Simulations of amphiphilic peptide assembly at the air-water interface

Dissertation
zur Erlangung des Grades

"Doktor der Naturwissenschaften"

im Promotionsfach Chemie

am Fachbereich Chemie, Pharmazie und Geowissenschaften
der Johannes Gutenberg-Universität
in Mainz

Mara Nikola Jochum

geb. in Ludwigshafen am Rhein

Mainz, den 22.11.12

Dekan:
1. Berichterstatter:
2. Berichterstatter:
Tag der mündlichen Prüfung: 22.1.13

D77 - Dissertation der Universität Mainz

To my parents

Zusammenfassung

Amphiphile Peptide, Pro-Glu-(Phe-Glu)$_n$-Pro, Pro-Asp-(Phe-Asp)$_n$-Pro, und Phe-Glu-(Phe-Glu)$_n$-Phe, können so aus n alternierenden Sequenzen von hydrophoben und hydrophilen Aminosäuren konstruiert werden, dass sie sich in Monolagen an der Luft-Wasser Grenzfläche anordnen. In biologischen Systemen können Strukturen an der organisch-wässrigen Grenzfläche als Matrix für die Kristallisation von Hydroxyapatit dienen, ein Vorgang der für die Behandlung von Osteoporose verwendet werden kann. In der vorliegenden Arbeit wurden Computersimulationen eingesetzt, um die Strukturen und die zugrunde liegenden Wechselwirkungen welche die Aggregation der Peptide auf mikroskopischer Ebene steuern, zu untersuchen.

Atomistische Molekulardynamik-Simulationen von einzelnen Peptidsträngen zeigen, dass sie sich leicht an der Luft-Wasser Grenzfläche anordnen und die Fähigkeit haben, sich in β-Schleifen zu falten, selbst für relativ kurze Peptidlängen ($n = 2$). Seltene Ereignisse wie diese (i.e. Konformationsänderungen) erfordern den Einsatz fortgeschrittener Sampling-Techniken. Hier wurde "Replica Exchange" Molekulardynamik verwendet um den Einfluss der Peptidsequenzen zu untersuchen. Die Simulationsergebnisse zeigten, dass Peptide mit kürzeren azidischen Seitenketten (Asp vs. Glu) gestrecktere Konformationen aufwiesen als die mit längeren Seitenketten, die in der Lage waren die Prolin-Termini zu erreichen. Darüber hinaus zeigte sich, dass die Prolin-Termini (Pro vs. Phe) notwendig sind, um eine 2D-Ordnung innerhalb der Aggregate zu erhalten. Das Peptid Pro-Asp-(Phe-Asp)$_n$-Pro, das beide dieser Eigenschaften enthält, zeigt das geordnetste Verhalten, eine geringe Verdrehung der Hauptkette, und ist in der Lage die gebildeten Aggregate durch Wasserstoffbrücken zwischen den sauren Seitenketten zu stabilisieren. Somit ist dieses Peptid am besten zur Aggregation geeignet. Dies wurde auch durch die Beurteilung der Stabilität von experimentnah-aufgesetzten Peptidaggregaten, sowie der Neigung einzelner Peptide zur Selbstorganisation von anfänglich ungeordneten Konfigurationen unterstützt.

Da atomistische Simulationen nur auf kleine Systemgrößen und relativ kurze Zeitskalen begrenzt sind, wird ein vergröbertes Modell entwickelt damit die Selbstorganisation auf einem größeren Maßstab studiert werden kann. Da die Selbstorganisation *an der Grenzfläche* von Interesse ist, wurden existierenden Vergröberungsmethoden erweitert, um nicht-gebundene Potentiale für inhomogene Systeme zu bestimmen. Die entwickelte Methode ist analog zur iterativen Boltzmann Inversion, bildet aber das Update für das Interaktionspotential basierend auf der radialen Verteilungsfunktion in einer Slab-Geometrie und den Breiten des Slabs und der Grenzfläche. Somit kann ein Kompromiss zwischen der lokalen Flüssigketsstruktur und den thermodynamischen Eigenschaften der Grenzfläche erreicht werden. Die neue Methode wurde für einen Wasser- und einen Methanol-Slab im Vakuum demonstriert, sowie für ein einzelnes Benzolmolekül an der Vakuum-Wasser Grenzfläche, eine Anwendung die von besonderer Bedeutung in der Biologie ist, in der oft das thermodynamische/Grenzflächenpolymerisations-Verhalten zusätzlich der strukturellen Eigenschaften des Systems erhalten werden müssen. Darauf basierend wurde ein vergröbertes Modell über einen Fragment-Ansatz parametrisiert und die Affinität des Peptids zur Vakuum-Wasser Grenzfläche getestet. Obwohl die einzelnen Fragmente sowohl die Struktur als auch die Wahrscheinlichkeitsverteilungen an der Grenzfläche reproduzierten, diffundierte das Peptid als Ganzes von der Grenzfläche weg. Jedoch führte eine Reparametrisierung der nicht-gebundenen Wechselwirkungen für eines der Fragmente der Hauptkette in einem *Trimer* dazu, dass das Peptid an der Grenzfläche blieb. Dies deutet darauf hin, dass die Kettenkonnektivität eine wichtige Rolle im Verhalten des Petpids an der Grenzfläche spielt.

Abstract

Amphiphilic peptides, Pro-Glu-(Phe-Glu)$_n$-Pro, Pro-Asp-(Phe-Asp)$_n$-Pro, and Phe-Glu-(Phe-Glu)$_n$-Phe, composed of n recurring sequences of alternating hydrophobic and hydrophilic amino acids can be designed such that they self-assemble into monolayers at the air-water interface. In biomimetic systems, these provide template matrices at the organic-aqueous interface to promote crystallization of hydroxyapatite, which can serve as a treatment for osteoporosis. In this work, computer simulations have been employed to investigate the structure and interactions which govern peptide self-assembly on the microscopic level.

Atomistic molecular dynamics simulations of single peptide strands show that they readily align at the air-water interface and have the ability to fold into β-hairpins, even for fairly short peptide lengths ($n = 2$). Such rare events (i.e. conformational changes) require the use of advanced sampling techniques. Here, replica exchange molecular dynamics has been used to study the conformational preferences of different peptide sequences. Simulation results revealed that peptides with shorter acidic side-chains (Asp vs. Glu) exhibit more extended conformations than those with longer side-chains which could reach the proline termini. Furthermore, studies suggest that the proline termini (Pro vs. Phe) are necessary to preserve the 2D order within the monolayer aggregates, as has been observed experimentally. The peptide Pro-Asp-(Phe-Asp)$_n$-Pro, which contains both of these features, shows the most ordered assembly, only a small twist in the backbone, and is able to stabilize the formed aggregates via hydrogen-bonding between acidic side-chains, making it the most suitable candidate for self-assembly. This has also been supported by assessing the stability of pre-assembled peptide aggregates as well as their tendency to self-assemble from initially disordered configurations.

As atomistic simulations are limited to small system sizes and relatively short simulation times, a coarse-grained model is developed to be able to study the self-assembly on a larger scale. Since self-assembly *at the interface* is of interest, existing coarse-graining (CG) methodology has been expanded to determine non-bonded potentials for inhomogeneous systems. The developed method is analogous to iterative Boltzmann inversion but constructs the update for the interaction potential based on the radial distribution function calculated in a slab geometry and the slab and interfacial widths, which allows a balance between the local liquid structure and the thermodynamic properties of the interface. The new method has been demonstrated for slabs of liquid water and methanol in vacuum, as well as a solute-solvent system of a single benzene molecule at the vacuum-water interface, which is of particular importance in biology, where the thermodynamic/interfacial behavior often needs to be included in addition to the structural properties of the system. Based on this, a CG model for the system was parametrized via a fragment-based approach and the peptide's affinity for the air-water interface has been tested. Although its individual fragments reproduced both the structure as well as the probability distributions to stay at the interface, the peptide as a whole diffused into the bulk. However, a reparametrization of the non-bonded interaction for one of the backbone beads in a *trimer* lead the CG peptide to remain at the interface. This indicates that the chain's connectivity plays an important role on the peptide's behavior at the interface.

Related publications

Parts of this thesis can also be found in:

- **Structure-based coarse-graining in liquid slabs**,
  M. Jochum, D. Andrienko, K. Kremer, and C. Peter,
  J. Chem. Phys., 137(6):064102–064102–9, 2012. (ch. 5)

- **Targeted coarse-graining using various algorithms**,
  S. Y. Mashayak, M. Jochum, K. Koschke, N. R. Aluru, K. Kremer, V. Rühle, and C. Junghans,
  [*To be submitted*] (ch. 4)

- **Structure-based coarse-graining in peptide systems**,
  M. Jochum, O. Bezkorovaynaya, K. Kremer, and C. Peter,
  [*Manuscript in preparation*] (ch. 6)

The code which implements the Downhill Simplex algorithm (ch. 4) and the modified IBI procedure for inhomogeneous systems (ch. 5) into the VOTCA CSG package (`http://www.votca.org/`) is available at:

`http://code.google.com/p/votca/`

Contents

Figures and Tables

Figures

Tables

1

Introduction

"Everything that living things do can be understood in terms of jigglings and wigglings of atoms." - Feynman [1]

Many phenomena in nature involve complex, non-equilibrium systems which undergo chemical reactions or transformations on various time and length scales. Although these biological processes often rely on sophisticated interplays of interactions or conformational changes, they continue to take place smoothly, adapting to changes in the environment, and generating optimally designed structures which in many cases are still far ahead of today's technologies. Understanding such processes does not only help scientists to discover the underlying mechanisms but can also inspire the development of new materials or pathways by artificially mimicking nature [2]. To gain fundamental insight, one needs to start by examining what is happening to a system on the atomic level, in other words, to analyse the "wigglings" and "jigglings" of individual atoms [1, 3]. Such motions are directly linked to the molecular interactions on the microscopic level which in turn affect the system's behaviour at the macroscopic level [4].

To analyse the behaviour of a biological system, experimental techniques such as X-ray crystallography, nuclear magnetic resonance (NMR) measurements, and other spectroscopic methods can be used to obtain structural information at various stages of a reaction. However, these may not always provide the full picture, as the number of degrees of freedom of a biological system (10^4-10^6 or more for a protein) largely outnumber the experimental data available for them. In addition, these measurements only provide *averages* over space and time [5]. It may also be that the conditions under which experiments have to be carried out are too difficult, costly, or in some cases even impossible to fulfil. Thus, to complement experimental results, the field of biomolecular simulation has evolved [6], attempting to use computational models to describe systems on various levels of resolution by combining the intuitive and conceptual knowledge of chemistry with the laws of physics. These models can in principle provide entire *distributions* of a measurable quantity, thereby filling in the missing gaps

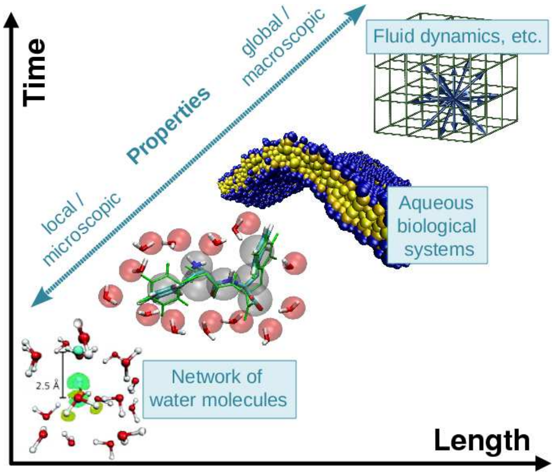


Figure 1.1: Hierarchy of length and time scales (as in [8]), depicting examples of biomolecular simulations from high (bottom left) to low levels of resolution (top right).

of the experimental picture [7].

When a new phenomenon is to be investigated, one must decide on an appropriate model for simulation which depends on the level of resolution at which the phenomenon of interest occurs (time and length scale) as well as the number of degrees of freedom one needs to describe. These can range from microscopic descriptions such as electrons on a quantum mechanical level for which electronic structure calculations are often performed, to macroscopic descriptions such as continuum models (fig. 1.1 as in [8]). In a classical atomistic description, interactions are prescribed by a force field, which consists of a set of interaction functions and parameters that determine the potential energy surface of the system. Before a new simulation model can be employed, it should be verified that the properties of interest of the experimental system can be adequately reproduced.

Modelling of an entire biological process often exceeds the range of a single model and it becomes infeasible (if not impossible) to continue on the same level of resolution, such that in practice, a set of models is usually required to cover the total range of interest [9]. For example, when studying peptide aggregation, one may want a high level of resolution to investigate the intermolecular hydrogen-bonding that takes place between individual peptide strands or with neighbouring water molecules, but one needs to resort to a lower level of resolution to be able to access the long times scales and high peptide concentrations necessary to determine the types of aggregates formed. In addition, the potential energy surfaces of such processes are often very rugged and peptides may become trapped in local energy minima, pronouncing this deficiency to be able to sample the entire phase space on an accessible time scale. Here, multiscale modelling can be employed, where advanced sampling techniques in addition to lower resolution simulations may help to understand equilibrium phenomena on experimental time scales.

In this thesis, amphiphilic peptides at the air-water interface are investigated on several levels of resolution to better understand the forces which drive their self-assembly. Initially, classical atomistic simulations and advanced sampling techniques are employed to model the system at the different stages of assembly on relatively short time intervals ($\approx$ 20-100 ns) on an atomistic level (chapter 3). Subsequently, a coarse-grained model is systematically developed to be able to reach longer time and length scales (chapters 4, 5, 6). The remainder of this introduction provides a brief motivation to this work by introducing the system from an experimental point of view, describing how it can be studied via biomolecular simulations, the experimental questions which will be addressed, and the computational methods employed to answer them.

From nature to experiment

With growing interest in the design of bio-inspired materials, many efforts have been put into solving current engineering problems by following nature's example, such as the creation of synthetic spider silk [10, 11], gecko-inspired adhesives [12], or the first artificial leaf [13]. However, going from understanding an existing biological material to designing a new synthetic material is not a trivial task. An important point to realize is that nature grows both, the material and the organism in which it functions, thereby providing a *dynamic* design strategy as opposed to a *static* one as is the case in engineering. This allows the biological material to adapt its (micro-)structure to changes in the environment, a feature which may lead to changes in the desired properties for the synthetic material. Therefore, one must have a full understanding of the *structure-function* relationship as well as the biological context in which the material exists before copying nature's design [14].

Another point to consider is which ingredients should be used to make the new material. Since many structure-function relationships of proteins as well as their assembly or folding behaviours are not sufficiently understood, researchers concentrate on using smaller biological units such as peptides, nucleic acids, or lipids as building blocks [15, 16]. Although these shorter units make it more difficult to steer the structural specificity and stability of the macromolecule, they also come with better understood design rules and reduced flexibility (e.g. lower entropy) which facilitates self-assembly and hence, the production of the new material. They are also preferred for practical reasons as they can be synthesized in large amounts and modified by decorating the peptide with functional elements that yield the desired chemical properties [17].

One way to fabricate macromolecular structures is via the *bottom-up* approach, starting with peptides as nanoscale molecular building blocks. These can be designed such that they self-assemble through weak, non-covalent bonds which, as a whole, also govern the conformation of the macromolecule. Engineering of such biomaterials is especially useful for applications in biomineralization, a process responsible for the creation of materials such as bone, dentine, enamel, and eggshells. Although vital for many organisms, its mechanisms are still poorly characterized on the molecular level and researchers continue to seek novel ways to modify and promote the growth of hydroxyapatite crystals (HA)[1] either by using peptides as additives to alter the further crystal growth or by employing pre-assembled peptide aggregates as templates to nucleate and control crystal growth [18, 19, 20].

A particularly promising group of peptides (fig. 1.2a) for the latter approach was designed by Rapaport et al. for applications in tissue engineering [21, 22, 23] and has been patented as a local injection for the treatment of osteoporosis [24, 25], a disease in which bone matrices slowly degenerate over time. These peptides are designed such that they can self-assemble into ordered β-pleated monolayers at the air-water interface (fig. 1.2b), an environment which can be considered analogous to the organic-aqueous interface [26] in a cell. These monolayers can in turn serve as scaffolds to instigate and promote HA formation upon the addition of ions to solution [27] (fig. 1.2c).

Each peptide follows the general sequence of X-Y-(Z-Y)$_n$-X, where X represents the terminus (Pro) and Y, Z denote the residues with hydrophilic, hydrophobic side-chains (Glu/Asp, Phe) respectively. Note, that throughout the text, the three letter codes Pro, Glu, Asp, and Phe are used to refer to the different amino acids (proline, glutamic acid, aspartic acid, and phenylalanine respectively) and the abbreviated names PGlu-n, PAsp-n, and PheGlu-n are used

[1] Hydroxyapatite, $Ca_5(PO_4)_3(OH)$, is a material very similar to the mineral component in bone and is thus used in experiments and modelling as a substitute.

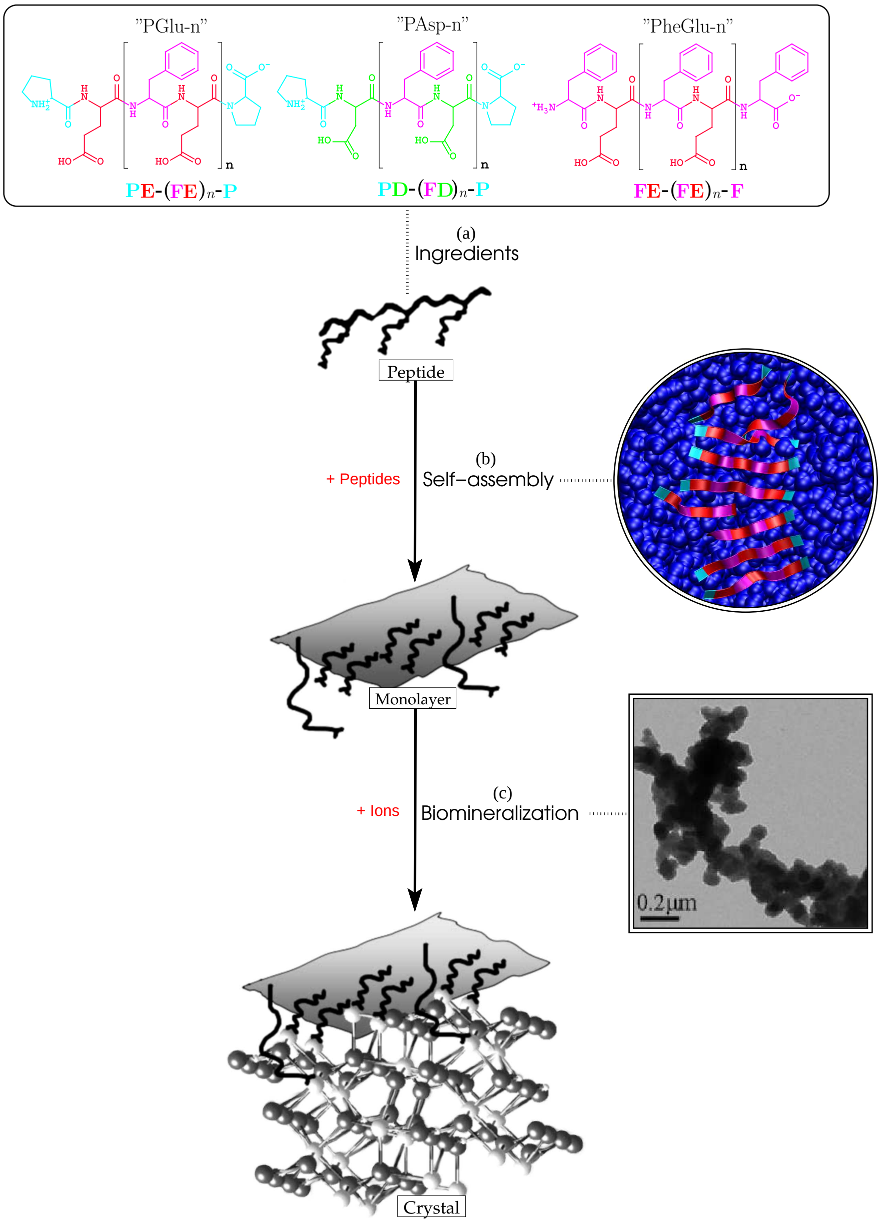


Figure 1.2: (a) Chemical structures of peptides PGlu-n, PAsp-n, and PheGlu-n as designed by Rapaport et al. (b) These self-assemble into ordered β-pleated monolayers at the air-water interface with their hydrophilic side-chains anchored in the water phase. A snapshot of an atomistic MD simulation of the self-assembly of PGlu-2 is shown on the right. (c) The formed monolayers can then serve as scaffolds to instigate and promote HA formation upon the addition of ions to solution [26]. The right side shows a transmission electron microscopy image of a calcium phosphate aggregate, which was grown on a PAsp-5 monolayer [27].

to refer to the different peptides (fig. 1.2a). By design, these peptides contain several features which make them amenable to self-assembly. First, their dual nature (amphiphilicity) allows them to align their backbone parallel to the air-water interface with the hydrophobic side-chains (Phe) pointing out of the water and the hydrophilic side-chains (Glu/Asp) pointing into the water phase, reducing the number of conformations which the peptide can adopt as compared to in a bulk water environment and essentially constricting the assembly to two dimensions. Second, the *protonated* acidic side-chains (Glu/Asp) promote self-assembly as they do not electrostatically repel each other and may also aid to stabilize the monolayer via interchain hydrogen bonding. Finally, the order of the growing monolayers is controlled by the nature of the termini. Since Pro residues are generally considered to be β-sheet breakers, employing them as end groups should guide the formation of hydrogen bonds along the backbone between parallel peptide strands, with their respective termini aligned.

Experimentally, these monolayers were prepared in Langmuir troughs by spreading a solution of 0.1 mg/mL peptide in trifluoroacetic acid/chloroform (1:9 v/v) onto a (deionized) water surface. Once the monolayer had been formed, surface pressure-area isotherms were constructed from Fourier Transform Infrared spectroscopy (FTIR) scans to obtain information about the rigidity of the monolayer, the peptides' orientation, as well as their tendency to stay at the air-water interface. In addition, Grazing-Incidence X-ray Diffraction (GIXD) was employed to get an idea about the peptides position (a, b lattice spacing) within the monolayer. Finally, the first molecular pictures proposed have been modelled and energy minimized via the CERIUS^2 computational package [21, 28].

Although experiments clearly help to understand what is happening on the macroscopic level, and to some extent also the microscopic level, computer simulations are required to confirm experimental hypotheses and shed light onto the origins of the behaviour of the system. In particular, this is important in providing information about the types of interactions that govern the overall structure of the aggregates observed. In this thesis, molecular modelling is employed on several levels of resolution to address these issues. It is probed to which extent the peptide concentration, length ($n = 2, 4, 5$), side-chain length (Glu/Asp), and the types of termini employed (Pro/Phe) have an effect on the monolayer's structure and stability. In addition, the peptides' conformations, their hydrogen bonding patterns (inter- and intrastrand), and hydrogen bonding bridges via water molecules are analysed to provide a microscopic/macroscopic model of the relevant interactions/phenomena that govern the structure formation process.

From experiment to modelling

When studying peptide aggregation, one usually tries to establish a connection between the peptide's sequence and its aggregation propensity. To do this, the conformational behaviour of a single peptide is monitored at the times of interest, for example during the transition from its soluble conformation to one that is later suited for aggregation [29]. Based on these simulations, predictions can be made, the quality of which depends on the size (and flexibility) of the peptide, with larger systems requiring more time and computer power to obtain sufficient statistics for an accurate description of the ensemble of peptide folds (i.e. minima in the free energy landscape). Here, molecular dynamics (MD) simulations can be employed. If, however, the peptide exhibits stable conformations other than its stable/rigid conformation in the β-sheet aggregate, one has to use more advanced sampling methods such as replica exchange molecular dynamics (REMD).

In the early stages of assembly where several peptides need to be simulated in concert, this sampling problem becomes even more evident and moving to a lower level of resolution is an efficient and cost-effective solution. One way to do this is by *coarse-graining*, a technique which groups several atoms into a single coarse-grained (CG) bead, thereby decreasing the number of degrees of freedom and hence the interactions which need to be computed, cutting down in computational cost. In addition, the smoother potential energy landscape of the CG model leads to reduced friction between CG beads and enables the use of larger integration time steps, resulting in a significant speed-up of the system dynamics and providing access to larger systems sizes and longer time scales [30]. Results from such simulations can be used to understand the main physical principles governing aggregation, provide insight into the underlying mechanisms, and yield results which are directly comparable to experiment.

In this thesis, the aggregation of peptides PGlu-n, PAsp-n, and PheGlu-n (fig. 1.2a) is investigated at two levels of resolution (united atom[1] and coarse-grained) to test the experimentalists' hypotheses about the aggregation behaviour of peptides as well as their picture of the underlying structure. In addition, first predictions are made about the types of interactions which govern peptide assem-

[1] In united-atom force fields, only polar hydrogen atoms are included explicitly.

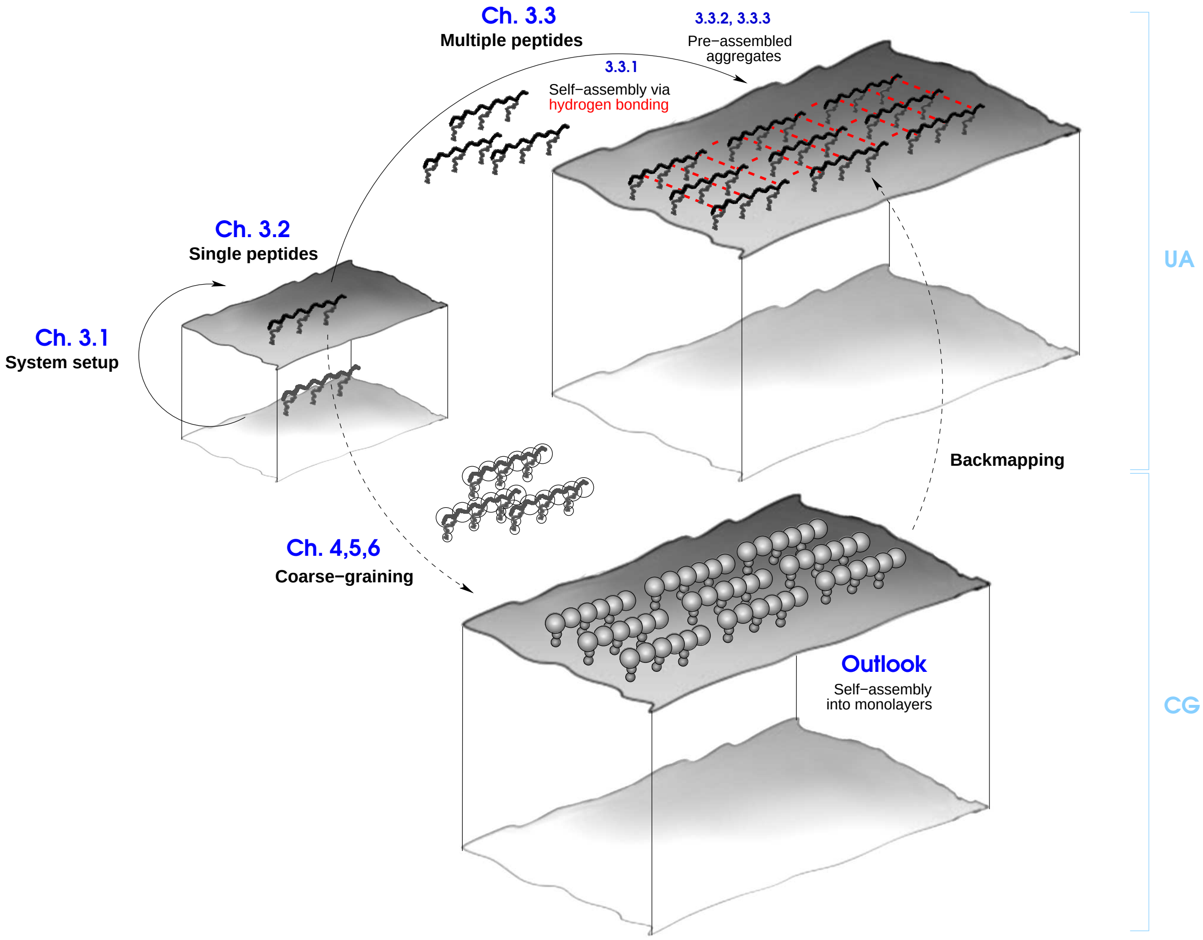


Figure 1.3: Systems overview for the multiscale study of peptide aggregation, with corresponding chapters highlighted in blue. Atomistic simulations (united-atom, UA) are performed (ch. 3), in which the system setup is first tested by checking whether peptides move to the air-water interface (sec. 3.1). Next, the conformations of single peptides at the interface are investigated (sec. 3.2) followed by simulations of multiple peptides, which are analysed during self-assembly (sec. 3.3.1) and as pre-assembled aggregates (sec. 3.3.2, 3.3.3). Finally, a coarse-grained model is developed based on a single peptide at the air-water interface (ch. 4, 5, 6), which can be used to simulate the peptide system on the coarse-grained (CG) level to reach longer time and length scales and backmapped to the UA description for the times of interest.

bly, monolayer stability, and the relationship between a peptide's sequence and its tendency to aggregate. Fig. 1.3 illustrates the steps taken towards the development of a multiscale model with which one can study the aggregation processes in these systems. After a brief introduction into the concepts of computational modelling (MD, REMD) and coarse-graining methodology (Boltzmann Inversion, Downhill Simplex algorithm, Iterative Boltzmann Inversion) in chapter 2, chapter 3 focuses solely on the results of atomistic simulations, making predictions about aggregation propensities based on the study of single peptides, multiple peptides at the early stages of aggregation, and pre-assembled aggregates which represent the final stages. In chapter 4, a solvent CG model for the air-water interface is derived using analytical pair potentials whose parameters are fitted to various system properties (radial distribution function, density profile, and surface tension) via the Downhill Simplex algorithm, with the aim of finding a compromise between reproducing structure and thermodynamic properties of the atomistic reference. In chapter 5, an alternative CG solvent model is developed using numerical pair potentials based on a novel approach which extends the Iterative Boltzmann Inversion procedure for homogeneous systems to inhomogeneous systems. The method is demonstrated on an SPC/E water slab and also tested on a slab of liquid methanol. Furthermore, the method is used to coarse-grain a single benzene molecule at the air-water interface, an example which is of particular importance as it shows how CG solute-solvent interactions, which not only account for the structure but also for the correct partitioning behaviour between two different media, can be derived. Chapter 6 then employs the proposed methodology to construct a CG peptide model via a fragment-based approach in the CG (compromise) water model derived (chapter 5). Finally, the complete CG model is tested by checking whether a peptide in simulation remains at the air-water interface and by comparing its bond and angle distributions to those of the atomistic reference. Chapter 7 outlines the conclusions which can be drawn from this study, briefly describing what can be predicted based on atomistic simulations, what needs to be studied via CG simulations, and the challenges/improvements/open questions that remain for further development in the multiscale modelling of biological systems.

2

Methodology

In materials science, the most fundamental way to describe a system stems from the field of quantum mechanics. Here, a wave function is introduced, which is a function of the coordinates of all electrons and nuclei in a system. Its time evolution can be obtained by solving the time-dependent Schrödinger equation [31].

In principle, such an accurate description can be used for any system, from a single hydrogen atom to an entire protein complex. In practise, however, it becomes infeasible to determine a many-body wave function for large systems which contain many electrons and nuclei, since computational costs scale as a higher power of the number of degrees of freedom included. Hence, simplifications must be made to be able to treat large systems in a more efficient manner.

The Born-Oppenheimer approximation, for example, relies on the fact that nuclei are much heavier than electrons and thus move much slower such that they are only considered as an external potential felt by the electrons. This allows one to employ quantum mechanical methods or density functional theory to solve the (electronic) Schrödinger equation for systems composed of up to thousands of atoms. For larger systems, however, this becomes too computationally demanding and further simplifications can be made as long as there is a clear separation between interaction energy scales. In case of a chemical reaction, where energies needed to break a chemical bond are of the order of a few eV (hundreds of $k_\mathrm{B}T$), corresponding vibrational modes are quantised and hence a quantum mechanical description is necessary to correctly describe the system. However, typical non-bonded interactions between molecules are of the order of several $k_\mathrm{B}T$ and so the density of states in the corresponding energy range can be considered continuous. In these cases, a simpler, classical mechanical description can be used [32].

For biological systems, whose temperatures range between room and body temperature ($T = 298\text{-}310\,\mathrm{K}$), typical types of interactions are hydrogen bonding, van der Waals, and screened Coulomb interactions, all of which are of the order of $k_\mathrm{B}T$. This allows one to construct an entirely classical description of the system, where quantum effects are modelled by classical formulas, for example, as constraints on the bond lengths or angles. Nuclei (atoms) interact via empirical classical potentials which are constructed from a collection of parameters and functions known as a force field, for example the AMBER [33], CHARMM [34], GROMOS [35], or the OPLS [36] force field. As such, force fields must be regarded as models which have limited accuracy in describing system properties as well as limited applicability as to which systems can be described, since the force field has been parameterized at a specific reference state point.

Consequently, the force field should always accommodate for the complexity of the system. For example, most biological force fields treat electrostatic interactions solely via a Coulombic term in the interaction potential, approximating effects such as polarizability implicitly via fixed partial charges on the atoms which overestimate molecular dipoles. In cases where the polarizabilities between the molecule and its environment differ significantly, polarizable force fields should be employed in which the standard interaction potential is extended to explicitly account for these effects. The main limitation of the force field based approach for biological systems, however, lies in the accessible time scales[1], since the time step used by the numerical molecular dynamics integrator is bounded by the typical interaction energies. Although with the approximations made, one can treat systems containing thousands of atoms, limited time scales still hinder their applicability to realistic systems significantly. In addition, the potential energy landscape of biological systems is often very complex and contains many local minima, such that one must resort to advanced sampling techniques to obtain a realistic picture of states available to a system on experimentally-relevant time scales.

The remainder of this chapter is divided into two parts which describe the simulation methodology and parametrisation techniques employed throughout this work. The first part introduces the theory behind Molecular dynamics simulations, including a more advanced sampling technique (Replica Exchange Molecular Dynamics), while the second part describes how the force fields/ interaction potentials described can be parametrised by different methods (Downhill Simplex algorithm, Boltzmann Inversion, Iterative Boltzmann Inversion).

2.1 Simulation methods

In this section, the basic simulation methodology is described, starting with Newton's laws of motion, introducing interaction potentials, the Verlet integrator,

[1] Depending on the system size, these are of the order of μs.

and briefly explaining some of the components necessary to be able to simulate a realistic system in a cost-efficient manner (periodic boundary conditions, thermostats, Ewald summation). This methodology will be used throughout this work (ch. 3-6). In addition, an advanced sampling technique known as replica exchange molecular dynamics technique is also introduced, which will be used solely for some of the atomistic simulations (ch. 3).

2.1.1 Molecular Dynamics

Molecular dynamics simulations are modelling techniques which consider particles as point masses, m_i, whose interactions are determined by potential energy terms (defined by the force field). Consider a system containing N such particles at position $\boldsymbol{r}_i$ with velocity $\dot{\boldsymbol{r}}_i$ $(i = 1, \ldots, N)$, which together make up its microscopic state. In classical mechanics, the motion of each particle satisfies Newton's equations of motion, such that its acceleration, $\ddot{\boldsymbol{r}}_i$, can be described by

$$m_i \ddot{\boldsymbol{r}}_i = \boldsymbol{f}_i \,, \tag{2.1.1}$$

where $\boldsymbol{f}_i$ represents the sum of all forces acting on the particle (both external and those arising due to interactions with neighbouring particles) [37]. It can be obtained by differentiating over the particle's potential energy, $U(\boldsymbol{r})$,

$$\boldsymbol{f}_i = -\frac{\partial U(\boldsymbol{r})}{\partial \boldsymbol{r}_i} \,, \tag{2.1.2}$$

which includes contributions from *bonded* as well as *non-bonded* interactions,

$$U = U_{\mathrm{non-bonded}} + U_{\mathrm{bonded}} \,. \tag{2.1.3}$$

Non-bonded interactions refer intermolecular interactions of pairs which are not connected via covalent bonds. The complexity of the interaction function depends on whether only 1-, 2-, or full 3-body terms should be included. In its most general form, $U_{\mathrm{non-bonded}}$ can be written as

$$U_{\mathrm{non-bonded}} = \sum_i u(\boldsymbol{r}_i) + \sum_i \sum_{j>i} v(\boldsymbol{r}_i, \boldsymbol{r}_j) + \ldots \;, \tag{2.1.4}$$

where the first term comprises all external forces on particle i and the second term includes all of the forces arising from interactions of particle i with other particles,

$$U = \overbrace{4\varepsilon\left[\left(\tfrac{\sigma}{r}\right)^{12} - \left(\tfrac{\sigma}{r}\right)^{6}\right] + \tfrac{Q_1 Q_2}{4\pi\epsilon_0 r}}^{\text{non-bonded}} + \overbrace{\frac{1}{2}\sum_{bonds} k^r_{ij}(r_{ij} - r_{eq})^2 + \frac{1}{2}\sum_{angles} k^\theta_{ijk}(\theta_{ijk} - \theta_{eq})^2 + \frac{1}{2}\sum_{torsions} k^\phi_{ijkl}(1 + \cos(m\phi_{ijkl} - \gamma_m))}^{\text{bonded}}$$

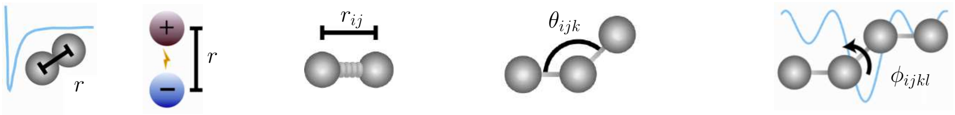


Figure 2.1: A typical interaction potential, depicting non-bonded (LJ12-6) and Coulomb) and bonded (bonds, angles, and torsions) contributions (as in [38]).

j. In the case of a pairwise potential, $v(\boldsymbol{r}_i, \boldsymbol{r}_j) = v(r_{ij})$ which is usually given by an analytical function, for example a Lennard-Jones 12-6 (LJ12-6) potential, plus a Coulombic term which takes into account all electrostatic interactions in the system.

$$v(r_{ij}) = v_{\text{LJ}}(r) + v_{\text{coulomb}}(r) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] + \frac{Q_1 Q_2}{4\pi\epsilon_0 r}\,. \tag{2.1.5}$$

Here, σ describes the van der Waals radius and ϵ is the well depth of the LJ12-6 potential. In the Coulombic term, Q_1 and Q_2 are the charges of the interaction sites and ϵ_0 is the permittivity of free space.

The second contribution to the potential energy, U (eq. 2.1.3), comes from the bonded interactions, which are intramolecular interactions of atom groups connected by covalent bonds. These are not exclusively pair interactions as they not only include bond stretching (2-body), but also angle bending (3-body) and torsional (4-body) contributions,

$$\begin{aligned} U_{bonded} = \quad & \frac{1}{2}\sum_{bonds} k^r_{ij}(r_{ij} - r_{\text{eq}})^2 \\ & + \frac{1}{2}\sum_{angles} k^\theta_{ijk}(\theta_{ijk} - \theta_{\text{eq}})^2 \\ & + \frac{1}{2}\sum_{torsions} k^\phi_{ijkl}(1 + \cos(m\phi_{ijkl} - \gamma_m))\,. \end{aligned} \tag{2.1.6}$$

Here, $r_{ij} = |\boldsymbol{r}_i - \boldsymbol{r}_j|$ is the distance between atom pairs ij, r_{eq} is the equilibrium distance, and θ_{ijk} and ϕ_{ijkl} denote the angles and torsions respectively [38]. Although the bond stretching is described by a harmonic potential here, it can also be represented by other types (i.e. LJ & FENE, Morse & FENE potential). Note, however, that most common classical computer simulations do not represent bonds by terms in the potential energy function, but treat them as having a fixed length by use of constraints via the Lagrangian of the Hamiltonian formalism [39, 40]. k_{ij}^{r}, k_{ijk}^{θ}, and k_{ijkl}^{ϕ} are the corresponding strength parameters as prescribed by the force field, m is an integer which describes the periodicity, and γ_m is the phase shift angle. Fig. 2.1 [38] gives an overview of the non-bonded and bonded interaction terms, with illustrations of their individual contributions.

The Verlet algorithm In order to propagate a system in space and time, one must integrate Newton's equations of motion, eq. 2.1.1. To achieve high accuracy at a low computational cost, several integrators have been designed. The most commonly used and computationally efficient integration algorithms are the *Verlet* algorithms [41]. These integrators have two properties important for physical systems, namely that they are symplectic integrators and that they are time-reversible. The basic Verlet algorithm can be derived by Taylor expanding the coordinate vector, $\boldsymbol{r}_i(t)$, forward and backward in time. Adding and subtracting the resulting expansions leads to

$$\begin{aligned} &\boldsymbol{r}_i(t+\delta t) = 2\boldsymbol{r}_i(t) - \boldsymbol{r}_i(t-\delta t) + \tfrac{1}{m_i}(\delta t)^2 \boldsymbol{f}_i(t) + O(\delta t^4) \\ &\boldsymbol{v}_i(t) = \tfrac{1}{2\delta t}\left[\boldsymbol{r}_i(t+\delta t) - \boldsymbol{r}_i(t-\delta t)\right] + O(\delta t^2)\ , \end{aligned} \tag{2.1.7}$$

where $\boldsymbol{v}_i(t)$ is the velocity and $\boldsymbol{f}_i(t)$ is the force obtained by substituting eq. 2.1.1 for the acceleration. Note, however, that the velocities are not directly generated since their update, $\boldsymbol{v}_i(t)$, relies on the coordinates of the previous time step, $\boldsymbol{r}_i(t-\delta t)$. A slight modification leads to the *velocity-Verlet* algorithm [42], which only relies on the velocities and positions at time t. It can be obtained upon adding the expansion of $\boldsymbol{r}_i(t)$ to its time-reversed form, which gives

$$\boldsymbol{v}_i(t+\delta t) = \boldsymbol{v}_i(t) + \tfrac{1}{2m_i}\delta t\left[\boldsymbol{f}_i(t) + \boldsymbol{f}_i(t+\delta t)\right]\ . \tag{2.1.8}$$

Mathematically equivalent to the velocity-Verlet algorithm is the *leap-frog* algorithm [43], which is used in this work. It calculates velocities at every half-time step, $\frac{\delta t}{2}$,

$$\boldsymbol{v}_i(t+\tfrac{\delta t}{2}) = \boldsymbol{v}_i(t-\tfrac{\delta t}{2}) + \tfrac{1}{2m_i}(\delta t)\left[\boldsymbol{f}_i(t-\delta t) + \boldsymbol{f}_i(t)\right]\ . \tag{2.1.9}$$

As the name suggests, $\boldsymbol{r}_i(t+\delta t)$ (here the same as in eq. 2.1.7) and $\boldsymbol{v}_i(t+\frac{\delta t}{2})$ are leaping over each other with the update of every time step.

Periodic boundary conditions In many computer simulations studies, one is interested in calculating the thermodynamic properties of a particular system. To do this accurately, one needs obtain averages for (ideally) infinitely large systems. Here, periodic boundary conditions can help to mimic such a system, without actually simulating it on such a large scale. Imagine a bulk liquid, a gas, or a macromolecule in large solvent environment where a significant amount of the particles in the system lie at the boundaries of the simulation box (unit cell). Without periodic boundary conditions, the surface effects would dominate and result in finite size effects. However, when periodic boundary conditions are employed, the simulation box is replicated in one or more (x, y, z)-directions and particles at the box boundaries are able to interact with the closest image of remaining particles in the system (minimum image convention). As a particle travels across the box boundary, the focus is shifted to the next image, such that it enters the simulation box from the opposite side, conserving the number of particles in the cell. One thus not only is able to obtain more accurate averages, but also avoids boundary effects, such as for example, a small bulk water system which could form a droplet due to surface tension without the use of periodic boundary conditions.

Thermostats Most biological systems of interest reside at a constant temperature, e.g. at body temperature of $T = 37°\mathrm{C}$. In simulations, this can be realised by employing a *thermostat*, which couples the model system to an external heat bath (a system with practically an infinite number of degrees of freedom) [44]. In addition to keeping the temperature constant, it also improves the stability of the integrator by removing energy drifts which may arise due to the accumulation of force truncation or integration errors throughout the MD simulation. Suppose a system of temperature T deviates from the reference temperature, T_0. The simplest way to correct for this temperature change, $\Delta T = T_0 - T$, is by rescaling the velocities by a factor λ at every time step, δt, such that

$$\begin{aligned} \Delta T &= \frac{1}{2}\sum_{i=1} \frac{m_i\,(\lambda v_i)^2}{N k_B} - \frac{1}{2}\sum_{i=1} \frac{m_i v_i^2}{N k_B} \\ \Delta T &= (\lambda^2 - 1)\,T(t) \\ \lambda &= \sqrt{T_0/T(t)}\;. \end{aligned} \tag{2.1.10}$$

This scaling, however, also prohibits any temperature fluctuations present in the *NVT* canonical ensemble. The *Berendsen* thermostat [45] is a weaker implementation of this approach. Here, the system is coupled to the heat bath via a coupling parameter τ, which determines how tightly they are coupled, and the velocities are rescaled according to

$$\lambda = \left[1 + \frac{\delta t}{\tau}\left\{\frac{T_0}{T\left(t - \frac{\delta t}{2}\right)} - 1\right\}\right]^{1/2}, \tag{2.1.11}$$

such that the rate of change of the temperature stays proportional to the difference in temperatures. For MD simulations of condensed phase systems, values of $\tau \approx 0.1$ ps are usually used. It is important to note that this method does not produce a canonical ensemble either, as the thermostat suppresses the fluctuations of the kinetic energy. This damping of velocity fluctuations is especially critical for simulations of small systems. For biological systems, however, where molecules in abundant solvent environments are treated, this thermostat can be assumed to suffice when studying most properties which do not rely on the systems dynamics. In addition, as this thermostat is very efficient in relaxing systems to the reference temperature, it is useful for equilibrating systems which are initially prepared far from equilibrium. Once the system has reached equilibrium and a correct canonical ensemble is required, one should either use the *velocity-rescaling* thermostat according to Bussi et al. [46], which is an extension of the Berendsen thermostat that uses an additional stochastic term for sampling a correct kinetic energy distribution, the *Nose-Hoover* thermostat [47, 48], or the *Langevin* thermostat [49], which both use the coordinates and velocities of artificial particles instead of stochastic collisions.

With motions being governed by eq. 2.1.1, the system described is isolated and hence its total energy is conserved (with periodic boundary conditions, this is the *NVE* or microcanonical ensemble). If one considers the system to be coupled to a thermostat to keep it at a constant temperature, the energy is no longer constant and one obtains the *NVT* (canonical) ensemble. Note, that as $\tau \to \infty$, the thermostat becomes inactive and the system samples a microcanonical ensemble.

Barostats Analogous to temperature coupling, pressure coupling can also be performed to simulate a system at constant pressure. This is done via a barostat, which rescales the coordinates at every step in the case of the *Berendsen* [45] barostat. Consider a system described by eq. 2.1.1. Multiplying by the coordinate vector, $\boldsymbol{r}_i$, and using the relation of the time derivative, $\frac{d}{dt}(\boldsymbol{r}_i\dot{\boldsymbol{r}}_i)$ to substitute for $\boldsymbol{r}_i\ddot{\boldsymbol{r}}_i$,

$$m_i\frac{d}{dt}\left(\boldsymbol{r}_i\dot{\boldsymbol{r}}_i - m_i\dot{\boldsymbol{r}}_i\right) = \boldsymbol{f}_i\boldsymbol{r}_i \; . \tag{2.1.12}$$

Averaging over all particles, the first term on the left-hand side disappears, the second term becomes the total energy, $-2E_{\text{kin}}$, and the term on the right-hand side becomes $\sum_{i=1}^{N}\boldsymbol{f}_i\boldsymbol{r}_i$, which is known as the virial term. Hence, eq. 2.1.12 becomes

$$-2E_{\text{kin}} = \sum\overline{\boldsymbol{f}_i\boldsymbol{r}_i} \; . \tag{2.1.13}$$

Suppose now that a wall element, *df*, in the simulation box exerts a force on the nearby particles, *pdf*, where p is the scalar pressure which can be calculated from the pressure tensor, $P_{\alpha\beta}$, as $p = \mathrm{tr}(P_{\alpha\beta})/3$. This force is known as the outer virial,

$$W_a = -p\int_V \nabla\cdot\boldsymbol{r}\,dV = -3pV \; . \tag{2.1.14}$$

The remaining force due to particle-particle interactions is called the inner virial, Ξ. Hence, the pressure due to the inner and outer virial can be computed as

$$p = \frac{2}{3V}\left(\Xi - E_{\text{kin}}\right) \; . \tag{2.1.15}$$

As the pressure in the system changes, the Berendsen barostat rescales the coordinates of the simulation box at every time step, δt, such that

$$\frac{dp}{dt} = \frac{p_0 - p}{\tau_p} \; , \tag{2.1.16}$$

where p_0 is the reference pressure and τ_p is the pressure coupling constant. Finally, a proportionate coordinate and volume scaling needs to be performed to minimise local disturbances such that the modified equation of motion becomes

$$\dot{\boldsymbol{r}}_i = \dot{\boldsymbol{r}}_i - \frac{\beta\left(p_0 - p\right)}{3\tau_p}\boldsymbol{r} \, . \tag{2.1.17}$$

Other barostats are the Anderson barostat [50] which rescales the coordinates of the system. This was later extended to the Parrinello-Rahman barostat[51] which also lets the simulation box changes its shape.

When coupled to a barostat to keep the pressure constant, the system now not only exchanges heat with the thermostat, but also volume (i.e. work) with the barostat. Although the total number of particles, the pressure, and the temperature now remain constant, the energy and volume of the system are allowed to fluctuate which yields the NpT ensemble. Note that in this work, the barostat is only employed briefly for the setup to bring the equilibrated system to a pressure of 1 atm (via a short NpT simulation) before performing simulations in the NVT ensemble.

Ewald summation When evaluating the forces and energies from the interaction potential (eq. 2.1.3), the long-ranged non-bonded contributions are the most time consuming. Their calculation scales as $\mathcal{O}(N^2)$ and hence they are especially slow to compute for systems with a large number degrees of freedom (already $\approx 10^5$ for small proteins). To improve the quadratic scaling, one can truncate the non-bonded potential at a cutoff distance, r_{cut}. Although the LJ term can be truncated easily since it is relatively short-ranged, truncating the Coulomb term leads to large inaccuracies in describing the systems electrostatics. However, the $\frac{1}{r}$ decay in the Coulomb term can be divided into two parts,

$$\begin{aligned}&\tfrac{1}{r} = \tfrac{\text{erfc}(\alpha r)}{r} + \tfrac{\text{erf}(\alpha r)}{r}\\ &\text{erf}(r) = \tfrac{2}{\pi}\int_0^r e^{-t^2}dt\\ &\text{erfc}(r) = 1 - \text{erfc}(r) = \tfrac{2}{\pi}\int_r^\infty e^{-t^2}dt\ ,\end{aligned} \tag{2.1.18}$$

where the first (short-ranged) term decays much faster than $\frac{1}{r}$ and the second (long-ranged) term decays as $\frac{1}{r}$ for large r. Realising that large r correspond to small k values in Fourier space, the latter term can be evaluated more efficiently. Hence, the Coulomb interaction potential is written as a sum of a short-ranged term, U^r, which is evaluated in real-space and a long-ranged term, U^m, which is evaluated in Fourier space, and a third term, U^0, which cancels all $i = j$

long-ranged particle interactions with itself,

$$
\begin{aligned}
U^r &= \frac{1}{2}\sum_{i,j}^{N'}\sum_{\boldsymbol{n}} Q_i Q_j \frac{\mathrm{erfc}(\alpha r_{ij,\boldsymbol{n}})}{r_{ij,\boldsymbol{n}}} \\
U^m &= \frac{1}{2\pi V}\sum_{i,j}^{N} Q_i Q_j \sum_{\boldsymbol{m}\neq\boldsymbol{0}} \frac{\exp(-(\pi\boldsymbol{m}/\alpha)^2 + 2\pi i\boldsymbol{m}\cdot(\boldsymbol{r}_i - \boldsymbol{r}_j))}{\boldsymbol{m}^2} \qquad (2.1.19) \\
U^0 &= \frac{-\alpha}{\sqrt{\pi}}\sum_{i=1}^{N} q_i^2 \ .
\end{aligned}
$$

Here, $\boldsymbol{n} = (n_1, n_2, n_3) = n_1 L\boldsymbol{x} + n_2 L\boldsymbol{y} + n_3 L\boldsymbol{z}$ is the cell coordinate vector with cell length L and the Cartesian coordinate unit vectors $\boldsymbol{x,y,z}$. $\boldsymbol{m} = (l, j, k)$ is the reciprocal space vector and V is the volume of the simulation box. Note, that the prime on the first sum in U^r indicates the omission of $i = j$ at $\boldsymbol{n} = \boldsymbol{0}$. In this work, the Particle-mesh Ewald (PME) method [52] is employed to improve the performance of the reciprocal sum, U^m. It does so by assigning all charges to a grid using a cardinal B-spline interpolation, which essentially *smears* them out over the entire lattice. The reciprocal sum can then be computed using a Fourier transform with convolutions and the forces can be obtained by analytically differentiating the potential energies at each of the grid points [53, 54]. Note, that this algorithm scales as $N\log(N)$ and is much faster than the conventional Ewald summation for medium to large system sizes [55].

2.1.2 Replica Exchange Molecular Dynamics

In some cases, the free energy landscape of a system may be too complex, with many local minima in which conformations can become trapped, such that the time scales of conventional MD simulations no longer suffice to adequately represent the equilibrium ensemble of conformations. To remedy this, the method of *replica exchange molecular dynamics* (REMD) [56] has been developed to enhance sampling.

By simulating M independent replicas of the same system at different temperatures $(T_1, T_2, \ldots, T_m)$ and allowing for exchanges between consecutive replicas, simulations which are trapped in one of the many local energy minima states at lower temperatures are able to exchange with those at higher temperatures (which usually sample large volumes of phase space), thereby overcoming high free energy barriers and sampling more of the phase space such that an equilibrium ensemble can be obtained. Here, T_{ref} denotes the reference temperature, which is the temperature of the system of interest. It is usually chosen to be the lowest temperature replica such that $T_{\text{ref}} < T_2 < \cdots < T_m$.

Consider a system at temperature T consisting of N atoms, each of position q, momentum p, and Hamiltonian $H(q,p)$. If one assumes a canonical ensemble, the Boltzmann factor or the relative probability that the system can be found in a state, $x(q,p)$, is

$$W(x;T) = e^{-\beta H(q,p)} \, , \tag{2.1.20}$$

where $\beta = 1/k_B T$. For M non-interacting replicas i in a generalised ensemble, the weighting factor for state $X = x_m^{[i]} = (q^{[i]}, p^{[i]})_m$ is given by the product of their Boltzmann factors

$$W(X) = e^{-\sum_{i=1}^{M} \beta_i H(q^{[i]},p^{[i]})} \, . \tag{2.1.21}$$

An exchange between replicas i and j must satisfy the *detailed balance* condition (ensuring that the reverse move is equally likely to occur), $W(X)w(X \to X') = W(X')w(X' \to X)$, such that the exchange process converges towards an equilibrium distribution. Hence, the exchange probability becomes a Metropolis criterion,

$$w(X \to X') = w(x_m^{[i]} | x_n^{[j]}) = \begin{cases} 1, & \Delta \leq 0 \\ e^{-\Delta}, & \Delta > 0 \, , \end{cases} \tag{2.1.22}$$

where $\Delta = (\beta_n - \beta_m) \left[U(q^{[i]}) - U(q^{[j]}) \right]$. Note, that only exchanges between replica of neighbouring temperatures are attempted, since the acceptance probability falls off exponentially with the difference in temperatures. In addition, temperature distributions should be chosen to increase exponentially, $T_i = T_{\text{ref}} e^{ik}$ where $k = \ln(T_i/T_{\text{ref}})$, and with a temperature step ($\Delta T = |T_i - T_1|$) small enough to allow for sufficient overlap between potential energy distributions in order to achieve the desired exchange probabilities [57].

2.2 Parametrisation methods

Now, that the simulation methodology has been described, several *parametrization techniques* are introduced with which one can obtain/parametrise the interaction functions described (sec. 2.1). In this chapter, they are described in a very general manner, irrespective of whether they are used to optimise atomistic force fields or CG interaction potentials. They are, however, employed to derive interaction potentials for coarse-grained (CG) models in later chapters, where the details with respect to their application to coarse-graining is given. First, the Downhill Simplex algorithm is described which is used to parametrise different CG water models via parameter fitting of analytical potentials to different structure/thermodynamic properties (sec. 4.2-4.4). Next, Boltzmann Inversion [58] is introduced which is employed to determine the internal conformational behaviour of the CG peptide (ch. 6),

and finally, Iterative Boltzmann Inversion is described which can be used to derive a purely structure-based CG water model (sec. 4.1) as well as provide initial guesses for the parametrisation of non-bonded interactions between the CG peptide and water beads, which are used in an extension of the method for inhomogeneous systems (ch. 5).

2.2.1 Downhill Simplex algorithm

The Downhill Simplex algorithm [59] is an optimisation procedure designed to minimise a function of $n+1$ variables, $y(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n+1})$. The starting simplex (fig. 2.2a) which is a geometric figure (polytope) composed of $n+1$ vertices (sets of initial guesses for minimising the function), is transformed by reflections, expansions, contractions, and reductions to search the parameter space for a set, $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n+1}$, which yields the global minimum, ε,

$$y(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n+1}) \leq \varepsilon \; . \tag{2.2.1}$$

Here, y is the function value at the current step and y^{ref} is the function value of the reference (target). At the beginning of every step, the vertices are sorted according to their function values (weights), $y_1 < y_2 < \cdots < y_{n+1}$ for vertices $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n+1}$, also known as *penalty* values, as they are a descriptor of how well a given parameter set reproduces the target property (X_{ref}), with bad parameter sets given higher weights ($y_i = |X_i - X_{\text{ref}}|^2 / X_{\text{ref}}$).

The first transformation performed is always a *reflection* (fig. 2.2b), in which the vertex corresponding to the worst point, $\boldsymbol{x}_{n+1}$, is reflected through the centroid of all remaining points, $\bar{\boldsymbol{x}}$, as

$$\boldsymbol{x}_r = (1+\alpha)\bar{\boldsymbol{x}} - \alpha \boldsymbol{x}_{n+1} \; , \; \alpha > 0 \; , \tag{2.2.2}$$

where $\boldsymbol{x}_r$ is the reflected point and α is the reflection coefficient. This transformation does not change the volume of the simplex and is used to move through

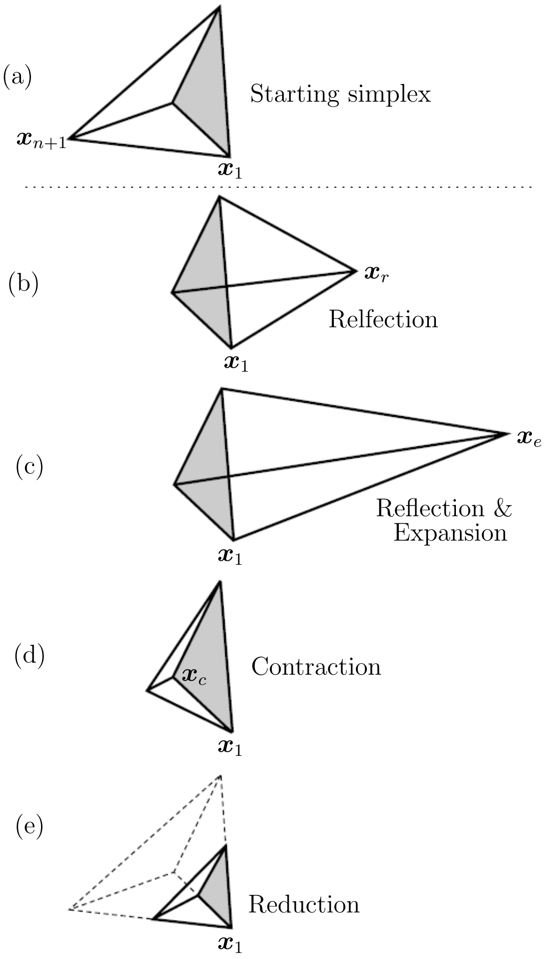


Figure 2.2: Diagram showing the starting simplex (a) and the possible transformations (b-e). The algorithm always starts by performing a reflection (b) in which the worst point ($\boldsymbol{x}_{n+1}$) is reflected through the centroid of the remaining points of the starting simplex. Depending on whether the reflected point ($\boldsymbol{x}_r$) is better or worse than the best point so far, an expansion ($\boldsymbol{x}_e$) (c) or a contraction ($\boldsymbol{x}_c$) (d) is performed. If one can not get rid of the worst point (via a contraction), a reduction (e) is performed in which the simplex contracts around the best point ($\boldsymbol{x}_1$) (as in [57]).

parameter space (away from bad parameter sets). If the function value of the reflected point (y_r) is better than the best value ($y_r < y_1$), an even larger move is performed in that direction to see if one can obtain an even better point. This is known as an *expansion* (fig. 2.2c)

$$\boldsymbol{x}_e = (1+\gamma)\bar{\boldsymbol{x}} - \gamma \boldsymbol{x}_{n+1} \ , \ \gamma > 1 \ , \tag{2.2.3}$$

where $\boldsymbol{x}_e$ is the expanded point and γ is the expansion coefficient, which is the distance from $\boldsymbol{x}_r$ through the centroid, $\bar{x}$, to $\boldsymbol{x}_e$. However, if the value of the reflected point is worse than the worst value ($y_r < y_{n+1}$), the move was too large and the simplex must have crossed a valley. In this case, a *contraction* (fig. 2.2d) is performed which takes smaller steps to find an intermediate lower point such as to move down the valley,

$$\boldsymbol{x}_c = (1+\beta)\bar{\boldsymbol{x}} - \beta \boldsymbol{x}_{n+1} \ , \ 0 > \beta > 1 \ , \tag{2.2.4}$$

where $\boldsymbol{x}_c$ is the contracted point and β is the contraction coefficient. If the value at the contracted point is still worse that then worst point in the current simplex ($y_c < y_{n+1}$), for example a situation in which the valley floor has been reached, a *reduction* (fig. 2.2e), i.e. a contraction in all directions, is performed which pulls the simplex in around the best point, $\boldsymbol{x}_i$, such that

$$\boldsymbol{x}_i \rightarrow \frac{(\boldsymbol{x}_i + \boldsymbol{x}_l)}{2} \ , \tag{2.2.5}$$

and the procedure is restarted (i.e. with another reflection). In all other cases, the highest point, $\boldsymbol{x}_{n+1}$, is first replaced by the new (reflected/ expanded/ contracted) point before the procedure is resumed [60].

This algorithm is, in general, very robust and efficient as it does not require the computation of derivatives. However, problems with the convergence of the algorithm may arise given bad initial guesses (i.e. a starting simplex which yields large penalty values), a minimisation function which is a poor descriptor of the properties of interest, or a function with too many parameters ($n > 10$).

2.2.2 Boltzmann inversion

The simplest structure-based method for coarse-graining is *Boltzmann inversion* [61]. Assuming that bonded and non-bonded interactions are separable (as in eq. 2.1.3), they can be used to obtain the bonded potentials for the CG model by inverting the probability distributions of all bonds, angles, and dihedral angles of the atomistic reference system.

Consider a system composed of q independent degrees of freedom at temperature, T, with a Boltzmann distribution of

$$P(q) = Z^{-1} e^{-\beta U(q)} \ , \tag{2.2.6}$$

where $Z = \int e^{-\beta U(q)} dq$ is the configurational partition function. A free energy can be obtained by inverting the probability density distribution, $P(q)$, such that

$$U(q) = -k_B T \ln P(q) \ , \tag{2.2.7}$$

where Z is omitted since it is just an additive constant to the potentials and the distributions are rescaled to represent volume normalised distribution functions. This is also known as the *potential of mean force* (PMF) [62] as it considers the force averaged over all conformations in a system. Considering the degrees of freedom to be bonds, angles, and torsions ($q = r, \theta, \phi$) and assuming that their probability distributions (eq. 2.2.7) are independent of one another, i.e. $P(r, \theta, \phi) = P(r)P(\theta)P(\phi)$, their potentials are given by

$$\begin{aligned} U^{\rm cg}(r,T) &= -k_{\rm B} T \ln \left(\frac{P^{\rm cg}(r,T)}{4\pi r^2} \right) \\ U^{\rm cg}(\theta,T) &= -k_{\rm B} T \ln \left(\frac{P^{\rm cg}(\theta,T)}{\sin\theta} \right) \\ U^{\rm cg}(\phi,T) &= -k_{\rm B} T \ln \left(P^{\rm cg}(\phi,T) \right) \ , \end{aligned} \tag{2.2.8}$$

Note, that eq. 2.2.7 is only exact for dilute systems (i.e. gases). For systems which are more dense, additional interactions arise from closely neighbouring particles and it no longer suffices to describe the potential energy based on a pairwise radial distribution function (RDF). However, the PMF can still be used as an initial guess in correction procedures such as Iterative Boltzmann inversion.

2.2.3 Iterative Boltzmann Inversion

Another parametrisation technique is *Iterative Boltzmann Inversion* (IBI) [58], which is a natural extension of Boltzmann inversion (eq. 2.2.8). It is a numerical scheme which can be used to refine CG bonded or non-bonded potentials, $U^{\rm cg}$, while attempting to match the reference (atomistic) and CG radial distribution functions, $g^{\mathrm{at}}(r)$ and $g^{\mathrm{cg}}(r)$, respectively. As an initial guess, it often uses the potential of mean force (eq. 2.2.7), given by $U^{(0)}(r) = -k_{\mathrm{B}}T \ln g^{\mathrm{at}}(r)$. At each subsequent iteration i, a correction is added to the interaction potential based on the differences in $\ln g(r)$,

$$U^{(i+1)}(r) = U^{(i)}(r) + \alpha \Delta U^{(i)}(r) \tag{2.2.9}$$

$$\Delta U^{(i)}(r) = k_{\mathrm{B}}T \ln \frac{g^{(i)}(r)}{g^{\mathrm{at}}(r)} \ , \tag{2.2.10}$$

where $g^{(i)}(r)$ is the radial distribution function of the CG system at iteration i and $\alpha \in (0, 1]$ is a scaling factor used to control the stability of the scheme. Convergence is monitored by evaluating the difference in RDFs,

$$\Delta g^{(i)} = \int_0^{r_{cut}} \left[g^{(i)}(r) - g^{\mathrm{at}}(r) \right]^2 dr \ , \tag{2.2.11}$$

where r_{cut} is the cutoff distance for the CG interaction potential. Once a given convergence criterion has been met (e.g. $\Delta g < 10^{-3}$), the update is terminated and one has obtained a CG potential which reproduces the RDF of the reference system.

Pressure correction Coarse-graining a system by fitting to the reference structure as described does not necessarily correspond to a match of the thermodynamic properties such as pressure. To do this, one needs to relax the convergence criterion in eq. 2.2.11 or, in other words, to seek potential updates which offer a compromise between structure and thermodynamic properties. In fact, the pressure changes as information is lost, for example by substituting water molecules by single beads. To adjust for this, a pressure correction [58, 63] can be performed, which is a linear correction to the potential

$$\Delta V(r) = A \left(1 - \frac{r}{r_{\mathrm{cut}}} \right) \tag{2.2.12}$$

where A is a constant usually given by $-0.1 k_B T$. In this work, however, it is estimated at every update step i by a factor of A_i [63], such that

$$-\left[\frac{2\pi N \rho}{3 r_{\mathrm{cut}}} \int_0^{r_{\mathrm{cut}}} r^3 g_i(r) dr \right] A_i \approx (p - p_{\mathrm{ref}}) \, V \ , \tag{2.2.13}$$

where p_{ref} is the correct pressure and p the pressure at step i. Note, that the pressure correction automatically implies a deviation from the reference structure [64] and also a deviation from the reference compressibility [63, 65], κ_T, since

$$\rho k_B T \kappa_T = 1 + 4\pi\rho \int r^2[g(r) - 1]dr \ . \tag{2.2.14}$$

Thus, the pressure correction should only be applied at every second potential update step such as not to completely destroy the converged structure, while still correcting for the high pressure of the CG system.

3

Atomistic simulations of amphiphilic peptides

In this chapter, atomistic simulations are employed to characterize the structural features and types of interactions which govern the self-assembly of amphiphilic peptides (fig. 1.2a) into monolayers. First, small systems (of single peptides) are simulated to understand the influence of a peptide's environment and structure on its conformational behavior. Such simple systems already demonstrate the sampling limitations which arise in the presence of a complex free energy landscape where rare events[1] (i.e. folding transitions such as β-hairpins) play an important role on the time scale of an MD simulation. In these cases, it is only feasible to use atomistic simulations for the analysis of specific stages of self-assembly or as a reference for a model with a lower level of resolution with which one can reach larger time and length scales (ch. 6). However, even for these purposes, advanced sampling techniques such as REMD [56] should be employed.

Initially, single peptide simulations are employed to test the system setup and to analyse the differences in conformational behaviour induced by variations of peptide chain lengths, peptide side-chain lengths, the types of termini, and the surrounding environment (bulk water vs. the air-water interface). Inter- and intramolecular interaction parameters for the system are taken from the GROMOS 53A6 force field [35], which is a commonly used biomolecular force field that has already been validated for many biological systems similar to the one of interest [67]. To test the simulation setup, it is checked whether the key characteristics of the experimental system are reproduced. In this case, peptides should diffuse to and remain at the air-water interface such that aggregation can take place. This is largely guided by the peptide's amphiphilicity (as prescribed by the force field) which determines whether the peptide has a high enough driving force to align with its backbone to the interface such that the hydrophobic/hydrophilic side-chains reside in the air/water phases respectively. Furthermore, peptides should have the ability to self-assemble via hydrogen bonding of neighbouring backbone residues to form aggregates which can be stabilised by additional hydrogen bond-

[1] Rare events are transitions from one metastable state to another as the system overcomes some energy barrier or goes through a sequence of correlated events. They are *rare* as such transitions happen very infrequently compared to the relaxation time of the system. [66]

ing between protonated side-chains and electrostatic interactions between the charged termini. In the absence of other peptides, this leads to peptide conformations such as β-hairpins as stable conformations. Since these are kinetically stable (in particular at an interface), REMD simulations (ch. 2.1.2) are performed to be able to sample all molecular conformations (i.e. well sampled and canonically distributed). Their results are compared to MD simulations according to the amount of extended backbone conformations observed (i.e. end-to-end distance distributions) and characteristic conformations are identified for specific peptide sequences (via a simple clustering of peptide conformations).

Subsequently, multiple peptides (9 chains) are simulated in concert to analyse their self-assembly behaviour and investigate the structure of the aggregates formed. These simulations address only the very early stages of self-assembly, where the formation of β-sheets commences, and the last stages, at which a perfectly ordered monolayer[1] has been formed, since it is infeasible to simulate an entire assembly process of realistic peptide concentrations on an atomistic time scale ($t \gg 100$ ns). The early stages of self-assembly are represented by an initial simulation setup of free, disordered peptides, distributed randomly across the air-water interface with the aim to correlate conformational preferences observed previously from single peptide simulations to current assembly behaviour. The final stages of self-assembly are represented by simulations in which peptides are arranged in pre-assembled, ordered aggregates. From these, structural features such as the peptide spacing, backbone extension, backbone twist, and the types of hydrogen bonding between individual peptide strands are investigated.

3.1 Testing the system setup

In all atomistic simulations presented in this chapter, the GROMOS 53A6 force field [35] is used to model the interactions in the system of interest. It is based on the GROMOS force field, a united atom force field which was originally optimised to reproduce the condensed phase properties of alkanes [68], which has undergone a series of reparametrisations since the development of the original force field [69, 70, 71]. The GROMOS 53A6 force field in particular was reparametrized to reproduce the free enthalpies of hydration and apolar solvation, both of which are key properties in biological processes such as protein folding, ligand binding, and

[1] Here, a pre-assembled aggregate of 3×3 peptide chains is simulated to describe an excerpt of such a monolayer.

membrane transport. As such, the force field has already been validated on typical biological systems, such as DNA, numerous proteins, e.g. the hen egg-white lyzosome (HEWL), and peptides such as the proteinogenic β^3- dodecapeptide [67].

The validation for the peptide is of particular importance here, since in contrast to DNA or proteins, smaller peptide simulations are able to reach their folding-unfolding equilibrium within shorter time scales. This folding-unfolding process is strongly influenced by the balance between the hydrophobic-hydrophilic interactions prescribed by the force field. NMR and CD[1] experiments have shown that the β^3-dodecapeptide formed a 3_{14}-helix in methanol but did not form a well-defined secondary structure in water. Simulations of the peptide actually found a more stable helical structure in water when the older GROMOS 45A3 [72] force field had been employed, whereas results agreed with experiment once the reparametrized version of the force field was used (GROMOS 53A6).

For the amphiphilic peptides studied here, the balance between the different solvation states and functional groups (polar and apolar) is bound to play a major role as well. From a chemical perspective, the alternating sequences of hydrophobic (Phe) and hydrophilic (Glu/Asp) residues should serve to align the peptide with its backbone parallel to the air-water interface as the side-chain polarity drives them to reside in like media (i.e. in the air/water partitions respectively). To test this, peptides of two lengths (PGlu-n, fig. 1.2a) of $n = 2$ and 5 are simulated. The system setup consists of a cubic water slab (simulated by the SPC/E water model) centerd in a box of twice its length in the z-direction with periodic boundary conditions applied in all (x,y,z)-directions. The smaller system (PGlu-2) consists of 5394 water molecules in a $5.50 \times 5.50 \times 11.00$ nm simulation box and the larger system (PGlu-5) of 15302 water molecules in a $7.68 \times 7.68 \times 15.37$ nm simulation box. The vacuum-water boundaries represent the air-water interface (two in each system). These are located at $z = -2.75, +2.75$ nm for the smaller system and at $z = -3.84, +3.84$ nm for the larger system.

Simulations are performed in the NVT ensemble with a temperature of $T = 300$ K, coupled with the Berendsen thermostat with a coupling constant of $\tau = 0.1\,\mathrm{ps}$, and a pressure of $p \approx 1\,\mathrm{atm}$, obtained via a short NpT simulation with the Berendsen barostat, prior to the extension of the simulation box. For electrostatic interactions, the smooth PME [73] method (sec. 2.1.1) was employed using a real space cutoff of 1 nm, while the van der Waals cutoff was set to 1.4 nm. Pe-

[1] Circular dichroism (CD) is an experimental technique used to study chiral molecules or identify the secondary structure of larger biomolecules.

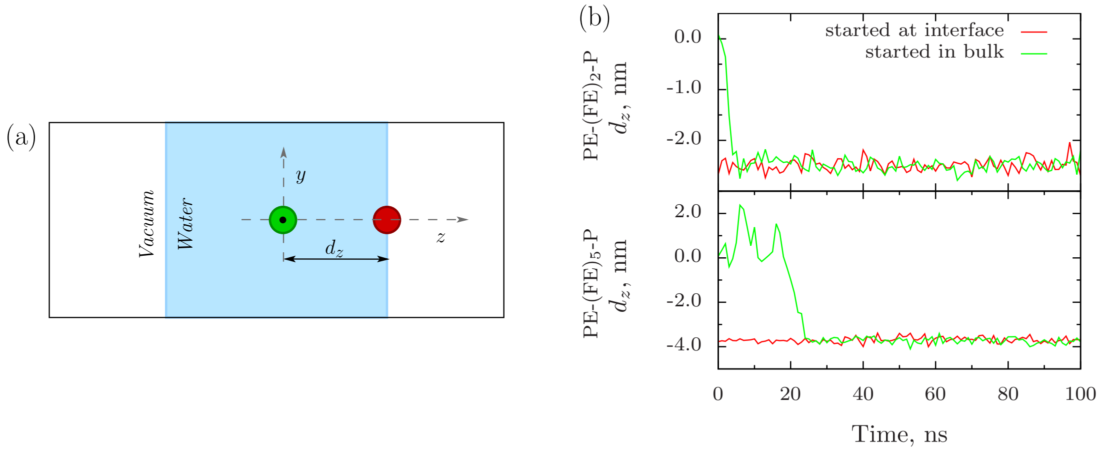


Figure 3.1: (a) Diagram depicting the air-water interface system setup with different peptide starting positions (green and red spheres). (b) Distances between the center-of-mass of the water slab and the peptide PGlu-n, for $n = 2$ (top) and $n = 5$ (bottom) as projected onto the z-axis, d_z. Green corresponds to the peptide being initially located at the slab's center (origin, denoted by the black dot) and red to being initially located at the slab's surface (interface).

riodic boundary conditions are applied in all (x, y, z)-directions. Each simulation box contains a single peptide, initially located at either the center of the water slab (green) or at the slab's surface (red) as depicted in fig. 3.1a. To analyse the peptides tendency to move towards the interface, the distance between the centers-of-mass of the peptide and the water slab (as projected onto the z-axis), d_z, is monitored over $t = 100\,\text{ns}$ of simulation with a time step of $\delta t = 2\,\text{ps}$.

Results show that when peptides are initially placed at the slab's center ($z = 0$), hydrophobic-hydrophilic interactions of the force field promote diffusion towards the interface. Note, that the shorter peptide ($n = 2$) is more rigid and diffuses much faster than the longer peptide ($n = 5$). When both peptides are started at the interface, however, no differences in behaviour can be observed, demonstrating that the interactions present are strong enough to keep peptides at the interface throughout the entire length of the simulation, even with a relatively small number of hydrophobic side-chains ($n = 2$).

3.2 Analysis of single peptides

Now, that the simulation setup has been tested, the properties of interest can be investigated. First, single peptide simulations are analysed to understand conformational changes induced by the interface environment as well as by variations in the peptide's length and sequence (i.e. side-chain length, types of termini).

This analysis is not only useful for estimating folding-unfolding times, but can also help to identify distinct peptide conformations for each sequence which can later be linked to their ability to self-assemble.

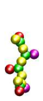
3.2.1 Effects of the environment

To analyse conformational changes induced by the interface environment, simulations of PGlu-n with $n = 2$ and 5 at the air-water interface (sec. 3.1) are compared to simulations of the same peptides in bulk water. The bulk system setup consists of a single peptide being placed at the center of a cubic water box (identical to the water slab from interface simulations) with periodic boundary conditions applied in all (x,y,z)-directions. All simulation settings are identical those of the slab system simulations (sec. 3.1). Comparisons of the effects induced by the two environments are made according to the peptide's backbone extension, measured by the end-to-end distance, $r_{\mathrm{end-to-end}}$, which is defined by the distance between the C_α atoms of the first and the last residue in the peptide chain.

Results for PGlu-n are shown in fig. 3.2a,b for $n = 2, 5$ respectively, for MD simulations in the bulk (blue dashed line) and at the interface (red solid line). For the shorter peptide, simulations in bulk water yield mostly extended conformations. At the interface, however, a β-hairpin[1] ($r_{\mathrm{end-to-end}} \approx 0.5\,\mathrm{nm}$) is formed which takes almost half of the simulation time ($\Delta t \approx 45$ ns) to unfold again. This indicates that much longer simulation times are needed to reach conformational equilibrium where multiple folding-unfolding events need to be sampled. Geometrically, this might be explained by the peptide's backbone alignment to the interface which confines its motions to a 2D plane, where interactions only have access to a cross-section of the potential energy surface. Hence, possible unfolding pathways (i.e. those favourable in energy) which might be found in the bulk may not be found at the interface, leading to more kinetically stable β-hairpins. It is, however, impossible to judge from these simulations if the β-hairpins are also thermodynamically more stable. For the longer peptide, β-hairpins are formed in both environments, which do not unfold again within the remaining simulation time.

To avoid getting trapped in such conformations (i.e. local minima in the free energy landscape) and to sample closer to the equilibrium (canonical) conforma-

[1] β-hairpins are secondary structures in which the backbone folds such that the C- and N-termini meet and intermediate backbone residues connect via hydrogen bonding.



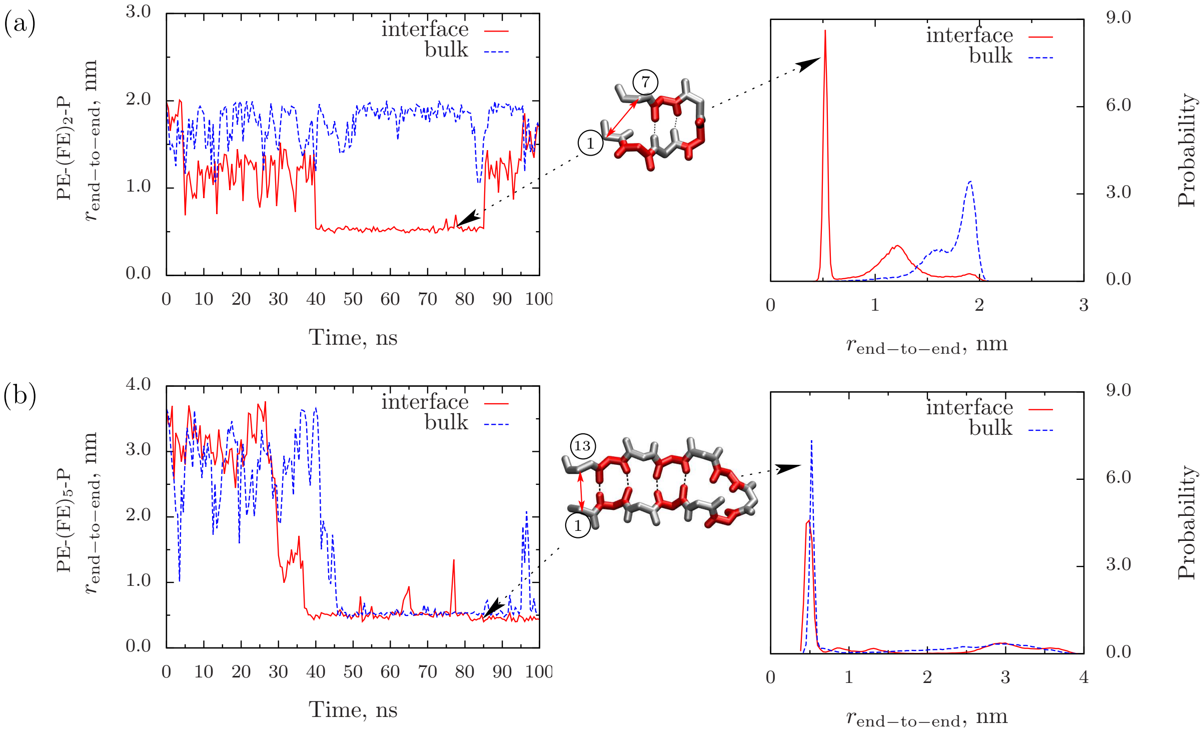


Figure 3.2: End-to-end distances and end-to-end distance distributions between C_α atoms of the first and last residue of peptides PGlu-2 (residues 1 and 7) and PGlu-5 (residues 1 and 13). Simulations of the peptides in bulk water (blue dashed line) are compared to simulations of the peptides at the interface (red solid line) for (a) $n = 2$ and (b) $n = 5$. The snapshots show β-hairpins for PGlu-2 and PGlu-5 (in a backbone plus hydrogen representation), with even/odd residues represented in grey/colour. The black dotted lines shows the hydrogen-bonding between the backbone residues.

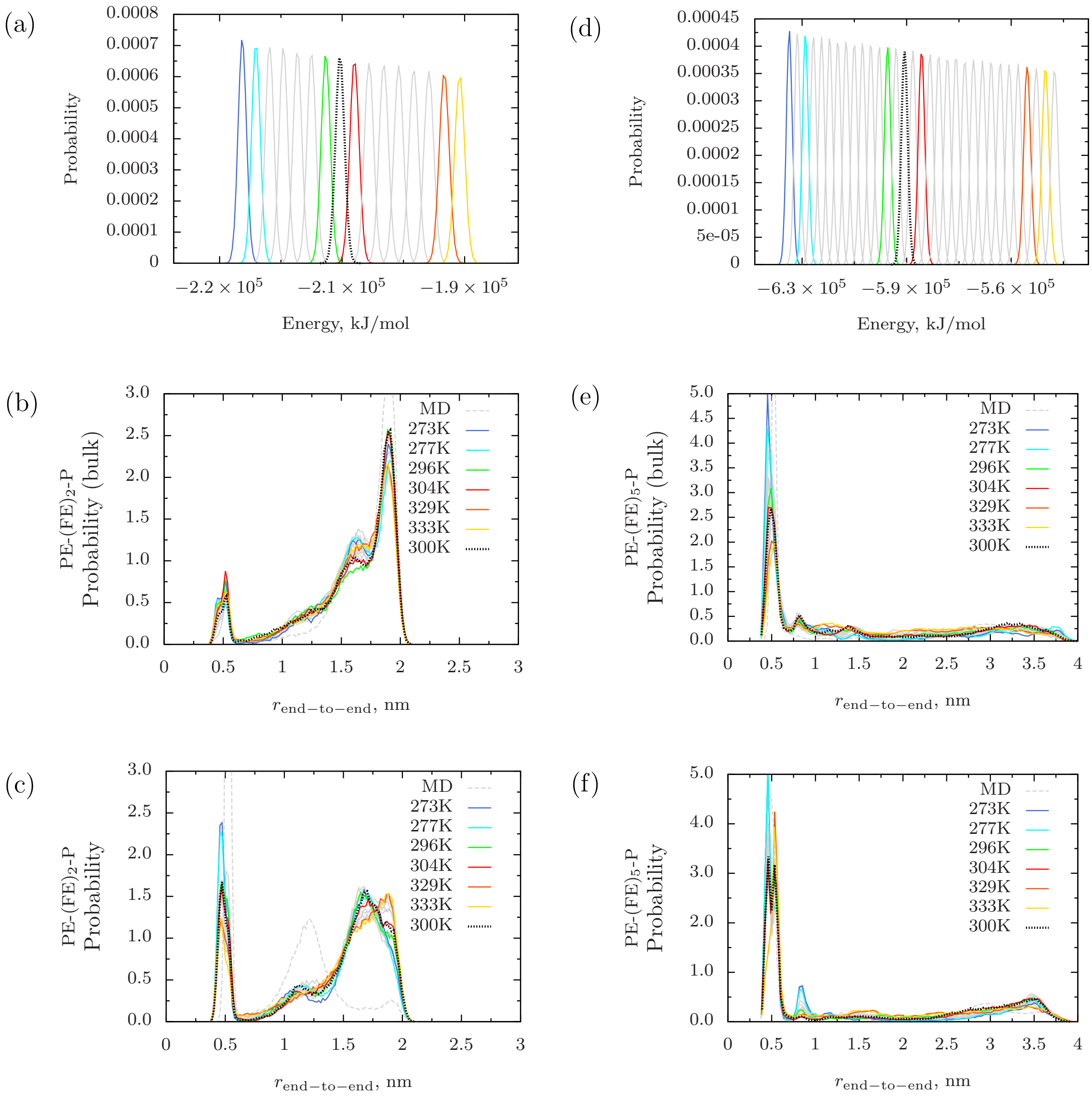


Figure 3.3: Results for REMD simulations of peptides PGlu-2 and PGlu-5, employing 16 and 32 replicas respectively. Plots show (a,d) the energy probability distributions, (b,e) the end-to-end distance distributions of REMD simulations in the bulk and (c,f) the end-to-end distance distributions of REMD simulations at the interface. For clarity, only the reference (black dotted line), its bounding temperatures, and the extremes (coloured) are highlighted, while all other replicas are plotted in grey. Results from MD simulations are shown by the grey dotted line for comparison.

tional distribution, REMD simulations are performed for the same systems. For PGlu-2, 16 replicas of temperatures $T = 273 - 333$K with $\Delta T \approx 4\,$K have been employed (fig. 3.3a), with a simulation time of $t = 100\,$ns per replica. For PGlu-5, 32 replicas were employed to cover the same temperature range with $\Delta T \approx 4\,$K (fig. 3.3d), with a simulation time of $t = 50\,$ns per replica. Note, that the reference temperature ($T_{\text{ref}} = 300\,$K) is chosen to lie approximately in the center ($i = 0$) of the temperature distributions, $T_i = T_{\text{ref}} e^{ik}$, such that replica indices are $i = -7, \ldots, 0, \ldots, 8$ for PGlu-2 and $i = -15, \ldots, 0, \ldots, 16$ for PGlu-5. All other simulation settings are the same as in sec. 3.1.

REMD results for simulations of PGlu-2 in the bulk and at the interface are shown in fig. 3.3b,c with average acceptance probabilities of $p_{\text{acc}} \approx 2.9\%, 2.7\%$ respectively. As compared to MD simulations (fig. 3.2a), a more equal amount of β-hairpins and extended conformations is observed in REMD simulations at the interface. In the bulk, however, the amount of extended conformations still outnumbers the β-hairpin formation events. Fig. 3.3e,f show results for REMD simulations in the bulk and at the interface of PGlu-5 with average acceptance probabilities of $p_{\text{acc}} \approx 7.8\%, 6.7\%$ respectively. Comparing end-to-end distributions to those obtained from MD simulations, no significant differences between the two environments can be observed. Note, however, that the end-to-end distribution is only a rough descriptor for the folding-unfolding equilibrium of a comparatively long peptide such as PGlu-5. Average acceptance ratios between neighbouring replicas as well as heat maps which show the temperature trajectory of each replica (16 for PGlu-2 and 32 for PGlu-5) can be found in the appendix (sec. A.1.1, fig. A.1).

In conclusion, the peptide's backbone alignment with the interface seems to increase the kinetic stability of conformations such as β-hairpins in the case of PGlu-2. For PGlu-5, one can not make any predictions as the peptides conformational transitions become even slower. Hence, to obtain sampling closer to the equilibrium conformational distributions, it is necessary to use advanced sampling techniques, where one should choose temperatures which are optimally distributed (i.e. similar acceptance ratios), a large enough number of replicas such that acceptance ratios are not too small, and most importantly, high enough temperatures such as to escape local energy minima traps.

3.2.2 Effects of peptide length

To analyse the effect of peptide length, MD and REMD simulations of peptides PGlu-n with $n = 2, 5$ at the air-water interface have been considered. In MD simulations, the peptides formed β-hairpins of different kinetic stabilities ($\Delta t \approx 45\,\text{ns}$ for $n = 2$ and $\Delta t > 65\,\text{ns}$ for $n = 5$), where the β-hairpin only unfolds again for the shorter peptide (within the simulated time of $t = 100\,\text{ns}$). Fewer recurring hydrophilic/hydrophobic residues provide less hydrogen bonding possibilities between the backbone residues, making the β-hairpin less stable and easier to unfold. These observations have been confirmed by REMD simulations.

3.2.3 Effects of peptide sequence

Having illustrated the effects of the environment and peptide length, small variations in the peptide's sequence can be considered to understand which conformational features might facilitate/hinder their self-assembly at the interface. Taking PGlu-n to be the reference peptide for comparison, effects of a shorter side-chain (in PAsp-n) and the absence of the proline termini (in PheGlu-n) are investigated. Note, that only the shorter peptides ($n = 2$) are simulated as they require less time to undergo conformational changes and hence results should be closer to conformational equilibrium than those for the longer peptides on the time scales employed. It will be shown that for these peptides of only 7 residues, a difference in conformational behaviour due to their particular sequences can already be observed.

REMD simulations for 8 replica of $t = 50\,\text{ns}$ each are carried out for the three types of peptides, ranging from temperatures of $T = 300 - 329$ K with $\Delta T \approx 4\,\text{K}$ (fig. 3.4a). The same simulation settings as in sec. 3.1 are employed. To compare the peptide's conformational behaviour, regions which showed significant differences in the backbone extension between the different types of peptides (b vs. c for side-chain comparison, and b vs. d for termini comparison) have been selected from end-to-end distance distributions and analysed via the GROMOS cluster algorithm [74] to identify predominant structures. This clustering algorithm sorts all structures according to their differences in RMSD[1] and clusters them according to a cutoff distance (r_{cut}), where structures of $r < r_{\text{cut}}$ are considered to be in the same cluster group and $r > r_{\text{cut}}$ are nearest neighbours.

[1] The root-mean-square deviation (RMSD) is a measure of the distances between atoms of a structure when superimposed onto a reference, $r_{\text{RMSD}} = \sqrt{(1/N)\sum_{i=1}^{N}\delta_i^2}$, where δ_i is the distance between pairs of equivalent atoms.

The group with the largest number of neighbours is then selected, eliminated from the cluster pool, and the process is repeated until all structures have been assigned to a cluster group. Here, the reference structure for the calculation of the RMSD is the peptide's structure at $t = 0$, including all atoms, with an RMSD distance of $r_{cut} = 0.15\,\text{nm}$ to define the same cluster group. The three most predominant cluster groups in regions 1 ($r_{\text{end-to-end}} = [1.00, 1.25]$ nm) and 2 ($r_{\text{end-to-end}} = [1.65, 1.75]$ nm) are considered for comparison of the different peptide sequences.

REMD results are shown in fig. 3.4b-d with replica exchange probabilities of $p_{\text{acc}} = 2.60\%$ (PGlu-2), 2.46% (PAsp-2), and 2.56% (PheGlu-2). Average acceptance ratios between neighbouring replicas as well as heat maps which show the replica's trajectory through temperature space are given in the appendix (sec. A.1.1, fig. A.2). For comparison, simulation results from $t = 100\,\text{ns}$ MD simulations are also plotted (grey, dotted line). Note, that the end-to-end distance distribution for PGlu-2 at the reference temperature ($T = 300\,\text{K}$, black dotted line) is similar to the one obtained from the longer sampled REMD simulations (fig. 3.3b) but differs significantly from the MD result. End-to-end distributions for PAsp-2 and PheGlu-2 do not show significant differences between REMD and MD simulations at the interface. Although an analysis of acceptance ratios indicates that longer simulation times are required to obtain a canonical distribution of the folding-unfolding equilibrium (app. A.1, A.2), REMD sampling results are still an improvement over MD simulations and can be used to make initial predictions about the different peptide behaviours observed based on their end-to-end distance distributions.

First, consider the shortening of the side-chain (fig. 3.4b,c). Comparing the end-to-end distribution to that of PGlu-2, it can be seen that there are more extended structures for PAsp-2, as indicated by the sharp peak in the end-to-end distribution around $r_{\text{end-to-end}} \approx 2\,\text{nm}$, whereas PGlu-2 peaks around $r_{\text{end-to-end}} \approx 1.7\,\text{nm}$. To find out which conformations might possibly hinder PGlu-2 from having a fully extended backbone, two regions are investigated. A cluster analysis of region 1, shows that both types of peptides have conformations in which one hydrophobic side-chain and a termini are close (Phe-Pro). In these conformations, however, the longer side-chain of PGlu-2 (cluster size ≈ 23%) is able to reach the P termini (Glu-Pro), which is not the case for the shorter (Asp) side-chain. Hence, the interaction of the longer side-chain with other residues

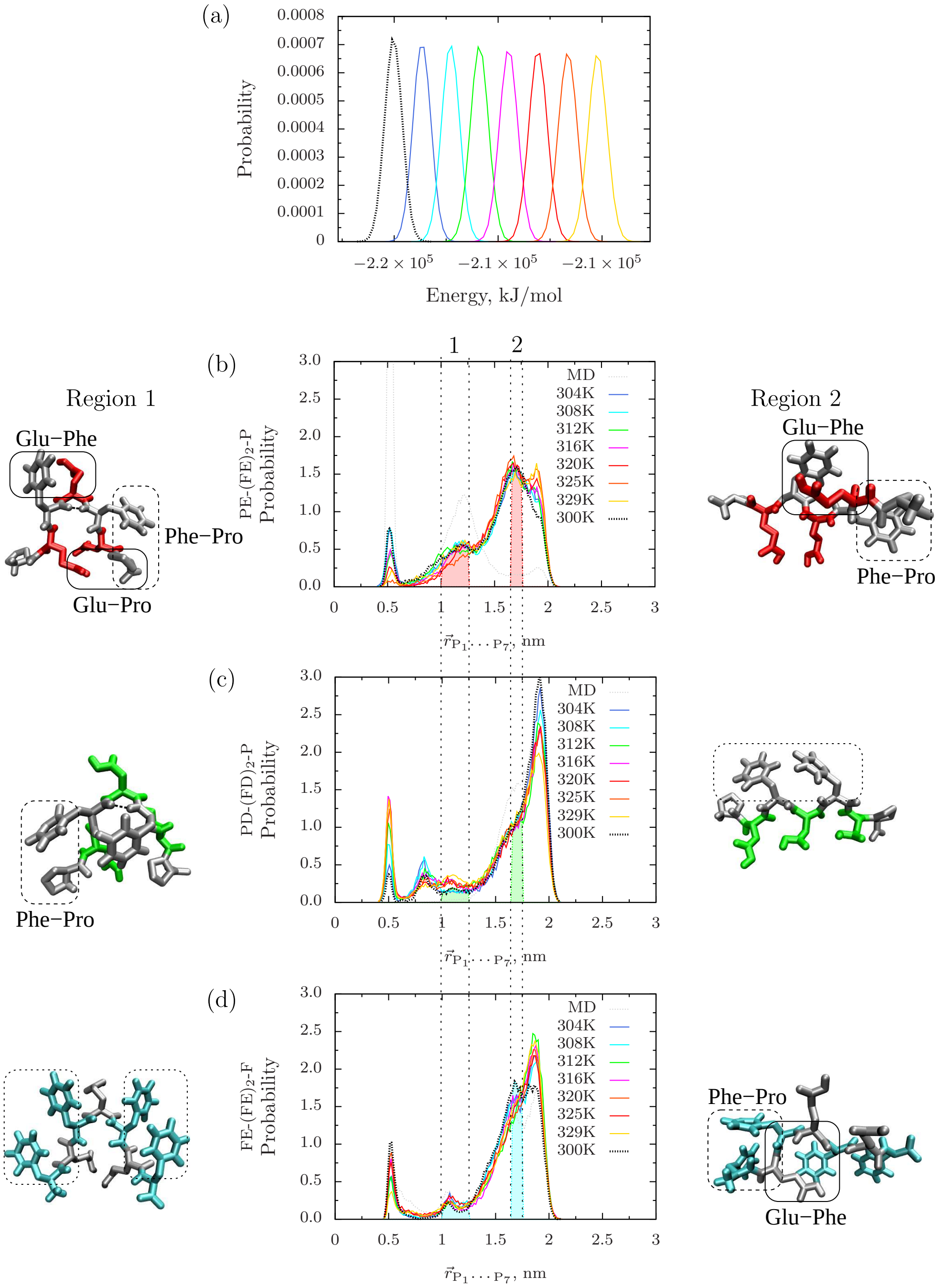


Figure 3.4: Results for REMD simulations of different peptides the air-water interface using 8 replicas of 50 ns each, with the reference at T=300K (black dotted line). (a) The top plot shows an example energy and temperature probability distribution (for the case of PGlu-2). The bottom plots show the end-to-end distance distributions measured between the first and last C_α of peptides (b) PGlu-2, (c) PAsp-2 and (d) PheGlu-2. Predominant clusters are shown for selected end-to-end regions (1 & 2), highlighting their distinct conformational features via solid/ dashed/ dotted windows. Note, that the dotted windows highlight the ordered regions (aligned residues) in the peptide's conformation.

can keep the peptide from sampling more extended conformations, which is also supported by the large peak observed in MD simulations in that region. Region 2, shows that the predominant conformations (cluster size ≈ 38%) correspond to conformations in which one of the hydrophilic side-chains bends towards the backbone (Glu-Phe), causing it to be slightly less extended compared to the PAsp-2 peptide, which shows no bending of the side-chain towards the backbone.

Next, consider the removal of the Pro termini (fig. 3.4b,d). In region 1, a cluster analysis of PheGlu-2 clearly shows that the added phenyl rings align (face-to-face) with the phenyl ring of the neighbouring residue. PGlu-2 do not exhibits this alignment, since one phenyl ring already aligns with the Pro terminus (Phe-Pro). This also illustrates the necessity of a second analysis criterion (i.e. the clustering algorithm) as quite different conformations (for different peptides) can correspond to the same end-to-end distance region. In region 2, PheGlu-2 shows similar features to PGlu-2 in that one of the hydrophobic side-chains bends towards the backbone, which brings the phenyl rings in close proximity to one another. However, in the absence of the proline termini, this effect is less pronounced.

Finally, differences between the peptide conformations can also be observed with temperature. Compared to PGlu-2, PAsp-2 shows a complete opposite effect with respect to a temperature increase in that it favours the formation of β-hairpins. This might be explained by the bonded interactions within the peptide which are only relaxed at higher temperatures, making the backbone more mobile such that it can allow for β-hairpin formation. Indeed, PAsp-2 in its extended form seems to display a more ordered behaviour (i.e. higher rigidity of the backbone) compared to PGlu-2 and may thus be more suitable for peptide self-assembly into ordered aggregates at lower temperatures.

In summary, it is observed that the longer side-chains of PGlu-2 peptides can reach and interact with other side-chain and backbone residues, contributing to many intermediate conformations (between β-hairpins and fully extended) which keep the peptide from sampling fully extended conformations at lower temperatures. PAsp-2 on the other hand seems to have a rather rigid backbone in comparison and displays less intermediate conformations. Very similar to this is the behaviour of PheGlu-2, which also shows more extended conformations, possibly due to more order created by the ring alignment of the hydrophobic side-chains, and less intermediate structures are observed. In addition, when considering the conformational changes with respect to an increase in temperature, PGlu-2 is seen to have the opposite behaviour than PAsp-2 in that it favours folded instead of extended conformations. This might indicate that the backbone becomes more mobile, and hence allows for the formation of more β- hairpins. Although longer simulation times and larger system sizes are required for more thorough sampling, the results obtained already indicate that conformational preferences vary between the different peptide sequences which are likely to affect their aggregation behaviour during self-assembly.

3.3 Analysis of multiple peptides

Now that the behaviour of single peptides has been illustrated, the interplay of multiple peptides can be considered to investigate their self-assembly behaviour. Although for self-induced peptide aggregation it is important to understand the detailed peptide conformations at the different stages of the assembly, difficulties in sampling the equilibrium conformational ensemble (due to the formation of β-hairpins, sec. 3.2.1) make it infeasible to study the entire self-assembly process on experimental time and length scales via atomistic simulations. Therefore, a small number of peptides (9 chains) is used to look at two specific stages of self-assembly. Namely, the first stage, which corresponds to initially free peptides[1] assembling into β-sheets, and the last stage, at which a perfectly ordered aggregate (3×3) has been formed. Intermediate stages should be investigated with a lower resolution model, developed in ch. 6, for which atomistic REMD simulations of single peptides (sec. 3.1-3.2) provide a reference, and atomistic MD simulations of multiple peptides (sec. 3.3) serve as a first approximation to compare to.

3.3.1 Aggregation into β-sheets

To simulate initial peptide aggregation, a water slab with 9 peptides at the interface is prepared with enough space around each peptide strand to rotate (in the z-axis) around its centre without touching its neighbours. The angle of the peptide's backbone with respect to the interface is chosen at random such as not to introduce a bias towards an assembly into parallel or anti-parallel β-sheets. System sizes of $12.00 \times 12.00 \times 24.00$ nm for peptides PGlu-n, PAsp-n, and PheGlu-n with $n = 2, 4, 5$ have been setup and simulated for $t = 50\,\mathrm{ns}$, each which a time step of $\delta t = 2\,\mathrm{ps}$. For comparison, identical systems were set up with

[1] Here, *free* refers to the fact that peptides are initially well-separated and non-interacting.

the peptides pre-assembled into ordered (3×3) aggregates to mimic an excerpt of a monolayer, with a, b lattice spacings as observed in experiment (≈ 0.5, $0.7\,\mathrm{nm}$ respectively) to represent the final stage of self-assembly. All simulations have been performed in the NVT ensemble, with the same simulation settings as in sec. 3.1.

To follow the progress of aggregation, the peptides' secondary structure has been monitored throughout the simulations. Note, that here only the results for the shorter peptides ($n = 2$) at a low peptide concentration (9 chains) are shown. Secondary structure analyses for the longer peptides ($n = 4, 5$) and higher peptide concentrations (16 chains on the same surface area) showed the same qualitative trends, but require much longer time scales to assemble into ordered aggregates, with the longest peptides ($n = 5$) producing predominantly β-hairpins (see appendix, sec. A.1.2). To classify the secondary structure of a given conformation, many algorithms exist which categorise 3D structures according to different criteria such as the inter- and intra-C_α-distances, angles, hydrogen bonding patterns, and backbone curvature. As such, there is no ideal method to use and the means of characterisation should be chosen according to the structural features of the system under investigation. The most widely used is the DSSP [75] algorithm, which assigns secondary structures according to hydrogen bonding patterns only. In this work, however, the *secondary STRuctural IDEndtification* (STRIDE) [76] algorithm is employed, which uses hydrogen bonding energies as well as backbone torsional angles to assign secondary structures.

Results for PGlu-2, PAsp-2, and PheGlu-2 are shown in fig. 3.5a-c. The white regions correspond to 'coils', which are random conformations that cannot be assigned to any secondary structure. Green represents the formation of 'turns', which are structures in which the backbone reverses its overall direction such that two C_α atoms are close (< 7 Å apart) but their corresponding residues do not match any secondary structure element. 'β-sheets' are depicted by the yellow regions, which correspond to structures with at least 2 hydrogen bonds between backbone residues. Hence, these arrangements can be parallel or anti-parallel. Very rare are 'β-bridges' (brown), which are a type of β-sheet with only one occurrence of hydrogen bonding between residues.

It can be seen that all three types of peptides exhibit a very fast formation of β-sheets ($t \approx 5 - 20$ ns). However, PGlu-2 seems to take the longest time to form more ordered structures which is probably due to the numerous side-chain

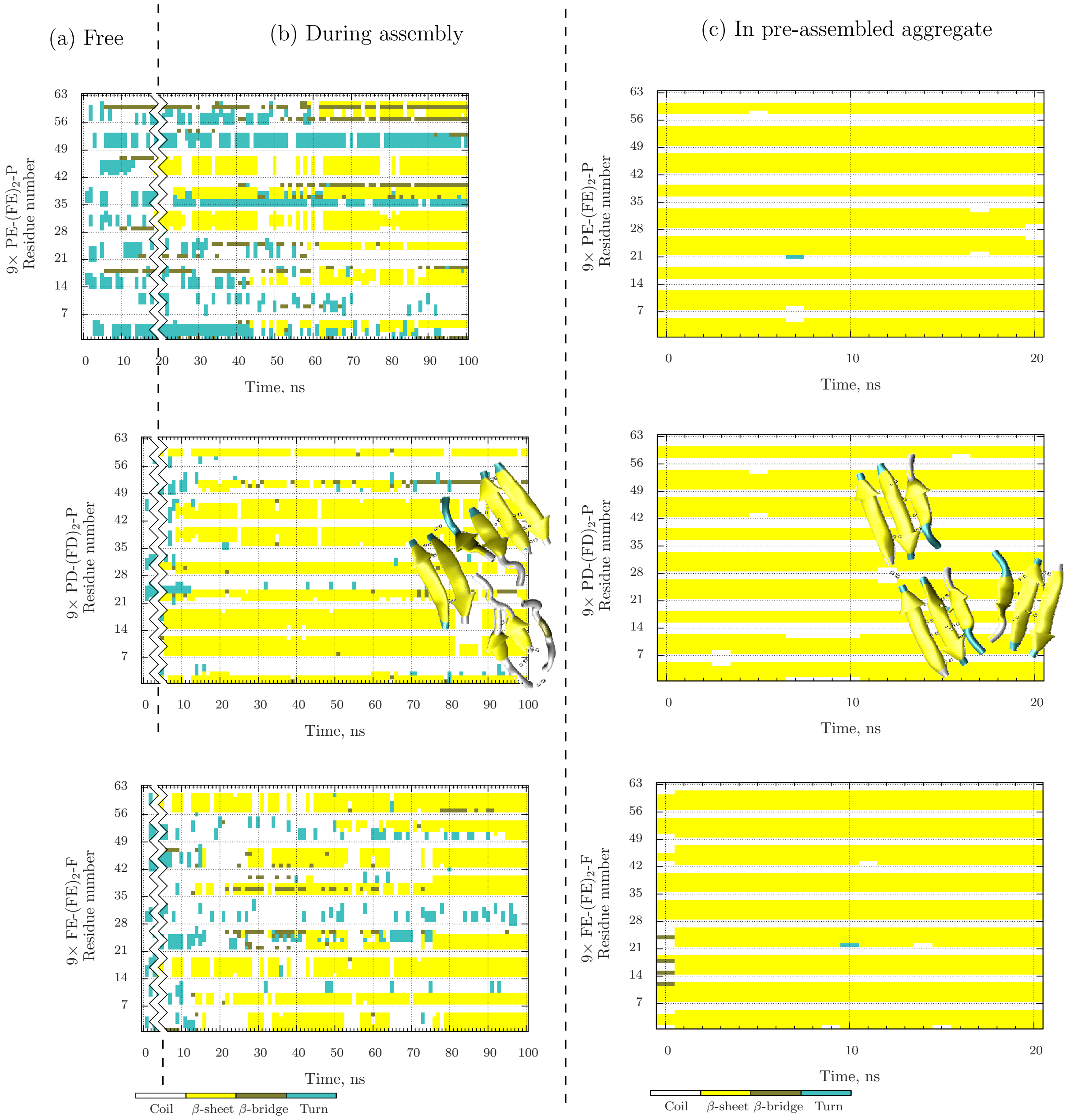


Figure 3.5: Secondary structure analyses for MD simulations of 9 peptides of PGlu-2, PAsp-2, and PheGlu-2 when initially started as free peptides ($t = 100\,\text{ns}$, left) or when started in a pre-assembled aggregate ($t = 20\,\text{ns}$, right). Based on the initial formation of β-sheets, the formed simulation is divided into into 2 stages, (a) free and (b) during assembly which represent the early and intermediate stages of self-assembly. (c) Peptides in pre-assembled aggregates which mimic an excerpt of an experimental monolayer represent the last stage of self-assembly. These stages also separate what is feasible to be analysed atomistically (a,c) and what should be studied with a lower resolution model (b).

- termini interactions which have already been observed in REMD simulations of single peptides (fig. 3.4). PAsp-2 has the most similar secondary structure plot when comparing the assembling stage to the final stage which was simulated by the pre-assembled aggregate (fig. 3.5b,c). This is probably due to the higher rigidity of its backbone induced by the shorter side-chains. PheGlu-2 falls in between the two as, on the one hand, it exhibits relatively ordered conformations due to the alignment of the faces of the additional hydrophobic Phe rings, but, on the other hand, it also creates a number of intermediate structures from interactions of the long hydrophilic side-chains (Glu) with backbone residues.

These observations can be confirmed by an analysis of the end-to-end distributions of the three stages of self-assembly (free: fig. 3.5a, during assembly: fig. 3.5b, in the pre-assembled aggregate: fig. 3.5c) which have been averaged over all peptide chains (fig. 3.6a-c). Here, PAsp-2 clearly shows the most extended conformations in the very early stage of assembly (free), with approximately the same peak positions as those observed in the last stage of self-assembly from (pre-assembled) aggregates. PGlu-2 can be seen to have the most intermediate structures during the assembling stage and PheGlu-2 seems to be an average between the two types of peptides. For longer peptides and higher peptide concentrations, the assembly shows a substantial amount of folded conformations and β-hairpins (sec. A.1.2, fig. A.3) and the systems require even longer simulation times to create ordered aggregates.

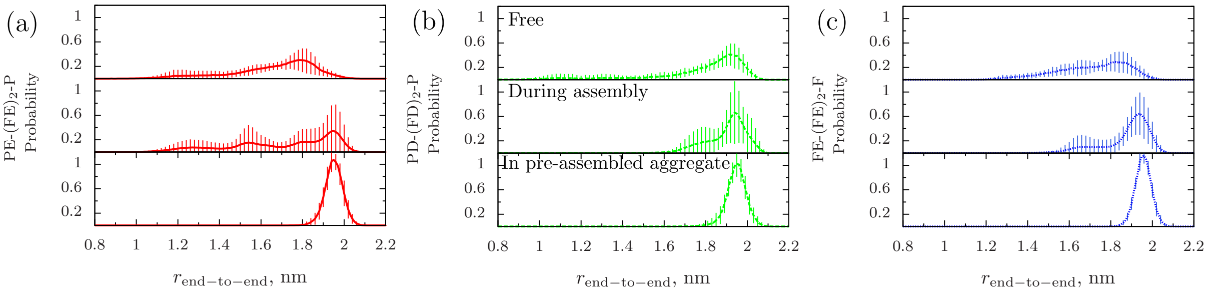


Figure 3.6: End-to-end distributions for MD simulations of peptides (a) PGlu-2, (b) PAsp-2, and (c) PheGlu-2, when initially started as free peptides: free (top) and during assembly (middle) or in the pre-assembled aggregate (bottom), as averaged over all 9 individual peptide chains.

3.3.2 Hydrogen bonding and stability

When studying peptide aggregates, it is important to understand the inter- and intra-chain hydrogen (H) bonding involved, not only to understand the self-assembly behaviour, but also to determine the stability of the aggregates. To do this, both the direct peptide-peptide H-bond interactions as well as the indirect H-bond interactions (i.e. those mediated by a single water molecule) have been analysed in the pre-assembled aggregates. An H-bond exists if the distance between the donor and the acceptor is $r_{\mathrm{HB}} \leq 0.35\,\mathrm{nm}$ and the hydrogen-donor-acceptor angle is $\alpha_{\mathrm{HB}} \leq 30°$.

First, the direct peptide-peptide H-bond interactions (depicted at the top of fig. 3.7 for the PAsp-2 aggregate) have been analysed by counting the occurrence of H-bonds between different residue categories (backbone 'BB', side-chain 'SC' of two types Phe and Glu/Asp, and terminus 'TM') throughout the simulations. In total, there are 10 different types of hydrogen bonding possible between these groups. On comparison of the three types of peptide aggregates, the most occurring H-bonds take place between backbone residues (BB-BB). They are on the order of 25-30 H-bonds for PGlu-2 and PAsp-2, and on the order of 30-35 H-bonds for PheGlu-2, but are not shown explicitly here such as to depict any differences between the peptides for less frequently occurring H-bonds. The BB-BB H-bonds occur between the secondary amine (N) and the Oxygen of the carbonyl group (mainly from Glu and rarely from Phe residues) and are believed to govern the stability of the monolayer in addition to the TM-TM H- bonds, which were observed to be the second most frequently occurring type of hydrogen bonding (fig. 3.7a-c).

Comparing PGlu-2 and PAsp-2 (fig. 3.7a,b), one sees that neither has many TM-SC bonds between the hydrophilic side-chains and termini. Earlier observations for PGlu-2 (sec. 3.2.3), which showed that many intermediate conformations exist in which the longer side-chain (Glu) bends towards the Pro terminus are more likely to interfere at an early stage of self-assembly where initial β-sheet formation takes place, not in the case of a stable pre-assembled aggregate (i.e. less BB-BB and TM-TM H-bonds). However, the SC-SC hydrogen bonding between hydrophilic side-chains seems to appear more frequently in the case of the PAsp-2 aggregate, which might also contribute to its stability. Hydrogen bonding bridged by water molecules (fig. 3.7d,e with examples depicted for the PAsp-2 aggregate) show no differences between the two types of peptide aggregates.

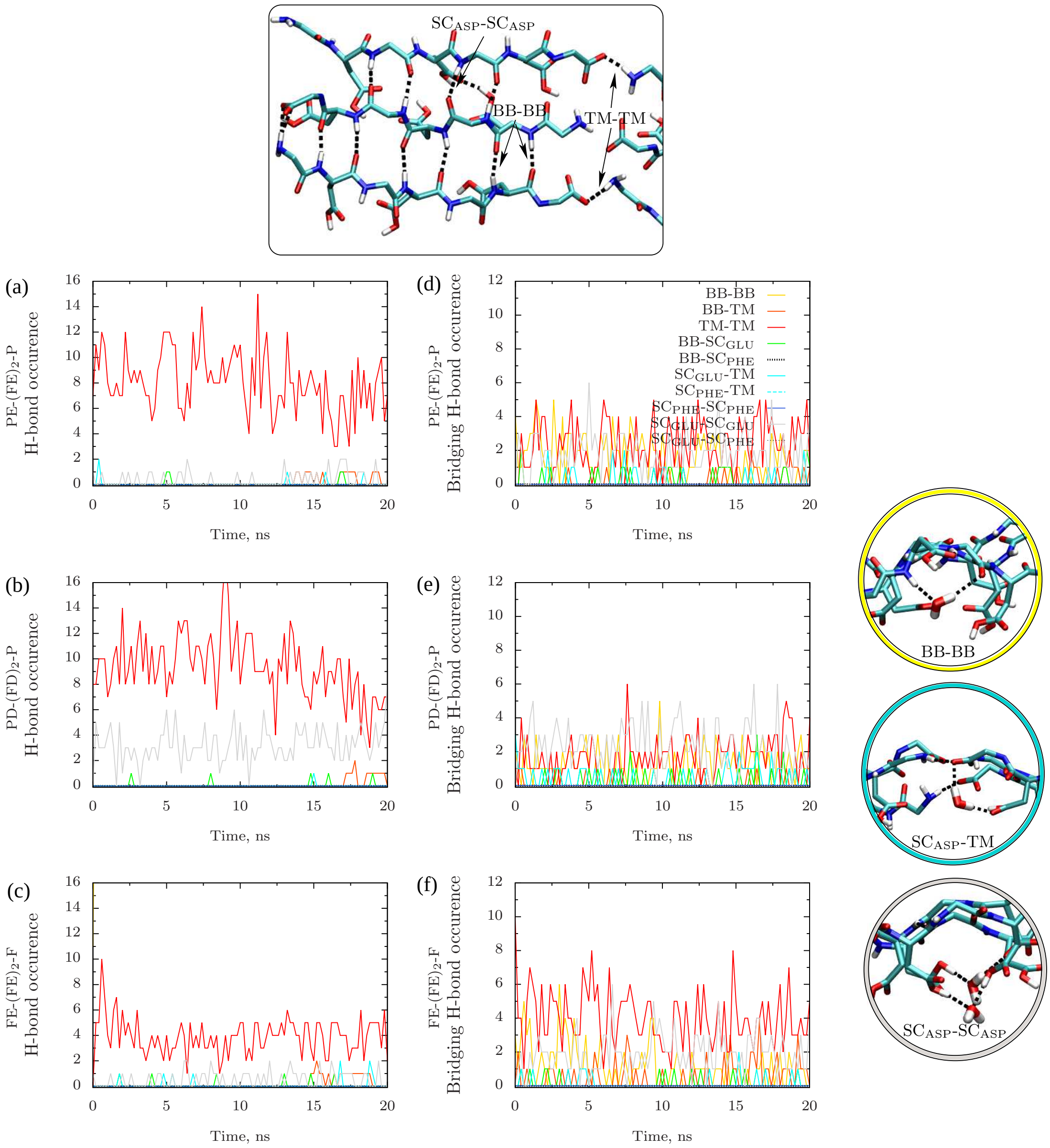


Figure 3.7: H-bonding analyses of (a) direct protein-protein H-bond interactions and (b) indirect protein-protein H-bond interactions (mediated by a single water molecule). These were assigned to separate the types of interactions, namely between groups such as the backbone (BB, residues 1-7, atoms backbone), the side-chains (SC, residues Phe, Glu or Asp, atoms not backbone) and the termini (TM, residues 1 and 8, atoms not backbone) to classify the type of interaction.

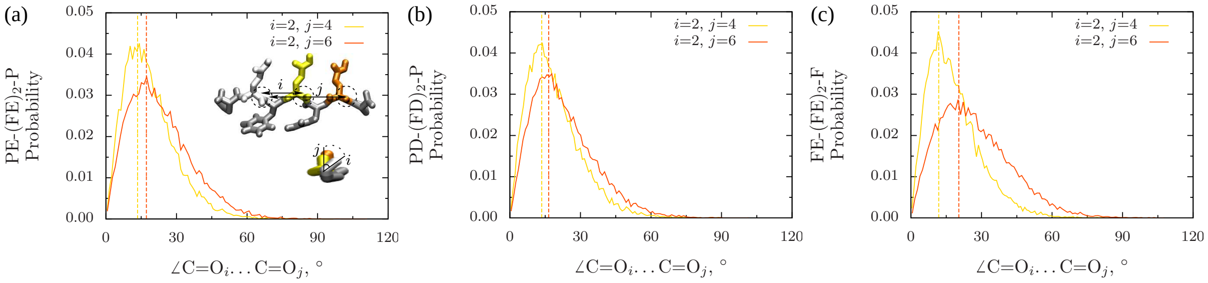


Figure 3.8: Twisting propensity of peptides PGlu-2, PAsp-2, and PheGlu-2 when sitting in the preassembled aggregates, showing the angle between the backbone C=O bonds i, j when looking down the backbone.

Considering the replacements of the Pro by Phe termini (fig. 3.7a,c), significant differences are observed. Indeed, in the absence of the Pro termini the TM-TM H-bond interactions are reduced, with some being partially bridged by water molecules. Throughout the simulation, this results in the aggregate losing contact between the termini much more readily and eventually drifting apart at the edges.

To further analyse the differences observed between the aggregates of different peptide sequences, an analysis of the backbone twist is performed which measured the angle between consecutive carbonyl groups along the peptide's backbone of residues i and j (as depicted in fig. 3.8a), averaging over all chains in the aggregate. The average distributions of the angles for peptides PGlu-2, PAsp-2, and PheGlu-2 are shown in fig. 3.8a,b,c respectively. From these results, it is apparent that PGlu-2 and PAsp-2 show a better overlap of the two (i,j) angles, while for PheGlu-2 the angle gets larger, indicating a twist of $\approx 10°$ along the backbone. Such a twist can result in a break in the β-sheet for larger aggregates sizes, making a peptide with a more rigid and straighter backbone the more suitable candidate to be used for the assembly into ordered monolayers.

3.3.3 2D order within the aggregate

Finally, one can consider the effect of the termini on the (2D) order of the monolayer. Experimentalists believe, that the Pro termini which are known to be β-sheet breakers induce 2D order within the monolayer. To study this effect, the pre-assembled aggregates (of PGlu-2, PAsp-2, and PheGlu-2) which represent excerpts of the respective monolayers have been analysed according to their diffusion and orientation of the individual peptide's backbones over time. To do this, the reference orientation of the aggregate was defined to be along the backbone of the central chain at $t = 0\,\mathrm{ns}$ (identified by the square in snapshots for 0 ns in fig. 3.9). After every 200 ps of simulation, a snapshot of the aggregate was taken and an order parameter was calculated via $S = (1/2) < 3\cos\theta - 1 >$, where $< \cdots >$ denotes the average over all peptide chains in the aggregate. If the Pro end groups do play a part in preserving the (2D) order of the aggregate, one should observe the aggregate to diffuse apart along the xy-plane as these are removed (i.e. in the PheGlu-2 aggregate).

Results are shown at the top of in fig. 3.9, with snapshots of the aggregates show after 0 ns, 10 ns, and 20 ns of simulation time. As can be seen, the PheGlu-2 aggregate displays the most disorder (i.e. lowest S) on average, in which the aggregate has separated at the termini and diffused away from a 2D ordered aggregate. This is probably due to the additional hydrophobic (Phe) groups which have already shown during the hydrogen bonding analysis to have a less H-bonds when employed as termini (sec. 3.3.2) with some mediated by bridging water, hence they are very weak. For peptides with Pro termini, however, there is also a deviation from a perfectly ordered aggregate structure. Here, the termini do not lose contact as readily as in the case of the PheGlu-2 aggregate, where the aggregate has already lost its 2D order at $t = 3\,\mathrm{ns}$. However, to have conclusive results, one needs to sample longer time scales, larger concentrations, as well as test different starting positions.

3.4 Conclusions

In conclusion, it has been demonstrated that differences between the peptides PGlu-2, PAsp-2, and PheGlu-2 can already be observed on atomistic time scales (for short peptides at small concentrations). The longer side-chains of PGlu-2 result in many intermediate structures between fully extended and folded (β-hairpins), which can hinder (or slow down) its self-assembly into ordered aggregates. Much more suitable is the behaviour of PAsp-2, which shows sharper distributions (i.e. fully extended with a only small amount of β-hairpins). Here, analysis of its internal structure shows that it not only has a fairly straight backbone (i.e. only a small twist) but also that the hydrogen-bonding between the hydrophilic side-chain residues can aid to stabilise the aggregate. Finally, PheGlu-2 shows the importance of employing the Pro termini as end groups in the peptide sequence, since without them, the aggregate does not have strong enough interactions between neighbouring termini and will quickly diffuse apart. For larger

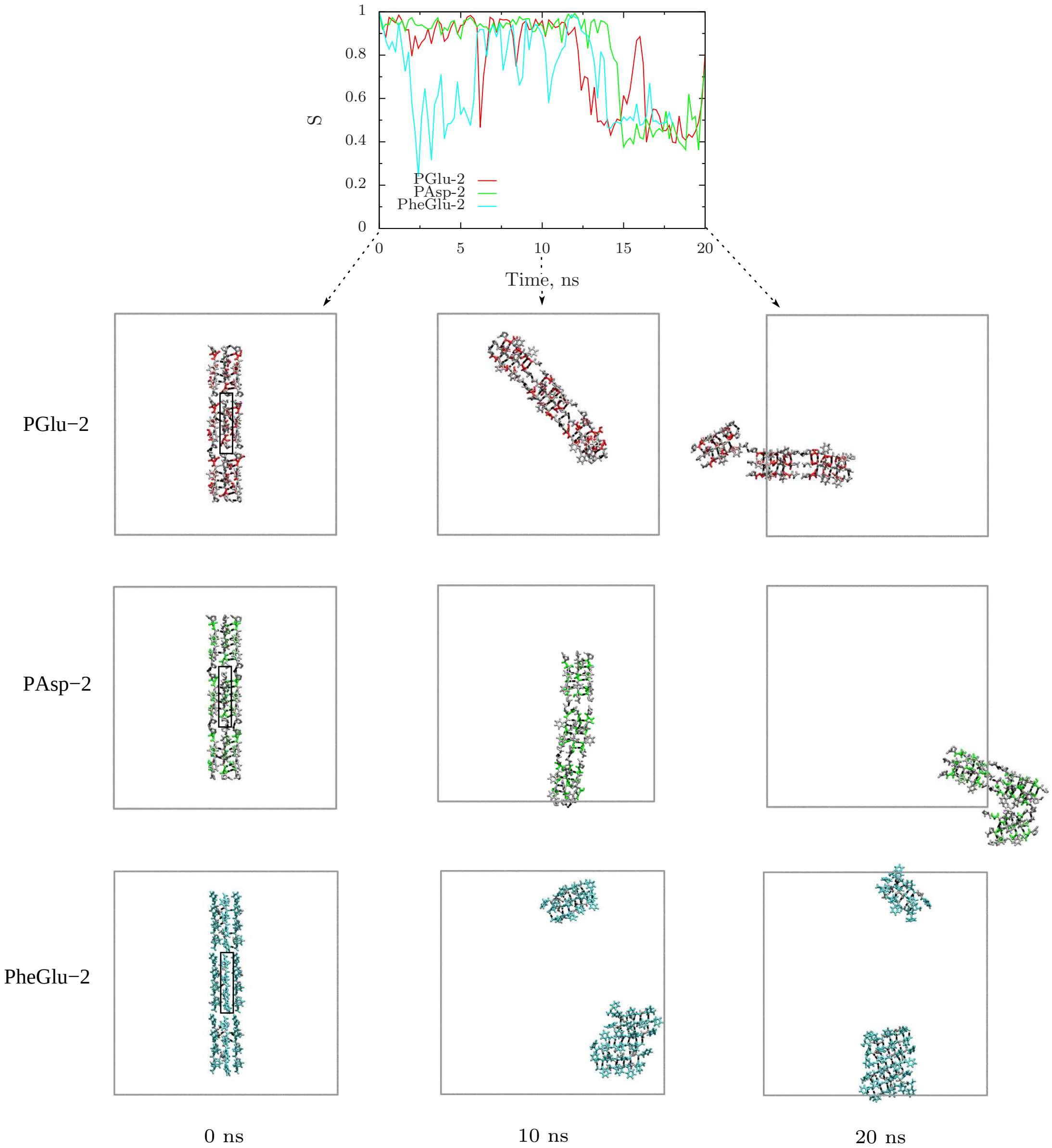


Figure 3.9: Analysis of order within the aggregate for peptides PGlu-2, PAsp-2, and PheGlu-2, depicting the average orientation of the peptides according to the central chain's backbone in the aggregate at $t = 0$ ns. Snapshots of the corresponding aggregates (top-view) are shown after 0 ns, 10 ns, and 20 ns of simulation time.

systems and longer time scales, it is already apparent that atomistic simulations do not suffice and one needs to resort to expensive REMD simulations to improve the sampling. As this limits the systems which can be studied, the time is better invested in developing a coarse-grained model with which one can address these scales more easily. One can then go back and forth between the scales (via back-mapping) to obtain the information required to aid experimentalists.

4

Coarse-graining by fitting analytical functions

While atomistic simulations are appealing to use for modeling biological systems as they provide access to full atomistic details at all times, they also consume vast amounts of computational power and time. In addition, most of that time is often spent on evaluating interactions not of interest, for example, the numerous solvent-solvent interactions in a protein environment, limiting biomolecular simulations on the atomistic level in time ($\approx \mu$s) and length scales ($\approx \mu$m). Hence, ways must be sought to reduce the number of interactions, or degrees of freedom, while preserving the key features which define the reference system (e.g. the local structure, bulk density, or the surface tension of a liquid).

One way to improve sampling efficiency is by *coarse-graining*, a procedure which groups several interaction sites (i.e. atoms) into coarse-grained (CG) beads. By introducing softer interaction potentials, it accelerates diffusion processes [30], increases intrinsic length scales, and consequently provides access to longer simulation times and larger system sizes. In the past, CG models have been developed for various types of systems, such as polymer melts [61, 77, 78, 79], organic solvents [80, 81, 82], lipid membranes [83, 84, 85], conjugated polymers [86, 87, 88], peptides [89, 90], surfactants [91], and proteins [92].

Often, coarse-graining is performed on the entire system as the interest lies in determining its macroscopic behaviour on extended time and length scales. If, however, the microscopic details of a particular region are required at all times, one may want to only coarse-grain part of the system while keeping the region of interest at a high level of resolution. An implementation of this is the particle-based adaptive resolution scheme (AdResS), which simulates regions of different resolution while allowing for particle exchange between them [93, 94, 95]. In this work, however, the aim is to simulate peptide self-assembly (simulated atomistically in ch. 3.3) under more realistic conditions (approaching experimental time and length scales). Hence, the entire system is coarse-grained but one can reintroduce atomistic coordinates at any point in time via a procedure known as *backmapping* [61, 96, 97].

To derive a CG potential, one needs to project a many-body potential of mean force onto a CG force field [98]. The projection operator is not unique and depends on the set of structure or thermodynamic properties of the system which should be preserved during coarse-graining. Force matching, for example, tries to approximate the entire distribution of states in a canonical ensemble [98, 99, 100, 101], while structure-based coarse-graining such as iterative Boltzmann inversion (IBI) [58] or inverse Monte Carlo [102] tries to match the radial distribution functions (RDFs) from the pair PMFs of the CG and reference system.

In the following two chapters (ch. 4,5) two different approaches for coarse-graining the air-water interface are described. Since atomistic water models are already fairly simple (e.g. an SPC/E-type water molecule consists of 3 points interacting via Lennard-Jones and Coulomb potentials), the CG models are made to consist of simple pair potentials, where long-range interactions are effectively taken into account to obtain a significant simulation speed-up. Initially, the CG potential is parametrised based on the structure of bulk water, leading to a very diffuse interface and hence motivating further development of systematic CG approaches for slab systems which yield sharper interfaces (sec. 4.1). The first CG approach then attempts to use analytical potentials to analyse the effect that the various potential regions (short- and long-range) and potential features (single versus double well) have on the air-water interface stability[1] and the structure of the liquid (i.e. RDF). This is done by testing a series of CG pair potentials, building up in complexity, in an attempt to find a compromise between matching the structure and the interfacial density profile of the atomistic reference (sec. 4.2-4.4). Chapter 5 on the other hand employs numerical potentials in a CG procedure which extends IBI for homogeneous systems to inhomogeneous systems.

4.1 Parametrisation in bulk water

As a starting point for the study, it is tested to which extent a coarse-grained pair potential obtained in a bulk water system is transferable to an interface system. The change in setting already hints at potential transferability issues, as the water molecules at the interface only have half as many interaction partners compared to those in a bulk water environment [103]. To illustrate the deficiencies of such an approach, the interface system is simulated by a potential obtained

[1] Here, a stable interface means a narrow transition region between the liquid phase and the vacuum, whose width is comparable to that of the atomistic reference.

via a structure-based CG method by matching the pair RDFs in bulk water.

Consider a bulk water system setup with a cubic box of length $l_b = 5.47\,\mathrm{nm}$, containing 5439 molecules, with a CG mapping of 1 bead per molecule (fig. 4.1a). To obtain the target RDF (evaluated between the two oxygen moieties), an atomistic NVT simulation of $t = 100\,\mathrm{ns}$ with a time step of $\delta t = 0.2\,\mathrm{ps}$ is performed at a temperature of $T = 300\,\mathrm{K}$, using the Berendsen thermostat with a coupling constant of $\tau = 0.1\,\mathrm{ps}$. Long-range electrostatics are treated with PME and periodic boundary conditions (PBC) are applied in all (x, y, z)-directions.

With this RDF as a target, the CG potential is parametrised by performing IBI (sec. 2.2.3) on the bulk water system for 300 iterations of $t = 100\,\mathrm{ps}$ each, with the same time step, temperature, and thermostat as the reference simulation and an interaction cutoff for CG beads of $r_{\mathrm{cut}} = 0.9\,\mathrm{nm}$. To remove the artificially high pressure introduced by coarse-graining, the obtained potential is pressure-corrected to yield a pressure of $p \approx 1\,\mathrm{atm}$. The resulting potential, $U^{\mathrm{cg}}_{\mathrm{ww}}(r)$, is then employed for simulations of a CG interface system (fig. 4.1b), obtained by expanding the bulk system by a factor of two in the z-direction ($l_s = 2l_b$), where w defines the width of the water slab ($w = l_b$ at $t = 0\,\mathrm{ps}$). The same procedure is repeated for two larger CG interaction cutoffs ($r_{\mathrm{cut}} = 1.4, 2.0$ nm) and the resulting density profiles for the CG interface systems are compared (fig. 4.2). Note, that all simulations are carried out via the GROMACS simulation package [54, 104] and the coarse-graining (CG mapping, IBI, pressure correction) is performed with the VOTCA package [105].

Results show that all of the obtained CG interfaces are too diffuse compared to the atomistic reference, demonstrating that by solely matching the bulk structure of water, $g(r)$, one does not recover important thermodynamic properties which contribute to the stability of the interface (e.g. surface tension). The microscopic reasons for this have already been discussed extensively in the literature [106], demonstrating that a simple CG pair potential cannot reproduce the structure as well as all of the thermodynamic properties of water simultaneously. Clearly, when moving from an atomistic to a CG water model, any orientational preference of the water molecule is removed as the molecules are replaced by isotropic spherical beads, which also leads to a different packing. In addition, CG beads do not carry charges and hence cannot take into account polarisation effects [107]. Furthermore, as hydrogen atoms are no longer simulated explicitly, the hydrogen bonding between the water molecules is lost which is especially critical for

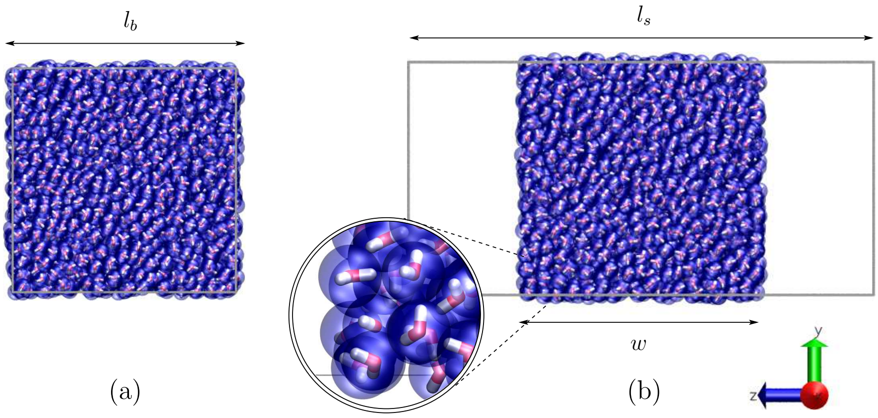


Figure 4.1: Snapshots of (a) the bulk system consisting of a cubic simulation box of length l_b and (b) the air-water interface system, which consists of a cubic water slab of width w and box length l_s in the z-direction ($l_s = 2l_b$ at $t = 0\,\text{ns}$). Both systems show the CG beads mapped onto the atomistic reference, with a mapping of 1 bead per water molecule as shown in the zoomed-in portion of the slab.

molecules at the surface which define the stability of the interface.

To understand how some of these properties can be preserved in a CG model, one needs to start with more simple potentials to identify which features are required.

4.2 A simple attractive potential (LJ12-4)

To start with, a simple Lennard-Jones 12-4 (LJ12-4) potential with only 2 parameters, σ and ϵ, is tested. The same potential has been previously employed by Shinoda et al. [108, 109] to develop a CG surfactant model, characterising water solely based on experimental surface tension (γ = 71.20 dyne/cm) and density (ρ = 999.57 kg/m^3) while keeping the functional form of the potential simple,

$$U_{\text{LJ12-4}}(r) = \tfrac{3\sqrt{3}}{2}\epsilon \left\{ \left(\tfrac{\sigma}{r}\right)^{12} - \left(\tfrac{\sigma}{r}\right)^{4} \right\}, \quad r \leq r_{\text{cut}} \,. \tag{4.2.1}$$

Here, σ is the distance between particles at $V = 0$ and ϵ is the energy minimum. Their findings showed that by choosing an appropriate cutoff distance (r_{cut}) and exponential index, such a simple model can reproduce the experimental density, surface tension, and compressibility of water, although at a cost of having a strongly over-structured liquid.

In this work, however, the aim is to employ the above potential to find a set of parameters which match the density *profile* ($X = \rho(z)$) of the atomistic reference

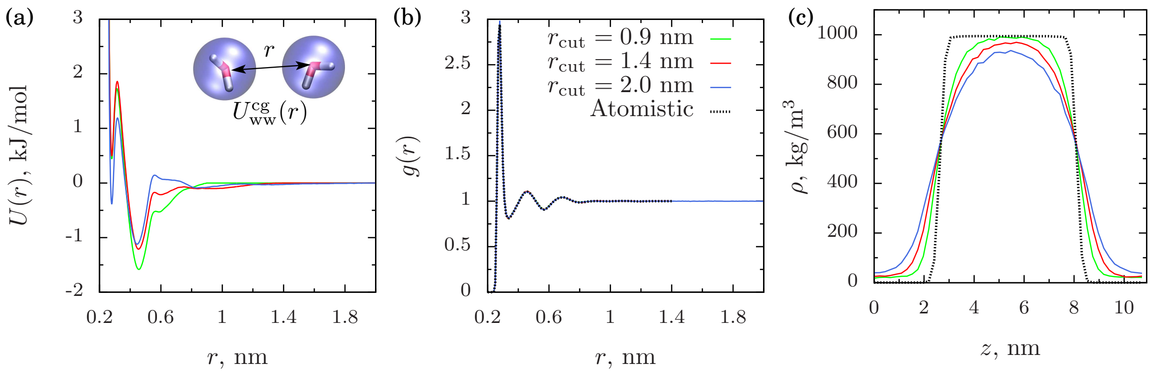


Figure 4.2: (a) CG pair potentials obtained from IBI parametrisation of (b) the bulk structure of water with different CG cutoffs, $r_{\mathrm{cut}} = 0.9$ nm (green), $r_{\mathrm{cut}} = 1.4$ nm (red), $r_{\mathrm{cut}} = 2.0$ nm (blue), and (c) their respective density profiles when employed for a simulation of a CG slab system.

simulation (fig. 4.2c, black dotted line) and see how well the other properties (i.e. RDF, surface tension) are matched. Note, that the second term of the LJ12-4 potential goes to zero as $-r^{-4}$, generating a potential which contains more long-range attraction than the conventional LJ12-6 potential. This may be especially important in reducing the diffusion of water molecules into the vacuum at the slab's surface (as observed for the CG bulk water model, sec. 4.1).

A broad range of bead sizes (σ =0.2-0.4 nm) and interaction energies (ϵ =1.0-5.0 kJ/mol) have been tested to find out which pairs of parameters provide density profile shapes comparable to that of the atomistic reference. To remain at the same state point (pressure) as the atomistic reference and avoid system size artifacts (i.e. droplet formation for small σ), box sizes have been rescaled to range from $3.6 \times 3.6 \times 7.3$ nm ($\sigma = 0.20$ nm) to $7.3 \times 7.3 \times 14.6$ nm ($\sigma = 0.40$ nm), with $\sigma = 0.30$ nm having the same box size as the atomistic reference, $5.5 \times 5.5 \times 11.0$ nm. For each parameter set, simulations were performed for 10 ns with an interaction cutoff of $r_{\mathrm{cut}} = 1.4$ nm (to have sufficient range for tuning the potential at a later stage) and the same settings as previous CG simulations (sec. 4.1). As one would expect, very low interaction energies result in volatile liquids (liquids with a high vapour pressure), while high interaction energies yield glasses (fig. 4.3a,b). Note, that beyond $\epsilon > 3.5$ kJ/mol, surface tensions become very high and no longer increase linearly with ϵ. The states of the individual simulations (volatile liquids → glasses, indicated by the colour bar) are assigned to range from the lowest to the highest (diffusion/ surface tension) values of the parameter sets (blue → red for diffusion and red → blue for surface tension), with the value of σ =0.30 nm and ϵ =3.0 kJ/mol, assigned to grey, defining the approximate of the region of interest. In the vicinity of this region, one finds a narrow range of parameter sets which yield systems of interfacial shapes close to that of the atomistic reference.

To characterise them, their RDFs, density profiles, and diffusion coefficients have been analysed. The RDF provides information about the structure and packing of the liquid, where the first r value provides an estimate for the particle size (σ). The density profile gives the bulk density[1] of the water slab, its width (w), and a measure of the thickness of the air-water transition layer. This last variable is closely related to the surface tension (γ), which describes the particles affinity for one another at the interface. Results for parameters σ = 0.3 nm with ϵ = 1.5-5.0 kJ/mol are shown in fig. 4.4. All other results (σ = 0.20, 0.25, 0.35, 0.40 nm with ϵ = 1.5-5.0 kJ/mol) can be found in the appendix (app. A.1.3, fig. A.4-A.7). Considering the RDFs from the different parameter sets, it can be seen that on the one hand, a stable interface comes at a cost of introducing long range order to the system. On the other hand, having interactions between particles which are too weak creates a very soft (diffuse) interface.

To obtain an optimum fit to the density profile of the atomistic reference, the Downhill Simplex algorithm (sec. 2.2.1, work-flow in app. A.1.4) is employed to avoid unnecessary searching of the parameter space, where the penalty function was calculated as $y_i = \int_0^{l_s} |\rho_i - \rho_{\text{ref}}|^2/\rho_{\text{ref}}\ dz$. This optimisation algorithm was originally employed for the development of atomistic force fields [110] but has since been successfully applied to coarse-graining applications [111, 112]. 3 sets of parameters which gave similar shapes to the atomistic density profile are used as initial guesses. For the Downhill simplex optimisation of the density profile ($X = \rho(z)$), 300 iterations of 100 ps each are performed. Results are shown in fig. 4.5 with initial guesses (grey) and the optimum parameters σ = 0.31 nm and ϵ = 3.65 kJ/mol (red), with only a very small deviation ($y \approx 2$) from the density profile of the atomistic reference. Although this is a good match of the thermodynamic properties which govern the shape of the interface, the corresponding RDF is over-structured, CG bead sizes are overestimated, and the packing between the first two coordination shells is different from that of the atomistic reference.

Hence, a simple potential with only two parameters (σ, ϵ) does not suffice to

[1] For a slab system, the bulk density is defined as the maximum of the density profile.

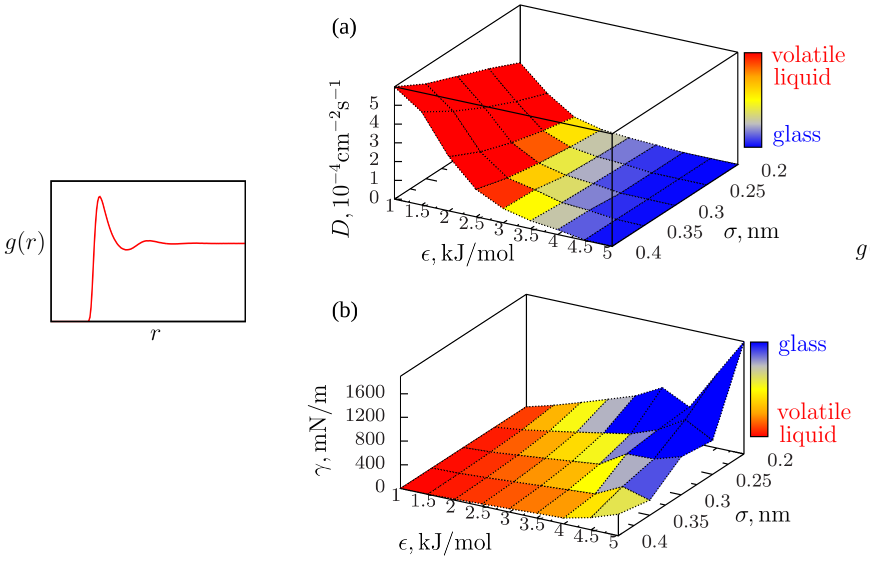


Figure 4.3: Diagram showing the (a) diffusion coefficients and (b) surface tensions for the sampled parameter space. The colour of the map's surface depicts the state of water, ranging from volatile liquids (red) to glasses (blue), with their characteristic radial distribution functions displayed on the left- and right-hand sides respectively.

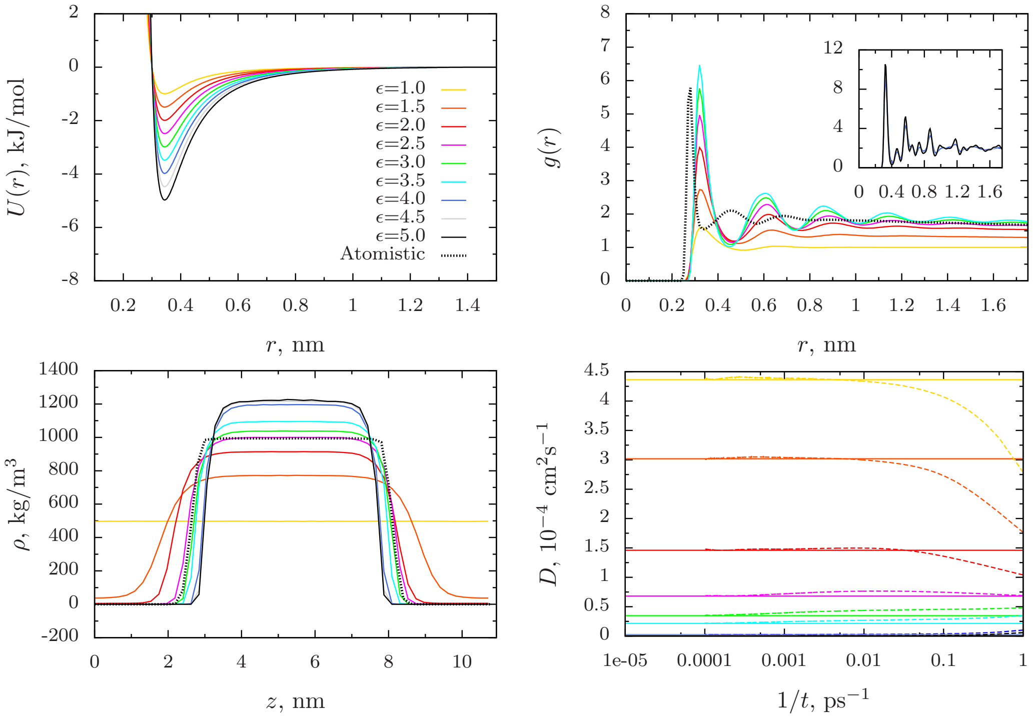
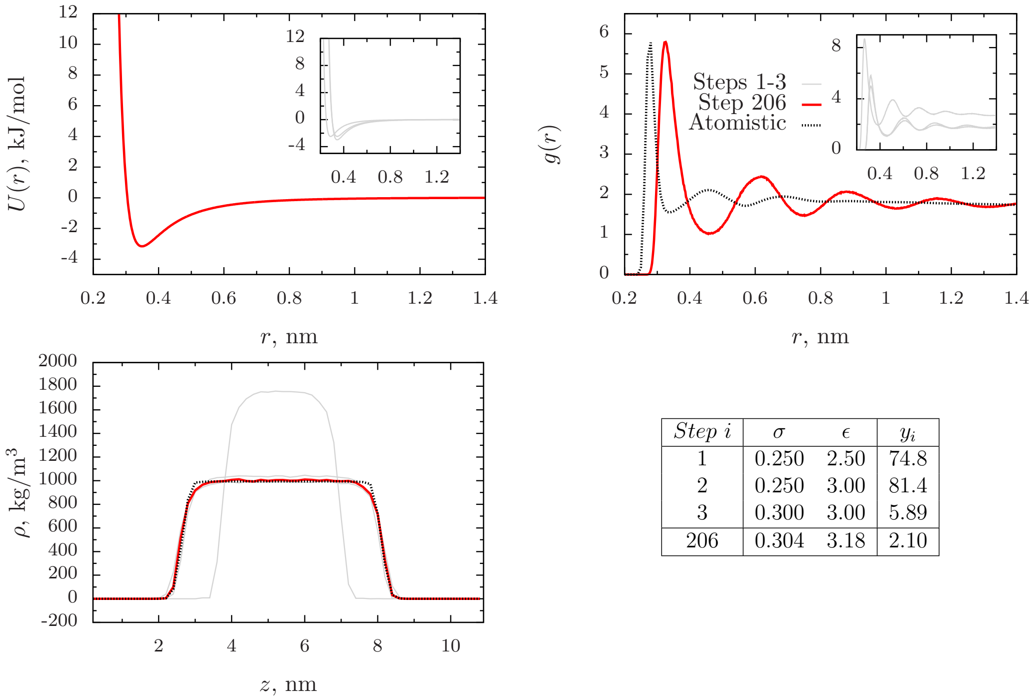


Figure 4.4: LJ12-4 results for parameters $\sigma = 0.30$ nm and $\epsilon = 1.0 - 5.0$ kJ/mol.



Step i	σ	ϵ	y_i
1	0.250	2.50	74.8
2	0.250	3.00	81.4
3	0.300	3.00	5.89
206	0.304	3.18	2.10

Figure 4.5: Downhill Simplex optimisation of the density profile ($X = \rho(z)$) with the LJ12-4 potential, showing the initial guesses (grey), the converged result (red) and the atomistic reference (black dotted line).

provide a good compromise between the structure and the density profile of the atomistic reference. Interactions between CG particles are either too weak to stabilise the interface or too strong, generating ordered phases which no longer preserve the liquid structure. However, the first issue can be tackled by modifying the attractive part of the potential to include even more long-range attraction. Such a model is introduced in the next section.

4.3 Tuning long-range attraction (CKD)

Having attempted to use a very simple form for the inter-particle potential, it is clear that although the density profile can be reproduced perfectly, the structure of the liquid is much too ordered. In previous work by Cooke, Kremer, and Deserno [113], a tunable model to study fluid bilayer membranes was designed, which could preserve a lipid bilayer without the presence of explicit solvent particles to aid stabilisation (referred to as the CKD potential). The key ingredient in this model is a range parameter with which one can tune the extent of long-range attraction[1] between the lipid tails. By doing so, the bilayer can be stabilised without the introduction of long-range order (and consequently freezing the system). The repulsive part of the potential consists of the Weeks-Chandler-Anderson potential,

$$U_{\mathrm{rep}}(r) = \begin{cases} 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} + \frac{1}{4}\right], & r \leq r_{\mathrm{c}} \\ 0\ , & r > r_{\mathrm{c}}\ , \end{cases} \tag{4.3.1}$$

where the repulsive cutoff is $r_{\mathrm{c}} = 2^{1/6}\sigma$ and ϵ is the energy minimum at this point. The attractive part of the potential is

$$U_{\mathrm{attr}}(r) = \begin{cases} -\epsilon\ , & r < r_{\mathrm{c}} \\ -\epsilon\cos^2\left(\frac{\pi(r-r_{\mathrm{c}})}{2w_{\mathrm{c}}}\right), & r_{\mathrm{c}} \leq r \leq r_{\mathrm{c}} + w_{\mathrm{c}} \\ 0\ , & r > r_{\mathrm{c}} + w_{\mathrm{c}}\ , \end{cases} \tag{4.3.2}$$

where w_{c} is the tuning parameter which determines the attractive range between the repulsive cutoff (r_{c}) and the cutoff of the potential ($r_{\mathrm{cut}} = r_{\mathrm{c}} + w_{\mathrm{c}}$) at which particles no longer interact (see potential shapes in fig. 4.7). With this additional handle on the tuning of the potential, a set of parameters ($\sigma = 0.30\,\mathrm{nm}$, $\epsilon = 3.0\,\mathrm{kJ/mol}$) which produced a CG stable interface with the LJ12-4 potential is tested for various ranges of w_{c} to find out which w_{c}/σ ratios generate a stable interface. Again, simulations are performed for 10 ns each, sampling a range of $w_{\mathrm{c}} = 0.2$-$1.0\,\mathrm{nm}$ in steps of 0.05, as well as $w_{\mathrm{c}} \approx 1.06\,\mathrm{nm}$ which is equivalent to CG particles being attractive until their cutoff (r_{cut}).

Results are shown in fig. 4.6, 4.7. It can be seen that a range of $w_{\mathrm{c}} = 0.35\,\mathrm{nm}$ has slightly softened the structure of the CG water without destabilising the interface. This is possible since, by increasing w_{c}, one is able to add more attraction to the tail of the potential without having to increasing ϵ (as in the case of the LJ12-4 potential) which also remedies some of the over-structuring in the higher order peaks of the RDF.

Next, the Downhill Simplex method is employed for 3 independent optimisations to match the density profile, the surface tension, and the density profile and the surface tension simultaneously with equal weights assigned to each property. These are started from 4 initial guesses which provide a good fit of the density profile ($X = \rho(z)$) and surface tension ($X = \gamma$) of the atomistic reference. Note,

[1] Long-range in the sense of the increased length scale over which attractions are effective, thereby making interactions *softer*.

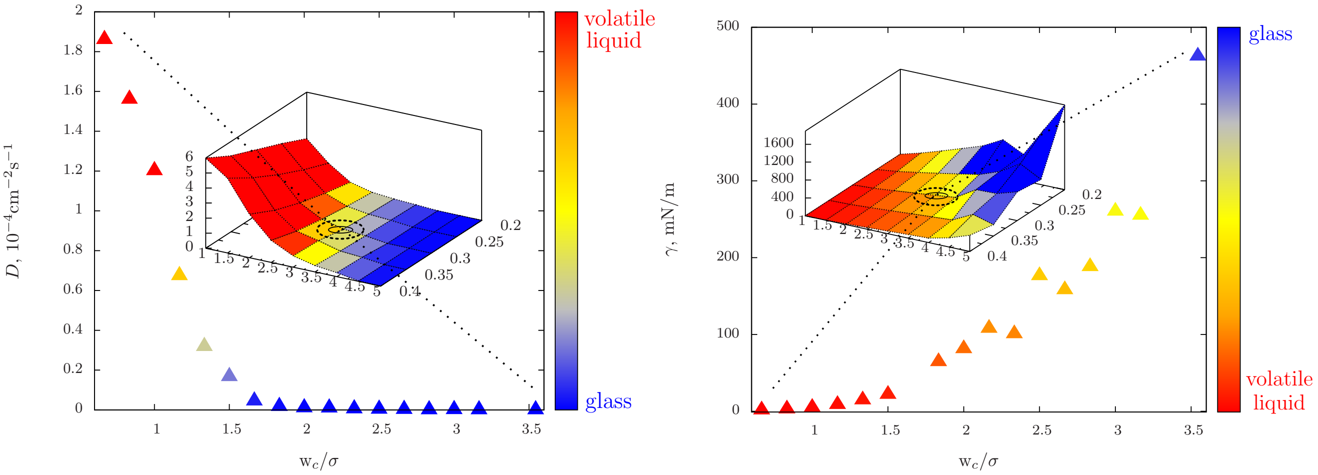


Figure 4.6: Diffusion coefficients and surface tensions for $\sigma = 0.30$ nm, $\epsilon = 3.0$ kJ/mol, and $w_c = 0.2 - 1.0, 1.06$ nm. Insets display previous results from LJ12-4, highlighting the point in phase space which was tested for various w_c/σ ratios.

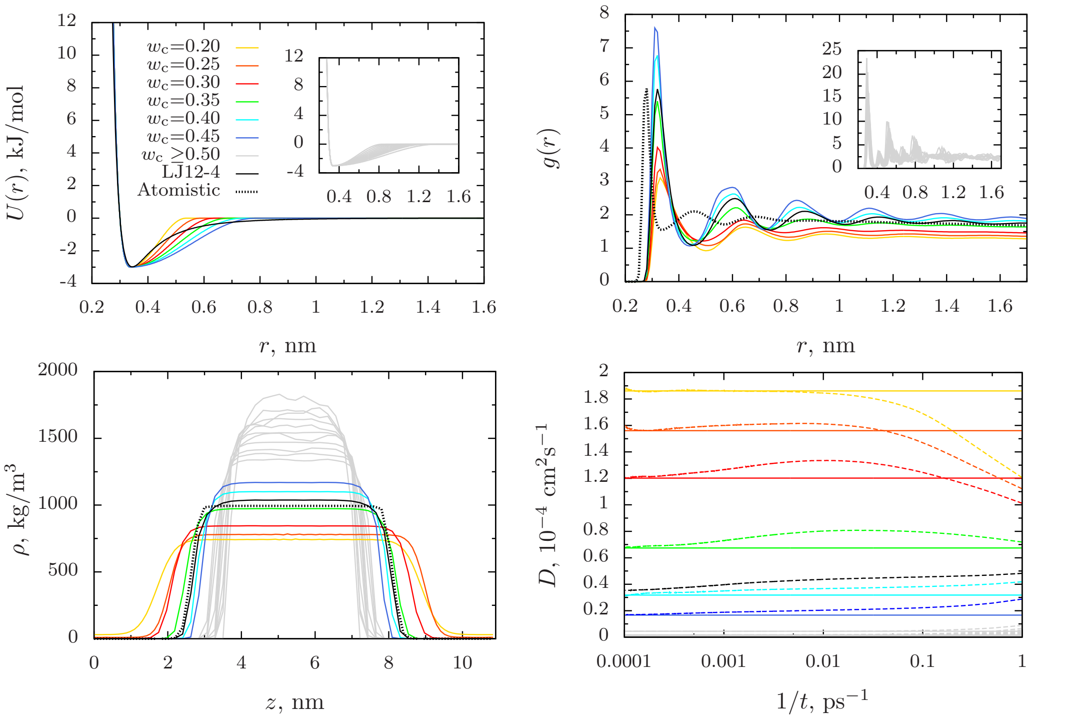
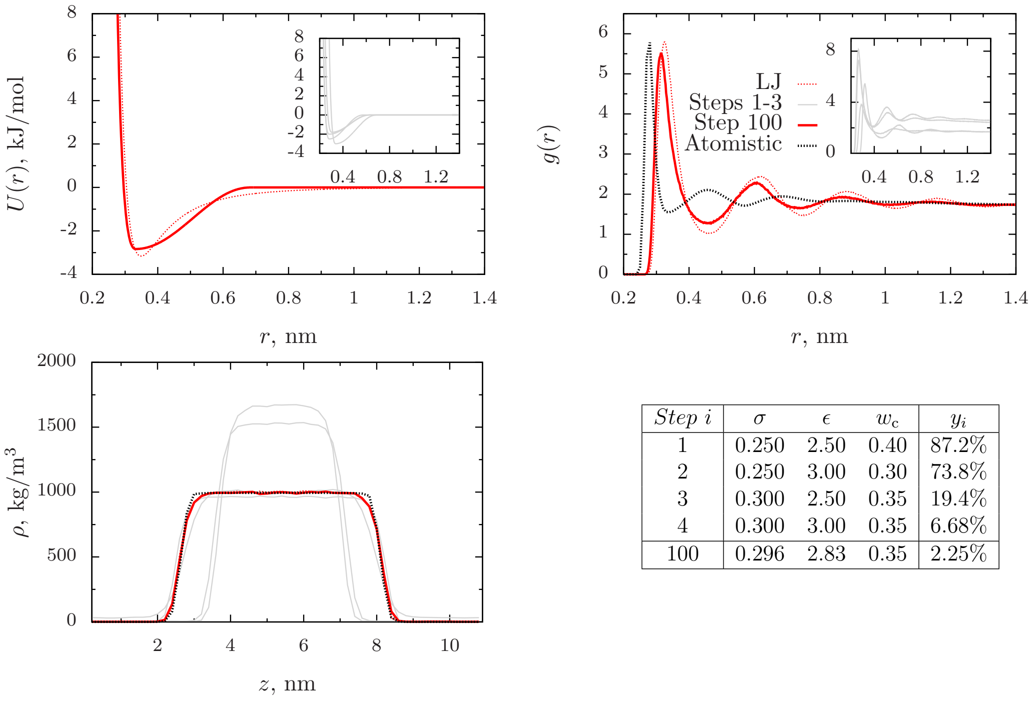


Figure 4.7: CKD results for parameters $\sigma = 0.30$ nm, $\epsilon = 3.0$ kJ/mol, and $w_c = 0.20 - 1.0, 1.06$ nm.



Step i	σ	ϵ	w_c	y_i
1	0.250	2.50	0.40	87.2%
2	0.250	3.00	0.30	73.8%
3	0.300	2.50	0.35	19.4%
4	0.300	3.00	0.35	6.68%
100	0.296	2.83	0.35	2.25%

Figure 4.8: Downhill Simplex optimisation of the density profile ($X = \rho(z)$) with the CKD potential, showing the initial guesses (grey), the converged result (red) and the atomistic reference (black dotted line). The optimised result from the LJ12-4 potential is also shown for comparison (red dotted line).

that as the ratio of w_c/σ remains constant, it can be used to estimate points which fall in between the grid of the already scanned region.

For the optimisation of the density profile ($X = \rho(z)$), 300 iterations of 100 ps each are performed. Results still show a small deviation ($y \approx 2$) from the density profile of the atomistic reference (fig. 4.8). Note, that the bulk density has been perfectly reproduced and that the slight discrepancy in the density profiles only arises from the edges, which are softer in the case of the CG interface. However, when looking at the RDF, one sees that the particle size determined by the algorithm has still been overestimated. In addition, the increased width of the first coordination shell indicates a different packing of particles at small distances ($r < 0.4$ nm).

Next, an optimisation for the surface tension ($X = \gamma$) is performed (app. A.1.5). Here, simulations are performed for 1 ns per iteration in order to obtain accurate surface tension averages. Only 100 iterations are performed. The surface tension is calculated from the pressure tensor, $P_{\alpha\beta}$, (sec. 2.1.1) as

$$\gamma = \int \left(P_{zz}(z) - \frac{P_{xx}(z) + P_{yy}(z)}{2} \right) dz \tag{4.3.3}$$

where $P_{zz}(z)$ is the perpendicular component and $P_{xx}(z)$ and $P_{yy}(z)$ are the transverse components with respect to the plane of the interface. Results show that by fitting to the surface tension of the atomistic reference ($\gamma^{\mathrm{at}} = 59.4 \pm 0.3$ mN/m), the resulting CG model ($\gamma^{\mathrm{CG}} = 59.1 \pm 0.7$ mN/m) neither produces a good fit of the density profile nor does it provide the correct CG particle size. The surface tension, however, is a value which can also be obtained by integrating an expression which includes the RDF, the slope of the density profile, and the derivative of the interaction potential [114]. As such, it is possible to have the same value of γ for very different $\rho(z)$ and $g(r)$. Therefore, interpreting results from an optimisation solely based on surface tension should be done with care.

Consequently, an optimisation for both the density profile and the surface tension is performed for 300 iterations of 1 ns each. Here, optimisation results fall directly in between those of the two individual (density/surface tension), optimisations with the density profile being reproduced well, a better estimate for the size for the CG beads, and a less ordered structure (app. A.1.5 with $\gamma^{\mathrm{CG}} = 59.3 \pm 0.8$ mN/m).

In conclusion, it is evident that although the CKD potential is an improvement over the LJ12-4 potential, a good compromise between the structure (i.e. RDF) and the thermodynamic properties (i.e. density profile and surface tension) has not yet been reached. The main problem is the overestimated bead size and the different packing of CG particles compared to the atomistic reference in the first two coordination shells. One way to correct for this might be the addition of a second minimum to the potential in order to absorb the excess CG particles now present in the first coordination shell into the second one. In the next section, this is implemented by the addition of a Gaussian to the current potential form (CKD).

4.4 Addition of a second minimum (CKDg)

In the past, numerous efforts have been made in deriving effective potentials for water which, in addition to reproducing the correct structure, could preserve its anomalities. Although it is known that it is impossible to correctly describe both, the thermodynamic properties and the structure of water with radially symmetric pair potentials [106, 115], several attempts have been made to recover them to a large extent. Work by Barraz et al. [116] employs two-length scale potentials which consist of a Lennard-Jones repulsive part and an attractive part composed of 4 Gaussians, creating shoulder-like potentials. It was discovered that as long as the first shoulder of the potential is not too deep compared to the second, some of the anomalities of water could be recovered and it would not be necessary to employ more expensive directional potentials. Here, a second minimum is created in the CKD potential (sec. 4.3) by addition of a single Gaussian (referred to as the CKDg potential). The repulsive part is

$$U_{\mathrm{rep}}(r) = \begin{cases} 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} + \frac{1}{4}\right] + he^{-\frac{(r-p)^2}{2s^2}}, & r \leq r_{\mathrm{c}} \\ 0\ , & r > r_{\mathrm{c}} \end{cases} , \tag{4.4.1}$$

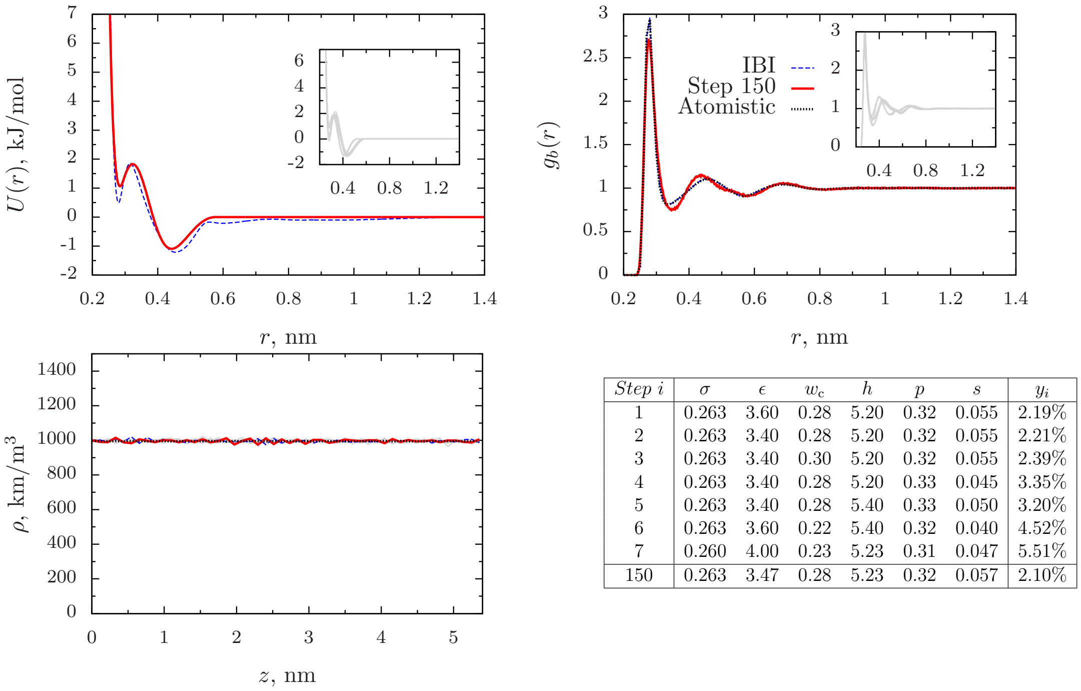
where $r_{\mathrm{c}} = 2^{1/6}\sigma$ and ϵ is the minimum energy at this point. h is the height of the Gaussian, p is the position of its centre, and s is its standard deviation. Similarly, the attractive part of the potential becomes

$$U_{\mathrm{attr}}(r) = \begin{cases} -\epsilon\ , & r < r_{\mathrm{c}} \\ -\epsilon\cos^2\left(\frac{\pi(r-r_{\mathrm{c}})}{2w_{\mathrm{c}}}\right) + he^{-\frac{(r-p)^2}{2s^2}}, & r_{\mathrm{c}} \leq r \leq r_{\mathrm{c}} + w_{\mathrm{c}} \\ he^{-\frac{(r-p)^2}{2s^2}}, & r > r_{\mathrm{c}} + w_{\mathrm{c}} \end{cases} . \tag{4.4.2}$$

With this new potential, the Downhill Simplex algorithm is employed in an attempt to improve the packing in the first coordination shell (i.e. first peak of the RDF) for the CG air-water interface system by fitting to the structure ($X = g(r)$). To facilitate optimisation, only the region of $r < 0.6$ nm has been considered when calculating the penalty function for the RDF ($y_i = \int_0^{r_{\mathrm{cut}}} |g_i - g_{\mathrm{ref}}|^2 / g_{\mathrm{ref}}\ dr$). First, a test run is performed on a CG bulk water system to see how closely one can recover the bulk structure ($g_b(r)$) with this potential. As there are now 3 additional parameters (needed to define the Gaussian), a total of 7 initial guesses is required. These are obtained by fitting (as closely as possible) the analytical CKDg potential to the pressure corrected tabulated potential, which was obtained by performing IBI in the bulk (sec. 4.1). The obtained parameter set is then varied slightly in the positions and depths of the potential minima to obtain the remaining sets of initial guesses. Similarly, sets of initial guesses are derived for the interface system, but with the parameter sets chosen to fluctuate more largely around the fitted IBI potential to explore the sensitivity of the system to changes in the potential. Note, that here IBI was performed on the slab system (i.e. matching the slab RDF, $g_s(r)$). Details about this can be found in the next chapter (sec. 5.1).

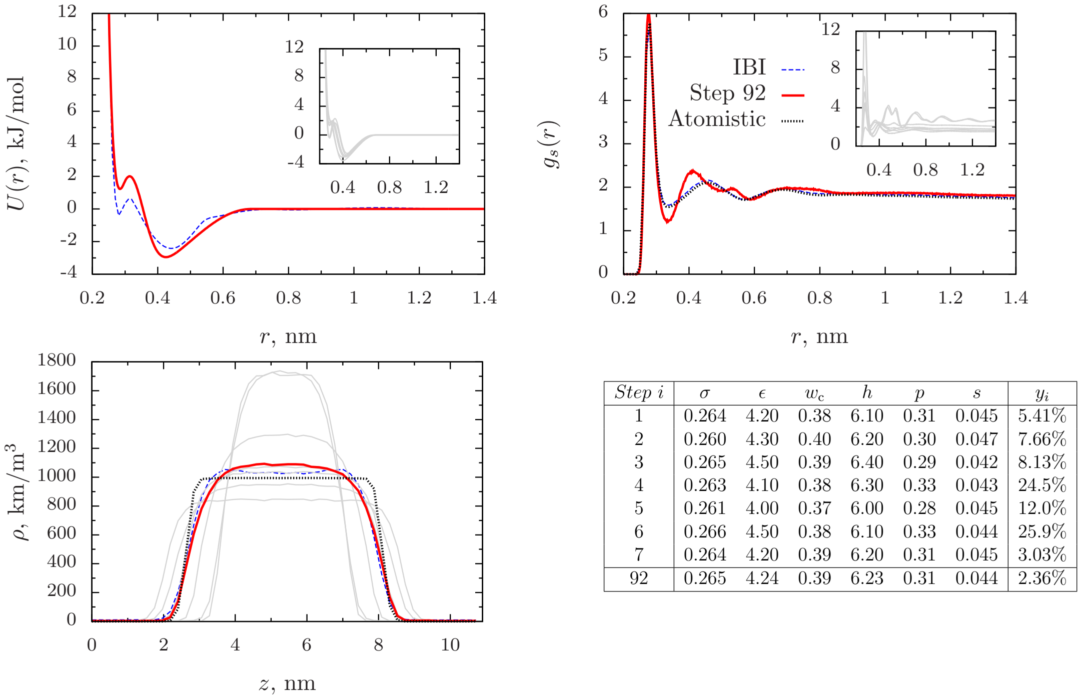
Fig. 4.9 and 4.10 show the IBI potential (blue dashed line) and the fitted initial guesses (grey), the IBI optimised result (red), and the atomistic reference (black dotted line) for the Downhill Simplex optimisations for the RDF of the bulk and the air-water interface systems respectively. For the CG bulk system optimisation, one obtains a good agreement with the structure of the atomistic reference. In the case of the slab system, however, the RDF cannot be perfectly reproduced (due to the simplicity of the analytical function compared to the numerical IBI potential) and one sees a deviation between the CG model and the atomistic reference in the first and second coordination shells. For both the bulk and the air-water interface systems, the CG potentials obtained now have two wells. With the new potential an identical optimisation was also performed to match the density profile ($X = \rho(z)$) of the slab system to compare the potential shapes. Here, however, only one potential well suffices to perfectly reproduce the density profile of the atomistic reference (app. A.1.6, fig. A.11).

Finding a compromise between the two properties ($X = g(r), \rho(z)$), however, is a difficult task as the path of convergence is heavily influenced by the choice of the penalty function employed. Although, the properties are given equal weights ($y_i = \int_0^{r_{\mathrm{cut}}} 0.5|g_i - g_{\mathrm{ref}}|^2/g_{\mathrm{ref}}\ dr + \int_0^{l_s} 0.5|\rho_i - \rho_{\mathrm{ref}}|^2/\rho_{\mathrm{ref}}\ dz$), the simplex proceeds to move towards the lowest sum of the penalty values, even if this corresponds to a region where the penalty of one property is significantly lower than the others. To try to avoid this and find a parameter set which reproduces both properties equally well, the property which has shown to converge more easily (i.e. the density profile) is optimised instead, starting from initial guesses which reproduce the other property (i.e. RDF). A total of 300 iterations have been performed, each of 100 ps, where both the penalty values of the RDF and density profile were monitored along the convergence path (fig. 4.11a). The optimised potential to represent the compromise CG model (sec. A.1.6, fig. A.12) has been chosen where the two lines cross (indicated by the grey arrow). This is indeed a good compromise as it approximately reproduces the bulk density and the size for the CG beads, while still remaining in the liquid state (i.e. no long-range order seen in the RDF).



Step i	σ	ϵ	w_c	h	p	s	y_i
1	0.263	3.60	0.28	5.20	0.32	0.055	2.19%
2	0.263	3.40	0.28	5.20	0.32	0.055	2.21%
3	0.263	3.40	0.30	5.20	0.32	0.055	2.39%
4	0.263	3.40	0.28	5.20	0.33	0.045	3.35%
5	0.263	3.40	0.28	5.40	0.33	0.050	3.20%
6	0.263	3.60	0.22	5.40	0.32	0.040	4.52%
7	0.260	4.00	0.23	5.23	0.31	0.047	5.51%
150	0.263	3.47	0.28	5.23	0.32	0.057	2.10%

Figure 4.9: Downhill Simplex optimisation of the bulk water system for the structure ($X = g_b(r)$) with the CKDg potential, showing the initial guesses (grey), the pressure-corrected IBI result (blue dashed line), the converged result (red line) and the atomistic reference (black dotted line).



Step i	σ	ϵ	w_c	h	p	s	y_i
1	0.264	4.20	0.38	6.10	0.31	0.045	5.41%
2	0.260	4.30	0.40	6.20	0.30	0.047	7.66%
3	0.265	4.50	0.39	6.40	0.29	0.042	8.13%
4	0.263	4.10	0.38	6.30	0.33	0.043	24.5%
5	0.261	4.00	0.37	6.00	0.28	0.045	12.0%
6	0.266	4.50	0.38	6.10	0.33	0.044	25.9%
7	0.264	4.20	0.39	6.20	0.31	0.045	3.03%
92	0.265	4.24	0.39	6.23	0.31	0.044	2.36%

Figure 4.10: Downhill Simplex optimisation of the slab system for the structure ($X = g_s(r)$) with the CKDg potential, showing the initial guesses (grey lines), the pressure-corrected IBI result (blue dashed line), the converged result (red line) and the atomistic reference (black dotted line).

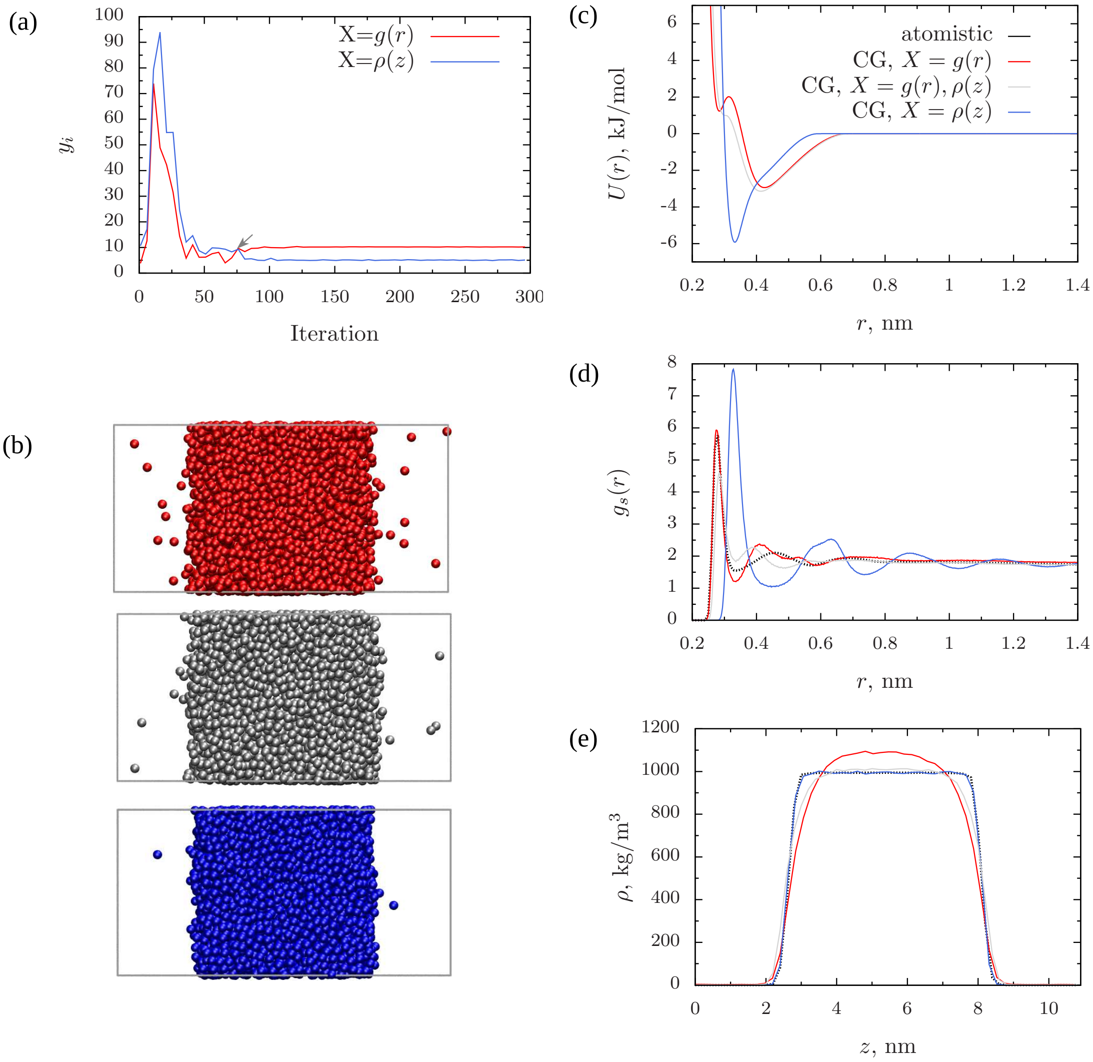


Figure 4.11: Summary of the Downhill Simplex optimisations for the structure ($X = g_s(r)$, red), the density profile ($X = \rho(z)$, blue), and a compromise between the two properties ($X = g_s(r), \rho(z)$, grey) showing (a) the convergence plot from which the compromise was selected (black dotted line), (b) snapshots of the slabs, as well as the (c) corresponding potentials, the RDFs (d), and the density profiles (e).

4.5 Conclusions

In conclusion, it has been seen that one can derive various CG models for the air-water interface via simple parameter fitting of analytical potentials to various system properties (e.g. RDF, density profile, and surface tension) as well as a model which provides a compromise between them (fig. 4.11). While it is easy to converge for properties like the density profile of the system, the structure of the interface is more complex to reproduce as one cannot fine tune the potential without introducing more complexity (i.e. more parameters) which eventually leads to problems in the convergence of the algorithm. Furthermore, for optimisations of multiple properties (simultaneously), it is difficult to derive a penalty function which takes into consideration the equal weighting of all properties at each iteration, without inhibiting the natural flow of the algorithm. In the next chapter, a similar CG model for the air-water interface is derived via an iterative procedure similar to IBI (i.e. IBI extended for inhomogeneous systems). The use of numerical potentials should improve the accuracy of the CG model as well as facilitate the coarse-graining procedure since neither an appropriate potential form nor as many initial guesses as for the Downhill Simplex algorithm are required to start the parametrisation.

5

Structure-based coarse-graining for slab systems

In principle, coarse-graining via parameter fitting (ch. 4) is a useful approach to develop a CG model based on a broad range of parameters with a minimum amount of manual interference. Here, the Downhill Simplex algorithm provides an attractive fitting tool as it is robust, easy to implement, and does not require the computation of derivatives. However, the amount of input required to start the algorithm can become problematic. First, one needs to provide an appropriate functional form and parameters to be optimized. The more complex the analytical potential is (i.e. the more parameters it contains), the slower the convergence of the algorithm will be. Second, accurate starting guesses which yield the property of interest need to be provided which may not always be available, and third, one requires an efficient penalty function to evaluate the quality of the results. As has been seen in sec. 4.4, some properties such as the slab radial distribution function (RDF), are very sensitive to changes in the interaction potential and can be difficult to optimize.

Hence, another technique of coarse-graining is sought which needs fewer input parameters, does not require a functional form for the interaction potential, and which can reproduce the structure of the slab more accurately. Here, an iterative scheme such as IBI is ideal as it only needs to compute the PMF for an initial guess, it employs numerical potentials, and it is intrinsically designed to converge to the RDF of the underlying reference. However, as it has been derived for homogeneous systems, some modifications need to be made to extend its applicability to slab (inhomogeneous) systems. Such a method could, in fact, be useful, for the systematic design of CG models for a wide range of phenomena which take place at interfaces (i.e. the aqueous/organic interface in biological cells) or in systems with phase boundaries.

Much debate exists about the true structure of water [117, 118, 119, 120, 121, 122]. Initially, it was believed that the structure in the bulk phase and that at the air-water interface are very different, since the hydrogen bonding network at the surface is interrupted. Shen et al. proposed that the region between the bulk and the vacuum phase consists of to two bands, one containing 'ice-like' water in which mainly tetrahedrally coordinated water can be found, and one containing 'gas-like' water [123, 124]. A new unified molecular view of the air-water interface, however, indicates that there is no 'ice-like' water but that the water molecules at the surface simply have stronger hydrogen bonding with their neighbors. These findings have indeed shown to have excellent agreement between theory and experiment [125] and as such, support the idea that a CG model for a water slab can be designed in which all particles (in bulk and at the interface) reproduce the local structure of the liquid.

In this chapter, a new coarse-graining method to develop CG models for liquid slabs is derived [126]. First, an analytical relation between the RDF of a (homogeneous) liquid and that of a slab (inhomogeneous system) is derived (sec. 5.1). One can use this transformation to design a new update for the IBI procedure which accounts for the system's inhomogeneity (sec. 5.2). The new method is demonstrated on a water slab and also tested on a slab of methanol. In addition, it can be applied to solute-solvent systems to obtain the correct partitioning (i.e. solute probability distribution in the solvent slab). An example for this is demonstrated by coarse-graining a single benzene molecule at the surface of a water slab (sec. 5.3).

5.1 Relation between bulk and slab RDFs

In a homogeneous system, the radial distribution function, $g(r)$, is calculated by counting the number of particles, $N_{\text{shell}}(r)$, in a spherical shell of a radius r and thickness Δr. This number is normalized by the shell volume, $V_{\text{shell}}(r) = 4\pi r^2 \Delta r$, and the number density, $\rho = N/V$, to ensure that $g(r) = 1$ at large r. To improve the accuracy, $g(r)$ is then averaged over all particles in the system

$$g(r) = \frac{\langle N_{\text{shell}}(r) \rangle}{\rho V_{\text{shell}}(r)} , \tag{5.1.1}$$

where $\langle \dots \rangle$ denotes an ensemble average.

If the same protocol is applied to a system consisting of a slab of thickness w, sandwiched between two vacuum layers in a box of length $l_s \geq 2w$ (see fig. 5.1a), two things change. First, the number density used for normalization, $\rho_s = N/V_s$, becomes smaller due to the larger size of the simulation box in the z-direction, l_s, and second, the number of particles in the shell is no longer uniform and becomes a function of the position of the shell center, a. This number can be estimated

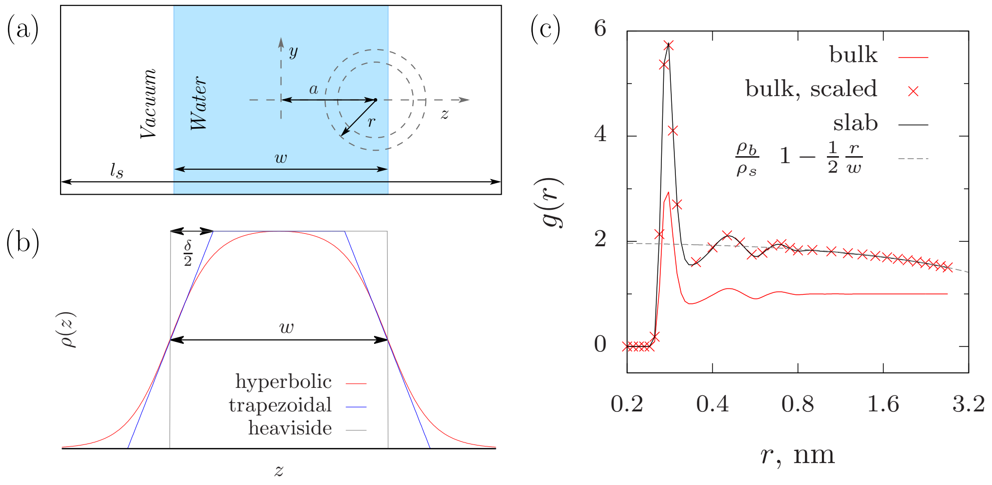


Figure 5.1: (a) Schematic diagram of a liquid slab of width w in a simulation box of length l_s. A partially filled shell used for RDF calculations is shown. (b) Density profiles, $\rho(z)$. The hyperbolic function, eq. 5.2.2, is used to fit simulation data to determine w, δ, and ρ_0. Trapezoidal and Heaviside functions are used to obtain the ratio of RDFs calculated in a slab and in bulk. (c) Radial distribution functions of atomistic water calculated in the bulk and slab systems, as well as the slab RDF obtained using eq. 5.1.4. The ratio between the RDFs is also shown for a density profile with sharp interfaces (eq. 5.1.4).

by assuming that only the density, $\rho(z)$, but not the local structure, $g_b(r)$, of the liquid changes within the slab (as a function of the z-coordinate).

Employing this assumption, the number of particles in the shell for a system with an interface can be estimated as

$$N_{\text{shell}}(r,a) = 2\pi r \Delta r g_b(r) \int_{a-r}^{a+r} dz \rho(z)\,, \tag{5.1.2}$$

where $g_b(r)$ is the RDF of a pure bulk system (without interfaces), $\rho(z)$ is the number density which depends on the z-coordinate, and a is the distance from the shell center to the symmetry plane of the slab (as shown in fig. 5.1b). Averaging over all particles, which is equivalent to integrating $N_{\text{shell}}(r,a)$ over $\rho(a)da$ and appropriately normalizing, one obtains

$$g_s(r) = g_b(r) \frac{\int_{-l_s/2}^{l_s/2} da \rho(a) \int_{a-r}^{a+r} dz \rho(z)}{2r\rho_s \int_{-l_s/2}^{l_s/2} da \rho(a)}\,. \tag{5.1.3}$$

Here, $g_s(r)$ is an object which one obtains by using a standard procedure of calculating an RDF for a system with a slab. In what follows, it is referred to as a *slab* RDF.

To illustrate how eq. 5.1.3 works in practice, consider first a slab with two sharp interfaces. In this case, $\rho(z)$ can be written as a sum of two Heaviside step functions (fig. 5.1b) and, for $r \leq w$, eq. 5.1.3 simplifies to

$$g_s(r) = g_b(r)\frac{\rho_b}{\rho_s}\left(1 - \frac{r}{2w}\right) \,, \tag{5.1.4}$$

where $w < l_s$ is the width of the slab and $\rho_b = N/V_b$ is the number density of the bulk system. One can see that apart from the local structure of the liquid, $g_s(r)$ also contains information about the slab width and total number density.

The transformation between the bulk and slab RDFs (eq. 5.1.4) is illustrated for atomistic SPC/E water in fig. 5.1c. Here, $g_b(r)$ obtained from simulations of bulk water, when scaled, lies exactly on top of $g_s(r)$ obtained from simulations of a slab of water. Note, that the slab system was prepared from the bulk system by increasing the box size in the z direction by a factor of two, such that $\rho_b/\rho_s = 2$.

Slice-resolved RDFs

To verify the assumption that only the density, $\rho(z)$, but not the local structure, $g_b(r)$, of the liquid changes within the slab, the RDFs of thin slices, $g_{\text{slice}}(r)$, of width Δz have been calculated in different regions of the water slab. Since each slice is simply a slab with two sharp interfaces, $g_{\text{slice}}(r)$ was rescaled by a factor

$$\kappa(r) = \frac{l_s}{\Delta z}\begin{cases} 1 - \frac{1}{2}\frac{r}{\Delta z}, & r \leq \Delta z \\ \frac{1}{2}\frac{\Delta z}{r}, & r > \Delta z \,, \end{cases} \tag{5.1.5}$$

which stems from the relation between the slab and bulk RDFs calculated for a slab with two sharp interfaces (tab. 5.2.2). These RDFs are shown in fig. 5.2 for a set of selected slices. One can see that in both the bulk and the slab cases, the deviation from the bulk RDF is only noticeable for the outermost interfacial layer ($\approx 1-2$ molecules thick). After this, it perfectly reproduces the bulk RDF.

5.2 IBI update for slab systems

5.2.1 Approximation for sharp interfaces

Employing the relation between the slab and bulk RDFs (eq. 5.1.4), one can rewrite the potential update of the IBI method for homogeneous systems,

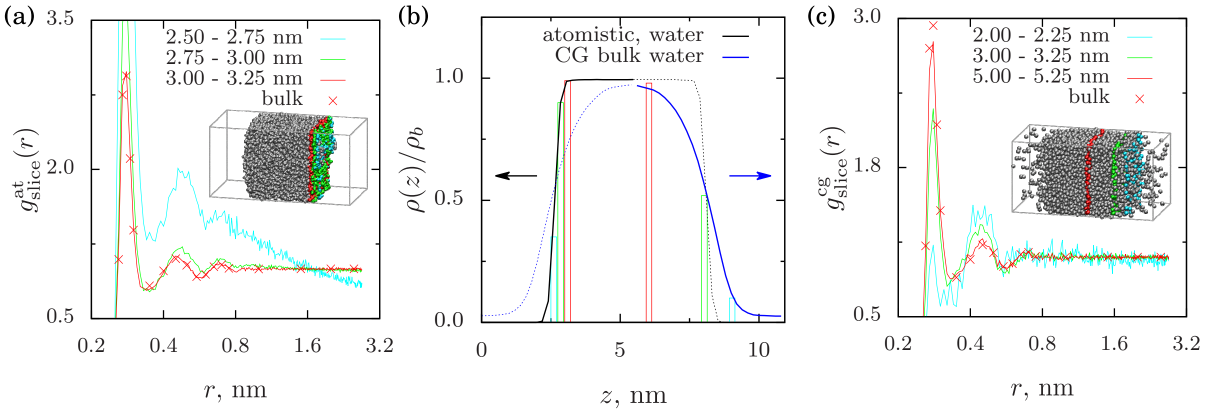


Figure 5.2: Slice-resolved and appropriately scaled radial distribution functions, $\kappa(r)g_{\text{slice}}(r)$ for (a) a slab of atomistic water (with sharp interfaces) and (c) a slab of coarse-grained, using IBI in bulk, water (wide interfaces). The corresponding density profiles are shown in (b), together with the slices, which are also depicted in the insets of (a) and (c). A slice width of $\Delta z = 0.25\,\text{nm}$ was used for RDF calculations.

eq. 2.2.10, for a slab system. Indeed, substituting eq. 5.1.4 one obtains

$$\frac{\Delta U^{(i)}}{k_{\text{B}}T} = \ln \frac{g_b^{(i)}(r)}{g_b^{\text{at}}(r)} = \ln \frac{g_s^{(i)}(r)}{g_s^{\text{at}}(r)} - \ln \frac{2w^{(i)} - r}{2w^{\text{at}} - r}\,, \tag{5.2.1}$$

where all constant offsets to the potential have been neglected. Here, one can see that $\Delta U^{(i)}$ splits into two contributions. The first depends on the ratio of the coarse-grained to atomistic *slab* RDFs, and the second is a function of the widths of the atomistic and coarse-grained slabs.

The derived update (eq. 5.2.1) might give the impression that it is sufficient to match *bulk* RDFs in order to satisfy $w^{\text{cg}} = w^{\text{at}}$. That is, however, incorrect, since structural coarse-graining does not preserve all thermodynamic properties of a system [63, 106, 127], among them the interfacial width of the atomistic reference. As already seen from conventional IBI in the bulk with pressure correction (sec. 4.1), simulations of a water slab using this model produced very wide and diffuse interfaces which, even though *bulk* RDFs were perfectly matched, resulted in different *slab* RDFs. In other words, $w^{\text{cg}} \neq w^{\text{at}}$ and $g_s^{\text{cg}}(r) \neq g_s^{\text{at}}(r)$. Note that this effect is especially severe in the case of a one-site coarse-grained *water* model, as it lacks three-body contributions to the potential of mean force [127]. A systematic solution to this problem is to either include a three-body interaction potential in the one-site water model [127] or to switch to a different representation which can reduce the role of such contributions, e.g. by mapping more than one water molecule onto a CG bead [84, 91, 109].

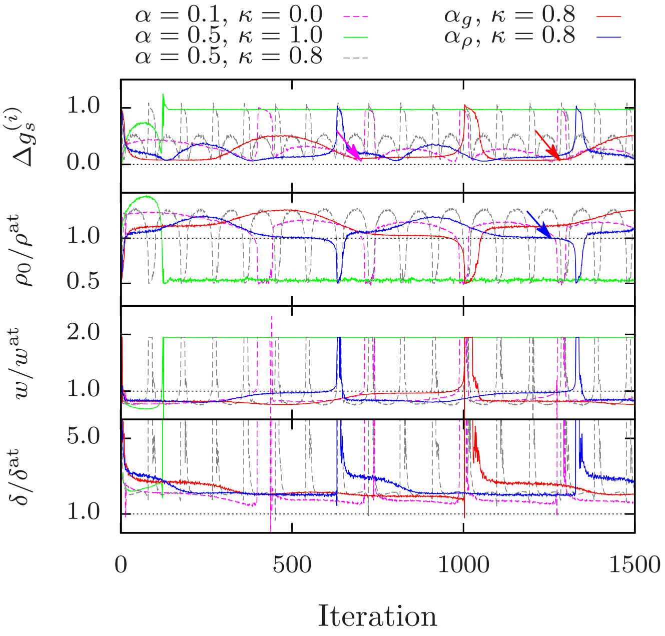


Figure 5.3: Convergence for the slab RDF, Δg_s, maximum of the density in a slab, ρ_0, width of the slab, w, and width of the interface, δ, all normalized by the reference values taken from atomistic simulations of SPC/E water, $\rho^{\text{at}} = 994.858\,\text{kg/m}^3$, $w^{\text{at}} = 5.4\,\text{nm}$, and $\delta^{\text{at}} = 0.43\,\text{nm}$.

If, however, a one-site representation with a pair interaction potential is still desired, one should try to find a balance between matching the pair distribution function and the thermodynamic properties of interest. Here, eq. 5.1.4 offers a practical solution as one can perform all simulations in a system with a slab, determine interfacial and slab widths, δ and w, calculate slab RDFs, $g_s(r)$, and subsequently change the interaction potential according to eq. 5.2.1. This idea, however, has both technical and conceptual issues. Conceptually, the Henderson theorem [64] states that there exists a unique (up to an additive constant) interaction potential which reproduces a given RDF of a homogeneous system. Based on experience, this is the stationary point of the IBI method. Since this statement does not depend on the density of the system, identical RDFs ($g_b(r)$) can correspond to different interaction potentials if the respective densities are different. The immediate implication is that IBI is now performed on an inhomogeneous system using the exact form of eq. 5.2.1, the system will eventually become homogeneous with the RDF reproducing that of the atomistic reference but an interaction potential which corresponds to a lower density (determined by the size of the simulation box). This is illustrated in fig. 5.3 (green solid line), where the convergence of the slab RDF (Δg), bulk density (ρ_0), interfacial and slab widths (δ, w) are shown as a function of iteration. These parameters have been extracted from simulations by fitting a sum of two hyperbolic tangent functions, adjusted to account for the slab drift, z_0, and a finite concentration of particles outside the slab, ρ_v,

$$\rho(z) = \rho_v + (\rho_0 - \rho_v)\frac{1 - \tanh\frac{2z-2z_0+w}{\delta}\tanh\frac{2z-2z_0-w}{\delta}}{1+\tanh^2\frac{w}{\delta}} \, . \tag{5.2.2}$$

By solving eq. 5.1.3, analytical expressions can be obtained to analyze density profiles and construct updates for the interaction potentials (see tab. 5.1). By following the convergence, it can be seen that independently from the initial density profile, the system quickly becomes homogeneous and the reference RDF is perfectly matched ($\Delta g = 0$ nm).

However, a spatially homogeneous density distribution is not what is desired. In this case, eq. 5.2.1 also offers a solution. Since its second term depends only on the density profile, one can introduce a bias which will make the homogeneous solution unstable. The simplest way of doing this is via a scaling factor for the density-dependent term, e.g. in the case of sharp interfaces, eq. 5.2.1, the potential update becomes

$$\frac{\Delta U^{(i)}}{k_\mathrm{B}T} = \ln\frac{g_s^{(i)}(r)}{g_s^{\mathrm{at}}(r)} - \kappa\ln\frac{2w^{(i)}-r}{2w^{\mathrm{at}}-r}, \quad \kappa < 1 \, . \tag{5.2.3}$$

This scaling effectively adds a long-range attractive term to the potential once the interface dissociates and hence destabilizes the homogeneous solution, leading to oscillations between the homogeneous and inhomogeneous density distributions. Such oscillations can be seen in fig. 5.3 for $\kappa = 0.8$ (gray dashed line) which show that the homogeneous density distribution is no longer a stationary point of the iterative scheme, since the structure- and density- dependent components of the update are unbalanced.

To find a compromise between optimizing the liquid structure and the density profile, one can slow down the oscillation dynamics by making a global scaling factor α in eq. 2.2.10 dependent on the convergence of the tail of the slab RDF, $\alpha_g = |1 - g(r_c)/g^{\mathrm{at}}(r_c)|$, where r_c is the cutoff distance. Similarly, one could ensure that the bulk densities of the slab match by introducing $\alpha_\rho = |1 - \rho_0/\rho_0^{\mathrm{at}}|$. The respective behavior of the iterative scheme and the properties of the CG models is shown in fig. 5.3 ($\kappa = 0.8$, α_g and α_ρ, red and blue solid lines). The

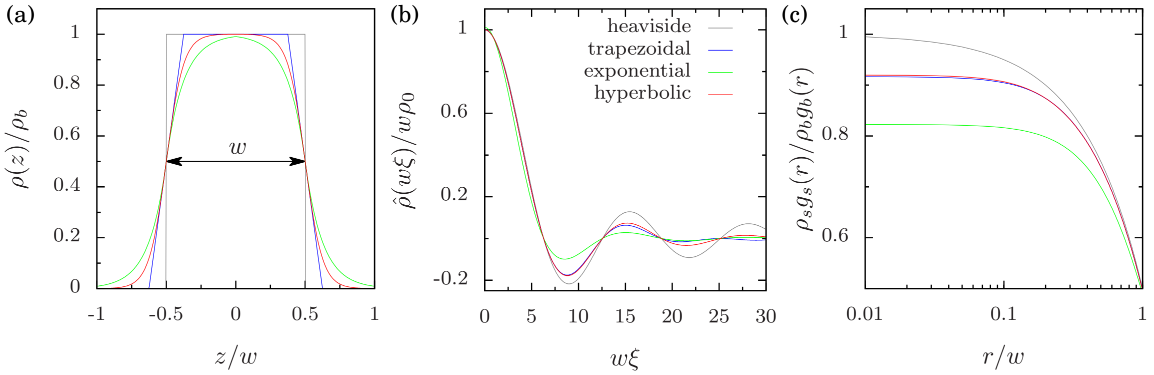


Figure 5.4: (a) Density profiles, (b) their fourier transforms, and (c) relations between the slab and bulk RDFs for all functions listed in tab. 5.1. $\delta/w = 0.25$.

"pseudo-stationary" points (depicted by the arrows) can be clearly identified and the corresponding potentials which, as will be seen in later examples (sec. 5.3), provide a good compromise between reproducing the thermodynamic and structural properties of a liquid.

5.2.2 Approximation for diffuse interfaces

It is not surprising that eq. 5.1.4 works perfectly for a water slab, since its interfaces are very sharp. For more diffuse interfaces it is still possible to obtain an analytical relation between the RDFs by approximating the slab density profile with more complicated functional forms (e.g. trapezoidal shape, fig. 5.1b, tab. 5.1, and fig. 5.4),

$$g_s(r) = g_b(r)\frac{\rho_b}{\rho_s}\begin{cases} 1 + \dfrac{r^3}{12w\delta^2} - \dfrac{r^2}{3w\delta} - \dfrac{\delta}{3w} & r < \delta \\ 1 - \dfrac{r}{2w} - \dfrac{\delta^2}{12rw} & \delta \le r < w \end{cases}, \tag{5.2.4}$$

where δ is the width of the interface. Note, that eq. 5.2.4 generalizes eq. 5.1.4 to a slab with diffuse interfaces.

5.2.3 Exact update using the Fourier transform

While approximations with the Heaviside of the Trapezoidal shapes for the density profile are useful to understand and test the two parts of the update, in practice it only suffices to use the Fourier update in which an exact representation of the density profile is employed. In the derived integral (eq. 5.1.3),

$$I(r) = \int_{-l_s/2}^{l_s/2} da \rho(a) \int_{a-r}^{a+r} dz \rho(z)\,, \tag{5.2.5}$$

it is assumed that the particle density vanishes sufficiently far from the slab interfaces, such that all particles are (on average) located within the slab. Under this condition, the integral over da can be extended to cover all space $[-\infty, \infty]$. Furthermore, a new variable, $t = z - a$, can be introduced. By changing the order of integration, one obtains

$$I(r) = \int_{-r}^{r} dt \int_{-\infty}^{\infty} da\, \rho(a)\rho(t+a) = \int_{-r}^{r} dt\, h(t)\,, \tag{5.2.6}$$

where $h(t)$ is the correlation of $\rho(a)$. Using the correlation (convolution) theorem, one can write

$$\hat{h}(\xi) = \hat{\rho}(\xi)^2\,, \tag{5.2.7}$$

where the hat notation implies the Fourier transform of the function

$$\hat{\rho}(\xi) = \int_{-\infty}^{\infty} da\, \rho(a) e^{-i\xi a}\,. \tag{5.2.8}$$

Back-transforming $\hat{h}(\xi)$ and integrating over t, one obtains

$$I(r) = \frac{1}{\pi} \int_{-\infty}^{\infty} d\xi\, \hat{\rho}^2(\xi) \frac{\sin(r\xi)}{\xi}\,. \tag{5.2.9}$$

Hence, the problem has now been reduced to simply calculating the Fourier transform of the density profile and evaluating the integral (eq. 5.2.9). These have been calculated for several common functional forms of the density profile, which are summarized in tab. 5.1 together with their Fourier transforms and the resulting relations between the bulk and slab RDFs.

Note, that it is implicitly assumed that the density of particles vanishes at the box boundaries, i.e. there is enough vacuum around the slab such that one can ignore effects arising from periodic boundary conditions.

Inclusion of periodic boundary conditions

On the technical side, eq. 5.2.1 is not exact and does not take into account

Type	$\rho(z)/\rho_0$	$\hat{\rho}(\xi)/\rho_0$	$\rho_s g_s(r)/\rho_b g_b(r)$
Heaviside	$\begin{cases} 0 & -\infty \le z < -\frac{1}{2}w \\ 1 & -\frac{1}{2}w \le z < \frac{1}{2}w \\ 0 & \frac{1}{2}w \le z \le \infty \end{cases}$	$\dfrac{2}{\xi} \sin \dfrac{w\xi}{2}$	$\begin{cases} 1 - \frac{1}{2}\frac{r}{w} & r \le w \\ \frac{1}{2}\frac{w}{r} & r > w \end{cases}$
Trapezoidal	$\begin{cases} 0 & -\infty \le z < -\frac{1}{2}(w+\delta) \\ \frac{1}{2}\left(1 + \frac{w+2z}{\delta}\right) & -\frac{1}{2}(w+\delta) \le z < -\frac{1}{2}(w-\delta) \\ 1 & -\frac{1}{2}(w-\delta) \le z < \frac{1}{2}(w-\delta) \\ \frac{1}{2}\left(1 + \frac{w-2z}{\delta}\right) & \frac{1}{2}(w-\delta) \le z < \frac{1}{2}(w+\delta) \\ 0 & \frac{1}{2}(w+\delta) \le z \le \infty \end{cases}$	$\dfrac{4}{\delta\xi^2} \sin \dfrac{\delta\xi}{2} \sin \dfrac{w\xi}{2}$	$\begin{cases} 1 + \frac{r^3}{12w\delta^2} - \frac{r^2}{3w\delta} - \frac{\delta}{3w} & r < \delta \\ 1 - \frac{r}{2w} - \frac{\delta^2}{12rw} & \delta \le r < w \end{cases}$
Exponential [114]	$\dfrac{e^{\frac{w}{\delta}}}{2e^{\frac{w}{\delta}} - 1} \begin{cases} \exp \frac{w+2z}{\delta} & -\infty \le z < -\frac{1}{2}w \\ 2 - \exp \frac{-w-2z}{\delta} & -\frac{1}{2}w \le z < 0 \\ 2 - \exp \frac{-w+2z}{\delta} & 0 \le z < \frac{1}{2}w \\ \exp \frac{w-2z}{\delta} & \frac{1}{2}w \le z \le \infty \end{cases}$	$\dfrac{4\left(4e^{\frac{w}{\delta}} \sin \frac{\xi w}{2} + \xi\delta\right)}{\left(2e^{\frac{w}{\delta}} - 1\right) \xi(\delta^2\xi^2 + 4)}$	
Hyperbolic	$\dfrac{1 - \tanh \frac{2z+w}{\delta} \tanh \frac{2z-w}{\delta}}{1 + \tanh^2 \frac{w}{\delta}}$	$\dfrac{\delta}{\tanh \frac{w}{\delta}} \dfrac{\sin\left(\frac{\xi w}{2}\right)}{\sinh\left(\frac{\xi\delta}{2}\right)}$	

Table 5.1: Density profiles, their Fourier transforms, and relations between RDFs calculated in a slab and in bulk. For the exponential profile the analytical expression for the relation between RDFs exists but is too lengthy to be shown here. From the practical point of view, fitting with a hyperbolic profile and using the relation between the RDFs provided by the trapezoidal profile suffices. All profiles, their Fourier transforms, and scaling relations are also shown in fig. 5.4. Here $\rho_0 = \rho(z = 0)$.

periodic boundary conditions present in MD simulations. This issue can, however, be improved by employing more realistic fitting functions for the density profile, or, ultimately, the Fourier series.

Suppose that during a single iteration step, the interface becomes so diffuse that the concentration of particles at the box boundaries ($z = -l_s/2$ and $z = l_s/2$) is no longer zero. In this case, the link between $g_b(r)$ and $g_s(r)$ must account for periodic boundary conditions (PBC). In other words, one can no longer change the range of integration from $[-l_s/2, l_s/2]$ to $[-\infty, \infty]$ in eq. 5.1.3 and Fourier *series* must be used. Expanding the density profile, one obtains

$$\rho(z) = \sum_{n=-\infty}^{\infty} a_n \exp\left(2\pi i n \frac{z}{l_s}\right) \,, \tag{5.2.10}$$

$$a_n = \frac{1}{l_s} \int_{-l_s/2}^{l_s/2} dz\, \rho(z) \exp\left(-2\pi i n \frac{z}{l_s}\right) \,. \tag{5.2.11}$$

By performing similar steps as in sec. 5.2.3 and taking into account that $\rho(-z) = \rho(z)$, the integral becomes

$$I(r) = l_s^2 \sum_{n=-\infty}^{\infty} \frac{a_n^2}{\pi n} \sin \frac{2\pi n r}{l_s} \,. \tag{5.2.12}$$

To illustrate the effect of having a finite concentration of particles in vacuum, the functional form for the slab with two sharp interfaces will be used

$$\rho(z) = \begin{cases} \rho_v & -l_s/2 \le z < -\frac{1}{2}w \\ \rho_0 - \rho_v \left(\frac{w_0 - w}{l_s - w}\right) & -\frac{1}{2}w \le z < \frac{1}{2}w \\ \rho_v & \frac{1}{2}w \le z \le l_s/2 \end{cases} \tag{5.2.13}$$

where ρ_v is the particle concentration in vacuum. Note that the total number of particles, that is, the integral over $z \in [-l_s, l_s]$ is $\rho_0 w_0$, where w_0 denotes the width of the slab when $\rho_v = 0$ (no particles in the vacuum).

Evaluating eq. 5.2.11, the corresponding Fourier coefficients are

$$a_n = \rho_0 \frac{l_s - w_0}{l_s - w} \frac{\sin(\pi n \frac{w}{l_s})}{\pi n} \,. \tag{5.2.14}$$

The relation between slab and bulk RDFs, $\rho_s g_s(r)/\rho_b g_b(r) = I(r)/(\rho_0^2 w_0 2r)$, is shown in fig. 5.5 for several widths of the slab, w/w_0, which correspond to different

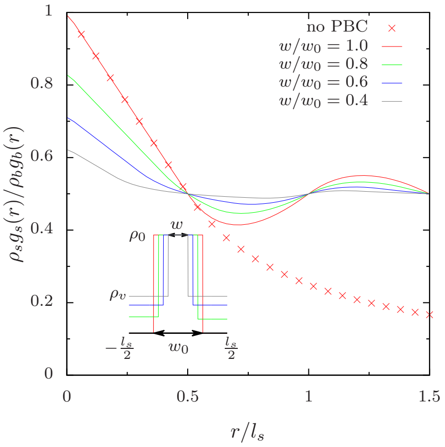


Figure 5.5: Ratio between the slab and bulk RDFs for a slab with two sharp interfaces, eq. 5.2.13. Two cases are shown: without taking into account periodic boundary conditions (no PBC), eq. 5.1.4, and with PBC for different concentrations of particles outside the slab, ρ_v (or different slab widths, w). The difference for $r > l_s/2$ is due to the slab periodic image and exist even if there are no particles outside the slab ($w = w_0$).

densities of particles in the vacuum. One can see that even when $w = w_0$ (i.e. all particles are part of the slab), the relation for a system with PBC starts to deviate from that with two infinite vacuum layers, starting at $r > l_s/2$. This is expected since at this point, the spherical shell of radius r starts to touch the periodic image of the slab. However, this is of no relevance for constructing the IBI update, as long as $g(r)$ is calculated only until the cutoff distance, $r_{\text{cut}} < l_s/2$.

If $w \neq w_0$, the deviation can already be seen at small r and, strictly speaking, one should calculate the update using the expansion in eq. 5.2.11, which is the simplest way of constructing the potential update as it only relies on the Fourier expansion coefficients of the density profile, ρ_n, does not require fitting of the simulated density profile, and avoids assumptions about the shape of the interface when constructing the update. This is especially useful in the case of solute-solvent systems, where the distribution of the positions of the solute, $p(z)$, is a tabulated function of the z coordinate. In this case, the update can be calculated as

$$\frac{g_s(r)}{g_b(r)} = \frac{l_s}{2rp_0\rho_s} \sum_{-\infty}^{\infty} \frac{\rho_n p_{-n}}{\pi n} \sin \frac{2\pi n r}{l_s} = \tag{5.2.15}$$

$$= \frac{\rho_0}{\rho_s} + \frac{l_s}{r\rho_s p_0} \sum_{n=1}^{\infty} \frac{\rho_n p_n^* + \rho_n^* p_n}{\pi n} \sin \frac{2\pi n r}{l_s} ,$$

where p_n are the Fourier expansion coefficients of $p(z)$.

5.3 Examples

5.3.1 Slab of SPC/E water

Putting this into practice, IBI was performed on a slab of SPC/E water using $g_s^{\mathrm{at}}(r)$ as a target and the pressure-corrected potential for bulk water as an initial guess. All simulations were performed with the GROMACS simulation package [54], while the coarse-graining tools (mapping, IBI) stem from the VOTCA package [105]. Atomistic simulations were performed in a box of approximately $5.5 \times 5.5 \times 11\,\mathrm{nm}$, containing 5439 type SPC/E water molecules. The GROMOS53a6 force field was used to describe inter- and intramolecular interactions. For electrostatic interactions, the smooth PME [73] method (sec. 2.1.1) was employed using a real space cutoff of 1 nm, while the van der Waals cutoff was set to 1.4 nm. The temperature of $T = 300\,\mathrm{K}$ was regulated by the Berendsen thermostat [45] with a coupling constant of $\tau = 0.1\,\mathrm{ps}$. Atomistic reference simulations were run for $t = 20\,\mathrm{ns}$ with a time step of $\delta t = 2\,\mathrm{fs}$, while CG simulations were performed for $t = 100\,\mathrm{ps}$ (per iteration) with the same time step. Fourier series and a scaling factor of $\kappa = 0.8$ were used to construct the potential updates (with α_g and α_ρ, sec. 5.2).

To analyze simulations, the density profile, $\rho(z)$, was fitted to a sum of two hyperbolic tangents (see fig. 5.1b) at each iteration step. Fitted parameters, namely the height of the slab density profile, ρ_0, the width of the slab, w, and the width of the interface, δ, were monitored over the course of the simulation and are shown in fig. 5.3, together with the convergence for the slab RDF, Δg_s. The points with matching slab RDFs at r_{cut} or bulk densities were identified (these are shown by the red and blue arrows) and the corresponding interaction potentials, RDFs, and density profiles are summarized in fig. 5.6. Note, that if a “naive” substitution of the IBI update is used (by simply matching the slab RDF, $\kappa = 0$) the convergence also oscillates. However, one can terminate iterations once the RDFs match and obtain a reasonable fit of the RDF as well as the density profile to those of the atomistic reference (fig. 5.3, 5.6, pink dashed line). Note, that this potential was used in sec. 4.4 as a fitting reference for the CKDg analytical potential to obtain an accurate starting guesses.

One can see that, depending on the potential update (α_g or α_ρ scaling), either the slab RDF or the density profile is better reproduced. This implies that a cer-

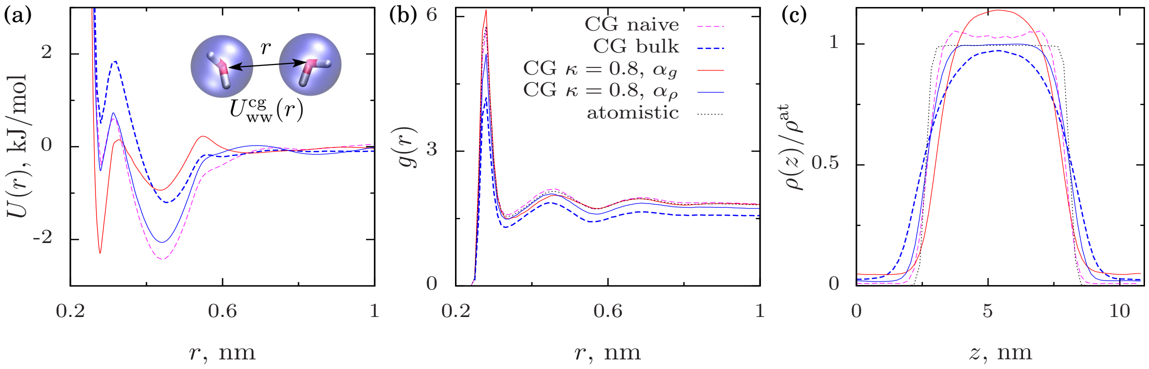


Figure 5.6: (a) Water-water pair interaction potentials, $U(r)$, (b) radial distribution functions, $g(r)$, and (c) slab density profiles, $\rho(z)$, for systems coarse-grained in the bulk (CG bulk), in a slab using the update from eq. 5.2.1 using slab RDFs (CG $\kappa = 0.8$, α_g), or densities (CG $\kappa = 0.8$, α_ρ) indicated by a red and blue arrows in fig.). For RDFs and density profiles the atomistic reference is also shown (atomistic).

tain compromise has been achieved between the local structure, characterized by $g_s(r)$, and the geometry of the interface, represented by w and δ. In other words, the two terms in the potential update (eq. 5.2.1) compensate each other within some numerical accuracy. Remarkably, in both cases the density profile and the slab RDF are significantly better in agreement with the atomistic reference than those obtained by using interaction potentials based on the IBI procedure in the bulk system.

We have also calculated the surface tension (eq. 4.3.3). The atomistic reference system has a surface tension of $\gamma^{\mathrm{at}} = 59.4 \pm 0.3$ mN/m. All CG models reproduce this value reasonably close, $\gamma_{\mathrm{bulk}}^{\mathrm{CG}} = 57.5 \pm 0.3$ mN/m, $\gamma_{\alpha_g,\kappa=0.8}^{\mathrm{CG}} = 45.6 \pm 0.9$ mN/m, and $\gamma_{\alpha_\rho,\kappa=0.8}^{\mathrm{CG}} = 31.7 \pm 1.3$ mN/m. However, as already mentioned, the surface tension, is a value which can also be obtained by integrating an expression which includes the RDF, the slope of the density profile, and the derivative of the interaction potential [114] and as such, it is possible to have the same value of γ for very different $\rho(z)$ and $g(r)$. Therefore, one should not validate the CG model on surface tension alone.

5.3.2 Slab of liquid methanol

Next, a slab of methanol is investigated which, in spite of being in a liquid state, exhibits significant differences in behavior as compared to water. First, its liquid-vacuum interface is much wider ($\delta^{\mathrm{at}} \approx 0.6$ nm) than that of water due to the substitution of one hydrogen by a methyl group. Second, a CG model with one site per molecule may be able to reproduce both structural and thermodynamic

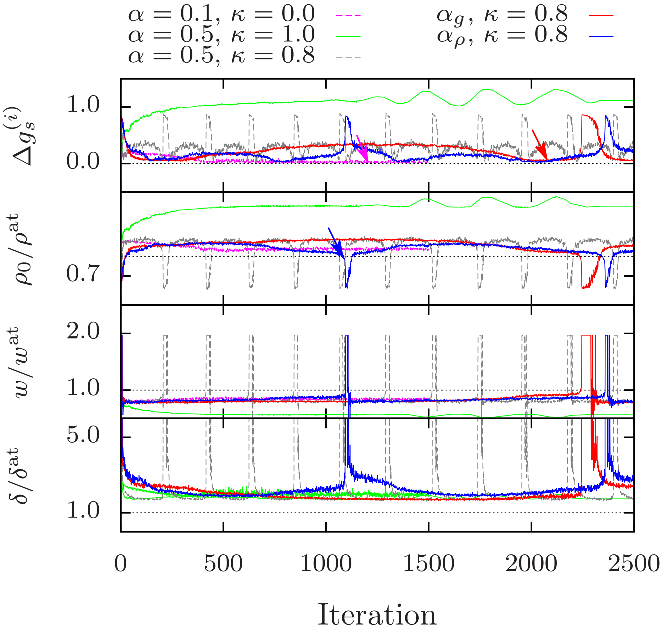


Figure 5.7: Convergence for the slab RDF, Δg_s, maximum of the density in a slab, ρ_0, width of the slab, w, and width of the interface, δ, all normalized by the reference values taken from atomistic simulations of SPC/E water, $\rho^{\mathrm{at}} = 994.858\,\mathrm{kg/m^3}$, $w^{\mathrm{at}} = 5.4\,\mathrm{nm}$, and $\delta^{\mathrm{at}} = 0.43\,\mathrm{nm}$.

properties of the bulk system better than the corresponding CG model of water, since similar coarse-grained potentials can be obtained from force-matching or conventional iterative Boltzmann inversion [105].

In spite of these differences, the dynamics of convergence is qualitatively very similar to water: stationary homogeneous density distribution for $\kappa = 1.0$, oscillations between homogeneous and heterogeneous (with a slab) states for $\kappa = 0.8$, and slow oscillations for both α_g and α_ρ (fig. 5.7). The respective interaction potentials, $U^{\mathrm{cg}}_{\mathrm{mm}}(r)$, radial distribution functions, and density profiles are shown in fig. 5.8a-c. Again, by coarse-graining in a slab, one can achieve much steeper density profiles, matching slab widths, while the reference and atomistic RDFs still agree with each other. Note, that if a naive substitution of the IBI update is used ($\kappa = 0$) and the iterations are terminated once the slab RDF is perfectly matched, the density profile is not reproduced well.

The surface tension of the atomistic reference, calculated using eq. 4.3.3, was $\gamma^{\mathrm{at}} = 21.3 \pm 0.4$ mN/m, while the CG models yielded values of $\gamma^{\mathrm{CG}}_{\mathrm{bulk}} = 20.3 \pm 0.2$ mN/m, $\gamma^{\mathrm{CG}}_{\alpha_g,\kappa=0.8} = 10.1 \pm 0.3$ mN/m, and $\gamma^{\mathrm{CG}}_{\alpha_\rho,\kappa=0.8} = 6.4 \pm 0.3$ mN/m. It has also been checked that the potentials obtained from the update using the Fourier series are transferable to systems of different slab widths (w) as can be

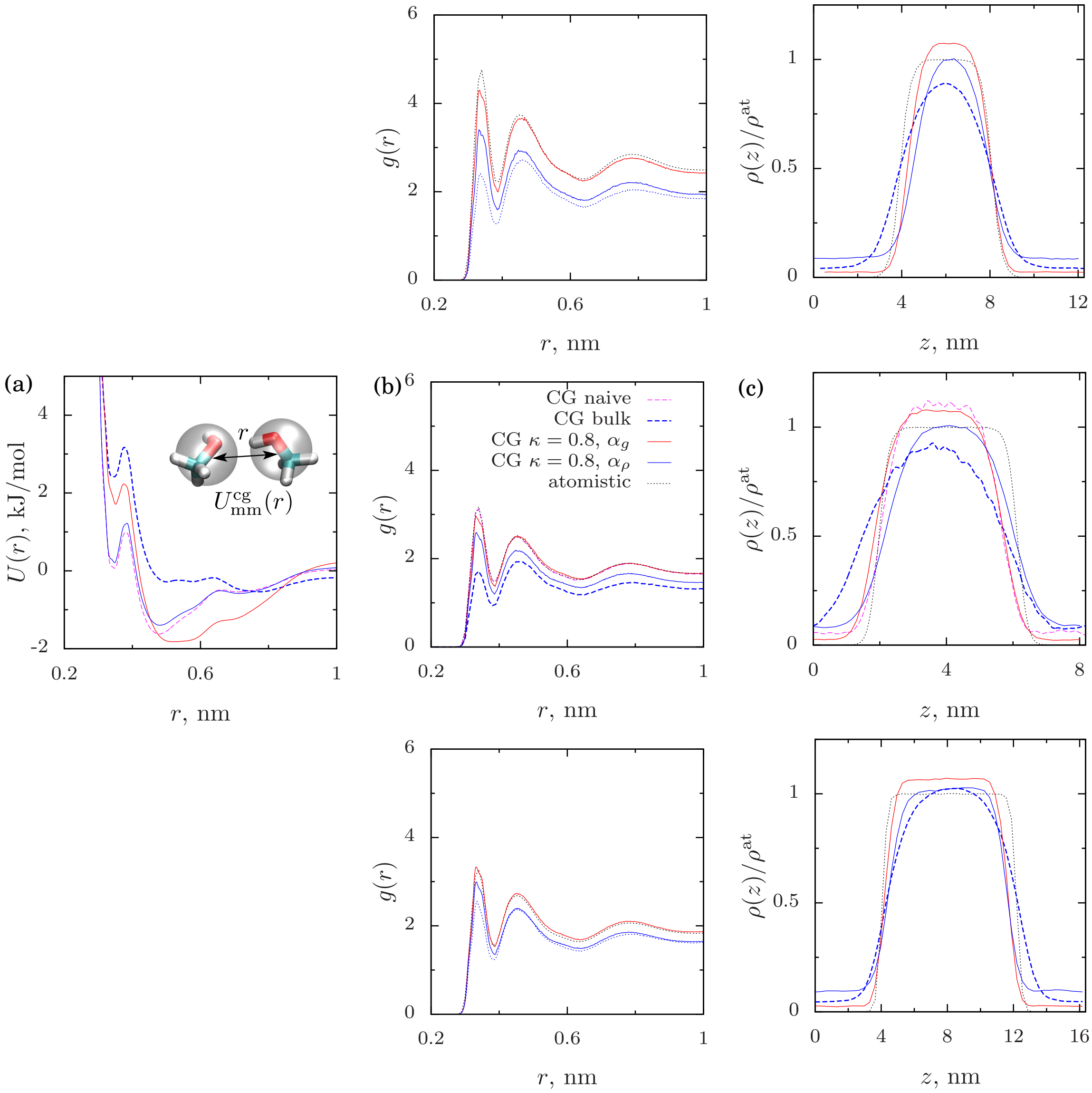


Figure 5.8: (a) Methanol-methanol (centers of mass) pair interaction potentials, $U(r)$, (b) radial distribution functions, $g(r)$, and (c) slab density profiles, $\rho(z)$ for systems coarse-grained in the bulk (CG bulk) and in a slab using the Fourier update (CG $\alpha = 0.5$, $\kappa = 1$). For RDFs and density profiles the atomistic reference is also shown (atomistic).

seen in fig. 5.8 (top and bottom for thinner and thicker slabs respectively). This is true as long as the interaction range is larger than w and there is a well defined plateau in the middle of the density profile.

5.3.3 Benzene in water

Now, a more complicated situation is described, that is, determining solute-solvent coarse-grained potentials. As an example, a single benzene molecule in water (i.e. dilute system) is considered. Benzene is a hydrophobic molecule and hence predominantly occupies the vacuum-water interface region. This has been confirmed by analyzing the distribution of its position, $p^{\mathrm{at}}(z)$, along the z coordinate from atomistic simulations, as shown in fig. 5.9c. One can see that this distribution is peaked around the positions of the two water-vacuum interfaces of the slab. Simulations details were the same as for the water slab, but with smaller simulation box sizes of $4 \times 4 \times 8$ nm. In addition, due to the low benzene concentration (1 benzene in 2175 water molecules), simulation times for both the atomistic references as well as the CG simulations have been increased to $t = 100$ ns and $t = 10$ ns (per iteration) respectively, to obtain better sampling. The benzene-water interaction was assigned a Van der Waals cutoff of 1.4 nm. If one now employs a usual IBI coarse-graining procedure (in bulk) and reproduces the bulk RDF between the benzene bead (which is treated as a single interaction site) and the surrounding water molecules, simulations in a slab with this potential result in benzene distributions which are peaked around the slab center (shown in fig. 5.9c).

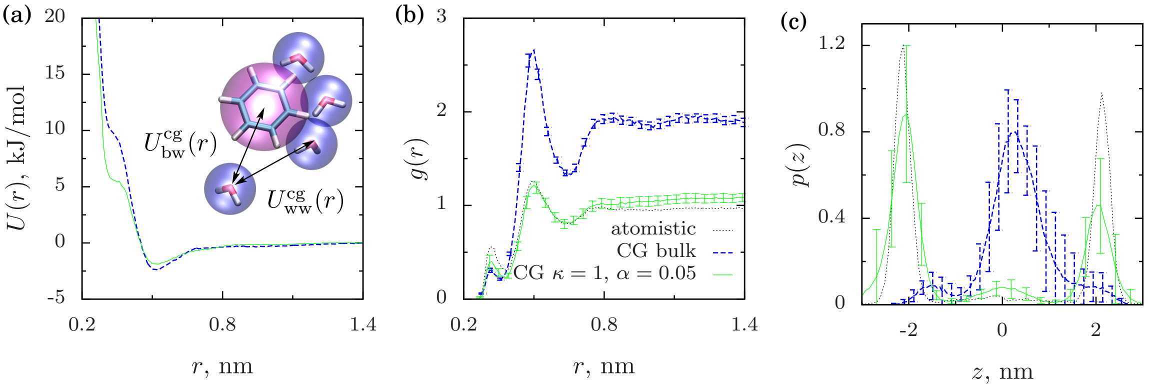


Figure 5.9: (a) Benzene-water pair interaction potentials, $U(r)$, (b) benzene-water radial distribution functions, $g(r)$ and (c) probability distribution to find a benzene molecule at a position z, $p(z)$. All plots show results of coarse-graining in the bulk (CG bulk) and in a slab using the Fourier-series update ($\kappa = 1$, $\alpha = 0.05$). For RDFs and probability distributions the atomistic reference is also shown (atomistic).

This means that the benzene bead mostly occupies the bulk water region, a circumstance which is directly reflected in the differences in *slab* RDFs of the coarse-grained and atomistic systems, as shown in fig. 5.9b. It arises due to the different probability distributions, $p^{\mathrm{cg}}(z)$ and $p^{\mathrm{at}}(z)$, of the benzene positions along the slab, since now, in order to average over the ensemble of benzene positions, one has to integrate N_{shell} over $p(a)da$ instead of $\rho(a)da$ in order to obtain the equivalent of eq. 5.1.3.

In other words, to perform an update of the interaction potential analogous to eq. 5.2.1, the benzene position along the slab was obtained and decomposed into a Fourier series in order to obtain the relation between g_s and g_b. Starting from an initial guess of a pressure-corrected potential for bulk water and applying this procedure we eventually recovered the shape of the atomistic slab RDF, as shown in fig. 5.9b.

By matching the slab RDFs, one automatically matches the distributions of the benzene position along the slab, as can be seen in fig. 5.9c. The benzene bead is now primarily located at the interfaces of the slab which coincides with its behavior in atomistic simulations. It is notable that the two coarse-grained potentials ($U^{\mathrm{cg}}_{\mathrm{bw}}(r)$) shown in fig. 5.9a, are remarkably similar: the one obtained by coarse-graining in a slab has a slightly more long-range attractive tail and smaller repulsive bump at the beginning than the one obtained from a bulk system. This small difference, however, is what leads to the distinct hydrophobic behavior of benzene when simulated in a slab.

5.4 Conclusions

Here, a scheme for obtaining coarse-grained potentials for inhomogeneous systems, analogous to the iterative Boltzmann inversion method has been proposed. The main idea is to construct an update for the interaction potential based on the radial distribution function calculated in a slab geometry. Apart from the local liquid structure, this update also carries information about the geometry of the system, namely the slab and interfacial widths. Since these geometric features are very sensitive to the thermodynamic properties of the system (surface tension, pressure tensor), a partially controllable balance between the local structure and thermodynamic properties can be achieved. This is of particular importance in solute-solvent systems in biology where the thermodynamic/interfacial behavior often needs to be included in addition to the structural properties of the system.

There are also drawbacks to the proposed method. First, simulation times for each iteration need to be increased to attain adequate accuracy when measuring thermodynamic properties of the system, such as a density profile or a distribution of positions of a solute within a solvent. Second, one needs to introduce a small scaling factor for the potential update in order to improve convergence and to avoid either evaporation or solidification of the system. Finally, while the form of the algorithm (i.e. two separate terms) would suggest that obtaining a controlled balance between the thermodynamic properties and the local structure is trivial, one cannot assign equal weights to both properties but needs to adjust the second term such that the update does not converge to the bulk result.

In spite of these limitations, the method offers a practical way of finding a compromise between the local structural and intensive thermodynamic properties of the system, and as such, can be useful for building simple coarse-grained models for inhomogeneous and open systems.

6

Coarse-grained model for an amphiphilic peptide

In recent years, the interest in the development of systematic coarse-graining approaches for modelling biological systems has grown rapidly, since processes such as protein folding or peptide aggregation involve large time and length scales which could not be addressed otherwise, for example, by atomistic MD simulations. One of the earliest reduced models designed to study protein folding are the Gō models [128] in which only those residues which correspond to contacts in the native state can interact, generating an energy landscape which resembles a weakly rugged funnel that points towards the native state. Although such coarse-grained models are very efficient in speeding up simulation times and in giving insight into the possible folding pathways of a protein, they provide a very rough description of the actual structure. For a more detailed description, so-called knowledge-based potentials [129] can be employed. Here, interaction energy functions can either be derived based on the analysis of known protein structures (e.g. crystal structures from the PDB[1]) [130, 131] to make secondary structure predictions or, if the protein structure is not available, based on atomistic simulations, as in the UNRES (UNited RESidue) model for polypeptide chains [132].

In this chapter, a coarse-grained (CG) model for the peptide PGlu-2 is developed based on atomistic simulations of smaller peptide *fragments*. These are determined according to a *mapping scheme* (fig. 6.1) which defines the positions and atomistic constituents of individual CG beads. Structure-based coarse-graining using this fragment-based approach[2] has already been applied in the conformational studies of diphenyl alanine [89] and oligoalanine [90] in bulk water. Hence, a similar mapping scheme is adopted here for the CG beads in common (i.e. the peptide group, the $C_{\alpha\beta}$ group, and the phenyl group). While bonded potentials are parametrized by Boltzmann inversion of the bond and angle distributions from atomistic MD simulations of the peptide at the air-water interface (sec. 6.1),

[1] The Protein data bank is a crystallographic database for proteins and nucleic acids.

[2] Fragments should have a comparable chemical and topological environments to their corresponding atomistic components. Hence, any bonds which extend to a neighboring CG bead are capped by methyl groups instead of hydrogens.

the parametrization of non-bonded potentials differs from that of previous studies as it tries to retain some of the thermodynamic properties in addition to the structure. It determines the radial distribution function (RDF) as well as the distribution of the bead's position from atomistic simulations of the peptide fragments in a *slab* system and tries to obtain a compromise between the two properties (sec. 6.2).

Once the CG model has been derived, a single CG peptide is simulated at the air-water interface and its affinity to diffuse away from the interface is monitored to test the fragment-based peptide model. A correct model should balance hydrophobic-hydrophilic interactions such that the peptide resides at the interface, as was the case in atomistic simulations (sec. 3.1).

6.1 Parametrization of bonded interactions

As initially it is only of interest whether the CG peptide with alternating hydrophobic-hydrophilic beads can align and stay at the air-water interface, the peptide PGlu-2 (of length $n = 2$) with its charged termini replaced by methyl groups (fig. 6.1) is simulated as an atomistic reference. This only requires 4 types of CG beads and should suffice to demonstrate to which extent a fragment-based coarse-graining approach is useful for these types of systems.

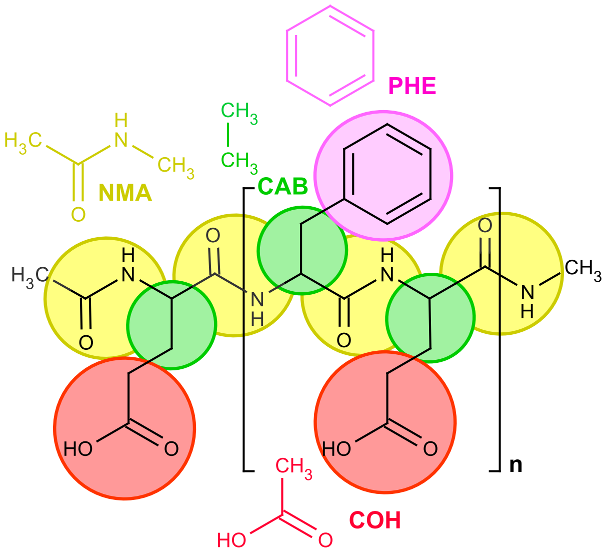


Figure 6.1: Mapping scheme for PGlu-n with the termini replaced by methyl groups, showing the backbone (NMA, CAB) and side-chain CG beads (PHE, COH), with their corresponding atomistic fragments in color.

With beads of NMA (**N**-**M**ethylacet**A**mide), CAB (**C**$_{\alpha\beta}$), PHE (**PHE**nyl), and COH (**COOH**, Carboxyl), 4 types of bond (NMA-CAB, CAB-NMA, CAB-PHE, CAB-COH), 6 types of angle (NMA-CAB-NMA, CAB-NMA-CAB, NMA-CAB-PHE, NMA-CAB-COH, PHE-CAB-NMA, COH-CAB-NMA), and 2 types of dihedral angle distributions [1] (NMA-CAB-NMA-CAB, CAB-NMA-CAB-NMA) are determined and the bonded potentials obtained via Boltzmann inversion (sec. 2.2.2).

However, as opposed to coarse-graining of, for example, polymers which are sampled in vacuum with exclusions, atomistic (reference) simulations of the peptide are performed in its solvent environment, as has been done in previous work for the developments of CG peptide models in the bulk [89, 90]. In this case, $t = 100\,\mathrm{ns}$ atomistic MD simulations were performed of a single peptide of length $n = 2$ as depicted in fig. 6.1 to calculate the distributions (fig. 6.4, black dotted lines).

6.2 Parametrization of non-bonded interactions

For the parametrization of the solute-solvent interaction potentials, atomistic MD simulations of $t = 100\,\mathrm{ns}$ with a time step of $\delta t = 2\,\mathrm{ps}$ have been performed in the NVT ensemble, at a temperature of $T = 300\,\mathrm{K}$ (Berendsen thermostat, $\tau = 0.1\,\mathrm{ps}$) with PME to treat electrostatics. Simulations have been performed in both, bulk water and a slab system (containing 2175 water molecules for the N-methylacetamide and benzene systems and 2176 water molecules for the ethane and acetic acid systems), with simulation box sizes of $4.0 \times 4.0 \times 4.0$ nm and $4.0 \times 4.0 \times 8.0$ nm respectively, to compare the two types of parametrizations schemes, CG$_{\mathrm{bulk}}$ and CG$_{\mathrm{slab}}$. As already seen from the parametrization of the hydrophobic bead (i.e. benzene in bulk water. sec. 5.3), reproducing the RDF in the bulk does not provide the correct density distribution of the bead which prefers to stay at the interface and a parametrization in a water slab is required. It is of interest whether this is also the case for the more hydrophilic beads.

To provide references for the CG beads NMA, CAB, PHE and COH, the atomistic fragments N-methylacetamide, ethane, benzene, and acetic acid have been used. For parameterizations in the bulk systems, iterations have been performed for $t = 5\,\mathrm{ns}$. Resulting potentials were used as initial guesses for the parametriza-

[1] Note, that only a minimum number of proper dihedrals has been used in the backbone to avoid cross-correlations between bonded potentials.

tion at the interface, which required $t = 10\,\mathrm{ns}$ per iteration to obtain sufficient statistics for an accurate calculation of the RDFs, $g(r)$, and the density distributions of the CG solute beads, $p(z)$, and solvent beads, $\rho(z)$. In order to avoid artifacts in the inversion of the radial distribution function, RDFs have only been calculated from r values at the onset of the first peak ($r_{\mathrm{min}} = 0.33\,\mathrm{nm}$ for NMA, $r_{\mathrm{min}} = 0.28\,\mathrm{nm}$ for CAB, $r_{\mathrm{min}} = 0.28\,\mathrm{nm}$ for PHE, and $r_{\mathrm{min}} = 0.29\,\mathrm{nm}$ for COH). A cutoff of $r_{\mathrm{cut}} = 1.4\,\mathrm{nm}$ was employed for all Van der Waals interactions between CG beads.

While parametrizations in the bulk employed the conventional IBI procedure (sec. 2.2.3), parametrizations at the interface used the extended version of IBI for inhomogeneous systems which was developed in ch. 5. This depends on the RDF calculated in a slab geometry, $g(r)$, and the density profiles of the solute bead and the water slab, $p(z)$ and $\rho(z)$, respectively. A compromise between the two properties can be steered by the scaling factors α and κ in the potential update,

$$\begin{aligned}\frac{\Delta U^{(i)}}{k_{\mathrm{B}}T} &= \alpha \left[\ln \frac{g_s^{(i)}(r)}{g_s^{\mathrm{at}}(r)} - \kappa \ln \frac{f^{(i)}}{f^{at}} \right] , \\ f &= \frac{\rho_0}{\rho_s} + \frac{l_s}{r \rho_s p_0} \sum_{n=1}^{\infty} \frac{\rho_n p_n^* + \rho_n^* p_n}{\pi n} \sin \frac{2\pi n r}{l_s} .\end{aligned} \tag{6.2.1}$$

Here, p_n and ρ_n are the Fourier expansion coefficients of $p(z)$ and $\rho(z)$ respectively, of which $n = 128$ were used. Each bead required between 30-50 iterations until a reasonable compromise between the structure and the density distribution was achieved. Note, that to speed up the parametrization process, a scaling of $\alpha = 0.05, \kappa = 1$ was used first, followed by a scaling of $\alpha = 0.05, \kappa = 0.4$ once the two update terms compensated one another.

Parametrization results are shown in fig. 6.2, with the obtained potentials (left), the RDFs (middle), and density distributions (right). Non-bonded potentials obtained from the parametrization in bulk water ($\mathrm{CG}_{\mathrm{bulk}}$, blue dashed line) show that although the atomistic structure of the bulk water is perfectly matched, the CG beads do not reproduce the correct $g(r)$ and $p(z)$ when placed in an interface environment. Here, all beads except for NMA reside mainly in the bulk water region (i.e. center of the slab).

For the parametrization at the interface ($\mathrm{CG}_{\mathrm{slab}}$, green solid line), however, the developed procedure is capable of finding a compromise between the structure

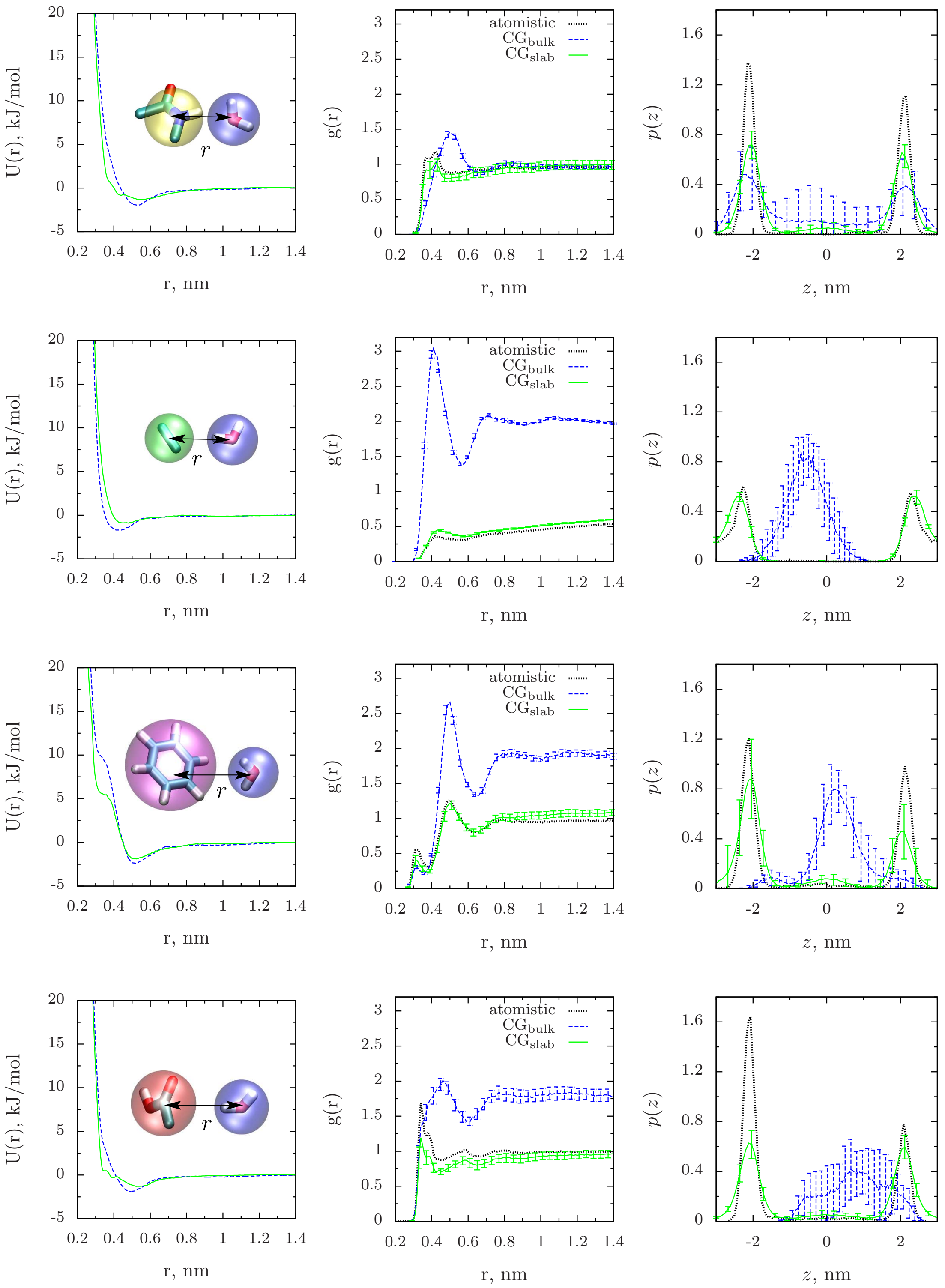


Figure 6.2: Parametrization results for non-bonded CG potentials, comparing the parametrization in bulk, CG_{bulk}, and at the interface, CG_{slab}, according to their potentials (left), their RDFs (center), and their density distributions (right). A snapshot of each CG bead (as mapped onto the atomistic fragment) is shown as it interacts with water.

and the affinity towards the interface.

Now, that both the bonded and non-bonded interaction potentials have been derived, the CG peptide can be assembled and its behavior in a water slab tested.

6.3 Testing the coarse-grained peptide model

Putting everything together, the first simulations can be performed to validate the new CG model. Here, a single PGlu-2 peptide is simulated in a water slab (with its initial positions at the interface) for $t = 10\,\text{ns}$ with a time step of $\delta t = 1$ ps, employing a Van der Waals cutoff of $r_{\text{cut}} = 1.4\,\text{nm}$ (fig. 6.3a). All other simulations settings are the same as in sec. 6.2.

Here, results show that the peptide diffused away from the interface into the bulk. This is at first sight surprising since even though all of the fragments stay at the interface when simulated individually, the combination of them in CG peptide does not do so. This indicates that the derived non-bonded potentials are not additive and a fragment-based parametrization as described can not be used to model amphiphilic peptides at interfaces on a coarse-grained level.

Since the presence of the interface is known to influence a fragment's orientation as well as a peptide's conformations, one can imagine that sampling of individual fragments in a water slab is not representative of the beads' orientations in a chain. To test this, one of the backbone beads (NMA) in the CG$_{\text{slab}}$ model has been re-parameterized in a *trimer* to obtain a better "intra-chain" representation for the CG bead. Here, N-ethylacetamide was simulated as an atomistic reference for $t = 100\,\text{ns}$ and the RDF was calculated in a slab geometry for the NMA fragment. Next, a coarse-grained trimer of CAB-NMA-CAB was prepared for parametrization of the NMA-water non-bonded interaction potential. Here, non-bonded potentials for CAB-water interactions were substituted from resulting potentials of the previous monomer parametrizations respectively. Starting from the initial guess of the NMA-water interaction (parametrized from a trimer in bulk water), approximately 50 iterations lead to a match in RDFs (fig. 6.3a). Results show, that the obtained interaction potential between the NMA bead and water has a deeper minimum than in the monomer case.

Substituting this new potential for the non-bonded NMA-water interaction and repeating the CG test simulation with the same simulation settings as before, the peptide now remained at the interface (fig. 6.3b). This implies that in addition to accounting for the bead's affinity towards the interface, the proper av-

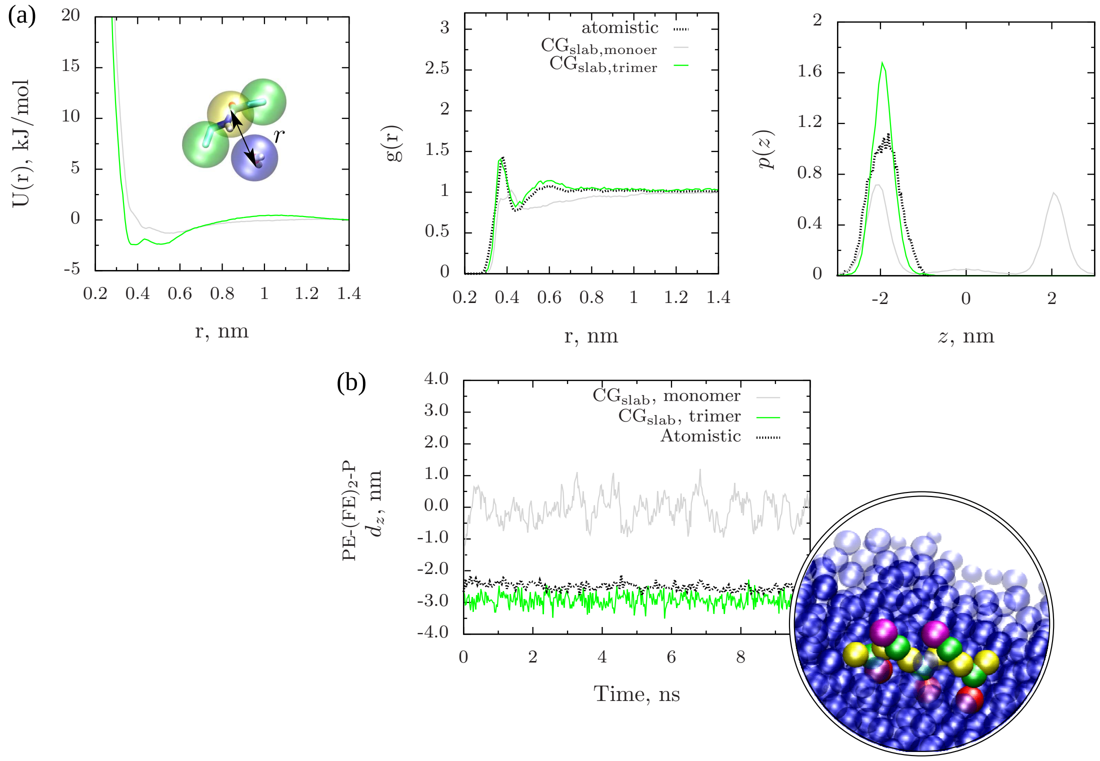


Figure 6.3: CG model results, showing the peptides affinity towards the interface. Here, the NMA bead which was initially parametrized by a monomer was reparametrized in a trimer CAB-NMA-CAB (a), which lead the peptide to remain at the air-water interface (b). A snapshot of the CG peptide at the interface is shown on the right.

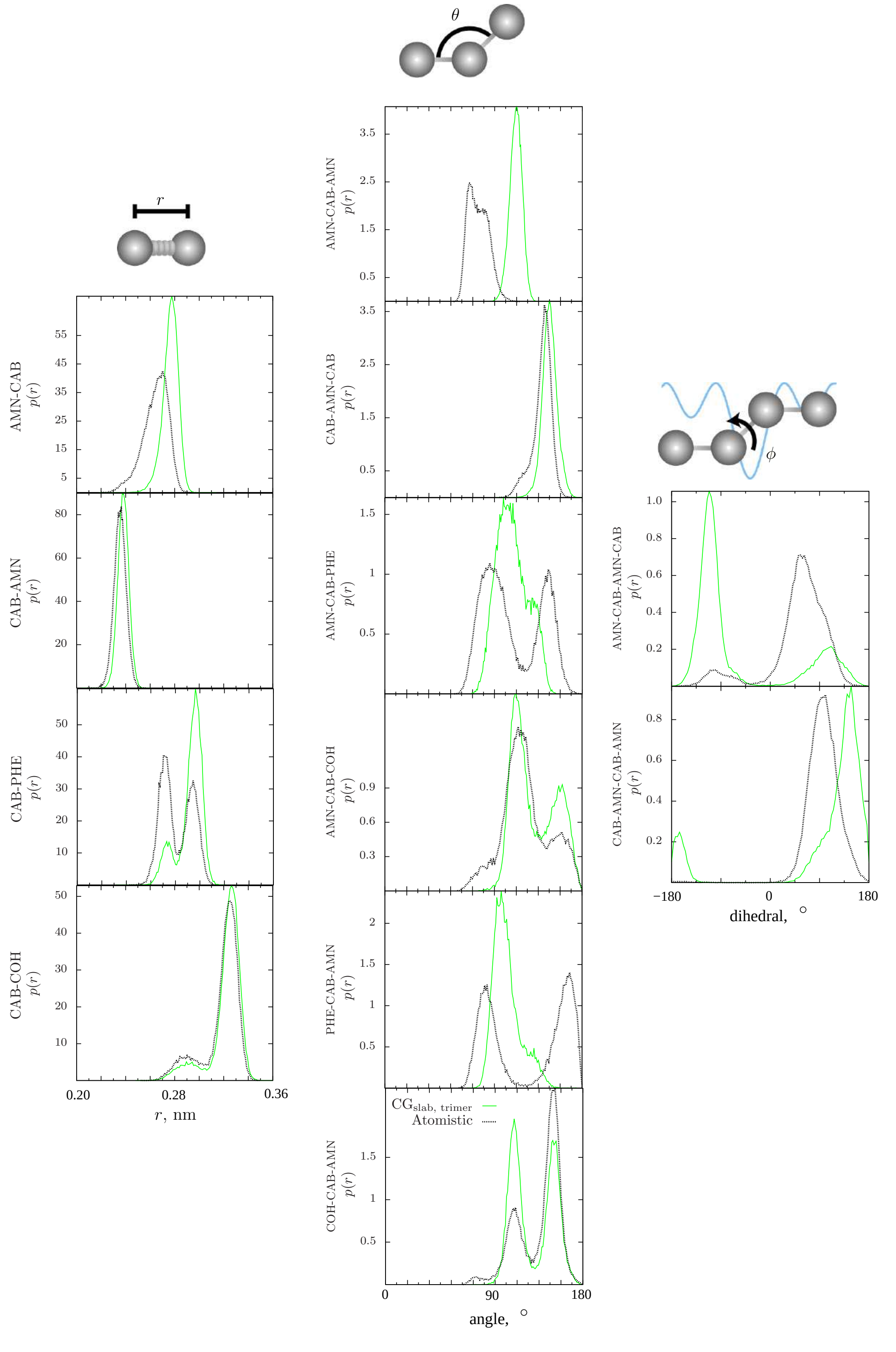


Figure 6.4: Bond (left), angle (middle), and dihedral (right) distributions obtained from the simulation of the CG peptide model ($\mathrm{CG}_{\mathrm{slab,trimer}}$, green solid line) which remained at the interface throughout the entire simulation time ($t = 10\,\mathrm{ns}$), compared to the distributions obtained from the atomistic reference simulation (black dotted line).

eraging over a bead's orientation during sampling needs to be considered. Finally, computing the bond, angle, and dihedral angle distributions, one sees that they only qualitatively agree with those of the atomistic reference (fig. 6.4). Hence, the model still needs to be refined, possibly by including bonded interactions into the iterative procedure, in order to be able to deduce any structural information from coarse-grained simulations.

6.4 Conclusions

In conclusion, a CG model for an amphiphilic peptide has been developed via fragment-based approach. For each single fragment, the procedure developed in ch. 5 yielded solute-solvent bead potentials which reproduced both the local structure as well as the solvation behavior of the atomistic fragment (otherwise not reproduced by conventional IBI in the bulk). When assembled in a CG peptide, however, the affinity towards the interface was not reproduced, with the CG peptide diffusing into the bulk. This indicated that the orientational sampling of the fragment (biased by the interface) needs to be taken into account. After reparametrizing one of the backbone beads (NMA) in a trimer, the CG peptide indeed stayed at the interface. Hence, one can conclude that a fragment-based approach needs to be adjusted to take into account the constraints imposed by a heterogeneous environment.

7 Conclusions

In this thesis, a series of amphiphilic peptides, PGlu-n, PAsp-n, and PheGlu-n, for $n = 2, 4, 5$ have been studied, which are designed to self-assemble into monolayers at the aqueous-organic interface. These can serve as template matrices to promote the crystallization of hydroxyapatite in the presence of ions and can thus be applied in the field of tissue engineering to treat diseases such as osteoporosis. To verify experimental hypotheses and obtain a better understanding of the structure and interactions which govern the self-assembly process on the microscopic level, computer simulations have been employed to complement existing experimental results.

In chapter 3, small peptide systems have been investigated via atomistic simulations. Here, MD simulations showed that peptides behaved very differently in terms of backbone extension when simulated in the bulk and at the air-water interface. While mostly extended backbone conformations could be observed in the bulk, stable β-hairpins were formed at the interface which greatly outnumbered the amount of extended conformations. However, these MD simulations were subject to severe sampling problems and when REMD simulations were employed, an approximately equal amount of β-hairpin and extended conformations could be found for short peptides ($n = 2$), while the longer peptides ($n = 5$) still displayed mainly β-hairpins at the interface since these become more stable as the length of the peptide (n) (i.e. the number of possible hydrogen bonds between backbone residues) increases. In addition, various peptide sequences have been compared by both MD and REMD simulations which displayed different conformational characteristics. These could later be linked to the peptide's tendency to aggregate from studies of assembling peptides and pre-assembled aggregates. It was discovered that peptides with longer acidic side-chains (i.e. Glu vs. Asp) were slower to self-assemble due to the sidechains' interactions with other backbone groups. Once shorter side-chains have been employed, the peptide has been seen to aggregate much faster and in a more ordered manner, where the aggregates have even been stabilized by hydrogen bonding between the acidic side-chains of neighboring peptides. The main hydrogen bonding contribution, however, was shown to arise from H-bonds between backbone residues (O and H of amide), which form a hydrogen bonding network whose regularity determines the stability of the monolayer. In addition, it was seen that an intrinsic twist in the peptide's backbone also restricts the length of the aggregates formed as it breaks the regular hydrogen bonding network. Similarly, the Proline termini (i.e. Pro vs. Phe), which have been said to be β-sheet breakers play an important role in ordering peptides in 2D within the aggregate and regulating the length of the aggregates. As any larger simulations (with realistic peptide concentrations and system sizes) would be infeasible an atomistic level of detail, a coarse-grained model has been developed to reach those time scales of interest.

In chapter 4, it was first tested whether a CG water model parametrized by conventional structure-based coarse-graining (IBI) in the bulk was transferable to a situation at the air-water interface. This, however, lead to very diffuse interfaces and hence, one needs other methods to derive a CG model which retains a stable interface. Next, it was attempted to systematically build up a CG strategy, starting from simple potentials in order to find out which features of the interaction potential preserve those atomistic properties which stabilize the interface (for a pair potential and a 1 bead/molecular mapping for water). To optimize analytical potentials, the Downhill Simplex method was employed in order to avoid unnecessary searching of the parameter space. The potential which was tested first was the LJ12-4 potential, which when optimized to fit the density profile of the atomistic reference showed that although one can perfectly reproduce the shape of the interface with the CG model, the resulting potential also leads to a system which is much too over-structured or interfaces which are too diffuse. To improve this, a long-range attractive tail was added to the potential, which demonstrated that the diffuse interface could now be stabilized by a tuning parameter for the attraction, resulting in a radial distribution function (RDF) closer to that of the atomistic reference. In a final attempt to fit the first two coordination shells (packing) and to obtain a correct bead size, a Gaussian function was added to the potential to produce a (structurally relevant) second minimum, which indeed led to a much better reproduction of the RDF in the first two coordination shells. Finally, to obtain a CG model which can compromise between reproducing the structure and the density profile of the atomistic reference, the Downhill Simplex algorithm was used to tune the potential. The resulting model reproduced the proper bead size and provided a good fit of the first coordination shell of the RDF as well as of the density profile.

However, as coarse-graining via parameter fitting is rather empirical, relies on a large set of fitting parameters, and is limited in terms of the flexibility of the shape of the potential, an alternative approach was sought in chapter 5. Here, the conventional IBI method for homogeneous systems was extended to inhomogeneous systems by deriving a transformation between the bulk and slab RDF, based on the assumption that as one moves to the slab's interface, only the density and not the local structure of the liquid changes. The derived potential update was found to consist of two parts, one which depends on the water structure in the slab and the second which depends on the density profile (i.e. slab and interfacial widths). Various approximations for the density profile have been made, starting from a simple Heaviside function for sharp interfaces, to a trapezoidal shape for more diffuse interface, to finally an exact (even with periodic boundaries) expression which employs the fourier transform coefficients to describe the density profile of the system. The state point to which this update converges, however, is that of a homogeneous bulk liquid. As it is not possible to stabilize the update without either ending in the gas or solid phase, different prefactors to the two parts of the update were employed which slow down the update at the desired state points yielding a stable interface with a suitable structure. This CG method was applied to coarse-grain a slab of water, a slab of methanol, as well as a benzene molecule (i.e. hydrophobic CG bead of the peptide) in a water slab. The last example is especially relevant for the solute/solvent interactions here, as the CG model should not only the reproduce the structure but also the correct partitioning in the solvent environment (i.e. affinity towards to interface). Employing the same procedure, all non-bonded interaction potentials for the remaining beads were parametrized in different water models (structure- and thermodynamics based), with bonded potentials derived via Boltzmann inversions of the distribution from REMD simulations at the interface.

Finally, in chapter 6, the new coarse-graining method is employed to derive a simple CG peptide model via a fragment-based approach. This was tested in combinations with different CG water models to find out which combination led the peptides to remain at the air-water interface. The reference (atomistic) peptide was similar to PGlu-2 (ch. 3), except that the charged termini have been replaced by methyl groups, as initially it is only of interest whether the CG peptide is able to align at the air-water interface. With parametrizations of non-bonded potentials purely based on monomers of individual peptide fragments, CG models eventually lead to the peptide diffusing into the bulk. However, when the peptide in the structure-based water model was reparametrized by using a molecular trimer (i.e. to obtain a better “in-chain” representation of the bead), the CG peptide showed the same interfacial behavior compared to atomistic simulations. Hence, a structure-based CG approach for peptides at the interface is capable of providing CG interaction potentials if the parametrization is performed in exactly the same environment (i.e. at the interface) and the affinity of the beads towards staying at the interface (monitored via the probability distributions to find a fragments at a specific position along the slab) are accounted for.

Future work still includes the parametrization of the charged termini as in experiment as well as fine tuning of the CG model by trying to achieve a better compromise between structural and thermodynamical properties. Once this has been achieved, effective peptide-peptide interactions will be derived based on PMF calculations of peptide fragments in a solvent environment. On a method development side, ways to stabilize the algorithm for inhomogeneous systems without driving the system into the gas or the solid phase need to be improved. Ideally, one would like to have a CG procedure which requires little user input and is transferable to other systems and applications.

With these tools at hand, one will then able to simulate the time and length scales relevant to experiment to study the formation and stability of the monolayers. By reintroducing atomistic details at the various stages (via a back-mapping procedure), one may then obtain further insight into the assembly behaviors of different peptide sequences and select the one most suitable for applications in tissue engineering.

Appendix

A.1 Supplementary material

A.1.1 REMD analysis

System	*Exchange pair*	*Acceptance ratio (bulk)*	*Acceptance ratio (interface)*
PGlu-2	1 ↔ 2	0.030	0.035
	2 ↔ 3	0.031	0.039
	3 ↔ 4	0.030	0.036
	4 ↔ 5	0.032	0.037
	5 ↔ 6	0.031	0.041
	6 ↔ 7	0.033	0.041
	7 ↔ 8	0.037	0.038
	8 ↔ 9	0.034	0.041
	9 ↔ 10	0.038	0.038
	10 ↔ 11	0.039	0.039
	11 ↔ 12	0.039	0.040
	12 ↔ 13	0.042	0.039
	13 ↔ 14	0.043	0.041
	14 ↔ 15	0.044	0.043
	15 ↔ 16	0.048	0.043
PGlu-5	1 ↔ 2	0.074	0.074
	2 ↔ 3	0.076	0.074
	3 ↔ 4	0.076	0.075
	4 ↔ 5	0.073	0.080
	5 ↔ 6	0.076	0.077
	6 ↔ 7	0.075	0.081
	7 ↔ 8	0.078	0.076
	8 ↔ 9	0.077	0.078
	9 ↔ 10	0.084	0.084
	10 ↔ 11	0.078	0.076
	11 ↔ 12	0.082	0.078
	12 ↔ 13	0.082	0.078
	13 ↔ 14	0.082	0.077
	14 ↔ 15	0.084	0.077
	15 ↔ 16	0.083	0.079
	16 ↔ 17	0.085	0.084
	17 ↔ 18	0.087	0.077
	18 ↔ 19	0.089	0.079
	19 ↔ 20	0.087	0.083
	20 ↔ 21	0.087	0.082
	21 ↔ 22	0.092	0.086
	22 ↔ 23	0.090	0.083
	23 ↔ 24	0.088	0.083
	24 ↔ 25	0.092	0.081
	25 ↔ 26	0.095	0.087
	26 ↔ 27	0.097	0.078
	27 ↔ 28	0.100	0.080
	28 ↔ 29	0.100	0.088
	29 ↔ 30	0.097	0.089
	30 ↔ 31	0.099	0.085
	31 ↔ 32	0.103	0.087

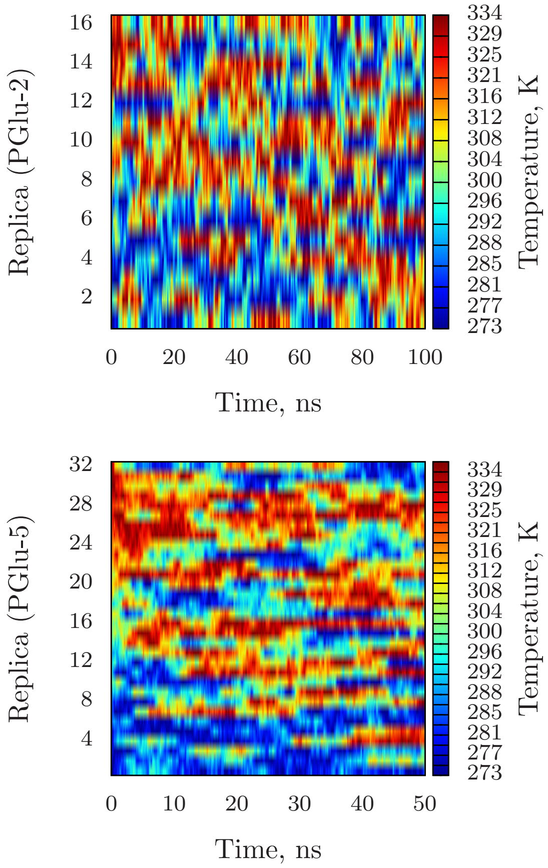
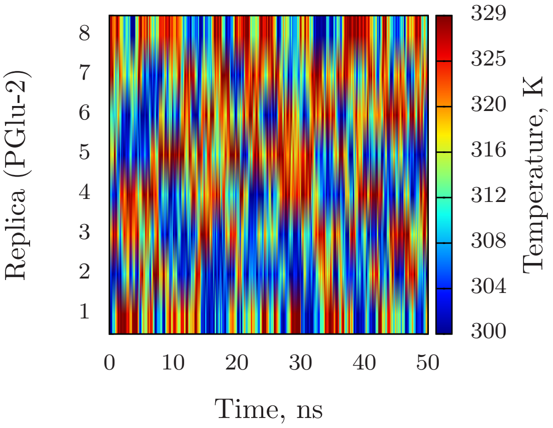
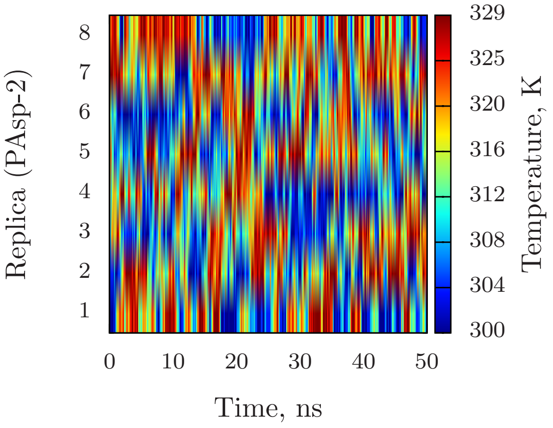


Figure A.1: Acceptance ratio for REMD simulations for PGlu-2 and PGlu-5, comparing conformations obtained from peptide simulations in the bulk and the air-water interface (fig. 3.3).

System	*Exchange pair*	*Acceptance ratio* $(interface)$
PGlu-2	1 ↔ 2	0.026
	2 ↔ 3	0.026
	3 ↔ 4	0.026
	4 ↔ 5	0.026
	5 ↔ 6	0.027
	6 ↔ 7	0.026
	7 ↔ 8	0.026
PAsp-2	1 ↔ 2	0.023
	2 ↔ 3	0.025
	3 ↔ 4	0.023
	4 ↔ 5	0.024
	5 ↔ 6	0.026
	6 ↔ 7	0.026
	7 ↔ 8	0.025
PGlu-2	1 ↔ 2	0.025
	2 ↔ 3	0.023
	3 ↔ 4	0.027
	4 ↔ 5	0.027
	5 ↔ 6	0.027
	6 ↔ 7	0.024
	7 ↔ 8	0.026





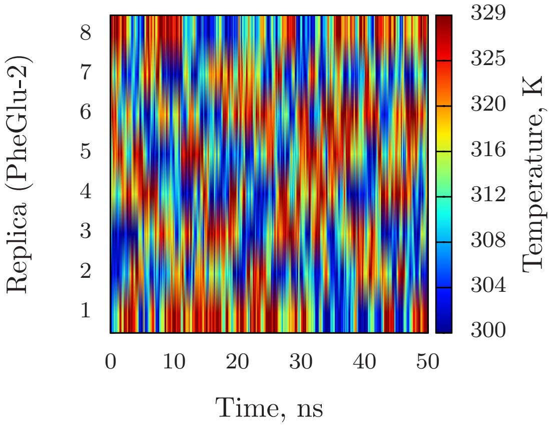


Figure A.2: Acceptance ratio for REMD simulations for PGlu-2, PAsp-2, and PheGlu-2 at the air-water interface, comparing conformations obtained for different peptide sequences (fig. 3.4).

A.1.2 Secondary structure analysis

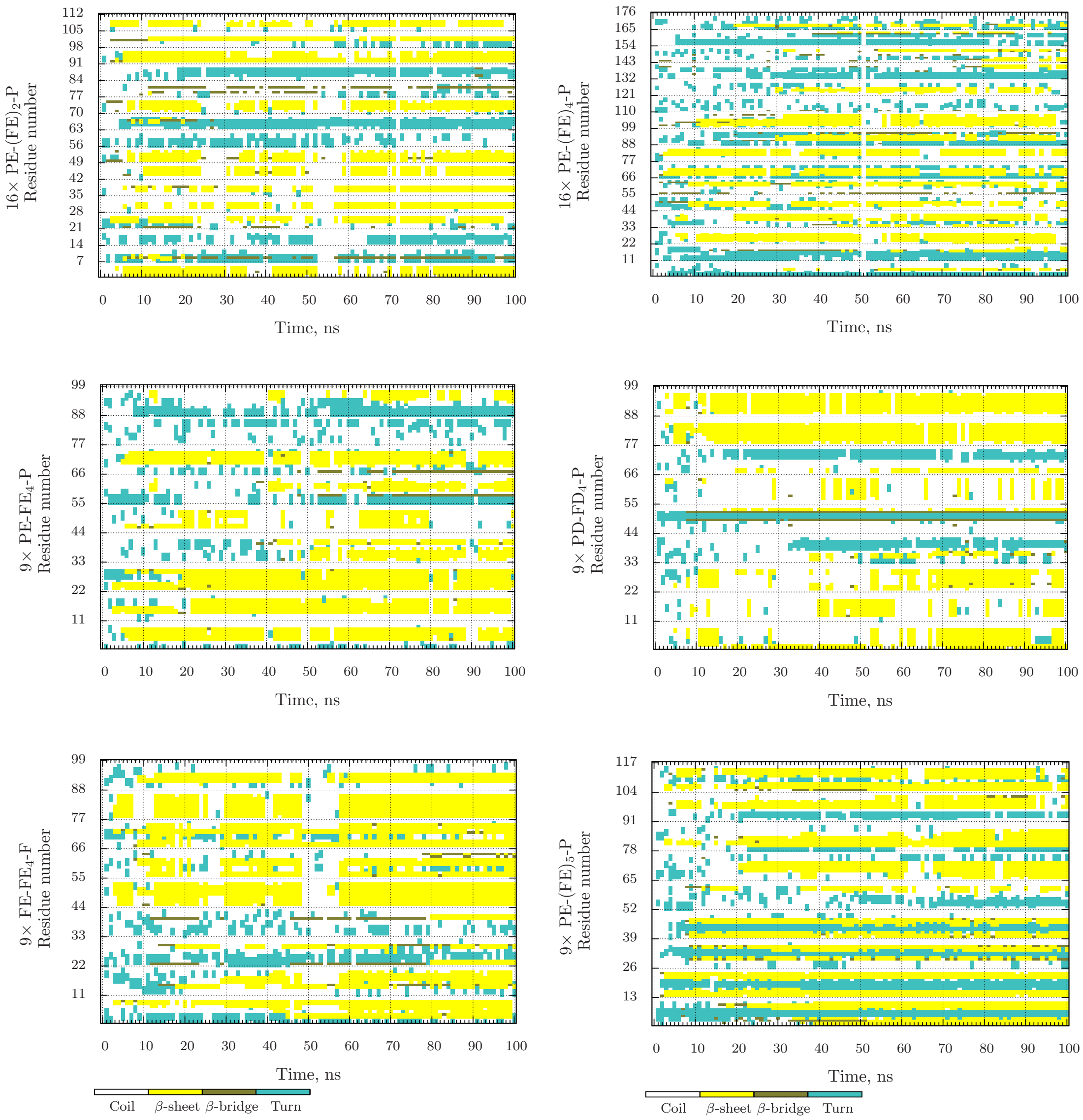


Figure A.3: Secondary structures analysis of systems with high peptide concentrations at the interface (16) for peptides PGlu-2 and PGlu-4, as well as for low peptide concentrations (9 as ch. 3, fig. 3.5) for longer peptides PGlu-4, PAsp-4, PheGlu-4 and PGlu-5.

A.1.3 LJ12-4 parametrisation results

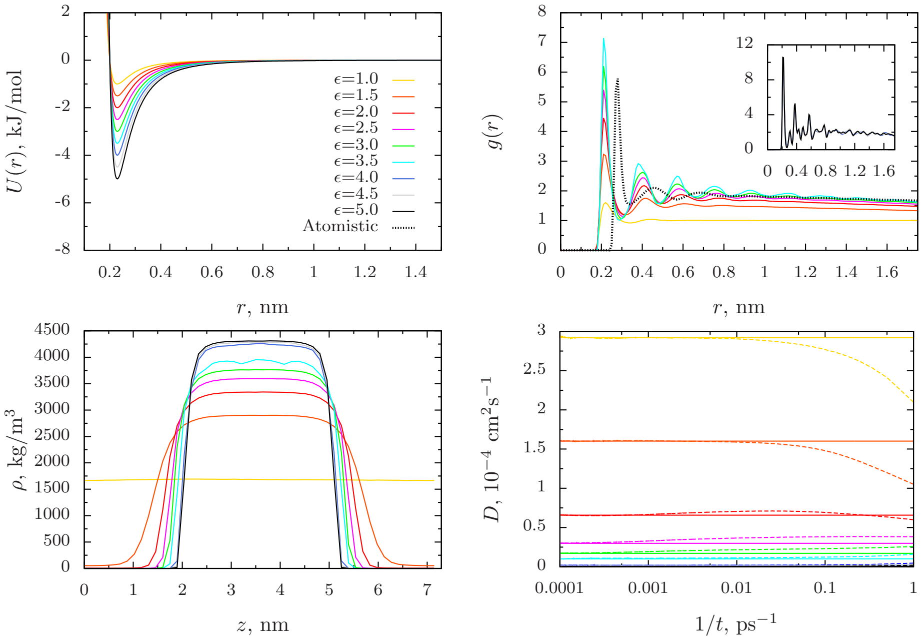


Figure A.4: Results for parameters $\sigma = 0.20$ nm with $\epsilon = 1.00 - 5.00$ kJ/mol.

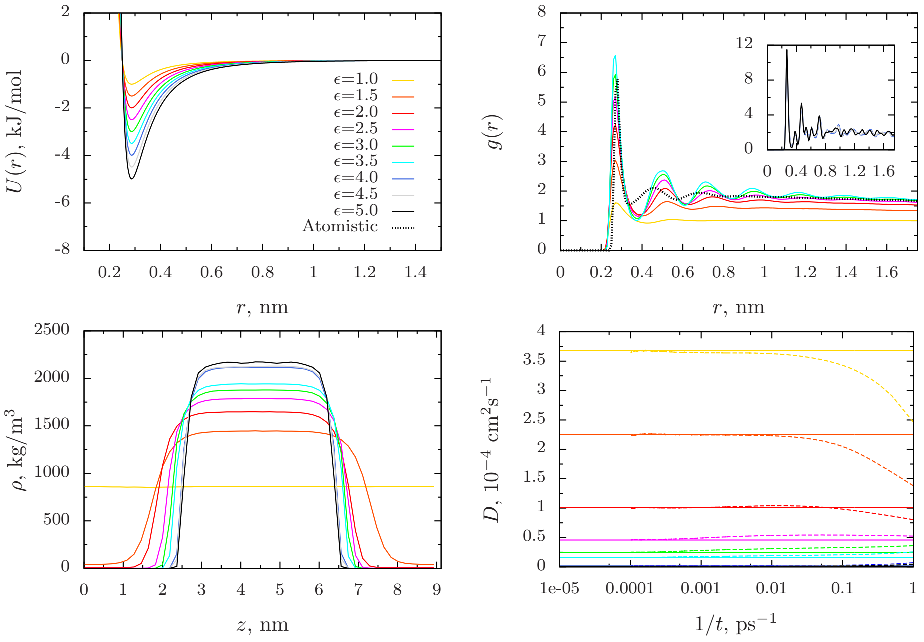


Figure A.5: Results for parameters $\sigma = 0.25$ nm with $\epsilon = 1.00 - 5.00$ kJ/mol.

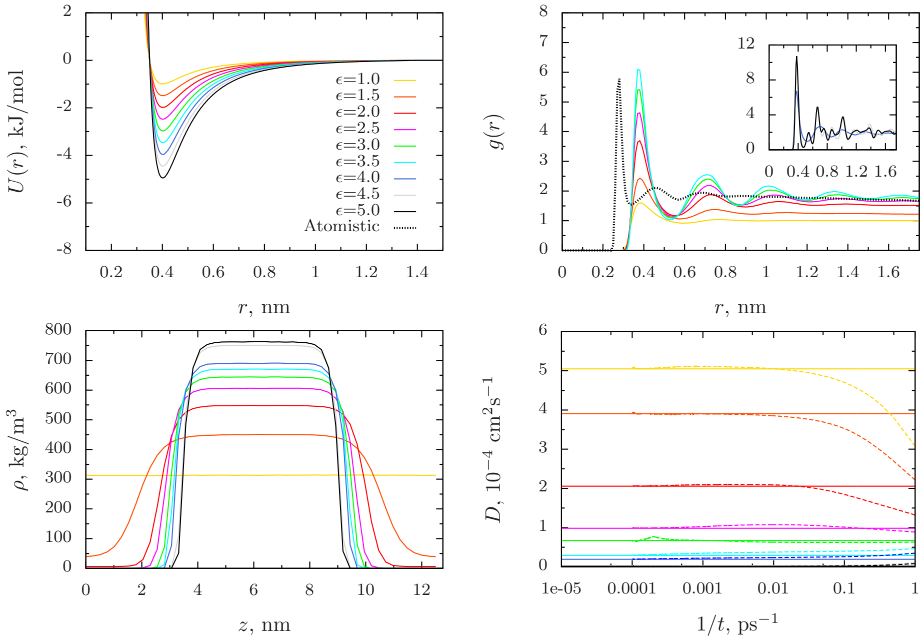


Figure A.6: Results for parameters $\sigma = 0.35$ nm with $\epsilon = 1.00 - 5.00$ kJ/mol.

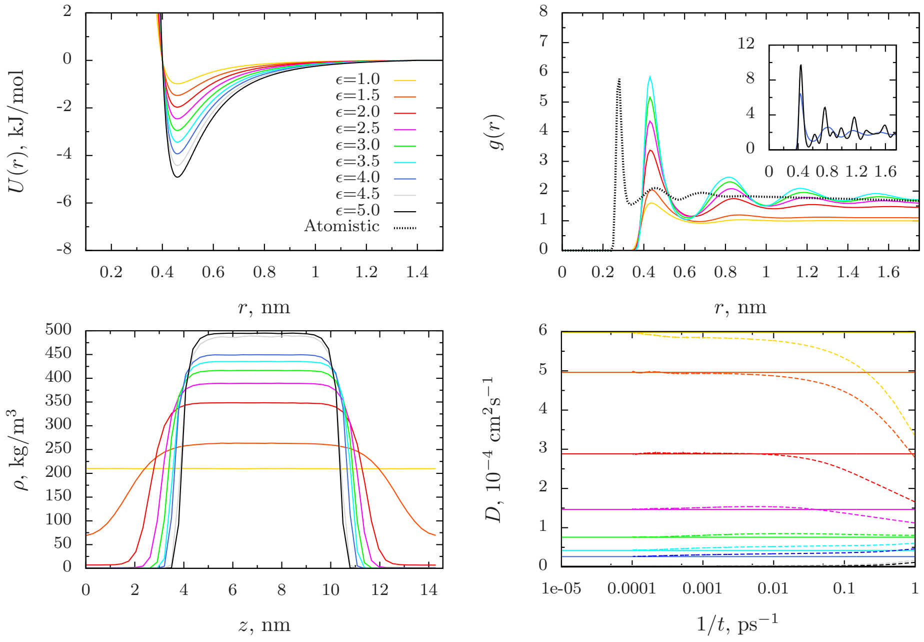


Figure A.7: Results for parameters $\sigma = 0.40$ nm with $\epsilon = 1.00 - 5.00$ kJ/mol.

A.1.4 Implementation of the Downhill Simplex algorithm

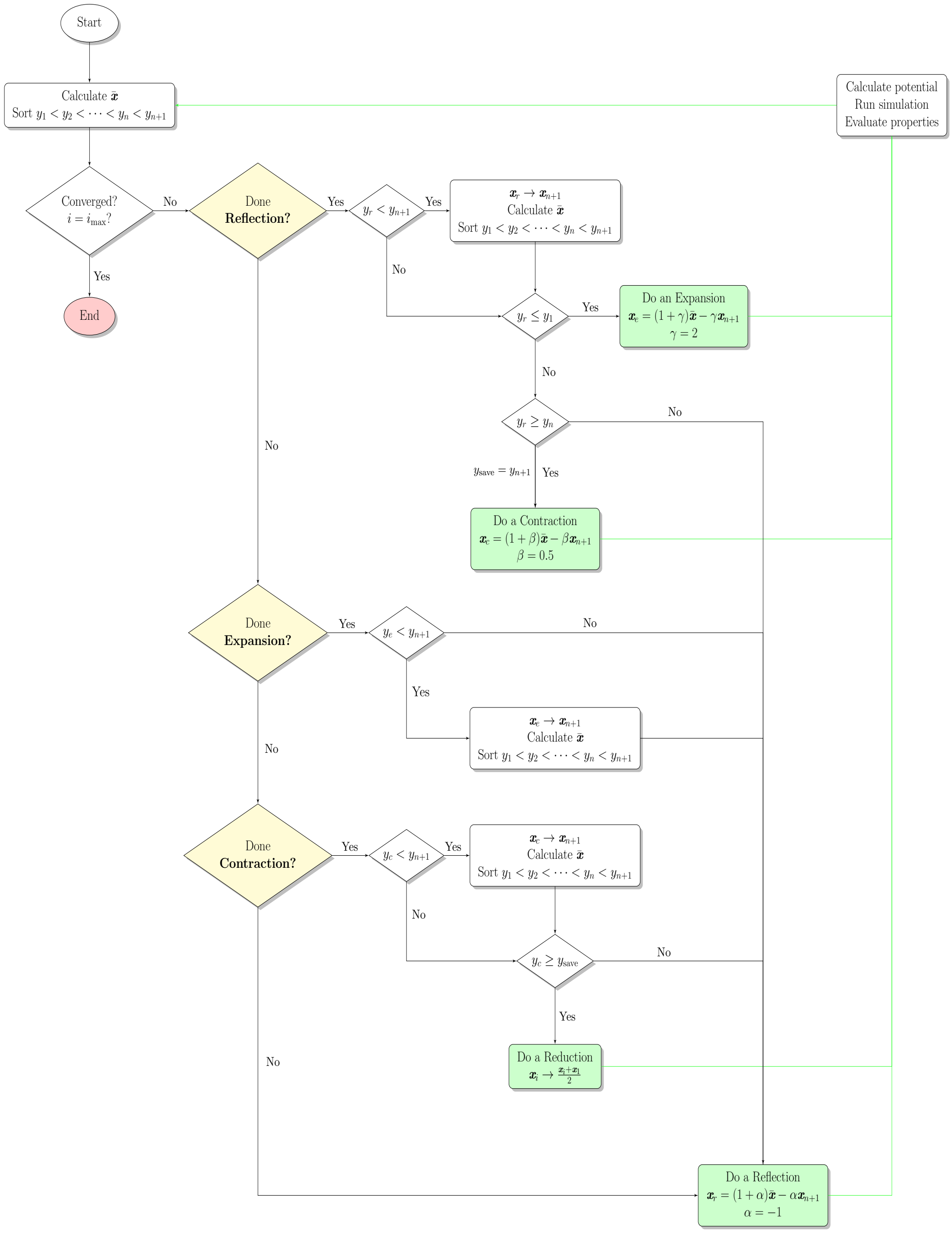
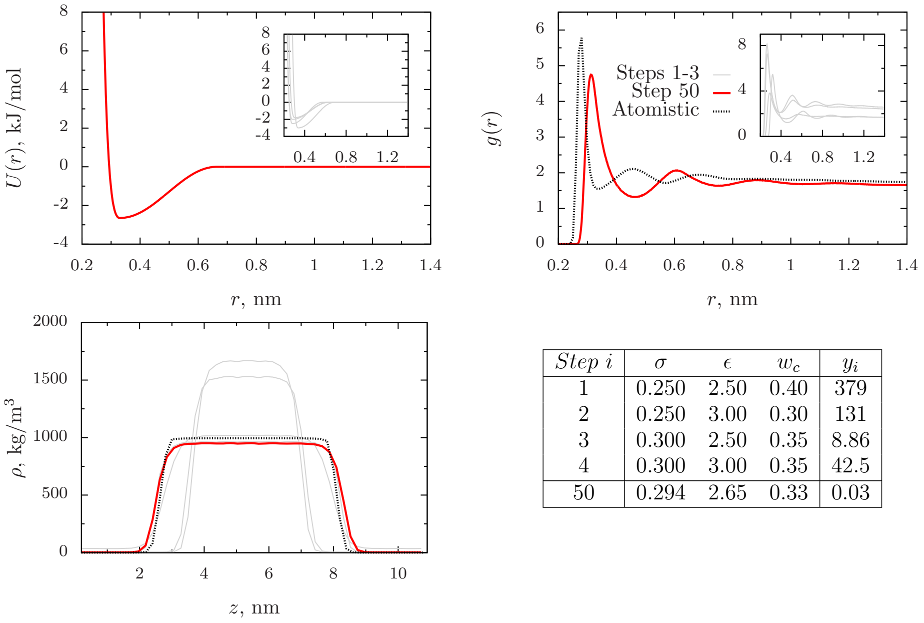

Start
Calculate $\bar{\boldsymbol{x}}$
Sort $y_1 < y_2 < \cdots < y_n < y_{n+1}$
Calculate potential
Run simulation
Evaluate properties
Converged?
$i = i_{\max}$?
No
Done
Reflection?
Yes
$y_r < y_{n+1}$
Yes
$\boldsymbol{x}_r \to \boldsymbol{x}_{n+1}$
Calculate $\bar{\boldsymbol{x}}$
Sort $y_1 < y_2 < \cdots < y_n < y_{n+1}$
Yes
End
No
$y_r \leq y_1$
Yes
Do an Expansion
$\boldsymbol{x}_e = (1+\gamma)\bar{\boldsymbol{x}} - \gamma\boldsymbol{x}_{n+1}$
$\gamma = 2$
No
$y_r \geq y_n$
No
No
$y_{\text{save}} = y_{n+1}$
Yes
Do a Contraction
$\boldsymbol{x}_c = (1+\beta)\bar{\boldsymbol{x}} - \beta\boldsymbol{x}_{n+1}$
$\beta = 0.5$
Done
Expansion?
Yes
$y_e < y_{n+1}$
No
Yes
$\boldsymbol{x}_e \to \boldsymbol{x}_{n+1}$
Calculate $\bar{\boldsymbol{x}}$
Sort $y_1 < y_2 < \cdots < y_n < y_{n+1}$
No
Done
Contraction?
Yes
$y_c < y_{n+1}$
Yes
$\boldsymbol{x}_c \to \boldsymbol{x}_{n+1}$
Calculate $\bar{\boldsymbol{x}}$
Sort $y_1 < y_2 < \cdots < y_n < y_{n+1}$
No
$y_c \geq y_{\text{save}}$
No
Yes
No
Do a Reduction
$\boldsymbol{x}_i \to \frac{\boldsymbol{x}_i + \boldsymbol{x}_1}{2}$
Do a Reflection
$\boldsymbol{x}_r = (1+\alpha)\bar{\boldsymbol{x}} - \alpha\boldsymbol{x}_{n+1}$
$\alpha = -1$

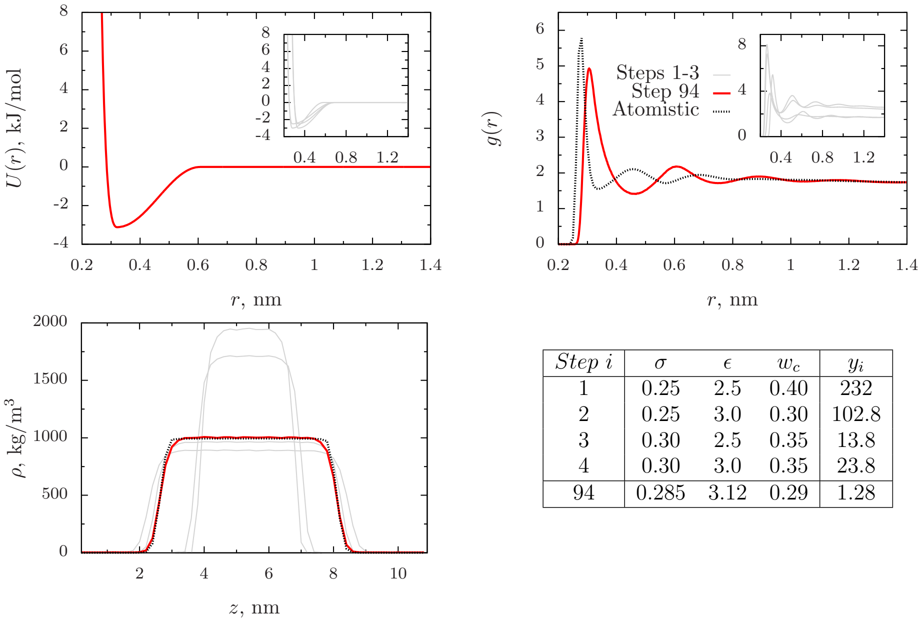
Figure A.8: Work-flow of the Downhill Simplex algorithm as implemented, which operates via a state machine. At every step, it uses the knowledge of the last transformation (yellow) to make a decision on which transformation to perform next (green). After the new transformation has bee performed, the convergence criterion is checked (in this case, if the maximum number of steps, $i_{\max}$, has been performed) and when met, the algorithm is terminated (red).

A.1.5 CKD Simplex results



Step i	σ	ϵ	w_c	y_i
1	0.250	2.50	0.40	379
2	0.250	3.00	0.30	131
3	0.300	2.50	0.35	8.86
4	0.300	3.00	0.35	42.5
50	0.294	2.65	0.33	0.03

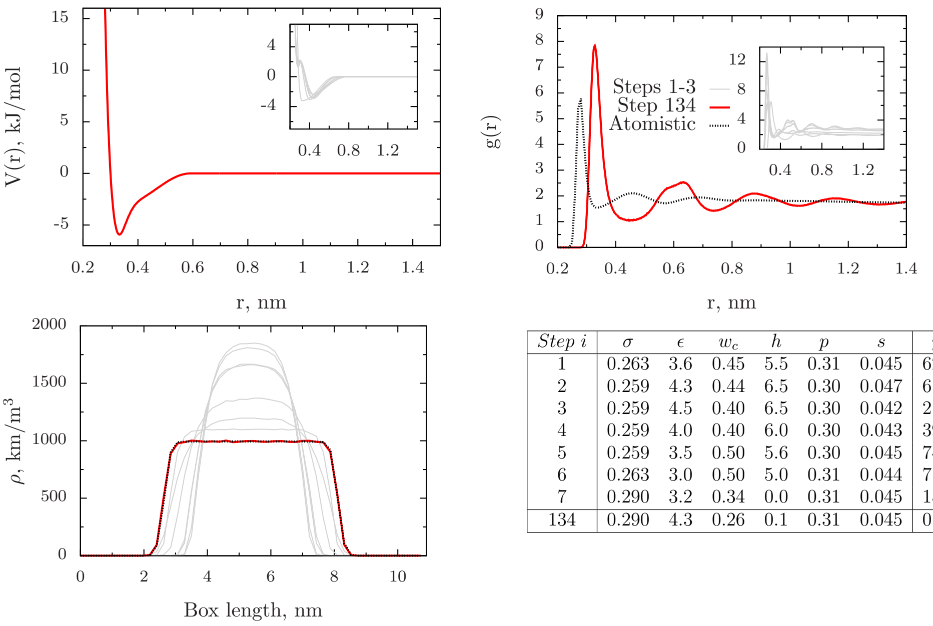
Figure A.9: Simplex optimisation of the slab system when fitting the surface tension with the CKD potential, showing the initial guesses (grey), the converged result (red) and the atomistic reference (black dotted line).



Step i	σ	ϵ	w_c	y_i
1	0.25	2.5	0.40	232
2	0.25	3.0	0.30	102.8
3	0.30	2.5	0.35	13.8
4	0.30	3.0	0.35	23.8
94	0.285	3.12	0.29	1.28

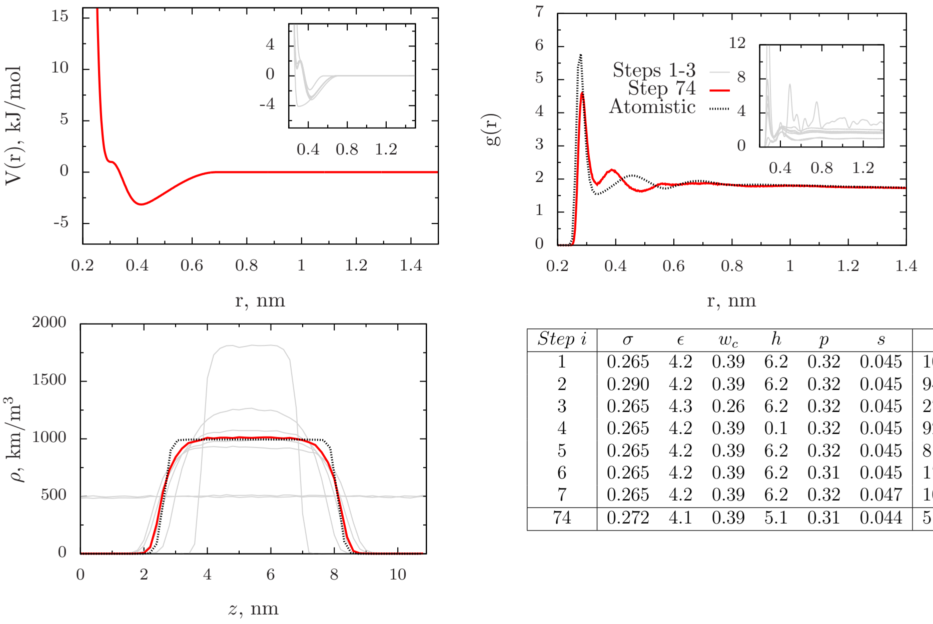
Figure A.10: Simplex optimisation of the slab system when fitting the density profile and the surface tension simultaneously with the CKD potential, showing the initial guesses (grey), the converged result (red) and the atomistic reference (black dotted line).

A.1.6 CKDg Simplex results



Step i	σ	ϵ	w_c	h	p	s	y_i
1	0.263	3.6	0.45	5.5	0.31	0.045	62.8
2	0.259	4.3	0.44	6.5	0.30	0.047	61.0
3	0.259	4.5	0.40	6.5	0.30	0.042	21.7
4	0.259	4.0	0.40	6.0	0.30	0.043	39.2
5	0.259	3.5	0.50	5.6	0.30	0.045	74.0
6	0.263	3.0	0.50	5.0	0.31	0.044	71.6
7	0.290	3.2	0.34	0.0	0.31	0.045	15.3
134	0.290	4.3	0.26	0.1	0.31	0.045	0.72

Figure A.11: Simplex optimisation of the slab system when fitting the density profile with the CKDg potential, showing the initial guesses (grey), the converged result (red) and the atomistic reference (black dotted line).



Step i	σ	ϵ	w_c	h	p	s	y_i
1	0.265	4.2	0.39	6.2	0.32	0.045	10.7%
2	0.290	4.2	0.39	6.2	0.32	0.045	94.3%
3	0.265	4.3	0.26	6.2	0.32	0.045	27.6%
4	0.265	4.2	0.39	0.1	0.32	0.045	92.5%
5	0.265	4.2	0.39	6.2	0.32	0.045	81.9%
6	0.265	4.2	0.39	6.2	0.31	0.045	17.1%
7	0.265	4.2	0.39	6.2	0.32	0.047	10.8%
74	0.272	4.1	0.39	5.1	0.31	0.044	5.98%

Figure A.12: Simplex optimisation of the slab system when fitting the structure (RDF) and the density profile with the CKDg potential simultaneously, showing the initial guesses (grey), the converged result (red) and the atomistic reference (black dotted line).

Bibliography

[1] R. P. Feynman. *The Feynman Lectures on Physics*. Addison-Wesley Longman, 1970.

[2] A. Rinaldi. Naturally better. Science and technology are looking to nature's successful designs for inspiration. *EMBO Rep.*, 8(11):995–999, 2007.

[3] A. J. Mulholland. Introduction. Biomolecular simulation. *J. R. Soc. Int.*, 5(3):169–172, 2008.

[4] M. Karplus, Y. Q. Gao, J. Ma, A. van der Vaart, and W. Yang. Protein structural transitions and their functional role. *Philos. R. R. Soc. A*, 363 (1827):331–356, 2005.

[5] 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. Edit.*, 45(25):4064–4092, 2006.

[6] M. W. van der Kamp, K. E. Shaw, C. J. Woods, and A. J. Mulholland. Biomolecular simulation and modelling: Status, progress and prospects. *J. R. Soc. Int.*, 5(Suppl 3):173–190, 2008.

[7] W. F. van Gunsteren, J. Dolenc, and A. E. Mark. Molecular simulation as an aid to experimentalists. *Curr. Opin. Struc. Biol.*, 18(2):149–153, 2008.

[8] K. Kremer, C. Peter. Multiskalensimulationen in der Materialwissenschaft. *Natur und Geist*, 25(1):14, 2009.

[9] G. S Ayton, W. G. Noid, and G. A. Voth. Multiscale modeling of biomolecular systems: In serial and in parallel. *Curr. Opin. Struc. Biol.*, 17(2):192–198, 2007.

[10] X. Xia, Z. Qian, C. S. Ki, Y. H. Park, D. L. Kaplan, and S. Y. Lee. Native-sized recombinant spider silk protein produced in metabolically engineered Escherichia coli results in a strong fiber. *P. Natl. Acad. Sci.*, 107(32):14059 –14063, 2010.

[11] W. A. Linke. Biomaterials: Spider strength and stretchability. *Nat. Chem. Biol.*, 6(10):702–703, 2010.

[12] L. Ge, S. Sethi, L. Ci, P. M. Ajayan, and A. Dhinojwala. Carbon nanotube-based synthetic gecko tapes. *P. Natl. Acad. Sci.*, 104(26):10792 –10795, 2007.

[13] D. G. Nocera and A. F. Heyduk. *Process for photocatalysis and two-electron mixed-valence complexes (Patent)*, 2005.

[14] P. Fratzl and R. Weinkamer. Nature's hierarchical materials. *Prog. Mater. Sci.*, 52(8):1263–1334, 2007.

[15] S. S. Santoso, S. Vauthey, and S. Zhang. Structures, function and applications of amphiphilic peptides. *Curr. Opin. Colloid In.*, 7(5–6):262–266, 2002.

[16] S. Cavalli, F. Albericio, and A. Kros. Amphiphilic peptides and their cross-disciplinary role as building blocks for nanoscience. *Chem. Soc. Rev.*, 39 (1):241, 2010.

[17] D. N. Woolfson and M. G. Ryadnov. Peptide-based fibrous biomaterials: Some things old, new and borrowed. *Curr. Opin. Chem. Biol.*, 10(6):559–567, 2006.

[18] S. Zhang. Fabrication of novel biomaterials through molecular self-assembly. *Nat. Biotech.*, 21(10):1171–1178, 2003.

[19] S. Segman-Magidovich and H. Rapaport. The effects of template rigidity and amino acid type on heterogeneous calcium-phosphate mineralization by langmuir films of amphiphilic and acidic beta-sheet peptides. *J. Phys. Chem. B*, 116(36):11197–11204, 2012.

[20] N. Amosi, S. Zarzhitsky, E. Gilsohn, O. Salnikov, E. Monsonego-Ornan, R. Shahar, and H. Rapaport. Acidic peptide hydrogel scaffolds enhance calcium phosphate mineral turnover into bone tissue. *Act. Biomat.*, 8(7): 2466–2475, 2012.

[21] H. Rapaport, K. Kjaer, T. R. Jensen, L. Leiserowitz, and D. A. Tirrell. Two-dimensional order in beta-sheet peptide monolayers. *J. Am. Chem. Soc.*, 122(50):12523–12529, 2000.

[22] H. Isenberg, K. Kjaer, and H. Rapaport. Elasticity of crystalline beta-sheet monolayers. *J. Am. Chem. Soc.*, 128(38):12468–12472, 2006.

[23] H. Rapaport, H. Grisaru, and T. Silberstein. Hydrogel scaffolds of amphiphilic and acidic beta-sheet peptides. *Adv. Funct. Mater.*, 18(19):2889–2896, 2008.

[24] H. Rapaport. *Amphiphilic peptide matrices for treatment of osteoporosis (Patent)*, 2010.

[25] H. Rapaport. *Amphiphilic peptides and hydrogel matrices thereof for bone repair (Patent)*, 2010.

[26] P. U. P. A. Gilbert, M. Abrecht, and B. H. Frazer. The organic-mineral interface in biominerals. *Rev. Mineral Geochem.*, 59(1):157–185, 2005.

[27] S. Segman-Magidovich, H. Grisaru, T. Gitli, Y. Levi-Kalisman, and H. Rapaport. Matrices of acidic beta-sheet peptides as templates for calcium phosphate mineralization. *Adv. Mater.*, 20(11):2156–2161, 2008.

[28] V. Vaiser and H. Rapaport. Compressibility and elasticity of amphiphilic and acidic beta-sheet peptides at the air-water interface. *J. Phys. Chem. B*, 115(1):50–56, 2010.

[29] G. Colombo, P. Soto, and E. Gazit. Peptide self-assembly at the nanoscale: A challenging target for computational and experimental biotechnology. *Trends Biotechnol.*, 25(5):211–218, 2007.

[30] D. Fritz, K. Koschke, V. A. Harmandaris, N. F. A. van der Vegt, and K. Kremer. Multiscale modeling of soft matter: Scaling of dynamics. *Phys. Chem. Chem. Phys.*, 13(22):10412–10420, 2011.

[31] P. W. Atkins and R. S. Friedman. *Molecular Quantum Mechanics.* Oxford University Press, USA, 5th edition, 2010.

[32] D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications.* Academic Press, 2nd edition, 2001.

[33] 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(19):5179–5197, 1995.

[34] A. D. MacKerell, N. Banavali, and N. Foloppe. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers*, 56(4): 257–265, 2001.

[35] 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(13):1656–1676, 2004.

[36] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, 118(45):11225–11236, 1996.

[37] M. Tuckerman. *Statistical Mechanics: Theory and Molecular Simulation.* Oxford University Press, 2010.

[38] J. D. Durrant and J. A. McCammon. Molecular dynamics simulations and drug discovery. *BMC Biology*, 9(1):71, 2011.

[39] J. 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(3):327–341, 1977.

[40] H. C. Andersen. Rattle: A “velocity” version of the shake algorithm for molecular dynamics calculations. *J. Comput. Phys.*, 52(1):24–34, 1983.

[41] L. Verlet. Computer ”experiments” on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.*, 159(1):98–103, 1967.

[42] W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.*, 76(1):637–649, 1982.

[43] R. W. Hockney, S. P. Goel, and J. W. Eastwood. Quiet high-resolution computer models of a plasma. *J. Comput. Phys.*, 14(2):148–158, 1974.

[44] P. Hünenberger. Thermostat algorithms for molecular dynamics simulations advanced computer simulation. In *Advanced Computer Simulation*, volume 173, pages 130–130. Springer Berlin/Heidelberg, 2005.

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

[46] G. Bussi, D. Donadio, and M. Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126(1):014101, 2007.

[47] S. Nosé. A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.*, 100(1):191–198, 2002.

[48] W. G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A*, 31(3):1695–1697, 1985.

[49] G. S. Grest and K. Kremer. Molecular dynamics simulation for polymers in the presence of a heat bath. *Phys. Rev. A*, 33(5):3628–3631, 1986.

[50] H. C. Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.*, 72(4):2384–2393, 1980.

[51] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.*, 52(12):7182–7190, 1981.

[52] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald: An NlogN method for ewald sums in large systems. *J. Chem. Phys.*, 98(12):10089–10092, 1993.

[53] M. Deserno and C. Holm. How to mesh up Ewald sums. I. A theoretical and numerical comparison of various particle mesh routines. *J. Chem. Phys.*, 109(18):7678–7693, 1998.

[54] 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(16): 1701–1718, 2005.

[55] A. Y. Toukmaji and J. A. Board Jr. Ewald summation techniques in perspective: A survey. *Comput. Phys. Commun.*, 95(2–3):73–92, 1996.

[56] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314(1-2):141–151, 1999.

[57] C. Tavan. Replica Exchange Molecular Dynamics Method (REMD) lehrseminar: Fortgeschrittene methoden in der simulation von biomolekülen, 2010.

[58] D. Reith, M. Pütz, and F. Müller-Plathe. Deriving effective mesoscale potentials from atomistic simulations. *J. Comput. Chem.*, 24(13):1624–1636, 2003.

[59] J. A. Nelder and R. Mead. A Simplex method for function minimization. *The Computer Journal*, 7(4):308 –313, 1965.

[60] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press, 3rd edition, 2007.

[61] W. Tschöp, K. Kremer, J. Batoulis, T. Bürger, and O. Hahn. Simulation of polymer melts. I. Coarse-graining procedure for polycarbonates. *Act. Polym.*, 49(2-3):61–74, 1998.

[62] J. G. Kirkwood. Statistical mechanics of fluid mixtures. *J. Chem. Phys.*, 3 (5):300, 1935.

[63] H. Wang, C. Junghans, and K. Kremer. Comparative atomistic and coarse-grained study of water: What do we lose by coarse-graining? *Eur. Phys. J. E*, 28(2):221–229, 2009.

[64] R. L. Henderson. A uniqueness theorem for fluid pair correlation functions. *Physics Letters A*, 49(3):197–198, 1974.

[65] J. Hansen and I. R. McDonald. *Theory of Simple Liquids.* Academic Press, 2006.

[66] W. E, W. Ren, and E. Vanden-Eijnden. Energy landscapes and rare events. *ICM*, pages 621–630, 2002.

[67] C. Oostenbrink, T. A. Soares, N. F. A. van der Vegt, and W. F. van Gunsteren. Validation of the 53A6 GROMOS force field. *Eur. Biophys. J.*, 34 (4):273–284, 2005.

[68] W. F. van Gunsteren and H. J. C. Berendsen. Groningen Molecular Simulation (GROMOS) library manual, 1987.

[69] L. Schuler, X. Daura, and W. F. van Gunsteren. An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *J. Comput. Chem.*, 22(11):1205–1218, 2001.

[70] 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(7): 725–737, 2005.

[71] N. Schmid, A. Eichenberger, A. Choutko, S. Riniker, M. Winger, A. Mark, and W. F. van Gunsteren. Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *Eur. Biophys. J.*, 40(7):843–856, 2011.

[72] L. D. Schuler, P. Walde, P. L. Luisi, and W. F. van Gunsteren. Molecular dynamics simulation of n-dodecyl phosphate aggregate structures. *Eur. Biophys. J.*, 30(5):330–343, 2001.

[73] 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, 1995.

[74] X. Daura, K. Gademann, B. Jaun, D. Seebach, W. F. van Gunsteren, and A. E. Mark. Peptide folding: When simulation meets experiment. *Angew. Chem. Int. Edit.*, 38(1-2):236–240, 1999.

[75] W. Kabsch and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[76] D. Frishman and P. Argos. Knowledge-based protein secondary structure assignment. *Proteins*, 23(4):566–579, 1995.

[77] Cameron F. Abrams and Kurt Kremer. Combined Coarse-Grained and Atomistic Simulation of Liquid Bisphenol A-Polycarbonate: liquid packing and intramolecular structure. *Macromolecules*, 36(1):260–267, 2003.

[78] C. F. Abrams, L. Delle Site, and K. Kremer. Dual-resolution coarse-grained simulation of the bisphenol-A-polycarbonate/nickel interface. *Phys. Rev. E*, 67(2):021807, 2003.

[79] C. Abrams, L. Delle Site, and K. Kremer. Multiscale Computer Simulations for Polymeric Materials in Bulk and Near Surfaces. In P. Nielaba, M. Mareschal, and G. Ciccotti, editors, *Bridging Time Scales: Molecular Simulations for the Next Decade*, volume 605, pages 143–164. Springer Berlin/Heidelberg.

[80] N. Zacharopoulos, N. Vergadou, and D. N. Theodorou. Coarse-graining using pretabulated potentials: Liquid benzene. *J. Chem. Phys.*, 122:244111, 2005.

[81] J. W. Mullinax and W. G. Noid. A generalized Yvon-Born-Green theory for determining coarse-grained interaction potentials. *J. Phys. Chem. C*, 114(12):5661–5674, 2009.

[82] J. W. Mullinax and W. G. Noid. Reference state for the generalized Yvon–Born–Green theory: Application for coarse-grained model of hydrophobic hydration. *J. Chem. Phys.*, 133:124107, 2010.

[83] T. Murtola, A. Bunker, I. Vattulainen, M. Deserno, and M. Karttunen. Multiscale modeling of emergent materials: Biological and soft matter. *Phys. Chem. Chem. Phys.*, 11(12):1869–1892, 2009.

[84] S. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries. The MARTINI force field: Coarse-grained model for biomolecular simulations. *J. Phys. Chem. B*, 111(27):7812–7824, 2007.

[85] Z. Wang and M. Deserno. Systematic implicit solvent coarse-graining of bilayer membranes: Lipid and phase transferability of the force field. *New J. Phys.*, 12(9):095004, 2010.

[86] D. M. Huang, R. Faller, K. Do, and A. J. Moule. Coarse-grained computer simulations of polymer/fullerene bulk heterojunctions for organic photovoltaic applications. *J. Chem. Theory Comput.*, 6(2):526–537, 2010.

[87] A. Lukyanov, A. Malafeev, V. Ivanov, H. Chen, K. Kremer, and D. Andrienko. Solvated poly-(phenylene vinylene) derivatives: Conformational structure and aggregation behavior. *J. Mater. Chem.*, 20(46):10475–10485, 2010.

[88] V. Rühle, J. Kirkpatrick, K. Kremer, and D. Andrienko. Coarse-grained modelling of polypyrrole morphologies. *Phys. Stat. Solidi B*, 245:844, 2008.

[89] A. Villa, N. F. A. van der Vegt, and C. Peter. Self-assembling dipeptides: Including solvent degrees of freedom in a coarse-grained model. *Phys. Chem. Chem. Phys.*, 11(12):2068–2076, 2009.

[90] O. Bezkorovaynaya, A. Lukyanov, K. Kremer, and C. Peter. Multiscale simulation of small peptides: Consistent conformational sampling in atomistic and coarse-grained models. *J. Comput. Chem.*, 33(9):937–949, 2012.

[91] W. Shinoda, R. DeVane, and M. L. Klein. Coarse-grained molecular modeling of non-ionic surfactant self-assembly. *Soft Matter*, 4(12):2454–2462, 2008.

[92] L. Monticelli, S. Kandasamy, X. Periole, R. Larson, P. Tieleman, and S Marrink. The MARTINI coarse-grained force field: Extension to proteins. *J. Chem. Theory Comput.*, 4(5):819–834, 2008.

[93] M. Praprotnik, L. Delle Site, and K. Kremer. Adaptive resolution molecular-dynamics simulation: Changing the degrees of freedom on the fly. *J. Chem. Phys.*, 123(22):224106–224106–14, 2005.

[94] M. Praprotnik, K. Kremer, and L. Delle Site. Adaptive molecular resolution via a continuous change of the phase space dimensionality. *Phys. Rev. E*, 75(1):017701, 2007.

[95] S. Fritsch, S. Poblete, C. Junghans, G. Ciccotti, L. Delle Site, and K. Kremer. Adaptive resolution molecular dynamics simulation through coupling to an internal particle reservoir. *Phys. Rev. Lett.*, 108(17):170602, 2012.

[96] B. Hess, S. León, N. van der Vegt, and K. Kremer. Long time atomistic polymer trajectories from coarse grained simulations: Bisphenol-A polycarbonate. *Soft Matter*, 2(5):409, 2006.

[97] V. A. Harmandaris, N. P. Adhikari, N. F. A. van der Vegt, and K. Kremer. Hierarchical modeling of polystyrene: From atomistic to coarse-grained simulations. *Macromolecules*, 39(19):6708–6719, 2006.

[98] W. G. Noid, J. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, and H. C. Andersen. The multiscale coarse graining method. 1. A rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.*, 128:244114, 2008.

[99] F. Ercolessi and J. B. Adams. Interatomic potentials from first-principles calculations: The force-matching method. *Europhys. Lett.*, 26:583–588, 1994.

[100] S. Izvekov and G. A. Voth. A multiscale coarse-graining method for biomolecular systems. *J. Phys. Chem. B*, 109(7):2469–2473, 2005.

[101] Gary S. Ayton, Will G. Noid, and Gregory A. Voth. Systematic coarse-graining of biomolecular and soft-matter systems. *MRS Bulletin*, 32(11): 929–934, 2007.

[102] A. P. Lyubartsev and A. Laaksonen. Calculation of effective interaction potentials from radial distribution functions: A reverse Monte Carlo approach. *Phys. Rev. E*, 52(4):3730, 1995.

[103] P. Gumbsch. Transferability of Interatomic Potentials in Strong Solids - What Properties Can Actually Be Represented, 2012.

[104] H. Berendsen. GROMACS: a message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.*, 91(1-3):43–56, 1995.

[105] V. Rühle, C. Junghans, A. Lukyanov, K. Kremer, and D. Andrienko. Versatile Object-oriented Toolkit for Coarse-graining Applications. *J. Chem. Theory Comput.*, 5(12):3211–3223, 2009.

[106] M. E. Johnson, T. Head-Gordon, and A. A. Louis. Representability problems for coarse-grained water potentials. *J. Chem. Phys.*, 126(14):144509–144509–10, 2007.

[107] S. O. Yesylevskyy, L. V. Schäfer, D. Sengupta, and S. Marrink. Polarizable water model for the coarse-grained MARTINI force field. *PLoS Comput. Biol.*, 6(6):e1000810, 2010.

[108] W. Shinoda, R. DeVane, and M. L. Klein. Multi-property fitting and parameterization of a coarse grained model for aqueous surfactants. *Mol. Simulat.*, 33(1):27, 2007.

[109] X. He, W. Shinoda, T. DeVane, and M. L. Klein. Exploring the utility of coarse-grained water models for computational studies of interfacial systems. *Mol. Phys.*, 108(15):2007, 2010.

[110] R. Faller, H. Schmitz, O. Biermann, and F. Müller-Plathe. Automatic parameterization of force fields for liquids by Simplex optimization. *J. Comput. Chem.*, 20:100–9, 1998.

[111] H. Meyer, O. Biermann, R. Faller, D. Reith, and F. Müller-Plathe. Coarse graining of nonbonded inter-particle potentials using automatic simplex optimization to fit structural properties. *J. Chem. Phys.*, 113(15):6264–6275, 2000.

[112] D. Reith, H. Meyer, and F. Müller-Plathe. Mapping atomistic to coarse-grained polymer models using automatic Simplex optimization to fit structural properties. *Macromolecules*, 34(7):2335–2345, 2001.

[113] I. R. Cooke, K. Kremer, and M. Deserno. Tunable generic model for fluid bilayer membranes. *Phys. Rev. E*, 72(1):011506, 2005.

[114] M. V. Berry, R. F. Durrans, and R. Evans. The calculation of surface tension for simple liquids. *J. Phys. A*, 5:166–170, 1972.

[115] T. Head-Gordon and F. H. Stillinger. An orientational perturbation theory for pure liquid water. *J. Chem. Phys.*, 98(4):3313, 1993.

[116] N. M. Barraz, E. Salcedo, and M. C. Barbosa. Thermodynamic, dynamic, and structural anomalies for shoulderlike potentials. *J. Chem. Phys.*, 131 (9):094504, 2009.

[117] P. Wernet, D. Nordlund, U. Bergmann, M. Cavalleri, M. Odelius, H. Ogasawara, L. Å. Näslund, T. K. Hirsch, L. Ojamäe, P. Glatzel, L. G. M. Pettersson, and A. Nilsson. The structure of the first coordination shell in liquid water. *Science*, 304(5673):995–999, 2004.

[118] C. Huang, K. T. Wikfeldt, T. Tokushima, D. Nordlund, Y. Harada, U. Bergmann, M. Niebuhr, T. M. Weiss, Y. Horikawa, M. Leetmaa, M. P. Ljungberg, O. Takahashi, A. Lenz, L. Ojamäe, A. P. Lyubartsev, S. Shin, L. G. M. Pettersson, and A. Nilsson. The inhomogeneous structure of water at ambient conditions. *P. Natl. Acad. Sci.*, 2009.

[119] B. M. Auer and J. L. Skinner. Water: Hydrogen bonding and vibrational spectroscopy, in the bulk liquid and at the liquid/vapor interface. *Chem. Phys. Lett.*, 470(1–3):13–20, 2009.

[120] J. Weeks and L. Pratt. Introduction to special issue on water and associated liquids. *J. Stat. Phys.*, 145(2):207–208, 2011.

[121] J. Dzubiella. How interface geometry dictates water's thermodynamic signature in hydrophobic association. *J. Stat. Phys.*, 145(2):227–239, 2011.

[122] W. Hujo, B. Shadrack Jabes, V. Rana, C. Chakravarty, and V. Molinero. The rise and fall of anomalies in tetrahedral liquids. *J. Stat. Phys.*, 145(2): 293–312, 2011.

[123] N. Ji and Y. Shen. Sum frequency vibrational spectroscopy of leucine molecules adsorbed at air–water interface. *J. Chem. Phys.*, 120(15):7107–7112, 2004.

[124] Y. Fan, X. Chen, L. Yang, P. S. Cremer, and Y. Q. Gao. On the structure of water at the aqueous/air interface. *The Journal of Physical Chemistry B*, 113(34):11672–11679, 2009.

[125] S. Nihonyanagi, T. Ishiyama, T. Lee, S. Yamaguchi, M. Bonn, A. Morita, and T. Tahara. Unified molecular view of the air/water interface based on experimental and theoretical spectra of an isotopically diluted water surface. *J. Am. Chem. Soc.*, 133(42):16875–16880, 2011.

[126] M. Jochum, D. Andrienko, K. Kremer, and C. Peter. Structure-based coarse-graining in liquid slabs. *J. Chem. Phys.*, 137(6):064102–064102–9, 2012.

[127] L. Larini, L. Lu, and G. A. Voth. The multiscale coarse-graining method. VI. implementation of three-body coarse-grained potentials. *J. Chem. Phys.*, 132:164107, 2010.

[128] H. Abe and N. Gō. Noninteracting local-structure model of folding and unfolding transition in globular proteins. II. application to two-dimensional lattice proteins. *Biopolymers*, 20(5):1013–1031, 1981.

[129] M. J. Sippl. Knowledge-based potentials for proteins. *Curr. Opin. Struc. Biol.*, 5(2):229–235, 1995.

[130] S. Miyazawa and R. L. Jernigan. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules*, 18(3):534–552, 1985.

[131] A. E. van Giessen and J. E. Straub. Monte Carlo simulations of polyalanine using a reduced model and statistics-based interaction potentials. *J. Chem. Phys.*, 122(2):024904, 2005.

[132] A. Liwo, M. Khalili, and H. A. Scheraga. Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *P. Natl. Acad. Sci.*, 102(7):2362–2367, 2005.

Acknowledgements