufficient for the description of amino acids. We briefly discuss the implications of our results for the field of protein science.

6.1 Introduction

No theory of biomolecular systems can be complete without understanding the role of water in one of its central paradigms, the hydrophobic effect [150]. With respect to proteins, water influences a wide spectrum of processes, including folding [151,152], stability [153] and dynamics [154]. Furthermore, it represents one of the main actors in ligand binding [155] and the selectivity of interactions [156]. From a biophysical point of view, it is, therefore, essential to understand the functional role of the solvent for complex environments such as biomolecules. Thus, it is not surprising that considerable effort has been invested in the study of protein solvation, especially in form of countless hydrophobicity scales [127]. In such scales, the hydrophobicity of a compound is commonly determined by the partitioning between an apolar phase and an aqueous phase, thus providing a ranking of the relative affinities for water. From an experimental point of view, the vapor phase can be regarded as the simplest and most rigorous apolar solvent since there are no interactions with the solute. Formally, the transfer of a solute from an ideal gas phase reference state into aqueous solution is quantified by its solvation free energy.

However, the solvation free energy of proteins or even amino acids cannot be measured experimentally [26]. Therefore, estimates of these solvation free energies were obtained from small molecules by adding contributions of model compounds. E.g., full amino acids were separated into a model compound representing the backbone (e.g., N-methylacetamide) [27] and the amino acid side chains (side chain analogs, e.g., methanol for Ser etc.) [28]. These estimates are based on the hypothesis that the solvation free energy is mostly additive. We note that the additivity assumption is inherently present in any hydrophobicity scale and also forms the basis of fragment based methods [29–31]. In particular, the side chain analog solvation free energies reported by Wolfenden and co-workers [28] are widely used as model systems to understand solvation properties of amino acids and proteins.

Obviously, the approach just outlined is a rather inadequate approximation of the solvation of a protein since amino acids in the interior will likely not make a significant contribution to its solvent affinity. We refer to this steric effect as *solvent exclusion*. Several techniques account for solvent exclusion by scaling the solvation free energy contribution of an atom or fragment by its solvent accessible surface area

[118–121], and also some implicit solvent models rely on this approximation [64,118, 119,122,123], often in conjunction with the side chain analog data by Wolfenden et al. [28]. However, even the consideration of the solvent accessible surface may not be enough to obtain satisfactory results since the side chains of polar and charged amino acids can form intramolecular interactions with the backbone or other polar groups in spatial proximity, thus reducing the effective solvation free energy [124–126,128,129]. This effect is called *self-solvation*.

In a recent publication [78], we computed relative solvation free energies for several pairs of N-acetyl-methylamide amino acids and compared them with the corresponding results for side chain analogs. The observed differences between side chain analog and amino acid solvation free energy differences were up to 66 percent (or, in absolute numbers, 4.9 kcal/mol) for the pair Ala–Ser. The major part of these non-additive effects was traced back to self-solvation. Thus, we concluded that side chain data do not suffice to estimate the solvation free energy in proteins, even accounting for the steric effect of solvent exclusion. However, this conclusion slightly contradicts recent computational experiments conducted by Chang et al. [33], who published a complete comparison of hydration free energies of non-zwitterionic and zwitterionic amino acids, as well as their corresponding side chain analogs. Although significant non-additivity was found for the zwitterionic form of some polar amino acids (Figure 3 in Ref. 33), the data of the non-zwitterionic amino acids (Figure 4 ibid.) correlates well with the side chain analog data. Therefore, Chang et al. concluded that "the hydration free energies of neutral *(i.e. non-zwitterionic)* amino acids can be reasonably approximated by adding the contributions of their side chains to that of the hydration of glycine".

The peculiar finding that zwitterionic amino acids show a high degree of non-additivity, while in non-zwitterionic amino acids side chain and backbone contributions are more or less additive stimulated further analysis on our side. In particular, the question whether solvation free energies can be considered to be additive for practical purposes is of fundamental importance for the evaluation of current simulation methods since the employment of additive approaches the computation of simplifies free energy differences considerably. However, aside from any principal reservations one may have, the usefulness of the additivity assumption is restricted by the error that can be tolerated. We will illustrate this with some thoughts based on Ref. 25. One minimum requirement for any computational method in biomolecular simulation is the ability to discriminate between meaningful and nonsensical protein structures. Therefore, the maximum error should be well below the free energy difference between the native and denaturated state, which is about 10 kcal/mol. Let us assume for arguments sake that an error of 10 kcal/mol is admissible. When considering a full protein consisting of 100 amino acids, random errors grow with the square root of the number of amino acids ($\sqrt{100} = 10$), which leads to an acceptable error of ~ 1 kcal/mol per residue. On the other hand, if the errors are not random but systematic, they do not compensate but simply add up. In such a case non-additivities should not be higher than 0.1 kcal/mol per monomer unit, otherwise our predictions would be useless for protein science. Thus, two questions follow: a.) What is the magnitude of potential errors when applying additivity principles to the computation of solvation free energies of amino acids by relying on side chain analog data ? b.) Are the associated errors systematic or random ?

In this work we present absolute solvation free energies for amino acids with N-acetyl-methylamide blocking groups. In contrast to the pure amino acids used by Chang et al., these blocking groups add peptide bonds to the two ends of the amino acid, thus resolving two problems: a.) While the zwitterionic amino acids of Chang et al. are representative for the situation found in solution, one is rather unlikely to encounter the zwitterionic form in the gas phase. The opposite is true for neutral amino acids, which reflect the most likely state in the gas phase, but are not the most favorable form for solution. Thus, the two kinds of simulations conducted separately by Chang et al. do not correspond to a real transfer process (with zwitterions in solution and the non-zwitterionic form in gas phase). However, since the occurrence of (de-)protonation processes can be ruled out in blocked amino acids, this issue is completely avoided in our study b.) Another benefit of the blocking groups is that the solutes start to resemble peptides, with the core of a peptide backbone present. Thus, the simulation results can account for possible interactions between the side chain and its backbone.

The remainder of this paper is organized as follows: First, we outline the methods (Sect. 6.2) employed in this study. We then present the results for the absolute solvation free energies of blocked amino acids (Sect. 6.3) and compare them with the results provided by Chang et al. for non-zwitterionic amino acids, as well as with the solvation free energies of the corresponding side chain analogs. In addition, we report some computational aspects which can influence the absolute solvation free energy, using the example of Gly. We conclude with a short discussion of our findings and their possible biological relevance in Section 6.4. A short comparison of absolute solvation free energies derived from several implicit solvent models with our explicit solvent results can be found in the Appendix.

6.2 Methods

We calculated the solvation free energies of all canonical neutral amino acids (Ala, Val, Leu, Ile, Ser, Thr, Cys, Met, Asn, Gln, His, Phe, Tyr, Trp) with N-acetyl-methylamide blocking groups attached. We did not include simulations of charged amino acids (Arg, Asp, Lys, Glu) since they require complex corrections for the finite-range treatment of electrostatic interactions [89]. Besides, we also omitted the simulation of the imino acid proline since there is no corresponding side chain analog. However, the two tautomeric states of neutral histidine were considered (referred to as Hid and Hie, where the proton is attached to the δ and ϵ nitrogen, respectively). To save computational costs, we first calculated relative solvation free energy differences of all amino acids to a pseudo Gly intermediate state (PG). This intermediate state resembles Gly, except for the atom type of the C_α carbon. In a second step, we calculated the absolute solvation free energy of PG. Thus, the absolute solvation free energy of each amino acid is the sum of one absolute and one relative solvation free energy difference (i.e., $\Delta A_{aa}^{solv} = \Delta A_{PG}^{solv} + \Delta\Delta A_{PG \to aa}^{solv}$).

Each relative solvation free energy was calculated with the standard thermodynamic cycle, which includes four kinds of calculations: 1.) Turning off the charges of the side chain in explicit solvent ($\Delta A_{aa \to unch.aa}^{H_2O}$) and 2.) mutating the uncharged side chains to PG ($\Delta A_{unch.aa \to PG}^{H_2O}$). Since CHARMM does not offer the separation of the nonbonded energies into solute-solute and solute-solvent interactions, also the corresponding gas phase corrections had to be computed ($\Delta A_{aa \to unch.aa}^{gas}$ and

$\Delta A^{gas}_{unch.aa \rightarrow PG}$). Thus, the relative solvation free energy can be calculated by:

$$\Delta\Delta A^{solv}_{PG \rightarrow aa} = \Delta A^{H_2O}_{aa \rightarrow unch.aa} + \Delta A^{H_2O}_{unch.aa \rightarrow PG} - \Delta A^{gas}_{aa \rightarrow unch.aa} - \Delta A^{gas}_{unch.aa \rightarrow PG}$$

For the absolute solvation free energy of PG, three simulations were conducted: a.) turning off all charges of PG in solution ($\Delta A^{H_2O}_{PG,unch}$) b.) turning off the van-der-Waals interactions between PG and water ($\Delta A^{H_2O}_{PG,vdw}$) c.) the gas phase correction for turning off the intramolecular interactions of PG (ΔA^{gas}_{PG}). Thus, the absolute solvation free energy of PG is given by:

$$\Delta A^{solv}_{PG} = \Delta A^{H_2O}_{PG,unch} + \Delta A^{H_2O}_{PG,vdw} - \Delta A^{gas}_{PG}$$

All free energy calculations were conducted with CHARMM [62, 104], using the CHARMM27 force field that includes the backbone cross term map (CMAP) correction [66, 106]. Most free energy differences were computed with the Bennett's Acceptance ratio method [9]. Only the gas phase corrections, $\Delta A^{gas}_{unch.aa \rightarrow PG}$ and ΔA^{gas}_{PGunch} were computed by thermodynamic integration (TI) [7] with the PERT module of CHARMM. In Table 6.1, we list the respective number of λ-points (second column) and simulation times (the third column shows the simulation time in nanoseconds per λ-point, and the fourth column the total simulation length of the respective free energy simulation) for each type of calculation.

Gas phase free energy differences were calculated using Langevin dynamics simulations with a friction coefficient of 5 ps^{-1} on all atoms. Random forces were applied according to the target temperature of 300 K, and hydrogen masses were set to 10 amu to justify a time step of 2 fs. For the BAR analysis, trajectories were written every 100 steps.

In all solvent simulations 862 TIP3P water molecules [140, 141] were present. The simulation box was a truncated octahedron. The side length L of the cube from which the octahedron was generated was $L = 37.25$ Å, which was the average boxsize over all selected amino acids (the optimal boxsize of each amino acid was determined from a 1 ns constant pressure simulation). For the determination of $\Delta A^{H_2O}_{PG,vdw}$ we used constant pressure simulations. Integration of the equations of motion was carried out with the velocity-Verlet algorithm as implemented in the TPCNTRL module of CHARMM [157]; the time step was 2 fs. The temperature was maintained at about 300 K using two separate Nosé-Hoover thermostats [73]

Table 6.1: Overview of the simulation protocols

Type of mutation	**# λ^a**	**ns/λ^b**	**Total time (ns)**
$\Delta A^{H_2O}_{aa\to unch.aa}$	3	10	30
$\Delta A^{H_2O}_{unch.aa\to PG}$	5-7	10	50-70
$\Delta A^{gas}_{aa\to unch.aa}$	3	84	252
$\Delta A^{gas}_{unch.aa\to PG}$	21	4	84
Total ($\Delta\Delta A^{solv}_{PG\to aa}$)	**32-34**		**416-436**
$\Delta A^{H_2O}_{PG,unch}$	3	10	30
$\Delta A^{H_2O}_{PG,vdw}$	9	10	90
ΔA^{gas}_{PG}	21	4	84
Total (ΔA^{solv}_{PG})	**33**		**204**

[a] Number of λ intermediate states

[b] Simulation time per λ-point

for solute and solvent. SHAKE [142] was used to keep the water geometry rigid. Lennard-Jones interactions were switched off between 10–12 Å, while electrostatic interactions were computed with the Particle Mesh Ewald method [74]. Coordinates obtained after 1 ns of equilibration served as the starting configuration for the free energy simulation. In addition, each system was equilibrated for 100 ps at every λ-value.

To overcome slow sampling of side chain rotamers when computing $\Delta A^{H_2O}_{aa \to unch.aa}$ and $\Delta A^{gas}_{aa \to unch\ aa}$, we lowered the energy barriers of χ_1 and χ_2 by deleting the corresponding dihedral potentials. To obtain correct free energies the data was reweighted with Non-Boltzmann Bennett (NBB) [158] according to the value of the dihedral potential.

Simulation lengths of all free energy protocols are given in Table 6.1. The standard deviations reported were determined by repeating each free energy simulation four times, starting with different initial random velocities. The energies of the respective states required for BAR and NBB were extracted from the trajectories using the EAVG command of the BLOCK module of CHARMM; the BAR/NBB analysis was carried out by a `Perl` program.

6.3 Results and Discussion

Absolute solvation free energies of the 15 blocked amino acids using the CHARMM27 [66, 106] force field are given in Table 6.2. Since the solvation free energies of side chain analogs and other small compounds often serve as a gauge for the accuracy of force fields [18, 89, 131, 132, 159], extensive data is available on the error margins of such free energy simulations. For example, in a blind test for 17 small molecules, free energy simulations yielded root mean square errors between 1.3 and 1.7 kcal/mol [18]. Shirts et al. [89] obtained a root mean square deviation of 1.3 kcal/mol for all amino acid side chain analogs, using the CHARMM force field (notably, also AMBER and OPLS-AA were employed in their study, yielding overall very similar results). Given the acute lack of experimental data for verification, we assume that our error margins will be comparable to the deviations found in these studies. The standard deviations of the absolute solvation free energies presented in this study range between 0.1 and 0.6 kcal/mol, which is comparable to the values reported by Chang et al. for calculations with full amino acids (their standard deviations lie between 0.2 and 1.0 kcal/mol).

In a recent study [78], we employed a less sophisticated protocol (using CHARMM22 with shorter simulation lengths, fewer water molecules and a shorter cut-off compared to the simulations conducted here) for the calculation of relative solvation free energy differences of blocked amino acids and their corresponding side chain analogs. In this earlier work, the results for the side chain analog solvation free energy differences were in good agreement with the experimental values, obtaining a root mean square deviation of 0.6 kcal/mol (which is a good check of our protocols). The absolute solvation free energies reported here agree well with the relative solvation free energy differences for blocked amino acids reported in Ref. 78 (see Table 6.3). In total, the root mean square deviation from the old results is just 0.3 kcal/mol, which can probably be traced back to the differences of the simulation protocols.

In Figure6.1, we compare the solvation free energy results for N-acetyl-methylamide amino acids (the dashed line represents a regression line of the data; individual data points are marked by crosses and the corresponding one-letter amino acid code) from our study with the results for non-zwitterionic amino acids as reported by Chang et al. [33](dotted regression line), as well as the side chain analog data, as calculated by Shirts et al. [89] with the CHARMM force field (continuous regression line). The computational results (ordinate) are plotted against the experimental data by Wolfenden et al. [28] (abscissa). Since the aim of this study is the assessment of the additivity hypothesis for solvation free energies, we show all results relative to the respective Gly reference state (i.e. a blocked Gly for the blocked amino acids, a pure Gly for the neutral amino acids and H_2 in the case of side chain analogs).

If solvation free energies were truly additive, all three lines should be identical and, in an ideal setting, form a perfect diagonal (i.e, for a regression line $f(x) = ax + b$, we should find a slope $a = 1.0$ and an axis intercept $b = 0.0$). However, as can be seen in Figure6.1, both the slopes and the axis intercepts of the three regression lines deviate from these ideal values. The steepest slope was found for the side chain analogs ($a = 0.95$), but the slope of the blocked amino acids is only

Table 6.2: Absolute solvation free energies of amino acids in kcal/mol

Mutation	$\Delta\Delta A_{PG\rightarrow aa}^{solv}$ [a]	ΔA_{aa}^{solv} [b]	σ [c]
Ala	2.9	-12.0	0.1
Asn	-2.6	-17.5	0.3
Cys	1.8	-13.1	0.2
Gln	-3.0	-17.9	0.1
Gly	0.4	-14.5	0.1
Hid	-6.2	-21.1	0.2
Hie	-3.3	-18.2	0.2
Ile	4.0	-10.9	0.5
Leu	3.6	-11.3	0.6
Met	2.4	-12.5	0.2
Phe	2.6	-12.3	0.4
Ser	0.1	-14.8	0.3
Thr	1.3	-13.6	0.5
Trp	-0.3	-15.2	0.2
Tyr	-2.0	-16.9	0.1
Val	3.3	-11.6	0.2

[a] Relative solvation free energies to Pseudo-Glycine (PG, $\Delta A_{PG}^{solv} = -14.9$ kcal/mol)

[b] Absolute solvation free energies derived from relative solvation free energies to PG

[c] Standard deviations

Table 6.3: Comparison of relative solvation free energy differences of blocked amino acids in kcal/mol with previous works

	$\Delta\Delta A_{abs}^{solv\,a}$	$\Delta\Delta A_{rel}^{solv\,b}$	Difference[c]
Ala-Ser	-2.8	-2.5	-0.3
Val-Thr	-2.0	-2.4	0.4
Leu-Asn	-6.2	-6.1	-0.1
Phe-Tyr	-4.6	-4.7	0.1
Val-Ala	-0.5	-1.0	0.5
Thr-Ser	-1.2	-1.3	0.1
Phe-Ala	0.3	0.0	0.3
Tyr-Ser	2.1	2.5	0.4
RMSD[d]			**0.3**

[a] Difference between absolute solvation free energies reported in Table 6.2

[b] Relative solvation free energy differences reported in Ref. 78

[c] Difference $\Delta\Delta A_{solv}^{abs} - \Delta\Delta A_{solv}^{rel}$

[d] Root mean square deviation of $\Delta\Delta A_{solv}^{abs}$ from $\Delta\Delta A_{solv}^{rel}$

$a = 0.54$, and the slope of the non-zwitterionic amino acids is in the middle between the two ($a = 0.78$). Since the abscissa denotes experimental solvation free energies for the side chain analogs, while the ordinate shows computational results, the small deviation (0.05) of the slope of the side chain analog data from the ideal slope can be attributed to imperfections of the force field. Such deviations can be expected for all regression lines, so this value gives us an idea of the acceptable incongruities between the three slopes; i.e. if the slopes of the three regression lines differ significantly more than by 0.05, the differences are unlikely to be caused by errors of the force field. However, the slopes of the non-zwitterionic and blocked amino acids deviate by 0.17 and 0.41 from the side chain analog data.

The slopes in Figure6.1 can be seen as a measure for the differing magnitude of group contributions of functional groups (e.g., an hydroxyl group) to the total solvation free energy. If the slope is very steep, adding functional groups to the molecule will change the solvation free energy drastically. On the other hand, if the slope were completely flat, the solvent affinity of the molecule would not be affected by the addition of functional groups. The different slopes in Figure6.1 clearly demonstrate that the solvation free energy difference associated with the addition of a functional group depends on the “scaffold“ it is attached to. For example, adding a hydroxyl group to methane (the side chain analog of alanine) changes the solvation free energy by -7.0 kcal/mol. If a hydroxyl group is attached to a non-zwitterionic alanine, its contribution is -3.1 kcal/mol. Finally, for the blocked alanine, the associated change of the solvation free energy by adding a hydroxyl is only -2.8 kcal/mol. This reduction of the relative solvation free energy differences corresponds exactly to what one would expect if interactions with the backbone weaken the solvent affinity of the side chain (and vice versa). Thus, the results presented here are in perfect agreement with the predictions made in our previous publication [78]. In addition, this comparison demonstrates, that even in the data for the non-zwitterionic amino acids by Chang et al. one can find considerable non-additivities.

Another interesting aspect in Figure6.1 are the different axis intercepts of the regression lines. The blocked amino acid results appear to be shifted to more positive solvation free energies. This effect can be explained by a change of the relative ranking of Gly in the list of solvation free energies. Since Gly (or its side chain analog

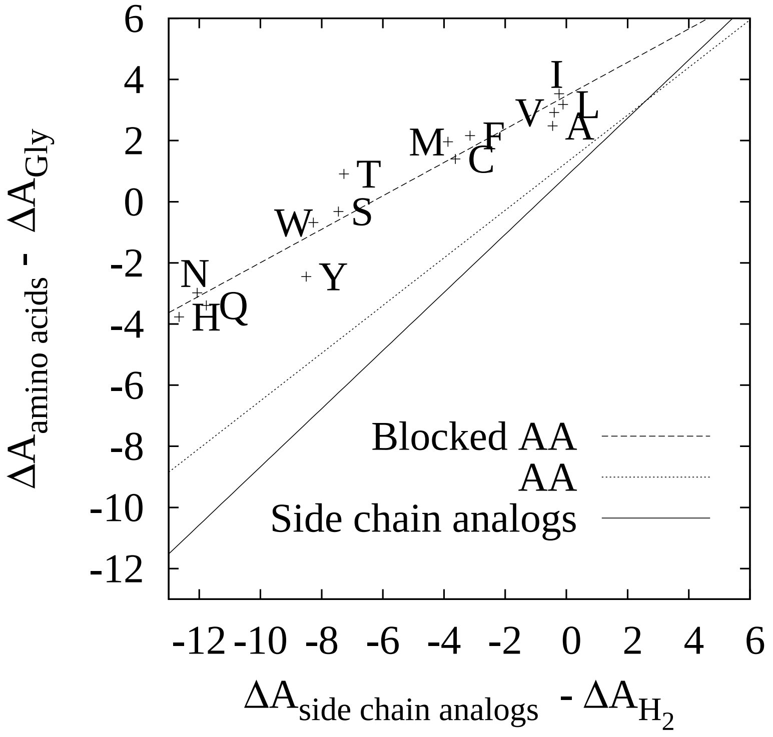


Figure 6.1: Comparison of computational results relative to Gly or its corresponding side chain analog (H_2). The abscissa denotes the experimental solvation free energies reported by Wolfenden et al. The dashed line represents a regression of the results for amino acids with blocking groups calculated in this paper. The corresponding individual results are indicated by crosses and the one-letter code for the amino acid. To avoid cluttering the figure, only linear regressions of the amino acids and side chain analog data are shown. The dotted line shows a linear regression of the solvation free energies of non-zwitterionic forms of amino acids as calculated by Chang et al. The continuous line represents a linear regression of the results for side chain analogs by Shirts et al.

H_2) was the reference state for all data points in the plot, this has a tremendous effect on the origin of the regression line. Though the relative rankings of most amino acids are changed only by one rank when side chain analogs, neutral amino acids and blocked amino acids are compared, the position of Gly differs dramatically: According to our blocked amino acid results, Gly is more hydrophilic than 8 out of 15 amino acids. In case of the non-zwitterionic amino acids data, four compounds (Ala, Leu, Ile and Val) are more hydrophobic than Gly. According to the side chain analog data, however, the analog of Gly (H_2) is most hydrophobic. Although the relative insolubility of such a volatile gas in water stands to reason, inferring the same for Gly is rather counterintuitive. We, therefore, are rather surprised to find that this notion (in conjunction with the very convenient assumption that solvation free energies are additive) has remained unchecked in the literature for almost 30 years. Our data (as well as the data provided by Chang et al.) clearly shows that H_2 (as well as all other side chain analogs) are adequate model systems for full amino acids.

Finally, we want to discuss the conformation dependency of solvation free energies. Most of the additive methods today simply ignore the conformation of the molecule when calculating solvation free energies (this is particularly the case for techniques based on atomic contributions). However, several studies indicate that the solvation free energy depends on the conformation of the molecule [24,33,65,91, 158]. In a previous work [78] we could demonstrate that the solvent affinity of amino acids is directly influenced by the conformation of the backbone since the solvation free energy depends on the distance between the functional groups of backbone and side chain. To illustrate the implications of this effect, we calculated the absolute solvation free energy of Gly without the backbone cross term map (CMAP) correction [106]. CMAP corrects the potentials of the ϕ and ψ backbone dihedrals in order to reproduce quantum mechanical potential energy surfaces. Since it only affects the dihedral potentials, but not the charges and Lennard-Jones parameters, one might expect its impact on the solvation free energy to be marginal. However, the removal of the CMAP potential actually reduces the solvent affinity of Gly by 1.9 kcal/mol (which is almost the same as the relative solvation free energy difference between Thr and Val). The finding that absolute solvation free energies do depend on the backbone conformation illustrates why molecular solvation is far too complex for a simple additive approach.

6.4 Conclusions

To evaluate whether solvation free energies are additive, we calculated absolute solvation free energies for 15 blocked amino acids. The results range between -11.3 and -21.5 kcal/mol, which is outside of the experimental detection range (+4 kcal/mol to -11 kcal/mol [26]). Thus, solvation free energies of these systems are currently only quantifiable by theoretical means. The presented results agree well with previously published relative solvation free energy differences [78] and, methodologically, free energy calculations in connection with present force fields have been demonstrated to yield root mean square errors < 2 kcal/mol [18, 89], depending on the simulation setup.

Using just the least complex amino acid, Gly, as an example, the supposed additivity of solvation free energies is easily refuted. We will illustrate this for a naive fragment based approach, which calculates the solvation free energy by adding the solvation free energies of small molecules that correspond to fragments of the molecule of interest. This is the most direct application of the additivity hypothesis, but it is rarely used anymore in this crude form without any correction terms. We readily admit that our comparison is not very sportive; however, it is motivated by Wolfenden, who recently asserted that "the few cases" of non-additivity can be explained in terms of electronic effects, which are unlikely "to alter the relative solvation properties of the different amino acid side chains significantly, as compared with the relative solvation properties of the corresponding amino acid residues" [116].

So just for the sake of the argument, let us assume that N-acetyl-methylamide Gly ($CH_3-CO-NH-CH_2-CO-NH-CH_3$) can be divided into two groups: a.) an acetamide group ($CH_3-CO-NH_2$, representing the N terminal blocking group) and b.) an N-methylacetamide group ($CH_3 - CO - NH - CH_3$ representing the C terminal rest). According to Wolfenden et al. [27, 28], the solvation free energies of acetamide and N-methylacetamide are -9.7 and -10.1 kcal/mol, respectively. The corresponding sum of the solvation free energies of the two fragments of Gly would be −19.8 kcal/mol. Compared to our Gly result of −14.5 kcal/mol, this would overestimate the solvation free energy by 5.3 kcal/mol (or $\sim$ 35%). (Theoretically, we still would have to subtract the effect of the two excessive hydrogen atoms in our calculation, which, in an atomistic fragment based approach, can be merged to H_2. However, since the hydration free energy of H_2 is +2.4 kcal/mol, the result of the additive fragment based approach would become −22.2 kcal/mol, which is even worse)

Since there is no side chain in Gly, this fragment based result does not even include any complications from the presence of the side chain, thus posing a best case scenario. Therefore, we will also consider a more complex system. As reported both by Chang et al. [33] and in our previous study [78], the non-additivity is strongest in polar amino acids. We will exemplify this with the most extreme case encountered in our work, Asn. The side chain analog of Asn is acetamide. Thus, by adding the solvation free energy of yet another acetamide (−9.7) to the fragment based solvation free energy of Gly we obtain an estimated solvation free energy for Asn of −29.5 kcal/mol. This result overestimates the Asn result obtained in our study (−17.5 kcal/mol) by 12 kcal/mol (or $\sim$ 70%), which is a distinctively higher error than in the Gly case. Generally, the root mean square error of the fragment based method over all amino acids would be 9.3 kcal/mol (data not shown). Although there are reported cases where the addition of a hyrophobic group can actually increase the solvent affinity [160], our data shows that the estimates of solvation free energies of amino acids based on side chain analog data were *always* overestimated if using a naive fragment based approach; thus, the errors are clearly systematic.

We note that this problem is (partly) known in the fragment-based community, where empirical “correction factors” are often employed to rectify the results for polyfunctional groups. However, there are several indications that even corrected fragment based approaches are inadequate. Fragment based methods are often employed in medicinal chemistry to determine the lipophilicity of a compound in form of its partition coefficient between n-1-octanol and water, the so-called log P. Since it is a common descriptor in the development of quantitative structure-activity relationships for drug candidates, there has been some interest in assessing the accuracy of methods to determine the log P. In a study of eight fragment based prediction programs, 340 peptides (varying in length from two to sixteen amino acids) were used to evaluate their accuracy [161]. While the correlation coefficients R^2 of most programs ranged between 0.1 and 0.5 (correlations < 0.5 are generally considered as weak), only one neural-network-based program achieved a relatively good R^2 of 0.8. This weak correlation of fragment based approaches with real log P's further supports our case that the additivity assumption is not valid for free energies.

Another example for the breakdown of the simple additivity hypothesis is the dependency of the solvation free energy on the conformation of the backbone. We illustrated this by calculating the solvation free energy of Gly once with the CMAP correction on the backbone and once without CMAP. Though all other parameters were equal and CMAP only affects the dihedral potentials of the backbone, the solvation free energies obtained in the free energy simulations differed by almost 2 kcal/mol. In our previous paper, we also described this conformation dependency for the solvation free energy difference between serine and alanine [78], tracing its origin back to self-solvation (i.e., interactions between the backbone and the sidechain, that weaken the interactions with water and stabilize the molecule in gas phase). If free energies were additive, the contribution of the backbone ought to cancel in such relative calculations. However, the results deviated by 3.5 kcal/mol, depending on the backbone conformation.

In macromolecular systems, such as proteins, several intra- and intermolecular interactions can lead to self-solvation. Especially polar amino acids can be stabilized by their own backbone, neighboring amino acids or other polar groups nearby, thus lowering their effective solvent affinity. However, methods based on the additivity hypothesis, such as fragment based methods, hydrophobicity scales (or, as shown in recent studies [78,162], even methods based solely on the solvent accessible surface) are not able to account for self-solvation effects in polar amino acids. Our results highlight that just the interactions with the backbone alone can reduce the solvation free energy per amino acid considerably and this effect can be amplified by backbone conformations that facilitate self-solvation (see results without CMAP). For a single amino acid, it may be tempting to accept such error margins, given the computational simplicity of the additivity hypothesis. However, since the resulting errors for each amino acid are systematic, they would simply add up in case of a hypothetical protein. Given that 10 kcal/mol are approximately the difference between the native and denaturated state in proteins, errors as outlined for Gly and Asn are clearly unacceptable [25].

Even a casual scan of the literature reveals that side chain analog results are still widely considered to be representative for full amino acids (a notion, which is closely intermingled with the additivity hypothesis) [116]. In particular, they are used to study the solvation of amino acids or trans-membrane helices in membranes. One widely used hydrophobicity scale in this context is the so-called hydropathy scale by Kyte and Doolittle [117], which was constructed using Wolfenden's side chain analog solvation free energies as one of the input parameters. However, even contemporary works concerning membrane proteins often rely on side chain analogs as model systems for full amino acids [163]. As we have shown, side chain analog data considerably overestimate the solvent affinity of most amino acids and even the relative ranking of some amino acids (such as Gly) is still a matter of debate. Thus, some modifications of prediction algorithms for transmembrane helices may be required.

Another field relying heavily on hydration free energies are the energetics of protein folding and stability, were intimate relationships with water play an significant role. Most studies concerning the contributions of solvation to folding were conducted in the early nineties and basically relied on group additivity or area based models to determine the effect of solvation [164]. Not surprisingly, they found strong correlations of the energy of unfolding with the surface areas of polar and nonpolar amino acids. However, since the number of experimental observables in proteins is miniscule in comparison to the number of constituents and interactions, it is relatively easy to find an interpretation that is consistent with the thermodynamic data. Therefore, critique of the additivity hypothesis followed swiftly, however relatively unheeded. E.g., in 1995, Lazaridis and Karplus demonstrated by theoretical means that the accessible surface area approximation breaks down for polar and charged groups [129]. In 1997, Robertson and Murphy [165] named the "non-additivity of energetic contributions from the various groups that make up polar and nonpolar surfaces" as the number one culprit for the deviation of 57% to 182% between calculated and experimental results for the ΔC_p of unfolding. However, concerning the supposed non-additivity they lamented that ”no straightforward approach is available yet for evaluating its role”. The results of our simulations now clearly make up for this lack of data, at least as far as hydration free energies are concerned.

Though our solvation free energy results themselves may be useful in other contexts, we want to stress that we are not advocating to use our results as yet another hydrophobicity scale or data for an improved fragment based approach. Except when dealing with casual qualitative comparisons (e.g., in bioinformatics), employing an additive approach would not be advisable (especially in biophysics). To borrow an analogy from Wittgenstein [166], our results concerning hydrophobicity scales should be regarded to be like a ladder that must be thrown away after one has climbed it. They are mainly to be used in order to recognize the relative uselessness of hydrophobicity scales in a macromolecular world with a broad range of possible interactions. Instead, given the availability of modern computer resources, the use of free energy simulations with explicit or, alternatively, Generalized Born based implicit solvent models [167, 168] (see the pertaining discussion in the Appendix) should be considered when dealing with solvation effects in biomolecules. Since preliminary results can be obtained even with normal desktop computers; and free software packages for biomolecular simulations are readily available, there is no excuse anymore for using hydrophobicity scales or side chain analog data for quantitative studies.

6.5 Some comments concerning implicit solvent models

In a previous study [78], we compared the relative solvent affinities of several amino acids both for explicit solvent as well as for the implicit solvent models ASP [119], EEF1 [64], SASA [123], GBMV [107], GBSW [136] and FACTS [68]. Our results demonstrated that several implicit solvent models are not able to account for self-solvation, with the notable exception of methods implementing the Generalized Born model (GBMV, GBSW, and, to some extend, also FACTS). Here, we compare explicit solvent results for the absolute solvation free energies (see Table 6.2) with results obtained from implicit solvent simulations in our recent study [78] (The data was not shown in Ref. 78, since we were only interested in relative solvation free energy differences at that time). The results of this comparison are shown in Table 6.4.

Notably, while the relative solvation free energy differences were quite similar for the three implicit solvent models [78], the absolute solvation free energies differ considerably. FACTS consistently underestimates the solvent affinity of the blocked amino acids and the GBSW results show a tendency to be too hydrophilic. The GBMV results, on the other hand, agree exceedingly well with our explicit solvent results (with a root mean square deviation of just 0.3 kcal/mol). Since GBMV is the most rigorous (and computationally expensive) method tested here, this is not very suprising. Thus, we conclude that implicit solvent models can indeed be a valuable tool for determing absolute solvation free energies.

Table 6.4: Comparison of absolute solvation free energies of blocked amino acids in kcal/mol with implicit solvent results from previous studies

	Explicit	FACTS	GBMV	GBSW
Ala	-12.0	-7.7	-11.9	-13.4
Asn	-17.5	-5.6	-18.2	-19.2
Leu	-11.3	-4.4	-11.2	-12.7
Phe	-12.3	-5.6	-11.9	-13.6
Ser	-14.8	-10.7	-14.7	-15.5
Thr	-13.6	-7.3	-13.4	-14.9
Tyr	-16.9	-8.7	-16.7	-17.5
Val	-11.6	-5.9	-11.3	-12.9
RMSD[a]		**7.2**	**0.3**	**1.3**

[a] Root mean square deviation from explicit solvent results

Final discussion

In this dissertation, we evaluated different methodological aspects of free energy simulations:

- We have applied Bennett's Acceptance Ratio method (BAR) to problems where standard methods to compute free energy differences such as TI or EF are not feasible. This was first demonstrated for the calculation of the free energy difference between ethane and methanol in aqueous solution, using the physical endpoints only. We then showed how BAR can be used to compute the free energy difference resulting from changing the cut-off, from switching the force field, or from using an implicit solvent model. Calculations of this kind should prove useful for force field development and the validation of implicit solvent methods.

- We demonstrated how Non-Boltzmann Sampling can be employed in connection with Bennett's Acceptance Ratio method. We refer to this technique as Non-Boltzmann Bennett. In particular, we illustrated how specially designed sampling states can be employed in connection with NBB to improve sampling, correct small errors in free energy simulations, or speed up the computation. All of these aspects can improve the efficiency of free energy simulations.

- We used free energy simulations to compute relative solvation free energies for pairs of N-acetyl-methylamide amino acids (Ala–Ser, Val–Thr, Phe–Tyr, Val–Ala, Thr–Ser, Phe–Ala, and Tyr–Ser) and compared the results with the relative solvation free energies of the corresponding pairs of side chain analogs. Our results showed that there are distinct discrepancies between the solvation free energy differences of blocked amino acids and side chain analogs. To rationalize these findings, we estimated separately contributions from what we refer to as solvent exclusion and self-solvation. While the former accounts for the reduction in solute–solvent interactions as one part of the solute occludes other parts of the solute, the latter turned out to be the determining contribution for small polar amino acids and could be shown to arise from interactions between the polar backbone and the polar functional group of the respective side chains. Our results indicate that the still widely used group additivity–solvent exclusion assumption to estimate solvation free energies for molecules such as peptides and proteins from model compound data is insufficient.

- To evaluate the additivity of side chain and backbone contributions, we calculated absolute solvation free energies for blocked N-acetyl-methylamide amino acids. By comparing our free energy results for blocked amino acids with solvation free energies for non-zwitterionic amino acids and side chain analogs, we demonstrate that methods which employ the additivity hypothesis clearly overestimate the solvation free energy of amino acids. We briefly discuss the implications of our results for the field of protein science, with a particular focus on the energetics of protein folding and stability.

Bibliography

[1] K. Popper, Logik der Forschung (The Logic of Scientific Discovery), Mohr Siebeck, Tübingen, 1994.

[2] A. Stephan, Emergentism, Irreducibility, and Downward Causation, Grazer Philosophische Studien 65 (2002) 77–93.

[3] Aristoteles, W. Nestle, Aristoteles Hauptwerke (Aristotle, Main works), Kröner, Stuttgart, 1977.

[4] J. Cohen, I. Stewart, Collapse of Chaos: Discovering Simplicity in a Complex World, Penguin Science, London, 2000.

[5] R. Feynman, R. Leighton, M. Sands, The Feynman Lectures on Physics, Addison-Wesley Longman, Amsterdam, 1970.

[6] A. M. Van der Sloot, C. Kiel, L. Serrano, F. Stricher, Protein design in biological networks: from manipulating the input to modifying the output, Protein. Eng. Des. Sel. 22 (9) (2009) 537–542.

[7] J. G. Kirkwood, Statistical mechanics of fluid mixtures, J. Chem. Phys. 3 (1935) 300–313.

[8] R. W. Zwanzig, High-temperature equation of state by a perturbation method. I. Nonpolar gases, J. Chem. Phys. 22 (1954) 1420.

[9] C. H. Bennett, Efficient estimation of free energy differences from Monte Carlo data, J. Comp. Phys. 22 (1976) 245–268.

[10] C. Jarzynski, Nonequilibrium equality for free energy differences, Phys. Rev. Lett. 78 (1997) 2690–2693.

[11] G. E. Crooks, Path-ensemble averages in systems driven far from equilibrium, Phys. Rev. E 61 (2000) 2361–2366.

[12] H. van de Waterbeemd, E. Gifford, ADMET in silico modelling: Towards prediction paradise ?, Nature Rev. Drug. Disc. 2 (2003) 192.

[13] J. Sikkema, J. A. de Bont, B. Poolman, Mechanisms of membrane toxicity of hydrocarbons, Microbiol. Rev. 59 (1995) 201.

[14] A. Curley, L. Fishbein, K. Gheorghiev, F. Korte, B. Paccagnella, M. Roberfroid, Y. Shirasu, E. M. B. Smith, D. C. Villeneuve, Environmental Health Criteria 9: DDT and its derivatives, World Health Organization, Geneva, Switzerland (1979).

[15] C. Oostenbrink, W. van Gunsteren, Free energies of ligand binding for structurally diverse compounds, Proc. Natl. Acad. Sci. U.S.A. 102 (19) (2005) 6750–6754.

[16] D. L. Mobley, A. P. Graves, J. D. Chodera, A. C. McReynolds, B. K. Shoichet, K. A. Dill, Predicting absolute ligand binding free energies to a simple model site, J. Mol. Biol. 371 (4) (2007) 1118–1134.

[17] J. Kästner, H. Senn, S. Thiel, N. Otte, W. Thiel, QM/MM free-energy perturbation compared to thermodynamic integration and umbrella sampling: Application to an enzymatic reaction, J. Chem. Theory Comput. 2 (2) (2006) 452–461.

[18] A. Nicholls, D. L. Mobley, J. P. Guthrie, J. D. Chodera, C. I. Bayly, M. D. Cooper, V. S. Pande, Predicting small-molecule solvation free energies: An informal blind test for computational chemistry, J. Med. Chem. 51 (4) (2008) 769–779.

[19] D. L. Mobley, C. I. Bayly, M. D. Cooper, M. R. Shirts, K. A. Dill, Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations, J. Chem. Theory Comp. 5 (2) (2009) 350–358.

[20] D. Seeliger, B. de Groot, Protein Thermostability Calculations Using Alchemical Free Energy Simulations, Biophys. J. 98 (10) (2010) 2309–2316.

[21] M. R. Shirts, V. S. Pande, Comparison of efficiency and bias of free energies computed by exponential averaging, the Bennett acceptance ratio, and thermodynamic integration, J. Chem. Phys. 122 (2005) 144107–1–144107–16.

[22] G. M. Torrie, J. P. Valleau, Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling, J. Comp. Phys. 23 (1977) 187.

[23] C. Bartels, M. Karplus, Probability distributions for complex systems: Adaptive umbrella sampling of the potential energy, J. Phys. Chem. B 102 (1998) 865–880.

[24] M. Leitgeb, C. Schröder, S. Boresch, Alchemical free energy calculations and multiple conformational substates, J. Chem. Phys. 122 (2005) 084109.

[25] K. Dill, Additivity principles in biochemistry, J. Biol. Chem. 272 (2) (1997) 701–704.

[26] A. Nicholls, D. L. Mobley, J. P. Guthrie, J. D. Chodera, C. I. Bayly, M. D. Cooper, V. S. Pande, Predicting small-molecule solvation free energies: An informal blind test for computational chemistry, J. Med. Chem. 51 (2008) 769–779.

[27] R. Wolfenden, Interactions of the peptide bond with solvent water: A vapor phase analysis, Biochemistry 17 (1978) 201–204.

[28] R. Wolfenden, L. Andersson, P. M. Cullis, C. C. B. Southgate, Affinities of amino-acid side-chains for solvent water, Biochemistry 20 (1981) 849–855.

[29] G. G. Nys, R. F. Rekker, Statistical analysis of a series of partition-coefficients with special reference to predictability of folding of drug molecules — introduction of hydrophobic fragmental constants (F-values), Chim. Ther. 8 (1973) 521–535.

[30] G. G. Nys, R. F. Rekker, Concept of hydrophobic fragmental constants (F-values). 2. extension of its applicability to calculation of lipophilicities of aromatic and heteroaromatic structures, Chim. Ther. 9 (1974) 361–375.

[31] C. Hansch, A. J. Leo, Substituent Constants for Correlation Analysis in Chemistry and Biology, Wiley Interscience, New York, 1979.

[32] Y. Nozaki, C. Tanford, The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions, J. Biol. Chem. 246 (1971) 2211–2217.

[33] J. Chang, A. M. Lenhoff, S. I. Sandler, Solvation free energy of amino acids and side-chain analogues, J. Phys. Chem. B 111 (2007) 2098–2106.

[34] A. R. Leach, Molecular Modelling - Principles and Applications, 2nd Edition, Pearson Education, 2001.

[35] C. Chipot, A. Pohorille, Free Energy Calculations: Theory and Applications in Chemistry and Biology, Springer, Berlin, 2007.

[36] D. Frenkel, B. Smit, Understanding Molecular Simulation. From Algorithms to Applications, Academic Press, London, New York, Sydney, 1996.

[37] C. D. Christ, A. E. Mark, W. F. van Gunsteren, Feature Article Basic Ingredients of Free Energy Calculations: A Review, J. Comp. Chem. 31 (8) (2010) 1569–1582.

[38] R. Hentschke, Statistische Mechanik, Wiley-VCH, 2004.

[39] H. Hellmann, Einführung in die Quantenchemie, Deuticke, 1937.

[40] R. P. Feynman, Forces in molecules, Phys. Rev. 56 (1939) 340.

[41] R. E. Stanton, Hellmann-feynman theorem and correlation energies, J. Chem. Phys. 5 (36) 1298.

[42] M. Born, K. Huang, Dynamical Theory of Crystal Lattices., Clarendon Press, 1954.

[43] B. J. Alder, T. E. Wainwright, Phase transition for a hard sphere system, J. Chem. Phys. 27 (1957) 1208.

[44] A. Rahman, Correlations in motion of atoms in liquid argon, A. Phys. Rev. 136 (1964) 405.

[45] A. Rahman, F. H. Stillinger, Improved simulation of liquid water by molecular dynamics, J. Chem. Phys. 60 (1974) 1545.

[46] J. A. Mccammon, B. R. Gelin, M. Karplus, Dynamics of folded proteins, Nature 267 (1977) 585.

[47] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, Equation of state calculations by fast computing machines, J. Chem. Phys. 6 (1953) 1087.

[48] M. Zacharias, T. P. Straatsma, J. A. Mccammon, Separation-shifted scaling, a new scaling method for lennard-jones interactions in thermodynamic integration, J. Chem. Phys. 100 (1994) 9025.

[49] S. Bruckner, S. Boresch, Efficiency of alchemical free energy simulations I: Practical comparison of the exponential formula, thermodynamic integration and bennett's acceptance ratio method, submitted to J. Comput. Chem. (2010).

[50] S. Bruckner, S. Boresch, Efficiency of alchemical free energy simulations II: Improvements for thermodynamic integration, submitted to J. Comput. Chem. (2010).

[51] B. L. Tembe, J. A. McCammon, Ligand-receptor interactions, Comput. Chem. 8 (1984) 281–283.

[52] H.-J. Woo, B. Roux, Chemical theory and computation special feature: Calculation of absolute protein–ligand binding free energy from computer simulations, Proc. Natl. Acad. Sci. USA 102 (2005) 6825–6830.

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

[54] T. Simonson, G. Archontis, M. Karplus, Free energy simulations come of age: Protein-ligand recognition, Accts. Chem. Res. 35 (2002) 430–437.

[55] T. Rodinger, R. Pomès, Enhancing the accuracy, the efficiency and the scope of free energy simulations, Curr. Opin. Struct. Biol. 15 (2005) 164–170.

[56] K. Raha, M. Kenneth, J. Merz, Calculating binding free energy in protein-ligand interaction, Annu. Rep. Comp. Chem. 1 (2005) 113–130.

[57] M. R. Shirts, D. L. Mobley, J. D. Chodera, Alchemical free energy calculations: Ready for prime time?, Annu. Rep. Comput. Chem. 3 (2007) 41–59.

[58] D. M. Zuckerman, T. B. Woolf, Theory of a systematic computational error in free energy differences, Phys. Rev. Lett. 89 (2002) 180602.

[59] N. Lu, J. K. Singh, D. A. Kofke, Appropriate methods to combine forward and reverse free-energy perturbation averages, J. Chem. Phys. 118 (2003) 2977–2984.

[60] N. Lu, D. Wu, T. B. Woolf, D. A. Kofke, Using overlap and funnel sampling to obtain accurate free energies from nonequilibrium work measurements, Phys. Rev. E 69 (2004) 05772.

[61] M. R. Shirts, E. Bair, G. Hooker, V. S. Pande, Equilibrium free energies from nonequilibrium measurements using maximum-likelihood methods, Phys. Rev. Lett. 91 (2003) 140601.

[62] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, M. Karplus, CHARMM: A program for macromolecular energy, minimization and dynamics calculations, J. Comput. Chem. 4 (1983) 187–217.

[63] B. Roux, T. Simonson, Implicit solvent models, Biophys. Chem. 78 (1999) 1–20.

[64] T. Lazaridis, M. Karplus, Effective energy function for proteins in solution, Proteins: Struct., Funct., Gen. 35 (1999) 133–152.

[65] D. L. Mobley, K. A. Dill, J. D. Chodera, Treating entropy and conformational changes in implicit solvent simulations of small molecules, J. Phys. Chem. B 112 (2008) 938–946.

[66] A. D. MacKerell, Jr., D. Bashford, M. Bellott, R. L. Dunbrack, Jr., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, III, B. Roux, M. Schlenkrich, J. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, M. Karplus, All-atom empirical potential for molecular modeling and dynamics studies of protein, J. Phys. Chem. B 102 (1998) 3586–3616.

[67] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, P. A. Kollman, A second generation force field for the simulation of proteins and nucleic acids, J. Am. Chem. Soc. 117 (1995) 5179–5197.

[68] U. Haberthür, A. Caflisch, FACTS: Fast analytical continuum treatment of solvation, J. Comput. Chem. 29 (2008) 701–715.

[69] E. Neria, S. Fischer, M. Karplus, Simulation of activation free energies in molecular systems, J. Chem. Phys. 105 (1996) 1902–1921.

[70] S. Boresch, M. Karplus, The role of bonded terms in free energy simulations. 2. calculation of their influence on free energy differences of solvation., J. Phys. Chem. A 103 (1999) 119–136.

[71] S. Boresch, M. Leitgeb, A. Beselman, A. D. MacKerell, Unexpected relative aqueous solubilities of a phosphotyrosine analogue and two phosphonate derivatives, submitted for publication (2004).

[72] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, Comparison of simple potential functions for simulating liquid water, J. Chem. Phys. 79 (2) (1983) 926–935.

[73] W. G. Hoover, Canonical dynamics: Equlibrium phase-space distributions, Phys. Rev. A 31 (1985) 1695–1697.

[74] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, L. G. Pedersen, A smooth particle mesh Ewald method, J. Chem. Phys. 103 (1995) 8577–8593.

[75] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, J. M. Rosenberg, The weighted histogram analysis method for free-energy calculations on biomolecules. I. the method, J. Comput. Chem. 13 (1992) 1011–1021.

[76] M. Souaille, B. Roux, Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations, Comp. Phys. Comm. 135 (2001) 40–57.

[77] M. Leitgeb, C. Schröder, S. Boresch, Alchemical free energy calculations and multiple conformational substates, J. Chem. Phys. 122 (2005) 084109.

[78] G. König, S. Boresch, Hydration Free Energies of Amino Acids: Why Side Chain Analog Data Are Not Enough, J. Phys. Chem. B 113 (26) (2009) 8967–8974.

[79] L. Nilsson, Efficient table lookup without inverse square roots for calculation of pair wise atomic interactions in classical simulations, J. Comput. Chem. 00 (2009) 000–000.

[80] D. M. Ferguson, On the use of acceptance ratio methods in free energy calculations, J. Chem. Phys. 99 (1993) 100860–10087.

[81] G. Hummer, L. R. Pratt, A. E. García, Hydration free energy of water, J. Phys. Chem. 99 (1995) 14188–14194.

[82] G. Hummer, J. C. Rasaiah, J. P. Noworyta, Water conduction through the hydrophobic channel of a carbon nanotube, Nature 414 (2001) 188–190.

[83] M. R. Shirts, J. D. Chodera, Statistically optimal analysis of samples from multiple equilibrium states, J. Chem. Phys. 129 (12).

[84] R. H. Swendsen, J. S. Wang, Replica monte-carlo simulation of spin glasses, Phys. Rev. Lett. 57 (1986) 2607.

[85] Y. Sugita, Y. Okamoto, Replica-exchange molecular dynamics method for protein folding, Chem. Phys. Lett. 314 (1999) 141.

[86] D. Hamelberg, J. Mongan, J. McCammon, Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules, J. Chem. Phys. 120 (24) (2004) 11919–11929.

[87] H. Grubmüller, Predicting slow structural transitions in macromolecular systems: Conformational flooding, Phys. Rev. E 52 (1995) 2893–2906.

[88] N. Madras, M. Piccioni, Importance sampling for families of distributions, Ann. Appl. Probab. 9 (4) (1999) 1202–1225.

[89] M. R. Shirts, J. W. Pitera, W. C. Swope, V. S. Pande, Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins, J. Chem. Phys. 119 (2003) 5740–5761.

[90] W. Gu, S. J. Rahi, V. Helms, Solvation free energies and transfer free energies for amino acids from hydrophobic solution to water solution from a very simple residue model, J. Phys. Chem. B 108 (2004) 5806.

[91] P. V. Klimovich, D. L. Mobley, Predicting hydration free energies using all-atom molecular dynamics simulations and multiple starting conformations, J. Comput.-Aided Mol. Des. 24 (4, Sp. Iss. SI) (2010) 307–316.

[92] Y. Deng, B. Roux, Calculation of standard binding free energies: Aromatic molecules in the T4 lysozyme L99A mutant, J. Chem. Theory Comp. 2 (5) (2006) 1255–1273.

[93] D. L. Mobley, J. D. Chodera, K. A. Dill, Confine-and-release method: Obtaining correct binding free energies in the presence of protein conformational change, J. Chem. Theory Comp. 3 (4) (2007) 1231–1235.

[94] C. A. F. De Oliveira, D. Hamelberg, J. A. McCammon, Coupling accelerated molecular dynamics methods with thermodynamic integration simulations, J. Chem. Theory Comp. 4 (9) (2008) 1516–1525.

[95] M. Fajer, D. Hamelberg, J. A. McCammon, Replica-Exchange Accelerated Molecular Dynamics (REXAMD) Applied to Thermodynamic Integration, J. Chem. Theory Comp. 4 (10) (2008) 1565–1569.

[96] T. P. Straatsma, J. A. McCammon, Treatment of rotational isomers in free energy evaluations. analysis of the evaluation of free energy differences by molecular dynamics simulations of systems with rotational isomeric states, J. Chem. Phys. 90 (1989) 3300–3304.

[97] A. Hodel, L. M. Rice, T. Simonson, R. O. Fox, A. T. Brünger, Proline *cis-trans* isomerization in staphylococcal nuclease: Multi-substate free energy perturbation calculations, Protein Sci. 4 (1995) 636–654.

[98] T. Straatsma, J. McCammon, Treatment of rotational isomeric states. III. The use of biasing potentials, J. Chem. Phys. 101 (6) (1994) 5032–5039.

[99] J. D. Chodera, W. C. Swope, J. W. Pitera, C. Seok, K. A. Dill, Use of the weighted histogram analysis method for the analysis of simulated and parallel tempering simulations, J. Chem. Theory Comp. 3 (2007) 26–41.

[100] M. R. Shirts, D. L. Mobley, J. D. Chodera, V. S. Pande, Accurate and efficient corrections for missing dispersion interactions in molecular simulations, J. Phys. Chem. B 111 (2007) 13052–13063.

[101] G. König, S. Bruckner, S. Boresch, Unorthodox Uses of Bennett's Acceptance Ratio Method, J. Comp. Chem. 30 (11) (2009) 1712–1718.

[102] D. Wu, D. Kofke, Phase-space overlap measures. I. Fail-safe bias detection in free energies calculated by molecular simulation, J. Chem. Phys. 123 (5) (2005) 054103.

[103] D. Wu, D. Kofke, Phase-space overlap measures. II. Design and implementation of staging methods for free-energy calculations, J. Chem. Phys. 123 (5) (2005) 084109.

[104] B. Brooks, C. Brooks III, A. Mackerell Jr., L. Nilsson, R. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. Pastor, C. Post, J. Pu, M. Schaefer, B. Tidor, R. Venable, H. Woodcock, X. Wu, W. Yang, D. York,

M. Karplus, CHARMM: The Biomolecular Simulation Program, J. Comp. Chem. 30 (10, Sp. Iss. SI) (2009) 1545–1614.

[105] D. A. Pearlman, A comparison of alternative approaches to free energy calculations, J. Phys. Chem. 98 (1994) 1487–1493.

[106] A. MacKerell, M. Feig, C. Brooks, Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations, J. Comp. Chem. 25 (11) (2004) 1400–1415.

[107] M. S. Lee, M. Feig, F. R. Salsbury, C. L. Brooks, III, New analytic approximation to the standard molecular volume definition and its application to generalized born calculations, J. Comput. Chem. 23 (2003) 1348–1356.

[108] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, R. J. Woods, The AMBER biomolecular simulation programs, J. Comp. Chem. 26 (2005) 1668–1688.

[109] B. Hess, C. Kutzner, D. van der Spoel, E. Lindahl, GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation, J. Chem. Theory Comp. 4 (3) (2008) 435–447.

[110] J. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. Skeel, L. Kale, K. Schulten, Scalable molecular dynamics with NAMD, J. Comp. Chem. 26 (16) (2005) 1781–1802.

[111] J. Wereszczynski, private communication.

[112] C. Christ, W. van Gunsteren, Enveloping distribution sampling: A method to calculate free energy differences from a single simulation, J. Chem. Phys. 126 (18).

[113] C. A. F. De Oliveira, D. Hamelberg, J. A. McCammon, Estimating kinetic rates from accelerated molecular dynamics simulations: Alanine dipeptide in explicit solvent as a case study, J. Chem. Phys. 127 (17).

[114] J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio, Jr., , M. Head-Gordon, G. N. I. Clark, M. E. Johnson, T. Head-Gordon, Current status of the AMOEBA polarizable force field, J. Phys. Chem. B 114 (2010) 2549–2564.

[115] J. A. V. Butler, Trans. Faraday. Soc. 33 (1937) 229–236.

[116] R. Wolfenden, Experimental measures of amino acid hydrophobicity and the topology of transmembrane and globular proteins, J. Gen. Physiol. 129 (5) (2007) 357–362.

[117] J. Kyte, R. F. Doolittle, A simple method for displaying the hydropathic character of a protein, J. Mol. Biol. 157 (1982) 105–132.

[118] T. Ooi, M. Oobatake, G. Nemethy, H. A. Scheraga, Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides, Proc. Natl. Acad. Sci. U.S.A. 84 (1987) 3086.

[119] L. Wesson, D. Eisenberg, Atomic solvation parameters applied to molecular dynamics of proteins in solution, Protein Science 1 (1992) 227.

[120] G. I. Makhatadze, P. L. Privalov, Energetics of protein structure, Adv. Prot. Chem. 47 (1995) 307–425.

[121] B. Lee, F. M. Richards, Interpretation of protein structures - estimation of static accessibility, J. Mol. Biol. 55 (1971) 379.

[122] D. Eisenberg, A. D. McLachlan, Solvation energy in protein folding and binding, Nature 319 (1986) 199.

[123] P. Ferrara, J. Apostolakis, A. Caflisch, Evaluation of a fast implicit solvent model for molecular dynamics simulations, Proteins 46 (2002) 24–33.

[124] P. A. Karplus, Hydrophobicity regained, Protein Sci. 6 (1997) 1302.

[125] L. M. Yunger, R. D. Cramer III, Measurement and correlation of partition-coefficients of polar amino acids, Mol. Pharmacol. 20 (1981) 602–608.

[126] M. A. Roseman, Hydrophilicity of polar amino-acid side-chains is markedly reduced by flanking peptide bonds, J. Mol. Biol. 200 (1988) 513–522.

[127] K. M. Biswas, D. R. DeVido, J. G. Dorsey, Evaluation of methods for measuring amino acid hydrophobicities and interactions, Chromatogr. A 1000 (2003) 637.

[128] W. C. White, T. P. Creamer, S. H. White, Solvation energies of amino acid side chains and backbone in a family of host–guest pentapeptides, Biochemistry 35 (1996) 5109–5124.

[129] T. Lazaridis, G. Archontis, M. Karplus, Enthalpic contribution to protein stability: Insights from atom-based calculations and statistical mechanics, Adv. Prot. Chem. 47 (1995) 231–306.

[130] E. Gasteiger, C. Hoogland, A. G. A. S. Duvaud, M. R. Wilkins, R. D. Appel, A. Bairoch, Protein identification and analysis tools on the expasy server, in: The Proteomics Protocols Handbook, Humana Press, 2005, pp. 571–607.

[131] A. Villa, A. E. Mark, Calculation of the free energy of solvation for neutral analogs of amino acid side chains, J. Comput. Chem. 23 (2002) 548–553.

[132] Y. Q. Deng, B. Roux, Hydration of amino acid side chains: Nonpolar and electrostatic contributions calculated from staged molecular dynamics free energy simulations with explicit water molecules, J. Phys. Chem. B 108 (2004) 16567–16576.

[133] U. C. Singh, F. K. Brown, P. A. Bash, P. A. Kollman, An approach to the application of free energy perturbation methods using molecular dynamics: Applications to the transformations of $CH_3OH \rightarrow CH_3CH_3$, $H_3O^+ \rightarrow NH_4^+$, Glycine → Alanine, and Alanine → Phenylalanine in aqueous solution and to $H_3O^+(H_2O)_3 \rightarrow NH_4^+(H_2O)_3$ in the gas phase, J. Am. Chem. Soc. 109 (1987) 1607–1614.

[134] Y. Sun, D. Spellmeyer, D. A. Pearlman, P. A. Kollman, Simulation of the solvation free energies for methane, ethane, and propane and corresponding amino acid dipeptides: A critical test of the "Bond PMF" correction, a new

set of hydrocarbon parameters, and the gas-phase-water hydrophobicity scale, J. Am. Chem. Soc. 114 (1992) 6798–6801.

[135] R. Staritzbichler, W. Gu, V. Helms, Are solvation free energies of homogeneous helical peptides additive?, J. Phys. Chem. B 109 (2005) 19000–19007.

[136] W. Im, M. S. Lee, C. L. Brooks, III, Generalized born model with a simple smoothing function, J. Comput. Chem. 24 (2003) 1691–1702.

[137] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, M. Karplus, All-atom empirical potential for molecular modeling and dynamics studies of proteins, J. Phys. Chem. B 102 (1998) 3586.

[138] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, M. Karplus, CHARMM - a program for macromolecular energy, minimization, and dynamics calculations, J. Comp. Chem. 4 (1983) 187.

[139] S. Boresch, M. Karplus, The role of bonded terms in free energy simulations: 1. theoretical analysis, J. Phys. Chem. A 103 (1999) 103–118.

[140] W. L. Jorgensen, H. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, Comparison of simple potential functions for simulating liquid water, J. Chem. Phys. 79 (1983) 926.

[141] E. Neria, S. Fischer, M. Karplus, Simulation of activation free energies in molecular systems, J. Chem. Phys. 105 (1996) 1902.

[142] W. F. Van Gunsteren, H. J. C. Berendsen, Algorithms for macromolecular dynamics and costraint dynamics, Mol. Phys. 34 (1977) 1311–1327.

[143] S. Jo, T. Kim, V. G. Iyer, W. Im, Charmm-gui: A web-based graphical user interface for charmm, J. Comput. Chem. 29 (2008) 1959–1865.

[144] M. Spichty, M. Karplus, private communication (2008).

[145] W. L. Jorgensen, D. S. Maxwell, J. Tirado-Rives, Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids, J. Am. Chem. Soc. 118 (1996) 11225–11236.

[146] S. Y. Sheu, D. Y. Yang, H. L. Selzle, E. W. Schlag, Energetics of hydrogen bonds in peptides, PNAS 100 (2003) 12683–12687.

[147] M. R. Shirts, V. S. Pande, Solvation free energies of amino acid side chain analogs for common molecular mechanics water models, J. Chem. Phys. 122 (2005) 134508.

[148] I.-M. Chu, W.-Y. Chen, Partition of amino acids and peptides in aqueous two-phase systems, in: R. Hatti-Kaul (Ed.), Aqueous Two-Phase Systems: Methods and Protocols, Vol. 11 of Methods in Biotechnology, Humana Press Inc., 2000, pp. 95–105.

[149] W. Im, J. Chen, C. L. Brooks, III, Peptide and protein folding and conformational equilibria: Theoretical treatment of electrostatics and hydrogen bonding with implicit solvent models, Adv. Prot. Chem. 72 (2006) 173–198.

[150] C. Tanford, Contribution of hydrophobic interactions to stability of globular conformation of proteins, J. Am. Chem. Soc. 84 (1962) 4240–4247.

[151] W. Kauzmann, Some factors in the interpretation of protein denaturation, Adv. Protein Chem. 14 (1959) 1.

[152] H. J. Dyson, P. E. Wright, H. A. Scheraga, The role of hydrophobic interactions in initiation and propagation of protein folding, Proc. Natl. Acad. Sci. U.S.A. 103 (35) (2006) 13057–13061.

[153] U. Langhorst, J. Backmann, R. Loris, J. Steyaert, Analysis of water mediated protein–protein interactions within RNase T1, Biochemistry 39 (2000) 6586.

[154] M. Tarek, D. Tobias, Environmental dependence of the dynamics of protein hydration water, J. Am. Chem. Soc. 121 (1999) 9740.

[155] J. Janin, Wet and dry interfaces: the role of solvent in protein–protein and protein–DNA recognition, Structure Fold. Des. 7 (1999) R277.

[156] A. Palomer, J. J. Pérez, S. Navea, O. L. amd J Pascual, L. García, D. Mauleón, Modeling cyclooxygenase inhibition. implication of active site hydration on the selectivity of ketoprofen analogues, J. Med. Chem. 43 (2000) 2280.

[157] G. Lamoureux, B. Roux, Modeling induced polarization with classical Drude oscillators: Theory and molecular dynamics simulation algorithm, J. Chem. Phys. 119 (6) (2003) 3025–3039.

[158] G. König, S. Boresch, Non-Boltzmann Sampling and Bennett's Acceptance Ratio Method: How free energy simulations can profit from bending the rules, J. Comp. Chem.

[159] D. L. Mobley, E. Dumont, J. D. Chodera, K. A. Dill, Comparison of charge models for fixed-charge force fields: Small-molecule hydration free energies in explicit solvent, J. Phys. Chem. B 111 (9) (2007) 2242–2254.

[160] R. Rizzo, W. Jorgensen, OPLS all-atom model for amines: Resolution of the amine hydration problem, J. Am. Chem. Soc. 121 (20) (1999) 4827–4836.

[161] S. J. Thompson, C. K. Hattotuwagama, J. D. Holliday, D. R. Flower, On the hydrophobicity of peptides: Comparing empirical predictions of peptide log P values, Bioinformation 1 (7) (2006) 237–41.

[162] R. Levy, L. Zhang, E. Gallicchio, A. Felts, On the nonpolar hydration free energy of proteins: Surface area and continuum solvent models for the solute-solvent interaction energy, J. Am. Chem. Soc. 125 (31) (2003) 9523–9530.

[163] A. C. V. Johansson, E. Lindahl, Protein contents in biological membranes can explain abnormal solvation of charged and polar residues, Proc. Natl. Acad. Sci. U.S.A. 106 (37) (2009) 15684–15689.

[164] N. Prabhu, K. Sharp, Heat capacity in proteins, Annu. Rev. Phys. Chem. 56 (2005) 521–548.

[165] A. Robertson, K. Murphy, Protein structure and the energetics of protein stability, Chem. Rev. 97 (5) (1997) 1251–1267.

[166] L. Wittgenstein, Tractatus logico-philosophicus, Suhrkamp, Berlin, 1963.

[167] M. Feig, C. L. Brooks III, Recent advances in the development and application of implicit solvent models in biomolecule simulations, Curr. Opin. Struct. Biol. 14 (2004) 217–224.

[168] J. Chen, C. L. Brooks, III, J. Khandogin, Recent advances in implicit solvent-based methods for biomolecular simulations, Curr. Opin. Struct. Biol. 18 (2) (2008) 140–148.

Declaration by the author

In this dissertation I presented four related methodological studies concerning free energy simulations. Two of the chapters have been previously published, one has been accepted for publication and the last chapter is about to be submitted:

1.) Unorthodox uses of Bennett's acceptance ratio method
Gerhard König, Stefan Bruckner, Stefan Boresch
Journal of Computational Chemistry, **30**(11), 1712-18 (2009)
(corresponds to Chapter 3)

2.) Non-Boltzmann Sampling and Bennett's Acceptance Ratio Method: How to profit from bending the rules
Gerhard König, Stefan Boresch
Accepted by Journal of Computational Chemistry (2010)
(corresponds to Chapter 4)

3.) Hydration Free Energies of Amino Acids: Why Side Chain Analog Data Are Not Enough
Gerhard König, Stefan Boresch
Journal of Physical Chemistry B, **113**(26), 8967-8974 (2009)
(corresponds to Chapter 5)

4.) Absolute hydration free energies of blocked amino acids: Are current estimates of protein solvation overvalued ?
Gerhard König, Stefan Bruckner, Stefan Boresch
To be submitted
(corresponds to Chapter 6)

Except for the last work, to which my colleague Stefan Bruckner contributed equally, I have been the first author of all works. The experiments were designed and analyzed together with my supervisor Stefan Boresch. However, all simulations and analyses presented here were performed by myself, with some exceptions in Chapters 3 (the data in Tables 3.2 and 3.3) and 6 ($\Delta A^{H_2O}_{unch.aa \rightarrow PG}$ and $\Delta A^{gas}_{unch.aa \rightarrow PG}$ were calculated by Stefan Bruckner).

Curriculum Vitae

Name: Dipl.-Ing. Mag. Gerhard König
Anschrift: Siegfried Esterl Gasse 15, A-8160 Weiz
Geburtsdatum: 18.05.1981
Geburtsort: Weiz

Eltern:
Gerhard Alexander König, geb. 16.10.1955, Automechanikermeister
Margareta König (geb. Städtler), geb. 05.02.1954, Bürokauffrau

Ausbildung:

1987-1991	Volkschule Weiz
1991-1999	Bundesrealgymnasium Weiz
2000	Präsenzdienst
2000-2005	Studium Molekulare Biologie an der Universität Wien mit den Schwerpunkten Biochemie, Strukturbiologie und Bioinformatik. Diplomarbeit bei Prof. Stefan Boresch am Institut für Theoretische Chemie und Molekulare Strukturbiologie.
15.12.2005	Abschluss Molekulare Biologie (mit Auszeichnung)
2006-2010	Studium Scientific Computing an der Universität Wien mit Schwerpunkt Computational Physics. Masterarbeit bei Prof. Wilfried Gansterer im Research Lab Computational Technologies and Applications.
15.01.2010	Abschluss Scientific Computing (mit Auszeichnung, erster Absolvent)
Seit 2006	Doktoratsstudium der Naturwissenschaften (Molekulare Biologie) am Institut für Computergestützte Biologische Chemie bei Prof. Stefan Boresch.