Disentangling Force Field and Sampling Issues in Biomolecular Systems

Von der Fakultät Energie-, Verfahrens- und Biotechnik der Universität Stuttgart und dem Stuttgart Center for Simulation Science (SC SimTech) zur Erlangung der Würde eines Doktors der Ingenieurwissenschaft (Dr.-Ing.) genehmigte Abhandlung

Vorgelegt von
Daniel Markthaler
aus Stuttgart-Bad Cannstatt

Hauptberichter: apl. Prof. Dr.-Ing. Niels Hansen
Mitberichter: Prof. Dr. rer. nat. Robin Ghosh
Prof. Dr. Chris Oostenbrink

Tag der mündlichen Prüfung: 23.07.2020

Institut für Technische Thermodynamik und Thermische Verfahrenstechnik der Universität Stuttgart

2020

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit mit dem Titel

"Disentangling Force Field and Sampling Issues in Biomolecular Systems"

selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, dass alle Stellen der Arbeit, die wörtlich oder sinngemäß aus anderen Quellen übernommen wurden, als solche kenntlich gemacht sind.
Ich versichere außerdem, dass die vorliegende Dissertation nur in diesem und keinem anderen Promotionsverfahren eingereicht wurde und dass diesem Promotionsverfahren keine endgültig gescheiterten Promotionsverfahren vorausgegangen sind.

I herewidth duly declare that I have authored the dissertation

"Disentangling Force Field and Sampling Issues in Biomolecular Systems"

independently and only with use of specified aids. I have mentioned all sources used and cited correctly according to scientific rules.

Stuttgart, 10.01.2020

Daniel Markthaler

Contents

1 Introduction 14

1.1 Fundamental Concepts of Protein Stability 17

1.2 Small Protein Domains 20

1.3 Impact of Co-Solvents 21

1.4 Protein-Ligand Binding 22

1.5 Purpose of this Work 24

1.5.1 TMAO Force Field Comparison 24

1.5.2 Calculation of Relative Folding Free Energies 25

1.5.3 Calculation of 1D-Potentials of Mean Force 25

2 Validation of Trimethylamine-N-oxide (TMAO) Force Fields Based on Thermophysical Properties of Aqueous TMAO Solutions 26

2.1 Introduction 27

2.2 Theory 29

2.2.1 Partial and Apparent Molar Volume 29

2.2.2 Dielectric Properties 29

2.3 Computational Details 31

2.3.1 TMAO Force Fields 31

2.3.2 Simulated Systems 33

2.3.3 Simulation Parameters 35

2.3.4 Trajectory Analysis 37

2.4 Results and Discussion 39

2.4.1 Volumetric Properties 39

2.4.2 Transport Properties 43

2.4.3 Structural Properties 43

2.4.4 Dielectric Properties 48

2.5 Conclusion 51

3 Overcoming Convergence Issues in Free-Energy Calculations of Amide-to-Ester Mutations in the Pin1-WW Domain 53
3.1 Introduction . . . 54
3.2 Methods . . . 55
3.2.1 Computational Approach . . . 55
3.2.2 Simulation Setup . . . 57
3.2.3 Analysis . . . 62
3.3 Results and Discussion . . . 62
3.3.1 Independence of $\Delta G^{\mathrm{u}}_{\mathrm{mw}}$ on the Peptide Size . . . 62
3.3.2 Dependence of $\Delta G^{\mathrm{f}}_{\mathrm{mw}}$ on the Starting Structure . . . 63
3.3.3 Hamiltonian Replica Exchange . . . 64
3.3.4 Correlation Between $\Delta G^{\mathrm{u}}_{\mathrm{mw}}$ and $\Delta G^{\mathrm{f}}_{\mathrm{mw}}$ . . . 66
3.3.5 Hydrogen Bond Analysis . . . 71
3.3.6 Comparison with Experimental Data . . . 71
3.4 Conclusion . . . 73

4 Lessons Learned from the Calculation of One-Dimensional Potentials of Mean Force 77
4.1 Introduction . . . 78
4.2 Theory . . . 80
4.3 Methods . . . 84
4.3.1 Host-Guest Systems . . . 84
4.3.2 Simulation Protocol . . . 84
4.3.3 Simulation Code and Parameters . . . 88
4.3.4 Free Energy Estimation . . . 91
4.4 Results . . . 92
4.4.1 Unpolar CNT / Methane . . . 92
4.4.2 Unpolar CNT / Multiatomic Ligand . . . 97
4.4.3 Polar CNT / Unpolar Ligand . . . 102
4.4.4 Polar CNT / Dipolar Ligand . . . 105
4.4.5 Cyclodextrin / Alcohols . . . 107
4.5 Discussion . . . 111
4.5.1 Enforcing PMF Periodicity . . . 111
4.5.2 Influence of the Host's / Ligand's Flexibility . . . 113
4.6 Conclusions . . . 113

5 Statistical Mechanical Considerations on the Calculation of the Binding Free Energy from Molecular Simulations 117

6 Conclusion and Outlook 126

6.1 TMAO Force Field Comparison . . . 127

6.2 Calculation of Relative Folding Free Energies . . . 127

6.3 Calculation of 1D-Potentials of Mean Force . . . 129

Appendices 158

A Validation of Trimethylamine-N-Oxide (TMAO) Force Fields Based on Thermophysical Properties of Aqueous TMAO Solutions 159

A.1 Detailed Simulation Results . . . 159

A.2 Radial Distribution Functions . . . 165

A.3 Densities and Apparent Molar Volumes . . . 193

A.4 Hydrogen Bond Analysis and Self-Diffusion Coefficient . . . 194

A.5 Enthalpy and Apparent Molar Volume at Infinite Dilution . . . 195

A.6 Dielectric Properties . . . 196

A.6.1 Calculation of Dielectric Spectra . . . 196

A.6.2 Noise Reduction . . . 196

A.6.3 Calculation of Static Permittivities . . . 199

A.6.4 Simulation Details . . . 200

A.6.5 Dielectric Spectra . . . 202

A.6.6 Static permittivities . . . 205

B Overcoming Convergence Issues in Free-Energy Calculations of Amide-to-Ester Mutations in the Pin1-WW Domain 206

B.1 EDS simulations . . . 206

C Lessons Learned from the Calculation of One-Dimensional Potentials of Mean Force 216

C.1 Double Decoupling . . . 216

C.2 Simulations with All-Atom Force Field . . . 218

Zusammenfassung

Das Verständnis von den thermodynamischen Grundlagen zu Proteinstabilität und deren Beeinflussung ist für die Grundlagenforschung gleichermaßen wichtig wie für die biotechnologische und pharmazeutische Anwendung. Trotz jahrzehntelanger Forschung sind viele Aspekte der Protein-Thermodynamik noch nicht im Detail verstanden. Ein Beispiel ist der molekulare Wirkungsmechanismus von stabilisierenden Osmolyten wie Trimethylamin-N-oxid (TMAO). Molekulardynamik- (MD) simulationen, basierend auf klassischen Kraftfeldern, erlauben die Beobachtung von Vorgängen auf atomarer Skala. Eine zentrale Voraussetzung, um aussagekräftige mechanistische Erkenntnisse mittels MD-Simulationen zu erhalten, ist die adäquate Beschreibung der relevanten Wechselwirkungen durch das molekulare Kraftfeld. Kapitel 2 befasst sich mit einem Kraftfeldvergleich für TMAO auf der Basis verschiedener thermo-physikalischer Eigenschaften von wässrigen TMAO-Lösungen.
Gleichermaßen wichtig wie die Qualität des Kraftfelds, ist die Robustheit des Simulationsprotokolls. Besonders im Kontext von freien Energieberechnungen werden Limitierungen des Kraftfelds oftmals von "Sampling-Problemen" überlagert. Die Etablierung eines robusten Simulationsprotokolls setzt die Auswahl geeigneter Modellsysteme voraus. Letztere müssen einfach genug sein, sodass eine Entkopplung der oben genannten Limitierungen möglich ist. Kapitel 3 behandelt die Auswirkung von Mutationen im Rückgrat eines kleinen β-Faltblatt-Proteins auf der Basis von alchemischen freien Energiedifferenzen. Die Studie zeigt, dass die Berechnung von robusten und von der Startstruktur unabhängigen freien Energieänderungen häufig herausfordernd sein kann, selbst im Fall von eher kleinen Proteinen. Kapitel 4 handelt von der Berechnung von Bindungsaffinitäten über das Potential der mittleren Kraft für eine Reihe von Wirt-Gast-Modellsystemen. Hierbei zeigt sich, dass selbst bei Anwendung von etablierten Methoden für vergleichsweise einfache Systeme, Simulationsartefakte auftreten können. In Kapitel 5 werden theoretische Hintergründe und Herleitungen der Arbeitsgleichungen des vorangegangenen Kapitels auf Basis statistisch-mechanischer Überlegungen präsentiert.

Summary

Understanding the thermodynamic basics of protein stability and its manipulation is equally important for fundamental research as well as biotechnological and pharmaceutical application. However, despite decades of research, many aspects of protein thermodynamics are still not fully understood. One example is the molecular mechanism of protecting osmolytes such as trimethylamine N-oxide (TMAO). Molecular dynamics (MD) simulations based on classical force fields, enable the observation of processes on the atomic scale. A central requirement to obtain meaningful mechanistic insights by MD simulations is the adequate description of the relevant interactions by the molecular force field. Chapter 2 deals with a force field comparison for TMAO, based on various thermophysical properties of aqueous TMAO solutions.

Equally important as the quality of the force field, is the robustness of the simulation protocol. Especially in the context of free energy calculations, force field limitations are often superimposed by "sampling issues". The establishment of a robust simulation protocol requires the selection of suitable model systems. The latter must be simple enough to enable the disentanglement of the limitations mentioned above. Chapter 3 is about the impact of mutations in the backbone of a small β-sheet protein based on alchemical free energy differences. The study shows that the calculation of robust free energy changes which are independent of the starting structure can often be challenging, even in case of rather small proteins. Chapter 4 is concerned with the calculation of binding affinities from the potential of mean force for a set of host-guest model systems. It is found that simulation artifacts can occur, even when applying established methods to rather simple systems. Chapter 5 presents the theoretical background and derivations of the working equations used in the previous chapter, based on statistical-mechanical considerations.

Journal publications

This thesis led to the following publications:

- Chapter 2: D. Markthaler, J. Zeman, J. Baz, J. Smiatek and N. Hansen: Validation of Trimethylamine-N-oxide (TMAO) Force Fields Based on Thermophysical Properties of Aqueous TMAO Solutions, *The Journal of Physical Chemistry B*, **121**, 10674-10688, 2017

- Chapter 3: D. Markthaler, H. Kraus and N. Hansen: Overcoming Convergence Issues in Free-Energy Calculations of Amide-to-Ester Mutations in the Pin1-WW Domain, *Journal of Chemical Information and Modeling*, **58**, 2305-2318, 2018

- Chapter 4: D. Markthaler, S. Jakobtorweihen and N. Hansen: Lessons Learned from the Calculation of One-Dimensional Potentials of Mean Force [Article v1.0], *Living Journal of Computational Molecular Science*, **1**, 11073, 2019

The chapters 2 to 4 present literal quotes of the published work. Any addition with respect to the published work is marked. Any deletion is indicated with square brackets as ,[...]'. Cross-references between chapters of this thesis, which are added to the published version of the text to increase readability, are marked by square brackets. The Supporting Informations to the single chapters are presented in the Appendix of this thesis.

Related publications:

- D. Markthaler, J. Gebhardt, S. Jakobtorweihen and N. Hansen: Molecular Simulations of Thermodynamic Properties for the System α-Cyclodextrin/Alcohol in Aqueous Solution, *Chemie Ingenieur Technik*, **89**, 1306-1314, 2017

- J. Baz, J. Gebhardt, H. Kraus, D. Markthaler and N. Hansen: Insights into Noncovalent Binding Obtained from Molecular Dynamics Simulations, *Chemie Ingenieur Technik*, **90**, 1864-1875, 2018

Danksagung

Im Folgenden möchte ich mich bei einer Reihe von Personen bedanken, die wesentlich zum Gelingen dieser Arbeit beigetragen haben, sei es auf wissenschaftlicher, wie auch menschlicher Ebene. Eine Danksagung birgt immer auch die Gefahr, dass die Menschen, denen besonderer Dank gebührt, nicht bzw. in nicht ausreichendem Maße gewürdigt werden. Unternimmt man andererseits den Versuch mit besonderen Stilmitteln zu arbeiten, um seinen Worten mehr Gewicht zu verleihen, begibt man sich schnell in fachfremdes Terrain. Ein Beispiel sind (unnötig) bedeutungsschwangere Zitate, von denen man als Wissenschaftler (besonders als Ingenieur) am besten die Finger lassen sollte. Ich habe mein Bestes gegeben, die richtige Balance zwischen Vollständigkeit, wissenschaftlicher Kompaktheit und Emotionalität zu finden. Gerade in Bezug auf letztere könnte ich mir keine passendere Passage innerhalb einer Dissertation vorstellen wie die Danksagung. Alle Leute, die ich im Folgenden vergessen habe, mögen nachsichtig mit mir sein.

Zunächst gilt mein besonderer Dank selbstverständlich meinem Doktorvater Niels Hansen. Als KIT'ler hatte ich mich ohne jegliche Vorkenntnisse in Molekularsimulation am ITT beworben. Trotz dieses "Makels", bist du das Wagnis mit mir eingegangen. Ob dies letztlich einzig auf meine Person oder den Mangel an Alternativen zu dieser Zeit zurückzuführen war, ist dabei eigentlich unerheblich. Du bist sicherlich einer der systematischsten und analytischsten Köpfe, die ich je kennen lernen durfte. Unsere wissenschaftlichen Diskussionen waren ein sehr wichtiger Bestandteil während der gesamten Zeit, von denen ich sehr profitiert habe. Besonders am Anfang hast Du mir die notwendige Führung gegeben und mir im weiteren Promotionsverlauf immer mehr Raum zur Entfaltung gegeben. Deine wissenschaftliche Hartnäckigkeit habe ich stets sehr geschätzt. Nichtsdestotrotz hoffe ich, dass, wenn ich in ferner Zukunft von der "Suche nach Artefakten" höre, dabei in erster Linie wieder an Indiana Jones denken kann und nicht mehr an Kapitel 4. Niels, vielen Dank für alles! Ich hoffe, es hat dir auch Spaß gemacht.

Desweiteren möchte ich mich bei Joachim Groß für die Einstellung zur Promotion am ITT bedanken. Danke, dass Du mir diese Chance gegeben hast Joachim! Ohne zu dick aufzutragen, kann ich behaupten, dass Du der eindrucksvollste Thermodynamik-Lehrerende bist, den ich jemals erlebt habe. Gleichauf mit Herrn Prof. Schaber, versteht sich. Ich habe dich stets als sehr diplomatischen und verständnisvollen Chef erlebt und ich kann mich glücklich schätzen, eine so lange Zeit Teil deines Teams gewesen zu sein.

Robin Ghosh habe ich sehr viel zu verdanken. Selten habe ich einen Wissenschaftler kennen gelernt, der so aufopfernd forscht. Deine Vorlesungen über die vielen faszinierenden Aspekte aus der Welt der Proteinforschung und unsere vielen Diskussionen, nicht zuletzt im Zuge unseres Paper Journal Clubs, waren Quellen für viel Inspiration und haben mir sehr geholfen über den Tellerrand hinaus zu blicken. Dank dir habe ich meine Begeisterung für experimentelle Arbeiten (wieder) entdeckt. Die Möglichkeit zwischen Simulations- und Experimentations-Welt wechseln zu können hat mir sehr gut gefallen und stets für Abwechslung gesorgt. Robin, ich danke dir für die vielen Stunden, die Du in meine experimentelle Ausbildung investiert hast, für die Chance an spannenden Projekten teilhaben zu dürfen und nicht zuletzt für die Übernahme des Co-Referats.

Bei Chris Oostenbrink bedanke ich mich für die Übernahme des zweiten Co-Referats und für den Forschungsaufenthalt in seiner Gruppe gleich nach Promotionsbeginn. Ich hätte mir als Fachfremder keinen besseren Einstieg ins Thema wünschen können. Egal ob Replica Exchange, Distance Field oder WHAM, bei euch ging es direkt von Tag 1 an richtig ab! Du und dein Team (wobei ich besonders Manuela, Matthias, Martina, Urban und alle Marias nennen möchte) haben mich so nett aufgenommen, so dass ich mich sofort voll integriert gefühlt habe. Außerdem konnte ich Wien für mich entdecken. Es war eine klasse Zeit, aus der ich auf verschiedensten Ebene sehr viel mitgenommen habe.

Ich danke Wilfred van Gunsteren, dass er mir 6 (!) Mal die Chance gegeben hat, beim BIOMOS-Meeting in Ausserberg dabei sein zu dürfen. Von all den Konferenzen und Meetings, die die Reisegruppe Hansen besucht hat (und das waren viele), war Ausserberg für mich immer etwas ganz besonderes. Niemals zuvor habe ich einen so intensiven und ungezwungen fachlichen Austausch mit vielen beeindruckenden Persönlichkeiten (allen voran Wilfred, Alan Mark und Philippe Hünenberger) erlebt wie in Ausserberg.

Eine besonders wichtige Komponente waren und sind meinen ITT-Kollegen, die aktuellen sowie die ehemaligen. Seid mir nicht böse, wenn ich euch nicht alle einzeln aufzähle. Als ich kam, habe ich großartige Kollegen vorgefunden. Ihr habt mich sehr nett aufgenommen (trotz des badischen Einschlags) und seid immer hilfsbereit gewesen. Wenn ich gehe, hoffe ich, dass die Kontakte auch weiterhin aufrecht erhalten werden. Leute, ich danke euch für die schöne und intensive Zeit!

Erwähnung finden müssen auch die hervorragenden studentischen Arbeiter(innen) - allen voran Max, David, Nadine, Timm und Hamzeh - die ich durch ihre Abschlussarbeiten am Institut begleiten durfte. Natürlich verhält sich die Situation hier wie beim Wein, dessen Qualität in erster Linie durch die Güte der Auslese bestimmt wird. Nadine, ich glaube man kann sagen, dass deine BA nicht nur zur Erweiterung unseres Know Hows über die dynamische Lichtstreuung geführt hat, sondern auch zum Ausbau unserer beiden Netflix-

Profile. Ich werde wahrscheinlich nie sattelfester in den Hintergründen der Drogenkartelle sein als während dieser Zeit. Es freut mich sehr, dass sich drei von euch zur Promotion entschieden und sich damit die (statistisch-mechanisch) harten Ketten des ITTs haben anlegen lassen. Die Vierte im Bunde kommt bestimmt noch?!

Kommen wir zu den Leuten an der Heimatfront. Meine Eltern ebenso wie meine Großeltern waren immer für mich da und haben mich mit all ihren Kräften immer unterstützt. Ich danke euch für diese ständige Unterstützung, allen voran meiner Mutter. Mama, Du bist die Beste! Selten war etwas ernst gemeinter und wahrer als das!
Da sich das Ende unter Umständen doch am besten mit einem gehaltvollen Zitat besiegeln lässt, habe ich mich hierzu bei einem der wahrscheinlich besten jemals produzierten Filme bedient: "Call me Snake." (Snake Plissken in "Die Klapperschlange" (1981))

Chapter 1

Introduction

Proteins form the fundamental building blocks of biological processes, ranging from enzymatic catalysis to molecular recognition and self assembly. Protein folding is a prerequisite for protein function. However, in 1966 Levinthal showed that protein folding is a very unlikely process[1]. Thus, for a random search, the average time τ for sampling all possible conformations of a polypeptide chain consisting of $N+1$ residues can be estimated as[2,3]:

$$\tau = (N k_v)^{-1} \, j^N \tag{1.1}$$

Assuming a maximal rate constant k_v for interconverting between conformations of $10^{13}\,\mathrm{s}^{-1}$ and $j = 2-8$ possible (backbone) conformations per residue[3], one calculates an estimated sampling time of $\tau = 10^7 - 10^{67}$ years, even for a small protein of only $N = 100$ residues. For comparison, the estimated age of the universe is around 10^{10} years[4]. Since the function of a protein is defined by its three-dimensional structure, life on earth would not be possible, if the estimation above would be valid. Strikingly, conformational sampling of real proteins is much faster with typical folding timescales between several micro- to milliseconds. The attempt to resolve this apparent conflict which is known as the "Levinthal paradox", was one of the initial questions of protein folding research. Protein folding research, which has evolved to one of the most actively studied fields of biophysics, is primarily concerned with two fundamental questions: (i) what are the thermodynamic driving forces of protein folding and (ii) what is the kinetic folding mechanism? Besides the fundamental interest, both questions have far-reaching implications for socially relevant fields such as biotechnological industry and medical research. For example, various diseases such as Alzheimer's disease, Parkinson's disease and Type 2 diabetes mellitus are caused by misfolded proteins[5]. Another problem which is closely related to the two addressed above, is the question if it is possible to predict the native (or folded) state structure of a protein only from the knowledge of the amino acid sequence. It has to be stressed that this question always refers to a given thermodynamic environment, defined by temperature, pressure and solution composition. One motivation for seeking to answer this question comes from rational drug design: while the development of high-throughput sequencing techniques enables rapid determination of protein sequences at moderate costs, the experimental determination of protein structures can still be very time consuming. Since the design of drug candidates usually starts from analysis of the structure of the target receptor, much effort has been put into the development of robust modeling approaches. Together with the questions (i) and (ii), this defines the so called **protein folding problem (PFP)**[6]. Despite decades of research which have elucidated many aspects of the PFP, the questions addressed above are by no means fully understood, even in 2020. A compact summary of the current understanding of protein folding is presented in Sec. 1.1.

While the research was initially driven by experimental investigations, modeling approaches such as **molecular dynamics (MD)** have gained ongoing importance. Computer-aided rational drug design is a nice illustration of the favorable synergetic effect between biomolecular simulation methods and experiments. Here, MD can be used to propose potential drug candidates or to limit the number of candidates to be synthesized[7,8]. MD simulations based on classical force fields provide a direct way for estimating averages of macroscopic observables (thermodynamic and dynamic) from microscopic data (coordinates, momenta). The theoretical framework for bridging these two worlds is delivered through application of statistical mechanics. Since the first protein simulations from the 1970s[9], significant theoretical, algorithmic and hardware developments enabled a continuous pushing of the boundaries towards bigger and more complex systems and longer timescales. One impressive example is the study of the structural dynamics of ubiquitin on the millisecond time scale using unbiased MD simulations[10].

As early as the mid-1970s, it was believed by some experts in the field that modeling approaches would soon reach a predictive quality such that experimental investigations can be reduced considerably or might even become redundant at all[1]. However, more than 30 years later we know that despite the fact that simulation techniques have become firmly established in biophysics as complementary tool and delivered impressive atomic-level insights, experimental approaches are still indispensable. Some major obstacles which prevented the big breakthrough of MD simulations in the early days were: (i) limited resolution of experimentally determined structures (> 2 Å); (ii) limited accessible simulation time scales; (iii) limited force field accuracy.

In the following, I will elaborate further on these challenges, the achieved improvements in recent years and their implications for biomolecular modeling. The availability of high (< 1.5 Å) and even ultra-high resolution (< 1.0 Å) X-ray structures enable the study of structure/function relationships at a novel quality level[11] together with precise estimation of bond lengths and angles. The latter is of great importance for force field development and refinement. Classical force fields approximate the total potential energy function by a sum of additive pairwise interaction terms. In many established force fields, electrostatic interactions are taken into account in the form of fixed partial charges associated with particular interaction sites. As a consequence, the notion of a particular pH value is primarily reflected in the assignment of fixed protonation states to components of the system such as amino acid residues. Examples for established biomolecular force fields are GROMOS[12,13], CHARMM[14,15], AMBER[16,17]. Despite differences in the underlying parametrization philosophy, the involved potential energy terms share a similar mathematical structure. To improve force field accuracy, one can increase the complexity of certain potential energy terms based on physical arguments. Examples are the inclusion of electronic polarizability[18] or generalized functional forms for dispersive interactions instead of the "standard" Lennard-Jones (12,6) potential[19]. The problem with such an approach is that it ultimately involves the inclusion of further parameters which once again have to be optimized in a consistent way. On the other hand, there are indications that within the huge parameter space of transferable force fields based on fixed partial charges, there is still room for improvement by (re-)optimizing the involved parameters[20,21]. Clearly, the terms "improvement" and "optimization" always refer to a particular application such as the prediction of vapor-liquid equilibria[19] or binding free energies[22,23]. Once a parametrized molecular model, i.e. force field, is "sufficiently" validated, it can be used to predict other physical properties which had not been incorporated into the parametrization process, or the model can be applied to different thermodynamic conditions (e.g. high pressure, other temperatures, solvent composition).

The current accessibility of nano-to microsecond timescales enables the validation of force fields on properties which require long sampling such as binding or folding free energies[24–26]. The results obtained from testing large sets of molecules can then be used to refine certain parameter subsets of the force field[27]. However, in case of complex molecules such as proteins, the assessment of the force field quality remains challenging. There are indications, based upon long unbiased MD simulations of proteins, that current force fields prefer folded conformations that are too stable compared to experimental reality[28]. This finding is not surprising considering that biomolecular force fields were parametrized in such a way as to preserve the experimental protein structures. Consequently, efforts were put into the improvement of torsional energetics in some of the main biomolecular force fields[29,30]. Due to the complexity of protein molecules, biomolecular simulations are still frequently restricted to the limiting case of infinite dilution, i.e. a single solvated protein or single host-guest pair. For this reason, comparison with results from bulk experiments, which are always based on populations of protein molecules, is often not straightforward. Moreover, this complexity complicates an unambiguous assessment of the quality of the force field, since limitations of the latter are typically superimposed by sampling issues.

This is especially problematic for the calculation of free energy changes such as binding or folding free energies.

In the scope of this thesis, special focus was given to the development of robust simulation protocols by identification and elimination of simulation artifacts through usage of enhanced sampling techniques. To minimize sampling-related issues, investigations were restricted to rather simple model systems such as a small β-sheet protein (see chapter 3) or idealized host molecules such as a carbon nano tube and cyclodextrin in case of binding studies (see chapter 4).

1.1 Fundamental Concepts of Protein Stability

Below, I outline some fundamental aspects of protein stability relevant for the questions addressed above. Therefore, I have focused on small single-domain proteins in dilute aqueous solution. For more general and in-depth reviews including cellular protein folding, I refer to Ref. 31.

Proteins are characterized by unique physicochemical properties in comparison to any other common (bulk) material: (i) compressibility measurements show that proteins have packing densities between 0.68 and 0.73, comparable to organic solids[32,33]; (ii) unfolding free energies $\Delta G_{F \rightarrow U}$ are low and found to be in the limited range of 20 to 40 $\mathrm{kJ\,mol^{-1}}$[3]; (iii) proteins are subject to significant statistical fluctuations[34]. The conflict resulting from the combination of apparently contrary properties above can be rationalized by considering that proteins are highly dynamic structures. While being tightly packed on average (time, ensemble), local unfolding events are possible and occur continuously.

The statistical-mechanical treatment of protein folding relies on the underlying free energy or rather **Gibbs energy landscape (GEL)** of the protein. The GEL conceptually represents the simultaneous decrease of the effective potential energy together with the configurational entropy of the amino acid chain upon folding[31,35]. Due to this correlation, the GEL is also denoted as a "folding funnel". According to energy landscape theory, a multiplicity of possible folding routes exists in the early stage of protein folding rather than a single pathway. At a later stage, discrete pathways emerge when much of the protein has already achieved a correct configuration[35]. The concept of the GEL originates from the theory of spin glasses and was further developed with the aid of coarse-grained lattice models based on simplified potentials. Since the GEL represents an average with respect to the solvent degrees of freedom, it is a temperature-dependent **potential of mean force (PMF)**. In the usual representation, it is depicted as function of one or two order parameters. Since the latter measure the progression of the system towards the native state, they are also called progress(ion) coordinates. Typical progress coordinates are geometrical quantities such as the radius of gyration or the fraction of native contacts or thermodynamic quantities. Finding the most informative coordinates is a central problem on its own[36]. Despite the plausibility of such plots, it should be kept in mind that they represent a simplified description which is dependent on the choice of order parameters. The "true" free energy surface in contrast is a high-dimensional rugged function featuring many local minima[37]. From a thermodynamic and kinetic perspective, the folding/unfolding transition must involve a continuum of states[34,37]. However, there is conclusive experimental evidence that the thermodynamic equilibrium ($F \rightleftharpoons U$) for most water-soluble monomeric single-domain proteins can described in a simplified manner, by considering only the folded (F) and unfolded (U) state[31,34]:

$$K_{\mathrm{eq}} = \exp\left(-\frac{\Delta G_{F\to U}}{RT}\right) = \frac{[U]}{[F]} \tag{1.2}$$

with equilibrium constant K_{eq} and thermal energy RT. Clearly, the F and U state can be represented by a distribution or ensemble of conformations rather than single structures. Fig. 1.1 shows the schematic GEL for a two-state folding protein as function of two progression coordinates.

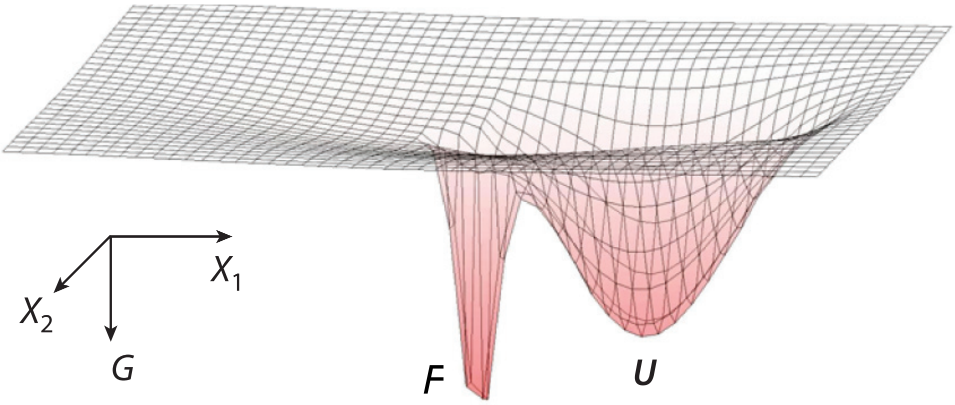


Figure 1.1: Schematic representation of the Gibbs energy (G) landscape of a two-state folding protein (picture taken from Ref. 31). The folded state (F) and unfolded state (U) are separated by a barrier. Axes illustrate two progression coordinates (x_1 and x_2).

For concentrated protein solutions, the molar concentrations in Eq. (1.2) (denoted as $[F]$ and $[U]$) would have to be replaced by thermodynamic activities. In the view of the two-state model, folding intermediates such as the molten globule state which is a partially folded structure are thought to be unimportant. Two-state behavior was not exclusively deduced from classical bulk experiments involving protein populations, but more recently also from single-molecule experiments[38]. It should be stressed that despite the popularity of the two-state model (not least because of its simplicity), its validation in general requires multiple thermodynamic as well as kinetic experiments[39,40]. Beyond that, there are many proteins for which the folding trajectories involve detectable folding intermediates along the folding pathway as well as proteins with continuous non-cooperative transitions such as the class of "downhill folders" which appear to lack a free energy barrier completely[31].

The narrow range of experimentally determined unfolding free energies mentioned above results from the combination of multiple weak non-bonded interactions[41,42]: (i) omnipresent van der Waals or London dispersion interactions, (ii) electrostatic interactions between dipolar and charged groups, (iii) hydrophobic interactions, (iv) intramolecular hydrogen bonding and (v) the gain in configurational entropy of the polypeptide chain upon unfolding. The importance of side chain packing (primarily mediated through van der Waals interactions) between buried residues within the hydrophobic core also has to be mentioned[43,44]. The marginal stability of proteins is the consequence of a delicate balance between this stabilizing and destabilizing interactions. For deeper insights into the thermodynamics of protein folding, the partitioning of the listed interactions into an enthalpic ($\Delta H_{F\to U}$) and entropic ($\Delta S_{F\to U}$) contribution according to: $\Delta G_{F\to U} = \Delta H_{F\to U} - T\Delta S_{F\to U}$ is of particular interest. The central role of the solvent environment for protein folding is primarily associated with the entropic term $\Delta S_{F\to U}$[45]. Typically the enthalpic and entropic differences are much higher in value than $\Delta G_{F\to U}$, very temperature-dependent and nearly compensate each other[3,46,47]. Based on current experimental evidence, it is assumed that hydrophobic interactions and intramolecular hydrogen bonding are of major importance for the stability of many proteins[3,48–50]. However, especially the energetic role of hydrogen bonds for protein stability is still under active discussion[48,50,51].

Hydrogen bonds (HBonds) are usually considered to be dominantly electrostatic interactions between permanent electrical dipoles such as -NH and -C=O groups, resulting in the partial sharing of a hydrogen atom between a HBond donor and acceptor. The HBond strength is critically dependent on the arrangement (distance, orientation) of the donor/acceptor pair as well as on the dielectric environment. The energetic scale for the required work to break a HBond can range from 25 $\mathrm{kJ\,mol^{-1}}$ in vacuum[51] to a value close to zero in water (depending on the position within the protein). Despite the rather small energetic differences between solute-solute, solute-solvent and solvent-solvent HBonding for proteins in aqueous solution, it has been shown experimentally that even a single unpaired HBond can destabilize the folded state significantly[48,52].

Hydrophobic interactions are intimately associated with the HBonding properties of water as well as other electrostatic effects due to the exceptionally high dipole moment of water[34,53]. The gain in orientational and translational degrees of freedom of water molecules upon folding is a dominant contribution, causing hydrophobic residues to be buried within the protein interior. While the thermodynamics of the hydrophobic effect were studied in detail for small organic molecules[54], the meaning of the term is different in the literature. According to the hydrophobic collapse mechanism[41,51,55], hydrophobic interactions are of primary importance for the initial step of the folding mechanism. At a later stage of the folding process, the formation of secondary structure together with correct packing is mainly accomplished by specific side-chain and backbone interactions.

1.2 Small Protein Domains

Typical water-soluble proteins, such as barnase[43], contain a rather hydrophobic interior while hydrophilic residues are located on the outer surface, accessible to the solvent. In recent years, increasing numbers of small independently folding protein domains have been discovered[56–58]. In general, the small protein size (< 8 kDa, < 100 residues) prevents an unambiguous definition of a classical hydrophobic core. Due to their simplicity which allows a focus at specific aspects of protein folding, these domains have become of increasing interest for experimentalists as well as theoreticians. It should be stressed that the transferability of such insights to more complex proteins has to be analyzed critically, since these proteins only deliver a "minimalistic" answer to the PFP[59].

The family of **WW domains** which is named according to two highly conserved tryptophan residues, represents one of the smallest ($34 - 44$ residues) native folds revealed to date[60]. WW domains appear to be mainly stabilized by intramolecular HBonding, which is why they have been established as excellent model systems for β-sheet folding proteins. They are involved in protein-protein recognition through binding of proline-rich ligands and possess a twisted three-stranded β-sheet configuration (see Fig. 3.1). These binding domains are thought to play a role in several human genetic disorders, such as Liddle's syndrome, muscular dystrophy, and Alzheimer's disease[61]. Their folding characteristics have been studied in great detail by numerous thermodynamic and kinetic measurements and extensive protein engineering studies. Main results can be summarized as follows[62–66]: (i) the folding/unfolding equilibrium for the majority of WW domains and experimental conditions can be described well by a two-state model; (ii) the unfolding free energies are typically very low, with values around 5 kJ mol^{-1} (~ 2 RT at $T = 300$ K); (iii) thermal unfolding transitions are broad, occurring over temperature intervals of $\sim 25 - 80$°C; (iv) for chemical unfolding, high denaturant concentrations (up to 7 M) are required for complete unfolding; (v) in contrast to typical two-state folders, $\Delta G_{F \to U}$ obtained from thermal and chemical unfolding do not correlate perfectly.

As demonstrated by the groups of Kelly and Gruebele, the free energy landscape of WW domains can be modified from three-state to two-state[67] and even downhill folding[68] by means of temperature, single amino acid mutations and small truncations at the C- and N-termini.

The position-dependent strengths of the backbone-backbone HBonds in case of the 34-residue Pin1-WW domain has been studied by means of so called amide-to-ester (A-to-E) mutations[64]. A-to-E mutations maintain the structure of the side chains and preserve the stereochemistry and backbone dihedral angles of the residue. Through substitution of an α-amino acid residue by the corresponding α-hydroxy acid residue, a particular HBond can be perturbed, either by elimination of the HBond donor (replacing an amide NH-group with an ester oxygen) or by weakening the HBond acceptor (replacing an amide carbonyl with an ester carbonyl)[69]. The impact of the perturbed backbone HBonds on protein stability was assessed by denaturation of the corresponding mutant (thermally and chaotropically) and subsequent comparison with the protein wildtype[64]. A significant dependence of the HBond strength on the microenvironment along the backbone was observed with variations up to several kJ mol^{-1}. In accordance with expectations, buried HBond donors showed the most significant destabilization effects, while the elimination of solvent-exposed HBonds was much less influential.

1.3 Impact of Co-Solvents

The sensitive folding/unfolding equilibrium of many proteins can be altered through addition of low molecular weight organic molecules, denoted as co-solutes or co-solvents. Co-solutes that shift the equilibrium towards the unfolded ensemble such as urea or guanidine hydrochloride are termed denaturants[70]. Thermodynamic experiments based on differential scanning calorimetry (DSC) and steady-state fluorescence show that increasing denaturant concentration leads to decreasing transition temperatures and unfolding free energies[39,70]. Experimental approaches have usually made assumptions on the dependence of the unfolding free energy on the denaturant concentration. For typical proteins, a linear dependence is observed, at least within a finite region around the transition[39,70]:

$$\Delta G_{F\to U}([C]) = \Delta G_{F\to U}(0) + m[C] \tag{1.3}$$

The slope m (denoted as the "m-value") measures the response of protein stability to the addition of co-solvent and is negative (positive) in a case of a denaturant (osmolyte), $[C]$ denotes the co-solvent concentration and the offset $\Delta G_{F\to U}(0)$ represents the estimated unfolding free energy extrapolated to zero co-solvent concentration. In case of classical two-state folding proteins such as barnase[43], the value $\Delta G_{F\to U}(0)$ is in very good agreement with the corresponding estimate obtained from thermal unfolding. It should be noted that alternative analysis procedures were proposed which do not require the assumption of a linear dependence of $\Delta G_{F\to U}$ on $[C]$[71].

Co-solutes that shift the equilibrium towards the folded state such as **trimethylamine N-oxide (TMAO)** are termed protecting osmolytes or protectants[70]. It is known that proteins denature at high pressure ($>$ 500 bar)[72]. Accumulation of TMAO in cells of deepsea fish enables their survival under such extreme conditions, by hindering the denaturing effect of high hydrostatic pressure[73]. In protein experiments *in vitro*, TMAO is a routinely used stabilizer due to its capability of counteracting destabilizing influences from high concentrations of salt and urea[74]. In thermal denaturation experiments it was shown that TMAO can have a stabilizing as well as destabilizing effect, depending on the pH considered with respect to the pKa value of TMAO[75].

Despite the usage of osmolytes such as TMAO since decades, the molecular-level mechanism is still not fully understood. In principle, one can think of two possibilities of how the stabilization is established, either by effectively stabilizing the folded state or by destabilizing the unfolded state. However, it is the detailed interplay between co-solvent, water and protein which is particularly controversial. Based on recent studies, the stabilization of the folded state results from preferential osmolyte exclusion or, equivalently, preferential hydration of the protein[70]. Other theoretical and experimental studies propose a direct mechanism between the osmolyte molecules and the polymer backbone[76,77].

1.4 Protein-Ligand Binding

The biological function of most proteins involves the recognition and binding of a substrate or ligand (in this context also called "guest"). As in case of protein folding, the binding of a ligand to a protein receptor (in this context also called "host") involves the sum of many weak non-bonded interactions such as HBonds or dispersion interactions between hydrophobic parts of the binding partners[78]. Further similarities include (i) the comparable value range of folding and binding free energies and (ii) the observed phenomenon of enthalpy-entropy compensation[79,80]. As mentioned previously, the accurate prediction of binding affinities is of great interest for the pharmaceutical industry in the early stage of rational drug design. The calculation of binding free energies is an active field of research for molecular simulation applications which led to the development of a variety of methods[81,82] and optimized force fields[23]. A sound statistical-mechanical basis for the treatment of non-covalent binding via molecular simulations has been elaborated[83,84]. Chapter 5 outlines some of this background which is relevant for the studies conducted in chapter 4. Compared to computationally faster approaches such as docking methods, molecular simulations based on all-atomistic force fields enable the treatment of protein and/or ligand flexibility and the explicit inclusion of the solvent[85]. The latter is of great importance to account for the entropy change upon binding[86]. Regular blind challenges such as "Modeling of Proteins and Ligands" (SAMPL), "Community Structure Activity Resource" (CSAR) or "Drug Design Data Resource" (D3R) are good opportunities to assess the current predictive performance of modeling approaches and to reveal their weaknesses[87].

The accurate prediction of standard binding free energies ($\Delta G^{\circ}_{\text{bind}}$) for realistic host-guest systems can pose various challenges for the quality of the force field as well as the applied sampling protocol[85,87]: (i) large scale conformational changes of ligand and/or protein, (ii) rearrangement of sidechains in the binding site, (iii) the choice of suitable protonation states, (iv) multiple binding orientations, (v) sensitivity towards buffer composition (e.g. pH, salt concentration), (vi) binding-site hydration, (vii) definition of the bound state and (viii) treatment of charged ligands. To alleviate these problems, the study of rather simple host-guest complexes such as the family of **cyclodextrins (CDs)** have come into the focus of computational scientists for testing binding free energy calculations methods. CDs are conically shaped oligosaccharides build from linked glucose units, featuring a hydrophilic outer surface and a hydrophobic inner cavity (see Fig. 4.2 (b)), which enables the binding of a diverse set of ligands. Binding to the hydrophobic cavity leads to the simultaneous desolvation of the CD interior and the ligand by stripping off its hydration shell. Thermodynamic studies such as isothermal titration calorimetry (ITC) measurements show that the binding process can be dominated by a change in enthalpy as well as entropy, depending on the involved CD and ligand species[87,88]. The availability of many binding data for various ligands from computational and experimental[89] studies (including different techniques and thermodynamic conditions) and conducted force field comparisons[90,91], makes CDs suitable benchmark systems for (computational) host-guest binding studies[87]. However, despite these advantages and their apparent simplicity, most of the challenges for predicting accurate binding free energies are also present for simulations of CDs, though somewhat less extensive. Due to the two non-equivalent rims of the cavity, asymmetric ligands can be bound in different binding orientations. To obtain binding affinities in case of bulky or elongated ligands (in comparison to the size of the cavity), it has to be taken into account that the sampling of the interchange between the two possible ligand orientations inside the cavity is a very rare event. As shown in chapter 4, this can lead to computational artifacts when physical pathway methods such as umbrella sampling[92] are applied to estimate $\Delta G^{\circ}_{\text{bind}}$.

1.5 Purpose of this Work

The purpose of this thesis can be viewed from two different directions. From an overall biophysical perspective, all topics treated herein touch an aspect which is related to protein thermodynamics:
(i) identification of a molecular model for the protecting osmolyte TMAO (chapter 2),
(ii) study of intramolecular HBonding via free-energy MD simulations (chapter 3),
(iii) computation of binding free energies for host-guest complexes (chapter 4).
All aspects are subjects of active research as outlined above and aim to contribute to the extension of our current understanding of protein stability.
From a methodological point of view, the thesis aims to disentangle the so called **force-field problem** and the **sampling problem**[93] in different contexts and applications of biomolecular simulations beyond pure toy models. In particular, for free energy calculations, where the target free energy difference typically results from the difference of two opposite and significantly larger values, this disentanglement can be a special challenge. In the theoretical case of infinitely long sampling, the estimated free energy difference is only dictated by the underlying force field and the system specifications, assuming a correctly implemented method and estimator[87,93]. However, in reality, where only finite sampling is possible, the obtained estimates will deviate from the theoretical value associated with infinite sampling. A meaningful assessment of the force field quality (and therefore comparison with experiments) can only be carried out, if the calculated estimates are sufficiently converged, free of artifacts and independent of the initial conditions. To achieve this goal, enhanced sampling methods are required, and molecular model or benchmark systems have to be established which are computationally tractable such that sampling issues are minimized[87]. The selection of suitable model systems is a critical aspect, since they should still be realistic enough to allow for the transfer of the results to more complex systems. In the following a brief introduction into the three topics is provided and the main research questions are highlighted.

1.5.1 TMAO Force Field Comparison

Elucidation of a molecular-level interaction mechanism of TMAO is still an active field of research[70,94], and in principle, a well-suited problem for molecular simulation studies. Over the recent years, several TMAO force fields have been proposed such as the models of Garcia[95], Netz[96], Shea[97] and Kast[98,99]. Each model was optimized with respect to particular target properties such as partial molar volumes, activity coefficients or osmotic coefficients. However, a force field comparison to identify the most suitable candidate(s) with respect to research question addressed above was missing. Chapter 2 is concerned with such a force field comparison which was conducted on the basis of aqueous TMAO solutions. Care was taken that the range of studied properties probe the balance of the co-solute/co-solute and co-solute/water interactions.

1.5.2 Calculation of Relative Folding Free Energies

Based upon many thermodynamic and kinetic experiments and supported by theoretical approaches, it was shown that the Pin1-WW domain, a small β-sheet protein, is mainly stabilized by intramolecular HBonding. So called amide-to-ester (A-to-E) mutations which are special kind of backbone mutations allow the perturbation of particular intramolecular HBonds. As shown previously, the impact of A-to-E mutations can also be obtained via free energy MD simulations[100]. A major drawback was that some mutations showed a considerable dependence on the used protein starting structure. Chapter 4 represents a follow-up study to this work. Through identification of the origins of the observed starting structure dependence and its alleviation through application of Hamiltonian replica exchange, a more meaningful and less ambiguous comparison with experimental data was possible.

1.5.3 Calculation of 1D-Potentials of Mean Force

Estimation of binding free energies is an important application and active field of research for biomolecular simulations. Over the years, several simulation protocols have been proposed based on solid thermodynamic and statistical-mechanical considerations. Umbrella sampling, which relies on estimating the potential of mean force (PMF) has been proven to be one of the most robust methods, which is why it is so popular and regularly used for validating new free energy techniques. An attempt to estimate binding free energies for complexes of primary alcohols and cyclodextrins from one-dimensional umbrella sampling, revealed artificial offsets between the bulk water regions in the estimated PMF. The occurrence of such PMF offsets has been observed in studies of more complex systems such as solute permeation through protein channels. However, usually they are only marginally discussed if at all. As shown in chapter 4, I could demonstrate that such artifacts may easily occur for much simpler systems. By systematic studies using host-guest systems of increasing complexity, different origins for the occurrence of these artifacts could be identified. The consequence of these offsets with respect to the estimation of binding free energies together with their prevention is discussed.

Chapter 2

Validation of Trimethylamine-N-oxide (TMAO) Force Fields Based on Thermophysical Properties of Aqueous TMAO Solutions

The content of this chapter is a literal quote of the publication

D. Markthaler, J. Zeman, J. Baz, J. Smiatek and N. Hansen, *The Journal of Physical Chemistry B*, **121**, 10674-10688, 2017.

The manuscript was written by Daniel Markthaler. All molecular simulations were conducted by Daniel Markthaler, except those used for calculating the dielectric spectra, which were conducted by Johannes Zeman. Jörg Baz performed preliminary simulations for the study and helped with the analysis of the data. Jens Smiatek and Niels Hansen supervised the project and manuscript writing.

Abstract

Five molecular models for trimethylamine N-oxide (TMAO) to be used in conjunction with compatible models for liquid water are evaluated by comparison of molecular dynamics (MD) simulation results to experimental data as functions of TMAO molality. The experimental data comprise thermodynamic properties (density, apparent molar volume and partial molar volume at infinite dilution), transport properties (self-diffusion

and shear viscosity), structural properties (radial distribution functions and degree of hydrogen bonding) and dielectric properties (dielectric spectra and static permittivity). The thermodynamic and transport properties turned out to be useful in TMAO model discrimination while the influence of the water model and the TMAO-water interaction are effectively probed through the calculation of dielectric spectra.

2.1 Introduction

Understanding the driving forces for protein folding and unfolding in aqueous solution is one of the major challenges in computational biochemistry. Many proteins in aqueous solution are marginally stable and the folding/unfolding equilibrium can easily be altered by the addition of small organic molecules known as co-solutes[101,102]. Co-solutes that shift the equilibrium toward the unfolded ensemble are termed denaturants, whereas those that favor the folded ensemble are known as protecting osmolytes.

Protecting osmolytes such as trimethylamine N-oxide (TMAO), glycerol and sugars that push the equilibrium toward the folded ensemble play a crucial role in maintaining the function of intra-cellular proteins in extreme environmental conditions. Since TMAO counteracts destabilizing effects from high salt concentrations, denaturants like urea, or elevated temperature and pressure[103], it is routinely used as a stabilizer in protein mutation experiments. However, a clear picture of the stabilization mechanism of TMAO has not yet been established[70,75]. Recent studies have featured the strong interactions of TMAO with water such that the stabilization of the folded state is a consequence of preferential osmolyte exclusion or, equivalently, preferential hydration of the protein[70]. Other theoretical and experimental studies suggest that osmolyte-induced stabilization of the folded state of polymers and proteins may involve a direct mechanism in which the osmolyte molecules interact with the macromolecules[76,77]. Recently, an additional mechanism was proposed, denoted as preferential attraction[104,105], which suggests a preferential accumulation of TMAO around folded state conformations in the absence of direct binding accompanied by preferential hydration of the solute mediated through TMAO[106].

During the past years, a number of TMAO models have been proposed for use in atomistic simulation of aqueous systems[95–99,107], followed by various studies comparing subsets of these models in terms of their impact on amino acids, polypeptides, proteins or hydrophobic polymers in solution[104,108–117]. The conclusions drawn from these studies about the mechanism of TMAO action remain controversial. The diversity of the existing TMAO models exemplifies how strong the parameters of pairwise additive fixed-charge force fields depend on the target properties used in the parametrization process. Although it is clear that the number of properties that can be simultaneously represented within a given model class is limited[118,119], there are indications that within the parameter space of pairwise additive fixed-charge force fields optimization is still possible[20,21]. Only if the resulting models are not sufficiently accurate, additional complexity such as accounting for polarizability may help to extend the scope of classical molecular models[120]. A comparison of models at a suboptimal level of parametrization may lead to the introduction of additional complexity which is not required[121–123].

Here, we analyse thermodynamic, structural, transport, and dielectric properties of aqueous TMAO solutions. A faithful representation of thermodynamic properties is important, as these constitute basic driving forces in a system. The effect of increasing TMAO concentrations however does not only influence the thermodynamics of the solution, manifested in altered chemical activities[124] for instance, but also dielectric properties[108,125]. This influence originates from strong interactions of TMAO with its surrounding water layer and illustrates that a proper modelling of the dielectric properties is also essential. Finally, from a physical perspective, structure follows energy. When calibrating models, ensuring the energetics are correct is crucial to maximize the applicability of a model. Transport properties may be quite sensitive to force field parameters[126] and are thus of further use in model discrimination. Regardless of the possible direct or indirect nature of the stabilizing mechanism of TMAO, we consider the accurate representation of these basic thermophysical properties of aqueous TMAO solutions over a broad range of compositions to be essential for atomistic studies of macromolecules in mixed solvent environments, including the correct description of the temperature dependence because these reflect the thermodynamic signature of the TMAO-water interactions[127]. The accuracy with which the molecular model mimics these interactions has an impact on the accumulation behavior of TMAO around proteins and thus on the thermodynamic equilibrium between different protein conformations observed in the simulation. These effects can be quantified using the potential distribution theorem or the Kirkwood-Buff theory[128–131]. If an atom-based co-solute model has been validated against independent data and turns out to reproduce in addition the macroscopic effect of interest, then it can tentatively be used to interpret this effect in physical terms[132–137]. Our work thus represents a valuable supplement to the recently performed TMAO force-field comparison reported by Rodríguez-Ropero et al.[115], which was mainly focused on the impact of TMAO on the folding equilibrium of a hydrophobic model polymer, but also investigated quantities of aqueous solutions such as surface tension, osmotic coefficient as well as the transfer free energy of neopentane from pure water to 1 molar TMAO solution.

Here, we present a comparative study of four different TMAO models reported in the literature[95–97,99] and one whose functional form is compatible with the GROMOS biomolecular force field[13,138–147], with parameters assigned automatically. A diverse set of thermophysical properties of aqueous TMAO solutions was selected in order to be able to sort out the strengths and weaknesses of the various parametrizations for TMAO.

2.2 Theory

2.2.1 Partial and Apparent Molar Volume

The total solution volume V is given as the sum of the partial molar volumes $\bar{V}_i = (\partial V/\partial n_i)_{p,T,n_{j\neq i}}$ of its compounds, weighted with the corresponding mole numbers n_i[148]. For a binary mixture of water (w) and a solute (s) it follows,

$$V = n_\mathrm{w}\bar{V}_\mathrm{w} + n_\mathrm{s}\bar{V}_\mathrm{s} \tag{2.1}$$

The definition of the solute's apparent molar volume ${}^{\phi}\bar{V}_\mathrm{s}$ follows from reformulating Eq. (2.1) in terms of the molar volume of pure water $\bar{V}_{\mathrm{w},0}$:

$$V = n_\mathrm{w}\bar{V}_{\mathrm{w},0} + n_\mathrm{s}^{\phi}\bar{V}_\mathrm{s} \tag{2.2}$$

In contrast to $\bar{V}_{\mathrm{w},0}$ which is a function of pressure and temperature only, ${}^{\phi}\bar{V}_\mathrm{s}$ depends in addition on the composition of the solution. The main difference of Eq. (2.2) compared to Eq. (2.1) is that the whole non-ideality of the mixing behavior is assigned to the solute, which makes ${}^{\phi}\bar{V}_\mathrm{s}$ an apparent quantity. The connection between the solute's apparent molar volume ${}^{\phi}\bar{V}_\mathrm{s}$ and its partial molar volume $\bar{V}_\mathrm{s}$ can be obtained in a straightforward manner,

$$\bar{V}_\mathrm{s} = {}^{\phi}\bar{V}_\mathrm{s} + m_\mathrm{s}\left(\frac{\partial^{\phi}\bar{V}_\mathrm{s}}{\partial m_\mathrm{s}}\right)_{p,T,n_\mathrm{w}} \tag{2.3}$$

where the molality m_s, is defined as the number of moles of solute per kg of solvent. For calculating the solute's apparent molar volume via the solution density ρ, Eq. (2.2) can be rewritten as

$${}^{\phi}\bar{V}_\mathrm{s} = \frac{M_\mathrm{s}}{\rho} + \frac{1}{m_\mathrm{s}}\left(\frac{1}{\rho} - \frac{1}{\rho_{\mathrm{w},0}}\right) \tag{2.4}$$

where M_s is the molar mass of the solute and $\rho_{\mathrm{w},0}$ the density of pure water at the same pressure and temperature. In the limit of infinite dilution, the apparent molar volume is equal to the partial molar volume, i.e.

$$\lim_{m_\mathrm{s}\to 0} {}^{\phi}\bar{V}_\mathrm{s} = \bar{V}_\mathrm{s} \tag{2.5}$$

2.2.2 Dielectric Properties

Dielectric spectra were computed following a fluctuation-based ("Green-Kubo") approach[149–152] described in the following. The complex frequency-dependent permittivity $\varepsilon_\mathrm{r}(\omega)$ is defined as

$$\varepsilon_\mathrm{r}(\omega) := \varepsilon_\mathrm{r}'(\omega) - i\varepsilon_\mathrm{r}''(\omega) \tag{2.6}$$

with its real part $\varepsilon_\mathrm{r}'(\omega)$, usually referred to as dielectric dispersion, and its negative imaginary part $\varepsilon_\mathrm{r}''(\omega)$, known as dielectric absorption or loss. The permittivity spectrum is calculated from that of the complex total conductivity $\sigma(\omega)$ as

$$\varepsilon_\mathrm{r}(\omega) = \varepsilon_\infty + \frac{i}{\varepsilon_0 \omega} \left(\sigma(\omega) - \sigma(0)\right) \tag{2.7}$$

where i is the imaginary number, ε_0 the permittivity of free space, and ω denotes the angular frequency of a hypothetically applied external electric field. The constant ε_∞ denotes the dielectric permittivity of the system for $\omega \to \infty$. Due to the fact that the molecular models investigated in this work do not incorporate explicit electronic polarizability and because of electrostatic tinfoil boundary conditions, this constant is equal to unity in our calculations. In order to facilitate the qualitative comparison with experimental data, we define the *reduced* permittivity $\bar{\varepsilon}_\mathrm{r}(\omega)$ as

$$\bar{\varepsilon}_\mathrm{r}(\omega) := \varepsilon_\mathrm{r}(\omega) - \varepsilon_\infty \tag{2.8}$$

The complex frequency-dependent total conductivity $\sigma(\omega)$ required to evaluate Eq. (2.7) is determined according to

$$\begin{aligned} \sigma(\omega) &= \frac{1}{3Vk_\mathrm{B}T} \int_0^\infty \langle \boldsymbol{J}_\mathrm{tot}(0) \boldsymbol{J}_\mathrm{tot}(t) \rangle \, e^{i\omega t} dt \\ &:= \frac{1}{3Vk_\mathrm{B}T} \langle \boldsymbol{J}_\mathrm{tot}(0) \boldsymbol{J}_\mathrm{tot}(t) \rangle_\omega \end{aligned} \tag{2.9}$$

where V represents the average volume of the simulation box, k_B the Boltzmann constant, and T the temperature of the system. The operator $\langle \cdot \rangle$ denotes the canonical average, $\langle \cdot \rangle_\omega$ its Fourier-Laplace transform, and $\boldsymbol{J}_\mathrm{tot}(t)$ is the fluctuating cumulative current

$$\boldsymbol{J}_\mathrm{tot}(t) = \sum_n q_n \boldsymbol{v}_n(t) \tag{2.10}$$

with the summation index n running over all atoms in the investigated system with partial charges q_n and corresponding velocities $\boldsymbol{v}_n(t)$. Note that our definition of the conductivity slightly differs from those given in Refs. 150,152 where only molecular net charges, i.e., translational contributions, are taken into account. However, the complex conductivity of a system in an electric field alternating at a finite, non-zero frequency also contains contributions from the reorientation of molecules in response to the field. In our definition, such rotational contributions are explicitly included. To stress this fact and to discriminate between the different definitions, we use the term *total* conductivity here.

Since the tails of the current autocorrelations $\langle \boldsymbol{J}_\mathrm{tot}(0) \boldsymbol{J}_\mathrm{tot}(t) \rangle$ generally suffer from statistical noise, we apply an integral-preserving noise reduction technique to the raw autocor-

relation data prior to computing the Fourier-Laplace transform. This procedure allows us to resolve the spectra with high precision on the entire frequency range accessible to the simulation without loss of spectral features and has already been successfully employed in a previous publication[151]. For an in-depth discussion of the technique, we refer the reader to the Supporting Information of this article.
Moreover, we employed the so-called "Einstein-Helfand" method[149,152] to directly evaluate the static dielectric permittivity $\varepsilon_{\text{static}} := \varepsilon_{\text{r}}'(\omega = 0)$. A detailed description of the method is given in the Supporting Information and can also be found in the literature[149,152].

2.3 Computational Details

2.3.1 TMAO Force Fields

Four of the five considered TMAO force fields, denoted according to their last authors as Garcia[95], Netz[96], Shea[97], and Kast 2016[99], originate from the same model developed by Kast et al.[98] in 2003 (Kast 2003). The parameters of the Kast 2003 force field were mainly derived from quantum chemical ab initio calculations to represent experimental crystal data and tested against measured densities of aqueous mixtures. While all four variants use the same intramolecular interactions from Kast 2003, except for the 1-3 Urey-Bradley terms, various authors optimized the nonbonded parameters with the purpose to better reproduce certain physical properties or other thermodynamic state points. It should be noted that alongside with the different nonbonded interactions the force fields differ in the used water models as well as the simulation parameters. The simulations conducted in the present work follow the original publications in terms of the employed water models in combination with long-range corrections (see Tab. 2.1). As a fifth candidate, an automatically parametrized[153] united atom (UA) model compatible with the GROMOS 54A7 force field[13] was considered without further refinement. A rough overview about the main features of the individual models will be presented in the following and in Table 2.1. We note that the dipole moments of the Shea and UA models are closer to the values TMAO exhibits in vacuum, while the dipole moments of the other models are closer to the values of TMAO in aqueous solution calculated from ab-initio MD simulation[99].

Garcia

Garcia and coworkers[95] optimized the Kast 2003 model with respect to the reproduction of osmotic pressures at various TMAO concentrations obtained from MD simulations with an applied restraining potential acting only on TMAO to imitate a semipermeable membrane. Their route was to scale all charges by a factor of 1.2 to make TMAO more hydrophilic and to use a modified combination rule for the cross-interactions between TMAO molecules to weaken the dispersion interactions. It was found that the Garcia force field is able to capture the experimentally measured preferential interaction of TMAO with proteins at least qualitatively. The model was parametrized using the TIP3P water model[154,155].

Netz

Netz and coworkers[96] optimized the Kast 2003 model with respect to the bulk activity coefficient derivative for TMAO and polyglycine m-values calculated via the Kirkwood-Buff theory. In a parametric study, the TMAO dipole strength was varied through the oxygen's partial charge q_O (compensated via q_N) together with the size of the hydrophobic groups by scaling the Lennard-Jones radii of carbon (C) and hydrogen (H) atoms. The optimized model, which shows enhanced dipole strength as well as hydrophobicity, yields results for the concentration dependence of the solution density closer to the experimental values as compared to Kast 2003. For the parametrization, the SPC/E water model[156] and the Lorentz-Berthelot combination rules were used.

Shea

Using an iterative bisection method, Larini and Shea[97] optimized only the Lennard-Jones parameters of Kast 2003 to reproduce experimental density data of aqueous mixtures at different concentrations and temperatures. They derived optimized parameter sets for SPC/E, TIP3P and TIP4P water[155]. To ensure compatibility with the OPLS-AA force field, pure geometric combination rules for the cross-interactions were chosen. A dynamic analysis revealed that the Shea model shows a slow-down of the water diffusion for increasing TMAO concentration, which is also found experimentally[157]. Here, we used the TMAO model compatible with SPC/E water.

Kast 2016

Kast and coworkers[99] slightly altered the Netz model to better reproduce the solution density at ambient conditions. They also proposed variants for simulation at elevated pressure. By combining quantum chemical calculations with liquid state integral equations theory for the pressure response of TMAO and water, respectively, it was possible to capture the whole pressure dependence by scaling the partial charges without the need of a complete reparametrization. It was found that strong compression of the solution leads to an increase in the TMAO dipole moment, which is manifested in a modified hydrogen bond pattern. The model uses the Lorentz-Berthelot combination rules for calculating cross-interactions. Simulations were conducted with the TIP4P/2005 water model[158], which was used for the parametrization, as well as with SPC/E water because the latter combination shows remarkable quantitative agreement for various properties studied in the present work. In contrast to the original work[99], the present simulations using the Kast 2016 model employed 1-4 interaction scaling compatible with the AMBER/GAFF force field definition[159]. However, the difference between properties calculated with and without 1-4 scaling was found to be negligible.

United atom (UA) model

The fifth molecular model considered relies on a UA representation of the CH_3 groups and a Lennard-Jones representation of the van der Waals interactions that is compatible with the GROMOS biomolecular force field[13,138–147], i.e. application of a geometric-mean combination rule[160,161] for pairwise Lennard-Jones interaction parameters, distinguishing between non-hydrogen-bonding, uncharged hydrogen-bonding, and oppositely charged interactions without the need to introduce additional atom types. In that sense, the UA representation of the methyl groups does not result in a reduction of the number of nonbonded parameters compared to an all-atom representation of TMAO. Atomic partial charges are treated as freely adjustable. In this work they were obtained from the Automated Topology Builder (ATB)[153], which also assigned Lennard-Jones atom types 16 (CH3), 8 (N) and 2 (O), corresponding to the GROMOS 54A7 force field[13]. Equilibrium bond lengths and bond angles along with the corresponding force constants were obtained from the quantum mechanical structure optimization underlying the ATB workflow[153] and are given in Table 1. An optimization of nonbonded parameters for the UA model against experimental data was not pursued. The results we obtain highlight the need of a further refinement for automatically parametrized models. The simulations were conducted with the SPC[162] and SPC/E water models.

2.3.2 Simulated Systems

MD simulations of the binary mixture TMAO-water were performed at a constant pressure of 1 bar and three different temperatures (278.15, 298.15, 323.15 K) to model experimental conditions. Nine different compositions as specified in Tab. 2.2 (denoted as C0-C8) were considered including pure water simulations. Tab. 2.2 also contains the five compositions considered in case of the dielectric spectra calculations (denoted as D0-D4). Though it is known that the solubility of TMAO in water strongly depends on temperature, there seems to be some discrepancy in the literature about the solubility limits. In Ref. 163 saturation molalities of 7.28 and 18.39 $\mathrm{mol\,kg^{-1}}$ were reported for 297.95 K and 336.15 K respectively, i.e. below the concentration C8 (and D4) for the two lower temperatures considered in this work. In contrast to this finding other groups reported monophasic aqueous solutions at 298 K of 10 $\mathrm{mol\,kg^{-1}}$ [125,164] . Despite this uncertainty the composition C8 was studied in the present work for the purpose of comparison to earlier simulation studies[97,98].

Table 2.1: Parameters specifying the nonbonded interactions in different TMAO models along with the methods to enforce constant temperature and pressure [a].

Model	Atom	σ/ nm	ϵ/ kJ mol^{-1}	q/e	r_c / nm	comb. rule	1-4 scaling	long-range corr.	water model	electrostatics[b]	T-coupling[c]	p-coupling[d]
Garcia [95]	C	0.3041	0.2826	-0.312	1.3	$\epsilon = (1 - k_{ij})(\epsilon_i \epsilon_j)^{0.5}$ [e] $\sigma = 0.5(\sigma_i + \sigma_j)$	1.0 for LJ 1.0 for Coulomb	energy and pressure	TIP3P	PME (1.3)	NH (0.1)	PR (1.0)
	H	0.1775	0.0773	0.132								
	O	0.3266	0.6379	-0.780								
	N	0.2926	0.8369	0.528								
Netz [96]	C	0.3600	0.2826	-0.260	1.2 [f]	$\epsilon = (\epsilon_i \epsilon_j)^{0.5}$ $\sigma = 0.5(\sigma_i + \sigma_j)$	0.5 for LJ 0.83 for Coulomb	none	SPC/E	PME (1.2)	NH (0.1)	PR (1.0)
	H	0.2101	0.0773	0.110								
	O	0.3266	0.6379	-0.910								
	N	0.2926	0.8360	0.700								
Shea [97]	C	0.3385	0.2800	-0.260	1.2	$\epsilon = (\epsilon_i \epsilon_j)^{0.5}$ $\sigma = (\sigma_i \sigma_j)^{0.5}$	0.5 for LJ 0.5 for Coulomb	energy and pressure	SPC/E [g]	PME (1.2)	NH (0.1)	PR (1.0)
	H	0.2319	0.0660	0.110								
	O	0.3109	0.6040	-0.650								
	N	0.3205	0.7430	0.440								
Kast 2016 [99]	C	0.3707	0.2830	-0.260	1.0	$\epsilon = (\epsilon_i \epsilon_j)^{0.5}$ $\sigma = 0.5(\sigma_i + \sigma_j)$	0.5 for LJ[h] 0.83 for Coulomb	energy and pressure	TIP4P/2005, SPC/E	PME (1.0)	NH (1.0)	PR (2.0)
	H	0.2130	0.0775	0.110								
	O	0.3266	0.6389	-0.815								
	N	0.2926	0.8374	0.605								
UA [i]	CH3	0.3748	0.8672	0.088	1.4	$\epsilon = (\epsilon_i \epsilon_j)^{0.5}$ $\sigma = (\sigma_i \sigma_j)^{0.5}$	N/A	none	SPC[j], SPC/E	RF (1.4)[j]	WC (0.1)[j]	WC (0.5)[j]
	O	0.2626	1.7250	-0.702								
	N	0.3136	0.6398	0.438								

[a] The parameters are the Lennard-Jones size and energy parameters, σ and ϵ, the partial charges, q, and the cut-off radius for the Lennard-Jones potential, r_c. The other columns specify the combination rules for nonbonded interactions, the scaling factors for 1-4 interactions, the type of long-range corrections used, the water model employed and the method to treat electrostatic interactions. The values of σ and ϵ of the various models may differ in the third digit across various publications, possibly a consequence of conversion between different units.
The average dipole moments (in Debye) calculated from the C1 simulations at 298.15 K simulations are 6.7, 7.3, 5.5, 6.8 and 5.3 for the Garcia, Netz, Shea, Kast 2016 and UA model respectively.

[b] The value in parentheses specifies the real-space cut-off in nm for the particle-mesh Ewald (PME) summation or the long-range cut-off for the reaction field (RF) scheme, respectively.

[c] The value in parentheses specifies the coupling constant in ps for the Nosé-Hoover (NH) or weak coupling (WC) method, respectively.

[d] The value in parentheses specifies the coupling constant in ps for the Parrinello-Rahman (PR) or the weak coupling (WC) method, respectively.

[e] $k_{ij} = 0.25$ for TMAO-TMAO interactions and zero otherwise.

[f] In the original work, $r_c = 1.0$ nm was used. Here we employed the same cut-off radius for the Lennard-Jones and real-space electrostatic interactions as recommended for the Verlet-buffered neighbor list in recent versions of GROMACS. The larger of the two values (1.0 vs. 1.2 nm) was used to achieve better agreement with experimental densities for the pure water simulations (C0).

[g] Reference 97 also reports simulations using the TIP3P and TIP4P water models, for which slightly different Lennard-Jones parameters are required for TMAO.

[h] In the original version of the model no 1-4 scaling was employed. As discussed in the main text the effect of 1-4 scaling was found to be negligible.

[i] These parameters only hold for self-interactions and for nonbonded interactions with CH_3. The other cross-terms are: $\sigma_{\mathrm{O,OW}} = 0.3272$ nm and $\epsilon_{\mathrm{O,OW}} = 0.4954$ kJ mol^{-1} for the interaction of the TMAO oxygen with the water oxygen, $\sigma_{\mathrm{N,OW}} = 0.3541$ nm and $\epsilon_{\mathrm{N,OW}} = 0.3202$ kJ mol^{-1} for the interaction of the TMAO nitrogen with the water oxygen and $\sigma_{\mathrm{O,N}} = 0.3986$ nm and $\epsilon_{\mathrm{O,N}} = 0.1464$ kJ mol^{-1} for the interaction of the TMAO oxygen with the TMAO nitrogen. This is a consequence of the differentiation of C12-parameters between non-hydrogen-bonding, uncharged hydrogen-bonding, and oppositely charged interactions in the GROMOS force field. Bond lengths are 0.15 nm and 0.138 nm for the N-CH3 and N-O bonds, respectively. All bond angles have an equilibrium value of 109° and a force constant of 1681 kJ mol^{-1} based on a potential energy function harmonic in the angle cosine.

[j] These are the standard settings in the GROMOS force field, referred to as setup RF in this work. Additional simulations were performed using PME electrostatics, long-range corrections for energy and pressure and the NH (0.1) and PR (1.0) methods to control temperature and pressure.

Table 2.2: System compositions studied in this work[a].

System	N_{TMAO}	$N_{\text{H}_2\text{O}}$	molality m [mol kg^{-1}]	mole fraction
C0	0	1000	0.0000	0.0000
C1	1	1000	0.0555	0.0010
C2	20	4508	0.2463	0.0044
C3	20	1895	0.5858	0.0104
C4	35	1000	1.9428	0.0338
C5	100	1500	3.7005	0.0625
C6	78	1000	4.3333	0.0724
C7	100	1000	5.5508	0.0909
C8	140	700	11.1016	0.1667
D0	0	2180	0.0000	0.0000
D1	100	2775	2.0003	0.0348
D2	100	1388	3.9991	0.0672
D3	100	694	7.9983	0.1259
D4	100	555	10.0015	0.1527

[a] Systems D0–D4 were exclusively used for the calculation of dielectric spectra. Simulation times depend on the measured observables and are given in the corresponding sections and in tables A4 and A7 of the Supporting Information.

2.3.3 Simulation Parameters

All simulations were performed under minimum image periodic boundary conditions based on cubic computational boxes containing aqueous TMAO solutions with the compositions specified in Tab. 2.2. The equations of motion were integrated using the leap-frog scheme[165] with a timestep of 1 fs. For all compositions an energy minimization followed by a constant-volume equilibration simulation of 1 ns and a successive constant-pressure equilibration simulation of 2 ns were conducted prior to the actual production simulations. All production simulations were performed at constant temperature and pressure except for those used to calculate the shear viscosity, which were performed at constant volume. Most MD simulations were performed using the GROMACS 5.0.5 program package[166–169] compiled in double precision while some additional simulations using the UA model in combination with the Barker-Watts reaction field scheme to treat electrostatic interactions were performed with the GROMOS11 program package[170–173].

In the simulations employing the GROMACS program package all bond lengths were kept fixed at their equilibrium values using either SETTLE[174] (for water) or LINCS[175,176] (for TMAO), except for the Kast 2016 model for which only the length of bonds involving an H atom were kept fixed. The bond lengths constrained by application of the LINCS procedure are using a LINCS-order of 4. The number of iterations to correct for rotational lengthening in LINCS was set to 2. The temperature was maintained close to its reference value with the Nosé-Hoover thermostat[177–179] with a coupling constant specified in Tab. 2.1. The pressure was set to 1 bar with the Parrinello-Rahman barostat[180,181] by isotropic coupling with a coupling constant specified in Tab. 2.1 and an isothermal compressibility of $4.5 \times 10^{-10}\,\mathrm{Pa}^{-1}$. The time constants for temperature and pressure coupling using the Garcia, Netz or Shea force fields, respectively, were taken from the work of Larini and Shea[97]. We note that other authors used larger time constants in liquid phase simulations[182]. Therefore simulations using the Shea force field at C0 and C1 were conducted with $\tau_T = 1$ ps and $\tau_p = 5$ ps, showing negligible differences to the other set up. Short-range electrostatic and Lennard-Jones interactions were treated with a Verlet-buffered neighbor list[183], with potentials shifted to zero at the cut-off. Analytical dispersion corrections for energy and pressure were included for the cases specified in Tab. 2.1. Long-range electrostatics were treated by the smooth particle-mesh Ewald (PME) summation[184,185] with a real-space cut-off as specified in Tab. 2.1.

The simulations for the calculation of dielectric spectra were performed with a modified version of GROMACS 4.6.5 for all TMAO models with system compositions D0–D4 as specified in Tab. 2.2. These simulations were conducted at a constant temperature of 298.15 K and a pressure of 1 bar using a Nosé-Hoover thermostat and a Parrinello-Rahman barostat with coupling constants $\tau_T = 1$ ps and $\tau_p = 2$ ps, respectively. All other parameters correspond to those of the other GROMACS simulations employing PME electrostatics as given above and in Tab. 2.1. For each of the investigated TMAO models and concentrations, eight independently generated systems were equilibrated for 5 ns (5×10^6 steps, $\Delta t = 1$ fs) followed by production runs of approximately 268.4 ns (2^{28} steps), yielding a total simulation time of approximately 2.15 μs per TMAO model and concentration. Cumulative currents of all atoms in the respective systems were computed according to Eq. (2.10) at every femtosecond during runtime and written to disk for offline analysis. A similar approach was also used in Ref. 151.

The MD simulations using the GROMOS11 program package were carried out with release version 1.3.0[186]. All bond lengths as well as the water hydrogen-hydrogen distances were constrained by application of the SHAKE procedure[187] with a relative geometric tolerance of 10^{-4}. The center of mass translation of the computational box was removed every 2 ps. The temperature was maintained close to its reference temperature by weak coupling to an external bath[188] with a relaxation time of 0.1 ps. Distinct temperature baths were used for the translational and for the rotational/internal degrees of freedom of the molecules. The pressure was calculated using a group-based virial and held constant at 1 bar using the weak coupling method with a relaxation time of 0.5 ps[188] and an isothermal compressibility κ_T of $7.513 \times 10^{-4}\,(\mathrm{kJ\,mol^{-1}\,nm^{-3}})^{-1}$ for water[189], equivalent to the value used in the GROMACS simulations. The effect of decreasing the value of κ_T as appropriate[190] for higher TMAO concentrations was found to be negligible. Van der Waals and electrostatic interactions were handled using the Lennard-Jones potential[191–193] and the Barker-Watts reaction field scheme[194] within a triple-range cut-off approach[195] applied on the basis of distances between charge group centers[138], with short- and long-range cut-off radii of 0.8 and 1.4 nm, respectively. The short-range interactions were calculated every time step using a group-based pairlist updated every fifth time step. The intermediate-range interactions were re-evaluated at each pairlist update and assumed constant in between. The reaction field scheme was applied using a relative dielectric permittivity of 78.5 for the dielectric continuum surrounding the cut-off sphere, corresponding to the experimental value for pure water[189]. The reaction field self-term and excluded-atom-term contributions[196] to the energy, forces, and virial were included as described in Ref. 197.

2.3.4 Trajectory Analysis

Simulations were performed for 50 ns (C2 - C8) or, in case for which higher precision was needed, 400 ns (C0 and C1). For calculating the shear viscosity and the dielectric spectra longer simulations were carried out as specified in the corresponding paragraphs below and above, respectively. Configurations were stored every 2 ps. Except for the dielectric spectra, all analyses were conducted using the post processing tools provided by the GROMACS and GROMOS program packages, respectively.

Partial Molar Volume at Infinite Dilution and Apparent Molar Volume

Two alternative methods can be employed to evaluate the (aqueous) partial molar volume at infinite dilution based on atomistic simulations[198–201]. The first method relies on the difference in average volume between two aqueous systems involving the same number of water molecules, either in the absence or in the presence of one solute molecule. The second method relies on the calculation of the hydration free enthalpy of the solute along with variations of this free enthalpy corresponding to finite pressure changes. In the present work, we followed the first route. The apparent molar volume was obtained according to Eq. (2.4).

The performance of the force fields with respect to the description of infinitely diluted solutions was further investigated by considering the relation[202,203]

$$\left(\frac{\partial \bar{C}_{p,\mathrm{s}}^{\infty}}{\partial p}\right)_T = -T\left(\frac{\partial^2 \bar{V}_{\mathrm{s}}^{\infty}}{\partial T^2}\right)_p \tag{2.11}$$

where $\bar{C}_{p,\mathrm{s}}^{\infty}$ denotes the partial molar isobaric heat capacity of the solute (TMAO) at infinite dilution. For evaluating the left-hand side the partial molar enthalpies $\bar{H}_{\mathrm{s}}^{\infty}$ at infinite dilution were calculated at temperatures of 278.15, 288.15, 298.15, 308.15 and 323.15 K and pressures of 1, 2500 and 5000 bar as the difference in average enthalpy between the systems C1 and C0, simulated for 400 ns. At each pressure $\bar{C}_{p,\mathrm{s}}^{\infty}$ was obtained from the slope of a linear least-squares fit through the $\bar{H}_{\mathrm{s}}^{\infty}(T)$ data. The pressure derivative was subsequently evaluated as the slope of a linear least-squares fit through the $\bar{C}_{p,\mathrm{s}}^{\infty}(p)$ data.

Shear Viscosity

The shear viscosity η was calculated from the Green-Kubo expression[204,205]

$$\eta = \frac{V}{k_{\mathrm{B}}T}\int_0^{\infty}\langle P_{\alpha\beta}(t)P_{\alpha\beta}(0)\rangle \mathrm{d}t \tag{2.12}$$

with the off-diagonal pressure tensor components $P_{\alpha\beta}$ ($\alpha, \beta = x, y, z, \alpha \neq \beta$). Constant-volume simulations of 5 ns production (1 ns equilibration) were performed at 298.15 K using the average density obtained from a preceding constant-pressure simulation at 1 bar and the same temperature. The pressure tensor elements were written out every 5 fs to the energy trajectory. Due to the high level of statistical noise in the running integral of Eq. (2.12) it was averaged over a series of at least 100 independent simulations. A double-exponential function $\eta_{\mathrm{fit}}(t)$ was then fitted to the averaged time-dependent running integral $\langle\eta(t)\rangle$ with the four fitting parameters η_{∞}, α, τ_1 and τ_2,

$$\frac{\eta_{\mathrm{fit}}(t)}{\eta_{\infty}} = \frac{\alpha\tau_1(1-\mathrm{e}^{-t/\tau_1}) + (1-\alpha)\tau_2(1-\mathrm{e}^{-t/\tau_2})}{\alpha\tau_1 + (1-\alpha)\tau_2} \tag{2.13}$$

To damp the effect of rather noisy values for larger times, the residuals $\langle\eta(t_i)\rangle - \eta_{\mathrm{fit}}(t_i)$ entering the objective function were weighted according to $1/t_i^b$ as described by Maginn et al.[206] The exponent b is the result of a preceding power law fit to the time-dependent standard deviation $s(t)$. In addition, the first two picoseconds of the data were discarded in the fitting procedure. The parameter η_{∞}, which defines the stationary plateau value of the double-exponential was taken as the zero-shear rate viscosity η. As the increase of the standard deviation follows a power law in time, the definition of an error in η is not straightforward. Here, we take $s(t_{99})$ according to the time t_{99} where the monotonically increasing function $\eta_{\mathrm{fit}}(t)$ reaches 99% of the plateau value η_{∞} as a conservative error estimate.

Self-Diffusion Coefficients

The self-diffusion coefficient of molecular species i, $D_{\mathrm{self,i}}$, was determined from a constant-pressure simulation as the slope of a linear fit to the mean-square displacement of the molecules in the long-time limit using the Einstein relation[207,208]

$$\lim_{t\to\infty} \langle (\mathbf{r}_i(\tau + t) - \mathbf{r}_i(\tau))^2 \rangle_{i,\tau} = 6D_{\mathrm{self,i}}\, t + \mathrm{const.} \tag{2.14}$$

where $\mathbf{r}_i$ is the instantaneous molecular position (following molecules across periodic boundaries) and $\langle \ldots \rangle_{i,\tau}$ stands for averaging over all molecules i and time origins τ. In practice, a least-squares fitting over 10 ns trajectory fragments was performed to obtain a set of diffusion coefficients from which a mean value was calculated as well as the corresponding standard deviation, used as error estimate. The correlation coefficients R^2 were at least 0.99 in all cases. A correction for finite-size effects[209] was not conducted.

Hydrogen Bonds

Hydrogen bonds between TMAO and water were identified based on a geometric criterion. A hydrogen bond was assumed to exist if the hydrogen-acceptor distance is smaller than 0.25 nm and the donor-hydrogen-acceptor angle is larger than 135°. Nitrogen atoms were explicitly precluded as hydrogen bond acceptors.

Radial Distribution Functions

The radial distribution functions $g_{ij}(r)$ were calculated in the usual way as the probability of finding a particle of type j at distance r from a central particle i relative to the same probability for a random distribution of particles j around i[208],

$$g_{ij}(r) = \frac{\rho_{ij}(r)}{\langle \rho_j \rangle} \tag{2.15}$$

where $\rho_{ij}(r)$ is the local density of atoms j at a distance r from atom i and $\langle \rho_j \rangle$ is the average bulk density of atom type j.

2.4 Results and Discussion

2.4.1 Volumetric Properties

Fig. 2.1 shows the ability of the considered force fields to reproduce the solution density and the apparent molar volume of aqueous TMAO mixtures over a broad concentration range (cf. Tab. 2.2) at 298.15 K. The corresponding results for 278.15 K and 323.15 K are reported in section A.3 of the Supporting Information. Section A.1 of the Supporting Information contains the corresponding raw data. Clearly, the Kast 2016 model shows the best agreement with experimental data. It is interesting that the combination with SPC/E water, which was not used in the parametrization process, yields even better results for the high concentration range (> 2 mol kg^{-1}) than the originally used TIP4P/2005 model. The Shea model in conjunction with SPC/E water slightly overestimates the solution density even though the gradient is reproduced quite well. The Netz model predicts a much too high density enhancement for increasing TMAO concentration which is even more dramatic in the case of the Garcia model. This density overestimation of the Netz and Garcia model seems to be reasonable from a molecular point of view since both of their models show an enhanced TMAO-water interaction compared to the original Kast 2003 parameter set. The UA model from this work in combination with SPC/E water yields reasonable agreement up to 1 mol kg^{-1} but systematically underestimates the density as well as its gradient at higher concentrations. It should be noted again that the corresponding force field parameters were not optimized against aqueous TMAO solution properties at all. A comparison of the setups with different treatments of the electrostatic interactions reveals that the usage of a reaction field scheme leads to an almost constant shift towards lower densities compared to PME, as has been reported before[210]. In Table S1 an offset is visible between the Netz model compared to Shea, Kast 2016 (SPC/E) and UA (PME, SPC/E) for the pure water density at zero TMAO concentration even though all models use SPC/E water. This offset can be attributed to the missing tail corrections which were not present in the corresponding simulation setup (cf. Tab. 2.1), following the setup in the original work[96]. We note that the solution densities calculated in the present work for the Netz model are slightly larger than those reported in the original work[96] (cf. Fig. S2 therein). This is a consequence of the increased van der Waals cut-off of 1.2 nm used in the present work compared to 1.0 nm used in the original work (see also Tab. 2.1). We also note that our implementation of the Garcia model produces densities that are slightly larger than those reported in the original work[95]. However, these differences are not responsible for the incorrect concentration dependence of the solution densities observed for these two models.

From Fig. 2.1b it follows that the ranking of the different models for the reproduction of the apparent molar volume stays the same as in the case of the solution density. This is obvious since the apparent molar volume is a derived quantity connected to the density through Eq. (2.4). However, the density representation in terms of the apparent molar volume is not redundant since it delivers a complementary perspective for the interpretation of the results. As an example, one can see that the models of Garcia and Netz yield too low values for the apparent molar volume. This again expresses the fact that the corresponding TMAO model is too hydrophilic and thus the volumetric effect of the insertion of additional TMAO to an existing solution (of known composition) is underestimated. For one particular composition and force field (C7 and Kast 2016) we studied the influence of the system size on density and apparent molar volume. The results are reported in Table A6 of the Supporting Information and show essentially no

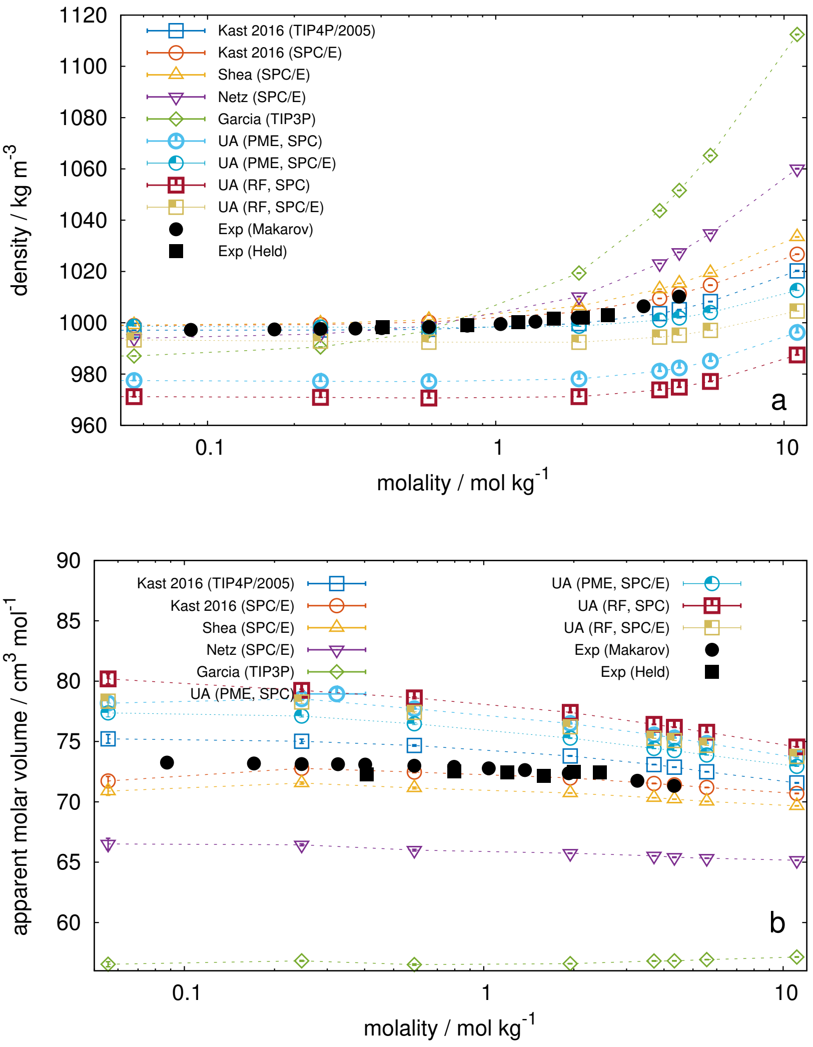


Figure 2.1: Density of aqueous TMAO solutions (a) and apparent molar volume of TMAO (b) as a function of TMAO molality at 298.15 K and 1 bar. Colored symbols represent simulation results, filled black symbols represent experimental data[163,190]. The dashed lines are used as guides to the eye. The numerical values of the densities including their statistical uncertainties are reported in Table A1 of the Supporting Information.

finite size effects.

Since the case of infinite dilution is of special interest, it was investigated separately in terms of the partial molar volume $\bar{V}_\mathrm{s}^\infty$ as a function of temperature which is depicted in Fig. 2.2. The partial molar volume at infinite dilution is an important quantity for characterizing the interactions of the solute with the solvent because it represents the balance between excluded-volume and electrostrictive effects which may act in the same or opposite direction[199]. The plot reveals that all regarded force fields yield approximately the same subtle increase in $\bar{V}_\mathrm{s}^\infty$ with rising temperature in accordance with the experimental data. As in the case for finite TMAO concentrations the results delivered by the Kast 2016 model in combination with TIP4P/2005 as well as SPC/E water are closest to the

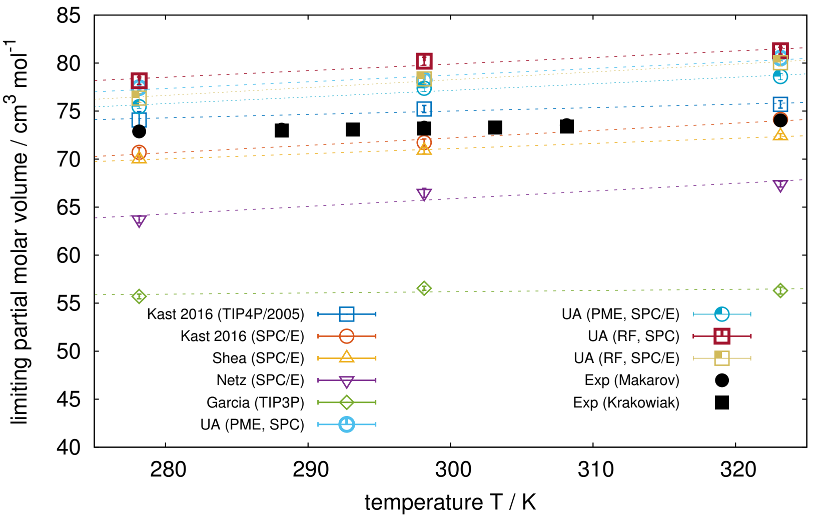


Figure 2.2: Temperature dependence of the partial molar volume at infinite dilution of TMAO at 1 bar. Colored symbols represent simulation results, filled black symbols represent experimental data[190,211]. The dashed lines represent the results of a linear least-squares fit. The numerical values of the box volumes including their statistical uncertainties are reported in Table A1 of the Supporting Information.

experiments. The models of Garcia and Netz predict quite too low values for the partial molar volume, again indicating a too favorable interaction between TMAO and water. In contrast, the UA model from this work seems to yield a too hydrophobic TMAO model since all setups systematically overestimate the partial molar volume.

For the force fields Kast 2016 (TIP4P/2005), Netz and UA (PME, SPC/E) Eq. (2.11) was evaluated. The results are shown in Fig. A35 to A37 in the Supporting Information. The analysis suggests a negative sign for the force fields Kast 2016 (TIP4P/2005) and UA (PME, SPC/E) while for the Netz force field no definite conclusion was possible due to the large error bars. A definite conclusion about the curvature of the function describing the temperature dependence of $\bar{V}_{\mathrm{s}}^{\infty}$ was difficult to achieve (cf. Fig. A35 to A37). However, for aqueous TMAO solutions also the experimental data are contradictory regarding the curvature of $\bar{V}_{\mathrm{s}}^{\infty}(T)$. While the data reported by Makarov et al.[190] suggest a positive value, those of Krakowiak et al.[211] suggest a negative one. A qualitative inspection of the calculated $\bar{V}_{\mathrm{s}}^{\infty}$ data presented in Fig. 2.2 leads us to conclude that the tested force fields indeed encode differences in these second derivative properties which are therefore valuable for further refinement of TMAO models in future studies. Note that for both $\bar{V}_{\mathrm{s}}^{\infty}$ and $\bar{H}_{\mathrm{s}}^{\infty}$ the statistical uncertainties are clearly small enough to distinguish the different TMAO models (cf. Fig. A35 to A37). For the Kast 2016 force field in combination with TIP4P/2005 water we studied the influence of the system size on $\bar{V}_{\mathrm{s}}^{\infty}$. The results are reported in Table A5 of the Supporting Information and show essentially no finite size effects.

2.4.2 Transport Properties

In order to study the influence of TMAO on the solution dynamics, the water self-diffusion coefficient (Fig. 2.3a) as well as the zero-shear rate viscosity (Fig. 2.3b) were calculated as a function of the TMAO molality at 298.15 K. To remove the bias of the water model, both quantities are normalized with the respective values for pure water ($D_{\mathrm{self,w,0}}$, $\eta_{\mathrm{w,0}}$) at the same temperature and pressure. The values of $D_{\mathrm{self,w,0}}$ and $\eta_{\mathrm{w,0}}$ are reported in Table A3 of the Supporting Information. Error bars for the reduced self-diffusion coefficient $D_{\mathrm{self,w}}/D_{\mathrm{self,w,0}}$ (and accordingly for the reduced viscosity) were calculated from the statistical uncertainties of $D_{\mathrm{self,w,0}}$ and $D_{\mathrm{self,w}}$ at a certain composition by the application of standard error propagation rules. The simulations performed in this work confirm the decrease of $D_{\mathrm{self,w}}$ with increasing TMAO molality independent of the force field. However, quantitative deviations between the models are visible even though they are not that distinct as in the case of the solution density. The results for the models of Netz and Kast 2016 are closest to the experimental data. The Garcia model and the different setups for the UA model underestimate the decrease in translational water diffusivity considerably, while the Shea model shows a slight underestimation.

Fig. 2.3b shows the reduced zero-shear rate viscosity as a function of the TMAO molality at 298.15 K. Like in the case of the self-diffusion coefficient, the different models agree at least on a qualitative basis with the experiments[157], which show a strong enhancement for increasing TMAO concentration. It can be noticed that the quantitative ranking of the models stays the same as in the previous case. Note however, that other authors[212] reported a value for the viscosity at a molality of $5.5\,\mathrm{mol\,kg^{-1}}$ that is closer to the simulation results obtained with the Shea and UA model, respectively. It should be noted that the shear viscosities for the different water models calculated in this work are in excellent agreement with values reported in the literature[213].

The example of the Netz model, which, together with the Kast 2016 model, performs best for the representation of transport properties even though it exhibits high deviations for the density, shows that these two properties can be used as independent probes for force field validation.

For one particular composition and force field (C7 and Kast 2016) we analyzed the influence of the system size on the self-diffusion coefficient and zero-shear rate viscosity. The results are reported in Table A6 of the Supporting Information and show that the reduced properties are essentially independent of the system size.

2.4.3 Structural Properties

For studying the difference of the force fields with respect to the structural solution properties, several site-site radial distribution functions (RDFs) were calculated over the

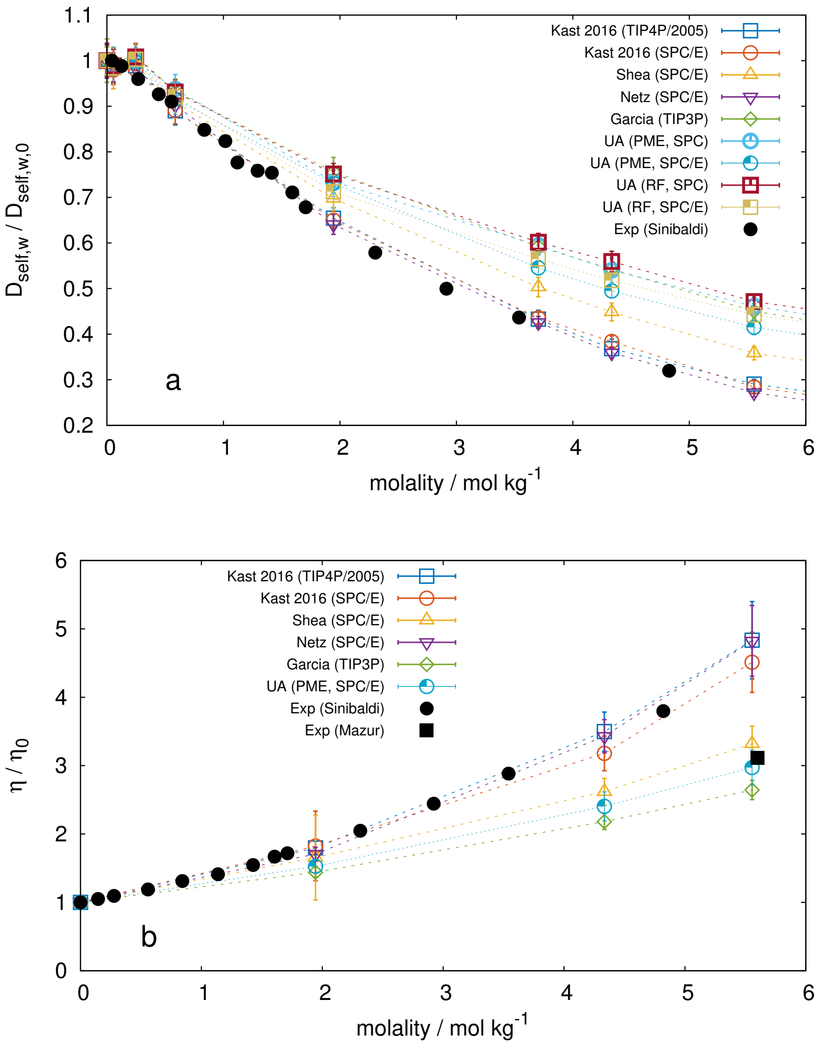


Figure 2.3: Self-diffusion coefficient of water (a) and zero-shear rate viscosity (b) in aqueous TMAO solutions relative to the values of pure water as function of TMAO molality at 298.15 K and 1 bar. Colored symbols represent simulation results, filled black symbols represent experimental data[157,212]. The dashed lines are used as a guide to the eye.

considered concentration and temperature range as well as the average number of hydrogen bonds per TMAO molecule.

In Fig. 2.4, the RDFs for the TMAO nitrogen atoms with itself denoted as N-N (a) as well as between TMAO nitrogen and water oxygen atoms denoted as N-OW (b) are shown for an intermediate TMAO molality (1.94 mol kg^{-1}) at 298.15 K. Fig. 2.5 contains the corresponding cumulative number distributions. The RDFs for N-N and N-OW represent the local ordering of TMAO around TMAO and water around TMAO, respectively. RDFs between other sites for the remaining concentrations and temperatures can be found in section A.2 of the Supporting Information.

The models of Netz and Kast 2016 yield very similar profiles for the N-N RDF with two peaks at around 0.6 nm and 0.8 nm. It is obvious from the high peak in the N-N RDF that

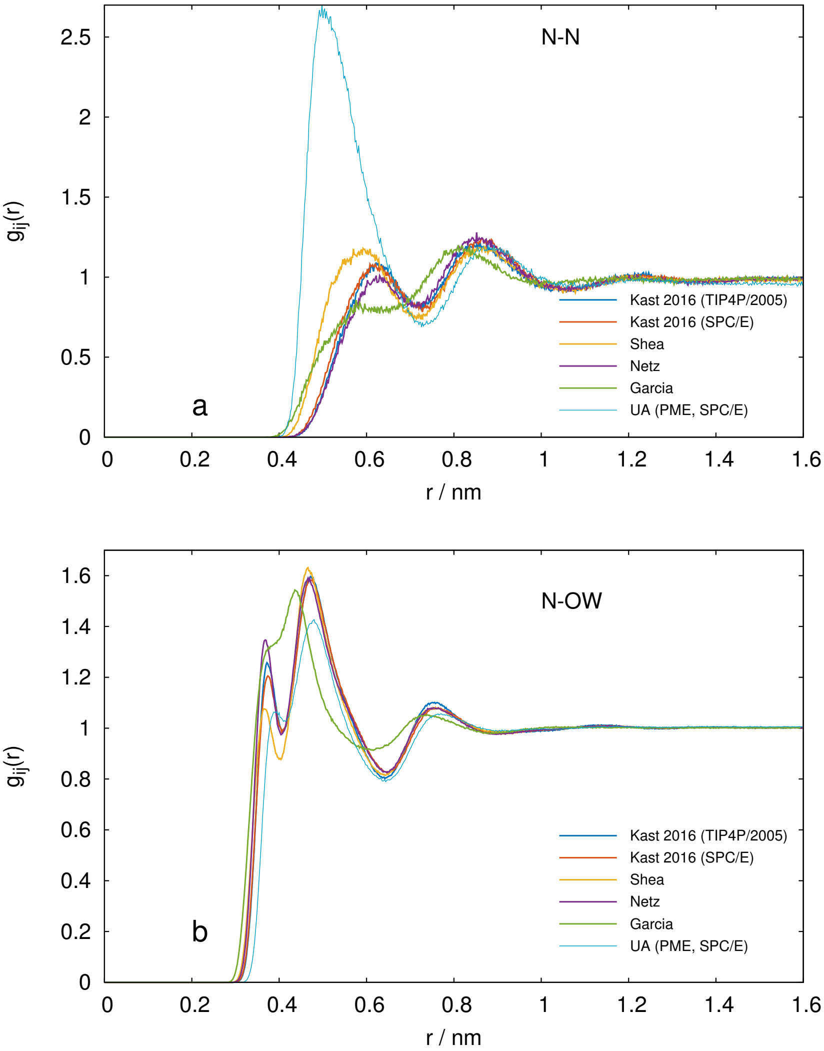


Figure 2.4: Radial distribution functions in 1.94 mol kg^{-1} TMAO solution at 298.15 K and 1 bar for different TMAO models. a: Distribution of TMAO nitrogen around TMAO nitrogen. b: Distribution of water oxygen around TMAO nitrogen.

the UA model shows stronlgy increased TMAO-TMAO attraction or too weak TMAO-water interaction, respectively, see e.g. Figs. A2, A3, A4 of the Supporting Information. Indeed, the interaction between TMAO oxygen and water oxygen and between TMAO nitrogen and water oxygen is much weaker than those of the other force fields. It has to be emphasized that through the manual choice of another oxygen atom type from the GROMOS 54A7 force field (type 1 instead of type 2) the properties of the UA model can be improved significantly. This has been tested by a limited set of simulations involving the compositions C1, C4 and C7 (data not shown) and shows the importance of validation of automatically parametrized molecular models. The Garcia model shows a decreased TMAO-TMAO attraction, which becomes visible by means of the strongly reduced height of the first peak in the N-N RDF consistent with the finding in the original work[95]. It is

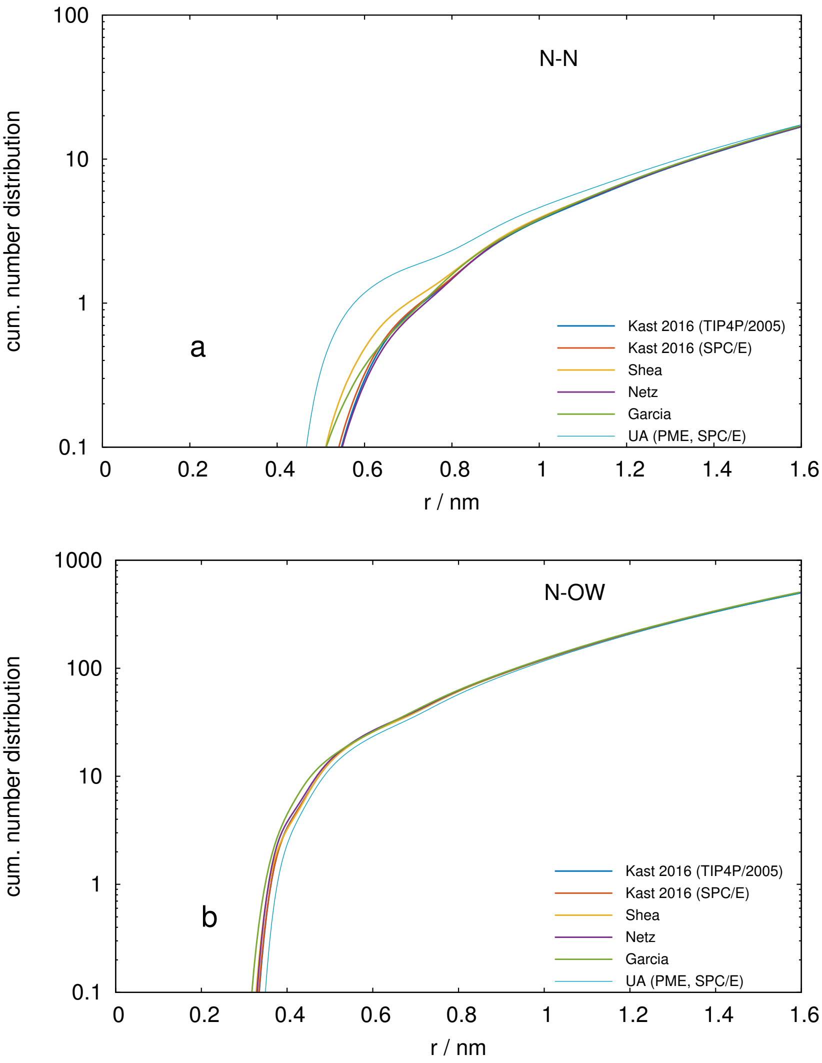


Figure 2.5: Cumulative number distributions in 1.94 mol kg^{-1} TMAO solution at 298.15 K and 1 bar for different TMAO models. a: Distribution of TMAO nitrogen around TMAO nitrogen. b: Distribution of water oxygen around TMAO nitrogen.

noteworthy to mention that not only the peak heights but also the locations of the peak maxima are slightly shifted to the left for the Garcia model compared to Netz and Kast 2016, possibly a consequence of the smaller Lennard-Jones size parameter for C and H in the Garcia model. The Shea model shows an enhanced TMAO-TMAO attraction, though not as distinct as in case of the UA model, which can be seen best in Fig. 2.5 based on the increased cumulative number of nitrogen atoms at small distances (below 0.8 nm) in comparison with the other models (except for UA). The corresponding N-OW RDFs reveal similar differences between the models but deliver complementary information in terms of TMAO-water interaction. It is conspicuous that all models, except for Garcia and UA, show two clearly separated peaks between 0.3 and 0.5 nm. For the UA model, the first peak is significantly smaller, whereas for the Garcia model, both peaks seem to

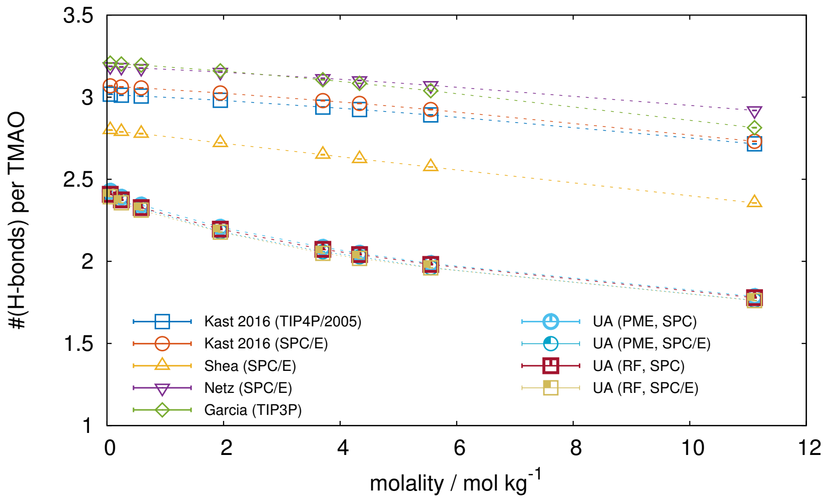


Figure 2.6: Average number of hydrogen bonds per TMAO molecule as function of TMAO molality at 298.15 K and 1 bar. A hydrogen bond is present when the distance H_w–O_{TMAO} is smaller than 0.25 nm and the angle O_w–H_w–O_{TMAO} is larger than 135°.

be smeared out to a single one which is located in between the two peaks exhibited by the other force fields.

Another property which can be used for probing the structural impact of TMAO on the surrounding water is the average number of hydrogen bonds per TMAO molecule as a function of TMAO concentration. Experimental findings based on density and activity coefficient data[214], dielectric and femtosecond mid-infrared spectroscopy[125], terahertz/far-infrared absorption measurements and Raman spectroscopy suggest stable TMAO-water complexes where the TMAO is bound to two, three or four water molecules[215]. Fig. 2.6 shows the result for the hydrogen bond analysis conducted in this work. First, it can be noticed that the concentration dependence follows an almost linearly decreasing trend for all models with similar slopes except for the UA model, which shows a stronger decrease at low concentrations. However, the actual numbers of hydrogen bonds reveal clear differences between the models. While the Kast 2016 model yields approximately 3 hydrogen bonds per TMAO molecule up to 4 mol kg^{-1}, the UA model gives (independent of the setup) less than 2.5. The Shea model delivers slightly less than 3 hydrogen bonds per TMAO in contrast to the models of Garcia and Netz with almost 3.3. We note that the numbers of hydrogen bonds per TMAO calculated for the Garcia model are larger than those reported by Rodríguez-Ropero et al.[115] This is possibly a consequence of the TIP3P water model used in the present work compared to SPC/E water used by Rodríguez-Ropero et al. Our results are consistent with the previously emerged picture namely that the UA model from this work represents a too hydrophobic TMAO model whereas the model of Garcia is too hydrophilic. The Netz model is a kind of special case since it seems to be too hydrophilic only with respect to some properties (number of hydrogen bonds, apparent molar volume) whereas the transport properties can be described almost perfectly. We assume that this is due to the two enhanced antagonist properties of the model, namely the larger dipole moment together with the increased hydrophobicity. To test the sensitivity of the results with respect to the applied hydrogen bond criterion, the definition suggested by Luzar and Chandler[216] assuming a hydrogen bond to exist if the donor-acceptor distance is smaller than 0.35 nm and the hydrogen-donor-acceptor angle is less than 30°, was additionally evaluated as well as the definition by Imoto et al.[217],

$$r_{\mathrm{H}\cdots\mathrm{A}} < -0.171\,\mathrm{nm}\cos(\Theta_{\mathrm{D-H}\cdots\mathrm{A}}) + 0.137\,\mathrm{nm} \tag{2.16}$$

where $r_{\mathrm{H}\cdots\mathrm{A}}$ and $\Theta_{\mathrm{D-H}\cdots\mathrm{A}}$ denote the intermolecular hydrogen-acceptor distance and the donor-hydrogen-acceptor angle, respectively. It was found that the deviations between the different criteria are marginal.

2.4.4 Dielectric Properties

In order to further evaluate the dynamic behavior of the different TMAO models, we also calculated dielectric spectra. These allow a very comprehensive assessment of cumulative dipolar reorientation dynamics over a wide frequency range and are directly comparable to experimental measurements. Furthermore, in systems where medium- to long-range electrostatic interactions play a role, it is important that the employed molecular models reproduce dielectric properties sufficiently well.

In order to discuss the agreement between simulation and experiment, the spectra obtained from simulations employing the Kast 2016 model with TIP4P/2005 water are displayed in Fig. 2.7 (solid lines) together with experimental data[125] (dashed lines) for TMAO molalities $m \approx \{0, 2, 4, 8, 10\}$ mol kg^{-1} as indicated in the legend. The results for the other TMAO models are reported in section A.6.5 of the Supporting Information. The dispersion spectra $\varepsilon_r'(\omega)$ are depicted in Fig. 2.7a and the corresponding absorption spectra $\varepsilon_r''(\omega)$ are shown in Fig. 2.7b. While both the dispersion and absorption spectra obtained from simulations reproduce the overall shape of the experimentally determined spectra well, there exist systematic quantitative discrepancies, which are discussed in the following.

In contrast to the experiment, the simulation underestimates the amplitude of the dispersion spectrum $\varepsilon_r'(\omega)$ on the whole range of concentrations by about 20% regardless of the TMAO molality. Furthermore, the peak frequencies in the absorption spectra are red-shifted with respect to experimental data. However, the frequency shifts of the loss peak maxima decrease with increasing TMAO molality m from approximately 32 GHz at $m = 0$ mol kg^{-1} down to about 0.4 GHz at $m = 10$ mol kg^{-1}. This decreasing shift is very likely due to the decreasing influence of bulk water to the spectrum, since an

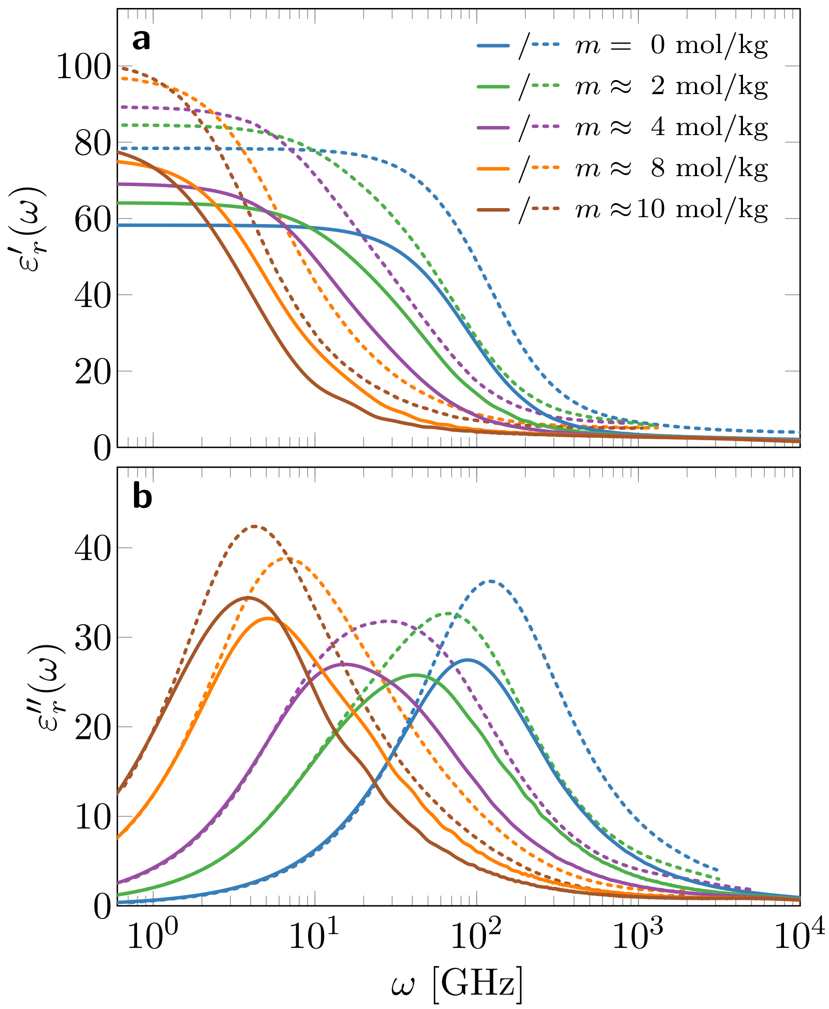


Figure 2.7: Dielectric spectra of aqueous TMAO solutions at different TMAO molalities m. **a**: dispersion $\varepsilon_r'(\omega)$. **b**: loss $\varepsilon_r''(\omega)$. Solid lines represent spectra obtained from simulations employing the Kast 2016 model with TIP4P/2005 water. Dotted lines are fits to experimentally measured spectra extracted from Hunger et al.[125] (Fig. 1 therein).

increasing fraction of water molecules are bound to TMAO molecules with increasing TMAO concentration. The drastic decrease of the loss peak frequency ω^* with increasing TMAO concentration from $\omega^* = 88.3$ GHz corresponding to a relaxation time of $\tau \approx 71.16$ ps (experiment: $\omega^* = 120.0$ GHz and $\tau \approx 52.36$ ps) at $m = 0$ mol kg^{-1} down to $\omega^* = 3.84$ GHz corresponding to $\tau \approx 1.64$ ns (experiment: $\omega^* = 4.27$ GHz and $\tau \approx 1.47$ ns) at $m = 10$ mol kg^{-1} TMAO further supports this interpretation. Due to their size, the collective dipolar reorientation of water-TMAO complexes is much slower than that of bulk water, and with increasing concentration of such complexes, the overall dielectric relaxation time therefore increases as well.

The amplitudes of the loss peaks are generally underestimated by about 20% in the simulation, as it was the case for the reduced static permittivities. Such quantitative discrepancies which are rather independent of TMAO concentration exist for several TMAO models and their magnitude seems to depend on the water model for which they have been developed. Thus, we rather attribute this mismatch to the employed water than to the TMAO model.

To facilitate the further comparison of concentration-dependent dielectric properties obtained from the different models in a concise manner, we chose to extract three key features from the data: the reduced static permittivity $\bar{\varepsilon}_{\text{static}}$ as defined by the zero-frequency limit of Eq. (2.8), the loss peak amplitude $\varepsilon_r''(\omega^*) := \max\left(\varepsilon_r''(\omega)\right)$, and the corresponding loss peak frequency ω^*. These features are depicted as a function of TMAO molality in Fig. 2.8a, b, and c, respectively.

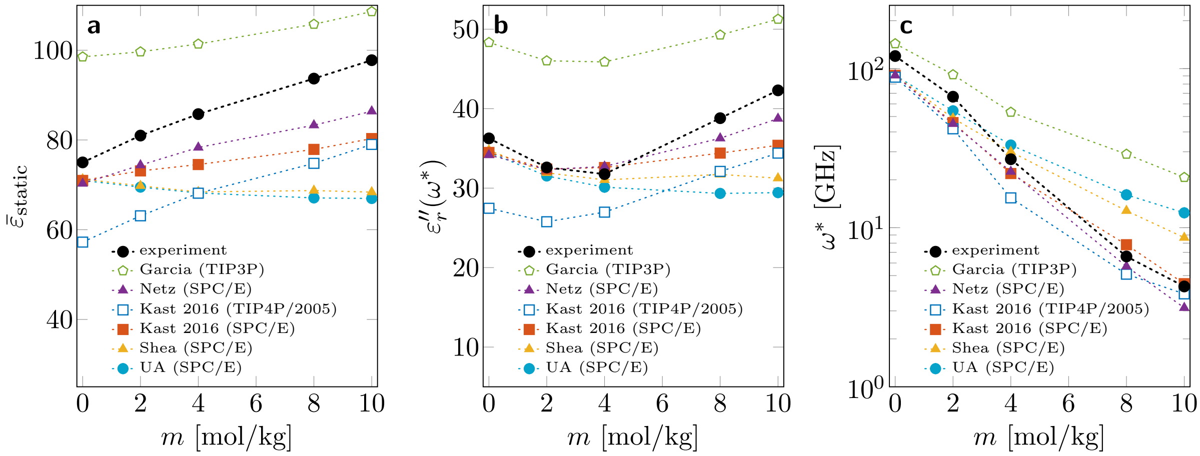


Figure 2.8: Key features of dielectric spectra of aqueous TMAO solutions at different TMAO molalities m. **a**: reduced static permittivity $\bar{\varepsilon}_{\text{static}}$. **b**: loss peak amplitude $\varepsilon_r''(\omega^*)$. **c**: loss peak frequency ω^* [GHz]. Experimental data were extracted from Hunger et al.[125] (Fig. 3 therein). Lines serve as a guide to the eye.

As already discussed based on the spectra shown in Fig. 2.7, the Kast 2016 model with TIP4P/2005 water exhibits good qualitative behavior for all key features displayed in Fig. 2.8. The quantitative mismatch can very likely be attributed to the TIP4P/2005 water model, which strongly underestimates the static permittivity of water. This proposition is also supported by the better quantitative agreement of the same TMAO model when used together with SPC/E water. However, especially for high TMAO concentrations, the qualitative behavior of the Kast 2016 model with SPC/E water deviates from the experiment, indicating that TMAO-water cross-interactions play an increasingly important role with rising TMAO concentration.

The strong quantitative influence of the employed water model is further highlighted by the data obtained from simulations of the Garcia model with TIP3P water. In contrast to SPC/E water, the TIP3P model is known to significantly underestimate rotational relaxation times[210], while both water models have the same molecular dipole moment of 2.35 D[156,218]. Thus, the faster reorientation dynamics of TIP3P water molecules lead to stronger and faster fluctuations of the system's total dipole moment, resulting in a larger amplitude and a blue-shift of the dielectric spectrum. This behavior is reflected in the dielectric properties of the Garcia model in terms of an overestimation of all three key features. Qualitatively, however, the Garcia model appears to reproduce the experimental trends comparatively well.

The dielectric properties of the Shea and UA models very much coincide for all dielectric key features. However, in contrast to all other investigated models, these models lead to a dielectric decrement with increasing TMAO concentration, and thus fail to reproduce the experimental trend observed in the reduced permittivity data. Even though the loss peak amplitudes follow the experiment in the low concentration regime, they suffer from the same deficiency for TMAO molalities exceeding 4 mol kg^{-1}. The only exception are the trends in the loss peak frequencies ω^* observed for these models. Nevertheless, in the high concentration regime, they are outperformed by all other models used with SPC/E water. The deviations of both the Shea and UA models from the experiment at higher concentrations of TMAO indicate that the agreement at low concentrations is probably mainly due to the good dielectric properties of SPC/E water, which is in agreement with the corresponding observations for the systems employing the Kast 2016 model.

Last but not least, the Netz model clearly exhibits the best overall agreement with experimental dielectric measurements for the entire concentration range, even though the Kast 2016 model with TIP4P/2005 water seems to reproduce qualitative trends slightly better.

2.5 Conclusion

The aim of the present study was to compare different molecular models for TMAO in terms of thermophysical properties probing solute-water and solute-solute interactions. The selected thermodynamic, transport, structural and dielectric properties proved to be a useful set for discriminating the various parametrizations. The analysis of concentration-dependent thermodynamic and dielectric properties is a good demonstration of the potential of MD simulations to accurately predict qualitative trends even for dynamic observables. However, it also demonstrates the vital importance of choosing an appropriate molecular model, since models which are parametrized to reproduce a specific observable may fail to predict the behavior of another. Models such as Kast 2016 and Netz showed essentially no difference in transport properties despite their differences in representing volumetric properties. Models such as Kast 2016 (SPC/E) and Shea revealed only little differences in volumetric properties but significant differences in structural properties. Other models such as the non-optimized united atom one showed acceptable volumetric properties at low molality despite being much too hydrophobic. Given the limits of pairwise additive fixed-charge force fields it was rather unexpected that the model Kast 2016 showed a good representation of all the properties considered over the entire concentration range, leading us to conclude that this model has achieved a good balance between solute-solute and solute-solvent interactions. In contrast, the automatically generated united atom model showed rather large deviations from most experimental data which could be explained in terms of too weak solute-water interactions, a consequence of non-optimal assignment of interaction parameters. The other models considered agree only with subsets of the investigated properties. For further refinement of TMAO models, the consideration of second-derivative properties may give valuable insight into the thermodynamic signature of the solute-water interactions. Regarding the findings of Rodríguez-Ropero et al.[115] it is likely that the transfer of the results for TMAO-water binary mixtures to the ternary case including proteins is not straightforward. Even though it is hard to say to what extent the force field performance found in the present work will be altered by the presence of a solute, we are convinced that a proper reproduction of binary aqueous solution data should be a demanded quality feature of co-solute-tailored force fields.

Chapter 3

Overcoming Convergence Issues in Free-Energy Calculations of Amide-to-Ester Mutations in the Pin1-WW Domain

The content of this chapter is a literal quote of the publication

D. Markthaler, H. Kraus and N. Hansen, *Journal of Chemical Information and Modeling*, **58**, 2305-2318, 2018

The manuscript was written by Daniel Markthaler. All molecular simulations were conducted by Daniel Markthaler. Hamzeh Kraus helped with the analysis in the context of a student research project supervised by Daniel Markthaler. Niels Hansen supervised the project and manuscript writing.

Abstract

Relative folding free energies for a series of amide-to-ester mutations in the Pin1-WW domain are calculated using molecular dynamics simulations. Special focus is given to the identification and elimination of a simulation-related bias which was observed in previous work (Eichenberger et al., Biochim. Biophys. Acta 1850 (2015) 983) by comparing simulation results obtained with two different starting structures. Subtle local variations in the protein starting structure may lead to substantial deviations in the calculated free-energy changes as a consequence of differences in the sampled ϕ/ψ-dihedral angle distributions of the mutated residue. It is found that the combination of alchemical transformation with Hamiltonian replica exchange for enhanced sampling reduces the

starting structure dependence considerably. Compared to previous work, the improved sampling of both the folded and unfolded state also improves the agreement between simulation and experiment.

3.1 Introduction

A detailed understanding of the thermodynamic principles of protein stability is crucial both from a fundamental point of view and for protein engineering. By using state-of-the-art alchemical free-energy simulations, large-scale protein thermostability estimates based on atomistic molecular dynamics (MD) simulations are now possible[219,220]. A residual discrepancy between simulation and experiment can often be assigned to either force field issues, sampling-related artefacts, errors in the experimental procedures or the reporting of experimental data[219]. Small, and experimentally well-characterized single-domain proteins such as the 34-residue Pin1-WW domain (see Fig. 3.1), represent attractive model systems for disentangling the causes for deviations between experiment and simulations. Moreover, small protein systems may focus on specific aspects of protein stability such as the importance of backbone hydrogen bonding. Specialized mutagenesis strategies, such as amide-to-ester (A-to-E) mutations[69], enable the impact of individual hydrogen bonds to be studied through substitution of the α-amino acid residue with the corresponding α-hydroxy-acid. A-to-E mutations facilitate a tailored probe for a particular hydrogen bond, either by elimination of the hydrogen bond donor (replacing an amide NH-group with an ester oxygen) or by weakening the acceptor of the corresponding residue (replacing an amide carbonyl with an ester carbonyl). Kelly and co-workers applied this approach for thermodynamic[64] and kinetic[221] analyses of the WW domain of human protein Pin1. To investigate the impact of the 11 hydrogen bonds as present in the X-ray crystal structure (see Fig. 3.1), 20 A-to-E variants were chemically designed from solid-phase synthesis. The mutation effects were measured via thermal and chemical denaturation as changes in melting temperature and free energies of chaotropic denaturation with respect to the protein wild-type. The data revealed that the degree of destabilization is extremely dependent on the location of the removed hydrogen bond within the backbone and can differ up to several kJ mol^{-1}. Mutations that removed buried donors inside the hydrophobic core showed the most significant destabilization effects while the perturbation of solvent-exposed hydrogen bonds was much less influential[64,221]. As demonstrated by Eichenberger et al.[100], relative free energies of folding of the considered A-to-E mutations can also be obtained from free-energy MD simulations. In their alchemical perturbation approach, the folded (or native) state is considered separately from the unfolded (or denatured) state. It was found that a proper representation of the unfolded state is essential for an adequate description of protein stability. Since correlation with experiments yielded only moderate agreement, it was suggested that the experimental and computational routes deliver similar but not identical quantities, and thus direct comparison is not straightforward. However, it remained unclear if the rather large discrepancies for some of the mutations arise solely from limitations of the force field, insufficient sampling, other simplifying assumptions within the approach (compared to experimental reality) or are actually a combination thereof. Another observation was that for some perturbations, calculated free-energy differences showed a considerable starting structure dependence which limits the significance of a comparison between simulation and experiment.

The present study represents a continuation and refinement of the previous work, and aims to provide a clearer picture regarding the quality assessment of the computational approach. To target the limiting situation in which the computed free energy values are purely determined by the applied physical model, i.e. the force field, all sampling-related issues have to be eliminated[222]. Therefore, we focus on two important aspects: (i) further investigation of an adequate description for the unfolded state and (ii) identification and elimination of effects due to the protein starting structure.

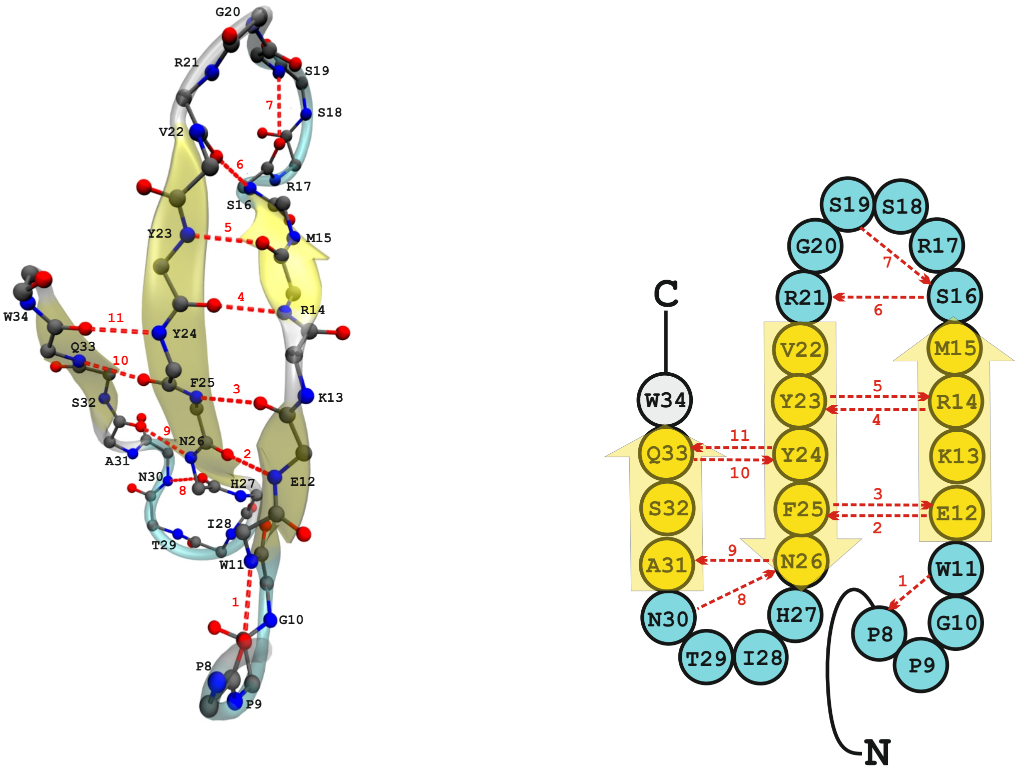
3.2 Methods

3.2.1 Computational Approach

Figure 3.2 illustrates the applied alchemical perturbation approach[100]. The physically meaningful free-energy differences ($\Delta G^{\mathrm{uf}}_{\mathrm{w}}$, $\Delta G^{\mathrm{uf}}_{\mathrm{m}}$), as accessed via thermal or chemical denaturation, are typically difficult to estimate adequately from computer simulations[224]. In contrast, the free-energy differences corresponding to the alchemical transformation from wild-type to mutant ($\Delta G^{\mathrm{f}}_{\mathrm{mw}}$, $\Delta G^{\mathrm{u}}_{\mathrm{mw}}$), can often be calculated with high precision. By treating the folded and unfolded state separately, it is implicitly assumed that the protein folds according to a two-state mechanism. It should be stressed that the two-state assumption was also made in the experimental analysis[64]. Due to the path-independence of the free energy, the experimental and the computational routes, should principally result in the same estimate of the relative free-energy difference $\Delta\Delta G$:

$$\underbrace{\Delta G^{\mathrm{uf}}_{\mathrm{w}} - \Delta G^{\mathrm{uf}}_{\mathrm{m}}}_{\text{Experiment}} = \Delta\Delta G = \underbrace{\Delta G^{\mathrm{f}}_{\mathrm{mw}} - \Delta G^{\mathrm{u}}_{\mathrm{mw}}}_{\text{Simulation}} \tag{3.1}$$

The latter quantity is closely related to the strength of the hydrogen bond affected by the concerned A-to-E mutation[64]. One practical limitation of the approach is related to the sampling of the unfolded state. Even for a small protein as used in this work, it is almost impossible to sample all relevant conformations contributing to the unfolded state ensemble with sufficient statistics[24]. Due to this limitation, the unfolded state is usually



(a)

(b)

Figure 3.1: (a) Protein backbone of the 34-residue Pin1-WW domain (X-ray structure, PDB code 1PIN[223]) in CPK-representation superimposed by cartoon-representation of secondary structure elements. Residue numbering of the sequence, given by K(6)LPPGWEKRMSRSSGRVYYFNHITNASQWERPSG(39), refers to the full-length Pin1 protein. Residues K6, L7 and E35 to G39 are not shown. The 11 backbone hydrogen bonds are highlighted as red dotted lines. The three β-strands are colored in yellow, loop regions in cyan. (b) Schematic representation of the covalent connectivity according to Fig. 1 in Ref. 62 using the same color scheme as in (a).

approximated by short peptides, which implicitly assumes a completely disordered and fully solvent-accessible denatured state. However, the suitability of this approximation is still controversial. For example, Pan and Daggett[225] compared models using sequence segments of a protein where the mutated residue X_i was present in the center ($X_{i-1}X_iX_{i+1}$) with a model where the same residue was surrounded by Ala residues (AX_iA). X_{i-1}, X_{i+1} denote the neighboring residues of X_i according to the protein sequence. The results for these models agreed neither with one another nor with different full-length denatured state models. On the other hand, Seeliger and de Groot[226] as well as Gapsys et al.[219] used generic GX_iG sequences in protein stability calculations, allowing to precompute and tabulate all possible side chain mutations. It was found that this approximation was sufficient to obtain quantitative agreement with experimental data, while Steinbrecher et al.[220] employed tripeptides having the real protein sequence. Among different approaches tested in Ref. 100, the modeling via individual tripeptides according to the protein sequence ($X_{i-1}X_iX_{i+1}$) yielded the best agreement with experiments. This finding emphasizes the importance of the local microenvironment along the backbone. Here, we investigate to what extent the incorporation of not only nearest but also second ($X_{i-2}X_{i-1}X_iX_{i+1}X_{i+2}$) and third nearest neighboring residues ($X_{i-3}X_{i-2}X_{i-1}X_iX_{i+1}X_{i+2}X_{i+3}$) influences the calculated free-energy differences.

We note that in the sections below, A-to-E mutations are denoted by the lowercase Greek letter of the perturbed amino acid in one-letter code, such as S16σ[64,221].

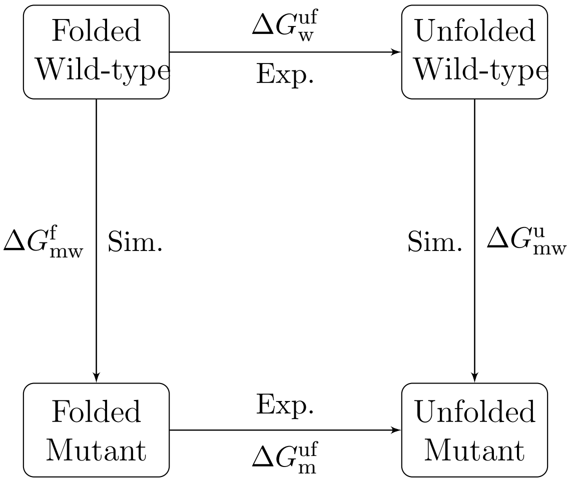

Folded Wild-type
$\Delta G^{\mathrm{uf}}_{\mathrm{w}}$
Exp.
Unfolded Wild-type
$\Delta G^{\mathrm{f}}_{\mathrm{mw}}$
Sim.
Sim.
$\Delta G^{\mathrm{u}}_{\mathrm{mw}}$
Folded Mutant
Exp.
$\Delta G^{\mathrm{uf}}_{\mathrm{m}}$
Unfolded Mutant

Figure 3.2: Thermodynamic cycle illustrating the different routes followed in experiment (Exp., horizontal arrows) and simulation (Sim., vertical arrows). Denaturation experiments deliver access to the free-energy differences separating the folded (f) from the unfolded (u) state ensemble (denoted as $\Delta G^{\mathrm{uf}}_{\mathrm{w}}$ and $\Delta G^{\mathrm{uf}}_{\mathrm{m}}$ in case of wild-type (w) and mutant (m)). MD simulations beneficially grant access to the free-energy difference corresponding to the alchemical transformation from wild-type to mutant (denoted as $\Delta G^{\mathrm{f}}_{\mathrm{mw}}$ and $\Delta G^{\mathrm{u}}_{\mathrm{mw}}$ in case of folded and unfolded state).

3.2.2 Simulation Setup

Force Field

As in previous work[100], a hybrid protein force field was used by combination of the GROMOS parameter sets 53A6[143] for the protein wild-type and $53A6_{\mathrm{OXY}}$[227] for the ester linkages. When ions were needed for system neutralization, parameters were taken from the GROMOS 54A7 force field[13]. The simple point charge (SPC) model[162] was applied to represent water.

Alchemical Perturbations

In previous work[100], a hybrid topology bearing both the wild-type amide and the mutated ester end state was used for the mutated residue. This topology was branched after the C_α-atom of the previous residue and reunified at the carbonyl C-atom of the mutated residue in order to avoid perturbations of bonded interaction terms, a feature that is currently not supported in the available enveloping distribution sampling (EDS)[228,229] implementation in the GROMOS program package[170–172]. For residues with larger side chains, this setup resulted in convergence issues. The reason was found to be an unfavorable overlap with solvent molecules, if the two non-interacting copies of the side chains were in different conformations during the simulation. As a consequence, unnecessarily low EDS smoothness parameters were required to decrease the energy barriers between the two end states. As described in the supporting information of the present work, this can be avoided by applying a distance restraint between the two side chain copies. In this case, a smoothness parameter of unity can be used for all perturbations which prevents any sampling issue caused by too strong solute-solvent overlap. However, particularly in the folded state, such a restraint may artificially restrict the conformational sampling of backbone dihedral angles in the two end states, which are described by different potential energy terms. To obtain a smooth transition between the amide and ester end states, we therefore used a coupling parameter approach in which the two physical end states correspond to the coupling parameter values of $\lambda = 0$ and $\lambda = 1$, respectively, whereas for the intermediate values the system is in a mixed unphysical state. The alchemical path between the two end states was divided into discrete intermediate states (also referred to as stratification) in which equilibrium simulations were carried out. To enhance configurational sampling[230], a Hamiltonian replica exchange (HRE) scheme was used in which the Hamiltonians representative of the various strata are swapped regularly[231,232]. A single-topology approach was applied, in which both bonded and non-bonded interactions were gradually perturbed from one end state into the other. To ensure comparability with previous work[100], it was confirmed that the change in the free-energy method including the change of the software package along with the recommended settings for treating electrostatic interactions (EDS as implemented in the GROMOS program package[170–172] in combination with reaction field treatment of electrostatic interactions vs. stratification as implemented in the GROMACS program package[166–169] in combination with particle-mesh Ewald) did not induce any systematic bias which hampers the comparison to previous work[100]. Figure B4 of the supporting information shows an almost perfect agreement between free-energy changes calculated with the EDS methodology and those calculated with the combination of HRE and stratification for tri- and pentapeptides, for which sampling issues are less relevant than for the protein. In particular, for the mutations L7λ, S16σ, S19σ, V22ϖ, A31α and S32σ for which no distance restraint was required in the EDS simulations, the agreement between the two methods was within the threshold of the thermal energy.

Preparation of Protein Simulations

Protein coordinates required for the initial configuration of the folded state simulations were obtained from two structures deposited in the Protein Data Bank (PDB)[233]: (i) a 1.35 Å resolution X-ray crystal structure (PDB code 1PIN[223]) and (ii) an NMR-solution structure (PDB code 2KCF[234]), reported as a set of 20 individual and slightly differing conformers, representing the folded state, which are compatible with the experimental NMR observables. In contrast to previous work[100], where only one particular structure of the NMR model set was considered, the HRE scheme used in this study (see below), allows the incorporation of all conformers simultaneously within the same simulation by distributing them among the various λ-points. In case of the X-ray structure, which represents the full-length two-domain Pin1 protein complexed with a dipeptide, only the coordinates of the WW domain were extracted. Protonation states of the amino acids were assigned according to pH 7, resulting in a positive net charge of four elementary charges, neutralized with four chloride ions. Protein topologies used for simulations of the X-ray and NMR structures, differ marginally in the protonation state assigned for His27: nitrogen atom N_δ was protonated in case of the NMR set (denoted as HisA), while N_ϵ was protonated in case of the X-ray structure (denoted as HisB). This choice was justified based on differences in the atomic positions of the X-ray and NMR structures, respectively[100]. To study the influence of the histidine protonation state on the estimated free-energy differences, we additionally prepared the X-ray structure with HisA and the NMR set with HisB, so that all structures with both histidine types are available. It was found, that the effect of the protonation state upon $\Delta G^{\mathrm{u}}_{\mathrm{mw}}$ and $\Delta G^{\mathrm{f}}_{\mathrm{mw}}$ for the mutation H27η, cancels out in the final estimate for $\Delta\Delta G$ (see Tab. 3.1). Neutralized protein structures were placed in cubic computational boxes of 6.3 nm box length and were first energy-minimized in vacuum. After solvation using an pre-equilibrated box of SPC water (leading to 7998/7993 solvent molecules for the X-ray/NMR structures), the solvent was also energy-minimized. Thermal equilibration was performed at constant box volume, by carefully raising the temperature in steps of maximal 60 K to the final target value of 278 K. Simultaneously, position restraints, acting on non-hydrogen protein atoms, were reduced from the initial value of 25000 to 0 kJ mol^{-1} nm^{-2} at 278 K. Initial velocities were assigned according to a Maxwell-Boltzmann distribution centered at 60 K. For pressure equilibration, the simulation was continued for 1 ns at 278 K and 1 bar. Both temperature and pressure were kept constant using the weak coupling scheme[188] with corresponding relaxation times of $\tau_{\mathrm{T}} = 0.1$ and $\tau_{\mathrm{p}} = 0.5$ ps. This pre-equilibrated system was then simulated for another 10 ns using the Nosé-Hoover thermostat[177–179] and Parrinello-Rahman barostat[180,181] with corresponding coupling constants of $\tau_{\mathrm{T}} = 1.0$

and $\tau_\mathrm{p} = 2.0$ ps. Protein and solvent (including ions) were coupled to separate heat baths. The value for the isothermal compressibility was set to $4.575 \cdot 10^{-5}\,\mathrm{bar}^{-1}$. All bond lengths were constrained using the LINCS algorithm[175,176] with a LINCS-order of 4. The number of iterations to correct for rotational lengthening in LINCS was set to 2. Short-range electrostatics and Lennard-Jones interactions were treated with a Verlet-buffered neighbor list[183], using a cutoff distance of 1.40 nm and potentials shifted to zero at the cutoff. Analytic dispersion corrections were applied for energy and pressure. The particle-mesh Ewald (PME) method[184,185] was used for treating long-range electrostatic interactions. Simulations were conducted under periodic boundary conditions using the leap-frog algorithm[235] for integrating Newton's equations of motion with a timestep of 2 fs.

Preparation of Peptide Simulations

Tripeptide topologies were prepared for the 20 A-to-E mutants with the mutated residue in the center, flanked by the adjacent residues according to the protein sequence. N- and C-termini were capped with neutral acetyl (MECO) and an N-methyl (NME) group respectively, except for the mutation L7λ which was terminated with NH_3^+ instead of MECO to resemble the situation in the physiological state of the protein. Moreover, it was treated as tetra- instead of a tripeptide to avoid the replacement of a C-terminal proline by the NME capping group, which would introduce an additional hydrogen-bond donor. The use of MECO instead of NH_3^+ at the N-terminus in previous work[100], was the reason for the artificially large $\Delta\Delta G$ value of the perturbation L7λ reported therein. In contrast, the treatment for L7λ described above, yields an estimate for $\Delta G^\mathrm{u}_\mathrm{mw}$ which is close to the value obtained for $\Delta G^\mathrm{f}_\mathrm{mw}$ as expected for the solvent-exposed N-terminal residue. Topology generation was performed by using the `make_top` program of the GROMOS++ software suite[173], followed by conversion to a GROMACS compatible format. Initial coordinates were obtained from the X-ray crystal structure of the full-length WW domain. The systems were neutralized if necessary, and placed in cubic boxes of 4.3 nm length. After energy minimization, the solvated peptides (number of water molecules varied between 2943 and 2958) were heated up to 1000 K for 1 ns at constant box volume to remove any possible bias due to the initial structure in the protein. Equilibration was performed in a subsequent NVT simulation at 278 K, followed by another isothermal-isobaric simulation at 278 K and 1 bar, each for 1 ns. Finally, the simulations were continued for another 2 ns. All simulation parameters were identical as for the protein simulations. Corresponding pentapeptides were prepared in analogous manner. Again, in case of the mutation L7λ, the N-terminus was capped with NH_3^+. The mutation W34ω was treated as hexapeptide, in order to avoid that the NME end-group replaces a proline at the C-terminus. To study the aforementioned influence of the histidine protonation state on the calculated free-energy changes, the affected mutations (N26ν, H27η in case of tripeptides and F25ϕ, N26ν, H27η in case of pentapeptides) were prepared with both histidine types (HisA, HisB). The extension to heptapeptides was only studied via the EDS methodology using the hybrid topology approach of previous work[100], but employing distance restraints between the two side chain copies in case of larger side chains (see supporting information).

Stratification and Hamiltonian Replica Exchange

Bonded and non-bonded parameters were transformed simultaneously within 20 equally spaced λ-states, from the amide ($\lambda = 0$) into the ester state ($\lambda = 1$). Details, concerning the affected force field parameters are reported in the supporting information of our previous work[100]. The influence of different setups with respect to calculated free-energy difference was tested (different λ-spacing, application of soft-core potentials, sequential transformation of electrostatic and Lennard-Jones interactions) and found to be negligible. For the peptide simulations, every λ-point was simulated for 40 ns, while for the protein simulations a simulation time per λ-point between 60 to 100 ns was required, depending on the considered mutation. Convergence was assessed by evaluating the free-energy difference as function of simulation time per λ-point. Potential energy differences between all λ-points together with the derivative of the coupled Hamiltonian H_λ with respect to the coupling parameter λ, $\partial H_\lambda(\lambda)/\partial\lambda$, were written to file every 500 steps. The configurational sampling was enhanced by allowing the Hamiltonians of λ-points to swap at predefined time intervals (here, every 1000 steps). The decision whether to accept or reject an attempted exchange between two replicas i and j is based on the Metropolis-Hastings criterion with a probability of[236]

$$\min\left\{1, e^{\left[\left(U_i\left(\mathbf{r}_i^N\right)-U_i\left(\mathbf{r}_j^N\right)\right)+\left(U_j\left(\mathbf{r}_j^N\right)-U_j\left(\mathbf{r}_i^N\right)\right)\right]/RT}\right\} \tag{3.2}$$

where RT denotes the thermal energy and U_i and U_j the potential energy component of the Hamiltonians for replica i and j, respectively, evaluated with instantaneous configurations of the replicas $\mathbf{r}_i^N$ or $\mathbf{r}_j^N$. When used with different initial conformations for each λ-point (i.e. different conformers of the NMR set representing the folded state), the configurational sampling of the HRE scheme is further improved and makes better use of the available experimental data compared to a single starting structure. The order of the assignment of a particular structure to a particular λ-point was found to have no influence on the calculated free-energy difference.

3.2.3 Analysis

Free-Energy Differences

Alchemical free-energy differences ($\Delta G_{\mathrm{w}}^{\mathrm{uf}}$, $\Delta G_{\mathrm{m}}^{\mathrm{uf}}$) were calculated from the sampled time series of the potential energy differences between all pairs of λ-states. Estimation was performed employing the multistate Bennett's acceptance ratio (MBAR)[237] estimator as implemented in the freely available python package pymbar[238]. Equivalence of MBAR to other free energy estimators can be shown for certain conditions, such as the Bennett's acceptance ratio (BAR) method[239] when only two states are considered or the weighted histogram analysis method[240] for the limiting situation of zero bin size in case of the latter[237]. Uncertainties can be computed from the covariance matrix of the free-energy differences, calculated beforehand[237,241]. The pymbar package delivers further comparison between different free-energy estimators (such as thermodynamic integration and BAR), additional tools for data decorrelation, equilibration detection and graphical inspection of the phase space overlap between the alchemical states in order to assess the spacing of λ-points[238,241]. Free energy differences in case of the EDS simulations were calculated as described in Ref. 100.

Hydrogen Bonds

Backbone hydrogen bonds between an amide hydrogen and oxygen were identified based on pure geometric criteria. A hydrogen bond was assumed to exist if the hydrogen-acceptor distance is smaller than 0.25 nm and the donor-hydrogen-acceptor angle is larger than 135°. Trajectory analysis was performed using the program `hbond` as part of the GROMOS++ software suite[173].

3.3 Results and Discussion

3.3.1 Independence of $\Delta G_{\mathrm{mw}}^{\mathrm{u}}$ on the Peptide Size

For the same mutation, no systematic influence of the peptide length in the unfolded state on $\Delta G_{\mathrm{mw}}^{\mathrm{u}}$ was observed. This result holds irrespectively of whether EDS simulations were employed (carried out for tri-, penta-, and heptapeptides, see Fig. B5) or the current setup (carried out for tri- and pentapeptides, see Fig. 3.3) was used. Therefore, we conclude that the dominant contributions to the free-energy changes in a fully water-exposed denatured state are of a local nature, mainly mediated by the interaction with the nearest neighboring residues along the backbone. In the following we will report only the results obtained from tripeptide simulations. Note, that a generic peptide sequence such as AX$_{\mathrm{i}}$A, where X$_{\mathrm{i}}$ represents the mutated residue, is not appropriate in the present

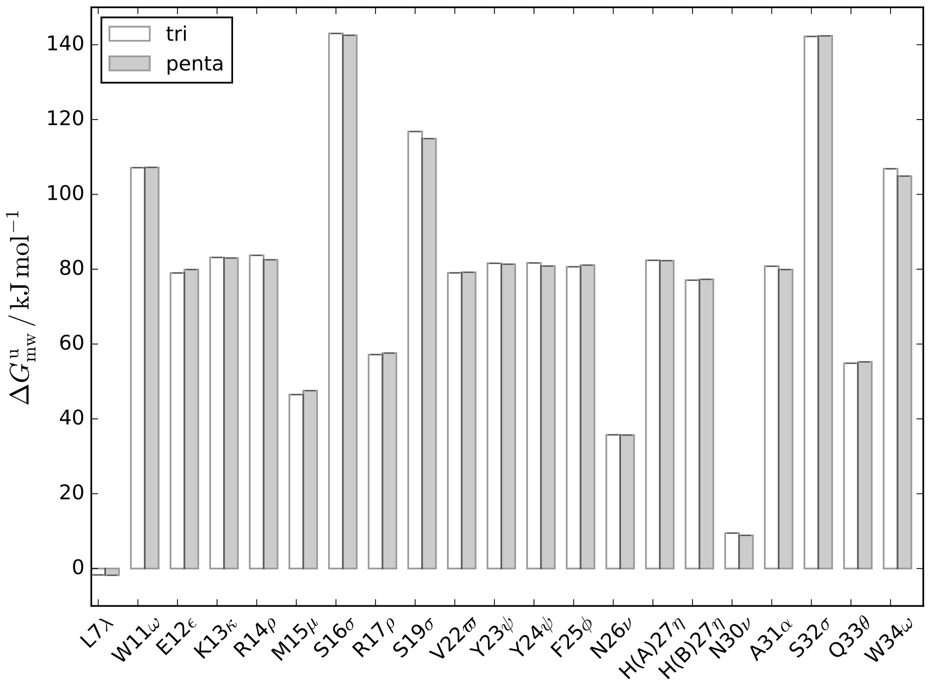


Figure 3.3: Dependence of the alchemical free-energy change ($\Delta G^{\mathrm{u}}_{\mathrm{mw}}$) obtained from HRE simulations on the type of the perturbed residue and on the length of the peptide as used to approximate the unfolded state (tripeptides: white, pentapeptides: grey). H(A) and H(B) denote the different protonation states tested for His27 as described in the main text. Statistical errors are of the order of 0.005 to 0.01 kJ mol^{-1}.

case. By comparing mutations that involve the same perturbed residue but with different neighboring residues, i.e. R14ρ with R17ρ, N26ν with N30ν and S16σ with S19σ or S32σ, respectively, the influence of the surrounding becomes clear immediately (see Tab. 3.1). An attempt to correlate the value of $\Delta G^{\mathrm{u}}_{\mathrm{mw}}$ with the difference in the total dipole moment between the amide and ester state did not show a significant correlation (data not shown).

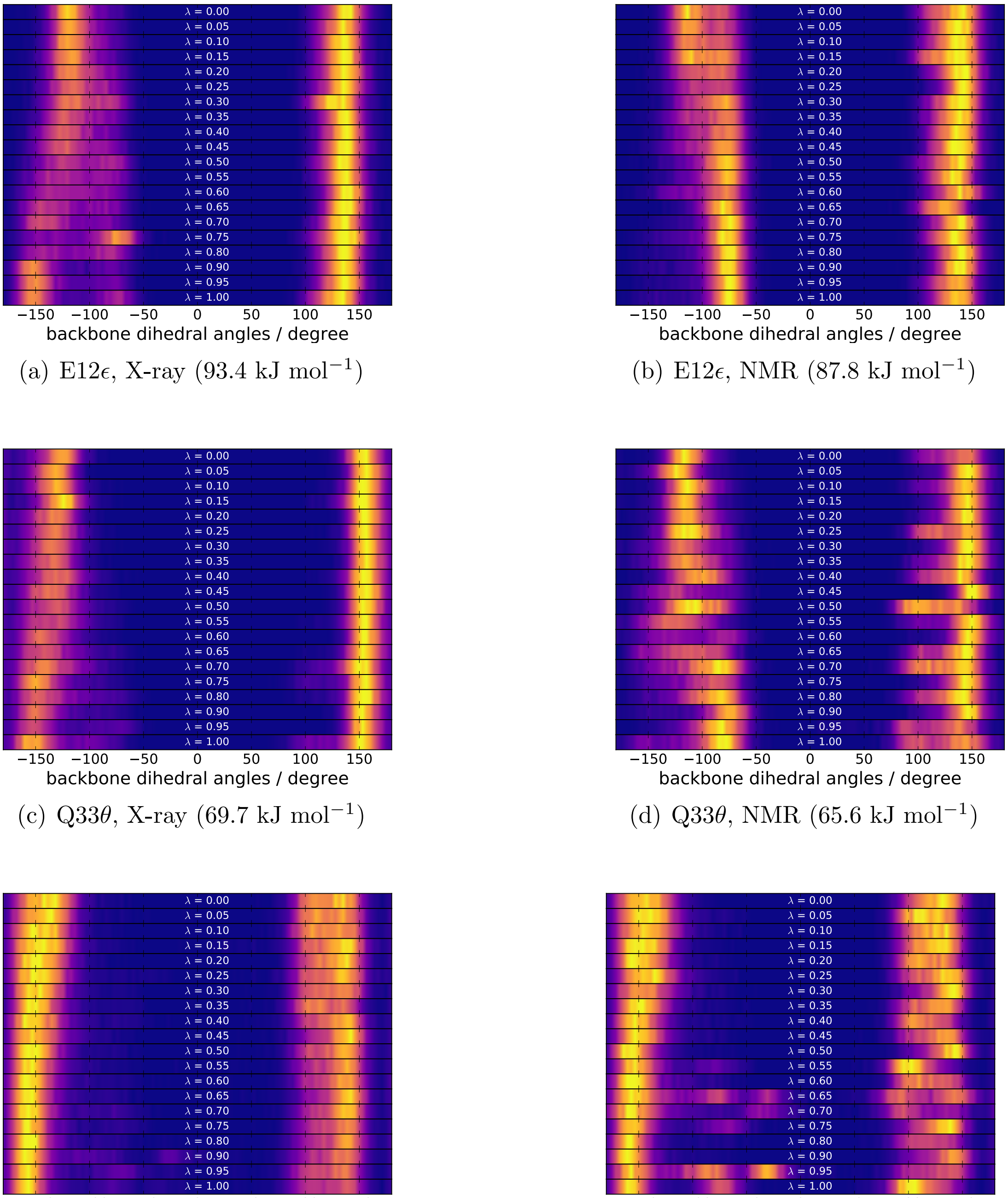
3.3.2 Dependence of $\Delta G^{\mathrm{f}}_{\mathrm{mw}}$ on the Starting Structure

Figure 3.4 shows the influence of the folded state starting structure (X-ray vs. NMR) in terms of ϕ/ψ-dihedral angle distributions of the mutated residue represented as heat map. The examples comprise two mutations (E12ϵ and Q33θ) which show a significant dependence of the alchemical free-energy difference $\Delta G^{\mathrm{f}}_{\mathrm{mw}}$ on the used protein starting structure and one case (A31α) which yields almost identical estimates of $\Delta G^{\mathrm{f}}_{\mathrm{mw}}$ for both starting structures. Sampled distributions, which were obtained from 20 ns per λ-point standard simulations, i.e. without replica exchange are given for all states along the alchemical path from the amide ($\lambda = 0$) to the ester end state ($\lambda = 1$). When considering the mutation E12ϵ (see Fig. 3.4 (a) and (b)) for $\lambda > 0.5$, it should be noted that ϕ-angles sampled within simulations initiated from the NMR set are considerably shifted to higher values compared to analogue simulations based on the X-ray structure. These results, which were equally observed for other cases, suggest that the major reason for the starting structure dependence in the folded state observed in Ref. 100, is related to insufficient sampling of backbone dihedral angles in the vicinity of the perturbed residue. As both mutations E12ϵ and Q33θ are located in stable secondary structures (see Fig. 3.1), simply prolonging the simulation time at each λ-state is unlikely to remedy this sampling problem. In contrast, mixing conformations during HRE simulations is an efficient approach to overcome the responsible sampling barriers without additional computational costs[242]. It should be stressed, that in the limiting case of infinite sampling, no influence of the starting structure is to be expected. In such a (theoretical) scenario, there would be no need for enhanced sampling methods such as HRE.

3.3.3 Hamiltonian Replica Exchange

The influence of the HRE scheme on the conformational sampling can be demonstrated in terms of ϕ/ψ-dihedral angle distributions of the mutated residue. Figure 3.5 shows corresponding histograms for the example S16σ initiated with the X-ray structure. In regular stratification (left column), a rather heterogeneous sampling can be observed with the system being trapped in states which are, according to the replica exchange simulation (right column), not correctly weighted (see e.g. $\lambda = 0.25$ and 0.60). Note, that both ϕ and ψ angles sample more than one state in the HRE simulation but with very different populations compared to regular stratification. This shows that HRE is the preferred methodology as long as the exchange frequency does not suppress the phase space exploration through the natural fluctuations at each λ-point. The latter means that states that would be populated within long independent equilibrium simulations, should also be visited in the HRE simulations.

In previous work[100], only the first structure of the NMR set has been used. HRE allows using different starting structures for different replicas to improve the conformational sampling. In case of the NMR structures, no significant difference in the calculated estimates for ΔG were found when using the first conformer of the NMR set for each replica or when making use of the full set of 20 NMR structures. However, the time to reach convergence may be significantly reduced when making use of the full set of NMR structures. For the X-ray structure, convergence can be improved significantly when generating a set of different starting configurations by assigning different velocities instead of using the same starting configuration for each replica. This is demonstrated for the mutation Y24ψ, which was found to be the most difficult case in the present work (see Fig. B6). When every replica is initiated from a different conformer of the generated synthetic structural set, the estimate of $\Delta G^{\mathrm{f}}_{\mathrm{mw}}$ is instantaneously much closer to the estimate obtained from simulations of the NMR set, compared to the situation



(a) E12ϵ, X-ray (93.4 kJ mol^{-1})

(b) E12ϵ, NMR (87.8 kJ mol^{-1})

(c) Q33θ, X-ray (69.7 kJ mol^{-1})

(d) Q33θ, NMR (65.6 kJ mol^{-1})

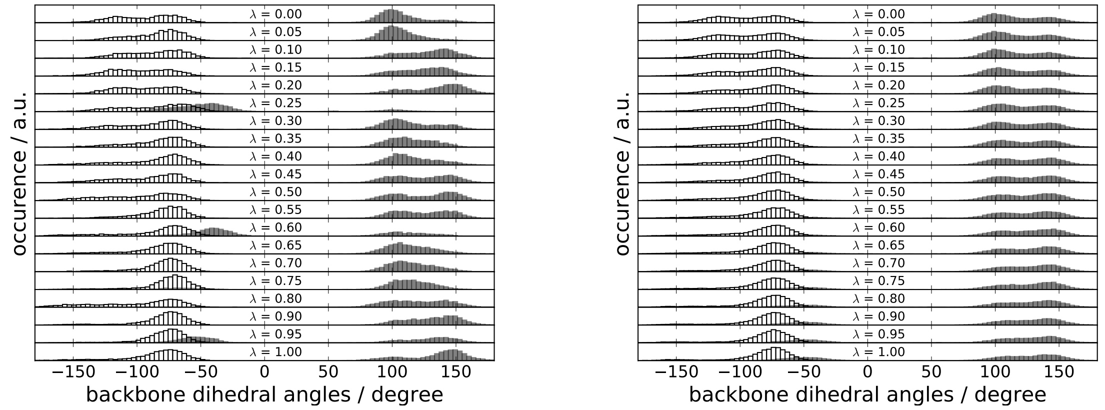
(e) A31α, X-ray (84.5 kJ mol^{-1})

(f) A31α, NMR (85.9 kJ mol^{-1})

Figure 3.4: Heat map representation of ϕ/ψ-dihedral angle distributions of the mutated residue in the folded state, obtained from standard stratification: E12ϵ (a,b), Q33θ (c,d) and A31α (e,f). $\lambda = 0.0$ corresponds to the native (amide) state and $\lambda = 1.0$ to the mutated (ester) state. Highly and low populated states are represented as "hot" light and "cold" dark regions respectively. Distributions of ϕ/ψ-angles are centered around negative/positive values. Left column: simulations initiated from the X-ray crystal structure, right column: every λ-point initiated from a different conformer of the NMR set. Estimates for $\Delta G^{\mathrm{f}}_{\mathrm{mw}}$ are given in parenthesis.

where only a single starting structure is used.

By providing a sufficient amount of sampling time per replica (between 60 and 100 ns), the starting structure dependence of the $\Delta G^{\mathrm{f}}_{\mathrm{mw}}$-values can essentially be removed up to an acceptable residual discrepancy on the order of the thermal energy RT (corresponding to 2.3 $\mathrm{kJ\,mol^{-1}}$ at 278 K) for the majority of mutations (see Fig. 3.6). Even for the cases showing the largest discrepancy between the various folded state starting structures (L7λ, E12ϵ, Y24ψ, H27η), the difference does not exceed 1.5 RT. The stability of the folded state, as judged by the conservation of secondary structure elements was verified, both for the amide and the mutated ester state (data not shown).



(a) S16σ, X-ray, without HRE

(b) S16σ, X-ray, with HRE

Figure 3.5: Distributions of backbone dihedral angles (ϕ: white, ψ: black) for S16σ in the folded state obtained from standard stratification (left) and the HRE approach (right). All λ-points were initiated with the same equilibrated X-ray structure.

3.3.4 Correlation Between $\Delta G^{\mathrm{u}}_{\mathrm{mw}}$ and $\Delta G^{\mathrm{f}}_{\mathrm{mw}}$

In previous work[100], a positive correlation between $\Delta G^{\mathrm{u}}_{\mathrm{mw}}$ and $\Delta G^{\mathrm{f}}_{\mathrm{mw}}$ was observed with correlation coefficients of 0.94 and 0.95 for the X-ray and NMR starting structures, respectively. With the improved conformational sampling of the present work, the correlation coefficient is increased to 0.99 (see Fig. B7), which suggests a nearly perfect linear relationship between $\Delta G^{\mathrm{u}}_{\mathrm{mw}}$ and $\Delta G^{\mathrm{f}}_{\mathrm{mw}}$. Equivalently, this also suggests a nearly perfect linear relationship between $\Delta\Delta G$ and $\Delta G^{\mathrm{u}}_{\mathrm{mw}}$ (or $\Delta G^{\mathrm{f}}_{\mathrm{mw}}$), in the form of $\Delta\Delta G = a\,\Delta G^{\mathrm{u}}_{\mathrm{mw}} + b$, with two adjustable parameters a, b.[100] Once these parameters are known, from optimization to the results of some subset or in the extreme case of only two mutations e.g., it would be possible to predict the $\Delta\Delta G$ of a particular mutation only with the knowledge of $\Delta G^{\mathrm{u}}_{\mathrm{mw}}$ (or $\Delta G^{\mathrm{f}}_{\mathrm{mw}}$) without the need to compute $\Delta G^{\mathrm{f}}_{\mathrm{mw}}$ (or $\Delta G^{\mathrm{u}}_{\mathrm{mw}}$) within additional simulations. However, the applicability of this approximation fails in practice, both with the data from Ref. 100 and the present work. From a numerical point of view, the rea-

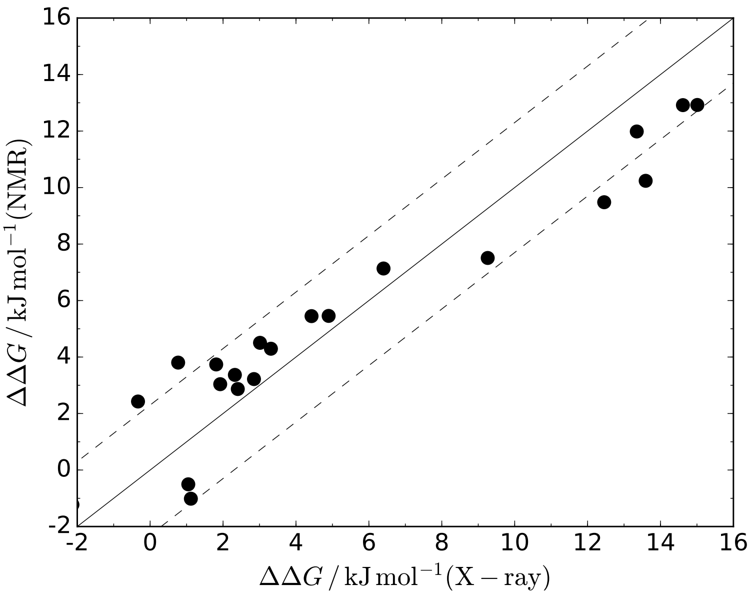
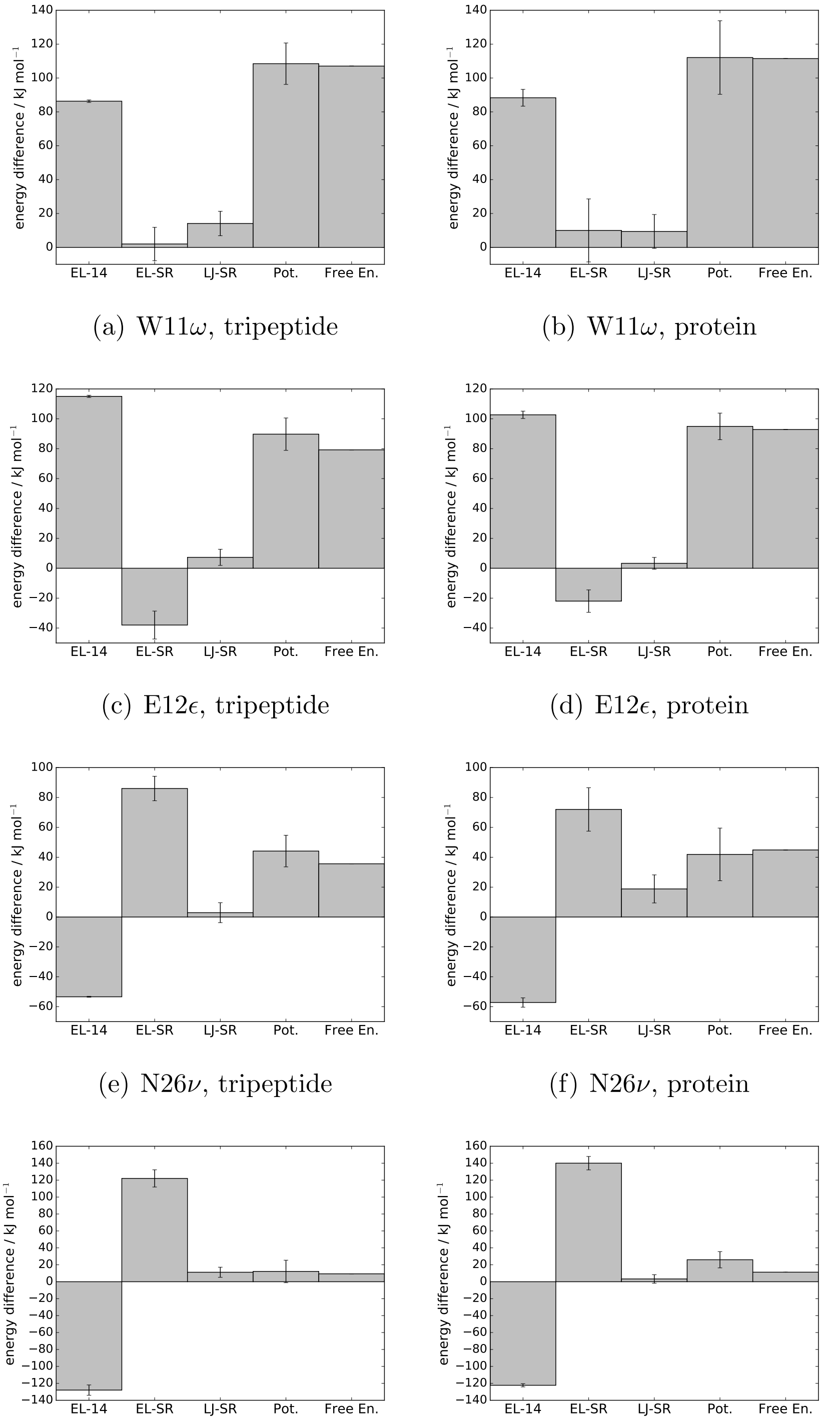


Figure 3.6: Influence of the protein starting structure (X-ray (1PIN) vs. NMR set (2KCF)) in the folded state on the calculated relative free-energy difference $\Delta\Delta G$. The two sets differ in the values for $\Delta G^{\mathrm{f}}_{\mathrm{mw}}$ while the values for $\Delta G^{\mathrm{u}}_{\mathrm{mw}}$, as obtained from the tripeptide simulations, are identical (see Tab. 3.1). Statistical errors are of the order of 0.005 to 0.01 kJ mol^{-1}. The solid line is intended as guide to the eye along $\Delta\Delta G$(X-ray) = $\Delta\Delta G$(NMR), while the dashed lines represent a corridor of ± 2.3 kJ mol^{-1}.

son lies in the rather high numerical values of $\Delta G^{\mathrm{u}}_{\mathrm{mw}}$ and $\Delta G^{\mathrm{f}}_{\mathrm{mw}}$ in comparison to the corresponding difference $\Delta\Delta G$ thereof (see Tab. 3.1 or Fig. 3.3). For the majority of mutations, the energetic contribution of a particular backbone hydrogen bond which is related to $\Delta\Delta G$ is very small compared to the alchemical free-energy differences in the folded ($\Delta G^{\mathrm{f}}_{\mathrm{mw}}$) and unfolded state ($\Delta G^{\mathrm{u}}_{\mathrm{mw}}$) itself. A model based on an approximated linear relationship between $\Delta\Delta G$ and $\Delta G^{\mathrm{u}}_{\mathrm{mw}}$ (or $\Delta G^{\mathrm{f}}_{\mathrm{mw}}$), will not be capable of reproducing the individual and highly context dependent character of a particular backbone hydrogen bond. This circumstance is also illustrated by the quasi-uniform distribution of data points above and below the line of best fit in Fig. B7. From a physical perspective, the observed strong correlation between $\Delta G^{\mathrm{u}}_{\mathrm{mw}}$ and $\Delta G^{\mathrm{f}}_{\mathrm{mw}}$ suggests some similarity between the folded and the unfolded state in terms of electrostatic interactions[243]. This hypothesis, which means that these interactions are qualitatively similar in the two states but differ in their magnitude, was investigated and is discussed below.

For further analysis, eight of the 20 mutations (W11ω, E12ϵ, S19σ, Y23ψ, Y24ψ, N26ν, N30ν, S32σ), spanning the entire range of values for ΔG_{mw} were studied in more detail. To decompose the total free-energy difference into an enthalpic and entropic component, additional simulations at three elevated temperatures of 298, 318 and 338 K were conducted in case of the unfolded state, represented by the corresponding tripeptides. The enthalpic contribution ($\Delta H^{\mathrm{u}}_{\mathrm{mw}}$) for each perturbation can be extracted from the slope of a linear least-squares fit to the data $\Delta G^{\mathrm{u}}_{\mathrm{mw}}(T)/T$ over $1/T$. The entropic component then follows from the Gibbs equation according to: $T_0\,\Delta S^{\mathrm{u}}_{\mathrm{mw}} = -\Delta G^{\mathrm{u}}_{\mathrm{mw}} + \Delta H^{\mathrm{u}}_{\mathrm{mw}}$ with $T_0 = 278$ K. In all cases, the enthalpy difference was dominant while $T_0\,\Delta S^{\mathrm{u}}_{\mathrm{mw}}$ was below $3RT_0$ (see Tab. B2). In order to gain further insight into the interactions dominating the A-to-E mutations, an energetic analysis was performed for the selected subset of mutations in the unfolded and folded state. The analyses for the folded state were based on protein simulations started from the X-ray (HisB) structure. Figures 3.7 and 3.8 show the results in terms of mean potential energy differences calculated from the corresponding time-series in the amide state at $\lambda = 0$ and the ester state at $\lambda = 1$. The major potential energy components are given by the change in electrostatic 1-4 interactions as part of the intra-protein bonded interactions (see W11ω, E12ϵ, Y23ψ, Y24ψ, S32σ) and/or the change in the short-range part of non-bonded electrostatic interactions (see S19σ, N26ν, N30ν). For the two mutations involving an asparagine (N26ν, N30ν), the contributions of the two electrostatic potential energy differences cancel out almost completely. The short-range component of non-bonded electrostatics is evaluated between all pairs of interaction sites within the cutoff except for pairs within the same molecule which are three bonds apart[244]. Interactions in case of the latter are treated separately within the 1-4 interactions term. In contrast to the short-range part, it was found that the long-range electrostatic component nearly cancels out completely for all studied mutations when considering the difference between the two end states. The dominant role of the change in electrostatic interactions during the alchemical mutation can be observed not only on the basis of potential energy but also free-energy differences. Therefore two different alchemical paths for connecting the amide with the ester state were compared by separate perturbation of Lennard-Jones (LJ) and electrostatic interactions in case of the tripeptides: in path A, LJ-parameters were perturbed first (together with masses and bonded interactions) followed by perturbation of partial charges and vice versa for path B. We are aware that since the free energy is a global system property, free-energy components associated with particular interactions (such as LJ- and electrostatic interactions) are only defined for a particular chosen path and are therefore, in general, not comparable for two different pathways (such as path A and B considered here)[245]. However, since the free energy contribution due to the change in LJ-interactions was below 2 kJ mol^{-1} for all mutations in both paths, it can be seen as a further confirmation of the dominant role of electrostatic interactions during the alchemical A-to-E mutation. We do not expect a significant different result in case of another established biomolecular force field.



(a) W11ω, tripeptide

(b) W11ω, protein

(c) E12ϵ, tripeptide

(d) E12ϵ, protein

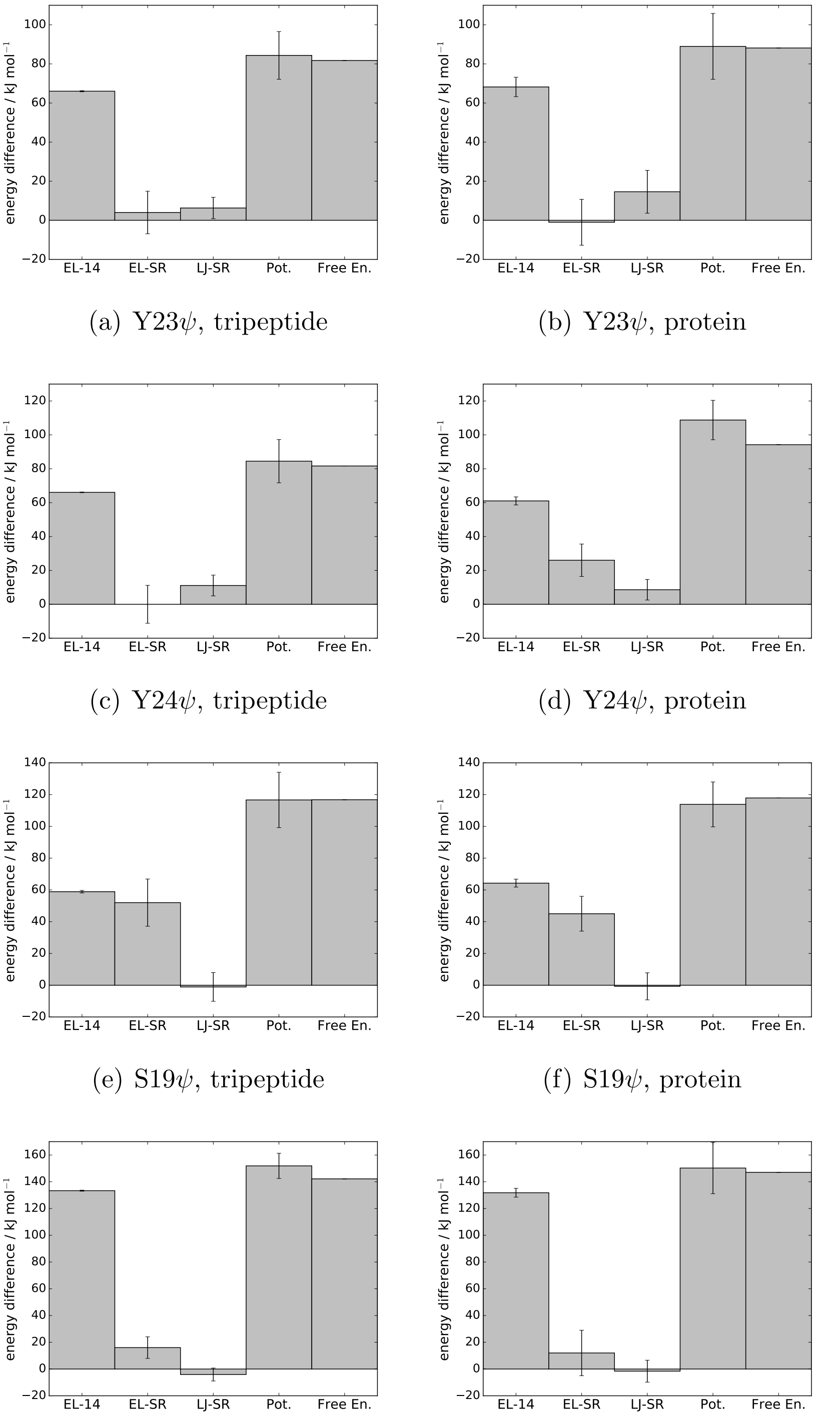
(e) N26ν, tripeptide

(f) N26ν, protein

(g) N30ν, tripeptide

(h) N30ν, protein

Figure 3.7: Energetic analysis in terms of mean potential energy differences between the amide ($\lambda = 0$) and ester state ($\lambda = 1$): (i) electrostatic 1-4 interactions (EL-14), short-range components of (ii) electrostatic (EL-SR) and (iii) Lennard-Jones (LJ-SR) non-bonded interactions, (iv) total potential energy (Pot.) and (v) free-energy difference (Free En.). Left column: tripeptides, right column: protein simulations based on the X-ray (HisB) initial structure.



(a) Y23ψ, tripeptide

(b) Y23ψ, protein

(c) Y24ψ, tripeptide

(d) Y24ψ, protein

(e) S19ψ, tripeptide

(f) S19ψ, protein

(g) S32ψ, tripeptide

(h) S32ψ, protein

Figure 3.8: Energetic analysis in terms of mean potential energy differences between the amide ($\lambda = 0$) and ester state ($\lambda = 1$): (i) electrostatic 1-4 interactions (EL-14), short-range components of (ii) electrostatic (EL-SR) and (iii) Lennard-Jones (LJ-SR) non-bonded interactions, (iv) total potential energy (Pot.) and (v) free-energy difference (Free En.). Left column: tripeptides, right column: protein simulations based on the X-ray (HisB) initial structure.

3.3.5 Hydrogen Bond Analysis

For differentiation between highly similar structures such as the 20 NMR conformers considered in this work, local structural parameters such as the (intra-molecular) hydrogen bonding pattern or secondary structure elements are expected to be more suitable than global ones such as the RMSD relative to some reference structure. Table S3 summarizes the results of a hydrogen-bond analysis applied for the wild-type proteins obtained from 10 ns MD simulations at constant temperature and pressure. While all NMR conformers completely lack hydrogen bond No. 1, it is found in 60 % of simulation time for the X-ray structure. Hydrogen bonds No. 7 and 8 are practically missing both in the X-ray structure and in the whole NMR set. Both of the observations were already made in previous simulations[100]. Focusing solely on the structures of the NMR set, it is clear that conformers 14 and 15 differ compared to the other 18 conformers by the absence of hydrogen bonds No. 2 and 9. For A-to-E mutations which incorporate these particular hydrogen bonds, this can have a significant effect on the calculated $\Delta G^{\mathrm{f}}_{\mathrm{mw}}$ when starting from a such an unfavorable perturbed initial structure. Therefore, HRE simulations were conducted in case of the mutations E12ϵ (H-bond No. 2) and N26ν (H-bond No. 9) where every λ-point was initiated with the same starting structure (data not shown). It was found that simulations that run from conformers 14 and 15, the values of $\Delta G^{\mathrm{f}}_{\mathrm{mw}}$ were up to 10 kJ mol^{-1} lower than those simulations initiated with the other conformers or simulations which incorporated the whole NMR set. Clearly this implies that in such a scenario, one would underestimate the corresponding relative free energy $\Delta\Delta G$ by up to 10 kJ mol^{-1}, leading to a $\Delta\Delta G$ or hydrogen bond strength close to zero, since the value for $\Delta G^{\mathrm{f}}_{\mathrm{mw}}$ approaches $\Delta G^{\mathrm{u}}_{\mathrm{mw}}$. From the temporal evolution of secondary structure elements (see Fig. B8), it can be seen that the third β-strand is absent for conformers 14 and 15 in contrast to the other conformers. Here, a further asset of the HRE scheme compared to standard stratification becomes evident. In cases, one may "accidentally" assign an unfavorable conformer (such as conformer 14 or 15) to the wild-type amide state at $\lambda = 0$, the formation of the affected hydrogen bonds (No. 2 and 9) would be prevented. In contrast, within HRE simulations both hydrogen bonds may be found even for $\lambda = 0$, due to the allowed replica exchanges, as long as these hydrogen bonds are present for the other conformers initially assigned to the remaining λ-points.

3.3.6 Comparison with Experimental Data

Table 3.1 summarizes the calculated free-energy differences for the unfolded state ($\Delta G^{\mathrm{u}}_{\mathrm{mw}}$) as obtained from the tripeptide simulations, the folded state ($\Delta G^{\mathrm{f}}_{\mathrm{mw}}$) and corresponding relative free-energy differences ($\Delta\Delta G$). Experimental reference data were taken from Ref. 64. Therein, A-to-E variants of the WW domain were thermodynamically characterized in two ways: (i) midpoint or melting temperatures (T_{m}) from thermal denaturation measured by far-UV CD spectroscopy and (ii) free-energy differences from chaotropic denaturation employing guanidine hydrochloride (GdnHCl) probed by far-UV CD and fluorescence spectroscopy. Estimation of chaotropic folding free energies at zero denaturant concentration ($\Delta G_{\mathrm{ch}}^{\mathrm{uf}}$) from the recorded denaturation curves, was based on the assumptions of two-state behavior and a linear dependence of the folding free energy on the denaturant concentration[64]. For every A-to-E mutant, both experimental quantities are reported as relative quantities towards the wild-type WW domain.

The overall agreement with chaotropic denaturation data is found to be reasonable, with Pearson correlation coefficients of 0.78, both for simulations based on the X-ray crystal structure and the NMR set (see Fig. 3.9). Compared to previous work[100], where correlation coefficients of 0.53 and 0.54 were determined for simulations based on the X-ray and NMR structures respectively, this represents a significant improvement. Especially for the mutations S19σ and Y23ψ, the starting structure dependence was considerably reduced, while the improved sampling for the unfolded state was particularly clear for the mutation Q33θ. Regarding the quality of the computed free-energy estimates subject to the applied protocol (long simulation times, application of HRE scheme, multiple initial structures), we are confident that the reported free-energy estimates are quite close to the limiting and unique values determined by the force field, and not dominated by sampling issues. Correlation coefficients of other *in silico* mutagenesis studies, employing similar computational approaches[219,220,225] are in the same range as the ones reported here. As can be seen from Figure 3.9, there are only four mutations (N26ν, E12ϵ, Y24ψ, N30ν) for which the deviations between the computed and experimental $\Delta\Delta G$ values considerably exceed the threshold of thermal noise. It is notable, that these cases were also found to be among the most destabilizing mutations in experiments (see Tab. 3.1), either in terms of $\Delta\Delta G_{\mathrm{ch}}$ (N26ν) or ΔT_{m} (E12ϵ, Y24ψ, N30ν). The mutation N26ν, which shows the largest deviation towards the experimental estimate, is worth special attention. For this case, no considerable dependence of $\Delta G_{\mathrm{mw}}^{\mathrm{f}}$ on the folded state starting structure was observed. The chaotropic relative folding free energy for N26ν, which exceeds the computational estimate by more than 10 kJ mol^{-1}, represents the highest $\Delta\Delta G_{\mathrm{ch}}$ value of all mutants. Figure 3.1 shows that N26 is involved in two hydrogen bonds, both as a donor (H-bond No. 9, pairing with A31) and acceptor (H-bond No. 8, pairing with N30). As already pointed out in the previous subsection, hydrogen bond No. 8 was absent in all of the considered folded state structures. However, it is unlikely that the absence of this particular hydrogen bond can explain the large discrepancy, since the same situation is found for S16σ (involved in two H-bonds, where the H-bond with S16 as acceptor is missing), for which very good agreement with experiments was achieved (see Tab. 3.1). Another possible explanation could be the limitation of the applied unfolded state approximation via short peptides. N26 represents the most buried residue in the Pin1-WW domain, as judged by the solvent accessible surface area, and it participates in the formation of one of the hydrophobic cores[62]. Experimental studies on WW domains belonging to the same protein family as the Pin1-WW domain studied here, revealed contradictory results regarding the existence of a residual structured hydrophobic cluster in the chaotropically denatured state when using urea and GdnHCl[63,65]. Considering the possibility that some residual structure remains in the denatured state, it can be expected that the peptide approximation will not be appropriate in cases for which the mutated residue is buried or only partially water-exposed. On the other hand, for mutations L7 and W11 which also participate in the hydrophobic cluster (in the folded state), our approach yields good agreement with the experimental estimates. Due to this uncertainty, it is difficult to assess whether the denatured state, as approximated by the peptide approach, shows closer resemblance to the thermally or chemically denatured state.

A further aspect, is the specific thermodynamic character of WW domains which is more complex compared to ideal two-state folders such as barnase[43] or barley chymotrypsin inhibitor 2 (CI2)[39] and which makes the interpretation of the experimental data more difficult. For WW domains, the free energies from chaotropic denaturation $\Delta G_{\mathrm{ch}}^{\mathrm{uf}}$ show poor agreement with results from thermal denaturation, both in terms of free energies $\Delta G_{\mathrm{th}}^{\mathrm{uf}}$[63,221] and melting temperatures T_{m}[64]. For A-to-E mutations, this poor agreement between experimental data[64] for $\Delta\Delta G_{\mathrm{ch}}$ with ΔT_{m} can be illustrated by means of the following examples (see Tab. 3.1): (i) W11ω, S16σ, Y24ψ, S32σ all show nearly identical estimates for $\Delta\Delta G_{\mathrm{ch}}$ but the values for ΔT_{m} differ by up to 20 °C; (ii) Y23ψ and Y24ψ have a very similar value for ΔT_{m} (38, 34 °C), but different $\Delta\Delta G_{\mathrm{ch}}$ values (9.2, 4.6 kJ mol^{-1}). The weak correlation between different experimental stability indicators for the A-to-E variants was already discussed previously[100] and explained by the inherent difference in the thermodynamic state points of these quantities. However, the aforementioned examples of barnase[43] and CI2[39] yield almost identical estimates for $\Delta G_{\mathrm{ch}}^{\mathrm{uf}}$ and $\Delta G_{\mathrm{th}}^{\mathrm{uf}}$ (when evaluated at the same temperature) and both quantities correlate perfectly with T_{m}. Even when the debate about the existence of residual structure in the chaotropically denatured state and its comparability to the thermally induced denatured state seems to have been resolved by experimental evidence[63], the above mismatch in case of the WW domain(s) is - to the best of our knowledge - still unexplained.

3.4 Conclusion

The current study revisited the alchemical perturbation approach presented in previous work[100] for the calculation of relative folding free energies of A-to-E mutations in the Pin1-WW domain. Therein, convergence issues for some of the mutations suggested that

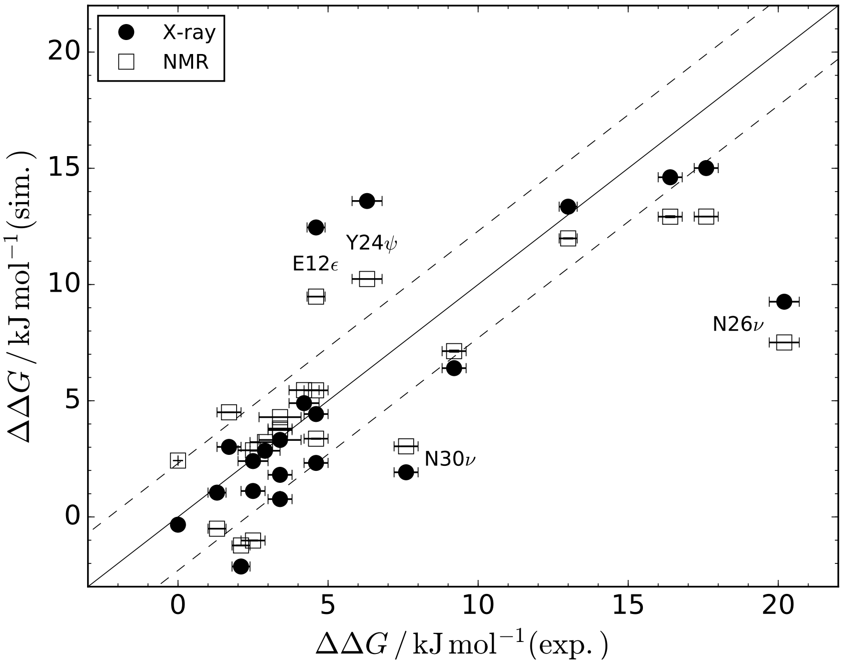


Figure 3.9: Correlation between relative free energies from chaotropic denaturation[64] with calculated values for $\Delta\Delta G$, based on simulations initiated with different protein starting structures (X-ray, NMR). Pearson correlation coefficients are found to be 0.78 for both starting structures. Statistical errors of computed $\Delta\Delta G$ estimates are of the order of 0.005 to 0.01 kJ mol^{-1}. The solid line is intended as guide to the eye along $\Delta\Delta G$(exp.) = $\Delta\Delta G$(sim.), while the dashed lines represent a corridor of ± 2.3 kJ mol^{-1}.

the reported moderate agreement with experiments does not purely reflect the accuracy of the force field and thus prevented an unambiguous quality assessment. The primary purpose of the present work was therefore to identify and eliminate major simulation-related inaccuracies that hamper the agreement with experimental data for linking a residual discrepancy to either force field inaccuracies or uncertainties in the experimental measurements.

The high sensitivity of the calculated free energy differences with respect to subtle conformational changes in vicinity of the perturbed residue could be mainly attributed to insufficient backbone dihedral angle sampling. It could be shown that the use of Hamiltonian replica exchange for enhanced sampling removes the starting structure dependence considerably, with a maximal residual discrepancy marginally larger than the thermal noise (see Fig. 3.6).

The resulting relative free energy differences obtained from the applied simulation setup, show a considerably improved agreement with experimental data (see Fig. 3.9) compared to previous work[100].

A thermodynamic and energetic analysis revealed that (i) most of the considered A-to-

E mutations are dominated by an enthalpy change and (ii) the origin for the strong correlation between the free-energy differences of the folded ($\Delta G^{\mathrm{f}}_{\mathrm{mw}}$) and unfolded state ($\Delta G^{\mathrm{u}}_{\mathrm{mw}}$) is founded in the dominating role of the change in electrostatic interactions.

It was found that individual tripeptides according to the real protein sequence were sufficient for the representation of a fully denatured state due to the dominance of interactions with nearest-neighbor residues. On the other hand, the results suggest that simpler approaches such as the usage of generic sequences would not be appropriate in this case.

The results suggest that the robustness of free-energy difference calculations may significantly depend upon the protein starting structure. The use of different starting structures may be beneficial. This can be done both in situations when multiple experimentally derived structures are available (such as NMR conformers), but also when only a single structure exists (e.g. from X-ray crystallography). In the latter case, a synthetic conformational set can be generated from independent equilibration runs by assigning new velocities.

It can be expected that the observed conformational sensitivity is not a particular property of the considered system but applies equally to other types of mutations and other proteins. Regarding the observed sensitivity, results of computational mutation studies derived from a single protein structure should be interpreted with care.

Studies such as the current one, are essential in order to disentangle limitations of the force field from sampling-related issues due to the applied simulation protocol and/or from the influence of the free-energy estimator[246]. However, the residual discrepancy between simulation and experiment suggest that more factors play a role, including, for example, differences in the thermodynamic state points, deviation from the two-state assumption or kinetic barriers for unfolding. For a future work, it would be interesting to systematically investigate the dependence of the alchemical free energy differences ($\Delta G^{\mathrm{u}}_{\mathrm{mw}}, \Delta G^{\mathrm{f}}_{\mathrm{mw}}, \Delta\Delta G$) on the polarity of the neighboring residues using the computational approach of the present work. Therefore, a similar strategy as proposed by Gao et al.[247] could be followed, where a combination of backbone A-to-E mutations and traditional side chain mutagenesis was applied.

Table 3.1: Summary of computed free-energy differences of all A-to-E mutations for the unfolded state $\Delta G^{\mathrm{u}}_{\mathrm{mw}}$, the folded state $\Delta G^{\mathrm{f}}_{\mathrm{mw}}$ and corresponding relative free energy differences $\Delta\Delta G$ [a].

Mutation	Affected	$\Delta G^{\mathrm{u}}_{\mathrm{mw}}$ [kJ mol^{-1}]		$\Delta G^{\mathrm{f}}_{\mathrm{mw}}$ [kJ mol^{-1}]				$\Delta\Delta G$ [kJ mol^{-1}]				$\Delta\Delta G_{\mathrm{ch}}$	$-\Delta T_{\mathrm{m}}$
	H-bond			X-ray		NMR		X-ray		NMR		[kJ mol^{-1}]	[°C]
	(D/A) [b]	HisA	HisB	HisA	HisB	HisA	HisB	HisA	HisB	HisA	HisB		
L7λ	- / -		-1.7		-2.0	0.7			-0.3	2.4		0.0 ± 0.0	2.3
W11ω	1 / -		107.0		111.5	112.5			4.4	5.5		4.6 ± 0.4	13.8
E12ϵ	2 / -		79.2		92.8	89.1			13.6	9.9		6.3 ± 0.5	36.8
K13κ	- / 3		83.0		85.8	86.2			2.9	3.2		2.9 ± 0.5	12.6
R14ρ	4 / -		83.5		98.1	96.4			14.6	12.9		16.4 ± 0.4	n.d.
M15μ	- / 5		46.5		49.5	51.0			3.0	4.5		1.7 ± 0.4	3.7
S16σ	6 / -		143.1		145.4	146.5			2.3	3.4		4.6 ± 0.4	16.8
R17ρ	- / 7		57.2		58.2	56.7			1.1	-0.5		1.3 ± 0.3	9.9
S19σ	7 / -		116.8		117.9	115.2			1.1	-1.6		2.5 ± 0.4	20.6
V22ϖ	-/ 6		79.5		81.9	82.3			2.4	2.9		2.5 ± 0.5	2.3
Y23ψ	5 / -		81.8		88.2	88.9			6.4	7.1		9.2 ± 0.4	38
Y24ψ	11 / 4		81.7		94.2	91.2			12.5	9.5		4.6 ± 0.3	34
F25ϕ	3 / 10		80.7	95.6	95.7	93.6	95.0	14.9	15.0	12.9	14.4	17.6 ± 0.4	n.d.
N26ν	9 / 2	35.9	35.6	45.0	44.9	43.1	44.0	9.0	9.3	7.2	8.4	20.2 ± 0.5	n.d.
H27η	- / 8	82.4	77.0	82.5	78.9	86.5	80.8	0.2	1.8	4.1	3.7	3.4 ± 0.4	20.3
N30ν	8 / -		9.4		11.3	12.4			1.9	3.0		7.6 ± 0.4	34.2
A31α	- / -		80.8		84.1	85.1			3.3	4.3		3.4 ± 0.7	13.7
S32σ	- / 9		142.2		147.1	147.6			4.9	5.5		4.2 ± 0.5	17.5
Q33θ	10 / -		54.7		68.0	66.6			13.4	12.0		13.0 ± 0.3	n.d.
W34ω	- / 11		106.8		104.7	105.6			-2.1	-1.2		2.1 ± 0.3	9.5

[a] Folded state simulations were initiated either from the X-ray structure[223] or the NMR set[234]. Statistical errors are of the order of 0.005 to 0.01 kJ mol^{-1}. Results are listed separately, according to the assigned protonation state (HisA, HisB) used for His27 as discussed in the main text. Experimental data[64] are reported as relative free-energy differences from chaotropic denaturation $\Delta\Delta G_{\mathrm{ch}}$ and changes in melting temperature ΔT_{m} towards the wild-type WW domain.

[b] Numbering of affected H-bonds according to Fig. 3.1. Concerned A-to-E mutation, either removes an H-bond donor (D) or weakens an H-bond acceptor (A).

Chapter 4

Lessons Learned from the Calculation of One-Dimensional Potentials of Mean Force

The content of this chapter is a literal quote of the publication

D. Markthaler, S. Jakobtorweihen and N. Hansen, *Living Journal of Computational Molecular Science*, **1**, 11073, 2019

The manuscript was written by Daniel Markthaler. All molecular simulations were conducted by Daniel Markthaler. Sven Jakobtorweihen and Niels Hansen supervised the project and manuscript writing.

Abstract

The origins of different computational artifacts that may occur in the calculation of one-dimensional potentials of mean force (PMF) via umbrella sampling molecular dynamics simulations and manifest as free energy offset between bulk solvent regions are investigated. By systematic studies, three distinct causes are elucidated: (i) an unfortunate choice of reference points for the umbrella distance restraint; (ii) a misfit in probability distributions between bound and unbound umbrella windows in case of multiple binding modes; (iii) artifacts introduced by the free energy estimator. Starting with a fully symmetric model system consisting of methane binding to a cylindrical host, complexity is increased through the introduction of dipolar interactions between the host and the solvent, the host and the guest molecule or between all involved species, respectively. The manifestation of artifacts is illustrated and their origin and prevention is discussed. Finally, the consequences for the calculation of standard binding free enthalpies is illustrated

using the complexation of primary alcohols with α-cyclodextrin as an example.

4.1 Introduction

The field of *in silico* pharmaceutical drug design impressively demonstrates the potential of state-of-the-art free energy molecular dynamics (MD) simulations[248]. However, despite a sound theoretical basis[83,84] and the emergence of best practices[230], reliable predictions of the standard binding free energy or rather free enthalpy[249] or Gibbs energy[250], for realistic host-guest systems from computer simulations are still far from routine. As revealed by different case studies, the discrepancy between computed and experimental estimates of the standard binding free enthalpy is often beyond the threshold of 4.2 kJ mol^{-1}, commonly referred to as chemical accuracy[251]. Such deviations may arise from three main simulation-related sources: (i) the force-field problem[22,23], (ii) the sampling problem[222] and (iii) the choice of the free-energy estimator[246]. In addition, experimental uncertainties also have to be considered[93] as well as incompatible thermodynamic state points[252], artifacts caused by the simulation method itself or an inappropriate use of it[253]. The present article covers two of these issues - the sampling problem and methodological artifacts.

In general, two different strategies can be utilized to compute the binding free enthalpy, related either to alchemical double decoupling, or to physical pathway methods such as potential of mean force (PMF) computations[254]. The latter class of methods requires an integration of the PMF over a bound and unbound region, corresponding to the reversible work to transfer the ligand (or guest molecule) from the bulk to the binding pose inside the host. In principle, the PMF-derived estimate of the binding free enthalpy can be validated by results from double decoupling[255,256] or, when possible, by direct counting estimation, based on long unbiased simulations[26,257]. PMF calculations for a specific binding process, are based upon either equilibrium methods such as umbrella sampling[92,258], local elevation[259] or metadynamics[260], adaptive biasing force[261], forward flux sampling[262] or on non-equilibrium methods such as steered MD[263]. In this article, we focus on one-dimensional PMFs obtained via umbrella sampling simulations for host molecules featuring a distinct hydrophobic cavity. Examples for these types of hosts range from rather low molecular-weight substances such as cyclodextrins[264] or cucurbiturils[265,266], up to large moieties such as micelles[267] or protein channels inside a membrane[268–271]. The cavity, enabling a ligand to be bound with high affinity and specificity, makes such host molecules attractive for applications in (computer-aided) drug design. However, it also poses various challenges regarding the applied simulation protocol. Studies of cucurbituril complexes revealed that thermodynamic irreversibilities can occur when certain guest atoms, that are not directly controlled via the bias potential, become stuck inside the host and then suddenly jump outside[265,266]. It was concluded that these dissipative conformational jumps might be a fundamental problem when applying steered MD but also umbrella sampling with fixed spring attachment points to flexible molecules. In typical applications of restrained MD simulations to such molecular systems, a one-dimensional PMF is evaluated by pulling or restraining the ligand along some (linear) path from the bulk at one side of the simulation box through the host to the bulk at the other side. Depending on the complexity of the system, it can be necessary to reduce the sampled space and thus to accelerate convergence by using auxiliary restraints in the simulation setup. The concrete choice of these auxiliary restraints is however non-trivial, since a rigorous way to estimate their effect on the calculated binding free enthalpy is required in order to remove it afterwards[84,272]. For vanishing interactions at large distances between the binding partners, the PMF becomes flat and approaches a constant value. The fact that this constant has to be the same for every ligand position within the bulk region (due to the isotropy of the bulk fluid in the absence of external potentials), can be used as a diagnostic test. In a couple of published examples[268–270,273–276], an artificial offset is visible in the free energy profile between the two bulk regions of the solvent which violates the state function property of the free energy. In some of the cases, this offset was interpreted as indication of insufficient simulation time[275]. However, systematic studies about the origins of these artifacts are scarce[266]. Hub et al.[270] considered solute permeation across a protein channel and found that limited sampling inside the channel in the presence of locally different correlation times can lead to PMF offsets up to 15 kJ mol^{-1}. In Ref. 268, the sampling problem was interpreted as a very small average force across the channel due to the accumulation of noise, originating from all degrees of freedom other than the chosen order parameter. To remedy this problem, the authors followed similar routes: Ref. 268 proposed a symmetrization procedure by creating duplicate umbrella windows on opposite sides of the channel, while Ref. 270 implemented a modified version of the weighted histogram analysis method (WHAM), featuring an additional constraint to enforce periodicity. While such pragmatic solutions may suppress the occurrence of PMF offsets, they do not solve the underlying sampling problem itself. The latter can be solved however using sampling times in excess of microseconds combined with a systematic variation of initial conformations[277] or enhanced sampling methods[278].

The purpose of the present contribution is to demonstrate that computational artifacts may easily occur on much simpler systems for which advanced sampling techniques are not necessarily applied. We elucidate various causes for PMF offsets and relate them to properties of the host-guest system and the applied simulation protocol. The difficulty for setting up free energy molecular dynamics simulations decreased a lot over the last decades, allowing also less experienced users to obtain binding free enthalpy estimates for realistic biomolecular systems. The critical assessment of the results including inspection of convergence and artifacts will always require advanced experience and knowledge, however. With the systematic discussion of the reported artifacts, we aim to sensitize especially newcomers and non-experts in the field in order to prevent time-consuming pitfalls in the context of binding free enthalpy calculations.

4.2 Theory

Although the main goal is to discuss PMF artifacts, for better evaluation of the results, it is advisable to calculate binding free enthalpies ($\Delta G^{\circ}_{\text{bind}}$) from the PMFs. More important, when using additional restraints on the ligand, the PMF depends on the details of these restraints but $\Delta G^{\circ}_{\text{bind}}$ can be calculated by taking into account the specific restraints. In other words, $\Delta G^{\circ}_{\text{bind}}$ should be independent of the concrete choice of restraints. Furthermore, $\Delta G^{\circ}_{\text{bind}}$ is directly related to the binding equilibrium constant and as such enables validation against experimentally determined equilibrium constants[255]. A detailed derivation how to calculate $\Delta G^{\circ}_{\text{bind}}$ from the PMF is beyond the scope of the current work. Therefore, we will just outline the central ideas and present the final expressions. For rigorous derivations, we refer to Refs. 254 and 279. The link between an appropriately defined PMF and $\Delta G^{\circ}_{\text{bind}}$ can be formulated as the ratio of two configurational integrals over a bound (b) and unbound (u) region:

$$\Delta G^{\circ}_{\text{bind}} = -RT \ln \left(\frac{\int_{\text{b}} \mathrm{e}^{-W(\boldsymbol{r}_{\text{HL}},\boldsymbol{\omega}_{\text{HL}})/RT} |\mathbf{J}| \mathrm{d}\boldsymbol{r}_{\text{HL}} \mathrm{d}\boldsymbol{\omega}_{\text{HL}}}{\int_{\text{u}} \mathrm{e}^{-W(\boldsymbol{r}_{\text{HL}},\boldsymbol{\omega}_{\text{HL}})/RT} |\mathbf{J}| \mathrm{d}\boldsymbol{r}_{\text{HL}} \mathrm{d}\boldsymbol{\omega}_{\text{HL}}} \right) - RT \ln \left(\frac{V_{\text{u}}}{V^{\circ}} \right) \tag{4.1}$$

where $V^{\circ} = 1.661\ \text{nm}^3$ is the standard state volume and RT is the thermal energy. The PMF W appearing in the Boltzmann factor, originally depends on the relative separation vector $\boldsymbol{r}_{\text{HL}}$ and the relative orientation vector $\boldsymbol{\omega}_{\text{HL}}$ between host and ligand. In particular, the PMF does not depend on the external degrees of freedom of the complex corresponding to the absolute position and the overall orientation inside the simulation box. Depending upon the choice of coordinates, a Jacobian determinant $|\mathbf{J}| = |\mathbf{J}(\boldsymbol{r}_{\text{HL}}, \boldsymbol{\omega}_{\text{HL}})|$ may arise in the configurational integrals of Eq. (4.1). The second term accounts for the free energy contribution of the volume change from the standard state volume V° to the unbound volume V_{u}. It should be noted that V_{u}, which depends on the size of the simulation box cancels from the final expression for $\Delta G^{\circ}_{\text{bind}}$[280]. At this point we want to emphasize the difference between a PMF and a free-energy curve (FEC). While these terms are often used synonymously in the literature, the FEC contains the Jacobian contribution, while the PMF does not. If, for example, umbrella sampling is applied to two non-interacting particles using the radial separation r as umbrella coordinate, the FEC decreases with $-2\,RT \ln r$, while the PMF becomes flat. For the one-dimensional setup as used in the present work (c.f. Sec. 4.3.2), the Jacobian contribution is equal to unity such that we mostly use the term PMF unless we refer to a three-dimensional calculation setup.

Due to the complexity of the systems, it is often necessary to use auxiliary restraints in the simulation setup. The effect of such additional restraints that limit the phase space to be sampled during the transfer of the ligand from the standard state volume to the binding pose of the host, can be incorporated by introduction of intermediate states into Eq. (4.1)[272]. The approach can be visualized in the form of a thermodynamic cycle as depicted in Fig. 4.1.

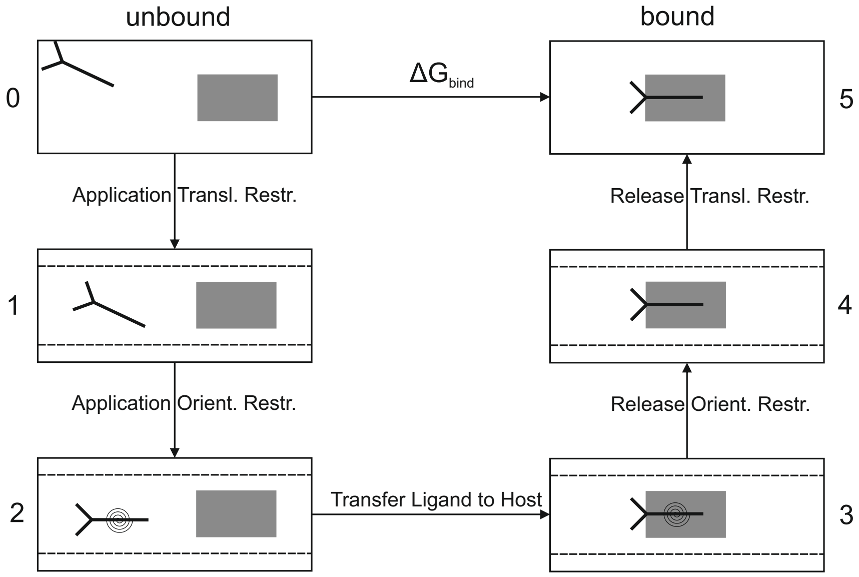


Figure 4.1: Thermodynamic cycle for the calculation of the binding free enthalpy ΔG_{bind}. The host and ligand molecule are represented by the grey rectangle and black structure, respectively. The free enthalpy difference between the unbound (point 0) and bound (point 5) state is given by ΔG_{bind}. When the volume in point 0 is given by the standard state volume V°, ΔG_{bind} corresponds to the standard binding free enthalpy $\Delta G^{\circ}_{\text{bind}}$. Due to the path-independence of ΔG_{bind}, it can be calculated not only from the direct path $0 \rightarrow 5$ as accessed experimentally but equivalently from an indirect path such as $0 \rightarrow 1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5$ as accessed from molecular simulations, including several intermediate states (see main text).

Application to the case of a PMF along a one-dimensional order parameter (ζ) including auxiliary translational and orientational restraints on the ligand (c.f. Sec. 4.3.2), finally yields the following expression for $\Delta G^{\circ}_{\text{bind}}$[255,281]:

$$\Delta G^{\circ}_{\text{bind}} = \Delta G_{\text{V}} + \Delta G_{\Omega} + \Delta W_{\text{R}} + \Delta G_{\theta} + \Delta G_{\rho} \tag{4.2}$$

with terms ordered according to the cycle in Fig. 4.1:

$$
\begin{aligned}
\Delta G_{\mathrm{V}} &= -RT\ln\left(\frac{l_{\mathrm{b}}A_{\mathrm{u},\rho}}{V^{\circ}}\right)\\
\Delta G_{\Omega} &= -RT\ln\left(\frac{\Omega}{8\pi^2}\right)\\
\Delta G_{\theta} &= RT\ln\left(\langle \mathrm{e}^{-U_{\theta}(\theta)/RT}\rangle_{\mathrm{b},k_{\theta}=0}\right)\\
\Delta G_{\rho} &= RT\ln\left(\langle \mathrm{e}^{-U_{\rho}(\rho)/RT}\rangle_{\mathrm{b},k_{\rho}=0}\right)
\end{aligned}
$$

ΔW_{R} denotes the thermally averaged depth of the one-dimensional PMF, i.e.

$$
\Delta W_{\mathrm{R}} = RT\ln\left(\frac{\int_{\mathrm{u}} \mathrm{e}^{-W_{\mathrm{R}}(\zeta)/RT}\mathrm{d}\zeta}{\int_{\mathrm{u}}\mathrm{d}\zeta}\right) \tag{4.3}
$$

If the PMF is constant in the unbound region and the global PMF minimum is defined to be zero as it was done in this work, ΔW_{R} corresponds to the negative PMF value in the bulk ($W_{\mathrm{R},\infty}$): $\Delta W_{\mathrm{R}} = -W_{\mathrm{R},\infty}$. The index "R" indicates that the PMF was evaluated in the presence of auxiliary restraints such as the orthogonal translational and orientational restraints (c.f. Sec. 4.3.2). That is, ΔW_{R} represents the step 2 → 3 in Fig. 4.1. The integration of the PMF over the bound region, is captured in the definition of the bound length l_{b}:

$$
l_{\mathrm{b}} = \int_{\mathrm{b}} \mathrm{e}^{-W_{\mathrm{R}}(\zeta)/RT}\mathrm{d}\zeta \tag{4.4}
$$

Due to the Boltzmann-weighting in Eq. (4.4), the lowest values of $W_{\mathrm{R}}(\zeta)$ contribute the most to the integral while larger values at increasing distances from the minimum have smaller weights. This makes the estimate of l_{b} and thus $\Delta G^{\circ}_{\mathrm{bind}}$ insensitive to the actual choice of the cut-off distance between bound and unbound region, in particular for rather tight binding situations as studied in the present work[83]. Here, the entire range of ζ-values between the flat parts of the PMF on both sides relative to the minimum were considered as the bound region. The free energy contribution due to the volume change from the standard state volume V° to $l_{\mathrm{b}}\,A_{\mathrm{u},\rho}$ is described by the term ΔG_{V} in Eq. (4.2), corresponding to the step 0 → 1 in Fig. 4.1. Therein, $A_{\mathrm{u},\rho}$ denotes the cross-sectional area which is accessible to the unbound ligand in orthogonal directions in the presence of the applied translational restraint. Its value can be calculated analytically from the partition function of the restraining potential $U_{\rho}(\rho)$ used for restricting the lateral movement of the ligand in the bulk solvent[255,281]:

$$
A_{\mathrm{u},\rho} = \int_{0}^{\infty} \mathrm{e}^{-U_{\rho}(\rho)/RT}\,2\pi\rho\,\mathrm{d}\rho \tag{4.5}
$$

where ρ is the orthogonal distance (c.f. Fig. 4.3). The term ΔG_ρ accounts for the free energy contribution of releasing the orthogonal translational restraint in the bound state (4 → 5 in Fig. 4.1). As indicated by the notation $\langle ... \rangle_{\mathrm{b},k_\rho=0}$, this contribution may be evaluated numerically by free energy perturbation[282] using exponential averaging from an additional simulation with the bound ligand at vanishing restraining force constant $k_\rho = 0$[281]. A more sophisticated way would be to perform the estimation within multiple simulations of decreasing values of k_ρ using thermodynamic integration[283] or the Multistate Bennett's Acceptance Ratio (MBAR) estimator[237]. The terms ΔG_Ω and ΔG_θ in Eq. (4.2) assess the free energy contributions from applying and releasing the angular restraint U_θ in the unbound and bound state, respectively (1 → 2 and 3 → 4 in Fig. 4.1). Ω denotes the rotational volume available to the ligand in the bulk under the influence of the angular restraint. For a given functional form of the restraining potential U_θ, its value can be evaluated from a three-dimensional integral over the Euler angles[284,285]:

$$\Omega = \int_0^{2\pi}\int_0^{2\pi}\int_0^{\pi} \mathrm{e}^{-U_\theta(\theta)/RT} \sin\theta \, \mathrm{d}\theta \mathrm{d}\phi \mathrm{d}\psi = 4\pi^2 \int_0^{\pi} \mathrm{e}^{-U_\theta(\theta)/RT} \sin\theta \, \mathrm{d}\theta \tag{4.6}$$

Depending on the functional form of U_θ, this integral may be solved analytically or numerically. If no angular restraint was applied, allowing the ligand to rotate freely in the bulk, Ω equals $8\pi^2$ and Eq. (4.2) becomes identical to Eq. (11) of Ref. 281. Therein, it is assumed that the change in rotational entropy for the bound ligand is included in ΔW_R. In the following we will study situations in which this assumption does not hold. The free energy term ΔG_θ has to be evaluated numerically by free energy perturbation or thermodynamic integration for example.

Application of standard error propagation rules to all involved quantities associated with an uncertainty in Eq. (4.2) gives:

$$\sigma^2\left\{\Delta G^\circ_\mathrm{bind}\right\} = \left(\frac{RT}{l_\mathrm{b}}\right)^2 \sigma^2\left\{l_\mathrm{b}\right\} + \sigma^2\left\{\Delta G_\Omega\right\} + \sigma^2\left\{\Delta W_\mathrm{R}\right\} + \sigma^2\left\{\Delta G_\theta\right\} + \sigma^2\left\{\Delta G_\rho\right\} \tag{4.7}$$

where $\sigma^2\{...\}$ denotes the variance. The uncertainty in ΔG_Ω arises from the numerical integration error associated with the applied quadrature scheme and is unnecessary if an analytical calculation is possible. Considering only the leading term in Eq. (4.7) which is given by $\sigma^2\left\{\Delta W_\mathrm{R}\right\}$ and delivered by the applied estimator (c.f. Sec. 4.3.4) yields the following simplified expression for the uncertainty estimate of $\Delta G^\circ_\mathrm{bind}$:

$$\sigma^2\left\{\Delta G^\circ_\mathrm{bind}\right\} \approx \sigma^2\left\{\Delta W_\mathrm{R}\right\} \tag{4.8}$$

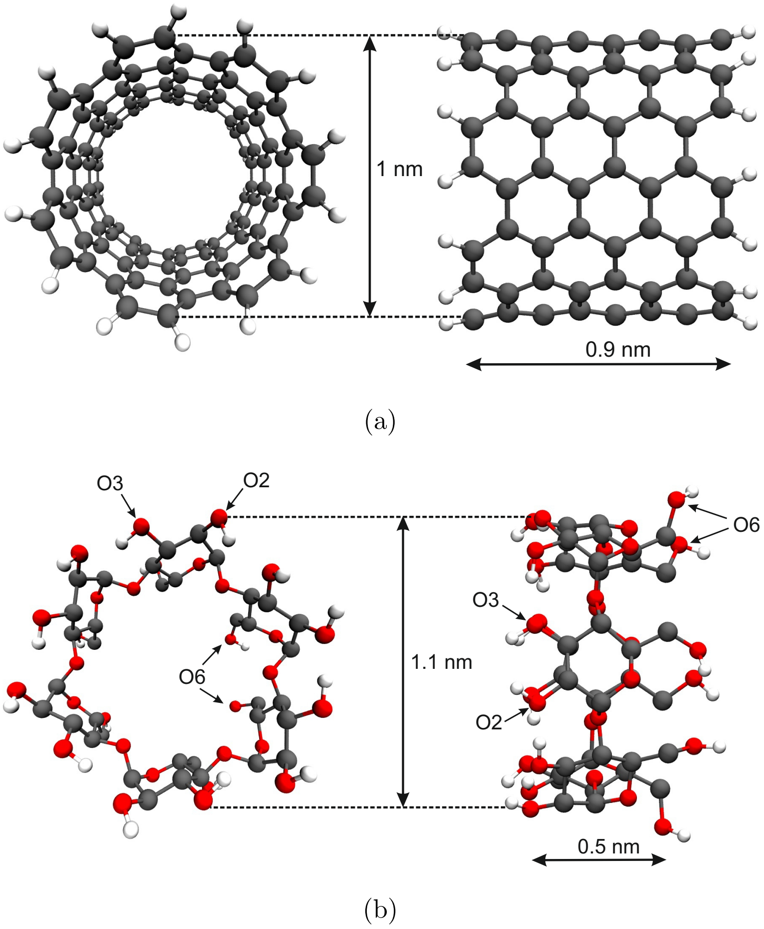
4.3 Methods

4.3.1 Host-Guest Systems

The search for suitable host-guest benchmarks which are simple enough to approach accurately by MD simulations within reasonable time scales yet complex enough to feature properties of protein-ligand systems is an ongoing and non-trivial problem[87,286]. The majority of simulations from the current work were based on a short carbon nanotube (CNT) host without partial charges (c.f. Fig. 4.2). This model system, featuring a hydrophobic, water-free cavity resembles the situation of an ideally symmetric and unpolar host molecule. On the other hand, it allows the effect of molecular "asymmetries" to be studied systematically. Here, such an asymmetry was introduced by distributing charge pairs on the terminal C-H atoms at one side of the CNT. The investigated ligands comprised united-atom models for methane, (elongated) ethane and hexane. The effect of dipolar ligands was modeled by placing a positive and neutralizing negative charge onto covalently bound neighboring carbon atoms in case of polyatomic ligands. To test the validity and transferability of the protocol in case of more realistic systems, it was applied to α-cyclodextrin (αCD, c.f. Fig. 4.2) as a host molecule of practical relevance, complexed with primary alcohols.

4.3.2 Simulation Protocol

The PMFs were constructed from the time series of a single order parameter sampled via umbrella sampling[92,258], similar to the approach proposed by Doudou et al.[281]. As illustrated in Fig. 4.3, the order parameter (ζ) used primarily in this work is given by the projection of the instantaneous separation vector ($\boldsymbol{r}_{\mathrm{HL}}$) between the centers of mass (COM) of the binding partners onto the host's instantaneous symmetry axis ($\boldsymbol{\omega}_{\mathrm{H}}$): $\zeta \equiv \boldsymbol{r}_{\mathrm{HL}} \cdot \boldsymbol{\omega}_{\mathrm{H}} = r_{\mathrm{HL}} \cos\varphi$. Here, the unity vector $\boldsymbol{\omega}_{\mathrm{H}}$ was defined by the connecting line through the geometric centers at both sides of the CNT. φ denotes the angle between $\boldsymbol{\omega}_{\mathrm{H}}$ and $\boldsymbol{r}_{\mathrm{HL}}$. Instead of the centers of mass, two other characteristic reference points of the host and ligand could be used instead (c.f. Sec. 4.4.2). The usage of the COM-COM radial distance ($r_{\mathrm{HL}} = |\boldsymbol{r}_{\mathrm{HL}}|$) itself as order parameter would lead to artifacts around $r_{\mathrm{HL}} = 0$[288]. It would further require to remove the Jacobian contribution from the free-energy profile in order to obtain the PMF[289], as discussed above. Such a Jacobian term, which is of pure entropic nature and accounts for the increase in the accessible configurational area at increasing distances, does not arise when ζ is used instead[281]. Lateral movement of the ligand at every umbrella window was restricted with the aid of a flat-bottom potential acting on the orthogonal displacement ($\rho = r_{\mathrm{HL}} \sin\varphi$, c.f. Fig. 4.3) of the ligand's COM



(a)

(b)

Figure 4.2: Carbon nanotube (CNT) model host (a) and α-cyclodextrin (αCD) molecule (b). The CNT is a (7,7) tube in armchair structure. Nomenclature of αCD-oxygen types according to Ref. 287. Host dimensions are depicted in front (left) and side view (right).

from the host's molecular axis:

$$U_\rho(\rho) = \begin{cases} k_\rho(\rho - \rho_{\mathrm{up}})^n, & \text{if } \rho > \rho_{\mathrm{up}} \\ 0, & \text{otherwise} \end{cases} \tag{4.9}$$

The flat-bottom potential (harmonic ($n = 2$) or quartic ($n = 4$), force constant k_ρ) is activated only when the actual displacement exceeds a certain threshold ρ_{up}. In this case, calculation of $A_{\mathrm{u},\rho}$ according to Eq. (4.5) yields[255]:

$$A_{\mathrm{u},\rho} = \pi\rho_{\mathrm{up}}^2 + \begin{cases} \frac{2\pi}{k_\rho^*} + \pi\rho_{\mathrm{up}}\frac{(2\pi)^{1/2}}{k_\rho^{*\,1/2}}, & \text{if } n = 2 \\ \frac{\pi^{3/2}}{2\,k_\rho^{*\,1/2}} + \pi\rho_{\mathrm{up}}\frac{\Gamma(1/4)}{2\,k_\rho^{*\,1/4}}, & \text{if } n = 4 \end{cases} \tag{4.10}$$

with the reduced restraining force constant $k_\rho^* \equiv k_\rho/RT$ and the Gamma function Γ. If the threshold value ρ_{up} for the flat-bottom potential is chosen to be large enough compared to the size of the host's cavity such that the ligand's dynamic is not affected in the bound state, the term ΔG_ρ in Eq. (4.2) makes no contribution. It should be stressed that while the PMF itself and the terms $\Delta W_\mathrm{R}, \Delta G_\mathrm{V}, \Delta G_\rho$ in Eq. (4.2) are influenced by the restraining parameters n, k_ρ and ρ_up, the estimate for $\Delta G^\circ_\mathrm{bind}$ should be independent when all contributions are evaluated adequately (c.f. Sec. 4.4.1). Major modifications compared to the original approach of Doudou et al.[281] can be summarized as follows: (i) the order parameters used for both the actual PMF calculation and for measuring the ligand's orthogonal movement are defined in a relative manner between ligand and host. Instead of using a particular Cartesian component such as the z-component of the COM-COM separation vector $\boldsymbol{r}_\mathrm{HL}$ with respect to an arbitrary external laboratory coordinate system, we look at projections of $\boldsymbol{r}_\mathrm{HL}$ onto axes of a body-fixed coordinate system which is centered inside the host. The usage of relative order parameters relaxes the requirement of a translationally and/or rotationally restrained host and allows the same approach to be used in case of a fully mobile host molecule without further modifications (c.f. Sec. 4.4.1); (ii) for the majority of ligands, an additional angular or orientational restraint in the form of a harmonic potential

$$U_\theta(\theta) = \frac{k_\theta}{2}(\theta - \theta_0)^2 \tag{4.11}$$

was applied, acting on the angle (θ) between the molecular axes of host ($\boldsymbol{\omega}_\mathrm{H}$) and ligand ($\boldsymbol{\omega}_\mathrm{L}$) in order to suppress flipping of the ligand relative to the host. The molecular axis of the ligand, expressed as unity vector $\boldsymbol{\omega}_\mathrm{L}$, was defined by the connecting line through two peripheric atoms of the ligand. The value of k_θ should be chosen high enough to prevent transitions between different ligand orientations. As in case of the translational restraint, the estimate for $\Delta G^\circ_\mathrm{bind}$ should be independent of the concrete choice of k_θ. In case of a translationally and rotationally restrained host aligned along the z-axis without orientational restraint on the ligand, the approach corresponds to the original setup described in Ref. 281. In this case, the order parameter ζ corresponds to the Cartesian z-component of $\boldsymbol{r}_\mathrm{HL}$ and ρ becomes $\rho = \sqrt{\Delta x^2 + \Delta y^2}$. Here, Δx and Δy denote the orthogonal displacements of the ligand's COM from the central z-axis.

The free energy contributions corresponding to the release of the translational and orientational restraint in the bound state (ΔG_ρ and ΔG_θ in Eq. (4.2)) were each calculated from a sequence of 20 simulations with the bound ligand located at the PMF minimum. The individual simulations were conducted at different scaled force constants $k_\rho(\lambda) = \lambda \cdot k_\rho$ with the scaling parameter λ equally distributed between 0 and 1 (analogously for k_θ). The endpoints correspond to the unrestrained case at $\lambda = 0$ and the actual force constant value as used for umbrella sampling at $\lambda = 1$, respectively. Using the configurations sampled from a particular state λ_i, all possible pairwise potential energy differences $\Delta_\mathrm{ij} U = U_\rho(\lambda_\mathrm{j}) - U_\rho(\lambda_\mathrm{i})$ towards the reference state potential $U_\rho(\lambda_\mathrm{i})$ were evaluated (analogously for U_θ). From these potential energy differences, the free energy calculation was performed via the MBAR estimator. For enhanced sampling, Hamiltonian Replica Exchange between neighboring λ-points was applied with attempted exchanges every 1000 steps.

Initial configurations for the production simulations of each umbrella window were generated within a prior equilibration phase (500 ps per window) in the following manner: starting in the bulk at one side of the CNT, the ligand was sequentially displaced in 0.1 nm increments along a linear path through the cavity, until the unbound ligand was located in the bulk again, but relative to the other side of the CNT. For production, all considered systems were simulated at least for 20 ns per window until converged PMF estimates were obtained. Specifications regarding the applied restraints in the protocol are summarized in Tab. 4.1.

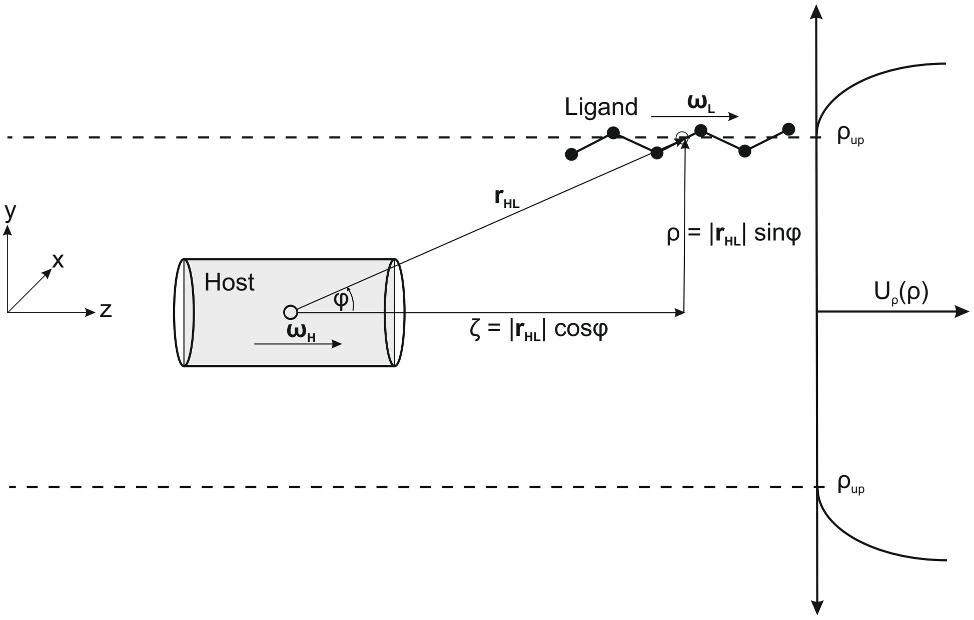


Figure 4.3: Schematic representation of the host-guest system and the relevant collective variables. The orientation of the ligand ($\boldsymbol{\omega}_\mathrm{L}$) may be aligned towards the orientation of the host ($\boldsymbol{\omega}_\mathrm{H}$) by the usage of an orientational restraint acting on the angle θ between $\boldsymbol{\omega}_\mathrm{L}$ and $\boldsymbol{\omega}_\mathrm{H}$ (see main text). φ denotes the angle between $\boldsymbol{\omega}_\mathrm{H}$ and the separation vector between the centers of mass of host and ligand ($\mathbf{r}_\mathrm{HL}$). The chosen order parameter (ζ) is the projection of $\mathbf{r}_\mathrm{HL}$ onto the host's molecular axis $\boldsymbol{\omega}_\mathrm{H}$. When the host itself is aligned along the z-axis of the laboratory coordinate system, as depicted here, the order parameter corresponds to the Cartesian z-component of $\mathbf{r}_\mathrm{HL}$. The ligand's movement orthogonal to the order parameter outside the host is restricted via a flat-bottom potential ($U_\rho(\rho)$) acting on the orthogonal distance (ρ) between the center of mass of the ligand and the molecular axis of the host. The flat-bottom potential is activated when the actual distance ρ exceeds a certain threshold distance (ρ_up), as depicted by the dashed lines.

Table 4.1: Default values for restraints specifying the umbrella sampling protocol as used for the majority of studies in the current work. In case of differing settings, the parameter choice is explicitly given. For all simulations involving a polar CNT, a value of $k_\zeta = 3000$ kJ mol^{-1} nm^{-2} was used for the distance restraint force constant. Lateral translational movement of the ligand (as measured by the orthogonal displacement ρ) was restrained using a flat-bottom potential (c.f. Eq. (4.9)). To restrain the ligand's orientation towards a specific bound state orientation, an orientational restraint acting on the angle θ between the molecular axes of host and ligand was applied (c.f. Eq. (4.11)).

Distance Restraint			
k_ζ	ζ_{min}	ζ_{max}	$\Delta\zeta$
[kJ mol^{-1} nm^{-2}]	[nm]	[nm]	[nm]
500	-2.5	2.5	0.1
Translational Restraint			
k_ρ	ρ_{up}	n	
[kJ mol^{-1} nm^{-n}]	[nm]	[$-$]	
500	0.4	2	
Orientational Restraint			
k_θ	θ_0		
[kJ mol^{-1} rad^{-2}]	[rad]		
500	0.0		

4.3.3 Simulation Code and Parameters

The GROMOS biomolecular force field was applied throughout this work using the 54A7[13] and 53A6$_{\mathrm{GLYC}}$[290] parameter sets for studies based on the CNT and αCD, respectively. The standard atom types 12 and 20 were used to represent the CNT carbon and hydrogen atoms, respectively. All systems were solvated in water based on the three-site simple point charge (SPC) water model[162]. Simulations were conducted under periodic boundary conditions using the leap-frog algorithm[235] for integrating Newton's equations of motion with a time step of 2 fs. The majority of simulations were performed with the GROMACS 2016.4 program package[166,169,291]. In the light of recent publications reporting on the sensitivity of simulation results on the choice of the pairlist algorithm, the electrostatics treatment, the cut-off scheme or other technical details[292–295], complementary simulations were conducted with the GROMOS11 program package (release version 1.5.0)[170–172] which has different recommended settings. In particular, GROMOS is usually used with a reaction field scheme for treating long-range electrostatic interactions. Since this approach is also used by other codes in the context of free energy simulations[296,297], it is interesting to study the effect on a PMF calculation. In the following, separated computational details are given for the two simulation codes.

GROMACS Simulations

Simulations using the particle-mesh Ewald (PME) method[184,185] for treating electrostatic interactions were conducted with the GROMACS 2016.4 program[166,169,291] patched to the free-energy library PLUMED 2.4.2[298] for restraints definition and biasing selected collective variables. The center of mass translation of the computational box was removed every 1000 steps. All bond lengths were constrained using the LINCS algorithm[175,176] with a LINCS-order of 4. The number of iterations to correct for rotational lengthening in LINCS was set to 2. SPC water was constrained using the SETTLE algorithm[174]. Equilibration of solvated energy-minimized systems was performed within a prior 100 ps constant-volume simulation at reference temperature of 300 K, followed by a 1 ns constant-pressure simulation at 300 K and 1 bar for pressure equilibration. Initial velocities were sampled from a Maxwell-Boltzmann distribution at 300 K. During the equilibration phase, both temperature and pressure were controlled by application of the weak coupling scheme[188] with corresponding relaxation times of $\tau_\mathrm{T} = 0.1$ ps and $\tau_\mathrm{p} = 0.5$ ps and an (isotropic) isothermal compressiblity of 4.5×10^{-5} bar^{-1}[189]. For production simulations, the Nosé-Hoover thermostat[177–179] and Parrinello-Rahman barostat[180,181] were applied with corresponding coupling constants of $\tau_\mathrm{T} = 1.0$ ps and $\tau_\mathrm{p} = 2.0$ ps. The solute (comprising the host and ligand molecule) and solvent were coupled to separate heat baths. A Verlet-buffered neighbor list[183] which was updated every 25 steps, was applied for the treatment of short-range electrostatic and van der Waals interactions with potentials shifted to zero at 1.4 nm. The latter were modeled by the Lennard-Jones potential. Analytic dispersion corrections were applied for energy and pressure calculation. Long-range electrostatic interactions were treated with the smooth particle-mesh Ewald (PME) method[184,185] using a real-space cut-off of 1.4 nm with a cubic splines interpolation scheme and a grid spacing of 0.12 nm. In most simulations reported here, the host's orientation was aligned along the z-axis of the simulation box (box dimensions: 3.4 x 3.4 x 12 nm) alongside with a translational restraint (500 kJ mol^{-1} nm^{-2}) to keep its COM close to the box center. The bias on the orientation was realized by an orientational restraint (500 kJ mol^{-1} rad^{-2}) acting on the angle between the host's symmetry axis and the external z-axis. Biased collective variables were written to file every 100 steps.

GROMOS Simulations

Simulations using the Barker-Watts reaction field (RF) scheme[194] for treating electrostatic interactions were conducted with the GROMOS11 program package (release version 1.5.0)[170–172]. The center of mass translation of the computational box was removed every 1000 steps. All bond lengths including the water hydrogen-hydrogen distances were constrained using the SHAKE algorithm[187] with a relative geometric tolerance of 10^{-4}.

Equilibration of solvated energy-minimized systems was performed within a prior 100 ps constant-volume simulation followed by a 1 ns constant-pressure simulation at 300 K and 1 bar for pressure equilibration. During the constant-volume equilibration, temperature was raised by increments of 60 K to the final value of 300 K with initial velocities assigned according to a Maxwell-Boltzmann distribution centered around 60 K. Temperature was maintained close to its reference value by weak coupling[188] to individual external baths for solute and solvent with relaxation times of 0.1 ps. Pressure was held constant at 1 bar by the weak coupling method with a relaxation time of 0.5 ps and an isothermal compressibility of $4.5 \times 10^{-5}\,\mathrm{bar}^{-1}$[189]. A Barker-Watts RF contribution[194] was applied to account for the long-range electrostatic effect beyond the (long-range) cut-off. The relative dielectric permittivity of the dielectric continuum outside the cut-off sphere was set to $\epsilon_{\mathrm{RF}} = 61$, as appropriate for SPC water[299]. In case of van der Waals interactions, no long-range correction was incorporated. Non-bonded interactions were either calculated using a single-range (SR) or a twin-range (TR) cut-off scheme[300]. In case of the TR scheme, interactions within the short-range cut-off radius of 0.8 nm were calculated every time step from a pairlist updated every five steps, while interactions between 0.8 and the long-range cut-off of 1.4 nm were reevaluated for each pairlist update and kept constant in between. In case of the SR scheme using a cut-off radius of 1.4 nm, the pairlist update was performed every time step. On top of the two cut-off schemes, the influence of different construction schemes for the non-bonded pairlist was further investigated, specifying whether the interactions are calculated based on distances between individual atoms (AT) or neutral charge groups (CG). In total, this results in four different non-bonded interaction setups that were tested in conjunction with the RF approach: (i) RF using a twin-range cut-off scheme based on charge groups (RF, TR-CG), (ii) RF using a twin-range and atomistic cut-off scheme (RF, TR-AT), (iii) RF using a single-range cut-off scheme based on charge groups (RF, SR-CG), (iv) RF using a single and atomistic cut-off scheme (RF, SR-AT). An overview of the systems which were treated with the different RF setups is given in Tab. 4.2.

Due to different implementations, the restraints handling was different in the GROMOS package compared to analogue simulations conducted with GROMACS/PLUMED (see above) and can be summarized as follows: (i) alignment of the host along the z-axis of the simulation box was realized by four individual position restraints (1000 kJ mol^{-1} nm^{-2}) imposed for two pairs of peripheric C-atoms located at opposing sides of the CNT; (ii) the coordinates ζ and ρ for measuring progression and lateral movement of the ligand, respectively, were defined by Cartesian components of the separation vector between the COM of the ligand and a fixed anchor point on the z-axis instead of using the separation vector between the COM of ligand and host (c.f. Fig. 4.3). It was verified that the difference in restraining the host's degrees of freedom does not affect the PMF (c.f. Sec. 4.4.1), while the usage of a translated reference point along the z-axis only shifts the whole PMF by the same offset along the range of ζ-values without affecting its shape or the barrier heights. Biased collective variables were written to file every 100 steps.

Table 4.2: Overview of simulated systems based on the reaction field treatment for long-range electrostatics using different cut-off schemes (SR, TR) and pairlist generation schemes (CG, AT) as specified in the main text (c.f. Sec. 4.3.3). All simulations were conducted with the GROMOS MD package. The CNT was aligned along the z-axis of the computational box such that the order parameter ζ corresponds to the z-component of the COM-COM separation vector between the binding partners. For ethane and hexane, no restraint was imposed on the orientation. Labels S1 to S3 refer to different box sizes - S1: 3.4 x 3.4 x 8.0 nm, S2: 4.0 x 4.0 x 8.0 nm, S3: 5.0 x 5.0 x 8.0 nm.

System	SR-AT			TR-AT			SR-CG			TR-CG		
	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3
unp. CNT / CH_4	√	-	-	√	√	-	√	-	-	√	√	-
unp. CNT / unp. C_2H_6	√	-	-	√	√	-	√	-	-	√	√	-
unp. CNT / unp. C_6H_{14}	-	-	-	√	-	-	-	-	-	√	√	-
pol. CNT / CH_4	√	√	-	√	√	√	√	√	-	√	√	√
pol. CNT / unp. C_2H_6	√	√	-	√	√	√	√	√	-	√	√	√

4.3.4 Free Energy Estimation

PMFs were evaluated employing three commonly used free-energy estimators or analysis methods: (i) the Weighted Histogram Analysis Method (WHAM)[240,301,302], (ii) Umbrella Integration (UI)[303–305] and (iii) the Multistate Bennett's Acceptance Ratio (MBAR) estimator[237]. For WHAM, the GROMACS implementation `g_wham`[270] was used, while in case of UI and MBAR, open source python packages[238,306] were employed. While each estimator aims to recover a statistically optimal estimate for the unbiased distribution function of the order parameter, differences become apparent from the underlying working equations and the uncertainty estimates. Detailed information regarding these aspects can be found in the specialized literature cited above. Both WHAM and MBAR result in a coupled set of non-linear equations for the free energy estimates which have to be solved iteratively in a self-consistent manner. This is avoided in the UI approach which was the primarily used estimator throughout this work. In UI, the biased distributions are approximated as normal distributions (fully characterized by the mean and variance) and the restraint forces from each window are combined instead of the unbiased distributions itself. As illustrated in Sec. 4.4.3, the assumption of normal distributions might not be fulfilled for certain conditions depending on the molecular system and simulation protocol. Analytic expressions for PMF uncertainties corresponding to the UI method involve a segment-based analysis (similar to block averaging) for mean and variance of the sampled biased distributions and follow from repeated application of error propagation as described in detail in Ref. 305. The resulting uncertainty over the interval $[\zeta_\mathrm{a}, \zeta_\mathrm{b}]$ refers to the 95% confidence interval such that the presented PMFs are reported in the form $\Delta W_\mathrm{R}(\zeta_\mathrm{b}; \zeta_\mathrm{a}) \pm 1.96\sqrt{\sigma^2 \left\{\Delta W_\mathrm{R}(\zeta_\mathrm{b}; \zeta_\mathrm{a})\right\}}$[305]. ζ_a denotes the minimal value of the order parameter (left border) and ζ_b some running upper value (right border). In that sense, the error bar represents a cumulative estimate with respect to a chosen reference point (ζ_a), resulting in larger error bars for increasing values of the order parameter $\zeta > \zeta_\mathrm{a}$[305].

4.4 Results

The results as presented in the following were obtained from systematic series of studies with the objective to analyze the influences of (i) restraining the host's degrees of freedom, (ii) restraining the ligand's degrees of freedom via translational and orientational restraints, (iii) the choice of reference points as used in the restraining setup, (iv) the treatment of electrostatic interactions (PME vs. RF) and (v) the free energy estimator. Issue (iv) also includes influences of the used cut-off scheme (SR vs. TR) as well as the underlying pairlist-generation scheme (AT vs. CG) in case of simulations based on the RF approach (c.f. Sec. 4.3.3). Except for the paragraphs considering different approaches for long-range electrostatics, all reported PMFs refer to simulations based on the PME approach. To separate the various influences, we started with united-atom methane binding to the completely symmetric and unpolar CNT host before studying polyatomic unpolar ligands. To investigate issues associated with intrinsically asymmetric systems, complexity was further increased by considering the binding of unpolar as well as dipolar ligands to a CNT with a polar pore mouth at one side. The consequences with respect to the calculation of the standard binding free enthalpy according to Eq. (4.2) are elucidated. For several cases, the PMF-derived estimates for $\Delta G^\circ_\mathrm{bind}$ were compared with results from alchemical double decoupling. Details about the double decoupling approach can be found in the appendix. Finally, the application to α-cyclodextrin (αCD) complexed with primary alcohols is presented. Special focus is given to the occurrence of computational artifacts which manifest as flawed PMFs featuring a significant offset between the two flat bulk regions. Specific parameters as used in the umbrella sampling setup are summarized in Tab. 4.1.

4.4.1 Unpolar CNT / Methane

This section reports PMFs between united-atom methane and the unpolar CNT. Since it was found that all three estimators (WHAM, UI, MBAR) yield indistinguishable PMFs within error bars, only the UI results will be reported in the following.

Restraining the Host

In the integrals of the PMF expression in Eq. (4.1), six external degrees of freedom, corresponding to overall rotation and translation of the host-guest complex were integrated out. In practical applications it is often desirable to restrain the position and orientation of the host molecule in order to limit the size of the computational box. Such a position restraint may influence the potential of mean force if conformational fluctuations of the host molecule are suppressed. For the CNT host studied here, we confirmed that the restraints acting on the external degrees of freedom of the host molecule do not influence the PMF. Therefore, five different setups were compared: (i) no external restraints applied for the host, (ii) a three-dimensional position restraint acting on the host's COM to keep it close to the box center, (iii) application of an axial restraint to keep the host aligned along the z-axis, (iv) a three-dimensional position restraint on the host's COM combined with an axial restraint (combination of (ii) and (iii)) and (v) three-dimensional position restraints acting on every host atom. In setup (i), the host-guest complex as a whole can translate and rotate in three dimensions. In setup (ii), the host (and thus the complex as a whole) can not translate, but it can rotate without hindrance. In setup (iii) in contrast, the axial restraint on the host restricts the rotation of the host-guest complex, but it can still translate in three dimensions. The setups (iv) and (v) hamper both, the translational and the rotational movement of the host molecule. Setup (v) even restricts a rotation of the host around its axis which is possible for setup (iv). From the perspective of a moving observer located in the host's COM, all setups are identical as long as the host's internal dynamic is not affected by the external restraints, which is only the case in setup (v). While the setups become more restrictive from (i) to (v), the system size (and thus the computational effort) is increased considerably for setup (i) and (ii), since a uniform simulation box is required in contrast to (iii), (iv) and (v). It should be stressed, that due to the relative formulation of the order parameter (guest relative to the host) and auxiliary quantities such as the angle φ and the orthogonal distance ρ (c.f. Fig. 4.3), identical restraints specifications between host and guest can be used for all setups without modifications.

It was found that all five setups yield indistinguishable PMFs within uncertainties (c.f. Fig. 4.4). The fact that even the very restrictive setup (v) has no effect on the PMF can be probably attributed to the rather rigid structure of the CNT cavity. For other more flexible host molecules, the effect of restraining so many degrees of freedom might be more pronounced and should be avoided as outlined above. We conclude, that the way we restrained the host's external degrees of freedom (overall rotation and translation) does not affect the calculated PMFs, as expected from theory such that the PMF artifacts reported below have a different cause. Unless explicitly stated otherwise, all results presented in the remainder of the article refer to setup (iv).

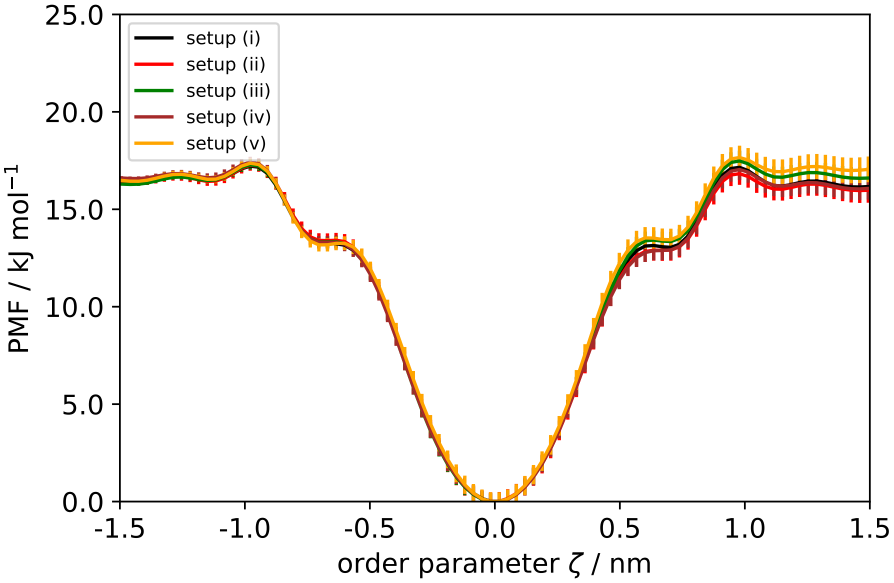


Figure 4.4: Effect of different restraining setups (i)-(v) as used to restrain the external degrees of freedom of the CNT host (c.f. Sec. 4.4.1). PMFs refer to the system methane / unpolar CNT.

Restraining the Ligand's Lateral Movement

While the flat-bottom potential U_ρ used for limiting the ligand's lateral movement influences the PMF, the final estimate of the standard binding free enthalpy $\Delta G^\circ_{\text{bind}}$ should be independent. Fig. 4.5 shows the PMFs for methane / CNT obtained for different restraining parameters in terms of the exponent n, the threshold ρ_{up} and force constant k_ρ (c.f. Eq. (4.9)). All PMFs show perfect symmetry as expected for such a system with a global minimum at $\zeta = 0.0$ nm, corresponding to configurations where methane is located at the cavity center. Different parameter combinations basically scale the PMFs while the overall shape remains very similar. Here, the usage of smaller threshold parameters (at constant k_ρ) as well as higher force constants (at constant ρ_{up}) leads to higher absolute numbers of ΔW_{R}. Corresponding estimates of $\Delta G^\circ_{\text{bind}}$ for every PMF according to Eq. (4.2) are summarized in Tab. 4.3. Since no orientational restraint was applied in this case, the terms ΔG_Ω and ΔG_θ in Eq. (4.2) make no contribution. While the estimates for ΔW_{R} and $A_{\text{u},\rho}$ are strongly influenced by the parameters of U_ρ and as such also ΔG_{V}, the bound length l_{b} is virtually independent. The free energy contribution associated with the orthogonal translational restraint in the bound state (ΔG_ρ) is close to zero due to the naturally restricted conformational space accessible to the bound ligand inside the host's cavity. In accordance with theoretical expectation, all PMFs yield very similar estimates for $\Delta G^\circ_{\text{bind}}$ independent of the choice of orthogonal restraining parameters (c.f. last column in Tab. 4.3). In addition, the PMF-based estimates are in reasonable agreement with the value of $\Delta G^{\circ}_{\text{bind}} = -13.0$ kJ mol^{-1} as obtained from alchemical double decoupling (c.f. Tab. C1).

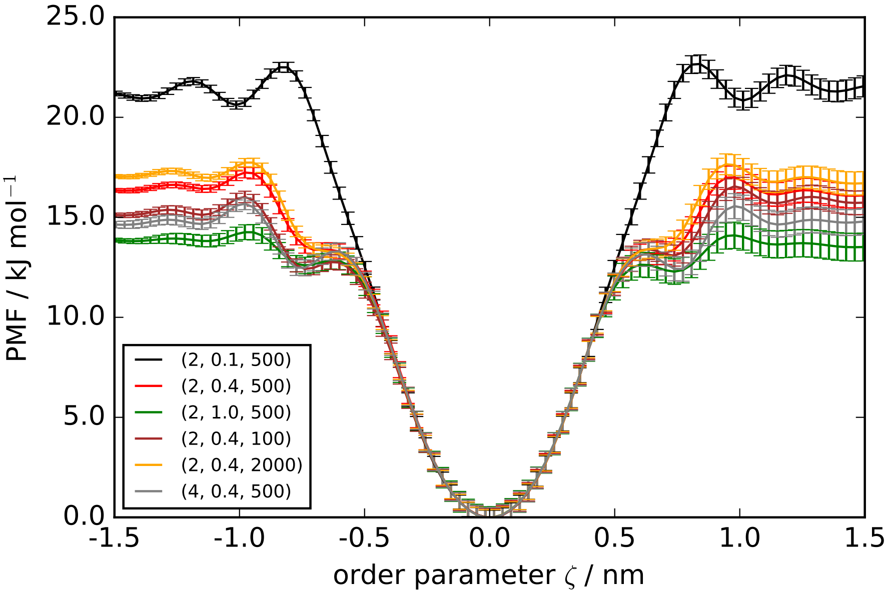


Figure 4.5: Effect of the orthogonal translational restraint on the PMF for methane / unpolar CNT. Restraining parameters $(n, \rho_{\text{up}}\,[\text{nm}], k_{\rho}\,[\text{kJ mol}^{-1}\ \text{nm}^{-n}])$ as used for the flat-bottom potential U_{ρ} according to Eq. (4.9), are given by number triplets in the legend.

Treatment of Electrostatic Interactions

Fig. 4.6 shows the influence of different treatments for long-range electrostatics (PME vs. RF) alongside with different cut-off schemes (SR vs. TR) and pairlist generation schemes (CG vs. AT) on the PMF. As can be seen, all setups yield very similar PMFs. In Ref. 307, a system size dependence of the PMF for ion association was observed in case of simulations based on the RF treatment. Therefore, additional simulations using box sizes of different x- and y-dimensions were conducted for the two TR-setups (c.f. Tab. 4.2) as well as for the PME treatment. In all cases, no system size dependence could be observed (data not shown).

Lesson Learned

For the CNT / methane system, consistent PMFs were obtained leading to binding free enthalpies within maximal statistical bounds of ±1.5 kJ mol^{-1}, regardless of how the host and the ligand's lateral movement was restrained (c.f. Tab. 4.3). This conservative estimate for the maximal error encompasses the PMF uncertainty as delivered by the UI estimator as well as the spread of $\Delta G^{\circ}_{\text{bind}}$ values obtained from the different setups. This

Table 4.3: Influence of the translational restraint settings on the calculated standard binding free enthalpy $\Delta G^{\circ}_{\mathrm{bind}}$ for united-atom methane / unpolar CNT (c.f. Sec 4.4.1). Corresponding PMFs are depicted in Fig. 4.5. First three columns specify the parameters used for the flat-bottom potential U_{ρ} (c.f. Eq. (4.9)). Calculations of l_{b}, $A_{\mathrm{u},\rho}$ and $\Delta G^{\circ}_{\mathrm{bind}}$ were performed according to Eq. (4.4), Eq. (4.10) and Eq. (4.2), respectively. The contribution of the translational restraint in the bound state ΔG_{ρ} was calculated using the MBAR estimator from a sequence of simulations in the bound state with force constants k_{ρ} varying from zero to the final value as given in the table. Error estimates refer to the UI result.

Setup								
n	ρ_{up}	k_{ρ}	ΔW_{R}	l_{b}	$A_{\mathrm{u},\rho}$	ΔG_{V}	ΔG_{ρ}	$\Delta G^{\circ}_{\mathrm{bind}}$
[–]	[nm]	[kJ mol^{-1} nm^{-n}]	[kJ mol^{-1}]	[nm]	[nm^2]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]
2	0.1	500	-21.37 ± 0.54	0.3827	0.1184	8.98	-0.06	-12.44 ± 0.54
2	0.4	500	-16.27 ± 0.62	0.3832	0.7565	4.35	-0.13	-12.05 ± 0.62
2	1.0	500	-13.80 ± 0.72	0.3869	3.7291	0.35	-0.14	-13.59 ± 0.72
2	0.4	100	-15.51 ± 0.61	0.3874	1.1569	3.27	-0.50	-12.74 ± 0.61
2	0.4	2000	-16.93 ± 0.60	0.3835	0.6217	4.84	-1.03	-13.12 ± 0.60
4	0.4	500	-14.84 ± 0.66	0.3856	1.3047	2.98	-0.64	-12.50 ± 0.66

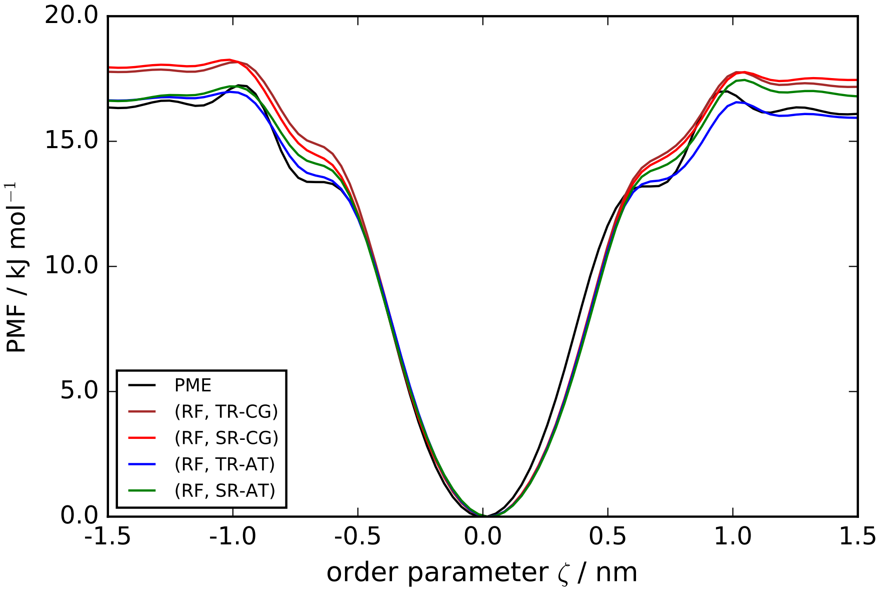


Figure 4.6: Effect of the treatment of electrostatic interactions (PME vs. RF), the cut-off scheme (SR vs. TR) and the pairlist generation scheme (CG vs. AT) on the PMF for unpolar methane / unpolar CNT (c.f. Sec. 4.3.3). Bounds for statistical uncertainties are below 1.0 kJ mol^{-1} and have been omitted in the interest of clarity.

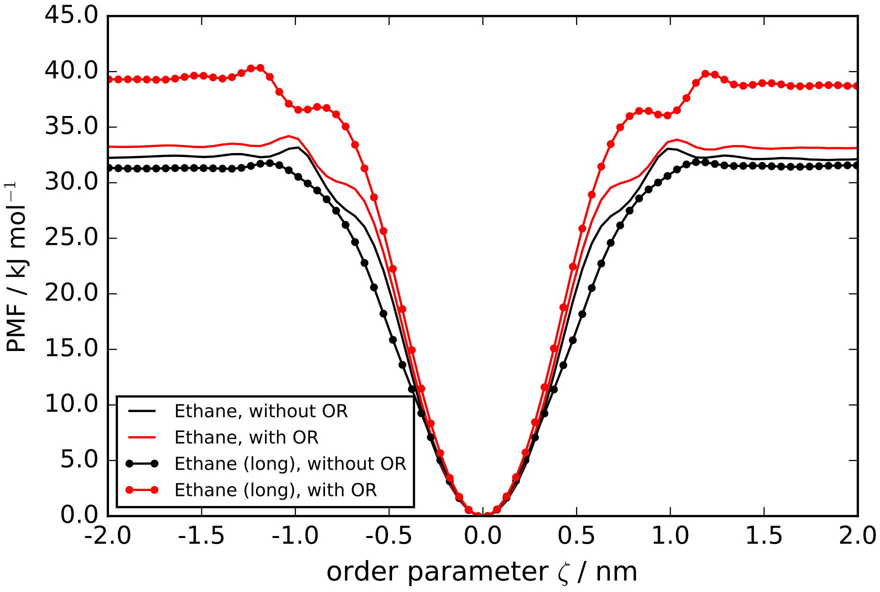
distribution of $\Delta G^{\circ}_{\mathrm{bind}}$ values also emphasizes however, that even for such a simple system, no perfect agreement can be expected. The treatment of electrostatic interactions and the pairlist generation scheme have an effect on ΔW_{R} on the order of ±1.5 kJ mol^{-1}. These results are an important basis to judge the artifact reported in Sec. 4.4.3.

4.4.2 Unpolar CNT / Multiatomic Ligand

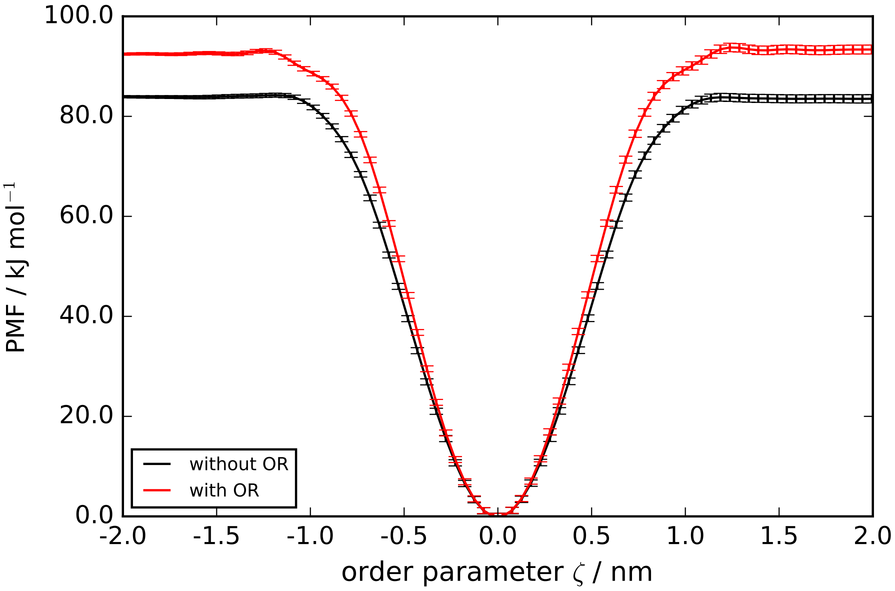
This section reports the PMFs for the unpolar CNT host complexed with different multiatomic unpolar ligands. Here, rigid diatomic ligands in the form of ethane and a modified model with increased bond length were studied as well as hexane. In contrast to (ordinary) ethane, the elongated variant (in the remainder denoted as "elongated ethane") is unable to rotate inside the CNT cavity once it is bound. This ligand selection enables to study the impact of the ligand's flexibility and rotational degrees of freedom inside the binding pose. Since it was found that all three estimators (WHAM, UI, MBAR) yield indistinguishable PMFs within errors bars, only the UI results will be reported in the following.

Restraining the Ligand's Orientation

In practice it can often be essential to restrain not only the translational movement of the ligand but also its orientation towards the host molecule (c.f. Sec. 4.4.4). Fig. 4.7 shows the PMFs for (a) ethane, elongated ethane and (b) hexane, obtained from the setup with (red curves) and without (black curves) orientational restraint. As can be seen, the restraint on the ligand's rotation leads to higher absolute numbers of ΔW_{R}. Comparison of the diatomic ligands shows that this increase is more pronounced for increasing bond lengths. Tab. 4.4 contains the calculated estimates for $\Delta G^{\circ}_{\mathrm{bind}}$. For each ligand, two estimates are provided, corresponding to the setup with and without orientational restraint. As revealed by the data, the application of an orientational restraint in case of ethane has a marginal effect on ΔW_{R} but the free energy contribution from releasing this restraint in the bound state is the highest for all ligands. The fact that this contribution is almost identical for the bound and unbound state shows that the confinement inside the host's cavity has no significant effect on the populated ligand orientations in this case, as expected. For elongated ethane and hexane, which are not able to rotate in the bound state, the free energy gain of releasing the restraint is much smaller. The good agreement of the corresponding values for $\Delta G^{\circ}_{\mathrm{bind}}$ from simulations with and without orientational restraint confirms consistency between the setups. Results from double decoupling which was performed for ethane and elongated ethane, was also found to be in good accordance with the PMF-based estimates (c.f. Tab. C1). We conclude that the effect of an orientational restraint included in the simulation protocol with respect to the calculation of $\Delta G^{\circ}_{\mathrm{bind}}$ is captured adequately by the terms ΔG_{Ω} and ΔG_{θ} in Eq. (4.2). Therefore, a variation of the restraining force constant k_{θ} was not performed in this work.



(a)



(b)

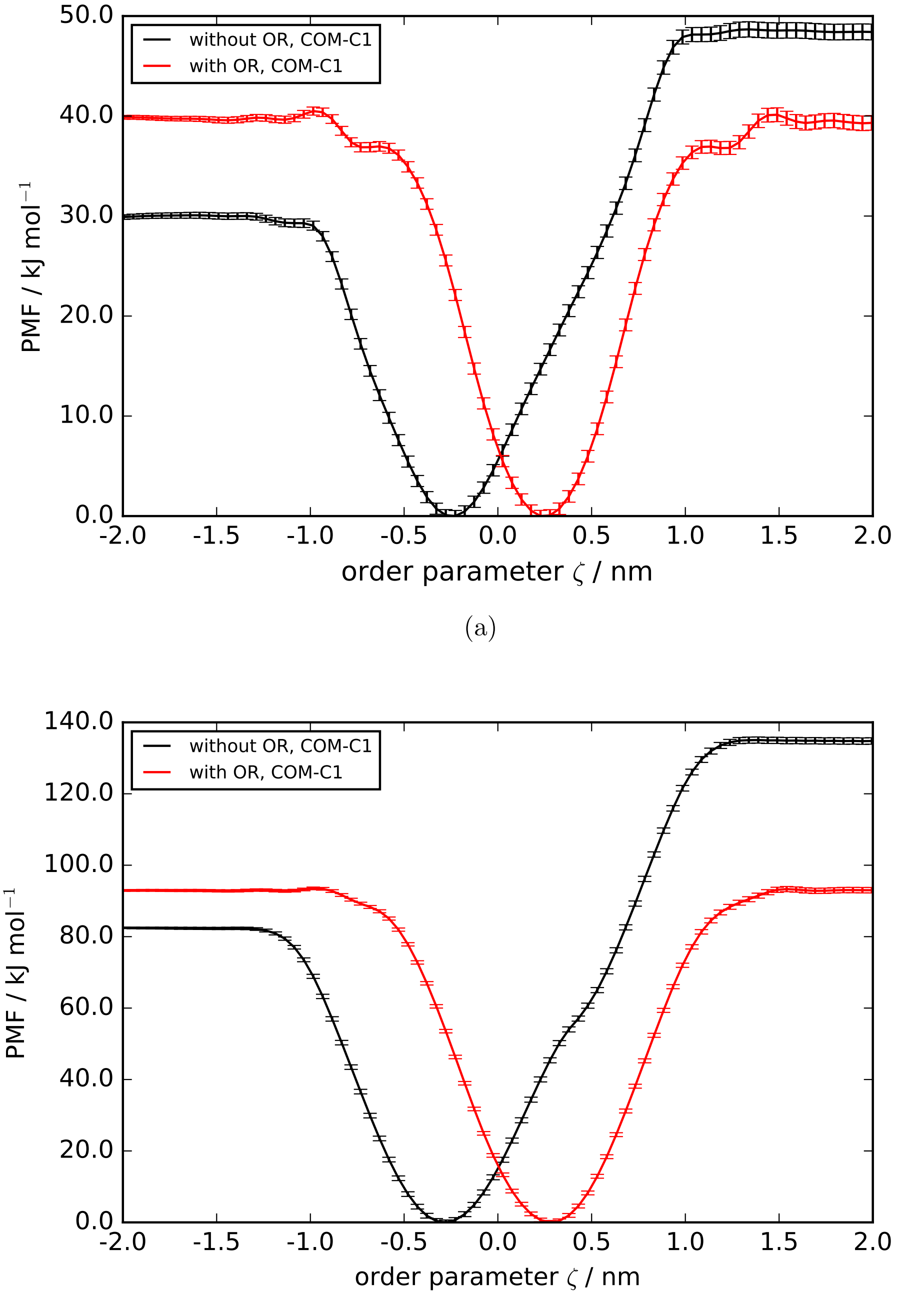
Figure 4.7: Effect of a restrained ligand orientation on the PMFs for (a) unpolar (elongated) ethane / unpolar CNT and (b) unpolar hexane / unpolar CNT. Red and black curves correspond to the setup with and without orientational restraint (OR), respectively. Error bars in graph (a) have been omitted in the interest of clarity.

Choice of Restraining Reference Points

The decision which (pseudo) atoms to choose in the host and ligand molecule to serve as reference or anchor points for the applied distance restraint in the umbrella sampling simulations is often not clear a priori. Though the centers of mass might be an intuitive choice (and were selected for the majority of studies of the current work), other choices might appear more suitable in practical application[281]. Fig. 4.8 (a) and (b) show the PMFs for elongated ethane / CNT and hexane / CNT, respectively, as obtained when a peripheric carbon atom was picked as reference point in the ligand. The COM of the CNT was chosen as reference point within the host as has been the case hitherto. Every graph contains two free energy profiles, corresponding to the PMF evaluated with (red curves) and without (black curves) orientational restraint. In the setup lacking an orientational restraint, a substantial free energy offset between 20 (elongated ethane) and 55 kJ mol^{-1} (hexane) is present in the PMF. The reason for this offset lies in the differences of the sampled configurational space at the two pore mouths. Depending on which part of the ligand is buried (the part with or without the anchor atom for the distance restraint), the configurational space accessible to the partly bound ligand is quite different. The estimation of $\Delta G^{\circ}_{\text{bind}}$ from such a PMF would lead to very different results depending on which branch of the PMF would have been taken as a basis for the analysis. In contrast, if the COM of the ligand is chosen as anchor atom, the rotational behavior of the ligand is symmetric at both CNT ends and no PMF offset is present, even when the ligand's orientation is not restrained (c.f. Fig. 4.7). However, as demonstrated in Fig. 4.8, even in case of such an "unfortunate" choice of anchor points, the PMF offset can be eliminated through the usage of an orientational restraint. Comparison of the corresponding profiles of Fig. 4.8 and Fig. 4.7 obtained from the setup including an orientational restraint (but different reference points in the ligand), shows that the PMFs are identical except for a marginal shift along the order parameter axis which will not affect the estimate for $\Delta G^{\circ}_{\text{bind}}$.

Treatment of Electrostatic Interactions

Fig. 4.9 shows the influence of the treatment for long-range electrostatics (PME vs. RF) alongside with the pairlist generation scheme (CG vs. AT) on the PMFs for two systems: (a) ethane / CNT and (b) hexane / CNT. In addition, the effect of different cut-off schemes (SR vs. TR) was tested in case of ethane / CNT. None of the cases included an orientational restraint. As can be seen, all setups yield almost indistinguishable PMFs for ethane / CNT, whereas for hexane / CNT, the two RF setups yield a slightly narrower PMF well compared to the PME result. Referring to the effect on the binding free enthalpy, such a different shape only affects the calculation of the bound length l_{b} (c.f. Eq. (4.4)) leading to a marginal discrepancy in the order of ±0.5 kJ mol^{-1} compared to the PME-based estimate. As in case of methane / CNT, no system size dependence for the PMFs could be observed (data not shown).



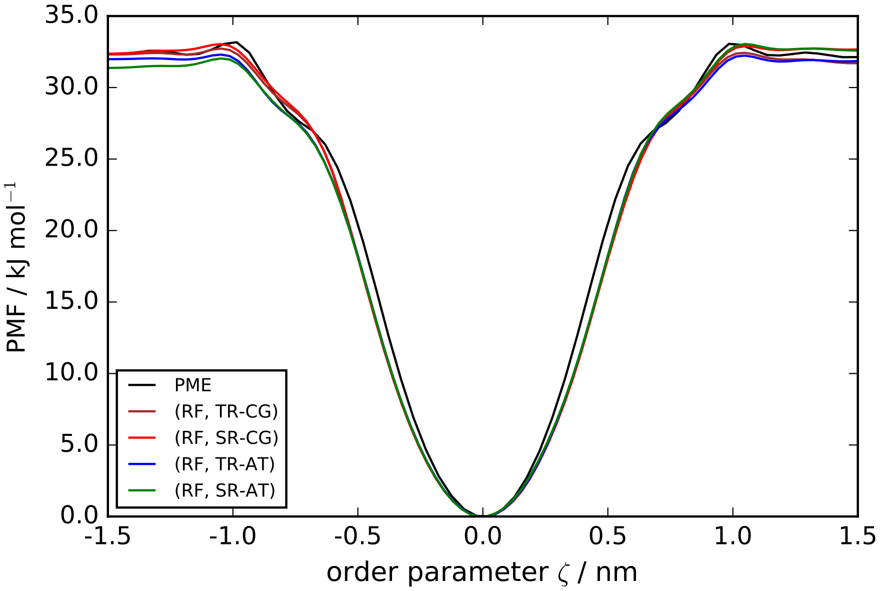
(a)

(b)

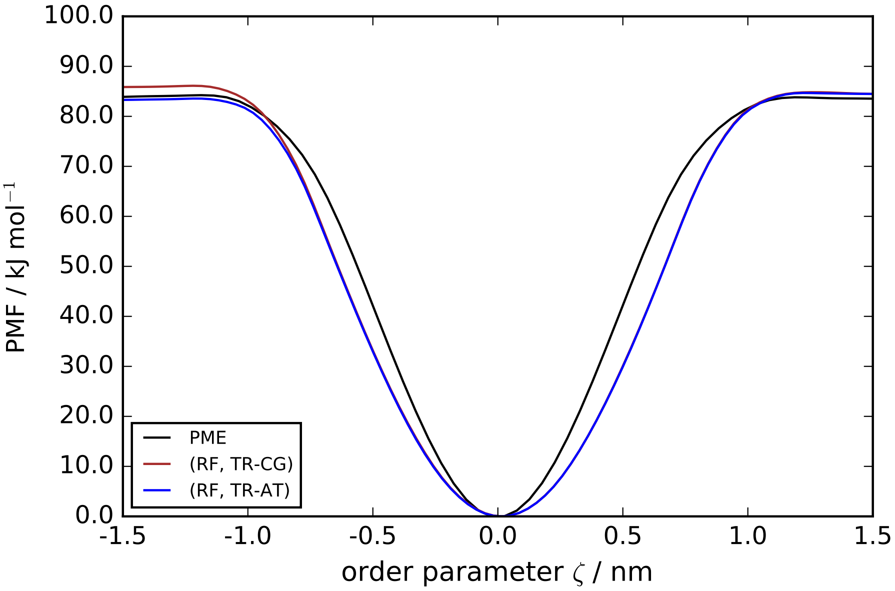
Figure 4.8: Effect of the choice of reference points for the distance restraint on the PMFs for (a) unpolar elongated ethane / unpolar CNT and (b) unpolar hexane / unpolar CNT. Here, the COM of the CNT and the C1 carbon atom of the ligand was chosen as reference points. Red and black curves correspond to PMFs obtained from the setup with and without orientational restraint (OR), respectively.

Lesson Learned

For the binding of symmetric unpolar multiatomic ligands to the unpolar CNT, the change in rotational entropy upon binding is included in the PMF if no orientational restraint



(a)



(b)

Figure 4.9: Effect of the treatment of electrostatic interactions (PME vs. RF), the cut-off scheme (SR vs. TR) and the pairlist generation scheme (CG vs. AT) on the PMF for (a) unpolar ethane / unpolar CNT and (b) unpolar hexane / unpolar CNT (c.f. Sec. 4.3.3). No orientational restraint was applied to the ligands. Bounds for statistical uncertainties are below 1.0 kJ mol^{-1} and have been omitted in the interest of clarity.

is used. The two setups (with and without orientational restraint) lead to standard binding free enthalpies which are indistinguishable within statistical uncertainties. An orientational restraint is required however, if the anchor points for the umbrella distance restraint in the ligand and in the CNT are chosen in such a way, that the configurational space accessible to the partly bound ligand at the both cavity entrances are different. The treatment of electrostatic interactions and the pairlist generation scheme have a marginal effect on ΔW_{R} on the order of ±1 kJ mol^{-1}.

4.4.3 Polar CNT / Unpolar Ligand

This section reports PMFs for the association of a polar CNT with different unpolar ligands. The polar CNT was modeled by distributing balancing charges to terminal pairs of C-H-atoms at one side of the CNT (C-atoms: -0.5 e, H-atoms: +0.5 e where ”e” denotes the elementary charge). Every balancing pair of C-H atoms was assigned to one neutral charge group in case of simulations based on the RF approach for long-range electrostatics in combination with the CG pairlist scheme. In contrast to the unpolar systems treated so far, care has to be taken in order to avoid a bias due to the applied PMF analysis method. This issue is discussed explicitly in a separate subsection.

Impact of the Host's Polarity

Fig. 4.10 shows PMFs for a set of unpolar ligands (methane, (elongated) ethane, hexane) binding to the polar CNT. No orientational restraint was imposed on the ligand. In contrast to the previous examples corresponding to the binding to an unpolar CNT, the resulting PMFs are highly asymmetric featuring a considerable barrier to be overcome by the ligand at the polar entrance of the CNT. This barrier which is caused by the modified water structure in proximity to the polar mouth, makes the binding path through that particular side energetically unfavorable.

Treatment of Electrostatic Interactions

Fig. 4.11 shows the influence of the treatment for long-range electrostatics (PME vs. RF) alongside with the pairlist generation scheme (CG vs. AT) and cut-off scheme (SR vs. TR) on the PMFs for two systems: (a) methane / polar CNT and (b) unpolar ethane / polar CNT. In case of ethane, no orientational restraint was applied. As can be seen, PMFs based on an atomistic interaction scheme (TR-AT, SR-AT) show a smaller well (measured by ΔW_{R}) and barrier at the polar entrance compared to analogue simulations based on charge groups (TR-CG, SR-CG). The combination of the RF approach with an atomistic interaction scheme also shows higher resemblance with the PME solution as judged by the value of ΔW_{R}. We found that typically much longer simulations times (more than 40 ns per window) were required compared to PME-based simulations (typically 20 ns per window were sufficient) in order to achieve converged estimates. No significant impact

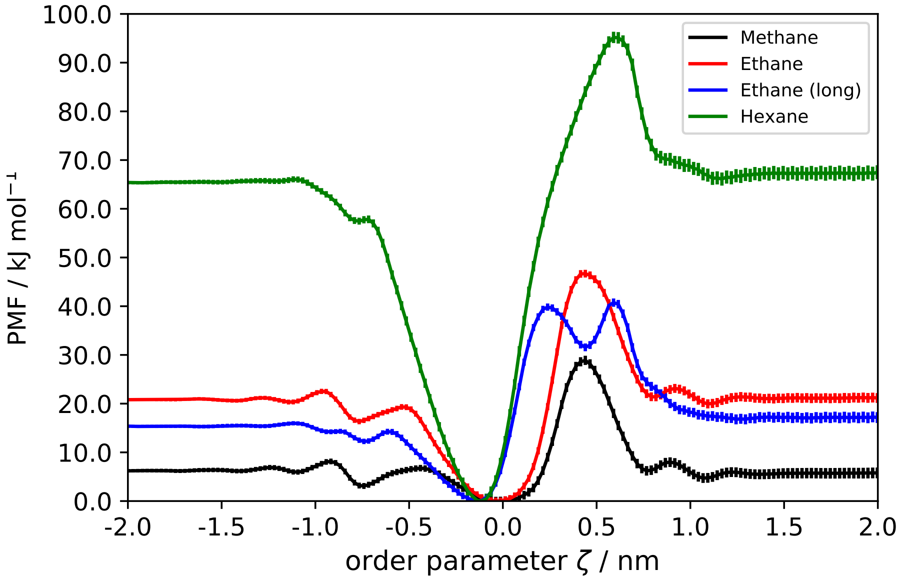
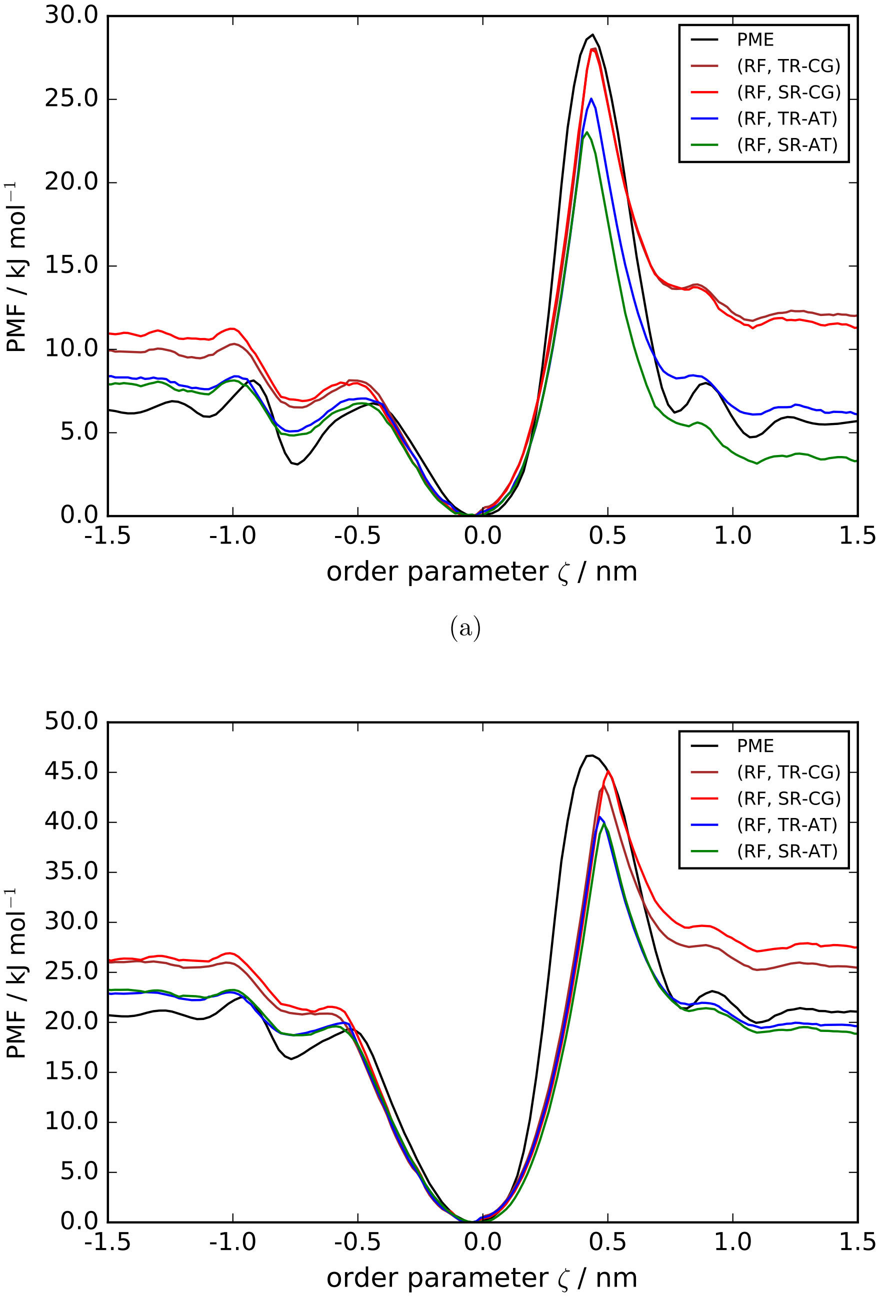


Figure 4.10: PMFs for the association of a polar CNT with different unpolar ligands. No orientational restraint was imposed on the ligand. Long-range electrostatic interactions were treated with the PME method. PMFs were estimated via the UI method.

of the underlying cut-off schemes (SR vs. TR) could be observed. As in the previous sections, no systematic system size dependence was found.

Impact of the Free Energy Estimator

For the considered systems unpolar ligand / polar CNT it was found that artifacts in the form of a PMF offset as detected previously in another context (c.f. Sec. 4.4.2), can be introduced by the analysis method. Fig. 4.12 shows a comparison between PMFs as obtained from different estimators (UI, WHAM, MBAR) for the example of unpolar ethane / polar CNT based on the (RF, TR-CG) setup. As can be seen, an offset between the flat bulk water regions of around 7 kJ mol^{-1} is present in the profile obtained from the UI method, which significantly exceeds bounds due to the statistical uncertainty. Its origin can be explained by means of the sampled biased distribution functions of the order parameter ζ (c.f. Fig. 4.13). As noted previously in Sec. 4.3.4, a central assumption in the UI approach is that the biased distributions can be approximated as Gaussian distributions. In Fig. 4.13 (a) it can be seen that the distribution sampled from window close to $\zeta = 0.5$ nm at the polar CNT entrance differs from the rest and is non-Gaussian in shape. The set of distributions from corresponding PME-based simulation in contrast, does not contain such a window (c.f. Fig. 4.13 (b)). Such an offset was exclusively observed for simulations based on the reaction field treatment (including different ligands and combinations of cut-off and pairlist generation schemes) which is more susceptible for



(a)

(b)

Figure 4.11: Effect of the treatment of electrostatic interactions (PME vs. RF), the cut-off scheme (SR vs. TR) and the pairlist generation scheme (CG vs. AT) on the PMF for (a) methane / polar CNT and (b) unpolar ethane / polar CNT (c.f. Sec. 4.3.3). No orientational restraint was applied for ethane. PMFs were estimated via the WHAM method. Error bars have been omitted in the interest of clarity.

cut-off artifacts in structural solvation properties[309] but not for PME-based simulations. Nonetheless, we stress that this artifact is only indirectly caused by the electrostatics treatment but actually results from application of an estimator to a situation for which

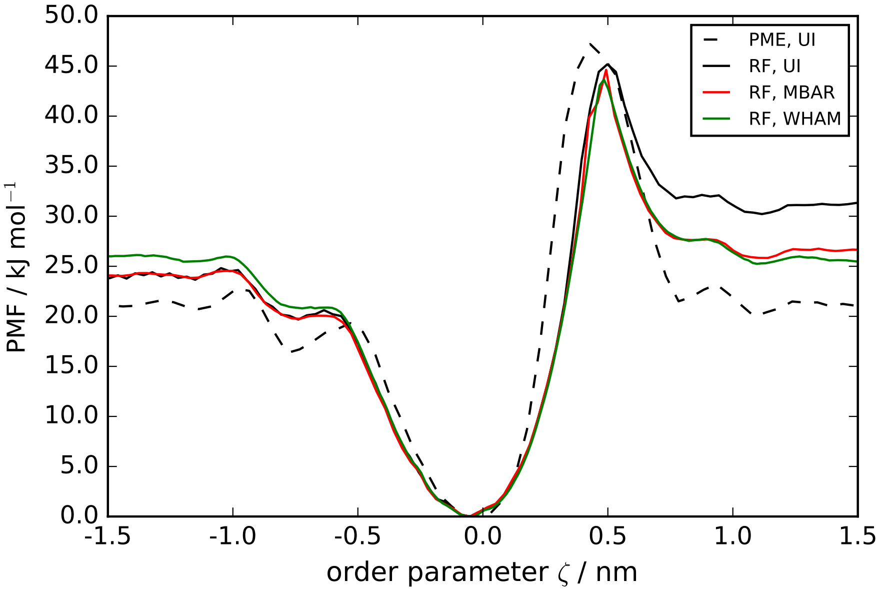


Figure 4.12: Influence of the free energy estimator (UI vs. MBAR vs. WHAM) on the PMF for unpolar ethane / polar CNT. Electrostatics treatment refers to the (RF, TR-CG) setup as described in the main text. The profile obtained from simulations using PME and evaluated via UI is shown for comparison (black dashed line). No orientational restraint was imposed on the ligand.

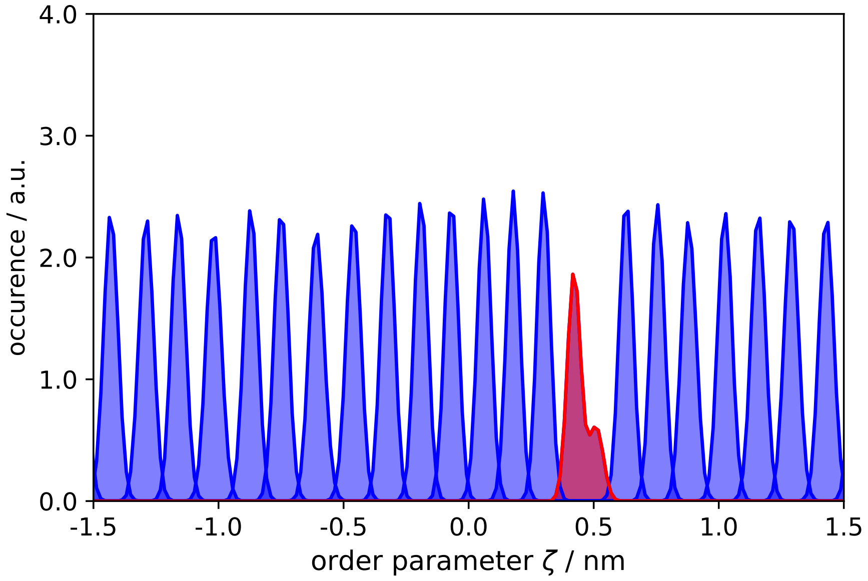
it was not designed for. The usage of a higher force constant for the umbrella distance restraint might probably remedy such a bias.

Lesson Learned

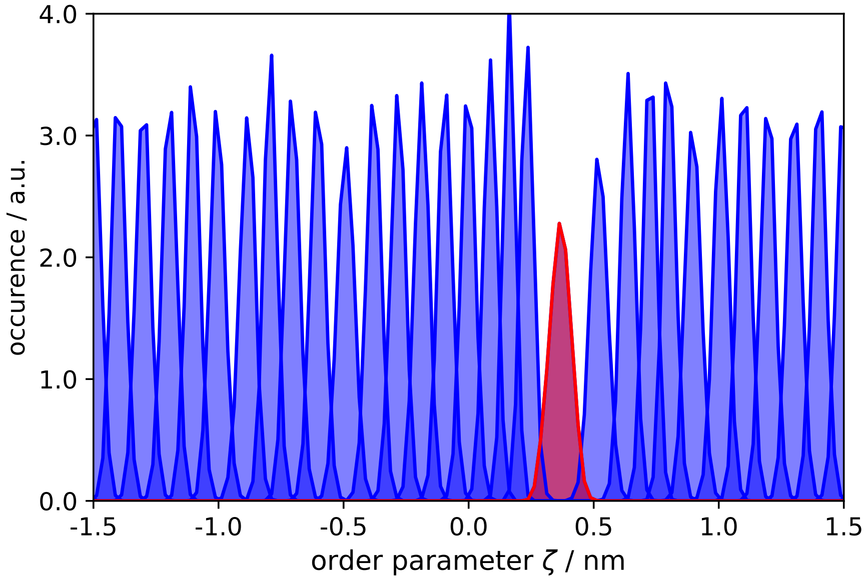
The examples show that also in the presence of considerable polar interactions between host and solvent, neither the differences in the treatment of electrostatic interactions nor in schemes for the cut-off or pairlist generation affect the estimated PMFs systematically. The artifact caused by the UI estimator demonstrates the benefit to compare different analysis methods on the same data set. Furthermore, if the UI estimator is used, the shape of the sampled distributions should be checked.

4.4.4 Polar CNT / Dipolar Ligand

This section reports PMFs for the association of a polar CNT with different dipolar ligands based on (elongated) ethane and hexane. The modeling of the polar CNT was described in the previous section. Dipolar ligands were modeled in a similar way by distributing a pair of balancing partial charges to the peripheric pair of covalently bound carbon atoms (C1-atom: +0.5 e, C2-atom: -0.5 e where ”e” denotes the elementary charge). For all simulations considered in this paragraph, the PME treatment for long-range electrostatics



(a)



(b)

Figure 4.13: Influence of electrostatics treatment on sampled biased distributions of the order parameter ζ along the considered path for unpolar ethane / polar CNT. (a): RF-based simulations using the TR-CG setup, (b): PME-based simulations. The distribution close to $\zeta = 0.5$ nm is highlighted in red to support the discussion in the main text. The abbreviation a.u. refers to arbitrary units.

was utilized. Since it was found that the different estimators (WHAM, UI, MBAR) yield indistinguishable PMFs within errors bars, all profiles reported in the following refer to the UI result.

Sampling of Ligand Orientations

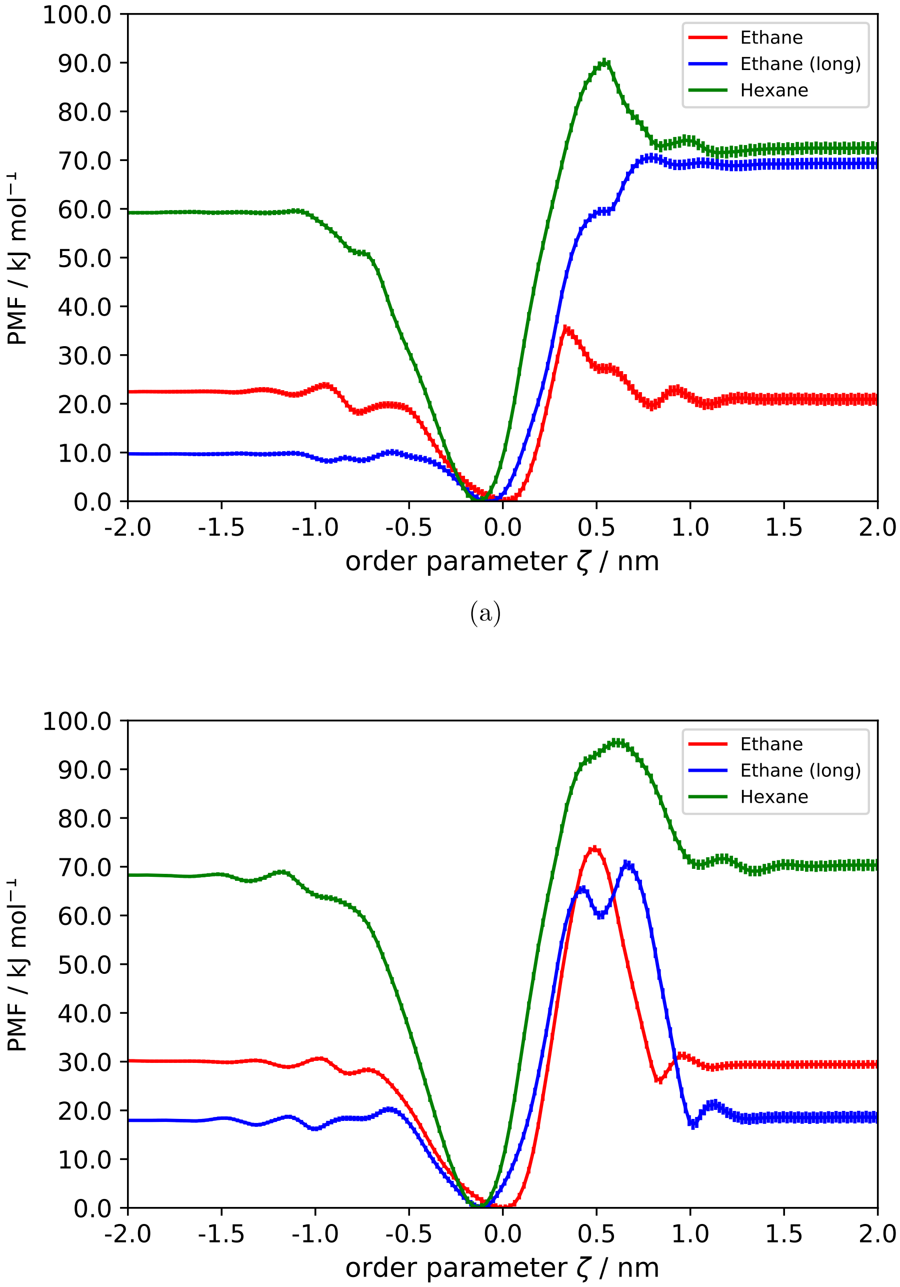
In contrast to the systems treated so far, two distinct binding configurations with different binding affinities can be distinguished. The bound configuration for which the positively charged ligand head (C1-atom) is facing (away from) the negatively charged C-atoms of the CNT is denoted as Conf. 1 (Conf. 2). PMFs for dipolar ethane, elongated ethane and hexane binding to the polar CNT are depicted in Fig. 4.14. Profiles in (a) and (b) were obtained without and with imposed orientational restraint on the ligand, respectively. For all simulations, the dipolar ligand was initially prepared in Conf. 1. Significant differences become apparent from comparison with corresponding profiles in Fig. 4.10 where the ligands "feel" the influence of the polar CNT only in an indirect manner mediated by the solvent. Fig. 4.14 (a) reveals substantial PMF offsets of 10 and 60 kJ mol^{-1} for hexane and elongated ethane, respectively, both of which are unable to rotate inside CNT. For ethane in contrast, which can rotate inside the CNT due to its small size, no offset is present. As demonstrated by Fig. 4.14 (b), such an offset can be removed for all considered ligands through inclusion of an orientational restraint in the simulation protocol.

Lesson Learned

The examples illustrate that in case of asymmetric ligands binding to an asymmetric host (which is probably the most common case in practice), the biased distributions sampled in umbrella windows outside the binding site in which the ligand is free to rotate does not fit to the biased distributions sampled in windows for the bound state in which the binding pose is prescribed. This misfit can be illustrated by excluding configurations exhibiting the "wrong" orientation from the analysis which reduces the offset considerably (data not shown). We point out that excluding states from the analysis was done just to support our findings, and is not meant to be a suitable method to avoid offsets. The use of an orientational restraint is therefore mandatory in such cases unless the umbrella sampling is combined with Hamiltonian Replica Exchange as discussed in the following section.

4.4.5 Cyclodextrin / Alcohols

This section reports PMFs for the association of α-cyclodextrin (αCD) with two different primary alcohols (1-butanol, 1-dodecanol). Results for further alcohols were reported in our previous work[255]. Since it was found that the different estimators (WHAM, UI, MBAR) yield indistinguishable PMFs within errors bars, all profiles reported in the following refer to the UI result.



(a)

(b)

Figure 4.14: PMFs for the association of a polar CNT with different dipolar ligands. All ligands were initially prepared in the same binding configuration (Conf. 1). Profiles in (a) and (b) refer to the setup without and with imposed orientational restraint on the ligand, respectively.

Sampling of Ligand Orientations

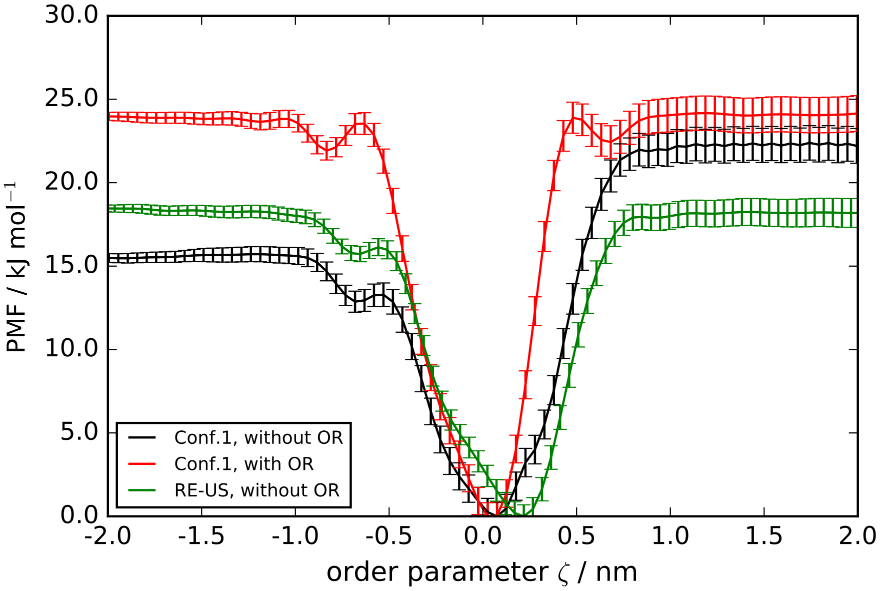
As in the case of dipolar ligands binding to the polar CNT, two different binding configurations can be distinguished which will be denoted as Conf. 1 and Conf. 2 according to

Refs. 255 and 310. Fig. 4.15 (a) and (b) show the PMFs for 1-butanol and 1-dodecanol binding to αCD, respectively. Each graph contains two PMFs, corresponding to the setup with (red curve) and without (black curve) orientational restraint with the ligand bound to αCD (Conf. 1). The third profile (green curve) in both graphs corresponds to the PMFs as obtained from umbrella sampling combined with Replica Exchange (RE-US)[311]. In RE-US, the Hamiltonians of neighboring windows defined by the individual values for the bias centers are allowed to swap after predefined time instances, based on the Metropolis-Hastings criterion. Here, an exchange was attempted every 1000 steps. For RE-US simulations, no orientational restraint was applied. A significant offset is visible for the PMFs lacking an orientational restraint. As observed previously in case of the dipolar ligand / polar CNT system (c.f. Sec. 4.4.4), this offset can be remediated by restricting the ligand orientation. The fact that also the RE-US approach (without orientational restraint) yields an offset-free PMF, further demonstrates that this artifact is caused by a bias introduced when prescribing the binding pose in standard umbrella sampling.

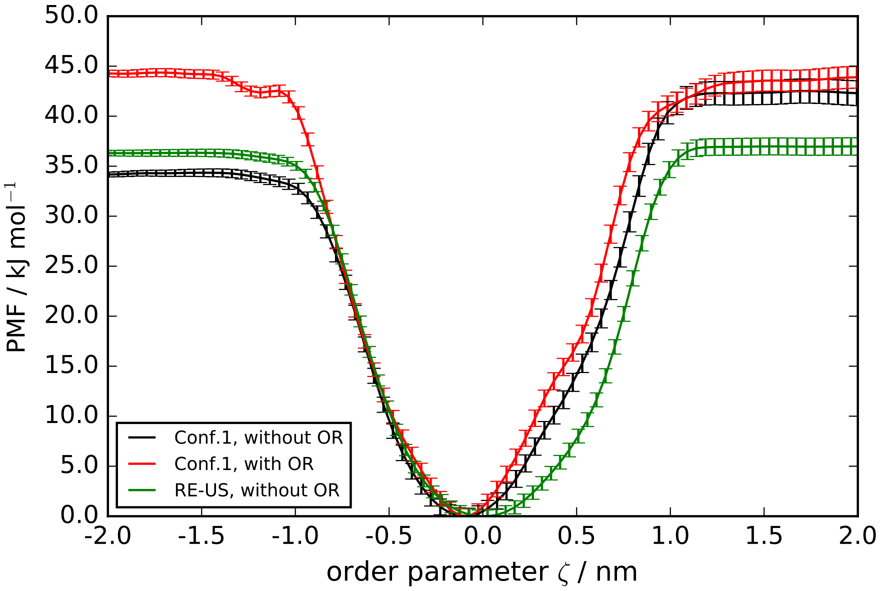
The estimate of the binding free enthalpy as obtained from the protocol including an orientational restraint, corresponds to one particular binding configuration (in this case Conf. 1) and should be therefore denoted as $\Delta G^{\circ}_{\mathrm{bind,Conf.1}}$. For comparison with experiments which measure a configurational average, the binding free enthalpy for the second binding configuration (obtained from additional simulations and denoted as $\Delta G^{\circ}_{\mathrm{bind,Conf.2}}$) can be combined with $\Delta G^{\circ}_{\mathrm{bind,Conf.1}}$ via exponential averaging[312]:

$$\Delta G^{\circ}_{\mathrm{bind}} = -RT \ln \left(\mathrm{e}^{-\Delta G^{\circ}_{\mathrm{bind,Conf.1}}/RT} + \mathrm{e}^{-\Delta G^{\circ}_{\mathrm{bind,Conf.2}}/RT} \right) \tag{4.12}$$

In RE-US simulations (without orientational restraint) in contrast, both binding configurations are sampled and the corresponding PMF can not be attributed to either Conf. 1 or Conf. 2 but already represents a configurational average. Thus, the estimate for $\Delta G^{\circ}_{\mathrm{bind}}$ inferred from such a PMF can be directly compared with the corresponding experimental value without the need of additional simulations. However, this gain in efficiency might be offset by an overhead in terms of hardware resources and (depending on the system) computation time for reaching convergence. Here, it was found that in case of butanol 20 ns per window were sufficient to obtain converged PMFs while 140 ns per window were required for dodecanol. For standard umbrella sampling including an orientational restraint in contrast, 20 ns per window were found to be sufficient for all systems, at least in case of simulations based on the PME treatment for long-range electrostatics as mentioned previously. The good agreement between the $\Delta G^{\circ}_{\mathrm{bind}}$ estimates obtained from the setup including an orientational restraint and the RE-US simulations (c.f. Tab. 4.5) indicates that the RE scheme not only removes the PMF artifact but especially samples both binding configurations with the correct weighting. Moreover, the results were found to be in good agreement with corresponding estimates from double decoupling (c.f. Fig. 5 in Ref. 255).



(a)



(b)

Figure 4.15: PMFs for (a) 1-butanol and (b) 1-dodecanol binding to αCD (Conf.1)[255,310]. Red and black profiles refer to the setup with and without imposed orientational restraint (OR), respectively. The PMF as obtained from the RE-US approach (green curve) represents a configurational average of Conf. 1 and Conf. 2.

Lesson Learned

The examples considering the binding of primary alcohols to the αCD-host show that in case of multiple binding configurations which are separated by significant energy barriers, artifacts in the form of a PMF offset might occur if the umbrella sampling protocol only includes a distance restraint. This artifact was already observed for the artificial model system dipolar ligand / polar CNT (c.f. Sec. 4.4.4) and is caused by insufficient sampling of ligand orientations in the binding site. The simulation protocol can be modified in two ways for such situations: (i) restraining the ligand's orientation to a specific binding configuration. For each binding configuration, one can calculate a binding free enthalpy and combine the distinct estimates during post processing (c.f Eq. (4.12)). (ii) combination of umbrella sampling with Replica Exchange to allow sampling of multiple ligand orientations in the binding site. In this case no orientational restraint is required and the estimate for $\Delta G^{\circ}_{\mathrm{bind}}$ represents a configurational average.

4.5 Discussion

4.5.1 Enforcing PMF Periodicity

In Sec. 4.4.2, 4.4.4 and 4.4.5, it was shown that PMF offsets due to an unfortunate choice of restraining reference points or insufficient sampling of ligand configurations in the bound state can be eliminated through application of a restraint acting on the ligand's orientation or Replica Exchange in the simulation protocol. However, to obtain more realistic PMF estimates also in case of existent simulation data, sampled from non-optimized protocols, Hub et al.[270] proposed another workaround. In their approach which focuses on the post processing estimation, a modified version of the WHAM algorithm was developed. Their method, denoted as `g_wham` as part of the GROMACS program collection, offers the calculation of integrated autocorrelation times (IACT) for reducing the bias from limited sampling as well as constraints for enforced FEC periodicity and / or symmetry. It should be kept in mind that imposing such a constraint will yield a solution for the free energy profile which - by design - satisfies the state function property by preventing an offset. On the other hand, it clearly does not reveal any information about the origin of this artifact, nor does it solve the actual sampling problem. Moreover, such an artificially generated FEC (and in consequence the derived estimate for $\Delta G^{\circ}_{\mathrm{bind}}$) might deviate significantly from the "true" profile, one would obtain in the absence of any sampling issues. To study the effect of the periodicity constraint, the simulation data for butanol / αCD (Conf. 1) without restraining the butanol orientation were reevaluated using `g_wham`. Resulting profiles with and without enforced periodicity are shown in Fig. 4.16. Estimation using standard WHAM without enforced periodicity yields a sig-

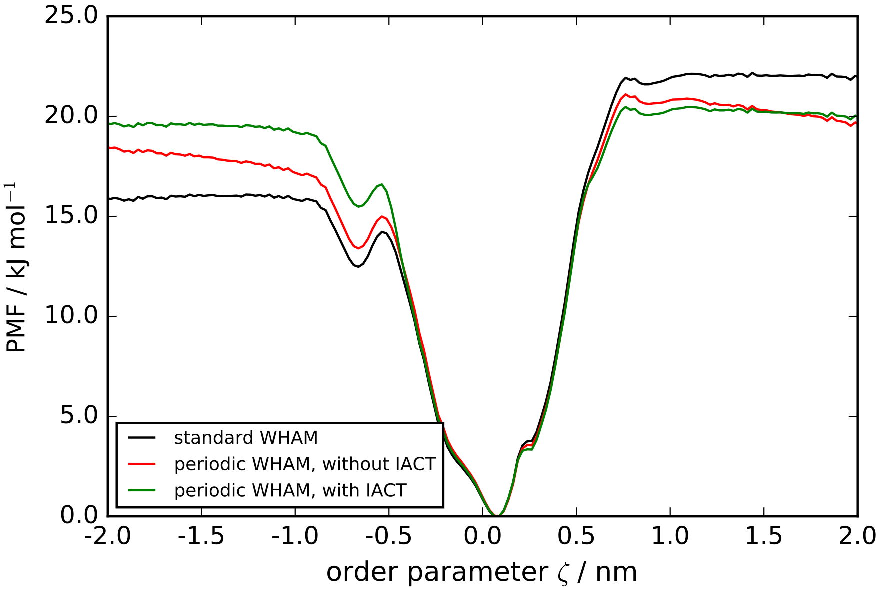


Figure 4.16: Effect of enforced periodicity (periodic) and integrated autocorrelation times (IACT) on the PMF for the system 1-butanol / αCD (Conf. 1). No orientational restraint was applied to the ligand. Calculation was performed using the g_wham method[270]. Error bars were neglected for clarity. Standard WHAM calculation (black curve) refers to estimation without IACT correction and without periodicity constraint.

nificant offset as shown beforehand (c.f. Fig. 4.15 (a)). As can be seen, the application of the periodicity constraint yields identical values at the end points of the considered order parameter interval but it induces artificial slopes in the bulk water regions. This artifact was also described in the original publication[270] where it was ascribed to the neglect of locally different IACT. Therefore, additional analysis was performed by incorporation of the distribution of local IACT into the analysis in addition to the enforced periodicity. The resulting periodic and IACT-corrected profile indeed shows flat bulk water regions. Estimation of the standard binding free enthalpy from the periodic / periodic and IACT-corrected profiles yields -13.9 / -14.7 kJ mol^{-1}, respectively, compared to -10.0 kJ mol^{-1} as obtained from standard WHAM estimation from the setup including an orientational restraint (c.f. Tab. 4.5). This examples illustrates that in the context of binding free enthalpy calculations, the usage of artificially constrained profiles might give a reasonable estimate for $\Delta G^{\circ}_{\text{bind}}$, however, one should be aware of that such a value does not purely reflect the precision of the force field. If very precise estimates are required (either for the profile itself or $\Delta G^{\circ}_{\text{bind}}$), we advise to focus on the elimination of possible sampling issues in the simulation protocol (if system complexity allows it) and to use non-constrained estimation.

4.5.2 Influence of the Host's / Ligand's Flexibility

If the host molecule is able to adopt multiple conformations, a bias might be introduced caused by the selection of initial conformations of the host or the method for generating starting configurations of the umbrella windows. As found by You et al.[288] from studies of βCD complexes, significantly different PMF depths can be obtained depending on the initial host conformation unless simulation time was sufficient. For such cases, discrepancies of the estimated binding free enthalpy compared to results from unbiased direct counting might be expected. Due to the insensitivity of the adopted CNT conformations upon ligand binding alongside with the insensitivity of the PMFs towards increasingly restrictive restraining setups (c.f. Sec. 4.4.1), we do not expect such a bias in this case. For further validation, a modified CNT was studied featuring decreased barriers for proper and improper dihedrals compared to the standard model. Despite increased conformational flexibility, the resulting PMF obtained from the association with hexane (data not shown) was identical with the profile as shown in Fig. 4.9 (b). For simulations based on αCD, we conclude from the good agreement between the PMF-based estimates for $\Delta G^{\circ}_{\mathrm{bind}}$ and the corresponding results from double decoupling[255,310] as well as direct counting[257] that simulation time was sufficient in order to remove any possible bias due to the initial host conformations. Moreover, the force field used in the present study does not show multiple conformations for αCD[91]. Considering host molecules which tend to undergo significant conformational changes upon ligand binding, the incorporation of a conformational restraint to bias the host conformation close to the bound state conformation might be advantageous[272]. Moreover, the ligand conformation could be biased analogously which might be of practical value for speeding up convergence, especially for very flexible ligands. The impact of such a conformational restraint with respect to the calculation of $\Delta G^{\circ}_{\mathrm{bind}}$ can be calculated rigorously[272]. In this case, Eq. (4.2) has to be complemented by the free energy contribution from rigidification of the non-complexed host (and / or unbound ligand) and the contribution from releasing the conformational restraint from the complexed host (and / or bound ligand) again. To obtain accurate results for this process, the force field has to capture the relative energies of the different conformers very accurately[313]. As judged by the good agreement for the $\Delta G^{\circ}_{\mathrm{bind}}$ estimates obtained from the PMF and double decoupling in case of αCD / dodecanol, we conclude that no conformational restraint is required for the flexible ligands considered in this work.

4.6 Conclusions

In this article, we studied the evaluation of one-dimensional potentials of mean force (PMF) of host-guest system obtained via umbrella sampling. A carbon nanotube (CNT) and α-cyclodextrin (αCD) were chosen as idealized model systems for pore- or channel-like protein host molecules featuring a hydrophobic cavity. A robust simulation protocol for the calculation of standard binding free enthalpies from such a PMF was established. From systematic studies of different CNT / ligand combinations of increasing complexity, we could identify distinct computational artifacts that may occur in the PMF calculation. Such artifacts which show up as PMF offset between the two flat bulk water regions prohibit an unambiguous estimation of the binding free enthalpy and have not been studied in detail so far. It was found that despite an identical manifestation, three different origins for PMF offsets can be distinguished: (i) an unfortunate choice of reference points for the umbrella distance restraint; (ii) a misfit in probability distributions between bound and unbound umbrella windows in case of multiple binding modes; (iii) offsets introduced by the UI estimator due to non-Gaussian-shaped biased distribution functions. It is important to distinguish these origins from possible primary reasons such as insufficient overlap between neighboring umbrella windows (which is especially critical when estimation is performed with WHAM) or insufficient sampling time. Neither the introduction of additional windows nor the extension of simulation time per window will eliminate the PMF artifacts in these cases. It was shown that offsets due to (i) and (ii) can be eliminated by either restraining the ligand orientation close to the bound state orientation or through combination of the umbrella sampling setup with Replica Exchange (RE-US). Application of two-dimensional umbrella sampling by incorporation of a second biased coordinate such as the orientational angle θ, might be an alternative to the application of restraints that may provide insight into the free energy surface at the rim region[268,269,314]. Offsets resulting from the analysis method can be identified by comparing PMF results from different estimators (UI, MBAR, WHAM). Such a comparison which serves as consistency check is always recommended. We note that comparative simulations for αCD / alcohol systems conducted with the CHARMM36 all-atom force field also lead to PMF offsets if the ligand orientation was not restrained (c.f. Fig. C1). This illustrates that the detected artifacts are force-field independent. Regarding the influence of the simulation protocol, it can be expected that artifacts due to issues (i) and (ii) also occur for alternative PMF-based protocols such as Forward Flux Sampling if the ligand orientation is not preserved or proper sampling of multiple orientations can not be guaranteed.

Table 4.4: Calculated standard binding free enthalpies $\Delta G^{\circ}_{\mathrm{bind}}$ for the binding of unpolar ethane, elongated ethane and hexane to the unpolar CNT (c.f. Sec 4.4.2). Two rows of data are associated with every ligand, corresponding to the setup with and without orientational restraint (OR). Corresponding PMFs are depicted in Fig. 4.7. Calculations of l_{b}, ΔG_{V} and ΔG_{Ω} were performed as described in Sec. 4.2. The joint contribution of the translational and orientational restraint in the bound state ($\Delta G_{\rho} + \Delta G_{\theta}$) was calculated using the MBAR estimator from a sequence of simulations in the bound state with force constants k_{ρ} and k_{θ} varying from zero to the final values as specified in Tab. 4.1. The estimate of $\Delta G^{\circ}_{\mathrm{bind,Conf.1}}$ as obtained from the setup including an orientational restraint corresponds to one distinct binding configuration and was corrected by a symmetry term of $-RT \ln 2$[284,308] to obtain $\Delta G^{\circ}_{\mathrm{bind}}$ in case of elongated ethane and hexane which are unable to rotate inside the CNT cavity in the absence of an orientational restraint (c.f. Sec. 4.4.5). Error estimates refer to the UI result.

System	Setup	ΔW_{R}	l_{b}	ΔG_{V}	ΔG_{Ω}	$\Delta G_{\rho} + \Delta G_{\theta}$	$\Delta G^{\circ}_{\mathrm{bind,Conf.1}}$	$\Delta G^{\circ}_{\mathrm{bind}}$
		[kJ mol^{-1}]	[nm]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]
Ethane	No OR	-32.16 ± 0.87	0.2989	4.97	0.00	0.00	-	-27.19 ± 0.87
	OR	-33.23 ± 0.74	0.2925	5.03	14.95	-14.41	-27.66 ± 0.74	-27.66 ± 0.74
Long Ethane	No OR	-31.38 ± 0.85	0.2864	5.08	0.00	0.00	-	-26.30 ± 0.85
	OR	-38.97 ± 0.81	0.2707	5.22	14.95	-5.65	-24.44 ± 0.81	-26.17 ± 0.81
Hexane	No OR	-83.65 ± 0.89	0.1995	5.98	0.00	0.00	-	-77.67 ± 0.89
	OR	-92.99 ± 0.92	0.1954	6.03	14.95	-4.72	-76.72 ± 0.92	-78.45 ± 0.92

Table 4.5: Calculated standard binding free enthalpies $\Delta G^{\circ}_{\text{bind}}$ for the binding of 1-butanol (BTL) and 1-dodecanol (DDL) to αCD (c.f. Sec. 4.4.5). The results as obtained from the setup with restrained ligand orientation (OR) are compared with corresponding results from Replica Exchange (RE) umbrella sampling which yield a configurationally averaged $\Delta G^{\circ}_{\text{bind}}$. Corresponding PMFs are depicted in Fig. 4.15. Calculations of l_{b}, ΔG_{V} and ΔG_{Ω} were performed as described in Sec. 4.2. The joint contribution of the translational and orientational restraint in the bound state $\Delta G_{\rho} + \Delta G_{\theta}$ was calculated using the MBAR estimator from a sequence of simulations in the bound state with force constants k_{ρ} and k_{θ} varying from zero to the final values as specified in Tab. 4.1. The estimate for $\Delta G^{\circ}_{\text{bind}}$ from the setup including an orientational restraint as reported in the last column follows from exponential averaging of the values $\Delta G^{\circ}_{\text{bind,Conf.1}}$ and $\Delta G^{\circ}_{\text{bind,Conf.2}}$ according to Eq. (4.12). Error estimates refer to the UI result.

System	Setup	Conf. X	ΔW_{R}	l_{b}	ΔG_{V}	ΔG_{Ω}	$\Delta G_{\rho} + \Delta G_{\theta}$	$\Delta G^{\circ}_{\text{bind,Conf.X}}$	$\Delta G^{\circ}_{\text{bind}}$
			[kJ mol^{-1}]	[nm]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]
BTL	OR	1	-24.17 ± 1.11	0.2218	5.72	14.93	-6.50	-10.02 ± 1.11	-13.86 ± 1.11
	OR	2	-26.92 ± 1.11	0.2017	5.96	14.95	-7.24	-13.25 ± 1.11	
	RE	1,2	-18.39 ± 0.90	0.3002	4.96	0.00	0.00	-	-13.43 ± 0.90
DDL	OR	1	-44.14 ± 1.11	0.2982	4.98	14.93	-7.01	-31.24 ± 1.11	-33.10 ± 1.11
	OR	2	-43.94 ± 1.11	0.2988	4.97	14.95	-7.49	-31.50 ± 1.11	
	RE	1,2	-36.66 ± 0.89	0.4842	3.77	0.00	0.00	-	-32.89 ± 0.89

Chapter 5

Statistical Mechanical Considerations on the Calculation of the Binding Free Energy from Molecular Simulations

This section intends to provide statistical-mechanical foundations and derivations of the central working equations used in the previous chapter 4 (compare Eq. (4.1), (4.2)). The nomenclature is mostly adopted from Ref. 279.

Thermodynamic Description of Binding Equilibria

Suppose an aqueous solution of receptor molecules A and ligand molecules B which form a non-covalently bound complex C in 1:1 stoichiometry: $A + B \rightleftharpoons C$. The special case of a dimerization process can be treated in the same way with $B = A$. Chemical equilibrium among all involved species imposes[148]:

$$\mu_A + \mu_B = \mu_C \tag{5.1}$$

where the chemical potential μ_i of each species i is of the form:

$$\mu_i = \mu_i^\circ + RT \ln\left(\frac{c_i \gamma_i}{c^\circ}\right) \tag{5.2}$$

c_i denotes the generic concentration of molecule type i which can be expressed in terms of different concentration scales such as molarity, molality or mole fraction. μ_i° and c° denote the chemical potential and the concentration associated with a particular thermodynamic standard or reference state. Thermodynamic deviations towards the chosen standard state are assessed by the dimensionless activity coefficient γ_i. Activity coefficients can be estimated via different routes, including equations of state[315], approaches based on modeling the excess Gibbs energy[315] or molecular simulations[316]. The latter requires the calculation of the solvation free enthalpy of i at two concentrations, one conducted at the reference state and one at the concentration of interest c_i. Possible reference states are the ideal gas or the real pure substance (i.e. based on Raoult's law). If (i) some or all of the pure components A, B and C exist as a solid or gas at the considered thermodynamic conditions, and (ii) the corresponding concentrations c_i in the aqueous solution are low compared to the concentration of the solvent, the reference state of infinite dilution is a common choice[317]. Experimental studies with protein solutions are typically conducted at low protein concentrations such that the second criterion is normally fulfilled [1]. For this infinitely diluted reference state, the generic activity coefficient has to be replaced by the rational activity coefficient $\gamma_i^r \equiv \gamma_i/\gamma_i^\infty$, where γ_i^∞ is the limiting activity coefficient with respect to the real pure substance[319]. Choosing the reference state of infinite dilution implies that component i obeys Henry's law in case of $\gamma_i^r = 1$. All considerations below refer to the reference state of infinite dilution. Concentrations ($[i] \equiv c_i$) are expressed in terms of the particle number density, with a standard concentration c° of one particle immersed into the standard state volume $V^\circ = 1.661\ \text{nm}^3$ (i.e. $c^\circ = 1/V^\circ$), or equivalently $c^\circ = 1\ \text{mol L}^{-1}$.

An expression for the association or binding equilibrium constant K_a is obtained through combination of Eq. (5.1) and (5.2):

$$K_\text{a} = \frac{[C]}{[A][B]} = c^{\circ -1} \exp\{-\beta(\mu_C^\circ - \mu_A^\circ - \mu_B^\circ)\} = \frac{Q_C}{Q_A Q_B} \tag{5.3}$$

with the inverse thermal energy $\beta = (RT)^{-1}$. In the definition above, K_a has the dimension of a volume. The expression for K_a in terms concentrations $[i]$ represents the well known law of mass action. The sum inside the exponential involving the standard chemical potentials defines the standard binding free energy: $\Delta G^\circ_\text{bind} = \mu_C^\circ - \mu_A^\circ - \mu_B^\circ$. For simplicity, a dilute solution with respect to all species involved into the binding process was assumed. In this case, all (rational) activity coefficients can be approximated as unity. The considerations above are not limited to binding processes based upon a 1:1 complex formation. Binding schemes with arbitrary stoichiometries can be treated analogously by inclusion of stoichiometric coefficients ν_i as exponents of $[i]$. The last equality of Eq. (5.3) relates the macroscopic equilibrium constant to microscopic quantities in the form of (effective) partition functions Q_i. This relationship founds on the formulation of the standard chemical potential μ_i° in terms of Q_i[83,320]. "Effective" partition functions means that the solvent is treated in an implicit way through definition of a normalized partition function: $Q_i \equiv Q_{N_s,1}/Q_{N_s,0}$[321]. Here, $Q_{N_s,1}$ and $Q_{N_s,0}$ represent the

[1] In the environment of biological cells, the situations can differ significantly[31,318].

partition functions of one solute molecule i immersed in a bath of N_s solvent molecules and of pure solvent, respectively. These two quantities are defined in the conventional statistical-mechanical way and refer to an appropriate thermodynamic ensemble such as the canonical or isobaric-isothermal ensemble.

Relating $\Delta G^\circ_{\mathrm{bind}}$ to the Potential of Mean Force (PMF)

The linkage of K_a with partition functions Q_i serves as starting point to derive formulations which can be evaluated by molecular simulations techniques such as the double decoupling method[83,84] or physical pathway methods[272,281]. Here, I focus on the latter. The (canonical) partition functions for receptor A and ligand B can be written as follows:

$$Q_i = C_i \, 8\pi^2 \, V \int \mathrm{d}\mathbf{x}_i \, \mathrm{J}_i(\mathbf{x}_i) \, \mathrm{e}^{-\beta U_i(\mathbf{x}_i)} \;\; \text{for} \;\; i = \{A, B\} \tag{5.4}$$

with normalization constant C_i and potential energy function U_i. The latter should be interpreted as an effective potential energy or potential of mean force (PMF), resulting from an averaging procedure over the solvent degrees of freedom (DOF). Here, the whole set of $3N_i$ spatial coordinates of the solute is split into six external DOF and $3N_i - 6$ internal DOF $\mathbf{x}_i$, which is always possible. Depending upon the choice for $\mathbf{x}_i$, a Jacobian determinant $\mathrm{J}_i(\mathbf{x}_i)$ may arise. Integration over the external DOF for overall translation and rotation (defined by the three Euler angles) contributes a factor of V and $8\pi^2$, respectively. The normalization constant C_i includes the kinetic contribution from integration over the particle momenta, which can be calculated analytically. Since the normalization constants cancel from the resulting expression for K_a (see Eq. (5.7)), one can focus on the configurational part of Q_i. The internal coordinates for the complex C ($\mathbf{x}_C$) consist of the internal coordinates of the individual binding partners as well as two vectors for describing the separation ($\mathbf{r}$) and relative orientation ($\boldsymbol{\omega}$) between A and B: $\mathbf{x}_C = (\mathbf{x}_A, \mathbf{x}_B, \mathbf{r}, \boldsymbol{\omega})^\mathrm{T}$. While $\mathbf{r}$ and $\boldsymbol{\omega}$ originate from the external DOF of ligand B for overall translation and orientation, respectively, they become internal DOF of C after appropriate coordinate transformation. The external DOF of receptor A again contribute a factor of $8\pi^2 \, V$. The partition function of the complex Q_C can then be written as:

$$Q_C = C_C \, 8\pi^2 \, V \int_\mathrm{b} \underbrace{\mathrm{d}\mathbf{x}_A \mathrm{d}\mathbf{x}_B \mathrm{d}\mathbf{r} \mathrm{d}\boldsymbol{\omega}}_{\mathrm{d}\mathbf{x}_C} \, \mathrm{J}_A(\mathbf{x}_A) \, \mathrm{J}_B(\mathbf{x}_B) \, \mathrm{J}(\boldsymbol{\omega}) \mathrm{e}^{-\beta U_C(\mathbf{x}_C)} \tag{5.5}$$

where the total energy of the complex U_C is given by:

$$U_C(\mathbf{x}_C) = U_A(\mathbf{x}_A) + U_B(\mathbf{x}_B) + w(\mathbf{x}_C) \tag{5.6}$$

The third energetic term $w(\mathbf{x}_C)$, arising from the interactions between A and B, vanishes for sufficiently large separations $r = |\mathbf{r}|$, independent of the relative orientation $\boldsymbol{\omega}$. The lower case "b" in Eq. (5.5) indicates that the integration is restricted to the phase space region for which configurations of A and B are considered to be bound. Through introduction of an indicator function $I(\mathbf{r}, \boldsymbol{\omega})$ which equals 1 within the bound region and 0 otherwise, integration can be extended over the whole configuration space: $\int_\mathrm{b} \mathrm{d}\mathbf{x}_C \cdots = \int \mathrm{d}\mathbf{x}_C\, I(\mathbf{r}, \boldsymbol{\omega}) \ldots$ [83]. Inserting the expressions for Q_A, Q_B (Eq. (5.4)) and Q_C (Eq. (5.5)) into Eq. (5.3) gives:

$$K_\mathrm{a} = \frac{1}{8\pi^2} \frac{\int_\mathrm{b} \mathrm{d}\mathbf{x}_C\, \mathrm{J}_A(\mathbf{x}_A)\, \mathrm{J}_B(\mathbf{x}_B)\, \mathrm{J}(\boldsymbol{\omega}) \mathrm{e}^{-\beta U_C(\mathbf{x}_C)}}{\int \mathrm{d}\mathbf{x}_A\, \mathrm{J}_A(\mathbf{x}_A)\, \mathrm{e}^{-\beta U_A(\mathbf{x}_A)} \int \mathrm{d}\mathbf{x}_B\, \mathrm{J}_B(\mathbf{x}_B)\, \mathrm{e}^{-\beta U_B(\mathbf{x}_B)}} \tag{5.7}$$

which is identical to Eq. (III.2) in the Supporting Information of Ref. 279. Using the identity $\int \mathrm{d}\boldsymbol{\omega}\, \mathrm{J}(\boldsymbol{\omega}) = 8\pi^2$ and the definition of the unbound volume $V_\mathrm{u} = \int_\mathrm{u} \mathrm{d}\mathbf{r}$, the equilibrium constant can be written as the ratio of two configurational integrals over the bound and unbound region:

$$\begin{aligned}
K_\mathrm{a} &= \frac{V_\mathrm{u}}{V_\mathrm{u}} \cdot \frac{\int_\mathrm{b} \mathrm{d}\mathbf{x}_C\, \mathrm{J}_A(\mathbf{x}_A)\, \mathrm{J}_B(\mathbf{x}_B)\, \mathrm{J}(\boldsymbol{\omega}) \mathrm{e}^{-\beta U_C(\mathbf{x}_C)}}{\int \mathrm{d}\boldsymbol{\omega} \mathrm{J}(\boldsymbol{\omega}) \int \mathrm{d}\mathbf{x}_A\, \mathrm{J}_A(\mathbf{x}_A)\, \mathrm{e}^{-\beta U_A(\mathbf{x}_A)} \int \mathrm{d}\mathbf{x}_B\, \mathrm{J}_B(\mathbf{x}_B)\, \mathrm{e}^{-\beta U_B(\mathbf{x}_B)}} \\
&= V_\mathrm{u} \frac{\int_\mathrm{b} \mathrm{d}\mathbf{x}_C\, \mathrm{J}_A(\mathbf{x}_A)\, \mathrm{J}_B(\mathbf{x}_B)\, \mathrm{J}(\boldsymbol{\omega}) \mathrm{e}^{-\beta U_C(\mathbf{x}_C)}}{\int_\mathrm{u} \mathrm{d}\mathbf{r} \int \mathrm{d}\boldsymbol{\omega} \mathrm{J}(\boldsymbol{\omega}) \int \mathrm{d}\mathbf{x}_A\, \mathrm{J}_A(\mathbf{x}_A)\, \mathrm{e}^{-\beta U_A(\mathbf{x}_A)} \int \mathrm{d}\mathbf{x}_B\, \mathrm{J}_B(\mathbf{x}_B)\, \mathrm{e}^{-\beta U_B(\mathbf{x}_B)}} \\
&= V_\mathrm{u} \frac{\int_\mathrm{b} \mathrm{d}\mathbf{x}_C\, \mathrm{J}_A(\mathbf{x}_A)\, \mathrm{J}_B(\mathbf{x}_B)\, \mathrm{J}(\boldsymbol{\omega}) \mathrm{e}^{-\beta U_C(\mathbf{x}_C)}}{\int_\mathrm{u} \underbrace{\mathrm{d}\mathbf{x}_A \mathrm{d}\mathbf{x}_B \mathrm{d}\mathbf{r} \mathrm{d}\boldsymbol{\omega}}_{\mathrm{d}\mathbf{x}_C}\, \mathrm{J}_A(\mathbf{x}_A)\, \mathrm{J}_B(\mathbf{x}_B)\, \mathrm{J}(\boldsymbol{\omega}) \mathrm{e}^{-\beta (U_A(\mathbf{x}_A) + U_B(\mathbf{x}_B) + 0)}} \\
&= V_\mathrm{u} \frac{\int_\mathrm{b} \mathrm{d}\mathbf{x}_C\, \mathrm{J}_A(\mathbf{x}_A)\, \mathrm{J}_B(\mathbf{x}_B)\, \mathrm{J}(\boldsymbol{\omega}) \mathrm{e}^{-\beta U_C(\mathbf{x}_C)}}{\int_\mathrm{u} \mathrm{d}\mathbf{x}_C\, \mathrm{J}_A(\mathbf{x}_A)\, \mathrm{J}_B(\mathbf{x}_B)\, \mathrm{J}(\boldsymbol{\omega}) \mathrm{e}^{-\beta U_C(\mathbf{x}_C)}}
\end{aligned} \tag{5.8}$$

The key to realize the equivalence of Eq. (5.7) and Eq. (5.8), is that the combined configurational integral of the complex (Eq. (5.5)) separates into four individual integrals over $\mathbf{x}_A$, $\mathbf{x}_B$, $\mathbf{r}$ and $\boldsymbol{\omega}$ when A and B are unbound, due to the aforementioned property $w(\mathbf{x}_C) = 0$ at large distances (c.f. Eq. (5.6)). Eq. (5.8) can be further simplified through definition of a six-dimensional PMF $W(\mathbf{r}, \boldsymbol{\omega})$ between receptor and ligand as function of their separation and relative orientation:

$$\begin{aligned}
W(\mathbf{r}', \boldsymbol{\omega}') \equiv\ & -RT \ln \Bigg(8\pi^2\, V \int \underbrace{\mathrm{d}\mathbf{x}_A \mathrm{d}\mathbf{x}_B \mathrm{d}\mathbf{r} \mathrm{d}\boldsymbol{\omega}}_{\mathrm{d}\mathbf{x}_C}\, \mathrm{J}_A(\mathbf{x}_A)\, \mathrm{J}_B(\mathbf{x}_B) \\
& \cdot \delta(\mathbf{r} - \mathbf{r}')\, \delta(\boldsymbol{\omega} - \boldsymbol{\omega}') \mathrm{e}^{-\beta U_C(\mathbf{x}_C)} \Bigg)
\end{aligned} \tag{5.9}$$

where W is evaluated at specific values $\mathbf{r}'$ and $\boldsymbol{\omega}'$. $\delta()$ denotes the Dirac delta distribution. By denoting the expression within the logarithm as $Z_C(\mathbf{r}', \boldsymbol{\omega}')$, it can be seen that $\int_\mathrm{b} \mathrm{d}\mathbf{r}\mathrm{d}\boldsymbol{\omega}\, \mathrm{J}(\boldsymbol{\omega})\, Z_C(\mathbf{r}, \boldsymbol{\omega})$ yields Q_C, except for the normalization constant C_C (c.f. Eq. (5.5)). It should be noted that the PMF definition in Eq. (5.9) slightly differs compared to Ref. 279. Therein, the authors basically normalize the expression above with the configurational integral of the complex in the unbound state, corresponding to the value of $W(\mathbf{r}, \boldsymbol{\omega})$ at large distances: $W_\infty \equiv \lim\limits_{r\to\infty} W(\mathbf{r}, \boldsymbol{\omega})$. This normalization ensures that the PMF goes to zero in the bulk. Using the PMF definition in Eq. (5.9), Eq. (5.8) can be further reduced:

$$\begin{aligned} K_\mathrm{a} &= V_\mathrm{u} \frac{\int_\mathrm{b} \mathrm{d}\mathbf{r}\mathrm{d}\boldsymbol{\omega}\, \mathrm{J}(\boldsymbol{\omega}) \cdot \mathrm{d}\mathbf{x}_A \mathrm{d}\mathbf{x}_B\, \mathrm{J}_A(\mathbf{x}_A)\, \mathrm{J}_B(\mathbf{x}_B)\, \mathrm{e}^{-\beta U_C(\mathbf{x}_C)}}{\int_\mathrm{u} \mathrm{d}\mathbf{r}\mathrm{d}\boldsymbol{\omega}\, \mathrm{J}(\boldsymbol{\omega}) \cdot \mathrm{d}\mathbf{x}_A \mathrm{d}\mathbf{x}_B\, \mathrm{J}_A(\mathbf{x}_A)\, \mathrm{J}_B(\mathbf{x}_B)\, \mathrm{e}^{-\beta U_C(\mathbf{x}_C)}} \\ &= V_\mathrm{u} \frac{\int_\mathrm{b} \mathrm{d}\mathbf{r}\mathrm{d}\boldsymbol{\omega}\, \mathrm{J}(\boldsymbol{\omega})\, \mathrm{e}^{-\beta W(\mathbf{r}, \boldsymbol{\omega})}}{\int_\mathrm{u} \mathrm{d}\mathbf{r}\mathrm{d}\boldsymbol{\omega}\, \mathrm{J}(\boldsymbol{\omega})\, \mathrm{e}^{-\beta W(\mathbf{r}, \boldsymbol{\omega})}} \end{aligned} \tag{5.10}$$

The expression for K_a in Eq. (5.10) seems to differ compared to the one reported in Ref. 279 at a first glance: $K_\mathrm{a} = (8\pi^2)^{-1} \int_\mathrm{b} \mathrm{d}\mathbf{r}\mathrm{d}\boldsymbol{\omega}\, \mathrm{J}(\boldsymbol{\omega})\, \mathrm{e}^{-\beta W(\mathbf{r}, \boldsymbol{\omega})}$. The latter only comprises the configurational integral over the bound region. This discrepancy originates again from the slightly different PMF definitions, mentioned before. Equivalence of both expressions follows from the constant value of $W(\mathbf{r}, \boldsymbol{\omega})$ (denoted as W_∞) for unbound A and B, which can therefore be taken out of the integral in the denominator of Eq. (5.10).

The standard binding free energy follows from Eq. (5.10):

$$\Delta G^\circ_\mathrm{bind} = -RT \ln(c^\circ K_\mathrm{a}) = -RT \ln\left(\frac{\int_\mathrm{b} \mathrm{d}\mathbf{r}\mathrm{d}\boldsymbol{\omega}\, \mathrm{J}(\boldsymbol{\omega})\, \mathrm{e}^{-\beta W(\mathbf{r}, \boldsymbol{\omega})}}{\int_\mathrm{u} \mathrm{d}\mathbf{r}\mathrm{d}\boldsymbol{\omega}\, \mathrm{J}(\boldsymbol{\omega})\, \mathrm{e}^{-\beta W(\mathbf{r}, \boldsymbol{\omega})}} \right) - RT \ln\left(\frac{V_\mathrm{u}}{V^\circ} \right) \tag{5.11}$$

where the relation $V^\circ = c^{\circ -1}$ has been used. The expression for $\Delta G^\circ_\mathrm{bind}$ consists of two terms: the first involves integration of the PMF (denoted as ΔG_PMF), while the second one represents an entropic term (denoted as ΔG_V), accounting for the volume change from the standard state volume to the unbound volume. This corresponds to Eq. (4.1) in the previous chapter.

If only the dependence on the separation between ligand and receptor is of interest, a three-dimensional PMF $W(\mathbf{r})$ can be obtained from Eq. (5.9) by integration over $\boldsymbol{\omega}$. In this case, Eq. (5.11) simplifies to:

$$\begin{aligned} \Delta G^\circ_\mathrm{bind} &= -RT \ln\left(\frac{\int_\mathrm{b} \mathrm{d}\mathbf{r}\, \mathrm{e}^{-\beta W(\mathbf{r})}}{\int_\mathrm{u} \mathrm{d}\mathbf{r}\, \mathrm{e}^{-\beta W(\mathbf{r})}} \right) - RT \ln\left(\frac{V_\mathrm{u}}{V^\circ} \right) && (5.12) \\ &= -RT \ln\left(\frac{\int_\mathrm{b} \mathrm{d}\mathbf{r}\, \mathrm{e}^{-\beta W(\mathbf{r})}}{\mathrm{e}^{-\beta W_\infty} \int_\mathrm{u} \mathrm{d}\mathbf{r}} \right) - RT \ln\left(\frac{V_\mathrm{u}}{V^\circ} \right) \\ &= -W_\infty - RT \ln\left(\frac{V_\mathrm{b}}{V^\circ} \right) && (5.13) \end{aligned}$$

with the bound volume $V_\mathrm{b} = \int_\mathrm{b} \mathrm{d}\mathbf{r}\, \mathrm{e}^{-\beta W(\mathbf{r})}$. In the second line, the property was used that the PMF takes a constant value W_∞ (i.e. it becomes flat) in the unbound region. It can be seen, that the final expression for the binding free energy is independent of the arbitrarily chosen unbound volume V_u.

Special Considerations on One-Dimensional PMFs

Frequently, one is interested in the analysis of the PMF along a single order parameter. One example is the scalar radial distance r between the centers-of-mass of ligand and receptor. In case of spherical coordinates, the volume element of integration is given by: $\mathrm{d}\mathbf{r} = 4\pi\, r^2\, \mathrm{d}r$. The pre-factor 4π arises from integration over the polar and azimuthal angle. From Eq. (5.12) it follows:

$$\Delta G^\circ_\mathrm{bind} = -RT \ln \left(\frac{\int_0^{r_\mathrm{b}} \mathrm{d}r\, 4\pi r^2\, \mathrm{e}^{-\beta W(r)}}{\int_{r_\mathrm{b}}^{r_\mathrm{u}} \mathrm{d}r\, 4\pi r^2\, \mathrm{e}^{-\beta W(r)}} \right) - RT \ln \left(\frac{V_\mathrm{u}}{V^\circ} \right) \tag{5.14}$$

with the unbound volume:

$$V_\mathrm{u} = \int_\mathrm{u} \mathrm{d}\mathbf{r} = \int_{r_\mathrm{b}}^{r_\mathrm{u}} \mathrm{d}r\, 4\pi r^2 = \frac{4\pi}{3} (r_\mathrm{u}^3 - r_\mathrm{b}^3) \tag{5.15}$$

and the bound volume:

$$V_\mathrm{b} = \int_\mathrm{b} \mathrm{d}\mathbf{r}\, \mathrm{e}^{-\beta W(\mathbf{r})} = \int_0^{r_\mathrm{b}} \mathrm{d}r\, 4\pi r^2\, \mathrm{e}^{-\beta W(r)} \tag{5.16}$$

r_b denotes a threshold or cutoff distance for the bound region and r_u some (arbitrary) upper integration limit. The binding free energy can be practically calculated from Eq. (5.13) with V_b obtained from numerical integration according to Eq. (5.16).

A slightly different but equivalent expression is obtained when the Jacobian determinant $4\pi r^2 = \exp\{-\beta[-(RT)\ln(4\pi r^2)]\}$ is absorbed into the exponent in Eq. (5.14):

$$\Delta G^\circ_\mathrm{bind} = -RT \ln \left(\frac{4\pi (r_\mathrm{u}^3 - r_\mathrm{b}^3)}{3V^\circ} \frac{\int_0^{r_\mathrm{b}} \mathrm{d}r\, \mathrm{e}^{-\beta F(r)}}{\int_{r_\mathrm{b}}^{r_\mathrm{u}} \mathrm{d}r\, \mathrm{e}^{-\beta F(r)}} \right) \tag{5.17}$$

where V_u according to Eq. (5.15) has been used. In contrast to the PMF $W(r)$, the resulting free energy curve (FEC) $F(r)$ contains the Jacobian contribution and therefore decreases as $-2RT \ln r$ for large distances instead of becoming flat. Analysis of umbrella sampling simulations (e.g. via the weighted histogram analysis method) delivers a FEC rather than a PMF.

Since the one-dimensional free energy profiles studied in chapter 4 were evaluated not as function of r but along a cartesian component of the receptor-ligand separation vector

$\mathbf{r} = (x,y,z)^{\mathrm{T}}$, this case will be investigated in the following. For this choice, PMF and FEC are identical. The original setup which was proposed by Doudou et al.[281] can be summarized as follows: (i) The receptor is positionally restrained such that the z-axis of the coordinate system of the simulation box points through its binding pose. The z-axis also represents the restraining pathway; (ii) The PMF is evaluated along the z-component of the separation vector $\mathbf{r}$; (iii) The sampled volume in orthogonal direction is limited by keeping the ligand close to the z-axis. This is realized via two harmonic axial restraints in x- and y-direction: $U_{xy}(x,y) = \frac{k_{xy}}{2}\left[(x-x_0)^2 + (y-y_0)^2\right]$, with force constant k_{xy} and reference values $x_0 = y_0 = 0$. In the original approach which is studied here, the (relative) orientation of the ligand towards the receptor is not restrained.

Derivation of an expression for $\Delta G^{\circ}_{\mathrm{bind}}$ starts from extending the key equation Eq. (5.8) through introduction of intermediate states (using the short-hand notation $\mathrm{J}_i \equiv \mathrm{J}_i(\mathbf{x}_i)$, $\mathrm{J}_\omega \equiv \mathrm{J}(\boldsymbol{\omega})$, $U_C \equiv U_C(\mathbf{x}_C)$, $U_{xy} \equiv U_{xy}(x,y)$ for simplification):

$$\begin{aligned} K_{\mathrm{a}} &= V_{\mathrm{u}} \frac{\int_{\mathrm{b}} \mathrm{d}\mathbf{x}_C\, \mathrm{J}_A\, \mathrm{J}_B\, \mathrm{J}_\omega\, \mathrm{e}^{-\beta U_C}}{\int_{\mathrm{u}} \mathrm{d}\mathbf{x}_C\, \mathrm{J}_A\, \mathrm{J}_B\, \mathrm{J}_\omega\, \mathrm{e}^{-\beta U_C}} \\ &= V_{\mathrm{u}} \underbrace{\frac{\int_{\mathrm{b}} \mathrm{d}\mathbf{x}_C\, \mathrm{J}_A\, \mathrm{J}_B\, \mathrm{J}_\omega\, \mathrm{e}^{-\beta U_C}}{\int_{\mathrm{b}} \mathrm{d}\mathbf{x}_C\, \mathrm{J}_A\, \mathrm{J}_B\, \mathrm{J}_\omega\, \mathrm{e}^{-\beta (U_C + U_{xy})}}}_{\langle \mathrm{e}^{-\beta U_{xy}} \rangle^{-1}_{\mathrm{b},U_C}} \cdot \underbrace{\frac{\int_{\mathrm{b}} \mathrm{d}\mathbf{x}_C\, \mathrm{J}_A\, \mathrm{J}_B\, \mathrm{J}_\omega\, \mathrm{e}^{-\beta (U_C + U_{xy})}}{\int_{\mathrm{u}} \mathrm{d}\mathbf{x}_C\, \mathrm{J}_A\, \mathrm{J}_B\, \mathrm{J}_\omega\, \mathrm{e}^{-\beta (U_C + U_{xy})}}}_{\mathrm{PMF-Term}} \\ &\quad \cdot \underbrace{\frac{\int_{\mathrm{u}} \mathrm{d}\mathbf{x}_C\, \mathrm{J}_A\, \mathrm{J}_B\, \mathrm{J}_\omega\, \mathrm{e}^{-\beta (U_C + U_{xy})}}{\int_{\mathrm{u}} \mathrm{d}\mathbf{x}_C\, \mathrm{J}_A\, \mathrm{J}_B\, \mathrm{J}_\omega\, \mathrm{e}^{-\beta U_C}}}_{\langle \mathrm{e}^{-\beta U_{xy}} \rangle_{\mathrm{u},U_C}} \end{aligned} \tag{5.18}$$

with $\langle \ldots \rangle_{\mathrm{b},U_C}$ and $\langle \ldots \rangle_{\mathrm{u},U_C}$ denoting the ensemble averages, sampled with the unbiased potential U_C within the bound (”b”) and unbound (”u”) region, respectively. The middle term corresponds to the work to transfer the orthogonally restrained ligand from the bulk to the binding pose. The first term represents the influence of the orthogonal restraints in the bound region and has to be evaluated numerically. Evaluation can be performed in different ways, using e.g. exponential averaging of the sampled time series for x and y (also known as the “Zwanzig”-formula[282]) or thermodynamic integration. The third term which measures the influence of the orthogonal restraints in the unbound region can be calculated analytically:

$$\begin{aligned}
\langle \mathrm{e}^{-\beta U_{xy}} \rangle_{\mathrm{u},U_C} &= \frac{\int_{\mathrm{u}} \mathrm{d}\mathbf{x}_A \mathrm{d}\mathbf{x}_B \mathrm{d}\boldsymbol{\omega}\, \overbrace{\mathrm{d}x\mathrm{d}y\mathrm{d}z}^{\mathrm{d}\mathbf{r}}\; \mathrm{J}_A\, \mathrm{J}_B\, \mathrm{J}_\omega\, \mathrm{e}^{-\beta U_C} \cdot \mathrm{e}^{-\beta U_{xy}}}{\int_{\mathrm{u}} \mathrm{d}\mathbf{x}_A \mathrm{d}\mathbf{x}_B \mathrm{d}\boldsymbol{\omega} \mathrm{d}x\mathrm{d}y\mathrm{d}z\, \mathrm{J}_A\, \mathrm{J}_B\, \mathrm{J}_\omega\, \mathrm{e}^{-\beta U_C}} \\
&= \frac{\int_{\mathrm{u}} \mathrm{d}\mathbf{x}_A \mathrm{d}\mathbf{x}_B \mathrm{d}\boldsymbol{\omega}\, \mathrm{J}_A\, \mathrm{J}_B\, \mathrm{J}_\omega\, \mathrm{e}^{-\beta U_C}}{\int_{\mathrm{u}} \mathrm{d}\mathbf{x}_A \mathrm{d}\mathbf{x}_B \mathrm{d}\boldsymbol{\omega}\, \mathrm{J}_A\, \mathrm{J}_B\, \mathrm{J}_\omega\, \mathrm{e}^{-\beta U_C}} \cdot \frac{\int_{\mathrm{u}} \mathrm{d}x\mathrm{d}y\, \mathrm{e}^{-\beta U_{xy}(x,y)}}{\int_{\mathrm{u}} \mathrm{d}x\mathrm{d}y} \cdot \frac{\int_{\mathrm{u}} \mathrm{d}z}{\int_{\mathrm{u}} \mathrm{d}z} \\
&= \frac{A_{\mathrm{u,R}}}{A_{\mathrm{u}}} \cdot \frac{l_{\mathrm{u}}}{l_{\mathrm{u}}} = \frac{V_{\mathrm{u,R}}}{V_{\mathrm{u}}}
\end{aligned} \tag{5.19}$$

with

$$\begin{aligned}
A_{\mathrm{u,R}} &= \int_{\mathrm{u}} \mathrm{d}x\mathrm{d}y\, \mathrm{e}^{-\beta U_{xy}(x,y)} \\
&= \int_{-\infty}^{\infty} \exp\left(\frac{-k_{xy}\, x^2}{2RT}\right) \mathrm{d}x \cdot \int_{-\infty}^{\infty} \exp\left(\frac{-k_{xy}\, y^2}{2RT}\right) \mathrm{d}y \\
&= \frac{2\pi RT}{k_{xy}}
\end{aligned} \tag{5.20}$$

l_{u}, A_{u} and $V_{\mathrm{u}} = A_{\mathrm{u}} \cdot l_{\mathrm{u}}$ denote the length, area and corresponding volume available to the unbound ligand in the absence of the orthogonal restraints. Whereas $A_{\mathrm{u,R}}$ and $V_{\mathrm{u,R}} = A_{\mathrm{u,R}} \cdot l_{\mathrm{u}}$ denote the corresponding unbound area and volume available to the ligand in the presence of the restraining potential $U_{xy}(x, y)$.

The middle term in Eq. (5.18) can be simplified by introduction of an appropriate one-dimensional PMF as done before (c.f. Eq. (5.9)):

$$W_{\mathrm{R}}(z') \equiv -RT \ln\left(8\pi^2\, V \int \underbrace{\mathrm{d}\mathbf{X}\; \overbrace{\mathrm{d}x\mathrm{d}y\mathrm{d}z}^{\mathrm{d}\mathbf{r}}}_{\mathrm{d}\mathbf{x}_C}\; \delta(z - z')\, \mathrm{e}^{-\beta(U_C(\mathbf{x}_C) + U_{xy}(x,y))}\right) \tag{5.21}$$

where the hyper volume element $\mathrm{d}\mathbf{X} \equiv \mathrm{d}\mathbf{x}_A \mathrm{J}_A(\mathbf{x}_A)\, \mathrm{d}\mathbf{x}_B \mathrm{J}_B(\mathbf{x}_B) \mathrm{d}\boldsymbol{\omega} \mathrm{J}(\boldsymbol{\omega})$ was introduced for notational simplification. The index "R" indicates that the PMF is generated from a potential energy function which includes the axial restraining potential U_{xy} in addition to physical potential U_C. With this PMF, the middle term in Eq. (5.18) can be written as:

$$\frac{\int_{\mathrm{b}} \mathrm{d}\mathbf{x}_C\, \mathrm{J}_A\, \mathrm{J}_B\, \mathrm{J}_\omega\, \mathrm{e}^{-\beta(U_C + U_{xy})}}{\int_{\mathrm{u}} \mathrm{d}\mathbf{x}_C\, \mathrm{J}_A\, \mathrm{J}_B\, \mathrm{J}_\omega\, \mathrm{e}^{-\beta(U_C + U_{xy})}} = \frac{\int_{\mathrm{b}} \mathrm{d}z\, \mathrm{e}^{-\beta W_{\mathrm{R}}(z)}}{\int_{\mathrm{u}} \mathrm{d}z\, \mathrm{e}^{-\beta W_{\mathrm{R}}(z)}} \tag{5.22}$$

In contrast to the previous case for which the radial distance r has been used as order parameter, no additional Jacobian factor arises in the integrals on the right hand side (c.f. Eq. (5.14)). For this reason, there is no need to distinguish between a PMF and a FEC as in the previous case. Combination of Eq. (5.18), Eq. (5.19) and Eq. (5.22) delivers the following expression for K_a:

$$K_\mathrm{a} = \langle \mathrm{e}^{-\beta U_{xy}} \rangle_{\mathrm{b},U_C}^{-1} \frac{\int_\mathrm{b} \mathrm{d}z\, \mathrm{e}^{-\beta W_\mathrm{R}(z)}}{\int_\mathrm{u} \mathrm{d}z\, \mathrm{e}^{-\beta W_\mathrm{R}(z)}} V_\mathrm{u,R} \tag{5.23}$$

or equivalently for $\Delta G^\circ_\mathrm{bind}$:

$$\Delta G^\circ_\mathrm{bind} = \underbrace{RT \ln \left(\langle \mathrm{e}^{-\beta U_{xy}} \rangle_{\mathrm{b},U_C} \right)}_{\Delta G_\mathrm{R}} \underbrace{-RT \ln \left(\frac{\int_\mathrm{b} \mathrm{d}z\, \mathrm{e}^{-\beta W_\mathrm{R}(z)}}{\int_\mathrm{u} \mathrm{d}z\, \mathrm{e}^{-\beta W_\mathrm{R}(z)}} \right)}_{\Delta G_\mathrm{PMF}} \underbrace{-RT \ln \left(\frac{V_\mathrm{u,R}}{V^\circ} \right)}_{\Delta G_\mathrm{V}} \tag{5.24}$$

which is exactly Eq. (6) in Ref. 281. The term ΔG_R accounts for the free energy contribution of releasing the orthogonal translational restraint in the bound state. The term ΔG_V represents the free energy contribution due to the volume change from the standard state volume V° to the unbound volume $V_\mathrm{u,R}$, available to the orthogonally restrained ligand. Eq. (5.25) can be rearranged to a slightly simpler and more practical form (c.f. Eq. (11) in Ref. 281):

$$\Delta G^\circ_\mathrm{bind} = \Delta W_\mathrm{R} - RT \ln \left(\frac{l_\mathrm{b} A_\mathrm{u,R}}{V^\circ} \right) + \Delta G_\mathrm{R} \tag{5.25}$$

with the bound length $l_\mathrm{b} \equiv \int_\mathrm{b} \mathrm{d}z\, \mathrm{e}^{-\beta W_\mathrm{R}(z)}$, defined as the integral of the Boltzmann factor of the 1D-PMF over the bound region, the PMF depth $\Delta W_\mathrm{R} \equiv -W_\infty$ and $A_\mathrm{u,R}$ according to Eq. (5.20).

If the ligand orientation would have been further restrained as it was done in chapter 4, a corresponding expression for K_a can be derived in a completely analogue manner through introduction of additional intermediate states into Eq. (5.18) which then leads to Eq. (4.2).

Chapter 6

Conclusion and Outlook

Protein research comprises so many facets, that even after decades of work by many groups, various fundamental aspects are not fully understood. Examples are the relative importance of intramolecular hydrogen bonding for protein stability or the molecular stabilization mechanism of protecting osmolytes. Molecular simulations based upon biomolecular force fields can provide structural, dynamic and energetic information in atomistic resolution, information about conformational distributions and averages of macroscopic thermodynamic observables. For these reasons, molecular simulations have established as a valuable tool in biophysical research to bridge the gap between experiment and theory. All topics treated within this thesis are related to specific aspects of protein stability:
(i) identification of a suitable molecular model for the protecting osmolyte TMAO (chapter 2),
(ii) study of intramolecular hydrogen bonding via free-energy MD simulations (chapter 3),
(iii) host-ligand binding (chapter 4).
Central problems which limit the application of biomolecular simulations are the so called **force-field problem** and the **sampling problem**[93]. The disentanglement of these two issues is of central importance to establish a robust and meaningful comparison with experiments and therefore an area of active research. While chapter 2 represents a force field comparison, the chapters 3 and 4 focus on the optimization of the simulation protocol (for a given force field) such that the computational results reflect primarily the quality of the force field and are not dominated by sampling issues.

6.1 TMAO Force Field Comparison

A detailed picture of how protein stabilization is mediated by protecting osmolytes such as TMAO is still not fully understood. Over recent years, several computational TMAO models, based on classical non-polarizable force fields have been developed. Those models which are most commonly used in molecular simulation studies, were derived from the same original parameter set[98] but have been trained to reproduce different physical properties (see Sec. 2.3.1). A well-parametrized and transferable osmolyte model from which one can hope to obtain fundamental insights about the interaction mechanism, requires calibration on the basis of binary aqueous mixtures, prior to the study of their impact on proteins. Since such a comparison was missing, these models were tested with respect to the reproduction of various thermophysical properties of aqueous TMAO solutions for a wide range of thermodynamic conditions (see chapter 2). It was found that one of the considered TMAO models (denoted as Kast 2016[99]) shows very good overall performance for all considered properties and under all conditions. It can be assumed that the Kast 2016 model is a promising candidate to deliver molecular insights of the interaction mechanism between TMAO and proteins, mentioned above. Thus, the door is now open for further investigations in the presence of proteins.

Studies such as the one presented in chapter 2 show the predictive power of classical force field models, despite the involved approximations of pairwise additive potentials and fixed partial charges. However, the force field parameters have to be optimized such that a good balance between solute-solute and solute-solvent interactions is achieved. The Kast 2016 model is a good example which demonstrates that such an optimization is possible in the high-dimensional parameter space of classical force fields. From the perspective of force field development, it is interesting to see whether the combination of liquid state integral equations theory coupled to a quantum-chemical treatment of the solute as it was applied in Ref. 99 (therein, in the form of the embedded cluster reference interaction site model (EC-RISM)[322]), performs equally well for other co-solvents. A recently developed force field for urea which was parametrized in the same way seems to perform equally well with respect to the description of aqueous solution properties[323].

6.2 Calculation of Relative Folding Free Energies

The relative importance of intramolecular hydrogen bonding (HBonding) for overall protein stability is still under debate. Small independently folding protein domains such as the Pin1-WW domain appear to be mainly stabilized by backbone HBonding, which is why they have become well-established model systems in this respect. So called amide-to-ester (A-to-E) mutations enable the perturbation of particular HBonds and to estimate their strengths on the basis of relative free energy differences from denaturation experiments[64,221]. As demonstrated in Ref. 100, these relative free energy differences can also be estimated from MD simulations by performing two alchemical free energy calculations, one conducted for the protein's folded state and one for the unfolded state. However, indications of non-converged free energy estimates resulting from sampling issues prevented an unambiguous evaluation of the force field quality at that time.

As part of the study presented in chapter 3, I observed that the free energy estimates are highly sensitive to subtle local variations in the protein starting structure. This finding, which was rather unexpected considering the small size of the Pin1-WW domain, could be attributed mainly to insufficient backbone dihedral angle sampling in the vicinity of the perturbed residue. The usage of multiple starting structures in combination with Hamiltonian replica exchange was found to reduce this starting structure dependence considerably, leading to an improved agreement and more meaningful comparison with experimental estimates.

With our optimized simulation protocol, future comparative simulations based upon other biomolecular force fields may provide valuable information about force field inaccuracies or even experimental uncertainties.

In a recent study, the same approach was applied to estimate relative free energy differences associated with side-chain mutations in case of the same protein[324]. The script-based workflow developed therein, allows the treatment of arbitrary mutations compatible with the GROMOS 54A7 biomolecular force field[13]. Comparison with experimental data from thermal denaturation of 76 single-point mutations[66] shows good overall agreement with only a few outliers. There are indications that once again, the usage of multiple starting structures in combination with Hamiltonian replica exchange is mandatory to obtain converged free-energy estimates. The results will be published in a future publication.

Analysis of the WW domain model system can be used for further studies of fundamental aspects of protein thermodynamics. The WW domain represents only the N-terminal part of the human protein Pin1. The C-terminus of Pin1 comprises the peptidyl-prolyl cis/trans isomerase (PPIase) domain. The crystal structure of Pin1 (PDB code 1PIN[223], Fig. 6.1) suggests that both domains interact, however, little data is available so far to support this hypothesis. In an ongoing collaboration project with R. Ghosh (IBBS, University of Stuttgart), a novel genetic construct was designed which enables expression of both domains and the whole Pin1 protein. Analysis of these constructs will yield insights into the thermodynamics of the inter-domain interactions.

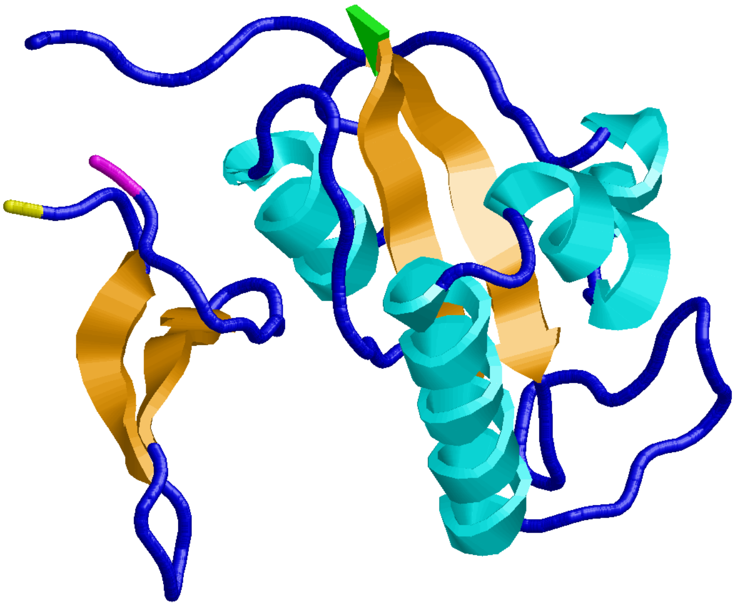


Figure 6.1: Structural domains of human Pin1 (PDB code 1PIN[223]). Left: N-terminal WW domain (34 residues); Right: C-terminal PPIase domain (119 residues).

6.3 Calculation of 1D-Potentials of Mean Force

Calculation of binding affinities is an important application of free energy MD simulations which led to the development of a variety of methods over the years. Within the class of physical pathway methods, the binding free energy is estimated from the potential of mean force (PMF) between the considered host-guest system. There are several examples of one-dimensional PMFs for symmetric systems such as membranes or nanopores which show an artificial offset between the two bulk regions. However, the origins of these offsets, which violate the state function property of the free energy, have been either ignored or only discussed marginally.

From systematic studies based upon suitable host-guest model systems, various reasons for PMF offsets could be identified (see chapter 4). The demonstration that such offsets can not exclusively occur for very complex systems, but equally in case of small and rather simple host molecules such as the considered short carbon nanotube or α-cyclodextrin has to be emphasized. It was found that offsets that result from insufficiently sampled ligand orientations in the host's cavity (bound state), can be eliminated by either (i) application of an orientational restraint or (ii) through combination of the sampling protocol with Hamiltonian replica exchange to enable the sampling of multiple bound state configurations.

As mentioned in the main text, it can be expected that PMFs obtained from alternative physical pathway protocols such as forward flux sampling will show the same offset.

However, it would be satisfying to prove our expectations by further simulation studies. A comparison of different one-dimensional order parameters might also be of interest[36]. It can be assumed that the main cause of the found artifacts is rooted in the projection of the free energy profile along a single order parameter. It would therefore be interesting to determine two-dimensional free energy profiles through biasing a second order parameter such as the relative orientational angle θ between ligand and host (see Fig. 4.3). Estimation of $\Delta G^{\circ}_{\mathrm{bind}}$ through integration from such a reweighted two-dimensional free energy landscape should be identical for both (asymmetric) ends of the host.

In summary, this work attempted to establish robust simulation protocols for different applications in biomolecular simulation, which can be used as guidelines for future work.

Bibliography

[1] C. Levinthal. Molecular model-building by computer. *Sci. Am.*, 214:42–53, 1966.

[2] R. Zwanzig, A. Szabo, and B. Bagchi. Levinthal's paradox. *Proc. Natl. Acad. Sci. USA*, 89:20–22, 1992.

[3] R. H. Pain, B. D. Hames, and D. M. Glover. *Mechanisms of protein folding.* IRL Press Oxford, 1994.

[4] P. Ade et al. Planck 2013 results. xvi. cosmological parameters. *Astron. Astrophys.*, 571:A16, 2014.

[5] G. M. Ashraf et al. Protein misfolding and aggregation in alzheimer's disease and type 2 diabetes mellitus. *CNS Neurol. Disord. – Drug Targets*, 13:1280–1293, 2014.

[6] K. A Dill and J. L. MacCallum. The protein-folding problem, 50 years on. *Science*, 338:1042–1046, 2012.

[7] A. Hillisch, N. Heinrich, and H. Wild. Computational chemistry in the pharmaceutical industry: from childhood to adolescence. *ChemMedChem*, 10:1958–1962, 2015.

[8] I. Muegge, A. Bergner, and J. M. Kriegl. Computer-aided drug design at boehringer ingelheim. *J. Comput. Aid. Mol. Des.*, 31:275–285, 2017.

[9] J. A. McCammon, B. R. Gelin, and M. Karplus. Dynamics of folded proteins. *Nature*, 267, 1977.

[10] K. Lindorff-Larsen, P. Maragakis, S. Piana, and D. E. Shaw. Picosecond to millisecond structural dynamics in human ubiquitin. *J. Phys. Chem. B*, 120:8313–8320, 2016.

[11] Y. Hirano, K. Takeda, and K. Miki. Charge-density analysis of an iron–sulfur protein at an ultra-high resolution of 0.48 Å. *Nature*, 534:281, 2016.

[12] W. F. Van Gunsteren and M. Karplus. Effect of constraints on the dynamics of macromolecules. *Macromolecules*, 15:1528–1544, 1982.

[13] N. Schmid, A.P. Eichenberger, A. Choutko, S. Riniker, M. Winger, A.E. Mark, and W.F. van Gunsteren. Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *Eur. Biophys. J.*, 40:843, 2011.

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

[15] A. D. MacKerell Jr et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102:3586–3616, 1998.

[16] P. K. Weiner and P. A. Kollman. AMBER: Assisted model building with energy refinement. a general program for modeling molecules and their interactions. *J. Comput. Chem.*, 2:287–303, 1981.

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

[18] Z. Jing, C. Liu, S. Y. Cheng, R. Qi, B. D. Walker, J. Piquemal, and P. Ren. Polarizable force fields for biomolecular simulations: Recent advances and applications. *Annu. Rev. Biophys.*, 48:371–394, 2019.

[19] A. Hemmen and J. Gross. Transferable anisotropic united-atom force field based on the mie potential for phase equilibrium calculations: n-alkanes and n-olefins. *J. Phys. Chem. B*, 119:11695–11707, 2015.

[20] C. Vega. Water: One molecule, two surfaces, one mistake. *Mol. Phys.*, 113:1145–1163, 2015.

[21] B. A. C. Horta, P. T. Merz, P. F. J. Fuchs, J. Dolenc, S. Riniker, and P. H. Hünenberger. A GROMOS-compatible force field for small organic molecules in the condensed phase: The 2016H66 parameter set. *J. Chem. Theory Comput.*, 12:3825–3850, 2016.

[22] G. J. Rocklin, D. L. Mobley, and K. A. Dill. Calculating the sensitivity and robustness of binding free energy calculations to force field parameters. *J. Chem. Theory Comput.*, 9:3072–3083, 2013.

[23] J. Yin, A. T. Fenley, N. M. Henriksen, and M. K. Gilson. Toward improved force-field accuracy through sensitivity analysis of host-guest binding thermodynamics. *J. Phys. Chem. B*, 119:10145–10155, 2015.

[24] D.E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R.O. Dror, M.P. Eastwood, J.A. Bank, J.M. Jumper, J.K. Salmon, Y. Shan, and W. Wriggers. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330:341–346, 2010.

[25] A. C. Pan, T. M. Weinreich, S. Piana, and D. E. Shaw. Demonstrating an order-of-magnitude sampling enhancement in molecular dynamics simulations of complex protein systems. *J. Chem. Theory Comput.*, 12:1360–1367, 2016.

[26] A. C. Pan, H. Xu, T. Palpant, and D. E. Shaw. Quantitative characterization of the binding and unbinding of millimolar drug fragments with molecular dynamics simulations. *J. Chem. Theory Comput.*, 13:3372–3377, 2017.

[27] D. R. Slochower, N. M. Henriksen, L.-P. Wang, J. D. Chodera, D. L. Mobley, and M. K. Gilson. Binding thermodynamics of host–guest systems with SMIRNOFF99Frosst 1.0. 5 from the open force field initiative. *J. Chem. Theory Comput.*, 15:6225–6242, 2019.

[28] J. Henriques, C. Cragnell, and M. Skepö. Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment. *J. Chem. Theory Comput.*, 11:3420–3431, 2015.

[29] G. R. Bowman. Accurately modeling nanosecond protein dynamics requires at least microseconds of simulation. *J. Comput. Chem.*, 37:558–566, 2016.

[30] M. J. Robertson, J. Tirado-Rives, and W. L. Jorgensen. Improved peptide and protein torsional energetics with the OPLS-AA force field. *J. Chem. Theory Comput.*, 11:3499–3509, 2015.

[31] M. Gruebele, K. Dave, and S. Sukenik. Globular protein folding in vitro and in vivo. *Annu. Rev. Biophys.*, 45:233–251, 2016.

[32] F. M. Richards. The interpretation of protein structures: total volume, group volume distributions and packing density. *J. Mol. Biol.*, 82:1–14, 1974.

[33] C. Chothia. Structural invariants in protein folding. *Nature*, 254:304, 1975.

[34] A. Cooper. Thermodynamics of protein folding and stability. *Protein: A comprehensive treatise*, 2:217–270, 1999.

[35] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes. Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.*, 48:545–600, 1997.

[36] R. B. Best and G. Hummer. Reaction coordinates and rates from transition paths. *Proc. Natl. Acad. Sci. USA*, 102:6732–6737, 2005.

[37] E. Shakhnovich. Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. *Chem. Rev.*, 106:1559–1588, 2006.

[38] E. Rhoades, M. Cohen, B. Schuler, and G. Haran. Two-state folding observed in individual protein molecules. *J. Am. Chem. Soc.*, 126:14686–14687, 2004.

[39] S.E. Jackson and A.R. Fersht. Folding of Chymotrypsin inhibitor 2. 1. Evidence for a two-state transition. *Biochemistry*, 30:10428–10435, 1991.

[40] D. Barrick. What have we learned from the studies of two-state folders, and what are the unanswered questions about two-state protein folding? *Phys. Biol.*, 6:015001, 2009.

[41] T. E. Creighton. Stability of folded conformations: Current opinion in structural biology 1991, 1: 5–16. *Curr. Opin. Struct. Biol.*, 1:5–16, 1991.

[42] K. A. Dill. Dominant forces in protein folding. *Biochemistry*, 29:7133–7155, 1990.

[43] J. T. Kellis Jr., K. Nyberg, and A. R. Fersht. Energetics of complementary side chain packing in a protein hydrophobic core. *Biochemistry*, 28:4914–4922, 1989.

[44] B. H. M. Mooers, D. Datta, W. A. Baase, E. S. Zollars, S. L. Mayo, and B. W. Matthews. Repacking the core of T4 lysozyme by automated design. *J. Mol. Biol.*, 332:741–756, 2003.

[45] H. Schäfer, X. Daura, A. E. Mark, and W. F. van Gunsteren. Entropy calculations on a reversibly folding peptide: changes in solute free energy cannot explain folding behavior. *Proteins*, 43:45–56, 2001.

[46] R. Lumry and S. Rajender. Enthalpy–entropy compensation phenomena in water solutions of proteins and small molecules: a ubiquitous properly of water. *Biopolymers*, 9:1125–1227, 1970.

[47] J. D. Dunitz. Win some, lose some: enthalpy-entropy compensation in weak intermolecular interactions. *Chem. Biol.*, 2:709–712, 1995.

[48] G. D. Rose, P. J. Fleming, J. R. Banavar, and A. Maritan. A backbone-based theory of protein folding. *Proc. Natl. Acad. Sci. USA*, 103:16623–16633, 2006.

[49] C. N. Pace et al. Contribution of hydrophobic interactions to protein stability. *J. Mol. Biol.*, 408:514–528, 2011.

[50] C. N. Pace et al. Contribution of hydrogen bonds to protein stability. *Protein Sci.*, 23:652–661, 2014.

[51] G. D. Rose and R. Wolfenden. Hydrogen bonding, hydrophobicity, packing, and protein folding. *Annu. Rev. Biophys. Biomol. Struct.*, 22:381–415, 1993.

[52] M. Blaber, J. D. Lindstrom, N. Gassner, J. Xu, D. W. Heinz, and B. W. Matthews. Energetic cost and structural consequences of burying a hydroxyl group within the core of a protein determined from ala→ser and val→thr substitutions in T4 lysozyme. *Biochemistry*, 32:11363–11373, 1993.

[53] D. Chandler. Interfaces and the driving force of hydrophobic assembly. *Nature*, 437:640, 2005.

[54] W. Kauzmann. Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.*, 14:1–63, 1959.

[55] V. Daggett and A. Fersht. The present view of the mechanism of protein folding. *Nat. Rev. Mol. Cell Bio.*, 4:497, 2003.

[56] R. F. Doolittle. The multiplicity of domains in proteins. *Annu. Rev. Biochem.*, 64:287–314, 1995.

[57] T. Pawson. Protein modules and signalling networks. *Nature*, 373:573, 1995.

[58] S. E. Jackson. How do small single-domain proteins fold? *Fold. Des.*, 3:R81–R91, 1998.

[59] E. I. Shakhnovich. Theoretical studies of protein-folding thermodynamics and kinetics. *Current opinion in structural biology*, 7:29–40, 1997.

[60] M. Socolich, S. W. Lockless, W. P. Russ, H. Lee, K. H. Gardner, and R. Ranganathan. Evolutionary information for specifying a protein fold. *Nature*, 437:512, 2005.

[61] L. Gamier, J. W. Wills, M. F. Verderame, and M. Sudol. WW domains and retrovirus budding. *Nature*, 381:744, 1996.

[62] M. Jäger, H. Nguyen, J.C. Crane, J.W. Kelly, and M. Gruebele. The folding mechanism of a β-sheet: The WW domain. *J. Mol. Biol.*, 311:373–393, 2001.

[63] N. Ferguson, C.M. Johnson, M. Macias, H. Oschkinat, and A.R. Fersht. Ultrafast folding of WW domains without structured aromatic clusters in the denatured state. *Proc. Natl. Acad. Sci. USA*, 98:13002–13007, 2001.

[64] S. Deechongkit, P.E. Dawson, and J.W. Kelly. Toward assessing the position-dependent contributions of backbone hydrogen bonding to β-sheet folding thermodynamics employing amide-to-ester perturbations. *J. Am. Chem. Soc.*, 126:16762–16771, 2004.

[65] E.K. Koepf, H.M. Petrassi, M. Sudol, and J.W. Kelly. WW: An isolated three-stranded antiparallel β-sheet domain that unfolds and refolds reversibly; Evidence for a structured hydrophobic cluster in urea and GdnHCl and a disordered thermal unfolded state. *Protein Sci.*, 8:841–853, 1999.

[66] K. Dave, M. Jäger, H. Nguyen, J. W. Kelly, and M. Gruebele. High-resolution mapping of the folding transition state of a ww domain. *J. Mol. Biol.*, 428:1617–1636, 2016.

[67] H. Nguyen, M. Jäger, A. Moretto, M. Gruebele, and J. W. Kelly. Tuning the free-energy landscape of a ww domain by temperature, mutation, and truncation. *Proc. Natl. Acad. Sci. USA*, 100:3948–3953, 2003.

[68] F. Liu, D. Du, A. A. Fuller, J. E. Davoren, P. Wipf, J. W. Kelly, and M. Gruebele. An experimental survey of the transition between two-state and downhill protein folding scenarios. *Proc. Natl. Acad. Sci. USA*, 105:2369–2374, 2008.

[69] C.J. Noren, S.J. Anthony-Cahill, M.C. Griffith, and P.G. Schultz. A general method for site-specific incorporation of unnatural amino acids into proteins. *Science*, 244:182, 1989.

[70] D. R. Canchi and A. E. Garcia. Cosolvent effects on protein stability. *Annu. Rev. Phys. Chem.*, 64:273–293, 2013.

[71] B. Ibarra-Molero and J. M. Sanchez-Ruiz. A model-independent, nonlinear extrapolation procedure for the characterization of protein folding energetics from solvent-denaturation data. *Biochemistry*, 35:14689–14702, 1996.

[72] J. L. Silva and G. Weber. Pressure stability of proteins. *Annu. Rev. Phys. Chem.*, 44:89–113, 1993.

[73] C. Krywka, C. Sternemann, M. Paulus, M. Tolan, C. Royer, and R. Winter. Effect of osmolytes on pressure-induced unfolding of proteins: A high-pressure SAXS study. *ChemPhysChem*, 9:2809–2815, 2008.

[74] P. H. Yancey, M. E. Clark, S. C. Hand, R. D. Bowlus, and G. N. Somero. Living with water stress: evolution of osmolyte systems. *Science*, 217:1214–1222, 1982.

[75] R. Singh, I. Haque, and F. Ahmad. Counteracting osmolyte trimethylamine n-oxide destabilizes proteins at ph below its pka. *J. Biol. Chem.*, 280:11035–11042, 2005.

[76] L. B. Sagle, Y. Zhang, V. A. Litosh, X. Chen, Y. Cho, and P. S. Cremer. Investigating the hydrogen-bonding model of urea denaturation. *J. Am. Chem. Soc.*, 131:9304–9310, 2009.

[77] J. Mondal, G. Stirnemann, and B. J. Berne. When does trimethylamine N-oxide fold a polymer chain and urea unfold it. *J. Phys. Chem. B*, 117:8723–8732, 2013.

[78] H.-J. Schneider. Binding mechanisms in supramolecular complexes. *Angew. Chem. Int. Ed.*, 48:3924–3977, 2009.

[79] W. Chen, C. Chang, and M. K. Gilson. Calculation of cyclodextrin binding affinities: energy, entropy, and implications for drug design. *Biophys. J.*, 87:3035–3049, 2004.

[80] K. Teilum, J. G. Olsen, and B. B. Kragelund. Protein stability, flexibility and function. *Biochim. Biophys. Acta*, 1814:969–976, 2011.

[81] Christophe Chipot and Andrew Pohorille. *Free energy calculations*. Springer, 2007.

[82] C. Chipot. Frontiers in free-energy calculations of biological systems. *Wiley Interdiscip. Rev.-Comput. Mol. Sci.*, 4:71–89, 2014.

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

[84] S. Boresch, F. Tettinger, M. Leitgeb, and M. Karplus. Absolute binding free energies: A quantitative approach for their calculation. *J. Phys. Chem. B*, 107:9535–9551, 2003.

[85] N. Hansen and W. F. Van Gunsteren. Practical aspects of free-energy calculations: a review. *J. Chem. Theory Comput.*, 10:2632–2647, 2014.

[86] H. Zhang, T. Tan, C. Hetényi, and D. Van Der Spoel. Quantification of solvent contribution to the stability of noncovalent complexes. *J. Chem. Theory Comput.*, 9:4542–4551, 2013.

[87] D. L. Mobley and M. K. Gilson. Predicting binding free energies: Frontiers and benchmarks. *Annu. Rev. Biophys.*, 46:531–558, 2017.

[88] Y. Matsui and K. Mochida. Binding forces contributing to the association of cyclodextrin with alcohol in an aqueous solution. *Bull. Chem. Soc. Jpn.*, 52:2808–2814, 1979.

[89] M. V. Rekharsky and Y. Inoue. Complexation thermodynamics of cyclodextrins. *Chem. Rev.*, 98:1875–1918, 1998.

[90] C. Cézard, X. Trivelli, F. Aubry, F. Djedaïni-Pilard, and F. Dupradeau. Molecular dynamics studies of native and substituted cyclodextrins in different media: 1. charge derivation and force field performances. *Phys. Chem. Chem. Phys.*, 13:15103–15121, 2011.

[91] J. Gebhardt, C. Kleist, S. Jakobtorweihen, and N. Hansen. Validation and comparison of force fields for native cyclodextrins in aqueous solution. *J. Phys. Chem. B*, 122:1608–1626, 2018.

[92] G. M. Torrie and J. P. Valleau. Monte Carlo free energy estimates using non-boltzmann sampling: Application to the sub-critical lennard-jones fluid. *Chem. Phys. Lett.*, 28:578–581, 1974.

[93] W. F. van Gunsteren, D. Bakowies, R. Baron, I. Chandrasekhar, M. Christen, X. Daura, P. Gee, D. P. Geerke, A. Glättli, P. H. Hünenberger, M. A. Kastenholz, C. Oostenbrink, M. Schenk, D. Trzesniak, N. F. A. van der Vegt, and H. B. Yu. Biomolecular modeling: Goals, problems, perspectives. *Angew. Chem. Int. Ed.*, 45:4064–4092, 2006.

[94] Z. Su and C. L. Dias. Individual and combined effects of urea and trimethylamine n-oxide (tmao) on protein structures. *J. Mol. Liq.*, 293:111443, 2019.

[95] D. R. Canchi, P. Jayasimha, D. C. Rau, G. I. Makhatadze, and A. E. Garcia. Molecular mechanism for the preferential exclusion of tmao from protein surfaces. *J. Phys. Chem. B*, 116:12095–12104, 2012.

[96] E. Schneck, D. Horinek, and R. R. Netz. Insight into the molecular mechanisms of protein stabilizing osmolytes from global force-field variations. *J. Phys. Chem. B*, 117:8310–8321, 2013.

[97] L. Larini and J.-E. Shea. Double resolution model for studying tmao/water effective interactions. *J. Phys. Chem. B*, 117:13268–13277, 2013.

[98] K. M. Kast, J. Brickmann, S. M. Kast, and R. S. Berry. Binary phases of aliphatic n-oxides and water: Force field development and molecular dynamics simulation. *J. Phys. Chem. A*, 107:5342–5351, 2003.

[99] C. Hölzl, P. Kibies, S. Imoto, R. Frach, S. Suladze, R. Winter, D. Marx, D. Horinek, and S. M. Kast. Design principles for high-pressure force fields: Aqueous tmao solutions from ambient to kilobar pressures. *J. Chem. Phys.*, 144:144104, 2016.

[100] A.P. Eichenberger, W.F. van Gunsteren, S. Riniker, L. von Ziegler, and N. Hansen. The key to predicting the stability of protein mutants lies in an accurate description and proper configurational sampling of the folded and denatured states. *Biochim. Biophys. Acta-General Subjects*, 1850:983–995, 2015.

[101] S. N. Timasheff. Control of protein stability and reactions by weakly interacting cosolvents: The simplicity of the complicated. *Adv. Prot. Chem.*, 51:355–432, 1998.

[102] S. N. Timasheff. Protein-solvent preferential interactions, protein hydration, and the modulation of biochemical reactions by solvent components. *Proc. Natl. Acad. Sci. USA*, 99:9721–9726, 2002.

[103] M. A. Schroer, Y. Zhai, D. C. Wieland, C. J. Sahle, J. Nase, M. Paulus, M. Tolan, and R. Winter. Exploring the piezophilic behavior of natural cosolvent mixtures. *Angew. Chem., Int. Ed.*, 50:11413–11416, 2011. *Angew. Chem.* 2011, *123*, 11615–11618.

[104] Y.-T. Liao, A. C. Manson, M. R. DeLyser, W. G. Noid, and P. S. Cremer. Trimethylamine n-oxide stabilizes proteins via a distinct mechanism compared with betaine and glycine. *Proc. Natl. Acad. Sci. USA*, 114:2479–2484, 2017.

[105] M. A. Schroer, J. Michalowsky, B. Fischer, J. Smiatek, and G. Grübel. Stabilizing effect of tmao on globular pnipam states: Preferential attraction induces preferential hydration. *Phys. Chem. Chem. Phys.*, 18:31459–31470, 2016.

[106] A. H. Gorensek-Benitez, A. E. Smith, S. S. Stadmiller, G. M. Perez Goncalvez, and G. J. Pielak. Cosolutes, crowding, and protein folding kinetics. *J. Phys. Chem. B*, 121:6527–6537, 2017.

[107] K. Usui, Y. Nagata, J. Hunger, M. Bonn, and M. Sulpizi. A new force field including charge directionality for tmao in aqueous solution. *J. Chem. Phys.*, 145:064103, 2016.

[108] A. Panuszko, P. Bruździak, J. Zielkiewicz, D. Wyrzykowski, and J. Stangret. Effects of urea and trimethylamine-n-oxide on the properties of water and the secondary structure of hen egg white lysozyme. *J. Phys. Chem. B*, 113:14797–14809, 2009.

[109] C. Y. Hu, G. C. Lynch, H. Kokubo, and B. M. Pettitt. Trimethylamine n-oxide influence on the backbone of proteins: An oligoglycine model. *Proteins*, 78:695–704, 2010.

[110] R. Sarma and S. Paul. The effect of aqueous solutions of trimethylamine-n-oxide on pressure induced modifications of hydrophobic interactions. *J. Chem. Phys.*, 137:094502, 2012.

[111] R. Sarma and S. Paul. Trimethylamine-n-oxide's effect on polypeptide solvation at high pressure: A molecular dynamics simulation study. *J. Phys. Chem. B*, 117:9056–9066, 2013.

[112] P. Ganguly, T. Hajari, J.-E. Shea, and N. van der Vegt. Mutual exclusion of urea and trimethylamine-N-oxide from amino acids in mixed solvent environment. *J. Phys. Chem. Lett.*, 6:581–585, 2015.

[113] P. Ganguly, T. Hajari, J.-E. Shea, and N. van der Vegt. Correction to "mutual exclusion of urea and trimethylamine-N-oxide from amino acids in mixed solvent environment". *J. Phys. Chem. Lett.*, 6:4728–4729, 2015.

[114] G. Borgohain and S. Paul. Model dependency of tmao's counteracting effect against action of urea: Kast model versus osmotic model of tmao. *J. Phys. Chem. B*, 120:2352–2361, 2016.

[115] F. Rodríguez-Ropero, P. Rötzscher, and N. F. A. van der Vegt. Comparison of different tmao force fields and their impact on the folding equilibrium of a hydrophobic polymer. *J. Phys. Chem. B*, 120:8757–8767, 2016.

[116] F. Rodríguez-Ropero, P. Rötzscher, and N. F. A. van der Vegt. Correction to "comparison of different tmao force fields and their impact on the folding equilibrium of a hydrophobic polymer". *J. Phys. Chem. B*, 121:1455–1455, 2017.

[117] N. Smolin, V. P. Voloshin, A. V. Anikeenko, A. Geiger, R. Winter, and N. N. Medvedev. Tmao and urea in the hydration shell of the protein snase. *Phys. Chem. Chem. Phys.*, 19:6345–6357, 2017.

[118] P. I. Nagy. Structure simulations of solutions with small organic solutes at ambient temperature and ph = 7. *Curr. Phys. Chem.*, 4:330–392, 2014.

[119] K. Stöbener, P. Klein, M. Horsch, K. Küfer, and H. Hasse. Parametrization of two-center lennard-jones plus point-quadrupole force field models by multicriteria optimization. *Fluid Phase Equilib.*, 411:33–42, 2016.

[120] I. Shvab and R. J. Sadus. Atomistic water models: Aqueous thermodynamic properties from ambient to supercritical conditions. *Fluid Phase Equilib.*, 407:7–30, 2016.

[121] C. Oostenbrink, D. Juchli, and W. F. van Gunsteren. Amine hydration: A united-atom force-field solution. *ChemPhysChem*, 6:1800–1804, 2005.

[122] N. M. Fischer, P. J. van Maaren, J. C. Ditz, A. Yildirim, and D. van der Spoel. Properties of organic liquids when simulated with long-range lennard-jones interactions. *J. Chem. Theory Comput.*, 11:2938–2944, 2015.

[123] E. A. Ploetz, A. S. Rustenburg, D. P. Geerke, and P. E. Smith. To polarize or not to polarize? charge-on-spring versus kbff models for water and methanol bulk and vapor-liquid interfacial mixtures. *J. Chem. Theory Comput.*, 12:2373–2387, 2016.

[124] J. Rösgen, B. M. Pettitt, and D. W. Bolen. Uncovering the basis for nonideal behavior of biological molecules. *Biochemistry*, 43:14472–14484, 2004.

[125] J. Hunger, K.-J. Tielrooij, R. Buchner, M. Bonn, and H. J. Bakker. Complex formation in aqueous trimethylamine-N-oxide (TMAO) solutions. *J. Phys. Chem. B*, 116:4783–4795, 2012.

[126] T. Kulschewski and J. Pleiss. A molecular dynamics study of liquid aliphatic alcohols: Simulation of density and self-diffusion coefficient using a modified opls force field. *Mol. Simul.*, 39:754–767, 2013.

[127] L. Sapir and D. Harries. Is the depletion force entropic? molecular crowding beyond steric interactions. *Curr. Opin. Colloid Interface Sci.*, 20:3–10, 2015.

[128] N. F. A. van der Vegt and D. Nayar. The hydrophobic effect and the role of cosolvents. *J. Phys. Chem. B*, 121:9986–9998, 2017.

[129] V. Pierce, M. Kang, M. Aburi, S. Weerasinghe, and P. E. Smith. Recent applications of kirkwood-buff theory to biological systems. *Cell Biochem. Biophys.*, 50:1–22, 2008.

[130] S. Shimizu and P. E. Smith. How osmolytes counteract pressure denaturation on a molecular scale. *ChemPhysChem*, 18:2243–2249, 2017.

[131] J. Smiatek. Aqueous ionic liquids and their effects on protein structures: an overview on recent theoretical and experimental results. *J. Phys. Condens. Matter*, 29:233001, 2017.

[132] M. Fioroni, K. Burger, A. E. Mark, and D. Roccatano. A new 2,2,2-trifluoroethanol model for molecular dynamics simulations. *J. Phys. Chem. B*, 104:12347–12354, 2000.

[133] D. Roccatano, G. Colombo, M. Fioroni, and A. E. Mark. Mechanism by which 2,2,2-trifluoroethanol/water mixtures stabilize secondary-structure formation in peptides: A molecular dynamics study. *Proc. Natl. Acad. Sci. USA*, 99:12179–12184, 2002.

[134] A. P. Eichenberger, W. F. van Gunsteren, and L. J. Smith. Structure of hen egg-white lysozyme solvated in tfe/water: A molecular dynamics simulation study based on nmr data. *J. Biomol. NMR*, 55:339–353, 2013.

[135] L. J. Smith, H. J. C. Berendsen, and W. F. van Gunsteren. Computer simulation of urea-water mixtures: A test of force field parameters for use in biomolecular simulation. *J. Phys. Chem. B*, 108:1065–1071, 2004.

[136] D. Trzesniak, N. F. A. van der Vegt, and W. F. van Gunsteren. Computer simulation studies on the solvation of aliphatic hydrocarbons in 6.9 M aqueous urea solution. *Phys. Chem. Chem. Phys.*, 6:697–702, 2004.

[137] L. J. Smith, R. M. Jones, and W. F. van Gunsteren. Characterization of the denaturation of human α-lactalbumin in urea by molecular dynamics simulations. *Proteins: Struct., Funct., Bioinf.*, 58:439–449, 2005.

[138] W. F. van Gunsteren, S. R. Billeter, A. A. Eising, P. H. Hünenberger, P. Krüger, A. E. Mark, W. R. P. Scott, and I. G. Tironi. *Biomolecular Simulation: The GROMOS96 Manual and User Guide.* Vdf Hochschulverlag AG an der ETH Zürich, Zürich, Groningen, 1996.

[139] X. Daura, A. E. Mark, and W. F. van Gunsteren. Parametrization of aliphatic chn united atoms of GROMOS96 force field. *J. Comput. Chem.*, 19:535–547, 1998.

[140] W. F. van Gunsteren, X. Daura, and A. E. Mark. GROMOS force field. In P. Schleyer, editor, *Encyclopedia of Computational Chemistry*, volume 2, pages 1211–1216. John Wiley & Sons, Chichester, 1998.

[141] L. D. Schuler, X. Daura, and W. F. van Gunsteren. An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *J. Comput. Chem.*, 22:1205–1218, 2001.

[142] I. Chandrasekhar, M. A. Kastenholz, R. D. Lins, C. Oostenbrink, L. D. Schuler, D. Tieleman, and W. F. van Gunsteren. A consistent potential energy parameter set for lipids: Dipalmitoylphosphatidylcholine as a benchmark of the GROMOS96 45A3 force field. *Eur. Biophys. J.*, 32:67–77, 2003.

[143] C. Oostenbrink, A. Villa, A.E. Mark, and W.F. van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.*, 25:1656–1676, 2004.

[144] T. A. Soares, P. H. Hünenberger, M. A. Kastenholz, V. Kräutler, T. Lenz, R. D. Lins, C. Oostenbrink, and W. F. van Gunsteren. An improved nucleic-acid parameter set for the GROMOS force field. *J. Comput. Chem.*, 26:725–737, 2005.

[145] R. D. Lins and P.H. Hünenberger. A new GROMOS force field for hexopyranose-based carbohydrates. *J. Comput. Chem.*, 26:1400–1412, 2005.

[146] D. Poger, W. F. van Gunsteren, and A. E. Mark. A new force field for simulating phosphatidylcholine bilayers. *J. Comput. Chem.*, 31:1117–1125, 2010.

[147] M. M. Reif, P. H. Hünenberger, and C. Oostenbrink. New interaction parameters for charged amino acid side chains in the GROMOS force field. *J. Chem. Theory Comput.*, 8:3705–3723, 2012.

[148] P. Atkins and J. de Paula, editors. *Atkins Physical Chemistry.* Oxford University Press, New York, 6th edition, 2006.

[149] M. Sega, S. S. Kantorovich, A. Arnold, and C. Holm. On the calculation of the dielectric properties of liquid ionic systems. In Yuri P. Kalmykov, editor, *Recent Advances in Broadband Dielectric Spectroscopy*, NATO Science for Peace and Security Series B: Physics and Biophysics, pages 103–122. Springer Netherlands, 2013.

[150] J. M. Caillol, D. Levesque, and J. J. Weis. Theoretical calculation of ionic solution properties. *J. Chem. Phys.*, 85:6645–6657, 1986.

[151] A. N. Krishnamoorthy, J. Zeman, C. Holm, and J. Smiatek. Preferential solvation and ion association properties in aqueous dimethyl sulfoxide solutions. *Phys. Chem. Chem. Phys.*, 18:31312–31322, 2016.

[152] C. Schröder and O. Steinhauser. On the dielectric conductivity of molecular ionic liquids. *J. Chem. Phys.*, 131:114504, 2009.

[153] A. K. Malde, L. Zuo, M. Breeze, M. Stroet, D. Poger, P. C. Nair, C. Oostenbrink, and A. E. Mark. An automated force field topology builder (ATB) and repository: Version 1.0. *J. Chem. Theory Comput.*, 7:4026–4037, 2011.

[154] W. L. Jorgensen. Quantum and statistical mechanical studies of liquids. 10. transferable intermolecular potential functions for water, alcohols, and ethers. application to liquid water. *J. Am. Chem. Soc.*, 103:335–340, 1981.

[155] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79:926–935, 1983.

[156] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma. The missing term in effective pair potentials. *J. Phys. Chem.*, 91:6269–6271, 1987.

[157] R. Sinibaldi, C. Casieri, S. Melchionna, G. Onori, A. L. Segre, S. Viel, L. Mannina, and F. De Luca. The role of water coordination in binary mixtures. a study of two model amphiphilic molecules in aqueous solutions by molecular dynamics and nmr. *J. Phys. Chem. B*, 110:8885, 2006.

[158] J. L. F. Abascal and C. Vega. A general purpose model for the condensed phases of water: TIP4P/2005. *J. Chem. Phys.*, 123:234505, 2005.

[159] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development and testing of a general Amber force field. *J. Comput. Chem.*, 25:1157–1174, 2004.

[160] A.T. Hagler, E. Huler, and S. Lifson. Energy functions for peptides and proteins. i. derivation of a consistent force field including the hydrogen bond from amide crystals. *J. Am. Chem. Soc.*, 96:5319–5327, 1974.

[161] S. Lifson, A. T. Hagler, and P. Dauber. Consistent force field studies of intermolecular forces in hydrogen-bonded crystals. 1. carboxylic acids, amides, and the c=o···h-hydrogen bonds. *J. Am. Chem. Soc.*, 101:5111–5121, 1979.

[162] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, and J. Hermans. Interaction models for water in relation to protein hydration. In B. Pullmann, editor, *Intermolecular Forces*, pages 331–342. Reidel, Dordrecht, the Netherlands, 1981.

[163] C. Held and G. Sadowski. Compatible solutes: Thermodynamic properties relevant for effective protection against osmotic stress. *Fluid Phase Equilib.*, 407:224–235, 2016.

[164] Y. L. A. Rezus and H. J. Bakker. Destabilization of the hydrogen-bond structure of water by the osmolyte trimethylamine n-oxide. *J. Phys. Chem. B*, 113:4038–4044, 2009.

[165] R. W. Hockney. The potential calculation and some applications. *Methods Comput. Phys.*, 9:136–211, 1970.

[166] H.J.C. Berendsen, D. van der Spoel, and R. van Drunen. GROMACS: a message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.*, 91:43–56, 1995.

[167] E. Lindahl, B. Hess, and D. van der Spoel. GROMACS 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Model.*, 7:306–317, 2001.

[168] D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A.E. Mark, and H.J.C. Berendsen. GROMACS: Fast, flexible, and free. *J. Comput. Chem.*, 26:1701–1718, 2005.

[169] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, 4:435–447, 2008.

[170] N. Schmid, C.D. Christ, M. Christen, A.P. Eichenberger, and W.F. van Gunsteren. Architecture, implementation and parallelisation of the GROMOS software for biomolecular simulation. *Comput. Phys. Commun.*, 183:890–903, 2012.

[171] A.E. Kunz, J.R. Allison, D.P. Geerke, B.A.C. Horta, P.H. Hünenberger, S. Riniker, N. Schmid, and W.F. van Gunsteren. New functionalities in the GROMOS biomolecular simulation software. *J. Comput. Chem.*, 33:340–353, 2012.

[172] S. Riniker, C.D. Christ, H.S. Hansen, P.H. Hünenberger, C. Oostenbrink, D. Steiner, and W.F. van Gunsteren. Calculation of relative free energies for ligand-protein binding, solvation, and conformational transitions using the GROMOS software. *J. Phys. Chem. B*, 115:13570–13577, 2011.

[173] A.P. Eichenberger, J.R. Allison, J. Dolenc, D.P. Geerke, B.A.C. Horta, K. Meier, C. Oostenbrink, N. Schmid, D. Steiner, D. Wang, and W.F. van Gunsteren. GROMOS++ software for the analysis of biomolecular simulation trajectories. *J. Chem. Theory Comput.*, 7:3379–3390, 2011.

[174] S. Miyamoto and P.A. Kollman. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.*, 13:952–962, 1992.

[175] B. Hess, H. Bekker, H.J.C. Berendsen, and J.G.E.M. Fraaije. LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.*, 18:1463–1472, 1997.

[176] B. Hess. P-LINCS: a parallel linear constraint solver for molecular simulation. *J. Chem. Theory Comput.*, 4:116–122, 2008.

[177] S. Nosé. A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.*, 52:255–268, 1984.

[178] W.G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A*, 31:1695, 1985.

[179] G.J. Martyna, M.E. Tuckerman, D.J. Tobias, and M.L. Klein. Explicit reversible integrators for extended systems dynamics. *Mol. Phys.*, 87:1117–1157, 1996.

[180] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.*, 52:7182–7190, 1981.

[181] S. Nosé and M.L. Klein. Constant pressure molecular dynamics for molecular systems. *Mol. Phys.*, 50:1055–1076, 1983.

[182] C. Caleman, P. J. van Maaren, M. Hong, J. S. Hub, L. T. Costa, and D. van der Spoel. Force field benchmark of organic liquids: Density, enthalpy of vaporization,

heat capacities, surface tension, isothermal compressibility, volumetric expansion coefficient, and dielectric constant. *J. Chem. Theory Comput.*, 8:61–74, 2012.

[183] S. Páll and B. Hess. A flexible algorithm for calculating pair interactions on SIMD architectures. *Comput. Phys. Commun.*, 184:2641–2650, 2013.

[184] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. *J. Chem. Phys.*, 98:10089–10092, 1993.

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

[186] http://www.gromos.net.

[187] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comput. Phys.*, 23:327–341, 1977.

[188] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, A. Di Nola, and J.R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81(8):3684–3690, 1984.

[189] G. S. Kell. Precise representation of volume properties of water at one atmosphere. *J. Chem. Eng. Data*, 12:66–69, 1967.

[190] D. M. Makarov, G. I. Egorov, and A. M. Kolker. Density and volumetric properties of aqueous solutions of trimethylamine N-oxide in the temperature range from (278.15 to 323.15) K and at pressures up to 100 MPa. *J. Chem. Eng. Data*, 60:1291–1299, 2015.

[191] G. Mie. Zur kinetischen Theorie der einatomigen Körper. *Ann. Phys.*, 316:657–697, 1903.

[192] J. E. Jones. On the determination of molecular fields. I. from the variation of the viscosity of a gas with temperature. *Proc. R. Soc. London, Ser. A*, 106:441–462, 1924.

[193] J. E. Jones. On the determination of molecular fields. ii. from the equation of state of a gas. *Proc. R. Soc. London, Ser. A*, 106:463–477, 1924.

[194] J. A. Barker and R. O. Watts. Monte Carlo studies of the dielectric properties of water-like models. *Mol. Phys.*, 26:789–792, 1973.

[195] H. J. C. Berendsen, W. F. van Gunsteren, H. R. J. Zwinderman, and R. G. Geurtsen. Simulations of proteins in water. *Ann. N. Y. Acad. Sci.*, 482:269–285, 1986.

[196] T. N. Heinz and P. H. Hünenberger. Combining the lattice-sum and reaction-field approaches for evaluating long-range electrostatic interactions in molecular simulations. *J. Chem. Phys.*, 123:034107, 2005.

[197] M. Christen, P. H. Hünenberger, D. Bakowies, R. Baron, R. Bürgi, D. P. Geerke, T. N. Heinz, M. A. Kastenholz, V. Kräutler, C. Oostenbrink, C. Peter, D. Trzesniak, and W. F. van Gunsteren. The GROMOS software for biomolecular simulation: GROMOS05. *J. Comput. Chem.*, 26:1719–1751, 2005.

[198] K. S. Shing and S.-T. Chung. Calculation of infinite-dilution partial molar properties by computer simulation. *AIChE J.*, 34:1973–1980, 1988.

[199] B. Dahlgren, M. M. Reif, P. H. Hünenberger, and N. Hansen. Calculation of derivative thermodynamic hydration and aqueous partial molar properties of ions based on atomistic simulations. *J. Chem. Theory Comput.*, 8:3542–3564, 2012.

[200] E. A. Ploetz and P. E. Smith. Infinitely dilute partial molar properties of proteins from computer simulation. *J. Phys. Chem. B*, 118:12844–12854, 2014.

[201] J. Z. Vilseck, J. Tirado-Rives, and W. L. Jorgensen. Determination of partial molar volumes from free energy perturbation theory. *Phys. Chem. Chem. Phys.*, 17:8407–8415, 2015.

[202] L. G. Hepler. Thermal expansion and structure in water and aqueous solution. *Can. J. Chem.*, 47:4613–4617, 1969.

[203] J. C. R. Reis. Theory of partial molar properties. *J. Chem. Soc. Faraday Trans. 2*, 78:1595–1608, 1982.

[204] P. E. Smith and W. F. van Gunsteren. The viscosity of SPC and SPC/E water at 277 and 300 K. *Chem. Phys. Lett.*, 215:315–318, 1993.

[205] B. Hess. Determining the shear viscosity of model liquids from molecular dynamics simulations. *J. Chem. Phys.*, 116:209–217, 2002.

[206] Y. Zhang, A. Otani, and E. J. Maginn. Reliable viscosity calculation from equilibrium molecular dynamics simulations: A time decomposition method. *J. Chem. Theory Comput.*, 11:3537–3546, 2015.

[207] A. Einstein. über die von der molekularkinetischen theorie der wärme geforderte bewegung von in ruhenden flüssigkeiten suspendierten teilchen. *Ann. Phys.*, 322:549–560, 1905.

[208] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Oxford University Press, New York, USA, 1987.

[209] I. C. Yeh and G. Hummer. System-size dependence of diffusion coefficients and viscosities from molecular dynamics simulations with periodic boundary conditions. *J. Phys. Chem. B*, 108:15873–15879, 2004.

[210] D. van der Spoel, P. J. van Maaren, and H. J. C. Berendsen. A systematic study of water models for molecular simulation: Derivation of water models optimized for use with a reaction field. *J. Chem. Phys.*, 108:10220–10230, 1998.

[211] J. Krakowiak, J. Wawer, and A. Panuszko. Densimetric and ultrasonic characterization of urea and its derivatives in water. *J. Chem. Thermodyn.*, 58:211–220, 2013.

[212] K. Mazur, I. A. Heisler, and S. R. Meech. Aqueous solvation of amphiphilic solutes: Concentration and temperature dependent study of the ultrafast polarisability relaxation dynamics. *Phys. Chem. Chem. Phys.*, 14:6343–6351, 2012.

[213] C. Vega and J. L. F. Abascal. Simulating water with rigid non-polarizable models: A general perspective. *Phys. Chem. Chem. Phys.*, 13:19663–19688, 2011.

[214] J. Rösgen and R. Jackson-Atogi. Volume exclusion and h-bonding dominate the thermodynamics and solvation of trimethylamine-n-oxide in aqueous urea. *J. Am. Chem. Soc.*, 134:3590–3597, 2012.

[215] L. Knake, G. Schwaab, K. Kartaschew, and M. Havenith. Solvation dynamics of trimethylamine-n-oxide in aqueous solution probed by terahertz spectroscopy. *J. Phys. Chem. B*, 119:13842–13851, 2015.

[216] A. Luzar and D. Chandler. Hydrogen-bond kinetics in liquid water. *Nature*, 379:55, 1996.

[217] S. Imoto, H. Forbert, and D. Marx. Water structure and solvation of osmolytes at high hydrostatic pressure: Pure water and tmao solutions at 10 kbar versus 1 bar. *Phys. Chem. Chem. Phys.*, 17:24224–24237, 2015.

[218] M. W. Mahoney and W. L. Jorgensen. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J. Chem. Phys.*, 112:8910–8922, 2000.

[219] V. Gapsys, S. Michielssens, D. Seeliger, and B.L. de Groot. Accurate and rigorous prediction of the changes in protein free energies in a large-scale mutation scan. *Angew. Chem. Int. Ed.*, 55:7364–7490, 2016. *Angew. Chem.*, 128: 7490–7494.

[220] T. Steinbrecher, C. Zhu, L. Wang, R. Abel, C. Negron, D. Pearlman, E. Feyfant, J. Duan, and W. Sherman. Predicting the effect of amino acid single-point mutations on protein stability—Large-scale validation of MD-based relative free energy calculations. *J. Mol. Biol.*, 429:948–963, 2017.

[221] S. Deechongkit, H. Nguyen, E.T. Powers, P.E. Dawson, M. Gruebele, and J.W. Kelly. Context-dependent contributions of backbone hydrogen bonding to β-sheet folding energetics. *Nature*, 430:101, 2004.

[222] D.M. Zuckerman. Equilibrium sampling in biomolecular simulations. *Annu. Rev. Biophys.*, 40:41–62, 2011.

[223] R. Ranganathan, K.P. Lu, T. Hunter, and J.P. Noel. Structural and functional analysis of the mitotic rotamase Pin1 suggests substrate recognition is phosphorylation dependent. *Cell*, 89:875–886, 1997.

[224] X. Periole and A.E. Mark. Convergence and sampling efficiency in replica exchange simulations of peptide folding in explicit solvent. *J. Chem. Phys.*, 126:014903, 2007.

[225] Y. Pan and V. Daggett. Direct comparison of experimental and calculated folding free energies for hydrophobic deletion mutants of chymotrypsin inhibitor 2: Free energy perturbation calculations using transition and denatured states from molecular dynamics simulations of unfolding. *Biochemistry*, 40:2723–2731, 2001.

[226] D. Seeliger and B.L. de Groot. Protein thermostability calculations using alchemical free energy simulations. *Biophys. J.*, 98:2309–2316, 2010.

[227] B.A.C. Horta, P.F.J. Fuchs, W.F. van Gunsteren, and P.H. Hünenberger. New interaction parameters for oxygen compounds in the GROMOS force field: Improved pure-liquid and solvation properties for alcohols, ethers, aldehydes, ketones, carboxylic acids, and esters. *J. Chem. Theory Comput.*, 7:1016–1031, 2011.

[228] C.D. Christ and W.F. van Gunsteren. Enveloping distribution sampling: A method to calculate free energy differences from a single simulation. *J. Chem. Phys.*, 126:184110, 2007.

[229] C.D. Christ and W.F. van Gunsteren. Multiple free energies from a single simulation: Extending enveloping distribution sampling to nonoverlapping phase-space distributions. *J. Chem. Phys.*, 128:174112, 2008.

[230] A. Pohorille, C. Jarzynski, and C. Chipot. Good practices in free-energy calculations. *J. Phys. Chem. B*, 114:10235–10253, 2010.

[231] H. Fukunishi, O. Watanabe, and S. Takada. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *J. Chem. Phys.*, 116:9058–9067, 2002.

[232] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314:141–151, 1999.

[233] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic Acids Res.*, 28:235–242, 2000.

[234] J.A. Kowalski, K. Liu, and J.W. Kelly. NMR solution structure of the isolated apo Pin1 WW domain: Comparison to the X-ray crystal structures of Pin1. *Biopolymers*, 63:111–121, 2002.

[235] J.W. Eastwood and R.W. Hockney. *Computer Simulation Using Particles.* N. Y.: McGraw-Hill, 1981.

[236] G. Bussi. Hamiltonian replica exchange in GROMACS:a flexible implementation. *Mol. Phys.*, 112:379–384, 2014.

[237] M.R. Shirts and J.D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.*, 129:124105, 2008.

[238] Python implementation of the multistate bennett acceptance ratio. https://github.com/choderalab/pymbar. Accessed: 2018-01-08.

[239] C. H. Bennett. Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.*, 22:245–268, 1976.

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

[241] P.V. Klimovich, M.R. Shirts, and D.L. Mobley. Guidelines for the analysis of free energy calculations. *J. Comput.-Aided Mol. Des.*, 29:397–411, 2015.

[242] M.M.H. Graf, M. Maurer, and C. Oostenbrink. Free-energy calculations of residue mutations in a tripeptide using various methods to overcome inefficient sampling. *J. Comput. Chem.*, 37:2597–2605, 2016.

[243] U. Börjesson and P.H. Hünenberger. Effect of mutations involving charged residues on the stability of Staphylococcal nuclease: A continuum electrostatics study. *Protein Eng.*, 16:831–840, 2003.

[244] *GROMACS Version 2016.5 Reference Manual.*

[245] A.E. Mark and W.F. van Gunsteren. Decomposition of the free energy of a system in terms of specific interactions: Implications for theoretical and experimental studies. *J. Mol. Biol.*, 240:167–176, 1994.

[246] C.D. Christ, A.E. Mark, and W.F. van Gunsteren. Basic ingredients of free energy calculations: A review. *J. Comput. Chem.*, 31:1569–1582, 2010.

[247] J. Gao, D.A. Bosco, E.T. Powers, and J.W. Kelly. Localized thermodynamic coupling between hydrogen bonding and microenvironment polarity substantially stabilizes proteins. *Nat. Struct. Mol. Biol.*, 16:684, 2009.

[248] S. P. Leelananda and S. Lindert. Computational methods in drug discovery. *Beilstein J. Org. Chem.*, 12:2694–2718, 2016.

[249] IUPAP. Symbols, units and nomenclature in physics. *Physica A*, 93:1–60, 1978.

[250] T. Renner. *Quantities, Units and Symbols in Physical Chemistry.* The Royal Society of Chemistry, 2007.

[251] J. A. Pople. Nobel lecture: Quantum chemical models. *Rev. Mod. Phys.*, 71:1267, 1999.

[252] G. König and B. R. Brooks. Predicting binding affinities of host-guest systems in the sampl3 blind challenge: The performance of relative free energy calculations. *J. Comput. Aided Mol. Des.*, 26:543–550, 2012.

[253] J. Wong-ekkabut and M. Karttunen. The good, the bad and the user in soft matter simulations. *Biochim. Biophys. Acta*, 1858:2529–2538, 2016.

[254] Y. Deng and B. Roux. Computations of standard binding free energies with molecular dynamics simulations. *J. Phys. Chem. B*, 113:2234–2246, 2009.

[255] D. Markthaler, J. Gebhardt, S. Jakobtorweihen, and N. Hansen. Molecular simulations of thermodynamic properties for the system α-cyclodextrin/alcohol in aqueous solution. *Chem.-Ing.-Tech.*, 89:1306–1314, 2017.

[256] J. C. Gumbart, B. Roux, and C. Chipot. Standard binding free energies from computer simulations: What is the best strategy? *J. Chem. Theory Comput.*, 9:794–802, 2012.

[257] J. Baz, J. Gebhardt, H. Kraus, D. Markthaler, and N. Hansen. Insights into non-covalent binding obtained from molecular dynamics simulations. *Chem. Ing. Tech.*, 90:1864–1875, 2018.

[258] G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.*, 23:187–199, 1977.

[259] T. Huber, A. E. Torda, and W. F. Van Gunsteren. Local elevation: A method for improving the searching properties of molecular dynamics simulation. *J. Comput. Aided Mol. Des.*, 8:695–708, 1994.

[260] A. Laio and M. Parrinello. Escaping free-energy minima. *Proc. Natl. Acad. Sci. USA*, 99:12562–12566, 2002.

[261] E. Darve, D. Rodríguez-Gómez, and A. Pohorille. Adaptive biasing force method for scalar and vector free energy calculations. *J. Chem. Phys.*, 128:144120, 2008.

[262] R. J. Allen, C. Valeriani, and P. R. ten Wolde. Forward flux sampling for rare event simulations. *J. Phys. Condens. Matter*, 21:463102, 2009.

[263] B. Isralewitz, M. Gao, and K. Schulten. Steered molecular dynamics and mechanical functions of proteins. *Curr. Opin. Cell. Biol.*, 11:224–230, 2001.

[264] E. M. M. Del Valle. Cyclodextrins and their uses: A review. *Process Biochem.*, 39:1033–1046, 2004.

[265] C. Velez-Vega and M. K. Gilson. Force and stress along simulated dissociation pathways of cucurbituril–guest systems. *J. Chem. Theory Comput.*, 8:966–976, 2012.

[266] C. Velez-Vega and M. K. Gilson. Overcoming dissipation in the calculation of standard binding free energies by ligand extraction. *J. Comput. Chem.*, 34:2360–2371, 2013.

[267] D. Yordanova, E. Ritter, T. Gerlach, J.-H. Jensen, I. Smirnova, and S. Jakobtorweihen. Solute partitioning in micelles: Combining molecular dynamics simulations, COSMOmic, and experiments. *J. Phys. Chem. B*, 121:5794–5809, 2017.

[268] T. W. Allen, O. S. Andersen, and B. Roux. Ion permeation through a narrow channel: Using gramicidin to ascertain all-atom molecular dynamics potential of mean force methodology and biomolecular force fields. *Biophys. J.*, 90:3447–3468, 2006.

[269] T. W. Allen, O. S. Andersen, and B. Roux. Molecular dynamics—potential of mean force calculations as a tool for understanding ion permeation and selectivity in narrow channels. *Biophys. Chem.*, 124:251–267, 2006.

[270] J. S. Hub, B. L. De Groot, and D. Van Der Spoel. g_wham - a free weighted histogram analysis implementation including robust error and autocorrelation estimates. *J. Chem. Theory Comput.*, 6:3713–3720, 2010.

[271] E. Flood, C. Boiteux, B. Lev, I. Vorobyov, and T. W. Allen. Atomistic simulations of membrane ion channel conduction, gating, and modulation. *Chem. Rev.*, 2019.

[272] H.-J. Woo and B. Roux. Calculation of absolute protein–ligand binding free energy from computer simulations. *Proc. Natl. Acad. Sci. USA*, 102:6825–6830, 2005.

[273] G. Filippini, F. Goujon, C. Bonal, and P. Malfreyt. Energetic competition effects on thermodynamic properties of association between β-CD and Fc group: A potential of mean force approach. *J. Phys. Chem. C*, 116:22350–22358, 2012.

[274] G. Filippini, C. Bonal, and P. Malfreyt. Why is the association of supramolecular assemblies different under homogeneous and heterogeneous conditions? *Phys. Chem. Chem. Phys.*, 14:10122–10124, 2012.

[275] T. Baştuğ, P.-C. Chen, S. M. Patra, and S. Kuyucak. Potential of mean force calculations of ligand binding to ion channels from Jarzynski's equality and umbrella sampling. *J. Chem. Phys.*, 128:155104, 2008.

[276] J. Krüger and G. Fels. Potential of mean force of ion permeation through alpha7 nachrion channel. In *International Workshop on Portals for Life Sciences*, volume 513 of *CEUR Workshop Proceedings*, 2009.

[277] C. Neale, J. C. Y. Hsu, C. M. Yip, and R. Pomès. Indolicidin binding induces thinning of a lipid bilayer. *Biophys. J.*, 106:L29–L31, 2014.

[278] R. Sun, Y. Han, J. M. J. Swanson, J. S Tan, J. P. Rose, and G. A. Voth. Molecular transport through membranes: Accurate permeability coefficients from multidimensional potentials of mean force and local diffusion constants. *J. Chem. Phys.*, 149:072310, 2018.

[279] H.-X. Zhou and M. K. Gilson. Theory of free energy and entropy in noncovalent binding. *Chem. Rev.*, 109:4092–4107, 2009.

[280] I. J. General. A note on the standard state's binding free energy. *J. Chem. Theory Comput.*, 6:2520–2524, 2010.

[281] S. Doudou, N. A. Burton, and R. H. Henchman. Standard free energy of binding from a one-dimensional potential of mean force. *J. Chem. Theory Comput.*, 5:909–918, 2009.

[282] R. W. Zwanzig. High-temperature equation of state by a perturbation method. i. nonpolar gases. *J. Chem. Phys.*, 22:1420–1426, 1954.

[283] J. G. Kirkwood. Statistical mechanics of fluid mixtures. *J. Chem. Phys.*, 3:300–313, 1935.

[284] J. Hermans and L. U. Wang. Inclusion of loss of translational and rotational freedom in theoretical estimates of free energies of binding. application to a complex of benzene and mutant T4 lysozyme. *J. Am. Chem. Soc.*, 119:2707–2714, 1997.

[285] I. Z. Steinberg and H. A. Scheraga. Entropy changes accompanying association reactions of proteins. *J. Biol. Chem.*, 238:172–181, 1963.

[286] R. G. Weiß, R. Chudoba, P. Setny, and J. Dzubiella. Affinity, kinetics, and pathways of anisotropic ligands binding to hydrophobic model pockets. *J. Chem. Phys.*, 149:094902, 2018.

[287] IUPAC-IUB. Symbols for specifying the conformation of polysaccharide chains, recommendations 1981. *Eur. J. Biochem.*, 131:5–7, 1983.

[288] W. You, Z. Tang, and C. A. Chang. Potential mean force from umbrella sampling simulations: What can we learn and what is missed? *J. Chem. Theory Comput.*, 15:2433–2443, 2019.

[289] D. Trzesniak, A.-P. E. Kunz, and W. F. van Gunsteren. A comparison of methods to compute the potential of mean force. *ChemPhysChem*, 8:162–169, 2007.

[290] L. Pol-Fachin, V. H. Rusu, H. Verli, and R. D. Lins. GROMOS 53A6GLYC, an improved GROMOS force field for hexopyranose-based carbohydrates. *J. Chem. Theory Comput.*, 8(11):4681–4690, 2012.

[291] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1:19–25, 2015.

[292] T. F. D. Silva, D. Vila-Viçosa, P. B. P. S. Reis, B. L. Victor, M. Diem, C. Oostenbrink, and M. Machuqueiro. The impact of using single atomistic long-range cutoff schemes with the GROMOS 54A7 force field. *J. Chem. Theory Comput.*, 14:5823–5833, 2018.

[293] Y. M. H. Gonçalves, C. Senac, P. F. Jimi Fuchs, P. H. Hünenberger, and B. A. C. Horta. Influence of the treatment of non-bonded interactions on thermodynamic and transport properties of pure liquids calculated using the 2016H66 force field. *J. Chem. Theory Comput.*, 15:1806–1826, 2019.

[294] S. Reißer, D. Poger, M. Stroet, and A. E. Mark. Real cost of speed: The effect of a time-saving multiple-time-stepping algorithm on the accuracy of molecular dynamics simulations. *J. Chem. Theory Comput.*, 13:2367–2372, 2017.

[295] H. H. Loeffler, S. Bosisio, G. Duarte Ramos Matos, D. Suh, B. Roux, D. L. Mobley, and J. Michel. Reproducibility of free energy calculations across different molecular simulation software packages. *J. Chem. Theory Comput.*, 14:5567–5582, 2018.

[296] M. Papadourakis, S. Bosisio, and J. Michel. Blinded predictions of standard binding free energies: Lessons learned from the sampl6 challenge. *J. Comput. Aided Mol. Des.*, 32:1047–1058, 2018.

[297] J. Åqvist. Cold adaptation of triosephosphate isomerase. *Biochemistry*, 56:4169–4176, 2017.

[298] G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, and G. Bussi. Plumed 2: New feathers for an old bird. *Comput. Phys. Commun.*, 185:604–613, 2014.

[299] T. N. Heinz, W. F. van Gunsteren, and P. H. Hünenberger. Comparison of four methods to compute the dielectric permittivity of liquids from molecular dynamics simulations. *J. Chem. Phys.*, 115:1125–1136, 2001.

[300] E. R. P. Zuiderweg, R. M. Scheek, R. Boelens, W. F. van Gunsteren, and R. Kaptein. Determination of protein structures from nuclear magnetic resonance data using a restrained molecular dynamics approach: The lac repressor dna binding domain. *Biochimie*, 67:707–715, 1985.

[301] A. M. Ferrenberg and R. H. Swendsen. Optimized Monte Carlo data analysis. *Phys. Rev. Lett.*, 63:1195–1198, 1989.

[302] B. Roux. The calculation of the potential of mean force using computer simulations. *Comput. Phys. Commun.*, 91:275–282, 1995.

[303] S. R. Billeter and W. F. van Gunsteren. Computer simulation of proton transfers of small acids in water. *J. Phys. Chem. A*, 104:3276–3286, 2000.

[304] J. Kästner and W. Thiel. Bridging the gap between thermodynamic integration and umbrella sampling provides a novel analysis method: "umbrella integration". *J. Chem. Phys.*, 123:144104, 2005.

[305] J. Kästner and W. Thiel. Analysis of the statistical error in umbrella sampling simulations by umbrella integration. *J. Chem. Phys.*, 124:234106, 2006.

[306] Python implementation of the umbrella integration method. https://github.com/martinstroet/umbrella_integration. Accessed: 2018-01-08.

[307] A. Baumketner. Removing systematic errors in interionic potentials of mean force computed in molecular simulations using reaction-field-based electrostatics. *J. Chem. Phys.*, 130:104106, 2009.

[308] M. K. Gilson and K. K. Irikura. Correction to "symmetry numbers for rigid, flexible, and fluxional molecules: Theory and applications". *J. Phys. Chem. B*, 117:3061–3061, 2013.

[309] M. M. Reif and P. H. Hünenberger. Computation of methodology-independent single-ion solvation properties from molecular simulations. iv. optimized lennard-jones interaction parameter sets for the alkali and halide ions in water. *J. Chem. Phys.*, 134:144104, 2011.

[310] J. Gebhardt and N. Hansen. Calculation of binding affinities for linear alcohols to α-cyclodextrin by twin-system enveloping distribution sampling simulations. *Fluid Phase Equilib.*, 422:1–17, 2016.

[311] Y. Sugita, A. Kitao, and Y. Okamoto. Multidimensional replica-exchange method for free-energy calculations. *J. Chem. Phys.*, 113:6042–6051, 2000.

[312] D. L. Mobley, J. D. Chodera, and K. A. Dill. On the use of orientational restraints and symmetry corrections in alchemical free energy calculations. *J. Chem. Phys.*, 125:084902, 2006.

[313] J. Tirado-Rives and W. L. Jorgensen. Contribution of conformer focusing to the uncertainty in predicting free energies for protein-ligand binding. *J. Med. Chem.*, 49:5880–5884, 2006.

[314] G. Singh, A. C. Chamberlin, H. R. Zhekova, S. Y. Noskov, and D. P. Tieleman. Two-dimensional potentials of mean force of nile red in intact and damaged model bilayers. application to calculations of fluorescence spectra. *J. Chem. Theory Comput.*, 12:364–371, 2015.

[315] J. M. Prausnitz, R. N. Lichtenthaler, and E. G. de Azevedo. *Molecular thermodynamics of fluid-phase equilibria.* Pearson Education, 1998.

[316] G. Raabe. *Molecular Simulation Studies on Thermophysical Properties.* Springer, 2017.

[317] H. J. Löffler. *Thermodynamik: Zweiter Band: Gemische und chemische Reaktionen.* Springer-Verlag, 2013.

[318] K. Burton and H.A. Krebs. The free-energy changes associated with the individual steps of the tricarboxylic acid cycle, glycolysis and alcoholic fermentation and with the hydrolysis of the pyrophosphate groups of adenosinetriphosphate. *Biochem. J.*, 54:94, 1953.

[319] J.-L. Burgot. *The notion of activity in chemistry.* Springer, 2017.

[320] T. L. Hill. *An introduction to statistical thermodynamics.* Courier Corporation, 1960.

[321] T. L. Hill. *Cooperativity theory in biochemistry: steady-state and equilibrium systems.* Springer Science and Business Media, 2013.

[322] T. Kloss, J. Heil, and S. M. Kast. Quantum chemistry in solution by combining 3d integral equation theory with a cluster embedding approach. *J. Phys. Chem. B*, 112:4337–4343, 2008.

[323] C. Hölzl, P. Kibies, S. Imoto, J. Noetzel, M. Knierbein, P. Salmen, M. Paulus, J. Nase, C. Held, G. Sadowski, D. Marx, S. M. Kast, and D. Horinek. Structure and thermodynamics of aqueous urea solutions from ambient to kilobar pressures: From thermodynamic modeling, experiments, and first principles simulations to an accurate force field description. *Biophys. Chem.*, 254:106260, 2019.

[324] M. Fleck. A molecular dynamics workflow for predicting the impact of amino acid side chain mutations on protein stability using the GROMOS force field. Master's thesis, University of Stuttgart, 2019.

[325] M. Neumann. Dipole moment fluctuation formulas in computer simulations of polar systems. *Mol. Phys.*, 50:841–858, 1983.

[326] W. Kabsch and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22:2577–2637, 1983.

[327] D. S. Dashti and A. E. Roitberg. Optimization of umbrella sampling replica exchange molecular dynamics by replica positioning. *J. Chem. Theory Comput.*, 9(11):4692–4699, 2013.

Appendices

Appendix A

Validation of Trimethylamine-N-Oxide (TMAO) Force Fields Based on Thermophysical Properties of Aqueous TMAO Solutions

A.1 Detailed Simulation Results

The raw simulation data used to calculate partial and apparent molar volumes for different TMAO models and their corresponding water models are presented in Tables A1 and A2. Table A3 reports self-diffusion coefficients of water in pure water as well as zero-shear rate viscosities of pure water for different water models and/or different simulation setups. Table A4 reports simulation times used while Table A5 and A6 present investigations of the system size dependence of various properties.

Table A1: Densities (in $\mathrm{kg\,m^{-3}}$) and simulation box volumes (in $\mathrm{nm^3}$) from constant pressure simulations at different temperatures and compositions.

system	T	Kast 2016				Kast 2016				Garcia				Netz				Shea			
	[K]	TIP4P/2005				SPC/E				TIP3P				SPC/E				SPC/E			
		ρ	$\delta\rho\,(\cdot 10^3)$	V_box	$\delta V_\mathrm{box}\,(\cdot 10^3)$	ρ	$\delta\rho\,(\cdot 10^3)$	V_box	$\delta V_\mathrm{box}\,(\cdot 10^3)$	ρ	$\delta\rho\,(\cdot 10^3)$	V_box	$\delta V_\mathrm{box}\,(\cdot 10^3)$	ρ	$\delta\rho\,(\cdot 10^3)$	V_box	$\delta V_\mathrm{box}\,(\cdot 10^3)$	ρ	$\delta\rho\,(\cdot 10^3)$	V_box	$\delta V_\mathrm{box}\,(\cdot 10^3)$
C0	278.15	1000.27	21.0	29.910	0.6	1007.26	18.0	29.702	0.5	1002.44	9.0	29.846	0.3	1001.92	5.5	29.860	0.2	1007.39	19.0	29.697	0.6
	298.15	997.07	5.8	30.006	0.2	998.73	19.0	29.955	0.6	985.98	4.9	30.344	0.2	993.44	18.0	30.115	0.6	998.87	14.0	29.951	0.4
	323.15	987.50	10.0	30.297	0.3	984.26	13.0	30.396	0.4	962.67	11.0	31.079	0.3	978.81	8.0	30.565	0.3	984.40	11.0	30.392	0.4
C1	278.15	1000.32	42.0	30.033	1.3	1007.47	21.0	29.819	0.6	1003.50	10.0	29.938	0.3	1002.54	19.0	29.966	0.6	1007.65	7.9	29.814	0.2
	298.15	997.07	19.0	30.131	0.6	998.92	9.1	30.075	0.3	987.03	10.0	30.438	0.3	993.94	18.0	30.225	0.6	999.11	13.0	30.069	0.4
	323.15	987.52	19.0	30.423	0.6	984.38	13.0	30.519	0.4	963.78	15.0	31.173	0.5	979.30	13.0	30.677	0.4	984.61	8.6	30.512	0.3
C2	278.15	1000.50	20.0	137.290	2.7	1007.98	30.0	136.267	4.1	1006.97	22.0	136.409	2.9	1004.35	30.0	136.759	4.0	1008.45	23.0	136.204	3.1
	298.15	997.14	42.0	137.753	5.7	999.31	17.0	137.449	2.3	990.55	7.4	138.670	1.0	995.63	5.5	137.957	0.8	999.75	17.0	137.389	2.4
	323.15	987.44	13.0	139.106	1.8	984.71	18.0	139.487	2.5	967.29	22.0	142.006	3.3	980.94	8.9	140.023	1.3	985.11	12.0	139.431	1.7
C3	278.15	1001.30	52.0	59.111	3.1	1009.12	43.0	58.651	2.5	1013.24	46.0	58.415	2.6	1007.65	43.0	58.736	2.5	1010.09	40.0	58.595	2.3
	298.15	997.44	33.0	59.340	2.0	1000.26	27.0	59.170	1.6	996.83	330.0	59.376	2.0	998.79	35.0	59.258	2.0	1001.14	39.0	59.119	2.3
	323.15	987.65	26.0	59.928	1.6	985.61	29.0	60.050	1.8	973.66	25.0	60.790	1.5	984.01	37.0	60.148	2.3	986.18	32.0	60.015	1.9
C4	278.15	1004.96	41.0	34.114	1.4	1013.85	39.0	33.814	1.3	1035.75	20.0	33.100	0.6	1019.84	48.0	33.615	1.6	1016.62	22.0	33.722	7.5
	298.15	999.65	36.0	34.295	1.2	1004.22	45.0	34.138	1.5	1019.38	24.0	33.632	0.8	1010.23	24.0	33.935	0.8	1006.49	29.0	34.061	1.0
	323.15	988.96	44.0	34.666	1.5	989.14	43.0	34.659	1.5	996.48	24.0	34.405	0.8	995.18	70.0	34.449	2.4	990.68	9.4	34.605	0.3
C5	278.15	1010.73	57.0	56.740	3.2	1020.01	31.0	56.222	1.7	1060.02	4.2	54.102	0.2	1033.51	57.0	55.488	3.1	1024.61	82.0	55.970	4.5
	298.15	1003.56	21.0	57.145	1.2	1009.48	27.0	56.808	1.5	1043.72	23.0	54.947	1.2	1023.13	35.0	56.051	1.9	1013.10	22.0	56.605	1.2
	323.15	991.69	14.0	57.829	0.8	993.87	29.0	57.701	1.7	1021.20	16.0	56.159	0.9	1007.68	13.0	56.901	0.7	996.16	33.0	57.568	1.9
C6	278.15	1012.95	59.0	39.139	2.3	1022.22	67.0	38.783	2.5	1067.77	27.0	37.130	0.9	1037.96	14.0	38.195	0.5	1027.29	57.0	38.592	2.2
	298.15	1005.11	43.0	39.445	1.7	1011.27	41.0	39.204	1.6	1051.59	30.0	37.702	1.1	1027.46	76.0	38.586	2.9	1015.36	38.0	39.046	1.5
	323.15	992.89	32.0	39.931	1.3	995.53	37.0	39.824	1.5	1029.07	27.0	38.527	1.0	1011.75	43.0	39.185	1.7	998.08	43.0	30.722	1.7
C7	278.15	1016.80	20.0	41.690	0.8	1025.94	58.0	41.317	2.3	1081.45	35.0	39.198	1.3	1045.93	76.0	40.527	2.9	1032.07	78.0	41.072	3.1
	298.15	1008.26	49.0	42.043	2.0	1014.67	31.0	41.776	1.3	1065.24	24.0	39.795	0.9	1034.84	50.0	40.962	2.0	1019.46	81.0	41.580	3.3
	323.15	995.18	32.0	42.596	1.4	998.60	31.0	42.449	1.3	1043.02	29.0	40.643	1.1	1019.07	47.0	41.596	1.9	1001.51	31.0	42.326	1.3
C8	278.15	1031.85	76.0	37.218	2.8	1039.49	120.0	36.945	4.4	1128.59	31.0	34.029	0.9	1072.23	270.0	35.816	9.1	1048.28	100.0	36.635	3.6
	298.15	1020.24	160.0	37.642	5.9	1026.74	49.0	37.404	1.8	1112.44	33.0	34.523	10.0	1060.08	120.0	36.227	4.2	1033.44	89.0	37.161	3.2
	323.15	1004.92	91.0	38.216	3.5	1009.46	97.0	38.044	3.7	1090.33	52.0	35.223	1.7	1043.47	53.0	36.804	1.9	1013.55	63.0	37.891	2.3

Table A2: Densities (in $\mathrm{kg\,m^{-3}}$) and simulation box volumes (in $\mathrm{nm^3}$) from constant pressure simulations at different temperatures and compositions.

system	T [K]	UA															
		setup 1 (RF)								setup 2 (PME)							
		SPC				SPC/E				SPC				SPC/E			
		ρ	$\delta\rho\,(\cdot 10^3)$	V_{box}	$\delta V_{\mathrm{box}}\,(\cdot 10^3)$	ρ	$\delta\rho\,(\cdot 10^3)$	V_{box}	$\delta V_{\mathrm{box}}\,(\cdot 10^3)$	ρ	$\delta\rho\,(\cdot 10^3)$	V_{box}	$\delta V_{\mathrm{box}}\,(\cdot 10^3)$	ρ	$\delta\rho\,(\cdot 10^3)$	V_{box}	$\delta V_{\mathrm{box}}\,(\cdot 10^3)$
C0	278.15	984.97	13.9	30.373	0.4	1002.45	19.7	29.844	0.6	990.68	11.0	30.198	0.3	1007.27	7.8	29.701	0.2
	298.15	971.33	14.6	30.800	0.5	993.45	19.1	30.114	0.6	977.54	12.0	30.604	0.4	998.77	17.0	29.953	0.5
	323.15	951.00	13.2	31.458	0.4	978.23	14.5	30.582	0.5	957.97	8.2	31.230	0.3	984.24	8.9	30.396	0.3
C1	278.15	984.87	15.7	30.503	0.5	1002.37	17.7	29.970	0.5	990.59	9.7	30.327	0.3	1007.21	5.6	29.826	0.2
	298.15	971.18	15.6	30.933	0.5	993.30	17.0	30.244	0.5	977.47	22.0	30.734	0.7	998.65	7.8	30.082	0.2
	323.15	950.88	12.0	31.593	0.4	978.05	12.4	30.715	0.4	957.86	9.0	31.364	0.3	984.12	15.0	30.526	0.5
C2	278.15	984.59	18.8	139.505	2.7	1002.14	34.3	137.062	4.7	990.39	21.0	138.685	2.9	1007.11	19.0	136.382	2.6
	298.15	970.89	17.9	141.473	2.6	992.81	19.4	138.350	2.7	977.15	18.0	140.564	2.6	998.31	12.0	137.584	1.6
	323.15	950.51	17.3	144.507	2.6	977.45	19.5	140.523	2.8	957.49	14.0	143.450	2.0	983.61	9.8	139.641	1.4
C3	278.15	984.38	27.1	60.125	1.7	1002.10	28.6	59.062	1.7	990.39	24.0	59.759	1.5	1007.21	32.0	58.761	1.8
	298.15	970.65	33.5	60.975	2.1	992.45	33.5	59.636	2.0	977.07	14.0	60.574	0.9	998.06	37.0	59.300	2.2
	323.15	950.07	23.9	62.296	1.6	976.65	24.8	60.600	1.5	957.27	29.0	61.827	1.9	983.03	20.0	60.207	1.2
C4	278.15	985.62	47.0	34.782	1.7	1003.60	58.5	34.159	2.0	992.00	51.0	34.558	1.8	1008.98	73.0	33.976	2.4
	298.15	971.19	47.2	35.299	1.7	992.42	62.8	34.544	2.2	978.10	71.0	35.050	2.5	998.63	70.0	34.329	2.4
	323.15	950.26	36.4	36.076	1.4	975.83	38.0	35.131	1.4	957.93	24.0	35.788	0.9	982.44	47.0	34.895	1.7
C5	278.15	988.80	37.1	57.996	2.2	1006.62	44.3	56.970	2.5	995.63	20.0	57.598	1.2	1012.64	34.0	56.630	1.9
	298.15	973.81	35.5	58.889	2.1	994.41	42.8	57.669	2.5	981.12	31.0	58.450	1.9	1001.03	57.0	57.287	3.3
	323.15	952.46	36.9	60.209	2.3	976.52	40.8	58.726	2.5	960.67	33.0	59.695	2.0	983.87	24.0	58.287	1.4
C6	278.15	990.08	57.0	40.042	2.3	1007.81	56.2	39.338	2.2	997.04	27.0	39.762	1.1	1014.07	55.0	39.095	2.1
	298.15	974.88	52.4	40.667	2.2	995.15	37.7	39.838	1.5	982.37	19.0	40.357	0.8	1001.88	95.0	39.570	3.7
	323.15	953.47	43.6	41.580	1.9	977.00	60.8	40.578	2.5	961.80	25.0	41.220	1.1	984.55	47.0	40.267	1.9
C7	278.15	992.74	41.9	42.699	1.8	1010.25	54.6	41.959	2.3	999.87	44.0	42.394	1.8	1016.67	54.0	41.693	2.2
	298.15	977.13	52.3	43.381	2.3	997.02	52.4	42.516	2.2	985.03	44.0	43.033	1.9	1003.97	41.0	42.221	1.7
	323.15	955.45	50.9	44.366	2.4	978.13	46.3	43.337	2.1	964.25	34.0	43.961	1.6	986.02	61.0	42.990	2.7
C8	278.15	1003.94	65.5	38.253	2.5	1019.64	78.6	37.664	2.9	1011.97	66.0	37.949	2.5	1027.28	48.0	37.383	1.7
	298.15	987.45	55.0	38.892	2.2	1004.56	65.1	38.229	2.5	996.21	110.0	38.549	4.2	1012.62	57.0	37.925	2.1
	323.15	964.95	42.4	39.798	1.8	984.34	50.7	39.014	2.0	974.80	83.0	39.397	3.4	993.26	70.0	38.664	2.7

Table A3: Pure water self-diffusion coefficient (in $10^{-9}\,\mathrm{m}^2\,\mathrm{s}^{-1}$) and zero-shear rate viscosity (in cp) from constant pressure and constant volume simulations, respectively, at different temperatures. These values were used for the reduced representation of the transport properties of aqueous TMAO solutions (C1 to C8) as described the main text.

system	T	Kast 2016				Kast 2016				Garcia				Netz				Shea			
	[K]	TIP4P/2005				SPC/E				TIP3P				SPC/E				SPC/E			
		$D_{\mathrm{self,w,0}}$	$\delta D_{\mathrm{self,w,0}}$	η_0	$\delta\eta_0$	$D_{\mathrm{self,w,0}}$	$\delta D_{\mathrm{self,w,0}}$	η_0	$\delta\eta_0$	$D_{\mathrm{self,w,0}}$	$\delta D_{\mathrm{self,w,0}}$	η_0	$\delta\eta_0$	$D_{\mathrm{self,w,0}}$	$\delta D_{\mathrm{self,w,0}}$	η_0	$\delta\eta_0$	$D_{\mathrm{self,w,0}}$	$\delta D_{\mathrm{self,w,0}}$	η_0	$\delta\eta_0$
C0	278.15	1.217	0.034	-	-	1.581	0.040	-	-	4.031	0.125	-	-	1.553	0.042	-	-	1.566	0.046	-	-
	298.15	2.099	0.071	0.845	0.026	2.544	0.060	0.722	0.027	5.493	0.115	0.320	0.011	2.521	0.062	0.715	0.026	2.519	0.083	0.724	0.027
	323.15	3.500	0.107	-	-	3.987	0.105	-	-	7.611	0.241	-	-	4.010	0.117	-	-	4.007	0.109	-	-

system	T	UA															
	[K]	setup 1 (RF)								setup 2 (PME)							
		SPC				SPC/E				SPC				SPC/E			
		$D_{\mathrm{self,w,0}}$	$\delta D_{\mathrm{self,w,0}}$	η_0	$\delta\eta_0$	$D_{\mathrm{self,w,0}}$	$\delta D_{\mathrm{self,w,0}}$	η_0	$\delta\eta_0$	$D_{\mathrm{self,w,0}}$	$\delta D_{\mathrm{self,w,0}}$	η_0	$\delta\eta_0$	$D_{\mathrm{self,w,0}}$	$\delta D_{\mathrm{self,w,0}}$	η_0	$\delta\eta_0$
C0	278.15	2.982	0.091	-	-	1.672	0.040	-	-	2.822	0.081	-	-	1.575	0.038	-	-
	298.15	4.303	0.115	-	-	2.671	0.088	-	-	4.098	0.113	-	-	2.524	0.069	0.710	0.025
	323.15	6.260	0.151	-	-	4.227	0.117	-	-	5.991	0.175	-	-	4.001	0.116	-	-

Table A4: Simulation times (in ns) used for the calculation of various properties.

Property	C0-C1	C2-C8
Density ρ	400	50
Apparent molar volume ${}^{\phi}\bar{V}_{\mathrm{s}}$	-	50
Partial molar volume $\bar{V}_{\mathrm{s}}^{\infty}$	400	-
Water self-diffusion coefficient $D_{\mathrm{self,w}}$	10 (× 40)	10 (× 5)
Zero-shear rate viscosity η	5 (× 100)	5 (× 100)
Number of hydrogen bonds N_{HB}	400	50
Radial distribution functions $g_{ij}(r)$	400	50

Table A5: System size dependence for the partial molar volume at infinite dilution (in $\mathrm{cm}^3\,\mathrm{mol}^{-1}$) in case of the Kast 2016 (TIP4P/2005) model at 298 K and 1 bar. Values for the box length and volume in nm and nm^3 respectively. The simulation time was 400 ns.

		C0				C1					
N_{W}	N_{T}	L_{box}	δL_{box}	V_{box}	δV_{box}	L_{box}	δL_{box}	V_{box}	δV_{box}	$\bar{V}_{\mathrm{s}}^{\infty}$	$\delta\bar{V}_{\mathrm{s}}^{\infty}$
1000	1	3.107	6.1e-6	30.006	1.8e-4	3.112	2.0e-5	30.131	5.8e-4	75.216	0.366
2000	1	3.915	1.6e-5	60.015	7.6e-4	3.918	2.1e-5	60.139	9.6e-4	74.434	0.737
4000	1	4.933	9.1e-6	120.033	6.7e-4	4.935	2.5e-5	120.159	1.8e-3	75.879	1.160

Table A6: System size dependence for the density (in $\mathrm{kg\,m^{-3}}$), apparent molar volume (in $\mathrm{cm^3\,mol^{-1}}$), water self-diffusion coefficient (in $10^{-9}\,\mathrm{m^2\,s^{-1}}$), zero-shear rate viscosity (in cp) and number of hydrogen bonds per TMAO in case of the Kast 2016 (TIP4P/2005) model at 298 K, 1 bar and TMAO molalities C0 and C7. Values for the box length in nm. Used system compositions ($N_\mathrm{W}/N_\mathrm{T}$) were (287/0), (1405/0) and (11433/0) for molality C0 and (205/20), (1000/100) and (8160/816) for C7. The simulation time was 50 ns.

	C0						C7											
L_box	ρ_0	$\delta\rho_0$	$D_\mathrm{self,w,0}$	$\delta D_\mathrm{self,w,0}$	η_0	$\delta\eta_0$	ρ	$\delta\rho$	${}^{\phi}\bar{V}_\mathrm{s}$	$\delta{}^{\phi}\bar{V}_\mathrm{s}$	$D_\mathrm{self,w}$	$\delta D_\mathrm{self,w}$	$\frac{D_\mathrm{self,w}}{D_\mathrm{self,w,0}}$	$\delta\left(\frac{D_\mathrm{self,w}}{D_\mathrm{self,w,0}}\right)$	η	$\delta\eta$	$\frac{\eta}{\eta_0}$	$\delta\left(\frac{\eta}{\eta_0}\right)$	N_HB	δN_HB
2.05	997.509	5.40e-2	2.0597	1.12e-1	0.8454	2.98e-2	1008.27	1.40e-1	72.57	3.65e-2	0.5839	2.15e-2	0.2835	1.86e-2	3.9294	3.60e-1	4.6481	4.56e-1	2.9544	6.79e-4
3.48	996.071	2.70e-2	2.1438	1.52e-2	0.8669	4.40e-2	1008.13	5.30e-2	72.34	1.42e-2	0.6055	1.19e-2	0.2824	5.90e-3	4.1281	5.33e-1	4.7621	6.61e-1	2.9467	3.08e-4
7.00	996.951	1.50e-2	2.2286	1.15e-2	0.8557	3.13e-2	1008.15	1.80e-2	72.50	5.28e-3	0.6234	4.20e-3	0.2797	2.37e-3	4.1202	5.17e-1	4.8151	6.29e-1	2.9457	1.07e-4

A.2 Radial Distribution Functions

Site-site radial distribution functions are presented for three different temperatures of 278.15 K, 298.15 K and 323.15 K and concentrations C1 to C8. The different sites are denoted as N (TMAO nitrogen), O (TMAO oxygen), C (TMAO carbon) and OW (water oxygen).

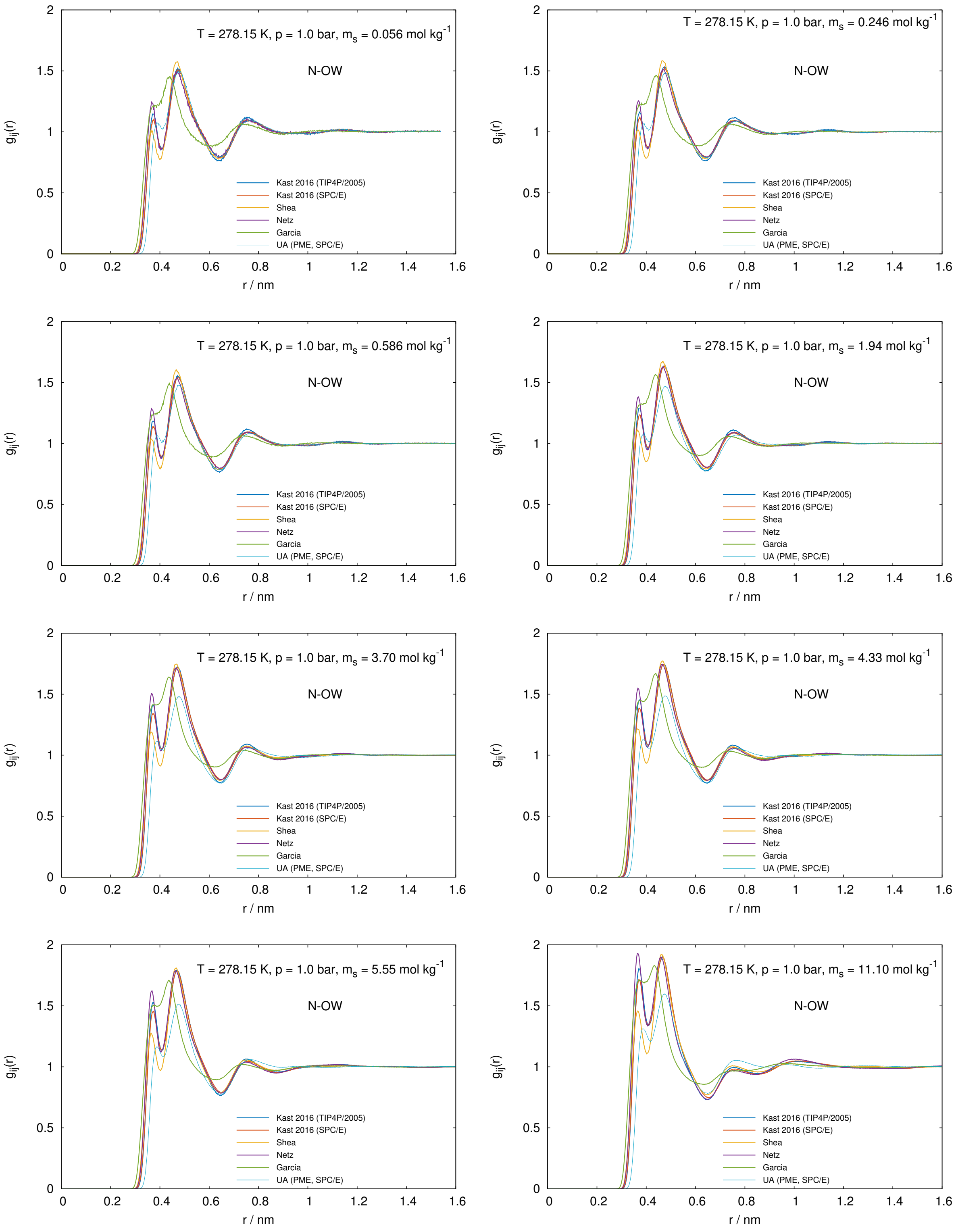


Figure A2: Radial distribution functions for N-OW at 278.15 K and 1.0 bar for different molalities.

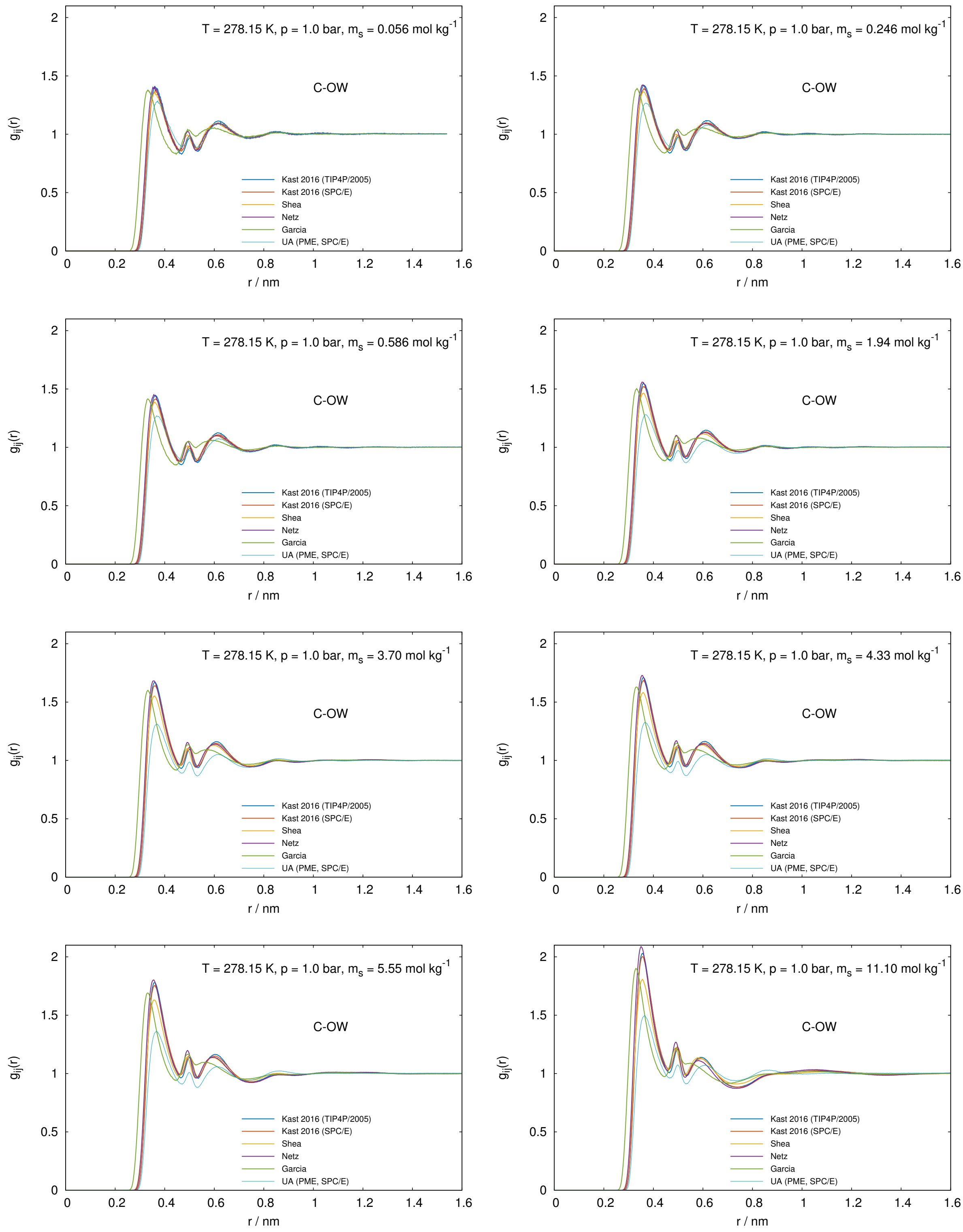


Figure A3: Radial distribution functions for C-OW at 278.15 K and 1.0 bar for different molalities.

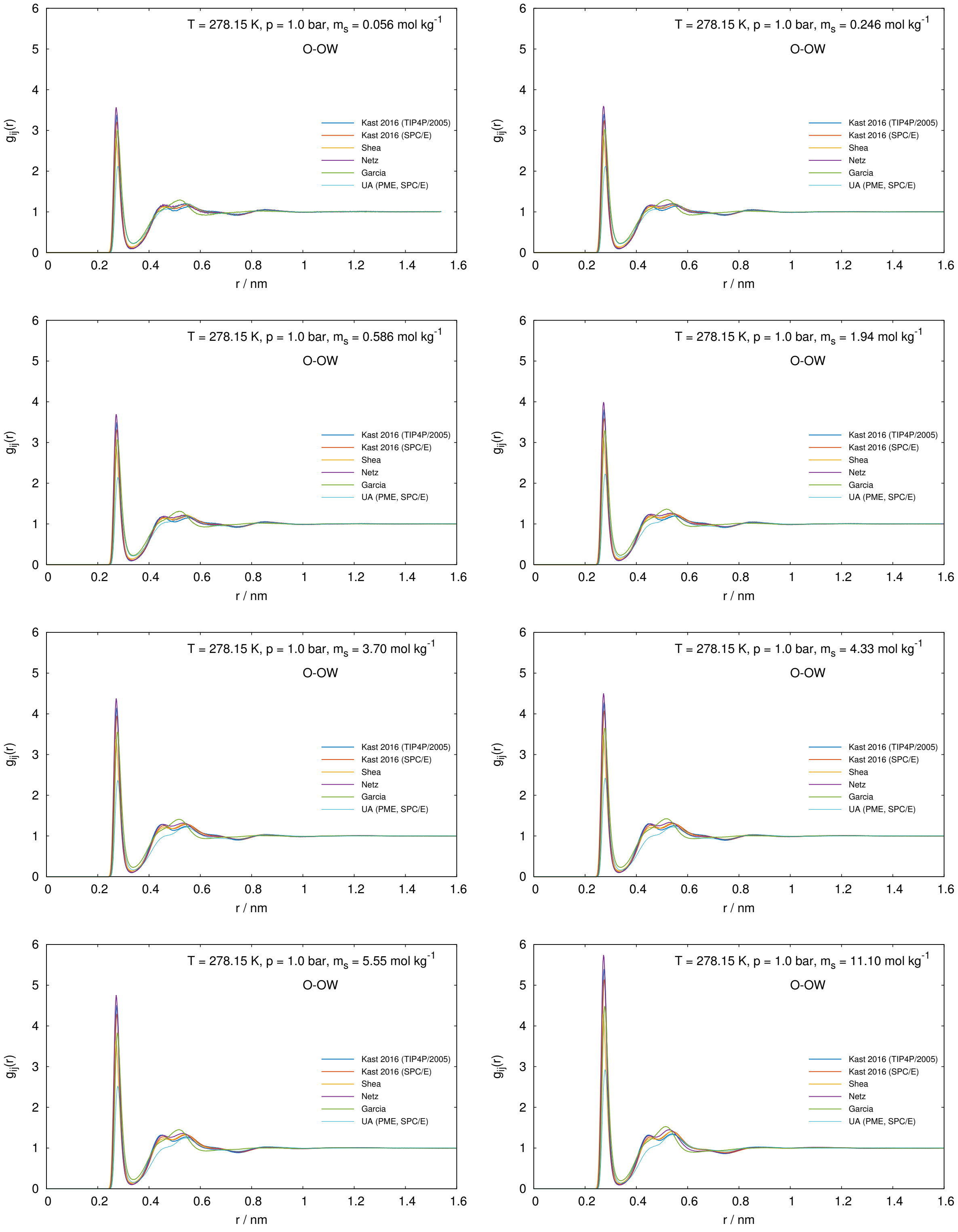


Figure A4: Radial distribution functions for O-OW at 278.15 K and 1.0 bar for different molalities.

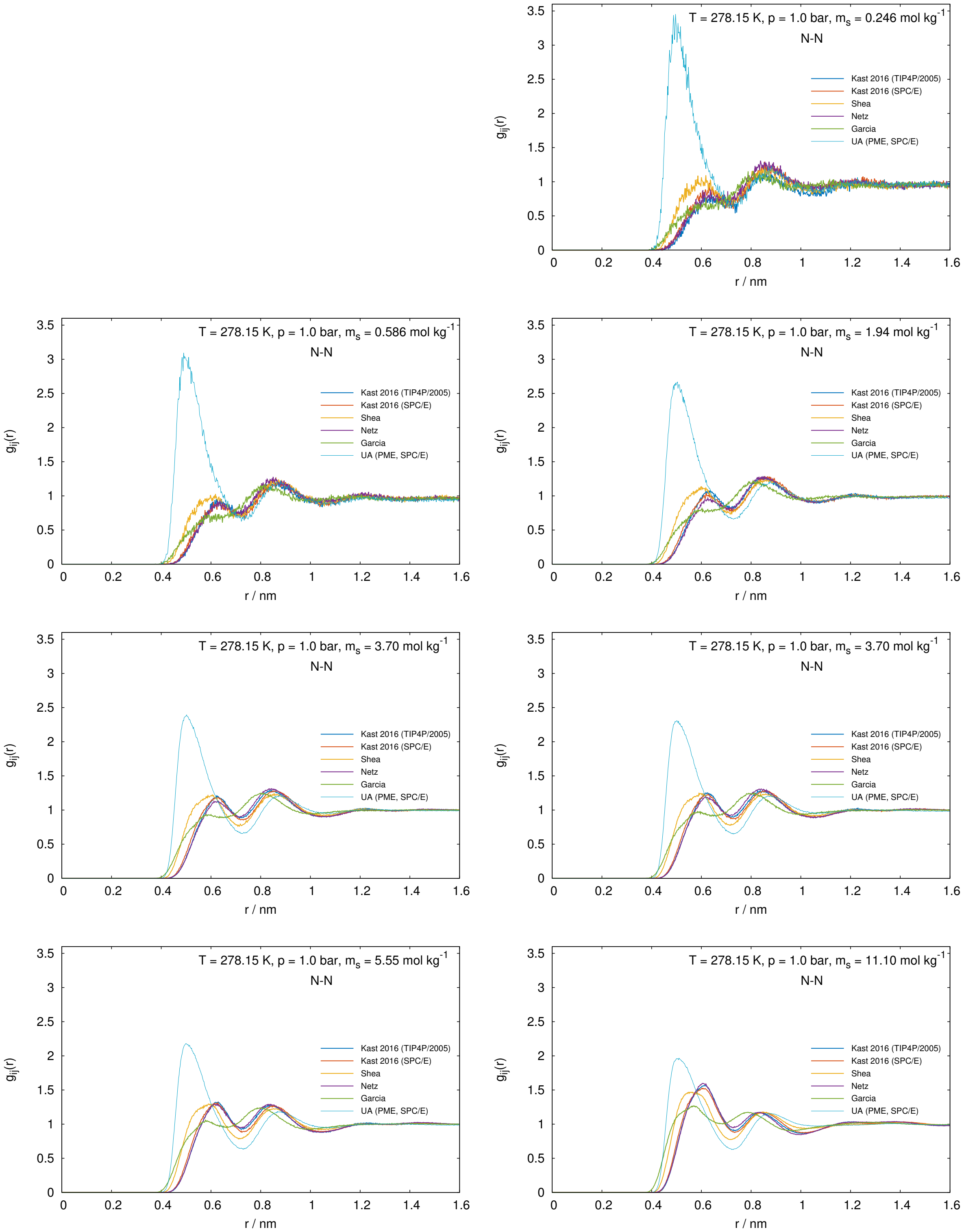


Figure A5: Radial distribution functions for N-N at 278.15 K and 1.0 bar for different molalities.

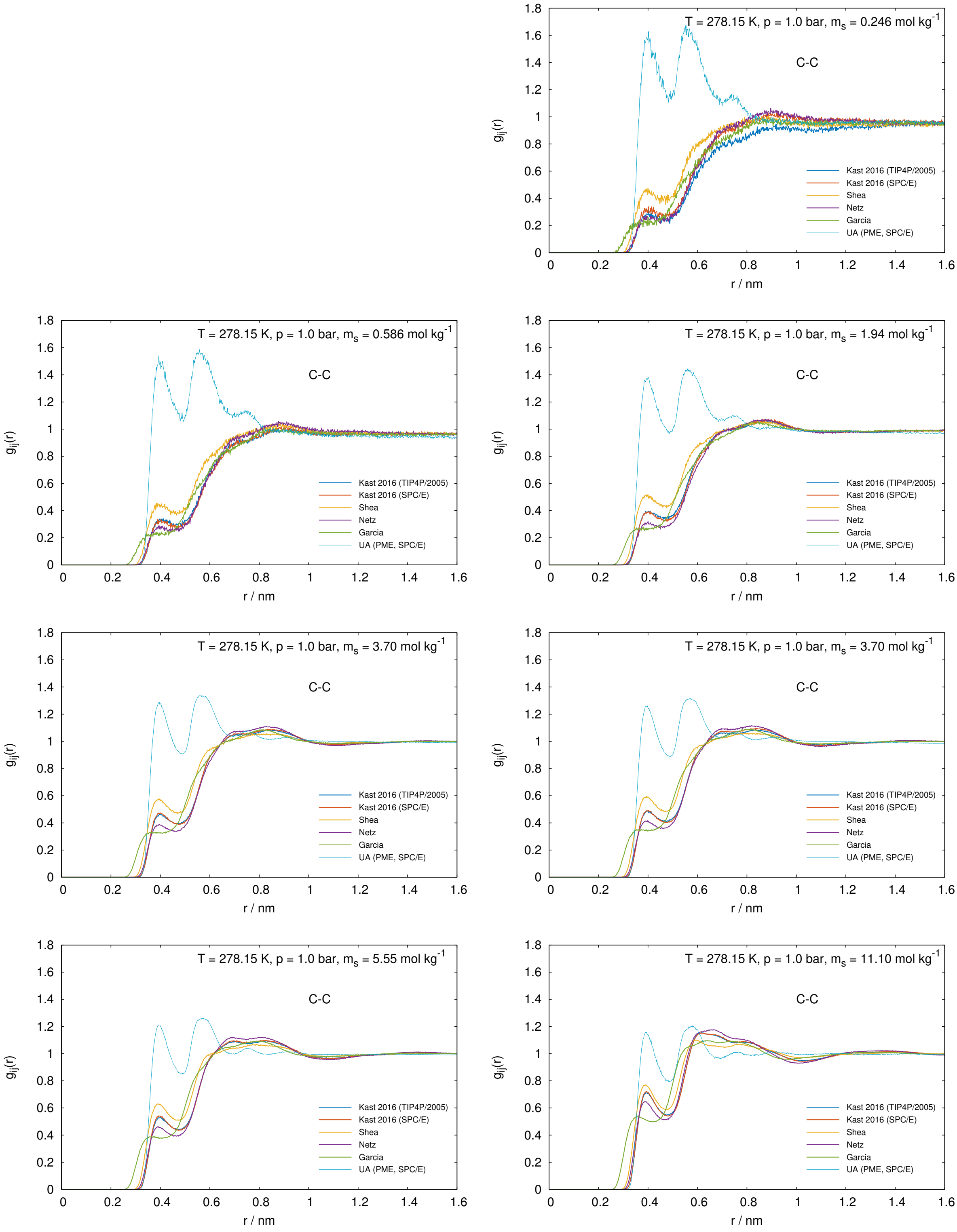


Figure A6: Radial distribution functions for C-C at 278.15 K and 1.0 bar for different molalities.

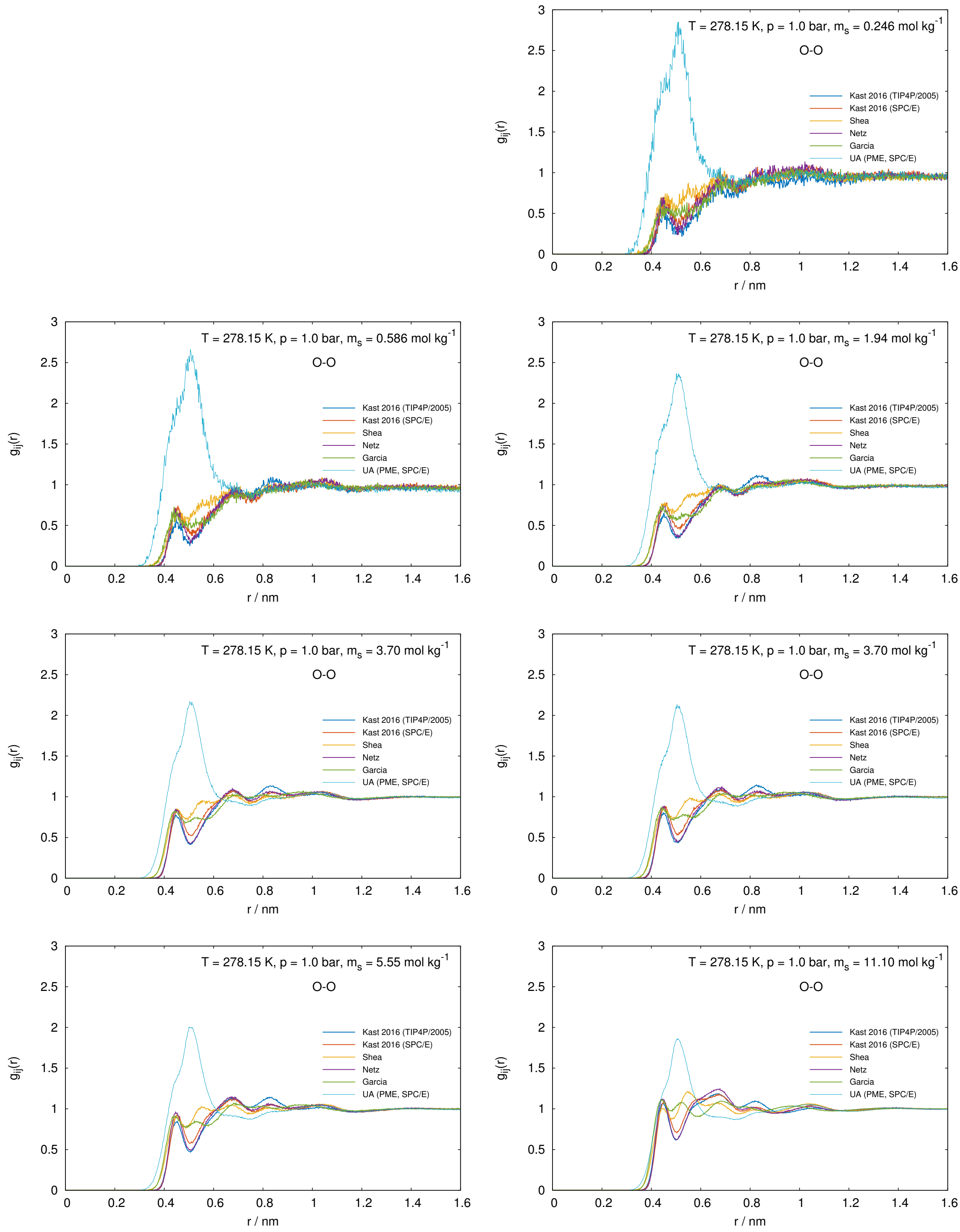


Figure A7: Radial distribution functions for O-O at 278.15 K and 1.0 bar for different molalities.

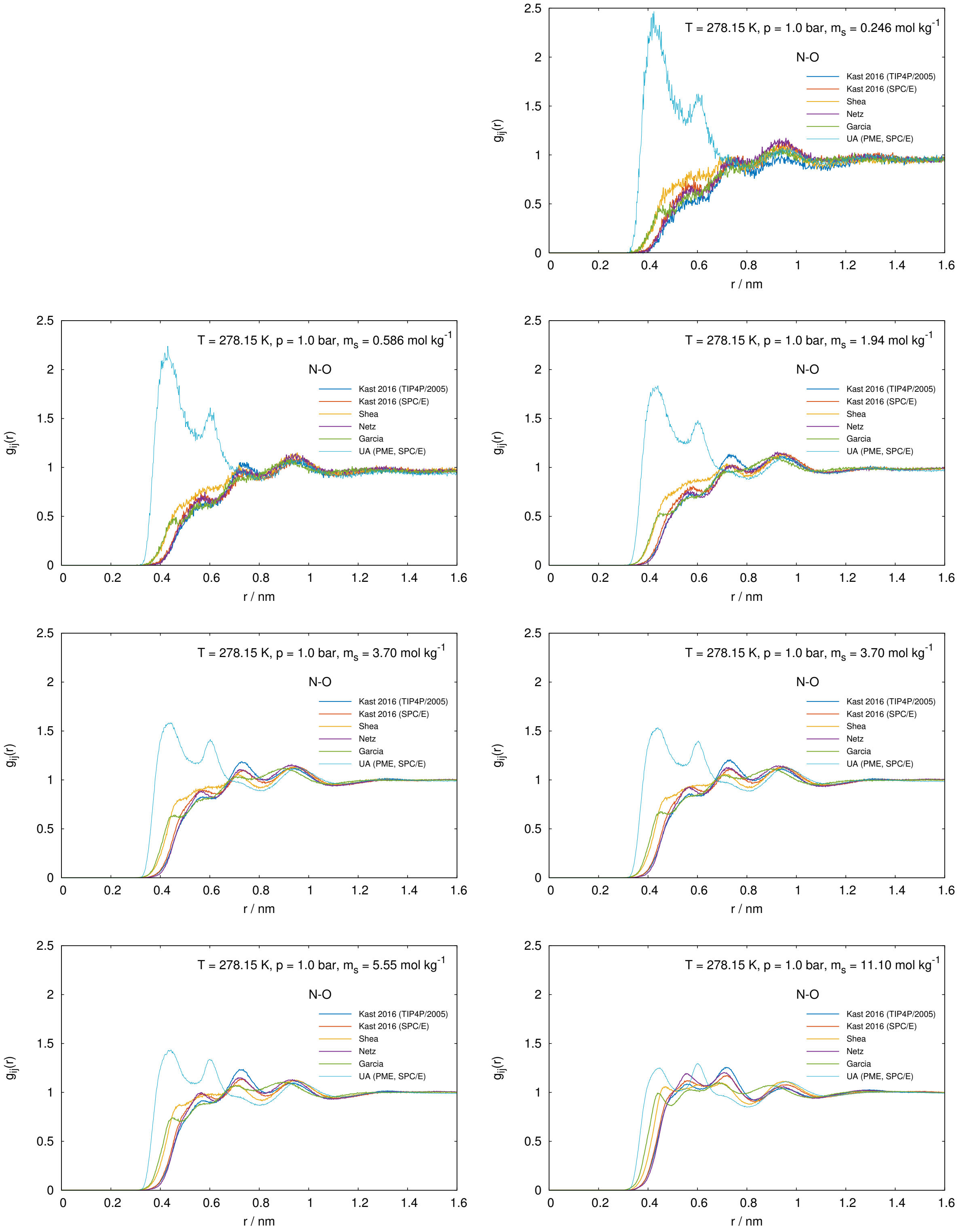


Figure A8: Radial distribution functions for N-O at 278.15 K and 1.0 bar for different molalities.

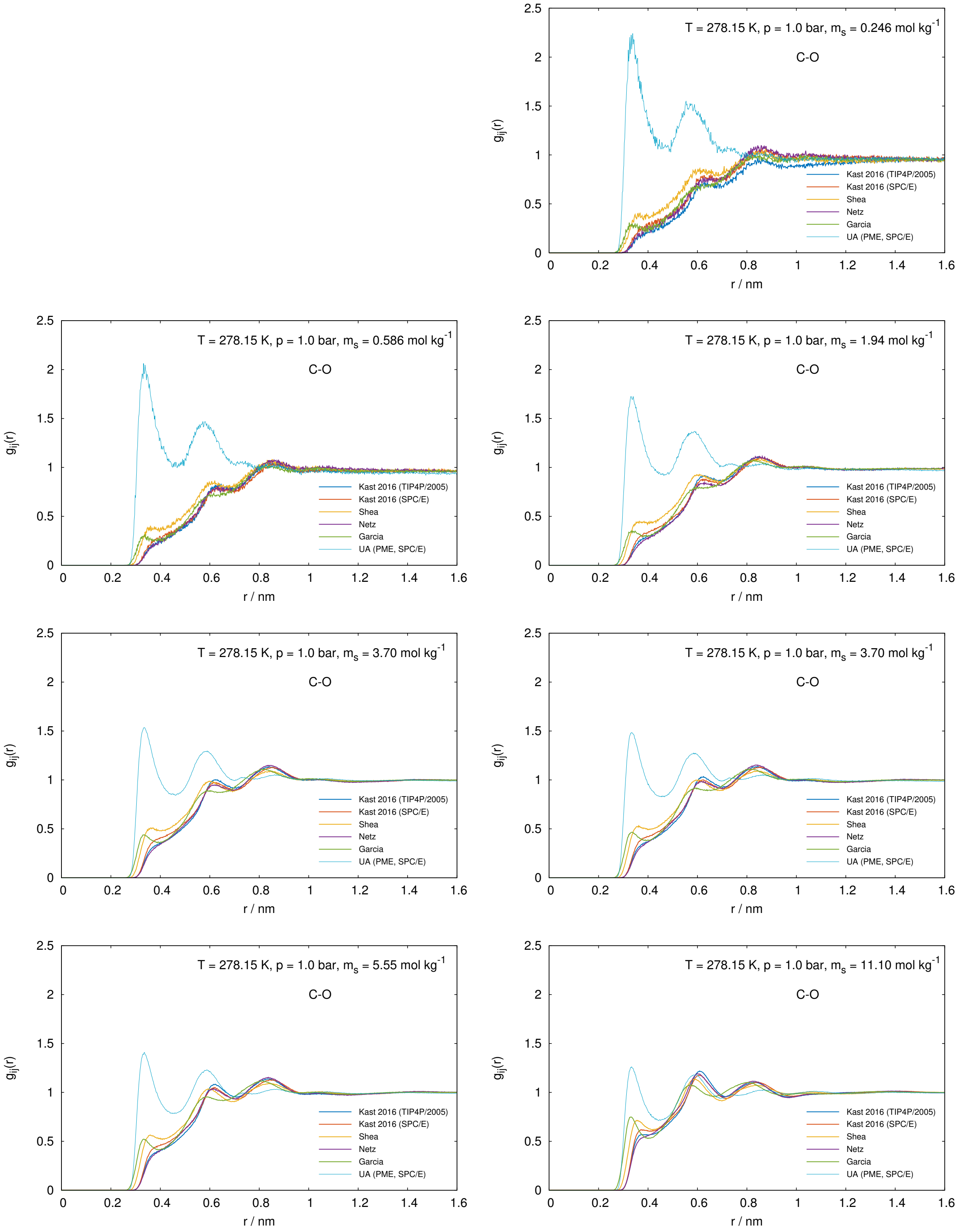


Figure A9: Radial distribution functions for C-O at 278.15 K and 1.0 bar for different molalities.

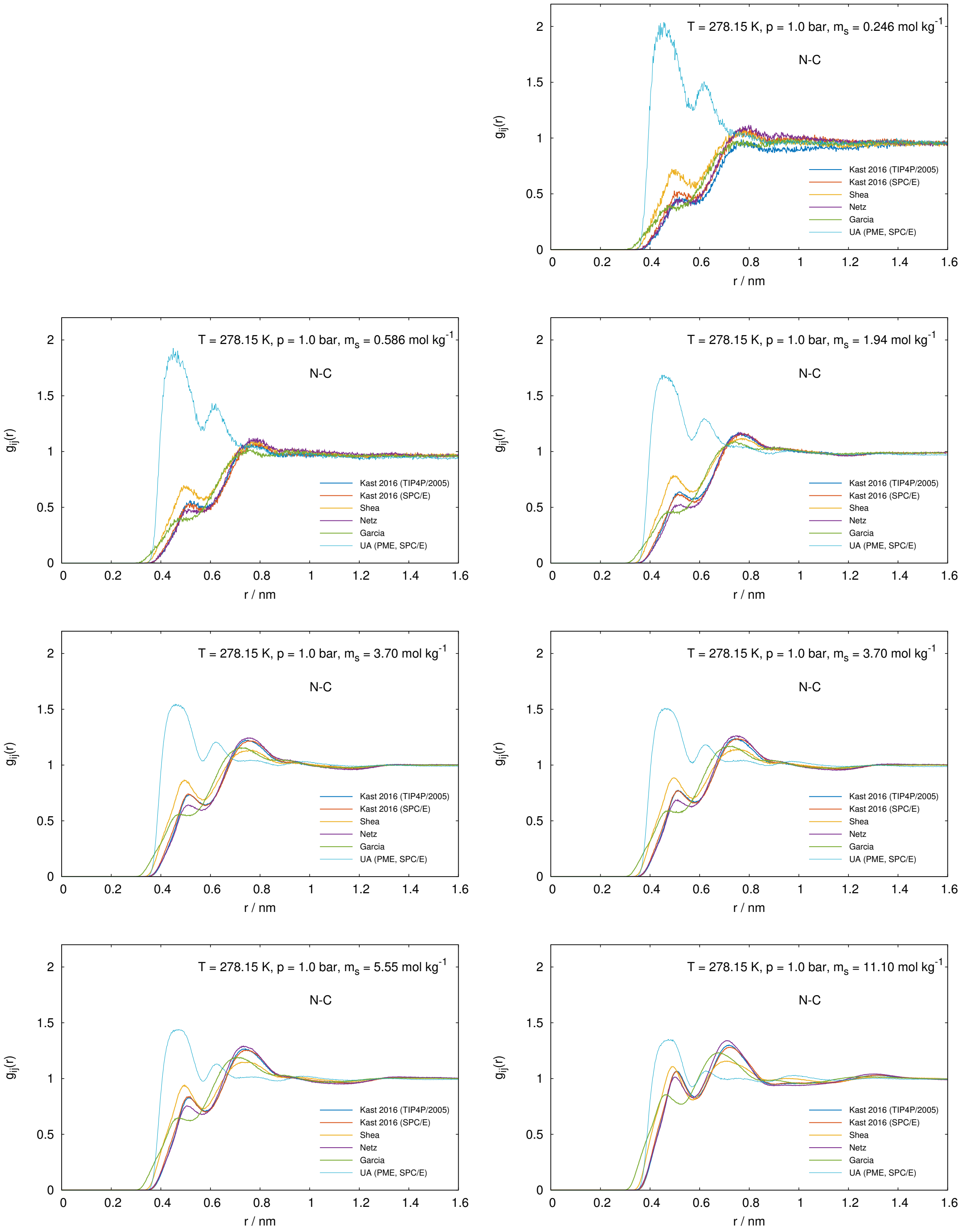


Figure A10: Radial distribution functions for N-C at 278.15 K and 1.0 bar for different molalities.

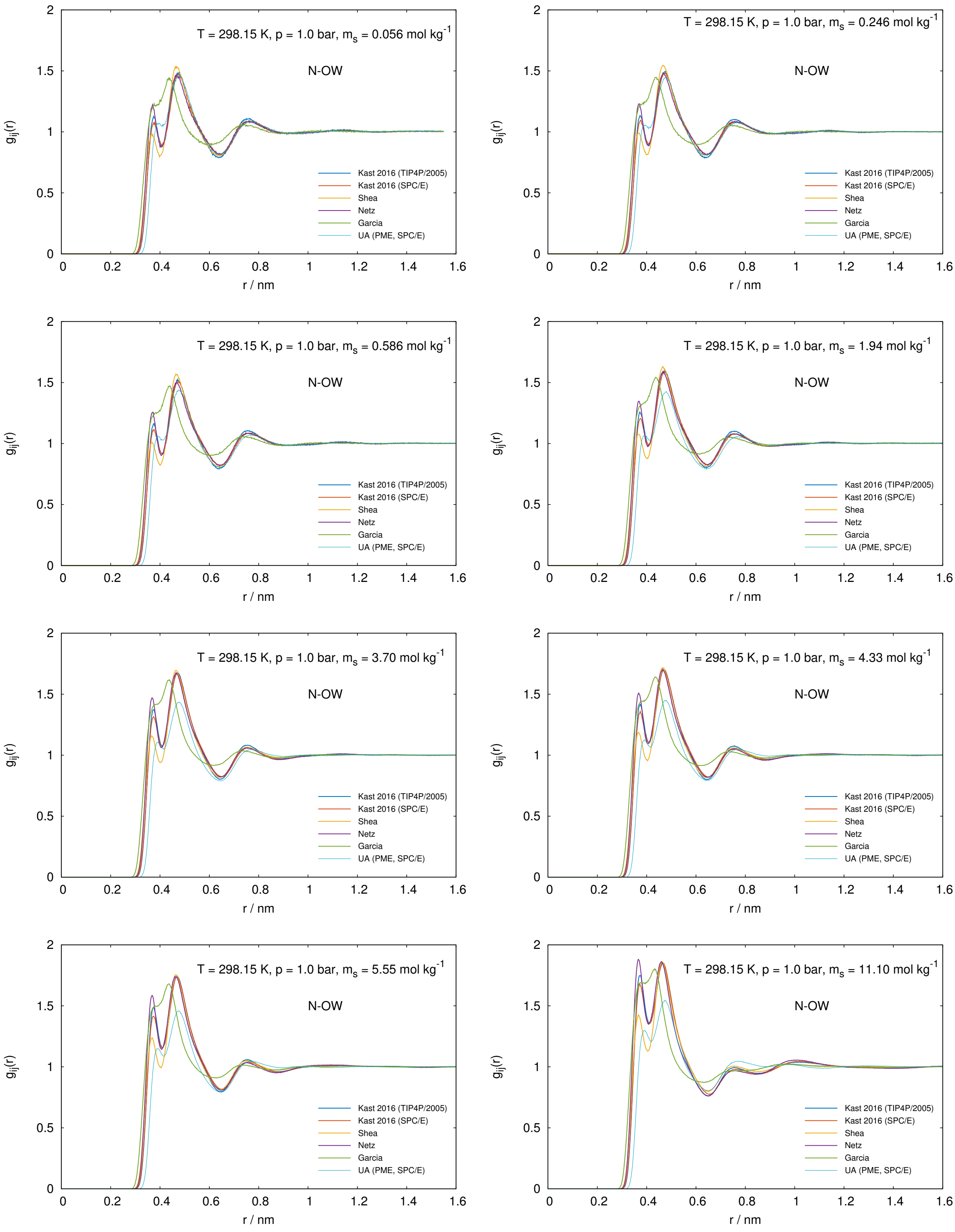


Figure A11: Radial distribution functions for N-OW at 298.15 K and 1.0 bar for different molalities.

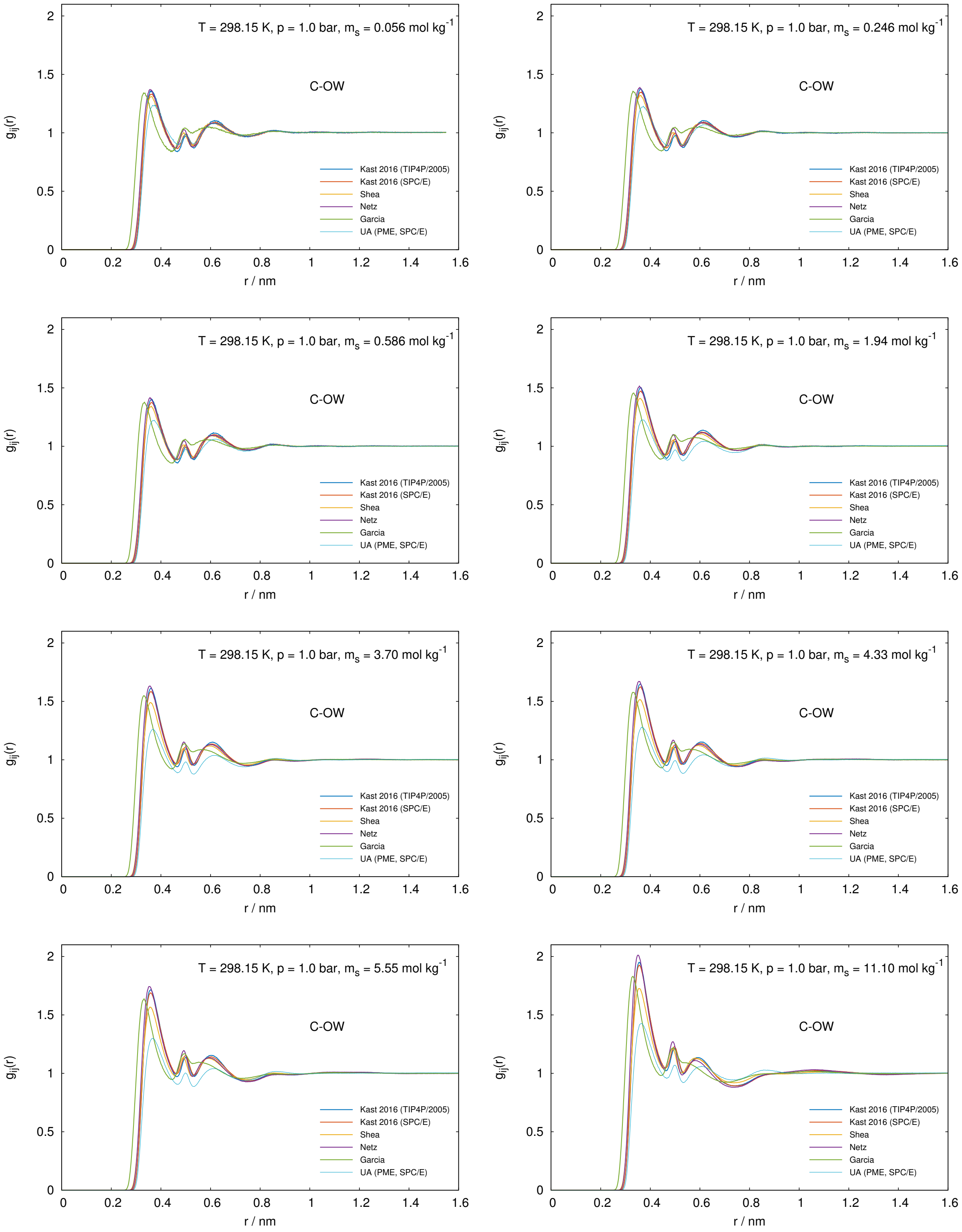


Figure A12: Radial distribution functions for C-OW at 298.15 K and 1.0 bar for different molalities.

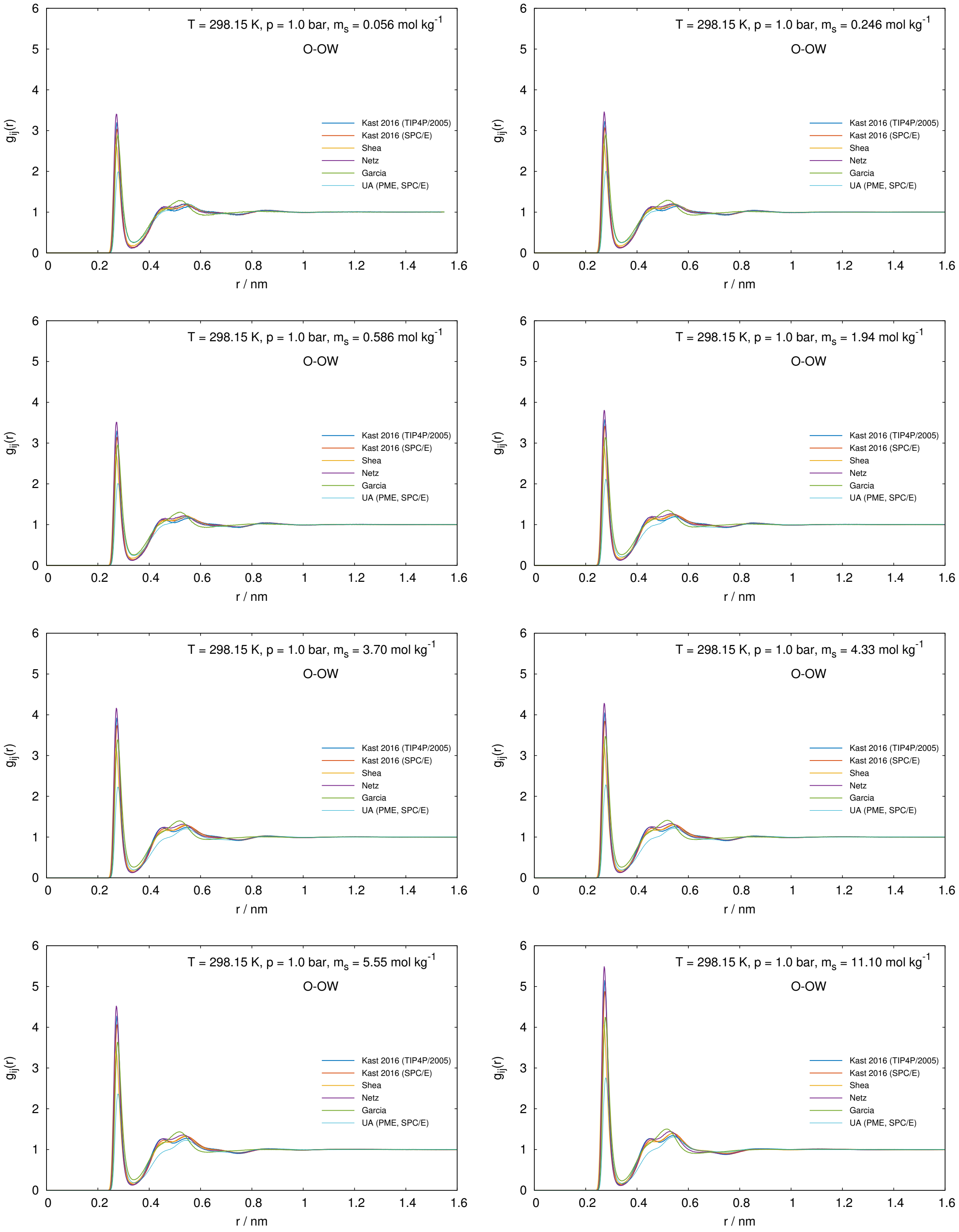


Figure A13: Radial distribution functions for O-OW at 298.15 K and 1.0 bar for different molalities.

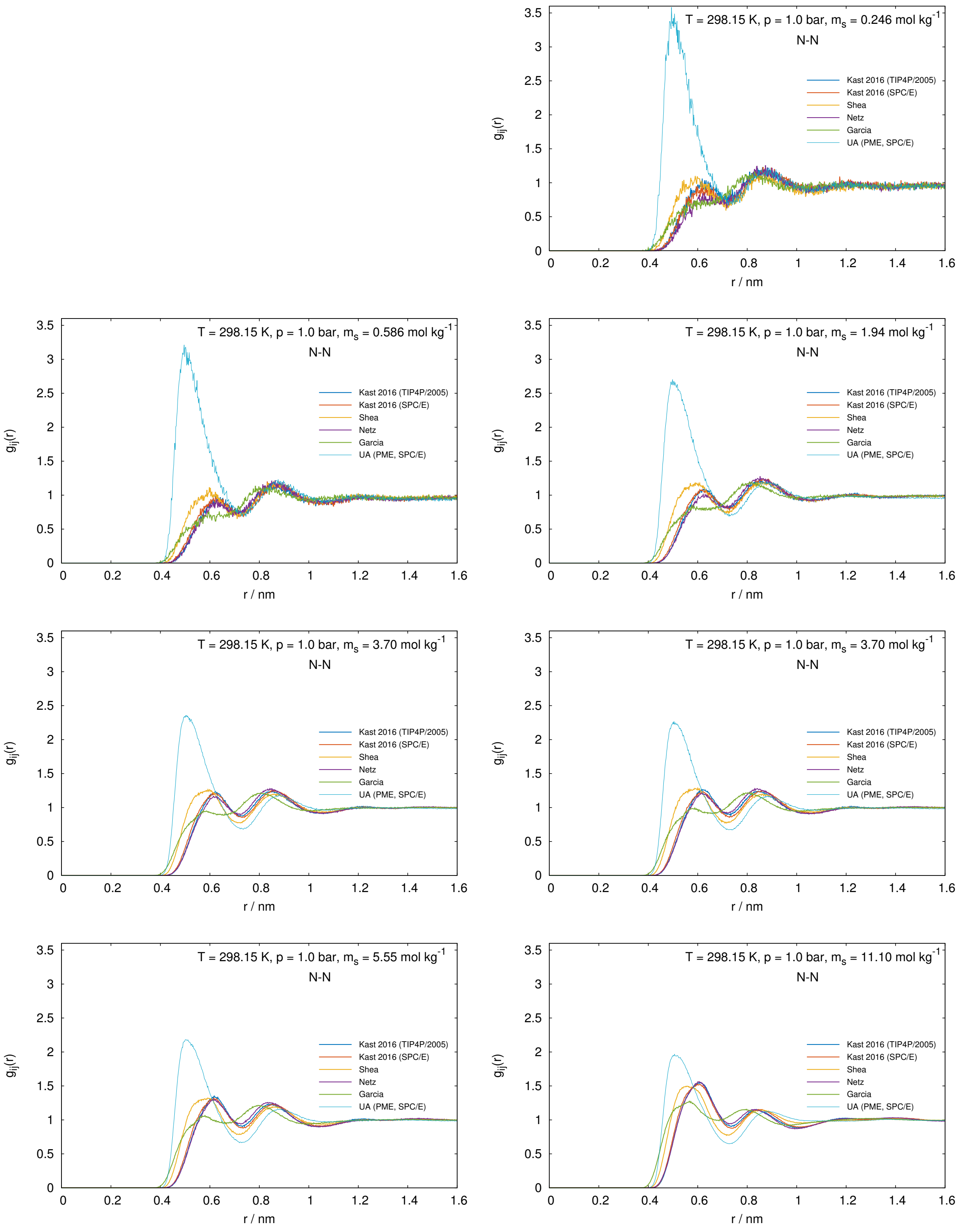


Figure A14: Radial distribution functions for N-N at 298.15 K and 1.0 bar for different molalities.

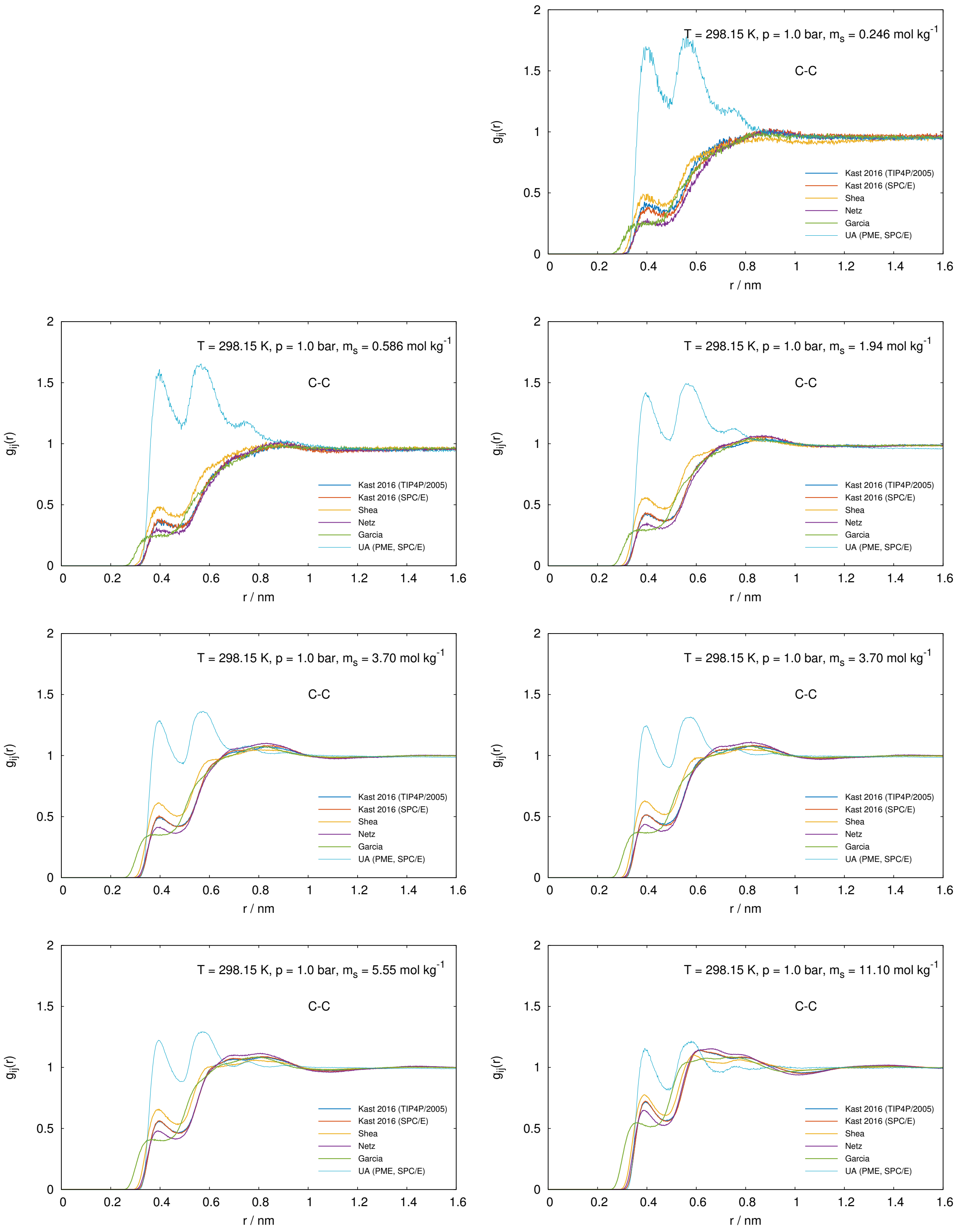


Figure A15: Radial distribution functions for C-C at 298.15 K and 1.0 bar for different molalities.

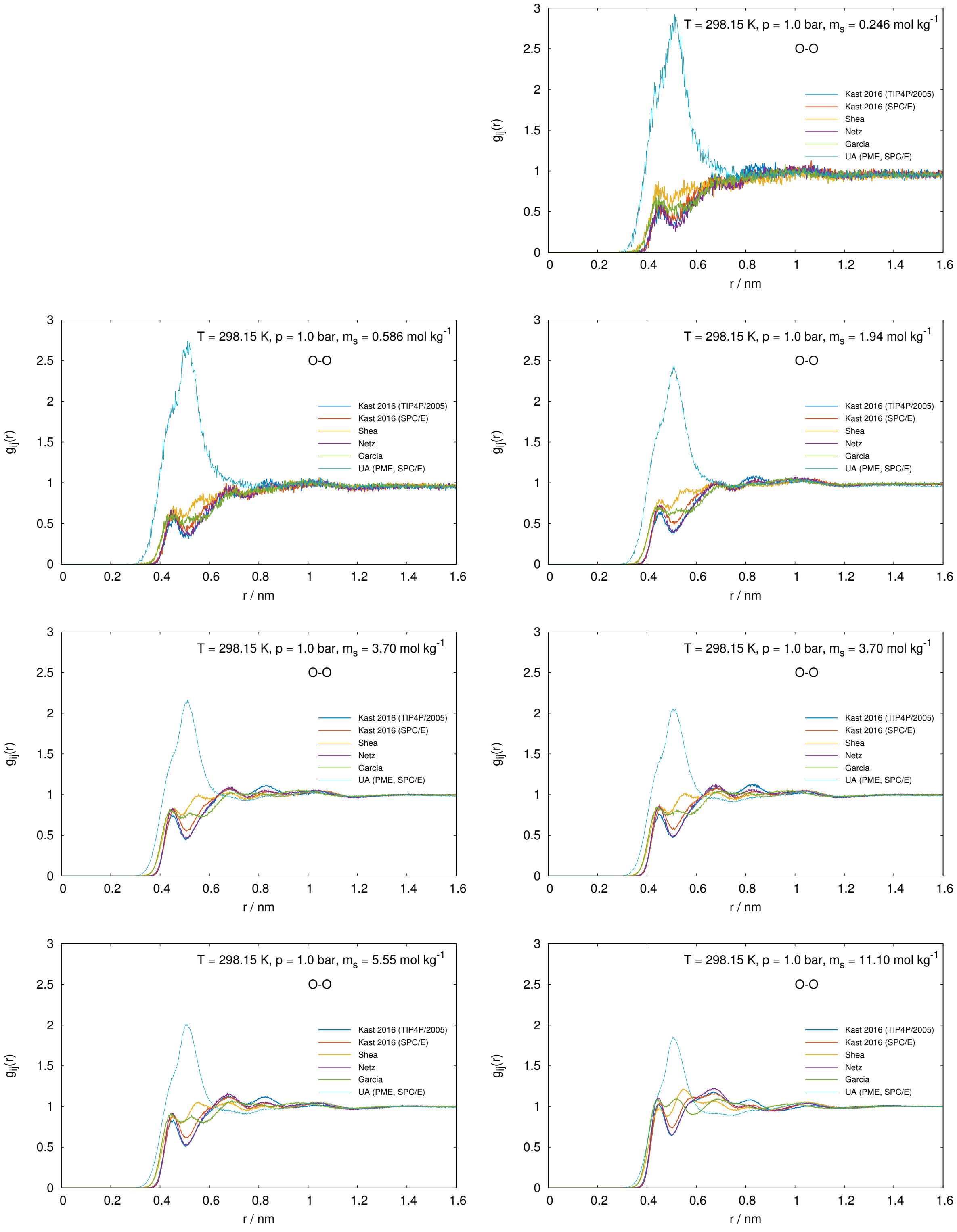


Figure A16: Radial distribution functions for O-O at 298.15 K and 1.0 bar for different molalities.

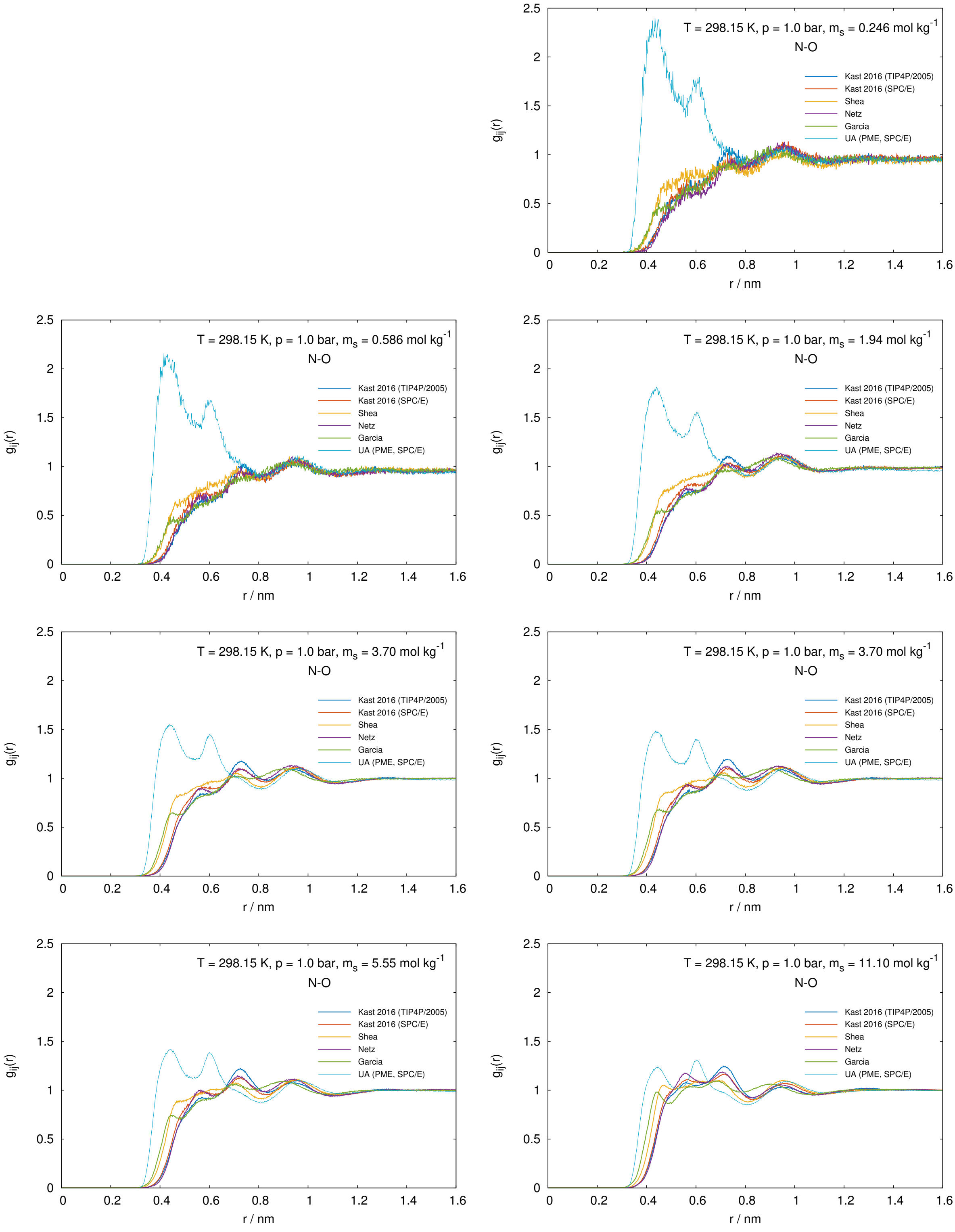


Figure A17: Radial distribution functions for N-O at 298.15 K and 1.0 bar for different molalities.

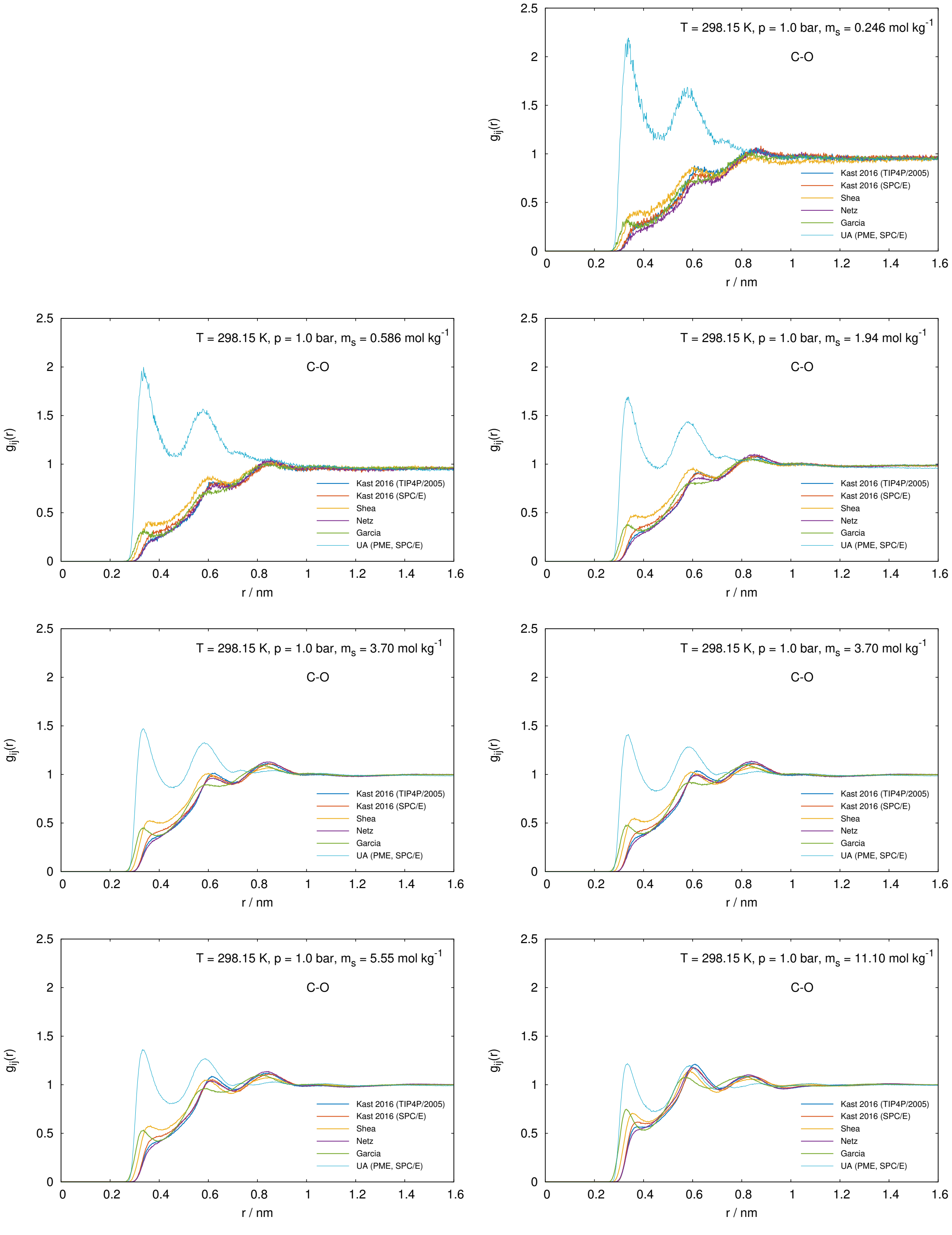


Figure A18: Radial distribution functions for C-O at 298.15 K and 1.0 bar for different molalities.

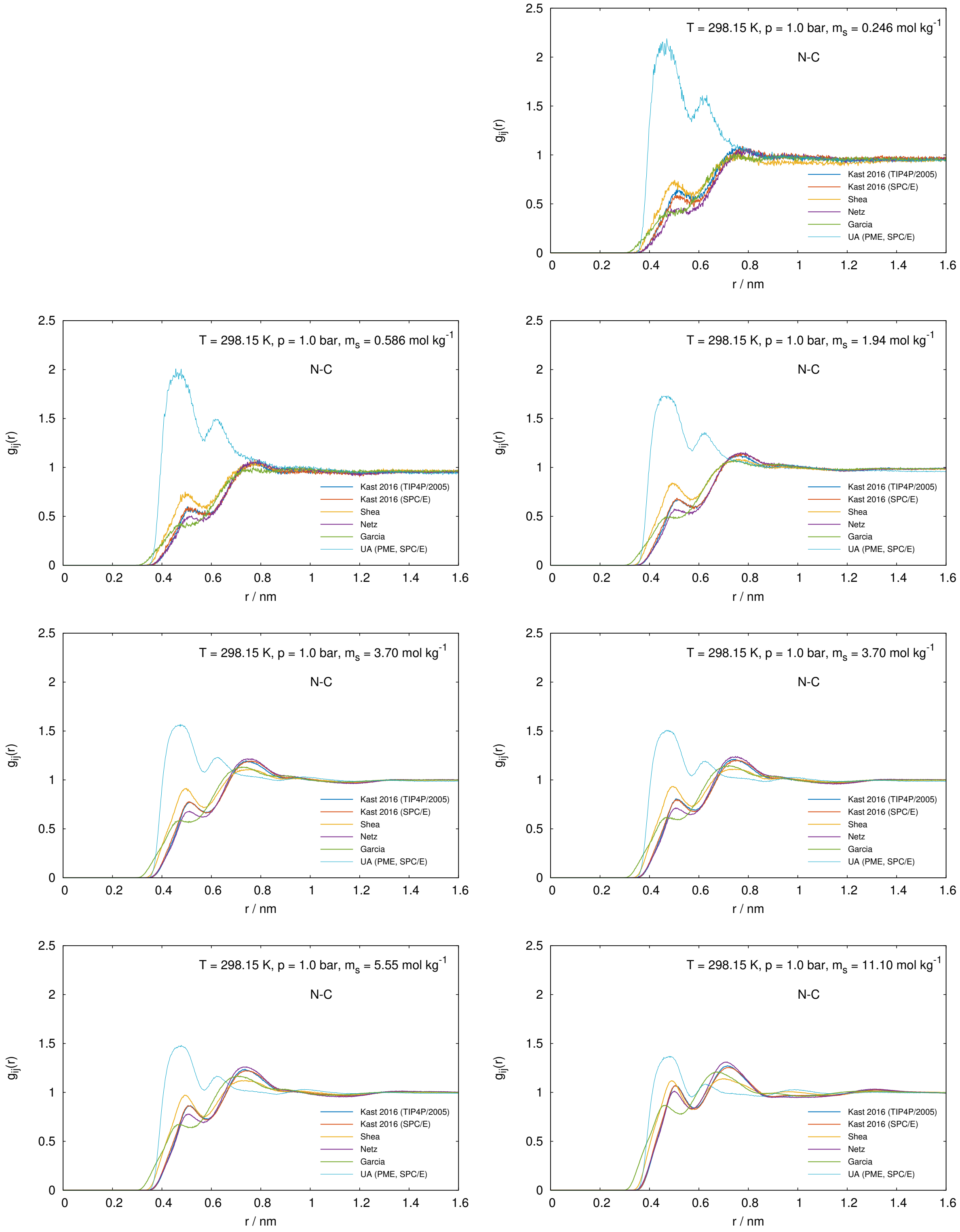


Figure A19: Radial distribution functions for N-C at 298.15 K and 1.0 bar for different molalities.

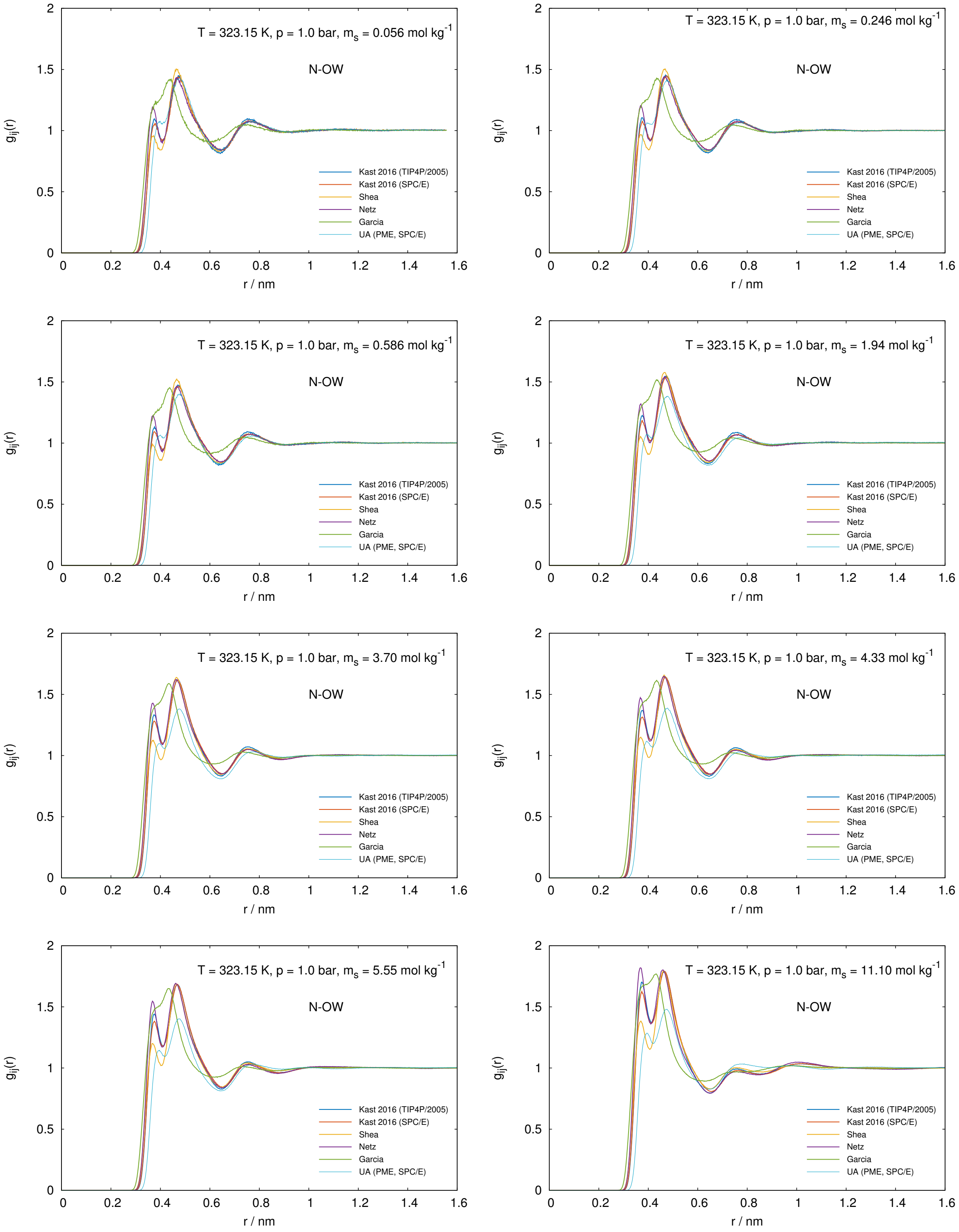


Figure A20: Radial distribution functions for N-OW at 323.15 K and 1.0 bar for different molalities.

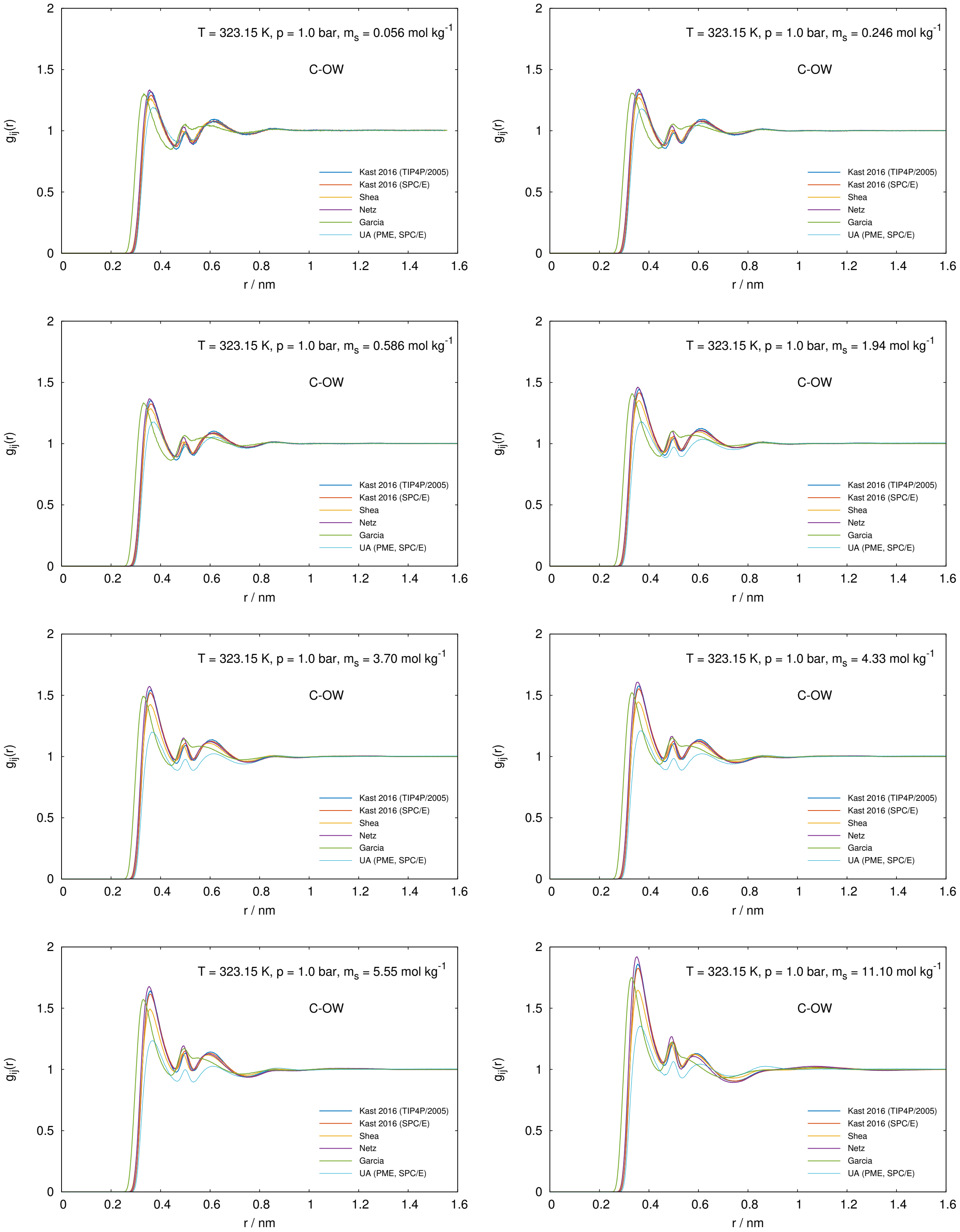


Figure A21: Radial distribution functions for C-OW at 323.15 K and 1.0 bar for different molalities.

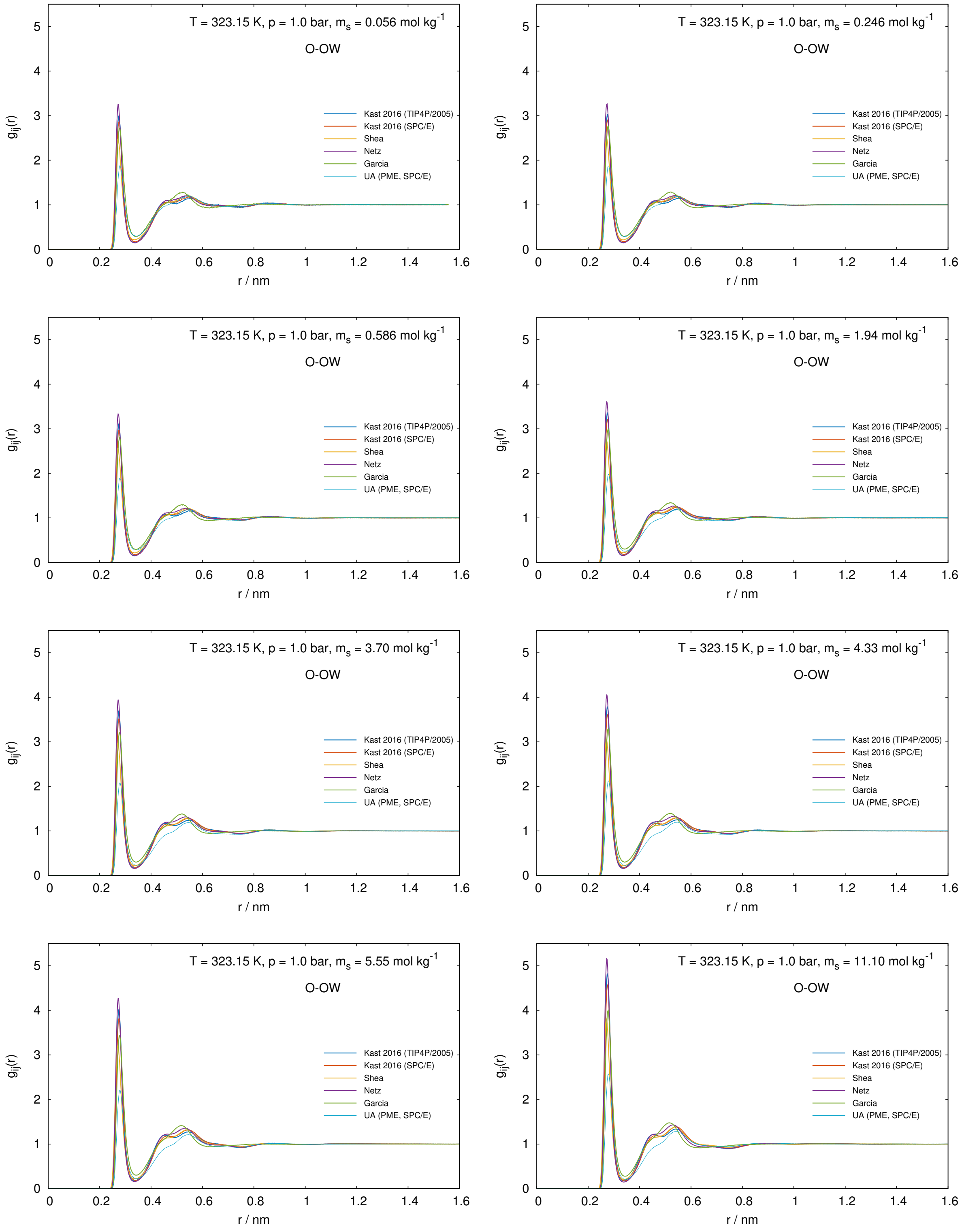


Figure A22: Radial distribution functions for O-OW at 323.15 K and 1.0 bar for different molalities.

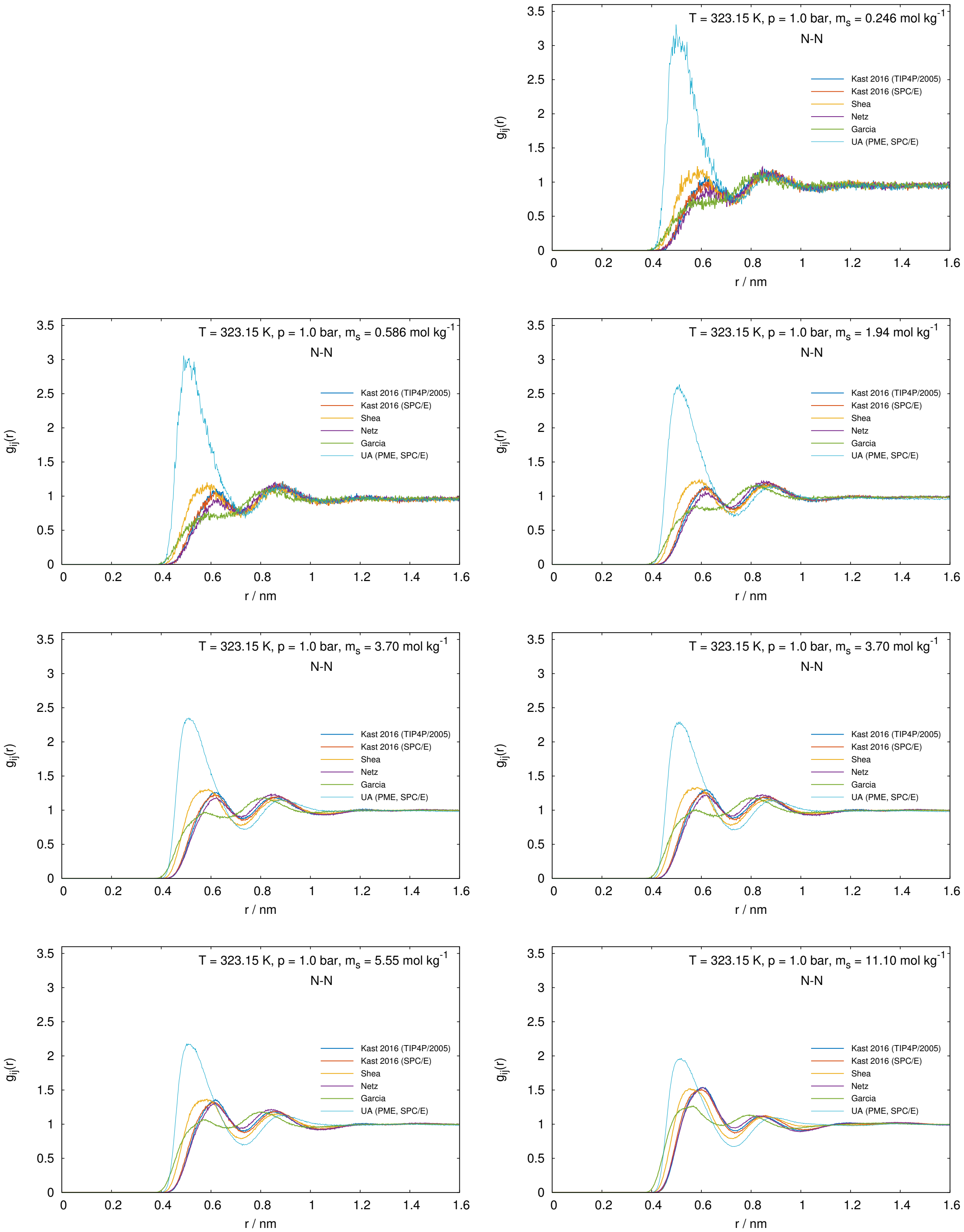


Figure A23: Radial distribution functions for N-N at 323.15 K and 1.0 bar for different molalities.

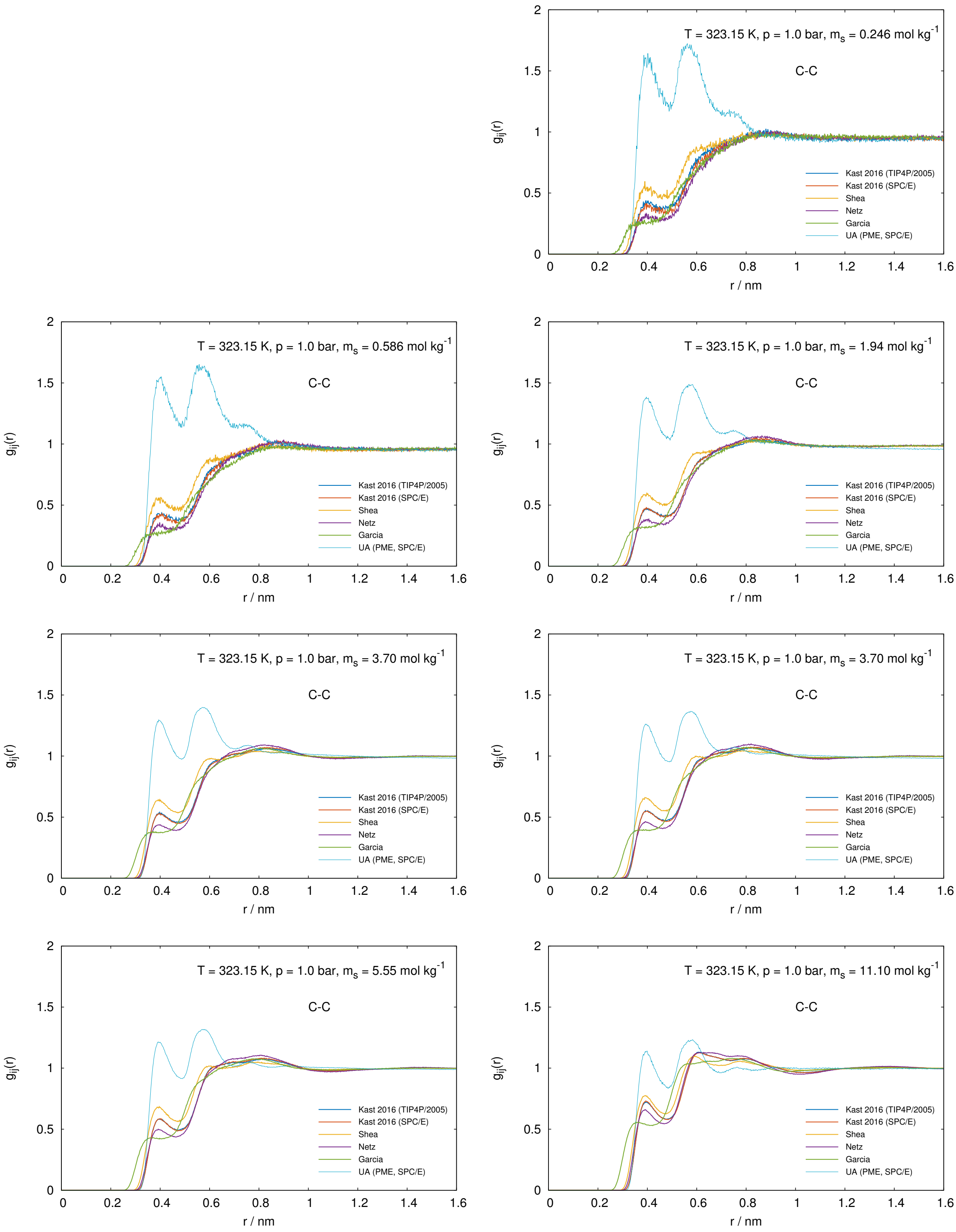


Figure A24: Radial distribution functions for C-C at 323.15 K and 1.0 bar for different molalities.

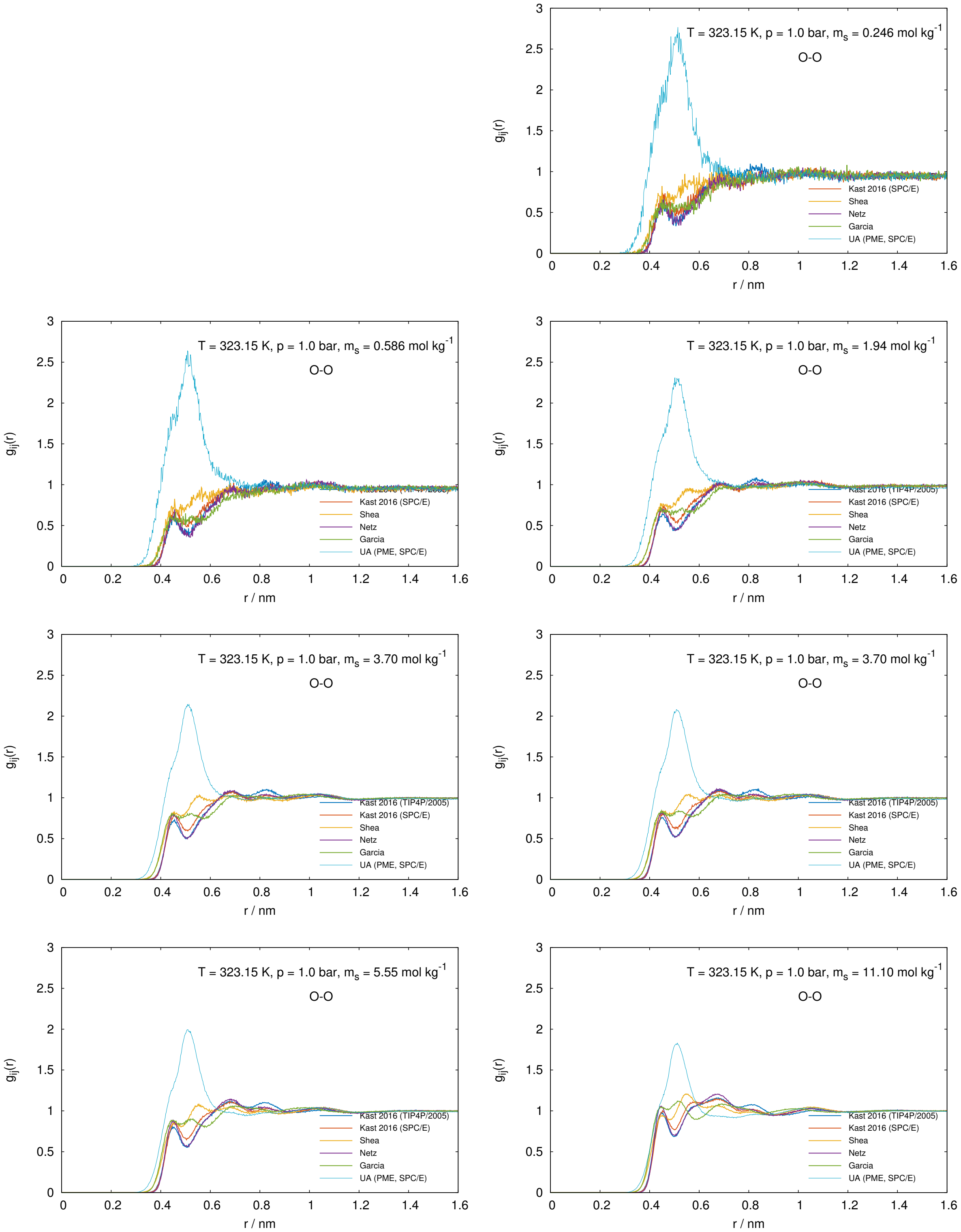


Figure A25: Radial distribution functions for O-O at 323.15 K and 1.0 bar for different molalities.

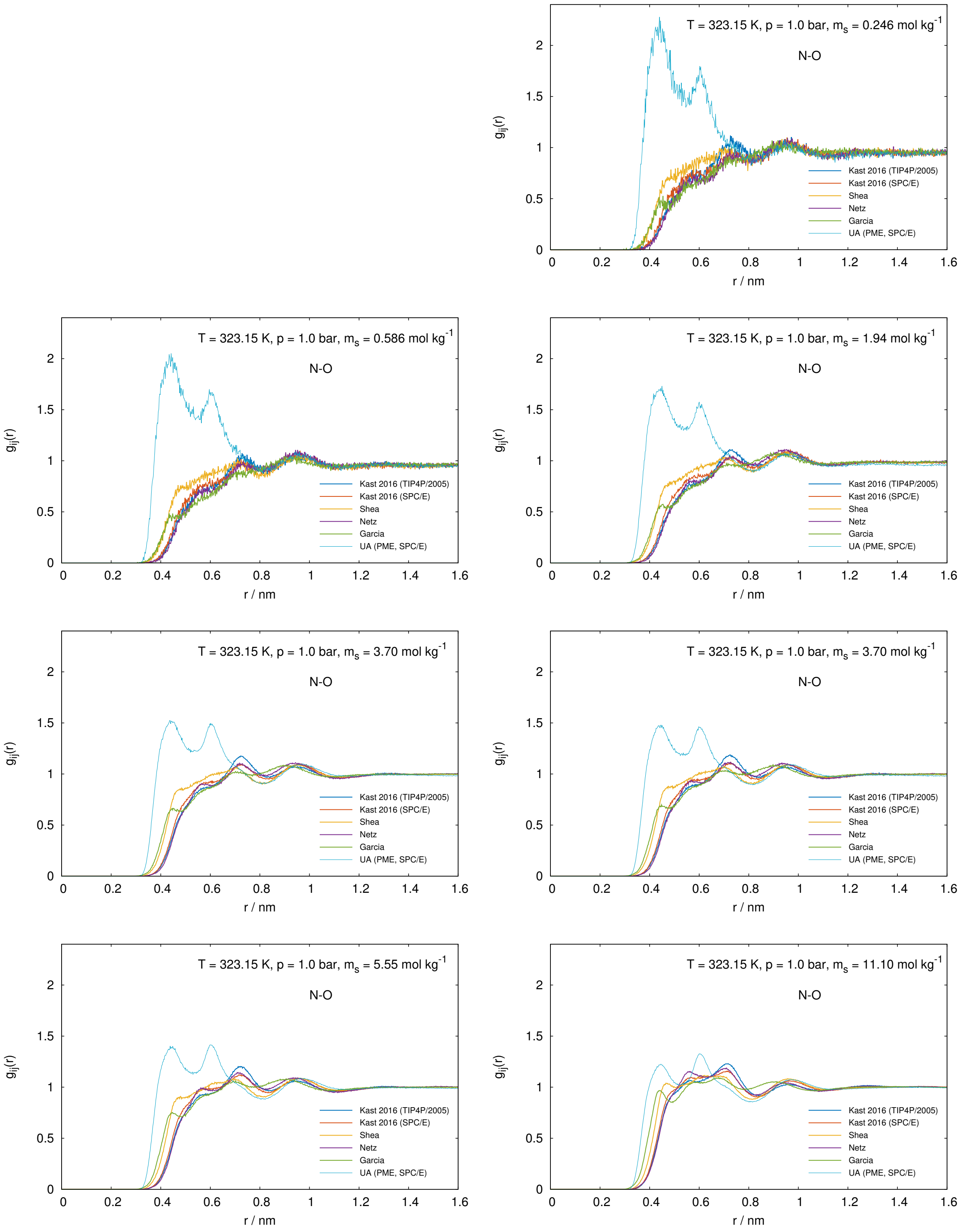


Figure A26: Radial distribution functions for N-O at 323.15 K and 1.0 bar for different molalities.

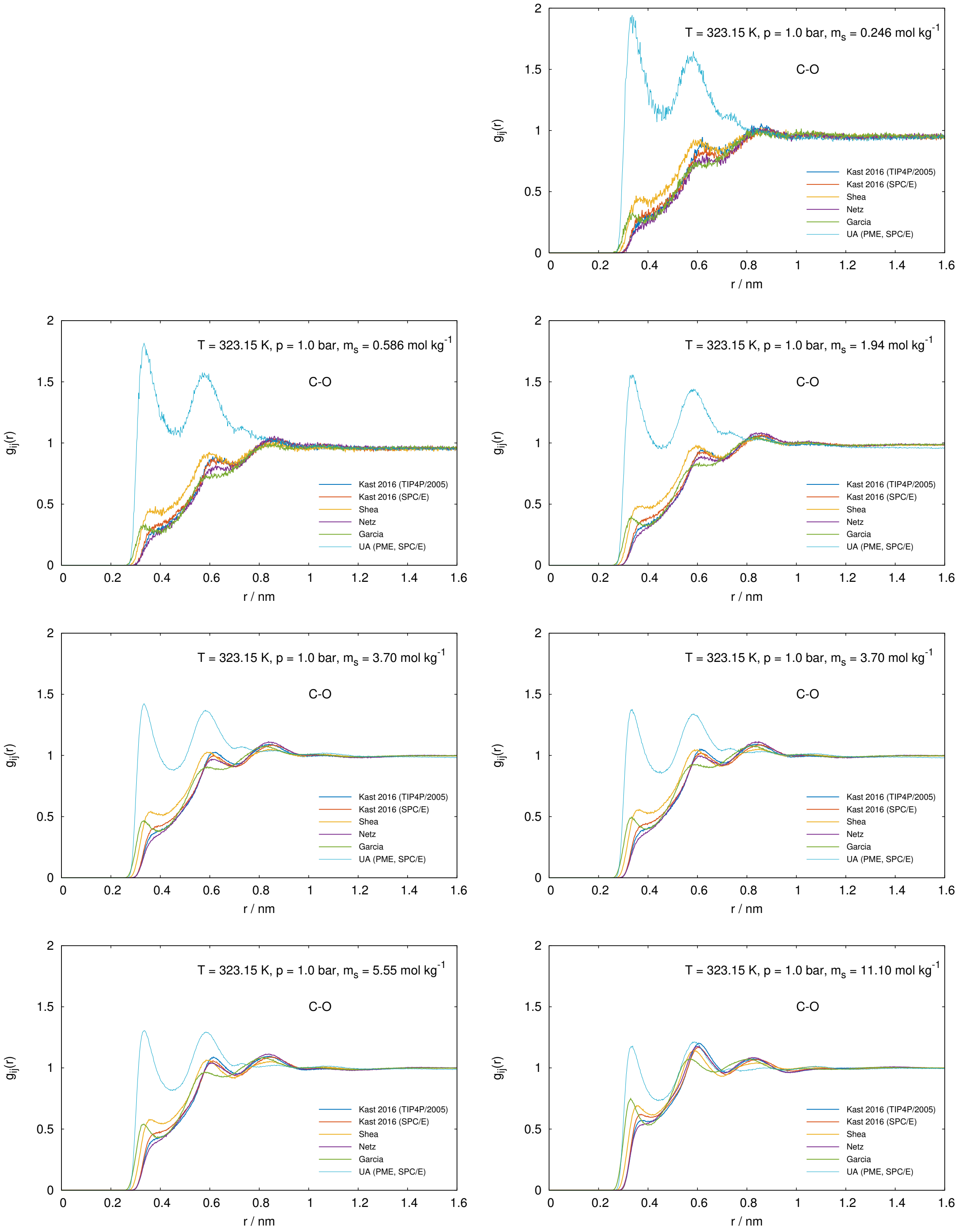


Figure A27: Radial distribution functions for C-O at 323.15 K and 1.0 bar for different molalities.

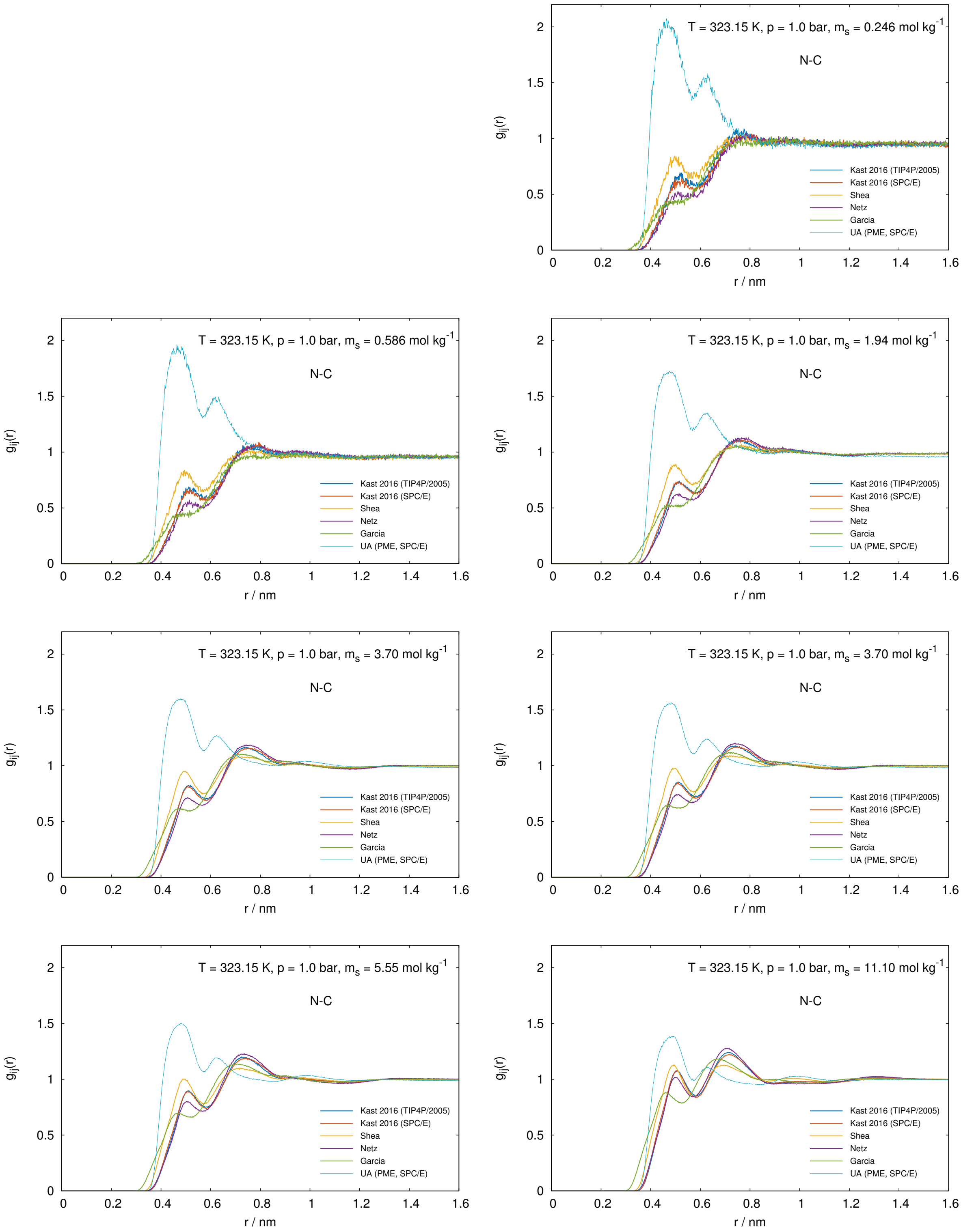


Figure A28: Radial distribution functions for N-C at 323.15 K and 1.0 bar for different molalities.

A.3 Densities and Apparent Molar Volumes

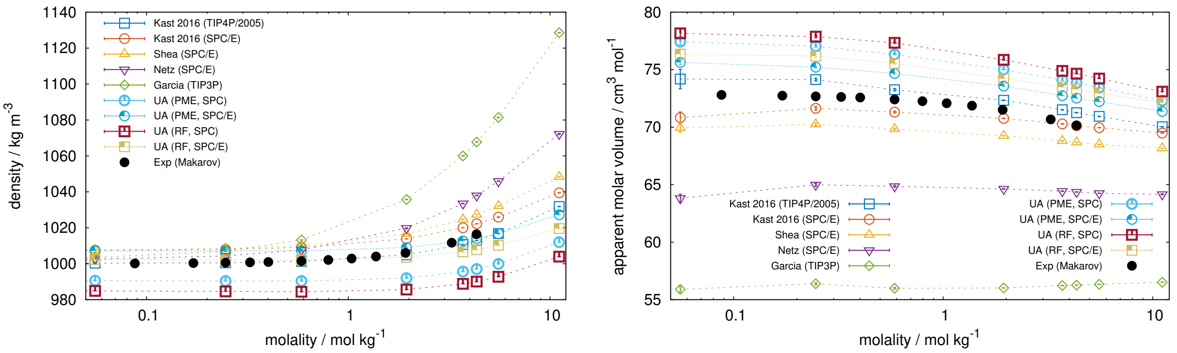


Figure A29: Densities and apparent molar volumes at 278.15 K and 1.0 bar for different force fields.

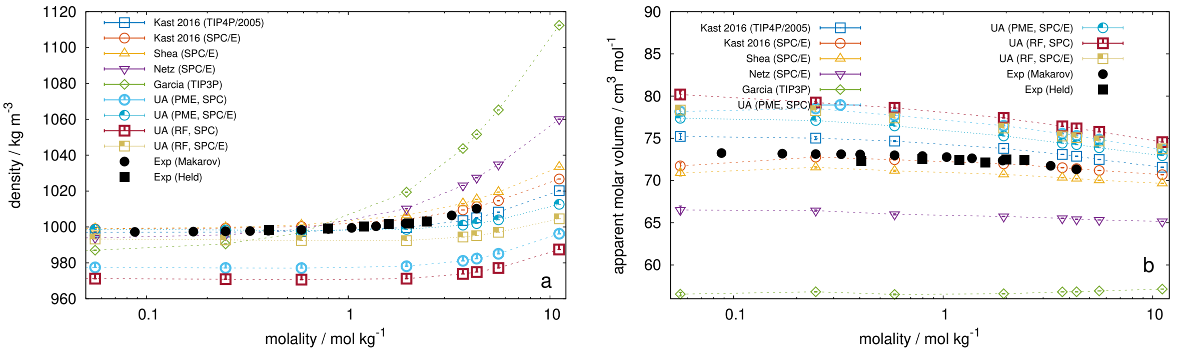


Figure A30: Densities and apparent molar volumes at 298.15 K and 1.0 bar for different force fields.

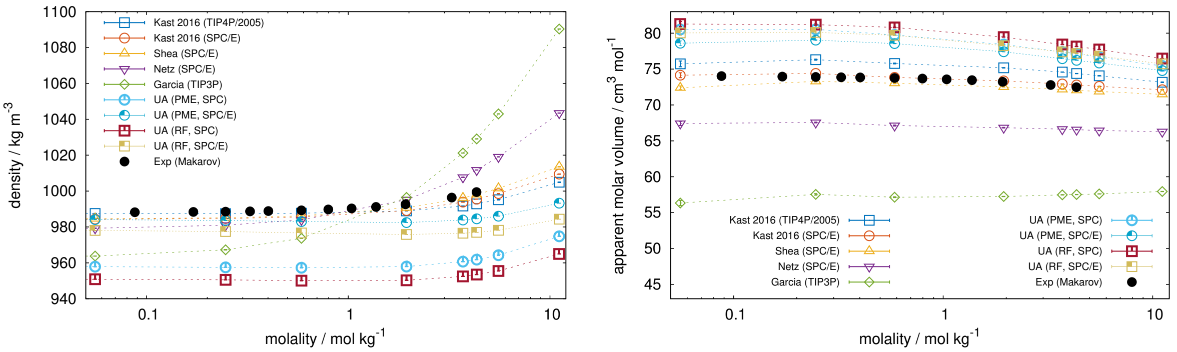


Figure A31: Densities and apparent molar volumes at 323.15 K and 1.0 bar for different force fields.

A.4 Hydrogen Bond Analysis and Self-Diffusion Coefficient

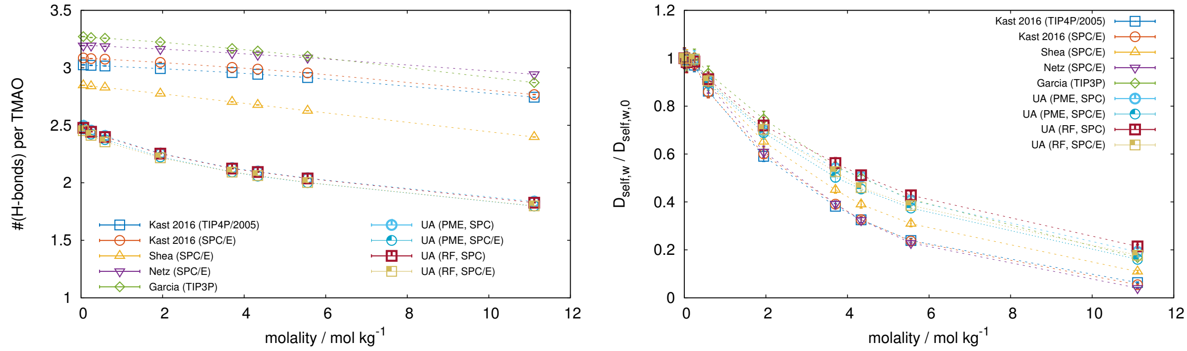


Figure A32: Number of hydrogen bonds per TMAO molecule and reduced self-diffusion coefficient of water at 278.15 K and 1.0 bar for different force fields.

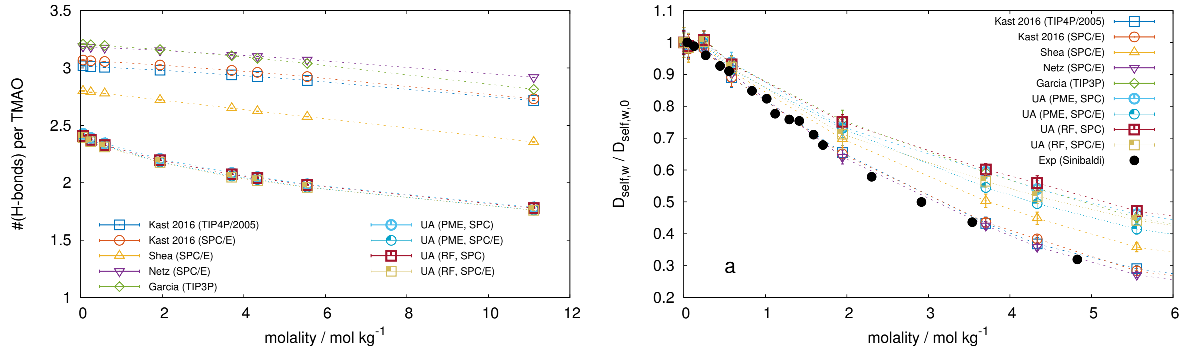


Figure A33: Number of hydrogen bonds per TMAO molecule and reduced self-diffusion coefficient of water at 298.15 K and 1.0 bar for different force fields.

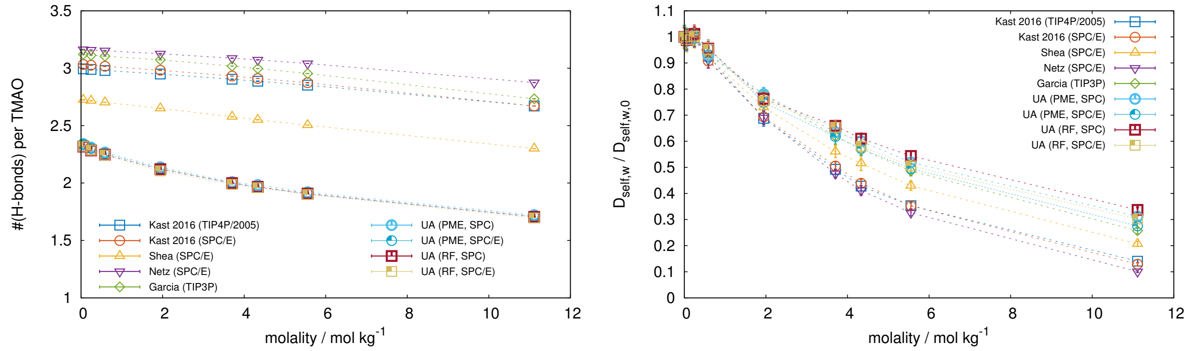


Figure A34: Number of hydrogen bonds per TMAO molecule and reduced self-diffusion coefficient of water at 323.15 K and 1.0 bar for different force fields.

A.5 Enthalpy and Apparent Molar Volume at Infinite Dilution

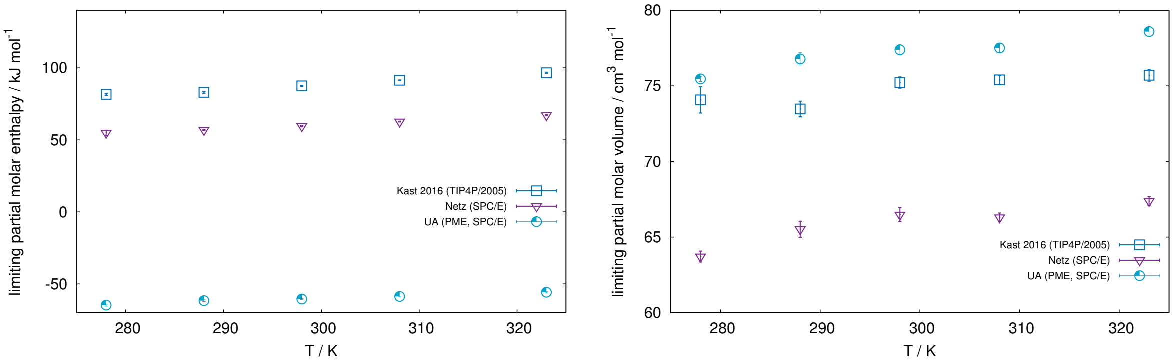


Figure A35: Enthalpies and apparent molar volumes at 1.0 bar for different force fields.

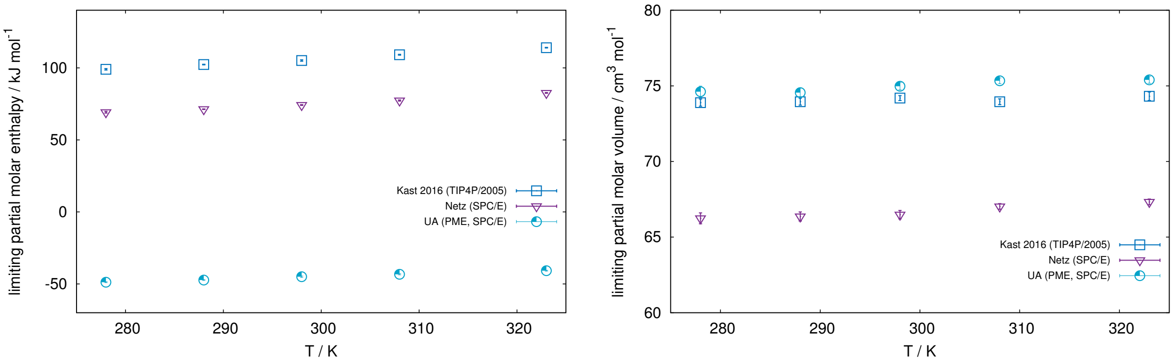


Figure A36: Enthalpies and apparent molar volumes at 2500.0 bar for different force fields.

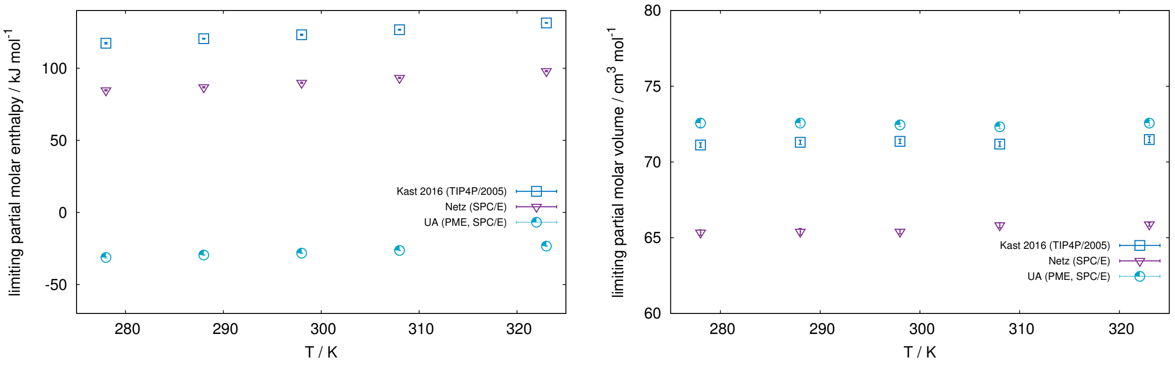


Figure A37: Enthalpies and apparent molar volumes at 5000.0 bar for different force fields.

A.6 Dielectric Properties

A.6.1 Calculation of Dielectric Spectra

All dielectric spectra have been computed following the fluctuation-based ("Green-Kubo") approach described in Refs. 149–151. The complex frequency-dependent permittivity $\varepsilon_r(\omega) := \varepsilon_r'(\omega) - i\varepsilon_r''(\omega)$ is calculated (in SI units) as

$$\varepsilon_r(\omega) = 1 + \frac{i}{\varepsilon_0 \omega} \left(\sigma(\omega) - \sigma(0)\right) , \tag{A.1}$$

where i denotes the imaginary number, ε_0 the permittivity of free space, ω the angular frequency of a hypothetically applied external electric field, and $\sigma(\omega)$ is the system's frequency-dependent total conductivity. The latter is computed as:

$$\begin{aligned} \sigma(\omega) &= \frac{1}{3Vk_BT} \int_0^\infty \langle \boldsymbol{J}_{\text{tot}}(0)\boldsymbol{J}_{\text{tot}}(t)\rangle \, e^{i\omega t} dt \\ &:= \frac{1}{3Vk_BT} \langle \boldsymbol{J}_{\text{tot}}(0)\boldsymbol{J}_{\text{tot}}(t)\rangle_\omega \ , \end{aligned} \tag{A.2}$$

where V is the volume of the simulation box, k_B the Boltzmann constant, T temperature, $\langle\cdot\rangle$ denotes the canonical average, $\langle\cdot\rangle_\omega$ its Fourier-Laplace transform, and $\boldsymbol{J}_{\text{tot}}(t)$ is the fluctuating cumulative current

$$\boldsymbol{J}_{\text{tot}}(t) = \sum_m \sum_\alpha q_{m,\alpha} \boldsymbol{v}_{m,\alpha}(t) \tag{A.3}$$

summed over all molecules m consisting of atoms α with partial charges $q_{m,\alpha}$ and corresponding velocities $\boldsymbol{v}_{m,\alpha}(t)$.

A.6.2 Noise Reduction

Since the long-time tails of the current autocorrelations $A_{\boldsymbol{J}_{\text{tot}}}(\tau) := \langle \boldsymbol{J}_{\text{tot}}(0)\boldsymbol{J}_{\text{tot}}(\tau)\rangle$ used for the calculations of dielectric spectra are generally noisy, a physically consistent noise reduction technique has been employed. An often-used method for noise reduction is to fit the tail of the autocorrelation with a suitable analytic function $f(\tau)$ starting from an appropriate lag time τ_{tail}. Then, in the analysis, the raw data is used for lag times $\tau < \tau_{\text{tail}}$, and the analytic fit for $\tau \geq \tau_{\text{tail}}$. However, when integral quantities are computed, it is a common misconception to apply a least-squares fit *directly* to the autocorrelation data. When employing the least-squares norm as a convergence criterion for the fit, the problem that arises is that this norm does *not* preserve the integral of the data. An additional problem specific to the calculation of spectra is the abrupt switch at τ_{tail} from the raw data to the fit funtion, which may lead to a small but sudden jump in the data. Such jumps can introduce artifacts in the Fourier transforms known as *spectral leakage*.

We overcome both problems in a mathematically and physically consistent way by determining the fit parameters from the integral of the autocorrelation data and by smoothly tapering the raw data to the fit function. The procedure is as follows:

1. Integrate raw autocorrelation data $A_{\boldsymbol{J}_{\text{tot}}}(\tau)$ to obtain $B_{\boldsymbol{J}_{\text{tot}}}(\tau) := \int A_{\boldsymbol{J}_{\text{tot}}}(\tau)\, d\tau$.
2. Fit an appropriate integrated analytic function $F(\tau) = \int f(\tau)\, d\tau + c$ to $B_{\boldsymbol{J}_{\text{tot}}}(\tau)$ starting from τ_{tail}.
3. From the obtained fit parameters of $F(\tau)$, determine the parameters of the fit function $f(\tau)$ that is supposed to fit the original $A_{\boldsymbol{J}_{\text{tot}}}(\tau)$.
4. On the interval $[\tau_{\text{tail}},\ \tau_{\text{tail}} + \Delta\tau_{\text{taper}}]$, let the autocorrelation data gradually approach the fit function $f(\tau)$ by means of an appropriate taper function $w(\tau)$ to obtain the noise-reduced autocorrelation data $\tilde{A}_{\boldsymbol{J}_{\text{tot}}}(\tau)$. A linear taper is usually sufficient.
5. Compute the integral $\tilde{B}_{\boldsymbol{J}_{\text{tot}}}(\tau) = \int \tilde{A}_{\boldsymbol{J}_{\text{tot}}}(\tau)\, d\tau$ and compare to $B_{\boldsymbol{J}_{\text{tot}}}(\tau)$. If the integral is not preserved, increase τ_{tail} and repeat from step 2.

As an example, we demonstrate the difference between our integral-preserving noise-reduction procedure ("integral-fitted/tapered") and the often-used direct approach ("directly fitted/tapered") for a system of pure SPC/E water. The raw current autocorrelation data $A_{\boldsymbol{J}_{\text{tot}}}(\tau)$ displayed in panel **a** of Fig. A38 (blue line) was obtained from an NPT simulation of 2180 SPC/E water molecules at 300 K and 1 bar with a length of ≈ 1.34 μs (2^{30} steps, $\Delta t = 1$ fs)).

We then followed the procedure as described above:

1. The integral $B_{\boldsymbol{J}_{\text{tot}}}(\tau)$ shown in panel **b** (blue line) was evaluated numerically by means of the composite trapezoidal rule.
2. We then performed a least-squares fit of the function $F(\tau) = a_F \exp\left(b_F(\tau - c_F)\right) + d_F$ to $B_{\boldsymbol{J}_{\text{tot}}}(\tau)$ for $\tau \geq \tau_{\text{tail}}$ with $\tau_{\text{tail}} = 2$ ps.
3. The obtained parameters a_F, b_F and c_F were used to determine the parameters of the function $f(\tau) = a_f \exp\left(b_f(\tau - c_f)\right) \overset{!}{=} \frac{d}{d\tau}F(\tau)$ as $a_f = a_F b_F$, $b_f = b_F$, and $c_f = c_F$.
4. On the interval $[\tau_{\text{tail}},\ \tau_{\text{tail}} + \Delta\tau_{\text{taper}}]$ with $\Delta\tau_{\text{taper}} = 10$ ps, we let the autocorrelation data approach $f(\tau)$ using a linear taper $w(\tau) = 1 - (\tau - \tau_{\text{tail}})/\Delta\tau_{\text{taper}}$ so that

$$\tilde{A}_{\boldsymbol{J}_{\text{tot}}}(\tau) = \begin{cases} A_{\boldsymbol{J}_{\text{tot}}}(\tau), & \tau < \tau_{\text{tail}} \\ f(\tau) + w(\tau)\left(A_{\boldsymbol{J}_{\text{tot}}}(\tau) - f(\tau)\right), & \tau_{\text{tail}} \leq \tau \leq \tau_{\text{tail}} + \Delta\tau_{\text{taper}} \\ f(\tau), & \tau > \tau_{\text{tail}} + \Delta\tau_{\text{taper}}\,. \end{cases}$$

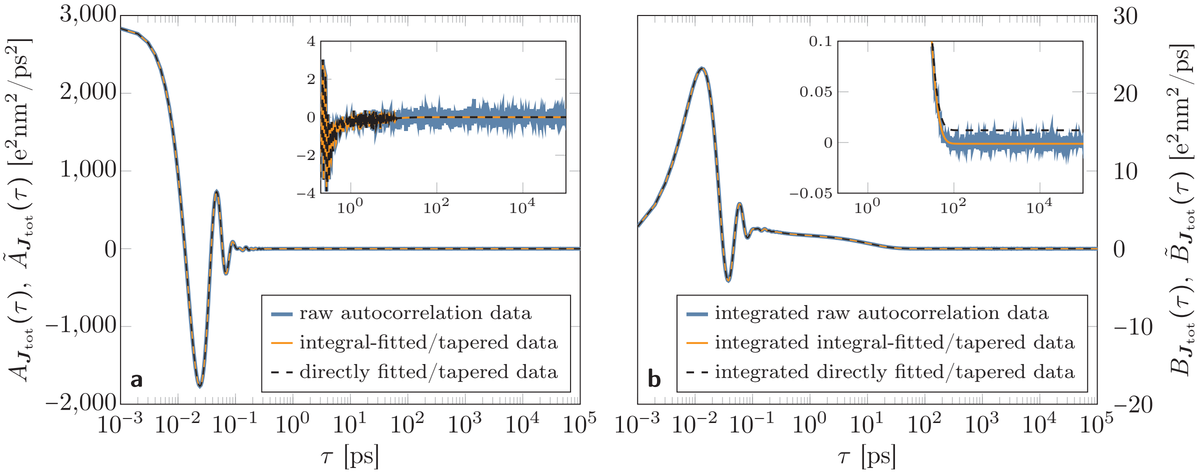


Figure A38: Noise reduction example. The data was obtained from an NPT simulation of 2180 SPC/E water molecules at 300 K and 1 bar with a length of ≈ 1.34 μs (2^{30} steps, $\Delta t = 1$ fs). **a**: Current autocorrelation data $A_{\boldsymbol{J}_{\text{tot}}}(\tau) = \langle \boldsymbol{J}_{\text{tot}}(0)\boldsymbol{J}_{\text{tot}}(\tau)\rangle$ (blue line). The orange line depicts the data obtained from the noise-reduction procedure ("integral-fitted/tapered") described above with an exponential fit function and a linear taper. For comparison, a direct fit to the data of the same functional form is shown as a dashed black line ("directly fitted/tapered"). Even in the magnified view of the autocorrelation tail provided in the inset, both fits seem to coincide since their difference is much smaller than the line width. **b**: Corresponding integrals of the raw autocorrelation data and fits from panel **a** (same color code). From the magnified view of the integrals' tails shown in the inset it becomes evident that the direct fit to the autocorrelation data does *not* preserve its integral (dashed black line).

The resulting data $\tilde{A}_{\boldsymbol{J}_{\text{tot}}}(\tau)$ is depicted in Fig. A38, panel **a** (orange line).

5. Once again using the trapezoidal rule, we computed the integral $\tilde{B}_{\boldsymbol{J}_{\text{tot}}}(\tau) = \int \tilde{A}_{\boldsymbol{J}_{\text{tot}}}(\tau)\, d\tau$ displayed in Fig. A38, panel **b** (orange line). The zoom-in on the tail shown in the inset confirms that $\tilde{B}_{\boldsymbol{J}_{\text{tot}}}(\tau)$ indeed follows $B_{\boldsymbol{J}_{\text{tot}}}(\tau)$.

In order to highlight the importance of this procedure, Fig. A38 also depicts $\tilde{A}_{\boldsymbol{J}_{\text{tot}}}(\tau)$ (panel **a**, dashed black line) and $\tilde{B}_{\boldsymbol{J}_{\text{tot}}}(\tau)$ (panel **b**, dashed black line), where $\tilde{A}_{\boldsymbol{J}_{\text{tot}}}(\tau)$ was obtained by fitting $f(\tau)$ directly to $A_{\boldsymbol{J}_{\text{tot}}}(\tau)$. In panel **a**, the data obtained by the two different procedures seems to perfectly coincide. However, the dashed black line in the inset of panel **b** clearly shows that the direct fit procedure fails to reproduce the integral of the original raw data.

Figure A39 shows the impact of the noise-reduction technique on dielectric spectra. The spectrum was computed from the current autocorrelation data shown above according to Eq. (A.1) and (A.2) (same color code as in Fig. A38). While the integral-fitting technique results in a smooth spectrum perfectly following the noisy spectrum obtained from the raw data, the direct fit approach yields a spectrum clearly underestimating the low-frequency part of the permittivity and the amplitude of the main loss peak.

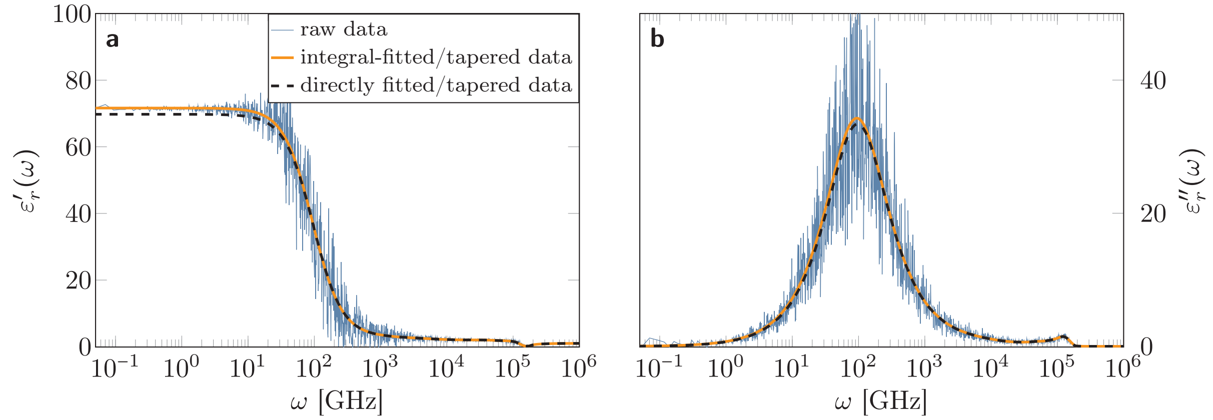


Figure A39: Dielectric spectrum of pure SPC/E water at 300 K and 1 bar. **a**: Permittivity $\varepsilon'_r(\omega)$. **b**: Loss $\varepsilon''_r(\omega)$. The integral-fitting technique results in a smooth spectrum (orange lines) perfectly following the spectrum computed from raw data (blue lines). The spectrum obtained from the direct fit approach clearly underestimates both the low-frequency part of $\varepsilon'_r(\omega)$ and the loss peak amlitude of $\varepsilon''_r(\omega)$.

A.6.3 Calculation of Static Permittivities

The static permittivities $\varepsilon_{\text{static}} := \varepsilon'_r(\omega = 0)$ have been computed according to the so-called "Einstein-Helfand" method[149,152]. This method exploits the long-term behavior of the mean square displacement (MSD) $\left\langle \Delta \boldsymbol{M}^2_{\text{tot}}(t) \right\rangle$ of the system's total itinerant dipole moment $\boldsymbol{M}_{\text{tot}}$:

$$\lim_{t\to\infty} \left\langle \Delta \boldsymbol{M}^2_{\text{tot}}(t) \right\rangle = 2 \left\langle \boldsymbol{M}^2_{\text{tot}} \right\rangle + 6Vk_BT\sigma t \tag{A.4}$$

Here, V denotes the average volume of the simulation box, k_B the Boltzmann constant, T temperature, and σ the *total* static conductivity[1]. The MSD of the total dipole moment is calculated as

$$\left\langle \Delta \boldsymbol{M}^2_{\text{tot}}(t) \right\rangle = 2t \int_0^t \left\langle \boldsymbol{J}_{\text{tot}}(0)\boldsymbol{J}_{\text{tot}}(\tau) \right\rangle d\tau - 2 \int_0^t \tau \left\langle \boldsymbol{J}_{\text{tot}}(0)\boldsymbol{J}_{\text{tot}}(\tau) \right\rangle d\tau \,, \tag{A.5}$$

where $\boldsymbol{J}_{\text{tot}}(t)$ is the the cumulative current as defined in Eq. (A.3). Once $\left\langle \Delta \boldsymbol{M}^2_{\text{tot}}(t) \right\rangle$ is computed for a sufficiently long time, the static permittivity $\varepsilon_{\text{static}}$ can be obtained according to Eq. (A.4) from the y-axis offset of a linear regression of $\left\langle \Delta \boldsymbol{M}^2_{\text{tot}}(t) \right\rangle$ together with the following fluctuation formula (see Ref. 325):

$$\varepsilon_{\text{static}} = 1 + \frac{\left\langle \boldsymbol{M}^2 \right\rangle}{3\varepsilon_0 V k_B T} \tag{A.6}$$

[1] It should be mentioned that the total static conductivity σ in Eq. (A.4) cannot be used as an estimate for the true static conductivity since it contains rotational contributions.

A.6.4 Simulation Details

Simulation Parameters

Table A7: Detailed list of parameters used in the simulations from which dielectric spectra and static permittivities were calculated.

TMAO force field	Garcia	Kast 2016	Netz	Shea	United-atom
Independent runs	8	8	8	8	8
Integration time step [ps]	0.001	0.001	0.001	0.001	0.001
Equilibration time [ps]	5000	5000	5000	5000	5000
Simulation time per run [ps]	268435.455	268435.455	268435.455	268435.455	268435.455
Van der Waals cut-off [nm]	1.3	1.0	1.2	1.2	1.4
Coulomb short-range cut-off [nm]	1.3	1.0	1.2	1.2	1.4
Dispersion correction	yes	yes	no	yes	yes
Electrostatics algorithm	PME	PME	PME	PME	PME
Ewald interpolation order	4	4	4	4	4
Relative Ewald tolerance	0.00001	0.00001	0.00001	0.00001	0.00001
Thermostat	Nosé-Hoover	Nosé-Hoover	Nosé-Hoover	Nosé-Hoover	Nosé-Hoover
Thermostat coupling constant [ps]	1.0	1.0	1.0	1.0	1.0
Temperature [K]	298.15	298.15	298.15	298.15	298.15
Barostat	Parrinello-Rahman	Parrinello-Rahman	Parrinello-Rahman	Parrinello-Rahman	Parrinello-Rahman
Barostat coupling constant [ps]	2.0	2.0	2.0	2.0	2.0
Pressure [bar]	1.0	1.0	1.0	1.0	1.0
Compressibility [bar^{-1}]	0.000045	0.000045	0.000045	0.000045	0.000045
Bond length constraints	all bonds	bonds with H atoms	all bonds	all bonds	all bonds
Constraint algorithm	LINCS (4-th order)	LINCS (4-th order)	LINCS (4-th order)	LINCS (4-th order)	LINCS (4-th order)

System Parameters

Table A8: Details of the simulations used to extract dielectric spectra and static permittivities for different TMAO and water models. The following quantities are listed: Molality m in mol(TMAO)/kg(H_2O), molarity c in [mol/l], number of TMAO molecules N_{TMAO}, number of water molecules N_{water}, average simulation box volume V in nm^3, average density ρ in kg/m^3.

TMAO force field	Water model	m [mol/kg]	c [mol/l]	N_{TMAO}	N_{water}	V [nm^3]	ρ [kg/m^3]
Garcia	TIP3P	0.0000	0.00	0	2180	66.16	985.9
		2.0003	1.77	100	2775	93.59	1020.3
		3.9991	3.22	100	1388	51.53	1047.9
		7.9983	5.44	100	694	30.50	1089.6
		10.0015	6.31	100	555	26.30	1105.8
Kast 2016	SPC/E	0.0000	0.00	0	2180	65.30	998.7
		2.0003	1.75	100	2775	95.07	1004.4
		3.9991	3.11	100	1388	53.44	1010.5
		7.9983	5.10	100	694	32.54	1021.3
		10.0015	5.86	100	555	28.35	1025.5
Kast 2016	TIP4P/2005	0.0000	0.00	0	1200	36.01	997.0
		2.0003	1.74	100	2775	95.51	999.8
		3.9991	3.09	100	1388	53.76	1004.5
		7.9983	5.07	100	694	32.76	1014.4
		10.0015	5.82	100	555	28.54	1018.8
Netz	SPC/E	0.0000	0.00	0	2180	65.65	993.3
		2.0003	1.76	100	2775	94.47	1010.7
		3.9991	3.15	100	1388	52.66	1025.4
		7.9983	5.24	100	694	31.70	1048.3
		10.0015	6.03	100	555	27.51	1057.1
Shea	SPC/E	0.0000	0.00	0	2180	65.30	998.8
		2.0003	1.75	100	2775	94.85	1006.8
		3.9991	3.12	100	1388	53.23	1014.4
		7.9983	5.13	100	694	32.35	1027.2
		10.0015	5.89	100	555	28.17	1032.2
United-atom	SPC/E	0.0000	0.00	0	2180	65.31	998.6
		2.0003	1.74	100	2775	95.62	998.6
		3.9991	3.08	100	1388	53.91	1001.6
		7.9983	5.04	100	694	32.95	1008.7
		10.0015	5.78	100	555	28.73	1012.2

A.6.5 Dielectric Spectra

In the main article, dielectric spectra are only shown for the Kast 2016 model with TIP4P/2005 water. Here, we provide the dielectric spectra of all systems. Note that the *angular* frequency $\omega = 2\pi\nu$ is plotted on the abscissa. The noisy pale lines correspond to spectra obtained from current autocorrelations without noise reduction.

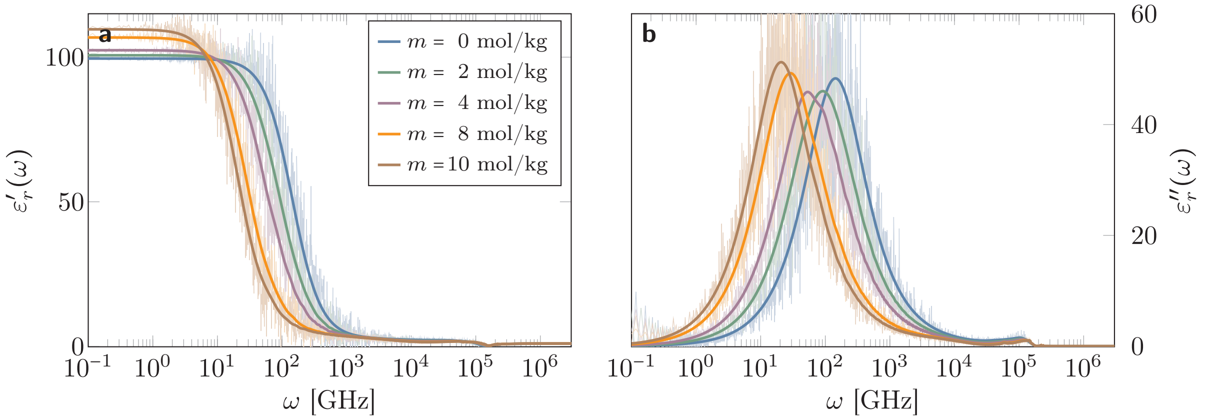


Figure A40: Dielectric spectra of aqueous TMAO solutions at different TMAO molalities for the Garcia model with SPC/E water. **a**: permittivity $\varepsilon'_r(\omega)$. **b**: loss $\varepsilon''_r(\omega)$.

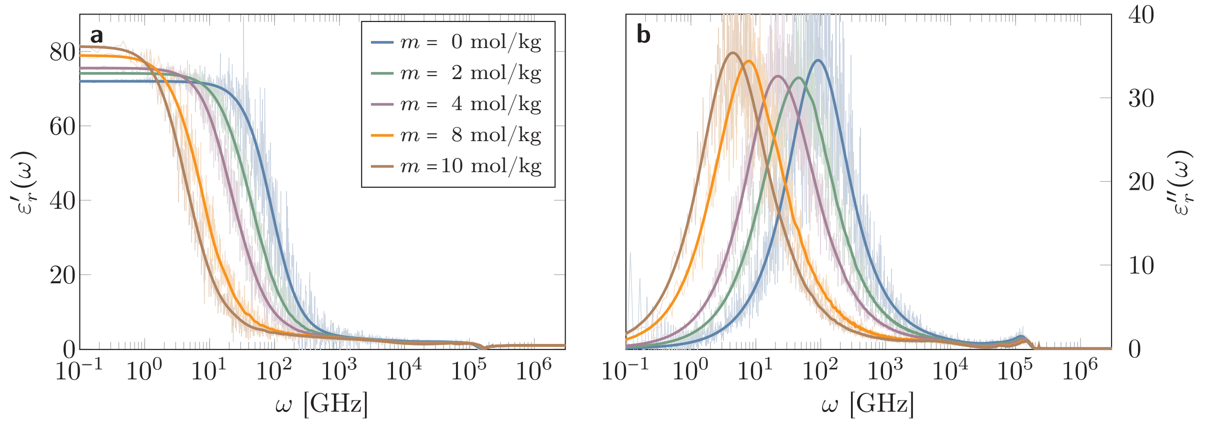


Figure A41: Dielectric spectra of aqueous TMAO solutions at different TMAO molalities for the Kast 2016 model with SPC/E water. **a**: permittivity $\varepsilon'_r(\omega)$. **b**: loss $\varepsilon''_r(\omega)$.

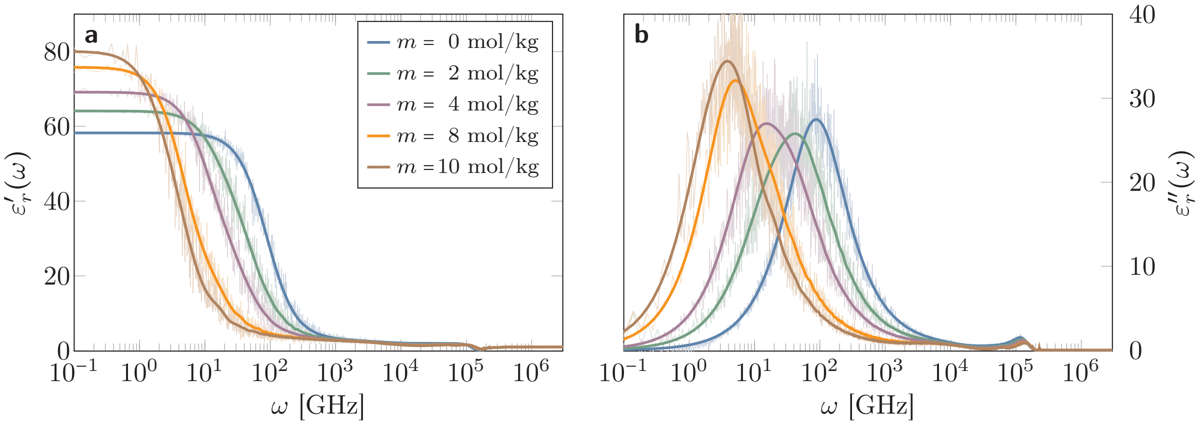


Figure A42: Dielectric spectra of aqueous TMAO solutions at different TMAO molalities for the Kast 2016 model with TIP4P/2005 water. **a**: permittivity $\varepsilon_r'(\omega)$. **b**: loss $\varepsilon_r''(\omega)$.

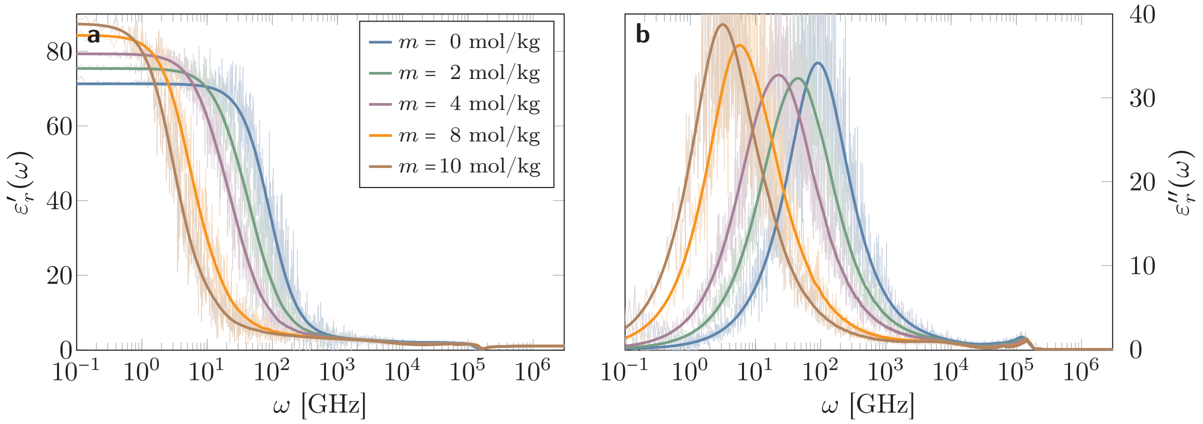


Figure A43: Dielectric spectra of aqueous TMAO solutions at different TMAO molalities for the Netz model with SPC/E water. **a**: permittivity $\varepsilon_r'(\omega)$. **b**: loss $\varepsilon_r''(\omega)$.

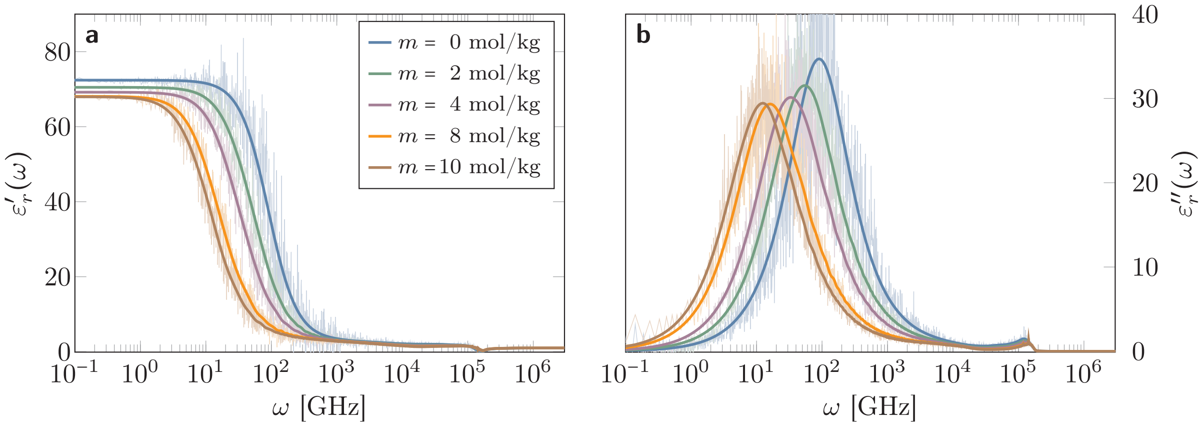


Figure A44: Dielectric spectra of aqueous TMAO solutions at different TMAO molalities for the Shea model with SPC/E water. **a**: permittivity $\varepsilon_r'(\omega)$. **b**: loss $\varepsilon_r''(\omega)$.

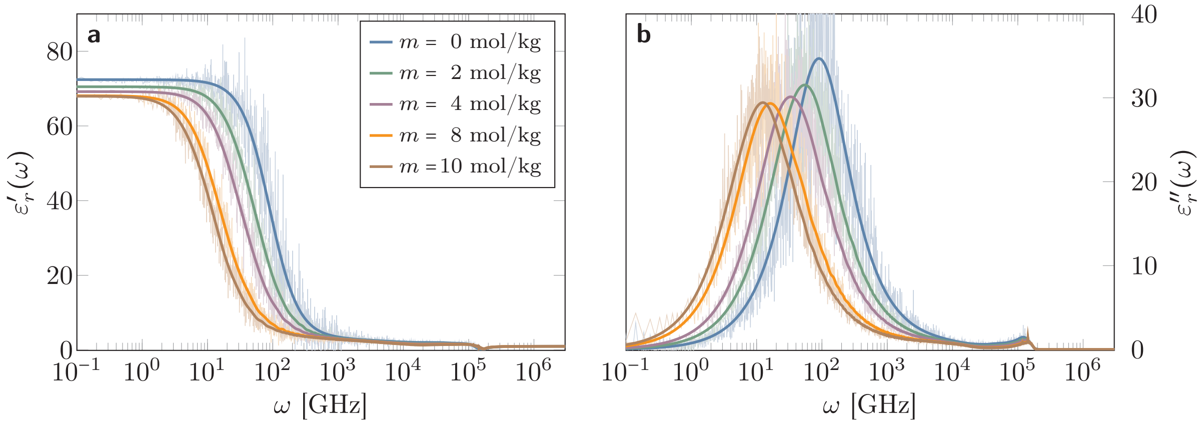


Figure A45: Dielectric spectra of aqueous TMAO solutions at different TMAO molalities for the united-atom model with SPC/E water. **a**: permittivity $\varepsilon_r'(\omega)$. **b**: loss $\varepsilon_r''(\omega)$.

A.6.6 Static permittivities

The static dielectric permittivities ϵ_{static} determined by the method described in Sec. A.6.3 are listed in Table A9. The linear regression of $\left\langle \Delta \boldsymbol{M}_{\text{tot}}^2(t) \right\rangle$ obtained from the evaluation of Eq. (A.4) has been performed on an interval ranging from 1 ns to 20 ns for all systems and TMAO concentrations.

Table A9: Static dielectric permittivities ϵ_{static} of the investigated systems for different TMAO molalities m.

TMAO force field	Water model	m [mol/kg]	ϵ_{static}
Garcia	TIP3P	0	99.43
		2	100.72
		4	102.20
		8	107.01
		10	110.27
Kast 2016	SPC/E	0	72.05
		2	74.20
		4	75.59
		8	79.01
		10	81.22
Kast 2016	TIP4P/2005	0	58.19
		2	64.03
		4	69.19
		8	75.24
		10	79.80
Netz	SPC/E	0	71.76
		2	75.55
		4	79.22
		8	83.97
		10	87.71
Shea	SPC/E	0	72.44
		2	70.67
		4	69.35
		8	69.68
		10	69.18
United-atom	SPC/E	0	72.16
		2	70.39
		4	69.28
		8	68.01
		10	67.81

Appendix B

Overcoming Convergence Issues in Free-Energy Calculations of Amide-to-Ester Mutations in the Pin1-WW Domain

B.1 EDS simulations

Enveloping distribution sampling (EDS) simulations[228,229] were conducted in the present work with the aim to remove a sampling issue related to the use of a dual topology in previous work[100], in which two copies of the side chain are present that do not interact with each other but with the environment. In many cases, these two copies (one belonging to the native and one to the ester end state) sampled different conformations at a particular time step leading to strong solute-solvent overlap and strong modifications of the potential energy landscape through the need to use a low EDS smoothness parameter see Table S1 in Ref. 100). Through the application of distance restraints that keep the two copies of the side chain within the same conformation (see Table B1), this sampling issue can be removed allowing an EDS smoothness parameter of unity to be used for all cases. The sampling quality is improved considerably as demonstrated in Figures B1 to B3 for the case of tri- to heptapeptides (see previous work[100] for computational details and Figures S4 to S9 therein for an analysis of sampling quality). However, as is discussed in the main text of the present work, the use of distance restraints does not remedy the observed starting structure dependence and might even hamper the sampling of conformational transitions in the backbone. Therefore, we resorted to a single topology approach in the present work combined with an enhanced sampling method.

Table B1: Overview of applied distance restraints between corresponding side chain atoms of the two non-interacting side chain branches (dual topology approach) in the EDS simulations. A force constant of 500 kJ mol^{-1} nm^{-2} and a zero reference distance was applied for the harmonic restraining potential.

Mutation	Residue Type [a]	Restrained Atoms [a]
L7λ	LEU	-
W11ω	TRP	CD1, CE3, CZ2
E12ϵ	GLU	CB, CG, CD
K13κ	LYSH	CE
R14ρ	ARG	CD, NH1, NH2
M15μ	MET	CG, CE
S16σ	SER	-
R17ρ	ARG	CD, NH1, NH2
S19σ	SER	-
V22ϖ	VAL	-
Y23ψ	TYR	CG, CE1, CE2
Y24ψ	TYR	CG, CE1, CE2
F25ϕ	PHE	CG, CE1, CE2
N26ν	ASN	CG
H27η	HISA/HISB	CG, CD2, CE1
N30ν	ASN	CG
A31α	ALA	-
S32σ	SER	-
Q33θ	GLN	CB, CG, CD
W34ω	TRP	CD1, CE3, CZ2

[a] Naming convention of residue types (protonation states) and atoms according to the GROMOS force field.

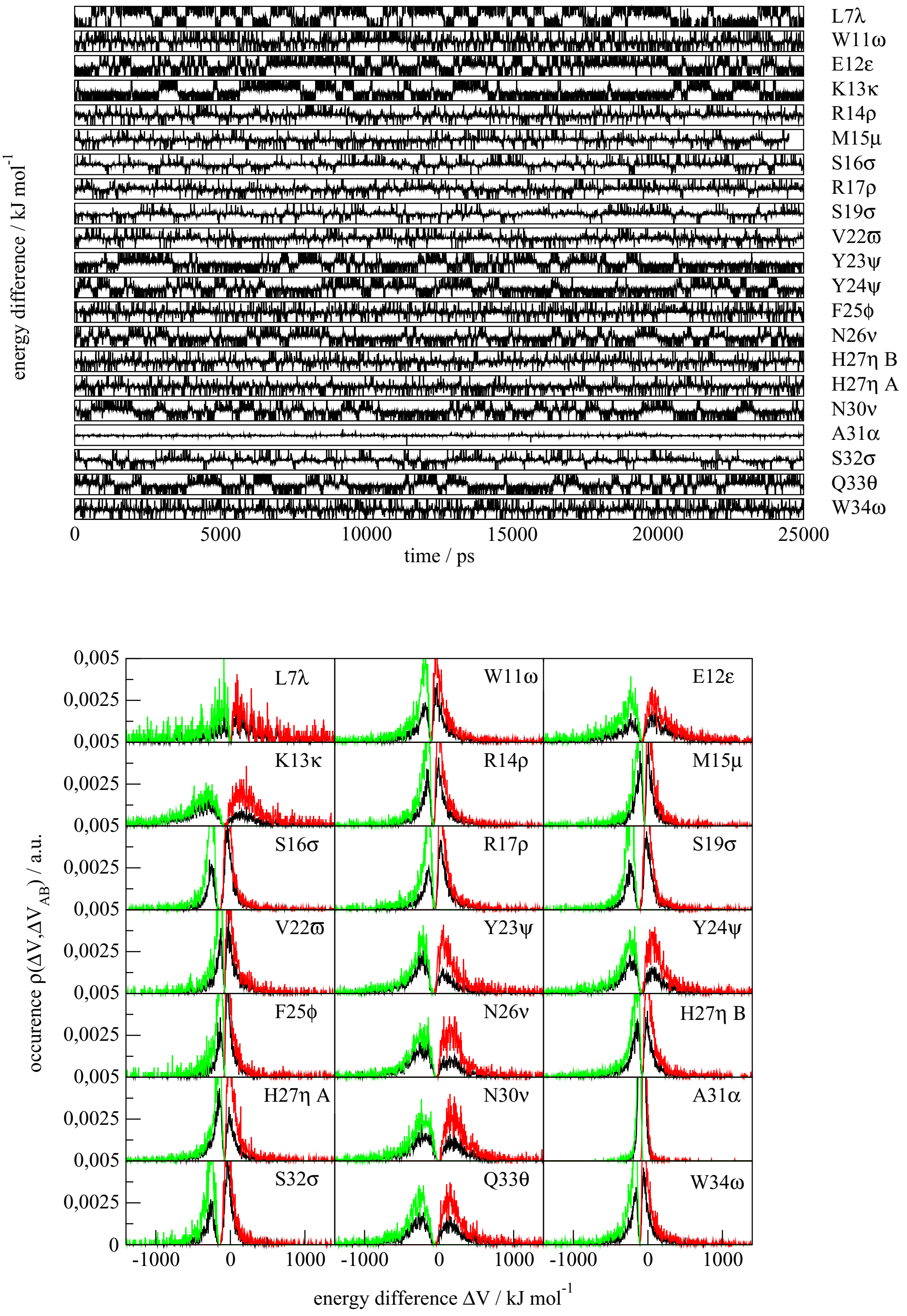


Figure B1: Upper panel: Time series of the potential energy differences, sampled from the EDS reference state simulations in case of tripeptides. The y-axes for all rows cover a range from -2000 to +2000 kJ mol^{-1}. For improved sampling, distance restraints between atoms of the two non-interacting copies of the side chains (dual topology approach) were applied, as specified in Tab. B1. Lower panel: Corresponding energy difference distributions for the reference state (black), the amide (green) and the ester state (red). Distributions of the two end states were obtained from reweighting of the reference state distribution[100].

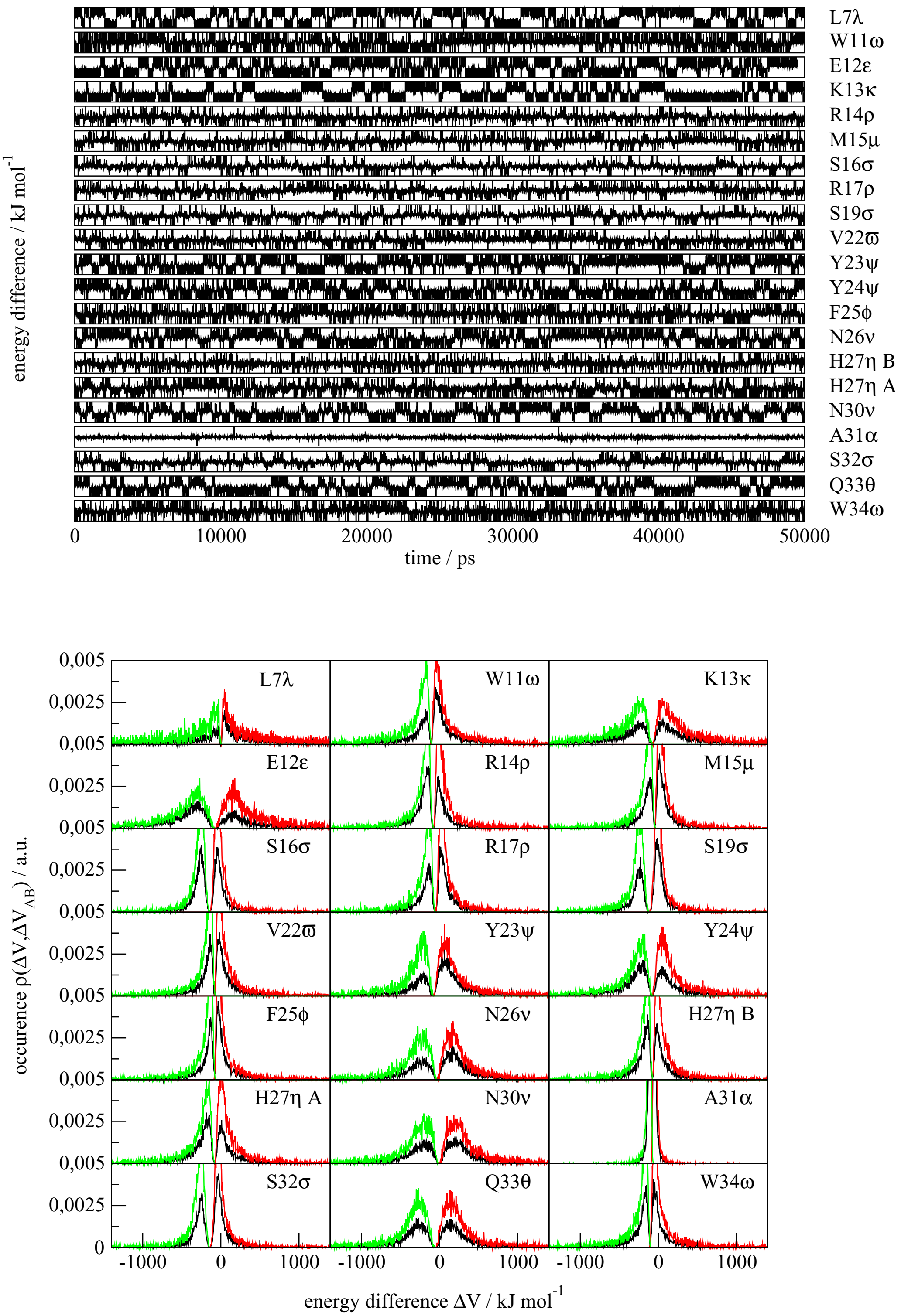


Figure B2: Upper panel: Time series of the potential energy differences, sampled from the EDS reference state simulations in case of pentapeptides. The y-axes for all rows cover a range from -2000 to +2000 kJ mol^{-1}. For improved sampling, distance restraints between atoms of the two non-interacting copies of the side chains (dual topology approach) were applied, as specified in Tab. B1. Lower panel: Corresponding energy difference distributions for the reference state (black), the amide (green) and the ester state (red). Distributions of the two end states were obtained from reweighting of the reference state distribution[100].

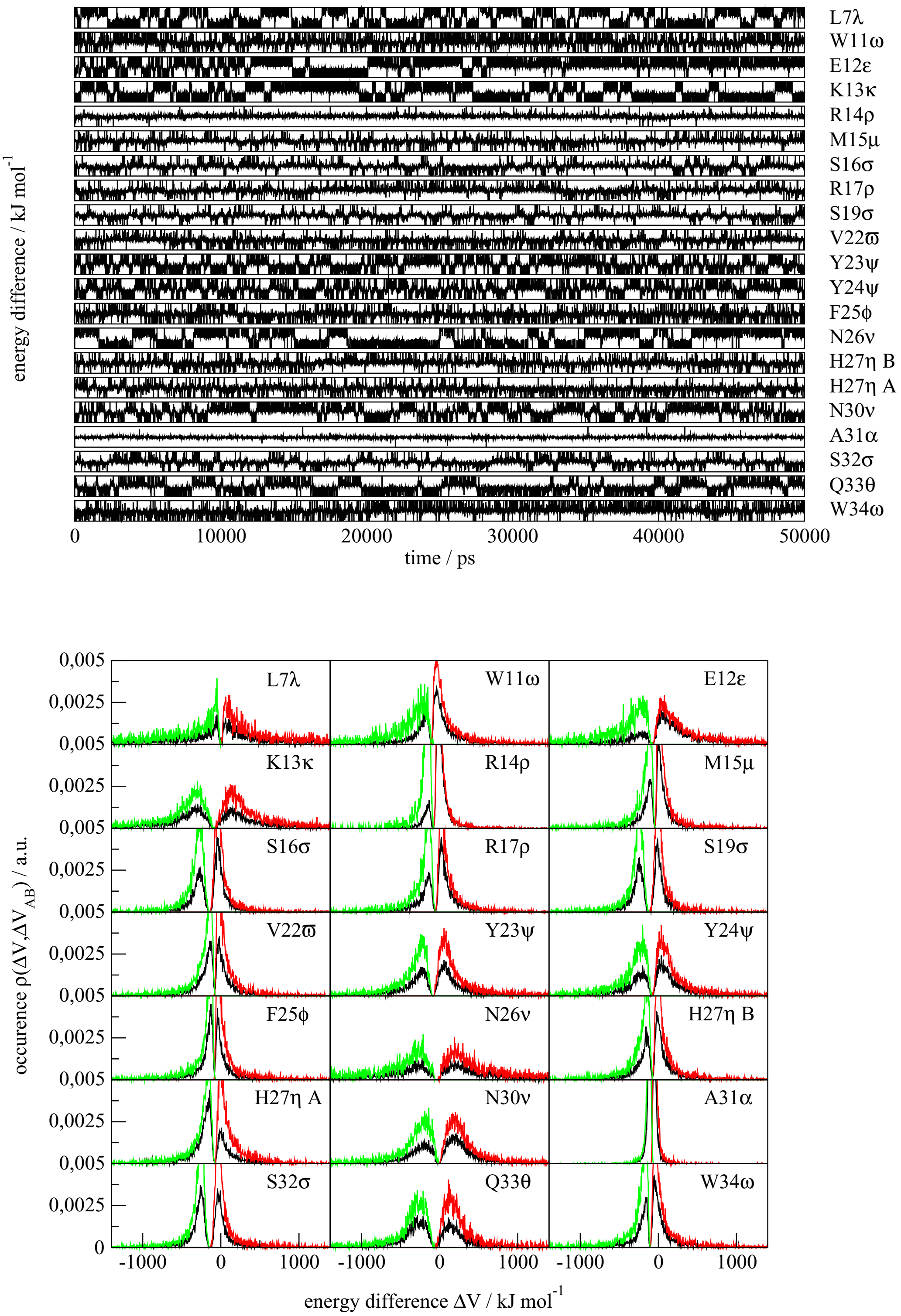


Figure B3: Upper panel: Time series of the potential energy differences, sampled from the EDS reference state simulations in case of heptapeptides. The y-axes for all rows cover a range from -2000 to +2000 kJ mol^{-1}. For improved sampling, distance restraints between atoms of the two non-interacting copies of the side chains (dual topology approach) were applied, as specified in Tab. B1. Lower panel: Corresponding energy difference distributions for the reference state (black), the amide (green) and the ester state (red). Distributions of the two end states were obtained from reweighting of the reference state distribution[100].

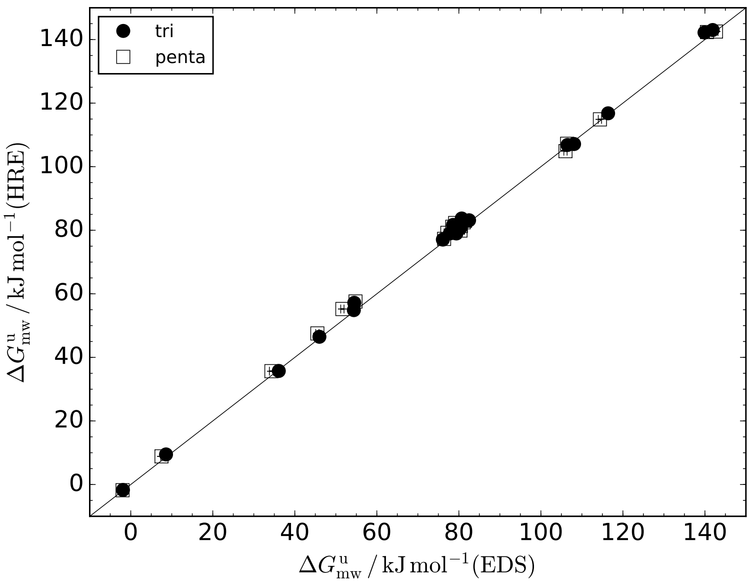


Figure B4: Agreement between alchemical free-energy differences for the unfolded states ($\Delta G^{\mathrm{u}}_{\mathrm{mw}}$), as obtained from EDS simulations and the combination of stratification and Hamiltonian replica exchange (HRE) simulations as used in the present study. The two data sets correspond to different peptide-lengths as used to approximate the unfolded state (tri-, pentapeptides). The solid line is intended as guide to the eye along $\Delta G^{\mathrm{u}}_{\mathrm{mw}}(\mathrm{EDS}) = \Delta G^{\mathrm{u}}_{\mathrm{mw}}(\mathrm{HRE})$.

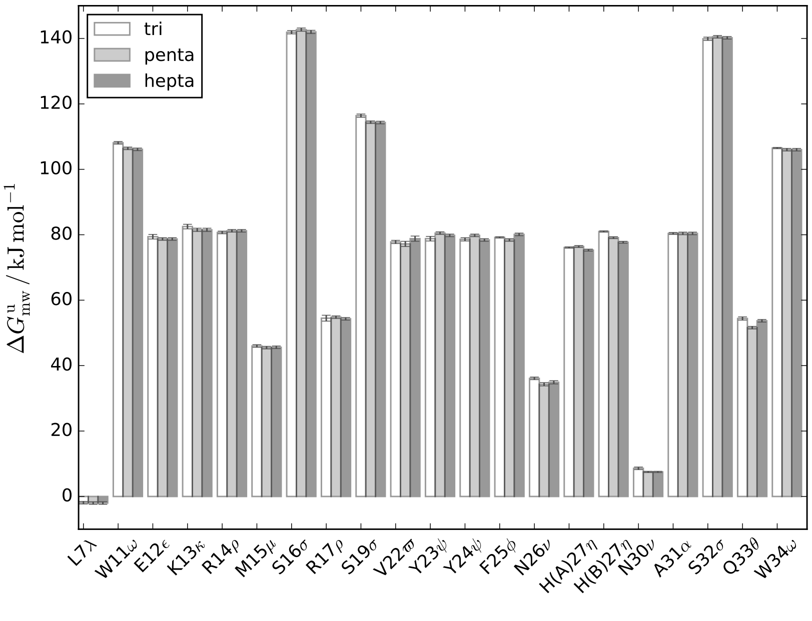


Figure B5: Dependence of the alchemical free-energy change ($\Delta G^{\mathrm{u}}_{\mathrm{mw}}$) obtained from EDS simulations on the type of the perturbed residue and on the length of the peptide as used to approximate the unfolded state (tri-, penta-, heptapeptides).

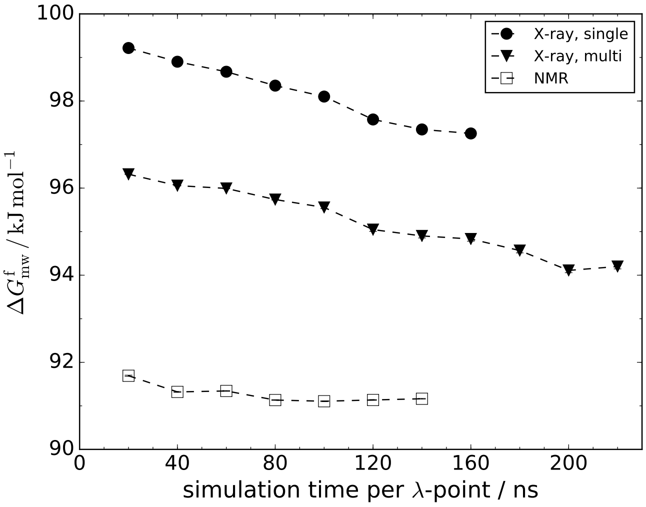


Figure B6: Alchemical free energy difference $\Delta G_{\mathrm{mw}}^{\mathrm{f}}$ of the mutation Y24ψ in the folded state, as function of the simulation time per λ-point. The different data sets correspond to different folded state starting structures (X-ray structure, NMR set). In case of the X-ray structure, further comparison is made between (i) simulations, where all λ-points are initiated from a single structure (single) and (ii) simulations, where each λ-point is initiated with a slightly perturbed structure from a synthetically generated conformational set (multi).

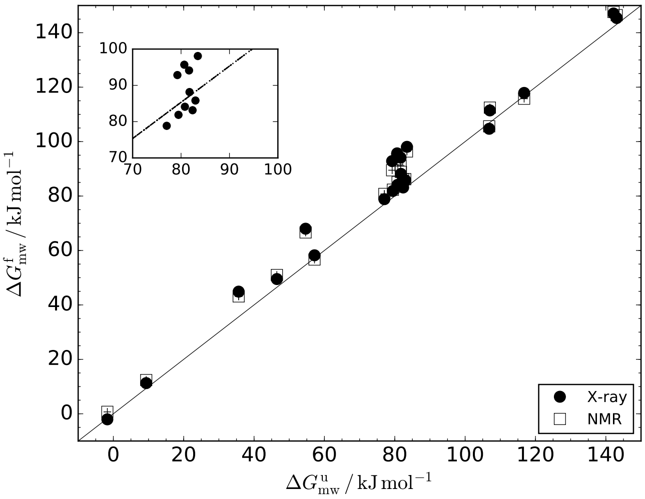


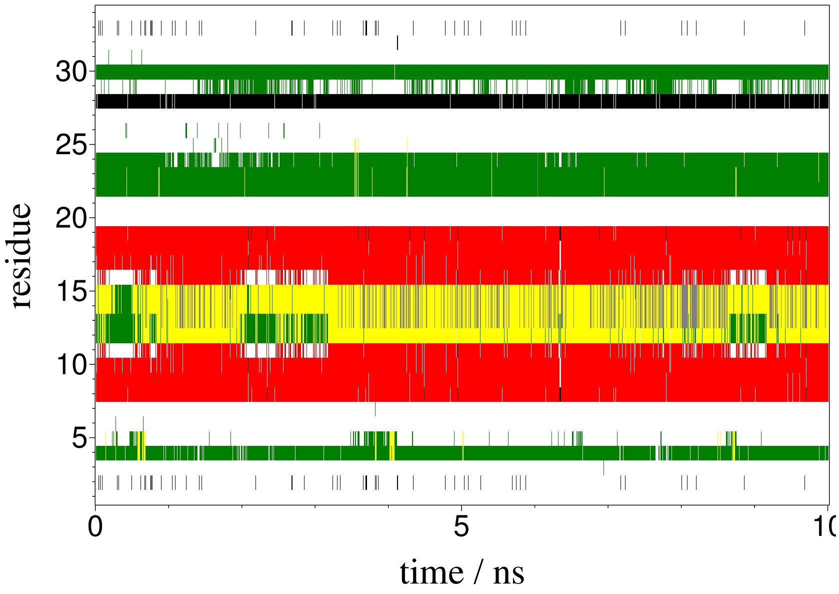
Figure B7: Correlation between alchemical free energy differences in the unfolded stated ($\Delta G_{\mathrm{mw}}^{\mathrm{u}}$ from tripeptide simulations) and the folded state ($\Delta G_{\mathrm{mw}}^{\mathrm{f}}$). The two data sets correspond to simulations based on different folded state starting structures (X-ray, NMR). The solid line is intended as guide to the eye along $\Delta G_{\mathrm{mw}}^{\mathrm{u}} = \Delta G_{\mathrm{mw}}^{\mathrm{f}}$. The inset shows a closer look at the highly populated medium-free energy region in case of the X-ray data set. The dashed line corresponds to a linear-least squares fit to the complete X-ray data set.

Table B2: Results from a thermodynamic analysis, conducted for tripeptides of a selected subset of A-to-E mutations, in terms of enthalpic ($\Delta H^{\mathrm{u}}_{\mathrm{mw}}$), entropic ($T_0\,\Delta S^{\mathrm{u}}_{\mathrm{mw}}$ with $T_0 = 278$ K) and free energy differences ($\Delta G^{\mathrm{u}}_{\mathrm{mw}}$).

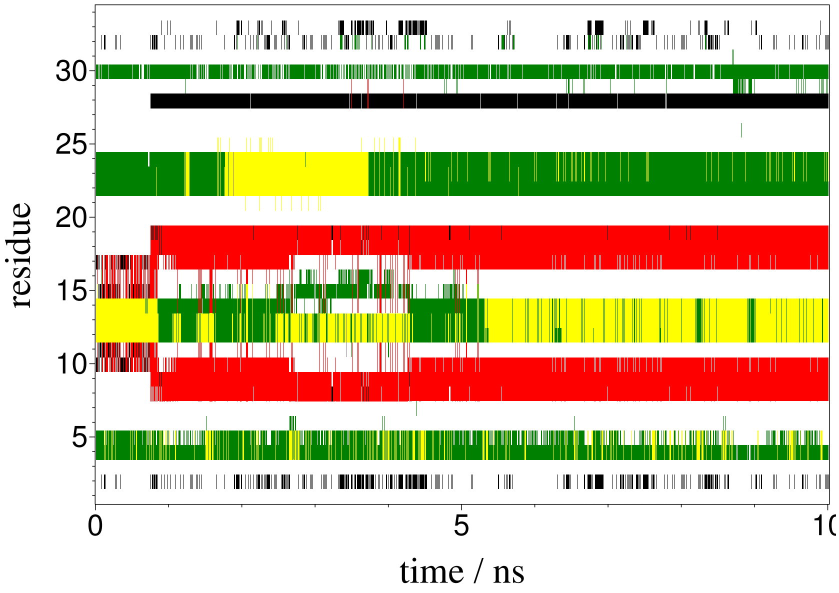
Mutation	$\Delta G^{\mathrm{u}}_{\mathrm{mw}}$ [kJ mol^{-1}]				$\Delta H^{\mathrm{u}}_{\mathrm{mw}}$	$T_0\Delta S^{\mathrm{u}}_{\mathrm{mw}}$
	278 K	298 K	318 K	338 K	[kJ mol^{-1}]	[kJ mol^{-1}]
W11ω	107.1	106.9	106.9	106.6	109.2	2.1
E12ϵ	79.0	79.5	78.9	78.6	81.0	2.0
N26ν	35.7	35.5	35.1	34.7	40.4	4.7
N30ν	9.5	9.4	9.0	9.0	12.0	2.5
Y23ψ	81.6	81.6	81.2	80.9	85.1	3.6
Y24ψ	81.7	80.7	80.6	80.3	87.7	6.0
S19σ	116.8	116.6	116.3	116.3	119.2	2.4
S32σ	142.2	141.8	142.0	141.8	143.7	1.4

Table B3: Occurrence (%) of the 11 hydrogen bonds (see Fig. 3.1 of the main text) for the protein wild-type within 10 ns MD simulations (isobaric, isothermal), initiated with the X-ray structure[223] and all 20 conformers (C1 to C20) of the NMR model set[234].

H-bond No.	Pairing residues	X-ray	NMR																			
	D→A		C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
1	W11 → P8	62.52	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	1.70	0.00	0.00	0.00	0.00	0.00
2	E12 → F25	97.80	95.35	94.40	95.95	96.40	94.70	93.05	93.65	90.05	95.30	95.85	75.96	97.20	90.50	0.00	0.00	91.85	32.98	95.85	73.01	93.15
3	F25 → E12	96.55	97.55	96.20	95.60	95.95	95.55	96.35	95.60	94.40	95.55	96.55	97.75	96.90	98.10	83.36	77.91	92.55	67.17	97.35	94.85	95.60
4	R14 → Y23	97.10	95.75	96.40	96.60	95.85	97.15	96.75	96.05	98.40	94.40	96.00	95.65	92.30	96.70	99.25	92.05	95.70	97.75	96.25	95.40	81.46
5	Y23 → R14	99.15	98.90	98.75	98.75	99.10	98.70	98.05	97.10	95.00	99.40	98.95	99.05	96.40	98.10	96.75	88.91	92.65	96.00	97.95	98.95	56.17
6	S16 → R21	89.56	83.81	72.01	82.01	90.90	89.31	75.41	74.41	26.84	83.81	82.91	77.41	70.06	87.36	89.91	68.27	63.27	88.91	91.10	91.20	26.59
7	S19 → S16	1.65	0.00	9.05	4.10	0.80	2.05	3.30	3.55	14.44	1.70	0.00	0.95	3.50	0.10	0.00	13.14	0.45	4.20	0.00	0.50	0.00
8	N30 → N26	0.10	1.75	7.90	12.99	3.90	0.00	0.00	0.05	0.00	0.00	4.00	0.00	0.00	0.30	0.05	18.74	0.00	0.00	2.45	0.00	2.35
9	N26 → A31	88.36	93.65	95.55	94.55	93.40	89.61	82.46	85.11	88.46	91.05	86.51	86.96	82.81	75.71	0.00	0.00	87.26	28.29	87.06	58.42	88.76
10	Q33 → Y24	96.70	97.55	96.00	93.50	96.80	95.40	97.75	94.55	95.65	93.95	93.35	87.41	97.55	95.40	76.16	76.86	97.40	39.33	95.65	92.65	97.95
11	Y24 → Q33	97.40	93.05	91.60	94.40	95.25	97.25	88.26	85.61	95.10	95.60	95.30	91.95	93.65	87.36	96.40	89.26	94.45	93.10	89.46	94.85	92.70



(a)



(b)

Figure B8: Time evolution of secondary structure elements, obtained from 10 ns MD simulations (isobaric, isothermal) of the protein wild-type, initiated from two conformers of the NMR model set ((a): conformer C14, (b): conformer C15). Every residue is assigned to a particular secondary structure element according to the DSSP algorithm[326]: coil (white), β-sheet (red), β-bridge (black), bend (green), turn (yellow).

Appendix C

Lessons Learned from the Calculation of One-Dimensional Potentials of Mean Force

C.1 Double Decoupling

According to the double decoupling method (DDM), the calculation of the standard binding free enthalpy reads as[254]:

$$\Delta G^{\circ}_{\mathrm{bind}} = \Delta G^{\mathrm{L}\rightarrow\mathrm{D}}_{\mathrm{u}} - \Delta G^{\mathrm{L}}_{\mathrm{b,0}\rightarrow\mathrm{R}} - \Delta G^{\mathrm{L}\rightarrow\mathrm{D}}_{\mathrm{b,R}} - \Delta G^{\mathrm{D}}_{\mathrm{b,R}\rightarrow 0} \tag{C.1}$$

$\Delta G^{\mathrm{L}\rightarrow\mathrm{D}}_{\mathrm{u}}$ refers to the free energy contribution for transforming the fully interacting unbound ligand (L) into its ideal gas or decoupled (D) state. $\Delta G^{\mathrm{L}\rightarrow\mathrm{D}}_{\mathrm{b,R}}$ represents the analogue contribution for the ligand bound to the CNT host. To prevent drifting of the decoupled ligand, an auxiliary translational (and possibly orientational) restraint (R) has to be applied. The translational and orientational restraints were implemented as harmonic potentials acting on the host-ligand COM-COM radial distance and the orientational angle θ, respectively. $\Delta G^{\mathrm{L}}_{\mathrm{b,0}\rightarrow\mathrm{R}}$ and $\Delta G^{\mathrm{D}}_{\mathrm{b,R}\rightarrow 0}$ refer to the contributions due to application and release of the auxiliary restraints for the fully interacting and decoupled ligand in the bound state, respectively. Decoupling of the ligand from the bulk solvent and host was conducted in a sequence of 20 discrete steps as controlled by the coupling parameter λ, equally distributed between $\lambda = 0$ (fully interacting) and $\lambda = 1$ (decoupled state). It should be stressed that since DDM was applied for systems unpolar CNT / unpolar ligand, the scaling with λ solely affects the dispersion interactions with the environment. Activation of the translational restraint in case of the fully interacting bound ligand was performed in 11 distinct simulations using uniformly increasing values for the force constant between 0 and 500 kJ mol^{-1} nm^{-2}. In case of an additional orientational restraint, it was activated simultaneously with the translational restraint using uniformly increasing values for the force constant between 0 and 500 kJ mol^{-1} rad^{-2}. The MBAR free energy estimator was used in all cases. The contribution $\Delta G^{\mathrm{D}}_{\mathrm{b,R}\to 0}$ was calculated analytically according to[284]:

$$\Delta G^{\mathrm{D}}_{\mathrm{b,R}\to 0} = -RT \ln \left(\frac{V^{\circ}\, 8\pi^2}{V_{\mathrm{tr}}\, \Omega} \right) \tag{C.2}$$

with the accessible translational and rotational volumes of

$$V_{\mathrm{tr}} = \left(\frac{2\pi RT}{k_{\mathrm{tr}}} \right)^{\frac{3}{2}} \tag{C.3}$$

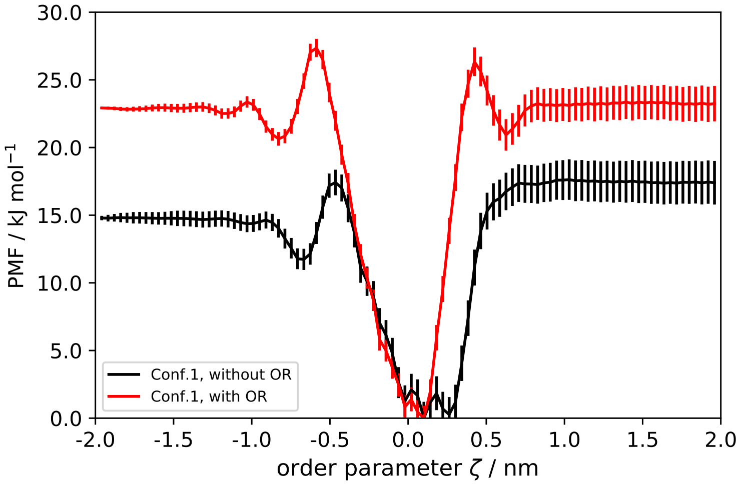
$$\frac{\Omega}{8\pi^2} = \frac{1}{2} \int_0^{\pi} \mathrm{e}^{-U_\theta(\theta)/RT} \sin\theta \, \mathrm{d}\theta \tag{C.4}$$

In case of a harmonic potential $U_\theta(\theta)$ according to Eq. (4.11), the rotational volume V_{rot} was calculated numerically while it reduces to unity in the absence of an orientational restraint. Calculated binding free enthalpies from DDM for systems unpolar ligand / unpolar CNT are summarized in Tab. C1. In all simulations, long-range electrostatics were treated with the particle-mesh Ewald (PME) method.

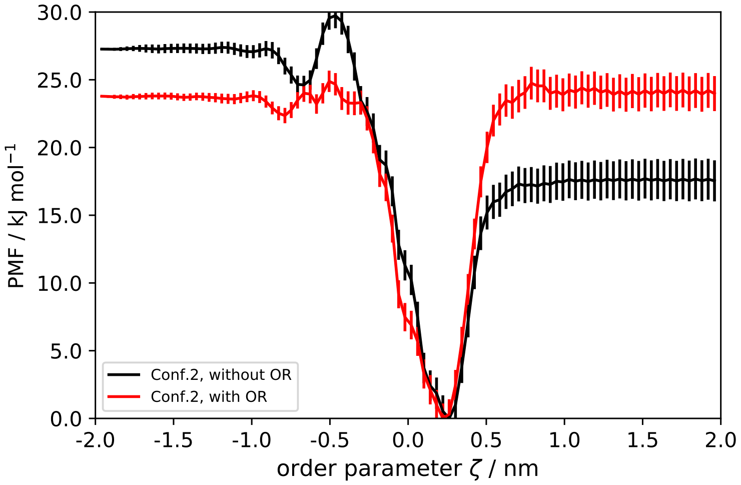
Table C1: Calculated standard binding free enthalpies $\Delta G^{\circ}_{\mathrm{bind}}$ from double decoupling for unpolar methane, ethane and elongated ethane binding to unpolar CNT. For each ethane model two data sets are presented, corresponding to the setup with and without orientational restraint (OR). Detailed description of the double decoupling approach can be found in the appendix. $\Delta G^{\mathrm{L}\to\mathrm{D}}_{\mathrm{u}}$, $\Delta G^{\mathrm{L}\to\mathrm{D}}_{\mathrm{b,R}}$ and $\Delta G^{\mathrm{L}}_{\mathrm{b,0}\to\mathrm{R}}$ were calculated using the MBAR estimator. The contribution for removing the restraints from the decoupled ligand ($\Delta G^{\mathrm{D}}_{\mathrm{b,R}\to 0}$) was calculated analytically according to Eq. (C.2). The estimate of $\Delta G^{\circ}_{\mathrm{bind,Conf.1}}$ as obtained from the setup including an orientational restraint corresponds to one distinct binding configuration and has to be corrected by an entropic symmetry term of $-RT\ln 2$[284,308] to obtain $\Delta G^{\circ}_{\mathrm{bind}}$ in case of elongated ethane. Estimates for statistical uncertainties of $\Delta G^{\circ}_{\mathrm{bind}}$ as obtained from application of standard error propagation to Eq. C.1 are below 0.5 kJ mol^{-1} where the statistical uncertainties of the individual free energy terms are delivered by the MBAR estimator[237].

System	Setup	$\Delta G^{\mathrm{L}\to\mathrm{D}}_{\mathrm{u}}$ [kJ mol^{-1}]	$\Delta G^{\mathrm{L}\to\mathrm{D}}_{\mathrm{b,R}}$ [kJ mol^{-1}]	$\Delta G^{\mathrm{L}}_{\mathrm{b,0}\to\mathrm{R}}$ [kJ mol^{-1}]	$\Delta G^{\mathrm{D}}_{\mathrm{b,R}\to 0}$ [kJ mol^{-1}]	$\Delta G^{\circ}_{\mathrm{bind,Conf.1}}$ [kJ mol^{-1}]	$\Delta G^{\circ}_{\mathrm{bind}}$ [kJ mol^{-1}]
Methane		-9.60	13.78	3.86	-14.22	-	-13.00
Ethane	No OR	-7.37	30.82	2.38	-14.22	-	-26.35
	OR	-7.37	31.99	17.03	-29.15	-27.23	-27.23
Long Ethane	No OR	-22.67	14.32	2.32	-14.22	-	-25.09
	OR	-22.67	22.23	8.31	-29.15	-24.06	-25.79

C.2 Simulations with All-Atom Force Field



(a)



(b)

Figure C1: PMFs for 1-butanol binding to αCD in Conf. 1 (a) and Conf. 2 (b). Red and black profiles refer to the setup with and without imposed orientational restraint (OR), respectively. Simulations are based on the CHARMM36 all-atom force field.