Estimating Orientational Water Entropy at Protein Interfaces

Dissertation

for the award of the degree

"Doctor rerum naturalium"

Division of Mathematics and Natural Sciences

of the Georg-August-Universität Göttingen

submitted by

Stephanus Michael Fengler

from Hildesheim

Göttingen 2011

Thesis Committee Members:

1. Prof. Dr. Helmut Grubmüller (Reviewer)
Max Planck Institute for Biophysical Chemistry
Department of Theoretical and Computational Biophysics

2. Prof. Dr. Annette Zippelius (Reviewer)
Georg-August-Universität Göttingen
Institute for Theoretical Physics

3. Prof. Dr. Bernd Abel
University Leipzig
Wilhelm-Ostwald-Institute for Physical and Theoretical Chemistry

Date of the oral examination: February 24, 2011

Hiermit bestätige ich, dass ich meine Dissertation selbständig und ohne unerlaubte Hilfe angefertigt habe.

Contents

List of Figures

1

Introduction

> "Any method involving the notion of entropy, the very existence of which depends on the second law of thermodynamics, will doubtless seem to many far-fetched, and may repel beginners as obscure and difficult of comprehension."
>
> Willard Gibbs, *Graphical Methods in the Thermodynamics of Fluids* (1873)

Water and sunlight are two crucial key components of the development of life on earth. In the course of evolution, in and around water species have occupied all niches in nature because they cannot live without this special molecule. It plays a central role because of its hydrogen bonding properties and polarity. This makes it a very good solvent for all types of salts and further, all major components of cells like proteins, deoxyribonucleic acid (DNA) or polysaccharides.

1.1 Water and Life

Water is vital both as a solvent in which many of the body's solutes dissolve and as an essential part of many metabolic processes. *Id est*, either water is removed from molecules by enzymatic chemical reactions to grow larger molecules (e.g. proteins) or it is used to break bonds to generate smaller molecules (e.g. amino acids). Water is thus essential and central to these processes. Without water, life as we know it would not exist. It also plays a central role in photosynthesis and respiration. Photosynthetic cells split water with the sun's energy into its basic elements hydrogen and oxygen. While oxygen is released, the hydrogen is combined with carbon dioxide (CO_2) to form glucose in the Calvin cycle [15]. All living cells use such processes to capture and store the sun's energy for further cellular respiration, which generates CO_2 and H_2O as waste products. This is a closed biological cycle.

Water is also central to acid-base neutrality and enzyme function. Its interactions play a mayor role in solvation free energies, protein folding and enzymatic reactions of proteins. Furthermore, most proteins function only in their native environment, folded in a free energy minimum, while they do not function in an unfolded state, which is induced in vitro in different ways,

e.g. by temperature increase, change in pH, adding certain denaturants like urea or *temperature decrease*.

Of course, physical properties of water are also very well studied and have been described in literature extensively. For example, the most prominent property is that the volume as a function of temperature behaves abnormally at normal pressure. When water is cooled from room temperature, it becomes increasingly dense as most other substances. At 4 degrees Celsius however, it reaches its density maximum and upon further cooling it expands again until it freezes. This negative expansion coefficient is due to strong orientation-dependent intermolecular dipole-dipole interactions.

1.2 Hydrophobic Effect

In biological systems the solubility of water is therefore important, because water is influenced by the chemical polarity of the dissolved molecule as is the molecule influenced by water, for example it effects protein stabilization and folding. This is known as the hydrophobic effect.

The hydrophobic effect occurs when a molecule is solvated and is the result of the free energy change upon solvation. In close vicinity of the molecule, the water entropy and enthalpy is changed depending on the solvated molecule.

If it is polar at the surface, water entropy is increased because the surrounding water molecules occupy additional orientational and translational states. Water enthalpy is slightly reduced because the original hydrogen bonding network is disturbed. But overall the free energy for solvation of a polar molecule is negative. Therefore, polar molecules like ethanol are easily dissolved in water.

The opposite effect occurs upon solvation of a non-polar molecule. Here, the water entropy is reduced [23] because the surrounding water molecules are more ordered around the non-polar molecule as depicted in figure 1.1. Although the enthalpy is increased because the molecules form a less disturbed hydrogen network, the loss of mobility outweighs and the solvation free energy is positive. Hence, the effect is named entropic repulsion.

At a higher temperature, the enthalpy, in particular the free energy contribution from hydrogen bonds is decreased, because water molecules are more mobile. The increase of entropy however, compensates this loss in free energy. Thus, the hydrophobic effect is only weakly temperature-dependent, but becomes smaller at a lower temperature, which effects the fold of biomolecules and is known as *cold denaturation*.

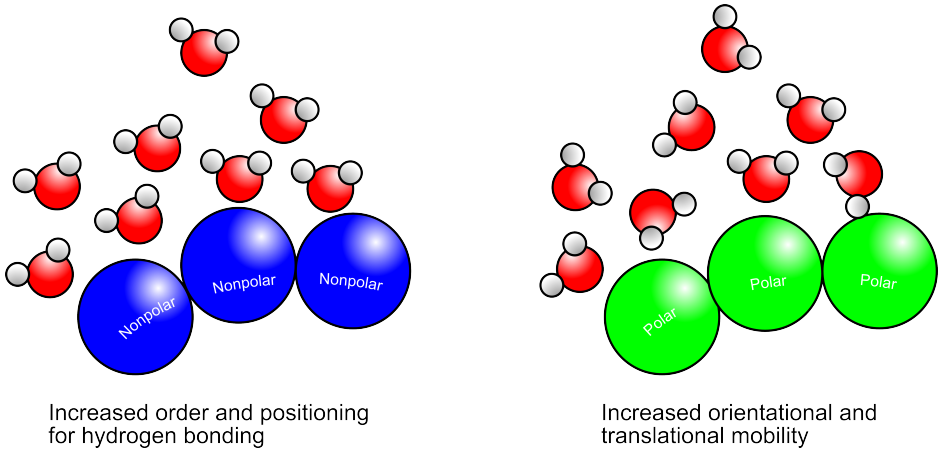


Figure 1.1 Iceberg model [23] according to Evans for non-polar molecules. On the left, the increased ordering of water molecules surrounding a nonpolar molecule is shown. This effect lowers the entropy of the system and therefore, the free energy of the system is increased. For comparison, on the right the entropy is increased, because the surrounding molecules occupy additional inaccessible states.

1.3 Cold Denaturation and Stabilization of the Native Protein Fold

Denaturation of proteins upon heat seems to be a natural and not surprising phenomenon, because any process, which is induced by increasing temperature, should proceed with heat absorption. That is an increase in enthalpy and entropy of the system. Therefore, an increase in the disorder of the system and finally unfolding is expected. However, by the same reasoning denaturation by temperature decrease seems to be counterintuitive because temperature decrease should lead to an increase of order, but the protein is already in its favorable most ordered state, its native fold. No significant change in the structure should occur. This point of view accounts for the free energy of the water and the protein but neglects interactions between them.

However, unfolding is observed for most globular proteins. As shown in figure (1.2), the free energy is composed of the entropy and enthalpy of the protein as well as the free energy of the solvent, given as the hydrophobic effect F_{HE}. Both the entropy and enthalpy difference contribute largely to the free energy difference between the denatured and native state, but they essentially compensate at native conditions. The determining factor for sta-

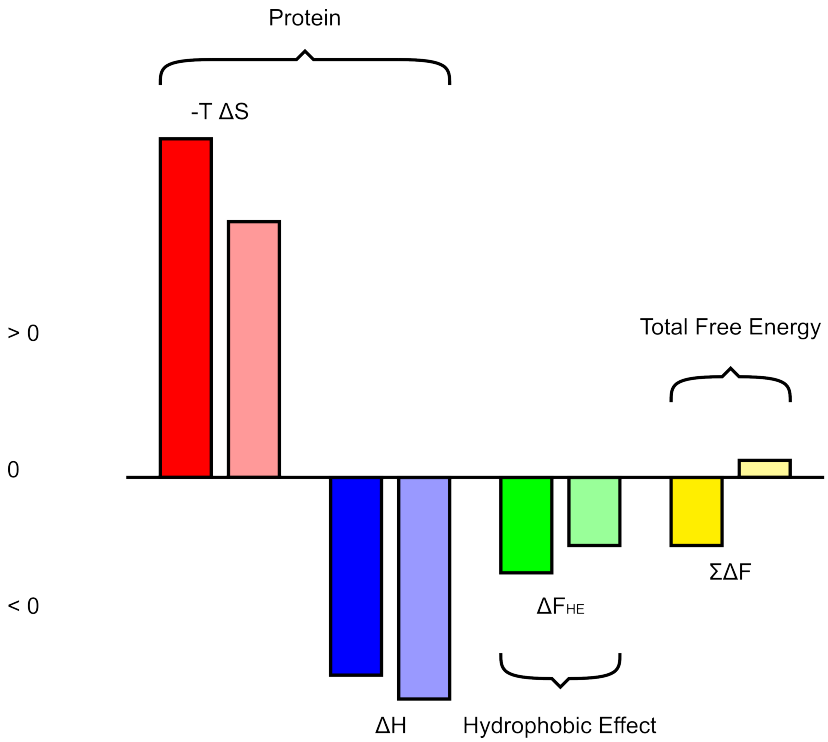


Figure 1.2 Free energy diagram of cold denaturation. The free energy difference between the denatured and native state of a protein before and after temperature decrease is shown. The solid bars in strong colors represent the different free energy contributions at normal conditions. The bars in light colors depict the situation upon temperature decrease. The hydrophobic effect decreases faster at lower temperatures than the net gain in the other contributions, hence the native state is destabilized. Therefore, the different temperature dependencies of all contributions lead to the unexpected effect that the denatured state is stabilized upon temperature decrease.

bilization of the protein comes from the hydrophobic effect, which lowers the free energy of the system just enough to favor the native state.

This delicate balance is shifted upon temperature decrease because entropy, enthalpy, and hydrophobic effect each show a different temperature dependency. First, the entropy difference decreases and the enthalpy difference increases, which increases the stability of the native state, as intuitively expected. But because the hydrophobic effect decreases, the native fold is destabilized. This decrease shifts the free energy balance towards the denatured state, which eventually leads to cold denaturation. As calorimetric studies [71] have shown, most globular proteins show a change in heat capacity as non-linear function of temperature while the protein unfolds. These

general examples of the versatile behavior of water in our environment indicates that it is a very complicated as well as interesting research topic.

1.4 Protein Folding

Besides "cold denaturation" of folded proteins, folding itself is already one of the most challenging tasks in modern biophysics. It gets even more complicated if structure and function for a protein is to be predicted from a given amino acid sequence. Since 1994 in the Critical Assessment of Techniques for Protein Structure Prediction (CASP) [48], state of the art techniques are applied to accomplish this task with modest success. One way for finding the protein structure are calculations of possible folding pathways. This process can be described as a complex sequence of moves along a high-dimensional free energy landscape. Although the high dimension suggests that finding the native fold by random sampling will take astronomical amount of time, in the real world a protein finds its native fold within seconds or less. This is known as the "Levinthal Paradox" [54]. The answer lies in the hydrophobic effect, in intra and inter protein and as well as protein solvent interactions. Though intra protein interactions are widely studied, the influence on the surrounding molecules is little or not fully understood [55] but acknowledged to play a substantial role in the folding process [19].

1.5 Molecular Dynamic Simulations

A classic molecular dynamics simulations (MD) is an atomistic simulation method, which treats each atom as a point mass and describes the interactions between atoms with a simple force field. Trajectories are generated by integrating Newton's equation of motions. Hence, molecular dynamics simulations are an adequate tool for studying the presented effects between water and proteins.

Since 1949, numerical techniques to explore the behavior of simple gases, liquids and solid states have been developed, with the long-range goal to describe living systems in terms of their chemistry and physics. During the last 60 years, great progress has been made in applying physical laws, representing protein systems, solving structures and chemical reactions. In fact, molecular dynamics simulations have become a valuable tool in crystallography or nuclear magnetic resonance spectroscopy to help interpreting experimental results or understanding protein motions.

The big advantage is that the simulations provide individual particle motions as a function of time in atomic resolution as they can answer detailed questions about the properties of a system, often more easily than experiments. For many aspects of biomolecule function, it is these details that are of interest.

Another feature is that, although the force field used in the simulations is approximate, it is completely under the user's control, so that removing or altering specific contributions to a potential can help in understanding the mechanics of the system. For example, because thermodynamic potentials are path-independent, one can choose an unphysical change of states like switching off electrostatic interactions in alchemical reactions to determine free energy differences.

Molecular dynamics simulations complement experiments because they provide a detailed atomistic picture of the complete system to experimental observables. However, the quality of this picture dependents on the quality of the force field, which determines how close simulations reflect the real world.

1.6 Aim of the Present Thesis

To gain insight into the balance between proteins and their solvent environments, especially the influence of the solvent entropy, my tool of investigation are molecular dynamics simulations, which have been proved to provide an accurate atomistic description of this relationship. One way to capture this relation between solvent and solute is the exploration of the configurational space of the involved particles. This is neither experimentally nor in molecular dynamics simulations an easy task. Experimentally, it is not possible to track individual water molecules, or computationally, it is not possible to sample the full configurational space of all molecules. Recently, progress has been made to investigate solvent entropies in molecular dynamics simulations, however none of them is capable of including proteins and capturing their effect on their environment.

The aim of my thesis is to contribute to and develop a technique for calculating water entropy in molecular dynamics simulations, which is applicable to water protein systems as well. I will develop a general framework for the entropy estimation but the focus of my work lies on the development of an estimator for the orientational entropy of the solvent.

My thesis is organized as follows. I reflect current methods for entropy estimation in the literature review (chapter 2) and explain the different ways how these methods estimate solvent entropies. Then, in chapter 3 I explain the essentials of molecular dynamics simulations and the used water models

in my study (3.2).

Chapter 5 embeds the orientational entropy within a general framework for solvent entropy estimation. The known problems for entropy estimations are addressed in section 5.1. It consists of three parts, which individually address sampling issues, the curse of dimensionality, and the curved space of orientations. I will first address the sampling problem by using the permutation reduction (5.2) technique [73]. The method exploits the fact that water molecules are indistinguishable, such that relabeling via permutations is possible. This increases the sampling of the water molecules by the Gibb's factor $N!$.

Second, to address the curse of dimensionality of the orientational entropy integral, I use a mutual information expansion [37,38,57] (5.3). This method approximates the total entropy by a truncated sum of mutual information terms of increasing order. For a weakly coupled system, it has been shown [5] that each additional expansion order improves the estimate, with the full expansion providing an exact result.

Finally, for evaluation of the terms I develop a kernel density estimator (5.4) with isotropic spherical kernels on the curved space of orientations SO(3), which described the orientational data. In addition to estimate mutual information of orientations between n molecules properly, I will expand the estimator on the tensor product space $(\mathrm{SO}(3)^n)$.

In the last chapter 6, I show and discuss how the influence of the surroundings of a water molecule is visible and measurable and effects the orientational entropy of the molecule.

2

Concepts for Entropy Estimation

2.1 Entropy

In the 1870s, the modern view of entropy was developed by Ludwig Boltzmann. He defined entropy as a measure of the number of accessible micro states of all particles in a system. This definition unified the experimentally measurable macro states of a system with the micro states of individual particles known to exist.

For canonical ensembles with known Hamiltonians this is done with the following definition of the partition function $Z(V,T)$:

$$Z(V,T) = \int \exp\left(-\frac{H(\omega)}{k_\mathrm{B}T}\right) \mathrm{d}\omega \tag{2.1}$$

$$p(\omega) = \frac{1}{Z(V,T)} \exp\left(-\frac{H(\omega)}{k_\mathrm{B}T}\right) \tag{2.2}$$

$$S = -k_\mathrm{B} \int_\Omega p(\omega) \log\left(p(\omega)\right) \mathrm{d}\omega\,. \tag{2.3}$$

Here, $H(\omega)$ is the Hamiltonian of the system, k_B the Boltzmann constant, T the temperature, and V the volume of the system. Derived of the partition function, $p(\omega)$ is the probability distribution of states, and S the entropy of the system. Ω denotes the phase space and $\mathrm{d}\omega$ an infinitesimal volume element in the phase space.

In general, for systems with known Hamiltonians an evaluation of the entropy integral is possible. However, for non trivial systems like one hundred atoms the configurational space is already extremely large, such that a numeric evaluation of the entropy integral fails to converge. In such a case, the phase space $\Omega = \mathbb{R}^{600}$ is 600 dimensional, which makes sampling difficult. Bellman coined this problem as the *curse of dimensionality* [7] in 1957. The reason is the exponential scaling of the number of data points with the dimensionality of the system, N^D, for constant accuracy.

Although in molecular dynamics simulations all potential functions are under the users control, and the Hamiltonian of the system is fully known, only a few methods exist to evaluate entropies. Hence, I will briefly revise several methods, which have been developed to calculate entropy differences or entropies of subsystems of the molecular dynamics systems.

2.2 Protein Entropy

In 1993, Schlitter [75] refined the quasi-harmonic approach [44] from Karplus and Kushik to calculate entropies of proteins. The method deals solely with

the configurational entropy of the protein and does not address the solvent. The technique takes advantage of the fact that the protein resides in a global free energy minimum. Therefore, a quasi-harmonic approximation to the minimum of the free energy function suffices and leads to a probability density function given by a multivariate Gaussian function. Instead of describing the motion of the protein by internal curve-linear coordinates as Schlitter's predecessors, he used the collective atomic fluctuations of all atoms. His estimated entropy turned out to be an upper limit, because high frequency motions contributed too much, e.g. C−H vibrations. Thus, he dropped them because of quantum-mechanical arguments (high frequency motions are either not properly captured by molecular dynamics simulations or quantum-mechanically frozen states). His well-considered physical arguments, which were not based on rigorous derivations, have been shown to lead to a good estimator, which has been applied on a lot of different protein systems [16, 17, 65].

The method works by estimating the probability density $p(\omega)$ of a protein with N atoms by

$$p(x) \quad = \quad (2\pi)^{-\frac{3N}{2}} \det \mathbf{A} \exp\left(-\frac{1}{2}\left(x - \langle x\rangle\right)^T \mathbf{A}\left(x - \langle x\rangle\right)\right) \tag{2.4}$$

with $\mathbf{A}^{-1} = \mathbf{C}$ derived from the covariance matrix $\mathbf{C} = \langle (x - \langle x\rangle)(x - \langle x\rangle)^T\rangle$ of the coordinates x of the protein. The entropy of the protein is then calculated by equation (2.3). For molecules undergoing large conformational changes, anharmonicities will contribute and the entropy is likely to be considerably smaller. Additionally, since the method only captures protein contributions, it is normally used to test for and report entropy differences of simulations where the protein structure has been altered.

Entropy estimates of solvent molecules using a quasi-harmonic approximation fail to converge, because unlike proteins water molecules sample a wide, shallow free energy basin by diffusion with no well defined minimum structure.

2.3 Free Energy Perturbation

Several techniques based on the perturbation theory have been developed. The most prominent ones are the equilibrium method by Zwanzig: *free energy perturbation* [85] and the non-equilibrium method: *thermodynamic integration* [40, 46]. Both methods calculate the free energy differences ΔG between two different systems. Since the free energy,

$$G = H - TS\,,$$

is composed of enthalpy H, the systems temperature T and entropy S, access to the enthalpy and free energy will allow an estimate of the entropy of the given system.

In Zwanzig's perturbation method, two systems are identified by the initial state A for the first and final state B for the second system. For both systems, their respective Hamiltonians H_A, H_B are known. Hence, the free energy difference is

$$G_A - G_B = \Delta G_{A,B} = -k_\mathrm{B}T \log \left[\frac{\int \exp(-\beta H_A(\omega_A))\, \mathrm{d}\omega_A}{\int \exp(-\beta H_B(\omega_B))\, \mathrm{d}\omega_B} \right] .$$

With the difference between the two Hamiltonian, $H_A - H_B = \Delta H_{A,B}$, the free energy is rewritten as the exponential average of the difference on the entire ensemble of system B,

$$\Delta G_{A,B} = -k_\mathrm{B}T \log \left\langle \exp(-\beta \Delta H_{A,B}) \right\rangle_B .$$

With the configurations of the ensemble of state B, one generates the exponential average of the difference Hamiltonian $\Delta H_{A,B}$ and obtains the free energy difference between the two states. Further, the accuracy of the free energy difference is increased by sampling of state B being exact for infinite sampling.

The method obtained its name from perturbations in $\Delta H_{A,B} \ll \beta = {}^1\!/_{k_\mathrm{B}T}$, because the exponential average converges poorly for systems where the energetic difference is not small.

To access entropy differences, a separate calculation of the enthalpy remains challenging for larger systems, because the enthalpy calculation involves an accurate estimate of an ensemble average that includes the complete Hamiltonian of the system. However, solvent molecules add large fluctuations to the enthalpy term such that it converges extremely slow.

The problem was partially circumvented by splitting the Hamiltonian into solute and solvent [69]. With the assumption that the intra-solvent interactions remain undisturbed, the entropy difference is determined solely by the solute-solute, solute-solvent entropy and solvent-solvent energy, which converge faster because the large fluctuations originating from water are not included.

Free energy perturbation and thermodynamic integration are normally used to calculate free energy differences in systems, for example an amino acid mutation in a protein or the binding free energy of a ligand. Entropy calculations of solvents however fail to converge because of the previously mentioned large fluctuations in the enthalpy term.

2.4 Cell methods for liquids

Cell methods are one of the earliest liquid-state models to calculate free energies. Their name comes from the idea, that each molecule is treated as if it moves in separate "cells" and the energy of the system can be decomposed into individual effective potentials, each dependent only on a single molecule. The partition function of the liquid is then approximated by integrals over these simpler effective potentials and consequently the different thermodynamic potentials like free energy, enthalpy or entropy are accessible.

Recently, a cell method [31] has been parametrized for water with the help of molecular dynamics simulations. In the model, each water molecule is described as a rigid body moving in a six-dimensional anisotropic harmonic potential defined by the three translational and three rotational degrees of freedom. The model parameters are obtained from the molecular dynamics simulation by measuring the average potential energy per molecule and the average magnitude of force per molecule along three orthogonal axes, and the average magnitude of torques about these same axes. These seven parameters describe the shape of the potential energy surface per molecule and are not affected by molecular diffusion. Additionally, the selection of the parameters splits the thermodynamic potential of interest into conformational, translational and rotational contributions.

The molecular dynamics simulations have shown that the translational modes needed in the model are about 2 to 3 times softer than the rotational ones. Therefore, two-thirds of the water entropy are contributed to translation and one-third to rotation.

The main difficulties of such an approach lie within a proper decomposition of the energy of the highly correlated system together with an accurate model of the liquid. In addition, every new entropy estimate of mixtures or mixed protein water systems requires a new parametrization, because all correlations have to be included in the parameters, which are given by the environment of the molecules. Also, the harmonic approximation might not suffice to describe all subtle effects of the correlations properly. Shortcomings of such models are usually corrected with additional parameters, which rise the level of complexity of the model greatly.

2.5 Entropy via Radial Distribution Functions

Another set of methods is based on the Ornstein-Zernike equation defined in the density functional theory. The integral equation describes the correlations of molecules within a homogeneous isotropic fluid dependent on

distance and orientation. The idea behind such an approach is to describe a thermodynamic potential, here entropy, as a functional of the density of the fluid.

Generally, the method is based on the N-body correlation function

$$g^{(N)}(r^N, \omega^N)\,,$$

which is a function of the positions and orientations of all molecules in the fluid. Here, the exponent indicates the order of the correlation function. However, this function is too complex to calculate and therefore the approach is to describe the structure of the fluid by a series of correlation functions of increasing complexity

$$g^{(2)}(r^2, \omega^2),\ \delta g^{(3)}(r^3, \omega^3),\ \delta g^{(4)}(r^4, \omega^4),\ \ldots\,.$$

$g^{(2)}$ describes the pair correlation function between molecules and $\delta g^{(n)}$ describe the remaining n particle correlation in the general $g^{(n)}$ correlation function. For example $\delta g^{(3)}$ of three particle system is defined by

$$g^{(3)}(r^3, \omega^3) = g^{(2)}(1,2)g^{(2)}(1,3)g^{(2)}(2,3)\delta g^{(3)}(r^3, \omega^3)\,.$$

Here, $(1,2)$ denotes the positions and orientations of molecules 1 and 2.

The advantage of such an approach is a hierarchical description of the correlations within a system where the importance of each term is expected to decrease rapidly with increasing correlation order N. For fluids most information is expected to reside in the pair correlation function and the triplet function $\delta g^{(3)}$ and it is generally presumed that functions $\delta g^{(N)}$ for $N \geq 4$ rapidly approach unity as N increases, meaning they contain no more information of the system.

Given a way of numerically evaluating the correlation terms, the entropy of the system is defined by

$$s = s^{\mathrm{id}} - \frac{1}{2} k_{\mathrm{B}} \frac{\rho}{\Omega^2} \int g^{(2)} \log(g^{(2)}) - g^{(2)} + 1 \,\mathrm{d}r\,\mathrm{d}\omega^2 - \cdots - \,,$$

where s^{id} is the entropy of an ideal gas, k_{B} the Boltzmann's constant, ρ the number density of the fluid and Ω the integral over the Euler angles of one molecule.

Under certain conditions these distribution functions are evaluated for pair correlation functions. For instance, with spherically symmetric interactions between molecules the pair correlation function can be obtained by standard integral equation theories [30]. Also, for systems with a Hamiltonian consisting only of one and two body terms, solutions have been

found [22, 26, 30, 45, 51, 84]. Approaches to calculate the third order correlation functions [3, 4] exist as well.

The difficulties in calculation of the correlation functions arise from the fact that they depend on both, the distance and the relative orientation of the molecules. Although formally the integral equations are extended to include orientational degrees of freedom, solution of these equations is very laborious and difficult to obtain. As an alternative, integral equations have been developed for the calculation of site–site correlation functions. Several such equations have been applied to liquid water with considerable success.

Still, enhancements to these correlation functions were made by Kinoshita *et al.*, who developed an angle-dependent sphere-based orientational distribution method, which they used to estimate the orientational entropy for the solvent molecules to about 35%-42%, which is in agreement with the other methods.

The most criticized aspect of this method is the assumption that the structure of an isotropic liquid can be described by spherical symmetric functions. It has been shown that the spherical symmetric character of the total correlation function derived from the experimentally observed diffraction pattern does not exclude non-spherical symmetric domains [77]. Thus, an application to mixed systems should be done with care and the proper treatment of orientational correlations is still open for further development.

3

Molecular Dynamics Simulations

3.1 Simulations at Atomic Resolution

In this section, I describe the theoretical framework of molecular dynamics (MD) simulations. A detailed description can be found in many textbooks [14, 27, 82]. Because my systems consists of proteins and water molecules, an accurate description of the real system would imply to solve the non relativistic time-dependent Schrödinger equation

$$\hat{H}\psi = \mathrm{i}\hbar\frac{\partial\psi}{\partial t}\,. \tag{3.1}$$

Herein the quantum mechanic energy operator $\hat{H}$ drives the multi-particle wave function ψ. $\hbar$ is Planck's constant. However, solving this equation is computationally not possible for time scales of protein motions, ranging from picoseconds to hours. Therefore, the following simplifications are used:

- Movements of nuclei and electrons are uncoupled.
- The electronic state of a molecule is constant and defines an effective potential for the nuclei motions.
- The effective potential is approximated by a semi-empirical function, a force field, which describes the interactions between the atoms.
- Dynamics of the atoms are described by solving Newton's equation of motion.

Four major reasons allow this simplification. First, in 1927 M. Born and J. R. Oppenheimer showed that due to the extreme relative mass ratio between nuclei and electrons, the nuclei move in an effective potential given by instantaneously reconfiguring electrons [63]. This is an adiabatic separation of the electronic and nuclear coordinates in the wave function, which is only valid as long as the electronic configuration is not influenced by nuclear positioning. In fact, even in small systems the electronic reconfiguration, e.g. like in amide photo-dissociation, is not calculated, because it is computationally expensive and the effect is only within a few time steps of the integrator. Therefore, electron interactions are only included as an effective potential for the nuclei motion and otherwise neglected.

Second, the electronic ground state of a molecule, which defines the effective potential for the nuclei, is approximated by a semi-empirical function, a simple force field. The force field defines a new potential energy as a function of atomic coordinates and approximates all bonded and non-bonded interactions. The parameters for the force field, which describe the different

interaction terms, are chosen such that thermodynamic properties of a simulation are close to experimental values or quantum mechanical calculations. This depends on the focus of its parametrization. However, this leads to problems for systems where electronic ground states are valid but fluctuations or transfers of charge happen. For example, the charge of an atom is not changed due to dipole interactions. Generally, fluctuations as such are only captured within interaction terms or as positional fluctuations of atoms but not in the form of tunable parameters of the force field.

Third, in reality most of the time a molecule is in the electronic ground state. Only during in some chemical reactions or photon absorption, higher unoccupied states are involved. Therefore, representing protein motions as thermal fluctuations in the electronic ground state is generally valid. However, chemical reactions in bio-molecules are very common and are associated electronic reconfigurations, e.g. bond formation. Since a force field is used, which does not allow electronic reconfigurations, this is neglected. Hence, to mention a few easily overlooked examples, proton transfer or fluctuations of sulfide bridges are not included in simulations. Recently, new methods have been developed, which allow the change of protonation states [52, 78] and therefore new bond formation during simulations.

Fourth, since Ehrenfest's theorem [21] states that for sufficiently narrow wave packets, the expectation value of the operator follows the classical equation of motion, the system is well described by the Newton's equation of motion. At atomic resolution, the time scale for a doubling the width of a Gaussian wave packet is about a pico second while the integrator (see 3.1.1) uses femto second time steps. Therefore, the force on a point mass approximates the expectation value of the force operator, respectively.

Hence, in molecular dynamics simulations the following equations of motions are computed.

$$\vec{F}_i = -\nabla_i V(\vec{R}_1, \ldots, \vec{R}_n) = m_i \frac{d^2 \vec{R}_i}{dt^2} = m_i \vec{a}_i(t)\,. \tag{3.2}$$

$\vec{R}_i$ defines the position of the i-th atom following the classical equation of motion within the time-independent force field $V(\cdot)$, which is given by the following effective interaction terms

$$\begin{aligned} V \;=\; & \sum_{\text{bonds i}} V_B^i + \sum_{\text{bond angles j}} V_\alpha^j + \sum_{\text{imp. dihedral k}} V_{\text{imp}}^k + \sum_{\text{dihedral l}} V_{\text{dihedral}}^l \\ & + \sum_{\text{pairs } \alpha,\beta} \left(V_{\text{vdW}}^{\alpha\beta} + V_{\text{Coulomb}}^{\alpha\beta} \right) . \end{aligned}$$

Bonds V_B, bond angles V_α, improper V_{imp} and proper dihedral angles V_{dihedral} are all harmonic potentials, van-der-Waals V_{vdW} and Coulomb interaction

terms V_{Coulomb} are given by physical laws like charge interactions, dipole-dipole and Pauli repulsion. Table 3.1 illustrates the origin and properties of the different terms. The parameters of the force fields were optimized on equilibrium distances in simulations, depending on the target properties the force field was developed for [42, 43, 81, 83].

In my study, I have used the OPLS [42] all atom force field together with SPC/e [8] and TIP4P [41] water models.

As mentioned in my enumeration, thermodynamic properties are given by sampling of the phase space of the given system. In the context of molecular dynamics simulations, this means that any macroscopic observable A of the system is given by the ensemble average $\langle A \rangle_{\text{Ensemble}}$. If one assumes the ergodicity of the systems, one can also use the time average

$$A = \lim_{T \to \infty} \frac{1}{T} \int_T A(t)\, \mathrm{d}t \tag{3.3}$$

to calculate ensemble averages. This fact holds in general for entropies and free energies as well, however computational difficulties arise, especially for solvent molecules, which I will discuss later.

3.1.1 Integrator

A number of algorithms has been devised to efficiently generate MD trajectories. In this work the leap frog modification of the Verlet scheme of the GROMACS package [35] was used,

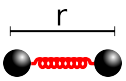
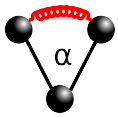
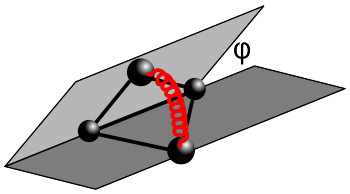
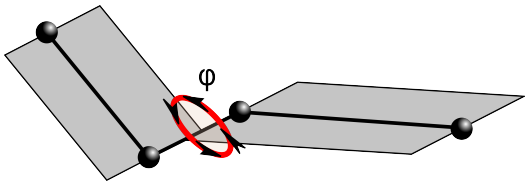
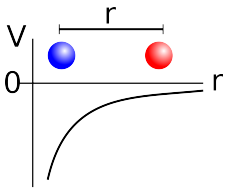
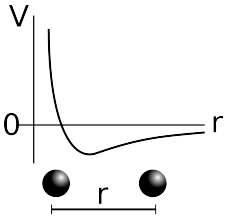
$$\begin{aligned} \vec{R}(t + \Delta t) &= \vec{R}(t) + \vec{v}(t + \frac{1}{2}\Delta t)\Delta t \\ \vec{v}(t + \frac{1}{2}\Delta t) &= \vec{v}(t - \frac{1}{2}\Delta t) + \vec{a}(t)\Delta t\,. \end{aligned}$$

The current position $\vec{R}(t)$ and acceleration $\vec{a}(t)$ of each atom is updated with mid-step velocities $\vec{v}(t + \frac{1}{2}\Delta t)$. This stable algorithm has the advantage that the expensive force calculations have to be done only once per integration step. Normally, the step size is two femto seconds. More details about the simulation setup is given in the method section.

3.2 Water models

I have used two different water models in my studies, which I want to introduce briefly and stress some differences of the models. Although the predictive nature [13, 80] of all water models has been questioned in the literature

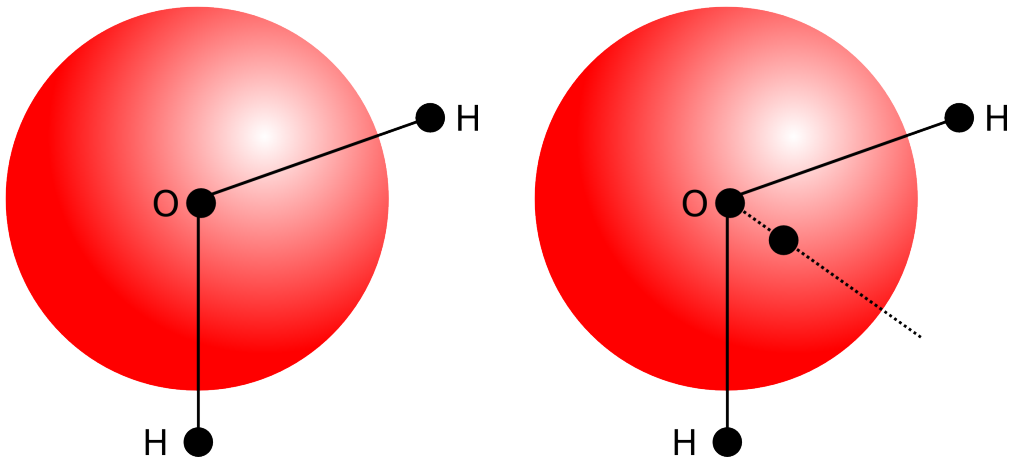
Table 3.1 Illustration of the different force field interactions.

Potential	Definition
	Bond potential between two atoms $V_B = {}^{k_B}/_2(r - r_0)^2$
	Bond angle potential $V_\alpha = {}^{k_\alpha}/_2(\alpha - \alpha_0)^2$
	Improper bond potential $V_{\text{imp}} = {}^{k_{\text{imp}}}/_2(\phi - \phi_0)^2$
	Proper bond potential $V_{\text{dih}} = k_{\text{dih}}(1 + \cos(m\phi - \delta))$
	Coulomb potential of two charges q_i and q_j $V_{\text{Coulomb}} = \frac{q_i q_j}{4\pi\epsilon_0 r}$
	Van-der-Waals potential with parameters A and B $V_{\text{vdW}} = \frac{B}{r^{12}} - \frac{A}{r^6}$

different times, the following two models are widely used and represent two typical non-polarizable force fields for water. They both reproduce density, dynamics, and dielectric and structural properties well.

3.2.1 SPC/e water model

The SPC/e model [8] (figure 3.1(a)) is an extension to its predecessor SPC [10] water model. The molecule is described by a simple point charge model positioned on the three atoms. The oxygen atom has a Lennard Jones interaction term. The two hydrogens are bonded with bond potentials to the oxygen. The structure of that model dictates a flat geometry. The extension introduced in 1987 accounts for the positive self-energy term given by induced dipole dipole interactions between molecules. The correction changes the charges and Lennard Jones parameters of the model such that polarizability effects are reflected in the electrostatic interactions of the system. Typical properties as the radial distribution function of oxygen-oxygen distances, heat of vaporization (which was wrong before the correction was introduced), and diffusion constant have been improved considerably.



(a) SPC/e (b) TIP4P

Figure 3.1 Illustration of the SPC/e and TIP4P water models. The black points denote the position of the point charges. The red sphere symbolizes the van-der-Waals potential of the oxygen atom. On the right, 3.1(b) the additional point charge in the HOH plane adds structural properties in the radial distribution function of the TIP4P model.

3.2.2 TIP4P water model

Two aspects of the SPC/e model have been criticized several times. First, the geometry of the molecule is not tetrahedral but rather flat and second, the radial oxygen oxygen distribution function does not show all experimentally observed properties. Although the geometry did not change in the TIP4P model (figure 3.1(b)), additional structure to its thermodynamic properties was given to the molecule by moving a part of the negative charge off the oxygen towards the hydrogens at a point M on the bisector of the HOH angle. This leads to an overall better radial distribution function and density of the system. Only with an additional charge out of the HOH plane, a tetrahedral structure can be obtained. This has been done in the TIP5P water model [56], where the additional charge represents the two lone electron pairs of the oxygen. Since most simulations of a protein water system consists of lots of water molecules, the TIP5P model is less used because it is computationally more expensive.

3.3 Simulation Details

To test the developed framework of entropy estimation for solvents, two different systems have been simulated, bulk water and a hydrophobic protein, Crambin. The simulation protocols are briefly summarized below.

3.3.1 Bulk Water

After equilibration 316 water molecules of SPCE/E [8] were simulated for 50 ns with the GROMACS [11, 35] package. The simulation was started from an equilibrated rectangular box with periodic boundary conditions and a 2 fs time step size was used in the integrator. The SETTLE [60] algorithm was applied to the water molecules, which enforces the distance constraints as discussed in section 4.3. The simulation system was coupled to a 300 K heat bath ($\tau_T = 0.1$ ps) with the velocity rescale algorithm [9]. The systems were isotropically coupled to a pressure bath at 1 bar with Parrinello-Rahman pressure coupling [64] ($\tau_P = 1$ ps). Short-range electrostatic and Lennard-Jones interactions were calculated within a cut-off of 1.1 nm and the neighbor list was updated every 10 steps. The particle mesh Ewald (PME) [18] method was used for the long range electrostatic interactions with a grid spacing of 0.12 nm. Every picosecond the configuration of the system was recorded.

3.3.2 Crambin

Crambin (1CBN) was solvated in 5,515 water molecules (TIP4P [41]) and simulated for 50 ns. For Crambin, the OPLS all atom force field [42] was used. The simulation was started from an equilibrated rectangular box with periodic boundary conditions and were integrated with 2 fs time step size. The SETTLE algorithm was applied to the water molecules and Lincs algorithm [34] to the protein. Three position restraints ($k = 5000$ kJ/(mol nm^2)) were applied to the C_α-atoms of the proline residues to remove the rotational freedom of the protein during the simulation. The water molecules were coupled to a 300 K heat bath ($\tau_T = 0.1$ ps) with the velocity rescale algorithm. The systems were isotropically coupled to a pressure bath at 1 bar with Parrinello-Rahman pressure coupling ($\tau_P = 1$ ps). Short-range electrostatic and Lennard-Jones interactions were calculated within a cut-off of 1.1 nm and the neighbor list was updated every 10 steps. The particle mesh Ewald (PME) method was used for the long range electrostatic interactions with a grid spacing of 0.12 nm. For physiological conditions of 150 mM sodium chloride, ions were added, respectively. Every picosecond the configuration of the system was recorded.

Electrostatic Surface Potential

All water molecules were stripped from the average structure of the simulation and the structure was processed with the PDB2PQR [20] package using the AMBER force field for charges. The protein dielectric was set to 2 and for water to 78. The Poisson-Boltzmann equation was solved with the linear solver of the ABPS package [6] and the solvent accessible surface potential evaluated for $T = 300$ K.

4

Properties of Orientations

4.1 Orientation of a Rigid Body

In my thesis, I present a method to estimate the orientational entropy of water obtained from molecular dynamics simulations. Therefore, I will briefly recap orientations and focus on the properties needed in the development.

Since a rigid body has three degrees of freedom for rotations, its orientation is given by three parameters. The parameters define uniquely a rotation matrix, which transforms the rigid body from his body fixed frame to the lab frame without changing its shape. Hence, the rotation matrix rotates a vector while preserving its length. The rotation matrices additionally obey

$$\mathcal{A} \in \mathbb{R}^{3\times 3},\ \mathcal{A}^T = \mathcal{A}^{-1} \quad \text{and} \quad \det(\mathcal{A}) = 1\,. \tag{4.1}$$

This set is also the matrix representation of the Lie group SO(3). That group is well studied and forms a curved, differentiable-connected Riemannian manifold. For this manifold and its tensor product spaces, I developed the density estimator. The key ingredients of the density estimator (see section 5.9) are distances between elements on the manifold and volumes as a function of this distance. Hence, I give an overview of the different charts on SO(3) used in the derivations of the volume integrals.

4.1.1 Euler angles

Since the Euler angles are widely discussed in many textbooks, e.g. [24], their parametrization is used as a reference in the calculations as well as in the derivation of volumes and the geodesic distances between elements of the group.

The Euler angles are defined by three pairwise different rotations around any of the coordinate axes $\{x, y, z\}$. Generally, the sequence can be arbitrarily chosen, but the most common sequence and the sequence used in my work is $\{z, x, z\}$. All three rotation matrices are given by

$$\mathcal{R}_x(\alpha) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & \sin(\alpha) \\ 0 & -\sin(\alpha) & \cos(\alpha) \end{pmatrix} \tag{4.2}$$

$$\mathcal{R}_y(\alpha) = \begin{pmatrix} \cos(\alpha) & 0 & \sin(\alpha) \\ 0 & 1 & 0 \\ -\sin(\alpha) & 0 & \cos(\alpha) \end{pmatrix} \tag{4.3}$$

$$\mathcal{R}_z(\alpha) = \begin{pmatrix} \cos(\alpha) & \sin(\alpha) & 0 \\ -\sin(\alpha) & \cos(\alpha) & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{4.4}$$

The domains of the Euler angles $\{\psi, \theta, \phi\}$ are $\psi, \phi \in (-\pi, \pi]$ and $\theta \in [0, \pi]$, which defines the rotation matrix

$$\mathcal{R}(\psi, \theta, \phi) = \mathcal{R}_z(\psi)\mathcal{R}_x(\theta)\mathcal{R}_z(\phi)\,. \tag{4.5}$$

The index of the rotation matrices denote the rotation axes.

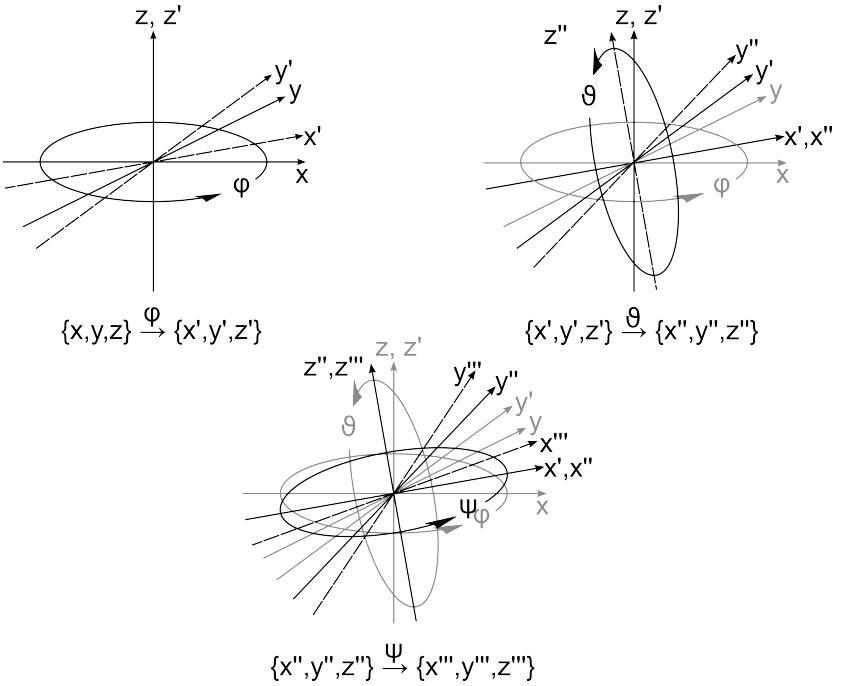


Figure 4.1 Definition of Euler angles. Three consecutive, pairwise different rotations represent any orientation of a rigid body. The sequence of rotations shown defines the Euler angles in my work. Starting in the upper left, the unprimed system is rotated around the z-axes with rotation angle ϕ, followed by a x-rotation with angle θ and again a z-rotation with rotation angle ψ.

Figure 4.1 illustrates the sequence of the three rotations. The advantage of Euler angles is, that they are easy to use but they lack computational stability and have singularities at the boundaries. This is known as *gimbal lock*. The first and third Euler angle are indistinguishable when θ is at 0 or π. The x-rotation becomes the identity rotation (for $\theta = 0$, $\theta = \pi$ leads to a similar situation) and the system is only given by two consecutive z-rotations, which can be explained by a single z-rotation. Therefore, the parameters ϕ and ψ are no longer well defined.

4.1.2 Skew-symmetric Matrices

The matrix representation of the corresponding Lie algebra $\mathfrak{so}(3)$ to the Lie group SO(3) form the set of skew-symmetric matrices,

$$\mathfrak{so}(3) = \left\{ \mathcal{A} \in \mathfrak{gl} | \mathcal{A}^T = -\mathcal{A} \right\} ,$$

where $\mathfrak{gl}$, the set of all linear transformation in $\mathbb{R}^3$, form the algebra of the general linear group GL(3).

The exponential of a skew-symmetric matrix $\mathcal{A}$ is a proper orthogonal matrix and element of SO(3) given by Rodrigues' formula as long as $a = \sqrt{\frac{1}{2}\mathrm{tr}\left(\mathcal{A}^T\mathcal{A}\right)} \in [0, \pi)$,

$$\mathrm{Exp}(\mathcal{A}) = \begin{cases} \mathcal{I}, & \text{if } a = 0, \\ \mathcal{I} + \frac{\sin(a)}{a}\mathcal{A} + \frac{1-\cos(a)}{a^2}\mathcal{A}^2, & \text{if } a \neq 0. \end{cases} \tag{4.6}$$

The inverse operation, the principal logarithm for a matrix $\mathcal{R}$ in SO(3) is the matrix in $\mathfrak{so}(3)$ given by

$$\mathrm{Log}(\mathcal{R}) = \begin{cases} \mathbf{0}, & \text{if } \theta = 0, \\ \frac{\theta}{2\sin(\theta)}\left(\mathcal{R} - \mathcal{R}^T\right), & \text{if } \theta \neq 0. \end{cases} , \tag{4.7}$$

where θ satisfies $\mathrm{tr}(\mathcal{R}) = 1 + 2\cos(\theta)$ and $|\theta| < \pi$ (this formula breaks down when $\theta = \pm\pi$).

The relationship between rotation matrices and skew-symmetric matrices allow a parametrization of SO(3) with parameters α, β, γ and the condition $\sqrt{\alpha^2 + \beta^2 + \gamma^2} \leq \pi$,

$$\mathcal{A}(\alpha, \beta, \gamma) = \begin{pmatrix} 0 & \alpha & \beta \\ -\alpha & 0 & \gamma \\ -\beta & -\gamma & 0 \end{pmatrix} \tag{4.8}$$

$$\mathcal{R}(\alpha, \beta, \gamma) = \mathrm{Exp}(\mathcal{A}(\alpha, \beta, \gamma)) . \tag{4.9}$$

The domain of the parameters α, β, γ is a sphere with radius $r \leq \pi$. A parametrization in the parameters of the skew-symmetric matrices is in particular useful for evaluating the norm on the manifold as well as calculating volumes.

4.2 Volume of a sphere on SO(3)

Having a parametrization of the group, measuring distances and volumes is completely given by the respective metric tensor g^{ij}. The shortest path

between two elements of the group is the length of their connecting geodesic γ, defined by $d(x,y) = \inf\{L(\gamma)\}, L(\gamma) = \int_0^1 \mathrm{d}t\sqrt{g^{ij}_{\gamma(t)}\dot{\gamma}_j(t)\dot{\gamma}_i(t)}$, and the volume of an area on the curved manifold is given by the integral over the respective domain $\int_\Omega \sqrt{g^{ij}g_{ij}}\,\mathrm{d}\omega$.

In the parametrization of Euler angles, the metric and geodesic equations on SO(3) are given by

$$g^{ij} = \begin{pmatrix} 2 & 0 & 2\cos(\theta) \\ 0 & 2 & 0 \\ 2\cos(\theta) & 0 & 2 \end{pmatrix} \tag{4.10}$$

$$\begin{aligned} \ddot{\psi} &= \dot{\theta}(\dot{\phi}\csc(\theta) - \dot{\psi}\cot(\theta)) \\ \ddot{\theta} &= -\dot{\phi}\dot{\psi}\sin(\theta) \\ \ddot{\phi} &= \dot{\theta}(-\dot{\phi}\cot(\theta) + \dot{\psi}\csc(\theta))\,. \end{aligned} \tag{4.11}$$

With the metric g^{ij}, the volume element of integration is evaluated to

$$\mathrm{d}\lambda_{\mathrm{SO(3)}} = \sqrt{g^{ij}g_{ij}}|\mathrm{d}\theta \wedge \mathrm{d}\psi \wedge \mathrm{d}\phi| \;= 2\sqrt{2}\sin\theta|\mathrm{d}\theta \wedge \mathrm{d}\psi \wedge \mathrm{d}\phi| \tag{4.12}$$

and the total volume to $V_{\mathrm{SO(3)}} = 16\sqrt{2}\pi^2$. Hence, the normalization gives the Haar measure for density estimation,

$$\mathrm{d}\mu_{\mathrm{SO(3)}} = \frac{1}{8\pi^2}\sin\theta|\mathrm{d}\theta \wedge \mathrm{d}\psi \wedge \mathrm{d}\phi|\,. \tag{4.13}$$

One key component of the entropy estimator (5.9) is the volume of a sphere centered at x_i as a function of the geodesic distance from x_i,

$$V(r) = \int_{\|\xi - x_i\| < r} \mathrm{d}\mu(\xi)\,. \tag{4.14}$$

For its derivation the parametrization 4.1.2 is used. The necessary calculations are limited to the neutral element $\mathcal{I}$, because the distance metric is bi-invariant in SO(3), i.e.

$$d_{\mathrm{SO(3)}}(\mathcal{P}\mathcal{R}_1\mathcal{Q}, \mathcal{P}\mathcal{R}_2\mathcal{Q}) = d_{\mathrm{SO(3)}}(\mathcal{R}_1, \mathcal{R}_2) \text{ for all } \mathcal{P}, \mathcal{Q} \in \mathrm{SO(3)}\,, \tag{4.15}$$

and thus the result is transferable to any point in SO(3). The domain of the parameters $\{\alpha, \beta, \gamma\}$ represents the volume of a sphere with radius $r \in [0, \pi)$. If the sphere is parametrized by spherical coordinates, it leads to simple distance and volume expressions. Figure 4.3 illustrates this double parametrization of the SO(3) group and figure 4.2 shows the volume of the sphere as a function of the geodesic distance from its center.

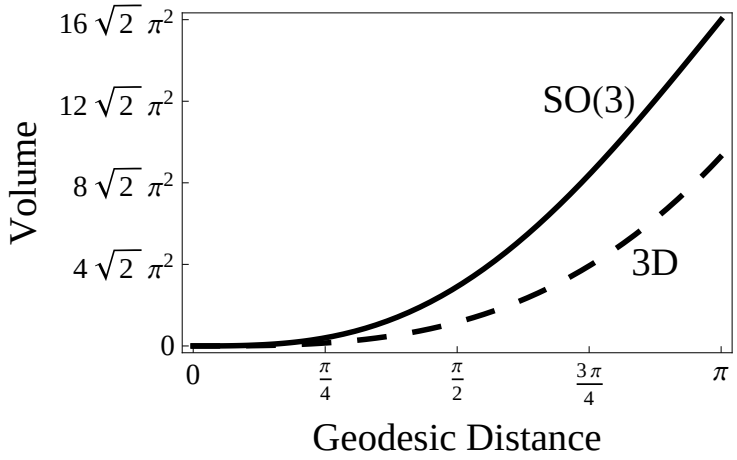


Figure 4.2 The volume of a sphere in SO(3) as a function of the geodesic distance from its center. The naive approach using a three-dimensional sphere as a volume is plotted with a dashed line for comparison. The deviation clearly shows that the volume needs to be correctly treated to obtain consistent results as estimates from the final density estimator.

$$
\begin{aligned}
d_{\mathrm{SO}(3)}(1, y) &= r_y && (4.16)\\
V_{\mathrm{SO}(3)}(r) &= \int_0^r \mathrm{d}r' \int_0^{2\pi} \mathrm{d}\phi \int_0^{\pi} \mathrm{d}\theta \, 8\sqrt{2} \sin\left(\frac{r'}{2}\right)^2 \sin(\theta) \\
&= 16\sqrt{2}\pi(r - \sin(r)) && (4.17)\\
\mu(V_{\mathrm{SO}(3)}(r)) &= \frac{1}{\pi}(r - \sin(r))\,. && (4.18)
\end{aligned}
$$

Here, $d_{\mathrm{SO}(3)}(1, y)$ is the geodesic distance between the neutral element and an element $y \in \mathrm{SO}(3)$, parametrized with spherical coordinates within the skew-symmetric parametrization, $V_{\mathrm{SO}(3)}(r)$ the volume of a sphere as a function of the geodesic from its center and $\mu(V_{\mathrm{SO}(3)}(r))$ the corresponding volume in the Haar measure.

For tensor product spaces $\mathrm{SO}(3)^n$, I demonstrate the necessary derivations only for the case $n = 2$, because the method is easily transferable to higher n. The tensor product of $\mathrm{SO}(3) \otimes \mathrm{SO}(3)$ is given by the Kronecker product of their representing matrices because the product is also the corresponding

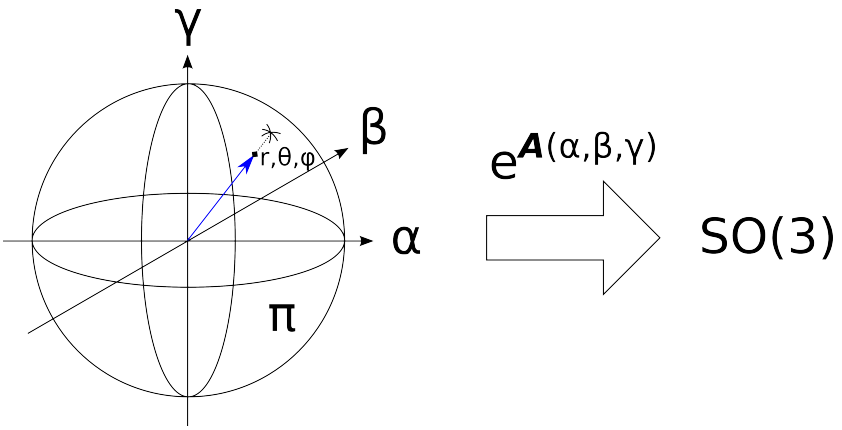


Figure 4.3 Parametrization of the domain of the skew symmetric matrices. The domain of $\{\alpha, \beta, \gamma\}$, a sphere with radius $r \leq \pi$ is parametrized with spherical coordinates $\{r, \theta, \phi\}$, which then represents a parametrization of the group SO(3).

matrix representation of the tensor product of the Lie groups.

$$\mathcal{A} \otimes \mathcal{B} = \begin{pmatrix} a_{1,1}\mathcal{B} & a_{1,2}\mathcal{B} & a_{1,3}\mathcal{B} \\ a_{2,1}\mathcal{B} & a_{2,2}\mathcal{B} & a_{2,3}\mathcal{B} \\ a_{3,1}\mathcal{B} & a_{3,2}\mathcal{B} & a_{3,3}\mathcal{B} \end{pmatrix} \tag{4.19}$$

$$= \begin{pmatrix} a_{1,1}b_{1,1} & a_{1,1}b_{1,2} & a_{1,1}b_{1,3} & a_{1,2}b_{1,1} & \dots \\ a_{1,1}b_{2,1} & a_{1,1}b_{2,2} & a_{1,1}b_{2,3} & a_{1,2}b_{1,1} & \\ a_{1,1}b_{3,1} & a_{1,1}b_{3,2} & a_{1,1}b_{3,3} & a_{1,2}b_{1,1} & \\ a_{2,1}b_{1,1} & a_{2,1}b_{1,2} & a_{2,1}b_{3,1} & a_{2,2}b_{1,1} & \\ \vdots & & & & \ddots \end{pmatrix} . \tag{4.20}$$

The derivation of the geodesic equations on the tensor product space is straight forward, because the equations decompose into two uncoupled sets of differential equations as in equation (4.11). Thus, the metric in the tensor product space $\mathrm{SO}(3)^n$ is given by the Euclidean sum of the metrics and the Haar measure is given by the product of the Haar measures of the subsystems

$$d_{\mathrm{SO}(3)^n} = \sqrt{\sum_{i=1}^{n} d^2_{\mathrm{SO}(3)_i}} \tag{4.21}$$

$$\mathrm{d}\mu_{\mathrm{SO}(3)^n} = \frac{1}{(8\pi^2)^n} \prod_{i=1}^{n} \sin\theta_i |\mathrm{d}\theta_i \wedge \mathrm{d}\psi_i \wedge \mathrm{d}\phi_i| \,, \tag{4.22}$$

which is equivalent to equation (4.32). For $n = 2$, this reads

$$d_{\mathrm{SO}(3)^2}\left((\mathcal{R}_1, \mathcal{R}_2), (\mathcal{P}_1, \mathcal{P}_2)\right) = \sqrt{d_{\mathrm{SO}(3)}(\mathcal{R}_1, \mathcal{P}_1)^2 + d_{\mathrm{SO}(3)}(\mathcal{R}_2, \mathcal{P}_2)^2}\,, \quad (4.23)$$

In the following paragraph, the volume of a sphere as a function of its radius in the tensor product space $\mathrm{SO}(3)^2$ is derived. Since as in equation (4.16) a parametrization in spherical coordinates of the skew-symmetric parameter domain, $\{r_i, \phi_i, \theta_i\}$, leads to

$$d_{\mathrm{SO}(3)^2}\left((\mathcal{R}_1, \mathcal{R}_2), (\mathcal{P}_1, \mathcal{P}_2)\right) = \sqrt{r_1^2 + r_2^2}\,, \quad (4.24)$$

an additional parametrization of the square, the domain of $\{r_1, r_2\}$ with polar coordinates $\{\rho, \phi_c\}$ yields a parametrization as a function of distance in the tensor product space. After applying all transformations and a short derivation, the integrand of the volume integral $V(r)$ reads

$$128\rho \sin(\theta_1)\sin(\theta_2)\sin\left(\frac{1}{2}\rho\sin(\phi_c)\right)^2 \sin\left(\frac{1}{2}\rho\cos(\phi_c)\right)^2\,, \quad (4.25)$$

where the parameters $\{\phi_1, \phi_2, \theta_1, \theta_2\}$ are integrated out to

$$2048\pi^2\rho\sin\left(\frac{1}{2}\rho\sin(\phi_c)\right)^2 \sin\left(\frac{1}{2}\rho\cos(\phi_c)\right)^2\,. \quad (4.26)$$

The boundary conditions for ϕ_c as a function of ρ if $\rho \in (\pi, \sqrt{2}\pi)$ are given by

$$r_1(\rho) = \arctan\left(\sqrt{\frac{\rho^2}{\pi^2} - 1}\right) \quad (4.27)$$

$$r_2(\rho) = \arctan\left(\frac{1}{\sqrt{\frac{\rho^2}{\pi^2} - 1}}\right)\,. \quad (4.28)$$

Using equations (4.26) to (4.28), the volume as a function of distance on the tensor product space is given by the integral:

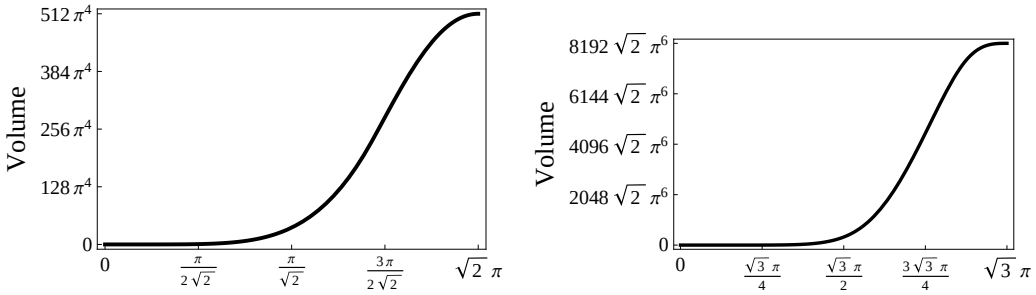
if $r \in [0, \pi]$

$$V_{\mathrm{SO}(3)^2}(r) = 2048\pi^2 \int_0^{\frac{\pi}{2}} \mathrm{d}\phi_c \int_0^r \mathrm{d}\rho\, \rho \sin\left(\frac{1}{2}\rho\sin(\phi_c)\right)^2 \sin\left(\frac{1}{2}\rho\cos(\phi_c)\right)^2$$

if $r \in (\pi, \sqrt{2}\pi)$

$$V_{\mathrm{SO}(3)^2}(r) = V_{\mathrm{SO}(3)^2}(\pi) + 2048\pi^2 \int_{r_1(\rho)}^{r_2(\rho)} \mathrm{d}\phi_c \int_\pi^R \mathrm{d}\rho\, \rho \sin\left(\frac{1}{2}\rho\sin(\phi_c)\right)^2 \sin\left(\frac{1}{2}\rho\cos(\phi_c)\right)^2\,.$$

I did not obtain a closed expression for that integral. Instead, I performed a numerical integration for one hundred equidistant nodes and implemented a B-spline interpolation algorithm for evaluation of the integral in the numerical analysis (Fig. 4.4).



(a) $\mathrm{SO}(3)^2$ (b) $\mathrm{SO}(3)^3$

Figure 4.4 The volume of a sphere in $\mathrm{SO}(3)^2$ and $\mathrm{SO}(3)^3$ as a function of the geodesic distance from its center.

The volume of a sphere as a function of its radius is similarly calculated for the tensor product space of $n \geq 2$. For example, for $n = 3$ the same parametrization is used as for $n = 2$ and the metric reads $d_{\mathrm{SO}(3)^3} = \sqrt{\rho^2 + r_3^2}$, $\rho \in [0, \sqrt{2}\pi]$ and $r_3 \in [0, \pi]$. Thus, if one more parametrization in polar coordinates $\{\tilde{\rho}, \tilde{\phi}\}$ is used to parametrize the rectangle of $\{\rho, r_3\}$, the integral as a function of $\tilde{\rho}$ is straight forward obtained. Care has to be taken with the new boundary functions for $\tilde{\phi}$ as in (4.28) for ϕ_c.

From a physics point of view the volume of sphere as a function from its geodesic distance has to be independent of its position, because orientations of a body in an isotropic, homogeneous space are independent from any underlying coordinate system. Thus, one can always choose the current orientation as the neutral element in $\mathrm{SO}(3)$. This holds true even for $n \geq 1$ in the tensor product spaces $\mathrm{SO}(3)^n$. Hence, the numerically computed volume expressions for $n = 2, 3$ are also transferable to any point within the tensor product spaces.

4.3 Change of Coordinates, Euler Angles and Center of Mass

Since in MD the interactions between water molecules are defined in Cartesian coordinates of the positions of their composing atoms, the transforma-

tion of the general entropy integral (2.3) to the Euler angle and center of mass coordinates is presented, because the orientational entropy(4.33) is defined on their orientational phase space. A transformation to center of mass coordinates introduces a molecule-specific constant, which is proportional to the moments of inertia of the molecule and adds a constant contribution to the total entropy of the molecule.

To avoid units in function arguments, e.g. in the logarithm, all physical quantities are replaced by its dimensionless numerical factors. For example, the position $\vec{r}$ is replaced by $\tilde{\vec{r}} = \frac{\vec{r}}{[\vec{r}]}$, where $[\cdot]$ returns the respective unit. Such a change in variables does not alter the definition of entropy.

Further, the transformation is demonstrated on the probability density of a single water molecule. The result is easily transferable to the entropy integral of multiple molecules. Since the focus is on the configurational entropy of the molecule, all dependencies on momenta have previously been integrated out.

In molecular dynamics simulations, water molecules (see 3.2) are normally treated as rigid bodies to speed up simulations. In Cartesian coordinates the positions of their composing atoms are defined by $\vec{r}_{\mathrm{O}}$, the position of the oxygen, and $\vec{r}_{\mathrm{H}_1}, \vec{r}_{\mathrm{H}_2}$, the positions of the two hydrogen atoms. To avoid indices, $\{r\}$ is defined as the set $\{r\} = \{\vec{r}_{\mathrm{O}}, \vec{r}_{\mathrm{H}_1}, \vec{r}_{\mathrm{H}_2}\}$. Further, the coordinates of the molecule are not independent of each other but have to satisfy three constraints, the angle between the oxygen and both hydrogen atoms $\sphericalangle\left(\vec{r}_{\mathrm{H}_2} - \vec{r}_{\mathrm{O}}, \vec{r}_{\mathrm{H}_1} - \vec{r}_{\mathrm{O}}\right) \stackrel{!}{=} \alpha_{\mathrm{H_2O}}$, and the distances between the oxygen and both hydrogen atoms $|\vec{r}_{\mathrm{O}} - \vec{r}_{\mathrm{H}_1}| \stackrel{!}{=} r_{\mathrm{OH}}$ and $|\vec{r}_{\mathrm{O}} - \vec{r}_{\mathrm{H}_2}| \stackrel{!}{=} r_{\mathrm{OH}}$ are constant. For systems with flexible water, the following transformation still holds but without the δ-distributions for the constraints.

Using the conditions outlined above, the configurational part of the partition function and the probability distribution of a single water are defined by

$$Z(V,T) = \int \mathrm{d}\{r\} e^{-\frac{H(\{r\})}{k_{\mathrm{B}}T}} \delta\left(|\vec{r}_{\mathrm{O}} - \vec{r}_{\mathrm{H}_1}| - r_{\mathrm{OH}}\right) \delta\left(|\vec{r}_{\mathrm{O}} - \vec{r}_{\mathrm{H}_2}| - r_{\mathrm{OH}}\right) \cdot \delta\left(\sphericalangle\left(\vec{r}_{\mathrm{H}_2} - \vec{r}_{\mathrm{O}}, \vec{r}_{\mathrm{H}_1} - \vec{r}_{\mathrm{O}}\right) - \alpha_{\mathrm{H_2O}}\right) \quad (4.29)$$

$$p(\{r\}) = \frac{1}{Z(V,T)} e^{-\frac{H(\{r\})}{k_{\mathrm{B}}T}} \delta\left(|\vec{r}_{\mathrm{O}} - \vec{r}_{\mathrm{H}_1}| - r_{\mathrm{OH}}\right) \delta\left(|\vec{r}_{\mathrm{O}} - \vec{r}_{\mathrm{H}_2}| - r_{\mathrm{OH}}\right) \cdot \delta\left(\sphericalangle\left(\vec{r}_{\mathrm{H}_2} - \vec{r}_{\mathrm{O}}, \vec{r}_{\mathrm{H}_1} - \vec{r}_{\mathrm{O}}\right) - \alpha_{\mathrm{H_2O}}\right), \quad (4.30)$$

where the δ-distributions enforce the constraints.

The bijective transformation to the center of mass coordinates $\vec{x}_{\mathrm{CoM}}$, Euler angles $\{\phi, \psi, \theta\}$ (following [66–68]), and three internal coordinates r_{OH_1}, r_{OH_2} and $\gamma_{\mathrm{H_2O}}$, which are chosen identically to the constraint definitions, change the probability density function to:

$$\begin{aligned} p\left(\{r\}\right) \mathrm{d}\{r\} = \left(\frac{M}{m_{\mathrm{O}}}\right)^3 & p\left(\vec{x}_{\mathrm{CoM}}, \theta, \phi, \psi, r_{\mathrm{OH}_1}, r_{\mathrm{OH}_2}, \gamma_{\mathrm{H_2O}}\right) r_{\mathrm{OH}_1}^2 r_{\mathrm{OH}_2}^2 \\ & \cdot \sin\left(\gamma_{\mathrm{H_2O}}\right) \sin\left(\theta\right) \delta\left(r_{\mathrm{OH}_1} - r_{\mathrm{OH}}\right) \delta\left(r_{\mathrm{OH}_2} - r_{\mathrm{OH}}\right) \\ & \cdot \delta\left(\gamma_{\mathrm{H_2O}} - \alpha_{\mathrm{H_2O}}\right) \mathrm{d}\vec{x}_{\mathrm{CoM}}\, \mathrm{d}\theta\, \mathrm{d}\phi\, \mathrm{d}\psi\, \mathrm{d}r_{\mathrm{OH}_1}\, \mathrm{d}r_{\mathrm{OH}_2}\, \mathrm{d}\gamma_{\mathrm{H_2O}}\,. \end{aligned} \tag{4.31}$$

The transformation into the center of mass coordinates introduces the mass term $\frac{M}{m_{\mathrm{O}}}$ with the total mass $M = 2m_{\mathrm{H}} + m_{\mathrm{O}}$. m_{O} and m_{H} are the masses of an oxygen and hydrogen atom. Integration over the constraint variables $\{r_{\mathrm{OH}_1}, r_{\mathrm{OH}_2}, \gamma_{\mathrm{H_2O}}\}$ leads to the final result:

$$p\left(\{r\}\right) \mathrm{d}\{r\} \;=\; C_{\mathrm{H_2O}} p\left(\vec{x}_{\mathrm{CoM}}, \theta, \phi, \psi\right) \sin\left(\theta\right) \mathrm{d}\vec{x}_{\mathrm{CoM}}\, \mathrm{d}\theta\, \mathrm{d}\phi\, \mathrm{d}\psi\,. \tag{4.32}$$

The factor $C_{\mathrm{H_2O}} = \left(\frac{M}{m_O}\right)^3 r_{\mathrm{OH}}^4 \sin\left(\alpha_{\mathrm{H_2O}}\right)$ is molecule-specific and given by the used force field, which defines the constraint distances, masses and angles.

With (4.32) the configurational entropy of a single water molecule is given by

$$S_{\mathrm{conf}} \;=\; -k_{\mathrm{B}} C_{\mathrm{H_2O}} \int p\left(\vec{x}_{\mathrm{CoM}}, \theta, \phi, \psi\right) \log(p\left(\vec{x}_{\mathrm{CoM}}, \theta, \phi, \psi\right)) \sin\left(\theta\right) \mathrm{d}\psi \mathrm{d}\phi \mathrm{d}\theta\,.$$

Therefore, integrating over the center of mass coordinates and dropping the molecule constants, the general term for the orientational entropy of n particles with $\Omega_i = \{\psi_i, \theta_i, \phi_i\}$ is given by

$$S \;=\; -k_B \int \cdots \int p\left(\Omega_1, \ldots, \Omega_n\right) \log(p\left(\Omega_1, \ldots, \Omega_n\right)) \prod_{i=1}^{n} \sin(\theta_i) \mathrm{d}\Omega_i\,. \tag{4.33}$$

5

Framework for Entropy Estimation of Solvents in MD

5.1 Entropy in Molecular Dynamics Simulations

For estimating solvent entropies in molecular dynamics simulations, all methods mentioned in chapter 2 have to overcome three major problems. First, because the configurational phase space is extremely large even for a small system consisting of a small number of molecules, sampling is poor. This is *the curse of dimensionality*. Second, solvent molecules reside in a much larger and shallow free energy basin, which impedes sampling as well. Third, solvent molecules have to sample the configurational space by diffusion, which is slow.

In this chapter, I will present the framework for entropy estimation of solvents and will address the orientational entropy of the solvent in a protein solvent system. Within the framework the slow and poor sampling of water molecules in simulations was solved by using the permutation reduction technique [73] (section 5.2), second, a mutual information expansion [37, 38, 57] was used to reduce the dimension of the entropy integral (4.33) and third, I developed an estimator for orientational entropies of multiple particles.

5.1.1 Orientational Entropy of Water as Part of the Total Entropy

To put the orientational entropy of water into the context of entropy estimates of solvated proteins, I start with a brief layout of notation and definitions. The total entropy of the system is separated into protein and water entropy and their mutual information (Fig. 5.1),

$$S = S_{\mathrm{P}}(\mathrm{P}) + S_{\mathrm{W}}(\mathrm{W}) - I(\mathrm{P},\mathrm{W}) \tag{5.1}$$

where P and W represent the protein and water system, respectively. This decomposition treats each entropy term independently and moves any correlations into the mutual information term. I will refer to $S_{\mathrm{W}}(\mathrm{W})$ as water entropy, $S_{\mathrm{P}}(\mathrm{P})$ as protein entropy and $I(\mathrm{P},\mathrm{W})$ as the mutual information between the water and protein system. For estimating the entropy of the protein, methods like Schlitter's formula [44, 75] (2.2) exist. The treatment of the mutual information term will be sketched in the outlook section of the estimator 7.3.

Similar to the above treatment of the total entropy, the water entropy is subdivided into translational and rotational entropy terms,

$$S_{\mathrm{W}}(\mathrm{W}) = S_{\mathrm{R}}(\mathrm{R}) + S_{\mathrm{T}}(\mathrm{T}) - I(\mathrm{R},\mathrm{T})\,, \tag{5.2}$$

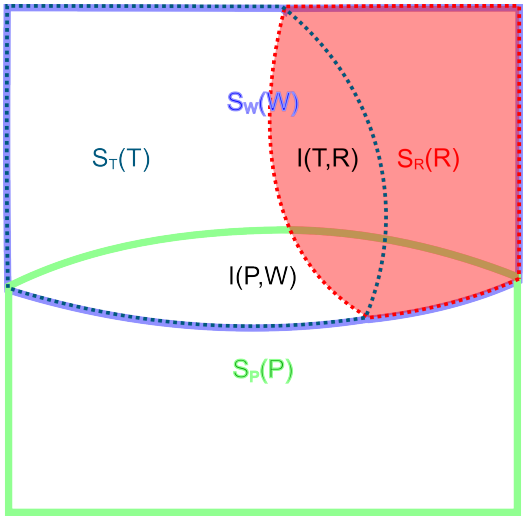


Figure 5.1 Illustration of entropy decomposition via mutual information $I(\cdot,\cdot)$. Light green denotes the entropy of the protein ($S_{\mathrm{P}}(\mathrm{P})$), light blue the entropy of the water ($S_{\mathrm{W}}(\mathrm{W})$), dark blue translational ($S_{\mathrm{T}}(\mathrm{T})$), and red rotational entropy of water ($S_{\mathrm{R}}(\mathrm{R})$). $I(\mathrm{P},\mathrm{W})$ is the area bounded by light green and light blue, $I(\mathrm{R},\mathrm{T})$ is the area bounded by red and dark blue. The presented entropy estimator addresses the area shaded in red.

where R and T represent subspaces spanned by the rotational and translational degrees of freedom of all water molecules.

The mutual information between these subsystems $I(\mathrm{R},\mathrm{T})$ is expected to have contributions of orientations and positions of the same molecule as well as correlations between the orientations of one molecule with the positions of its neighboring molecules, and therefore the term is generally not negligible. The entropy of translation and its mutual information with rotations are out of the scope of my thesis and are partially discussed elsewhere [32, 72], additionally the estimation of this mutual information term is still ongoing research in the department of my supervisor.

The entropy originating from internal vibrations can be accounted for by standard techniques [58] and is therefore also not considered. Hence, I focus on the orientational entropy term, which contains all orientational correlations of all water molecules. Describing each water molecule by its center of mass and three rotational Euler angles (see section 4.3), abbreviated with $\Omega_i = \{\psi_i, \theta_i, \phi_i\}$, the orientational water entropy reads

$$S_{\mathrm{R}}(\mathrm{R}) = -k_B \int \cdots \int \rho\left(\Omega_1, \ldots, \Omega_m\right) \log\left(\rho\left(\Omega_1, \ldots, \Omega_m\right)\right) \prod_{i=1}^{m} \sin(\theta_i) \mathrm{d}\Omega_i \,.$$

The aim of the next sections is to develop a framework for entropy estimation, which tackles this high dimensional entropy integral and allows further advancement to the mutual information term $I(P, W)$ between protein and water. Accordingly, the following section introduces a method for enhancing the phase space sampling of water molecules in molecular dynamics simulations.

5.2 Permutation Reduction

The first method in the framework addresses the sampling problem of water in molecular dynamics simulations. In contrast to solvated biological macromolecules, which often adopt a well-defined average structure, water molecules sample the entire simulation volume by diffusion and therefore their phase space densities converge too slowly to obtain results. The molecule has to sample the entire simulation volume as a part of its phase space, because all water molecules are distinguishable. If they are no longer distinct, each water molecule can contribute to the phase space of all water molecules and phase space sampling is enhanced.

This is done by exploiting the permutation symmetry of the Hamilton function of the system [73]. The physics is left invariant under the permutations of the indices of the water molecules, but a permutation translates the water configuration $\mathbf{x}(t_i)$ in phase space. Now, by choosing a permutation for every water configuration $\mathbf{x}(t_i)$ such that the root mean square deviation between the current configuration $\mathbf{x}(t_i) = \{\vec{r}_1(t_i), \ldots, \vec{r}_N(t_i)\}$ and a reference configuration $\mathbf{x}(t_0) = \{\vec{r}_1(t_0), \ldots, \vec{r}_N(t_0)\}$ is minimized,

$$\pi\left(\mathbf{x}(t_i)\right) = \{\vec{r}_{\pi(1)}(t_i), \ldots, \vec{r}_{\pi(N)}(t_i)\} \tag{5.3}$$

$$\left|\pi\left(\mathbf{x}(t_i)\right) - \mathbf{x}(t_0)\right| \stackrel{!}{=} \min, \tag{5.4}$$

the water configurations are so translated that the phase space of each molecule is more dense than before. Consequently, every molecule contributes to the phase space sampling of each molecule, which leads to a sampling increase by the Gibbs factor $N!$. The condition of the permutation does not restraint the mobility of a molecule. Rather it collects the molecules' contributions to reference positions.

Figure 5.2 shows the permutation reduction technique on an example of three water molecules. On the left the reference configuration and in the center the configuration before the permutation is shown. If the molecules remain distinguishable, each molecule has to sample the entire box. However, after application of the permutation reduction technique the effective volume

of each molecule is decreased and sampling is enhanced. Figure 5.3 shows the result of a sample trajectory of 316 water molecules, where 200 configurations of 4 randomly chosen molecules are shown before and after the permutations.

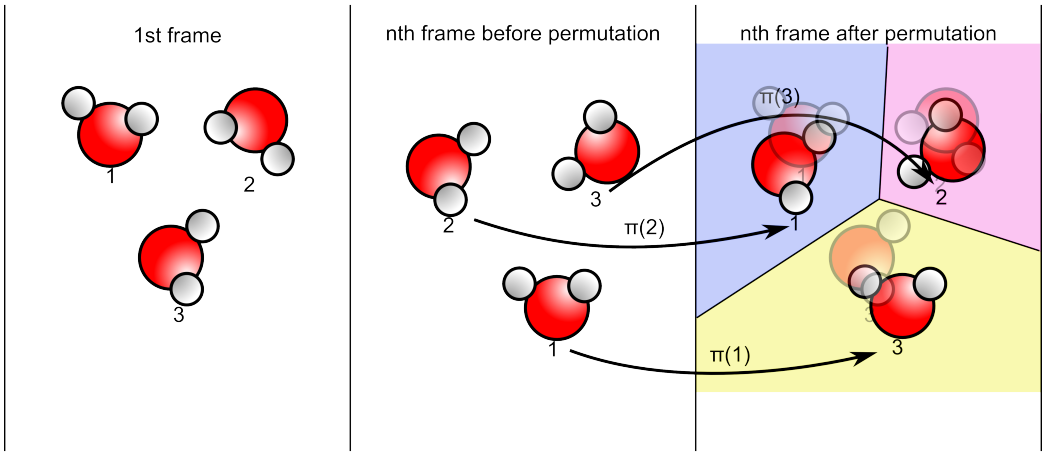


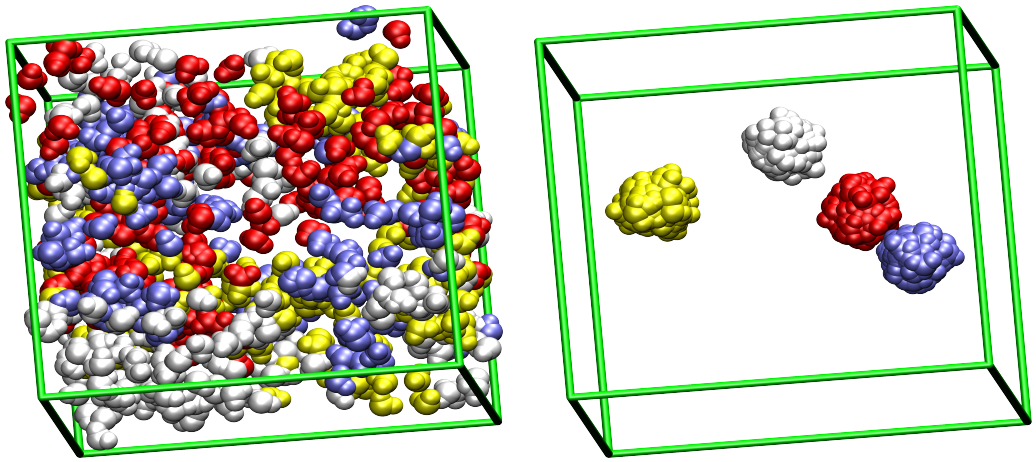
Figure 5.2 Illustration of the permutation reduction technique. The configurations of the first frame, which also serves as the reference configuration, and the nth frame are illustrated. The labeling of the molecules in the nth frame is changed by a permutation such that the root mean square deviation of the relabeled molecules in respect to the first frame is minimized. The molecules loose their distinctness in exchange for a reduced phase space and consequently improved sampling. The phase space reduction is indicated by the light color-coded background of the molecules.

By construction the diffusive motion of each molecule is converted into localized fluctuations around its reference position $\mathbf{x}(t_0)$, where the sampling per volume is highly increased. Accordingly, I will refer to the mean positions of the molecules during the simulation as spatial resolution.

5.3 Mutual Information Expansion

The second method in the framework, the truncated mutual information expansion reduces the high dimensionality of the entropy integral (see eq. (4.33)).

The mutual information expansion splits the numerically intractable high dimensional entropy integral (4.33) into subsystems of lower dimensionality capturing correlations of increasing order. Assuming that the contribution of higher order correlations to the entropy is small, a truncation of this



(a) before permutations (b) after permutations

Figure 5.3 Effect of the permutation reduction technique on phase space sampling of water molecules. 200 configurations of four water molecules color-coded by their index are displayed. They illustrate the effective trajectory of a molecule of "constant" index before and after applying the permutation reduction technique. The trajectories of the permuted system on the right are localized to their reference positions given by the first frame. Since the index is changed every time a molecule passes the reference positions,the molecules do not diffuse through the simulation box. The density of the sampling points is increased after applying the permutation reduction technique.

expansion reduces the problem to a series of numerically tractable integrals of sufficiently low dimensionality. As a subsystem, I label subspaces spanned by any combination of degrees of freedom of the atoms in our simulation box. Examples are all rotations of all water molecules, all coordinates of protein atoms, projection of the protein motions on a specific linear combination of atomic coordinates or the set of coordinates of three water molecules.

In the context of orientational entropy I consider the rotational freedom of each water molecule as the smallest subsystem, which has a dimension of three. All other subsystems in the mutual information expansion are combinations of the smallest subsystem and therefore have a dimension of multiples of three.

The idea of the mutual information expansion is illustrated in Venn diagrams in figure 5.1 and 5.4, which represent the entropy of subsystems and their mutual information as sets and subsets. The union of the translational and rotational subsystem and protein subsystem $S_{\mathrm{T}}(\mathrm{T}), S_{\mathrm{R}}(\mathrm{R})$ and $S_{\mathrm{P}}(\mathrm{P})$ represents the total system $S(\mathrm{T}, \mathrm{R}, \mathrm{P})$. But the sum of their entropies over-

estimates the total entropy because the mutual information among them $I(\mathrm{T},\mathrm{P})$, $I(\mathrm{T},\mathrm{R})$ and $I(\mathrm{R},\mathrm{P})$, which are the intersections of the sets, has been counted multiple times. In figure 5.4, the next order in the expansion is il-

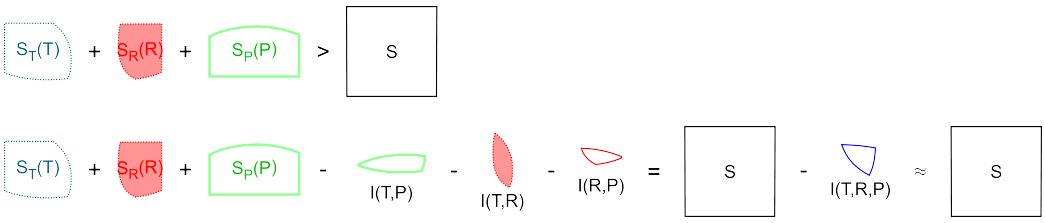


Figure 5.4 Illustration of the entropy expansion of a system with three subsystems. For a detailed description, see the text in 5.3. To reduce the dimensionality of the integral, the last mutual information term $I(\mathrm{T},\mathrm{R},\mathrm{P})$ is dropped to approximate the true entropy with lower dimensional entropy integrals.

lustrated in the second line, where the mutual information terms $I(\mathrm{T},\mathrm{P})$, $I(\mathrm{T},\mathrm{R})$ and $I(\mathrm{R},\mathrm{P})$ are included. The total entropy of the system is then approximated by neglecting the mutual information term $I(\mathrm{T},\mathrm{R},\mathrm{P})$, whose integral has the same dimensionality as the total system. I emphasize that the dimension of each entropy integral $S(\cdot)$ and mutual information integral $I(\cdot,\cdot)$ on the left side of the equation is lower than the integral of the full space.

In weakly correlated systems, mutual information terms of higher order are small and therefore the total entropy is approximated by neglecting them. This reduces the high dimensional numerically intractable integral to a sum of mutual information terms of lower dimensionality.

For a system of s subsystems, the mutual information expansion [37,38,57] reads

$$S(1,\ldots,s) = \underbrace{\sum_{m=1}^{m_t}(-1)^{m+1}\sum_{i_1<\cdots<i_m} I_m(i_1,\ldots,i_m)}_{\text{entropy estimate with correlation order } m_t} + \underbrace{\sum_{m=m_t+1}^{s}(-1)^{m+1}\sum_{i_1<\cdots<i_m} I_m(i_1,\ldots,i_m)}_{\text{correlations of order higher than } m_t}. \quad (5.5)$$

Here, $I_m(i_1,\ldots,i_m)$ are mutual information terms of m subsystems. The summation over the ordered index set $i_j \in \{1,\ldots,s\}$ guarantees the proper

recurrence of the mutual information terms,

$$I_m(1,\ldots,m) = \sum_{l=1}^{m}(-1)^{l+1} \sum_{i_1<\cdots<i_l} S(i_1,\ldots,i_l)\,. \tag{5.6}$$

Also in this general case the truncation order m_t lowers the dimensionality of the entropy integral by dropping higher correlations than m_t in (5.5). This leaves a summation of entropy integrals over subsystems of lower dimensionality and therefore allows a numerical treatment of the problem.

As an illustration, two extreme cases are considered. First, for $m_t = 1$ all correlations including pair correlations are neglected, $S(1,\ldots,s) = \sum_{m=1}^{s} S(m)$. Second, for $m_t = s$ the alternating sum of the right side in (5.5) reduces to $S(1,\ldots,s)$ and is therefore exact.

5.3.1 Fill Modes

For the discretized data provided by molecular dynamics simulations, only a finite set of configurations is available as an approximation for the configurational space density, implying sparse sampling of the high dimensional space. For this reason, the above introduced dimension reduction is essential. However, when combined with the MIE, two issues arise.

First, according to equation (5.6), the expansion is a sum of entropy estimates derived from subsystems with different dimensionality. Therefore, working with a finite, fixed sample size, the statistical accuracy of each term is different, and the error is dominated by the term with highest dimensionality.

The second issue arises from the third method in the framework, the density estimator for the entropy terms in the MIE (presented in detail in section 5.4). The estimator rests on the straightforward assumption that the local density at sample point x_i is approximately given by

$$\rho(x_i) \approx \frac{k}{NV(r_i(k))},$$

where N is the sample size and $V(r_i(k))$ denotes a volume of a sphere with radius $r_i(k)$, which is chosen such that the sphere is centered at x_i and contains k sample points.

The estimator introduces a systematic bias when combining configurational sets of different dimensionality. Although the estimators converges asymptotically for infinite sampling points [25] to the true entropy, in figure 5.5 the numerical experiments show that a bias for finite sampling sizes exists, which is not negligible. This bias is due to two possible reasons, either the length scale [47] to the k-nearest neighbor or the shape of the volume

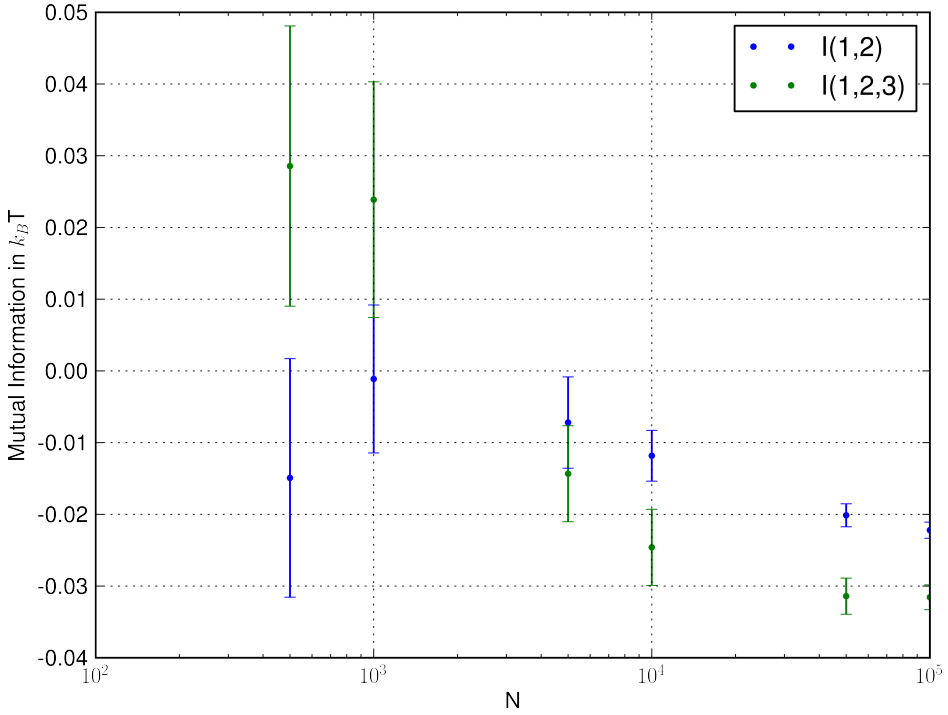


Figure 5.5 Mutual information as a function of sampling points of uncorrelated systems. To demonstrate the bias introduced by the density estimator, the mutual information of two and three independent, one-dimensional systems is shown. For each subsystem, N samples were drawn from a normal distribution $\mathcal{N}(0, 1)$ and the mutual information of the combined system estimated. For each shown N, the mutual information estimate was repeated 100 times and the standard deviation of the mutual information distribution is shown as error bars. The expected mutual information of the systems is zero. However, with increasing sample size an increasing bias is observed for both mutual information estimates, which is due to different length scales or shape of estimator volumes.

elements. The latter is less well documented in the literature and therefore illustrated in figure 5.6.

I consider the mutual information term between two independent systems, for which the mutual information is zero. For finite N, the mutual information is estimated according to (5.9) via

$$\tilde{I}(1,2) = -k_{\mathrm{B}} \left(\frac{1}{N} \sum_{i=1}^{N} \log \left(\frac{k}{N} \frac{V_{(1,2)}(r_{i,k})}{V_{(1)}(r_{i,k}) V_{(2)}(r_{i,k})} \right) + f(k) \right) , \quad (5.7)$$

where N is the number of configurations, $r_{i,k}$ the distance between the ith data point and its kth neighbor, $V_{(\cdot)}(r)$ the volume of sphere as a function of its radius r in the specific subsystem and $f(k)$ the bias correction term of a single entropy estimator defined in equation (5.9).

To distinguish the bias in the length scale and the shape of the volume in the subsystems, I assume that for each subsystem, N sampling points are drawn from a homogeneous density, and that both samples have equal spatial resolution. With fixed k in the systems (1) and (2) the size of an estimator volume is constant. Consequently, in the combined system $(1,2)$ $k' \approx k^2/N$ points are observed within the product volume of system (1) and (2), and the mutual information term is approximately zero. For a homogeneous distribution, the shape of the volume in the system $(1,2)$ does not play a role because the underlying distribution is gradient free. However, in an inhomogeneous distribution the change in density becomes important because the product volume from system (1) and (2) adapts differently to the local density than the volume in $(1,2)$.

To circumvent these issues, all terms are estimated with the dimensionality given by the term with the highest considered correlation in the system. Terms with the highest considered correlation do not need special treatment, but lower dimensional terms do. Thus, in the following I explain how fill modes [33] allow an estimate of lower dimensional terms within a higher dimension. To that aim, fill modes introduce additional artificial terms, which are combined with lower dimensional terms such that the resulting term matches the desired dimensionality.

Two cases are considered. First, lower dimensional terms are combined such that their dimensionality matches the given dimensionality and no artificial terms are required. Since the combination has to reflect the respective summation in the MIE, any mutual information between the terms has to be removed before the entropy is estimated on their combined data in the desired dimensionality. By permuting the data sets of each respective term, the mutual information between their sets is destroyed [39] but their marginal

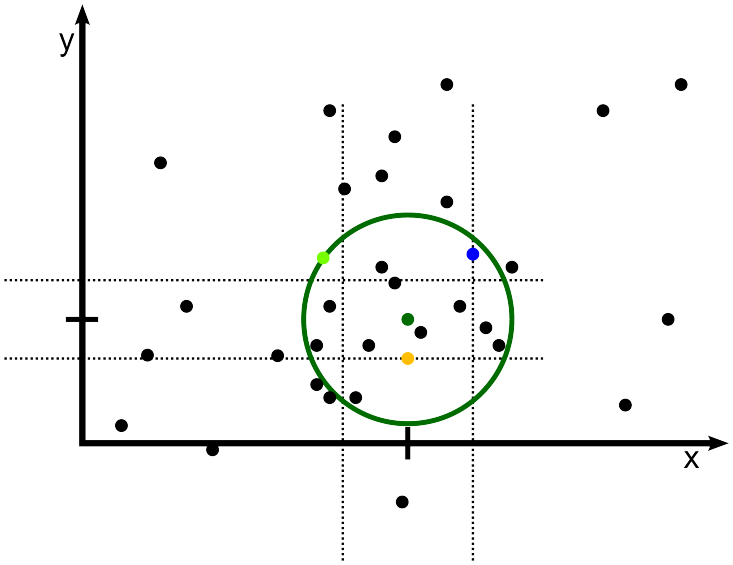


Figure 5.6 The difference in the shape of the estimator volume of one-dimensional and two-dimensional k-NN estimators. The mixing of two-dimensional density estimation and one-dimensional density estimation introduces a bias. The k-nearest neighbor for system x in blue, system y in yellow and system $x \otimes y$ in green in the respective subsystems are shown. The bias is due to the shape difference between the volume given by the product space (dotted rectangle) and the tensor product space (green circle).

entropies are kept,

$$\begin{aligned} S(1,2) &\leq S(1) + S(2) \\ \text{but } S(\hat{1},\hat{2}) &= S(\hat{1}) + S(\hat{2})\,, \end{aligned}$$

where the hat marks permuted sets.

Second, when combinations of lower dimensional terms do not lead to the correct dimensionality, fill modes are added. These are artificial terms of subsystems with removed mutual information to other terms in the expansion. They extend the MIE such that all terms can be estimated in the desired dimensionality as in the first case.

To illustrate the concept of fill modes in an example, let me consider a mutual information term of a system with three, dimensional identical

subsystems given by (5.6):

$$I(1,2,3) = \underbrace{S(1)+S(2)+S(3)}_{S(\hat{1},\hat{2},\hat{3})} - S(1,2) - S(1,3) - S(2,3) + S(1,2,3)$$

$$= S(\hat{1},\hat{2},\hat{3}) \underbrace{-S(1,2) - S(1,3) - S(2,3) + \overbrace{S(\hat{1},\hat{2},\hat{3}) - S(\hat{1},\hat{2},\hat{3})}^{0}}_{S(\hat{1},\hat{2},\hat{3}) - S(1,2,\hat{3}) - S(1,\hat{2},3) - S(\hat{1},2,3)} + S(1,2,3)$$

$$= 2S(\hat{1},\hat{2},\hat{3}) - S(1,2,\hat{3}) - S(1,\hat{2},3) - S(\hat{1},2,3) + S(1,2,3)\,. \quad (5.8)$$

The entropies of all three subsystems $S(1)+S(2)+S(3)$ are estimated on their common data with removed mutual information, $S(\hat{1},\hat{2},\hat{3})$, which represents the first case of fill modes. For the combined subsystems $S(1,2), S(1,3)$ and $S(2,3)$, fill modes are added in the form of the entropy of the marginal systems $S(\hat{1},\hat{2},\hat{3})$ to rise their dimensionality, which is the second case of fill modes.

The example demonstrates that fill modes increase the dimensionality of all terms to the desired dimensionality but they also decrease the number of terms in the mutual information expansion as seen at the final expression that consists of fewer terms than before. The number of fill modes, and their multiplicity and permutations needed to rise all entropy terms to the desired dimension depends on the truncation order of the MIE.

Therefore, I emphasize that the expansion order m_t not only defines the highest correlation in the system but also the maximum dimensionality all entropy estimates have to be evaluated on, which determines the statistical error of the estimate.

5.4 The Nearest Neighbor Method

The last method in the framework for entropy estimation of solvents in molecular dynamics simulations is the density estimator for the terms in the MIE. The numerical estimation of the individual terms given in equation (5.5) dimensionally extended by fill modes is done by a non-parametric k-nearest neighbor kernel density estimator (k-NN KDE). The parameter-free estimator does not make any assumptions about the functional form of the underlying distribution and is therefore well-suited for density estimation of unknown distributions.

The density estimator rests on the straightforward assumption that the local density at sample point x_i is approximately given by

$$\rho(x_i) \approx \frac{k}{NV(r_i(k))}\,,$$

as introduced in the previous section. The estimator is locally adaptive and has been successfully applied in Euclidean and non-Euclidean space up to forty dimensions [25, 33, 37, 38, 59, 76].

The estimator for the entropy S for the terms in equation (5.5) is defined in [38] as

$$\tilde{S}_{k,N} = -k_\mathrm{B} \left(\frac{1}{N} \sum_{i=1}^{N} \log \left(\frac{k}{N} \frac{1}{V_{(s)}(r_{i,k})} \right) + \underbrace{L_{k-1} - \gamma - \log(k)}_{f(k)} \right) , \tag{5.9}$$

where $V_{(s)}(r_{i,k})$ is the volume of a sphere in system s with radius $r_{i,k}$ which is the distance between point i and its kth neighbor, $f(k) = L_{k-1} - \gamma - \log(k)$ a correction for the asymptotic bias for large N consisting of a part of the harmonic series $L_k = \sum_{i=1}^{n} \frac{1}{i}$ and the Euler-Mascheroni constant $\gamma = 0.5772\ldots$.

In 2004 Kraskov et al. [47] generalized the kNN estimator in a general metric, however expressions for the volume are only known for the maximum and Euclidean metric. They did not provide formal proofs of asymptotic unbiasedness and consistency. However, after revisiting the proof for the Euclidean metric in $\mathbb{R}^n$ in [76], I presume that the given bias correction term also applies to the differential manifold $\mathrm{SO}(3)^n$. The reason is that the proof does not use any special properties of the space but the invertibility of the volume as function of its radius, which holds in $\mathrm{SO}(3)^n$, too.

In section 4.2, I have derived the metric tensor and the volume of a sphere as a function of the geodesic distance between two elements of the group $\mathrm{SO}(3)$ and its tensor product space $\mathrm{SO}(3)^n$. In [61] the geodesic distance between two elements $x, y \in \mathrm{SO}(3)$ of the group is given by

$$d_{\mathrm{SO}(3)}(x, y) = \frac{1}{\sqrt{2}} \left\| \mathrm{Log}(\mathcal{R}(x)^T \mathcal{R}(y)) \right\|_F , \tag{5.10}$$

which is in accordance to the geodesic equations (4.11). Here, $\mathcal{R}(\cdot)$ is the matrix representation of an orientation, $\| \cdot \|_F$ denotes the Frobenius norm, which is induced by the Euclidean inner product of the embedding space $(\mathbb{R}^{3\times3})$, $\mathrm{Log}(\cdot)$ stands for the principal logarithm of a matrix. Together with the volume calculations in section 4.2, I have a density estimator on $\mathrm{SO}(3)$ and its tensor product spaces at hand.

The remaining difficulty is to find the k-nearest neighbors of a point on $\mathrm{SO}(3)^n$ as efficient as possible because it is a $O(N)$ problem. The solution is to use the Euclidean norm of the embedding space and the ANN algorithm [2], which reduces the computational effort to $O(\log(N)^3)$, because the Euclidean

metric describes the chordal length between two elements of SO(3) and is always smaller than their geodesic distance

$$d_{E(\mathrm{SO}(3))}(x, y) = \|\mathcal{R}(x) - \mathcal{R}(y)\|_F \leq d_{\mathrm{SO}(3)}(x, y)\,. \tag{5.11}$$

The obtained order with respect to their neighbors is metric-independent as long as both metrics are smooth. Thus, the expensive calculation of geodesic distances is done for each kth neighbor only once.

6

Results and Discussion

In the previous chapter, a framework for entropy estimation on the space of orientations as well as an entropy estimator have been developed. This final chapter serves to test the accuracy of the method for realistic macromolecular systems. To this end, I proceed in three steps. First, to test the accuracy and variance of the entropy estimator, several simple synthetic test distributions will be studied, for which the entropy can be calculated analytically. Second, to test how the entropy estimator performs on correlated systems, a box of 316 water molecules at room temperature will be analyzed including higher order terms of the mutual information expansion. Third, I will apply the framework to the solvated hydrophobic protein Crambin to study how the local entropy of the first water layer depends on the hydrophobicity of the surface amino acids.

One further aim is to test the main assumption that the water orientations are sufficiently weakly correlated, on which the mutual information expansion critically rests on.

6.1 Testing Simple Distributions

First, to test the accuracy and variance of the entropy estimator on SO(3), 50,000 random samples were drawn from the uniform distribution using a Euler angle parametrization. To test a discontinuous distribution, the domain of one Euler angle was restrained, e.g. $\phi \in [0, 2\pi)$ was restrained to $\phi \in [0, \pi)$.

Since for water molecules close to a hydrophobic surface a weak orientational preference is expected, 50,000 random samples were drawn from a "normal distribution" of orientations, the matrix Fisher–von Mises distribution [50, 53, 70], with a mean direction $\mathcal{M}$ and a concentration parameter $\kappa \geq 0$. The distribution is rotationally symmetric and becomes successively more concentrated as κ approaches infinity. The distribution is given by

$$p(\mathcal{X}|\kappa, \mathcal{M}) = \frac{\sqrt{\pi}\Gamma(\kappa + 2)}{2^{2\kappa}\Gamma(\kappa + 1/2)} \left(1 + \operatorname{tr}\left(\mathcal{X}\mathcal{M}^T\right)\right)^{\kappa} . \tag{6.1}$$

All entropy estimates presented in table 6.1 were repeated 100 times and three standard deviations is reported as their error. The neighbor parameter k was set to 40 for all estimates.

The table 6.1 shows the results of the different estimates in comparison to the exact values of the distributions. In all cases the estimator slightly underestimates the exact value, but the overall obtained accuracy is very good.

Uniform Distribution	Estimate in $k_\mathrm{B}\mathrm{T}$
exact value	$\log(8\pi^2) \approx 4.37$
estimate	4.34 ± 0.01
Restricted Distributions	
exact value	$\log(4\pi^2) \approx 3.68$
$\theta \in [0, \pi/2)$	3.70 ± 0.01
$\phi \in [0, \pi)$	3.69 ± 0.01
$\psi \in [0, \pi)$	3.69 ± 0.01
Fisher–von Mises Distribution with $\kappa = 3, \mathcal{M} = \mathcal{I}$	
exact value	$\log(40\pi^2) - {}^{59}/_{20} \approx 3.03$
estimate	3.02 ± 0.01

Table 6.1 Testing simple distributions. 50,000 random samples from several synthetic probability distributions were drawn and their entropy estimated. The exact value of the distribution is given for comparison. If not stated otherwise, the domains of the Euler angles were $\psi, \phi \in [0, 2\pi)$ and $\theta \in [0, \pi)$. The matrices for the Fisher–von Mises distribution were generated by using equation(6.1).

The real interesting tests are distributions on the tensor product space $\mathrm{SO}(3)^n$, but since correlated distributions for rotation matrices on the tensor product spaces are not given in literature, the exact reference values are not known. However, an estimate of combinations of uncorrelated distributions still characterizes the accuracy of the density estimator. According to the same protocol mentioned for the simple distributions, from the combined distribution, Fisher–von Mises and uniform distribution for $\mathrm{SO}(3)^2$ and from the combined distribution, Fisher–von Mises, uniform and ϕ-restricted distribution for $\mathrm{SO}(3)^3$ samples were drawn. I found that the estimator deviates from the exact value by about $-0.1\ k_\mathrm{B}\mathrm{T}$ on $\mathrm{SO}(3)^2$ and by $-0.3\ k_\mathrm{B}\mathrm{T}$ on $\mathrm{SO}(3)^3$. Such an increase of the deviation is expected because the dimensionality of the space has risen from three to six and nine, respectively.

In summary, the estimator tends to slightly underestimate the entropy. However, in regard to the mutual information terms of the expansion and the fill modes, the deviation might cancel if all terms are estimated within the same dimensionality.

6.2 Bulk Water in MD

To test how the entropy estimator performs on correlated systems, a box of 316 water molecules has been simulated with the protocol given in 3.3.1. The trajectory of the water molecules was permuted using the first frame as the reference configuration as described in section 5.2.

6.2.1 Orientational Entropy of Individual Water Molecules

First, the orientational entropy of individual water molecules was investigated, which is the first truncation order $m_t = 1$ in the mutual information expansion as described in section 5.3.

The autocorrelation time for orientational tumbling is about 1 ps (data not shown). Consequently, all orientations of the water molecule were extracted yielding 50,000 nearly independent samples per molecule in total.

The orientational entropy of each water molecule was estimated and the distribution of the estimates plotted in figure 6.1. As a reference, the exact value together with the same amount of samples drawn from a uniform distribution was included.

The mean of the uniform distribution differs about 0.6% from the exact value as reported in the previous section. The entropy distribution of bulk water molecules is clearly close to the uniform distribution. Thus, single water molecules in bulk water sample the entire orientational space as expected.

6.2.2 Orientational Correlations

With the analysis of the mutual orientation of water molecules, I have addressed the second order in the mutual information expansion and the spatial range of "orientational influence" between bulk water molecules.

Since water molecules fluctuate around a reference position after the permutation reduction (5.2), I have calculated their mean positions in the permuted trajectory. Subsequently, the orientational entropy was estimated between all pairs, which were closer than 1 nm; pairs further apart were neglected. Their mutual information was calculated using fill modes as described in section 5.3.1 and linked with their spatial distance.

Figure 6.2 shows the mutual information as a function of distance, the mean mutual information and its 2σ-confidence bands. The confidence bands and the mean were calculated with boot strapping, in which 300 random samples were drawn from the original data, and subsequently averaged with

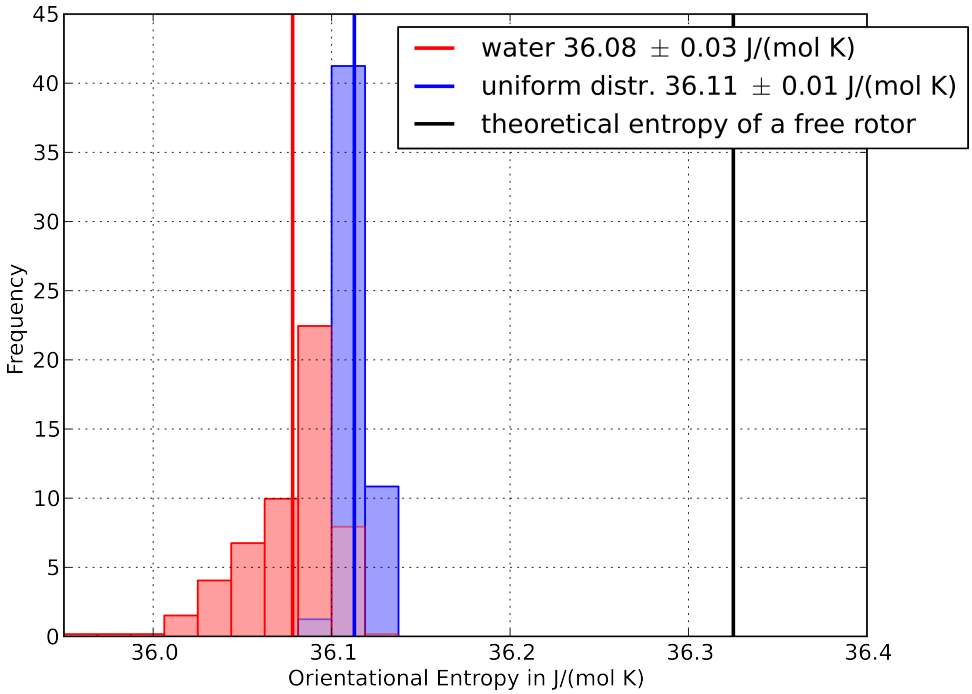


Figure 6.1 Distribution of single molecule entropies in comparison to a uniform distribution and theoretical expectation value. The histogram of water in red is close to the uniform distribution in blue. As a reference, the exact value is included (black line). Single water molecules rotate freely in bulk water as expected.

a running average ($N = 500$). Additionally, the mutual information without using fill modes is included.

The main result is that at 300 K the "orientational influence" between water molecules is quite short-ranged. It drops to zero within the third water shell and therefore, molecules further apart can be treated as independent in orientational phase space. Moreover, figure 6.2 indicates that water molecules are ordered up to 0.5 – 0.6 nm around the given water molecule and beyond this limit the ordering diminishes. Additionally, without fill modes the mutual information between molecules decreases below zero, although it should be either positive or zero, which demonstrates the bias of shape and length scale of the estimator if it is not corrected for different dimensions.

The conclusion is that the interactions between water molecules clearly induce orientational correlations within the first three water shells diminishing further apart, and that the second order mutual information is not negligible in the mutual information expansion. This finding is in agreement with reported orientational correlation functions, which have been calculated

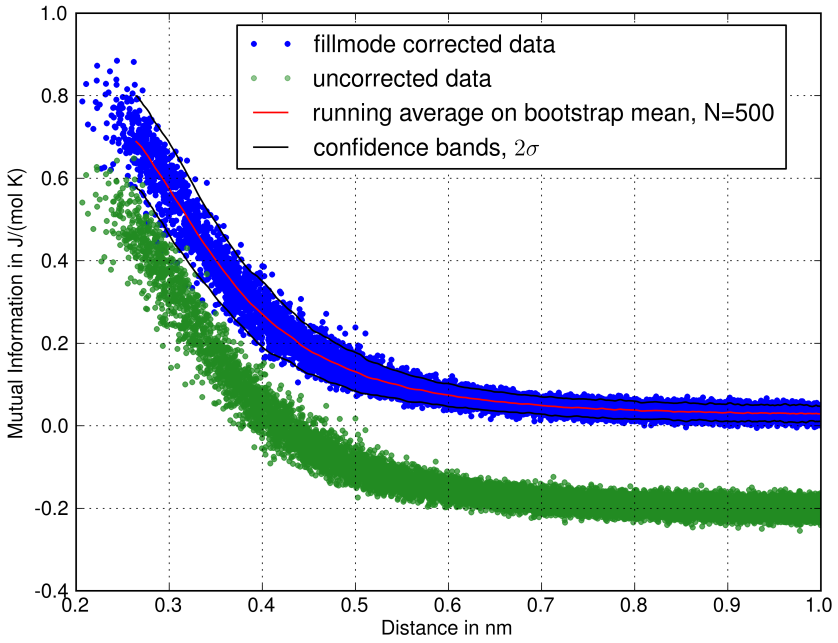


Figure 6.2 Mutual information of orientations as a function of distance. The pair mutual information of orientations (blue dots) decays to zero within 0.5 – 0.6 nm. Thus, water molecules influence their neighbors within the first three water shells. To show the effect of fill modes, the uncorrected mutual information (light green dots) is included and a bias is observed.

for entropy estimates with the Ornstein-Zernike equation [45, 84].

The contribution to the orientational entropy arising from the mutual information of pairs is about $\sum_{i<j} I(i,j) = 5.5$ J/(mol K), which is not negligible. Thus, the total orientational entropy of water at 300 K considering pair correlations is reduced to 31 J/(mol K).

6.2.3 Higher Order Correlations

To see how much triple and higher order correlations contribute to the orientational entropy of water in comparison to pair correlations, and if the orientational entropy of bulk water is indeed a weakly correlated system, the next order in the mutual information expansion $m_t = 3$ was estimated.

Since the pair correlations show, that the "orientational influence" diminishes outside the third water shell, and the number of possible combinations of subsystems increases with each truncation order with $\binom{N}{m_t}$, I limited the

analysis to molecules, which were closer than 0.5 nm. Each three molecules within a mutual information estimate form a triangle with their spatial positions as vertexes. Using the area and the smallest acute angle of this triangle as coordinates, I present the relative arrangement of the molecules in space as a function of mutual information in figure 6.3. The selection criterion

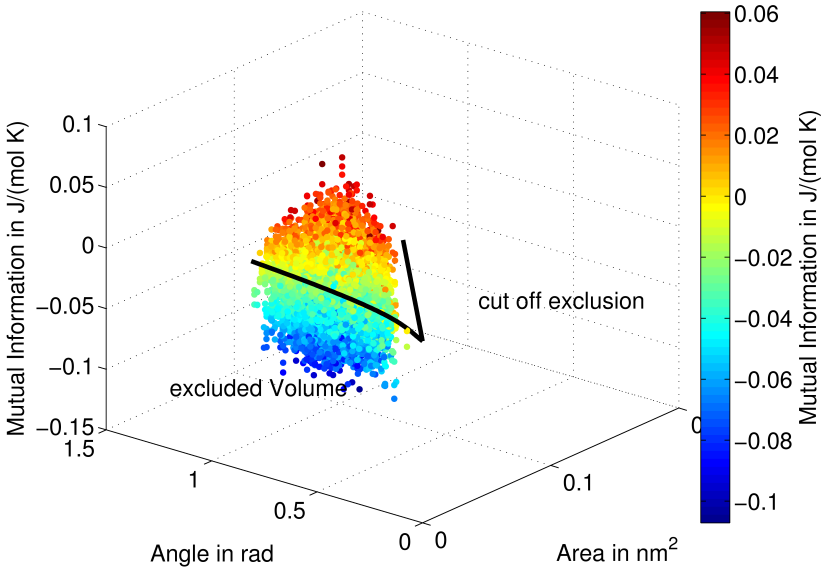


Figure 6.3 Mutual information between three water molecules. The three water molecules used in an mutual information estimate, form a triangle. The area and smallest acute angle of this triangle are describe relative arrangement of the molecules in space. The regions marked with black lines are excluded for mutual information estimates, because either the molecules' own volumes or the selection criterion prevents these angle-area combinations. The plot indicates that the mutual information is independent of the arrangement of the water molecules and fluctuates around −0.02 J/(mol K).

that molecules have to be within 0.5 nm, and the molecules' own volumes, exclude combinations in the figure with small angle and large area and with large angle and small area, respectively. These regions are marked with black lines and labeled "cut off exclusion" and "excluded volume".

The figure indicates that the mutual information of three molecules is close to zero $\left(\overline{I}_3 = -0.02 \text{ J/(mol K)}\right)$ and fluctuates $(\sigma_{I_3} = 0.02 \text{ J/(mol K)})$ independently of the relative arrangement of the molecules. This is supported by the histogram of their mutual information in figure 6.4.

Whether this mutual information estimate is a result of noise in the data or the effect of "orientational influence" is not yet determined, because there

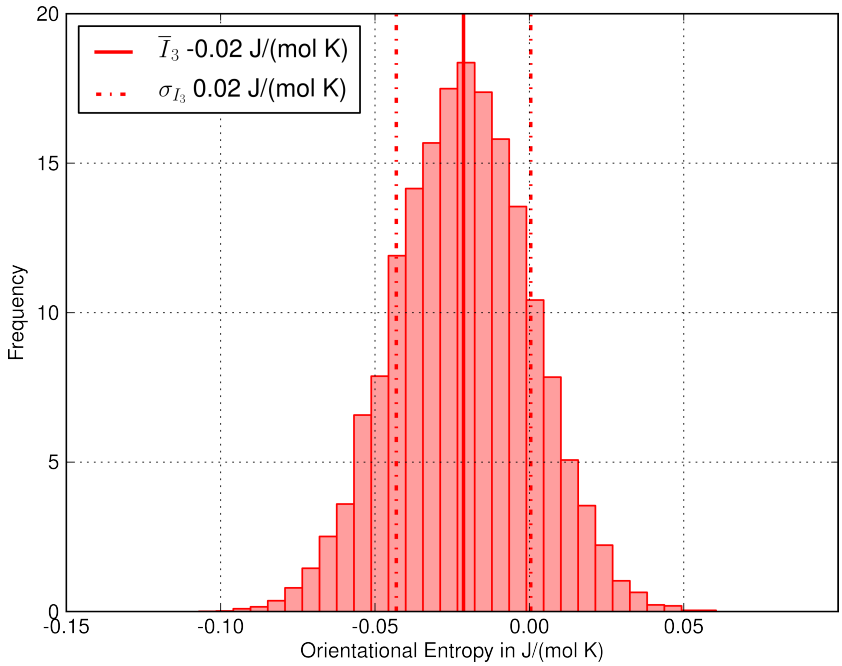


Figure 6.4 Histogram of the mutual information between three water molecules. The distribution is normal, centered at $\left(\overline{I}_3 = -0.02 \text{ J/(mol K)}\right)$ with standard deviation $(\sigma_{I_3} = 0.02 \text{ J/(mol K)})$.

are several reasons, why conclusions should be drawn carefully. First, the magnitude of the estimates is within the observed deviation of the combined simple distributions, which might not cancel completely in the summation of the mutual information. Second, the density of the points is reduced because the dimension has increased to nine, which increases the noise in the estimate. Third, the fact that the mutual information is independent of the arrangement between the molecules, hints to noise, but for "orientational influence", because one would rather expect that the mutual information decreases with increasing area, as it does for increasing distance in pair correlations.

If the mutual information of the next truncation order $m_t = 4$, which has not been estimated, also decreases and higher truncation orders can be neglected, is therefore difficult to conclude. But the fact that the influence is dropped by one order of magnitude between pair and triple correlation, the chance might be high that it is indeed small.

In summary, whether orientational entropy of bulk water represents a weakly correlated system is still an open question, the estimates hint that the system might be sufficiently weak correlated such that at some higher truncation order $m_t > 3$ further mutual information terms can be neglected.

6.3 Water at a Protein Surface

To study the effect of the hydrophobicity of surface amino acids on water molecules, the protein Crambin was chosen as an example, because although the surface of the crystal structure consists about 60% hydrophobic and 40% hydrophilic residues, which is similar to other proteins, it shows a very low water solubility. Since the iceberg model [23] predicts an entropy loss in vicinity of a hydrophobic molecule, it is an excellent test system for our entropy framework.

The protein Crambin (PDB-code: 1CBN) [79] was obtained from the protein data base [12]. It is a very small protein consisting of 46 residues and belongs to the family of membrane-active plant toxins. Its was found in the seeds of *Crambe abyssinica* and its biological function is unknown. The crystal structure of the protein is well resolved to 0.83 Å resolution. Its secondary structure is nearly 45% α-helix together with one β-sheet consisting of three β-strands. According to section 3.3.2, the protein was prepared and simulated.

For the permutation reduction technique (5.2), it is essential that the phase space blocked by the protein remains inaccessible for water molecules, because otherwise the permutation translates that phase space and an estimate with relation to the surface is no longer possible. To guarantee that the protein did not move or rotate during the simulation, three position restrains were applied to the proline residues of the protein. These residues are positioned in the loops connecting the different secondary structure elements of the protein and they are far enough apart that they stabilize the protein properly.

To demonstrate the stability of the protein during the simulation, the root mean square deviation of the protein was calculated shown in figure 6.5.

6.3.1 Orientational Entropy of Water Molecules at the Surface of Crambin

From the permuted trajectory, 50,000 nearly independent orientations for each water molecule were extracted and the mean positions of all water molecules calculated. To investigate and quantify the influence of the protein on the water molecules within the first water shell, all molecules within 0.3 nm of the protein surface were selected resulting in 190 water molecules. The orientational entropy of these surfaces molecules was estimated.

Figure 6.6 shows the distribution of orientational entropies of the surface

Figure 6.5 Root mean square deviation of Crambin. Molecule is stable over the simulation time and did not change more than 0.2 nm from the starting structure.

molecules. As seen, when compared to figure 6.1, the average orientational entropy is reduced by 2 J/(mol K). Most of the molecules still show their orientational freedom, but a small fraction loses in orientational entropy, up to 10 J/(mol K).

Further, to compare the entropy loss with the hydrophobicity of the surface amino acids, the electric potential at the solvent accessible surface was calculated with the Poisson-Boltzmann equation as described in section 3.3.2. The surface potential ranges from -27 mV in red over white to $+27$ mV in blue ($[-1, 1]$ in $k_{\mathrm{B}}\mathrm{T}/\mathrm{e}$, in the figure). The entropy loss of the water molecules was color-coded and mapped on the average positions of the water molecules. To visualize the spread in the entropy loss, the same color bar was used and the molecules with maximal entropy loss mapped on red (21.89 J/(mol K)) and those with no entropy loss on blue (36.03 J/(mol K)), respectively.

To identify the shown hydrophobic surface within the protein structure, the protein is displayed three times in figure 6.7, first, in the upper left a cartoon presentation of secondary structure elements for orientation, second in the upper right, the electric potential on the solvent accessible surface and third with overlaid entropy loss of the water molecules.

As seen, molecules close to the surface lose orientational entropy depend-

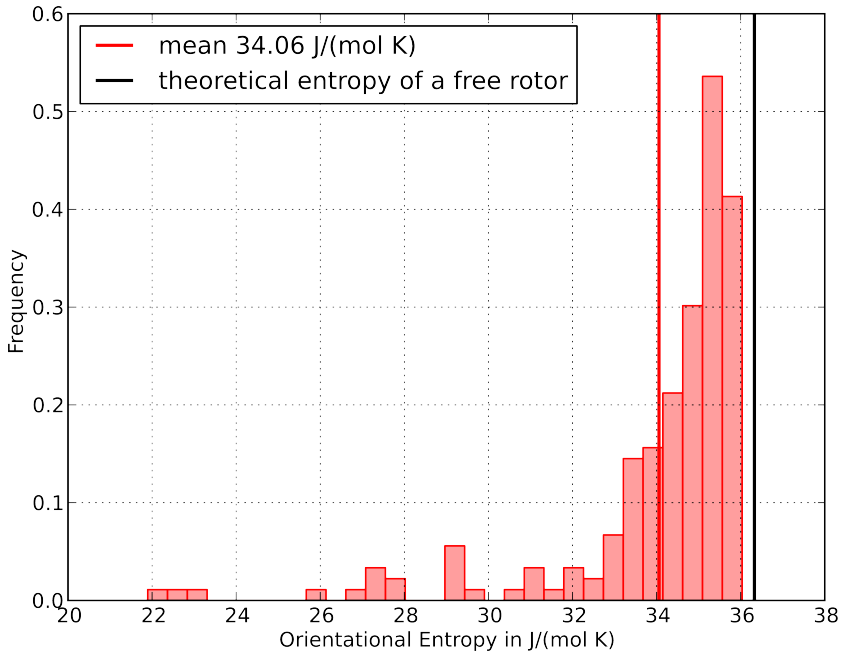


Figure 6.6 Distribution of orientational entropies of water molecules located within 0.3 nm of the protein surface. Besides freely rotating molecules, a significant number of water molecules show an entropy loss up to 10 J/(mol K).

ing on their position and the surface potential. Regions with a electric potential close to 0 mV (in white), correlate with the entropy loss of nearby water molecules (in light blue to red), although this needs to be further investigated and quantified. The highest entropy loss is observed between both helices, which is due to both steric restriction of the surrounding amino acids and low surface potential. In regions with high electric potential (red or blue), all water molecules rotate freely.

In summary, the hydrophobic effect was observed as a an entropy loss in the orientational entropy of water molecules at the protein surface, but further investigations are necessary to quantify the spatial correlation with the surface potential.

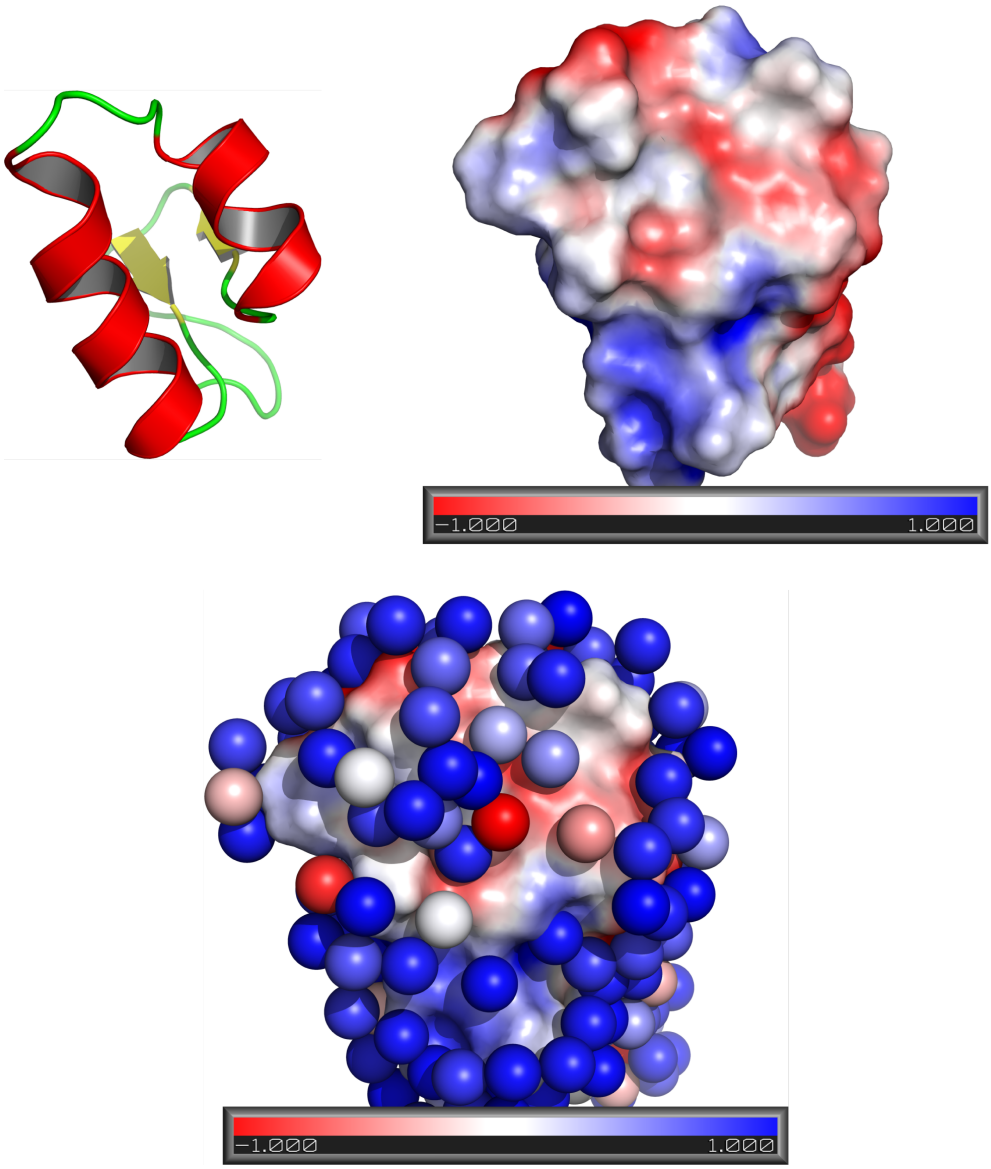


Figure 6.7 Electric potential at the solvent accessible surface of Crambin. For orientation, the protein is displayed three times. In the upper left, a cartoon presentation of secondary structure elements defines the point of view. In the upper right, the electric potential on the solvent accessible surface is presented ($[-1, 1]$ in $k_{\mathrm{B}}\mathrm{T}/\mathrm{e}$) and last overlaid with the entropy loss of the water molecules, which is mapped on the same color bar with red (21.89 J/(mol K)) to blue (36.03 J/(mol K)). The entropy loss (light blue to red) of the water molecules tends to correlate with regions of low surface potential (in white).

7

Summary, Conclusion and Outlook

Entropy effects of the surrounding water layer at the protein surface have been studied for a long time, and their relevance e.g. for protein folding is well recognized.

In molecular dynamics simulations, for proteins, progress has been made to quantify entropies because they reside in a free energy minimum, which is fairly well sampled within simulation times. However, solvents reside in a much larger and shallow free energy minimum rather than a localized minimum and therefore all statistical averages converge poorly. Additionally, the dimensionality of the phase space becomes very large even for small systems.

In molecular dynamics simulations entropy estimates for explicit water molecules surrounding a protein are difficult to calculate with established methods such as thermodynamic integration or radial distribution functions, because either they fail to converge or refer to a homogeneous systems.

Hence, the central aim of this thesis was to develop a general framework for the orientational entropy of water molecules as part of the total entropy of the system, in particular focusing on its application to protein water systems in molecular dynamics simulations. The developed framework will allow insights into protein-water interactions in further studies as e.g. anti-freeze proteins. Although the framework is not complete because the translational part of the solvent entropy is still ongoing research, it is already applicable.

7.1 The Orientational Entropy Framework

Several methods have been combined into an entropy estimator for orientational entropies in molecular dynamics simulations.

In a first step, the entropy integral of orientations of all water molecules has been decomposed by the mutual information expansion into entropies of its subsystems, the orientational space of each molecule. Under the assumption that the orientational entropy is a weakly correlated system, the expansion provides an approximation to the entropy integral of all orientations. Thus, the high dimensional, numerically intractable entropy integral is now approximated by a series of lower dimensional, numerically tractable integrals and the quality of the approximation can be systematically improved by adding more mutual information terms with increasing dimensionality.

In a second step, the sampling problem of solvents in molecular dynamics simulations was addressed. Solvent molecules sample the simulation box by diffusion, so all statistical averages converges slowly due to the resulting bad sampling. The permutation reduction technique solves this problem. It exploits the fact that molecules are indistinguishable, thus physics is left invariant under permutations of the molecules' indices. A specific choice of

permutation translates the phase space of each molecule, such it is more dense than before and in the resulting trajectory the molecules fluctuates around a reference position. This enhances the sampling by the Gibbs factor $N!$.

The last step addressed the entropy estimation of the terms provided from the mutual information expansion. A non-parametric k-NN neighbor algorithm was tailored for the curved manifold of orientations, which locally adapts to the underlying distribution. The key components were the development of the distance and volume functions on the curved manifold of orientations as a function of the geodesic distance on SO(3) to $\mathrm{SO}(3)^3$. Additionally, the method can be easily extended to $\mathrm{SO}(3)^n$.

The combination of all three methods provides a framework for entropy estimation on the high dimensional phase space of orientations of solvents in molecular dynamics simulations.

7.2 Testing the Entropy Estimator

The introduced estimator on the rotation group SO(3) and its tensor product spaces has been thoroughly tested on simple, synthetic distributions, bulk water and water at a protein surface.

Testing the accuracy and variance on the simple, synthetic distributions showed that the estimator slightly underestimates the exact values but obtains an overall good accuracy.

In bulk, water several truncation orders of the mutual information expansion have been analyzed. The first truncation order shows that bulk water molecules rotate freely as expected. Though, in the second truncation order their "orientational influence" is observed up to the third water shell and diminish if the molecules are further apart. This finding is in accordance to those obtained from other techniques in literature. The third truncation order seemed to be unaffected by the relative arrangement of the water molecules and fluctuated around about -0.02 J/(mol K). Several reasons hint that this estimate needs further investigation and the current estimate is too much affected by noise for proper conclusions. Whether the orientational entropy is a weakly correlated system remains as an open question, but with the developed framework higher truncation orders are open for investigation.

The framework is not limited to bulk water systems but can be applied to molecular dynamics simulations containing proteins as well. This has been shown in a proof of concept on Crambin, a hydrophobic protein. Herein, the orientational entropy loss of water molecules at the protein surface has

been investigated. The superposition of the electric potential at the solvent accessible surface with the entropy loss of the water molecules at the surface shows, how the protein affects its surrounding water molecules.

7.3 Outlook to both, Entropy of Protein and Water

I noted that the above framework can be extended to also estimate the mutual information entropy term $I(\mathrm{P},\mathrm{W})$ in (5.1) between water and the protein.

To that aim, I would combine this framework with a method, which addresses the protein entropy and decomposes the configurational phase space of the protein into subspaces, keeping the mutual information expansion in mind.

Protein entropy can be estimated either with a principal component analysis (PCA) [1], or more appropriately with a full correlation analysis (FCA) [49] or with a minimally coupled subspace approach (MCSA) [33]. All these methods provide subspaces as linear combination of protein coordinates, which minimize either linear correlation between pairs of subspaces (PCA) or their mutual information (FCA, MCSA).

The configurational subspaces of the protein estimator provide a decomposition in subsystems, which depending on the used method have already low mutual information. The solvent phase space is tackled as in orientational framework by applying the permutation reduction technique.

Further, the mutual information in (5.6) is composed of entropies of its subsystems, which can be approximated with the truncated MIE as well. Hence recursively, the entropy estimate of the total system can be systematically improved.

Thus, the high dimensionality of the mutual information integral is reduced as in the orientational entropy framework. Also, most of the low order mutual information terms in the MIE have already been computed either in the water entropy or protein entropy (in FCA and MCSA but not in PCA). Therefore, if one focuses on those combinations of subsystems, which consist of protein subspaces and water molecules on the surface of their presenting protein modes, one might capture all relevant mutual information terms and can neglect the rest.

The permutation reduction technique enhances the phase space sampling of water molecules and allows to refer to their position in space. Therefore, with the assumption that the mutual information between the protein and water molecules decreases as a function of distance to the protein surface, one

can lower the computational costs by neglecting bulk water far away from the protein surface.

Moreover, the protein motions in subspaces of great variance (or great anharmonicity), which represent global protein motions involving all protein atoms, are expected to contribute to the mutual information between the subspace and all water molecules on the surface. Motions in subspaces of small variance usually represent local fluctuations of side chains and therefore contribute together with those water molecules closest to their position of highest fluctuation. Therefore, a clever selection of water molecules depending on distance and subspace of the protein motion helps to reduce computations as well.

For the actual density estimation, I would define the tensor product space of protein subspaces $\mathbb{R}^k$ and water molecules

$$\begin{aligned} T &= \left\{\mathrm{SO}(3) \times \mathbb{R}^3\right\}^m , \\ \mathrm{MW} &= \left\{\mathbb{R}^k \times T^m\right\} . \end{aligned}$$

I propose to advance similar to the orientational entropy estimator and use isotropic spherical kernels defined on that space with the metric

$$d_{\mathrm{MW}}(A, B) = \sqrt{d_{\mathrm{E}}(A_{\mathrm{E}}, B_{\mathrm{E}}) + d_{\mathrm{W}}(A_{\mathrm{W}}, B_{\mathrm{W}})} ,$$

with $A, B \in \mathrm{MW}$ and $d_{\mathrm{E}}(\cdot, \cdot)$ the metric of the subspaces and $d_{\mathrm{W}}(\cdot, \cdot)$ of the water molecules, respectively. The appropriate function of volume $V_{\mathrm{MW}}(r)$ has to be calculated.

The strength of correlation between the protein and its surrounding water molecules, which has to be analyzed for each protein individually, determines if such an approach will be successful.

8

Appendix

In the appendix, I present a grid-based approach to a density estimator on SO(3). During its development it turned out to be not as efficient as the k-NN algorithm and was not used in the results section nor further developed for tensor product spaces beyond $n = 2$. Nevertheless, it is worth presenting and therefore, I shortly introduce the parametrization of the group with quaternions followed by the description of a singularity-free grid on SO(3).

8.1 Unit Quaternions

Quaternions [28, 29] have been developed by Hamilton in 1843 as a description of rotations of rigid bodies. They are defined by elements on the unit sphere in four-dimensional space, $\mathbb{S}^3$ with an additional algebra, the quaternion multiplication. I have used quaternions for the development of a grid on SO(3), because they simplify the transfer of a grid from a sphere in three dimensions, $\mathbb{S}^2$ to $\mathbb{S}^3$.

Since quaternions are also discussed in detail in the literature, I refer the reader to standard textbooks e.g. [36, 62], especially the representation as the even part of the Clifford Algebra $\mathcal{C}l^+_{(3,0)}(\mathbb{R})$. Here, I will limit my review to some important properties, which have been used in the development of the grid.

- The set of unit quaternions form the unit sphere in four dimensions, $\mathbb{S}^3$.
- Quaternion unit sphere is a double cover of the group SO(3), $2 \wedge \mathrm{SO}(3) \cong \mathbb{S}^3$.
- Quaternions form a differentiable chart on SO(3).
- The Quaternion cover has no singularities.
- Quaternion multiplication is numerically more stable then matrix multiplication and needs less operations.

These properties are useful in several aspects. For a construction of a grid on the four dimensional unit sphere, I can rely on a visualization in three dimensions and transfer the result to the higher dimension. Since the smooth differentiable double cover of quaternions on SO(3) has no singularities, calculations do not encounter the *gimbal lock*. Also, any grid defined on a hemisphere $\mathbb{S}^3/\{1\}$ is also a grid on SO(3).

8.2 Grid-based entropy estimation on SO(3)

In addition to the k-NN estimator, which was finally used in the evaluation of the entropy terms, I have developed a grid on SO(3) for two purposes. First, collecting sets of data points by grid coordinates reduces the computational costs of finding the next neighbors, which was also solved differently in the k-NN estimator. Second, a grid can be used as an histogram-based estimator for the given probability density of orientations. However, histogram-based estimators are usually used in one or two dimensions and avoided for higher dimensional spaces because the necessary bin sizes and number of bins to reasonably estimate densities increases dramatically. Since SO(3) is three-dimensional, the development was still useful but was not further implemented for its tensor product spaces.

To show, which problems arise for grids on curved manifolds, I briefly recap the geographic coordinate system on the sphere in three dimensions, $\mathbb{S}^2$. In the geographic coordinate system, the parametrization of the grid coordinates is by latitude and longitude (figure 8.1). A grid built up by

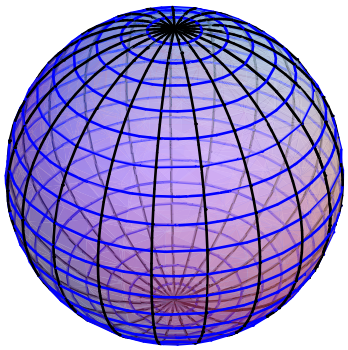


Figure 8.1 The geographic coordinate system on the sphere $\mathbb{S}^2$. Latitudinal grid lines are plotted in blue, longitudinal grid lines in black. At both poles all longitudinal grid lines join at the singularity of the chart. Note also that the grid areas close to the poles are much smaller than around the equator.

these coordinates, suffers from the following problems. First, at every grid point four grid edges join, but at two singular grid points, the poles, all

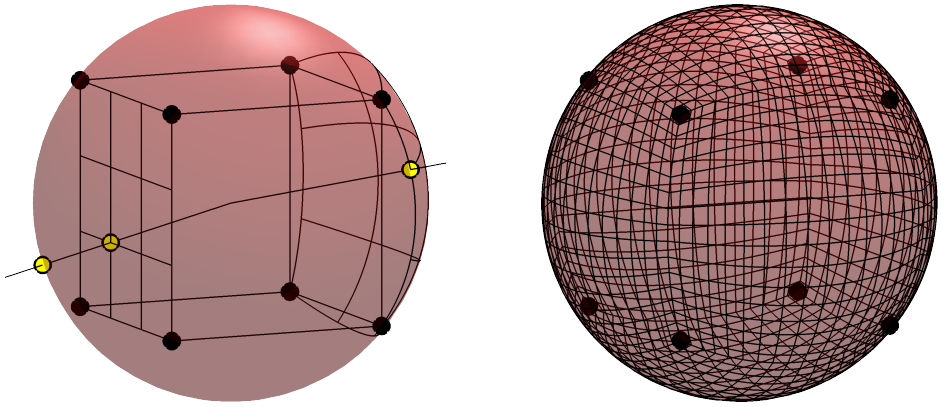


Figure 8.2 Generation of a singularity free grid on $\mathbb{S}^2$. On the left, the grid on the left side of the circumscribed cube is projected on to the surface of the sphere. The result of that projection is shown with the grid on the right side of the cube. On the right, a refined grid is projected onto the sphere. Although the grid areas have different sizes, they all shrink uniformly and the number of grid edges joining at a grid point stays constant.

longitudinal grid lines join. This difference in the number of joining grid lines let the size of a grid area at the pole shrink much faster than at the equator if the grid is refined. The reason for that behavior is the singularity in the chart at the pole. Second, the size of a grid area at the pole is much smaller than of a grid area at the equator. Ideally all grid areas would have the same size, shape and shrink uniformly under refining. However, such a grid does not exist for spheres but a grid called "the cubed sphere" meets most of these requirements.

Since charts on SO(3) with singularities have the same problem and two charts and a transition map doubles the work, the best choice for a chart on SO(3) is the double cover with quaternions. They represent the unit sphere $\mathbb{S}^3$, which eases the transfer of the following grid from $\mathbb{S}^2$ to $\mathbb{S}^3$ and then onto SO(3).

The grid on $\mathbb{S}^2$, "the cubed sphere" was developed by Ronchi et al. in 1996 [74] for solving differential equations on a unit sphere. The grid decomposes the sphere into six identical regions, which are obtained by projecting the sides of the circumscribed cubed onto the spherical surface. Each side of the cube is the basis of one coordinate system. By choosing the coordinate lines on the regions as arcs of great circles one obtains 6 coordinate systems on the sphere, which are free of any singularity. Figure 8.2 illustrates the grid on $\mathbb{S}^2$.

This concept is applicable in higher dimensions as well. The hypercube

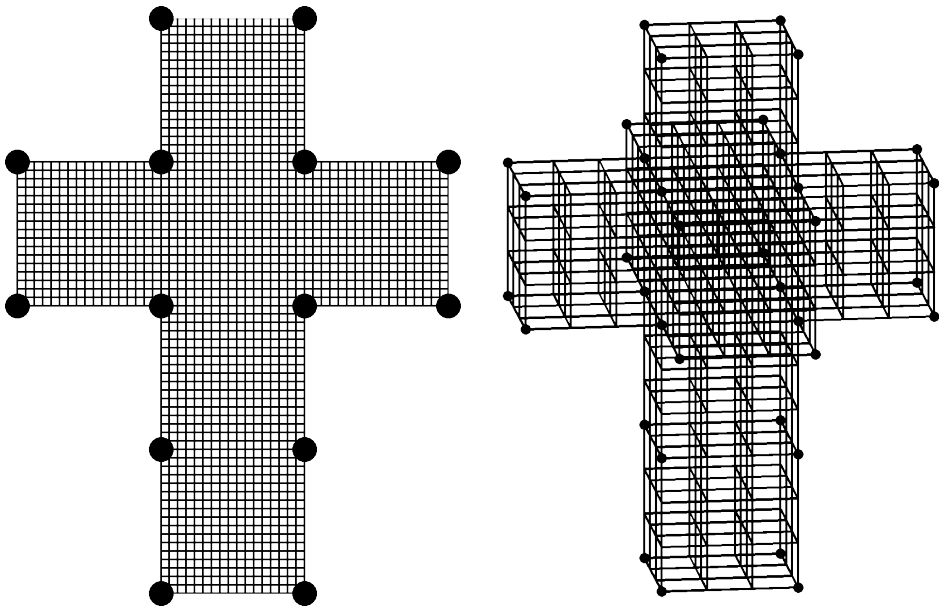


Figure 8.3 The grid on the net of the three- and four-dimensional cubes. On the left the grid of each surface is shown, which is projected onto $\mathbb{S}^2$. On the right the grid on each side of the hypercube, the tesseract, is shown, which is projected on $\mathbb{S}^3$.

in four dimensions, also called tesseract has 12 identical sides. Projecting the sides on the circumscribed hypersphere dissects the hypersphere in 12 identical regions. The projection of a grid on each side with great circles as grid lines defines the refined grid on the surface. The concept of the grid on each side is shown on the net of the cube (figure 8.3), which is a unfolded representation of the cube. Since the quaternions represent a two-one-mapping to the group, the grid on one hemisphere of the hypersphere represents a singularity free grid on SO(3).

Danksagung

Zur erfolgreichen Durchführung meiner Doktorarbeit haben viele Menschen beigetragen, bei denen ich mich an dieser Stelle bedanken möchte.

Zuerst gebührt mein besonderer Dank Prof. Dr. Helmut Grubmüller, unter dessen Leitung die vorliegende Arbeit entstand. Er hat mich bei meinen Vorschlägen und Ideen stets mit physikalische Bildern, Argumenten und wertvollen Diskussionen unterstützt und war mir während meiner Promotion eine große Hilfe.

Insbesondere bedanke ich mich bei Christian Blau und Ulrike Gerischer für das Lesen meiner Doktorarbeit. Des weiteren möchte ich mich bei Ulf Hensen, Maik Götte, Martin Höfling, Lars Schäfer, Martin Stumpe und Martin Vesper für fachliche wie auch weniger fachliche Diskussionen innerhalb als auch außerhalb des Instituts bedanken, so wie allen weiteren Mitgliedern der drei Arbeitsgruppen von Helmut Grubmüller, Bert de Groot und Gerrit Groenhof.

Ich danke Eveline Heinemann, Antje Erdmann, Ansgar Esztermann und Martin Fechner für ihr Unterstützung in organisatorischen wie auch technischen Dingen rund um die Promotion und in der Abteilung.

Besonderer Dank gebührt meinen Eltern und Brüdern für ihre Unterstützung, sowie ihre Ideen und Anregungen an unseren weihnachtlichen und österlichen Diskussionsrunden. Ganz besonders danke ich Bärbel und Fred für ihre Liebe, Unterstützung und Geduld.

Bibliography

[1] A. Amadei, A.B.M. Linssen, and H.J.C. Berendsen. Essential dynamics of proteins. *Proteins: Structure, Function, and Bioinformatics*, 17(4):412–425, 1993. 74

[2] S. Arya and D.M. Mount. Approximate nearest neighbor queries in fixed dimensions. In *Proceedings of the fourth annual ACM-SIAM Symposium on Discrete algorithms*, pages 271–280. Society for Industrial and Applied Mathematics Philadelphia, PA, USA, 1993. 57

[3] P. Attard. Spherically inhomogeneous fluids. i. percus–yevick hard spheres: Osmotic coefficients and triplet correlations. *The Journal of Chemical Physics*, 91:3072, 1989. 23

[4] P. Attard. Spherically inhomogeneous fluids. ii. hard-sphere solute in a hard-sphere solvent. *The Journal of Chemical Physics*, 91:3083, 1989. 23

[5] P. Attard, O.G. Jepps, and S. Marčelja. Information content of signals using correlation function expansions of the entropy. *Physical Review E*, 56(4):4052–4067, 1997. 16

[6] N.A. Baker, D. Sept, S. Joseph, M.J. Holst, and J.A. McCammon. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences of the United States of America*, 98(18):10037, 2001. 32

[7] R. Bellman. Dynamic programming, princeton. *NJ: Princeton UP*, 1957. 18

[8] H.J.C. Berendsen, J.R. Grigera, and T.P. Straatsma. The missing term in effective pair potentials. *Journal of Physical Chemistry*, 91(24):6269–6271, 1987. 28, 30, 31

[9] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, A. DiNola, and JR Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81:3684, 1984. 31

[10] H.J.C. Berendsen, J.P.M. Postma, W.F. Van Gunsteren, and J. Hermans. Interaction models for water in relation to protein hydration. *Intermolecular forces*, 331, 1981. 30

[11] HJC Berendsen, D. van der Spoel, and R. Van Drunen. Gromacs: a message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, 91(1-3):43–56, 1995. 31

[12] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000. 67

[13] A. Brodsky. Is there predictive value in water computer simulations? *Chemical Physics Letters*, 261(4-5):563–568, 1996. 28

[14] C.L. Brooks, M. Karplus, and B.M. Pettitt. *Proteins: a theoretical perspective of dynamics, structure and thermodynamics*. Wiley-Interscience, 1990. 26

[15] M. Calvin and P. Massini. The path of carbon in photosynthesis. *Cellular and Molecular Life Sciences*, 8(12):445–457, 1952. 10

[16] C.A. Chang, W. Chen, and M.K. Gilson. Ligand configurational entropy and protein binding. *Proceedings of the National Academy of Sciences*, 104(5):1534, 2007. 19

[17] C.E. Chang, W. Chen, and M.K. Gilson. Evaluating the accuracy of the quasiharmonic approximation. *J. Chem. Theory Comput*, 1(5):1017–1028, 2005. 19

[18] T. Darden, D. York, and L. Pedersen. Particle mesh ewald: An $n \cdot \log(n)$ method for ewald sums in large systems. *The Journal of Chemical Physics*, 98:10089, 1993. 31

[19] K.A. Dill. Dominant forces in protein folding. *Biochemistry*, 29(31):7133–7155, 1990. 14

[20] T.J. Dolinsky, P. Czodrowski, H. Li, J.E. Nielsen, J.H. Jensen, G. Klebe, and N.A. Baker. Pdb2pqr: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Research*, 35(Web Server issue):W522, 2007. 32

[21] P. Ehrenfest. Bemerkung über die angenäherte gültigkeit der klassischen mechanik innerhalb der quantenmechanik. *Zeitschrift für Physik A Hadrons and Nuclei*, 45(7):455–457, 1927. 27

[22] R. Esposito, F. Saija, A.M. Saitta, and P.V. Giaquinta. Entropy-based measure of structural order in water. *Arxiv preprint cond-mat/0603764*, 2006. 23

[23] H.S. Frank and M.W. Evans. Free volume and entropy in condensed systems iii. entropy in binary liquid mixtures; partial molal entropy in dilute solutions; structure and thermodynamics in aqueous electrolytes. *The Journal of Chemical Physics*, 13:507, 1945. 11, 12, 67

[24] H. Goldstein, C.P. Poole, and J.L. Safko. *Klassische Mechanik*. Akad. Verl.-Ges., 1972. 34

[25] MN Goria, NN Leonenko, VV Mergel, and P.L.N. Inverardi. A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *Journal of Nonparametric Statistics*, 17(3):277–297, 2005. 52, 57

[26] CG Gray and KE Gubbins. Theory of molecular liquids. *Clarendon: Oxford*, 1984. 23

[27] M. Griebel, S. Knapek, G. Zumbusch, and A. Caglar. *Numerische Simulation in der Molekü ldynamik*. Springer, 2004. 26

[28] S.W.R. Hamilton. *Lectures on quaternions*. Hodges and Smith Dublin, 1853. 78

[29] W.R. Hamilton and C.J. Joly. *Elements of quaternions*. Longmans, Green, and co., 1901. 78

[30] J.P. Hansen. *Theory of Simple Liquids: By Jean Pierre Hansen and Ian R. Mcdonald*. Academic Press, 1976. 22, 23

[31] R.H. Henchman. Free energy of liquid water from a computer simulation via cell theory. *The Journal of chemical physics*, 126:064504, 2007. 21

[32] M. Hennig. Entropy-preserving transformation method. Master's thesis, University of Göttingen, 2007. 47

[33] U. Hensen, O.F. Lange, and H. Grubmüller. Estimating absolute configurational entropies of macromolecules: The minimally coupled subspace approach. *PloS one*, 2010. 54, 57, 74

[34] B. Hess, H. Bekker, H.J.C. Berendsen, and J.G.E.M. Fraaije. Lincs: a linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18(12):1463–1472, 1997. 32

[35] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput*, 4(3):435–447, 2008. 28, 31

[36] D. Hestenes. *New foundations for classical mechanics*. Kluwer Academic Publishers, 1999. 78

[37] V. Hnizdo, E. Darian, A. Fedorowicz, E. Demchuk, S. Li, and H. Singh. Nearest-neighbor nonparametric method for estimating the configurational entropy of complex molecules. *Journal of computational chemistry*, 28(3):655, 2007. 16, 46, 51, 57

[38] V. Hnizdo, J. Tan, B.J. Killian, and M.K. Gilson. Efficient calculation of configurational entropy from molecular simulations by combining the mutual-information expansion and nearest-neighbor methods. *Journal of computational chemistry*, 29(10):1605, 2008. 16, 46, 51, 57

[39] S. Hoory, A. Magen, S. Myers, and C. Rackoff. Simple permutations mix well. *Theoretical Computer Science*, 348(2-3):251–261, 2005. 54

[40] C. Jarzynski. Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78(14):2690–2693, 1997. 19

[41] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, and M.L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79:926, 1983. 28, 32

[42] 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. 28, 32

[43] M. Karplus. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *J. comput. Chem*, 4:187–217, 1983. 28

[44] M. Karplus and J.N. Kushick. Method for estimating the configurational entropy of macromolecules. *Macromolecules*, 14(2):325–332, 1981. 18, 46

[45] M. Kinoshita, N. Matubayasi, Y. Harano, and M. Nakahara. Pair-correlation entropy of hydrophobic hydration: Decomposition into translational and orientational contributions and analysis of solute-size effects. *The Journal of chemical physics*, 124:024512, 2006. 23, 64

[46] John G. Kirkwood. Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics*, 3(5):300–313, 1935. 19

[47] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical Review E*, 69(6):66138, 2004. 52, 57

[48] A. Kryshtafovych, K. Fidelis, and J. Moult. Casp8 results in context of previous experiments. *Proteins*, 77(Suppl 9):217–228, 2009. 14

[49] OF Lange and H. Grubmüller. Full correlation analysis of conformational protein dynamics. *Proteins*, 70(4):1294, 2008. 74

[50] K. Lassen, J. Jensen, and K. Conradsen. On the statistical analysis of orientation data. *Acta Crystallographica Section A: Foundations of Crystallography*, 50(6):741–748, 1994. 60

[51] T. Lazaridis and M. Karplus. Orientational correlations and entropy in liquid water. *The Journal of Chemical Physics*, 105:4294, 1996. 23

[52] M.S. Lee, F.R. Salsbury Jr, and C.L. Brooks III. Constant-ph molecular dynamics using continuous titration coordinates. *Proteins: Structure, Function, and Bioinformatics*, 56(4):738–752, 2004. 27

[53] C.A. León, J.C. Massé, and L.P. Rivest. A statistical model for random rotations. *Journal of Multivariate Analysis*, 97(2):412–430, 2006. 60

[54] C. Levinthal. Are there pathways for protein folding. *J. Chim. Phys*, 65(1):44–45, 1968. 14

[55] R. Lumry and S. Rajender. Enthalpy-entropy compensation phenomena in water solutions of proteins and small molecules: A ubiquitous properly of water. *Peptide Science*, 9(10):1125–1227. 14

[56] M.W. Mahoney and W.L. Jorgensen. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *The Journal of Chemical Physics*, 112:8910, 2000. 31

[57] H. Matsuda. Physical nature of higher-order mutual information: Intrinsic correlations and frustration. *Physical Review E*, 62(3):3096–3102, 2000. 16, 46, 51

[58] D.A. McQuarrie. *Statistical thermodynamics*. HarperCollins Publishers, 1973. 47

[59] N. Misra, H. Singh, and V. Hnizdo. Nearest neighbor estimates of entropy for multivariate circular distributions. *Entropy*, 12(5):1125–1144, 2010. 57

[60] S. Miyamoto and P.A. Kollman. Settle: an analytical version of the shake and rattle algorithm for rigid water models. *Journal of Computational Chemistry*, 13(8):952–962, 1992. 31

[61] M. Moakher. Means and averaging in the group of rotations. *SIAM Journal on Matrix Analysis and Applications*, 24(1):1–16, 2002. 57

[62] W. Neutsch. *Koordinaten*. Spektrum, Akad. Verl., 1995. 78

[63] J. R. Oppenheimer and M. Born. Zur quantentheorie der molekeln. *Annalen der Physik*, 84:457, 1927. 26

[64] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12):7182–7190, 2009. 31

[65] C.S. Pereira, D. Kony, R. Baron, M. Müller, W.F. Van Gunsteren, and P.H. Hünenberger. Conformational and dynamical properties of disaccharides in water: a molecular dynamics study. *Biophysical journal*, 90(12):4337–4344, 2006. 19

[66] J. Pesonen. Vibrational coordinates and their gradients: A geometric algebra approach. *The Journal of Chemical Physics*, 112:3121, 2000. 43

[67] J. Pesonen. Vibration–rotation kinetic energy operators: A geometric algebra approach. *The Journal of Chemical Physics*, 114:10598, 2001. 43

[68] J. Pesonen and L. Halonen. Volume-elements of integration: A geometric algebra approach. *The Journal of Chemical Physics*, 116:1825, 2002. 43

[69] C. Peter, C. Oostenbrink, A. van Dorp, and W.F. van Gunsteren. Estimating entropies from molecular dynamics simulations. *The Journal of chemical physics*, 120:2652, 2004. 20

[70] M.J. Prentice. Orientation statistics without parametric assumptions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 214–222, 1986. 60

[71] P.L. Privalov. Cold denaturation of protein. *Critical reviews in biochemistry and molecular biology*, 25(4):281–306, 1990. 13

[72] F. Reinhard. Computation of solvent entropies from molecular dynamics simulations. Master's thesis, University of Göttingen, 2005. 47

[73] F. Reinhard and H. Grubmüller. Estimation of absolute solvent and solvation shell entropies via permutation reduction. *The Journal of chemical physics*, 126:014102, 2007. 16, 46, 48

[74] C. Ronchi, R. Iacono, and P.S. Paolucci. The" cubed sphere": a new method for the solution of partial differential equations in spherical geometry. *Journal of Computational Physics*, 124(1):93–114, 1996. 80

[75] J. Schlitter. Estimation of absolute and relative entropies of macromolecules using the covariance matrix. *Chemical Physics Letters, 215*, 6, 1993. 18, 46

[76] H. Singh, N. Misra, V. Hnizdo, A. Fedorowicz, and E. Demchuk. Nearest neighbor estimates of entropy. *American Journal of Mathematical and Management Sciences*, 23(3-4):301–21, 2003. 57

[77] B. Steffen and R. Hosemann. Critical remarks on the ornstein-zernike integral equation. *Berichte der Bunsengesellschaft für physikalische Chemie*, 80(8):712–715, 1976. 23

[78] H.A. Stern. Molecular simulation with variable protonation states at constant ph. *The Journal of chemical physics*, 126:164112, 2007. 27

[79] M.M. Teeter, S.M. Roe, and N.H. Heo. Atomic resolution (.83 Å) crystal structure of the hydrophobic protein crambin at 130 k. *Journal of molecular biology*, 230(1):292–311, 1993. 67

[80] D. van der Spoel, P.J. van Maaren, and H.J.C. Berendsen. A systematic study of water models for molecular simulation: Derivation of water models optimized for use with a reaction field. *The Journal of Chemical Physics*, 108:10220, 1998. 28

[81] W.F. Van Gunsteren and H.J.C. Berendsen. Groningen molecular simulation (gromos) library manual. *Biomos, Groningen*, 1987. 28

[82] W.F. van Gunsteren and P.K. Weiner. *Computer simulation of biomolecular systems: theoretical and experimental applications*. ESCOM Science Publishers, 1989. 26

[83] S.J. Weiner, P.A. Kollman, D.T. Nguyen, and D.A. Case. An all force field for simulatins of proteins and nucleid acids. *J. Comp. Chem*, 7:230–252, 1986. 28

[84] J. Zielkiewicz. Structural properties of water: Comparison of the spc, spce, tip4p, and tip5p models of water. *The Journal of chemical physics*, 123:104501, 2005. 23, 64

[85] Robert W. Zwanzig. High-temperature equation of state by a perturbation method. i. nonpolar gases. *The Journal of Chemical Physics*, 22(8):1420–1426, 1954. 19

Curriculum Vitae

Persönliche Daten

Name	Stephanus Michael Fengler
Geburtsort	Hildesheim
Geburstag	24.04.1980
Staatsang.	deutsch
Familienstand	ledig
Anschrift	Prinzenstraße 20a 37073 Göttingen

Ausbildung

1986–1990	Dohnser Schule Alfeld
1990–1992	Orientierungsstufe Alfeld
1992–1999	Gymnasium Alfeld Abitur im Juni 1999
10/99–09/06	Studium der Physik an der Universtität Ulm Thema der Diplomarbeit: “Molecular Dynamics Simulations of Neuroglobin – Ligand Binding and Protein Dynamics”
02/07–03/11	Promotion an der Georg-August-Universität Göttingen im Max Planck Institut für biophysikalische Chemie Thema der Doktorarbeit: “Estimating Orientational Water Entropy at Protein Interfaces”

Göttingen, February 8. 2012