Diplomarbeit

Aggregation of Mesoscopic Protein-like Heteropolymers

Christoph Junghans

November 2006
(with corrections from Feb 15 2007)

Betreuer: Prof. Dr. Wolfhard Janke
Dr. Michael Bachmann

Gutachter: Prof. Dr. Wolfhard Janke
Prof. Dr. Ulrich Hansmann

Institut für Theoretische Physik
Fakultät für Physik und Geowissenschaften
Universität Leipzig

Abstract

The studies presented in this thesis are deeply biologically motivated. The main request was to understand a bit more about the statistical mechanics of proteins and protein aggregation, which have been a research field for years. We have used a very simple hydrophobic-polar aggregation model on a mesoscopic level, where a protein is represented by a chain of monomers of the hydrophobic A type or the polar (hydrophilic) B type. We have performed Monte Carlo simulations for up to four chains with different sequences of monomers, hetero- and homopolymers. Even in this simple model the energy landscape is "rough", thus the applied methods were mostly generalized ensemble methods. These sophisticated methods like multicanonical sampling, parallel tempering and multicanonical replica exchange provided more reliable and accurate results than common canonical Monte Carlo simulations. Despite that the interchain interaction was modeled by a weak Lennard-Jones like potential, we found a first-order like transition of the aggregation. In a microcanonical analysis we also found a negative microcanonical specific heat, which is a little bit "exotic".

Zusammenfassung

Proteine spielen in der Natur und im menschlichen Körper eine extrem wichtige Rolle. Die vorliegende Arbeit beschäftigt sich mit der Untersuchung von Proteinen und deren Aggregation.

Es wurde ein sehr stark vereinfachtes Modell, das so genannte AB Modell, verwendet. In diesem Modell sind Proteine nur Ketten von Monomeren des hydrophoben A-Typs oder des polaren B-Typs. Bis zu vier dieser Ketten wurden mit verschiedensten Monte-Carlo-Methoden untersucht. Die kanonischen Monte-Carlo-Simulationen lieferten keine zufriedenstellenden Ergebnisse. Die Ursache dafür ist die sehr stark zerklüftete Energielandschaft. Um verlässlichere Daten zu erhalten, wurden erweiterte Ensemble-Methoden wie multikanonische Simulationen, Parallel Tempering und multikanonisches Replica Exchange verwendet. Trotz der eher schwachen Lennard-Jones-artigen Wechselwirkung zwischen den verschiedenen Ketten wurde ein Phasenübergang ähnlich dem erster Ordnung beobachtet. Eine mikrokanonische Analyse dieses Phasenübergangs zeigte den eher seltenen Effekt einer negativen mikrokanonischen spezifischen Wärmekapazität.

Contents

Overview 1

1 Introduction 3
- 1.1 Proteins 3
- 1.2 The Protein Folding Problem 3
- 1.3 Protein Docking 4
- 1.4 Models for Proteins 4
  - 1.4.1 All-Atom Models 4
  - 1.4.2 Coarse-Grained Models 5

2 Models & Methods 7
- 2.1 Models 7
  - 2.1.1 AB Model 7
  - 2.1.2 Interaction Model 9
  - 2.1.3 Boundary Conditions 10
- 2.2 Monte Carlo Methods 11
  - 2.2.1 Basics of Monte Carlo 11
  - 2.2.2 Random Statistical Physics 12
  - 2.2.3 Markov Chain Process 12
- 2.3 Updates 13
  - 2.3.1 Spherical Update 14
  - 2.3.2 Rotation Update 16
- 2.4 Simulation Methods 16
  - 2.4.1 The Metropolis Update 17
  - 2.4.2 Canonical Simulations 17
  - 2.4.3 Multicanonical Simulations 18
  - 2.4.4 Energy Landscape Paving 19
  - 2.4.5 Parallel Tempering 20
  - 2.4.6 Multi-Histogram Reweighting 21
  - 2.4.7 Multicanonical Replica Exchange 22

2.5 Error Estimation 22
    2.5.1 Autocorrelation Function 22
    2.5.2 Blocking Jackknife Technique 26

3 Conventions **29**
3.1 Units 29
3.2 Common Simulation Settings 29
3.3 Distance Measurement 31
3.4 Measured Thermodynamic Quantities 32
    3.4.1 Energy Quantities 32
    3.4.2 Aggregation Parameter 32

4 Results **35**
4.1 Single-Chain Simulations 35
    4.1.1 Verification of Known Results 35
    4.1.2 Sequence 13.1 36
    4.1.3 Autocorrelation Analysis 37
    4.1.4 Sequences 20.X 38
    4.1.5 Homopolymers 39
4.2 Solvents 40
    4.2.1 Polymers in Solvents 41
4.3 Two Polymer Chains 42
    4.3.1 General Facts 43
    4.3.2 Dual Seq. 13.1 43
    4.3.3 Dependence on the Size of the Periodic Box 46
    4.3.4 Homopolymers 48
    4.3.5 Microcanonical Interpretation 50
4.4 Three and More Chains 53
    4.4.1 Triple Sequence 13.1 53
    4.4.2 Type of Aggregation 54
    4.4.3 Scaling 55
    4.4.4 Additional Studies 56

5 Summary **57**

Appendix **61**

A Details about the Updates **61**
A.1 Spherical Update 61
A.2 Rotation Update 63

B The Multicanonical Recursion **65**

C Multi-Histogram Reweighting **69**

D Usage of ABSimT **71**
- D.1 Function Overview 71
- D.2 Structure Overview 72
- D.3 How to Do a Simulation 73
- D.4 How to Create an Init-File 73
- D.5 Technical Details 73
  - D.5.1 ABlib 73
  - D.5.2 Global Variables 74
  - D.5.3 Main Update Function 75
  - D.5.4 Further Details 75

E Calculations **77**
- E.1 Rotation Matrix 77
- E.2 Radius of Gyration 78
- E.3 Thermal Fluctuations Equation 79
- E.4 Two Particles in Periodic Box 80

Bibliography **90**

List of Figures

1.1 Secondary structure of goose lysozyme 3
1.2 Sketch of two proteins coupling by rigid-body docking 4

2.1 Scheme of a polymer chain in the AB model 8
2.2 Sketch of energy values of the two terms of the AB model 8
2.3 Scheme of the interaction for multiple chains 9
2.4 Distance measurement in a periodical box 10
2.5 Scheme of the spherical update 14
2.6 Scheme of the rotation update 16
2.7 Parallel tempering scheme for even number of processes 21
2.8 Parallel tempering scheme for odd number of processes 21
2.9 Muca replica exchange scheme for even number of processes . . . 23
2.10 Theoretical behavior of the autocorrelation function 24
2.11 Sketch of the blocking jackknife method 26

3.1 Periodic distance of two points in a 1D box 32
3.2 Sketch of the problem of the center of mass in a periodic box I . 33
3.3 Sketch of the problem of the center of mass in a periodic box II . 33

4.1 Thermodynamics of Seq. 13.1 36
4.2 Example configuration of Seq. 13.1 37
4.3 Autocorrelation function for Seq. 13.1 at $T = 0.15$ and $T = 0.40$ 38
4.4 Specific heat of the 20-monomer sequences 39
4.5 Thermodynamics of Seq. A_{13} 40
4.6 Ground states of 2, 3, 4 and 5 “water” monomers 41
4.7 Specific heat and lowest energy found for Seq. 8.1 42
4.8 Thermodynamics of twice Seq. 13.1 43
4.9 Example configuration of twice Seq. 13.1 44
4.10 Multicanonical distribution of twice Seq. 13.1 45
4.11 Distribution of E and Γ for twice Seq. 13.1 at T=0.201 45
4.12 Distribution of Γ for twice Seq. 13.1 at low energies 46

4.13 Configuration of twice Seq. 13.1 in the area of lowest energy . . . 47
4.14 Thermodynamics of twice Seq. 13.1 for different box sizes 48
4.15 Muca distribution of twice Seq. 13.1 for different box sizes 49
4.16 Thermodynamics of twice Seq. A_{13} 49
4.17 Lowest energy found for twice Seq. A_{13} 50
4.18 Microcanonical quantities of twice Seq. 13.1 51
4.19 Reweighted canonical distribution of twice Seq. 13.1 52
4.20 Thermodynamics of three times Seq. 13.1 54
4.21 Histograms of three times Seq. 13.1 at $T \approx 0.212$ 54
4.22 Scaling of different amount of Seq. 13.1 55

A.1 Sketch of the spherical update 62
A.2 Sketch for the determination of the factors a and b I 62
A.3 Sketch for the determination of the factors a and b II 63

E.1 Partition of the integration area for two particles in a box 80

List of Tables

2.1 Simple Monte Carlo integration 12

3.1 Table of the available updates 30
3.2 Table of sequences for a multi chain system 31

4.1 Table of the used heteropolymer sequences 36
4.2 Table of the autocorrelation times for Seq. 13.1 37

D.1 Directory structure of ABSimT 72
D.2 Parameters in the init-file 74

E.1 Monte Carlo program for the mean distance of two particles . . . 81

Overview

This thesis deals with the aggregation of protein-like heteropolymers within the frame of a mesoscopic, coarse-grained model.

The first chapter explains the motivation for choosing this problem. Some basic ideas about proteins and their folding will be illuminated in Sect. 1.1. Also two different simulation approaches, an all-atom approach (see Sect. 1.4.1) and the coarse-grained approach (see Sect. 1.4.2) used here, will be introduced.

Following, the second chapter explains the employed aggregation model (see Sect. 2.1.1) and the applied methods. Basically, Markov chain Monte Carlo simulations, which will be briefly reviewed in Sect. 2.2.3, were used, but because of the "rough" energy landscape, more sophisticated generalized ensemble methods like multicanonical simulations (see Sect. 2.4.3), parallel tempering (see Sect. 2.4.5) and a combination of both, the multicanonical replica exchange (see Sect. 2.4.7), were applied as well.

The third chapter discusses the conventions used in the simulations. Special emphasis has been put on the distance measurement in Sect. 3.3 and the related problem of the periodic box and the, therefore, necessary changes in the thermodynamic quantities.

Then the results will be presented in detail. In the first part of the fourth chapter some known results will be verified (see Sect. 4.1.1). The second part deals with the single monomer aggregation, which is comparable to solvents on a very basic level (see Sect. 4.2) and polymers in solvents. After that, the aggregation behavior of two and more polymer chains will be described in Sect. 4.3. It turns out that the transition from the fragmented to the aggregated phase is first-order like. Also a microcanonical interpretation of the aggregation (see Sect. 4.3.5) will be given which reveals a negative microcanonical specific heat and led to the discovery of a second, weaker transition. This second transition would have been easily overlooked in canonical calculations.

Last but not least the main facts will be compiled in the summary chapter where the most significant information will be presented in a short form.

Chapter 1

Introduction

This chapter gives a short summary of proteins and their functions as motivation for this thesis. Also the two possible variants of modeling, an all-atom model and a mesoscopic coarse-grained model, are briefly reviewed.

1.1 Proteins

From the technical point of view proteins are just “an organic compound that consists of amino acids joined by peptide bonds” [1]. The interesting thing is that most of the proteins fold into a unique three-dimensional structure which is directly correlated to its biological function in the cells (see example in Fig. 1.1). The functions of proteins cover very different areas. Some of them are catalyzing biochemical reactions, transporting molecules and also fighting infections. That is why the study of proteins and their folding is so interesting and important.

1.2 The Protein Folding Problem

Following this diversity of functions, another question directly appears: If one has a given sequence (*primary structure*) to which unique structure (*secondary* and *tertiary structure*) will it fold? Or in other words, if we have a given

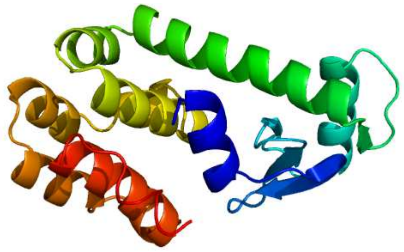


Figure 1.1: Secondary structure of the protein PDB:153L (Goose lysozyme) [2] which has 185 amino acids.

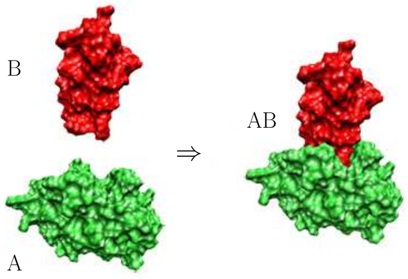


Figure 1.2: Sketch of two proteins coupling by rigid-body docking [3], i.e. without changing bond and torsion angles.

sequence which function in the cell will it have? This is the so called *direct protein folding problem*.

There is another aspect to mention: In some cases the protein will not fold to its unique structure but to some other one. Such mis-folding events will cause diseases like cancer and a deeper understanding of those incidents would obviously be highly desirable.

But from the pharmaceutic point of view the search of the primary structure for a given secondary structure is much more interesting for developing new drugs. This *inverse protein folding problem* needs an even deeper understanding of the folding mechanics.

1.3 Protein Docking

The idea from above has to be generalized to more than one protein, because some proteins will achieve their function only in a compound of proteins. Before having such a complex the single proteins have to dock (see Fig. 1.2) and this process is not totally understood. There are two kinds of docking: The rigid-body docking where the torsion and bond angles stay constant and the flexible docking where these can change.

1.4 Models for Proteins

1.4.1 All-Atom Models

A common way to model a protein is to take all single atoms of the protein and all different interaction forces into account. Different models are circulating, one of them is the ECEPP/3 potential [4] which comprises the four main interactions

$$E_{\text{tot}} = E_{\text{LJ}} + E_{\text{el}} + E_{\text{hb}} + E_{\text{tors}} \ ,$$

with:

- Lennard-Jones potential $E_{\text{LJ}} = \sum_{j>i} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right)$

- Electrostatic potential $E_{\mathrm{el}} = \sum_{i,j} \frac{332 q_i q_j}{\epsilon r_{ij}}$
- Hydrogen-bond energy $E_{\mathrm{hb}} = \sum_{j>i} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right)$
- Torsion energy $E_{\mathrm{tors}} = \sum_l U_l (1 \pm \cos(n_l \chi_l))$

This potential is a mixture of attractive and repulsive interactions that leads to a multiple minima problem. As the number of terms of this potential scales with N^2, where N is the number of atoms, even for small proteins with around a hundred atoms computer simulation are demanding.

The potential comprises empirical constants (e.g. A_{ij}, B_{ij}), which characterizes the force field. These constants are a priori unknown and have to be determined out of quantum-chemical calculations and experimental measurements, which have a certain measurement uncertainty and partly unknown dependences. Because of the complicated form of the potential, small changes in the empirical constant cause great changes in the result of the simulation. To read off of a “basic” physical law is very hard, if not impossible.

In previous work [5, 6] we have done several studies with the ECEPP/3 potential for a small protein with five amino acids and showed that the results produced by different simulation methods (see Sect.2.4) are consistent. In contrast, we used a coarse-grained model for the simulations of this work.

1.4.2 Coarse-Grained Models

To understand the basic physics, we used a coarse-graining approach which models the proteins on a mesoscopic scale. Several models on lattices were used in the last decades [7, 8]. Due to the evolution of the computer power [9] off-lattice approaches also became possible. We used the very simple AB model [10, 11], in which every amino acid is seen as a monomer and not every attribute is taken into account. The property to remain is hydrophobicity and polarity; a monomer can be hydrophobic or polar. The bond lengths are fixed and set to unity, so this can be seen as the length scale of this mesoscopic approach. More technical details will be discussed in the next chapter.

Chapter 2

Models & Methods

This chapter discusses the used model and methods. Starting with an introduction to the AB model, we then give an overview about the methods, which are in general Monte Carlo methods.

2.1 Models

In this section the AB model will be described in detail and our extension for multiple chain interactions will be introduced.

2.1.1 AB Model

The AB model [10, 11] is a coarse-grained heteropolymer model, where coarse-graining means modeling at mesoscopic length scales. The model provides two kinds of monomers, the hydrophobic A and polar B type. The energy function has only two terms:

$$E^{\text{total}} = E_{\text{bend}} + E_{\text{LJ}} \ , \tag{2.1}$$

the bending energy E_{bend} and a Lennard-Jones interaction energy E_{LJ}. For the latter, the heteropolymer character comes into play, because the Lennard-Jones interaction differs for the different kinds of monomers. A sketch of a polymer chain in the AB model can been seen in Fig. 2.1. The coordinate vector of the kth monomer is called $\vec{r}_k$ and the distance between the neighboring single monomers of the chain is fixed and chosen to be in good relation to the distances in the Lennard-Jones energy. Due to this fact we set the bond lengths to unity:

$$|\vec{r}_k - \vec{r}_{k+1}| = 1 \quad \forall \quad 0 < k < N \ , \tag{2.2}$$

where N is the number of monomers in the chain. Originally, the model was designed for 2D [10, 11] but can naturally be expanded into 3D [12].

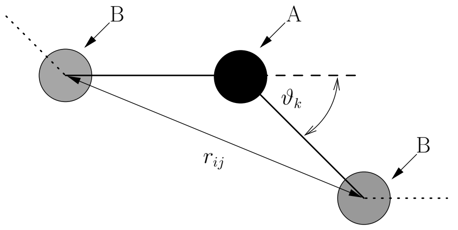


Figure 2.1: A polymer in the AB model can be seen as a chain of two different types of monomers, where the neighbors have a fixed distance, we used distance 1. The bonding angles at the $(k+1)$th atom are denoted by ϑ_k and the distance between the ith and jth atom is called r_{ij}.

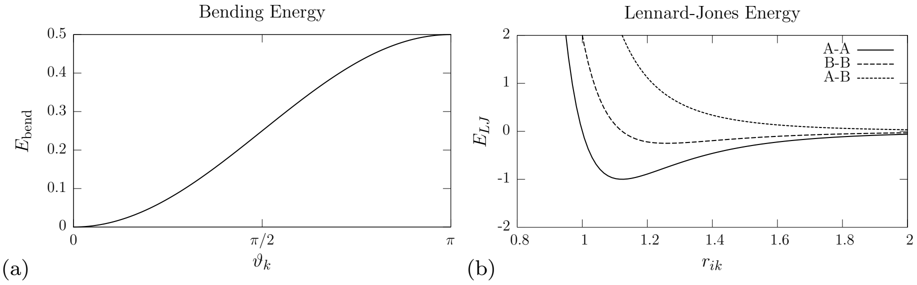


Figure 2.2: (a) The bending energy for a chain with three monomers, e.g. one bonding angle. This part of the total energy has its minimum at angles of 0 and maximums at $-\pi$ and π. (b) The Lennard-Jones energy for the different kinds of interactions, the A-A interaction has a minimum at $r_{\text{min}}^{\text{AA}} = \sqrt[6]{2} \approx 1.122$ and the B-B interaction at $r_{\text{min}}^{\text{BB}} = \sqrt[6]{4} \approx 1.26$ but the A-B potential has no minimum.

The first term in the energy function is the bending energy:

$$E_{\text{bend}} = \frac{1}{4} \sum_{k=1}^{N-2} (1 - \cos \vartheta_k) \ , \tag{2.3}$$

where ϑ_k is the bending angle at the $(k+1)$th monomer, defined by:

$$\cos \vartheta_k = \frac{(\vec{r}_k - \vec{r}_{k+1}) \cdot (\vec{r}_{k+1} - \vec{r}_{k+2})}{|\vec{r}_k - \vec{r}_{k+1}| \cdot |\vec{r}_{k+1} - \vec{r}_{k+2}|} \overset{(2.2)}{=} (\vec{r}_k - \vec{r}_{k+1}) \cdot (\vec{r}_{k+1} - \vec{r}_{k+2}) \ , \tag{2.4}$$

that is why there are $(N-2)$ bonding angles in a chain with N monomers. Obviously ϑ_k is in the interval $[0, \pi)$. The highest bending energy of 1/2 is taken at $\vartheta_k = \pi$ (see Fig. 2.2(a)).

In contrast to the bending energy, the Lennard-Jones energy depends on the types of interacting monomers:

$$E_{\text{LJ}} = 4 \sum_{i=1}^{N-2} \sum_{j=i+2}^{N} \left(\frac{1}{r_{ij}^{12}} - \frac{C(\sigma_i, \sigma_j)}{r_{ij}^{6}} \right) \ , \tag{2.5}$$

where

$$C(\sigma_i, \sigma_j) = \begin{cases} +1 & : \ \sigma_i = \sigma_j = A \ , \\ +1/2 & : \ \sigma_i = \sigma_j = B \ , \\ -1/2 & : \ \sigma_i \neq \sigma_j \end{cases} \tag{2.6}$$

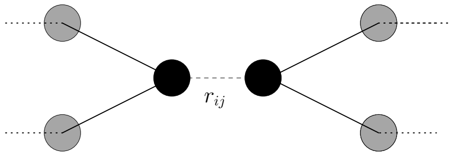


Figure 2.3: Scheme of the interaction for multiple chains. The Lennard-Jones interaction is also assumed for monomers in different chains.

and the distances between the ith and jth monomer is:

$$r_{ij} = |\vec{r}_i - \vec{r}_j|. \tag{2.7}$$

There is no energetic coupling between adjacent monomers in the Lennard-Jones potential due to the fixed bond lengths (see Eqn. (2.2)). The attractive A-A and B-B interactions have a minimum at $r_{\mathrm{min}}^{\mathrm{AA}} = \sqrt[6]{2}$ and $r_{\mathrm{min}}^{\mathrm{BB}} = \sqrt[6]{4}$. Respectively the A-B interaction has no minimum and decreases with the distance generating a repulsive force. The minimum of the A-A interaction at $E_{\mathrm{min}}^{\mathrm{AA}} = -1$ lies much deeper than the minimum of the B-B interaction at $E_{\mathrm{min}}^{\mathrm{BB}} = -1/4$ (see Fig. 2.2(b)). That is why the polymer chain in ground-state prefers to form A-A rather than B-B contacts. In comparison to nature, the A type can be seen as hydrophobic monomers and the B type as hydrophilic or polar monomers.

2.1.2 Interaction Model

The model described in Sect. 2.1.1 is a single-chain model which we expanded by extending the Lennard-Jones interaction to monomers in different chains. So the total energy for a system of chains is thus given by:

$$E_{\mathrm{system}} = \sum_{i=1}^{K} E_i^{\mathrm{total}} + E_{\mathrm{interact}} \ , \tag{2.8}$$

where E_i^{total} is the energy of the ith chain (see Eqn. (2.1)) out of a system of K chains.

The interaction energy can be calculated by:

$$E_{\mathrm{interact}} = \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} E_{ij}^{\mathrm{interact}} \ , \tag{2.9}$$

where $E_{\mathrm{interact}}^{i,j}$ is the interaction energy between the ith and jth polymer:

$$E_{ij}^{\mathrm{interact}} = 4 \sum_{k=1}^{N_i} \sum_{l=1}^{N_j} \left(\frac{1}{r_{kl}^{12}} - \frac{C(\sigma_k, \sigma_l)}{r_{kl}^{6}} \right) \ , \tag{2.10}$$

where $C(\sigma_i, \sigma_j)$ is defined in Eqn. (2.6). N_i is the number of monomers in the ith chain respectively N_j in the jth chain (see Fig. 2.3). As can easily be seen, this is just one possible expansion of the model but a natural one. One

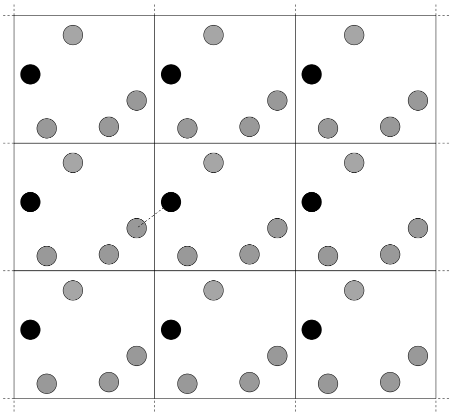


Figure 2.4: System in a box with periodical boundary conditions. Of all possible distances one defines the shortest distance to be the measured distance.

could argue that this is not the best choice, because inner- and outer polymer interactions often differ in experimental analysis. Also the bending energy alone cannot compensate this difference. But in order to keep the model as simple as possible and free of additional interaction parameters, this expansion is most suitable.

2.1.3 Boundary Conditions

Along with the simulation of multiple chains, a boundary condition has to be chosen. The force caused by the potential described in Sect. 2.1.1 and Sect. 2.1.2 is very short-ranged, so it can happen that the chains never interact and each one folds in its single chain ground state, which is also some metastable system state. That is why the introduction of a box is useful to study aggregation. For single chain simulation this is not necessary, one can also take the center of mass as point of origin.

For the examined system, two kinds of boundary conditions are possible, periodical and hard walls. To avoid effects of hard walls which will obviously affect the folding process in the edges of the box, periodical boundary conditions are used. Along with this choice of boundary conditions one also has to define a new measurement for the distances between two monomers as it is required for the calculation of the energy and nearly all other quantities (see Eqn. (2.7)). Out of all possible distances one defines the shortest one to be the distance to measure (see Fig. 2.4). This is called *minimum image convention* [13]. The periodic distance between two points $\vec{p}$, $\vec{q}$ can be written as:

$$d^{\text{per}}\left(\vec{p},\vec{q}\right) = \min_{\text{all boxes}} \left|\vec{p}-\vec{q}\right| \ . \tag{2.11}$$

The box size has to be chosen carefully. A too large box size means sampling of single chain properties but too small sizes make you lose the stretched configurations and the fragmented phase. A detailed study of the dependence on the box size can be found in Sect. 4.3.3.

From now on all simulations which take place in a periodic box will use the periodic distance measurement. For a system of one chain and a box that is large enough (see Sect. 4.1.1) the periodic box will not change the thermodynamics because of the short-ranged interaction (Eqn. (2.5)).

2.2 Monte Carlo Methods

Monte Carlo simulations were performed to obtain the thermodynamic quantities of the system. The basic ideas and physical applications of the Monte Carlo method are shortly explained in the following section. A good overview can also be found in standard literature [14, 15].

2.2.1 Basics of Monte Carlo

In connection to the words "Monte Carlo" [16] one always thinks about randomness which comes into play by drawing random numbers inside the simulation. These numbers are generated by special algorithms, so called *random number generators* [17, 18]. In the following section the drawing of random numbers is hidden in the word *probability.* If something is done with a certain probability, we perform this step only in certain amount of cases. Whether something is done or not is decided by the drawn random number. Random numbers are of the interval $[0, 1)$ and probabilities are naturally of the interval $[0, 1]$. Further on, if the random number is smaller than the probability the step is executed, else it is not. When performing this decision several times the result is distributed according to the given probabilities.

One of the first applications was the Monte Carlo integration. For example the function

$$0 \leq f(x) \leq 1 \tag{2.12}$$

should be integrated over the x range $[0, 1]$. The integral

$$I = \int_0^1 \mathrm{d}x \; f(x) \tag{2.13}$$

is the value to calculate. The scheme of the Monte Carlo integration can be seen in Table 2.1. The main point why this works is because:

$$p = \frac{N_{\text{hits}}}{N_{\text{total}}} = \frac{I}{1 \cdot 1} \; . \tag{2.14}$$

The probability p to draw a random point in the area under the function is equal to the ratio of integral I to the area of the whole region ($[0, 1] \times [0, 1]$). And this probability p can be measured as the ratio of the number of hits N_{hits} to the total number N_{total} of drawn points. But now back to physics.

```
H=0
for ( i = 1 to N )
  x=random()
  y=random()
  if ( y < f(x) )
    H=H+1
end for
I= H/N
print I
```

Table 2.1: A simple Monte Carlo program to integrate a function `f(x)` ($0 \leq f(x) \leq 1$) over the interval $[0, 1]$. `random()` gives a random number, which is uniformly distributed in the interval $[0, 1)$. `N` gives the number of iteration steps.

2.2.2 Random Statistical Physics

As known from statistical physics [19], the expectation value of a quantity is given as:

$$\langle O \rangle(T) = \frac{\sum_{\mu \in \mathcal{C}} O_\mu \mathrm{e}^{-E_\mu/\mathrm{k_B}T}}{\sum_{\mu \in \mathcal{C}} \mathrm{e}^{-E_\mu/\mathrm{k_B}T}} \;, \tag{2.15}$$

where the sum goes over all possible configurations μ of the system. Now the basic idea of Monte Carlo is to take a random subset $\mathcal{R}$ instead of the whole configuration space $\mathcal{C}$. And using the following mean value as estimator for the expectation value (see Eqn. (2.15)):

$$\hat{O}(T) = \frac{\sum_{\mu \in \mathcal{R}} O_\mu \mathrm{e}^{-E_\mu/\mathrm{k_B}T}}{\sum_{\mu \in \mathcal{R}} \mathrm{e}^{-E_\mu/\mathrm{k_B}T}} \;. \tag{2.16}$$

This random subset has to be chosen carefully, so that all important states are sampled. This is called importance sampling [15].

2.2.3 Markov Chain Process

The usual way to produce a set of important states is a Markov chain process

$$\mu \overset{P(\mu \to \nu)}{\longrightarrow} \nu \overset{P(\nu \to \lambda)}{\longrightarrow} \lambda \;. \tag{2.17}$$

Starting from a configuration μ one goes to another configuration ν with a certain probability $P(\mu \to \nu)$ and so on, that is why it is called chain process. The subset produced by this procedure is distributed with p_μ which will be defined in Eqn. (2.22). This distribution obviously depends on the transition probabilities $P(\mu \to \nu)$ which have to fulfill three conditions:

- Ergodicity = every transition should at least be possible in N_S steeps

$$P(\nu \to \mu) = P(\nu \to \lambda_1) \cdot \prod_{i=1}^{N_S} P(\lambda_i \to \lambda_{i+1}) \cdot P(\lambda_N, \mu) \;, \tag{2.18}$$

- Normalization
$$0 \leq P(\nu \to \mu) \leq 1 \tag{2.19}$$
  and
$$\sum_{\mu} P(\nu \to \mu) = 1 \ , \tag{2.20}$$
- Balance = the probability p_μ of a configuration μ in this random subset is given by:
$$p_\mu = \sum_{\nu} p_\nu P(\nu \to \mu) \ . \tag{2.21}$$

To avoid so called random cycles [15], which are subsets which do not consist of all regions of the configuration space, the last condition has to be changed to the stronger condition of *detailed balance*

$$p_\mu P(\mu \to \nu) = p_\nu P(\nu \to \mu) \ . \tag{2.22}$$

The choice of $P(\mu \to \nu)$ depends on the specific problem. In general one can use methods like the Metropolis [20], heatbath [21], Glauber [21] update or cluster algorithms [22, 23]. The transition probabilities can be broken down to [15]:

$$P(\mu \to \nu) = A(\nu \to \mu) g(\nu \to \mu) \ , \tag{2.23}$$

where $A(\nu \to \mu)$ is the *acceptance ratio* and $g(\nu \to \mu)$ is the *choose probability*, which is indirectly chosen by the update.

2.3 Updates

For the chosen model, the update plays an essential role to generate new configurations. In the case of spin systems one never notices this, because a single spin flip is a simple but permitted update. For polymer systems like the one examined here it quickly gets much more complicated.

The choose probabilities $g(\nu \to \mu)$ must fulfill two important properties. First, the update or at least the combination of updates has to be ergodic. This is important to make sure that every configuration in the configuration space could be reached in a finite number of updates. Second, the ratio of probabilities to get the new from the old configuration and backward must be calculable. In detail this means:

- Ergodicity = every configuration can be reached from every configuration in N_S steeps
$$g(\nu \to \mu) = g(\nu \to \lambda_1) \cdot \prod_{i=1}^{N_S} g(\lambda_i \to \lambda_{i+1}) \cdot g(\lambda_{N_S}, \mu) \ , \tag{2.24}$$

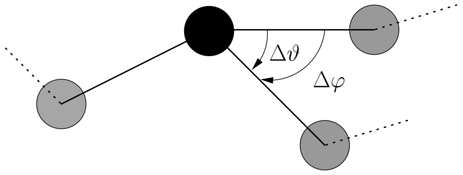


Figure 2.5: Spherical update. One monomer is moved on the surface of the spherical sector with opening angle $\vartheta_{\max} = 5°$. The change of the bonding angle is called $\Delta\vartheta$ and the rotation angle is named $\Delta\varphi$. All following monomers are moved by the same difference.

- Transition ratio:

$$\frac{g(\nu \to \mu)}{g(\mu \to \nu)} = \text{const.}(\nu, \mu) \ , \tag{2.25}$$

- Normalization

$$0 \leq g(\nu \to \mu) \leq 1 \tag{2.26}$$

and

$$\sum_{\mu} g(\nu \to \mu) = 1 \ . \tag{2.27}$$

For the following updates the transition ratio is always one:

$$g(\nu \to \mu) = g(\mu \to \nu) \ , \tag{2.28}$$

i.e. the "forward" and "backward" moves between two configurations have the same probabilities. This can be seen in the following sections. If the backward and forward transitions don't have the same transition probabilities this has to be taken into account in the transition probability (Eqn. (2.44)), and is then called *biased* update[1].

2.3.1 Spherical Update

This update changes the tail of a chain by moving one monomer of a sphere, that is why it is called spherical update.

First it picks one arbitrary monomer, e.g. the jth one, then the $(j+1)$th monomer is changed to a position on the cap of a spherical sector. The center of the associated sphere is given by the jth monomer and the opening angle of this spherical sector is limited by $\vartheta_{\max} = 5°$. As the bonding lengths are fixed to unity, the sphere has obviously the radius one. For better understanding see Fig. 2.5. The old connection vector $\vec{r}$ between the jth and $(j+1)$th monomer

---

[1] Originally this was called Metropolis-Hasting update [24, 25]. In recent times there are several new examples like Biased Metropolis-heatbath algorithm [26], biased multicanonical sampling [27] and Multiple Gaussian modified ensemble [28].

$$\vec{r} = \vec{r}_{j+1} - \vec{r}_j \tag{2.29}$$

is replaced by a new vector $\vec{r}\,'$ given as:

$$\vec{r}\,' = r\cos\Delta\vartheta\ \vec{e}_r(\vec{r}\,) + r\sin\Delta\vartheta\sin\Delta\varphi\ \vec{e}_\varphi(\vec{r}\,) + r\sin\Delta\vartheta\cos\Delta\varphi\ \vec{e}_\vartheta(\vec{r}\,)\ , \tag{2.30}$$

which simplifies with the help of Eqn. (2.2) to

$$\vec{r}\,' = \cos\Delta\vartheta\ \vec{e}_r(\vec{r}\,) + \sin\Delta\vartheta\sin\Delta\varphi\ \vec{e}_\varphi(\vec{r}\,) + \sin\Delta\vartheta\cos\Delta\varphi\ \vec{e}_\vartheta(\vec{r}\,)\ . \tag{2.31}$$

For details about the choice of $\vec{e}_r$, $\vec{e}_\varphi$ and $\vec{e}_\vartheta$ and the dependence on $\vec{r}$ see Sect. A.1. The angles $\Delta\varphi$ and $\Delta\vartheta$ are random variables. But as $\vec{r}\,'$ should be distributed uniformly on the cap of the spherical sector, the angles have to be chosen in a special way. Every infinitesimal small part of the area $\mathrm{d}A$ on the cap of the spherical sector should have the same probability:

$$\mathrm{d}P \propto \mathrm{d}A\ . \tag{2.32}$$

With the help of ordinary spherical coordinates one gets:

$$\mathrm{d}A = \cos\vartheta\ \mathrm{d}\vartheta\mathrm{d}\varphi = \mathrm{d}(\cos\vartheta)\mathrm{d}\varphi\ , \tag{2.33}$$

that is why the $\Delta\varphi$ is chosen uniformly from the interval $[0, 2\pi)$, but for $\Delta\vartheta$ the cosine $\cos\Delta\vartheta$ must be chosen uniformly from the interval $(\cos\vartheta_{\max}, 1]$. So Eqn. (2.31) can be written as:

$$\begin{array}{rcl} \vec{r}\,' & = & (1 - a_r(1-\cos\vartheta_{\max}))\ \vec{e}_r(\vec{r}\,) \\ & & +\sqrt{1-(1-a_r(1-\cos\vartheta_{\max}))^2}\sin(2\pi b_r)\ \vec{e}_\varphi(\vec{r}\,) \\ & & +\sqrt{1-(1-a_r(1-\cos\vartheta_{\max}))^2}\cos(2\pi b_r)\ \vec{e}_\vartheta(\vec{r}\,)\ , \end{array} \tag{2.34}$$

where a_r and b_r are uniformly distributed random numbers of the interval $[0, 1)$ which can be generated by every common random number generator [18, 17].

All the following monomers $(j+2 \ldots N)$ are changed by the difference vector

$$\Delta\vec{r} = \vec{r}\,' - \vec{r}\ . \tag{2.35}$$

As the direction of numbering the monomers is free to choose one can do this update in backward and forward direction, here called forward and backward spherical update. Obviously for a system of a single polymer both updates are identical but not for two or more polymers in a system.

As mentioned above, the probability for this update to get from an old to a new configuration and backward is equal (see Eqn. (2.28)). This can easily be proven by replacing $\vec{r}$ by $\vec{r}\,'$. (And remember that $\Delta\varphi$ and $\Delta\vartheta$ are random angles.)

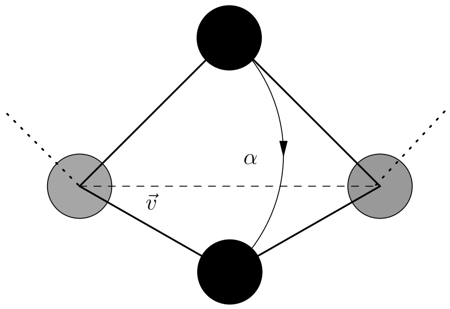


Figure 2.6: Rotation update, which rotates a single monomer by random angle α around the axis $\vec{v}$ defined by the two neighboring monomers. This is a pivot-like update, which itself alone is not ergodic, because the end monomers can not be updated.

2.3.2 Rotation Update

The rotation update is a very simple update which changes only the position of one single monomer. On a 2D-lattice this update is a corner flip update, but due to the continuous space there is more than one possibility to flip. First, a monomer is chosen randomly, e.g. the jth monomer. As axis $\vec{v}$ for the rotation the connection between the $(j-1)$th and the $(j+1)$th monomer is used:

$$\vec{v} = \frac{\vec{r}_{j-1} - \vec{r}_{j+1}}{|\vec{r}_{j-1} - \vec{r}_{j+1}|} \ . \tag{2.36}$$

Next, the connection vector $\vec{r} = \vec{r}_j - \vec{r}_{j-1}$ is rotated by a random angle $\alpha \in [0, 2\pi)$ around this axis (see Fig. 2.6). So the new position of the jth monomer is given by:

$$\vec{r}_j^{\,\prime} = R(\vec{v}, \alpha)\vec{r} + \vec{r}_{j-1} \ . \tag{2.37}$$

For details about the rotation matrix $R(\vec{v}, \alpha)$ see Sect. A.2. Obviously this update is not ergodic, so the sequence of updates together with the spherical update have to be chosen to be ergodic, but at least the probability to come from the old to the new configuration and backward is equal (see Eqn. (2.28)).

2.4 Simulation Methods

In addition to the updates this section gives a short overview of the used Monte Carlo methods, which are mostly *generalized ensemble methods* [29]. These methods are necessary because of the mixture of repulsive and attractive interaction forces which make the energy landscape "rough"[2]. To overcome this multiple-minima problem, more sophisticated simulation techniques, like parallel tempering and multicanonical simulation, have to be used to obtain reliable and accurate results. All these could also be combined with other models if another update is given.

---

[2] The phrase "rough energy landscape" is often used together with spin glass and biophysical systems and became a dictum for the problems in simulating such systems with multiple-minima in the energy.

2.4.1 The Metropolis Update

We used the Metropolis update [20]:

$$P_{\text{Metro}}(\nu \to \mu) = A_{\text{Metro}}(\nu \to \mu) g(\nu \to \mu) + \left(1 - \sum_{\lambda} A(\nu \to \lambda) g(\nu \to \lambda)\right) \delta_{\nu\mu} \tag{2.38}$$

with the famous Metropolis acceptance ratio:

$$A_{\text{Metro}}(\nu \to \mu) = \min\left(1, \frac{p_\mu}{p_\nu}\right) \ . \tag{2.39}$$

The second term on the r.h.s. of Eqn. (2.38) is simply to satisfy the normalization (see Eqn. (2.20)). The condition of detailed balance (see Eqn. (2.22)) is also fulfilled:

$$\begin{aligned}
\frac{P_{\text{Metropolis}}(\nu \to \mu)}{P_{\text{Metropolis}}(\mu \to \nu)} &= \frac{A(\nu \to \mu)}{A(\mu \to \nu)} \frac{g(\nu \to \mu)}{g(\mu \to \nu)} (1 - \delta_{\mu\nu}) + \delta_{\mu\nu} && (2.40)\\
&\overset{(2.28)}{=} \frac{A(\nu \to \mu)}{A(\mu \to \nu)} (1 - \delta_{\mu\nu}) + \delta_{\mu\nu} && (2.41)\\
&= \frac{A(\nu \to \mu)}{A(\mu \to \nu)} + \delta_{\mu\nu} \underbrace{\left(1 - \frac{A(\mu \to \mu)}{A(\mu \to \mu)}\right)}_{=0} && (2.42)\\
&= \frac{A(\nu \to \mu)}{A(\mu \to \nu)} && (2.43)\\
&= \frac{p_\mu}{p_\nu} \ . && (2.44)
\end{aligned}$$

The normalization condition in Eqn. (2.19) can be satisfied by the choice of $g(\nu \to \mu)$.

We used the Metropolis update because one does not need to know all possible transitions like for the heatbath algorithm. And cluster algorithms are so far unknown for polymers.

After performing a simulation with those conditions one obtains a set $\mathcal{M}$ of configurations distributed with p_μ. So Eqn. (2.16) changes to:

$$\hat{O}(T) = \frac{\sum_{\mu \in \mathcal{M}} O_\mu {p_\mu}^{-1} \mathrm{e}^{-E_\mu/\mathrm{k_B}T}}{\sum_{\mu \in \mathcal{M}} {p_\mu}^{-1} \mathrm{e}^{-E_\mu/\mathrm{k_B}T}} \ . \tag{2.45}$$

This is the estimator of the expectation value (see Eqn. (2.15)) and a reweighted mean value, this is why the equation is the so called *master reweighting equation* which is important for every one of the following methods.

2.4.2 Canonical Simulations

As the name implies, this kind of simulation tries to produce a random subset of states distributed according to the canonical Gibbs-Boltzmann ensemble. That is why one uses:

$$p_\mu = \mathrm{e}^{-E_\mu/T} \ , \tag{2.46}$$

where $\mathrm{k_B}$ was set to 1 (see Sect. 3.1). Especially in Eqn. (2.39) of Sect. 2.4.1 this has to be inserted. The sampled distribution of energies is given by:

$$P_{\mathrm{can}}(E,T) = \Omega(E)\mathrm{e}^{-E/T} \ , \tag{2.47}$$

which is the canonical distribution and $\Omega(E)$ the density of states defined by:

$$\Omega(E) = \sum_{\mu\in\mathcal{M}} \delta_{E,E_\mu} \ . \tag{2.48}$$

Also the reweighting equation simplifies to

$$\hat{O}(T) = \frac{\sum_{\mu\in\mathcal{M}} O_\mu}{\sum_{\mu\in\mathcal{M}} 1} = \frac{1}{n}\sum_{i=1}^{n} O_i = \bar{O}(T) \ , \tag{2.49}$$

which is just the common mean value $\bar{O}(T)$ of n different measurements.

This sampling was the conventional method for years but has several problems. Because of the fixed temperature certain configurations have very low probabilities, that is why in a simulation of finite length they will never appear. This leaves only a very small interval of reweighting the histogram to other temperatures (Eqn. (2.45) and [30]).

$$H_{\mathrm{can}}(E,T') = H_{\mathrm{can}}(E,T')\mathrm{e}^{E/T}\mathrm{e}^{-E/T'} \ . \tag{2.50}$$

Also mean values can be calculated in very small intervals around the simulation temperature. More simulations at different temperatures are necessary and have to be combined by multi histogram reweighting (see Sect. 2.4.6). But because of the rough free-energy landscape the system can get trapped during the simulation and it is not sure that an important region is sampled.

2.4.3 Multicanonical Simulations

As the canonical simulations have problems in sampling configurations with low probabilities, it seems useful to sample another distribution:

$$P_{\mathrm{muca}}(E) = \Omega(E)W^{\mathrm{muca}}(E) \approx \mathrm{const.} \tag{2.51}$$

This is the multicanonical distribution [31]. Obviously this distribution does not have any problems with low probabilities, because it is flat. The state distribution is given by:

$$p_\mu \propto W^{\mathrm{muca}}(E_\mu) \ , \tag{2.52}$$

which has to be inserted in Eqn. (2.39). But a new problem arises: The density of states $\Omega(E)$ is unknown and so the multicanonical weights $W^{\mathrm{muca}}(E)$ also have to be identified. One way to solve this problem is to use the estimator of the density of states from the Wang-Landau algorithm [32]. Here another way described by W. Janke [33] was used. The so called *multicanonical recursion* is explained in detail in Sect. B. In summary it works as follows [34]:

1. Set start values

$$n = 1 \quad , \quad W_1^{\text{muca}}(E) = 1 \quad \forall \quad E \tag{2.53}$$

2. Perform multicanonical simulation with $p_\mu = W_n^{\text{muca}}(E_\mu)$ and gain histogram $H_n(E)$ of energies

3. Calculate the old ratios of neighboring energies out of the old weights

$$R_n(E) = \frac{W_n^{\text{muca}}(E + \Delta E)}{W_n^{\text{muca}}(E)} \tag{2.54}$$

4. Calculate the new ratios

$$\ln R_{n+1}(E) = \ln R_n(E) + \frac{q_n(E)}{p_n(E)} \ln\left(\frac{H_n(E)}{H_n(E + \Delta E)}\right) \tag{2.55}$$

   with

$$q_n(E) = \frac{H_n(E + \Delta E) H_n(E)}{H_n(E + \Delta E) + H_n(E)} \ , \tag{2.56}$$

   and

$$p_n(E) = \sum_{i=1}^{n} q_i(E) \tag{2.57}$$

5. Calculate the new weights out of the new ratios (see Eqn. (2.54))

6. Increase n

7. If $n \leq n_{\text{end}}$ goto 2

After n_{end} steps the recursion is aborted, then the obtained weights are used as input

$$W^{\text{muca}}(E) = W_{n_{\text{end}}}^{\text{muca}}(E) \tag{2.58}$$

to perform a long multicanonical simulation with high statistics. The means of a multicanonical simulation can be calculated according to Eqn. (2.45).

2.4.4 Energy Landscape Paving

This minimizer[3] [37] became very popular in the last years because it is so simple. The only thing to change is to replace the energy by an effective energy

$$E \to E_{\text{ef}} = E + H(E, t) \ , \tag{2.59}$$

---

[3] A minimizer is an algorithm to find the lowest energy state. There are several approaches like quasi-Newton [35], conjugated gradients [36] and energy-landscape-paving [37].

where $H(E)$ is the histogram of the energies. The rest is just a normal canonical simulation at a low temperature. This algorithm works quite well because the effective energy is time dependent. If the system gets trapped in a local energy minimum, the histogram $H(E)$ for this E will increase, so the effective energy increases and states with this energy become less attractive. The system will leave the state with this energy. But because of this time dependence, this algorithm only finds the lowest energy but not the thermodynamics. Another problem is that this algorithm cannot distinguish between different structures at the same energy. And configurations which are not generated by the updates can not be found either.

2.4.5 Parallel Tempering

This method [38] is also often called *replica exchange method.* This was coined by K. Hukushima and K. Nemoto [39] in the field of spin glasses, but was independently invented 10 years earlier by R.S. Swendsen and J.-S. Wang [40]. The idea is as simple as brilliant. The probability to be in a configuration μ at temperature T is

$$p_\mu(T) = \mathrm{e}^{-E_\mu/T} \ . \tag{2.60}$$

If you have two copies μ_1 and μ_2 of the system at different temperatures T_1 and T_2 the probability for the combined system to be in that state is:

$$p_{\mu_1}(T_1) \cdot p_{\mu_2}(T_2) = \mathrm{e}^{-(E_{\mu_1}/T_1 + E_{\mu_2}/T_2)} = P(\{\mu_1, T_1\}, \{\mu_2, T_2\}) \ . \tag{2.61}$$

If one exchanges μ_1 and μ_2 the probability changes to:

$$p_{\mu_1}(T_2) \cdot p_{\mu_2}(T_1) = \mathrm{e}^{-(E_{\mu_1}/T_2 + E_{\mu_2}/T_1)} = P(\{\mu_1, T_2\}, \{\mu_2, T_1\}) \ . \tag{2.62}$$

Now one can choose the acceptance ratio as in the Metropolis update (see Eqn. (2.39)):

$$\begin{aligned}
P(\mu_1 \overset{\text{exchange}}{\longleftrightarrow} \mu_2) &= P(\{\mu_1, T_1\}, \{\mu_2, T_2\} \rightarrow \{\mu_1, T_2\}, \{\mu_2, T_1\}) && (2.63)\\
&= \min\left(1, \frac{p_{\mu_1}(T_2) \cdot p_{\mu_2}(T_1)}{p_{\mu_1}(T_1) \cdot p_{\mu_2}(T_2)}\right) && (2.64)\\
&\overset{(2.61)}{=} \min\left(1, \frac{\mathrm{e}^{-(E_{\mu_1}/T_2 + E_{\mu_2}/T_1)}}{\mathrm{e}^{-(E_{\mu_1}/T_1 + E_{\mu_2}/T_2)}}\right) && (2.65)\\
&= \min\left(1, \frac{\mathrm{e}^{-E_{\mu_1}(1/T_2 - 1/T_1)}}{\mathrm{e}^{-E_{\mu_2}(1/T_1 - 1/T_2)}}\right) && (2.66)\\
&= \min\left(1, \mathrm{e}^{-(E_{\mu_1} - E_{\mu_2})(1/T_2 - 1/T_1)}\right) && (2.67)\\
&= \min\left(1, \mathrm{e}^{\Delta E \cdot \Delta\beta}\right) && (2.68)
\end{aligned}$$

This simple equation defines the acceptance ratio. The algorithm can be generalized for any number of copies. But as the acceptance drops exponentially

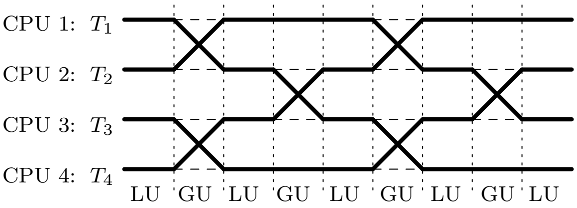


Figure 2.7: A possible scheme of a system with even number of processes (copies of the system) in a parallel tempering simulation. "GU" stands for Global Update and "LU" for Local Update.

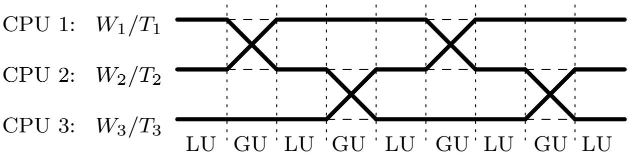


Figure 2.8: A possible scheme of a system with odd number of processes (copies of the system) in a parallel tempering or multicanonical replica exchange simulation. "GU" stands for Global Update and "LU" for Local Update.

with the difference of the temperatures one should only allow exchanges with neighboring temperatures. A good way to choose the temperatures is to look at the histograms at fixed temperature if they overlap. Normally, it turns out that equidistant on β-scale is a good choice. Between the temperature exchange update a certain number of normal canonical updates were performed to change the energy. So the total course is a mixture of global and local updates. This process can easily be distributed on multiple CPU's. The schemes of the different updates (local and global) for odd and even number of copies that have been used can be seen in Fig. 2.7 and Fig. 2.8.

2.4.6 Multi-Histogram Reweighting

After a parallel tempering simulation or multiple canonical simulations one has multiple histograms at different temperatures and wants to combine them to use the overall statistic for calculations at every temperature. The common way to do this is multi-histogram reweighting [41, 42]. Details can be read in Sect. C. The final result is:

$$\hat{\Omega}(E) = \frac{\sum_i H_{\mathrm{can}}(T_i, E)}{\sum_i N_i \hat{Z}^{-1}(T_i) \mathrm{e}^{-E/T_i}} \tag{2.69}$$

where N_i are just the number of entries in the ith histogram

$$N_i = \sum_E H_{\mathrm{can}}(T_i, E) \tag{2.70}$$

and the partition function $\hat{Z}(T)$ has to be determined self-consistently

$$Z(T_k) = \sum_E \hat{\Omega}(E) \mathrm{e}^{-E/T_k} \overset{(2.69)}{=} \sum_E \mathrm{e}^{-E/T_k} \frac{\sum_i H_{\mathrm{can}}(T_i, E)}{\sum_i N_i Z^{-1}(T_i) \mathrm{e}^{-E/T_i}} \ . \tag{2.71}$$

2.4.7 Multicanonical Replica Exchange

There are different approaches [43] of combining the advantages of parallel tempering and multicanonical simulation. The simplest version to use is a multicanonical distribution in every copy of the system instead of the canonical distribution. The distribution of the configurations is given by (like in Eqn. (2.51)):

$$p_{\mu_i} = W_i^{\mathrm{muca}}(E_{\mu_i}) \ , \tag{2.72}$$

where i now numbers the different threads in contrast to Sect. 2.4.5 where i labeled the recursion level. The acceptance ratio can be calculated similar as in Eqn. (2.68).

$$P(\mu_1 \overset{\text{exchange}}{\longleftrightarrow} \mu_2) = \min\left(1, \frac{W_1^{\mathrm{muca}}(E_{\mu_2}) W_2^{\mathrm{muca}}(E_{\mu_1})}{W_1^{\mathrm{muca}}(E_{\mu_1}) W_2^{\mathrm{muca}}(E_{\mu_2})}\right) \ . \tag{2.73}$$

This algorithm behaves like a multicanonical simulation in every single process but finds new states much faster because of the multiple processes. Also it does not get trapped as often as a normal single multicanonical simulation because of the exchanges. Another big advantage is to get more statistics in less time[4]. The scheme of local and global updates can be seen in Fig. 2.8 and Fig. 2.9.

2.5 Error Estimation

The error estimation is an important fact in the field of simulations because a good result with a big error is as worse as a wrong result with a too small error. It is not always possible to do this estimation exactly, but at least for the common mean values it is possible. This is described in the first subsection. For the other cases the blocking jackknife techniques was used, which can be found in the second subsection.

2.5.1 Autocorrelation Function

The update plays an important role for generating new configurations. As mentioned in Sect. 2.3 the sequence of updates has to be chosen ergodic to satisfy

---

[4] This is only half of the truth: If one would perform the multiple (single process) simulation with different seeds on several machines, one would get the same statistics, but would not use the advantages of the exchanges. On the other hand one would get rid of the communication between the processes, which costs lots of computer time. This is why this kind of simulation strongly depends on the equipment and the preference of the researcher.

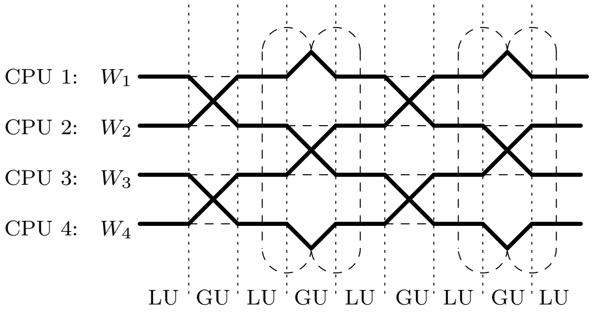


Figure 2.9: A possible scheme of a system with even number of processes (copies of the system). In contrast to the parallel tempering scheme (Fig. 2.7) periodical boundary conditions for the exchanges are introduced. "GU" stands for Global Update and "LU" for Local Update.

that every configuration could be reached in N_S update steps. It is obvious that after one single update not every configuration could be reached, i.e. the configuration after the update looks a little bit like the configuration before the update. This is called *autocorrelation*. The normal *correlation* between two quantities is described by the cross expectation value:

$$\langle A\ B\rangle \ , \tag{2.74}$$

where A and B are some observables. For the uncorrelated case it holds that:

$$\langle A\ B\rangle = \langle A\ \rangle\langle B\rangle \ . \tag{2.75}$$

In the case of autocorrelation we set

$$A = O_k \tag{2.76}$$

and

$$B = O_{k+i} \ , \tag{2.77}$$

where the subscript stands for the discrete time. The values of observable O at time k is correlated with its value at some other time $k+i$:

$$\langle O_k O_{k+i}\rangle - \langle O_k\rangle\langle O_{k+i}\rangle \neq 0 \ . \tag{2.78}$$

In equilibrium the expectation value should be time independent:

$$\langle O_k\rangle = \langle O_j\rangle \ . \tag{2.79}$$

So we define the autocorrelation function [44] which should only depend on the time difference:

$$A(i) = \frac{\langle O_k O_{k+i}\rangle - \langle O_k\rangle^2}{\langle {O_k}^2\rangle - \langle O_k\rangle^2} \ . \tag{2.80}$$

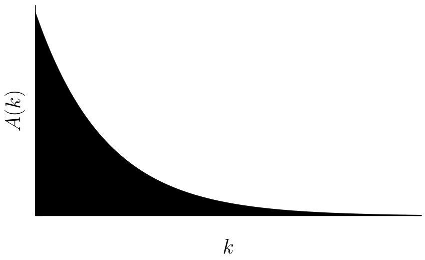


Figure 2.10: The autocorrelation function $A(k)$ drops exponentially with the difference of the time k. This behavior is only accurate in theory, normally, for large difference k the autocorrelation function will start to get noisy. The hatched area is approximately the integrated autocorrelation time τ_{int} (see Eqn. (2.84)).

The denominator was only introduced for normalization. So the autocorrelation function is theoretically of the interval

$$A(i) \in [0,1] \ . \tag{2.81}$$

With "theoretically" we mean in the case of measurements without noise. As we can see later, values smaller than 0 are possible for measurements with noise.

For good[5] processes the autocorrelation function behaves like:

$$A(i) \overset{i\to\infty}{\longrightarrow} a\mathrm{e}^{-i/\tau_{\text{exp}}} \ , \tag{2.82}$$

where τ_{exp} is the *exponential* autocorrelation time. The theoretical behavior can be seen in Fig. 2.10. Another appearance of the autocorrelation function is inside the *integrated* autocorrelation time

$$\tau_{\text{int}} = \frac{1}{2} + \sum_{k=1}^{N} A(k) \left(1 - \frac{k}{N}\right) \ , \tag{2.83}$$

where N is the number of measurements of a quantity. This equation can be simplified by neglecting the last term due to the exponential decay of $A(k)$ (see Eqn. (2.82)) and large N:

$$\tau_{\text{int}} = \frac{1}{2} + \sum_{k=1}^{N} A(k) \ . \tag{2.84}$$

Now these two terms are just the area below the autocorrelation curve, which is easy to measure (see Fig. 2.10). But because of the noisy tail of the autocorrelation function a cutoff k_{max} has to be introduced:

$$\tau_{\text{int}}(k_{\text{max}}) = \frac{1}{2} + \sum_{k=1}^{k_{\text{max}}} A(k) \ . \tag{2.85}$$

---

[5] The definition of good is very hard. Mostly the autocorrelation function is a product of multiple autocorrelation functions with different autocorrelation times. And only the behavior of the strongest autocorrelation time can be seen. The details about autocorrelation can be found in a work by W. Janke [44].

The choice of the cutoff[6] is arbitrary, but has to be done consistently for all measurements. Here the cutoff was chosen to be the point where the autocorrelation function first subtends the base line, because this is obviously the value where two measurements are totally uncorrelated. This choice also helps to avoid problems caused by autocorrelation functions with a “fat tail”.

So the autocorrelation time to measure is

$$\tau_{\text{int}} = \frac{1}{2} + \sum_{k=1}^{N^*} A(k) \ , \tag{2.86}$$

where N^* is the first zero of the autocorrelation function $A(k)$.

As seen in Eqn (2.83), the integrated autocorrelation time plays an important role in the error calculation of correlated mean values $\bar{O}$ of the observable O. This error $\sigma_{\bar{O}}$ is given by [44]:

$$\sigma^2_{\bar{O}} = \frac{\sigma^2_{O_i}}{N} 2\tau_{\text{int}} \ , \tag{2.87}$$

where N is just the number of correlated measurements of the observables O and σ_{O_i} is the common variance of O. In contrast to the uncorrelated error of the mean value

$$\sigma^2_{\bar{O}} = \frac{\sigma^2_{O_i}}{N} \tag{2.88}$$

the factor $2\tau_{\text{int}}$ was added. This implies the definition of the *effective* number of measurements:

$$N_{\text{eff}} = N/2\tau_{\text{int}} < N \ . \tag{2.89}$$

The effective number of measurements is the number of measurements to be carried out in order to have the same error as for the uncorrelated case. It is obvious that as the autocorrelation time gets bigger, N_{eff} gets smaller and smaller.

There are unattended facts about the autocorrelation. First the autocorrelation function depends on the quantity to measure, which means e.g. the autocorrelation time for energy and radius of gyration can be totally different. This comes from the fact that energy is not directly correlated to the radius of gyration. Thus only autocorrelation times of the same quantity are comparable.

Another important point is that the autocorrelation time strongly depends on the kind of simulation. The autocorrelation time of a canonical and a multicanonical simulation can not be compared easily. However, an upper border of the autocorrelation has been compared by the author [5] for multicanonical simulations and parallel tempering simulations with an all-atom protein model.

Only in the case of canonical simulations the autocorrelation time has a simple meaning [44]. And even for this case it depends on the temperature due to

---

[6] Another popular self-consistent choice is to cut off the summation once $k_{\max} \geq \tau_(k_{\max})$.

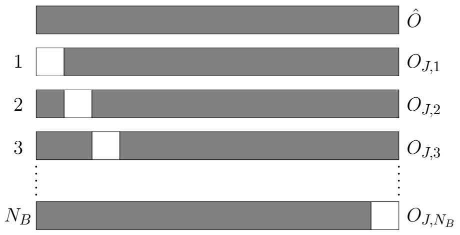


Figure 2.11: The scheme of the block Jackknife error estimation. Out of the total data a estimator $\hat{O}$ is calculated. The N_B Jackknife blocks are formed by leaving out one block of the N_B blocks of the complete data. Every Jackknife block defines a new estimator $O_{J,i}$ out of which the jackknife error (see Eqn. (2.90)) is calculated.

the fact that for higher temperatures the acceptance rate (see Eqn. (2.39)) is higher and so the autocorrelation time gets smaller. And for low temperatures more proposed configurations will be rejected which lead to a higher autocorrelation time.

2.5.2 Blocking Jackknife Technique

The calculation from the last section can be done easily for common mean values, but for the other quantities, e.g. specific heat, the Blocking Jackknife technique was used, because it is much simpler.

This technique [44] is a combination of the Jackknife analysis [45, 46] and the blocking analysis. First an estimator of observable O is calculated out of the complete data of the simulation, called $\hat{O}$ (see Eqn. (2.45)). For this estimator one wants to evaluate the error, too. So the second step is to divide the complete data in N_B blocks, i.e. *blocking* Jackknife techniques. The N_B Jackknife blocks (see Fig. 2.11) are formed by leaving out one of the N_B blocks. (A common mistake is to leave out one of every N_B values of the complete data, but this will not work!) The next step is to calculate the other N_B estimators of O out of the N_B Jackknife blocks which then are called $O_{J,i}$ with $i = 1 \dots N_B$. The Jackknife error is now defined by:

$$\epsilon^2_{J,\hat{O}} = \frac{N_B - 1}{N_B} \sum_{i=1}^{N_B} \left(O_{J,i} - \hat{O} \right)^2 \tag{2.90}$$

For the simplest case of the common mean value:

$$\hat{O} = \bar{O} = \frac{1}{N} \sum_{i=1}^{N} O_i \ , \tag{2.91}$$

where N is the number of measurement. We show that the Jackknife error is equal to the exact error of uncorrelated measurement (see Eqn. (2.87)) under one strong assumption for the number of Jackknife blocks N_{B}:

$$N_{\mathrm{B}} = N \ . \tag{2.92}$$

The jackknife error simplifies to:

$$\begin{aligned}
\epsilon^2_{J,\bar{O}} &\overset{(2.90)}{=} \frac{N-1}{N} \sum_{i=1}^{N} \left(O_{J,i} - \bar{O}\right)^2 && (2.93)\\
&= \frac{N-1}{N} \sum_{i=1}^{N} \left(\frac{1}{N-1} \sum_{j \neq i} O_j - \bar{O} \right)^2 && (2.94)\\
&= \frac{1}{N(N-1)} \sum_{i=1}^{N} \left(\sum_{j=1}^{N} O_j - O_i - (N-1)\bar{O} \right)^2 && (2.95)\\
&\overset{(2.91)}{=} \frac{1}{N(N-1)} \sum_{i=1}^{N} \left(N\bar{O} - O_i - (N-1)\bar{O}\right)^2 && (2.96)\\
&= \frac{1}{N(N-1)} \sum_{i=1}^{N} \left(\bar{O} - O_i\right)^2 && (2.97)\\
&= \frac{\epsilon^2_{O_i}}{N} && (2.98)\\
&= \epsilon^2_{\bar{O}} \ . && (2.99)
\end{aligned}$$

This is the same error as given in Eqn. (2.87), i.e. the error of uncorrelated measurements. Motivated from this result the blocking jackknife technique is the method of choice, because of its simplicity.

Chapter 3

Conventions

3.1 Units

The units are chosen as usual in computational statistical physics, i.e.$\mathrm{k_B} = 1$. So all formulas are free of the $\mathrm{k_B}$'s. For example, the specific heat c_V simplifies to:

$$c_V = \frac{1}{N_\mathrm{G} T^2} \left(\langle E^2 \rangle - \langle E \rangle^2 \right) \ , \tag{3.1}$$

where N_G is the total number of monomers in the system (see Eqn. (3.15)). The global definitions of the quantities measured in Chap. 4 are described in detail in Sect. 3.4.

3.2 Common Simulation Settings

The parameters given here are used in all simulations if not mentioned otherwise in the special section.

The box size has to be chosen carefully. In a big box the single chains will rarely find each other and in a too small box the stretched configurations are not possible. That is why taking

$$L_\mathrm{box} = 40 \tag{3.2}$$

is a good choice, because up to 40 monomers are possible in the stretched configuration in all directions. Also up to 69 monomers are possible in the stretched configuration along the room diagonal of the cube. One should mention that the stretched configuration is less important for the thermodynamics.

The smallest energy difference for the histograms and the multicanonical weights was

$$\Delta E = 0.01 \ . \tag{3.3}$$

Name of the update	Shortcut	ABSimT Number
Nothing	N	0
Forward spherical update	F	1
Backward spherical update	B	2
Rotation update	R	3
Multiple rotation update	Q	4
Move update	M	5

Table 3.1: Table of the available updates. The details are explained in Sect. 2.3. "Nothing" was added just for technical reasons. The ABSimT numbers are helpful to generate an initfile for ABSimT (see Sect. D).

This is a good choice between to big arrays in the program and losing to much accuracy because of the binning. The time series with continuous values are saved as well for data analysis.

As mentioned above, the chosen sequence of updates has to be ergodic (see Sect. 2.3), which leads to the used sequence of update (see Table 3.2). The types of updates implemented in the used program (see Appendix D) are listed in Table 3.1. The update, which does nothing was added only for technical reasons. For the spherical update a maximum opening angle of

$$\vartheta_{\text{max}} = 5^\circ \tag{3.4}$$

for the spherical sector was used. The rotation update chooses the random rotation angle

$$\alpha \in [0, 2\pi) \tag{3.5}$$

out of the full possible interval. In the simulations mostly the ergodic sequence FRBR was used. The monomer on which every single chain update operates is chosen randomly, that means in average every monomer of a chain is updated for the same time. This part is performed as often as there are monomers in the system. Normally, this would be called a sweep but as it is combined with the sequence of updates it is called *sweep sequence* here. The length of such a sweep sequence can be calculated easily:

$$L_{\text{swseq}} = L_{\text{seq}} \cdot \sum_{i=i}^{k} N_i \ , \tag{3.6}$$

where N_i is the number of monomers in the ith chain. The length of the simulations will be measured in units of sweep sequences. For a system with more than one chain the order of touching the different chains is important as well. In the majority of cases the chains are updated in fixed order, one after

Number of Chains	Order of chain to update	Sequence of the updates
1	1111	FRBR
2	12121212	FFRRBBRR
3	123123123123	FFFRRRBBBRRR
4	1234123412341234	FFFFRRRRBBBBRRRR

Table 3.2: The different chains are always updated one after another with the same type of updates and then the type is changed in the order FRBR.

another, at a fixed kind of update, then the update is changed. An example for one to four chains can be seen in Table 3.2.

3.3 Distance Measurement

The distance measurement is changed according to the length of the box (see Eqn. (3.2)), that is why

$$|\cdot - \cdot| \longrightarrow d^{\text{per}}(\cdot,\cdot) \ , \tag{3.7}$$

where d^{per} was defined in Eqn. (2.11) of Sect. 2.1.3. For technical reasons the whole simulations was done in the "first" box:

$$\mathcal{B}_1 = [0, L_{\text{Box}}) \times [0, L_{\text{Box}}) \times [0, L_{\text{Box}}) \ . \tag{3.8}$$

In this case also the shortest difference vector between two points ($\vec{p}, \vec{q} \in \mathcal{B}_1$) can be calculated accurately:

$$\vec{d}_{\text{per}}(\vec{p},\vec{q}) = \begin{pmatrix} d_1^{\text{per}}(p_1,q_1) \\ d_2^{\text{per}}(p_2,q_2) \\ d_3^{\text{per}}(p_3,q_3) \end{pmatrix} = \sum_{i=1}^{3} d_i^{\text{per}}(p_i,q_i)\,\vec{e}_i \ , \tag{3.9}$$

where the ith component of this vector is:

$$d_i^{\text{per}}(p_i,q_i) = \begin{cases} (p_i - q_i) + L_{\text{Box}} & : \quad (p_i - q_i) < -L_{\text{Box}}/2 \ , \\ (p_i - q_i) & : \quad -L_{\text{Box}}/2 < (p_i - q_i) < L_{\text{Box}}/2 \ , \\ (p_i - q_i) - L_{\text{Box}} & : \quad L_{\text{Box}}/2 < (p_i - q_i) \ . \end{cases} \tag{3.10}$$

So the periodical distance simplifies to:

$$d_{\text{per}}(\vec{p},\vec{q}) = \sqrt{\sum_{i=1}^{3} \left(d_i^{\text{per}}(p_i,q_i)\right)^2} \ . \tag{3.11}$$

This is now the shortest distance between all periodic boxes as it was mentioned in Sect. 2.1.3. A plot of the distance of two points in a one dimensional box can be seen in Fig. 3.1.

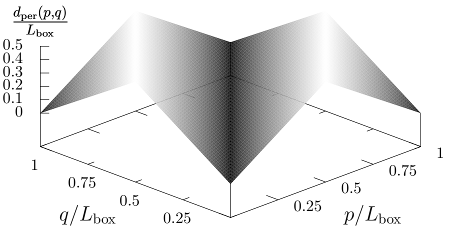


Figure 3.1: Periodic distance of two points in a 1D box scaled with the box length L_{box}.

3.4 Measured Thermodynamic Quantities

3.4.1 Energy Quantities

First of all the Lennard-Jones energy term of Eqn. (2.5) changes a bit because of the periodic box.

$$E_{\mathrm{LJ}} = 4 \sum_{i=1}^{N-2} \sum_{j=i+2}^{N} \left(\frac{1}{d_{\mathrm{per}}(\vec{r}_i, \vec{r}_j)^{12}} - \frac{C(\sigma_i, \sigma_j)}{d_{\mathrm{per}}(\vec{r}_i, \vec{r}_j)^{6}} \right) , \tag{3.12}$$

where $C(\sigma_i, \sigma_j)$ was defined in Eqn. (2.6). The mean value of the energy can be calculated with the help of Eqn. (2.45).

The specific heat is defined as:

$$c_V = \frac{1}{N_{\mathrm{G}}} \frac{\mathrm{d}\hat{E}}{\mathrm{d}T} , \tag{3.13}$$

which can be calculated with the thermal fluctuation equation (see Sect. E.3):

$$c_V = \frac{1}{N_{\mathrm{G}} T^2} \left(\widehat{E^2} - \hat{E}^2 \right) , \tag{3.14}$$

where N_{G} is the total number of monomers in the system with K chains:

$$N_{\mathrm{G}} = \sum_{i=1}^{K} N_i , \tag{3.15}$$

where N_i is the number of monomers in the ith chain.

3.4.2 Aggregation Parameter

To distinguish between aggregated and fragmented configurations of polymers another parameter has to be introduced. Recalling the well-known radius of gyration:

$$r_{\mathrm{gyr}}{}^2 = \frac{1}{N} \sum_{i=1}^{N} (\vec{r}_i - \vec{r}_M)^2 . \tag{3.16}$$

The radius of gyration can be understood as the mean difference from the center of mass $\vec{r}_M$ defined by:

$$\vec{r}_M = \frac{1}{N} \sum_{i=1}^{N} \vec{r}_i . \tag{3.17}$$

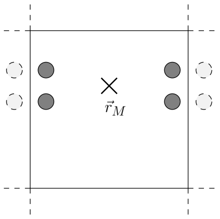


Figure 3.2: The center of mass $\vec{r}_M$ (marked with ×) is not well-defined for particles in a periodic box when using Eqn. (3.17). It should be on the left or right border line, which is the same point in the periodic box.

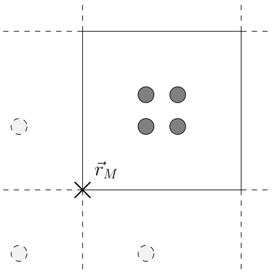


Figure 3.3: The center of mass (marked with ×) is not well-defined for particles in a periodic box when using Eqn. (3.18) because a fixed zero point implies the problem of the global coordinate system, which does not always exist for more than 3 particles.

The center of mass in a periodical box is not well-defined and has problems with certain arrangements of particles (see Fig. 3.2). So the first guess is to change the definition to

$$\vec{r}_M{}' = \frac{1}{N}\sum_{i=1}^{N} \vec{d}_{\text{per}}(\vec{r}_i, \vec{0}) \ . \tag{3.18}$$

But there are also problems as can be seen in Fig. 3.3. The difficulties come from the fact that for more than two particles a global coordinate system cannot always be defined. Luckily Eqn. (3.16) can be rewritten as (see Sect. E.2):

$$r_{\text{gyr}}{}^2 = \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} (\vec{r}_i - \vec{r}_j)^2 \ . \tag{3.19}$$

This also is a good definition in a periodical box after using the periodical distance measurement:

$$r_{\text{gyr}}{}^2 = \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \left(\vec{d}_{\text{per}}(\vec{r}_i, \vec{r}_j)\right)^2 \ , \tag{3.20}$$

where the sum goes over all monomers in the system. One problem remains, this quantity does not distinguish between different chains, that is why we need to go over to the following description:

$$\Gamma^2 = \frac{1}{2K^2}\sum_{i=1}^{K}\sum_{j=1}^{K} \left(\vec{d}_{\text{per}}(\vec{r}_{M,i}, \vec{r}_{M,j})\right)^2 \ , \tag{3.21}$$

where the local centers of mass of the K chains in the system are defined by:

$$\vec{r}_{M,i} = \frac{1}{N_i}\sum_{j=1}^{N_i} \vec{d}_{\text{per}}(\vec{r}_i, \vec{r}_1) + \vec{r}_1 \ . \tag{3.22}$$

For simulations without a periodic box this would just be the common center of mass definition as in Eqn. (3.17). In this definition every other monomer could be possible as reference point as well, as long as $L_{\text{Box}} > 2N_i$, where N_i is the number of monomers in the ith chain. Such measurements with local centers of reference are common in simulations with periodic boundary conditions [47].

This aggregation parameter now distinguishes between the aggregated and fragmented phase. The smaller values belong to the aggregated phase and higher values are related to the fragmented phase. We can say that the aggregation parameter is the *radius of gyration of the centers of mass.* To have a better understanding of this one may take a look from an other point of view at the aggregation parameter (see Eqn. (3.21)). For the simple case of two polymer chains the parameter simplifies to:

$$\begin{aligned} \Gamma^2 &= \frac{1}{2 \cdot 2^2} \sum_{i=1}^{2} \sum_{j=1}^{2} \left(\vec{d}_{\text{per}}(\vec{r}_{M,i}, \vec{r}_{M,j}) \right)^2 && (3.23) \\ &= \frac{1}{4} \cdot \left(\vec{d}_{\text{per}}(\vec{r}_{M,1}, \vec{r}_{M,2}) \right)^2 \ . && (3.24) \end{aligned}$$

That is why the periodic distance between the centers of mass of the two chains is just:

$$d_{\text{per}}(\vec{r}_{M,1}, \vec{r}_{M,2}) = 2\Gamma \ . \tag{3.25}$$

Obviously this is a good quantity to see the aggregation.

The mean value of the aggregation parameter can be calculated from the master reweighting equation (Eqn. (2.45)) and its thermal fluctuation from the thermal fluctuation equation (see Sect. E.3):

$$\frac{\mathrm{d}\hat{\Gamma}}{\mathrm{d}T} = \frac{1}{T^2} \left(\widehat{E\Gamma} - \hat{E} \cdot \hat{\Gamma} \right) \ . \tag{3.26}$$

Chapter 4

Results

This chapter gives a detailed overview of simulations done in this thesis starting with single-chain simulations and going over to two, three and more chains. For details about the usual simulation settings see Sect. 3.2. If there are special settings it will be mentioned in the relevant section. In combination with aggregation we will often use the words "phase", "order parameter" or "phase transition", so we want to emphasize that it should be called *pseudo* phase, *pseudo* order parameter and *pseudo* phase transition, but for simplification we omit the "pseudo".

4.1 Single-Chain Simulations

Several works have been performed on single-chain properties of hydrophobic-polar heteropolymer models, like thermodynamics [48, 49, 50] and folding behavior [51]. This data should be free of systematic errors, because of the very different programs and simulation techniques applied. The first task was to reproduce some of those known results by using my own program ABSimT (see Chap. D).

4.1.1 Verification of Known Results

In comparison to former, similar studies [48, 51], periodic boundary conditions and other updates were used. This can lead to some differences. The periodic box should be no problem for

$$L_{\mathrm{Box}} > 2N \;, \tag{4.1}$$

because in that case the periodic distance in Eqn. (3.11) is just the common distance which was also used in the former studies. For the latter the spherical update was the used update, but in this work the new rotation update

Name	Sequence of monomers
Seq. 13.1	$AB_2AB_2ABAB_2AB$
Seq. 20.1	$BA_6BA_4BA_2BA_2B_2$
Seq. 20.2	$BA_2BA_4BABA_2BA_5B$
Seq. 20.3	$A_4B_2A_4BA_2BA_3B_2A$
Seq. 20.4	$A_4BA_2BABA_2B_2A_3BA_2$
Seq. 20.5	$BA_2B_2A_3B_3ABABA_2BAB$
Seq. 20.6	$A_3B_2AB_2ABAB_2ABABABA$
Seq. 34.1	$AB_2AB_2ABAB_2AB_2ABAB_2ABAB_2AB_2ABAB_2AB$

Table 4.1: Table of the used heteropolymer sequences [48].

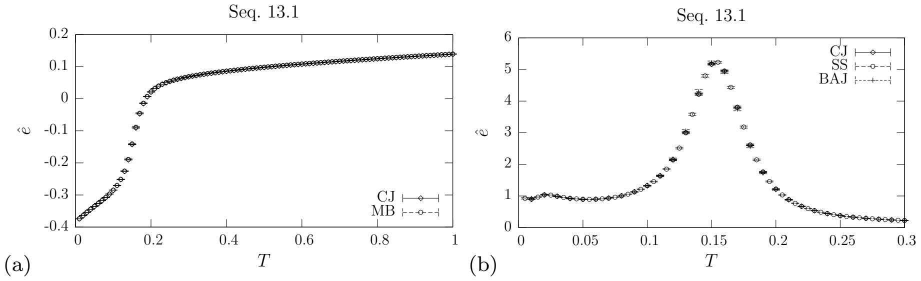


Figure 4.1: (a) The mean of the inner energy of Seq. 13.1 normalized on the number of monomers ($e = E/L$). (b) The specific heat of Seq. 13.1 with the typical peak. The data for this thesis called "CJ", "BAJ"[48] and "SS"[51] show very good consistence.

was applied. But the choice of the update should make no difference for the thermodynamics as long as the sequence of updates is ergodic. The used order of updates can be found in Table 3.2. The selected polymer sequences [48] are listed in Table 4.1.

4.1.2 Sequence 13.1

A comparison of specific[1] inner energy and the specific heat of Seq. 13.1 (see Table 4.1) with data from M. Bachmann *et al.* [48] and S. Schnabel [51] can be seen in Fig. 4.1. The data were obtained by a multicanonical simulation with 120 recursions of 10^6 sweep sequences and a final simulation with 10^8 sweep sequences. The consistence is very good, so it seems that the new updates do not lead to a systematic error and the periodic boundary conditions do not influence the results noticeably. The error bars are a little bit smaller because

---

[1] With "specific" the normalization on the number of monomers is meant, this is a common way to avoid the growing of all quantities with the size of the system.

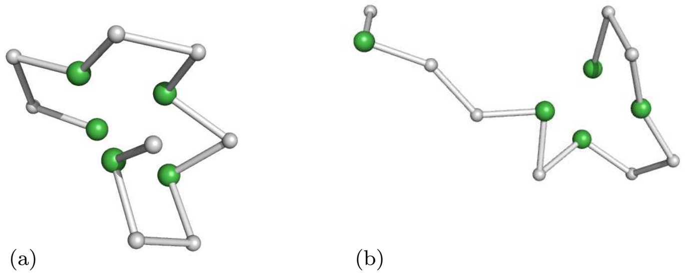

(a) (b)

Figure 4.2: (a) A configuration near the ground state, which belongs to the folded phase. (b) A configuration from the high energy phase which shows the typical random coils.

τ_{int}		
	Sequence of updates	
T	FFFF	FRBR
0.15	741	374
0.40	9.03	3.67

Table 4.2: Table of the different integrated autocorrelation times for Seq. 13.1.

of the lower autocorrelation time which will be discussed in Sect. 4.1.3.

Also it is obvious that there is a peak in the heat capacity, which is caused by a fundamental structural transition, which is also known from lattice heteropolymers [52]. In the low temperature region the folded configurations with their characteristic hydrophobic (A type) core (see Fig. 4.2(a)) are more probable, in contrast to the high-energy phase, where the configurations with random coils are dominating (see Fig. 4.2(b)). The transition point strongly depends on the sequence of the heteropolymer. The peak is shown as the first point in the research of finite-size scaling effects (see Sect. 4.4.3).

4.1.3 Autocorrelation Analysis

We compared the autocorrelation time of the energy for a canonical simulation at two different temperatures, $T = 0.15$ and $T = 0.40$. The investigated sequence was Seq. 13.1 (see Table 4.1). The measured autocorrelation function $A(k)$ and the cut-off integrated autocorrelation time $\tau_{\text{int}}(k)$ are shown in Fig. 4.3. As expected, the autocorrelation time for the lower temperature was in general higher because of the lower acceptance rate (see Table 4.2). The result is that the choice of the sequence of updates gives a factor 2 in a direction that would not have been expected. The mixed sequence FRBR (see Table 3.1) has a smaller autocorrelation time than the sequence FFFF, which was used by nearly all

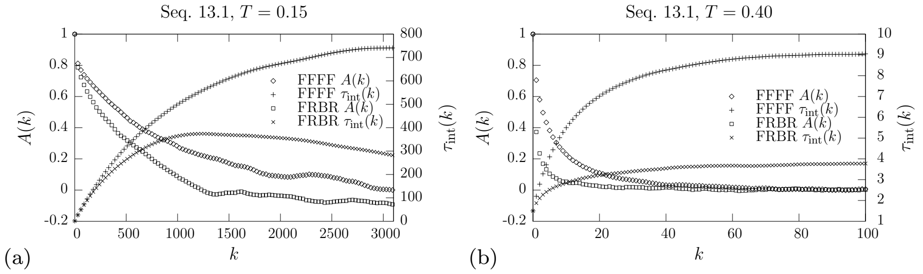


Figure 4.3: (a) The autocorrelation function for Seq. 13.1 at temperature $T = 0.15$. (b) The autocorrelation function for Seq. 13.1 at temperature $T = 0.40$. As expected, the autocorrelation time increases in general with decreasing temperature, from $T = 0.15$ to $T = 0.40$ by a factor 100. Also the autocorrelation time for the mixed update sequence FRBR (see table 3.1) is a factor 2 smaller than for the sequence FFFF.

previous works [48, 50, 51]. An explanation for the results at low temperature may be that the local pivot-like rotation update R is accepted much more often than the global spherical update (F and B), because in this temperature region the folded phase with very compact configurations and hydrophobic cores is dominating.

4.1.4 Sequences 20.X

Also the well-researched sequences with 20 monomers (see Table 4.1) were tested for agreement with data from M. Bachmann *et al.* [48] (see Fig. 4.4), because these sequences are the first check point for every new heteropolymer simulation program or algorithm, even comparisons with Molecular Dynamic simulations [49] are possible.

The similarity of the results from the different simulations is apparent for all 6 sequences. Multicanonical simulations with about 100 recursions and 10^5 sweep sequences each and a final simulation with 10^7 sweep sequences were performed. We also tested different kinds of boundary conditions. Simulations with and without periodic boxes lead to the same results due to the fact, that for a well chosen box length the periodic distance measurement is the same as the common measurement (see Sect. 4.1.1). Around 10 different simulations, like parallel tempering and multicanonical replica exchange simulations, for every chain led to equal results, which satisfy the trustfulness of the program.

From the physical point of view, it can be seen that thermodynamics depends strongly on the sequence of monomers. The heat capacities shown in Fig. 4.4 have one or two peaks at different temperatures which result from the different folding pathways [51] from the random-coil phase to the hydrophobic core region. The small deviations in the lower energy region appear due to less statistics

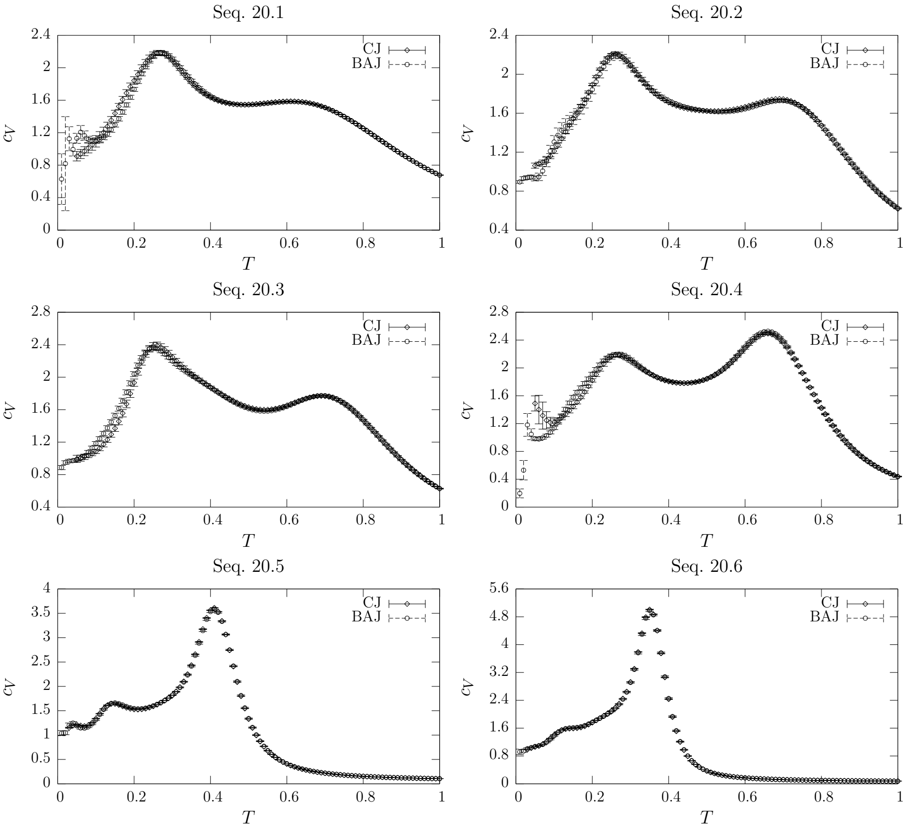


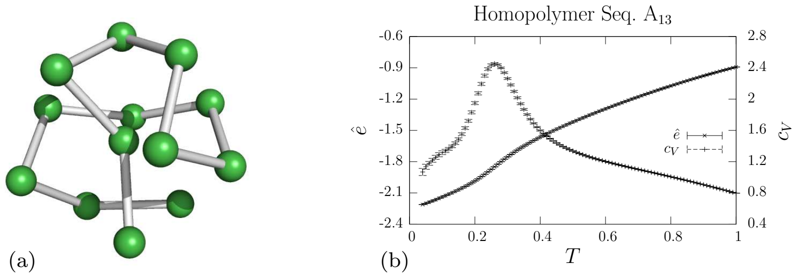
Figure 4.4: Specific heat of sequence with 20 monomers (see Tab. 4.1). The data from this thesis, called “CJ” and data from Bachmann et al. [48] “BAJ” show very good consistence. Depending on the number of hydrophobic A type monomers the maximum of the heat capacity can vary between 2 and 5.

there.

As the number of hydrophobic monomers differs in the sequences, the size of the core is also different. That is why the height of the peaks varies between 2 and 5. The higher peaks were obtained for the sequences with less of the A type (Seq. 20.5 (see Fig. 4.4(e)) and Seq. 20.6 (see Fig. 4.4(f))), which automatically form a more compact core than the others.

4.1.5 Homopolymers

The AB model can also be used for homopolymers by simple means. We chose the A type to be the homomer. One could expect that there will be no hydrophobic cores due to the fact that there are no hydrophilic monomers. So the ground states will have a totally different structure, to provide as much optimal distance between the monomers as possible.



(a) (b)

Figure 4.5: (a) A configuration close to the ground state. The onset of the outer helix can be seen. (b) The average of the inner energy (normalized on the number of monomers) and the specific heat for Seq. A_{13}.

The lowest energy found ($E = -29.25$) in a multicanonical simulation with 50 recursions of 10^5 sweep sequences and a final run with 10^7 sweep sequences can be seen in Fig. 4.5(a). This structure is typical for homopolymers with Lennard-Jones interaction [53, 54].

The region of interesting energies is slightly bigger than for a heteropolymer of the same length. This comes from the fact that the B type monomers are missing and the ground states have a much lower energy. But in general, the graphs remain the same, the mean value of the inner energy is always monotonically increasing and the specific heat has at least one peak at the transition point between the random coil and the folded phase.

4.2 Solvents

Understanding fluids has been a big research field for years that is why we also tried to simulate water in the AB model, which is some kind of Lennard-Jones fluid. But we *extremely* simplified the water to a single B type monomer which is then the simplest case of aggregation, whose energy function simplifies to:

$$E_{\text{system}} = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(\frac{1}{r_{ij}^{12}} - \frac{1}{2r_{ij}^{6}} \right) , \tag{4.2}$$

where N is just the number of monomers in the system. As there are no bonding constraints there is no bending energy. This Lennard-Jones energy function has its minimum at

$$r_{\text{min}}^{\text{BB}} = \sqrt[6]{4} \approx 1.26 \ . \tag{4.3}$$

This distance is the optimal distance for two B type monomers to form the ground state. Every one of those optimal distances gives an energy of

$$E_{\text{min}}^{\text{BB}} = -0.25 \ , \tag{4.4}$$

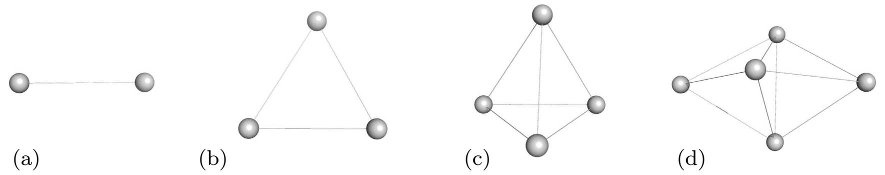
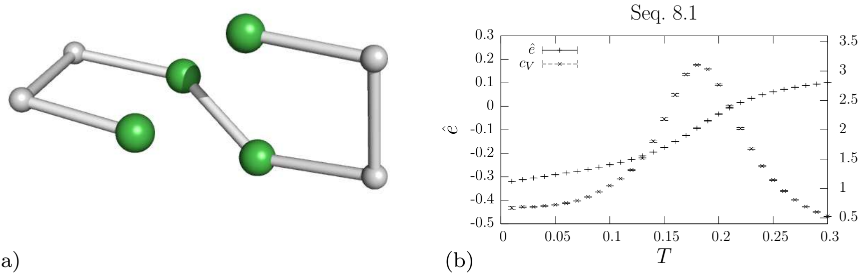

(a) (b) (c) (d)

Figure 4.6: Ground states of 2, 3, 4 and 5 “water” molecules. (a) The ground state of two B type monomers is exactly a line with distance $r_{\mathrm{min}}^{\mathrm{BB}}$ (see Eqn. (4.3). (b) The ground state of three B type monomers is exactly a equilateral triangle. (c) The ground state of four B type monomers is exactly a regular tetrahedrons. (d) For five B type monomers it is not possible to calculate the ground state analytically, but the configuration of the lowest energy found looks like two tetrahedrons sharing the same base.

which is also the ground state energy of 2 B type monomers (see Fig. 4.6(a)). Obviously, the ground state of 3 B type monomers will form a equilateral triangle with $r_{\mathrm{min}}^{\mathrm{BB}}$ as side length (see Fig. 4.6(b)). The energy of this ground state is -0.75. Four molecules form a regular tetrahedron with side length $r_{\mathrm{min}}^{\mathrm{BB}}$ (see Fig. 4.6(c)). This ground state has an energy of -1.5. As four molecules are the maximum that can be treated analytically, we also simulated this system with ELP and multicanonical simulations. The ELP simulation found a configuration with energy of -1.499, which is very close to the exact value. The configuration of lowest energy found in the multicanonical simulation was $E = -1.498$ which is also very close to exact value. The found configurations are regular tetrahedrons. For five B type monomers it is not possible to predict the ground state analytically. One could argue that the ground state should look like two tetrahedrons sharing the same base (see Fig. 4.6(d)). This formation has an energy of -2.25, but in various ELP simulations we found configurations with energies around -2.273. This configuration was a non-regular double tetrahedron. In the literature, such aggregates are known as Lennard-Jones crystals or Lennard-Jones clusters [55].

4.2.1 Polymers in Solvents

The next step was to investigate some small polymers in solvent. The Seq. 8.1 ($AB_2A_2B_2A$) was first studied without solvent. The configuration with the lowest energy found in a vacuum can be seen in Fig. 4.7(a). The energy was approximately -2.6. The configuration is symmetric because of the symmetry in the sequence of monomers. For this reason, a similar configuration exists where the middle angle is twisted, which has also the same energy. The heat capacity (see Fig. 4.7(b)) was obtained by a multicanonical simulation with 100 recursions and 10^5 sweep sequences each and a final simulation with 10^7 sweep sequences.



(a) (b)

Figure 4.7: (a) The lowest energy found for Seq. 8.1 ($AB_2A_2B_2A$) was around -2.6. The configuration is symmetric and a similar configuration with the same energy, but an other symmetry exist. (b) The inner energy and specific heat was obtained by a multicanonical simulation.

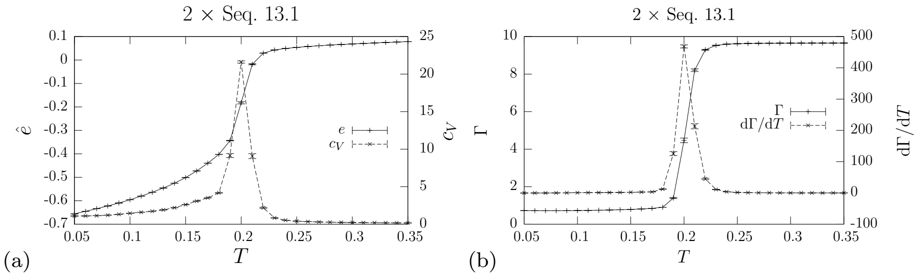
The amount of “water” monomers mixed together with polymer Seq. 8.1 was also an interesting question, we tried 4, 8, 12 and 16 “water” monomers. This is obviously insufficient to speak of a realistic simulation of a polymer in solvent, but a bigger amount of “water” monomers was not possible because of the effort of computer time. Even a simulation with 16 “water” monomers at a fixed temperature takes around 12 times longer than a simulation without “water” monomers which results from the fact that every water molecule has to be updated.

But all simulations lead to problems with fixed box size. A small box size acts like high pressure, but even more complicated, a big box size leads to vacuum comparable state what causes a crystallization of the “water” monomers. That is why in the case of mixed simulations we always found a combination of “water” crystals with tetrahedron structures and the ground state of Seq. 8.1 as global system ground.

For many particle systems it is known [13] that simulations at constant pressure are much more effective than simulations with fixed volume like done here. That is why one should go over from the $\{N, V, T\}$ ensemble to the $\{N, p, T\}$ ensemble. This is technically harder and not implemented in the program until now. Fortunately for two up to four particles the simulation with fixed volume works also, if the box size is chosen well.

4.3 Two Polymer Chains

The investigation of two chains is the simplest non-trivial aggregation of two smaller systems. We start with two heteropolymers with the Seq. 13.1 to come



(a) (b)

Figure 4.8: (a) The specific inner energy and heat capacity of 2 × Seq. 13.1. (b) The mean of the aggregation parameter Γ and its fluctuation of 2 × Seq. 13.1. The specific heat as well as the derivation dΓ/dT of the aggregation parameter Γ have a peak at $T = 0.2$, which can be explained by the transition from the fragmented to the aggregated phase.

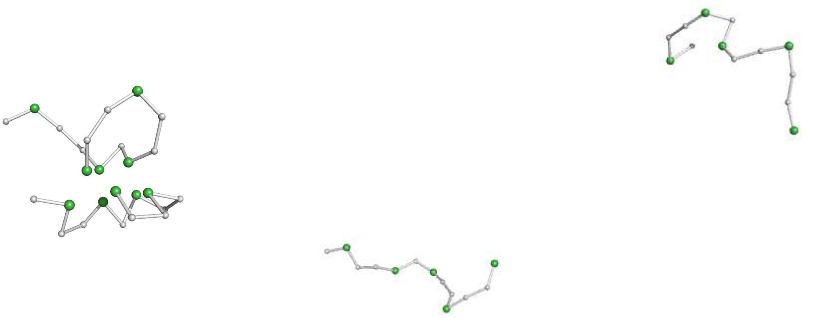
into the region of polymers with length 20 monomers. The Seq. 13.1 is just one possible example and as arbitrary as any other sequence. Later we also study homopolymers with length 13, which turned out to be technically a little bit demanding.

4.3.1 General Facts

There are some general facts about aggregation included in this model. It is obvious that the ground state is in the aggregated phase, because it is not possible to form as much low energetic A-A and B-B contacts in the fragmented phase as in the aggregated phase. But there are also high-energy configurations in the aggregated phase for too small distances of AA and BB pairs. The barrier between aggregated and fragmented phase comes from the fact that the configurations in both phases have completely different geometries. In the fragmented phase configurations with separated hydrophobic cores are dominating and in the aggregated a joined core is formed. For homopolymers this works as well because of the structure with outer helices at the single chains. Two phases with a barrier in between will behave like a first order phase transition.

4.3.2 Dual Seq. 13.1

We performed a multicanonical simulation with around 180 recursions with 10^5 sweep sequences and a final run with 10^8 sweep sequences. Mean energy and specific heat are shown in Fig. 4.8. We have done several cross-checks ($\approx$ 50 simulations) with other multicanonical simulations starting from different seeds and multicanonical weights. We also compared the results with parallel tempering simulations over various temperature intervals and canonical simulations at different temperatures. In the heat capacity one pronounced peak can



(a) (b)

Figure 4.9: (a) A configuration in the aggregated phase with some random coils, but joint hydrophobic core. (b) A configuration from the fragmented phase which shows the typical random coils. The single chains do not interact which each other due to the large distance. The scale of the picture is changed due to a better view.

be observed which is four times larger than the one for the single chain (see Fig. 4.1(b)). This is the second point in the finite-size scaling analysis, it seems like the scaling law goes with the square of the number of polymer chains. The surprising thing is that there is only one peak. One would expect two peaks due to the second transition from random coils to folded phase. It seems that they coincide with each other.

The plot of the aggregation parameter Γ (see Fig. 4.8(b)) shows that the expected transition from the fragmented to aggregated phase happened at a temperature around 0.2. In the lower temperature regions (smaller than 0.2) the aggregated configurations (see Fig. 4.9(a)) are dominating and for the high-temperature regions the fragmented configurations (see Fig. 4.9(b)).

Also it seems that the average of Γ goes against some limit (≈ 9.8). So the distance between the centers of mass (see Eqn. (3.25)) goes against 19.6, which is nearly $L_{\mathrm{Box}}/2$. This is also the mean distance of two free particles in a periodic box (see Sect. E.4). The deviation comes from the fact that the chains are not ideal particles without a volume expansion. So one can say that for high temperatures both polymers behave like they were separated.

The phases are well separated which can be seen in the multicanonical distribution (see Fig. 4.10(a)). There is a small dip between them. The multicanonical distribution could be reweighted to a canonical distribution, but as the multicanonical weights range over 100 orders of magnitude (see Fig. 4.10(b)) this would not be that meaningful. Because of this dip, which marks the barrier,

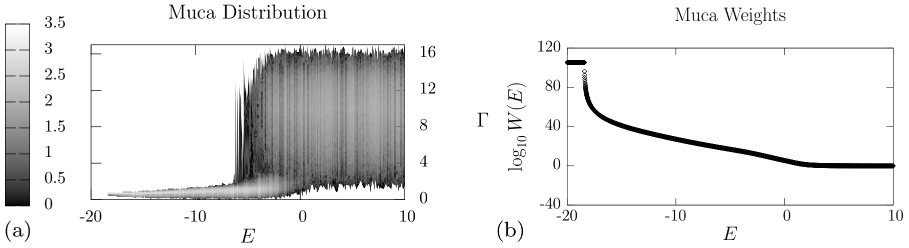


Figure 4.10: (a) Multicanonical distribution $P_{\mathrm{muca}}(E,\Gamma)$ (see Eqn. (2.51)). The transition from the fragmented to the aggregated phase can be seen clearly between the energy of $-5\ldots0$. (b) Multicanonical weights $W^{\mathrm{muca}}(E)$ reach over 100 orders of magnitude.

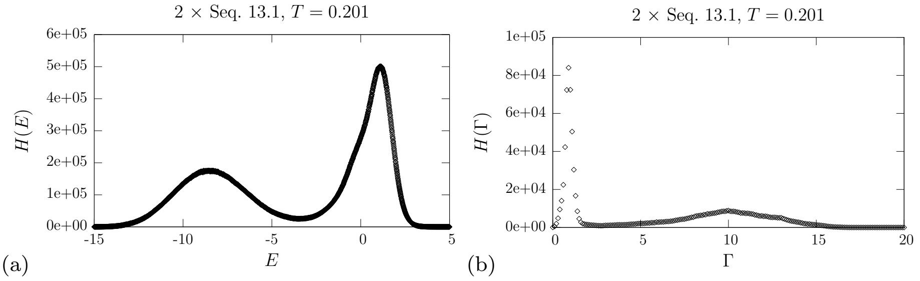


Figure 4.11: (a) The energy distribution close to the transition point. (b) The distribution of the aggregation parameter close to the transition point. The characteristic behavior of a first order like transition can be seen in the distribution of the energies, but also in the distribution of the aggregation parameter Γ.

the transition is first order like. The variations of the aggregation parameter in both phases are very different. In the aggregated phase, the variation is of the order one and in the fragmented phase the variation is of the order ten, because the peak for low values of Γ is much higher than the peak at larger ones (see Fig. 4.11(b)) for equilibrium between both phases.

This first-order-like behavior can also be seen in the histogram of a canonical simulation at this transition temperature ($T \approx 0.201$). This is observable as the well-known double-peak structure which can be seen in Fig. 4.11(a). Also the distribution of Γ shows such a double peak structure (see Fig. 4.11(b)). Example configurations from both peaks can be seen in Fig. 4.9.

One could expect two ways of getting from the aggregated to the fragmented phase. The first possibility is that every single chain forms a hydrophobic core and then they dock to each other and form a global hydrophobic core without totally unfolding again. The second way is to first unfold, form a global hydrophobic core and then fold again. So the question is if the chains have to unfold to aggregate.

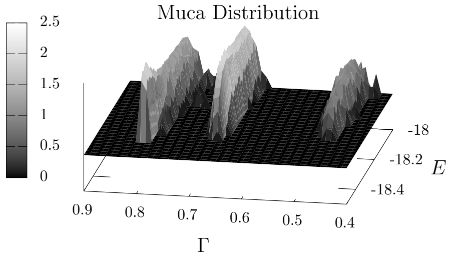


Figure 4.12: The multicanonical distribution of the aggregation parameter Γ at low energies for twice Seq. 13.1 shows several peaks. The peak at $\Gamma = 0.8$ are the half sphere configurations and at $\Gamma = 0.47$ are the entangled configurations. Intermediate configurations exist, but a clear classification is not possible.

We observed the latter possibility for several aggregation events. The reason is probably that in the aggregated phase, which is more closely packed than the fragmented phase, large structural changes like forming a global hydrophobic core out of two single cores is much less probable. In contrast, it is always possible to form a hydrophobic core at higher temperature because of the random coils which can easily interact.

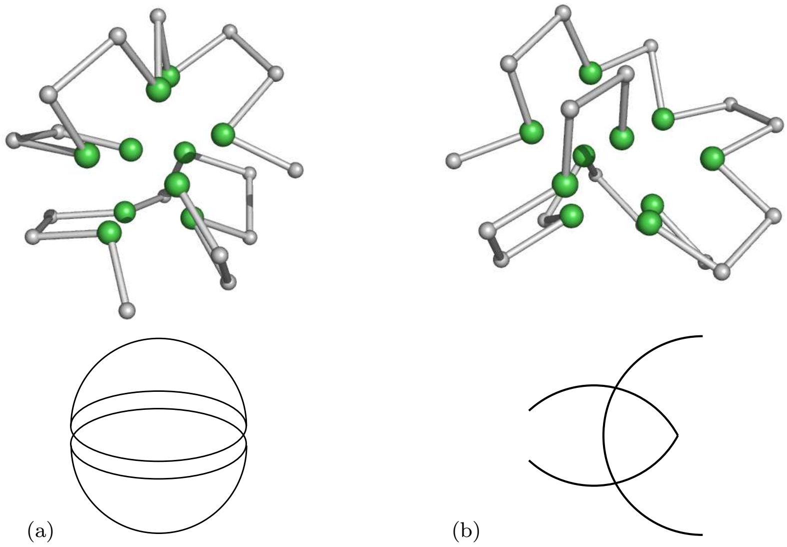
It has to be mentioned that in our simulations this process is *no* real dynamic, because of the Monte Carlo simulation. However, the way of aggregation shows that rigid-body docking is more suppressed than flexible docking.

It should be noted that the individual conformations in the aggregate strongly differ from the single-polymer ground states ($E_{\mathrm{min}}^{\mathrm{total}} \approx -4.967$). Their respective energies in the aggregate are different and much larger ($E_{1,\mathrm{min}}^{\mathrm{total}} \approx -3.197$ and $E_{2,\mathrm{min}}^{\mathrm{total}} \approx -3.798$). The strongest contribution is due to the interaction between the heteropolymers ($E_{\mathrm{interact,min}} \approx -11.412$).

In the area of the lowest energies (≈ -18.3) we found two different kinds of structures. The first kind has values of Γ around 0.8 and the second has values around 0.47 (see Fig. 4.12). The configurations with higher values consist of two half spheres (see Fig. 4.13(a)), the others look like entangled structures (see Fig. 4.13(b)). Obviously the entangled structure has a shorter distance between the single chain centers of mass than the configuration in the half sphere structure. But the energy of the half sphere configurations is a little smaller (≈ 0.1) than the others. There are also intermediate states between these two extrema, these arise from the interaction of the random coils of configurations at higher energy, but a classification of those is not possible. A more detailed picture of the lowest energy area from Fig. 4.10 can be seen in Fig. 4.12, where also a peak of an intermediate structure can be observed.

4.3.3 Dependence on the Size of the Periodic Box

To be certain that the aggregation effects and thermodynamics do not depend so much on the box we tried different box sizes. Normally, we set $L_{\mathrm{box}} = 40$, but we also used half and twice of this length. With doubling the length the



(a) (b)

Figure 4.13: (a) Configuration with the structure of two half spheres having the energy $E = -18.34$ and $\Gamma = 0.8$. (b) Configuration with the entangled structure having the energy $E = -18.27$ and $\Gamma = 0.47$.

volume gets bigger by a factor 8, so the volume of $L_{\text{box}} = 80$ was 64 times bigger than the box of $L_{\text{box}} = 20$.

First, the thermodynamics was checked (see Fig. 4.14). The thermodynamics in a single phase was equal for all sizes, but in the region of the phase transition ($T \approx 0.2$) small deviations appear. These come mainly from the very different probabilities (see Fig. 4.15) of the chains to find each other. The difference between the minimum and maximum in the canonical histogram (see Fig. 4.11(a)) is much bigger[2]. So we see that the behavior of a system which does not change the phase is the same for all box sizes. But for a system which does a transition between the phases, the thermodynamics changes a bit for the different box sizes. A small variation of the box size will not change the thermodynamics very much, but in our studies we increased the volume by a factor 8 and 64.

The finite-size scaling of the deviation of the aggregation parameter works quite well, but 3 points is too small to obtain an accurate scaling law. Obviously for all 3 box sizes the limit of the normalized aggregation parameter Γ/L_{box} for higher temperatures is 0.25, which is the proposed value as it will be calculated in Sect. E.4. The small deviation from this value results from the fact that the polymers are real particles and have a dilatation. The deviation is of the order of the radius of gyration of single chains, which shortens the accessible area.

---

[2] Often the logarithm of the ratio of the probabilities at the minimum and the maximum, $\ln(P_{\text{min}}/P_{\text{max}})$ is called *interface tension*, which has a special scaling behavior and was one reason for the invention of multicanonical simulation [14].

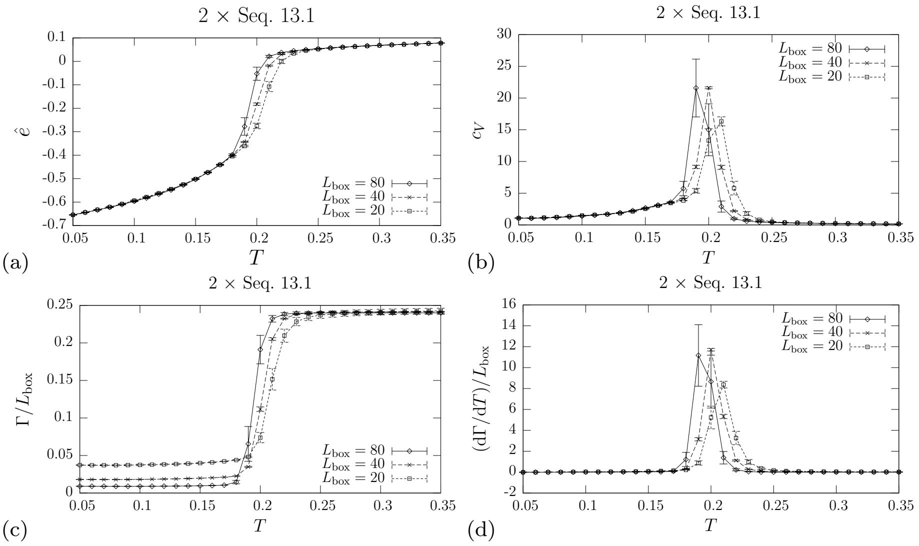


Figure 4.14: Thermodynamics of twice Seq. 13.1 for different box sizes. (a) Mean value of energy normalized on the number of monomers ($e = E/N_G$) for twice Seq. 13.1. (b) Specific heat for twice Seq. 13.1. (c) Aggregation parameter Γ for twice Seq. 13.1 normalized on the box length L_{box}. (d) The derivation of the aggregation parameter for twice Seq. 13.1 normalized on the box length L_{box}. In all plots small variations can be seen around temperature $T = 0.2$, which is also the transition temperature from fragmented to aggregated phase.

We also compared the multicanonical distribution of energy and aggregation parameter, which can be seen in Fig. 4.15. It is obvious that the two phases still have a sink between each other, but the phases are not equally distributed anymore. For smaller box lengths the aggregated phase is more dominating, also for a larger box length the fragmented phase is dominating. This is a hint to the fact that the only thing which is changing for varying box size is the behavior close to the phase transition, because the probabilities for this transition are changing. The smallest box size also leads to problems with the stretched configuration. It would be possible to put two stretched configurations beside each other, but not behind each other. So an unfolding of the complete system is not possible in every direction. We have to mention that these stretched configurations play a less important role for the thermodynamic properties.

4.3.4 Homopolymers

To understand the essentials behind the process of aggregation we also investigated aggregation of homopolymers of the same length. Because of the missing of hydrophobic monomers we expected a different type of ground state devoid of a hydrophobic core, but with a maximum of optimal distances between the

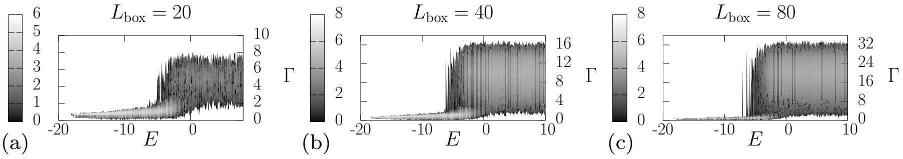


Figure 4.15: Multicanonical distribution of twice Seq. 13.1 for different box sizes. (a) $L_{\mathrm{box}} = 20$ shows a small fragmented phase (high values of Γ). (b) $L_{\mathrm{box}} = 40$ shows a equipartition between the fragmented and aggregated phase. (c) $L_{\mathrm{box}} = 80$ shows a small aggregated phase (low values of Γ).

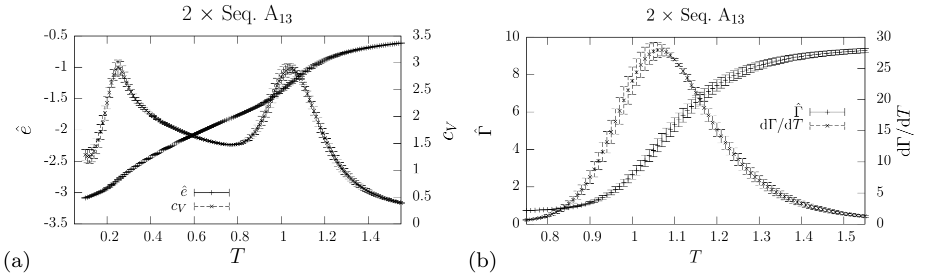


Figure 4.16: (a) The specific inner energy and heat capacity of twice Seq. A_{13}. (b) The mean of the aggregation parameter Γ and its fluctuation of 2 x sequence a_{13}. The peak in the heat capacity at $T \approx 1$ comes from the aggregation transition. The other peak at $T \approx 0.2$ should mark the collapse to the ground state.

monomers, as it was also the case for solvent simulations (see Sect. 4.2).

We performed a multicanonical simulation with 115 recursions of 10^5 sweep sequences and a final simulation with 10^7 sweep sequences to obtain the thermodynamics shown in Fig. 4.16. The results were also cross-checked with parallel tempering and multicanonical replica exchange, which show no significant deviation.

The specific heat (see Fig. 4.16(a)) has two peaks which show that the transition ($T \approx 1$) from fragmented to aggregation phase and the transition ($T \approx 0.2$) from random coil to folded phase does not happen at the same temperature. For the transition at high temperature the derivation of aggregation parameter also shows a peak (see. Fig. 4.16(b)), which clearly identifies this transition. For the first peak it is not that easy, but we think this is a transition to the folded phase due to the other kind of ground state (see Fig. 4.17).

As for heteropolymers the pathway to the ground state is similar, they aggregate before they collapse to a compact configuration. In contrast to the heteropolymer ground states which have a hydrophobic core, these configurations have a “well”-like structure [54] like the singe chains, with a stretched part inside and a surrounding helix (see Fig. 4.17). And the fold to this ground state

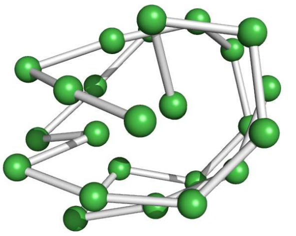


Figure 4.17: The configuration of twice Seq. A_{13} with the lowest energy found in a multicanonical simulation with 10^7 sweep sequences. This shows some "well" like structure with a stretched part inside and a surrounding helix.

is a large structural change in comparison to the random coil phase.

4.3.5 Microcanonical Interpretation

In this section, we analyze the aggregation transition from the microcanonical perspective [56]. The multicanonical simulations give a sufficiently nice estimator for the density of states $\Omega(E)$ which is also the microcanonical partition function supplying the key to all microcanonical information.

The microcanonical entropy is defined as[3]:

$$S(E) = \mathrm{k_B} \ln \Omega(E) \ , \tag{4.5}$$

where the Boltzmann constant was set to one due to Sect. 3.1. Due to the first fundamental law of thermodynamics

$$\mathrm{d}U = T\mathrm{d}S - p\mathrm{d}V \tag{4.6}$$

the first derivation of the entropy is the inverse microcanonical temperature:

$$T^{-1}(E) = \left.\frac{\partial S(E)}{\partial E}\right|_{V,N} \tag{4.7}$$

and the inverse microcanonical heat capacity:

$$\begin{aligned} c_V^{-1}(E) &= \left.\frac{\partial T(E)}{\partial E}\right|_{V,N} && (4.8)\\ &= -T^2(E)\left.\frac{\partial T^{-1}(E)}{\partial E}\right|_{V,N} && (4.9)\\ &= -\frac{\left.\frac{\partial^2 S(E)}{\partial E^2}\right|_{V,N}}{\left(\left.\frac{\partial S(E)}{\partial E}\right|_{V,N}\right)^2} \ . && (4.10)\end{aligned}$$

---

[3] This is only true in the *thermodynamic limit*. The Hertz definition of the entropy is $\mathcal{S} = \mathrm{k_B} \ln \Gamma(E)$, where $\Gamma(E) = \int_{E<E'} \mathrm{d}E\, g(E')$ is the phase-space volume. In the thermodynamic limit both definitions become equal, but for finite systems the two entropies are not necessarily identical [57]. However, for our finite system we have checked out both definitions and did not find noticeable deviations in the transition region.

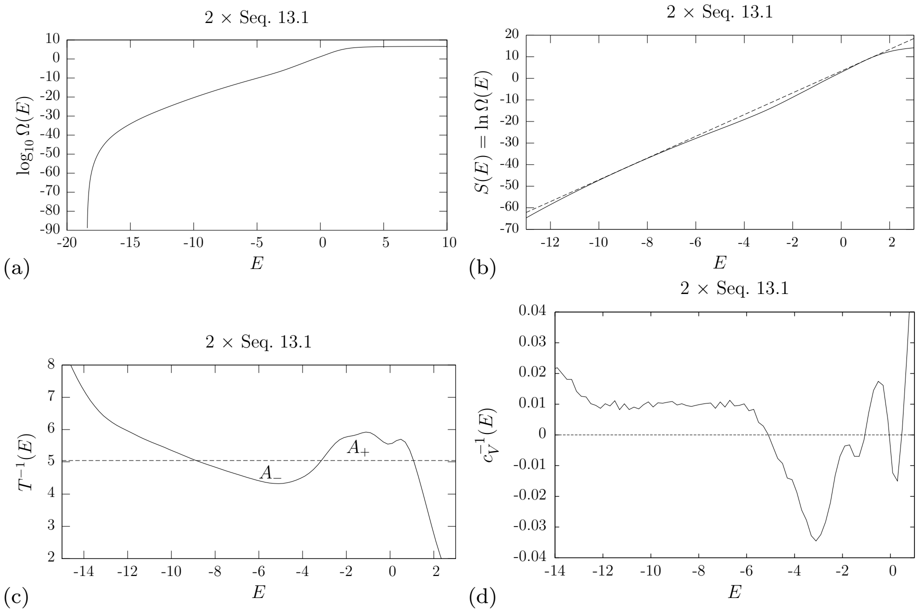


Figure 4.18: (a) The density of states $\Omega(E)$ ranges over around 100 order of magnitude. (b) The microcanonical entropy shows a small dip in the energy range $-8.85 \ldots 1.05$ (also see Footnote 3). The tangent has the slope of the inverse aggregation temperature $T_{\mathrm{agg}}^{-1} \approx 5.04$. (c) The inverse microcanonical temperature $T^{-1}(E)$ shows the characteristic backbending effect. The dashed line is the inverse aggregation temperature $T_{\mathrm{agg}}^{-1}(E)$, which is defined by the area criterion $A_- = A_+$. (d) The inverse microcanonical specific heat has four zero points, which are singularities in the specific heat. The limit for high and low energies is infinity, which means the specific heat equals zero.

In the thermodynamic limit one would expect that $S(E)$ is a concave function, that is why $T^{-1}(E)$ is a monotonically decreasing function and $\left.\frac{\partial^2 S(E)}{\partial E^2}\right|_{V,N}$ is negative. So the microcanonical specific heat is always positive, which is a well-known fact.

But there are several finite systems which show opposite behavior in certain energy regions. This phenomenon has long been known from astrophysical systems [58] and spin systems on finite lattices [59, 60], but also from experiments with sodium clusters [61].

We also observed this convex behavior of the density of states for a system with two polymers with Seq. 13.1 [56]. The density of states was an outcome of a multicanonical simulation with 180 recursions and 10^8 sweep sequence to accumulate reliable statistics. The results are shown in Fig. 4.18(a).

The most interesting region

$$E_{\mathrm{agg}} \approx -8.85 \leq E \leq 1.05 \approx E_{\mathrm{frag}} \tag{4.11}$$

can be found in Fig. 4.18(b), where also the concave hull is shown. This hull

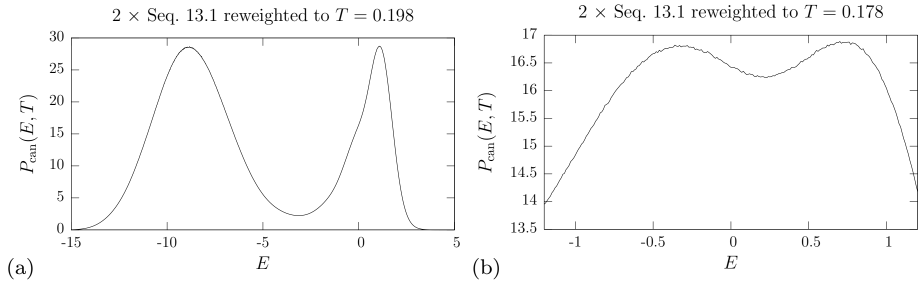


Figure 4.19: (a) The histogram of twice Seq. 13.1 at $T = 0.198$, which is the aggregation temperature obtained by the Gibbs construction (see Fig. 4.18(b)). Obviously the two peaks have the same height. (b) The histogram at $T = 0.178$, which is the second transition temperature from the fragmented to meta stable phase (see the dip in A_+ at Fig. 4.18(c)). The peak at lower energies from (a) can not be seen in this plot.

is called *Gibbs construction* and has the slope of the inverse aggregation temperature $T_{\mathrm{agg}}^{-1} \approx 5.043$. The two intersection points of the Gibbs line and the entropy are called E_{agg} and E_{frag}. The interval

$$\Delta Q = E_{\mathrm{frag}} - E_{\mathrm{agg}} - T_{\mathrm{agg}} \left(S(E_{\mathrm{frag}}) - S(E_{\mathrm{agg}}) \right) \approx 9.90 \tag{4.12}$$

is the latent heat required to release inter-chain contacts at the aggregation temperature T_{agg}. The first derivative is the inverse temperature shown in Fig. 4.18(c) again together with T_{agg}^{-1}, which is not just a constant line. The inverse temperature function gives the definition of T_{agg}^{-1} by an area criterion: T_{agg}^{-1} is the temperature where A_- is equal to A_+. This is the so called *Maxwell line*. We determined T_{agg} to be around 0.198, which is near to the value we found with the canonical calculations (see Sect. 4.3.2 on page 44). The inverse temperature shows the backbending effect, which is typical for a first-order-like phase transition. Normally, this effect vanishes in the thermodynamic limit, but because of the lack of a limit for heteropolymers it does not vanish here. This special behavior of the inverse temperature leads to a negative specific heat as a function of energy (see Fig. 4.18(d)), which is a little bit "exotic".

But the more important point of the backbending effect is the loss of the temperature as an external control parameter. The interpretation of the canonical formalism in that region is not generic. The canonical formalism makes sense for a monotonic increasing inverse microcanonical temperature, because for this case mapping from temperature to energy and backwards is unique. In temperature regions of the aggregation this is violated. In simple words it means that in certain temperature regions the system gains energy by cooling. Also the obtained aggregation temperature T_{agg} can be used to find a reweighted canonical histogram with equal peak heights (see Fig. 4.19(a)).

Obviously in Fig. 4.18(c) there is a second transition in the area A_+ (close to $E \approx -0.32$), which has a transition temperature of $T_{\mathrm{agg},2} \approx 0.178$, which can easily be overlooked in a canonical calculation.

For energies close to E_{frag} the system is in a fragmented state, and the population of aggregated polymers in this energy region is extremely small. The situation is different for energies $E < 0.22$ (see Fig. 4.19(b)), where weakly stable aggregated conformations and polymer fragments coexist. Only for much smaller energies ($E < E_{\mathrm{agg}}$), compact aggregates dominate. Having this in mind, the transition can also be understood from the canonical view. For temperatures below $T_{\mathrm{agg},2} \approx 0.178$, stable aggregates (solids) of low energies ($E < E_{\mathrm{agg}}$) dominate. Approaching $T_{\mathrm{agg},2}$, the system enters the subphase of coexisting unstable pre-molten aggregates of comparatively high energies ($E \approx -0.32$) and already fragmented polymers.

Compared with the distribution at the aggregation transition in Fig. 4.19(a), the ratio between maximum and minimum is small and, therefore, also the interface tension. In consequence, the transition between the solid and the pre-molten, unstable aggregates is, compared to the aggregation transition, negligibly weak. It should be noted that the peak of the aggregation phase, not shown in Fig. 4.19(b), is much more pronounced than the peaks of the pre-molten aggregates near $E \approx -0.32$ and the fragments close to $E \approx 0.73$.

We have also observed these subphases in the studies of the pathways of the aggregation event. As mentioned in Sect. 4.3.2 on page 45 the single chains first unfold completely before aggregating.

4.4 Three and More Chains

The studies of three and more chains are a little shortened in this thesis for several reasons. First we have not done as much cross-checks as for two chains. Secondly the simulations for more chains need a much bigger effort of computer time. The number of single updates scales with the square of number of chains (see Eqn. (3.6)) and the number of distances in the Lennard-Jones energy (see Eqn. (2.5)) also scales with the square of number of chains. But the results in the following section can be trusted anyway.

4.4.1 Triple Sequence 13.1

We simulated a three chain system in the same way as for two chains, we only changed the sequence of updates (see Table 3.2), but nothing else.

With the help of a multicanonical simulation we found some similar behavior in the thermodynamics (see Fig.4.20) as for two chains (see Fig. 4.8). This

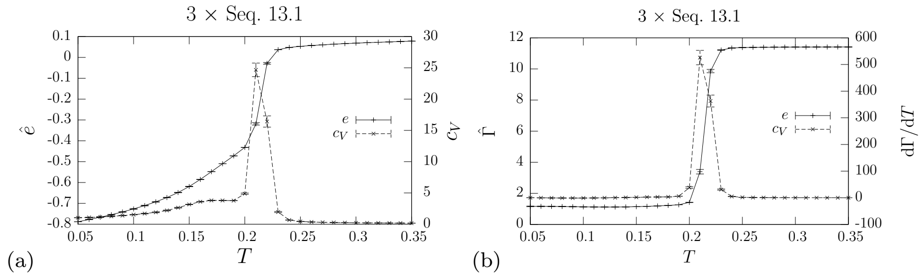


Figure 4.20: (a) The specific inner energy and heat capacity of 3 × Seq. 13.1. (b) The mean of the aggregation parameter Γ and its fluctuation of 3 × Seq. 13.1. The specific heat as well as the derivation $\mathrm{d}\Gamma/\mathrm{d}T$ of the aggregation parameter Γ have a peak at $T = 0.2$, which can be explained by the transition from the fragmented to the aggregated phase. The similarity to the thermodynamics of 2 × Seq. 13.1 (see Fig.4.8) is unmistakable.

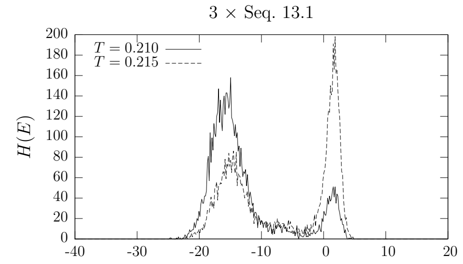


Figure 4.21: The canonical histogram at $T = 0.210$ and $T = 0.215$. At higher temperature the fragmented phase (right peak) is dominating. The right wing of the left peak shows a transition to the "2 of 3" aggregation. The histogram of Γ looks like Fig. 4.11(b).

is confirmed by the fact that the thermodynamics strongly depends on the sequence of monomers. But if the sequences of monomers are comparable, the thermodynamics should be as well.

4.4.2 Type of Aggregation

The question is, how is the way of many body aggregation in comparison to a system of two chains? For the two chains this question was senseless, but for at least three chains it would be interesting if two chains will aggregate first and then the third accrue or if all three aggregate in one step. For that reason one can expect one or two peaks in the derivation of the aggregation parameter, but there is only one (see Fig. 4.20(b)). This is the first hint that there is only one transition from the fragmented to the aggregated phase, but as mentioned above (see Sect. 4.3.5) sometimes there are small dips which could easily be overlooked.

However, from Fig. 4.20 we extracted the transition temperature to be around $T \approx 0.2$. A canonical simulation at the transition region confirmed this hint. The histogram (see Fig. 4.21) shows a straight transition from fragmented to aggregated phase. A temperature change from $T = 0.210$ to $T = 0.215$ let the

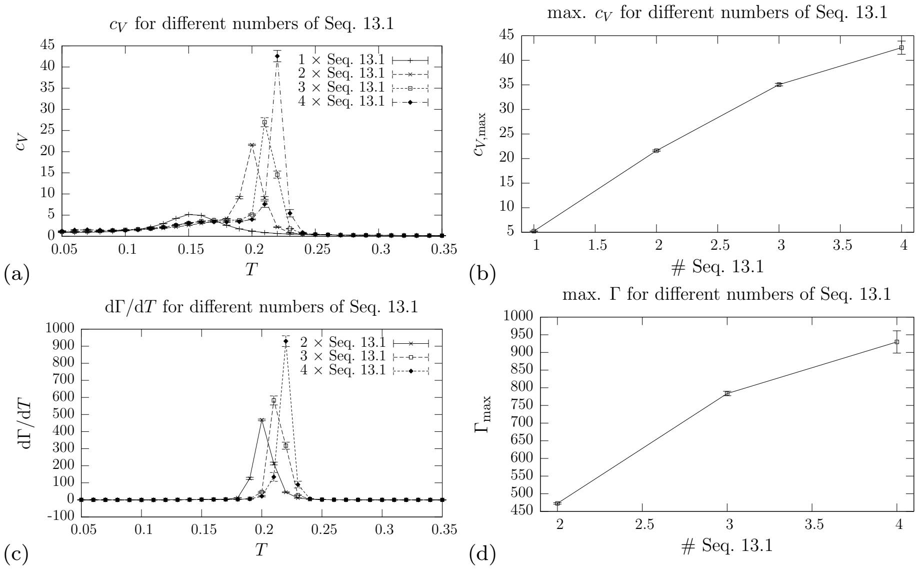


Figure 4.22: (a) Specific heats for different number of chains of Seq. 13.1. (b) The maxima of the specific heat show an increasing behavior with the number of chains, but a clear law is not visible. (c) The derivation of the aggregation parameter dΓ/dT for different amounts of Seq. 13.1. (d) The maxima of dΓ/dT also show an increasing behavior with the number of chains, but three points are to few to obtain a scaling law.

dominating phase go from aggregated to fragmented phase. The intermediate “2 of 3” phase exists, but never alone. This intermediate phase can be seen in the left wing of the right peak of Fig. 4.21. This concludes in the fact that the aggregation of three chains goes through an intermediate phase, but this phase is never dominating.

4.4.3 Scaling

Obviously the height of the peaks of the heat capacity (see Fig. 4.22(a)) and derivation of the aggregation parameter (see Fig. 4.22(c)) is increasing with the number of chains, so we tried to obtain a scaling law (see Fig. 4.22(b) and Fig. 4.22(d)) for which we made multicanonical simulations of comparable lengths with one, two, three and four times Seq. 13.1. Because of the great effort of computer time for four chains it was not possible to obtain satisfying data for more chains.

But the number of points is much too small to identify a scaling law correctly. However, any obtained scaling law cannot be trusted, because reasonable finite-size scaling cannot be performed with small numbers of chains[4].

---

[4] The common scaling law is $c_{V,\mathrm{max}} \sim L^{\alpha/\nu}$, but for small L there is a correction term

4.4.4 Additional Studies

Additionally we studied homopolymer systems with 3 and 4 chains, but as mentioned above (see Sect. 4.3.4) the ground states always have a structure like a “well”. This structural difference between the random coil phase and the folded phase respectively the fragmented and aggregated phase leads to a bigger barrier than for the heteropolymers. This barrier is more a technical problem due to the updates. The acceptance ratio in the more compact phase drops nearly to zero. For the spherical update only the end monomers have a reasonable chance to be updated. This leads to higher autocorrelations in the low temperature phase.

---

like $c_{V,\max} = c_0 + aL^{\alpha/\nu}(1 + bL^{-\omega})$ [62, 63], which can not be neglected. Anyway, there are too many unknown parameters to perform a nice fitting.

Chapter 5

Summary

In this thesis we studied a simplified mesoscopic aggregation model for hydrophobic-polar peptides. In that model, a protein is just a chain of monomers of type A or B, where A are hydrophobic monomers and B are polar or hydrophilic monomers.

However, even for this simple model the complexity is high enough so that common canonical Monte Carlo gave no satisfying results, that is why we used more sophisticated generalized ensemble methods. The simplest one was the multicanonical simulation technique, which changes the sampled distribution to a flat one. Secondly, we used parallel tempering, which simulates multiple temperature runs in parallel to shorten the autocorrelation time. Thirdly, the not so common multicanonical replica exchange was used, which is a combination of the both methods mentioned above.

The single-chain simulations with various methods showed that the physical quantities of the chains strongly depend on the sequence of the chains. The mean energy is always monotonically increasing with the temperature, which is normal. In contrast, the specific heat showed, depending on the sequence, one or two peaks of different height, which are caused by a transition between three "pseudo" phases. The phase at high temperatures is dominated by stretched configurations with random coils. With decreasing temperature the system enters the globular phase. The third phase, at very low temperatures, is the hydrophobic core region, where a core of hydrophobic monomers has formed. The transitions can fall together as it happened for Seq. 13.1, which is the central sequence studied in this work. The ground state was observed to have a hydrophobic core and a hydrophilic surrounding.

Homopolymers showed a totally different kind of ground state conformations due to the missing interaction between hydrophobic and hydrophilic monomers. They exhibited a "well"-like structure with a stretched part in the middle and the onset of an outer helix.

As first sample for the aggregation study we took twice the Seq. 13.1 in order to have a system with around 20 monomers, which can be simulated in acceptable CPU times. For this system we found an aggregation transition in the region of $T \approx 0.2$. The ground state is observed to lie in the aggregation phase and to have a global hydrophobic core, like the single-chain ground state.

We first introduced periodic boundary conditions, which were also tested to not contain any systematic errors in single-chain simulations. This boundary condition was chosen to increase the probability of the single chains to find each other. We have not used fixed boxes to avoid effects resulting from the hard walls. The dependence on the size of the box was also studied and it turned out that there are only small variations in the region of the aggregation due to the different probabilities of the single chains to find each other.

The aggregation transition is indicated by a peak in the specific heat. We defined the aggregation parameter as the radius of gyration of the centers of mass of the single chains. The centers of mass are measured with respect to the first monomer of the chain due to the periodic boundary condition. This parameter behaved like a common order parameter, its derivative showed a peak at the same temperature ($T \approx 0.2$) as the specific heat. Higher values are connected to the fragmented phase and higher energies. And low values belong to the aggregated phase.

The transition itself is first-order like which can be seen from the distribution of energies and the distribution of the aggregation parameter. These distributions showed the well-known double-peak structure. Due to only one peak in the specific heat the transition from the random coils to the hydrophobic core region and the aggregation transition nearly fell together.

A microcanonical analysis was used to identify the transition temperature. Surprisingly, the analyses showed a second transition at a little higher temperature. This was identified to be a transition from random coil to a metastable phase which does not have a global hydrophobic core. This transition is so weak that it would have been easily overlooked in canonical analyses. This also led to the detection that the single chains have to unfold first before aggregation, which was also observed in canonical analyses.

In contrast to heteropolymers, the transition from random coil to hydrophobic core region and the aggregation transition for homopolymers did not fall together for this category of polymers. The aggregation transition is at a much higher temperatures than the other transition, which came from the fact that the structural difference between the ground state and the random coil phase was much bigger for homopolymers than for heteropolymers.

We also studied three and more heteropolymer chains which showed an equal aggregation transition. We found that the transition takes place in one step, the phase with "2 of 3" chains aggregated is never dominating, thus we came to the conclusion that there is only one transition from fragmented to aggregated phase. In a finite-size-scaling analysis we showed that the size of the peaks grows with the system size, but a scaling law could not be obtained due to the small number of chains.

Simulating three and more homopolymers was more demanding, because of the much bigger structural difference to the ground state, which also has a "well"-like structure. However, the aggregation transition did not fall together with the transition between random coil and compact phase.

There are still a lot of interesting open questions concerning these topics for a future work. The transitions of single chains are not completely understood, especially the dependence on the sequence of monomers. A small-scale microcanonical analysis also showed a first-order like transition.

Most parts of the aggregation process are well understood, but aspects concerning the aggregation pathway have to be researched in more details. Also the study of bigger systems and systems with more sequences would be interesting.

Improved methods with shorter autocorrelation time have to be used to study homopolymers to obtain sufficient statistics. Because the "well"-like structure will stay for even bigger systems this will make the simulations more and more difficult.

Finally, a study of more than four chains would be desirable to obtain the scaling law more accurately and to understand the process of forming a hydrophobic core.

Appendix A

Details about the Updates

A.1 Spherical Update

The old vector $\vec{r}$ is given by:

$$\vec{r} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = r\vec{e}_r \ , \tag{A.1}$$

where

$$\vec{e}_r = \frac{1}{r}\begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{A.2}$$

and r is the spherical radius

$$r = \sqrt{x^2 + y^2 + z^2} \ . \tag{A.3}$$

Perpendicular to this vector one can choose two other vectors

$$\vec{e}_\varphi = \frac{1}{\sqrt{x^2 + y^2}}\begin{pmatrix} -y \\ x \\ 0 \end{pmatrix} \tag{A.4}$$

and

$$\vec{e}_\vartheta = \frac{\vec{e}_r \times \vec{e}_\varphi}{|\vec{e}_r \times \vec{e}_\varphi|} = \frac{1}{r\sqrt{x^2 + y^2}}\begin{pmatrix} -xz \\ -yz \\ (x^2 + y^2) \end{pmatrix} \ . \tag{A.5}$$

Because these three vectors make an orthonormal system one can write the new vector $\vec{r}\,'$ in terms of this:

$$\vec{r}\,' = a\vec{e}_r + b\vec{e}_\vartheta + c\vec{e}_\varphi \tag{A.6}$$

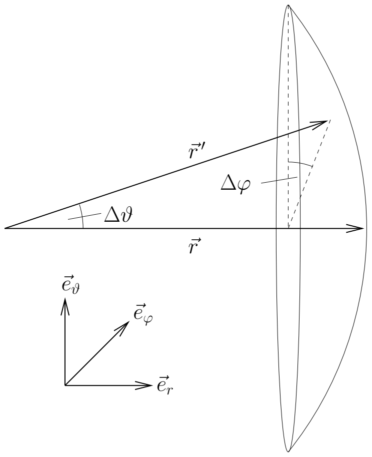

$\vec{r}'$
$\Delta\varphi$
$\Delta\vartheta$
$\vec{r}$
$\vec{e}_\vartheta$
$\vec{e}_\varphi$
$\vec{e}_r$

Figure A.1: The spherical sector with the local coordinate system $\vec{e}_r, \vec{e}_\varphi, \vec{e}_\vartheta$ and the angle changes $\Delta\vartheta$, $\Delta\varphi$.

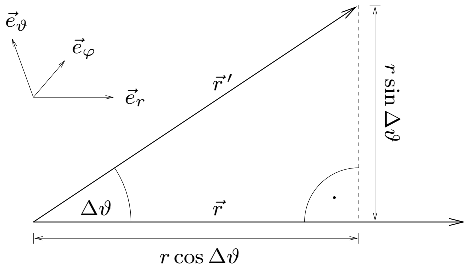

$\vec{e}_\vartheta$
$\vec{e}_\varphi$
$\vec{e}_r$
$\vec{r}'$
$r \sin\Delta\vartheta$
$\Delta\vartheta$
$\vec{r}$
$r \cos\Delta\vartheta$

Figure A.2: The length along the $\vec{e}_r$ direction is $r\cos\Delta\vartheta$ and the distance in the plane of $\vec{e}_\varphi$ and $\vec{e}_\vartheta$ is $r\sin\Delta\vartheta$.

with the constraint

$$|\vec{r}'| = |\vec{r}| = r \ . \tag{A.7}$$

As the angle between $\vec{r}$ and $\vec{r}'$ is defined to be $\Delta\vartheta$ (see Fig. 2.5)

$$\vec{r} \cdot \vec{r}' = r^2 \cos\Delta\vartheta = ar \tag{A.8}$$

one gets:

$$a = r\cos\Delta\vartheta \ . \tag{A.9}$$

For better understanding see Fig. A.1. The factors b and c can be distinguished with the help of Fig. A.2 and Fig. A.3.

So the new vector $\vec{r}'$ is given by:

$$\vec{r}' = r\cos\Delta\vartheta\vec{e}_r + r\sin\Delta\vartheta\sin\Delta\varphi\vec{e}_\varphi + r\sin\Delta\vartheta\cos\Delta\varphi\vec{e}_\vartheta \ . \tag{A.10}$$

The problem appears for $x = y = 0$ because then $\vec{e}_\vartheta$ and $\vec{e}_\varphi$ are not well

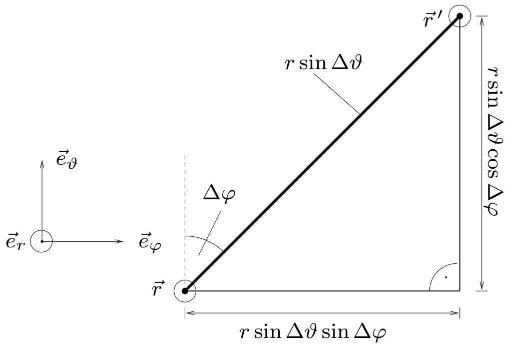

$\vec{r}'$
$r\sin\Delta\vartheta$
$r\sin\Delta\vartheta\cos\Delta\varphi$
$\vec{e}_\vartheta$
$\Delta\varphi$
$\vec{e}_r$
$\vec{e}_\varphi$
$\vec{r}$
$r\sin\Delta\vartheta\sin\Delta\varphi$

Figure A.3: The length along the $\vec{e}_\vartheta$ direction is $r\sin\Delta\vartheta\cos\Delta\varphi$ and along the $\vec{e}_\varphi$ direction $r\sin\Delta\vartheta\sin\Delta\varphi$.

defined. A solution is to use two other base vectors in a coordinate system:

$$\vec{e}_\varphi = \frac{1}{\sqrt{x^2+z^2}}\begin{pmatrix} z \\ 0 \\ -x \end{pmatrix} \tag{A.11}$$

and

$$\vec{e}_\vartheta = \frac{\vec{e}_r \times \vec{e}_\varphi}{|\vec{e}_r \times \vec{e}_\varphi|} = \frac{1}{r\sqrt{x^2+z^2}}\begin{pmatrix} -xy \\ \left(z^2+x^2\right) \\ -yz \end{pmatrix} \tag{A.12}$$

and the new vector $\vec{r}'$ is defined equally:

$$\vec{r}' = r\cos\Delta\vartheta\vec{e}_r + r\sin\Delta\vartheta\sin\Delta\varphi\vec{e}_\varphi + r\sin\Delta\vartheta\cos\Delta\varphi\vec{e}_\vartheta \ . \tag{A.13}$$

A.2 Rotation Update

The matrix R for rotation of angle α around a axis $\vec{v}$ with components

$$\vec{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \tag{A.14}$$

with

$$|\vec{v}| = 1 \tag{A.15}$$

is given by:

$$R_{ij} = v_i v_j + \cos\alpha(\delta_{ij} - v_i v_j) + \sin\alpha\epsilon_{ilj}v_l \tag{A.16}$$

which can also be written as:

$$R = \begin{pmatrix} \cos\alpha + v_1^2(1-\cos\alpha) & v_1v_2(1-\cos\alpha) - v_3\sin\alpha & v_1v_3(1-\cos\alpha) + v_2\sin\alpha \\ v_2v_1(1-\cos\alpha) + v_3\sin\alpha & \cos\alpha + v_2^2(1-\cos\alpha) & v_2v_3(1-\cos\alpha) - v_1\sin\alpha \\ v_3v_1(1-\cos\alpha) - v_2\sin\alpha & v_3v_2(1-\cos\alpha) + v_1\sin\alpha & \cos\alpha + v_3^2(1-\cos\alpha) \end{pmatrix} \tag{A.17}$$

To proof that this is a rotation matrix, one just has to show that:

$$R \cdot R^T = \mathbb{1} \tag{A.18}$$

where

$$R^T_{jk} = v_j v_k + \cos\alpha(\delta_{jk} - v_j v_k) + \sin\alpha(\epsilon_{kmj} v_m) \ . \tag{A.19}$$

This calculation is done in detail in Sect. E.1. In the case of the rotation update the axis $\vec{v}$ is

$$\vec{v} = \frac{\vec{r}_{j-1} - \vec{r}_{j+1}}{|\vec{r}_{j-1} - \vec{r}_{j+1}|} \ . \tag{A.20}$$

The next connection vector $\vec{r} = \vec{r}_j - \vec{r}_{j-1}$ is rotated by a random angle $\alpha \in [0, 2\pi)$ around this axis. So the new connection vector is given by:

$$\vec{r}_j' = R(\vec{v}, \alpha)\vec{r} + \vec{r}_{j-1} \ . \tag{A.21}$$

Appendix B

The Multicanonical Recursion

The idea is as follows: Normally, in the beginning, all multicanonical weights are set to 1. A "short" simulation is done, and the weights are updated by:

$$W^{\mathrm{muca}}(E) \propto 1/\Omega(E) \propto 1/H(E) \ . \tag{B.1}$$

This is useful because the histogram of a random simulation is naturally an estimator for the density of states which is inverse proportional to the multicanonical weights (Eqn. (2.51)).

Then another "short" simulation is done and from the gained histogram more accurate weights are calculated by:

$$W_{n+1}^{\mathrm{muca}}(E) = \frac{W_n^{\mathrm{muca}}(E)}{H_n(E)}. \tag{B.2}$$

This is the *"naive" recursion.* The better way is to add the old weights and the new weights according to their errors. This is the so called *multicanonical recursion.* To derive the formula a little more notation (motivated by W. Janke [33, 34]) has to be introduced.

The histogram of the nth recursion is called $H_n(E)$ and the weights are $W_n^{\mathrm{muca}}(E)$. Now we can introduce the ratio of the weights of the neighboring energy bins $R_n(E)$. (If the energy is not discrete, binning will be necessary anyway.)

$$R_n(E) = \frac{W_n^{\mathrm{muca}}(E+\Delta E)}{W_n^{\mathrm{muca}}(E)} \ . \tag{B.3}$$

After the nth recursion step a more accurate estimator for the multicanonical weight is given by:

$$\tilde{W}_{n+1}^{\mathrm{muca}}(E) = \frac{W_n^{\mathrm{muca}}(E)}{H_n(E)}. \tag{B.4}$$

The corresponding ratio of the neighbors is:

$$\tilde{R}_{n+1}(E) \overset{(B.3)}{=} \frac{\tilde{W}^{\mathrm{muca}}_{n+1}(E+\Delta E)}{\tilde{W}^{\mathrm{muca}}_{n+1}(E)} \overset{(B.4)}{=} \frac{W^{\mathrm{muca}}_{n}(E+\Delta E)}{W^{\mathrm{muca}}_{n}(E)} \frac{H_n(E)}{H_n(E+\Delta E)} \ . \tag{B.5}$$

After taking the logarithm this simplifies to:

$$\ln \tilde{R}_{n+1}(E) = \ln R_n(E) + \ln\left(\frac{H_n(E)}{H_n(E+\Delta E)}\right) \ . \tag{B.6}$$

The neighboring ratios for the $(n+1)$th recursion should be calculated by adding $\ln R_n(E)$ and $\ln \tilde{R}_{n+1}(E)$ weighted with their errors:

$$\ln R_{n+1}(E) = \kappa_n(E) \ln \tilde{R}_{n+1}(E) + \xi_n(E) \ln R_n(E) \ , \tag{B.7}$$

where $\kappa_n(E)$ is inverse proportional to the square of the error of $\ln \tilde{R}_{n+1}(E)$

$$\kappa_n(E) \propto 1/\epsilon^2_{\ln \tilde{R}_{n+1}(E)} = q_n(E) \tag{B.8}$$

and $\xi_n(E)$ is inverse proportional to the sum of the square of the errors of all previous estimators of $\ln R(E)$

$$\xi_n(E) \propto 1/\sum_{i=1}^{n-1} \epsilon^2_{\ln R_i(E)} = p_{n-1}(E) \ . \tag{B.9}$$

Eqn. (B.8) can also be written as

$$\kappa_n(E) = a q_n(E) \tag{B.10}$$

and Eqn. (B.9) as

$$\xi_n(E) = b p_{n-1}(E) \ . \tag{B.11}$$

For reasons of normalization

$$\kappa_n(E) + \xi_n(E) = 1 \tag{B.12}$$

has to be satisfied. A solution for these three equations is

$$a = b = \frac{1}{q_n + p_{n-1}} = \frac{1}{p_n} \ . \tag{B.13}$$

Everything is inserted in Eqn. (B.7):

$$\begin{aligned} \ln R_{n+1}(E) &\overset{(B.12)}{=} (1-\kappa_n(E)) \ln R_n(E) + \kappa_n(E) \ln \tilde{R}_{n+1}(E) && \text{(B.14)} \\ &\overset{(B.6)}{=} \ln R_n(E) + \kappa_n(E) \ln\left(\frac{H_n(E)}{H_n(E+\Delta E)}\right) && \text{(B.15)} \\ &\overset{(B.13)}{=} \ln R_n(E) + \frac{q_n(E)}{p_n(E)} \ln\left(\frac{H_n(E)}{H_n(E+\Delta E)}\right) \ . && \text{(B.16)} \end{aligned}$$

So the only thing missing is the estimation of the error $\epsilon_{\ln \tilde{R}_{i+1}(E)}$, because it is necessary for $q_n(E)$ and $p_n(E)$ (see Eqn. (B.8) and Eqn. (B.9)):

$$\epsilon^2_{\ln \tilde{R}_{i+1}(E)} \overset{(B.6)}{=} \underbrace{\epsilon^2(\ln R_i(E))}_{=0} + \epsilon^2(\ln H_i(E+\Delta E)) + \epsilon^2(\ln H_i(E)) \ . \tag{B.17}$$

The first term on the r.h.s. vanishes because $R_i(E)$ was fixed during the nth recursion. The other terms can easily be calculated by assuming that the a priori error of the histogram H_i develops with the number of entries $\sqrt{H_i(E)}$:

$$\begin{aligned}
\epsilon^2\left(\ln H_i(E)\right) &= \left(\left.\frac{\partial \ln x}{\partial x}\right|_{x=H_i(E)} \epsilon(H_i(E)) \right)^2 && \text{(B.18)}\\
&= \left(\frac{1}{H_i(E)} \cdot \epsilon(H_i(E)) \right)^2 && \text{(B.19)}\\
&= \left(\frac{\sqrt{H_i(E)}}{H_i(E)} \right)^2 && \text{(B.20)}\\
&= \frac{1}{H_i(E)} \ . && \text{(B.21)}
\end{aligned}$$

So finally we get:

$$\begin{aligned}
\epsilon^2_{\ln \tilde{R}_{i+1}(E)} &\overset{(B.17)}{=} \epsilon^2(\ln H_i(E+\Delta E)) + \epsilon^2(\ln H_i(E)) && \text{(B.22)}\\
&\overset{(B.21)}{=} \frac{1}{H_i(E+\Delta E)} + \frac{1}{H_i(E)} && \text{(B.23)}\\
&= \frac{H_i(E+\Delta E) + H_i(E)}{H_i(E+\Delta E) H_i(E)} \ . && \text{(B.24)}
\end{aligned}$$

One could argue, that for $H_i(E) = 0$ or $H_i(E+\Delta E) = 0$ a singularity appears, but then $q_n(E)$ will vanish due to the fact that ϵ^2 is infinity. So there is no real problem. The complete formula can be found in Sect. 2.4.3 on page 19.

Appendix C

Multi-Histogram Reweighting

This derivation is motivated by A.M. Ferrenberg und R.H. Swendsen [41, 42]. After performing k canonical simulations (sequential or in parallel) at k different temperatures one gets k histograms $H_i(E)$ with $i = 1, \ldots, k$. The distribution on the ith simulation at a fixed temperature T_i is given by:

$$H_i(E) \propto P_{\text{can}}(E, T_i) = \alpha_i H_i(E) \ . \tag{C.1}$$

The missing constant α_i can be obtained by summing up both sides:

$$\begin{aligned} \sum_E P_{\text{can}}(E, T_i) &= \sum_E \alpha_i H_i(E) && \text{(C.2)} \\ Z(T_i) &= \alpha_i N_i \ , && \text{(C.3)} \end{aligned}$$

where $Z(T_i)$ is the (unknown) partition function at temperature T_i and N_i is the number of entries in the ith histogram. An estimator for the density of states $\Omega(E)$ out of the ith simulation is given by:

$$\begin{aligned} \Omega_i(E) &\overset{(2.45)}{=} P_{\text{can}}(E, T_i)\mathrm{e}^{E/T_i} && \text{(C.4)} \\ &\overset{(\text{C}.1)}{=} \alpha_i H_i(E)\mathrm{e}^{E/T_i} && \text{(C.5)} \\ &\overset{(\text{C}.3)}{=} H_i(E) Z(T_i) {N_i}^{-1} \mathrm{e}^{E/T_i} \ . && \text{(C.6)} \end{aligned}$$

A much better estimator can be obtained by using the statistics of all k simulations, so we add them weighted inverse to their error:

$$\hat{\Omega}(E) = \sum_{i=1}^{k} W_i(E)\Omega_i(E) \ , \tag{C.7}$$

where

$$W_i(E) \propto 1/\epsilon^2(\Omega_i(E)) \tag{C.8}$$

and the normalization

$$\sum_{i=1}^{N} W_i(E) = 1 \tag{C.9}$$

has be to fulfilled. A solution is

$$W_i(E) = \frac{1/\epsilon^2(\Omega_i(E))}{\sum_{i=1}^{k} 1/\epsilon^2(\Omega_i(E))} \; . \tag{C.10}$$

The error of Ω at the energy E can easily be calculated by assuming that the error of the histogram H_i at energy E grows with the square root of the number of entries $\sqrt{H_i(E)}$:

$$\begin{aligned}
\epsilon(\Omega_i(E)) &\overset{(C.6)}{=} \epsilon(H_i(E)Z(T_i){N_i}^{-1}\mathrm{e}^{E/T_i}) && \text{(C.11)}\\
&= Z(T_i){N_i}^{-1}\mathrm{e}^{E/T_i}\epsilon(H_i(E)) && \text{(C.12)}\\
&= Z(T_i){N_i}^{-1}\mathrm{e}^{E/T_i}\sqrt{H_i(E)} && \text{(C.13)}\\
&\overset{(C.6)}{=} \frac{\Omega_i(E)}{\sqrt{H_i(E)}} && \text{(C.14)}
\end{aligned}$$

So the estimator $\hat{\Omega}(E)$ can be calculated out of Eqn. (C.7):

$$\begin{aligned}
\hat{\Omega}(E) &\overset{(C.10)}{=} \frac{\sum_{i=1}^{k} 1/\epsilon^2(\Omega_i(E))\Omega_i(E)}{\sum_{i=1}^{k} 1/\epsilon^2(\Omega_i(E))} && \text{(C.15)}\\
&\overset{(C.14)}{=} \frac{\sum_{i=1}^{k} \frac{H_i(E)}{\Omega_i(E)^2}\Omega_i(E)}{\sum_{i=1}^{k} \frac{H_i(E)}{\Omega_i(E)^2}} && \text{(C.16)}\\
&= \frac{\sum_{i=1}^{k} \frac{H_i(E)}{\Omega_i(E)}\Omega(E)}{\sum_{i=1}^{k} \frac{H_i(E)}{\Omega_i(E)^2}\Omega(E)} && \text{(C.17)}\\
&\approx \frac{\sum_{i=1}^{k} H_i(E)}{\sum_{i=1}^{k} \frac{H_i(E)}{\Omega_i(E)}} && \text{(C.18)}\\
&\overset{(C.6)}{=} \frac{\sum_{i=1}^{k} H_i(E)}{\sum_{i=1}^{k} N_i Z^{-1}(T_i)\mathrm{e}^{-E/T_i}} && \text{(C.19)}
\end{aligned}$$

The unknown partition function can be calculated self-consistently

$$Z(T_k) = \sum_E \hat{\Omega}(E)\mathrm{e}^{-E/T_k} \overset{(C.19)}{=} \sum_E \mathrm{e}^{-E/T_k} \frac{\sum_i H_i(E)}{\sum_i N_i Z^{-1}(T_i)\mathrm{e}^{-E/T_i}} \; . \tag{C.20}$$

Appendix D

Usage of ABSimT

ABSimT is the short form of "AB Simulation Tools". This package of programs can be used to reproduce the data of this thesis completely. The programs are written in **C** [64] and the source code is freely available [65].

D.1 Function Overview

The functions are modular, most routines were used by different main programs. The methods implemented are:

- Canonical simulation
- Multicanonical simulation
- Energy-Landscape-Paving
- Parallel tempering
- Multicanonical replica exchange
- Analysis of all of the above by:
  - Histogram analysis
  - Time series analysis
  - Error estimation with Blocking-Jackknife method

On the one side we wanted to keep the program as simple as possible to easily allow changes of the model and the updates. On the other hand we wanted to keep the program free of doublings.

All kinds of simulation use the same update procedure `system_update()` which chooses the needed weight automatically and allows several types of updates (see Sect. D.5.3).

Directory	Kind of simulation and comment
`ABcan`	canonical simulation
`ABcan_analy`	histogram analysis of `ABcan`
`ABcan_analy_time`	timeseries analysis of `ABcan`
`ABcan_p`	parallel tempering simulation (swap temperatures)
`ABcan_p_analy_time`	timeseries analysis of `ABcan_p`
`ABcan_p2`	parallel tempering simulation (swap configuration)
`ABcan_p2_analy`	histogram analysis of `ABcan_p2`
`ABcan_p2_analy_time`	timeseries analysis of `ABcan_p2`
`ABelp`	ELP simulation
`ABinit`	init-file maker
`ABlib`	procedure library
`ABmuca`	multicanonical simulation
`ABmuca_analy`	histogram analysis of `ABmuca`
`ABmuca_analy_time`	timeseries analysis of `ABmuca`
`ABnuca_p`	multicanonical replica exchange simulation
`ABmuca_p_analy`	histogram analysis of `ABmuca_p`
`ABmuca_p_analy_time`	timeseries analysis of `ABmuca_p`
`ranMARS`	random number generator library [17]

Table D.1: The directory structure of the ABSimT package.

D.2 Structure Overview

The main directory displays the structure as shown in Table D.1 which includes the three single-CPU programs `ABcan`, `ABmuca` and `ABelp` and the three parallel programs `ABcan_p`, `ABcan_p2`, `ABmuca_p` for which a working installation of **MPI** [66] is necessary.

All programs except `ABelp` and `ABcan_p` have a histogram analysis routine which is denoted by the program name plus the ending `_analy`. For the ELP simulation `ABelp` an analysis is not necessary because the only useful outcome is the configuration with the lowest energy found and the histogram of energies which does not need to be analyzed.

Parallel tempering has two implementations, one which exchanges the temperatures and one which exchanges the configuration. The first does not need any histogram analysis, because counting histograms for a fixed temperature would be an additional effort of computer time.

Additionally, every program except `ABelp` has a timeserie analysis routine ending with `_analy_time`.

The remaining items are the global ABSimT library `ABlib`, the random number generator library `ranMARS` [17] and the init-file creator `ABinit`.

D.3 How to Do a Simulation

One has to choose the kind of simulation (canonical, multicanonical or ELP) and with this the associated version of implementation (serial or parallel) can be chosen according to the available computers.

Then one has to create an appropriate init-file by hand or by using `ABinit` (see Sect. D.4). The init-file should be given as first parameter to the program.

The simulation does backup at reasonable points, like the end of a recursion run or the end of a thermalization run, which makes it possible to restart aborted or crashed simulations.

After the simulation on should run the associated analysis program to calculate the common mean values like energy, specific heat and aggregation parameter. The common filenames can be change in the library `ABlib.h` (see Sect. D.5.1).

D.4 How to Create an Init-File

The init-file contains the essential information for the program. In general, the init-file is named `init` plus the program name, but any other init-file can be given as first parameter of the main program. A list of the required options in *correct* order can be found in Table D.2. Another option is to use the init-file maker `ABinit`, which is the best way to create the first init-file, but this would be a bit complicated for single-option change.

D.5 Technical Details

This section is just for people who want to understand the way of programming or want to change the source code, basically, to simulate another model.

D.5.1 ABlib

The library `ABlib.h` contains all routines used by the main programs. The detailed functions of the routines are documented in the source code.

The routines are sorted in four groups. The first group are the "elementary" functions like scalar product and lengths. The second group are functions of the standard variables, e.g. input and output of the histograms and weights. The next group contains all kind of update routines and their subroutines (see

Parameter	Comment	Used for simulation
`E_start`	Min. energy of the histogram	All
`E_end`	Max. energy of the histogram	All
`delta_E`	Energy difference of the histogram	All
`size`	Number of processes	`ABcan_p`, `ABcan_p2`, `ABmuca_p`
`T`	Temperature	All except `ABcan_p`, `ABcan_p2`, `ABmuca_p`
`T00`, `T01`,...	Temperatures	`ABcan_p`, `ABcan_p2`, `ABmuca_p`
`L_rec`	Length of recursion or thermalization run	All except `ABelp`
`nr_start`	Number of start recursion	`ABmuca`, `ABmuca_p`
`nr_rec`	Number of recursion to do	`ABmuca`, `ABmuca_p`
`L_final`	Length of final run	All
`r_cut`	Cutoff radius for contact measurement	All
`L_seq`	Length of update sequence	All
`seq_upd`	Sequence of updates	All
`seq_poly`	Sequence of polymers to update	All
`Nr_polys`	Number of polymers	All
`poly00`,`poly01`,...	Sequence of polymers	All

Table D.2: Parameters in the init-file on ABSimT.

Sect. D.5.3). The last group is the *polyconf* group, which provides all functions for the polymer configuration, like measurement, input and output. The polyconf type is explained in the next section.

D.5.2 Global Variables

Some variables are defined globally to make them accessible in every part of the program. The global variables vary with the kind of simulation, but in general the list of the configurations (`polylist`), all simulation parameters (`parameter`) and the histogram of energies (`E_histo`) are global. For multicanonical simulations additionally the weights (`lnW`) are global.

The list of polymers is a new defined type of variable. From the technical point of view it is an array of polymers, which is also a new defined variable. A polymer itself is defined as an array of monomers, where a monomer is a struct which includes a 3D vector and the type of the monomer. We think that this is the natural way of building up a system of multiple chains.

D.5.3 Main Update Function

The heart of every simulation is the update routine, which should be highly optimized. The procedure `system_update()` automatically chooses the weight according to the kind of simulation. The new configurations are created by the procedure `allupds_polyconf()`, which also calculates the energy difference. This is much faster than calculating the complete energy of the system. One of the arguments, which are passed to the function, are the consecutive number of the update sequence (`seq upd`) and of the sequence of the polymer to be updated (`seq poly`) which can be defined in the init-file (see Sect. D.4). This number (see table 3.1) allows to choose one of the following updates:

- Spherical update
  - Forward
  - Backward
- Rotation update
- Move update
- "Nothing" update

D.5.4 Further Details

The source code of `ABlib.h` is very well documented and can be understood quiet easily. Several example simulations of Seq. 13.1 were added to the package to test the functionality. We hope to publish a detailed documentation and manual of the project soon on the homepage [65].

Appendix E

Calculations

E.1 Rotation Matrix

Eqn. (A.16) gives:

$$
\begin{align*}
R_{ij} &= v_i v_j + (\delta_{ij} - v_i v_j)\cos\alpha + \epsilon_{ilj} v_l \sin\alpha \tag{E.1}\\
R^T_{jk} &= v_j v_k + (\delta_{jk} - v_j v_k)\cos\alpha + \epsilon_{kmj} v_m \sin\alpha\ , \tag{E.2}
\end{align*}
$$

with (see Eqn. (A.15))

$$v_i v_i = 1 \tag{E.3}$$

We evaluate:

$$
\begin{align*}
R_{ij}R^T_{jk} &= v_i v_k + (\underbrace{v_i v_k - v_i v_k}_{=0})\cos\alpha + (\underbrace{\epsilon_{kmj} v_m v_j}_{=0}\, v_i)\sin\alpha \tag{E.4}\\
&\quad + (\underbrace{v_i v_k - v_i v_k}_{=0})\cos\alpha + (\delta_{ij} - v_i v_k)\cos^2\alpha \\
&\quad + (\epsilon_{kmi} v_m - \underbrace{\epsilon_{kmj} v_m v_j}_{=0}\, v_i)\sin\alpha\cos\alpha \\
&\quad + (\underbrace{\epsilon_{ilj} v_l v_j}_{=0}\, v_k)\sin\alpha + (\epsilon_{ilk} v_l - \underbrace{\epsilon_{ilj} v_l v_j}_{=0}\, v_k)\sin\alpha\cos\alpha \\
&\quad + (\underbrace{\overbrace{\epsilon_{ilj}\epsilon_{kmj}}^{\delta_{ik}\delta_{lm}-\delta_{im}\delta_{lk}}\quad v_l v_m}_{\delta_{ik}-v_i v_k})\sin^2\alpha \\
&= v_i v_k + (\delta_{ik} - v_i v_k)(\underbrace{\cos^2\alpha + \sin^2\alpha}_{=1}) \tag{E.5}\\
&\quad + (\underbrace{\epsilon_{kmi} + \overbrace{\epsilon_{imk}}^{=-\epsilon_{kmi}}}_{=0}) v_m \sin\alpha\cos\alpha \\
&= v_i v_k + \delta_{ik} - v_i v_k \tag{E.6}\\
&= \underline{\underline{\delta_{ik}}} \tag{E.7}
\end{align*}
$$

E.2 Radius of Gyration

$$\begin{aligned}
{r_{\mathrm{gyr}}}^2 &= \frac{1}{N}\sum_{i=1}^{N}(\vec{r}_i - \vec{r}_M)^2 && \text{(E.8)}\\
&= \frac{1}{N}\sum_{i=1}^{N}(\vec{r}_i \cdot \vec{r}_i - 2\vec{r}_i \cdot \vec{r}_M + \vec{r}_M \cdot \vec{r}_M) && \text{(E.9)}\\
&= \frac{1}{N}\left(\sum_{i=1}^{N}\vec{r}_i \cdot \vec{r}_i\right) - \vec{r}_M \cdot \left(\frac{2}{N}\sum_{i=1}^{N}\vec{r}_i - \vec{r}_M \frac{1}{N}\sum_{i=1}^{N}1\right) && \text{(E.10)}\\
&= \frac{1}{N}\left(\sum_{i=1}^{N}\vec{r}_i \cdot \vec{r}_i\right) - \vec{r}_M \cdot (2\vec{r}_M - \vec{r}_M) && \text{(E.11)}\\
&= \frac{1}{N}\left(\sum_{i=1}^{N}\vec{r}_i \cdot \vec{r}_i\right) - \vec{r}_M \cdot \vec{r}_M \ , && \text{(E.12)}
\end{aligned}$$

which is the last global representation of the radius of gyration.

$$\begin{aligned}
{r_{\mathrm{gyr}}}^2 &= \frac{1}{N}\left(\sum_{i=1}^{N}\vec{r}_i \cdot \vec{r}_i\right) - \left(\frac{1}{N}\sum_{i=1}^{N}\vec{r}_i\right)^2 && \text{(E.13)}\\
&= \frac{1}{2N}\left(\sum_{i=1}^{N}\vec{r}_i \cdot \vec{r}_i + \sum_{j=1}^{N}\vec{r}_j \cdot \vec{r}_j\right) && \text{(E.14)}\\
&\quad - \left(\frac{1}{N}\sum_{i=1}^{N}\vec{r}_i\right) \cdot \left(\frac{1}{N}\sum_{j=1}^{N}\vec{r}_j\right) && \text{(E.15)}\\
&= \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\vec{r}_i \cdot \vec{r}_i + \vec{r}_j \cdot \vec{r}_j) - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\vec{r}_i \cdot \vec{r}_j) && \text{(E.16)}\\
&= \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\vec{r}_i \cdot \vec{r}_i + \vec{r}_j \cdot \vec{r}_j - 2\vec{r}_i \cdot \vec{r}_j) && \text{(E.17)}\\
&= \underline{\underline{\frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\vec{r}_i - \vec{r}_j)^2}} \ , && \text{(E.18)}
\end{aligned}$$

which is the relative representation of the radius of gyration.

E.3 Thermal Fluctuations Equation

Eqn. (2.45) gives:

$$\hat{O}(T) = \frac{\sum_{\mu\in\mathcal{M}} O_\mu {p_\mu}^{-1} \mathrm{e}^{-E_\mu/\mathrm{k_B}T}}{\sum_{\mu\in\mathcal{M}} {p_\mu}^{-1} \mathrm{e}^{-E_\mu/\mathrm{k_B}T}} \ . \tag{E.19}$$

We calculate:

$$\begin{aligned}
\frac{\mathrm{d}\hat{O}}{\mathrm{d}T} &= \frac{\mathrm{d}\beta}{\mathrm{d}T}\frac{\mathrm{d}\hat{O}}{\mathrm{d}\beta} && \text{(E.20)}\\
&= -\frac{1}{\mathrm{k_B}T^2}\frac{\mathrm{d}}{\mathrm{d}\beta}\left(\frac{\sum_{\mu\in\mathcal{M}} O_\mu {p_\mu}^{-1} \mathrm{e}^{-E_\mu/\mathrm{k_B}T}}{\sum_{\mu\in\mathcal{M}} {p_\mu}^{-1} \mathrm{e}^{-E_\mu/\mathrm{k_B}T}}\right) && \text{(E.21)}\\
&= -\frac{\frac{\mathrm{d}}{\mathrm{d}\beta}\sum_{\mu\in\mathcal{M}} O_\mu {p_\mu}^{-1} \mathrm{e}^{-E_\mu/\mathrm{k_B}T}}{\mathrm{k_B}T^2\sum_{\mu\in\mathcal{M}} {p_\mu}^{-1} \mathrm{e}^{-E_\mu/\mathrm{k_B}T}} && \text{(E.22)}\\
&\quad -\frac{\sum_{\mu\in\mathcal{M}} O_\mu {p_\mu}^{-1} \mathrm{e}^{-E_\mu/\mathrm{k_B}T}}{\mathrm{k_B}T^2}\cdot\frac{\mathrm{d}}{\mathrm{d}\beta}\left(\frac{1}{\sum_{\mu\in\mathcal{M}} {p_\mu}^{-1} \mathrm{e}^{-E_\mu/\mathrm{k_B}T}}\right)\\
&= -\frac{\sum_{\mu\in\mathcal{M}} O_\mu(-E_\mu) {p_\mu}^{-1} \mathrm{e}^{-E_\mu/\mathrm{k_B}T}}{\mathrm{k_B}T^2\sum_{\mu\in\mathcal{M}} {p_\mu}^{-1} \mathrm{e}^{-E_\mu/\mathrm{k_B}T}} && \text{(E.23)}\\
&\quad -\frac{\sum_{\mu\in\mathcal{M}} O_\mu {p_\mu}^{-1} \mathrm{e}^{-E_\mu/\mathrm{k_B}T}}{\mathrm{k_B}T^2}\cdot\frac{-\sum_{\nu\in\mathcal{M}}(-E_\nu) {p_\nu}^{-1} \mathrm{e}^{-E_\nu/\mathrm{k_B}T}}{(\sum_{\mu\in\mathcal{M}} {p_\mu}^{-1} \mathrm{e}^{-E_\mu/\mathrm{k_B}T})^2}\\
&= \frac{\sum_{\mu\in\mathcal{M}} O_\mu E_\mu {p_\mu}^{-1} \mathrm{e}^{-E_\mu/\mathrm{k_B}T}}{\mathrm{k_B}T^2\sum_{\mu\in\mathcal{M}} {p_\mu}^{-1} \mathrm{e}^{-E_\mu/\mathrm{k_B}T}} && \text{(E.24)}\\
&\quad -\frac{\sum_{\mu\in\mathcal{M}} O_\mu {p_\mu}^{-1} \mathrm{e}^{-E_\mu/\mathrm{k_B}T}}{\mathrm{k_B}T^2\sum_{\mu\in\mathcal{M}} {p_\mu}^{-1} \mathrm{e}^{-E_\mu/\mathrm{k_B}T}}\cdot\frac{\sum_{\nu\in\mathcal{M}} E_\nu {p_\nu}^{-1} \mathrm{e}^{-E_\nu/\mathrm{k_B}T}}{\sum_{\mu\in\mathcal{M}} {p_\mu}^{-1} \mathrm{e}^{-E_\mu/\mathrm{k_B}T}}\\
&= \underline{\underline{\frac{1}{\mathrm{k_B}T^2}\left(\widehat{OE} - \hat{O}\cdot\hat{E}\right)}}\ . && \text{(E.25)}
\end{aligned}$$

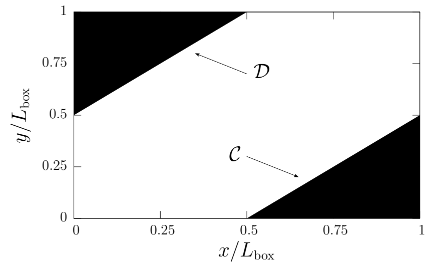


Figure E.1: The area of integration can be divided in three parts, the hatched parts are the areas ($|x-y|>0.5$) $\mathcal{C}$ and $\mathcal{D}$. The complete area is called $\mathcal{A}$.

E.4 Two Particles in Periodic Box

The mean distance of two particles ($\vec{x},\vec{y}\in\mathcal{B}_1$, see Eqn. (3.8)) in a periodic box can be calculated by:

$$\begin{aligned}
\langle \vec{d}_{\mathrm{per}}(\vec{x},\vec{y})^2\rangle &= \iiint_{\mathcal{B}_1}\mathrm{d}^3x\iiint_{\mathcal{B}_1}\mathrm{d}^3y\;\vec{d}_{\mathrm{per}}(\vec{x},\vec{y})^2\rho(\vec{x})\rho(\vec{y}) &\text{(E.26)}\\
&\overset{(3.9)}{=} \sum_{i=1}^{3}\int_0^{L_{\mathrm{box}}}\mathrm{d}x_i\int_0^{L_{\mathrm{box}}}\mathrm{d}y_i\;d_i^{\mathrm{per}}(x_i,y_i)^2\rho(x_i)\rho(y_i) &\text{(E.27)}\\
&= 3\int_0^{L_{\mathrm{box}}}\mathrm{d}x_1\int_0^{L_{\mathrm{box}}}\mathrm{d}y_1\;d_1^{\mathrm{per}}(x_1,y_1)^2\rho(x_1)\rho(y_1) &\text{(E.28)}\\
&= 3\langle d_1^{\mathrm{per}}(x_1,y_1)^2\rangle\;, &\text{(E.29)}
\end{aligned}$$

where $\rho(x_i)=1/L_{\mathrm{per}}$ is the probability density of x_i, which is just constant. This intermediate result is not surprising, but shows the isotropy of the 3D space. For simplicity we omit the subscript 1 and call the integration area $\mathcal{A}$ in the next steps. We divide the integration area according to Fig. E.1 in three parts due to the three different parts of the function $d^{\mathrm{per}}(x,y)$ (see Eqn. (3.10)). So Eqn. (E.29) simplifies to:

$$\begin{aligned}
\langle d^{\mathrm{per}}(x,y)^2\rangle &= \iint_{\mathcal{A}}\mathrm{d}x\mathrm{d}y\;d(x,y)^{\mathrm{per}}(x,y)^2\rho(x)\rho(y) &\text{(E.30)}\\
&= \iint_{\mathcal{A}\setminus\mathcal{C}\setminus\mathcal{D}}\mathrm{d}x\mathrm{d}y\;d^{\mathrm{per}}(x,y)^2\rho(x)\rho(y) &\text{(E.31)}\\
&\quad+\iint_{\mathcal{C}}\mathrm{d}x\mathrm{d}y\;d^{\mathrm{per}}(x,y)^2\rho(x)\rho(y)\\
&\quad+\iint_{\mathcal{D}}\mathrm{d}x\mathrm{d}y\;d^{\mathrm{per}}(x,y)^2\rho(x)\rho(y)\\
&= \iint_{\mathcal{A}\setminus\mathcal{C}\setminus\mathcal{D}}\mathrm{d}x\mathrm{d}y\;\frac{(x-y)^2}{L_{\mathrm{box}}^2} &\text{(E.32)}\\
&\quad+\iint_{\mathcal{C}}\mathrm{d}x\mathrm{d}y\;\frac{(x-y-L_{\mathrm{box}})^2}{L_{\mathrm{box}}^2}\\
&\quad+\iint_{\mathcal{D}}\mathrm{d}x\mathrm{d}y\;\frac{(x-y+L_{\mathrm{box}})^2}{L_{\mathrm{box}}^2}
\end{aligned}$$

```
md=0
for ( i = 1 to N )
  d=random()-random()
  if ( d > 0.5 )
    d = d - 1
  if (d < -0.5 )
    d = d + 1
  md = md + (d * d ) / N
end for
print md
```

Table E.1: Monte Carlo program to determine the mean distance `md` of two particles in a periodic box with length 1. `random()` gives a random number, which is uniformly distributed in the interval $[0,1)$. `N` gives the number of iteration steps, in this case the number of measured distances `d`.

Now we can scale out L_{box} by replacing x and y by $x' = x/L_{\text{box}}$ and $y' = y/L_{\text{box}}$. The rescaled integration areas are called $\mathcal{A}'$, $\mathcal{C}'$ and $\mathcal{D}'$.

$$\begin{aligned}
\langle d^{\text{per}}(x,y)^2\rangle &= L_{\text{box}}^2 \iint_{\mathcal{A}'\setminus\mathcal{C}'\setminus\mathcal{D}'} \mathrm{d}x'\mathrm{d}y'\ (x'-y')^2 && \text{(E.33)}\\
&\quad + L_{\text{box}}^2 \iint_{\mathcal{C}} \mathrm{d}x'\mathrm{d}y'\ (x'-y'-1)^2 \\
&\quad + L_{\text{box}}^2 \iint_{\mathcal{D}'} \mathrm{d}x'\mathrm{d}y'\ (x'-y'+1)^2 \\
&= L_{\text{box}}^2 \iint_{\mathcal{A}'} \mathrm{d}x'\mathrm{d}y'\ \tilde{d}^{\text{per}}(x',y')^2 && \text{(E.34)}\\
&= L_{\text{box}}^2 \langle \tilde{d}^{\text{per}}(x',y')^2\rangle && \text{(E.35)}
\end{aligned}$$

This intermediate result is also not surprising; it is just the mean values of two particles in a periodic box with length 1, where

$$\tilde{d}^{\text{per}}(x,y) = \begin{cases} (x-y)+1 & : \quad (x-y) < -1/2 \\ (x-y) & : \quad -1/2 < (x-y) < 1/2 \\ (x-y)-1 & : \quad 1/2 < (x-y) \end{cases}, \qquad \text{(E.36)}$$

is the distance function on a 1D dimensional box length 1. We see that the mean distance scales with the box length, which seems natural[1].

The last integral can be calculated by Monte Carlo integration. A possible program with `N` iteration steps can be found in Tab. E.1. The program gives 1/12 as result, which can also be calculated analytically. First we can use the symmetry of the square of the distance measurement (see Fig. 3.1).

$$d^{\text{per}}(x,y)^2 = d^{\text{per}}(y,x)^2\ . \qquad \text{(E.37)}$$

---

[1] For a box with hard walls this is totally clear, because the distance measurement is the same function in the whole cube, but for a periodic box it is not that obvious.

$$\begin{aligned}\langle \tilde{d}^{\mathrm{per}}(x,y)^2\rangle &= \iint_{\mathcal{A}'} \mathrm{d}x\mathrm{d}y\ \tilde{d}^{\mathrm{per}}(x,y)^2 &\text{(E.38)}\\ &= \int_0^1 \mathrm{d}x \int_0^1 \mathrm{d}y\ \tilde{d}^{\mathrm{per}}(x,y)^2 &\text{(E.39)}\\ &= 2\int_0^1 \mathrm{d}x \int_0^x \mathrm{d}y\ \tilde{d}^{\mathrm{per}}(x,y)^2 &\text{(E.40)}\\ &= 2\iint_{\mathcal{A}''} \mathrm{d}x\mathrm{d}y\ \tilde{d}^{\mathrm{per}}(x,y)^2 &\text{(E.41)}\end{aligned}$$

where $\mathcal{A}''$ is the half of the square $[0,1]\times[0,1]$.

$$\begin{aligned}\frac{1}{2}\langle \tilde{d}^{\mathrm{per}}(x,y)^2\rangle &= \iint_{\mathcal{A}''\backslash\mathcal{C}} \mathrm{d}x\mathrm{d}y\ \tilde{d}^{\mathrm{per}}(x,y)^2 + \iint_{\mathcal{C}} \mathrm{d}x\mathrm{d}y\ \tilde{d}^{\mathrm{per}}(x,y)^2 &\text{(E.42)}\\ &= \iint_{\mathcal{A}''\backslash\mathcal{C}} (x-y)^2 + \iint_{\mathcal{C}} \mathrm{d}x\mathrm{d}y\ (x-y-1)^2 &\text{(E.43)}\\ &= \iint_{\mathcal{A}''} \mathrm{d}x\mathrm{d}y\ (x-y)^2 - \iint_{\mathcal{C}} \mathrm{d}x\mathrm{d}y\ (x-y)^2 &\text{(E.44)}\\ &\quad + \iint_{\mathcal{C}} \mathrm{d}x\mathrm{d}y\ (x-y-1)^2 \\ &= \iint_{\mathcal{A}''} \mathrm{d}x\mathrm{d}y\ (x-y)^2 &\text{(E.45)}\\ &\quad + \iint_{\mathcal{C}} \mathrm{d}x\mathrm{d}y\ \left[(x-y+1)^2-(x-y)^2\right] \\ &= \int_0^1 \mathrm{d}x \int_0^x \mathrm{d}y\ \left(x^2-2xy+y^2\right) &\text{(E.46)}\\ &\quad + \int_{\frac{1}{2}}^1 \mathrm{d}x \int_0^{x-\frac{1}{2}} \mathrm{d}y\ (1-2x+2y) \\ &= \int_0^1 \mathrm{d}x\ \frac{x^3}{3} - \int_{\frac{1}{2}}^1 \mathrm{d}x\ \left(x^2-x+\frac{1}{4}\right) &\text{(E.47)}\\ &= \frac{1}{12}-\frac{1}{24}\ . &\text{(E.48)}\end{aligned}$$

So the final result is:

$$\langle \tilde{d}^{\mathrm{per}}(x,y)^2\rangle = 1/6 - 1/12 = 1/12\ , \tag{E.49}$$

which is the same as for the Monte Carlo program. The first term (1/6) in Equ. (E.49) is the square mean distance of two particles in a box with hard walls. The second term is negative, which shows that a periodic box "looks" smaller than a normal box. The result for the 3D dimensional box with length L_{box} follows instantly:

$$\langle \vec{d}_{\mathrm{per}}(\vec{x},\vec{y})^2\rangle = 3\cdot L_{\mathrm{box}}^2/12 = L_{\mathrm{box}}^2/4\ . \tag{E.50}$$

So the mean distance of two particles is:

$$\langle \Delta r_M\rangle = \sqrt{\langle \vec{d}_{\mathrm{per}}(\vec{x},\vec{y})^2\rangle} = \underline{\underline{L_{\mathrm{box}}/2}}\ . \tag{E.51}$$

In the previous chapters we often used the aggregation parameter:

$$\langle \Gamma \rangle = \langle \Delta r_M \rangle / 2 = \underline{\underline{L_{\mathrm{box}}/4}} \ . \tag{E.52}$$

Bibliography

[1] T.E. Creighton. *Proteins: Structures and Molecular Properties.* W.H. Freeman and Company, New York, 2nd ed. edition, 1993.

[2] http://www.rcsb.org. Homepage of the Protein Database.

[3] http://fred.bioinf.uni-sb.de:4711/DFG-protein-protein-docking/index .shtml. Protein docking project of the university Saarbrücken.

[4] G. Némethy, K.D. Gibson, K.A. Palmer, C.N Yoon, G. Paterlini, A. Zagari, S. Rumsey, and H.A. Scheraga. Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. *Journal of Physical Chemistry*, 96(15):6472–6484, 1992.

[5] C. Junghans and U.H.E. Hansmann. Numerical comparison of Wang Landau sampling and parallel tempering for met-enkephalin. *International Journal of Modern Physics C: Physics and Computers*, 17(6):817–824, 2006.

[6] C. Junghans and U.H.E. Hansmann. Cross-check methods in protein simulations. In J. Meinke, S. Mohanty, O. Zimmermann, and U.H.E. Hansmann, editors, *From Computational Biophysics to Systems Biology 2006*, volume 34 of *NIC Series*, pages 157–160, Jülich, 2006.

[7] K.F. Lau and K.A. Dill. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, 22(10):3986–3997, 1989.

[8] K.A. Dill. Polymer principles and protein folding. *Protein Science*, 8(6):1166–1180, 1999.

[9] Moore's law. http://en.wikipedia.org/wiki/Moore%27s_Law.

[10] F.H. Stillinger, T. Head-Gordon, and C.L. Hirshfeld. Toy model for protein folding. *Physical Review E*, 48(2):1469–1477, 1993.

[11] F.H. Stillinger and T. Head-Gordon. Collective aspects of protein folding illustrated by a toy model. *Physical Review E*, 52(3):2872–2877, 1995.

[12] A. Irbäck, C. Peterson, F. Potthast, and O. Sommelius. Local interactions and protein folding: A three-dimensional off-lattice approach. *Journal of Chemical Physics*, 107(1):273–282, 1997.

[13] R. Haberland, S. Fritzsche, G. Peinel, and K. Heinzinger. *Molekulardynamik*. Vieweg, 1994.

[14] B.A. Berg. *Markov Chain Monte Carlo Simulations and Their Statistical Analysis*. World Scientific, Singapore, 2004.

[15] M.E.J. Newman and G.T. Barkema. *Monte Carlo Methods in Statistical Physics*. Oxford University Press, 2002.

[16] N. Metropolis and S. Ulam. The Monte Carlo method. *Journal of the American Statistical Association*, 44(247):335–341, 1949.

[17] G. Marsaglia and A. Zaman. Toward a universal random number generator. *Statistics & Probability Letters*, 9(1):35–39, 1990.

[18] W. Janke. Pseudo random numbers: Generation and quality checks. In J. Grotendorst, D. Marx, and A. Maramatsu, editors, *Quantum Simulations of Complex many-Body Systems: From Theory to Algorithms*, volume 10 of *NIC Series*, pages 423–445, Jülich, 2002. John von Neumann Institute for Computing.

[19] K. Huang. *Introduction to Statistical Physics*. Taylor & Francis, 2001.

[20] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[21] R.J. Glauber. Time-dependent statistics of the ising model. *Journal of Mathematical Physics*, 4(2):294–307, 1963.

[22] U. Wolff. Collective Monte Carlo updating for spin systems. *Physical Review Letters*, 62(4):361–364, 1989.

[23] R.H. Swendsen and J.-S. Wang. Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters*, 58(2):86–88, 1987.

[24] W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[25] C. Chib and E. Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

[26] A. Bazavov, B.A. Berg, and U.M. Heller. Biased metropolis-heatbath algorithm for fundamental-adjoint SU(2) lattice gauge theory. *Physical Review D*, 153(11):117501, 2005.

[27] T. Lu and D. Yewick. Biased multicanonical sampling. *Photonics Technology Letters*, 7(2):1420–1422, 2005.

[28] T. Neuhaus and J.S. Hager. Free-energy calculations with multiple Gaussian modified ensembles. *Physical Review E*, 74(3):036702, 2006.

[29] G.M. Torrie and J.P. Valleau. Nonphysical sampling distribution in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, 1977.

[30] A.M. Ferrenberg and R.H. Swendsen. New Monte Carlo technique for studying phase transitions. *Physical Review Letters*, 61(23):2635–2638, 1988.

[31] B.A. Berg and T. Neuhaus. Multicanonical ensemble: A new approach to simulate first-order phase transitions. *Physical Review Letters*, 68(1):9–12, 1992.

[32] F. Wang and D.P. Landau. Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram. *Physical Review E*, 64(5):056101, 2001.

[33] W. Janke. The recursion. unpublished notes, 1997.

[34] W. Janke. Histograms and all that. In B. Dünweg, D.P. Landau, and A.I. Milchev, editors, *Computer Simulations of Surfaces and Interfaces*, volume 114 of *NATO Science Series*, pages 137–157, Albena, Bulgaria, 2003. NATO Advances Study Institute.

[35] C.G. Broyden. Quasi-Newton methods and their application to function minimization. *Mathematics of Computation*, 21(99):368–381, 1967.

[36] M.R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49(6):409–436, 1952. Later known as National Institute of Standards and Technology (NIST).

[37] U.H.E. Hansmann and L.T. Wille. Global optimization by energy landscape paving. *Physical Review Letters*, 88(4):068105, 2002.

[38] U.H.E. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chemical Physics Letters*, 281(1-3):140–150, 1997.

[39] K. Hukushima and K. Nemeto. Exchange Monte Carlo method and application to spin glass simulations. *Journal of Physical Society of Japan*, 65(6):1604–1608, 1996.

[40] R.H. Swendsen and J.-S. Wang. Replica Monte Carlo simulation of spin-glasses. *Physical Review Letters*, 57(21):2607–2609, 1986.

[41] A.M. Ferrenberg and R.H. Swendsen. Optimized Monte Carlo data analysis. *Physical Review Letters*, 63(12):1195–1198, 1989.

[42] A.M. Ferrenberg and R.H. Swendsen. Errata of “New Monte Carlo technique for studying phase transitions”. *Physical Review Letters*, 63(16):1658, 1989.

[43] Y. Sugita and Y. Okamoto. Replica-exchange multicanonical algorithm and multicanonical replica-exchange method for simulating systems with rough energy landscape. *Chemical Physics Letters*, 329(3-4):261–270, 2000.

[44] W. Janke. Statistical analysis of simulations: Data correlations and error estimation. In J. Grotendorst, D. Marx, and A. Maramatsu, editors, *Quantum Simulations of Complex many-Body Systems: From Theory to Algorithms*, volume 10 of *NIC Series*, pages 423–445, Jülich, 2002. John von Neumann Institute for Computing.

[45] B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans.* SIAM, Philadelphia, 1982.

[46] R.G. Miller. The Jackknife – a Review. *Biometrika*, 61(1):1–15, 1974.

[47] W. Janke. Private communications. This was also the case for the measurement of the radius of gyration of clusters of an Ising Model in a periodic box.

[48] M. Bachmann, H. Arkın, and W. Janke. Multicanonical study of coarse-grained off-lattice models for folding heteropolymers. *Physical Review E*, 71(3):031906, 2005.

[49] J. Schluttig. Molecular mechanics of coarse-grained protein models. Diploma thesis, University of Leipzig, 2005.

[50] A. Kallias. Thermodynamics and folding kinetics of coarse-grained protein models. Diploma thesis, University of Leipzig, 2005.

[51] S. Schnabel. Thermodynamische Eigenschaften und Faltungskanäle von coarse-grained Heteropolymeren. Diploma thesis, University of Leipzig, 2005.

[52] M. Bachmann and W. Janke. Thermodynamics of lattice heteropolymers. *Journal of Chemical Physics*, 120(14):6779–6791, 2004.

[53] T. Vogel, S. Schnabel, M. Bachmann, and W. Janke, 2006. to be published.

[54] T. Vogel. Low-temperature behaviour of short polymers. Seminar Talk, 2006. The ground states always look like a " well" with a stretched part in the middle and surrounding helix.

[55] J. Lee, I.-H. Lee, and J. Lee. Unbiased global optimization of Lennard-Jones clusters for $n \leq 201$ using the conformational space annealing method. *Physical Review Letters*, 91(8):080201, 2003.

[56] C. Junghans, M. Bachmann, and W. Janke. Microcanonical analyses of peptide aggregation processes. *Physical Review Letters*, 2006. in press.

[57] S. Hilbert and J. Dunkel. Nonanalytic microscopic phase transitions and temperature oscillations in the microcanonical ensemble: An exactly solvable one-dimensional model for evaporation. *Physical Review E*, 74(1):011120, 2006.

[58] W. Thirring. Systems with negative specific heat. *Zeitschrift für Physik A Hadrons and Nuclei*, 235(4):339–352, 1970.

[59] W. Janke. Canonical versus microcanonical analysis of first-order phase transitions. *Nuclear Physics B (Proceedings Supplements)*, 63(A-C):631–633, 1998.

[60] H. Behringer and M. Pleimling. Continuous phase transitions with a convex dip in the microcanonical entropy. *Physical Review E*, 74(1):011108, 2006.

[61] M. Schmidt, R. Kusche, T. Hippler, J. Donges, W. Kronmüller, B. v. Issendorff, and H. Haberland. Negative heat capacity for a cluster of 147 sodium atoms. *Physical Review Letters*, 86(7):1191–1194, 2001.

[62] V. Privman, editor. *Finite-Size Scaling and Numerical Simulations of Statistical Systems*, Singapore, 1990. World Scientific.

[63] K. Binder. Computational methods in field theory. In H. Gausterer and C.B. Lang, editors, *Schladming Lecture Notes*, page 59, Berlin, 1992. Springer.

[64] B.W. Kerrighan and D. Ritchie. *C. Programming Language.* Prentice Hall, 1988.

[65] http://www.physik.uni-leipzig.de/ junghans/absimt/. Homepage of ABSimT.

[66] W. Gropp, E. Lusk, and A. Skjellum. *Using MPI: portable parallel programming with the message-passing interface.* MIT Press Cambridge, 1994.

Danksagung

Selbständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Diplomarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Alle Stellen der Arbeit, die wörtlich oder sinngemäß aus Veröffentlichungen oder aus anderweitigen fremden Äußerungen entnommen wurden, sind als solche kenntlich gemacht. Ferner erkläre ich, dass diese Arbeit noch nicht in einem anderen Studiengang als Prüfungsleistung verwendet wurde.

Christoph Junghans

Ich bin einverstanden, dass diese Arbeit nach positiver Begutachtung in der Universitätsbibliothek zur Verfügung steht.

Christoph Junghans