Fakultät für Physik und Geowissenschaften
Institut für theoretische Physik

UNIVERSITÄT LEIPZIG

Diplomarbeit

Molecular Mechanics of Coarse-Grained Protein Models

vorgelegt von

Jakob Schluttig

Oktober 2005

Betreuer:	Prof. Dr. Wolfhard Janke
	Dr. Michael Bachmann
Gutachter:	Prof. Dr. Wolfhard Janke
	Prof. Dr. Ulrich Behn

Contents

Introduction **1**

1 The “AB” Heteropolymer Model **3**
1.1 Introduction 3
1.2 Detailed Description 4
1.3 Influence of Flexible Bonds 7
1.4 Cartesian Formulation of the Potential Terms 8
1.5 Derivation of Forces 9

2 Molecular Dynamics at Finite Temperature **13**
2.1 Liouville Operator and Trotter Factorisation 13
2.1.1 Liouville Formalism 14
2.1.2 Propagator 14
2.1.3 Factorisation 16
2.1.4 Time Reversibility 17
2.2 Simulations in the Microcanonical Ensemble 19
2.3 Thermodynamics of Selected 1d Systems 21
2.3.1 Harmonic Oscillator 21
2.3.2 The Quartic Double Well 23
2.4 Molecular Dynamics in the Canonical Ensemble 26
2.4.1 Andersen Thermostat – Stochastic Molecular Dynamics 26
2.4.2 Overview of the Nosé-Hoover-Chain Thermostat 29
2.4.3 Algorithmic Details of NHC 30
2.4.4 Exemplary Simulations of Selected 1d Systems with the NHC 34

3 Monte Carlo Simulation Methods and Data Analysis **43**
3.1 Simulation 43
3.1.1 Metropolis Sampling 43
3.1.2 Parallel Tempering 44
3.2 Statistical Analysis 44
3.2.1 Error Analysis 44
3.2.2 Reweighting 46
3.2.3 Fluctuation Quantities in Molecular Dynamics Simulations 49
3.3 Quantities 51
3.3.1 Energy and Specific Heat 51
3.3.2 End-To-End-Distance and Radius of Gyration 52

3.3.3 Comparing Two Configurations 52

4 Examination of the Model System 57

4.1 Thermodynamics with Monte Carlo simulations 57

4.1.1 General Facts about the Simulations 57

4.1.2 Thermodynamics 62

4.1.3 Impact of Different Bond Strengths on the System 65

4.2 Comparison with Molecular Dynamics simulations 69

4.2.1 Simulation Fundamentals 69

4.2.2 Thermodynamics 71

4.2.3 Time Scales 75

4.3 Free-Energy Landscapes 78

5 Torsion – Generalisation of the Model 83

5.1 The Torsion Potential 83

5.1.1 Cartesian Transcription 85

5.1.2 Derivation of the Torsional Force 86

5.2 Monte Carlo Simulations 88

5.3 Realizing and Tackling the Problem in Molecular Dynamics 93

5.3.1 Explanation and Deeper Analysis of the Divergence 94

5.3.2 Possible Solutions 98

Summary 105

A Selected Source Codes 109

A.1 Andersen Thermostat 109

A.2 Nosé-Hoover-Chain Thermostat 110

List of Figures

1.1 Schematic view of a heteropolymer 4
1.2 Lennard-Jones potential 6
1.3 Low energy structure of the 55mer 6
1.4 Specific heat contribution of flexible bonds 8
1.5 Specific heat of sequence 20.4 without bonds contribution 8

2.1 Microcanonical trajectory of the 1d harmonic oscillator and 1d quartic double well 20
2.2 Time series of the 1d harmonic oscillator and 1d quartic double well 20
2.3 Mean energy of the quartic double well 24
2.4 Specific heat of the quartic double well 24
2.5 Distribution of positions and potential energy for the quartic double well . . . 25
2.6 Distribution of positions and potential energy for the quartic double well . . . 25
2.7 Trajectory of harmonic oscillator with Andersen thermostat 27
2.8 Trajectory of quartic double well with Andersen thermostat 27
2.9 Relative error from x distribution of harmonic oscillators with Andersen thermostat 28
2.10 Relative error from specific heat of one-dimensional harmonic oscillators with Andersen thermostat 28
2.11 Different choices for the virtual masses for a NHC of length 2 32
2.12 Velocity autocorrelation function 33
2.13 Fourier transform of velocity autocorrelation function 33
2.14 Long-time trajectory of harmonic oscillator with $M = 1$ 35
2.15 Distributions of x, p and v_ξ for harmonic oscillator with $M = 1$ 35
2.16 Long-time trajectory of harmonic oscillator with $M = 1$ and two different choices for the Nosé-Hoover time scale ($\tau = 0.2$ and $\tau = 5$) 35
2.17 Frequency spectrum of harmonic oscillator at $\omega = 1\mathrm{s}^{-1}$ 36
2.18 Frequency spectrum of quartic double well 36
2.19 Frequency spectra of harmonic oscillators at $\omega = 0.2\mathrm{s}^{-1}$ and $\omega = 5\mathrm{s}^{-1}$ 36
2.20 Frequency spectra of harmonic oscillator at $\omega = 1\mathrm{s}^{-1}$ with Nosé-Hoover time scale $\tau = 0.2\mathrm{s}$ and $\tau = 5\mathrm{s}$ 37
2.21 Distributions of x and p of harmonic oscillator at $\omega = 1\mathrm{s}^{-1}$ with Nosé-Hoover time scale $\tau = 0.2\mathrm{s}$ and $\tau = 5\mathrm{s}$ 37
2.22 Projections of short-time trajectory of the harmonic oscillator thermostated with a NHC with $M = 2$ 38
2.23 Long-time trajectory of harmonic oscillator at $T = 0.1$ and $T = 1.0$ 39

2.24 Long-time trajectory of quartic double well at $k_BT = 0.1$J and $k_BT = 1.0$J . 39
2.25 Distributions of x and p for the harmonic oscillator 40
2.26 Distributions of x 40
2.27 Specific heat of harmonic oscillator and quartic double well 40
2.28 Distributions of kinetic and potential energies for harmonic oscillator 41
2.29 Distributions of kinetic and potential energies for quartic double well 41
2.30 Comparison of the convergence of moments for harmonic oscillator and quartic double well 41

4.1 Monte Carlo update scheme 58
4.2 Temperature distribution in parallel tempering 59
4.3 Energy histograms in parallel tempering 59
4.4 Autocorrelation functions and error estimations 60
4.5 Mean energy and heat capacity of six 20mers 63
4.6 Fluctuation of end-to-end-distance radius of gyration of six 20mers 64
4.7 Impact of bond strength on specific heat, overview 65
4.8 Impact of bond strength on specific heat, without bond contribution 65
4.9 Fluctuation of end-to-end-distance for different bond strengths 66
4.10 Fluctuation of radius of gyration for different bond strengths 66
4.11 Impact of bond strength on specific heat, detail 66
4.12 Overlap and rmsd of ground-states with different bond strengths 67
4.13 Ground-state snapshots for different bond strengths 67
4.14 Contact maps for different bond strengths 68
4.15 Frequency spectrum of sequence 20.2 71
4.16 Distributions of velocities and bond lengths 71
4.17 Kinetic energy in a MD simulation of sequence 20.2 71
4.18 Mean energy and specific heat of 20.2 in MD 72
4.19 q parameter for two different MD runs 72
4.20 Fluctuation of end-to-end-distance measured with MD 73
4.21 Fluctuation of radius of gyration measured with MD 73
4.22 Mean energy and specific heat of 20.2 in MD with Andersen thermostat . . . 75
4.23 q parameter for MD run with Andersen thermostat 75
4.24 Fluctuation of end-to-end-distance in the simulation with Andersen thermostat 76
4.25 Fluctuation of radius of gyration in the simulation with Andersen thermostat 76
4.26 Comparison of autocorrelation function for MC and MD 77
4.27 Schematic free-energy landscape of a two-state folder 79
4.28 q parameter of sequence 20.6 for two definitions 80
4.29 Reweighted fluctuation of the q parameter for two definitions 80
4.30 Free-energy landscape for sequence 20.6 around folding temperature 81
4.31 Time series of overlap Q at $T = 0.08$ 81
4.32 Reweighted free-energy landscape for sequence 20.6 around folding temperature 82

5.1 Cross product definition of the torsion angle 84
5.2 Scalar product definition of the torsion angle 84
5.3 Explanation for the algebraic sign of the torsion angle 84
5.4 Potential energy and heat capacity of several sequences for the GAB model . 89

5.5 End-to-end-distance, radius of gyration and respective fluctuations for the GAB model 90
5.6 Ground-state of sequence 20.2 for AB and GAB model 91
5.7 Contact maps of sequence 20.2 for AB and GAB model 92
5.8 Impact of bond strength on specific heat, overview 92
5.9 Impact of bond strength on specific heat, without bond contribution 92
5.10 Impact of bond strength on GAB model, detail 93
5.11 "Conserved" energy trajectories 94
5.12 Frequency spectra of the GAB model 94
5.13 Time series of energies 95
5.14 Time series of velocities 95
5.15 The critical configuration 95
5.16 Example for divergence of the torsion force 96
5.17 Conserved energy trajectory for the homo-four-mer 102
5.18 Heat capacity of the homo-four-mer 102
5.19 Conserved energy trajectory for sequence 20.4 102
5.20 Heat capacity of sequence 20.4 102
5.21 Fluctuation of end-to-end-distance for sequence 20.4 with torsion 103
5.22 Fluctuation of radius of gyration for sequence 20.4 with torsion 103

List of Tables

1.I Observed AB-Sequences 5

4.I Exponential autocorrelation times 61
4.II Ground-state energies 65
4.III Lennard-Jones contribution to the potential energy 68
4.IV Comparison of MD and MC ground-states 74
4.V Comparison of ground-states of MC and MD run with Andersen thermostat . 75
4.VI Exponential autocorrelation times of MC and MD 77

5.I Overlap and rmsd of GAB and AB ground-states 91

A.I Variables used in source code of Andersen thermostat 110
A.II Variables used in source code of Nosé-Hoover-Chain thermostat 113
A.III Higher order Trotter factorisation 113

Introduction

Proteins are highly specialised macromolecules performing essential functions in a biological system, such as controlling transport processes, stabilisation of the cell structure, enzymatic catalysation of chemical reactions, etc. Chemically, proteins are build up as sequences of $N \sim 50, \ldots, 3000$ amino acid residues linked by peptide bonds. All proteins are composed of 20 different types of amino acids. The particular consecutiveness of amino acids in a protein, also referred to as *primary structure*, is encoded in the DNA of the organism [1]. This sequence is responsible for the formation of a stable and unique native conformation. The three-dimensional structure itself, also called the *tertiary structure*, determines the biological function of a protein. Anfinsen's refolding experiments [2] showed that the native conformation is *not* a result of the synthetisation process of the protein. Rather, it is an intrinsic property of the amino acid sequence. Therefore, a certain sequence of amino acids must uniquely lead to the three-dimensional, biologically relevant structure of the associated protein. I.e. physical forces, e.g. complex electrostatic and Van-der-Waals interactions between atoms, molecules, and the surrounding solvent, are primarily responsible for the native fold.

Since 2001, the whole human genome is sequenced [3]. According to the hitherto considerations, this information is sufficient for gaining an outstanding deep understanding of the human body, at least in principle. The consequences are of essential significance, e.g., for drug designing. Unfortunately, the connection of the primary structure and the tertiary structure is not yet understood. Since proteins are very complex macromolecules consisting of hundreds to thousands of atoms, the free-energy landscape of a protein is expected to be very rugged, with many local minima and, for stability and uniqueness, a deep, funnel-like global energy minimum [4]. The *folding* process of a protein into this minimum takes milliseconds to seconds and raises many questions. Currently, even with enormous computer power it is only possible to cover time scales of the order of nanoseconds in Molecular Dynamics simulations including all atoms of a protein.

Although it is known that characteristic atomic interactions are very important for protein folds, e.g., the formation of hydrogen bonds which are responsible for *secondary structures* (helices, sheets, hairpins), *coarse-grained* models can play an important role in understanding qualitative properties of complex heteropolymers. One of these coarse-grained models, the AB model [5, 6], is mainly studied within this work. However, it is slightly modified from stiff to flexible bonds for technical reasons when applying Molecular Dynamics simulations. The model and the mentioned alteration are motivated and described in chapter 1. Furthermore, the potential is differentiated to derive the forces which act in the system, which is necessary to apply Molecular Dynamics simulation techniques.

There are two big classes of computer simulations, which are extensively employed to study protein folding: Monte Carlo and Molecular Dynamics simulations. Both have assets and drawbacks. However, it is not clear whether the results of these different types of simulations are really comparable, since the dynamics of the employed algorithms is significantly different. One of the main goals of the work at hand is to thoroughly compare the outcome of Monte Carlo and Molecular Dynamics simulations.

Unlike Monte Carlo, Molecular Dynamics in its simplest form leaves the total energy of a system constant. Thus, the simulation would be carried out in the *microcanonical ensemble*. For the comparison with Monte Carlo, the concept of *temperature* has to be introduced, which leads to a simulation in the *canonical ensemble*, with given volume V, number of particles N, and temperature T (the so-called NVT ensemble). For this purpose, *thermostat* algorithms are applied to the classic simulation. Chapter 2 first explains the basic principles of Molecular Dynamics at finite temperature. Afterwards, two common *thermostats*, the Andersen thermostat [7] and the Nosé-Hoover-Chain thermostat [8, 9, 10] are explained and applied to simple model systems for testing purposes.

Over the years, a variety of sophisticated improvements of the simplest Monte Carlo method, the *Metropolis* algorithm [11], has been developed. This work utilises the so-called *parallel tempering* approach [12], which is briefly introduced in chapter 3. Furthermore, important concepts in the analysis of statistical data are treated therein. Finally, quantities are defined, which are important when investigating the AB protein model.

Chapter 4 covers the detailed examination of the model system with both Monte Carlo and Molecular Dynamics simulations. After the adjustment of the Monte Carlo method and an analysis of the statistical properties, several investigations are carried out concerning the general thermodynamic behaviour of the model system and the impact of certain parameters. Afterwards, similar computer experiments are performed with Molecular Dynamics by applying a preliminarily adjusted thermostat. The initial question of the comparableness of Monte Carlo and Molecular Dynamics is investigated concerning thermodynamic quantities. Furthermore, general statements about dynamics and time scales of the two classes of simulations with respect to the considered systems are made. Finally, the structural behaviour of the model in Molecular Dynamics is examined by studying free-energy landscapes.

From experiment it is known that three successive bonds in polymers are likely to have certain alignments. In particular, there are preferred values for the *torsion angle* formed by such three bonds. It is possible to introduce an additional potential term in the considered model to take these experiences into account. It turns out that while the integration of this additional potential does not yield any problem in the Monte Carlo simulation, it implies serious difficulties in Molecular Dynamics. An in-depth description of torsion and the implementation to the already applied simulation methods can be found in chapter 5. Afterwards, the extended system is studied with Monte Carlo methods. Again, the most important aspect is the comparability of the results from Monte Carlo and Molecular Dynamics, which concludes the considerations.

In the concluding summary, the most important findings of the work are presented. Finally, interesting starting points for further investigations are pointed out.

Chapter 1

The “AB” Heteropolymer Model

In this chapter, the mainly considered coarse-grained protein “AB” model shall be introduced. Since all forthcoming considerations like the applied simulation techniques or the measured quantities strongly depend on the type of examined system, this introduction is brought forward to the beginning of the thesis. After describing the involved potential energy contributions, the technical details like the cartesian transcription of the energy terms and the calculation of the forces for the Molecular Dynamics simulations are approached.

1.1 Introduction

The investigation of protein folding has been a challenging topic of research for decades. Up to now, real proteins are far too complex to involve every aspect in a whole folding simulation, although the computer power is growing exponentially. Therefore, researchers have always considered simplified models which are suitable for the available computer power and implied some major properties of proteins. One of these simplified models is the HP-model, which was first described by Dill [13]. It features a chain of two types of monomers, where the H type depicts a hydrophobic amino acid and the P type is a hydrophilic amino acid. The chain is simulated on a lattice. The hydrophobic and hydrophilic effect is simply induced by energetically favouring configurations, where non-bonded H monomers are residing on neighbouring lattice sites. Therefore it can be expected (and is observed) that for low temperatures a core of H monomers evolves, which is surrounded by P monomers. In nature, hydrophobic amino acids will also form a dense core and the hydrophilic amino acids will provide a shell. However, this effect is not intrinsically induced, but by the presence of the surrounding medium: water.

The introduction of the AB model by Stillinger [5, 6] made it possible to examine the properties of systems similar to the HP-model, which are not restricted to a lattice. Again, the system consists of a chain of two types of monomers: hydrophobic (A) and hydrophilic (B). Next neighbours along the chain are considered to be chemically bound and have a fixed distance (unit length for simplicity). In the original papers, the model was discussed in two dimensions. But without any alteration, it could be adapted to three dimensions [14, 15]. The adaption of the model which is utilised within this work, is described in detail in the following.

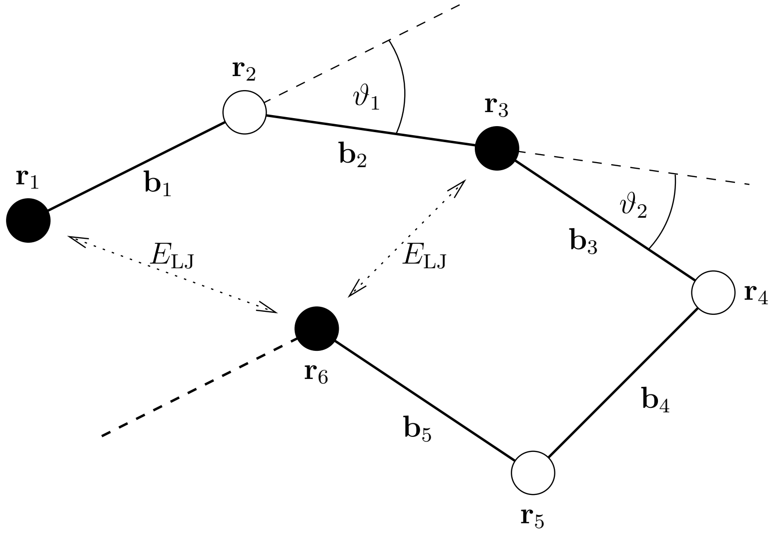


Figure 1.1: Sketch of a heteropolymer in two dimensions. The two types of monomers of the AB model are drawn with hollow and solid circles. The chemical bonds are the thick solid lines. The correct measurement of the bond angles is denoted. The symbols of the monomer positions ($\mathbf{r}_i$), bond vectors ($\mathbf{b}_i$) and bond angles (ϑ_i) are given. Furthermore, the action of the Lennard-Jones potential contribution described in the text is indicated for two cases.

1.2 Detailed Description

As already mentioned, in the AB model a chain of two types of monomers – A and B – is considered. The number of monomers is denoted with N in the following. In Table 1.I some sequences are given, which have been already investigated in several references and were thus chosen to do the studies in this work.

Figure 1.1 visualises such a chain in two dimensions for the purpose of clarity, but the scheme can be adapted to three dimensions without alteration. The position of each monomer $i \in \{1, \dots, N\}$ is written as $\mathbf{r}_i$ in the following. The whole structural information about a configuration is given, if all position vectors $\mathbf{r}_i$ are known. A certain structure is denoted with $\mathbf{X} = (\mathbf{r}_1, \dots, \mathbf{r}_N)$.

Two successive monomers are considered to be chemically bound. In nature, these bonds are peptide bonds. Since each monomer in the model stands for a whole amino acid, the whole peptide bond is described by one bond vector between two successive monomers. The $N-1$ bond vectors are $\mathbf{b}_i = \mathbf{r}_{i+1} - \mathbf{r}_i$. The length of these bonds will be simply referred to as b_i. In the original work all bond lengths b_i are fixed to unit length. In the work at hand, this is *not* the case. The reason for this modification will be explained later. To point out this difference, the total potential is denoted with V_{ABFB} below, where ABFB means “AB model with Flexible Bonds”.

The angle between two successive bond vectors is referred to as *bond angle*, and the symbol is $\vartheta_i = \angle(\mathbf{b}_i, \mathbf{b}_{i+1})$. Straightforward, the number of bond vectors is $N - 2$.

The total potential energy of the model consists of several contributions, which will be explained in detail in the following:

$$V_{\mathrm{ABFB}}(\mathbf{X}) = V_{\mathrm{bond}}(\mathbf{X}) + V_{\mathrm{bend}}(\mathbf{X}) + V_{\mathrm{LJ}}(\mathbf{X}) \ . \tag{1.1}$$

Flexible Bonds

The main difference between the classical AB model and the considered one in this work is, as already mentioned, the introduction of flexible bonds:

$$V_{\text{bond}}(\mathbf{X}) = \alpha_r \sum_{k=1}^{N-1} (b_k - r_0)^2 \ , \tag{1.2}$$

where α_r denotes the strength of the bond and r_0 the equilibrium bond length. In analogy to the original model the choice for the latter is $r_0 = 1$. The reason for this modification is that one of the goals of this project is to simulate the AB model with both Monte Carlo simulations (see chapter 3) and Molecular Dynamics (see chapter 2). In Monte Carlo simulations, the introduction of constraints – like fixed bond lengths – is not a big problem. This is explained in more detail in section 4.1. In Molecular Dynamics, the motion of the particles of a system is guided by the Newtonian forces, i.e. by the potential energy gradient. In cartesian coordinates (see section 1.4), the gradient of $V_{\text{bend}} + V_{\text{LJ}}$ of some monomer $\mathbf{r}_i$ will not be tangential with respect to a spherical shell around $\mathbf{r}_{i-1}$, i.e., the bond lengths will change in each step. The only analytical method to prevent this is to use a different type of coordinate system, where the bond lengths are intrinsically fixed, but this is extremely complicated. A short discussion of this issue can be found in section 5.3.2.

However, it is possible to introduce constraints in Molecular Dynamics simulations. It is inevitable to use an iterative algorithm after each MD time step to, e.g., keep the bond lengths fixed. The two most popular methods are the *Shake* [18] and the slightly improved *Rattle* [19] algorithm. As already stated, both require an iterative loop after each MD step and do thus not only seriously complicate the whole MD implementation, but also provoke an extensive slow-down.

Therefore, it was decided to replace the stiff bonds from the original model by strong, but flexible bonds. The impact of this alteration especially for large α_r (i.e. for rather strong bonds) is studied in more detail in the next section.

Table 1.I: Sequences, as they were already used in several references [15, 16, 17].

No.	Sequence	#A
20.1	BAAAAAABAAAABAABAABB	14
20.2	BAABAAAABABAABAAAAAB	14
20.3	AAAABBAAAABAABAAABBA	14
20.4	AAAABAABABAABBAAABAA	14
20.5	BAABBAAABBBABABAABAB	10
20.6	AAABBABBABABBABABABA	10
34	ABBABBABABBABBABABBABABBABBABABBAB	13
55	BABABBABABBABBABABBABABBABBABABBABBABABBABABBABBABABBAB	21

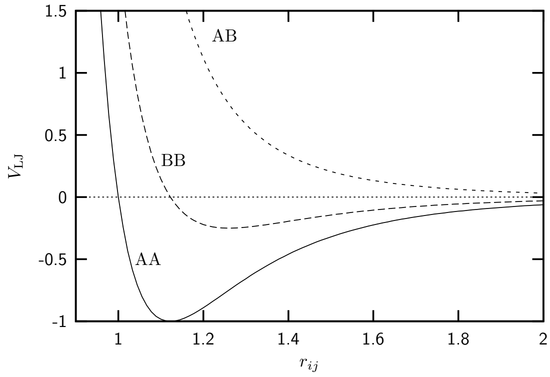


Figure 1.2: The Lennard-Jones potential is plotted for the three types of interaction as given in (1.5). If two monomers are of equal type, the interaction is attractive except for the steric repulsion for very small distances. Unequal monomers act completely repulsive.

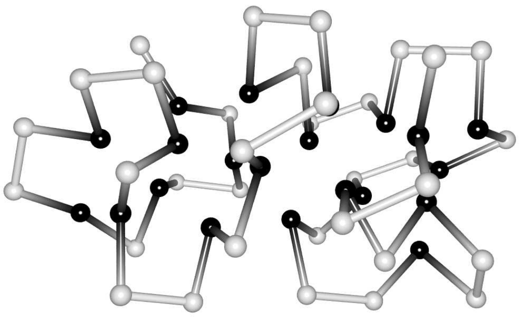


Figure 1.3: Minimal energy structure of the 55mer as listed in Table 1.I. Dark spheres depict A monomers, while B monomers are visualised by light spheres. The evident property of the AB model is that A monomers form a dense core, while B monomers provide a shell. This is comparable to hydrophobic and hydrophilic amino acids in real proteins.

Bending Potential

The AB model implies a potential in dependence of the bond angles ϑ_i:

$$V_{\rm bend}(\mathbf{R}) = \frac{1}{4}\sum_{k=1}^{N-2}\left(1-\cos\vartheta_k\right) \ . \tag{1.3}$$

With respect to a single bond angle, the domain is $V_{\rm bond} \in [0, 1/2]$. The energetically most preferable state is $\vartheta = 0$, i.e. the two bonds are parallel. Thus, the bending potential favours the elongated chain.

Lennard-Jones Interaction

The character of the AB model mainly originates from a Lennard-Jones-like interaction:

$$V_{\rm LJ}(\mathbf{R}) = 4\sum_{k=1}^{N-2}\sum_{l=k+2}^{N}\left(\frac{1}{r_{kl}^{12}} - \frac{C(\sigma_k,\sigma_l)}{r_{kl}^{6}}\right) \ , \tag{1.4}$$

$$C(\sigma_k,\sigma_l) = \begin{cases} +1, & \text{if } \sigma_k = \sigma_l = \text{A}, \\ +1/2, & \text{if } \sigma_k = \sigma_l = \text{B}, \\ -1/2, & \text{if } \sigma_k \neq \sigma_l \ . \end{cases} \tag{1.5}$$

It acts between any two monomers which are not bound, i.e. which are not next neighbours in the chain (in Fig. 1.1 only the A-A interactions are drawn). The function $C(\sigma_k, \sigma_l)$ alters the potential with respect to the types of the two monomers.

Figure 1.2 shows the three cases of interaction. The potential is attractive for monomers of equal type and repulsive otherwise. However, the fact that two A monomers attract each other more than two B monomers is the cause for the arising low energy structures. The A monomers form a dense core, while the B monomers arrange like an outer shell. Exemplary, Fig. 1.3 shows the structure with the lowest potential energy that has been found during a parallel tempering simulation (see chapter 3) of the sequence with 55 monomers from Table 1.I.

For a deeper analysis of the arising low energy structures it is expedient to evaluate the minima of V_{LJ}, and the distances where they are found:

$$0 \overset{!}{=} \frac{\partial}{\partial r} 4\left(r^{-12} - Cr^{-6}\right)\bigg|_{r=r_{\min}} = 4\left(-12 r_{\min}^{-13} + 6C r_{\min}^{-7}\right) \ ,$$

$$\Rightarrow \qquad r_{\min} = \left(\frac{2}{C}\right)^{\frac{1}{6}} \ , \tag{1.6}$$

$$V_{\mathrm{LJ}}(r_{\min}) = 4\left(\frac{C^2}{4} - \frac{C^2}{2}\right) = -C^2 \ . \tag{1.7}$$

For a A-A interaction, the minimum is thus $V_{\mathrm{LJ}}(r_{\min}) = -1$ at $r_{\min} = 2^{\frac{1}{6}} \approx 1.12$. For a B-B interaction, the minimum is $V_{\mathrm{LJ}}(r_{\min}) = -0.25$ at $r_{\min} = 4^{\frac{1}{6}} \approx 1.26$.

1.3 Influence of Flexible Bonds

As explained, the flexible bonds are introduced to get around technical problems that constraints induce in a Molecular Dynamics simulation. Therefore, the impact of V_{bond} in general has to be taken as artificial. Thus, it is interesting, what the particular effect of the flexible bonds is, compared to the case with fixed bonds. The hope is that for strong bonds (large α_r), the behaviour of the whole system is similar to the fixed bond case.

Although the flexible bonds are realized by introducing a harmonic bond potential, the particular case is different from the one-dimensional harmonic oscillator as calculated in section 2.3: V_{bond} denotes a three-dimensional harmonic oscillator. Therefore, the contribution of one bond to the heat capacity is not $C_V = k_B$ like for the one-dimensional case (compare (2.37)). By assuming independent bonds, it is enough to treat one bond:

$$\begin{aligned} Z_{\mathrm{bond}} &= \int\limits_{-\infty}^{\infty} \mathrm{d}b_x \int\limits_{-\infty}^{\infty} \mathrm{d}b_y \int\limits_{-\infty}^{\infty} \mathrm{d}b_z \, \exp\left[-\beta\alpha_r\left(r_0 - \sqrt{b_x^2 + b_y^2 + b_z^2}\right)^2\right] \\ &= \underbrace{\int\limits_{0}^{2\pi} \mathrm{d}\varphi \int\limits_{0}^{\pi} \mathrm{d}\vartheta \, \sin\vartheta}_{4\pi} \int\limits_{0}^{\infty} \mathrm{d}r \, r^2 \exp\left[-\beta\alpha_r\left(r_0 - r\right)^2\right] \\ &= \frac{\pi}{\alpha_r^2\beta^2}\left[2\alpha_r\beta r_0 \exp\left[-\alpha_r\beta r_0^2\right] + \sqrt{\pi\alpha_r\beta}(1 + 2\alpha_r\beta r_0^2)(1 + \mathrm{erf}\left[\sqrt{\alpha_r\beta r_0^2}\right])\right] \ , \end{aligned} \tag{1.8}$$

$$\begin{aligned} C_{V,\mathrm{bond}} &= k_B\beta^2 \frac{\partial^2}{\partial\beta^2} \ln Z_{\mathrm{bond}} \\ &= k_B \frac{h(x)\left[10\sqrt{x} + 4x^{3/2} + h(x)\left(3 + 4x(3 + x)\right)\right]}{2\left[2\sqrt{x} + h(x)\left(1 + 2x\right)\right]^2} \ , \\ x &= \alpha_r\beta r_0^2 \ , \qquad h(x) = e^x\sqrt{\pi}(1 + \mathrm{erf}\left[\sqrt{x}\right]) \ . \end{aligned} \tag{1.9}$$

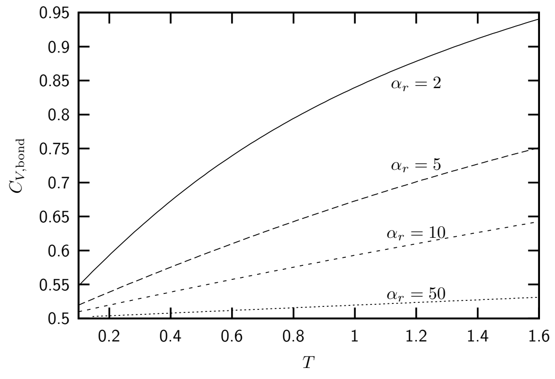


Figure 1.4: Analytic result of the specific heat contribution of V_{bond} according to (1.9) for different bond strength α_r, $r_0 = k_B = 1$. The asymptotic behaviour for rather strong bonds can be seen ($C_V \to k_B/2$).

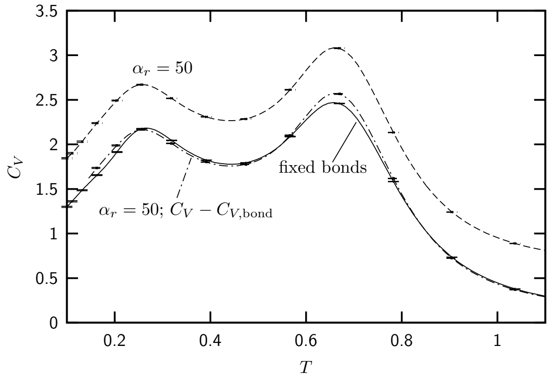


Figure 1.5: Specific heat of an AB model system with sequence 20.4 from Table 1.I (also see chapter 4). The simulation with fixed bonds is visualised with the solid line and flexible bonds ($\alpha_r = 50$) with the long-dashed line. The chain dotted line shows the subtraction of the bond contribution from the total specific heat for flexible bonds. The comparison with the solid line (fix bonds) shows only minor discrepancies.

The asymptotic behaviour for strong bonds, i.e. large α_r, is interesting. Because $\beta > 0$ and $r_0 > 0$ this means: $x = \alpha_r \beta r_0^2 \to \infty$. It holds:

$$\lim_{x\to\infty} \mathrm{erf}\left[\sqrt{x}\right] = 1 \ . \tag{1.10}$$

Considering only the leading contribution in each term, the asymptotic behaviour can be found:

$$\begin{aligned} h(x) &\overset{\alpha_r\to\infty}{\longrightarrow} 2\sqrt{\pi}e^x \\ C_{V,\mathrm{bond}} &\overset{\alpha_r\to\infty}{\longrightarrow} \frac{k_B}{2}\frac{h(x)\left(h(x)4x^2\right)}{(h(x)2x)^2} = \frac{k_B}{2} \ . \end{aligned} \tag{1.11}$$

It is interesting that even for an infinitely large bond strength, the fact that the bonds are systematically flexible can be clearly seen by an increase of the specific heat by $k_B/2$. Figure 1.4 visualises the outcome of (1.9) and (1.11). The stronger the bonds are, the closer is the specific heat to $k_B/2$ ($k_B = 1$ in this case). From Fig. 1.5 it can be seen that for large α_r the specific heat of the system with fixed bonds is reproduced acceptably, subtracting the pure contribution of the bonds to the specific heat. Therefore it is reasonable to examine the system with flexible bonds (with an adequate choice of α_r, e.g. $\alpha_r = 50$) and compare the data to results of simulations with conceptually rigid bonds. This topic is discussed in more detail in section 4.1.3.

1.4 Cartesian Formulation of the Potential Terms

In the Molecular Dynamics simulation the only significant information is the set of instantaneous monomer positions and velocities. The potential energy depends on the steric config-

uration **R** only. Thus all contributions of the potential energy and the corresponding forces must be expressed in terms of the position vectors of the monomers $\mathbf{r}_i$.
First the harmonic bond potential (1.2) is treated:

$$\begin{aligned} V_{\text{bond}}(\mathbf{R}) &= \alpha_r \sum_{k=1}^{N-1} (b_k - r_0)^2 = \alpha_r \sum_{k=1}^{N-1} \left(\sqrt{\mathbf{b}_k \cdot \mathbf{b}_k} - r_0\right)^2 \\ &= \alpha_r \sum_{k=1}^{N-1} \left(\sqrt{(\mathbf{r}_{k+1} - \mathbf{r}_k) \cdot (\mathbf{r}_{k+1} - \mathbf{r}_k)} - r_0\right)^2 \ . \end{aligned} \tag{1.12}$$

The Lennard-Jones like potential term (1.4) can be written as:

$$\begin{aligned} V_{\text{LJ}}(\mathbf{R}) &= 4 \sum_{k=1}^{N-2} \sum_{l=k+2}^{N} \left(\frac{1}{r_{kl}^{12}} - \frac{C(\sigma_k, \sigma_l)}{r_{kl}^{6}}\right) \\ &= 4 \sum_{k=1}^{N-2} \sum_{l=k+2}^{N} \left(\frac{1}{\sqrt{(\mathbf{r}_l - \mathbf{r}_k) \cdot (\mathbf{r}_l - \mathbf{r}_k)}^{12}} - \frac{C(\sigma_k, \sigma_l)}{\sqrt{(\mathbf{r}_l - \mathbf{r}_k) \cdot (\mathbf{r}_l - \mathbf{r}_k)}^{6}}\right) \\ &= 4 \sum_{k=1}^{N-2} \sum_{l=k+2}^{N} \frac{1}{((\mathbf{r}_l - \mathbf{r}_k) \cdot (\mathbf{r}_l - \mathbf{r}_k))^3} \left(\frac{1}{((\mathbf{r}_l - \mathbf{r}_k) \cdot (\mathbf{r}_l - \mathbf{r}_k))^3} - C(\sigma_k, \sigma_l)\right) \ . \end{aligned} \tag{1.13}$$

For the bond angle potential term (1.3) the following property of the scalar product of two vectors is used:

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}| \cos\left(\angle\left(\mathbf{a}, \mathbf{b}\right)\right) \ . \tag{1.14}$$

With this it is possible to rewrite the $\cos\vartheta$ as the normalised scalar product of the two successive bond vectors that enclose ϑ. The normalisation is necessary because, as mentioned above, the bond length is not fixed to unit length in this work:

$$\begin{aligned} V_{\text{bend}}(\mathbf{R}) &= \frac{1}{4} \sum_{k=1}^{N-2} (1 - \cos\vartheta_k) = \frac{1}{4} \sum_{k=1}^{N-2} \left(1 - \frac{\mathbf{b}_k \cdot \mathbf{b}_{k+1}}{\sqrt{(\mathbf{b}_k \cdot \mathbf{b}_k)(\mathbf{b}_{k+1} \cdot \mathbf{b}_{k+1})}}\right) \\ &= \frac{1}{4} \sum_{k=1}^{N-2} \left(1 - \frac{(\mathbf{r}_{k+1} - \mathbf{r}_k) \cdot (\mathbf{r}_{k+2} - \mathbf{r}_{k+1})}{\sqrt{((\mathbf{r}_{k+1} - \mathbf{r}_k) \cdot (\mathbf{r}_{k+1} - \mathbf{r}_k))\left((\mathbf{r}_{k+2} - \mathbf{r}_{k+1}) \cdot (\mathbf{r}_{k+2} - \mathbf{r}_{k+1})\right)}}\right) . \end{aligned} \tag{1.15}$$

1.5 Derivation of Forces

While a Monte Carlo Markov chain simulation is driven by the potential energy only, for a Molecular Dynamics simulation the Newtonian forces have to be known (see chapter 2), which are the negative gradients of the implied potential terms:

$$\mathbf{F} = -\nabla \mathcal{U}(\mathbf{r}^N) \ . \tag{1.16}$$

In the following, the forces of the previously introduced potential contributions shall be derived in Cartesian coordinates. Some preparations will help with the explicit calculations later. From section 1.4 it can be seen that all the potential terms are fully dependent on differences of position vectors of monomers. So while differentiating the potentials with respect to these position vectors, terms of the following forms will be frequently faced:

$$\nabla_{\mathbf{a}}\left(\mathbf{b}-\mathbf{a}\right)\cdot\left(\mathbf{b}-\mathbf{a}\right)=\nabla_{\mathbf{a}}\left(\mathbf{b}-\mathbf{a}\right)^2=-2\left(\mathbf{b}-\mathbf{a}\right)\ , \tag{1.17}$$

$$\nabla_{\mathbf{b}}\left(\mathbf{b}-\mathbf{a}\right)\cdot\left(\mathbf{b}-\mathbf{a}\right)=\nabla_{\mathbf{b}}\left(\mathbf{b}-\mathbf{a}\right)^2=2\left(\mathbf{b}-\mathbf{a}\right)\ , \tag{1.18}$$

$$\nabla_{\mathbf{a}}\left(\mathbf{b}-\mathbf{a}\right)\cdot\left(\mathbf{d}-\mathbf{c}\right)=-\left(\mathbf{d}-\mathbf{c}\right)\ , \tag{1.19}$$

$$\nabla_{\mathbf{b}}\left(\mathbf{b}-\mathbf{a}\right)\cdot\left(\mathbf{d}-\mathbf{c}\right)=\left(\mathbf{d}-\mathbf{c}\right)\ , \tag{1.20}$$

$$\nabla_{\mathbf{b}}\left(\mathbf{b}-\mathbf{a}\right)\cdot\left(\mathbf{c}-\mathbf{b}\right)=\left(\mathbf{c}-\mathbf{b}\right)-\left(\mathbf{b}-\mathbf{a}\right)\ . \tag{1.21}$$

While V_{bond} and V_{LJ} are two-body potentials, it is easy to see that V_{bend} is a three-body potential, since all bond angles incorporate the positions of three monomers. Therefore the positions of the monomers $i \in \{3,\dots,N-2\}$ arise in 3 terms of the whole sum, which leads to the conclusion that also the corresponding force will consist of three more or less independent terms. Analogously, it is expected that the bond potential and the Lennard-Jones potential will have the least complicated derivatives and would be a good point to start. The different terms of the force always arise from the fact that for different summands k of a potential, one specific monomer i can appear in a different position within the actual potential term. Thus the derivative has to be calculated with respect to every included monomer $k+n$.
First the bond potential contribution to the total force shall be derived:

$$\begin{aligned}
-\nabla_{\mathbf{r}_k}\left[\alpha_r\left(\sqrt{\left(\mathbf{r}_{k+1}-\mathbf{r}_k\right)^2}-r_0\right)^2\right] &= -\alpha_r 2\left(\sqrt{\mathbf{b}_k^2}-r_0\right)\frac{1}{2\sqrt{\mathbf{b}_k^2}}\left(-2\left(\mathbf{r}_{k+1}-\mathbf{r}_k\right)\right)\\
&= 2\alpha_r\left(\sqrt{\mathbf{b}_k^2}-r_0\right)\frac{1}{\sqrt{\mathbf{b}_k^2}}\mathbf{b}_k\\
&= 2\alpha_r\mathbf{b}_k\left(1-\frac{r_0}{b_k}\right)\ , \qquad (1.22)\\
-\nabla_{\mathbf{r}_{k+1}}\left[\alpha_r\left(\sqrt{\left(\mathbf{r}_{k+1}-\mathbf{r}_k\right)^2}-r_0\right)^2\right] &= -\alpha_r 2\left(\sqrt{\mathbf{b}_k^2}-r_0\right)\frac{1}{2\sqrt{\mathbf{b}_k^2}}\left(2\left(\mathbf{r}_{k+1}-\mathbf{r}_k\right)\right)\\
&= -2\alpha_r\left(\sqrt{\mathbf{b}_k^2}-r_0\right)\frac{1}{\sqrt{\mathbf{b}_k^2}}\mathbf{b}_k\\
&= -2\alpha_r\mathbf{b}_k\left(1-\frac{r_0}{b_k}\right)\ . \qquad (1.23)
\end{aligned}$$

The expressions in (1.22) and (1.23) are similar except for the algebraic sign. This property corresponds to the *“Actio=Reactio”* principle of the Newtonian mechanics with respect to the two involved monomers for every bond. It can be used later to speed up the calculation in the implementation. Actually only one kind of term has to be evaluated. Summing up both equations with $i=k$ and $i=k+1$ respectively, the bond force acting on monomer i can be calculated:

$$\mathbf{F}_{\mathrm{bond}\,i}=2\alpha_r\Bigg(\underbrace{\mathbf{b}_i\left(1-\frac{r_0}{b_i}\right)}_{i\leq N-1}-\underbrace{\mathbf{b}_{i-1}\left(1-\frac{r_0}{b_{i-1}}\right)}_{i\geq 2}\Bigg)\ . \tag{1.24}$$

Now the Lennard-Jones part of the potential shall be tackled:

$$
\begin{aligned}
&-\nabla_{\mathbf{r}_k}\left[4\left(\left((\mathbf{r}_l-\mathbf{r}_k)\cdot(\mathbf{r}_l-\mathbf{r}_k)\right)^{-6}-C(\sigma_k,\sigma_l)\left((\mathbf{r}_l-\mathbf{r}_k)\cdot(\mathbf{r}_l-\mathbf{r}_k)\right)^{-3}\right)\right]\\
&\quad=-4\left(-6\left((\mathbf{r}_l-\mathbf{r}_k)^2\right)^{-7}-C(\sigma_k,\sigma_l)\left(-3\left((\mathbf{r}_l-\mathbf{r}_k)^2\right)^{-4}\right)\right)\left(-2\left(\mathbf{r}_l-\mathbf{r}_k\right)\right)\\
&\quad=-4\left(\mathbf{r}_l-\mathbf{r}_k\right)\frac{6}{\left((\mathbf{r}_l-\mathbf{r}_k)^2\right)^4}\left(\frac{2}{\left((\mathbf{r}_l-\mathbf{r}_k)^2\right)^3}-C(\sigma_k,\sigma_l)\right)\ , \qquad (1.25)\\
&-\nabla_{\mathbf{r}_l}\left[4\left(\left((\mathbf{r}_l-\mathbf{r}_k)\cdot(\mathbf{r}_l-\mathbf{r}_k)\right)^{-6}-C(\sigma_k,\sigma_l)\left((\mathbf{r}_l-\mathbf{r}_k)\cdot(\mathbf{r}_l-\mathbf{r}_k)\right)^{-3}\right)\right]\\
&\quad=-4\left(-6\left((\mathbf{r}_l-\mathbf{r}_k)^2\right)^{-7}-C(\sigma_k,\sigma_l)\left(-3\left((\mathbf{r}_l-\mathbf{r}_k)^2\right)^{-4}\right)\right)\left(2\left(\mathbf{r}_l-\mathbf{r}_k\right)\right)\\
&\quad=4\left(\mathbf{r}_l-\mathbf{r}_k\right)\frac{6}{\left((\mathbf{r}_l-\mathbf{r}_k)^2\right)^4}\left(\frac{2}{\left((\mathbf{r}_l-\mathbf{r}_k)^2\right)^3}-C(\sigma_k,\sigma_l)\right)\ . \qquad (1.26)
\end{aligned}
$$

Again the only difference between (1.25) and (1.26) is the algebraic sign and only one term of the specific kind will have to be calculated for each pair of non-bonded monomers in the implementation, although it is still important to correctly collect the contributions to the Lennard-Jones force for each monomer coming out of (1.25) and (1.26). Calculating the force acting on monomer i means that in (1.25) k is fixed to $k=i$. Thus the first sum in (1.13) is reduced to one summand $k=i$, whereas the second sum over l remains. The opposite is the case for (1.26), where $l=i$:

$$
\begin{aligned}
\mathbf{F}_{\mathrm{LJ}\,i}&=4\Bigg(\underbrace{\sum_{j=1}^{i-2}\tilde{\mathbf{F}}_{\mathrm{LJ}\,ji}}_{i\geq 3}-\underbrace{\sum_{j=i+2}^{N}\tilde{\mathbf{F}}_{\mathrm{LJ}\,ij}}_{i\leq N-2}\Bigg)\ ,\\
\tilde{\mathbf{F}}_{\mathrm{LJ}\,ij}&=(\mathbf{r}_j-\mathbf{r}_i)\frac{6}{\left((\mathbf{r}_j-\mathbf{r}_i)^2\right)^4}\left(\frac{2}{\left((\mathbf{r}_j-\mathbf{r}_i)^2\right)^3}-C(\sigma_j,\sigma_i)\right)\ . \qquad (1.27)
\end{aligned}
$$

Now the gradient of the three-body bending potential term is of interest:

$$
\begin{aligned}
&-\nabla_{\mathbf{r}_k}\left[\frac{1}{4}\left(1-\frac{(\mathbf{r}_{k+1}-\mathbf{r}_k)\cdot(\mathbf{r}_{k+2}-\mathbf{r}_{k+1})}{\sqrt{(\mathbf{r}_{k+1}-\mathbf{r}_k)^2(\mathbf{r}_{k+2}-\mathbf{r}_{k+1})^2}}\right)\right]\\
&\quad=-\frac{1}{4}\left(\frac{-(\mathbf{r}_{k+2}-\mathbf{r}_{k+1})}{\sqrt{\mathbf{b}_k^2\mathbf{b}_{k+1}^2}}-\frac{\mathbf{b}_k\cdot\mathbf{b}_{k+1}}{2\sqrt{\mathbf{b}_k^2\mathbf{b}_{k+1}^2}^{3}}\left(-2\left(\mathbf{r}_{k+1}-\mathbf{r}_k\right)\mathbf{b}_{k+1}^2\right)\right)\\
&\quad=-\frac{1}{4}\frac{1}{b_k}\left(\frac{\mathbf{b}_k\cdot\mathbf{b}_{k+1}}{b_k b_{k+1}}\frac{\mathbf{b}_k}{b_k}-\frac{\mathbf{b}_{k+1}}{b_{k+1}}\right)=\tilde{\mathbf{F}}_{\mathrm{bend}\,1\,k}\ , \qquad (1.28)
\end{aligned}
$$

$$\begin{aligned}
&- \nabla_{\mathbf{r}_{k+2}} \left[\frac{1}{4}\left(1 - \frac{(\mathbf{r}_{k+1}-\mathbf{r}_k)\cdot(\mathbf{r}_{k+2}-\mathbf{r}_{k+1})}{\sqrt{(\mathbf{r}_{k+1}-\mathbf{r}_k)^2(\mathbf{r}_{k+2}-\mathbf{r}_{k+1})^2}}\right)\right] \\
&\qquad = -\frac{1}{4}\left(\frac{(\mathbf{r}_{k+1}-\mathbf{r}_k)}{\sqrt{\mathbf{b}_k^2\mathbf{b}_{k+1}^2}} - \frac{\mathbf{b}_k\cdot\mathbf{b}_{k+1}}{2\sqrt{\mathbf{b}_k^2\mathbf{b}_{k+1}^2}^3}\left(\mathbf{b}_k^2\, 2\,(\mathbf{r}_{k+2}-\mathbf{r}_{k+1})\right)\right) \\
&\qquad = -\frac{1}{4}\frac{1}{b_{k+1}}\left(\frac{\mathbf{b}_k}{b_k} - \frac{\mathbf{b}_k\cdot\mathbf{b}_{k+1}}{b_k b_{k+1}}\frac{\mathbf{b}_{k+1}}{b_{k+1}}\right) = \tilde{\mathbf{F}}_{\mathrm{bend}\,2\,k}\ ,
\end{aligned} \tag{1.29}$$

$$\begin{aligned}
&- \nabla_{\mathbf{r}_{k+1}} \left[\frac{1}{4}\left(1 - \frac{(\mathbf{r}_{k+1}-\mathbf{r}_k)\cdot(\mathbf{r}_{k+2}-\mathbf{r}_{k+1})}{\sqrt{(\mathbf{r}_{k+1}-\mathbf{r}_k)^2(\mathbf{r}_{k+2}-\mathbf{r}_{k+1})^2}}\right)\right] \\
&\qquad = -\frac{1}{4}\left(\frac{\mathbf{b}_{k+1}-\mathbf{b}_k}{\sqrt{\mathbf{b}_k^2\mathbf{b}_{k+1}^2}} - \frac{\mathbf{b}_k\cdot\mathbf{b}_{k+1}}{2\sqrt{\mathbf{b}_k^2\mathbf{b}_{k+1}^2}^3}\left(2\mathbf{b}_k\mathbf{b}_{k+1}^2 - 2\mathbf{b}_{k+1}\mathbf{b}_k^2\right)\right) \\
&\qquad = -\left(\tilde{\mathbf{F}}_{\mathrm{bend}\,1\,k} + \tilde{\mathbf{F}}_{\mathrm{bend}\,2\,k}\right)\ .
\end{aligned} \tag{1.30}$$

It makes sense that the force on monomer $k+1$ coming from the k^{th} bond angle is equal to the sum of the negative corresponding forces on monomer k and $k+2$. (1.28) – (1.30) result in the total bending force acting on monomer i:

$$\mathbf{F}_{\mathrm{bend}\,i} = \underbrace{\tilde{\mathbf{F}}_{\mathrm{bend}\,1\,i}}_{i\leq N-2} - \underbrace{\left(\tilde{\mathbf{F}}_{\mathrm{bend}\,1\,(i-1)} + \tilde{\mathbf{F}}_{\mathrm{bend}\,2\,(i-1)}\right)}_{i\in[2,N-1]} + \underbrace{\tilde{\mathbf{F}}_{\mathrm{bend}\,2\,(i-2)}}_{i\geq 3}\ , \tag{1.31}$$

with the definitions for $\tilde{\mathbf{F}}_{\mathrm{bend}\,1\,k}$ and $\tilde{\mathbf{F}}_{\mathrm{bend}\,2\,k}$ from (1.28) and (1.29).

Chapter 2

Molecular Dynamics at Finite Temperature

The basic idea behind Molecular Dynamics is to numerically integrate the Newtonian equations of motion. From that point of view Molecular Dynamics should be capable to visualise the exact evolution in time of the considered system. However, in practice it is hard to verify that the trajectory calculated with a Molecular Dynamics simulation is similar to the behaviour the system would show in reality. This is due to numerical errors during the simulation on the one hand and crucially linked to the utilised algorithm on the other. Therefore, Molecular Dynamics at finite temperature is to be seen as another computer-aided method of *statistical* mechanics.

To gain experience in Molecular Dynamics, it was expedient to test some of the common algorithms and get familiar with its behaviour. For this purpose the one-dimensional harmonic oscillator was chosen as a trial system. Although it is a very simple system and does not hold any potential energy traps, it has the big advantage that it is possible to calculate the exact solution for most of the dynamic and thermodynamic quantities. This makes it perfect for checking the correctness of results from certain computer simulations. For a more thorough investigation, several simulations were carried out with the quartic double well potential. It is also possible to analytically calculate solutions for several dynamic and thermodynamic characteristics for this system. Still rather simple it "provides" a potential energy barrier, which is illuminative for testing algorithms.

The first section is a collection of detailed thoughts about a frequently used approach to derive Molecular Dynamics algorithms. Afterwards some exemplary microcanonical simulations are performed with an algorithm that is obtained in this way - the Störmer-Verlet algorithm. In the third part, thermodynamic quantities of the already mentioned testing systems are calculated analytically. Thereafter two common approaches for thermostating – the Andersen and the Nosé-Hoover algorithm – are described the and simulations are carried out utilising it to obtain the canonical ensemble.

2.1 Liouville Operator and Trotter Factorisation

Before starting to discuss several simulation methods in detail, a brief introduction to Liouville's formulation of classical mechanics shall be given. In combination with the Trotter identity this approach can be used as a general tool to derive algorithms for the solution of coupled differential equations of motion. There are several very good references, where this topic is explained [20, 21, 22]. However, some detailed aspects shall be considered here.

2.1.1 Liouville Formalism

Let $f(\mathbf{X}(t))$ be some general, well-behaved function without an explicit time dependency, where $\mathbf{X}(t)$ is the phase space vector of a system at time t. Since f is only implicitly dependent on t, the total derivative with respect to time can written as:

$$\frac{\mathrm{d}}{\mathrm{d}t} f = \underbrace{\frac{\partial}{\partial t} f}_{\equiv 0} + \frac{\partial \mathbf{X}}{\partial t} \cdot \frac{\partial}{\partial \mathbf{X}} f = \dot{\mathbf{X}} \cdot \frac{\partial}{\partial \mathbf{X}} f = \imath \mathcal{L} f \ . \tag{2.1}$$

The "·" denotes a formal scalar product. In a molecular dynamics simulation the interesting function f is the trajectory of the system $\mathbf{X}(t)$ itself. The definition of the Liouville operator $\imath\mathcal{L}$ can be extracted from (2.1):

$$\imath\mathcal{L} = \dot{\mathbf{X}} \cdot \frac{\partial}{\partial \mathbf{X}} \ . \tag{2.2}$$

The $\imath$ is convention and has the effect of making $\mathcal{L}$ a Hermitian operator. For example the Liouville operator of a one-dimensional system would have the following form:

$$\mathbf{X}(t) = \begin{pmatrix} r(t) \\ p(t) \end{pmatrix} \qquad \Rightarrow \qquad \imath\mathcal{L} = \dot{r}\frac{\partial}{\partial r} + \dot{p}\frac{\partial}{\partial p} \ . \tag{2.3}$$

Generally it is possible to split $\imath\mathcal{L}$ into two or more parts. For the further considerations $\imath\mathcal{L}$ is separated into one partition containing the position-dependent terms and one covering the momenta:

$$\begin{aligned} \imath\mathcal{L} &= \imath\mathcal{L}_r + \imath\mathcal{L}_p \ , \\ \imath\mathcal{L}_r &= \dot{\mathbf{r}} \cdot \frac{\partial}{\partial \mathbf{r}} = \sum_{i=1}^{N'} \dot{r}_i \frac{\partial}{\partial r_i} = \sum_{i=1}^{N'} \imath\mathcal{L}_{r_i} \ , \\ \imath\mathcal{L}_p &= \dot{\mathbf{p}} \cdot \frac{\partial}{\partial \mathbf{p}} = \sum_{i=1}^{N'} \dot{p}_i \frac{\partial}{\partial p_i} = \sum_{i=1}^{N'} \imath\mathcal{L}_{p_i} \ . \end{aligned} \tag{2.4}$$

In (2.4) N' is the number of degrees of freedom of the considered system (e.g., $N' = dN$ for a system with N particles in d dimensions).

2.1.2 Propagator

Since $\imath\mathcal{L}$ does not explicitly depend on t, (2.1) can be formally integrated:

$$\begin{aligned} \frac{\mathrm{d}f}{\mathrm{d}t} &= \imath\mathcal{L} f & &\Rightarrow & \int \mathrm{d}f \, \frac{1}{f} &= \int_0^t \mathrm{d}t' \, \imath\mathcal{L} \ , \\ \ln f(\mathbf{X}(t)) + \mathcal{C}_{\text{initial}} &= \imath\mathcal{L} t & &\Rightarrow & f(\mathbf{X}(t)) &= e^{\imath\mathcal{L} t} f(\mathbf{X}(0)) \ . \end{aligned} \tag{2.5}$$

Because $\mathcal{L}$ depends on the derivatives of positions $\dot{\mathbf{r}}(t)$ and momenta $\dot{\mathbf{p}}(t)$, (2.5) is still to be considered as a system of coupled differential equations. It is just an integrated form of the Liouville equation (2.1). If (2.5) would be used to calculate the trajectory $\mathbf{X}(t)$ numerically starting from the initial values $\mathbf{X}(0)$, a systematic error of $\mathcal{O}(t^2)$ would arise from the fact that $\imath\mathcal{L}$ would be considered as temporally constant by just taking $\dot{\mathbf{r}}(0)$ and $\dot{\mathbf{p}}(0)$ into account:

$$\begin{aligned}
\int \mathrm{d}f \, \frac{1}{f} &= \int\limits_0^t \mathrm{d}t' \, \imath\mathcal{L}(t') = \int\limits_0^t \mathrm{d}t' \, \left(\imath\mathcal{L}_{\text{const.}}(0) + \mathcal{O}(t')\right) \ , \\
\ln f(\mathbf{X}(t)) + \mathcal{C}_{\text{initial}} &= \imath\mathcal{L}_{\text{const.}}(0)t + \mathcal{O}(t^2) \ , \\
f(\mathbf{X}(t)) &= e^{\imath\mathcal{L}_{\text{const.}}(0)t} f(\mathbf{X}(0)) + \mathcal{O}(t^2) \ . \qquad (2.6)
\end{aligned}$$

In (2.5) a propagator can be read off:

$$\mathcal{U}(t) = e^{\imath\mathcal{L}t} \ . \qquad (2.7)$$

Analogically it is possible to define two fractional propagators $\mathcal{U}_r(t)$ and $\mathcal{U}_p(t)$:

$$\mathcal{U}_r(t) = e^{\imath\mathcal{L}_r t} \ , \qquad \mathcal{U}_p(t) = e^{\imath\mathcal{L}_p t} \ . \qquad (2.8)$$

The impact of $\mathcal{U}_r$ on an initial state $\mathbf{X}(0)$ shall be exemplarily derived for a one-dimensional system. Therefore it is assumed that the derivatives $\dot{r}$ and $\dot{p}$ are formulated independently from r and p respectively. In that case $\partial/\partial r$ and $\dot{r}$ commutate and higher terms of the form $(\imath\mathcal{L}_r\delta t)^n = (\dot{r}(\partial/\partial r)\delta t)^n$ can be written as $\dot{r}^n\delta t^n\partial^n/\partial r^n$:

$$\begin{aligned}
\mathcal{U}_r(\delta t)\mathbf{X}(0) &= e^{\imath\mathcal{L}_r\delta t} \begin{pmatrix} r(t) \\ p(t) \end{pmatrix}\Bigg|_{t=0} \\
&= \begin{pmatrix} r(0) \\ p(0) \end{pmatrix} + \left(\dot{r}(t)\delta t \frac{\partial}{\partial r}\right) \begin{pmatrix} r(t) \\ p(t) \end{pmatrix}\Bigg|_{t=0} + \frac{1}{2}\left((\dot{r}(t)\delta t)^2 \frac{\partial^2}{\partial r^2}\right) \begin{pmatrix} r(t) \\ p(t) \end{pmatrix}\Bigg|_{t=0} + \ldots \\
&= \begin{pmatrix} r(0) \\ p(0) \end{pmatrix} + \begin{pmatrix} \dot{r}(0)\delta t \\ 0 \end{pmatrix} = \begin{pmatrix} r(0) + \dot{r}(0)\delta t \\ 0 \end{pmatrix} \ . \qquad (2.9)
\end{aligned}$$

The terms of order $\mathcal{O}((\imath\mathcal{L}_r\delta t)^2)$ and higher do not give a contribution due to the fact that the second derivative $(\partial^2/\partial r^2)r$ vanishes:

$$\mathcal{U}_r(\delta t)\mathbf{X}(0) \ : \qquad \mathbf{r}(0) \longrightarrow \mathbf{r}(0) + \dot{\mathbf{r}}(0)\delta t = \mathbf{r}(0) + \dot{\mathbf{r}}\left(\mathbf{p}(0)\right)\delta t \ . \qquad (2.10)$$

Thus, the application of $\mathcal{U}_r(\delta t)$ shifts $\mathbf{r}$. This result is exact and does hence *not* suffer from the problem described in (2.6). It makes sense that $\mathcal{U}_r$ propagates $\mathbf{r}$ for all times with $\dot{\mathbf{r}}(0)$, since $\mathcal{U}_r$ has no influence on the momenta. So if the momentum does not change, the velocity remains constant. Therefore, $\imath\mathcal{L}_r$ is considered constant in that sense. Thus the error $\mathcal{O}(t^2)$ does not occur. An analogous outcome is obtained for $\mathcal{U}_p(\delta t)$:

$$\mathcal{U}_p(\delta t)\mathbf{X}(0) \ : \qquad \mathbf{p}(0) \longrightarrow \mathbf{p}(0) + \dot{\mathbf{p}}(0)\delta t = \mathbf{p}(0) + \dot{\mathbf{p}}\left(\mathbf{r}(0)\right)\delta t \ . \qquad (2.11)$$

2.1.3 Factorisation

It is important to note that, for the following proposal, the time derivatives of the positions $\dot{\mathbf{r}}$ may only depend on the conjugate momenta $\mathbf{p}$, which does not seem to be a strict requirement. Actually it holds $\dot{\mathbf{r}} = \mathbf{p}/m$. Vice versa the time derivatives of the conjugate momenta $\dot{\mathbf{p}} \sim \mathbf{F}$ must not depend on anything else but the positions $\mathbf{r}$. That is, terms like friction will lead to difficulties in this approach. Starting from these assumptions it can be shown that the already mentioned problems with higher order terms arise considering the full propagator $\mathcal{U}(t)$:

$$
\begin{aligned}
\mathcal{U}(\delta t)\mathbf{X}(0) &= e^{\imath\mathcal{L}\delta t}\begin{pmatrix} r(t) \\ p(t)\end{pmatrix}\Bigg|_{t=0} \\
&= \begin{pmatrix} r(0) \\ p(0)\end{pmatrix} + \underbrace{\left(\dot{r}\frac{\partial}{\partial r} + \dot{p}\frac{\partial}{\partial p}\right)\begin{pmatrix} r(t) \\ p(t)\end{pmatrix}\Bigg|_{t=0}}_{\begin{pmatrix} \dot{r}(0) \\ \dot{p}(0)\end{pmatrix}} \delta t + \frac{1}{2}\left(\dot{r}\frac{\partial}{\partial r} + \dot{p}\frac{\partial}{\partial p}\right)^2 \begin{pmatrix} r(t) \\ p(t)\end{pmatrix}\Bigg|_{t=0} \delta t^2 + \ldots \\
&= \begin{pmatrix} r(0) \\ p(0)\end{pmatrix} + \begin{pmatrix} \dot{r}(0) \\ \dot{p}(0)\end{pmatrix}\delta t + \frac{\delta t^2}{2}\Bigg[\underbrace{\dot{r}\frac{\partial}{\partial r}\dot{r}\frac{\partial}{\partial r}\begin{pmatrix} r(t) \\ p(t)\end{pmatrix} + \dot{p}\frac{\partial}{\partial p}\dot{p}\frac{\partial}{\partial p}\begin{pmatrix} r(t) \\ p(t)\end{pmatrix}}_{\equiv 0} \\
&\qquad + \underbrace{\dot{r}\frac{\partial}{\partial r}\dot{p}\frac{\partial}{\partial p}\begin{pmatrix} r(t) \\ p(t)\end{pmatrix}}_{=\dot{r}\frac{\partial \dot{p}}{\partial r} = \dot{r}\frac{\partial F(r)}{\partial r}} + \underbrace{\dot{p}\frac{\partial}{\partial p}\dot{r}\frac{\partial}{\partial r}\begin{pmatrix} r(t) \\ p(t)\end{pmatrix}}_{=\dot{p}\frac{\partial \dot{r}}{\partial p} = \dot{p}\frac{1}{m} = \ddot{r}}\Bigg]\Bigg|_{t=0} + \mathcal{O}(\delta t^3)\ . \qquad (2.12)
\end{aligned}
$$

By writing $\dot{r}$ as $\partial r/\partial t$ and $\dot{p}$ as $\partial p/\partial t$ (2.12) could be written more simply and the $\mathcal{O}(\delta t^3)$ term would vanish. But as postulated before, the derivatives $\dot{\mathbf{r}}$ and $\dot{\mathbf{p}}$ are considered to be functions formulated independently from $\mathbf{r}$ and $\mathbf{p}$, respectively. This is reasonable as long as there is no compact solution for $\mathbf{r}(t)$. During the numerical integration of the equations of motion in a simulation the only analytical correlation is $\dot{\mathbf{r}} = \mathbf{p}/m$.

Considering the result (2.12) from an algorithmic point of view, it is crucial to factorise $\mathcal{U}(t)$ in terms of $\mathcal{U}_r(t)$ and $\mathcal{U}_p(t)$. It is easy to see that the summands of $\imath\mathcal{L}_r$ from (2.4) do commutate pairwise, as well as the summands of $\imath\mathcal{L}_p$. However, pairs of terms from $\imath\mathcal{L}_r$ and $\imath\mathcal{L}_p$ do *not* commutate due to the dependencies described above:

$$
\begin{aligned}
\left[\imath\mathcal{L}_{r_i}, \imath\mathcal{L}_{p_j}\right] &= \left[\dot{r}_i\frac{\partial}{\partial r_i}, \dot{p}_j\frac{\partial}{\partial p_j}\right] = \dot{r}_i\left(\frac{\partial \dot{p}_j}{\partial r_i}\right)\frac{\partial}{\partial p_j} + \dot{r}_i\dot{p}_j\frac{\partial^2}{\partial r_i \partial p_j} - \dot{p}_j\left(\frac{\partial \dot{r}_i}{\partial p_j}\right)\frac{\partial}{\partial r_i} - \dot{p}_j\dot{r}_i\frac{\partial^2}{\partial p_j \partial r_i} \\
&= \dot{r}_i\left(\frac{\partial F_j}{\partial r_i}\right)\frac{\partial}{\partial p_j} - \left\{\begin{matrix} \dot{p}_i\left(\frac{1}{m}\right)\frac{\partial}{\partial r_i} & \text{if} & i = j \\ 0 & \text{if} & i \neq j \end{matrix}\right\} \neq 0\ , \qquad (2.13)
\end{aligned}
$$

i.e., the propagator $\mathcal{U}$ cannot be trivially factorised:

$$
\mathcal{U}(t) = e^{\imath\mathcal{L}t} = e^{\imath(\mathcal{L}_r + \mathcal{L}_p)t} \overset{!}{\neq} e^{\imath\mathcal{L}_r t} \times e^{\imath\mathcal{L}_p t} \qquad \Rightarrow \qquad \mathcal{U}(t) \neq \mathcal{U}_r(t) \times \mathcal{U}_p(t)\ . \qquad (2.14)
$$

For two non-commutating operators $\mathcal{A}$ and $\mathcal{B}$ there is the Trotter identity

$$e^{(\mathcal{A}+\mathcal{B})t} = e^{\mathcal{A}t/2} \times e^{\mathcal{B}t} \times e^{\mathcal{A}t/2} + \mathcal{O}(t^3) \ , \tag{2.15}$$

or more generally

$$\begin{aligned} \Rightarrow \qquad \exp\left[\sum_{i=1}^{n} \mathcal{A}_i t\right] &= e^{\mathcal{A}_1 t/2} \times \exp\left[\sum_{i=2}^{n} \mathcal{A}_i t\right] \times e^{\mathcal{A}_1 t/2} + \mathcal{O}(t^3) = \dots \\ &= \prod_{i=1}^{n-1} e^{\mathcal{A}_i t/2} \times e^{\mathcal{A}_n t} \times \prod_{i=n-1}^{1} e^{\mathcal{A}_i t/2} + \mathcal{O}(t^3) \ . \end{aligned} \tag{2.16}$$

In case of $\mathcal{A} = \imath\mathcal{L}_r$ and $\mathcal{B} = \imath\mathcal{L}_p$ the obtained formulae are well known as the velocity Störmer-Verlet algorithm. The correlation with the original formulation of Verlet [23] is shown in [20] in section 4.3. It shares a big advantage for implementation purposes with all algorithms that are derived in this way: it is time reversible and thus energy conserving. Actually it has been tested that the short-time energy drift of the Störmer-Verlet-algorithm is rather large, while the long-time drift is very small. The method has some self-healing behaviour which leads to outstanding stable trajectories.

To have a better imagination of why this form of factorisation is so powerful it shall be compared to the trivial one claimed as unprofitable in (2.14). Recalling the derivation of (2.9), terms with $\mathcal{A}^2 = (\imath\mathcal{L}_r)^2$ and $\mathcal{B}^2 = (\imath\mathcal{L}_p)^2$ on the right-hand side (i.e., acting first) do not contribute. Therefore the exact expansion of $\exp[(\mathcal{A}+\mathcal{B})t]$ is:

$$\begin{aligned} e^{(\mathcal{A}+\mathcal{B})t} &= 1 + (\mathcal{A}+\mathcal{B})t + \frac{1}{2}\left(\mathcal{A}\mathcal{B} + \mathcal{B}\mathcal{A}\right)t^2 \\ &\quad + \frac{1}{6}\left(\mathcal{A}^2\mathcal{B} + \mathcal{A}\mathcal{B}\mathcal{A} + \mathcal{B}\mathcal{A}\mathcal{B} + \mathcal{B}^2\mathcal{A}\right)t^3 + \mathcal{O}(t^4) \ . \end{aligned} \tag{2.17}$$

Obviously the trivial factorisation in (2.14) is wrong for order $\mathcal{O}(t^2)$ since one mixed term is missing while the other one has a factor 2 compared to the correct result:

$$e^{\mathcal{A}t} \times e^{\mathcal{B}t} = (1 + \mathcal{A}t) \times (1 + \mathcal{B}t) = 1 + (\mathcal{A}+\mathcal{B})t + \mathcal{A}\mathcal{B}t^2 \ . \tag{2.18}$$

The Trotter formula expands to:

$$\begin{aligned} e^{\mathcal{A}\frac{t}{2}} \times e^{\mathcal{B}t} \times e^{\mathcal{A}\frac{t}{2}} &= \left(1 + \mathcal{A}\frac{t}{2}\right) \times (1 + \mathcal{B}t) \times \left(1 + \mathcal{A}\frac{t}{2}\right) \\ &= 1 + \left(\frac{\mathcal{A}}{2} + \mathcal{B} + \frac{\mathcal{A}}{2}\right)t + \left(\frac{\mathcal{A}}{2}\mathcal{B} + \mathcal{B}\frac{\mathcal{A}}{2}\right)t^2 + \frac{\mathcal{A}\mathcal{B}\mathcal{A}}{4}t^3 \\ &= 1 + (\mathcal{A}+\mathcal{B})t + \frac{1}{2}\left(\mathcal{A}\mathcal{B} + \mathcal{B}\mathcal{A}\right)t^2 + \frac{1}{4}\mathcal{A}\mathcal{B}\mathcal{A}t^3 \ , \end{aligned} \tag{2.19}$$

which is obviously in agreement with (2.17) up to the order $\mathcal{O}(\delta t^2)$.

2.1.4 Time Reversibility

Again, an exemplified one-dimensional system shall be considered to clarify the fact of time reversibility of the formulation derived from the Trotter identity. First, the evolution of $\mathbf{X}(0)$

in time, δt, is treated as it is described in analogy to by the approximated time evolution operator:

$$\tilde{\mathcal{U}}(t) = \mathcal{U}_p(t) \times \mathcal{U}_r(t) = \exp[\imath L_p t] \times \exp[\imath L_r t] \ . \tag{2.20}$$

With (2.10) and (2.11) it can be calculated:

$$\mathcal{U}_r(\delta t)\mathbf{X}(0) = \begin{pmatrix} r(0) + \dot{r}(0)\delta t \\ p(0) \end{pmatrix} = \tilde{\mathbf{X}}(\delta t') \ , \tag{2.21}$$

$$\mathcal{U}_p(\delta t)\mathbf{X}(\delta t') = \begin{pmatrix} r(\delta t') \\ p(\delta t') \end{pmatrix} + \begin{pmatrix} 0 \\ \dot{p}(\delta t')\delta t \end{pmatrix} =^{\{p(\delta t')=p(0)\}} \begin{pmatrix} r(\delta t') \\ p(0) + \dot{p}(r(\delta t'))\delta t \end{pmatrix} = \tilde{\mathbf{X}}(\delta t) \ . \tag{2.22}$$

The resulting phase space vector $\tilde{\mathbf{X}}(\delta t)$ is indicated with a tilde, because it is only an approximation of the exact result $\mathbf{X}(\delta t)$ and depends on the utilised time evolution operator $\tilde{\mathcal{U}}$. If $\tilde{\mathcal{U}}(t)$ is time reversible, it would be possible to go back to $\mathbf{X}(0)$ by applying $\tilde{\mathcal{U}}(-\delta t)$ to $\tilde{\mathbf{X}}(\delta t)$. But as expected some different result is obtained by the arithmetics analogous to (2.21) and (2.22). This time the intermediate step is indicated by "*", not to confuse with the quantities from (2.21) marked with an inverted comma which represent a different state:

$$\begin{aligned}
\mathcal{U}_r(-\delta t)\tilde{\mathbf{X}}(\delta t) &= \begin{pmatrix} r(\delta t) \\ p(\delta t) \end{pmatrix} + \begin{pmatrix} \dot{r}(\delta t)(-\delta t) \\ 0 \end{pmatrix} =^{\{r(\delta t)=r(\delta t')\}} \begin{pmatrix} r(0) + \dot{r}(0)\delta t + \dot{r}(\delta t)(-\delta t) \\ p(\delta t) \end{pmatrix} \\
&= \begin{pmatrix} r(0) + (\dot{r}(0) - \dot{r}(\delta t))\delta t \\ p(\delta t) \end{pmatrix} = \tilde{\mathbf{X}}(\delta t^*) \ ,
\end{aligned} \tag{2.23}$$

$$\begin{aligned}
\mathcal{U}_p(-\delta t)\tilde{\mathbf{X}}(\delta t^*) &= \begin{pmatrix} r(\delta t^*) \\ p(\delta t^*) \end{pmatrix} + \begin{pmatrix} 0 \\ \dot{p}(\delta t^*)(-\delta t) \end{pmatrix} \\
&=^{\{p(\delta t^*)=p(\delta t)\}} \begin{pmatrix} r(\delta t^*) \\ p(0) + \dot{p}(r(\delta t'))\delta t + \dot{p}(r(\delta t^*))(-\delta t) \end{pmatrix} \\
&= \begin{pmatrix} r(0) + [\dot{r}(p(0)) - \dot{r}(p(\delta t))]\delta t \\ p(0) + [\dot{p}(r(\delta t')) - \dot{p}(r(\delta t^*))]\delta t \end{pmatrix} = \tilde{\mathbf{X}}(0) \overset{!}{\neq} \mathbf{X}(0) \ .
\end{aligned} \tag{2.24}$$

(2.24) shows that the propagation by $\tilde{\mathcal{U}}(t)$ is *not* time reversible, because doing one step forward and one backward, the system does not arrive at its initial conditions. Now the approach derived by the Trotter factorisation shall be tested. The corresponding time evolution operator is

$$\hat{\mathcal{U}}(t) = \mathcal{U}_r(t/2) \times \mathcal{U}_p(t) \times \mathcal{U}_r(t/2) = \exp[\imath L_r t/2] \times \exp[\imath L_p t] \times \exp[\imath L_r t/2] \ . \tag{2.25}$$

Again, the two intermediate steps in the forward direction are indicated with one and two inverted commas, while the backward direction is marked with one and two "*" respectively:

$$\mathcal{U}_r\left(\tfrac{\delta t}{2}\right)\mathbf{X}(0) = \begin{pmatrix} r(0) + \dot{r}(0)\frac{\delta t}{2} \\ p(0) \end{pmatrix} = \hat{\mathbf{X}}(\delta t')\ , \tag{2.26}$$

$$\mathcal{U}_p(\delta t)\hat{\mathbf{X}}(\delta t') = \begin{pmatrix} r(\delta t') \\ p(0) + \dot{p}(\delta t')\delta t \end{pmatrix} = \hat{\mathbf{X}}(\delta t'')\ , \tag{2.27}$$

$$\mathcal{U}_r\left(\tfrac{\delta t}{2}\right)\hat{\mathbf{X}}(\delta t'') = \begin{pmatrix} r(\delta t') + \dot{r}(\delta t'')\frac{\delta t}{2} \\ p(\delta t'') \end{pmatrix} = \begin{pmatrix} r(0) + (\dot{r}(0) + \dot{r}(\delta t''))\frac{\delta t}{2} \\ p(0) + \dot{p}(\delta t')\delta t \end{pmatrix} = \hat{\mathbf{X}}(\delta t)\ . \tag{2.28}$$

It is interesting that for the reverse time direction $\hat{\mathcal{U}}(-\delta t)$ the intermediate states do exactly agree with those from the forward propagation:

$$\begin{aligned} \mathcal{U}_r\left(-\tfrac{\delta t}{2}\right)\hat{\mathbf{X}}(\delta t) &= \begin{pmatrix} r(\delta t) + \dot{r}(p(\delta t))\left(-\frac{\delta t}{2}\right) \\ p(\delta t) \end{pmatrix} \\ &\overset{\{p(\delta t)=p(\delta t'')\}}{=} \begin{pmatrix} r(0) + (\dot{r}(0) + \dot{r}(p(\delta t'')) - \dot{r}(p(\delta t'')))\frac{\delta t}{2} \\ p(\delta t'') \end{pmatrix} \\ &= \begin{pmatrix} r(0) + \dot{r}(0)\frac{\delta t}{2} \\ p(\delta t'') \end{pmatrix} = \hat{\mathbf{X}}(\delta t^*) = \hat{\mathbf{X}}(\delta t'')\ , \end{aligned} \tag{2.29}$$

$$\begin{aligned} \mathcal{U}_p(-\delta t)\hat{\mathbf{X}}(\delta t^*) &\overset{\{r(\delta t^*)=r(\delta t'')=r(\delta t'),p(\delta t^*)=p(\delta t'')\}}{=} \begin{pmatrix} r(\delta t') \\ p(\delta t'') + \dot{p}(r(\delta t'))(-\delta t) \end{pmatrix} \\ &= \begin{pmatrix} r(\delta t') \\ p(0) + [\dot{p}(r(\delta t')) - \dot{p}(r(\delta t'))]\delta t \end{pmatrix} = \hat{\mathbf{X}}(\delta t^{**}) = \hat{\mathbf{X}}(\delta t')\ , \end{aligned} \tag{2.30}$$

$$\begin{aligned} \mathcal{U}_r\left(-\tfrac{\delta t}{2}\right)\hat{\mathbf{X}}(\delta t^{**}) &\overset{\{r(\delta t^{**})=r(\delta t'),p(\delta t^{**})=p(\delta t')=p(0)\}}{=} \begin{pmatrix} r(\delta t') + \dot{r}(p(0))\left(-\frac{\delta t}{2}\right) \\ p(0) \end{pmatrix} \\ &= \begin{pmatrix} r(0) + [\dot{r}(p(0)) - \dot{r}(p(0))]\frac{\delta t}{2} \\ p(0) \end{pmatrix} = \hat{\mathbf{X}}(0) \overset{!}{=} \mathbf{X}(0)\ . \end{aligned} \tag{2.31}$$

From (2.31) it can be seen that the evolution obtained by $\hat{\mathcal{U}}$ as defined in (2.25) is actually time reversible.

2.2 Simulations in the Microcanonical Ensemble

In the following the standard implementation of the already mentioned Störmer-Verlet algorithm [20, 23] shall be used. Due to its explicit time reversibility (see section 2.1.4), it is

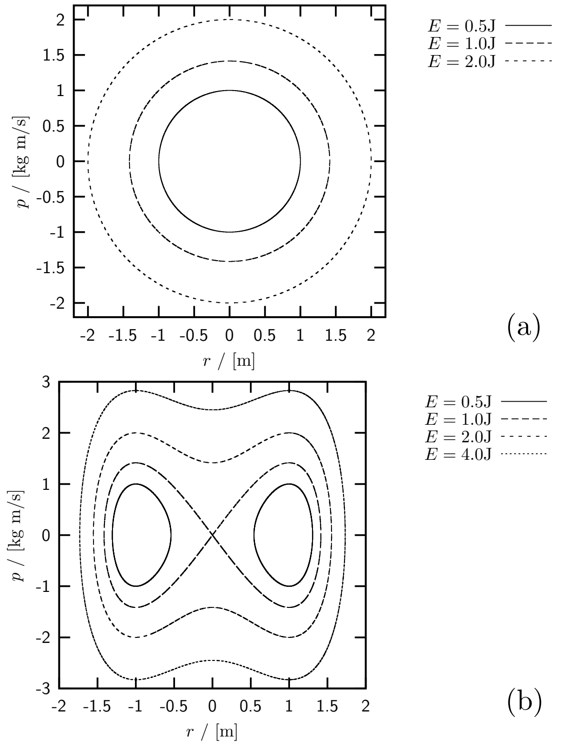


Figure 2.1: Trajectories of (a) the 1d harmonic oscillator ($m = 1$kg, $\omega = 1\text{s}^{-1}$) and (b) the 1d quartic double well ($a = 1$m, $D_0 = 1\text{Jm}^{-4}$) in a microcanonical simulation at different energies.

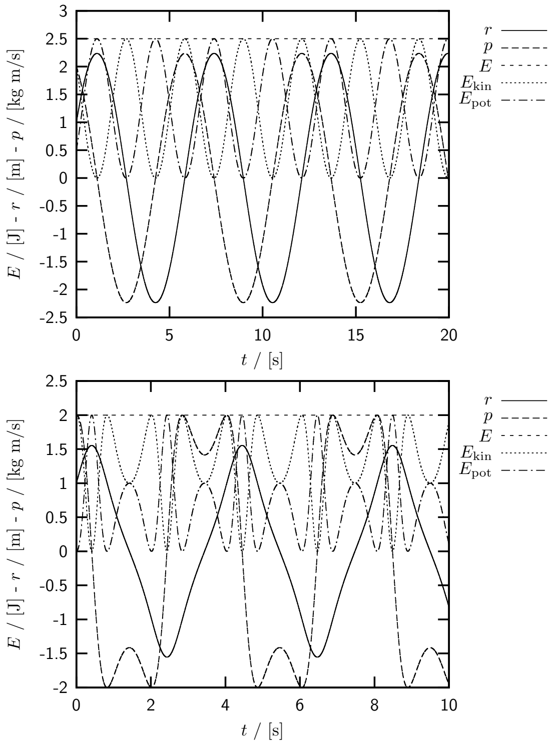


Figure 2.2: Time series of dynamic quantities for the same systems. The data was obtained by simulating with time steps $\delta t = 10^{-4}$s.

possible to obtain trajectories with a minuscule long-time energy drift. For the propagation of the system with the Störmer-Verlet algorithm, the force is needed. This can of course be calculated from the potential, $F = -\partial U/\partial x$. In pseudo code, the method looks like this:

```
rnew = r    + ( p        + force(r)*dt/2.0 )/m*dt
p    = p    + ( force(r) + force(rnew)     )  *dt/2.0
r    = rnew
```

Figure 2.1 shows trajectories in phase space of two simple one-dimensional systems, the harmonic oscillator and the quartic double well. The thermodynamic properties of these systems will be discussed in detail in the next section 2.3. The potentials are defined as follows:

$$E_{\text{pot, ho}} = \frac{1}{2}m\omega^2x^2 \ , \qquad E_{\text{pot, qdw}} = D_0(a^2 - x^2)^2 \ . \tag{2.32}$$

The respective forces are:

$$F_{\text{ho}} = -\frac{\partial}{\partial x}E_{\text{pot, ho}} = -m\omega^2x \ , \qquad F_{\text{ho}} = -\frac{\partial}{\partial x}E_{\text{pot, qdw}} = 4D_0x(a^2 - x^2) \ . \tag{2.33}$$

The quartic double well potential has some interesting characteristics. A particle with an energy of $E < a^4D_0$ (e.g. $E = 0.5$J in the figure) would of course stay on one side of the positional space and never see the other one, due to the potential energy barrier a^4D_0. For simulations in the microcanonical ensemble this is not a problem. But for the canonical ensemble, the algorithm will have to ensure that the system also samples the side, where it did not reside initially. The "trajectory" for $E = a^4D_0$ ($E = 1$J in Fig. 2.1) is also something special: it is a separatrix. That is it separates trajectories like the previously described that belong fully to one half of the phase space, and trajectories of particles with an energy $E > a^4D_0$, which are capable of overcoming the potential energy barrier. This also means a particle with a kinetic energy of $E_{\text{kin}} = a^4D_0$ at $r = \pm a$ would need infinitely long to reach $r = 0$. At $r = 0$ there is an instable fix point.

2.3 Thermodynamics of Selected 1d Systems

In the following, only purely position-dependent potentials will be discussed that is all systems are conservative ($\mathbf{F} = \nabla f(\mathbf{x})$). Therefore any quantities like moments of p or the kinetic energy distribution, which only depend on momentum, are equal for any considered type of system.

2.3.1 Harmonic Oscillator

The first system to be considered is the harmonic oscillator:

$$U = E_{\text{pot}} = \frac{1}{2}m\omega^2x^2 \ . \tag{2.34}$$

The partition function can be calculated, as well as some thermodynamic quantities:

$$\begin{aligned} Z_{\text{ho}} &= \int\limits_{-\infty}^{\infty} \mathrm{d}p \int\limits_{-\infty}^{\infty} \mathrm{d}x \, \exp\left[-\beta\left(\frac{1}{2m}p^2 + \frac{1}{2}m\omega^2x^2\right)\right] \\ &= \underbrace{\int\limits_{-\infty}^{\infty} \mathrm{d}p \, \exp\left[-\left(\beta\frac{1}{2m}\right)p^2\right]}_{\sqrt{\frac{\pi}{\beta\frac{1}{2m}}}} \underbrace{\int\limits_{-\infty}^{\infty} \mathrm{d}x \, \exp\left[-\left(\beta\frac{1}{2}m\omega^2\right)x^2\right]}_{\sqrt{\frac{\pi}{\beta\frac{1}{2}m\omega^2}}} = \frac{2\pi}{\beta\omega} \ . \end{aligned} \tag{2.35}$$

The mean energy and the heat capacity can be directly calculated from the partition function:

$$\langle E\rangle_{\text{ho}} = -\frac{\partial}{\partial\beta}\ln Z_{\text{ho}} = -\frac{\beta\omega}{2\pi}\left(-\frac{2\pi}{\beta^2\omega}\right) = \frac{1}{\beta} \ , \tag{2.36}$$

$$C_{v\,\text{ho}} = \underbrace{\frac{\partial}{\partial T}}_{-k_B\beta^2\frac{\partial}{\partial\beta}} \langle E\rangle_{\text{ho}} = -k_B\beta^2\frac{\partial}{\partial\beta}\frac{1}{\beta} = k_B \ . \tag{2.37}$$

Also, for checking the convergence of an algorithm, it is reasonable to calculate some moments of x and p. A general type of integral, which appears in every derivation of these moments is:

$$\int\limits_0^\infty \mathrm{d}x\, x^n e^{-ax^2} = \frac{\Gamma\left(\frac{n+1}{2}\right)}{2a^{\left(\frac{n+1}{2}\right)}} \ , \tag{2.38}$$

$$\begin{aligned}
\int\limits_{-\infty}^\infty \mathrm{d}x\, x^n e^{-ax^2} &= \int\limits_{-\infty}^0 \mathrm{d}x\, x^n e^{-ax^2} + \int\limits_0^\infty \mathrm{d}x\, x^n e^{-ax^2} = \int\limits_0^\infty \mathrm{d}x'\,(-x')^n e^{-ax'^2} + \int\limits_0^\infty \mathrm{d}x\, x^n e^{-ax^2} \\
&= (1+(-1)^n)\,\frac{\Gamma\left(\frac{n+1}{2}\right)}{2a^{\left(\frac{n+1}{2}\right)}} \\
&= \begin{cases} 0 & \text{for } n \in \{1,3,5,\ldots\} \ , \\ \sqrt{\frac{\pi}{a}}(2a)^{-n/2}\prod\limits_{l=1}^{n/2}(2l-1) & \text{for } n \in \{0,2,4,\ldots\} \ . \end{cases}
\end{aligned} \tag{2.39}$$

Thus, any moments of an odd order do vanish. For moments of the momentum p arises from (2.39):

$$\begin{aligned}
\langle p^{2k}\rangle &= \frac{1}{Z_{\text{ho}}}\left(\int\limits_{-\infty}^\infty \mathrm{d}x\ \exp\left[-\left(\beta\frac{1}{2}m\omega^2\right)x^2\right]\right)\left(\int\limits_{-\infty}^\infty \mathrm{d}p\, p^{2k}\exp\left[-\left(\beta\frac{1}{2m}\right)p^2\right]\right) \\
&= \frac{\beta\omega}{2\pi}\sqrt{\frac{2\pi}{\beta m\omega^2}}\left(\sqrt{\frac{\pi}{(\beta\frac{1}{2m})}}\left(2\left(\beta\frac{1}{2m}\right)\right)^{-k}\prod_{l=1}^{k}(2l-1)\right) \\
&= \left(\frac{\beta}{m}\right)^{-k}\prod_{l=1}^{k}(2l-1) \ .
\end{aligned} \tag{2.40}$$

Analogically the result for moments of the positions x is obtained:

$$\langle x^{2k}\rangle_{\text{ho}} = \left(\beta m\omega^2\right)^{-k}\prod_{l=1}^{k}(2l-1) \ . \tag{2.41}$$

With this knowledge, it is easy to calculate the mean kinetic and potential energy:

$$\langle E_{\text{kin}}\rangle = \left\langle \frac{1}{2m}p^2\right\rangle = \frac{1}{2m}\langle p^2\rangle \overset{\{(2.40)\}}{=} \frac{1}{2\beta} \ , \tag{2.42}$$

$$\langle E_{\text{pot}}\rangle_{\text{ho}} = \left\langle \frac{1}{2}m\omega^2 p^2\right\rangle = \frac{1}{2}m\omega^2\langle p^2\rangle \overset{\{(2.41)\}}{=} \frac{1}{2\beta} \ . \tag{2.43}$$

To verify the reproduction of the canonical ensemble by an algorithm, it is crucial to study distributions of certain quantities. Especially the distribution of the momentum p is interesting, since this is equal for all conservative systems:

$$\begin{aligned}P(p_0) &= \frac{1}{Z_{\text{ho}}}\int\limits_{-\infty}^{\infty}\mathrm{d}p\int\limits_{-\infty}^{\infty}\mathrm{d}x\,\exp\left[-\beta\left(\frac{1}{2m}p^2+\frac{1}{2}m\omega^2x^2\right)\right]\delta(p-p_0)\\ &= \frac{\beta\omega}{2\pi}\exp\left[-\beta\frac{1}{2m}p_0^2\right]\underbrace{\left(\int\limits_{-\infty}^{\infty}\mathrm{d}x\,\exp\left[-\left(\beta\frac{1}{2}m\omega^2\right)x^2\right]\right)}_{\sqrt{\frac{2\pi}{m\omega^2\beta}}}\\ &= \sqrt{\frac{\beta}{2\pi m}}\exp\left[-\beta\frac{1}{2m}p_0^2\right]\ , \qquad (2.44)\\ P(x_0)_{\text{ho}} &= \sqrt{\frac{\beta m\omega^2}{2\pi}}\exp\left[-\beta\frac{1}{2}m\omega^2x_0^2\right]\ . \qquad (2.45)\end{aligned}$$

Finally the distributions of the kinetic and the potential energy shall be given. It is not possible to calculate the density of states, which is actually the distribution of the total energy:

$$\begin{aligned}P(E_{\text{kin}0}) &= \frac{1}{Z_{\text{ho}}}\int\limits_{-\infty}^{\infty}\mathrm{d}p\int\limits_{-\infty}^{\infty}\mathrm{d}x\,\exp\left[-\beta\left(\frac{1}{2m}p^2+\frac{1}{2}m\omega^2x^2\right)\right]\delta\left(\frac{p^2}{2m}-E_{\text{kin}0}\right)\\ &\overset{\{z=\frac{p^2}{2m}\}}{=} \sqrt{\frac{\beta}{2\pi m}}\,2\int\limits_{0}^{\infty}\mathrm{d}z\,\underbrace{\frac{m}{\sqrt{2mz}}}_{\frac{\mathrm{d}x}{\mathrm{d}z}}e^{-\beta z}\delta(z-E_{\text{kin}0})\\ &= \sqrt{\frac{\beta}{\pi}}\frac{1}{\sqrt{E_{\text{kin}0}}}\exp\left[-\beta E_{\text{kin}0}\right]\ , \qquad (2.46)\\ P(E_{\text{pot}0})_{\text{ho}} &= \sqrt{\frac{\beta}{\pi}}\frac{1}{\sqrt{E_{\text{pot}0}}}\exp\left[-\beta E_{\text{pot}0}\right]\ . \qquad (2.47)\end{aligned}$$

2.3.2 The Quartic Double Well

A little more interesting than the harmonic oscillator is the quartic double well, which is generally described by the potential:

$$U = E_{\text{pot}} = D_0(a^2-x^2)^2\ . \qquad (2.48)$$

While in the case of the harmonic oscillator, the system has only one minimum, the quartic double well has two local minima. This can lead to problems at very low temperatures, compared to the height of the potential barrier D_0a^2. In this case, the system is likely to be trapped in one of the two minima, and it is up to the algorithm to push it into the other one to provide an acceptable sampling of the whole phase space. This is crucial for the statistics. The partition sum can be calculated as:

$$Z_{\text{qdw}} = \sqrt{\frac{\pi^3ma^2}{2}}\frac{1}{\sqrt{\beta}}\exp\left[-\frac{1}{2}a^4D_0\beta\right]\left(I_{-\frac{1}{4}}\left(\frac{1}{2}a^4D_0\beta\right)+I_{\frac{1}{4}}\left(\frac{1}{2}a^4D_0\beta\right)\right)\ , \qquad (2.49)$$

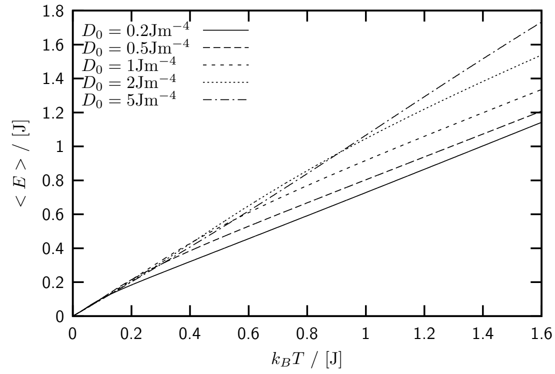


Figure 2.3: The mean energy with respect to thermal energy for different parameters D_0 ($a = 1$m).

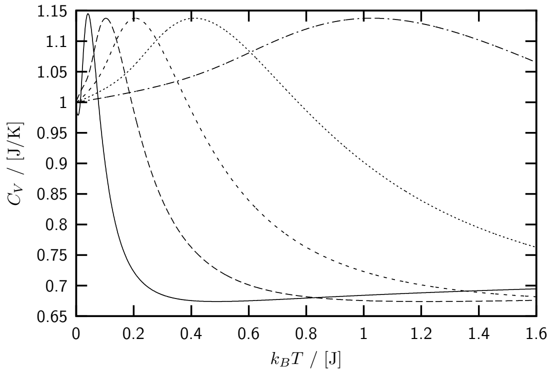


Figure 2.4: The respective specific heat. The key is the same as in Fig. 2.3.

where $I_\nu(x)$ is the modified Bessel function of first kind. The mean energy and the heat capacity can be calculated straightforward, but the expressions are rather lengthy:

$$\begin{aligned}\langle E\rangle_{\mathrm{qdw}} &= -\frac{\partial}{\partial\beta}\ln Z_{\mathrm{qdw}}\\ &= \frac{\frac{1}{2}+x}{\beta} - \frac{x\left(I_{-\frac{5}{4}}(x)+I_{\frac{5}{4}}(x)+I_{-\frac{3}{4}}(x)+I_{\frac{3}{4}}(x)\right)}{2\beta\left(I_{-\frac{1}{4}}(x)+I_{\frac{1}{4}}(x)\right)} \qquad \text{with} \qquad x = \frac{1}{2}a^4D_0\beta\ . \end{aligned} \tag{2.50}$$

The mean potential energy is an even more complicated function. Again, the abbreviation $x = a^4D_0\beta/2$ is used:

$$\begin{aligned}\langle E_{\mathrm{pot}}\rangle_{\mathrm{qdw}} = \frac{2x}{\beta} + &\left(\frac{e^{-x}(2x)^{\frac{1}{4}}}{\pi\beta\left(I_{-\frac{1}{4}}(x)+I_{\frac{1}{4}}(x)\right)}\right)\\ &\times\left[2\Gamma\left(\frac{7}{4}\right)\,{}_1F_1\left(\frac{7}{4},\frac{3}{2},2x\right)+2\Gamma\left(\frac{3}{4}\right)\,{}_1F_1\left(\frac{3}{4},\frac{1}{2},2x\right)\right.\\ &\left.\quad+\beta^2\Gamma\left(\frac{5}{4}\right)\,{}_1F_1\left(\frac{5}{4},\frac{1}{2},2x\right)-4\sqrt{2x}\Gamma\left(\frac{5}{4}\right)\,{}_1F_1\left(\frac{5}{4},\frac{3}{2},2x\right)\right]\ ,\end{aligned} \tag{2.51}$$

where ${}_1F_1(a,b,z)$ is the Kummer confluent hypergeometric function. The distribution of the positions arises from (2.49):

$$\begin{aligned}P(x_0)_{\mathrm{qdw}} &= \frac{1}{Z_{\mathrm{qdw}}}\underbrace{\int_{-\infty}^{\infty}\mathrm{d}p\ \exp\left[-\left(\beta\frac{1}{2m}\right)p^2\right]}_{\sqrt{\frac{2\pi m}{\beta}}}\int_{-\infty}^{\infty}\mathrm{d}x\ \exp\left[-\beta D_0(a^2-x^2)^2\right]\delta(x-x_0)\\ &= \frac{2\exp\left[\frac{1}{2}a^4D_0\beta\right]}{\pi a\left(I_{-\frac{1}{4}}\left(\frac{1}{2}a^4D_0\beta\right)+I_{\frac{1}{4}}\left(\frac{1}{2}a^4D_0\beta\right)\right)}\exp\left[-\beta D_0(a^2-x_0^2)^2\right]\ .\end{aligned} \tag{2.52}$$

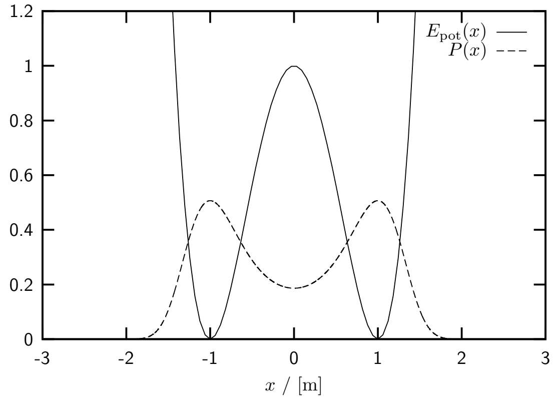


Figure 2.5: Potential and distribution of the positions of a quartic double well with the parameters: $a = 1\text{m}$, $D_0 = 1\text{Jm}^{-4}$ and $k_B T = 1\text{J}$.

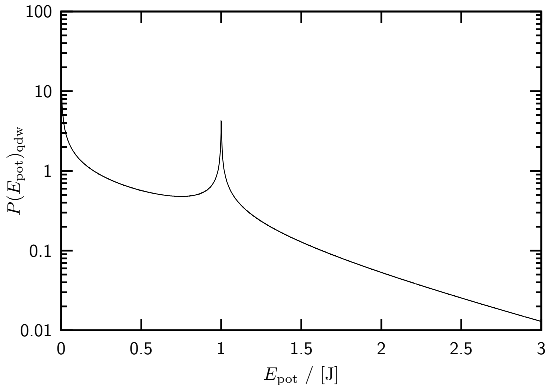


Figure 2.6: Logarithmic plot of the potential energy distribution for the same system as in Fig. 2.5.

With the general formula

$$\int_{-\infty}^{\infty} \mathrm{d}x\, \phi(x)\delta(g(x)) \overset{\{z=g(x)\}}{=} \int_{-\infty}^{\infty} \mathrm{d}z \underbrace{\frac{\mathrm{d}x}{\mathrm{d}z}}_{\displaystyle\frac{1}{g'(x)}} \phi(x)\delta(z) = \sum_i \frac{\phi(x_i)}{g'(x_i)} \ , \tag{2.53}$$

where x_i are single roots of $g(x)$, it is also possible to derive the potential energy distribution:

$$\begin{aligned}
P(E_{\text{pot}_0})_{\text{qdw}} &= \frac{1}{Z_{\text{qdw}}}\sqrt{\frac{2\pi m}{\beta}} \int_{-\infty}^{\infty} \mathrm{d}x \, \exp\left[-\beta D_0 (a^2 - x^2)^2\right] \delta(D_0(a^2 - x^2)^2 - E_{\text{pot}_0}) \ , \\
g(x) &= D_0(a^2 - x^2)^2 - E_{\text{pot}_0} \ , \\
g'(x) &= D_0 2(a^2 - x^2)(-2x) = -4x\sqrt{D_0 E_{\text{pot}}(x)} \ , \\
x_0 &= \pm\frac{\sqrt{\left|\sqrt{E_{\text{pot}_0} D_0} \pm a^2 D_0\right|}}{\sqrt{D_0}} \ , \\
P(E_{\text{pot}_0})_{\text{qdw}} &= \frac{2\exp\left[\frac{1}{2}a^4 D_0 \beta\right]}{\pi a \left(I_{-\frac{1}{4}}\left(\frac{1}{2}a^4 D_0\beta\right) + I_{\frac{1}{4}}\left(\frac{1}{2}a^4 D_0\beta\right)\right)} \exp\left[-\beta E_{\text{pot}_0}\right] \\
&\quad \times \left(\frac{2}{4\sqrt{E_{\text{pot}}(x)}\sqrt{\sqrt{E_{\text{pot}_0} D_0} + a^2 D_0}}\right. \\
&\qquad \left. + \frac{2}{4\sqrt{E_{\text{pot}}(x)}\sqrt{\left|\sqrt{E_{\text{pot}_0} D_0} - a^2 D_0\right|}}\right) \ .
\end{aligned} \tag{2.54}$$

2.4 Molecular Dynamics in the Canonical Ensemble

As described in section 2.2, the numerical integration of the standard Newtonian equations of motion of a system is equivalent to consider the system in a microcanonical ensemble with given energy. Especially time-reversible algorithms as described in section 2.1 do conserve the energy. Thus there are two ways for a simulation in the canonical ensemble:

1. Either the algorithm has to be changed in some way so that it does not conserve the energy anymore and somehow creates a canonical ensemble. This is handled by the stochastic Andersen thermostat [7] for example. The price to pay is that the dynamics is not deterministic anymore, which is undesirable.
2. Or the system has to be coupled to some kind of heat bath by means of deterministic degrees of freedom, it has to be *thermostated.* The system gets more complex in doing so, but it will turn out that the loss in efficiency is negligible, especially compared to the possibility of observing continuous trajectories in the canonical ensemble.

2.4.1 Andersen Thermostat – Stochastic Molecular Dynamics

As stated above, the Andersen thermostat [7] reproduces the canonical ensemble by stochastically changing the amount of kinetic energy. Actually, it can be understood as a kind of hybrid algorithm between MC and MD. The procedure is described in more detail in Ref. [20]:

- perform constant energy Molecular Dynamics (microcanonical ensemble) as described in the first two sections of this chapter,
- chose each particle at a certain collision frequency ν, which is equivalent to the strength of the heat bath coupling,
- assign a new velocity according to the Boltzmann distribution to this particle.

The second point in this list has to be a little more illuminated. Although introducing a collision frequency, the collisions should still happen stochastically of course. Therefore, it is not possible to simply carry out point three in the upper list every $1/\nu$ MD steps. Rather a probability of time intervals between two collisions is assessed, so that successive collisions are uncorrelated. This is of the Poisson form [24, 25]:

$$P(\nu, t) = \nu \exp\left[-\nu t\right] \ . \tag{2.55}$$

According to the references given above, the probability that a particle is selected in a certain time step δt, is $p = \nu \delta t$.

The choice of a velocity from a Boltzmann distribution, or the Maxwell-Boltzmann distribution in three dimensions, can be done by the Box-Müller method described in [26]. Therewith, from two uniformly distributed random numbers $R_1, R_2 \in [0, 1)$, two random numbers x and y from a Gaussian distribution with width σ can be calculated:

$$r = \sqrt{-2\sigma^2 \ln(1 - R_1)} \ , \qquad \Theta = 2\pi R_2 \ , \tag{2.56}$$

$$x = r \cos(\Theta) \ , \qquad y = r \sin(\Theta) \ . \tag{2.57}$$

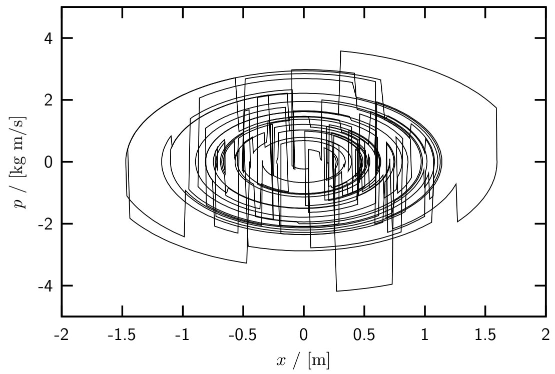


Figure 2.7: Trajectory over the first $5 \cdot 10^4$ time steps of a one-dimensional harmonic oscillator ($\omega = 1\mathrm{s}^{-1}$, $m = 2\mathrm{kg}$) with Andersen thermostat at $k_B T = 1.0\mathrm{J}$ and collision frequency $\nu = 1\mathrm{s}^{-1}$.

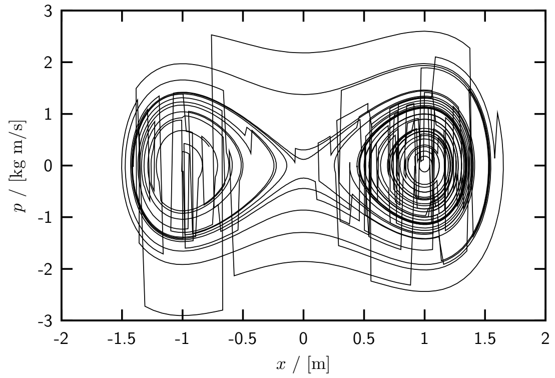


Figure 2.8: Trajectory as in Fig. 2.7 for a one-dimensional quartic double well ($D_0 = 1\mathrm{Jm}^{-4}$, $a = 1\mathrm{m}$, $m = 1\mathrm{kg}$).

For simulations in more than one dimension, it is still enough to choose every component of the velocity from a Gaussian distribution. Nothing else does, e.g., produce the Maxwell-Boltzmann distribution in three dimensions. The correct width σ can be found by comparing the standard Gaussian distribution and the Boltzmann distribution:

$$P(x) = \frac{1}{\sqrt{\sigma^2 2\pi}} \exp\left[-\frac{x^2}{2\sigma^2}\right] , \tag{2.58}$$

$$P(v_i) = \sqrt{\frac{\beta m}{2\pi}} \exp\left[-\frac{\beta m}{2} v_i^2\right] , \tag{2.59}$$

$$\Rightarrow \qquad \sigma^2 = \frac{1}{\beta m} = \frac{k_B T}{m} . \tag{2.60}$$

In appendix A.1, an example pseudo source code is given for the implementation of the Andersen thermostat.

Exemplary Simulations

For the purpose of testing the thermostat and especially different choices of the collision frequency τ, several trial runs with the one dimensional harmonic oscillator and the quartic double well are carried out. These systems were already treated analytically in the previous section, therefore, the outcome of the simulations can be directly compared to the exact results. All runs were performed with step size $\delta t = 0.005\mathrm{s}$ and had an equilibration phase of 10^5 steps. The measurements were carried out over 10^8 time steps at temperatures $k_B T = \{0.1, 0.2, \ldots, 1.0\}$.

The general behaviour of the system due to the Andersen thermostating is shown in Figs. 2.7 and 2.8. There, trajectories of a one-dimensional harmonic oscillator and a one-dimensional quartic double well are plotted. The trajectories look very similar to those in Fig. 2.1. This is due to the fact that constant energy MD simulations are performed. Although, due to the stochastic changes in the kinetic energy, the system moves to a different energy shell from time to time.

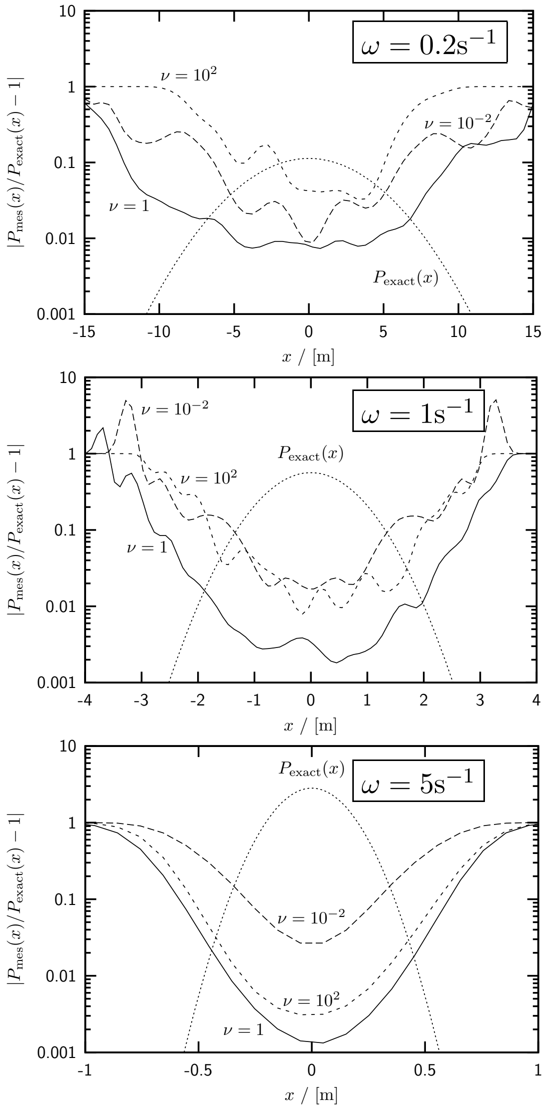


Figure 2.9: Relative error of x distributions of harmonic oscillators measured with collision frequency $\nu = \{0.01\mathrm{s}^{-1}, 1\mathrm{s}^{-1}, 100\mathrm{s}^{-1}\}$.

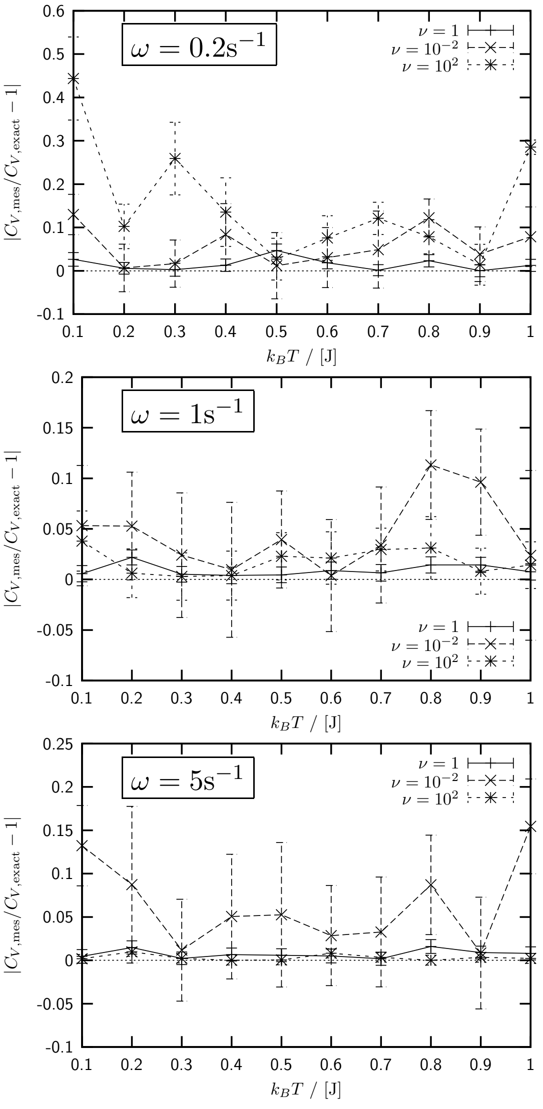


Figure 2.10: Relative error of specific heat ($C_V = k_B$) of one-dimensional harmonic oscillators for the same set-ups as in Fig. 2.9.

To have an impression of the effect of altering ν, three different harmonic oscillators ($\omega = \{0.2\mathrm{s}^{-1}, 1\mathrm{s}^{-1}, 5\mathrm{s}^{-1}\}$) are measured with three adjustments of the collision frequency ($\nu = \{0.01\mathrm{s}^{-1}, 1\mathrm{s}^{-1}, 100\mathrm{s}^{-1}\}$). The relative errors of the position distributions and the specific heat are shown in Figs. 2.9 and 2.10. Especially from the former, $\nu = 1\mathrm{s}^{-1}$ seems to be a good choice, because for this adjustment, the relative deviations are the smallest for all trial systems. This is emphasised by the fact that the reproduction of the specific heat (k_B, according to eq. (2.37)) is also very good for this choice. The general choice $\nu = 1\mathrm{s}^{-1}$ is additionally supported by the fact that Ref. [20] suggests a minor role of the adjustment of ν.

2.4.2 Overview of the Nosé-Hoover-Chain Thermostat

A very well-known and frequently used class of methods to manage thermostating is the Nosé algorithm [8] and certain extensions which shall be briefly reviewed.

The original approach of Nosé uses an extended Lagrangian formulation to derive the Hamiltonian for the system and the equations of motion:

$$\mathcal{L}_{\mathrm{Nose}} = \sum_{i=1}^{N} \frac{m_i}{2} s^2 \dot{\mathbf{r}}_i{}^2 - \mathcal{U}(\mathbf{r}^N) + \frac{Q}{2}\dot{s}^2 - (f+1) k_B T_{\mathrm{eq}} \ln s \ . \tag{2.61}$$

This results in non-equidistant time steps and is fixed by the introduction of a virtual time scale. In (2.61) T_{eq} is the desired equilibrium temperature. f is the number of degrees of freedom and s can be understood as a dimensionless scaling factor. A detailed discussion can be found in section 6.1 of Ref. [20].

Hoover showed that it is equivalent to introduce one additional degree of freedom to the considered system [9]. The connection to Nosé's formulation is:

$$\frac{Q\dot{s}}{s} = Q\frac{\mathrm{d}\ln s}{\mathrm{d}t} = p_\xi \ . \tag{2.62}$$

The resulting equations of motion are the following, where ξ is the virtual position of the thermostat "particle", which is obviously dimensionless like s, and $p_\xi = Qv_\xi = Q\dot{\xi}$:

$$\dot{\mathbf{r}}_i = \frac{\mathbf{p}_i}{m_i} \ , \tag{2.63}$$

$$\dot{\mathbf{p}}_i = \mathbf{F}_i - v_\xi \mathbf{p}_i = -\nabla_{\mathbf{r}_i}\mathcal{U}(\mathbf{r}^N) - \frac{p_\xi}{Q}\mathbf{p}_i \ , \tag{2.64}$$

$$\dot{p_\xi} = \left(\sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{m_i} - f k_B T_{\mathrm{eq}} \right) = 2\left(E_{\mathrm{kin,instantaneous}} - E_{\mathrm{kin,equilibrium}}\right) \ . \tag{2.65}$$

Q has the dimension [Energy×Time^2 =Mass×Length^2] and is a thermal inertia parameter, which determines the rate of the heat transfer. However, it appears in the equations of motion as a virtual mass. This parameter will be discussed in more detail later. The advantage of the Nosé-Hoover (NH) thermostat is firstly that it is possible to derive equations of motion in "real time", and furthermore that the according equations of motion are easier to understand. Specifically from (2.65) it is easy to see the operating principle of the NH thermostat. The Nosé-Hoover "particle" at the virtual position ξ is driven by the difference of the instantaneous and the desired kinetic energy. However, the system is *not* Hamiltonian anymore. A certain energy term is conserved, but it is not possible to derive the equations of motion from it. This problem was tackled by Tuckerman [27], who deduced the basis for the statistical mechanics of non-Hamiltonian systems. Also Joannopoulos showed [28] that the canonical ensemble is only obtained by a NH thermostat, if the virtual total momentum is not conserved or constant zero ($\dot{\mathbf{P}} \neq 0$ or $\mathbf{P} \equiv 0$).

A much more critical issue is the fact that both the original Nosé and the Nosé-Hoover algorithm do not reproduce the correct canonical statistics for the one-dimensional harmonic oscillator, which is the simplest nontrivial system. Hoover already pointed out this problem in his publication [9]. It was claimed that this could be connected to the problem found by Joannopoulos. The solution was the introduction of the *Nosé-Hoover-Chain* method (NHC)

[10]. Actually the idea is to control the kinetic energy of the Nosé-Hoover "particle" by yet another NH thermostat and so forth. The equations of motion for a system with N particles, f degrees of freedom and a chain of M NH thermostats are therefore:

$$\dot{\mathbf{r}}_i = \frac{\mathbf{p}_i}{m_i} \ , \tag{2.66}$$

$$\dot{\mathbf{p}}_i = \mathbf{F}_i - \frac{p_{\xi_1}}{Q_1}\mathbf{p}_i = -\nabla_{\mathbf{r}_i}\mathcal{U}(\mathbf{r}^N) - \frac{p_{\xi_1}}{Q_1}\mathbf{p}_i \ , \tag{2.67}$$

$$\dot{p_{\xi_1}} = \left(\sum_{i=1}^{N}\frac{\mathbf{p}_i^2}{m_i} - fk_BT_{\text{eq}}\right) - \frac{p_{\xi_2}}{Q_2}p_{\xi_1} \ , \tag{2.68}$$

$$\dot{p_{\xi_k}} = \left(\frac{p_{\xi_{k-1}}^2}{Q_{k-1}} - k_BT_{\text{eq}}\right) - \frac{p_{\xi_{k+1}}}{Q_{k+1}}p_{\xi_k} \ , \tag{2.69}$$

$$\dot{p_{\xi_M}} = \left(\frac{p_{\xi_{M-1}}^2}{Q_{M-1}} - k_BT_{\text{eq}}\right) \ . \tag{2.70}$$

The conserved energy for this general system is:

$$\mathcal{H}_{\text{NHC}} = \mathcal{H}(\mathbf{r},\mathbf{p}) + \sum_{k=1}^{M}\frac{p_{\xi_k}^2}{2Q_k} + fk_BT_{\text{eq}}\xi_1 + \sum_{k=2}^{M}k_BT_{\text{eq}}\xi_k \ . \tag{2.71}$$

As already mentioned, it is not possible to derive the equations of motion (2.66) – (2.70) from (2.71). Thus $\mathcal{H}_{\text{NHC}}$ is *not* a Hamiltonian.

It turned out that it was not only possible to obtain the correct canonical ensemble for the harmonic oscillator by coupling two NH-thermostats ($M = 2$), but also the pitfall concerning the conservation of the total momentum was not a problem anymore [29]. According to Liu and Tuckerman [30] there is still a concern, the NHC method is only capable of maintaining adequate temperature control in equilibrium. Any perturbation away from equilibrium, for example caused by the presence of external fields or by motion over a high barrier, causes the method to break down.

2.4.3 Algorithmic Details of NHC

Numerical Integration in Particular

It is possible in principle, to use the Liouville/Trotter formalism as it was explained in section 2.1, to derive an explicit algorithm. However it is expected that there will occur some problems, since the system is not strictly conservative anymore! As it can be seen from (2.67) and (2.64) respectively, the force has now a velocity dependence. So the original approach has to be modified somewhat.

The separation of the Liouville operator is done in a slightly different way. While for a simple microcanonical approach, where the number of particles N, the volume of the system V and the total energy E are given and fixed (see section 2.2), $\imath\mathcal{L}$ is divided into two parts $\imath\mathcal{L} = \imath\mathcal{L}_r + \imath\mathcal{L}_p$, now the terms depending on the thermostat variables are separated from the variables of the stand-alone system [21]:

$$
\begin{aligned}
\imath\mathcal{L}_{\mathrm{NHC}} &= \sum_{i=1}^{N} \mathbf{v}_i \nabla_{\mathbf{r}_i} + \sum_{i=1}^{N} \left(\frac{\mathbf{F}_i}{m_i}\right) \nabla_{\mathbf{v}_i} - \sum_{i=1}^{N} v_{\xi_1} \mathbf{v}_i \nabla_{\mathbf{v}_i} + \sum_{k=1}^{M} v_{\xi_k} \frac{\partial}{\partial \xi_k} + \sum_{k=1}^{M} \underbrace{\left(\frac{\dot{p}_{\xi_k}}{Q_k}\right)}_{=\ddot{\xi}_k} \frac{\partial}{\partial \dot{\xi}_k} \\
&= \underbrace{\sum_{i=1}^{N} \mathbf{v}_i \nabla_{\mathbf{r}_i}}_{\imath\mathcal{L}_r} + \underbrace{\sum_{i=1}^{N} \left(\frac{\mathbf{F}_i}{m_i}\right) \nabla_{\mathbf{v}_i}}_{\imath\mathcal{L}_p} \underbrace{- \sum_{i=1}^{N} v_{\xi_1} \mathbf{v}_i \nabla_{\mathbf{v}_i}}_{\imath\mathcal{L}_{C_v}} + \underbrace{\sum_{k=1}^{M} v_{\xi_k} \frac{\partial}{\partial \xi_k}}_{\imath\mathcal{L}_\xi} \\
&\quad + \underbrace{\sum_{k=1}^{M-1} G_k \frac{\partial}{\partial v_{\xi_k}}}_{\imath\mathcal{L}_{G_{k<M}}} \underbrace{- \sum_{k=1}^{M-1} v_{\xi_{k+1}} v_{\xi_k} \frac{\partial}{\partial v_{\xi_k}}}_{\imath\mathcal{L}_{v_\xi}} + \underbrace{G_M \frac{\partial}{\partial \dot{\xi}_M}}_{\imath\mathcal{L}_{G_M}} \ ,
\end{aligned}
\tag{2.72}
$$

$$
\text{with} \quad G_1 = \frac{1}{Q_1} \left(\sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{m_i} - f k_B T_{\mathrm{eq}} \right) \ , \qquad G_{k>1} = \frac{1}{Q_k} \left(\frac{p_{\xi_{k-1}}^2}{Q_{k-1}} - k_B T_{\mathrm{eq}} \right) \ , \tag{2.73}
$$

$$
\imath\mathcal{L}_C = \imath\mathcal{L}_{C_v} + \imath\mathcal{L}_\xi + \imath\mathcal{L}_{G_{k<M}} + \imath\mathcal{L}_{v_\xi} + \imath\mathcal{L}_{G_M} \ . \tag{2.74}
$$

$\imath\mathcal{L}_C$ represents the parts of the Liouville operator, which are connected to the thermostat. The structure of $\imath\mathcal{L}_{C_v}$ and $\imath\mathcal{L}_{v_\xi}$ is in principle different from the form of the Liouville operator as considered in section 2.1 (see Ref. [20], appendix E):

$$
\begin{aligned}
\exp\left[ax\frac{\partial}{\partial x}\right] f(x) &= \exp\left[a \underbrace{\left(\frac{\partial \ln x}{\partial x}\right)^{-1}}_{x} \frac{\partial}{\partial x}\right] f(x) = \exp\left[a\frac{\partial}{\partial \ln x}\right] f(x) \\
&= \exp\left[a\frac{\partial}{\partial \ln x}\right] f\left(\exp[\ln x]\right) = f\left(\exp[\ln x + a]\right) = f(x \exp[a]) \ .
\end{aligned}
\tag{2.75}
$$

Therefore, actually the considerations based on the certain form in section 2.1 do not hold for $\imath\mathcal{L}_{C_v}$ and $\imath\mathcal{L}_{v_\xi}$. Tuckerman derives the analytical propagation for a more general case of dependency in Ref. [31]. However, in practice higher-order Trotter schemes and a multiple time step approach are used to deal with this problem for the propagation of the thermostat "particles", which comes out of $\imath\mathcal{L}_C$. That is, the propagator is firstly separated with the Trotter scheme as discussed in section 2.1, (2.16):

$$
\hat{\mathcal{U}}_{\mathrm{NHC}}(\delta t) = \exp\left[\imath\mathcal{L}_C \tfrac{\delta t}{2}\right] \times \exp\left[\imath\mathcal{L}_r \tfrac{\delta t}{2}\right] \times \exp\left[\imath\mathcal{L}_p \delta t\right] \times \exp\left[\imath\mathcal{L}_r \tfrac{\delta t}{2}\right] \times \exp\left[\imath\mathcal{L}_C \tfrac{\delta t}{2}\right] \ . \tag{2.76}
$$

Afterwards, for the propagation of $\exp\left[\imath\mathcal{L}_C \delta t/2\right]$, the time step δt is divided once more, particularly into n_c equidistant steps. For the integration of $\exp\left[\imath\mathcal{L}_C \delta t/(2n_c)\right]$ the already mentioned higher-order Trotter formulae [32, 33] are applied. This results in:

$$
\exp\left[\imath\mathcal{L}_C \tfrac{\delta t}{2}\right] = \prod_{i=1}^{n_c} \left(\prod_{j=1}^{m} \exp\left[\imath\mathcal{L}_C \frac{w_j \delta t}{2n_c}\right] \right) \ . \tag{2.77}
$$

The Trotter factorisation of $\imath\mathcal{L}_C$ is discussed in detail by Martyna [21]. According to Windiks [22], it is standard to use a 5th order integration scheme ($m = 3$) for the thermostat propagation, which is correct up to the 6th order as proven in Ref. [32], and set $n_c = 1$. Only if

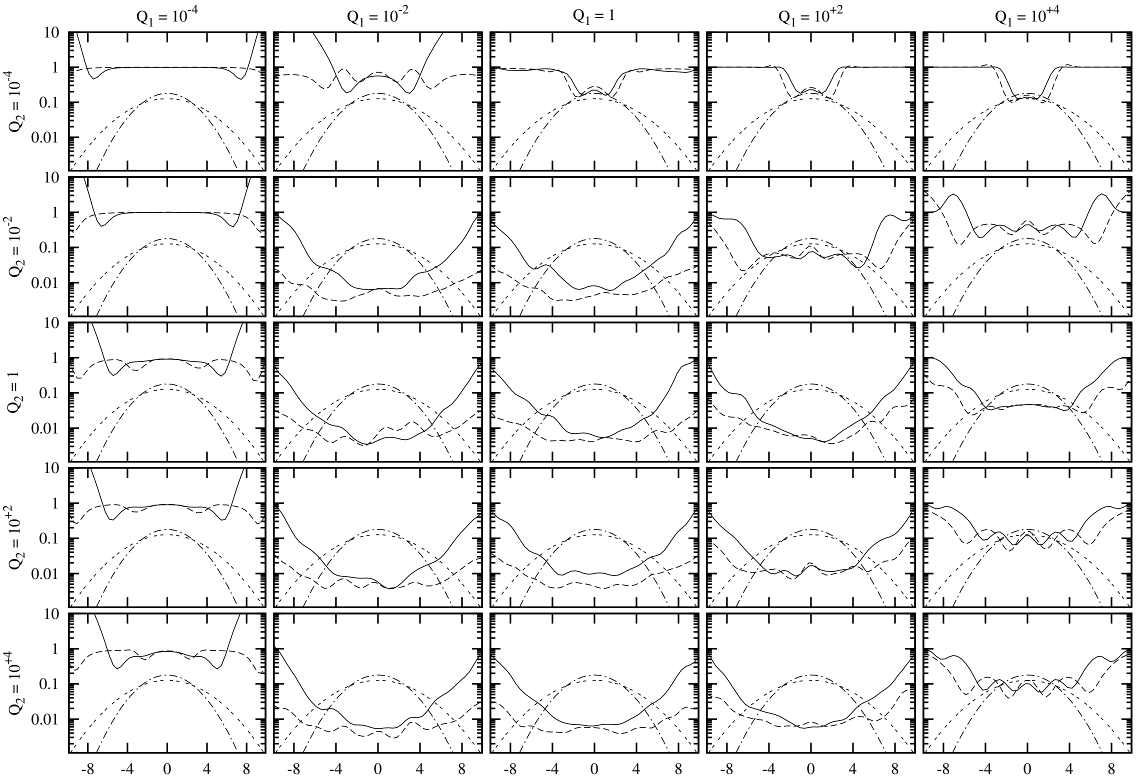


Figure 2.11: The chain dotted line and the twofold dashed line are the theoretical distributions of r and p of a one-dimensional harmonic oscillator. The solid line and the dashed line are the relative errors ($|P_{\text{mes}}(x)/P_{\text{analyt}}(x) - 1|$) of the measured histograms over r and p respectively for different choices of Q_1 and Q_2, given in Js^2. All quantities are plotted logarithmically. The data is obtained by simulating 10^7 time steps of length $\delta t = 0.01$s of a harmonic oscillator with the following parameters: $k_B T = 5$J, $m = 2$kg, $\omega^2 = 1/2\text{s}^{-2}$.

the typical time scales of fluctuations in the system are very short, it is useful to increase n_c. Also a detailed description of the Yoshida-Suzuki approach and the required numerical values for the 3rd, 5th, 7th, 9th and 27th order schemes can be found in Ref. [22]. There, as well as in Ref. [21], a general implementation of a Nosé-Hoover-Chain thermostat in pseudo code can be found, each with one straightforward optimisation, which can be combined. The full example pseudo source of the Nosé-Hoover-Chain thermostat with both optimisations and some minor spelling fixes is given in appendix A.2.

Choice of the Virtual Masses

It turns out that the choice of the thermal inertia parameters Q_i is crucial for the reproduction of the canonical ensemble. From Fig. 2.11 it is obvious that not only for unfavourable ratios, but also for very small or very large Q the obtained distributions are totally wrong. Martyna give an estimate for the masses in appendix B of Ref. [10]. According to these considerations, it follows:

$$Q_1 = f k_B T_{\text{eq}} \tau^2 \ , \qquad\qquad Q_{i>1} = k_B T_{\text{eq}} \tau^2 \ , \qquad (2.78)$$

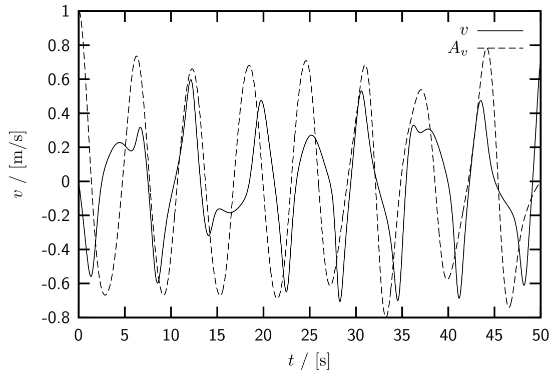


Figure 2.12: Time series of the velocity and the respective velocity autocorrelation function for a harmonic oscillator with the following parameters: $k_BT = 1$J, $m = 1$kg, $\omega = 1\text{s}^{-1}$. The simulation is carried out with time steps of length $\delta t = 0.005$s and a NHC of length $M = 2$ ($m = 3$, $n_c = 1$). The time scale is chosen to be $\tau = 1/\omega = 1$s.

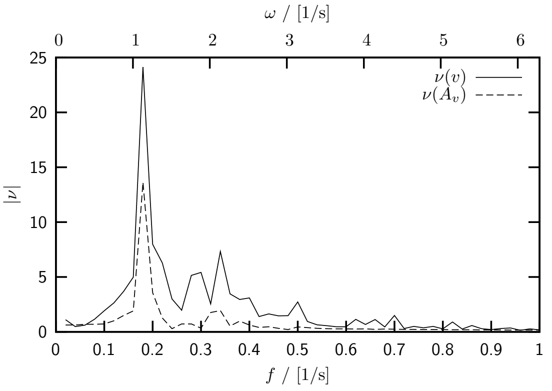


Figure 2.13: Discrete Fourier transform of the velocity time series and the velocity autocorrelation function. The scale below shows the real frequencies $f = 1/T$, where T is the duration of one oscillation. The scale above shows the circular frequency ω.

where $\tau = 1/\omega$ is a typical time scale for the considered stand-alone system. That is, larger thermal inertia Q result in a larger time scale for the fluctuations of the bath particles. Thus the thermostat is not capable of following the system fluctuations fast enough. On the other hand, making Q very small can lead to sampling problems due to high-frequency oscillations in v_ξ. For a correct propagation of the thermostat, the overall time step δt should be roughly a factor 20 – 40 smaller than τ. But this can also be fixed with increasing n_c as described above.

But how to define τ if the considered system is rather complex? Windiks [22] suggests to measure velocity autocorrelation functions. The main peak of the Fourier transform gives a typical frequency of the system. The expression for a normalised autocorrelation function is:

$$A_v(\Delta t) = \frac{\langle v(t)v(t+\Delta t)\rangle - \langle v(t)\rangle^2}{\langle v(t)^2\rangle - \langle v(t)\rangle^2} \ . \tag{2.79}$$

How can a velocity autocorrelation function be understood? Figure 2.12 shows a velocity time series of a harmonic oscillator. Contrary to the time series from the microcanonical simulation in Fig. 2.2, it is not a smooth sine. Due to the heat bath coupling by the thermostat, there are fluctuations on the major oscillation. It is expected that in a complex system these fluctuations will increase because of interactions with other particles. Therefore it is possible that the major oscillation mode will not be as obvious as in the exemplary system. The autocorrelation function will have maxima at the major oscillation times nT, where the data of t and $t + \Delta t = t + nT$ fit best. In Fig. 2.12 it is obvious that the autocorrelation function does effectively smooth the velocity time series. This will emphasise the major oscillation peak in the Fourier transform.

The discrete Fourier transform of a list u_r of length n is calculated as follows:

$$\nu_s = \frac{1}{\sqrt{n^{(1-\alpha)}}} \sum_{r=1}^{n} u_r \exp\left[\frac{2\pi\imath}{n} b(r-1)(s-1)\right] \ . \tag{2.80}$$

The parameters a and b can be set to $a = 0$, $b = 1$. In Fig. 2.13, the Fourier transform of the velocity time series is compared to the Fourier transform of the velocity autocorrelation function. The signal-to-noise ratio is not significantly different, but this is also due to the simplicity of the system. It is not astonishing that the main peak is found at about $\omega = 1$ as for the harmonic oscillator the circular frequency is explicitly defined to be $\omega = 1$.

2.4.4 Exemplary Simulations of Selected 1d Systems with the NHC

For the purpose of testing the algorithm and the various statements about its correct use, several simulations are carried out. The chosen test systems are the one-dimensional harmonic oscillator and the one-dimensional quartic double well, as they were already treated analytically in section 2.3.

In contrast to MC simulations, there is a time scale in Molecular Dynamics. Therefore, it is possible to directly compare the results of a MD simulation with experimental data by choosing SI units for all intrinsic quantities in the simulation. This is the reason, why in the following the mass is given in [kg], energy in [J], length in [m] and momentum in [kg m/s]. However, the choice may be misleading, since a harmonic oscillator with a mass of 2kg would never show measurable thermodynamic effects at room temperature, and a thermal energy of $k_B T$ ∼[J] is also not realistic. Actually, it would be necessary to rescale all units to a reasonable range. This is forgone here for keeping the simplicity of the values. The *one-dimensional* systems observed here are artificial anyway, so it is not worthy to consider complex scaling factors for all the following measurements.

The chosen parameters for the test systems are given in the following. The mass of the harmonic oscillator (HO) is chosen to be $m = 2$kg. If not differently noted, the circular frequency is $\omega = 1\text{s}^{-1}$. The parameters of the quartic double well (QDW) are $D_0 = 1\text{Jm}^{-4}$ and $a = 1$m. The mass that is moving in the QDW potential is chosen to be $m = 1$kg.

Only a Single Nosé-Hoover Thermostat

As mentioned before, it is not possible to simulate the one-dimensional HO with a single Nosé-Hoover thermostat, which is equivalent to a NHC thermostat with chain length $M = 1$. Figure 2.14 shows a long-time trajectory of the simulation of a HO with such a thermostat. It is obvious that the system is trapped in a fix area of phase space. Therefore, the obtained distributions of position x and momentum p of the HO as well as the velocity of the Nosé-Hoover particle v_ξ, which are plotted in Fig. 2.15 and should be Gaussian, are completely wrong. From Fig. 2.16 it is clear that a different choice of the Nosé-Hoover mass does not solve this problem.

Choice of the Virtual Masses

Before starting to look at thermodynamic quantities, the idea for choosing the virtual masses as described in the previous section shall be tried out. This is done by measuring frequency spectra for simulations with several parameters.

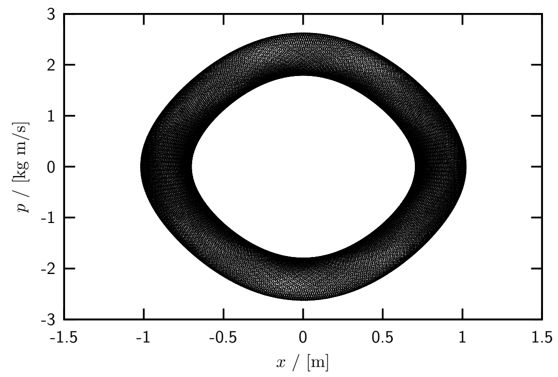


Figure 2.14: Trajectory of a simulation of the HO for 10^8 time steps $\delta t = 0.005$s at $k_BT = 1.0$J. Every 10^3 time steps, a dot is drawn. Only a single NH thermostat is used ($M = 1$), and the chosen time scale is $\tau = 1/\omega = 1$s.

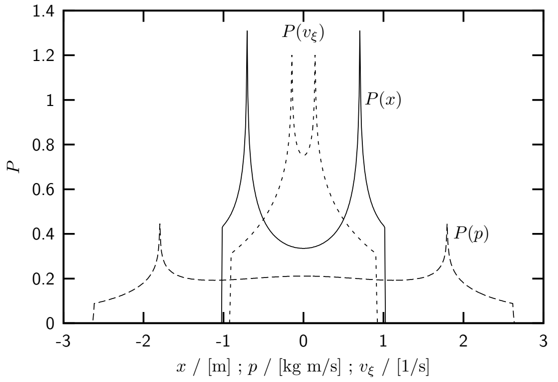


Figure 2.15: Distributions of position x and momentum p of the HO as well as the velocity of the NH particle for the same simulation as described in Fig. 2.14. Theoretically, all three histograms should have a Gaussian form.

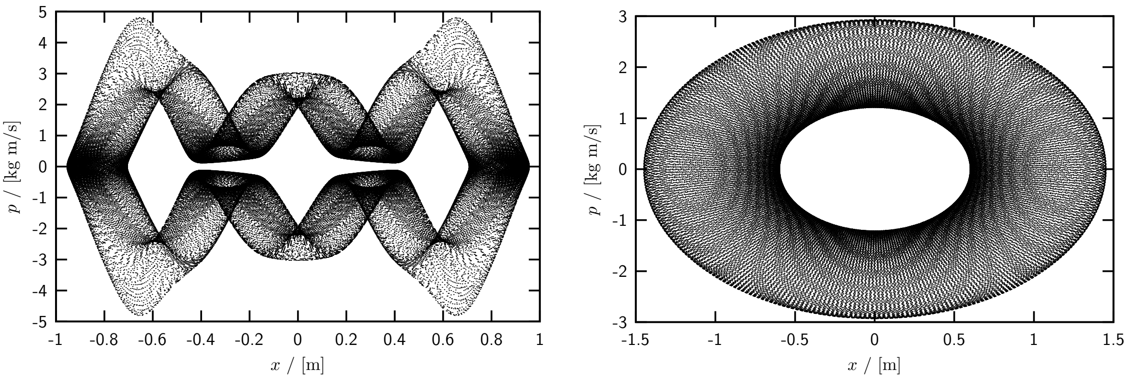


Figure 2.16: Trajectories of the HO, similar to the one in Fig. 2.14. Again a single Nosé-Hoover thermostat ($M = 1$) is utilised. The difference is that the time scale τ, which is used to calculate the virtual mass of the Nosé-Hoover particle, is chosen differently: (a) $\tau = 0.2$ and (b) $\tau = 5.0$. The plots are comparable to the pictures in Ref. [9].

Therefore, time series of the momentum p of both the HO and the QDW are measured. The MD step size is $\delta t = 0.005$s. The measurement is carried out at temperature $k_BT = 1.0$J. After equilibrating for 10^6 steps, a time series of 10^6 steps is calculated. Thereafter, a momentum autocorrelation function is calculated up to a distance of $5 \cdot 10^5$ steps. The Fourier transform of these functions gives the desired frequency spectra. The frequencies f in the following spectra of the HO are scaled by the expected frequency $f_0 = \omega/(2\pi)$ for better comparability. This is known, since the circular frequency ω can be directly selected for this system.

In Fig. 2.17 the result of choosing $\omega = 1\text{s}^{-1}$ and $\tau = 1/\omega = 1$s, which should lead to acceptable results in combination with (2.78) according to Ref. [10] as explained in the last section. Indeed, there is a definite peak at the expected main frequency of the HO. The finite

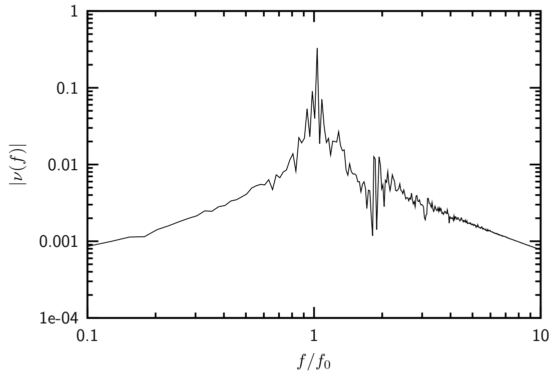


Figure 2.17: Frequency spectrum of HO at $\omega = 1\text{s}^{-1}$. The exact value $1/\omega = \tau = 1\text{s}$ is used for determining the virtual masses of the NHC.

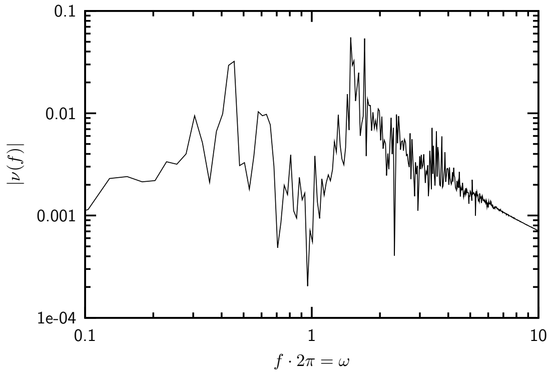


Figure 2.18: Frequency spectrum of QDW. As in Fig. 2.17, the NHC time scale is set to $\tau = 1\text{s}$, whereas there is no expectation about the frequencies of the QDW.

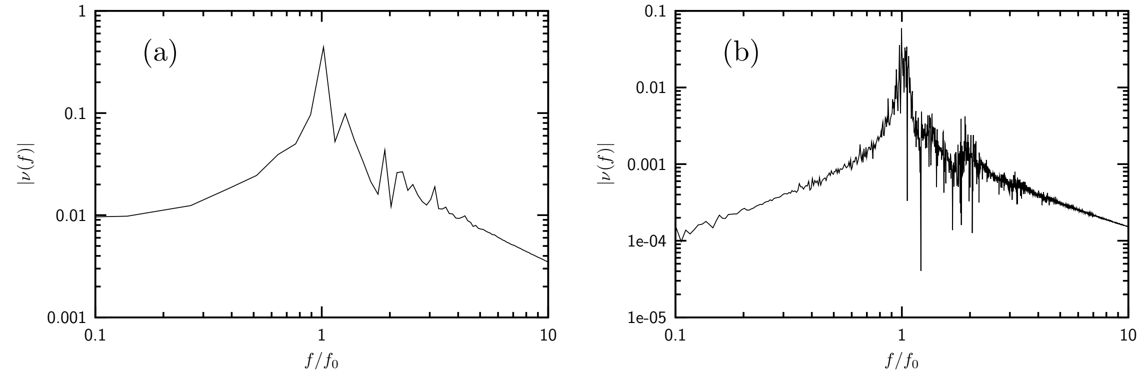


Figure 2.19: Frequency spectra of HO with (a) $\omega = 0.2\text{s}^{-1}$ and (b) $\omega = 5\text{s}^{-1}$. The time scale τ for the Nosé-Hoover thermostat is chosen "correctly" $\tau = 1/\omega$.

width of the peak is caused by the thermal fluctuations and is thus an expected effect rather than an artificial result from statistical errors.

Analogously, in Fig. 2.18, the outcome of the same simulation of the QDW is plotted. Since the frequency structure of the quartic double well is not obvious, this can be only taken as a measure of the main frequency to adjust the thermostat for further simulations of this system. The main peak is found at $\omega = 1.5\text{s}^{-1}$, which suggests a better choice of $\tau = 0.67\text{s}$.

From Fig. 2.19 it is obvious that the instructions of how to adjust the thermostat also work for $\omega \neq 1\text{s}^{-1}$. Again, a clear peak arises at the expected frequency f_0. The different look of the plots in Figs. 2.17 and 2.19 is mainly caused by the different normalisation of the abscissa, while the frequency resolution is equal in all the three measurements.

Up to now, it is not clear, if another choice of τ really causes worse results. This question is answered in Figs. 2.20 and 2.21. There, the Nosé-Hoover time scale is deliberately chosen different from $1/\omega$. The upper pictures show the result of $\tau = 0.2\text{s}$ and the lower belong to $\tau = 5\text{s}$ respectively. From the frequency spectra in Fig. 2.20 it looks as if $\tau = 5\text{s}$ is even a

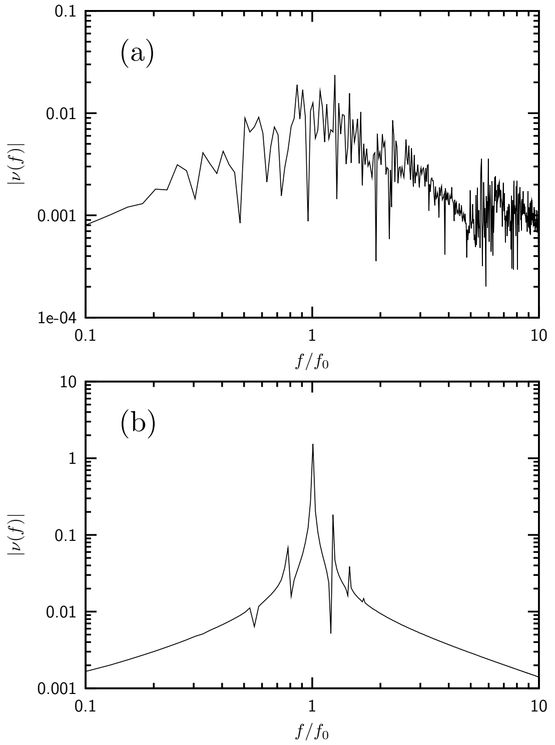


Figure 2.20: Frequency spectra of harmonic oscillators at $\omega = 1\mathrm{s}^{-1}$ with Nosé-Hoover time scale (a) $\tau = 0.2$s (above) and (b) $\tau = 5$s (below).

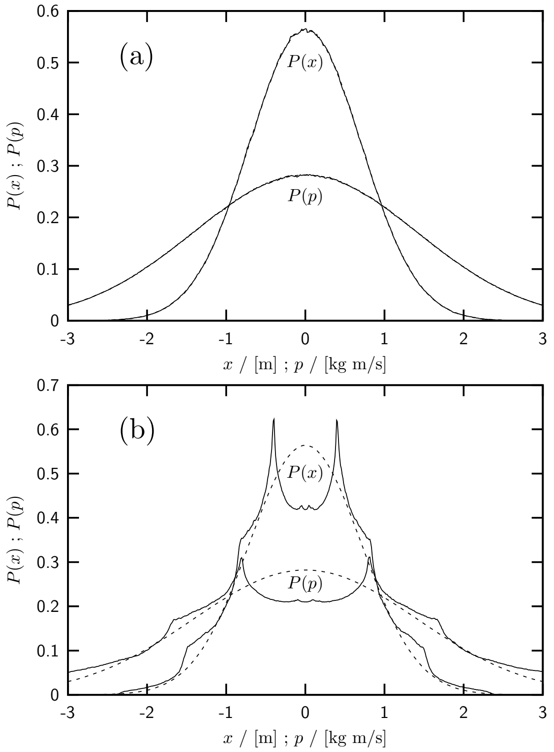


Figure 2.21: Distributions of x and p of harmonic oscillator at $\omega = 1\mathrm{s}^{-1}$ with Nosé-Hoover time scale (a) $\tau = 0.2$s (above) and (b) $\tau = 5$s (below). The measured data is drawn with the solid line, the dashed lines depict exact results.

better choice, compared to $\tau = 1$s from Fig. 2.17, where the spectrum is more noisy. Whereas, for $\tau = 0.2$ the spectrum is seriously altered and there is no clear peak at $f = f_0$ anymore. However, the distributions of position and velocity, which are shown in Fig. 2.21, completely disprove the latter assumptions. The distributions which are obtained by using a thermostat with $\tau = 0.2$s matches the exact results very well, *although* the spectrum (Fig. 2.20) is so noisy. On the other hand, the outcome of the simulation with $\tau = 0.5$s is totally wrong.

This can be taken as a proof that the adjustment of the Nosé-Hoover thermostat explained in the last section (according to [10]) works correctly.

Thermodynamics with Correct Thermostating

After being confident that the proposed method for adjusting the thermostat (2.78) works correctly, it is now examined, if the thermostat is really capable of reproducing the canonical ensemble. Many thermodynamic quantities have been calculated for the two test systems in section 2.3. Therefore, all the produced data can be compared to exact results, which is a big advantage.

All the following simulations used a Nosé-Hoover-Chain thermostat with $M = 2$. The observed systems are as described in the beginning of the section. The time scale τ is $\tau =$

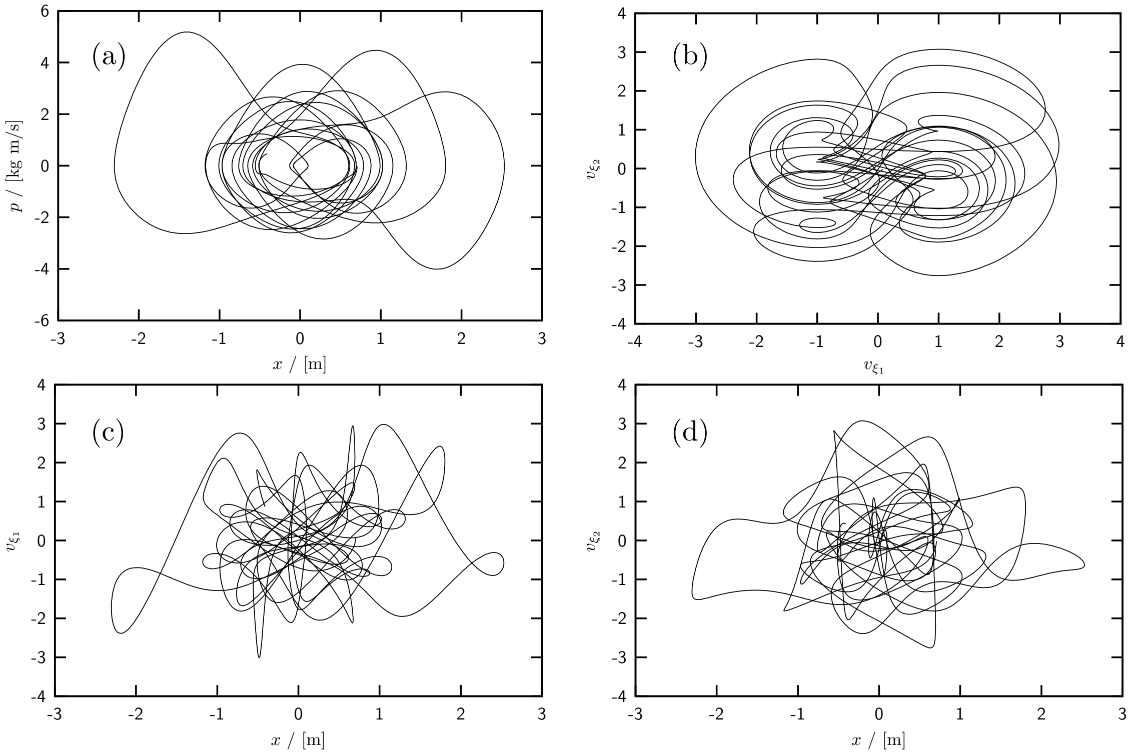


Figure 2.22: Several projections of the trajectory of a harmonic oscillator with a NHC thermostat ($M = 2$) at $k_B T = 1.0$J. The trajectory shows a length of 10^5 MD time steps. The projections are: (a) x - p, (b) v_{ξ_1} - v_{ξ_2}, (c) x - v_{ξ_1}, (d) x - v_{ξ_2}.

$1/\omega = 1$s for HO simulations and $\tau = 0.67$s for simulations of the QDW, according to Fig. 2.18. After equilibrating the system for 10^6 steps, a measurement of 10^8 steps is carried out. The chosen step size is $\delta t = 0.005$s. This makes a total length of the run of $t = 5 \cdot 10^5$s.

To get an impression of the action of a NHC, several projections of the six-dimensional[1] trajectory of a harmonic oscillator with a NHC ($M = 2$) are drawn in Fig. 2.22.

In Figs. 2.23 and 2.24, long time trajectories of both the HO and the QDW are shown. It is obvious, that the deterministic changes on a short-time scale as seen in Fig. 2.22 lead to a correct sampling of the phase space for a reasonably long trajectory. The comparison with the respective distributions in Fig. 2.25 and 2.26 shows that the sampling really matches the theoretical predictions, except for the low temperature simulation ($k_B T = 0.1$J) of the QDW.

For the QDW, the probability to change from one of the two potential minima to the other one is suppressed exponentially due to the Boltzmann factor with respect to the height of the potential energy barrier at $x = 0$ normalised by the thermal energy of the system. Thus, for low temperatures the system remains in one of the two potential valleys for a long time, and exchanges between the valleys are extremely seldom. This is the reason, why one of the two peaks of the distribution in Figs. 2.26 is more pronounced than the other one. Also from the upper picture in Fig. 2.24 it can be assumed that the right valley is more populated than the left. This is a statistical error that vanishes for very long run times. In other words, the

[1] There are three degrees of freedom in the whole system: the one-dimensional harmonic oscillator and the two Nosé-Hoover particles. Each of these has a position and a velocity, spanning a six-dimensional phase space.

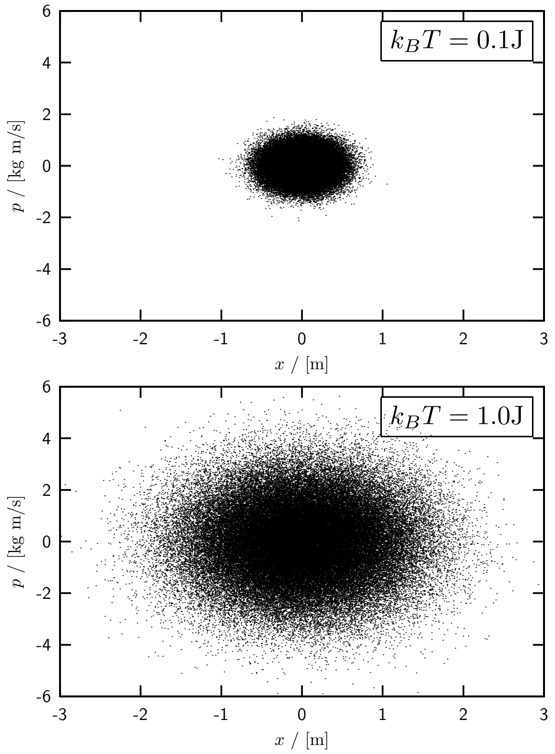


Figure 2.23: Trajectory of a simulation of the HO for 10^8 time steps $\delta t = 0.005$s at (a) $k_BT = 0.1$J and (b) $k_BT = 1.0$J. A dot marks the position of the system after every 10^3 time steps.

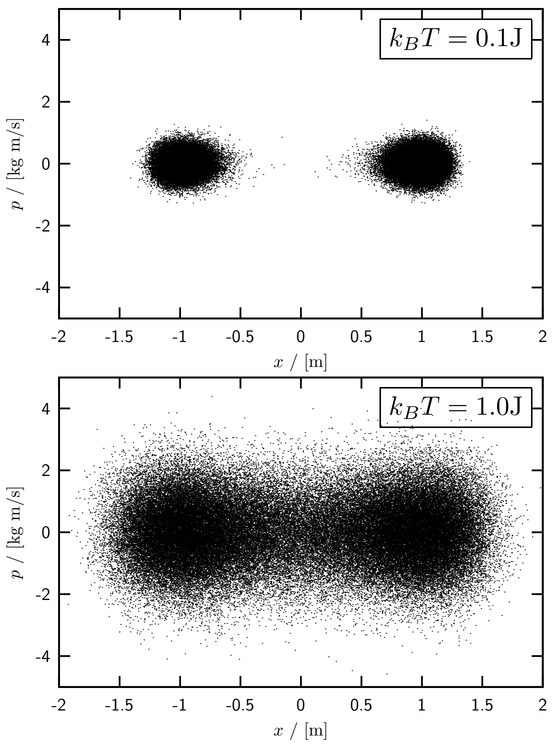


Figure 2.24: Trajectories of the QDW at (a) $k_BT = 0.1$J and (b) $k_BT = 1.0$J obtained similarily as in Fig. 2.23.

autocorrelation (see chapter 3) of the system is very large at low temperatures.

There is nothing remarkable about the specific heat plots in Fig. 2.27. The error bars cover the analytically calculated heat capacity.

Besides the distributions of position and velocity, it was also possible to derive energy distributions analytically for the two test systems (see section 2.3). The comparisons of the measured histograms of potential and kinetic energy are shown in Figs. 2.28 and 2.29. For the HO, the potential and the kinetic energy are equivalent. Both histograms (solid lines) match the theoretic prediction (dashed lines). This is also the case for the kinetic energy histogram of the QDW, which is the same as for the HO.

The only remarkable detail about these three plots is that the measured histogram entry for low energies is a little bit too large, compared to the exact result. This could be caused by the fact that equidistant histogram bins were used. Thus, the plot is obtained by connecting the points at the middle of each histogram bin. This presumes that the plotted function (the energy histogram in this case) can be approximated by calculating averages with the trapezoidal rule ($f(x) \approx (f(x + \epsilon/2) + f(x - \epsilon/2))/2$). For low energies, there is a divergence in the energy distribution, and thus this assumption does not hold.

The only result, which shows some serious deviation from the analytic prediction, is the potential energy distribution of the QDW (Fig. 2.29 (a)). For both the simulation at $k_BT =$

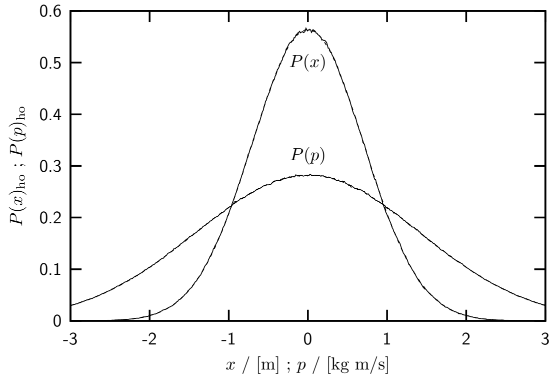


Figure 2.25: Distributions of x and p for the HO at $k_BT = 1.0$J. The solid line depicts the measured result. The exact results are also drawn with dashed lines, but the match is perfect, so they are overlapped by the simulation data.

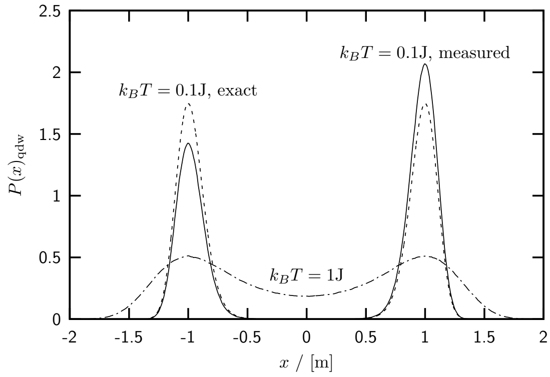


Figure 2.26: Distribution of x for the QDW at $k_BT = 0.1$J and $k_BT = 1.0$J. These are the same temperatures as in Fig. 2.23 above, where the density of the plot in the x-direction is proportional to $P(x)_{\text{qdw}}$. The solid line depicts the measured result, the dashed line shows the analytic finding.

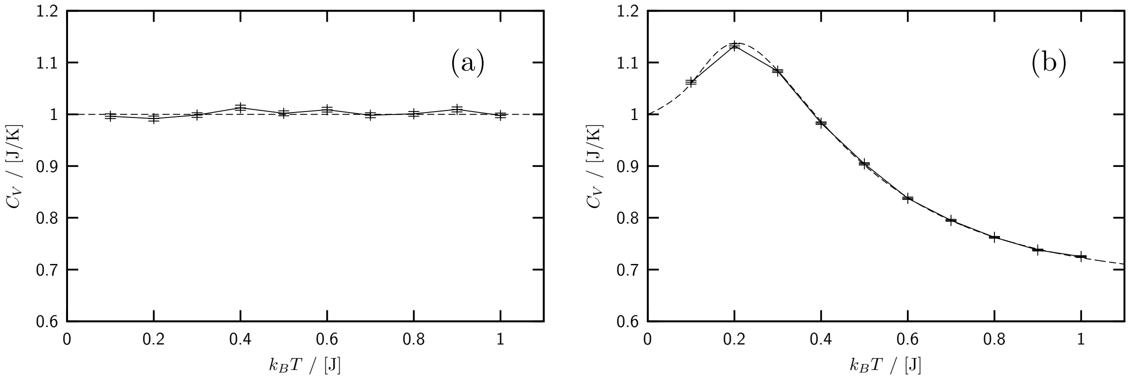


Figure 2.27: Specific heats of (a) HO and (b) QDW. The symbols are the average results from the measurements, with error bars from a Jackknife analysis (see chapter 3). The dashed lines denote the exact result.

0.1J and $k_BT = 1.0$J, the histogram obtained by simulation drops quickly at $E_{\text{pot}} = 1$J. This is certainly caused by some anomaly of the Nosé-Hoover thermostat. However, since all the other tests showed outstanding results, this is not taken too serious here.

The convergence of the algorithm is assessed by measuring time series of several moments of the momentum p. The outcome shown in Fig. 2.30 is as expected. The relative deviations from the exact result are larger for higher moments, but these deviations are decaying quickly. At least at $t = 5000$s, which is equivalent to 10^6 time steps, the errors are acceptably small. This is the length of the equilibration phase in the other simulations. Because for this measurement, the behaviour from the first time step on was interesting, no equilibration was applied to the system.

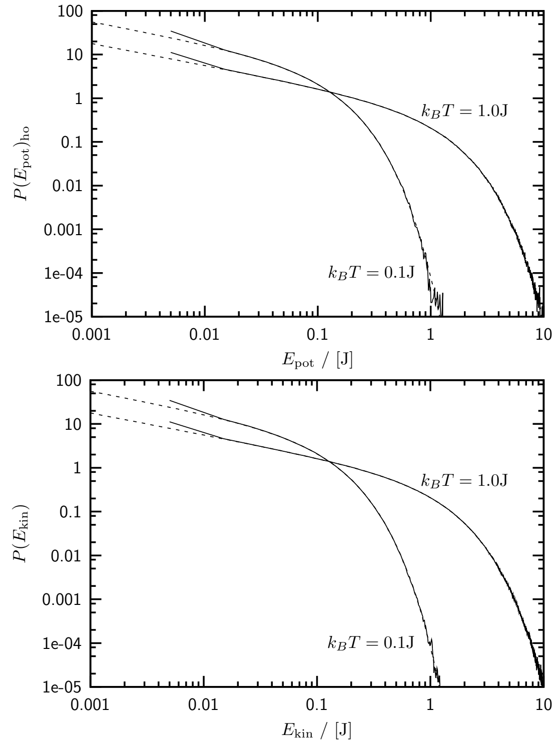


Figure 2.28: Distributions of E_{pot} and E_{kin} of the HO, each at $k_BT = 0.1$J and $k_BT = 1.0$J. Solid lines depict the measurements, the exact results are indicated with a dashed line. The double logarithmic scale is chosen to visualise the deviations both at the divergence near $E = 0$ and in the tails.

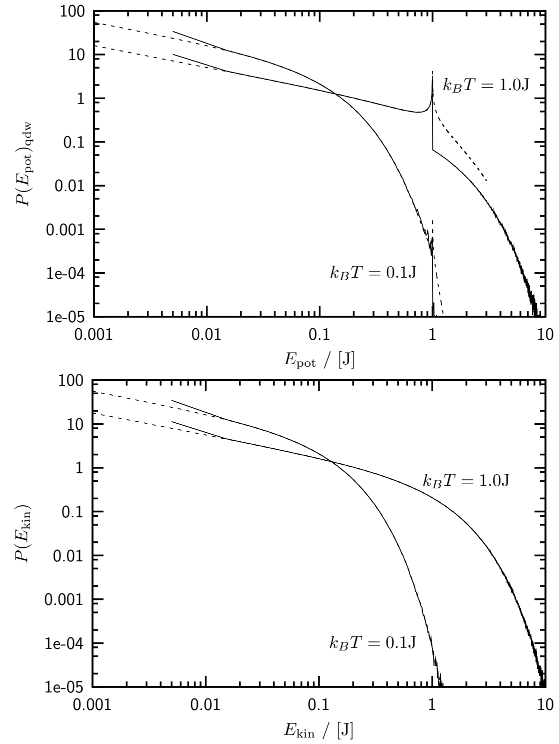


Figure 2.29: Distributions of E_{pot} and E_{kin} similar to Fig. 2.28 for the quartic double well (QDW).

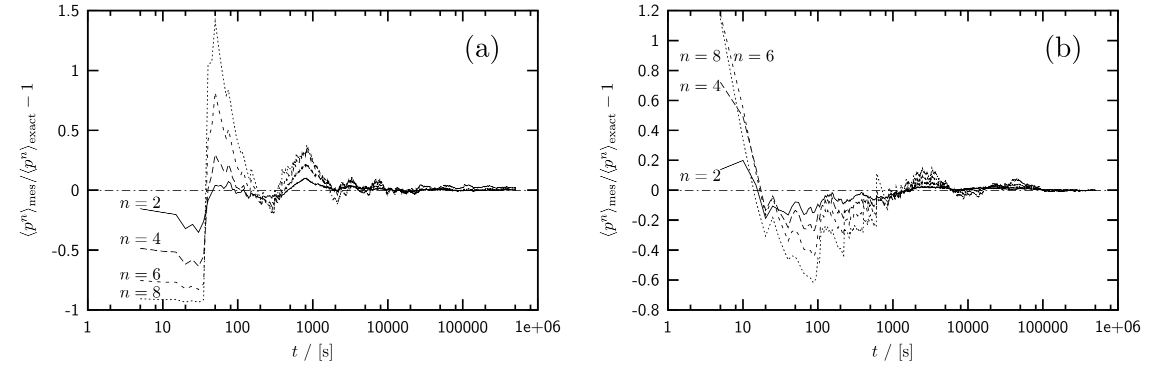


Figure 2.30: Time series of the relative deviation of the running averages of several non-vanishing moments of the momentum from the exact result according to (2.40) for (a) HO and (b) QDW. For the purpose of checking the convergence especially the first part of the simulation is of interest. Thus, the averages are already recorded during the equilibration phase. A logarithmic scale of the abscissa is chosen to make it possible to see both the strong fluctuations at the very beginning of the run and the long-time relaxation behaviour.

Chapter 3

Monte Carlo Simulation Methods and Data Analysis

3.1 Simulation

One of the goals of the work at hand is to get a basic understanding of Molecular Dynamics simulations at finite temperature. This topic is covered in an individual chapter 2. Another big class of computer experiments are Monte Carlo simulations. In this section, the utilised Monte Carlo methods shall be only briefly introduced.

The principle of Monte Carlo simulations is to sample the phase space by a random walk at finite temperature. This is done in order to collect information about the density of states $\Omega(E)$, from which all thermodynamic quantities can be calculated. The Boltzmann distribution that gives the probability of sampling states $\mathbf{X}$ with a certain energy $E = E(\mathbf{X})$:

$$P_\beta(E) = \Omega(E) e^{-\beta E} \ . \tag{3.1}$$

Here, $\beta = 1/k_B T$ is the inverse temperature.

The simplest approach would be to do "simple sampling" and just randomly choose states and average them with respect to the Boltzmann weight $\exp[-\beta E(\mathbf{X})]$. However, for any nontrivial system, most of the randomly generated states will have a high potential energy and thus do virtually *not* contribute to the partition sum. Therefore, the "importance sampling" was invented, which is suitable to find relevant states.

3.1.1 Metropolis Sampling

The first method with general validity was the Metropolis algorithm, published in 1953 [11]. The simulation starts from any random configuration. Afterwards, configurational changes, so called "updates", are proposed to the system. The update is accepted with a probability depending on the potential energy of the old and the new configuration:

$$\begin{aligned} w(\mathbf{X}^{\text{old}} \to \mathbf{X}^{\text{new}}) &= \begin{cases} 1 \ , & \text{if} \quad E^{\text{old}} > E^{\text{new}} \ , \\ \exp\left[-\beta(E^{\text{new}} - E^{\text{old}})\right] \ , & \text{if} \quad E^{\text{old}} < E^{\text{new}} \end{cases} \\ &= \min\left(1, \exp\left[-\beta(E^{\text{new}} - E^{\text{old}})\right]\right) \ . \end{aligned} \tag{3.2}$$

The big advantage of the Metropolis algorithm is its universality. It is possible to use it with *any* model system, i.e., any rule of calculating a potential energy from a given state of the system. A serious drawback is in contrast that the probability of overcoming barriers E_{bar} in the free-energy landscape is proportional to the Boltzmann factor $\exp[-\beta E_{\text{bar}}]$. Therefore, sampling can be trapped in local minima especially at low temperatures.

3.1.2 Parallel Tempering

Over the years a number of sophisticated algorithms have been developed to get rid of this problem. One of these is the parallel tempering (PT) algorithm introduced by Hukushima and Nemoto in 1996 [12]. In this generalised-ensemble method, several replica of the system are simulated in parallel with, e.g., a Metropolis algorithm at different temperatures. After a number τ_{PT} of independent Metropolis sweeps, an exchange between two configurations $\mathbf{X}$ and $\mathbf{X}'$ at the inverse temperatures β and β' is suggested and accepted with the following exchange probability:

$$\begin{aligned} w(\mathbf{X} \leftrightarrow \mathbf{X}'; \beta, \beta') &= \begin{cases} 1 \ , & \text{if} \quad \Delta < 0 \ , \\ e^{-\Delta} \ , & \text{if} \quad \Delta > 0 \end{cases} \\ &= \min\left(1, e^{-\Delta}\right) \ , \\ \Delta &= (\beta' - \beta)(E(\mathbf{X}) - E(\mathbf{X}')) \ . \end{aligned} \tag{3.3}$$

The key idea is that due to frequent exchanges of the configurations to low and high temperatures, valleys of the free-energy landscape are sampled thorough enough to gain information for the low temperature range and high barriers can be circumvented. Due to the dependency of the transition probability on energy and temperature, a suitable acceptance rate of exchanges is obtained for acceptably overlapping potential energy histograms for the different temperatures. The overlap of the energy histograms has another benefit as it allows the application of reweighting methods (see section 3.2.2). On the other hand, to make use of the advantages of the PT algorithm, the temperature range should be as broad as possible. Thus, some thought has to be put into the choice of the number of replica N_{R}, the temperature range and the distribution of the single replica over the latter. This is discussed in detail in section 4.1.1.

The character of the algorithm makes the parallelisation with, e.g., MPI easy and fast, even for distributed memory systems. To save communication, each replica is covered on one single processor. When actually exchanging configurations, it is also possible to only exchange the temperature.

3.2 Statistical Analysis

3.2.1 Error Analysis

The error analysis in methods of statistical physics is a very delicate task. In [34] this is discussed in detail. Here, only a brief summary shall be presented to explain the applied method in the work at hand, the Jackknife error analysis.

Autocorrelation

Purely stochastic data is uncorrelated and thus the statistical error of the average $\overline{\mathcal{O}}$ of a number of N measurements $\mathcal{O}_i$ is the root of the variance of the single events:

$$\sigma^2_{\overline{\mathcal{O}}} = \frac{\sigma^2_{\mathcal{O}_i}}{N} = \frac{\langle \mathcal{O}_i^2 \rangle - \langle \mathcal{O}_i \rangle^2}{N} \ . \tag{3.4}$$

In a Monte Carlo simulation an intrinsic memory exists, resulting from the fact that each configuration is evolved from the previous one by a certain alteration – the update. Thus, the obtained data are *correlated.* The explicit memory does only reach to the previous state. However, the previous state covers information about the last but one and so forth, which leads to a decaying long-range correlation. The correlation can be measured with the normalised autocorrelation function:

$$A(k) = \frac{\langle \mathcal{O}_i \mathcal{O}_{i+k} \rangle - \langle \mathcal{O}_i \rangle^2}{\langle \mathcal{O}_i^2 \rangle - \langle \mathcal{O}_i \rangle^2} \ . \tag{3.5}$$

$A(k)$ denotes the correlation of two individual measurements with a time separation of k simulation iterations. For $k = 0$ it holds $A(k) = 1$ due to the normalisation. To have suitable data for $A(k)$, the simulation has to be much longer than the considered time separations $k \ll N$. From the autocorrelation function the integrated autocorrelation time $\tau'_{\mathcal{O},\mathrm{int}}$ can be calculated:

$$\tau'_{\mathcal{O},\mathrm{int}} = \frac{1}{2} + \sum_{k=1}^{N} A(k) \left(1 - \frac{k}{N} \right) \ . \tag{3.6}$$

This goes directly into the error estimation:

$$\sigma^2_{\overline{\mathcal{O}}} = \frac{\sigma^2_{\mathcal{O}_i}}{N} 2\tau'_{\mathcal{O},\mathrm{int}} \ . \tag{3.7}$$

Therefore, the number of measurements is effectively decreased due to the correlations by a factor $2\tau'_{\mathcal{O},\mathrm{int}}$. The calculation of $\tau'_{\mathcal{O},\mathrm{int}}$ is not trivial, since the statistical noise in the tail of the autocorrelation function always adulterates the result.

For large time separations k, the autocorrelation function decays exponentially. From this behaviour, the exponential autocorrelation time $\tau_{\mathcal{O},\mathrm{exp}}$ can be estimated:

$$A(k) \overset{k\to\infty}{\longrightarrow} a \exp\left[-\frac{k}{\tau_{\mathcal{O},\mathrm{exp}}} \right] \ . \tag{3.8}$$

It can be shown that $\tau'_{\mathcal{O},\mathrm{int}} \leq \tau_{\mathcal{O},\mathrm{exp}}$. Therefore, fitting the measured autocorrelation function by the exponential behaviour (3.8) can give a confident insight in the statistical effect of the correlations in a simulation.

Autocorrelation times are especially large for low temperatures and at critical points, where transitions between different conformational areas of the state space take place. Hence, an analysis of autocorrelation at a low temperature, or at a known transition temperature, gives a “worst case” estimation that vaguely holds for the whole temperature range.

Jackknife Analysis

Since it would be quite exhausting to employ the whole analysis described above for every single measurement, general but less accurate methods have been developed to estimate the error of a Monte Carlo run. Both the Binning and the Jackknife analysis divide the data of a simulation $\mathcal{O}_i$ in blocks. The Binning analysis considers a number $N_B = N/k$ of equidistant blocks of length k.

$$\mathcal{O}_{B,n} = \frac{1}{k} \sum_{i=1}^{k} \mathcal{O}_{(n-1)k+i} \ , \qquad n = 1, \ldots, N_B \ . \tag{3.9}$$

In contrast, a Jackknife block $\mathcal{O}_{J,n}$ contains all data except for the data in one certain Binning block:

$$\mathcal{O}_{J,n} = \frac{N\overline{\mathcal{O}} - k\mathcal{O}_{B,n}}{N-k} \ . \tag{3.10}$$

The benefit of the Jackknife method is that the statistics of each block is better than in the Binning approach. Actually, for linear quantities like the energy, both methods give an analytically equal result. Deviations do only occur, when nonlinear quantities like the mean square energy are treated, as it is necessary for evaluating the specific heat. Considering bias effects and the raised statistics of the longer Jackknifing blocks, the error of the estimator $\overline{\mathcal{O}}$ can be evaluated:

$$\sigma^2_{\overline{\mathcal{O}}} = \frac{N_B - 1}{N_B} \sum_{n=1}^{N_B} \left(\mathcal{O}_{J,n} - \overline{\mathcal{O}}\right)^2 \tag{3.11}$$

The whole methods presumes that the data in different blocks is uncorrelated, i.e. $k \gg \tau_{\text{int}}$. However, an empiric rule is that $k \approx 6\tau_{\text{exp}}$ gives acceptable results. Also, if k is chosen too small, the resulting error is also too small. Therefore, calculating the estimated error for different block sizes k can help choosing a reasonable bin size. If increasing k does not alter σ^2 considerably, the bin size is large enough. Furthermore, an interesting feature is that for large k the product $k\sigma^2$ converges against $2\tau_{\text{int}}$. So the observation of the evolving plateau of $k\sigma^2$ for large k allows the estimation of τ_{int}.

3.2.2 Reweighting

In the canonical ensemble, i.e. at a given volume V, particle number N and temperature T, the thermodynamic properties are guided by the Boltzmann distribution. The probability to observe states with a certain energy E is (as already given in (3.1)):

$$P_\beta(E) = \Omega(E) e^{-\beta E} \ . \tag{3.12}$$

Thus, if the *exact* energy distribution at any temperature $T = 1/k_B\beta$ is known, the density of states $\Omega(E)$ can be simply calculated from this single energy histogram. If $P_\beta(E)$ is not known exactly, it still covers information about the density of states for an energy range, where it has reliable values, i.e. around its maximum. By the knowledge of $P_\beta(E)$, the average of any function of E can be evaluated at β:

$$\langle f(E) \rangle_\beta = \frac{\sum\limits_E f(E) P_\beta(E)}{\sum\limits_E P_\beta(E)} \ . \tag{3.13}$$

Reweighting means the calculation of energy histograms for various temperatures from a single measured distribution $P_{\beta_0}(E)$. Actually, the method is better referred to as "single histogram reweighting", since all the information input comes from a single histogram at temperature β_0. In doing so, in combination with (3.15) it gets possible to calculate $\langle f(E)\rangle$ at temperatures $\beta \neq \beta_0$. The method is described in detail in [35]. Considering (3.12) it is easy to see that from a given histogram $P_{\beta_0}(E)$ at inverse temperature β_0, the energy distribution $P_\beta(E)$ at inverse temperature β can be calculated as follows:

$$P_\beta(E) = P_{\beta_0}(E)e^{-(\beta-\beta_0)E} \ , \tag{3.14}$$

which leads to an expression for $\langle f(E)\rangle_\beta$:

$$\langle f(E)\rangle_\beta = \frac{\sum\limits_E f(E)P_{\beta_0}(E)e^{-(\beta-\beta_0)E}}{\sum\limits_E P_{\beta_0}(E)e^{-(\beta-\beta_0)E}} \ . \tag{3.15}$$

This approach is only possible in a narrow temperature range, where the erroneous tails of the reweighted histograms have a negligible influence on the results.

Multiple Histogram Reweighting

Ferrenberg and Swendsen derived a technique to use a number of m overlapping histograms for virtually extending the range where reweighting gives acceptable results as desired [36]. The key point is that the density of states is calculated from any of the implied histograms $P_{\beta_i}(E)$ by inverting (3.12). By trivial averaging over the so obtained $\Omega_i(E)$, the errors from the tails of $P(E)$, which are multiplied by gigantic values $e^{\beta E}$ due to the inverted equation (3.12), would lead to a chaotic guess of $\overline{\Omega}(E)$. Therefore, the error of each value $P_{\beta_i}(E)$ is approximated by $\sim \sqrt{P_{\beta_i}(E)}$. The error weighted average of x_i with variance σ_i^2 is [37]:

$$\overline{x} = \frac{\sum\limits_i x_i/\sigma_i^2}{\sum\limits_i 1/\sigma_i^2} \ . \tag{3.16}$$

Thus, the error weighted combined histogram gives:

$$\Omega(E) = \frac{\sum\limits_{i=1}^m P_{\beta_i}(E)}{\sum\limits_{i=1}^m Z(\beta_i)^{-1}e^{-\beta_i E}} \ . \tag{3.17}$$

The values of the partition function $Z(\beta_i)$ are not known, but can be self-consistently determined from the density of states:

$$Z'(\beta_i) = \sum_E \Omega(E)e^{-\beta_i E} = \sum_E e^{-\beta_i E}\frac{\sum\limits_{k=1}^m P_{\beta_k}(E)}{\sum\limits_{k=1}^m Z(\beta_k)^{-1}e^{-\beta_k E}} \ . \tag{3.18}$$

The whole reweighting procedure is thus an iterative algorithm which consisting of the following steps:

1. start from any choice for all $Z(\beta_i)$ (e.g. $Z(\beta_i) \equiv 1 \,\forall\, i$),
2. calculate $\Omega(E)$ according to (3.17),
3. with the so obtained $\Omega(E)$ calculate $Z'(\beta_i)$ according to (3.18),
4. if $Z(\beta_i) \approx Z'(\beta_i) \,\forall\, i \rightarrow$ ready, else $Z(\beta_i) := Z'(\beta_i) \rightarrow$ return to step 2.

According to [37] a reasonable interruption condition is to check the relative deviation of $Z(\beta_i)$ and $Z'(\beta_i)$:

$$\Delta^2 = \sum_i \left[\frac{Z'(\beta_i) - Z(\beta_i)}{Z'(\beta_i)} \right]^2 \ . \tag{3.19}$$

If Δ^2 under-runs a limit value ϵ^2, the iteration has reached the desired accuracy.

Usually, the partition function $Z(\beta_i)$ spans many orders of magnitudes. Even when normalising it to values around 1, it happens frequently that the exponent overrides the provided range even for double-precision variables. Therefore, it can be useful to work with the logarithm of the partition function $\ln Z$ instead of Z. In doing so, the product $Z^{-1}e^{-\beta E}$ simply gets $-\ln(Z) - \beta E$. Unfortunately, there is also a *sum* running over $Z^{-1}e^{-\beta E}$. A solution of the problem how to calculate $\ln C = \ln(A+B)$ by only knowing $\ln A$ and $\ln B$ is provided in [38]:

$$\begin{aligned} \ln C = \ln(A+B) &= \ln \left[\max(A,B) \left(1 + \frac{\min(A,B)}{\max(A,B)} \right) \right] \\ &= \max(\ln A, \ln B) + \ln \left\{ 1 + \exp\left[\min(\ln A, \ln B) - \max(\ln A, \ln B) \right] \right\} \ . \end{aligned} \tag{3.20}$$

Reweighting of Non-Energy Quantities

So far, it is only possible to calculate averages over functions of the energy E from the density of states as obtained by the procedure described above:

$$\langle f(E) \rangle_\beta = \frac{\sum\limits_E f(E)\Omega(E)e^{-\beta E}}{\sum\limits_E \Omega(E)e^{-\beta E}} \ . \tag{3.21}$$

By simply introducing an additional degree of freedom for the density of states, any quantity can be reweighted by an analogous procedure. Let $P_{\beta_i}(E, \mathcal{A})$ be two-dimensional histograms covering the number of events that matched a certain range of energy E and any quantity $\mathcal{A}$ at inverse temperature β_i. The above equations (3.17) and (3.18) can be rewritten as:

$$\Omega(E, \mathcal{A}) = \frac{\sum\limits_{i=1}^{m} P_{\beta_i}(E, \mathcal{A})}{\sum\limits_{i=1}^{m} Z(\beta_i)^{-1} e^{-\beta_i E}} \ , \tag{3.22}$$

$$Z'(\beta_i) = \sum_E e^{-\beta_i E} \frac{\sum\limits_{k=1}^{m} \overbrace{\sum_{\mathcal{A}} P_{\beta_k}(E, \mathcal{A})}^{P_{\beta_k}(E)}}{\sum\limits_{k=1}^{m} Z(\beta_k)^{-1} e^{-\beta_k E}} \overset{(3.18)}{=} Z(\beta_i) \ . \tag{3.23}$$

Therefore, once the partition function $Z(\beta)$ has been iterated, it can also be used for calculating higher dimensional densities of states as seen in (3.22). Now, histograms of $\mathcal{A}$ and mixed functions $f(E,\mathcal{A})$ can be evaluated at any temperature within a reasonable range that is spanned by β_i:

$$P_\beta(\mathcal{A}) \sim \sum_E \Omega(E,\mathcal{A}) e^{-\beta E} \ , \tag{3.24}$$

$$\langle f(E,\mathcal{A})\rangle_\beta = \frac{\sum\limits_E \left(\sum\limits_{\mathcal{A}} f(E,\mathcal{A})\Omega(E,\mathcal{A})\right) e^{-\beta E}}{\sum\limits_E \underbrace{\left(\sum\limits_{\mathcal{A}} \Omega(E,\mathcal{A})\right)}_{\Omega(E)} e^{-\beta E}} \ . \tag{3.25}$$

3.2.3 Fluctuation Quantities in Molecular Dynamics Simulations

In conservative systems, the whole energy $E = \mathcal{H}(\mathbf{r},\mathbf{p})$ can be fully separated into a potential energy term E_{pot} which depends on the positions only $E_{\text{pot}} = \mathcal{H}_{\text{pot}}(\mathbf{r})$, and the kinetic energy, which is of course only dependent on the momenta: $E_{\text{kin}} = \mathcal{H}_{\text{kin}}(\mathbf{p})$. The fluctuation of a quantity $\mathcal{A}$ with respect to temperature can be written as:

$$\frac{\partial}{\partial T}\langle\mathcal{A}\rangle = k_B\beta^2 \left(\langle\mathcal{A}E\rangle - \langle\mathcal{A}\rangle\langle E\rangle\right) \ . \tag{3.26}$$

In Monte Carlo simulations the kinetic energy is usually not considered, thus "E" in (3.26) always means E_{pot}. In a Molecular Dynamics simulation kinetic energy and velocities appear. Now it shall be shown that both the whole system energy E and the potential energy E_{pot} are suitable to calculate fluctuation quantities as in (3.26), provided that the system's potential energy *and* the quantity do not have any velocity dependence ($\mathcal{A} \equiv \mathcal{A}(\mathbf{r})$, $E_{\text{pot}} \equiv \mathcal{H}_{\text{pot}}(\mathbf{r})$):

$$\begin{aligned}
\frac{\partial}{\partial T}\langle\mathcal{A}\rangle &= \underbrace{\frac{\partial\beta}{\partial T}}_{-k_B\beta^2} \frac{\partial}{\partial\beta} \frac{\int \mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{p}\,\mathcal{A}e^{-\beta\mathcal{H}(\mathbf{r},\mathbf{p})}}{\int \mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{p}\,e^{-\beta\mathcal{H}(\mathbf{r},\mathbf{p})}} \\
&= -k_B\beta^2 \Bigg[\underbrace{\frac{\int \mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{p}\,\mathcal{A}\left(-\mathcal{H}(\mathbf{r},\mathbf{p})\right) e^{-\beta\mathcal{H}(\mathbf{r},\mathbf{p})}}{\int \mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{p}\,e^{-\beta\mathcal{H}(\mathbf{r},\mathbf{p})}}}_{-\langle\mathcal{A}E\rangle} \\
&\qquad - \underbrace{\frac{\int \mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{p}\,\mathcal{A}e^{-\beta\mathcal{H}(\mathbf{r},\mathbf{p})}}{\int \mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{p}\,e^{-\beta\mathcal{H}(\mathbf{r},\mathbf{p})}}}_{\langle\mathcal{A}\rangle} \underbrace{\frac{\int \mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{p}\,\left(-\mathcal{H}(\mathbf{r},\mathbf{p})\right) e^{-\beta\mathcal{H}(\mathbf{r},\mathbf{p})}}{\int \mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{p}\,e^{-\beta\mathcal{H}(\mathbf{r},\mathbf{p})}}}_{\langle -E\rangle} \Bigg] \\
&= k_B\beta^2 \left(\langle\mathcal{A}E\rangle - \langle\mathcal{A}\rangle\langle E\rangle\right) \ , \qquad (3.27)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial T}\langle\mathcal{A}\rangle &= -k_B\beta^2 \frac{\partial}{\partial\beta} \frac{\int \mathrm{d}\mathbf{r}\,\mathcal{A}e^{-\beta\mathcal{H}_{\text{pot}}(\mathbf{r})} \int \mathrm{d}\mathbf{p}\,e^{-\beta\mathcal{H}_{\text{kin}}(\mathbf{p})}}{\int \mathrm{d}\mathbf{r}\,e^{-\beta\mathcal{H}_{\text{pot}}(\mathbf{r})} \int \mathrm{d}\mathbf{p}\,e^{-\beta\mathcal{H}_{\text{kin}}(\mathbf{p})}} = -k_B\beta^2 \frac{\partial}{\partial\beta} \frac{\int \mathrm{d}\mathbf{r}\,\mathcal{A}e^{-\beta\mathcal{H}_{\text{pot}}(\mathbf{r})}}{\int \mathrm{d}\mathbf{r}\,e^{-\beta\mathcal{H}_{\text{pot}}(\mathbf{r})}} \\
&= -k_B\beta^2 \Bigg[\underbrace{\frac{\int \mathrm{d}\mathbf{r}\,\mathcal{A}\left(-\mathcal{H}_{\text{pot}}(\mathbf{r})\right) e^{-\beta\mathcal{H}_{\text{pot}}(\mathbf{r})}}{\int \mathrm{d}\mathbf{r}\,e^{-\beta\mathcal{H}_{\text{pot}}(\mathbf{r})}}}_{-\langle\mathcal{A}E_{\text{pot}}\rangle}
\end{aligned}$$

$$
\left. - \underbrace{\frac{\int \mathrm{d}\mathbf{r}\,\mathcal{A}e^{-\beta\mathcal{H}_{\mathrm{pot}}(\mathbf{r})}}{\int \mathrm{d}\mathbf{r}\,e^{-\beta\mathcal{H}_{\mathrm{pot}}(\mathbf{r})}}}_{\langle\mathcal{A}\rangle}\underbrace{\frac{\int \mathrm{d}\mathbf{r}\,\left(-\mathcal{H}_{\mathrm{pot}}(\mathbf{r})\right)e^{-\beta\mathcal{H}_{\mathrm{pot}}(\mathbf{r})}}{\int \mathrm{d}\mathbf{r}\,e^{-\beta\mathcal{H}_{\mathrm{pot}}(\mathbf{r})}}}_{\langle -E_{\mathrm{pot}}\rangle}\right]
$$

$$
= k_B\beta^2\left(\langle\mathcal{A}E_{\mathrm{pot}}\rangle - \langle\mathcal{A}\rangle\langle E_{\mathrm{pot}}\rangle\right) \ . \tag{3.28}
$$

However, for the special case of the heat capacity $C_V = \partial\langle E\rangle/\partial T$, the considered fluctuation quantity is the whole energy: $E(\mathbf{r},\mathbf{p}) = E_{\mathrm{pot}}(\mathbf{r}) + E_{\mathrm{kin}}(\mathbf{p})$. Due to this momentum dependence it does *not* fulfil the criterion stated above and (3.28) does not hold:

$$
\begin{aligned}
C_V &= \frac{\partial}{\partial T}\langle E\rangle = \frac{\partial\langle E_{\mathrm{kin}}\rangle}{\partial T} + \frac{\partial\langle E_{\mathrm{pot}}\rangle}{\partial T}\\
&= -k_B\beta^2\frac{\partial}{\partial\beta}\frac{\int \mathrm{d}\mathbf{r}\,e^{-\beta\mathcal{H}_{\mathrm{pot}}(\mathbf{r})}\int \mathrm{d}\mathbf{p}\,\mathcal{H}_{\mathrm{kin}}(\mathbf{p})e^{-\beta\mathcal{H}_{\mathrm{kin}}(\mathbf{p})}}{\int \mathrm{d}\mathbf{r}\,e^{-\beta\mathcal{H}_{\mathrm{pot}}(\mathbf{r})}\int \mathrm{d}\mathbf{p}\,e^{-\beta\mathcal{H}_{\mathrm{kin}}(\mathbf{p})}} + \frac{\partial\langle E_{\mathrm{pot}}\rangle}{\partial T}\\
&= -k_B\beta^2\frac{\partial}{\partial\beta}\frac{\int \mathrm{d}\mathbf{p}\,\mathcal{H}_{\mathrm{kin}}(\mathbf{p})e^{-\beta\mathcal{H}_{\mathrm{kin}}(\mathbf{p})}}{\int \mathrm{d}\mathbf{p}\,e^{-\beta\mathcal{H}_{\mathrm{kin}}(\mathbf{p})}} + \frac{\partial\langle E_{\mathrm{pot}}\rangle}{\partial T}\\
&= -k_B\beta^2\left[\underbrace{\frac{\int \mathrm{d}\mathbf{p}\,\mathcal{H}_{\mathrm{kin}}\left(-\mathcal{H}_{\mathrm{kin}}(\mathbf{p})\right)e^{-\beta\mathcal{H}_{\mathrm{kin}}(\mathbf{p})}}{\int \mathrm{d}\mathbf{p}\,e^{-\beta\mathcal{H}_{\mathrm{kin}}(\mathbf{p})}}}_{-\langle E_{\mathrm{kin}}^2\rangle}\right.\\
&\qquad\left. - \underbrace{\frac{\int \mathrm{d}\mathbf{p}\,\mathcal{H}_{\mathrm{kin}}e^{-\beta\mathcal{H}_{\mathrm{kin}}(\mathbf{p})}}{\int \mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{p}\,e^{-\beta\mathcal{H}_{\mathrm{kin}}(\mathbf{p})}}}_{\langle E_{\mathrm{kin}}\rangle}\underbrace{\frac{\int \mathrm{d}\mathbf{p}\,\left(-\mathcal{H}_{\mathrm{kin}}(\mathbf{p})\right)e^{-\beta\mathcal{H}_{\mathrm{kin}}(\mathbf{p})}}{\int \mathrm{d}\mathbf{p}\,e^{-\beta\mathcal{H}_{\mathrm{kin}}(\mathbf{p})}}}_{\langle -E_{\mathrm{kin}}\rangle}\right] + \frac{\partial\langle E_{\mathrm{pot}}\rangle}{\partial T}\\
&= k_B\beta^2\left(\langle E_{\mathrm{kin}}^2\rangle - \langle E_{\mathrm{kin}}\rangle^2\right) + \frac{\partial\langle E_{\mathrm{pot}}\rangle}{\partial T} \ .
\end{aligned} \tag{3.29}
$$

Since the kinetic energy for a system of N particles in d dimensions is always given as $E_{\mathrm{kin}} = \sum_{i=1}^{N}\sum_{j=1}^{d} p_{ij}^2/2m_i$, it is possible to generally integrate the kinetic energy part. For brevity it shall be written as $E_{\mathrm{kin}} = \sum_{i=1}^{dN} p_i^2/2\mu_i$, where $\mu_i = m_{(i-i \bmod d)/d}$:

$$
\begin{aligned}
\langle E_{\mathrm{kin}}\rangle &= \frac{1}{Z_{\mathrm{kin}}}\int\limits_{-\infty}^{\infty}\mathrm{d}p^{dN}\left(\sum_{k=1}^{dN}\frac{p_k^2}{2\mu_k}\right)\exp\left[-\beta\sum_{i=1}^{dN}\frac{p_i^2}{2\mu_i}\right]\\
&= \sum_{k=1}^{dN}\frac{\int_{-\infty}^{\infty}\mathrm{d}p^{dN}\,\frac{p_k^2}{2\mu_k}\exp\left[-\beta\sum_{i=1}^{dN}\frac{p_i^2}{2\mu_i}\right]}{\int_{-\infty}^{\infty}\mathrm{d}p^{dN}\,\exp\left[-\beta\sum_{i=1}^{dN}\frac{p_i^2}{2\mu_i}\right]} = \sum_{k=1}^{dN}\underbrace{\frac{\int_{-\infty}^{\infty}\mathrm{d}p_k\,\frac{p_k^2}{2\mu_k}\exp\left[-\beta\frac{p_k^2}{2\mu_k}\right]}{\int_{-\infty}^{\infty}\mathrm{d}p_k\,\exp\left[-\beta\frac{p_k^2}{2\mu_k}\right]}}_{\frac{1}{2\beta}} = \frac{dN}{2\beta} \ ,
\end{aligned} \tag{3.30}
$$

$$
\begin{aligned}
\langle E_{\mathrm{kin}}^2\rangle &= \frac{1}{Z_{\mathrm{kin}}}\int\limits_{-\infty}^{\infty}\mathrm{d}p^{dN}\left(\sum_{k=1}^{dN}\frac{p_k^2}{2\mu_k}\right)\left(\sum_{l=1}^{dN}\frac{p_l^2}{2\mu_l}\right)\exp\left[-\beta\sum_{i=1}^{dN}\frac{p_i^2}{2\mu_i}\right]\\
&= \sum_{k=1}^{dN}\sum_{l=1}^{dN}\frac{\int_{-\infty}^{\infty}\mathrm{d}p^{dN}\,\frac{p_k^2}{2\mu_k}\frac{p_l^2}{2\mu_l}\exp\left[-\beta\sum_{i=1}^{dN}\frac{p_i^2}{2\mu_i}\right]}{\int_{-\infty}^{\infty}\mathrm{d}p^{dN}\,\exp\left[-\beta\sum_{i=1}^{dN}\frac{p_i^2}{2\mu_i}\right]}
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{k=1}^{dN}\sum_{l=1}^{dN}\Bigg(\delta_{k,l}\underbrace{\frac{\int\limits_{-\infty}^{\infty}\mathrm{d}p_k\,\frac{p_k^4}{(2\mu_k)^2}\exp\left[-\beta\frac{p_k^2}{2\mu_k}\right]}{\int\limits_{-\infty}^{\infty}\mathrm{d}p_k\,\exp\left[-\beta\frac{p_k^2}{2\mu_k}\right]}}_{\frac{3}{4\beta^2}} \\
&\quad + (1-\delta_{k,l})\underbrace{\frac{\int\limits_{-\infty}^{\infty}\mathrm{d}p_k\,\frac{p_k^2}{2\mu_k}\exp\left[-\beta\frac{p_k^2}{2\mu_k}\right]}{\int\limits_{-\infty}^{\infty}\mathrm{d}p_k\,\exp\left[-\beta\frac{p_k^2}{2\mu_k}\right]}}_{\frac{1}{2\beta}}\cdot\underbrace{\frac{\int\limits_{-\infty}^{\infty}\mathrm{d}p_l\,\frac{p_l^2}{2\mu_l}\exp\left[-\beta\frac{p_l^2}{2\mu_l}\right]}{\int\limits_{-\infty}^{\infty}\mathrm{d}p_l\,\exp\left[-\beta\frac{p_l^2}{2\mu_l}\right]}}_{\frac{1}{2\beta}}\Bigg) \\
&= dN\left(\frac{3}{4\beta^2}-\frac{1}{4\beta^2}\right)+(dN)^2\frac{1}{4\beta^2}=\frac{dN}{2\beta^2}+\left(\frac{dN}{2\beta}\right)^2 \; . \qquad (3.31)
\end{aligned}
$$

With (3.30) and (3.31) C_V can now be evaluated:

$$
\begin{aligned}
C_V &= k_B\beta^2\left(\frac{dN}{2\beta^2}+\left(\frac{dN}{2\beta}\right)^2-\left(\frac{dN}{2\beta}\right)^2\right)+k_B\beta^2\left(\langle E_{\mathrm{pot}}^2\rangle-\langle E_{\mathrm{pot}}\rangle^2\right) \\
&= k_B\frac{dN}{2}+k_B\beta^2\left(\langle E_{\mathrm{pot}}^2\rangle-\langle E_{\mathrm{pot}}\rangle^2\right) \; . \qquad (3.32)
\end{aligned}
$$

3.3 Quantities

For the observation of the considered protein models, several thermodynamic and structural quantities are measured. These shall be briefly summarised within this section. In the following chapters, most of the time a certain normalisation (e.g. to the number of atoms N or the number of bonds $N-1$) is used for most of the quantities. The particularly chosen normalisations are also given here.

3.3.1 Energy and Specific Heat

The potential energy E_{pot} is elementary for any simulation. It is responsible for the specific behaviour of a system. The description of a considered model always means to introduce the expression for evaluating the potential energy of a certain configuration of the system. This has been done in chapter 1.

Talking of the *energy* E in a Monte Carlo simulation always means the internal (potential) energy E_{pot}, since the kinetic energy is already integrated out by the algorithm. Within the Molecular Dynamics simulation, there is an explicit kinetic energy E_{kin} besides the potential energy. Furthermore, when applying the Nosé-Hoover-Chain (NHC) thermostat, the virtual heat bath particles also have a potential and kinetic energy. The sum of all these four energies (2.71) is theoretically kept constant over the simulation.

However, as explained in the previous section, the specific heat of a MC simulation can be compared to the fluctuation of the potential energy in a MD simulation. Therefore, within this work heat capacities from MD simulations *always* mean the contribution of the potential energy according to the second summand of (3.32). The fixed part $k_B dN/2$ arising from the kinetic energy is not taken into account. One good reason to do so is that in a finite simulation the average will never match the exact value and the variance will not vanish. Thus, considering the total energy of the stand-alone system for measuring the specific heat would only increase the statistical errors.

In the following, the values of energy and specific heat are normalised to the number of monomers N if not denoted otherwise. To consider the energy per monomer makes sense as far as most of the potential contributions can be linked to single monomers.

3.3.2 End-To-End-Distance and Radius of Gyration

For protein folding, structural aspects of a certain configuration are very important. Therefore, it is necessary to define suitable quantities to describe the structure.

A straightforward approach to describe a configuration is to measure the distance of the first and the last monomer – the end-to-end-distance:

$$R_{\mathrm{ee}}(\mathbf{X}) = |\mathbf{r}_N - \mathbf{r}_1| \ . \tag{3.33}$$

It is expected that for rather linear or "unfolded" structures R_{ee} will have explicitly larger values than for more compact ones, that could be called "folded".

For a chain with N monomers in an equal distance r_0 the largest end-to-end-distance will be $(N-1) \cdot r_0$. Thus, the obvious normalisation is the number of *bonds* $N-1$, since the equilibrium bond length is set to $r_0 = 1$. If the end-to-end-distance is normalised by the number of bonds, it is possible to compare the findings of sequences with unequal length.

Although the evaluation of R_{ee} is very simple, the information content is questionable. If, for example, a long chain has a sharp bend a monomer in the middle and the two tails are nearly parallel and very close to each other, the end-to-end-distance could be very small, while the structure is actually quite wide-stretched.

Therefore, a more sophisticated quantity shall be used to describe the structural properties of a configuration. The radius of gyration is defined as the average square distance from the centre of mass:

$$R_{\mathrm{gyr}}(\mathbf{X}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (\mathbf{r}_i - \mathbf{r}_S)^2} \ , \qquad \mathbf{r}_S = \frac{1}{N}\sum_{i=1}^{N} \mathbf{r}_i \ . \tag{3.34}$$

This value will be still big, if e.g. a large loop is evaluated as described above. Only if a configuration is really dense, R_{gyr} will have small values, since then most of the monomers are close to the centre of mass.

The normalisation is not as obvious as for R_{ee}. To normalise by the maximal value of R_{gyr} for a given number of monomers N implies a rather complicated expression and it is not clear, what the real differences of R_{gyr} are for different N. Therefore, in the following R_{gyr} will be simply normalised by the number of monomers N.

3.3.3 Comparing Two Configurations

Proteins live at room temperature in nature and thus have to cope with rather large thermal energies. Still, a protein has to exist in a relatively stable state at these temperatures, since the function is mainly induced by its three-dimensional structure. Therefore, in the complicated and rugged free-energy landscape of a protein there must be some steep and deep valley, so that it is possible to surely stay within this valley at room temperature. The whole thermal energy has to be accumulated by fluctuations which are not crucial for the biological behaviour of the protein.

Therefore, in a realistic protein model the hope is to also find minimal energy states which can be explicitly distinguished from the rest of the ensemble of configurations not only by especially low energies, but also by means of the structural uniqueness. Once such a potential ground-state has been found, the terms "folded" and "unfolded" can be identified by comparing with this ground-state structure $\mathbf{X}^{(0)}$. Several ways of accomplishing the comparison are described in the following.

The methods can of course be used to compare any two configurations, no matter how they have been obtained. For example it is also possible to compare two ground-state structures that have been found for different simulation parameters.

q Parameter

As described in chapter 1, by the Lennard-Jones-like interaction it is induced that A monomers have a strong attraction. Indeed, the only type of energy that is negative, are the attractive A-A and B-B interactions from the Lennard-Jones potential. Therefore, it is expected that a lot of A monomer pairs will have a distance close to that, where the Lennard-Jones potential has its minimum ($r_{ij} \approx 1.1$ for the given parameters). Such a pair of non-neighbouring nearby monomers is called a *contact*. In a ground-state structure with very low energy, there will be a high number of such contacts. Since the ground-state is considered to be the native configuration $\mathbf{X}^{(0)}$ of the system as explained above, a contact within the ground-state structure is called *native contact*.

By counting, how many of the native contacts are formed in an instantaneous structure $\mathbf{X}$, it can be compared to the ground state:

$$q(\mathbf{X}, \mathbf{X}^{(0)}) = \frac{\#\text{ contacts formed in } \mathbf{X}}{\#\text{ contacts in } \mathbf{X}^{(0)}} \; . \tag{3.35}$$

The implementation leaves some choices. The first is the exact definition of what a contact is. Mostly the following criteria are used in this work: (a) $r_{ij} < r_{0,\text{cont}} = 1.7$, (b) $|i-j| > 2$. The latter fact is motivated by the fact that even next but one neighbours get quickly in contact even for smaller choices of $r_{0,\text{cont}}$. This can yield relatively high values for q although two configurations are really disparate. However, there is some space for adjustment.

The second question is: When is a contact considered formed? The rule applied in this work is determined following [39]. A contact is formed, if the instantaneous distance r_{ij} of two monomers which are in contact in the ground-state configuration $\mathbf{X}^{(0)}$ is smaller than $\gamma \cdot r_{ij}^{(0)}$. According to [39] the cut-off factor γ does not have a crucial influence on the results and is chosen $\gamma = 1.2$.

The range of q is $q \in [0,1]$. If $q = 1$, the $\mathbf{X}$ is very similar to $\mathbf{X}^{(0)}$. Analogously, if $q = 0$, none of the native contacts is formed and the configuration is far from the ground state structure. A drawback of the q parameter is that it is highly discrete. A typical number of native contacts in a ground-state of a sequence with $N = 20$ monomers is $25 \ldots 35$. Also, since the presented criterion for considering a contact as formed is quite imprecise, the variety of configurations that lead to the same value for q is enormous.

Overlap Parameter

The two drawbacks brought up in the last paragraph are dissolved with the introduction of the overlap parameter in Ref. [17]:

$$\begin{aligned} Q(\mathbf{X},\mathbf{X}^{(0)}) &= 1 - \frac{d(\mathbf{X},\mathbf{X}^{(0)})}{N_t + N_b} \ , \qquad (3.36) \\ d(\mathbf{X},\mathbf{X}^{(0)}) &= \frac{1}{\pi}\left(\sum_{i=1}^{N_t} d_t(\Phi_i,\Phi_i^{(0)}) + \sum_{i=1}^{N_b} d_t(\Theta_i,\Theta_i^{(0)})\right) \ , \\ d_t(\Phi_i,\Phi_i^{(0)}) &= \min\left(|\Phi_i - \Phi_i^{(0)}|, 2\pi - |\Phi_i - \Phi_i^{(0)}|\right) \ , \\ d_b(\Theta_i,\Theta_i^{(0)}) &= |\Theta_i - \Theta_i^{(0)}| \ . \end{aligned}$$

In (3.36), N_t is the number of torsion angles (see chapter 5) and N_b is the number of bond angles. With all torsion angles Φ_i and bond angles Θ_i, a configuration is uniquely defined. Therefore, comparing the deviation of all the angles as it is done when evaluating $Q(\mathbf{X},\mathbf{X}^{(0)})$ gives thus a definite statement about the equality of two configurations. If $Q = 1$, i.e. if all angles are the same and therefore d vanishes, the two configurations are exactly the same. For lower values of Q the match is less well. However, values lower than $Q = 0.5$ are hardly observed. The reason is that e.g. if the difference of two bond angles should be $|\Theta_i - \Theta_i^{(0)}| = \pi$, which would be the maximum contribution to $d(\mathbf{X},\mathbf{X}^{(0)})$, one of the bond angles would have to be 0 and the other one π. While, $\Theta = 0$ is unlikely, $\Theta = \pi$ is virtually impossible, since the strong Lennard-Jones repulsion for small distances excludes configurations, where two monomers are close.

There are two special cases, which have to be considered when calculating Q. Firstly, one configuration can be a mirrored image of the other one if $\Phi_i = -\Phi_i^{(0)}$ holds for all torsion angles. Furthermore, in the special case of a symmetric sequence, it can of course happen that the configurations are nearly equal, but running in opposite directions. Hence, also a sequence reversal has to be calculated in case of symmetric sequences by comparing Φ_{N+1-i} and $\Phi_i^{(0)}$ as well as Θ_{N+1-i} and $\Theta_i^{(0)}$. The maximum of all Q resulting from the four possible combinations of these two assumptions is the overall overlap of the two sequences.

In principle, Q gives a reasonable information over the equality of two configurations for $Q \approx 1$. For lower values, it is not clear where the deviation comes from. On the one hand, it would be possible that all angles are altered a little bit but the configurations are still very similar. On the other hand, it is as likely that only a small number of angles is completely different from the original structure, which can lead to an explicitly different overall configuration.

Root Mean Square Deviation

Finally, a very useful but extremely costly method of comparing two configurations is to calculate the root mean square deviation:

$$D_{\mathrm{rms}} = \min\sqrt{\frac{1}{N}\sum_{i}^{N}\left(\tilde{\mathbf{r}}_i - \tilde{\mathbf{r}}_i^{(0)}\right)^2}\ , \tag{3.37}$$

$$\tilde{\mathbf{r}}_i^{\{(0)\}} = \mathbf{r}_i^{\{(0)\}} - \mathbf{r}_S^{\{(0)\}} = \mathbf{r}_i^{\{(0)\}} - \frac{1}{N}\sum_{i=1}^{N}\mathbf{r}_i^{\{(0)\}}\ .$$

Actually, the structures are shifted to the same position by considering the positions of the monomers relative to the centre of mass $\mathbf{r}_S$. The critical point is finding the minimum of the root mean square deviation, since the alignment of the structures in terms of rotation is free. This implies a complicated calculation and thus cannot be done for every step within the simulation. If the structures are identical, than $D_{\mathrm{rms}} = 0$. For unequal structures of length 20, values of $D_{\mathrm{rms}} \approx 1.0$ are typical.

Chapter 4

Examination of the Model System

One of the key points of the present project is to study possible differences in the behaviour and the results of Monte Carlo and Molecular Dynamics simulations. The systems described in the first chapter have been studied in detail with several methods of Monte Carlo within the group. Thus, there are many reliable data for the original model available, where the bonds are fixed to unit length. However, flexible bonds have never been examined before. Therefore, it is necessary to create data for this particular case both with MC and MD. The results of the MC studies are presented in this section.

4.1 Thermodynamics with Monte Carlo simulations

4.1.1 General Facts about the Simulations

MC Update

For testing the MC program and comparing the flexible bonds case with the original model, both alternatives were implemented. In contrast to Molecular Dynamics, constraints can be easily implemented in a Monte Carlo simulation by choosing a suitable update that follows the requirements. One method of updating the positions of monomers by keeping all next-neighbour distances constant is to rotate one bond vector. Afterwards, the positions of the monomers can be calculated from successively summing up all bond vectors, starting from the very first monomer.

Figure 4.1 shows such an update scheme, the monomers are numbered. The positions of the monomers at the beginning of the chain stay unchanged, until monomer number i. Since $\mathbf{b}_i$ is rotated, all succeeding monomers are moved by the difference of the new and original bond vector $\mathbf{b}_i$. The bond vector is rotated by an angle θ with respect to its original direction, which leaves an azimuthal degree of freedom φ, which can be randomly selected out of $\varphi \in [0, 2\pi)$. The proposed update is done in spherical coordinates, which means the Jacobian of spherical coordinates $r^2 \sin\theta$ has to be taken into account. The radius r means the bond length in this case, which is fixed to unit length. Thus the only relevant term is $\sin\theta$. Choosing $\cos\theta$ instead of θ means that not explicitly spherical coordinates are used, but r, φ and $\cos\theta$. This choice makes the Jacobian r^2, so no particular distribution for selecting $\cos\theta$ has to be followed. Thus, it is possible to simply select $\cos\theta \in [\cos_0, 1.0)$, where $\cos_0$ is a freely selectable lower bound. The smaller $\cos_0$ is chosen, the wider is the possible update cone, from which the new bond vector can be chosen. This value should be selected carefully

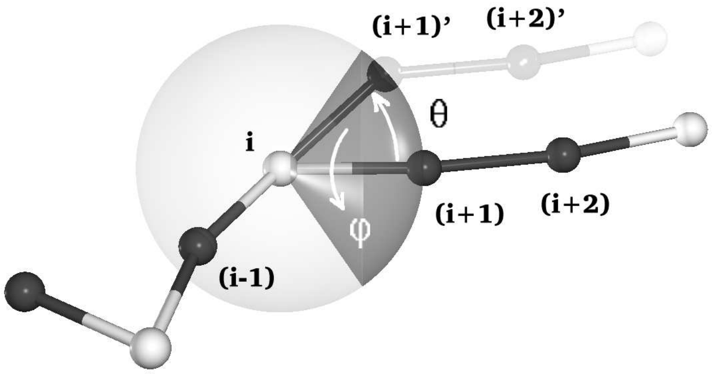


Figure 4.1: Visualisation of the utilised update scheme for simulating the AB model with fixed bonds. The picture is taken from Ref. [17].

for obtaining reasonable acceptance rates. If $\cos_0$ is too small, the suggested updates will be rejected most of the time, since the configurational change is too large. If $\cos_0$ is too large, the movement in phase space is very slow and the acceptance rate is near to 1, since the suggested update does rather change anything about the current configuration.

For the case of flexible bond, things are simpler. Here it is possible to suggest a small cartesian movement of one or more monomers as and update. Indeed, it is reasonable to chose between two different types of cartesian updates: a local one, where the position of only one monomer is changed, and a global one, analogous to the previously described case, where one *bond vector* is altered, and thus the whole succeeding chain is moved. The first case has an advantage considering the implementation: Since only one monomer is moved, only potential terms which imply this particular monomer must be recalculated to obtain the energy of the new configuration. This is especially crucial for the Lennard-Jones potential, which has a complexity $\mathcal{O}(N^2)$. By using a local update scheme, the complexity can be reduced to $\mathcal{O}(N)$. However, it will turn out that the autocorrelation time for the local update is significantly larger than for the global update. Therefore the gain in computation time is lost by the need of more statistics. For that reason and also because in the fixed bond case also a global update is used, the local update scheme is rejected.

Specific Set-Up for Parallel Tempering

As explained earlier, a parallel tempering simulation uses several replica of a system, which are simulated coinstantaneously at different temperatures. The probability that two replica at neighbouring temperatures are exchanged depends on the difference of the inverse temperatures, as well as the current energies. Therefore, for having a suitable acceptance rate of such exchanges, the ensemble of temperatures should feature an acceptable overlap of the energy histograms. This means that more replica should be placed at low temperatures in principle, where the energy histograms are sharp. One way to reach this would be to chose temperatures by hand. Both in [12] and in [40] rather sophisticated methods for the choice of the simulation temperatures are described. However, the effort is questionable and the results were not satisfying in some test simulations. Yet, to stay flexible in terms of temperature range $(T_1 \dots T_2)$ and number of replica N_{R}, a general formula for distributing the temperature

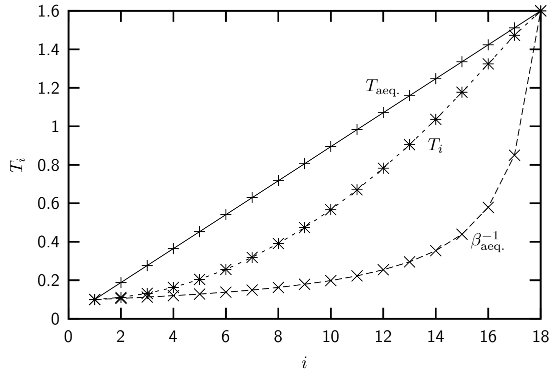


Figure 4.2: Several possibilities of distributing the temperature range from $T_1 = 0.1$ to $T_2 = 1.6$ over $N_{\rm R} = 18$ processors are plotted as defined in (4.1) – (4.3).

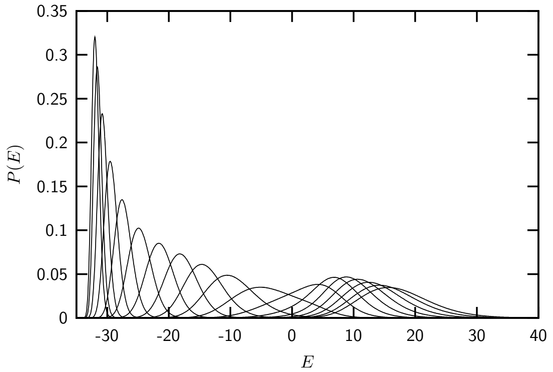


Figure 4.3: Normalised potential energy histograms for the finally chosen scheme of temperature distribution.

range over the different replica i is developed. Using a linear interpolation between the two bordering temperatures,

$$T_{\text{aeq.},i} = T_1 + (T_2 - T_1)\frac{i}{N_{\rm R} - 1} \ , \tag{4.1}$$

would surely imply problems in the low temperature range, because there would be too few replica. However, a linear interpolation between the inverse temperatures (implying $k_B \equiv 1$),

$$\beta_{\text{aeq.},i} = T_1^{-1} + (T_2^{-1} - T_1^{-1})\frac{i}{N_{\rm R} - 1} \ , \tag{4.2}$$

would stress the low temperature range far too much. Therefore, a quadratic interpolation is chosen, which is close to $\beta_{\text{aeq.},i}$ for small i, and adapts more to $T_{\text{aeq.,i}}$ for larger i:

$$\begin{aligned} T_i &= \beta^{-1}_{\text{aeq.},i} + (T_{\text{aeq.},i} - \beta^{-1}_{\text{aeq.},i})\frac{i}{N_{\rm R} - 1} \\ &= \frac{(N_{\rm R}-1)^2(iT_1 + (N_{\rm R} - i - 1)T_2)}{T_1T_2(N_{\rm R}-1)^3 + i^2(N_{\rm R}-1)(T_1-T_2)^2 - i^3(T_1-T_2)^3} \ . \end{aligned} \tag{4.3}$$

Figure 4.2 compares the temperature “functions” for the parameters that are used in *all* simulations, if not specified differently. I.e., 18 replica ($N_{\rm R} = 18$) cover a temperature range from $T_1 = 0.1$ to $T_2 = 1.6$. T_2 is considered high enough to quickly sample the rough energy landscapes of the observed systems and circumvent large potential energy barriers. Whereas, T_1 seems to be low enough to thoroughly sample the valleys of the energy landscape to find low-energy ground-states. The number of processors is large enough to have a suitable potential energy histogram overlap (see Fig. 4.3). Additionally, the variance of the replica number that is currently simulated at a particular temperature is accumulated for each temperature. This makes it possible to access how often each replica visits the lowest and highest temperature. The outcome was always satisfactory.

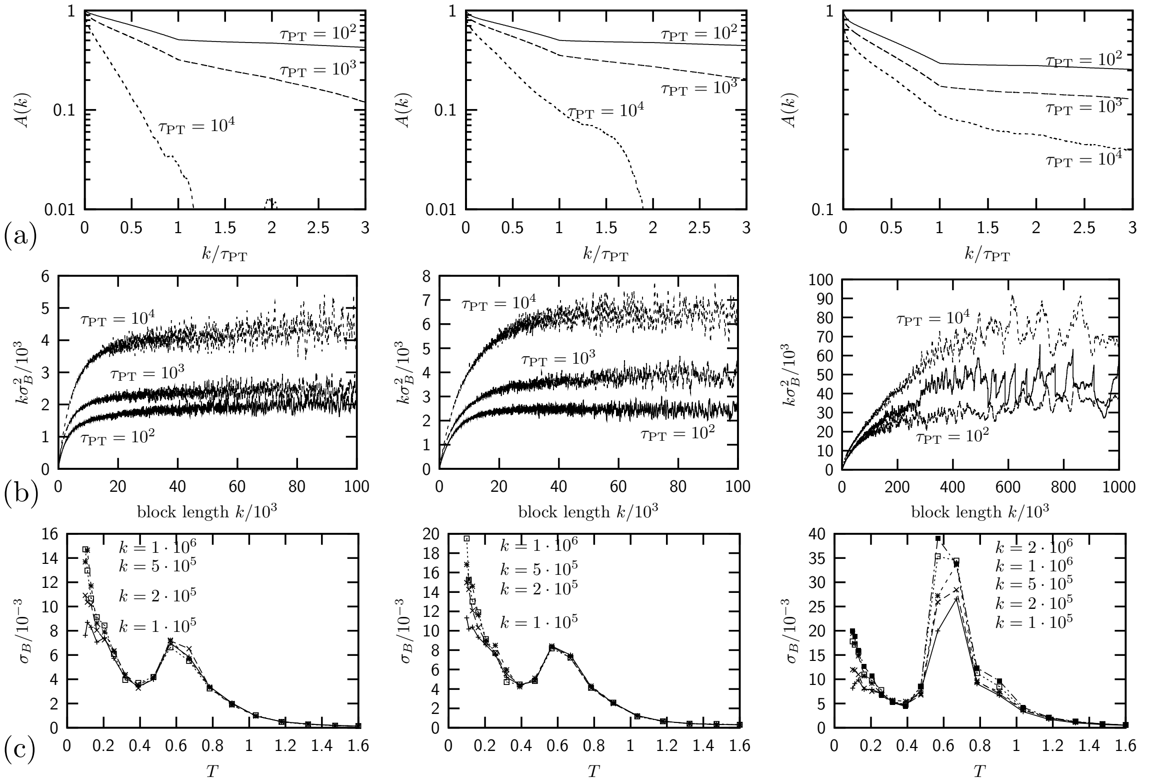


Figure 4.4: From the left to the right, the different update types are visualised: The update with fixed bond length (compare Fig. 4.1), the global cartesian update (also one bond vector is altered) and the local cartesian update, where only one monomer is moved.
(a) In the first row a logarithmic plot of the autocorrelation function $A(k)$ is shown. The x-coordinate is normalised by τ_{PT}. For the largest choice of τ_{PT}, the noise is too strong to have confident results for $k > \tau_{\mathrm{PT}}$. (b) Beneath, the product of variance and Jackknife block size is plotted over the Jackknife block size for the respective measurements (see text for details). In [34] it is proven that a plateau shows up at $k\sigma_B^2 = 2\tau_{\mathrm{int}}$. Therefore these plots can be used to estimate τ_{int}. (c) In the bottom row, the estimated errors of a simulation of sequence 20.4 are plotted for several Jackknife block sizes.

Autocorrelation Functions

The study of potential energy autocorrelation functions is important for two major reasons. First, for the whole error analysis machinery it is crucial to have a rough estimation of the autocorrelation time. Secondly, the simulation should be capable of running on parallel systems with distributed memory, using MPI. Therefore, it is to expect that the more steps τ_{PT} are made between two trial replica exchanges, the more efficient would be the parallelisation. On the other hand, the efficiency of the parallel tempering algorithm is increased by frequently proposing such exchanges, i.e. for small τ_{PT}. The influence of τ_{PT} on the autocorrelation times should be an important argument for the choice of a reasonable medium value for τ_{PT}.

Since the autocorrelation time is always larger at low and especially at critical temperatures, the following simulation is chosen to compare the three different MC update types: 10 replica of sequence 20.6 are simulated on a temperature range from $T_1 = 0.35$ to $T_2 = 1.0$. Thus, the lowest temperature is lying at a strong peak of the specific heat (compare Fig.

4.5). After 10^5 MC sweeps of equilibration, the potential energy time series at the lowest temperature is measured over 10^7 sweeps. After $\tau_{\mathrm{PT}} = \{10^2, 10^3, 10^4\}$ steps an exchange of the replica is suggested. For the previously described update with fixed bond length (see Fig. 4.1), $\cos\theta$ is chosen from $\cos\theta \in [0.99, 1.0)$, thus $\cos_0 = 0.99$. For the two types of cartesian updates, a random displacement is chosen out of a cubic box with edge length 0.1.

The results of the measurements are shown in Fig. 4.4. The first row (a) shows the autocorrelation functions for the three update types and for several choices of τ_{PT}. The exponential autocorrelation times are listed in Table 4.I. It is very remarkable that the autocorrelation time seems to rise by a factor $2 \ldots 10$ for time separations $k > \tau_{\mathrm{PT}}$! For showing this effect more clearly, the abscissa is normalised to τ_{PT}. At $k/\tau_{\mathrm{PT}} = 1$ is a definite sharp bend, and the exponential autocorrelation time changes its value noticeably (compare Table 4.I). However, the normalisation of the abscissa leads to the impression that the autocorrelation time is smaller for large τ_{PT} which would be the contrary of the expectation. But this is *not* the case, it is only an effect of the chosen normalisation of the x-axis. The speculation about the form of $A(k)$ would have been that only for $k > \tau_{\mathrm{PT}}$ the advantage of parallel tempering is really seen and thus the autocorrelation time is much *smaller* than for $k < \tau_{\mathrm{PT}}$. Obviously, the opposite is the case. The reason for this astonishing effect is maybe not trivial, but could result from the fact that for small k the system samples only one valley of the potential energy landscape, and only for $k > \tau_{\mathrm{PT}}$ the real roughness of the landscape is noticed. Since detailed parallel tempering studies are not the main goal of the project, no further effort is made to investigate this effect in more detail.

From the pictures in the second row of Fig. 4.4 an assumption about τ_{int} can be made. Obviously the difference between $\tau_{\mathrm{PT}} = 10^2$ and $\tau_{\mathrm{PT}} = 10^3$ is smaller, than between $\tau_{\mathrm{PT}} = 10^3$ and $\tau_{\mathrm{PT}} = 10^4$. Since $\tau_{\mathrm{PT}} = 10^2$ implies 10 times more latency than $\tau_{\mathrm{PT}} = 10^3$ considering the parallelisation, it is chosen to use $\tau_{\mathrm{PT}} = 10^3$ in all the following simulations. Also, for the local cartesian update, it gets clear that the autocorrelation time is about one order of magnitude larger than for the two more “global” update schemes, where whole parts of the polymer are shifted against each other. Unfortunately, the statistical noise gets too strong for large block sizes, so that an estimation of τ_{int} is impossible. Therefore, another couple of simulations is carried out.

Sequence 20.4 is simulated for 10^8 sweeps after an equilibration phase of 10^5 sweeps. The

Table 4.I: Exponential autocorrelation times obtained by fitting the plots from Fig. 4.4.

Update type	Range	$\tau_{\mathrm{PT}} = 10^2$	$\tau_{\mathrm{PT}} = 10^3$	$\tau_{\mathrm{PT}} = 10^4$
b_i =const.	$k < \tau_{\mathrm{PT}}$	$1.59 \cdot 10^2$	$1.02 \cdot 10^3$	$2.80 \cdot 10^3$
global cartesian	$k < \tau_{\mathrm{PT}}$	$1.75 \cdot 10^2$	$1.21 \cdot 10^3$	$4.79 \cdot 10^3$
local cartesian	$k < \tau_{\mathrm{PT}}$	$1.99 \cdot 10^2$	$1.47 \cdot 10^3$	$1.20 \cdot 10^4$
b_i =const.	$k > \tau_{\mathrm{PT}}$	$1.20 \cdot 10^3$	$2.17 \cdot 10^3$	–
global cartesian	$k > \tau_{\mathrm{PT}}$	$1.76 \cdot 10^3$	$3.72 \cdot 10^3$	$1.08 \cdot 10^4$
local cartesian	$k > \tau_{\mathrm{PT}}$	$3.13 \cdot 10^3$	$1.52 \cdot 10^4$	$5.05 \cdot 10^4$

temperature range is chosen as in the following runs from $T_1 = 0.1$ to $T_2 = 1.6$, and $N = 18$ processors are used. The absolute error of the heat capacity obtained by the use of different Jackknife block sizes k is plotted in the last row of Fig. 4.4. The correct choice of k is reached, when larger k do not increase the error substantially anymore. While for low temperatures the differences between the three update types are marginal, it is clear that at the larger peak ($T \approx 0.6$, compare Fig. 4.5) the local update is explicitly worse than the other methods. This is analogous to the general finding that global update schemes lead to seriously shorter autocorrelation times at transition temperatures. Therefore, the global cartesian update is chosen for all runs with flexible bonds, by accepting a slower computation of about a factor 3 for systems with 20 monomers. Since the deviations of the estimated error for Jackknife bin lengths $k = 5 \cdot 10^5$ and $k = 10^6$ sweeps are negligible even for low temperatures, the former selection $k = 5 \cdot 10^5$ is chosen for all following simulations, yielding a better statistics of Jackknife bins.

4.1.2 Thermodynamics

For the six 20mer sequences given in Table 1.I, several thermodynamic quantities are measured. Also, the configurations with the minimal observed energy are written out after every simulation. Since the data should be used for comparison with the Molecular Dynamics results, flexible bonds with a bond strength of $\alpha_r = 50$ are used. All runs have an equilibration phase of 10^5 sweeps, to which a parallel tempering run of $3 \cdot 10^8$ sweeps is affiliated. The Jackknife bin size and the parallel tempering exchange rate τ_{PT} are chosen as described in the previous section.

Specific Heat

The measurements of mean potential energy and specific heat are shown in Fig. 4.5. All values are normalised to the length of the sequence, i.e., divided by 20. First of all, the results match very well with data for fixed bond length in [17], recalling that the heat capacity is raised by about $\approx 1/2$ because of V_{bond}. For a graphical illustration, see Fig. 1.5 above, where the stiff bond case is compared to the flexible bond case. The mean energy 20.1 – 20.4 lies in the same range from $\bar{E}(T = 0.1) \approx -1.5$ to $\bar{E}(T = 1.6) \approx 1.0$. For 20.5 and 20.6, which are also distinctive for the specific heat, the energy range is smaller and starts at about $E = -1.0$ for low temperatures. The specific heat of the observed systems always shows a double peak, except for sequence 20.6, where only a shoulder is left over from the first peak. This indicates two transitions between different configurational states. In [41], the three states are referred to as the *ground-state* domain, the *globule* domain and the *random-coil* domain. In 20.1 – 20.4, the temperatures, where the peak arises, are very similar. For 20.5 and 20.6, there are also similarities. The main difference is the height of the peaks. Therefore, the general behaviour seems to be much more dependent on the fraction of A monomers. This is the main difference between the two sequence groups 20.1 – 20.4 and 20.5 – 20.6, as it can be seen from Table 1.I. Both peaks of the specific heat are at lower temperatures for the latter two sequences in comparison to 20.1 – 20.4. This is analogous to the finding that the energies are larger at low temperatures. Thus, the amount of thermal energy that destabilises the ground-state is lower, and the transitions occur at lower temperatures.

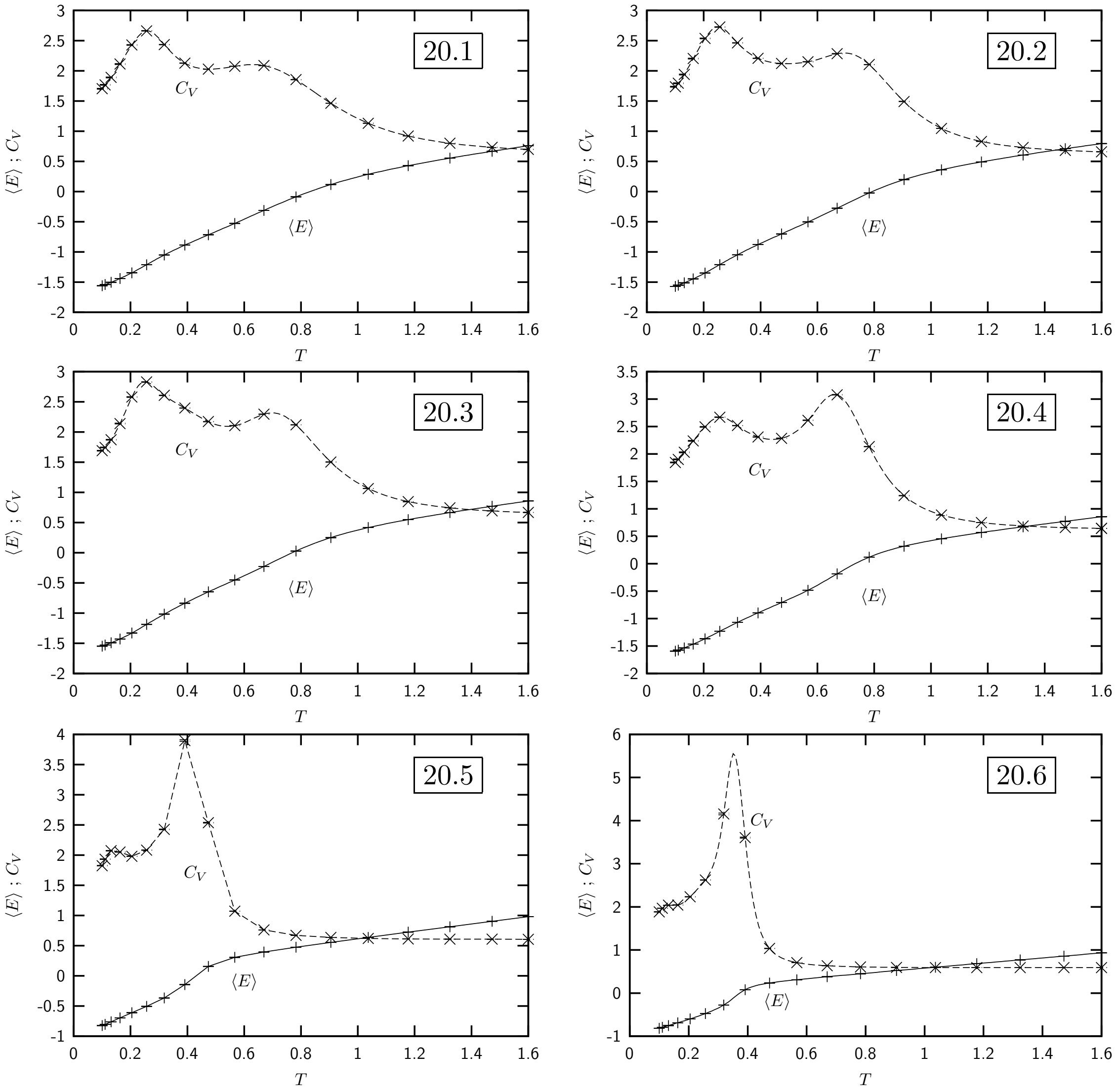


Figure 4.5: Mean potential energy and heat capacity of the six 20mers listed in Table 1.I. The lines are obtained by multiple histogram reweighting (except for 20.5, where the reweighting procedure did not work properly). The error bars result from a Jackknife analysis.

End-To-End-Distance and Radius of Gyration

Since the fluctuations are more interesting, as their extrema indicate temperature ranges, where possibly transitions take place, only the fluctuations of end-to-end-distance and radius of gyration are plotted in Fig. 4.6. In general, the fluctuation of the radius of gyration is not as strong as for the end-to-end-distance. Obviously, the end-to-end-distance is more dependent on the several configurational domains that have already been identified by the specific heat peaks. For the first three sequences, there is also a double peak in both of the fluctuation quantities, although it is not as pronounced as for the specific heat. For 20.4-20.6, actually only the second peak is left, which is much more pronounced in exchange. Maybe this deals with the fact that for these three systems the second peak of the specific heat is

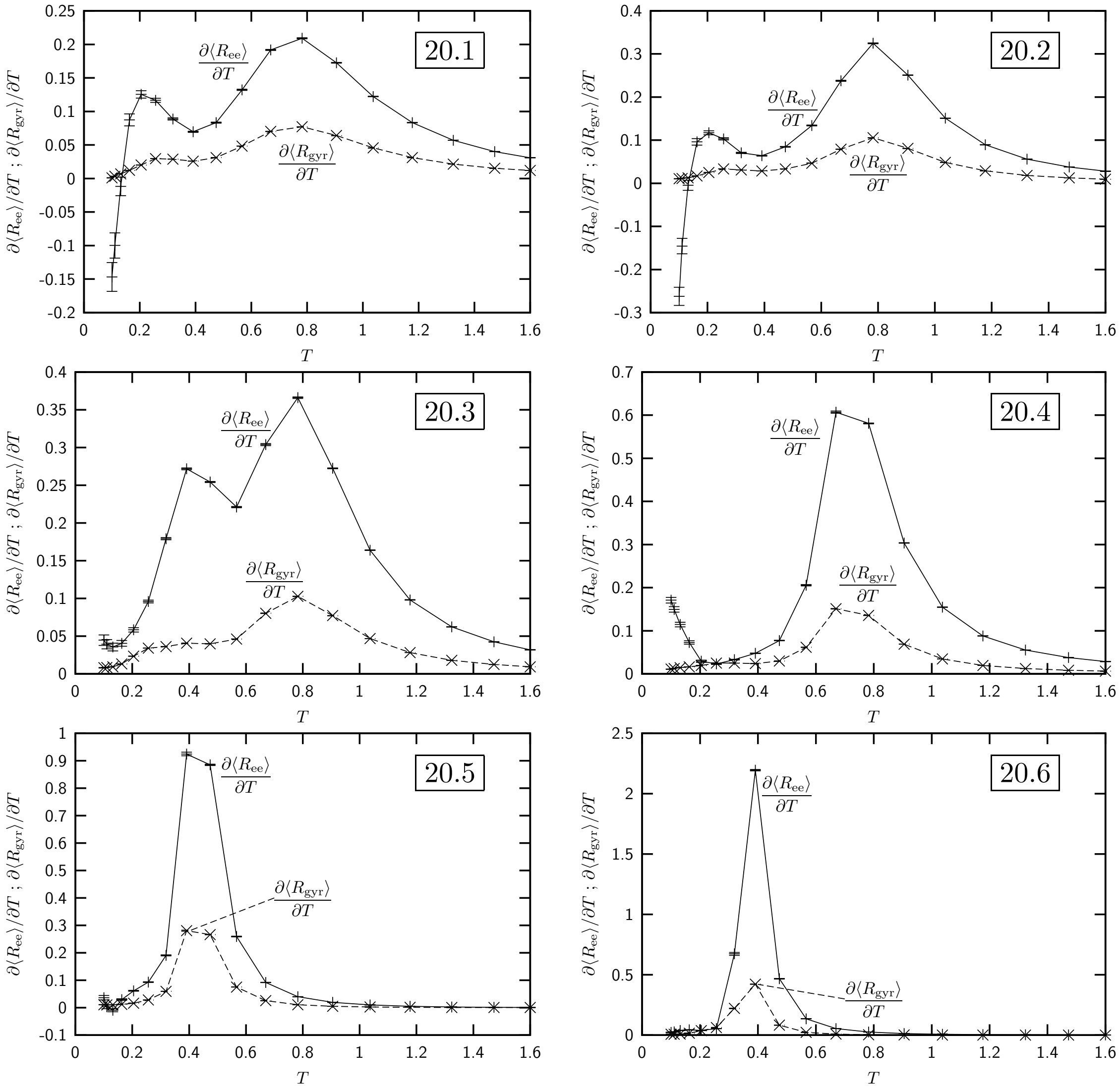


Figure 4.6: Fluctuation of end-to-end-distance radius of gyration of the six 20mers listed in Table 1.I. Unfortunately no reweighting data is available, so the results are obtained by simple averaging.

higher than the first peak. It is remarkable that the temperatures of the peaks of the R_{ee} and R_{gyr} fluctuations correlate very well with those from the specific heat.

Ground-States

In Table 4.II, the minimum energies are listed as they are found during the simulations. It is remarkable that the minimal energies do obviously also rather depend on the particular sequence, but on the number of B monomers (compare Table 1.I). This is reasonable considering the fact that a close A-A interaction gives an negative energy contribution that is four times as low as a B-B interaction. Thus, the more A monomers are present, the more A-A contacts will arise and lead to a seriously lower ground-state energy. This can be also seen from the mean energy plots in Fig. 4.5.

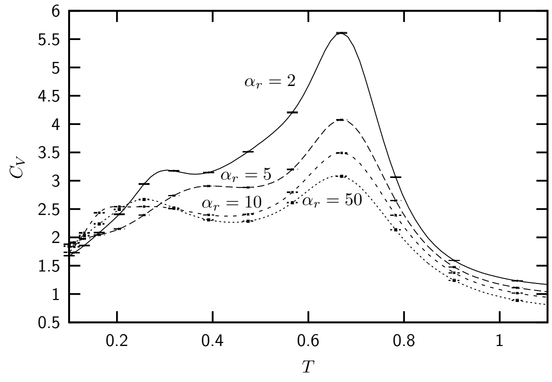


Figure 4.7: Heat capacities of sequence 20.4 for different bond strengths α_r, normalised to the number of monomers. Again, the lines are obtained by multiple histogram reweighting, while the error bars result from a Jackknife analysis of the measured data.

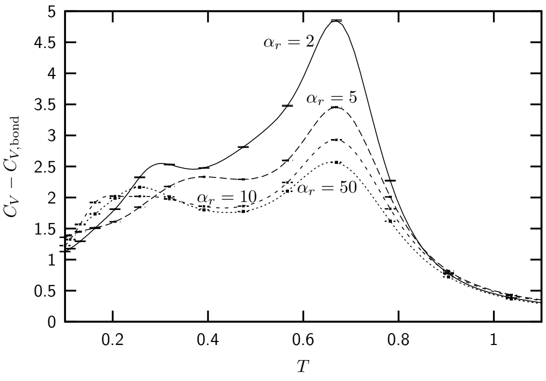


Figure 4.8: To clarify that the observed effects do not qualitatively result from the bond term, the analytic effect of V_{bond} as given in (1.9) is subtracted. The different heights of the tails in Fig. 4.7 are obviously only caused by the larger contributions of $C_{V,\text{bond}}$ for weaker bonds.

4.1.3 Impact of Different Bond Strengths on the System

The question of what the impact of different bond strengths α_r is, is exemplarily examined for sequence 20.4. Figures 4.7 and 4.8 show an overview of the heat capacities of sequence 20.4 for several bond strength α_r. The second peak is continuously lowered by increasing the bond strength. For small α_r, there is a much larger ensemble of accessible configurations in principle, compared to rather strong bonds. The second peak denotes the transition between globule and random-coil configurations. It is easy to see that within the globule domain – the ground-state domain is considered as a special case of the globule domain here – there are strong constraints due to the Lennard-Jones potential. Within the globule domain, even a system with quite weak bonds can thus not take advantage of the wider range of possible configurations. Only after reaching the random-coil domain, the steric repulsions lose importance in favour of the bond fluctuations. Therefore the difference in the density of states within the globule and the random-coil domain is the larger the smaller α_r is. This can be taken as an argument for the higher peak in the heat capacity. Nevertheless it is remarkable that the temperature where the second peak arises seems to be independent from α_r. The fluctuations of end-to-end-distance in Fig. 4.9 and of the radius of gyration in Fig. 4.10 do not show any distinctive feature. The big peak correlates acceptably with the second peak of the heat capacity and is pronounced for weak bonds as well. Also the tails show small deviations,

Table 4.II: Energies of the ground-states observed by the described MC parallel tempering simulations.

Sequence	20.1	20.2	20.3	20.4	20.5	20.6
E_{min}	−33.4	−33.7	−33.1	−34.2	−18.9	−18.6

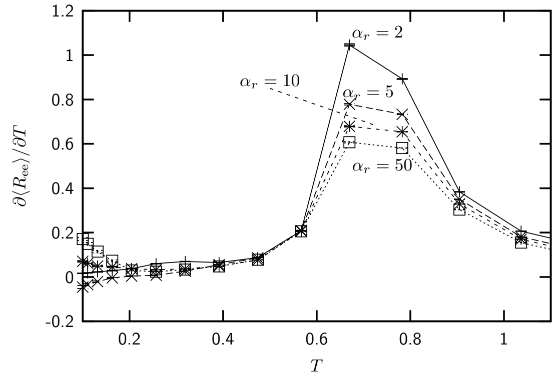


Figure 4.9: Fluctuation of end-to-end-distance for sequence 20.4 for several bond strengths.

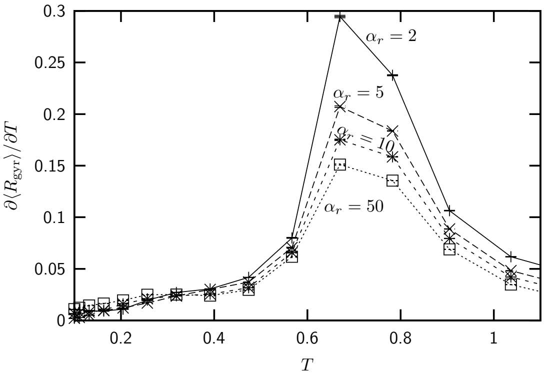


Figure 4.10: Fluctuation of radius of gyration for the same parameters.

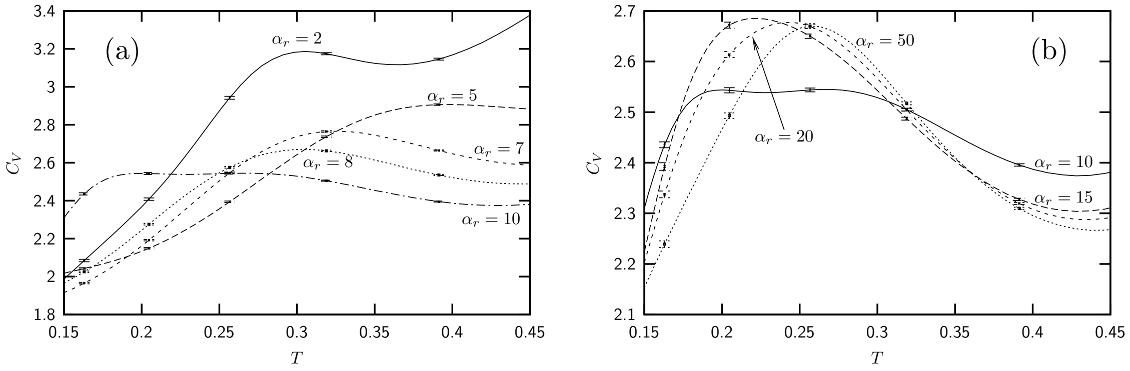


Figure 4.11: Closer look on the first peak of the heat capacities of sequence 20.4 for different bond strengths α_r. (a) Low bond strength, (b) higher bond strength.

which can be probably explained by the larger bond length fluctuations for low α_r.

For the first peak in the heat capacity, things are more difficult. A closer look is provided in Fig. 4.11. When increasing α_r coming from weak bonds (see Fig. 4.11(a)), the height of the first peak decreases. The specific temperature moves first towards higher temperatures and is decreasing again afterwards. For higher values of α_r (Fig. 4.11(b)), the peak gets stronger again and is finally *again* slowly decaying and slightly moving towards higher temperatures. A special case is observed for $\alpha_r = 10$. There is a double-peak structure in the heat capacity, where generally only one peak arises. However, the effect is very small compared to the usual height of the peak, thus it is not unlikely that it is an artifact of the reweighting procedure. As mentioned earlier, the first peak of the heat capacity is thought to constitute the transition from ground-state like structures to globules. The variety of different heights and locations of the first peak for different bond strength indicates that the ground-state structure is highly dependent on α_r.

Therefore, the minimum energy states found during the parallel tempering runs should be investigated more thoroughly. Figure 4.12 shows the overlap and the root mean square deviations (see section 3.3.3) of pairwise compared ground-state configurations for several

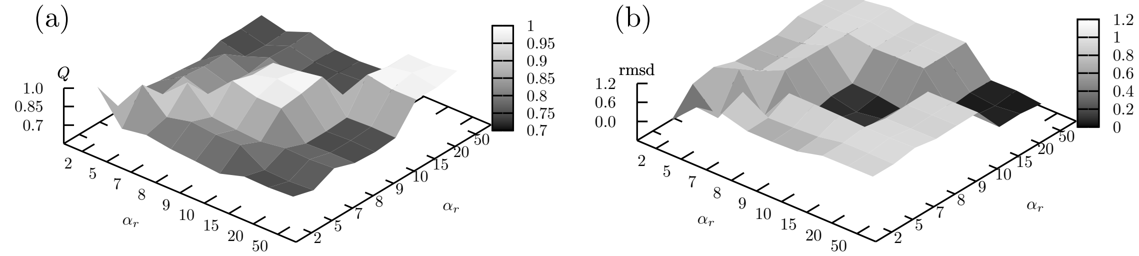


Figure 4.12: The result of the pairwise comparison of the ground-state structures for the observed bond strengths is plotted here. (a) Overlap Q, (b) root mean square deviation.

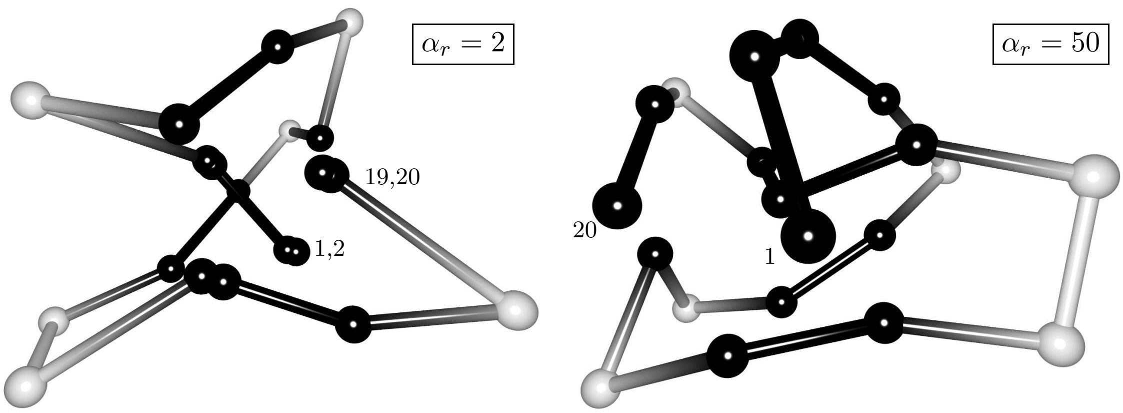


Figure 4.13: Pictures of the ground-states for (a) $\alpha_r = 2$, (b) $\alpha_r = 50$. The dark spheres depict monomers of type A, the light ones are B monomers. The numbers of the first and last monomers are given for better orientation.

values of α_r. It is clear that only structures of equal or relatively close α_r match adequately. All other pairwise comparisons show that there is a wide range of independent structures for the different observed bond strengths.

An interesting feature of weak bonds can be directly seen and qualitatively understood from a ground-state snapshot. Figure 4.13 shows a minimum energy configuration of sequence 20.4 for (a) the weakest ($\alpha_r = 2$), and (b) the strongest bond ($\alpha_r = 50$) that has been studied. The potential energy penalty for two neighbouring monomers that get very close comes only from the bond potential, the Lennard-Jones potential does not act between next neighbours. Particularly, according to the definition of the bond potential (eq. (1.2)), the potential energy is raised by $\alpha_r r_0^2$ if two neighbouring monomers reside in the same position. In the considered case $r_0 = 1$, the energy penalty is actually α_r. Therefore, for low values of α_r it is possible that the benefit from the Lennard-Jones contribution by forming more contacts is greater than the energetic loss from the bond potential. This is exactly what can be seen in not less than four cases in Fig. 4.13 (a). The potential energy of this configuration is $E_{\text{pot}} = -43.9$, while the potential energy of the configuration with $\alpha_r = 50$ in Fig. 4.13 (b) is $E_{\text{pot}} = -34.3$. So the weaker bonds make it really possible to gain a lower ground-state energy. It is conceivable that for higher bond strengths, the monomer pairs break up one by one. This will lead to new types of ground state like configurations again and again and can be taken as an argument

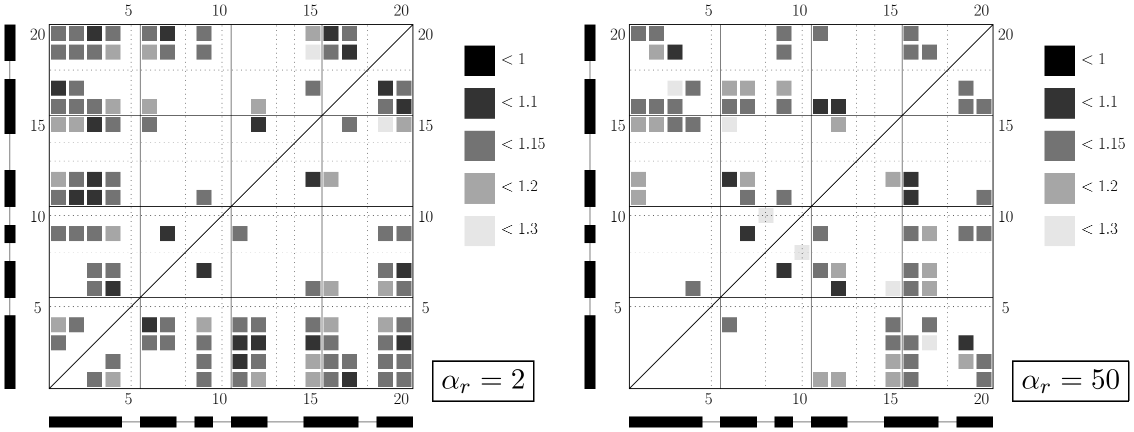


Figure 4.14: A contact map for the ground-states of sequence 20.4 for (a) $\alpha_r = 2$, (b) $\alpha_r = 50$. It shows where the distance between non-neighbouring monomers r_{ij} is rather small. In this case these two monomers are considered to be “in contact”. It is clear that the map is symmetric with respect to the diagonal because $r_{ij} = r_{ji}$. The key at the right shows the range of distances that the certain gray tones mark. Since a contact between A monomers is energetically much more favourable than a contact between B monomers (which can be seen only once in (b) for that reason), the sequence is important for the interpretation of the map. Therefore, the sequence is illustrated at the left and below the map. Thick parts of the line depict an A monomer. Furthermore, the B monomers are marked with dashed lines within the contact map.

for the large variety of forms of the first heat capacity peak. Additionally Fig. 4.13 (a) clearly shows that due to the weak bonds and the unaltered strong Lennard-Jones repulsion, the equilibrium distance for the B monomers is definitely above $r_0(= 1)$.

A look onto the contact maps in Fig. 4.14 makes it possible to roughly estimate the contributions from the Lennard-Jones potential. In Table 4.III the numbers of A-A contacts within a certain range of distances are collected for the two considered cases (a) $\alpha_r = 2$ and (b) $\alpha_r = 50$. Additionally the average value of the Lennard-Jones potential for the particular ranges are given, which makes it possible to estimate the energy contribution of close A-A interactions. The formation of the four monomer pairs for $\alpha_r = 2$ makes roughly $V_{\text{bond}} = 4 \cdot \alpha_r = 8$. Therefore the final approximation neglecting farther A-A and B-B interactions as

Table 4.III: The average Lennard-Jones potential energy for A-A interaction is given within certain distance ranges, as well as the number of contacts in these ranges for the two considered cases (a) $\alpha_r = 2$ and (b) $\alpha_r = 50$ (compare Fig. 4.14). Additionally the total estimate of the close A-A interaction is calculated in the last column.

Range	$r_{ij} \in (1.0, 1.1)$	$r_{ij} \in (1.1, 1.15)$	$r_{ij} \in (1.15, 1.2)$	$r_{ij} \in (1.2, 1.3)$	Total
$\oslash V_{LJ}$	-0.67	-0.99	-0.94	-0.77	
# (a)	12	32	11	1	≈ -50.8
# (b)	5	20	10	2	≈ -34.1

well as the bending term and small bond length deviations is: (a) $V \approx -51 + 8 = -43$ and (b) $V \approx -34$, which is very close to the exact result given before.

In conclusion, weak bond strengths are artificial in principle. Experiment shows that especially for proteins the bonds are nearly rigid. Also the observed monomer-pairs for low α_r are a heavy contrast to the fact that the "monomers" in the model depict whole amino acids in nature, which have a strong steric repulsion of course. Therefore, the found effects in the weak bond regime do not have any relevance in protein folding. Nevertheless, the examination has shown, that there is no systematic impact of flexible but strong bonds, besides that the heat capacity is raised by $\approx 1/2$ due to the bond fluctuations.

4.2 Comparison with Results from Molecular Dynamics simulations at finite temperature

4.2.1 Simulation Fundamentals

Starting Configuration

When choosing a random configuration with unit bond length as the only requirement, the potential energy of this configuration will be generally very large. This comes from the strong steric repulsion (r_{ij}^{-12}) of the Lennard-Jones energy. If two non-neighbouring monomers have a distance of 0.6, the potential energy contribution is about 10^3. Therefore, a random configuration will have a potential energy that is orders of magnitude higher than the expected mean potential energy at a certain temperature. In a Monte Carlo simulation, this is not really a problem. After a small number of updates, the potential energy reaches a reasonable range. In a MD simulation, however, a very high initial energy will lead to high potential gradients, which is equivalent to very large forces. Large forces result in large velocities, so the potential energy is quickly translated into kinetic energy. Thus, before the thermostat is able to bring the system into an equilibrium state by assimilating the excessive amount of energy, the system will simply "explode". To prevent this, a nearly linear starting configuration is chosen randomly, by selecting bond vectors of the following form:

$$\mathbf{b}_i = \begin{pmatrix} (\mathrm{RAN} - 0.5) \cdot 0.1 \\ (\mathrm{RAN} - 0.5) \cdot 0.1 \\ 1 \end{pmatrix} . \tag{4.4}$$

RAN denotes a random number from a uniform distribution in the range $\mathrm{RAN} \in [0, 1)$. The small intrinsic bond length deviations from $r_0 \equiv 1$ do not have a noticeable effect for the utilised bond strength $\alpha_r = 50$. If a really linear configuration would be chosen, there would be only one explicit direction in the system and no force could point away from the linearity, i.e., a linear configuration is an *instable fixed point.* The potential energy of such an elongated configuration is around $E = 0$. It would also be possible to start from a linear configuration and randomly chose non-zero velocities for each monomer. However, in the selected method the dynamics is purely driven by forces and the thermostat, which is more analogous to the desire to observe *deterministic* dynamics.

Adjustment of the Thermostat

The most important source of information when adjusting a Molecular Dynamics simulation with the Nosé-Hoover thermostat are frequency spectra. On the one hand, it is crucial to know the typical, fastest time scale of the system to correctly chose the virtual masses of the NHC, on the other hand only the knowledge of the fastest fluctuation can ensure that the time step δt has not been chosen to large. Furthermore, if the system shows a broad range of frequencies, it is possible that a chain of more than two thermostats $M > 2$ is needed. However, for measuring a frequency spectrum, a MD run has to be performed, i.e. an initial guess of the respective constants is needed. The first assumption is therefore that the bond fluctuations arising from V_{bond} are the fastest within the system. Also, the bonds are approximated as harmonic oscillators by neglecting the influence of the other potential terms. Finally, all monomers are considered to have the same mass $m = 1$ for simplicity. In doing so, the frequency of the bond fluctuation can be calculated in dependency of the bond strength α_r:

$$\alpha_r = \frac{1}{2} m \omega_{\text{bond}}^2 = \frac{1}{2} m \frac{1}{\tau_{\text{bond}}^2} \ ,$$

$$\Rightarrow \qquad f_{\text{bond}} = \frac{\omega_{\text{bond}}}{2\pi} = \frac{1}{2\pi}\sqrt{\frac{2\alpha_r}{m}} \ , \tag{4.5}$$

$$\tau_{\text{bond}} = \sqrt{\frac{m}{2\alpha_r}} \ . \tag{4.6}$$

To check the assumptions for their validity, a whole MD run is performed with sequence 20.2. As a first guess, the length of the Nosé-Hoover-Chain is set to $M = 2$, as required for the harmonic bonds. For being sure that the thermostat can handle potential fluctuation of higher frequencies, the time step is chosen to be $\delta t = 0.001$. The system is simulated at temperature $T = 1.0$. A long equilibrium phase of 10^8 MD steps is performed to make *sure* that the system really is in equilibrium when starting the measurement of $3 \cdot 10^8$ time steps. As a first guess, the size of the Jackknife bins for the error analysis is set to 10^6. At temperature 1.0, the histograms of velocities and bond lengths are measured for comparison with the exact results. Also the time series of the same quantities are used to study the frequency spectrum.

The results can be seen in Figs. 4.15 and 4.16. The average frequency spectrum of the bond length fluctuations (the lower plot in Fig. 4.15) has – as expected – one maximum at $f = f_0$. The velocity spectrum (upper plot) analogously shows the same maximum. At lower frequencies, there are further maxima, which are much more pronounced. These can surely be ascribed to the rest of the interactions. The velocity (component) histograms in Fig. 4.16 perfectly match with the theoretic prediction. The velocities of the thermostat particles should also have a Gaussian distribution, and live up to expectations, which is exemplarily shown for the first thermostat particle in the figure. The coincidence is nearly perfect, so that the measured and theoretic graphs cannot be visually distinguished. Also, histograms of the bond length deviations have been measured. The magnification in the top left corner of the plot shows that the measured distribution behaves very similar to a harmonic oscillator, but the average is slightly larger than the equilibrium bond length. This could be caused by the Lennard-Jones interaction, which does in principle favour distances larger than r_0, although not for neighbouring monomers. However, e.g. from Fig. 4.13 (a) it can be seen that the equilibrium distance especially of B monomers can be definitely larger than r_0 due to the Lennard-Jones potential.

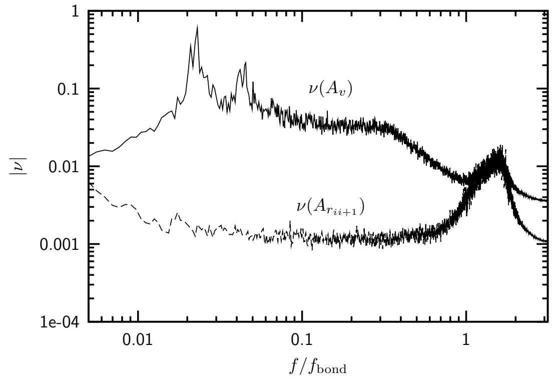


Figure 4.15: Double logarithmic plot of the average frequency spectra of the velocities (upper, solid line) and the bond lengths (lower, dashed line). The frequencies are normalised to the expected bond length fluctuations f_{bond}.

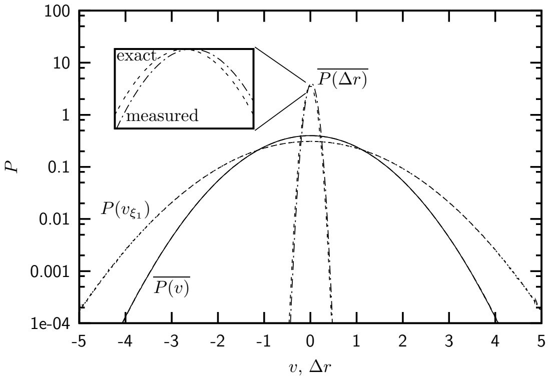


Figure 4.16: Logarithmic plot of the average distributions of the monomer velocity, the velocity of the first NH particle and the difference of the bond length from r_0 (unit length). Since the deviations of the measured distribution and the theoretical is smaller than the line width for the observed velocities, a closer view of the bond length plot is given in the top left corner.

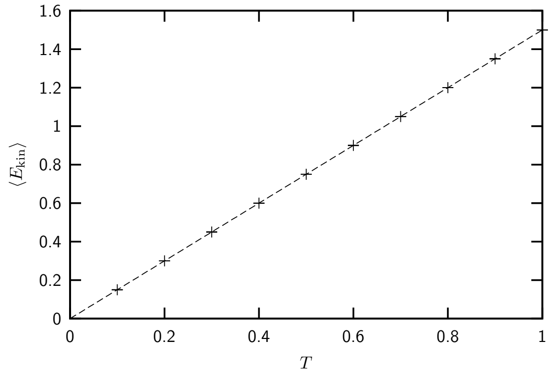


Figure 4.17: Plot of the average kinetic energy (cross symbols) and the theoretic prediction (dashed line) for the MD simulation of sequence 20.2. The error bars are smaller than the line thickness.

Another good test is of course, to compare the average kinetic energy at each temperature with the theoretic prediction $E_{\text{kin}} = (3/2)k_BT$. This comparison is shown in Fig. 4.17. But as it could be expected from the very good results of the two more sensitive previous checks, the agreement of the measured values and the analytic results is great. Obviously, all made choices have been reasonable and can be kept up.

4.2.2 Thermodynamics

So far, only purely kinetic and thus thermally induced effects have been inspected. But up to now it is not clear, whether the results are comparable with those obtained from the Monte Carlo simulations. Therefore, the thermodynamic quantities that have already been

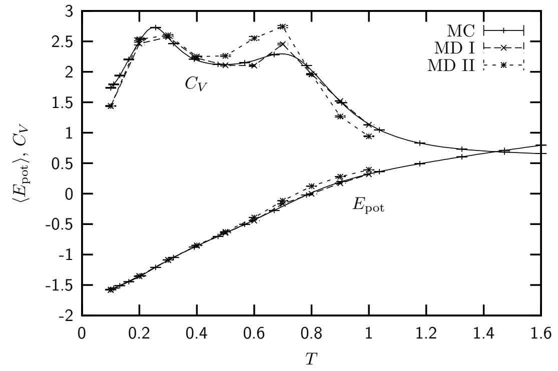


Figure 4.18: Mean energy and specific heat of sequence 20.2, normalised to the number of monomers. The solid line denotes the MC data, while the long- and medium-dashed lines belong to the two MD simulations as described in the text.

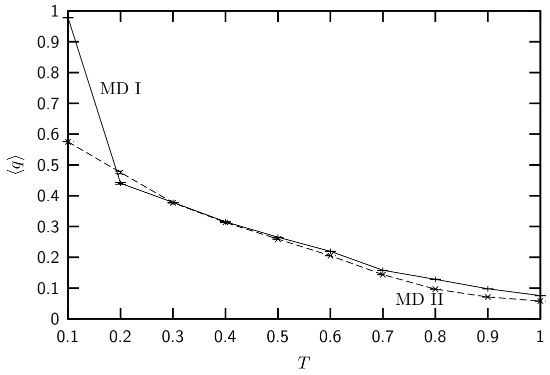


Figure 4.19: Average value of the q parameter for the two simulation types. It is defined as the quotient of the number of instantaneously formed native contacts and the total number of native contacts.

calculated with MC are also measured within Molecular Dynamics. The observed temperature range is $T_1 = 0.1$ to $T_2 = 1.0$ in equidistant temperature steps of $\Delta T = 0.1$. Since MD as it is implemented here is best compared with Metropolis in MC, it is not sure if the phase space is sampled well especially in the low temperature range and at critical temperatures, essentially everywhere, where the autocorrelation times are expected to be large. Thus, two different types of simulations are performed. In one case – MD I – the simulation starts with the ground-state as found with parallel tempering at each temperature. This means, the structure has to "unfold" from its minimum energy state. The other one – MD II – uses a random starting configuration as explained before. Thus, the chain is "folding" to lower energy states at low temperatures. Both use 10^8 MD steps for equilibration at each of the simulated temperatures. The measurement is again performed by simulating 300 Jackknife bins of 10^6 MD steps.

Figure 4.18 shows the mean potential energy and the specific heat as it has been measured in the two different MD runs in comparison to the MC data. Strikingly, in the *low* temperature range the data match very well. Deviations only arise at higher temperatures.

For the energy plots it is remarkable that the MD II "unfolding" simulation shows a slightly higher potential energy than the MC comparison data, the error bars do not overlap. This deviation is negligible in the other simulation. A trivial explanation would be that the thermostat does not manage to drive the system into the relevant low energy domain. But as everything seems to work fine for very *low* temperatures, where the ground-state domain is definitely most relevant, this is obviously not the case. During the observations it seemed as if the effect vanishes with longer equilibration and simulation times. On the other hand it cannot be a purely statistical error, since remarkably it has *never* been observed that the energy in an MD run is lower than in the according MC simulation. It is always systematically higher except for low temperatures. The phenomenon is not really understood, although a small additional investigation is carried out by the use of the Andersen thermostat in an equivalent simulation (see below).

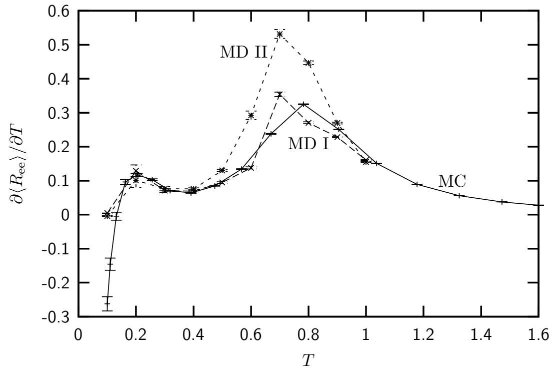


Figure 4.20: Fluctuation of the end-to-end-distance for the two MD runs in comparison to MC.

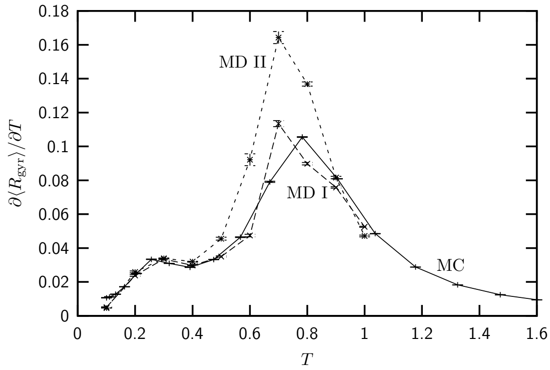


Figure 4.21: Fluctuation of the radius of gyration is plotted for the two MD runs in comparison to MC.

The measurement of the specific heat, which is more sensitive to statistical errors than the plain average energy, confirms the previous result. The MD I simulation, which started from the ground-state at each temperature, perfectly matches the MC comparison data. Although, here the error bars do not overlap at all measured temperatures. However, the effect is small and can be suspected to be due to too small Jackknife bins or too few statistics in general. The MD II run shows larger deviations at temperatures $T > 0.4$, which is not surprising since already the mean energy showed discrepancies. It does not seem as if this error results from the fact that around $T = 0.7$ there is some thermodynamic transition, which is indicated by a peak in the specific heat. If so, the high temperature tail would be reproduced correctly, which is not the case.

In Figs. 4.20 and 4.21, the fluctuations of end-to-end-distance and radius of gyration, respectively, are shown. The same observations can be made. The MD I plot shows the largest difference to the MC data at $T = 0.7$, where also the specific heat showed the largest deviation. The fact that again the error bars do not cover all the discrepancies, is not especially noteworthy. This is considered to be a statistical problem. The behaviour of the second MD simulation is again seriously deviating from the MC results.

Finally, Fig. 4.19 shows the measurement of the q parameter for the two MD simulations. Unfortunately, no data are available from parallel tempering for comparison. However, it can be clearly seen that for the lowest temperature, where the configuration is expected to be close to the ground-state structure, only the simulation which *starts* from the ground-state samples the respective domain of the configuration space correctly. In the second case, the structurally relevant domain is not found at all. This can be proven by comparing the state of lowest energy with the ground-state found by parallel tempering, as it is done in Table 4.IV. The comparison of the parallel tempering ground-state and the minimum energy structure found by a multicanonical simulation in ref. [42] shows a very well agreement, which supports the assumption, that it belongs to the *global* energy minimum. The lowest energy that is found by the unbiased MD run, which means that no structural information is put into the simulation, is about $\Delta E = 0.66$ higher than in the parallel tempering run. This is a serious difference. Also, the measurements of overlap and root mean square deviation show clearly, that the structures are definitely not equivalent.

In conclusion, the results of the measurement showed that the qualitative thermodynamic and structural behaviour of the considered system in a Molecular Dynamics simulation, where the canonical ensemble is provided by a Nosé-Hoover-Chain thermostat with the described adjustments, is very similar to what is seen in sophisticated Monte Carlo simulations. However, there are quantitative deviations beyond the error bars, whose origin is not fully understood. One possible reason is that the evolution of the system within MD is obviously much slower than in a comparable MC run. Therefore, for producing data with an acceptable error analysis, a thorough investigation of the statistical parameters like equilibration time, Jackknife bin size and of course simulation time has to be carried out. Some thoughts considering this issue can be found in the next subsection.

Cross-Check with Andersen Thermostat

The question of how the effect of the systematically higher mean energies in the MD simulation could be explained, is still open. It is interesting to observe what happens when changing the thermostat. The Andersen thermostat can be considered to be a hybrid algorithm between MC and MD. Therefore, if the considered problem does not occur when utilising the Andersen thermostat, this would be a definite sign that the problem has to do with the specifics of the Nosé-Hoover-Chain thermostat. If the effect is only decreased but still exists, it might be possible that it is a general problem of MD. All parameters of the run are chosen as for the previous MD II simulation. As before, the starting configuration is an elongated but not linear chain. The collision frequency is selected to be $\nu = 1$, which seems to be a reasonable choice in general (compare section 2.4.1).

Indeed, the plots of the average potential energy and the specific heat in Fig. 4.22 show a perfect agreement with the parallel tempering MC data. The error bars overlap over the whole temperature range. No deviation can be seen in the lower potential energy plot. This supports the assumption that the erroneous behaviour that is observed in the MD NHC simulation is really caused by the specific type of thermostat.

Interestingly the q parameter, which is shown in Fig. 4.23, is again too small for low temperatures. But in contrast to the observations for the MD NHC simulation, the ground-state is found! As Table 4.V shows, the minimum energy configuration of the Andersen run is even $\Delta E = 0.62$ *lower* than the state that is found in the parallel tempering comparison simulation. But, as the overlap and root mean square deviation shows the two ground-state configurations are systematically equal. This finding means that obviously the ground-state is found too late within the simulation, when the measurement had already been started. Thus, a longer equilibration time would be necessary to prevent this problem. The fluctuations of end-to-end-distance and radius of gyration are also reproduced well in the particular simulation as seen in Figs. **??** and 4.25.

Table 4.IV: Difference of the lowest energy, overlap and root mean square deviation for the observed minimum energy states by parallel tempering and an MD run which started from an elongated chain.

ΔE_{min}	Overlap	rmsd
0.66	0.74	1.22

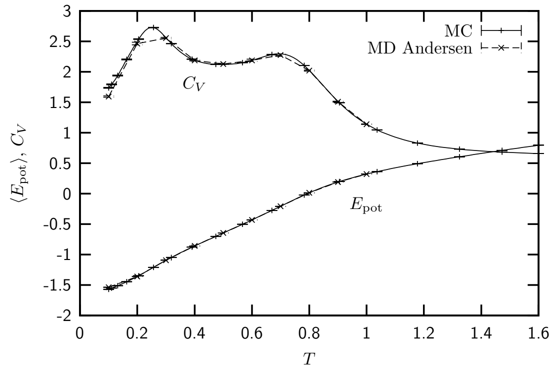


Figure 4.22: Mean energy and specific heat of sequence 20.2, normalised to the number of monomers. The solid line denotes the MC data, while the dashed line belong to the MD simulation with the applied Andersen thermostat.

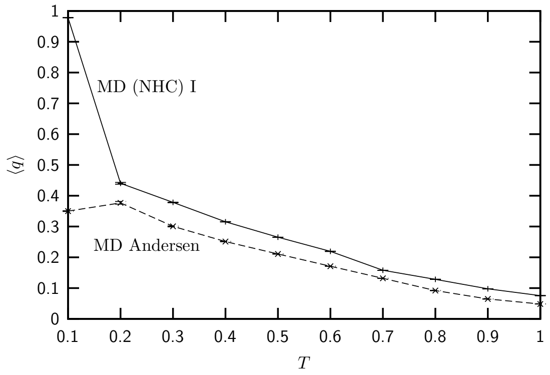


Figure 4.23: Average value of the q parameter in the simulation with Andersen thermostat is compared to the previously described MD I simulation, where the starting configuration is the lowest energy state as found by parallel tempering.

4.2.3 Time Scales

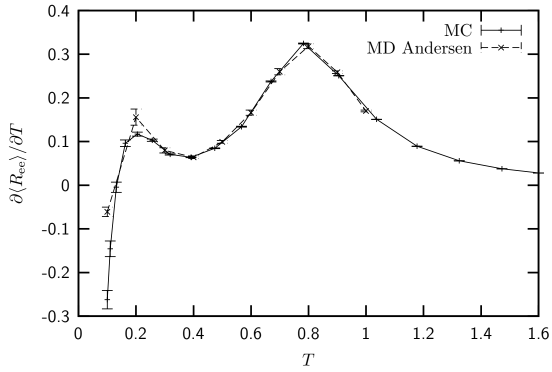
Several interesting questions concerning time scales can be asked when performing continuous, deterministic Molecular Dynamics simulations.

1. The most important: In contrast to Monte Carlo simulations, a *physical time* is implied. By translating quantities and scales of the model into the "real world", or vice versa, the MD time can be identified with a real time scale in seconds. This is not that easy possible in a Monte Carlo simulation. However, the estimation of the real time scale for the model at hand is difficult if possible at all. The reason is the artificial character of the model, which impedes the identification of energy and temperature scales.

2. Although the dynamics in a Monte Carlo simulation is not deterministic, there is of course still Markovian *dynamics*. The comparison of kinetic effects in MC and MD is thus very interesting. For the considered systems, a good opportunity would be so called Chevron plots [43, 44, 45], where folding kinetics are analysed. The incorporated measurements are quite exhausting, thus this is not done in this work. The expectation is that the dynamic behaviour in MC and MD is similar and the MC time scale can be translated into a MD time scale. The question is: How many MD time steps can be compared to one MC sweep?

Table 4.V: Difference of the lowest energy, overlap and root mean square deviation for the observed minimum energy states by parallel tempering and an MD run which started from an elongated chain and is thermostated by an Andersen thermostat with a collision frequency $\nu = 1$.

ΔE_{min}	Overlap	rmsd
-0.62	0.97	0.07



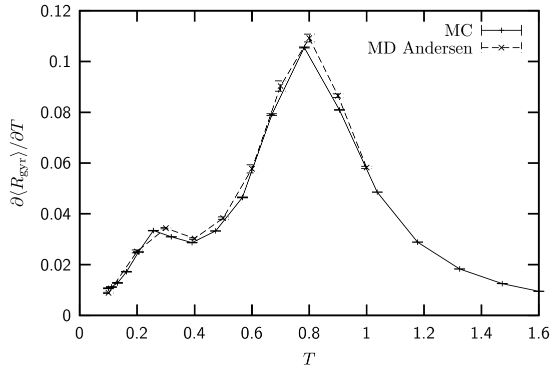


Figure 4.24: Shows the fluctuation of the end-to-end-distance for the MD run with the Andersen thermostat in comparison to MC.

Figure 4.25: The fluctuation of the radius of gyration is plotted for the MD run with the Andersen thermostat in comparison to MC.

3. In Markov processes, there is an autocorrelation time, which is induced by the fact that each of the instantaneous configurations of a system has been generated from the previous state by some update procedure. Hence, there is an intrinsic memory. In Molecular Dynamics, the successive states are highly correlated, since there is a set of deterministic rules – the Newtonian equations of the system – defining how each configuration is developed from the previous one. Therefore, for the correct error analysis it is important to evaluate these autocorrelations.

Autocorrelation Times

Contrary to Monte Carlo dynamics, in MD there are explicit *fluctuations*, e.g. the bond fluctuations in the considered system. Therefore, it would be possible that the relatively quick changes in energy due to these fluctuations cause an artificially rapid fall-off of the autocorrelation function. Thus, it could be necessary to measure autocorrelation functions for each potential energy contribution independently and take the highest autocorrelation time found by this procedure as the autocorrelation time of the whole system. This is not done here.

Instead, Metropolis and MD simulations are carried out with sequence 20.4. In contrast to parallel tempering, a single Metropolis simulation is more comparable to MD, since the system cannot jump to a completely different phase space area within one sweep – which happens quite often in parallel tempering, when replica are exchanged. The update and step sizes are chosen as in the simulations in the previous sections. In the Metropolis simulation the update of a bond vector means a displacement of every component between -0.1 and 0.1. For MD the time step is $\delta t = 0.001$ again. The autocorrelation function is measured at $T = 0.25$, where the specific heat has a maximum (see Fig. 4.5), and at $T = 1.0$. In each case a time series of 10^8 successive potential energy values is measured after an equilibration of 10^7 sweeps or steps respectively. Only in the MD simulation at $T = 0.25$, where the exponential behaviour of the autocorrelation function is only observed for large time separations k, a longer time series of 10^9 MD steps is necessary to have enough statistics for large k.

The results are shown in Fig. 4.26, and the exponential autocorrelation times extracted

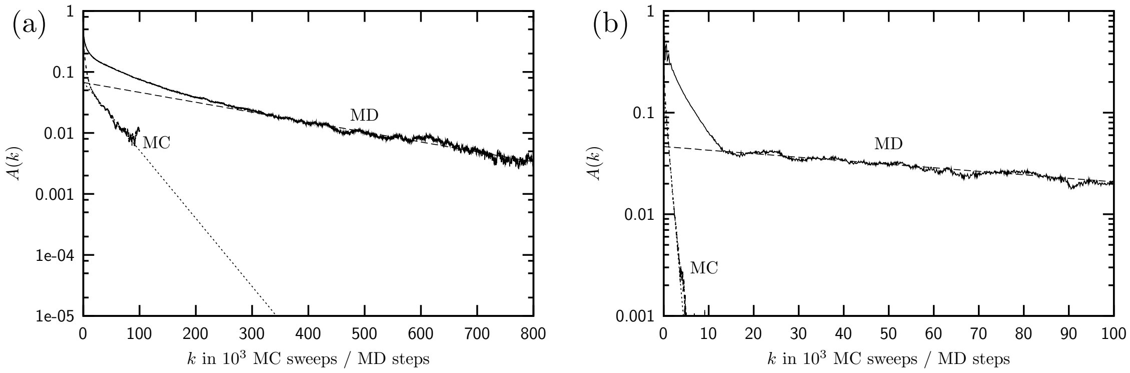


Figure 4.26: Logarithmic plot of autocorrelation functions for a Metropolis and a MD simulation of sequence 20.4 at (a) $T = 0.25$ and (b) $T = 1.0$.

from the functions by fitting are listed in Table 4.VI. The MD autocorrelation function for $T = 0.25$ could easily and mistakenly be considered to behave already exponential for much lower values of k. Only the $A(k)$ plot for large time separations k shows that it converges extremely slow against the final exponential fit. For $T = 1.0$ the change in the behaviour is more like a kink and can be identified easily.

The listing of τ_{exp} in Table 4.VI shows that at the specific heat peak ($T = 0.25$) the autocorrelation times in terms of *simulation iterations* (MC sweeps/MD steps) are in the same range. The exponential autocorrelation time of the Metropolis simulation is 3 times smaller than τ_{exp} in MD. In contrast a serious difference can be seen for $T = 1.0$. There, the autocorrelation times differ by more than two orders of magnitude. If it is possible to generalise this finding this would mean that contrary to Metropolis, the autocorrelation time in a Molecular Dynamics with a NHC thermostat is not as susceptible for transition regions. This would also mean that the identification of MC and MD time scales is not generally possible. But this should be checked by further examinations. A definite outcome of the explicit values of τ_{exp} in Table 4.VI is that *if* the error analysis in a MD simulation can be carried out as for a Markov process, the chosen Jackknife interval of 10^6 time steps does *not* guarantee uncorrelated data and is thus too small.

Table 4.VI: Exponential autocorrelation times of MC and MD.

Simulation	T	τ_{exp} in sweeps/steps
MC	0.25	$40 \cdot 10^3$
MC	1.0	850
MD	0.25	$250 \cdot 10^3$
MD	1.0	$125 \cdot 10^3$

Search for States with Minimal Energy

From the previous paragraph it is already apparent that because of the faster evolution of the system in Markov dynamics, a Monte Carlo simulation should be generally faster in finding configurations with minimal energy. However, there is also an objective technical reason, for which MC simulations should be preferred for ground-state searches. In MC it does not cause any problem to choose high temperatures and propose large updates for having a fast random walk through the phase space. It is even possible to use sophisticated techniques like *energy landscape paving* [46] to drive the system into regions of minimal potential energy. Molecular Dynamics is much more sensitive with respect to the choice of big time steps or other technical tricks to accelerate the overall dynamics. Actually, while in Monte Carlo maybe the ensemble is not sampled correctly anymore (like when using energy landscape paving), a wrong choice of the parameters in Molecular Dynamics can quickly lead to an "explosion" of the system, e.g. when large forces arise due to too large step sizes. Therefore, MD is much more restricted compared to Markov dynamics. Searching for states with minimal energy will thus be always better carried out with a Monte Carlo program.

Computer Time

The question whether a Monte Carlo sweep or a Molecular Dynamics step is faster with respect to the computational effort, is easy to answer. In most of all cases, Molecular Dynamics is faster. The reason is that a Monte Carlo sweep usually includes N updates for a system with N particles. In particular, all the necessary quantities for evaluating the potential energy have to be calculated N times every MC sweep. When performing Molecular Dynamics, once the forces of the system have been evaluated for a certain configuration, the application of e.g. the Störmer-Verlet algorithm handles the evolution of all particles at once. The calculation of the forces is usually more costly than the calculation of the potential energy. However, for the case at hand, the calculation of the pairwise monomer distances r_{ij} is most expansive, since it has the complexity $\mathcal{O}(N^2)$. This has to be done for calculating both the potential energy and the forces. The rest of the calculations are of complexity $\mathcal{O}(N)$. Therefore, even the fact that for the force three components have to be calculated, does not change the finding: One Molecular Dynamics step is less exhausting than one Monte Carlo sweep, especially for large N. On the other hand, recalling the results of the analysis of autocorrelation times, it is expected, that for large numbers of particles, the autocorrelation still decays much faster for one Monte Carlo sweep. Particularly, because for large N, a large number of updates is proposed every sweep. Therefore, from the statistical point of view, Monte Carlo surely provides more uncorrelated data in the same computer time compared to Molecular Dynamics.

4.3 Free-Energy Landscapes

Fundamentals

In protein folding research interest is directed towards the path to fold, i.e. the question: How does the protein find its native fold starting from a more or less random configuration. Proteins are very complex and have thus a very high-dimensional free-energy landscape. In general, it is thus very difficult to make statements about the path to fold, since a trivial localisation of the protein within the phase space is not possible. Therefore, one of the key problems is finding a suitable reaction coordinate, in whose dependence the free-energy

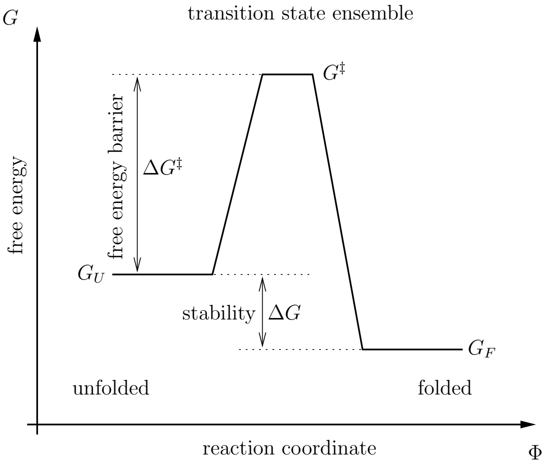


Figure 4.27: Schematic view of the free-energy landscape of a two-state folding protein. This scheme uses the symbols as they are usual in biochemistry, i.e. G is used for the *Gibbs* free-energy and Φ is some general reaction coordinate. The Gibbs free-energy G is considered to be equal to the free-energy F for reactions at constant temperature and pressure.

landscape can be observed. However, once the ground-state has been found, it is common to define the reaction coordinate as the degree of equality of the instantaneous structure and this ground-state [39].

In chapter 3 several methods are shown, how structures can be compared. In the following the according quantities, the q parameter and the overlap Q, are considered as reaction coordinates. The partition sum of a system can be written in terms of the free-energy:

$$Z = e^{-\beta F(T)} = \int \mathrm{d}^N \mathbf{X}\, e^{-\beta E(\mathbf{X})} \ . \tag{4.7}$$

Asking for the free-energy of a system with respect to a certain parameter, e.g. the reaction coordinate Q, the ensemble of possible configurations has to be constrained to configurations matching a certain value of the parameter $Q = Q_0$:

$$e^{-\beta F(T,Q_0)} = \int \mathrm{d}^N \mathbf{X}\, e^{-\beta E(\mathbf{X})} \delta(Q(\mathbf{X}) - Q_0) \sim P(Q_0) \ . \tag{4.8}$$

Therefore, the free-energy can be expressed in dependence of Q as:

$$F(Q,T) \sim -k_B T \ln P(Q) \ . \tag{4.9}$$

A very simple model of protein folding, which is suitable for small proteins [47], is the two-state folding. Near the *folding temperature* T_f, the free-energy landscape is expected to look schematically like the one shown in Fig. 4.27. The state space is divided into two parts, the *unfolded* and the *folded* state, which leads to the notation two-state folding. The two domains are separated by a *transition state ensemble*, which are states with a relatively high free-energy. In particular, folded states can be identified by having a value of $Q \approx 1$. On the other hand, unfolded states will have significantly lower values of Q.

In Ref. [45], it is shown that simple protein models like the AB model can also have a two-state folding behaviour. Since in [45] all simulations are done with Monte Carlo methods,

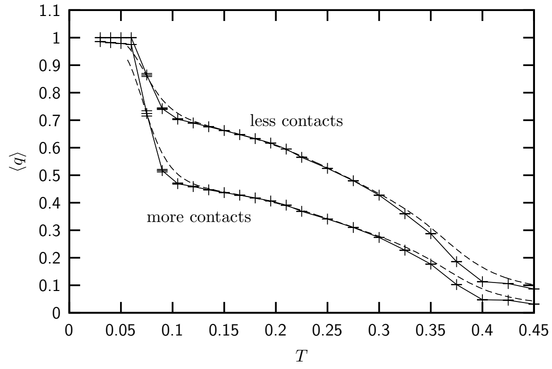


Figure 4.28: The q parameter of sequence 20.6 for the two definitions given in the text. The solid lines with symbols and error bars depict the measured averages, the dashed lines result from reweighting.

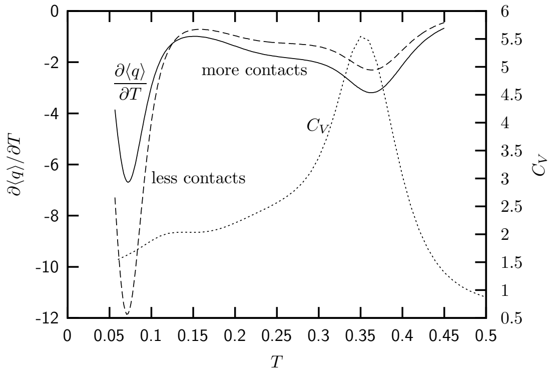


Figure 4.29: Fluctuation of the q parameter for the two different definitions (solid and dashed line) as obtained by reweighting. The dotted line shows the specific heat of the sequence 20.6. The respective scale is found on the right.

it is interesting to investigate, whether the same can also be seen with Molecular Dynamics. This is proven for another type of model, the GōC$^\alpha$ model in Ref. [39].

Because sequence 20.6 (see Table 1.I) has a very sharp peak in the heat capacity (see Fig. 4.5), the expectation is that it behaves like a two-state folder. Because the folding temperature is very low, which will be shown in the next paragraph, the system will move slowly through the configuration space. Therefore, the time step is chosen larger than in the previous section $\delta t = 0.006$, by accepting larger numerical errors. Based on the previous finding that the native state is hardly found by Molecular Dynamics, it is decided to start the simulations from the ground-state configuration. Because of the larger time step, the system is equilibrated for only 10^6 steps. Thereafter, a measurement of $6 \cdot 10^8$ steps is performed. This procedure is repeated at each measured temperature.

Folding Temperature

The question of how to define the folding temperature is not trivial. In Ref. [48] it is proposed that the folding temperature is, where the population of the ground-state is $P_{\text{nat}}(T_f) \approx 1/2$. Another attempt is to define T_f as the temperature, where half of the native contacts are formed on average [39], i.e. $\langle q \rangle \approx 1/2$ (compare section 3.3.3). However, both definitions depend very much on the used method to determine the degree of equality of the instantaneous and the native conformation.

Figure 4.28 shows the average q parameter for two different definitions. The first one, denoted with “less contacts”, is analogous to the description in section 3.3.3. According to this definition, monomers are considered as native contact, if their distance in the ground-state is $r_{ij} < 1.7$ and if they have at least two monomers between them in the chain $|i-j| \geq 3$. The second definition, denoted with “more contacts”, excludes only next neighbours (which are considered as chemically bound) $|i-j| \geq 2$ from being counted as native contacts, and instead has a sharper distance criterion: $r_{ij} < 1.3$. The most important observation about the plots in Fig. 4.28 is, that the temperature where $\langle q \rangle \approx 1/2$ is very different for the two

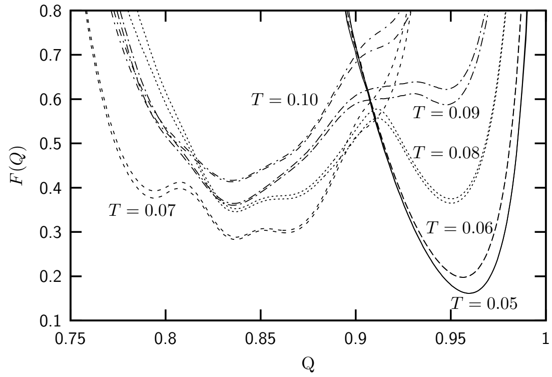


Figure 4.30: Free-energy landscape for sequence 20.6 around the folding temperature. Performing a Jackknife error analysis for every histogram bin of $P(Q)$ gives an error estimate for $F(Q)$. The two lines belonging to each temperature denote the upper and lower boundary of $F(Q)$ according to the error bars.

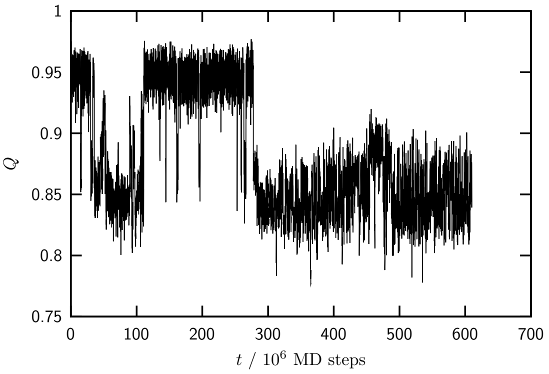


Figure 4.31: Time series of overlap Q at $T = 0.08$.

definitions. Thus it is not a good criterion to define the folding temperature.

In Fig. 4.29, the fluctuations of q are shown for both definitions. It is clear that the fluctuations have minimum peaks at the same temperatures. Therefore, it is reasonable to correlate the folding temperature with the temperatures, where the peaks of the q fluctuation arise. Furthermore, the specific heat of the sequence 20.6 is shown in the figure, which makes clear that the peaks of the specific heat and the minima of $\partial\langle q\rangle/\partial T$ collapse. In particular, the minimum of the fluctuation of q at lower temperature is only seen as a shoulder in the specific heat. Also, a small deviation in the extremum temperatures can also be ascribed to the finite size of the system. Hence, the explanation of the peaks in the specific heat can be translated to the minima in the fluctuation of q. Two-state folding behaviour considers folded and unfolded domains, i.e. it is expedient to define the first transition temperature as the folding temperature, since there the system switches from the ground-state to the globule domain. This is a reasonable definition of the folding temperature. In the example of sequence 20.6, the folding temperature is thus about $T_f = 0.07$. The result here is very similar to the Monte Carlo data in [45].

Two-State Folding

In the following, the overlap parameter Q as introduced in section 3.3.3 is used as reaction coordinate instead of the q parameter. Figure 4.30 shows the result of a measurement of the free-energy landscape at $T = \{0.05, 0.06, \ldots, 0.1\}$ around the folding temperature $T_f \approx 0.07$. Recalling that the free-energy is linked to the population of certain states, the plots are as expected. For the lower temperatures, only states with $Q \approx 1$ are pronounced, which gives a deep valley in $F(Q)$. For higher temperatures, a second valley evolves, which is due to the increasing population of states in the globular domain. The energy landscape at the estimated folding temperature $T = T_f = 0.07$ is remarkable. The ground-state domain is not populated at all, which gives only one broad valley at lower values of Q. Instead, the plot for $T = 0.08$

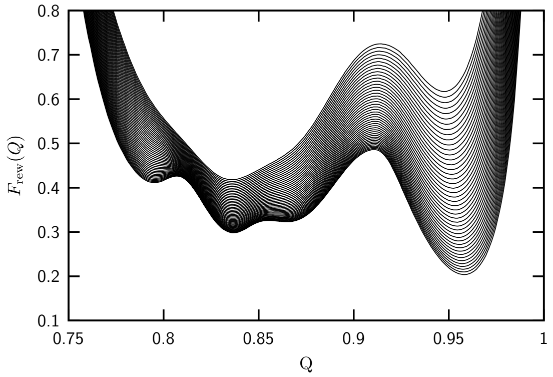


Figure 4.32: Reweighted free-energy landscape for sequence 20.6 around folding temperature $T \approx T_f$.

shows the expected two-state behaviour, since it clearly has two valleys separated by an area of higher free-energy – the transition state ensemble.

The reason for the observed misbehaviour at $T = T_f = 0.07$ gets clear from Fig. 4.31. Although the MD runs are twice as long as for the measurements in the previous section, the structural state space is still not sampled well enough. In the *whole* simulation of $6 \cdot 10^8$ steps, the system undergoes only three changes between the ground-state domain to the globule domain. Actually, this can be considered as three independent events, which has no statistical significance. This is a serious problem, since it is not easily possible to magnify the number of steps of the MD run by a serious factor. For this reason, using a single MD run for measuring the free-energy landscape at a certain temperature is questionable.

However, it is possible to use reweighting methods as explained in section 3.2.2 to enhance the statistical significance of the *sum* of all performed runs. This technique is also utilised in Ref. [39], where simulations of the GōC^α model are performed with Molecular Dynamics. Since data was collected at $T < T_f$ as well as $T > T_f$, reweighting gives a good estimate of what the free-energy landscape looks like at $T \approx T_f$. The result is shown in Fig. 4.32. For the low-temperature range, several intermediate states cause buckles in the valley of unfolded states. These buckles can also be seen at several temperatures in Fig. 4.30. Again, the comparison to the Monte Carlo data in [45] shows a good qualitative agreement.

Chapter 5

Torsion – Extending Towards a Generalised Coarse-Grained Heteropolymer Model

From experiment it is known that not only the bond angles ϑ play a role in the energy function of a polymer, but also *torsion* angles, denoted with ϕ in the following. For example the so-called "trans"-conformation, where $\phi = \pm\pi$, is well-known to be energetically favourable. This property of natural occurring polymers was incorporated into a coarse-grained heteropolymer model in Ref. [49]. In the following it is denoted "GAB model". Unfortunately, while this extension does not pose any difficulties in Monte Carlo simulations, there arise systematic problems in a Molecular Dynamics simulation - besides technical challenges, since the terms of the potential gradient get very large and complex.

In this chapter this problem will be discussed in detail. First the potential term will be introduced and the *torsion angle* shall be illustrated. Furthermore, two possible definitions for the angle will be shown to be analytically identical in the cartesian transcription. Afterwards, the torsional force will be derived. Finally the occurrence and treatment of the already mentioned systematic problem will be described and some words of outlook concerning it will conclude the considerations.

5.1 The Torsion Potential

A torsional angle is built up by three successive bond vectors $\mathbf{b}_{1,2,3}$. It is defined to be $\phi \in [-\pi, \pi]$. There are two analogous possibilities of defining it. One is to describe ϕ by the angle between the pairwise cross products $\mathbf{b}_1 \times \mathbf{b}_2$ and $\mathbf{b}_2 \times \mathbf{b}_3$. It is also possible to appoint $\phi \pm \pi$ as the angle between the projections of the $\mathbf{b}_1$ and $\mathbf{b}_3$ on the plane perpendicular to $\mathbf{b}_2$. Figures 5.1 and 5.2 shall illustrate the two definitions.

It is easy to see that so far it is not clear, how the algebraic sign of ϕ should be determined, since the enclosed angle of two vectors is defined to be in $[0, \pi]$. In Figs. 5.1 and 5.2 $\mathbf{b}_3$ is supposed to point at the left half space with respect to the plane spanned by $\mathbf{b}_1$ and $\mathbf{b}_2$ if the viewer looks into the direction of the latter. If $\mathbf{b}_3$ would be mirrored by this plane, the absolute value of ϕ would stay the same for both definitions, but $\mathbf{b}_3$ would point at the right half space. The two configurations should be distinguished by the algebraic sign of ϕ. An easy way to achieve this is to check, whether the projection of $\mathbf{b}_3$ on the normal vector of the

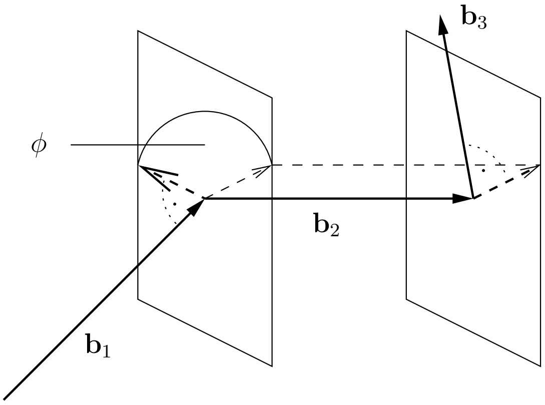


Figure 5.1: The cross product definition of the torsion angle. Two planes perpendicular to $\mathbf{b}_2$ are hinted for a better understanding, because the resulting cross product vectors will lie in such a plane.

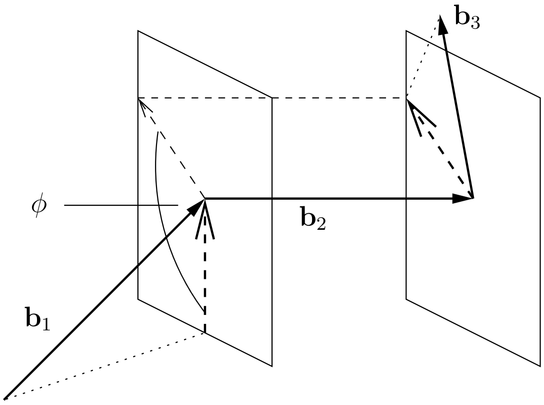


Figure 5.2: The scalar product definition of the torsion angle. The perpendicular (projection) planes of $\mathbf{b}_2$ are drawn again. As it can be seen from the picture, ϕ is not the *enclosed* angle between the two projection vectors, thus π has to be subtracted or added according to the torsional direction.

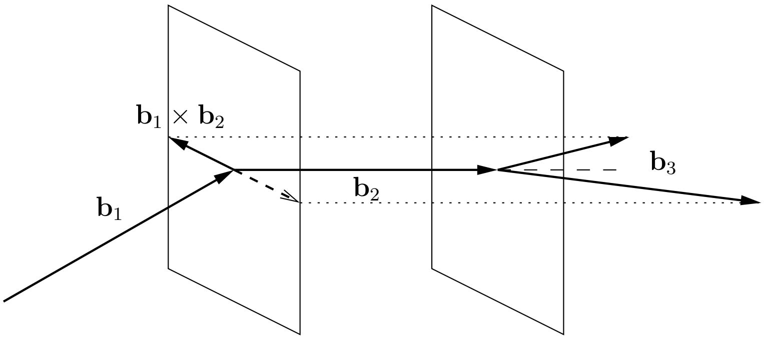


Figure 5.3: Two different cases for $\mathbf{b}_3$ which will lead to the same value for ϕ. But the projection on $\mathbf{b}_1 \times \mathbf{b}_2$ is either positive or negative. Thus the two configurations can be distinguished by an algebraic sign for ϕ according to the sign of the described projection.

plane spanned by $\mathbf{b}_1$ and $\mathbf{b}_2$ – which is the cross product of these two vectors – has a positive or negative algebraic sign. This is equivalent to distinguishing a “right” or “left” half space. Figure 5.3 shall help to make this point less confusing.

Finally, the definition of the torsion potential is:

$$V_{\text{tors}}(\mathbf{R}) = \frac{1}{2} \sum_{k=1}^{N-3} \left(1 + \cos 3\phi_k\right) \ . \tag{5.1}$$

From this definition it can be seen that the algebraic sign of the torsional angles ϕ does not play a role in calculating the potential energy of a configuration, since the cosine is an even function ($\cos\phi = \cos(-\phi)$). Also it is clear that the values $\phi = \pm\pi/3$ and $\phi = \pm\pi$ are energetically favoured, which is motivated by results from experimental polymer research.

5.1.1 Cartesian Transcription

For translating (5.1) into a cartesian formulation, the definitions from chapter 1 are used. Furthermore a trigonometric identity is utilised:

$$\cos 3\phi = 4\cos^3\phi - 3\cos\phi \ . \tag{5.2}$$

From (5.2) it can be seen that for calculating the torsional contribution to the potential energy it is again only necessary to know the cosine of the torsional angle ϕ. Thus (1.14) can be used for expressing the according potential:

$$V_{\mathrm{tors}}(\mathbf{R}) = \frac{1}{2}\sum_{k=1}^{N-3}(1+\cos 3\phi_k) = \frac{1}{2}\sum_{k=1}^{N-3}\left(1+(4\cos^2\phi_k - 3)\cos\phi_k\right) \ . \tag{5.3}$$

It will be shown that the two definitions of ϕ explained in Figs. 5.1 and 5.2 lead to analytically equivalent expressions in the cartesian transcription.

First, the cross product definition will be faced. The Binet-Cauchy identity will help to rewrite the appearing cross products with scalar products, which are mostly also used in the bending potential (compare section 1.4):

$$(\mathbf{a}\times\mathbf{b})\cdot(\mathbf{c}\times\mathbf{d}) = (\mathbf{a}\cdot\mathbf{c})(\mathbf{b}\cdot\mathbf{d}) - (\mathbf{a}\cdot\mathbf{d})(\mathbf{b}\cdot\mathbf{c}) \ . \tag{5.4}$$

The consequential scalar product of the two plane normal vectors has to be normalised:

$$\begin{aligned}\cos\phi_k &= \frac{(\mathbf{b}_k\times\mathbf{b}_{k+1})\cdot(\mathbf{b}_{k+1}\times\mathbf{b}_{k+2})}{\sqrt{[(\mathbf{b}_k\times\mathbf{b}_{k+1})\cdot(\mathbf{b}_k\times\mathbf{b}_{k+1})]\,[(\mathbf{b}_{k+1}\times\mathbf{b}_{k+2})\cdot(\mathbf{b}_{k+1}\times\mathbf{b}_{k+2})]}} \\ &= \frac{(\mathbf{b}_k\cdot\mathbf{b}_{k+1})(\mathbf{b}_{k+1}\cdot\mathbf{b}_{k+2}) - \mathbf{b}_{k+1}^2(\mathbf{b}_k\cdot\mathbf{b}_{k+2})}{\sqrt{\left[\mathbf{b}_k^2\mathbf{b}_{k+1}^2 - (\mathbf{b}_k\cdot\mathbf{b}_{k+1})^2\right]\left[\mathbf{b}_{k+1}^2\mathbf{b}_{k+2}^2 - (\mathbf{b}_{k+1}\cdot\mathbf{b}_{k+2})^2\right]}} \ .\end{aligned} \tag{5.5}$$

Second, the scalar product definition will be used to derive a cartesian expression for $\cos\phi$. The projection of $\mathbf{b}_k$ on a plane with the normal vector $\mathbf{b}_{k+1}$ can be calculated by subtracting the contribution of $\mathbf{b}_k$ that is parallel to $\mathbf{b}_{k+1}$ and will be denoted as $\mathbf{b}_{k,\perp}$. Therefore, a unit vector in direction of $\mathbf{b}_{k+1}$ is needed:

$$\mathbf{b}_{k,\perp} = \mathbf{b}_k - \left(\mathbf{b}_k\cdot\frac{\mathbf{b}_{k+1}}{b_{k+1}}\right)\frac{\mathbf{b}_{k+1}}{b_{k+1}} = \mathbf{b}_k - \left((\mathbf{b}_k\cdot\mathbf{b}_{k+1})\,\mathbf{b}_{k+1}\right)b_{k+1}^{-2} \ . \tag{5.6}$$

Even if $\mathbf{b}_i$ are unit length vectors, $\mathbf{b}_{k,\perp}$ does *not* have unit length anymore in the general case. Analogous to (5.6) the projection of $\mathbf{b}_{k+2}$ can be evaluated. As described earlier, ϕ is not the *enclosed* angle of $\mathbf{b}_{k,\perp}$ and $\mathbf{b}_{k+2,\perp}$. Thus, the enclosed angle has to be shifted by π. Since $\cos(\phi\pm\pi) = -\cos\phi$ it is also possible to just multiply one of the two vectors by -1 (ϕ is actually the enclosed angle of one projection vector and the negative other projection as described above). Utilising (1.14) and recalling that it is *necessary* to normalise $\mathbf{b}_{k,\perp}$ and

$\mathbf{b}_{k+2,\perp}$ it is obtained:

$$
\begin{aligned}
\cos\phi_k &= -\frac{\left(\mathbf{b}_k - \left(\left(\mathbf{b}_k\cdot\mathbf{b}_{k+1}\right)\mathbf{b}_{k+1}\right) b_{k+1}^{-2}\right)\left(\mathbf{b}_{k+2} - \left(\left(\mathbf{b}_{k+2}\cdot\mathbf{b}_{k+1}\right)\mathbf{b}_{k+1}\right) b_{k+1}^{-2}\right)}{\sqrt{\left[\mathbf{b}_k - \left(\left(\mathbf{b}_k\cdot\mathbf{b}_{k+1}\right)\mathbf{b}_{k+1}\right) b_{k+1}^{-2}\right]^2 \left[\mathbf{b}_{k+2} - \left(\left(\mathbf{b}_{k+2}\cdot\mathbf{b}_{k+1}\right)\mathbf{b}_{k+1}\right) b_{k+1}^{-2}\right]^2}} \\
&= \frac{\left(b_{k+1}^2\mathbf{b}_k - \left(\mathbf{b}_k\cdot\mathbf{b}_{k+1}\right)\mathbf{b}_{k+1}\right)\left(-b_{k+1}^2\mathbf{b}_{k+2} + \left(\mathbf{b}_{k+2}\cdot\mathbf{b}_{k+1}\right)\mathbf{b}_{k+1}\right)}{\sqrt{\left[b_{k+1}^2\mathbf{b}_k - \left(\mathbf{b}_k\cdot\mathbf{b}_{k+1}\right)\mathbf{b}_{k+1}\right]^2 \left[b_{k+1}^2\mathbf{b}_{k+2} - \left(\mathbf{b}_{k+2}\cdot\mathbf{b}_{k+1}\right)\mathbf{b}_{k+1}\right]^2}} \\
&= \frac{-b_{k+1}^4(\mathbf{b}_k\cdot\mathbf{b}_{k+2}) + b_{k+1}^2(\mathbf{b}_k\cdot\mathbf{b}_{k+1})(\mathbf{b}_{k+1}\cdot\mathbf{b}_{k+2})}{\sqrt{\left[b_{k+1}^4\mathbf{b}_k^2 - b_{k+1}^2\left(\mathbf{b}_k\cdot\mathbf{b}_{k+1}\right)^2\right]\left[b_{k+1}^4\mathbf{b}_{k+2}^2 - b_{k+1}^2\left(\mathbf{b}_{k+2}\cdot\mathbf{b}_{k+1}\right)^2\right]}} \\
&= \frac{(\mathbf{b}_k\cdot\mathbf{b}_{k+1})(\mathbf{b}_{k+1}\cdot\mathbf{b}_{k+2}) - \mathbf{b}_{k+1}^2(\mathbf{b}_k\cdot\mathbf{b}_{k+2})}{\sqrt{\left[\mathbf{b}_k^2\mathbf{b}_{k+1}^2 - (\mathbf{b}_k\cdot\mathbf{b}_{k+1})^2\right]\left[\mathbf{b}_{k+1}^2\mathbf{b}_{k+2}^2 - (\mathbf{b}_{k+1}\cdot\mathbf{b}_{k+2})^2\right]}}\ . \qquad (5.7)
\end{aligned}
$$

Obviously (5.7) and (5.5) are analytically equal. Astonishingly, in [50] the authors pass over the normalisation of the two projection vectors for some undocumented reason and thus obtain a much simpler expression. Since the projection vectors do definitely *not* have unit length except from very specific cases, even if all bond vectors are fixed to unit length (that would be if the denominator would trivially get 1 by introducing $b_i \equiv 1$ – which is not the case!), it is neither trivial nor obvious why it should be possible to calculate $\cos\phi$ without doing the normalisation.

The combination of (5.3) and (5.5) or (5.7), respectively, leads to the final result:

$$
\begin{aligned}
V_{\text{tors}}(\mathbf{R}) &= \frac{1}{2}\sum_{k=1}^{N-3}\left(1+\cos 3\phi_k\right) = \frac{1}{2}\sum_{k=1}^{N-3}\left(1 + \left(4\cos^2\phi_k - 3\right)\cos\phi_k\right)\ , \\
\cos\phi_k &= \frac{(\mathbf{b}_k\cdot\mathbf{b}_{k+1})(\mathbf{b}_{k+1}\cdot\mathbf{b}_{k+2}) - \mathbf{b}_{k+1}^2(\mathbf{b}_k\cdot\mathbf{b}_{k+2})}{\sqrt{\left[\mathbf{b}_k^2\mathbf{b}_{k+1}^2 - (\mathbf{b}_k\cdot\mathbf{b}_{k+1})^2\right]\left[\mathbf{b}_{k+1}^2\mathbf{b}_{k+2}^2 - (\mathbf{b}_{k+1}\cdot\mathbf{b}_{k+2})^2\right]}}\ . \qquad (5.8)
\end{aligned}
$$

The transcription of the bond vectors in differences of $\mathbf{r}_i$ is foregone. Recalling the definition of the algebraic sign for ϕ connected with Fig. 5.3, it is now possible to give a closed expression for ϕ:

$$
\begin{aligned}
\phi = {} & \operatorname{sign}\left(\left(\mathbf{b}_k\times\mathbf{b}_{k+1}\right)\cdot\mathbf{b}_{k+2}\right) \\
& \times \arccos\left(\frac{(\mathbf{b}_k\cdot\mathbf{b}_{k+1})(\mathbf{b}_{k+1}\cdot\mathbf{b}_{k+2}) - \mathbf{b}_{k+1}^2(\mathbf{b}_k\cdot\mathbf{b}_{k+2})}{\sqrt{\left[\mathbf{b}_k^2\mathbf{b}_{k+1}^2 - (\mathbf{b}_k\cdot\mathbf{b}_{k+1})^2\right]\left[\mathbf{b}_{k+1}^2\mathbf{b}_{k+2}^2 - (\mathbf{b}_{k+1}\cdot\mathbf{b}_{k+2})^2\right]}}\right)\ . \qquad (5.9)
\end{aligned}
$$

Here “sign” means the signum function:

$$
\operatorname{sign}(x) : \mathbb{R}\to\mathbb{R} \qquad \operatorname{sign}(x) = \begin{cases} -1 & \text{when } x < 0 \\ 0 & \text{when } x = 0 \\ 1 & \text{when } x > 0 \end{cases}\ . \qquad (5.10)
$$

5.1.2 Derivation of the Torsional Force

Before going into the systematic details, the pure torsion force arising from the gradient of V_{tors} in (5.8) shall be calculated. Several expressions will be used for abbreviation in the following, to make the formulae clearer. In (5.8) the cosine of the torsional angle ϕ_k is already defined in cartesian coordinates, especially in terms of $\mathbf{b}_i$. The numerator of $\cos\phi_k$ according to this formulation will be marked as ν_k, while the denominator will be called ρ_k for a certain reason which will be explained later:

$$\nu_k = (\mathbf{b}_k \cdot \mathbf{b}_{k+1})(\mathbf{b}_{k+1} \cdot \mathbf{b}_{k+2}) - \mathbf{b}_{k+1}^2(\mathbf{b}_k \cdot \mathbf{b}_{k+2}) \ , \tag{5.11}$$

$$\rho_k = \sqrt{\left[\mathbf{b}_k^2\mathbf{b}_{k+1}^2 - (\mathbf{b}_k \cdot \mathbf{b}_{k+1})^2\right]\left[\mathbf{b}_{k+1}^2\mathbf{b}_{k+2}^2 - (\mathbf{b}_{k+1} \cdot \mathbf{b}_{k+2})^2\right]} \ . \tag{5.12}$$

As a further abbreviation $\xi_{1,k}$ and $\xi_{2,k}$ shall be introduced as the two factors in the square root of ρ_k:

$$\xi_{1,k} = \left[\mathbf{b}_k^2\mathbf{b}_{k+1}^2 - (\mathbf{b}_k \cdot \mathbf{b}_{k+1})^2\right] \ , \qquad \xi_{2,k} = \left[\mathbf{b}_{k+1}^2\mathbf{b}_{k+2}^2 - (\mathbf{b}_{k+1} \cdot \mathbf{b}_{k+2})^2\right] \ . \tag{5.13}$$

The differentiation of the “outer” part of V_{tors} (an expression equivalent to one summand in the first line of (5.8)) is quite simple:

$$-\nabla_{\mathbf{r}_l}\left[\frac{1}{2}\left(1 + (4\cos^3\phi_k - 3\cos\phi_k)\right)\right] = -\frac{1}{2}\left(12\cos^2\phi_k - 3\right)\left(\nabla_{\mathbf{r}_l}\cos\phi_k\right) \ . \tag{5.14}$$

Large expressions arise, when processing $\nabla\cos\phi$ in the cartesian representation:

$$\begin{align}
\nabla_{\mathbf{r}_l}(\cos\phi_k) &= \nabla_{\mathbf{r}_l}\left(\nu_k\,\rho_k^{-1}\right) = \rho_k^{-1}\nabla_{\mathbf{r}_l}(\nu_k) - \nu_k\rho_k^{-2}\nabla_{\mathbf{r}_l}(\rho_k) \ , \tag{5.15}\\
\nabla_{\mathbf{r}_k}(\nu_k) &= -\mathbf{b}_{k+1}(\mathbf{b}_{k+1} \cdot \mathbf{b}_{k+2}) - \mathbf{b}_{k+1}^2(-\mathbf{b}_{k+2}) = \mathcal{N}_{0,k} \ , \tag{5.16}\\
\nabla_{\mathbf{r}_{k+3}}(\nu_k) &= (\mathbf{b}_k \cdot \mathbf{b}_{k+1})\mathbf{b}_{k+1} - \mathbf{b}_{k+1}^2\mathbf{b}_k = \mathcal{N}_{3,k} \ , \tag{5.17}\\
\nabla_{\mathbf{r}_{k+1}}(\nu_k) &= (\mathbf{b}_{k+1} - \mathbf{b}_k)(\mathbf{b}_{k+1} \cdot \mathbf{b}_{k+2}) + (\mathbf{b}_k \cdot \mathbf{b}_{k+1})(-\mathbf{b}_{k+2}) \notag\\
&\quad - \left[(-2\mathbf{b}_{k+1})(\mathbf{b}_k \cdot \mathbf{b}_{k+2}) + \mathbf{b}_{k+1}^2(\mathbf{b}_{k+2})\right] \notag\\
&= -\mathcal{N}_{0,k}\underbrace{-\mathbf{b}_k(\mathbf{b}_{k+1} \cdot \mathbf{b}_{k+2}) - \mathbf{b}_{k+2}(\mathbf{b}_k \cdot \mathbf{b}_{k+1}) + 2\mathbf{b}_{k+1}(\mathbf{b}_k \cdot \mathbf{b}_{k+2})}_{=\mathcal{N}_{12,k}} \ , \tag{5.18}\\
\nabla_{\mathbf{r}_{k+2}}(\nu_k) &= \mathbf{b}_k(\mathbf{b}_{k+1} \cdot \mathbf{b}_{k+2}) + (\mathbf{b}_k \cdot \mathbf{b}_{k+1})(\mathbf{b}_{k+2} - \mathbf{b}_{k+1}) \notag\\
&\quad - \left[2\mathbf{b}_{k+1}(\mathbf{b}_k \cdot \mathbf{b}_{k+2}) + \mathbf{b}_{k+1}^2(-\mathbf{b}_k)\right] \notag\\
&= -\mathcal{N}_{3,k} + \mathbf{b}_k(\mathbf{b}_{k+1} \cdot \mathbf{b}_{k+2}) + \mathbf{b}_{k+2}(\mathbf{b}_k \cdot \mathbf{b}_{k+1}) - 2\mathbf{b}_{k+1}(\mathbf{b}_k \cdot \mathbf{b}_{k+2}) \notag\\
&= -\mathcal{N}_{3,k} - \mathcal{N}_{12,k} \ , \tag{5.19}\\
\nabla_{\mathbf{r}_k}(\rho_k) &= \frac{1}{2\rho_k}\xi_{2,k}\left[-2\mathbf{b}_k\mathbf{b}_{k+1}^2 - 2(\mathbf{b}_k \cdot \mathbf{b}_{k+1})(-\mathbf{b}_{k+1})\right] \notag\\
&= \frac{1}{\rho_k}\xi_{2,k}\left[\mathbf{b}_{k+1}(\mathbf{b}_k \cdot \mathbf{b}_{k+1}) - \mathbf{b}_k\mathbf{b}_{k+1}^2\right] = \rho_k^{-1}\mathcal{D}_{0,k} \ , \tag{5.20}\\
\nabla_{\mathbf{r}_{k+3}}(\rho_k) &= \frac{1}{2\rho_k}\xi_{1,k}\left[\mathbf{b}_{k+1}^2(2\mathbf{b}_{k+2}) - 2(\mathbf{b}_{k+1} \cdot \mathbf{b}_{k+2})(\mathbf{b}_{k+1})\right] \notag\\
&= \frac{1}{\rho_k}\xi_{1,k}\left[-\mathbf{b}_{k+1}(\mathbf{b}_{k+1} \cdot \mathbf{b}_{k+2}) + \mathbf{b}_{k+2}\mathbf{b}_{k+1}^2\right] = \rho_k^{-1}\mathcal{D}_{3,k} \ , \tag{5.21}
\end{align}$$

$$\begin{aligned}\nabla_{\mathbf{r}_{k+1}}(\rho_k) &= \frac{1}{2\rho_k}\Big(\xi_{2,k}\left[2\mathbf{b}_k\mathbf{b}_{k+1}^2 + \mathbf{b}_k^2(-2\mathbf{b}_{k+1}) - 2(\mathbf{b}_k\cdot\mathbf{b}_{k+1})(\mathbf{b}_{k+1}-\mathbf{b}_k)\right] \\ &\qquad + \xi_{1,k}\left[-2\mathbf{b}_{k+1}\mathbf{b}_{k+2}^2 - 2(\mathbf{b}_{k+1}\cdot\mathbf{b}_{k+2})(-\mathbf{b}_{k+2})\right]\Big) \\ &= -\frac{\mathcal{D}_{0,k}}{\rho_k} \\ &\qquad + \frac{1}{\rho_k}\underbrace{\Big(\xi_{2,k}(\mathbf{b}_k(\mathbf{b}_k\cdot\mathbf{b}_{k+1}) - \mathbf{b}_{k+1}\mathbf{b}_k^2) + \xi_{1,k}(\mathbf{b}_{k+2}(\mathbf{b}_{k+1}\cdot\mathbf{b}_{k+2}) - \mathbf{b}_{k+1}\mathbf{b}_{k+2}^2)\Big)}_{=\mathcal{D}_{12,k}} \\ &= \rho_k^{-1}(-\mathcal{D}_{0,k} + \mathcal{D}_{12,k}) \ ,\end{aligned} \tag{5.22}$$

$$\begin{aligned}\nabla_{\mathbf{r}_{k+2}}(\rho_k) &= \frac{1}{2\rho_k}\Big(\xi_{2,k}\left[\mathbf{b}_k^2(2\mathbf{b}_{k+1}) - 2(\mathbf{b}_k\cdot\mathbf{b}_{k+1})(\mathbf{b}_k)\right] \\ &\qquad + \xi_{1,k}\left[2\mathbf{b}_{k+1}\mathbf{b}_{k+2}^2 + \mathbf{b}_{k+1}^2(-2\mathbf{b}_{k+2}) - 2(\mathbf{b}_{k+1}\cdot\mathbf{b}_{k+2})(\mathbf{b}_{k+2}-\mathbf{b}_{k+1})\right]\Big) \\ &= -\frac{\mathcal{D}_{3,k}}{\rho_k} + \frac{1}{\rho_k}\Big(\xi_{2,k}(-\mathbf{b}_k(\mathbf{b}_k\cdot\mathbf{b}_{k+1}) + \mathbf{b}_{k+1}\mathbf{b}_k^2) \\ &\qquad\qquad + \xi_{1,k}(-\mathbf{b}_{k+2}(\mathbf{b}_{k+1}\cdot\mathbf{b}_{k+2}) + \mathbf{b}_{k+1}\mathbf{b}_{k+2}^2)\Big) \\ &= \rho_k^{-1}(-\mathcal{D}_{3,k} - \mathcal{D}_{12,k}) \ .\end{aligned} \tag{5.23}$$

Equations (5.16) – (5.23) show that the gradient of $\cos\phi_k$ with respect to all included monomers $i \in [k, k+3]$ can be described by three terms, which appear with both positive and negative algebraic signs and thus fulfil Newton's law of *"Actio=Reactio"*. These terms are a combination of $\mathcal{N}_{0,12,3}$ and $\mathcal{D}_{0,12,3}$ according to (5.15), whose definitions are given in (5.16) – (5.18) and (5.20) – (5.22):

$$\frac{\mathcal{N}_{l,k}}{\rho_k} - \frac{\nu_k\mathcal{D}_{l,k}}{\rho_k^3} \equiv \mathcal{C}_{l,k} \ . \tag{5.24}$$

With this knowledge the whole torsional force acting on monomer i can be expressed in the previously defined abbreviations:

$$\begin{aligned}\mathbf{F}_{\mathrm{tors}\,i} = -\frac{1}{2}\Big[&\underbrace{(12\cos^2\phi_i - 3)\,\mathcal{C}_{0,i}}_{i\le N-3} + \underbrace{(12\cos^2\phi_{i-1} - 3)\,(-\mathcal{C}_{0,i-1} + \mathcal{C}_{12,i-1})}_{i\in[2,N-2]} \\ &+ \underbrace{(12\cos^2\phi_{i-2} - 3)\,(-\mathcal{C}_{3,i-2} - \mathcal{C}_{12,i-2})}_{i\in[3,N-1]} + \underbrace{(12\cos^2\phi_{i-3} - 3)\,\mathcal{C}_{3,i-3}}_{i\ge 4}\Big] \ .\end{aligned} \tag{5.25}$$

5.2 Monte Carlo Simulations

As before for the AB model, parallel tempering Monte Carlo simulations are carried out to obtain reliable data to compare with the Molecular Dynamics results. The set of parameters is left completely unchanged. The simulations are observed with the same number of replica (temperatures), the same number of sweeps and all the error analysis is done fully analogous to section 4.1. The quantities of interest are again the potential energy E, the end-to-end-distance R_{ee}, the radius of gyration R_{gyr} as well as the respective fluctuation quantities. The most important question is to observe the impact of the additional potential term on

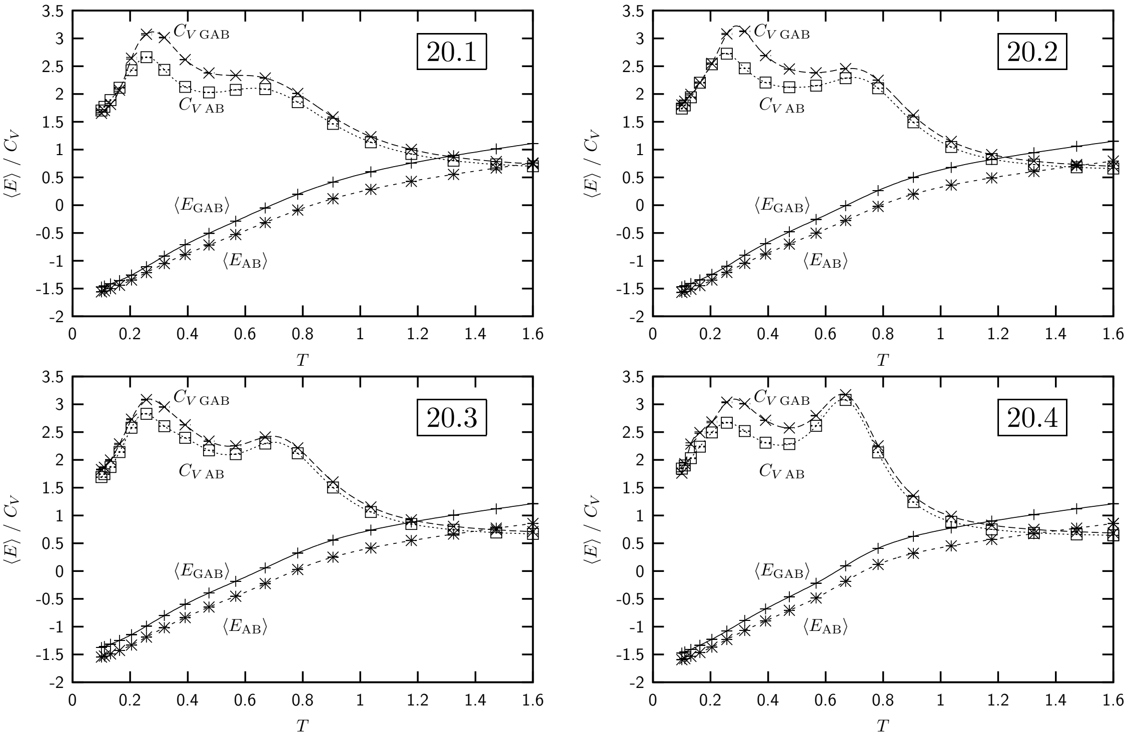


Figure 5.4: Potential energy and heat capacities of the GAB model, represented by the solid and the long-dashed line, compared to the AB model, where the same quantities are plotted with medium and short dashes respectively. The lines are obtained by multiple histogram reweighting (compare section 3.2.2), while the error bars result from the Jackknifing analysis of the measured data. The sequences are as follows: (a) 20.1, (b) 20.2, (c) 20.3, (d) 20.4.

the results obtained for the AB model. Therefore, all results for the GAB model will be immediately compared to the AB model.

General Thermodynamics

As in section 4.1, the system is first observed with bond strength $\alpha_r = 50$. In Fig. 5.4 the potential energy and the heat capacity is plotted for both the GAB and the AB model. As it is to be expected, the energy is systematically higher for the GAB model, since the additional torsion potential gives always a positive contribution if any. For the heat capacity, the change is less obvious. But it seems that it is also higher in the GAB case, except for very low temperatures. Furthermore, the introduction of the torsion term seems to stress the first peak of the specific heat. This might be an indicator for a higher potential energy barrier between the ground-state like structures and globular configurations, supposed the peaks can be understood as it is discussed earlier in section 4.1. Also, the first peak of the heat capacity is slightly shifted to higher temperatures by the introduction of the torsion potential. This could signify that the ground-state structures of the GAB model are more resistant against thermal heating, which is analogous to the conclusion from the increased height of the peak. On the other hand, the potential energy of the ground state is less negative, which would lead to the assumption that it is less stable and the peaks of the specific heat move to lower

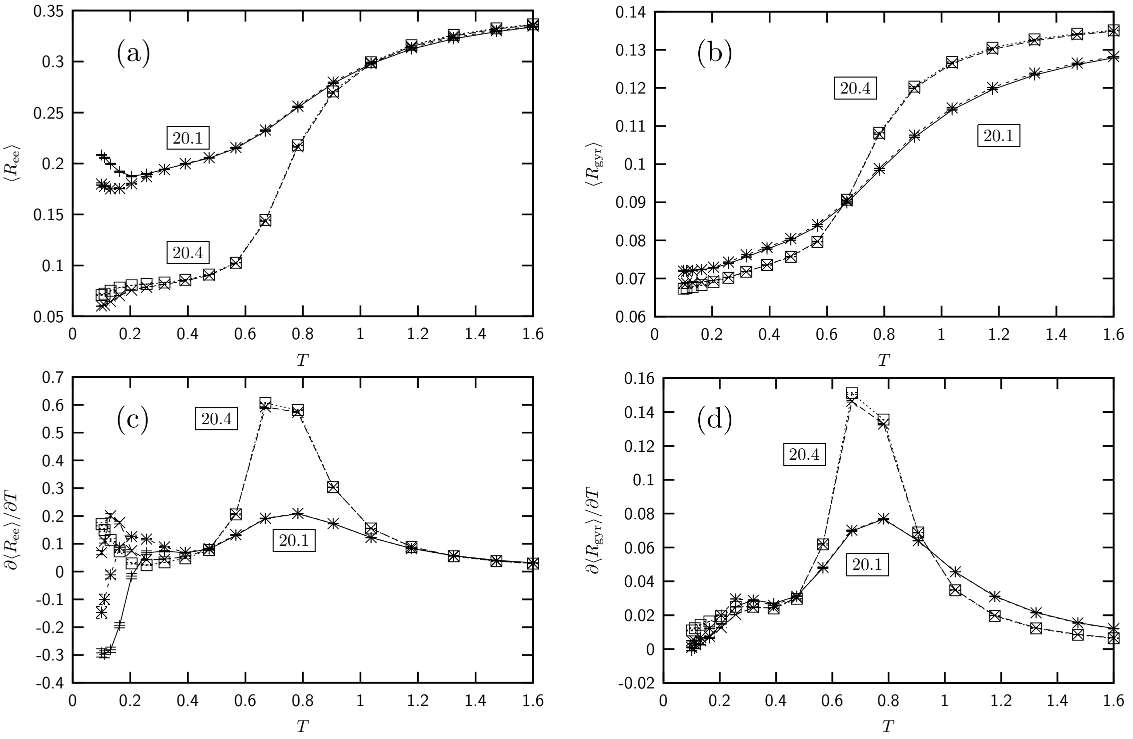


Figure 5.5: It is obvious from the pictures that the plots for the AB and GAB model can be hardly distinguished. The vertical dashes and and diagonal crosses belong to the latter, while the star and hollow square symbols represent the results from the AB model. The assignment of the respective sequences 20.1 and 20.4 is easier and therefore given within the figures. The plotted quantities are: (a) End-to-end-distance, (b) radius of gyration, (c) fluctuation of the end-to-end-distance, (d) fluctuation of the radius of gyration.

temperatures. Obviously this thought is misleading.

Structural Behaviour

Since for protein folding the structural behaviour of a system is most important, it is interesting, how the GAB model differs from the AB model considering quantities like the end-to-end-distance or the radius of gyration. Figure 5.5 shows the respective plots for sequences 20.1 and 20.4. It is very remarkable that, except for low temperatures, the two considered, structurally meaningful quantities as well as the fluctuations do scarcely differ for the two models. Especially considering the radius of gyration, the deviations are minimal. For low temperatures, the ground-state of a sequence is dominating the ensemble of configurations. Therefore, the torsion potential does only take effect on the ground-state of a sequence, but not on the structural behaviour for higher temperatures. Table 5.I gives the results of the comparison between the ground-states of the AB and GAB model for several sequences. This emphasises the assumption that the torsion potential is crucial for the structure of the ground-state. While the overlap in the range of 0.8 suggests an acceptable uniformity, the root mean square deviation clearly shows that the found ground-state structures differ noticeably.

Figure 5.6 shows the ground-state for sequence 20.2, where both the overlap parameter

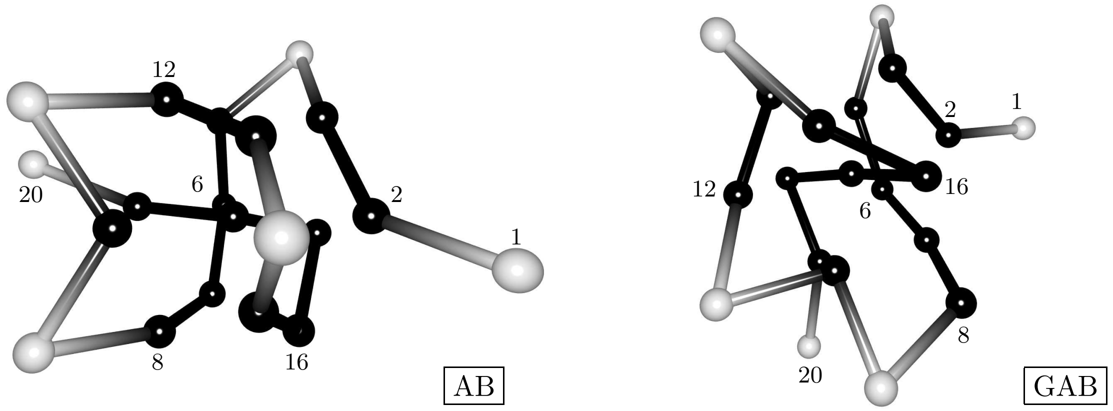


Figure 5.6: The ground-state of sequence 20.2 for (a) the AB and (b) the GAB model.

and the rmsd have a nearly optimal value, with respect to the examined cases. To make the identification easier, some of the monomers are numbered. The loop between monomer number 2 and 6, as well as the structure between monomer 8 and 12 are obviously very similar for both models. The main difference is located at the tail of the chain, where the last four monomers (17-20) are nearly lying in one line in the AB model, but not in the GAB case. This is also obvious from the contact maps in Fig. 5.7, where monomer number 18 has most contacts in the AB case, and monomer number 17 in the GAB case. I.e., these respective monomers reside in the centre of the ground-state structure. The rest of the contact maps undergoes only minor changes. Also, in the GAB ground-state the first monomer (type B) has only repulsive Lennard-Jones interactions. Therefore it directs straight away from the core of the structure, which is not the case in the AB model. There it has an attractive interaction with monomer number 14.

Although deviations like the two explained ones seem to be negligible, they can have a considerable influence especially on the end-to-end-distance because they are both located in the beginning and the end of the chain. This makes it more believable that the deviations can be especially seen when considering R_{ee}. However, the torsion term seems to quickly lose importance when going to higher temperatures.

Table 5.I: Overlap and root mean square deviation (compare section 3.3.3) of ground-states in the AB and GAB model.

Sequence	20.1	20.2	20.3	20.4
Overlap	0.71	0.83	0.73	0.78
rmsd	1.12	0.89	0.84	1.20

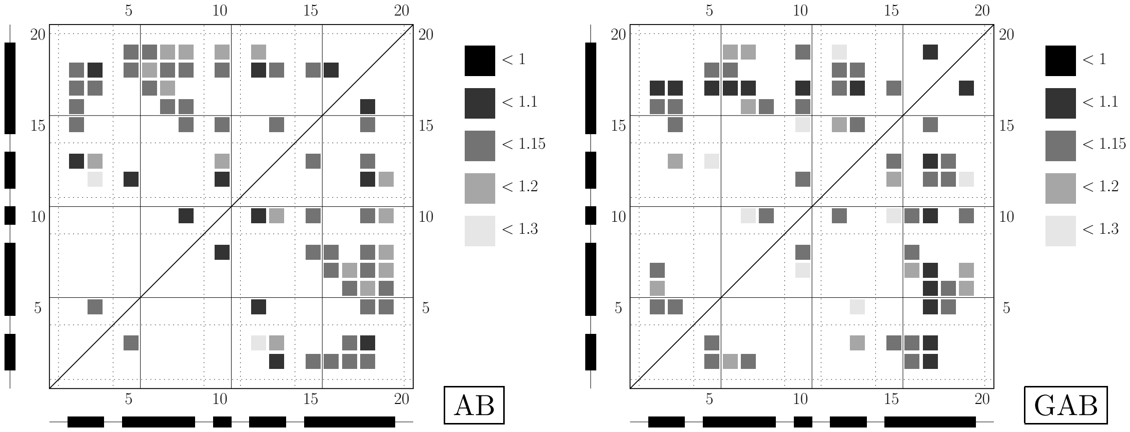


Figure 5.7: Contact maps of sequence 20.2 for (a) the AB and (b) the GAB model. For a description of contact maps see Fig. 4.14.

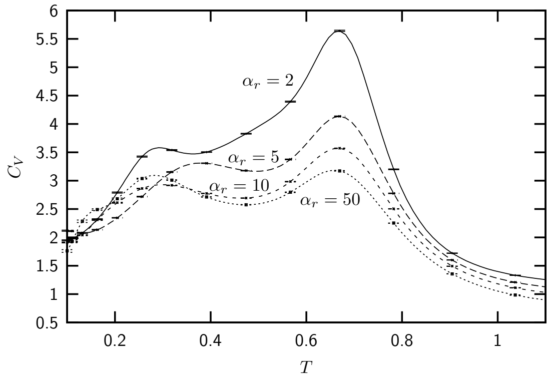


Figure 5.8: Heat capacities of sequence 20.4 for different bond strength α_r in the GAB model (i.e. with torsion).

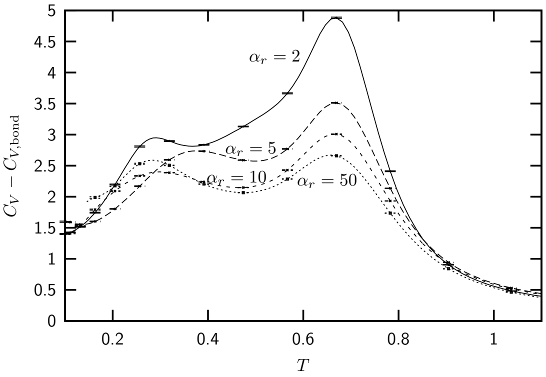


Figure 5.9: To clarify that the observed effects do not qualitatively result from the bond term, the analytic effect of V_{bond} as given in (1.9) is subtracted. The different heights of the tails in Fig. 5.8 are obviously only caused by the larger contributions of $C_{V,\text{bond}}$ for weaker bonds.

Impact of Different Bond Strengths

The observations for different bond strengths α_r can be shortened here. The outcome is shown in Figs. 5.8, 5.9 and 5.10. It is easy to see that everything is very similar to the results from section 4.1.3. So all the detailed considerations concerning reasons for the specific appearance can be adapted analogously. The only remarkable point is that in Fig. 5.10 (b) the first peak of the specific heat is growing with larger α_r, which is not the case in the AB model (see Fig. 4.11 (b)). However, since the behaviour cannot be generally explained in that details, it can be taken as an effect of the several formed monomer pairs and its stepwise breakup for growing α_r as explained in the already mentioned section 4.1.3.

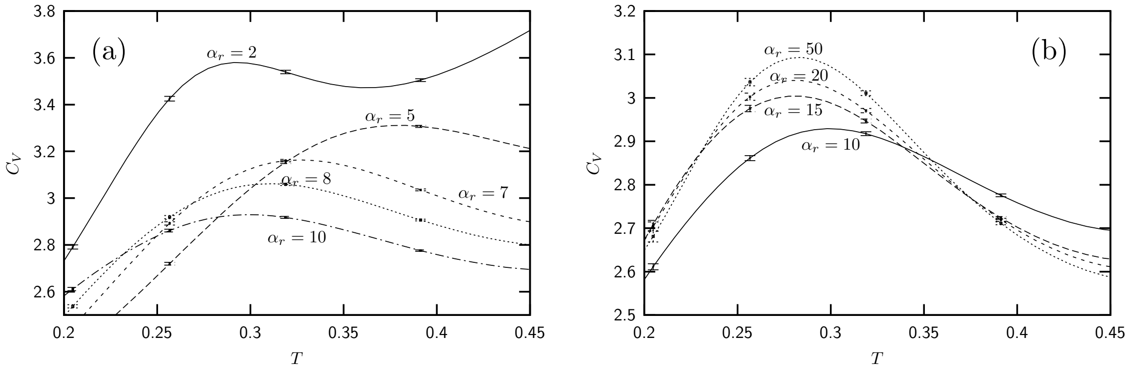


Figure 5.10: Closer look on the first peak of the heat capacities of sequence 20.4 for different bond strength α_r in the GAB model. (a) Low bond strength, (b) higher bond strength.

5.3 Realizing and Tackling the Problem in Molecular Dynamics

After including the torsion potential in the already existing Molecular Dynamics program it turned out that the theoretically conserved energy $\mathcal{H}_{\mathrm{NHC}}$ (see (2.71)) is fluctuating rather strongly, whereas in a simulation without the torsion potential the fluctuations of this quantity had been very small. More precisely $\mathcal{H}_{\mathrm{NHC}}$ seems to jump from time to time during the simulation, if the torsion potential is switched on.

To study this in detail, the system is simplified as much as possible. So the homo-four-mer (sequence AAAA) is chosen. In Fig. 5.11 this behaviour is plotted for several trial runs with different parameters. It is easy to see from this picture that elongating the Nosé-Hoover-Chain does not seem to solve the problem. However for the simulation with a smaller time step, the deviations in $\mathcal{H}_{\mathrm{NHC}}$ seem to be less crucial. So it is an obvious consideration, whether the torsion potential implies some fluctuations at a higher frequency as the bond length fluctuations, wherefore the time step and the Nosé-Hoover coupling would have to be adapted. In Fig. 5.12 it can be seen that this is obviously not the case. After the expected peak in the frequency spectrum at the bond length fluctuations, the frequency spectra of the simulations with as well as without the torsion potential are quickly decreasing.

After double-checking the correctness of the force calculations and the implementation it shall be checked, whether the problem is a numerical error or some problem with the Nosé-Hoover thermostating. Therefore it is crucial to find out, whether the deviations are really "jumps" in that sense that the conserved energy is more or less constant over a long-time and then, in one time step, is changed dramatically. Figures 5.13 and 5.14 show the result of a closer look at the time series. Indeed the deviations turn out to be events that arise from one single time step. The whole problem is initiated by a sudden rise in the kinetic energy of the system. E_{cons} is increased by the same amount of energy. After that time step E_{cons} stays constant again, apart from very small numerical fluctuations. The rest of the total system included the thermostat relaxes quickly by asymptotically absorbing the additional amount of energy in the potential energy of the Nosé-Hoover thermostat. That makes sense, since the latter is the only type of energy that is free of any requirements or expectations. Whereas the kinetic energies of both the classical system and the Nosé-Hoover thermostat

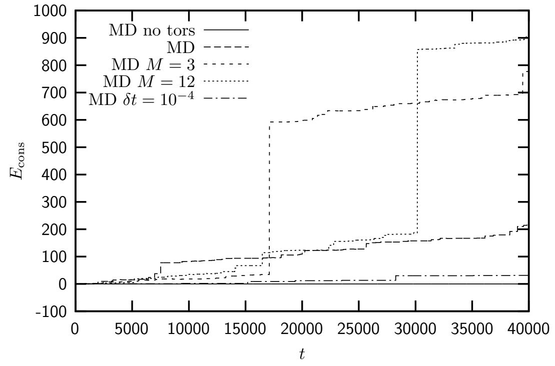


Figure 5.11: Trajectories of the quantity $\mathcal{H}_{\text{NHC}}$ for several trial simulations of the homo-four-mer (sequence AAAA). The solid line is obtained while simulating without torsion energy. The dashed lines are simulations with the typical time step $\delta t = 10^{-3}$ and different lengths of the Nosé-Hoover-Chain: $M = 2$ is long-dashed, $M = 3$ is medium dashed and $M = 12$ (massive thermostating: one thermostat for every degree of freedom of the system) is the short-dashed line. Finally another trial is done with $M = 2$ and a smaller time step $\delta t = 10^{-4}$, plotted with the chain dotted line.

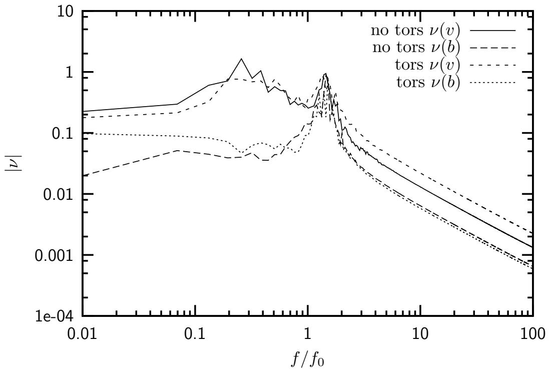


Figure 5.12: Here frequency spectra are plotted double logarithmic. The solid and long-dashed lines are the spectra obtained by measuring the velocity and bond length fluctuations respectively without the torsion potential. The medium dashed and short-dashed lines are the velocity and bond length fluctuations including the torsion potential. The frequencies are again normalised to the expected bond length fluctuations $f_0 = 2\pi\sqrt{m/(2\alpha_r)}$. There is obviously no qualitative difference, especially for high frequencies, which would explain the need of a smaller time step when simulating the model with torsion potential.

are guided by temperature and the type of the observed system guides the potential energy of the stand-alone system of course.

Figure 5.14 shows exemplary trajectories of velocity components of the homo-four-mer at the time, where the energy jump is observed. Obviously for some reason the velocities are changed drastically within the respective time step. Figure 5.15 shows the configuration one time step before the jump. To exclude the possibility that the problem is caused by numerical precision, the forces acting on the respective configuration are calculated with Mathematica and compared to the results from the simulation program. The relative deviations are about 10^{-4}, but it is remarkable that the absolute value of the torsional force is about three orders of magnitude higher than the expected values from the other potential contributions. This finally leads to the correct conclusion: The torsional force is *divergent* for the case shown in Fig. 5.15.

5.3.1 Explanation and Deeper Analysis of the Divergence

The divergence is not obvious at first sight. The torsion potential is very well behaved for any configuration. From (5.8) it is clear that $V_{\text{tors}}(\mathbf{R}) \in [0, 1]\,\forall\,\mathbf{R}$. The divergence must be somehow correlated with the property that three monomers nearly form a straight line. No matter, which of the two introduced definitions of the torsion angle is used, the angle between the two planes spanned by $\mathbf{b}_2$ and $\mathbf{b}_1$ or $\mathbf{b}_3$ respectively is measured (compare Figs.

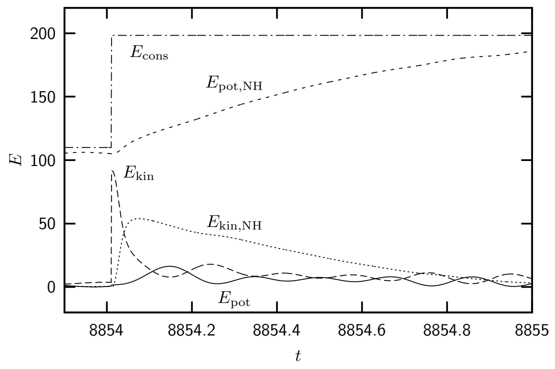


Figure 5.13: Time series of the theoretically conserved energy (E_{cons}, the chain dotted line) and the contributions of the potential and kinetic energy of the stand-alone system, plotted with a solid and long-dashed line, as well as the potential and kinetic energy of the Nosé-Hoover thermostat, visualised by the medium and short-dashed line, respectively.

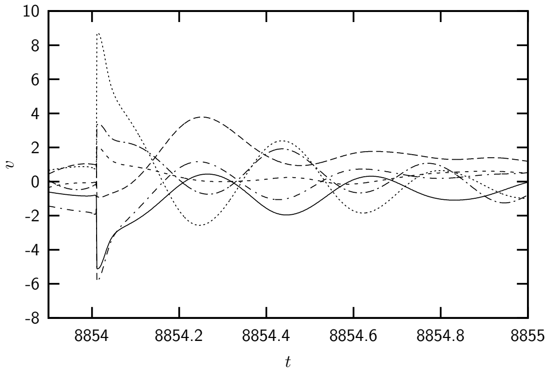


Figure 5.14: Time series of the x, y and z velocity components of two monomers. The only important information is the velocity jump at $t \approx 8854$, so a detailed key is not instructive.

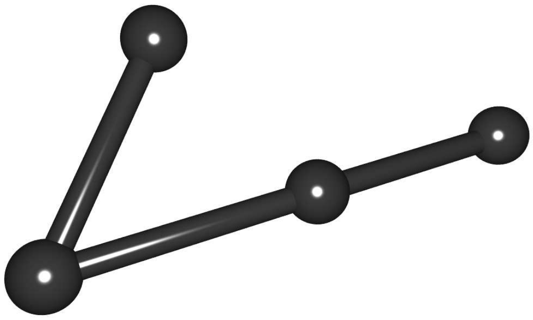


Figure 5.15: After reaching this somehow *critical* configuration of the homo-four-mer, the velocities undergo a sudden drastic change which leads to a jump in the kinetic energy and – which is more noteworthy – in the theoretically conserved energy of the total system. The only remarkable property of the configuration is, that three monomers nearly form a straight line, i.e., the middle and one of the two other bond vectors are nearly parallel.

5.1 and 5.2). Figure 5.16 shall help to describe the critical case. Assume that $\mathbf{b}_1$ and $\mathbf{b}_2$ are nearly parallel. In this case, $\mathbf{b}_2$ and $\mathbf{b}_3$ span a plane that is relatively stable against small movements of the three included monomers. Whereas, the plane spanned by $\mathbf{b}_1$ and $\mathbf{b}_2$ could easily rotate around the two included bond vectors by disturbing the position of one of the three monomers a little bit. The angle of such a rotation would be directly reflected in the torsion angle. Recalling the definition of the torsion potential (5.1) this means that a small movement of one of the three considered monomers would be able to drastically change the torsion potential. In a formal writing, a force $\mathbf{F}$ is defined as the change of the potential energy ΔV over a length – the change of position Δx: $\mathbf{F} = -\Delta V/\Delta x$. Although $\Delta V_{\mathrm{tors}} \leq 1$,

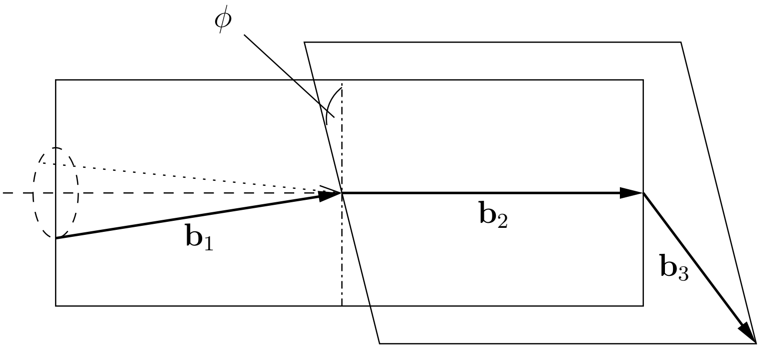


Figure 5.16: In case $\mathbf{b}_2$ and one of the other bond vectors (here $\mathbf{b}_1$) is nearly parallel, a small movement of one monomer can lead to a complete change of the torsion potential energy. The rotation of the first monomer of $\mathbf{b}_1$ around the small dashed circle is completely responsible for the torsion angle of the configuration in the figure.

Δx can get very small as described above, which will lead to large forces $\mathbf{F}$.

A closer look at (5.8) shows that for $\mathbf{b}_1 \approx \mathbf{b}_2$ the denominator of $\cos\phi$ vanishes (assuming $k = 1$):

$$\begin{aligned}\rho_1 &= \sqrt{\left[\mathbf{b}_1^2\mathbf{b}_2^2 - (\mathbf{b}_1 \cdot \mathbf{b}_2)^2\right]\left[\mathbf{b}_2^2\mathbf{b}_3^2 - (\mathbf{b}_2 \cdot \mathbf{b}_3)^2\right]} \\ &\approx \sqrt{\left[\mathbf{b}_2^2\mathbf{b}_2^2 - (\mathbf{b}_2 \cdot \mathbf{b}_2)^2\right]\left[\mathbf{b}_2^2\mathbf{b}_3^2 - (\mathbf{b}_2 \cdot \mathbf{b}_3)^2\right]} = 0\ . \end{aligned} \tag{5.26}$$

Of course the expression for $\cos\phi$ cannot diverge. On the other hand, the numerator also converges against 0 in the considered case, which countervails the vanishing denominator $\cos\phi$:

$$(\mathbf{b}_1 \cdot \mathbf{b}_2)(\mathbf{b}_2 \cdot \mathbf{b}_3) - \mathbf{b}_2^2(\mathbf{b}_1 \cdot \mathbf{b}_3) \approx (\mathbf{b}_2 \cdot \mathbf{b}_2)(\mathbf{b}_2 \cdot \mathbf{b}_3) - \mathbf{b}_2^2(\mathbf{b}_2 \cdot \mathbf{b}_3) = 0\ . \tag{5.27}$$

The denominator of $\cos\phi_k$ which is shown to be vanishing in (5.26), was previously denoted as ρ_k. From (5.24) it is obvious that indeed ρ_k appears in the denominator of summands of the torsion force, which would explain the divergence. However, the structures of the various numerators of the summands are too complex to proof the general type of divergence. Thus the force for a simplified case shall be calculated exemplarily. Similar to Fig. 5.16, the three bond vectors are chosen as follows:

$$\mathbf{b}_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} - \begin{pmatrix} r\cos\varphi \\ r\sin\varphi \\ 0 \end{pmatrix} = \begin{pmatrix} -r\cos\varphi \\ -r\sin\varphi \\ 1 \end{pmatrix}\ , \qquad \mathbf{b}_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}\ , \qquad \mathbf{b}_3 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}\ . \tag{5.28}$$

This makes $\mathbf{b}_2$ and $\mathbf{b}_3$ perpendicular, spanning a plane that is stable against small distortions of one of the included monomers. $\mathbf{b}_1$ is parallel to $\mathbf{b}_2$ except for a small deviation with a distance r from the extension of $\mathbf{b}_2$ and some freely selectable angle φ. Before calculating the force acting on the first monomer $\mathbf{F}_{\text{tors}\,1}$ according to (5.25), some of the needed quantities shall be determined:

$$\rho_1 = \sqrt{\xi_{1,1} \cdot \xi_{2,1}} = \sqrt{(1 + r^2 - 1) \cdot (1 - 0)} = r \ , \tag{5.29}$$

$$\cos\phi = \frac{1 \cdot 0 - 1 \cdot (r\cos\varphi)}{\rho_1} = \frac{r\cos\varphi}{r} = \cos\varphi \ . \tag{5.30}$$

From (5.29) it gets clear that the denominator of $\cos\phi$, denoted with ρ earlier, is equal to the radius of the deviation r. (5.30) shows that in this special case, the torsion angle is exactly the same as the freely selectable displacement angle φ of the first monomer. This makes it more obvious that no matter how small r is chosen, the torsion potential depends completely on φ. For a very small r, a deviation in φ is a negligible displacement of the first monomer, which can lead to a potential energy change as big as possible within the domain:

$$\begin{aligned}
\mathcal{N}_{0,1} &= -\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \cdot 0 - 1 \cdot \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \mathbf{b}_3 \ , \\
\mathcal{D}_{0,1} &= \underbrace{\xi_{2,1}}_{=1} \left(\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \cdot 1 - \begin{pmatrix} -r\cos\varphi \\ -r\sin\varphi \\ 1 \end{pmatrix} \cdot 1 \right) = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} - \begin{pmatrix} -r\cos\varphi \\ -r\sin\varphi \\ 1 \end{pmatrix} = \mathbf{b}_2 - \mathbf{b}_1 \ , \\
\mathcal{C}_{0,1} &= \frac{r^2}{r^3}\mathbf{b}_3 - \frac{r\cos\varphi}{r^3}(\mathbf{b}_2 - \mathbf{b}_1) = r^{-3} \begin{pmatrix} r^2 - r\cos\varphi(r\cos\varphi) \\ 0 - r\cos\varphi(r\sin\varphi) \\ 0 - r\cos\varphi(1-1) \end{pmatrix} \\
&= r^{-3} \begin{pmatrix} r^2(1 - \cos^2\varphi) \\ -r^2\cos\varphi\sin\varphi \\ 0 \end{pmatrix} = r^{-1}\sin\varphi \begin{pmatrix} \sin\varphi \\ -\cos\varphi \\ 0 \end{pmatrix} \ , \\
\Rightarrow \qquad \mathbf{F}_{\text{tors}\,1} &= -\frac{1}{2}(12\cos^2\varphi\sin\varphi - 3\sin\varphi)\frac{1}{r} \begin{pmatrix} \sin\varphi \\ -\cos\varphi \\ 0 \end{pmatrix} \\
&= -\frac{3}{2}\sin 3\varphi \frac{1}{r} \begin{pmatrix} \sin\varphi \\ -\cos\varphi \\ 0 \end{pmatrix} \ .
\end{aligned} \tag{5.31}$$

Two things are remarkable about $\mathbf{F}_{\text{tors}\,1}$. First, the divergence is of first order with respect to r. Second, the force is acting perpendicular to the considered deviation from "parallelity". It just drives the monomer to a position, where the torsion potential energy is small by effectively changing φ – and thus the torsion angle ϕ. The radius r, which is responsible for the divergence, remains unchanged. What makes the situation explicitly bad is the fact that for example the bending energy favours configurations, where the bond vectors are as parallel as possible. A good example would be the simple sequence AABB. There even the Lennard-Jones potential acts exclusively repulsive, which will strongly favour straight configurations energetically.

Now it is also obvious, how the problem happens within the simulation. As soon as a configuration is reached, where two bond vectors are nearly parallel, the forces increase extensively. As shown above, the torsional forces are acting circular, around the middle bond vector of a torsion angle. But since the Molecular Dynamics simulation uses cartesian coordinates, during the next time step the velocities are crucially updated in the *tangential* direction. This causes the jump in the kinetic energy and is provoked by the fact that the torsion force cannot be represented in cartesian coordinates in a sufficient manner. As the update of the velocities is calculated by $\mathbf{F} \times \delta t$, the effect gets less important for smaller time steps, which is the reason why in Fig. 5.11 the simulation with a smaller time step behaves better.

5.3.2 Possible Solutions

There are three different ways to tackle the problem:

1. In the previous paragraph it is explained, why a smaller time step actually solves the appearing problems with the wrongly directed, strong force. A very easy trial is thus to make the time step dependent of the instantaneously strongest force. In doing so, the velocity update could be adjusted by some freely selectable upper limit. To conserve the correct canonical statistics, it would only be necessary to do all required statistical measurements in equidistant time segments. As mentioned earlier, the torsion force can easily get orders of magnitudes as strong as the rest of the force terms. To have a nearly constant velocity update, orders of magnitudes more time steps would be necessary in the respective configurational situation. For a sequence of considerable length this happens quite often. In conclusion, the whole simulation time could be easily raised by a factor of 10^3. Obviously this is not the method of choice.

2. As described in the previous section, the problem is caused by the fact that the simulation uses cartesian coordinates so far, but the torsion force would better be expressed in terms of angles of the configuration. Indeed, it is theoretically possible to describe a whole configuration by another set of degrees of freedom than the cartesian coordinates of every monomer. For N monomers, there are $3N$ cartesian coordinates. Considering $N-1$ bond lengths, $N-2$ bond angles and $N-3$ torsion angles, there is a lack of 6, which are of course the 3 translational and 3 rotational degrees of freedom with respect to the whole configuration.

   As explained in section 2.1, it is possible to split up the Hamiltonian of the system into several parts and do the integration of these parts independently of each other. This is for example utilised for the Nosé-Hoover thermostat, where the heat-bath particles are treated with higher order integration schemes, while the classical system is integrated by the Störmer-Verlet algorithm. So one thought is to separate the torsion part of the Hamiltonian from the rest of the forces of the model and integrate only the torsion part in angular degrees of freedom. Unfortunately, this is *not* possible, since the condition for such a separation is that the different parts are independent from each other. But a torsion angle is of course not independent from for example the r_{ij} distances, which are responsible for the Lennard-Jones contribution.

This conclusion means that the integration has to be carried out in one step and one fixed coordinate system. Therefore, all required quantities would have to be expressed in angular degrees of freedom. Of course bond lengths, bond angles and torsion angles *are* such degrees of freedom, thus the expression is trivial. But the already mentioned example of the distances r_{ij} of non-neighbouring monomers in terms of angles and bond lengths of the configuration only, is very complicated. Even worse, after finding a way of transcribing the potential, this construct would have to be differentiated. Thus, this is a probably very effective but extremely complicated way of getting around the respective difficulties.

3. In all-atom Molecular Dynamics simulations it is common to cut off the long-range Coulomb interactions smoothly by altering the $1/r$ behaviour by an exponential decay factor $\exp[-r/r_0]$. As proved above in (5.31) there is also a $1/r$ behaviour in the existing case – but for the force, not for the potential. Furthermore, the problem is not the long-range tail, but the divergence for very small r. Nevertheless, it is possible to introduce an additional factor $(1 - \exp[-r/r_0])$ in the torsion potential, which will be shown to fix the divergence of the force. Since this is the most practical method of overcoming the considered difficulties, it will be studied more thoroughly. Unfortunately, it will turn out that this way of solution implies other systematic problems concerning the structural behaviour of the system.

Modified Potential

In the example in section 5.3.1, the divergence is shown to be of first order with respect to some distance r of $\mathbf{b}_1$ from an axis parallel to $\mathbf{b}_2$. This deviation r is intrinsically chosen by the type of example. It is a value for the "parallelity" of two successive bond vectors. In case $r = 0$, the two bond vectors are exactly parallel. But how can r be determined for a general configuration? Any four successive monomers build up a torsion angle. The distance of the first or the last of these four monomers from the infinitely elongated second bond vector (the vector between the second and the third monomer), is such a parallelity value r. Another way of describing this would be: r is the length of the projection of the first or last bond vector ($\mathbf{b}_1$ or $\mathbf{b}_3$) to the plane perpendicular to $\mathbf{b}_2$. This reminds one of the second definition of the torsion angle in the beginning of the chapter. (5.6) gives the form of such a projection vector. The length can easily be obtained by calculating $\sqrt{\mathbf{b}_{k,\perp}^2}$:

$$\begin{aligned}\sqrt{\mathbf{b}_{k,\perp}^2} &= \sqrt{\left(\mathbf{b}_k - \left(\left(\mathbf{b}_k \cdot \mathbf{b}_{k+1}\right)\mathbf{b}_{k+1}\right) b_{k+1}^{-2}\right)^2} \\ &= \sqrt{\mathbf{b}_k^2 - 2\left(\mathbf{b}_k \cdot \mathbf{b}_{k+1}\right)^2 b_{k+1}^{-2} + \left(\mathbf{b}_k \cdot \mathbf{b}_{k+1}\right)^2 \mathbf{b}_{k+1}^2 b_{k+1}^{-4}} \\ &= \sqrt{\mathbf{b}_k^2 - \left(\mathbf{b}_k \cdot \mathbf{b}_{k+1}\right)^2 b_{k+1}^{-2}} = b_{k+1}^{-1}\sqrt{\mathbf{b}_k^2\mathbf{b}_{k+1}^2 - \left(\mathbf{b}_k \cdot \mathbf{b}_{k+1}\right)^2}\ . \end{aligned} \tag{5.32}$$

The b_{k+1}^{-1} factor is not crucial for the behaviour, since $b_i \approx 1\,\forall\, i \in \{1, N\}$. So what lasts as the important part of (5.32) is:

$$r_{k,\perp} = \sqrt{\mathbf{b}_k^2\mathbf{b}_{k+1}^2 - \left(\mathbf{b}_k \cdot \mathbf{b}_{k+1}\right)^2}\ . \tag{5.33}$$

But not only the case if $\mathbf{b}_k$ and $\mathbf{b}_{k+1}$ are parallel is critical for $\mathbf{F}_{\mathrm{tors}\,k}$, the same problem arises for $\mathbf{b}_{k+1} \parallel \mathbf{b}_{k+2}$. An analogous term can be used to evaluate $r_{k+2,\perp}$. For the extension of the potential it will be sufficient to control $r_k = r_{k,\perp} \cdot r_{k+2,\perp}$ and guarantee that the potential will vanish for small r_k. From (5.33) it is obvious that the form of r_k is exactly the same as the definition of ρ_k in (5.12). Thus, the nomenclature of the denominator of $\cos\phi$ as ρ_k – like a radius – is belatedly justified. ρ_k indicates how far the two outer monomers are away from the axis through the middle bond vector.

After these considerations, it is clear how the torsion potential from (5.8) has to be modified:

$$\begin{aligned} V_{\mathrm{tors}}(\mathbf{R}) &= \frac{1}{2}\sum_{k=1}^{N-3} \left(1+\cos 3\phi_k\right)\left(1-\exp\left[-\rho_k/r_0\right]\right) \\ &= \frac{1}{2}\sum_{k=1}^{N-3} \big(1+\left(4\cos^2\phi_k-3\right)\cos\phi_k\big)\left(1-\exp\left[-\rho_k/r_0\right]\right) \ . \end{aligned} \tag{5.34}$$

With r_0 the falloff of the potential for small ρ_k can be adjusted. For $\rho_k = r_0$ the additional factor is $1 - e^{-1} \approx 0.63$, i.e. the falloff gets effective for $\rho_k \ll r_0$.

Before discussing some details of the impact of r_0, the modified torsion force shall be derived. Fortunately, the only effective difference to the derivation in section 5.1.2 is that (5.14) is altered to:

$$\begin{aligned} -\nabla_{\mathbf{r}_l} V_{\mathrm{tors}\,k}(\mathbf{R}) &= -\nabla_{\mathbf{r}_l}\left[\frac{1}{2}\big(1+\left(4\cos^3\phi_k - 3\cos\phi_k\right)\big)\left(1-\exp\left[-\rho_k/r_0\right]\right)\right] \\ &= -\frac{1}{2}\big(12\cos^2\phi_k - 3\big)\left(\nabla_{\mathbf{r}_l}\cos\phi_k\right)\left(1-\exp\left[-\rho_k/r_0\right]\right) \\ &\quad - \frac{1}{2}\big(1+\left(4\cos^3\phi_k - 3\cos\phi_k\right)\big)\exp\left[-\rho_k/r_0\right] r_0^{-1}\left(\nabla_{\mathbf{r}_l}\rho_k\right) \ . \end{aligned} \tag{5.35}$$

The respective calculations for $\nabla_{\mathbf{r}_l}\cos\phi_k$ and $\nabla_{\mathbf{r}_l}\rho_k$ are given in (5.15) - (5.24). The complete expression for $\mathbf{F}_{\mathrm{tors}\,i}$ analogous to (5.25) is still too complex to give a considerable insight if it would be stated here. Instead, referring to section 5.3.1, the force acting on the first monomer of the arbitrarily selected configuration shall be exemplarily calculated, using the results from (5.31):

$$\begin{aligned} \mathbf{F}_{\mathrm{tors}\,1} &= -\frac{3}{2}\sin 3\varphi \frac{1}{r}\left(1-\exp\left[-r/r_0\right]\right)\begin{pmatrix}\sin\varphi \\ -\cos\varphi \\ 0\end{pmatrix} \\ &\quad -\frac{1}{2}\left(1+\cos 3\varphi\right)\exp\left[-r/r_0\right] r_0^{-1}\begin{pmatrix}\cos\varphi \\ \sin\varphi \\ 0\end{pmatrix} \ . \end{aligned} \tag{5.36}$$

As expected, the amplitude of the tangential part of the force is limited by the exponential factor $1 - \exp[-r/r_0]$, thus it does not diverge anymore:

$$\lim_{r\to 0}\frac{1-e^{-r/r_0}}{r} = \lim_{r\to 0}\frac{e^{-r/r_0}}{r_0} = r_0^{-1} \ . \tag{5.37}$$

Even adjusting a limit for the tangential force is possible, because the absolute value is limited by $3/2r_0^{-1}$ (compare (5.36) and (5.37)).

Furthermore, an additional *radial* part is induced by the fact, that the potential has now an explicit radial dependency. The absolute value of the radial part is also guided by r_0^{-1} for small r, so the upper limit for the tangential contribution also holds for the radial force. Unfortunately, the radial contribution brings along several systematic problems. Since the torsion energy is suppressed for smaller r, this radial part of the force points to the origin of the circle that r describes ($\mathbf{F}_{\mathrm{rad}} \sim (-\cos\varphi, -\sin\varphi, 0)$). Thus r tends to get smaller. This means that the system now prefers configurations that would have been critical in the unmodified model. By now, the simulation will work well in these situations as the force converges. On the other hand, the originally desired torsion potential $(1+\cos 3\phi)/2$ is described especially bad by the modified potential for the former critical configurations. So the modification itself tends to enlarge its effect. Secondly, the absolute value of the radial contribution has its maximum for $r = 0$ (the definition forbids $r < 0$). So it is expected that for very small r, the radial contribution leads to a turn-down of the monomer, since it is pushed too far in the central direction because of the finite time step. This leads to an artificial change of ϕ to $\phi \pm \pi$, which very probably causes serious structural misbehaviour. However, all unwanted effects for $r \ll r_0$ can now be considered as of less weight.

For small r_0 the modified potential is similar to the original potential of the GAB model, while a big r_0 implies a strong falloff and thus better prevents large forces: $F_{\mathrm{tang,max}} \sim r_0^{-1}$. However, the second case has the drawback that the potential converges against the one of the AB model since the torsion potential is suppressed. But that is of course not desirable, since all the hitherto effort was motivated by the *introduction* of this additional contribution. Actually, it is possible to continuously select between the AB and GAB model. It can be shown that while the thermodynamic properties of the GAB model are very similar to those of the AB model, the additional torsion potential has a considerable impact on the structural behaviour of the system. This is best seen by comparing ground-state structures of the same sequence for the GAB and the AB model. The conclusion of this is that the freely selectable fall-off threshold r_0 leads to different structural results for any different adjustment. This conclusion is very unsatisfactory, since the structural behaviour is crucial in protein folding. But it would be possible to think of more sophisticated and better adapted modifications to the original potential (5.1) to get around this problem.

Simulations with Modified Potential

After the theoretical part, the modifications should be tested. Therefore several simulations are carried out both for the homo-four-mer and the sequence 20.4 (see Table 1.I). All the usual parameters are used and the modified torsion potential is included. The considered problem causes always a *rise* of E_{cons}, since the kinetic energy is always spontaneously *increased* and the excessive energy is accumulated in the potential energy of the NHC particles. Therefore, the general expectation would be that still a slow increase of the “conserved” energy is observed, but not as crucial as in Fig. 5.11. Also, the smaller the chosen time step δt and the bigger the effective cut-off radius r_0, the smaller should be the increase of E_{cons}.

A time series of $T = 40 \cdot 10^3$ is simulated for the homo-four-mer with different selections of δt and r_0. The resulting time series are shown in Fig. 5.17. Interestingly from the picture it could be assumed that E_{cons} is really equilibrating at a constant level. But this cannot be true and is thus an artifact that originates from numerical errors, which are the only possible

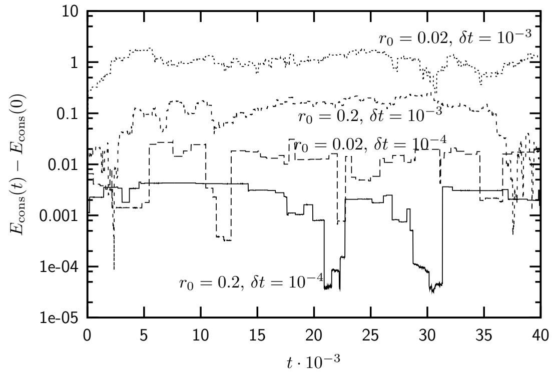


Figure 5.17: Logarithmic plot of the conserved energy for several simulation setups of the homo-four-mer, all at $T = 1.0$.

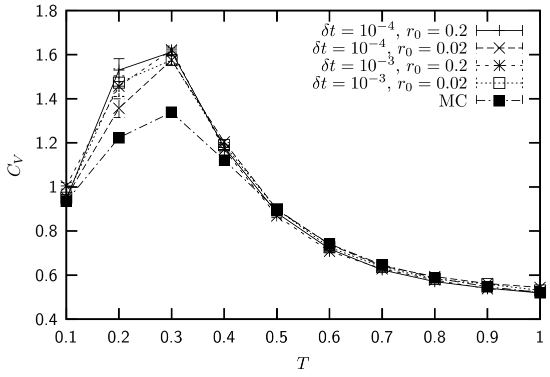


Figure 5.18: Heat capacities for the same simulations as in Fig. 5.17 are compared to the Monte Carlo result, which is represented by the chain dotted line.

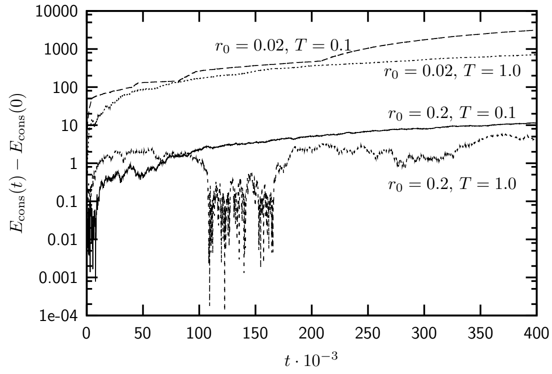


Figure 5.19: Logarithmic plots of the conserved energy for several simulation setups of the sequence 20.4, all with time step $\delta t = 10^{-3}$.

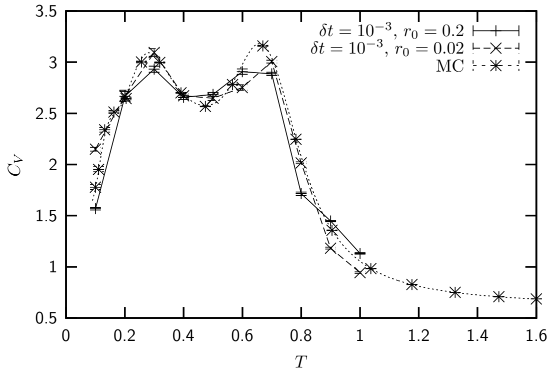


Figure 5.20: Heat capacities for $r_0 = 0.2$ (solid line) and $r_0 = 0.02$ (long-dashed line) are compared to the Monte Carlo result, which is represented by the dotted line.

cause for a decrease of E_{cons}. However, the general finding is that the assumptions concerning the effect of the choice of δt and r_0 are correct. The smaller the time step and the larger the cut-off radius is, the more accurate is the simulation in the critical situations, which causes E_{cons} to be less acceding.

The plot of the specific heat in Fig. 5.18 suggests that fortunately the effect of the considered problem does not have a systematic influence on the thermodynamic behaviour of the system. Compared to the Monte Carlo Metropolis data, there is a clear systematic deviation for all the MD simulations at the peak of the specific heat. The error bars of the MD measurements overlap in the whole temperature range.

Unfortunately it is only possible to perform the simulation of sequence 20.4 with $\delta t = 10^{-3}$, since a smaller time step caused technical problems due to too much memory consumption. The results of the investigation with $r_0 = 0.2$ and $r_0 = 0.02$ are shown in Figs. 5.19 and

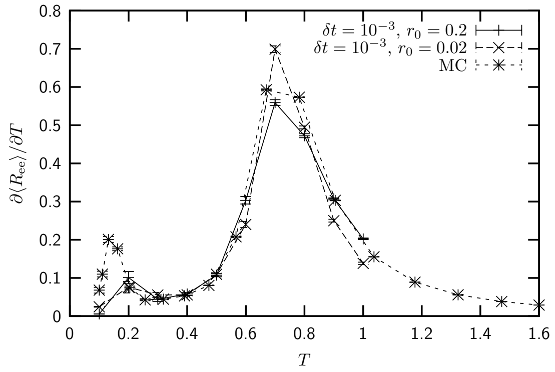


Figure 5.21: Fluctuation of end-to-end-distance for $r_0 = 0.2$ (solid line) and $r_0 = 0.02$ (long-dashed line) are compared to the Monte Carlo result, which is represented by the dotted line.

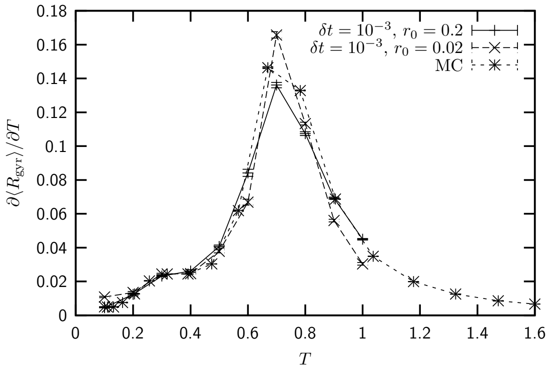


Figure 5.22: Fluctuation of radius of gyration for $r_0 = 0.2$ (solid line) and $r_0 = 0.02$ (long-dashed line) are compared to the Monte Carlo result, which is represented by the dotted line.

5.20. As expected, the total energy $E_{\rm cons}$ is slowly growing, but slower for the bigger choice of r_0. Interestingly, the increase is a little larger for $T = 0.1$, while for low temperatures the fluctuations should be smaller than for higher temperatures. But the effect is not as explicit as the difference between the two runs with unequal cut-off radii. Again, the impact of the different choices for r_0 on the heat capacity cannot be systematically identified. Both measurements show an acceptable agreement with the MC comparison data, except for small deviations as already observed in section 4.2. However, the run with the bigger cut-off radius $r = 0.2$, which should be closer to the AB case, really has slightly lower peaks, which is analogous to the comparison of GAB and AB specific heats in Fig. 5.4.

Since the fluctuations of the end-to-end-distance and the radius of gyration are nearly equal for the AB and the GAB model (see Fig. 5.5), the impact of r_0 should be of a purely numerical kind for these quantities. In fact, in Figs. 5.21 and 5.22 the deviations of the simulation with $r_0 = 0.2$ from the Monte Carlo data is smaller than those measured for $r_0 = 0.02$. However, in both cases the error bars do not overlap, which is analogous to the findings for the AB model without torsion (see Figs. 4.20 and 4.21).

In conclusion, the observations of the testing simulations have shown that it is really possible to decrease the unwanted effect by applying the modified torsion potential instead of the originally defined one. The adjustment of r_0 can help to obtain desired results. The most important finding is that the described problem obviously does not have a crucial influence on the measurement of thermodynamic quantities. Still, Molecular Dynamics simulations of models implying the torsion term should be handled carefully.

Summary

Within this work, a coarse-grained protein model has been studied, where amino acids are represented by two types of single monomers – A (hydrophobic) and B (hydrophilic) – like in a C_α model. Two successive monomers are linked by virtual peptide bonds. Therefore, the model can be considered as a continuous, three-dimensional ball-and-stick model. The most important properties of this model originate from an effective Lennard-Jones-like interaction between the monomers. At low temperatures, this leads to conformations with a dense core of A monomers, while the B monomers arrange as an outer shell. Naturally, proteins reside in an aqueous environment, which causes a behaviour for hydrophobic and hydrophilic amino acids comparable to the previously described effect in the AB model. The basic difference is that in the model this external effect is implied as an intrinsic feature.

The key issue of this project was to study the above model with both Monte Carlo simulations and Molecular Dynamics, representing the two big classes of computer simulations which are frequently used in protein folding research. Since usually different problems are treated with only one type of these techniques, this suggested the non-trivial question, whether the obtained results can be considered as equivalent. In the work at hand, an in-depth comparison is carried out and general statements are made.

First of all it was necessary to gain a fundamental understanding of both methods. Molecular Dynamics numerically integrates the Newtonian equations of motion. There are various ways to perform this numerical integration. The most common class of algorithms is derived from the *Liouville* formalism of classical mechanics. This approach has been reconstructed within this work. Since the classical Newtonian mechanics leaves the total energy of a system constant, modifications, regarding the coupling to the heat bath are required to make it possible to sample a system within the canonical ensemble. Therefore, two common *thermostats*, the Andersen thermostat and the Nosé-Hoover-Chain thermostat were tested for the harmonic oscillator and the quartic double well as simple model systems, which were also treated analytically.

The AB model system was first investigated with *parallel tempering* Monte Carlo simulations. The several considered thermodynamic properties such as specific heat or the fluctuations of end-to-end-distance and radius of gyration clearly show that the system undergoes two conformational transitions. The three separate domains can be denoted with *ground-state-like*, for the lowest temperatures, *globule*, and *random-coil* for high temperatures. The ground-state-like domain includes only structures which are close to few or even single states with very small potential energy.

Performing Molecular Dynamics with constraints covers technical difficulties. Therefore, flexible bonds were introduced in contrast to the original model. The impact of the usage of flexible bonds was studied analytically as well as in computer simulations. It turned out that in comparison to the classical model, the modification mainly constitutes a different behaviour of the transition from the ground-state-like domain to the globule domain. The according differences were only observed for small bond strengths. Therefore, the choice of a reasonable large bond strength allows to compare the results of the modified model with the original one.

Similar computer experiments were carried out with Molecular Dynamics. The utilised Nosé-Hoover-Chain thermostat was adjusted combining the experiences from the tests with a harmonic oscillator and approximations according to the flexible bond potential. The adjustment showed very good results in comparison with theoretic expectations about kinetic quantities. It was found that the thermodynamic properties as they had been measured with Monte Carlo were acceptably reproduced by the Molecular Dynamics simulation. A small systematic drift of the average potential energy towards higher values was noticed. By using the Andersen thermostat in further measurements, this deviation could be uncovered to be caused by characteristics of the Nosé-Hoover-Chain thermostat.

Additionally, the overlap parameter Q as well as the fraction of formed native contacts q were considered as reaction coordinates for the measurement of free-energy landscapes with Molecular Dynamics. It was shown that although the statistical properties of Molecular Dynamics are unsatisfying, the reweighting of the data makes it possible to obtain qualitative agreement with the results of Monte Carlo experiments. However, the measurements of structural quantities in general with Molecular Dynamics showed quantitative deviations from the Monte Carlo data.

Furthermore, the autocorrelation times of Molecular Dynamics simulations were found to be orders of magnitudes larger than for Metropolis, especially for non-critical, higher temperatures, where autocorrelations usually tend to be smaller. Also, the error analysis in Molecular Dynamics runs presumes large Jackknife bins and long run times to be certain. Therefore, Monte Carlo is superior to Molecular Dynamics with respect to the production of statistical data. This is also the case for the special purpose of finding structures with minimal energy, since Molecular Dynamics is much more sensitive against the implementation of sophisticated technical extensions.

To imply further empirical chemical constraints into the AB model, an additional potential term with respect to the *torsion angles* was included. With Monte Carlo methods, the deviations of the extended (GAB) and the original system were examined. The torsion potential was found to mainly influence the ground-state-like domain, which is slightly stabilised. Also, the minimal energy structures of the AB and GAB model are different. For higher temperatures, on the other hand, the impact of the torsion potential does not seem to play a remarkable role.

The application of the torsion potential to Molecular Dynamics was found to be non-trivial. In particular, while the potential itself is well-behaved, its gradient, which is proportional to the force, is divergent for nearly linear configurations. Unlike for, e.g., the repulsion of the Lennard-Jones potential, the divergence in the torsion force is important because linear configurations occur frequently. Several approaches to eliminate the misbehaviour caused by this problem were presented. The most promising, a modification of the torsion potential, was discussed in detail and implemented. The result of according measurements was that it was really possible to control the unwanted effect. Particularly, the impact of the divergence and the modification of the torsion potential both did not show a remarkable deviation with respect to the thermodynamic properties of the model system.

This work has shown that the observation of protein folding with simple model systems can be carried out with Monte Carlo as well as with Molecular Dynamics, where the kinetic properties of the system are driven by deterministic laws. Therefore, many interesting starting points for further investigations of especially kinetic effects are apparent.

For a more thorough identification of Monte Carlo and Molecular Dynamics time scales, the measurement of Chevron plots [43, 44, 45] would be expedient. These are directly connected to kinetics, as the basic is to measure the number of simulation iterations, until the sequence is unfolded and folded respectively up to a certain degree. The comparison of the obtained results should be even more significant than the comparison of autocorrelation times. Being sure that the dynamics of Monte Carlo and Molecular Dynamics is comparable in principle, it would then be possible to do all measurements with the statistically superior Monte Carlo methods and appropriately rescale the kinetic properties.

The impact of mutation on the path to fold can be evaluated by measuring Φ-values [51, 52], which compare the change of stability and free-energy-barrier caused by a certain mutation of the sequence. Natural proteins can be considered to be optimised, i.e. the path to fold will usually be decelerated by mutation. However, the sequences used for the AB model are artificial. Thus, it would not be a surprise, if the path to fold could be accelerated by mutation. This way, it would be even possible to perform sequence designing and find sequences with especially stable ground states.

A more technical aspect could be the implementation of the model in non-cartesian coordinates, where the angles and bond lengths are the degrees of freedom. This would make it possible to study e.g. the GAB model as well as the G$\bar{\text{o}}$C$^{\alpha}$ model [39] without the need of a cut-off for the torsion potential. Additionally, this would enable the use of stiff bonds and thus allow larger time steps. Furthermore, the observation of the autocorrelation behaviour could give an insight, if it is possible to accelerate the evolution of the system in phase space by altering the coordinate system.

Appendix A

Selected Source Codes

A.1 Andersen Thermostat

A pseudo-C source code of the Andersen thermostat with use of the Box-Müller method for calculating Gaussian distributed random numbers could look like this (see section 2.4.1 for a description of the algorithm):

```
void thermostating() {
  for (i=1; i<=Np; i=i+1) {                      // for each particle
    if (RAN01()<nu*dt) {                         // check for collision
      delta_ekin=0.0;
      sigma_sq=k_B*T/m[i];                       // calculate sigma^2
      for (j=1; j<Nd; j=j+2) {                   // for each dimension
        gaussian(sigma_sq,&x,&y);
        delta_ekin=delta_ekin-v[i,j]*v[i,j];
        v[i,j]=x;
        delta_ekin=delta_ekin+v[i,j]*v[i,j];
        delta_ekin=delta_ekin-v[i,j+1]*v[i,j+1];
        v[i,j+1]=y;
        delta_ekin=delta_ekin+v[i,j+1]*v[i,j+1];
      }
      if (j<=Nd) {                               // if Nd is odd
        gaussian(sigma_sq,&x,&y);
        delta_ekin=delta_ekin-v[i,j]*v[i,j];
        v[i,j]=x;
        delta_ekin=delta_ekin+v[i,j]*v[i,j];
      }
      ekin=ekin+delta_ekin*m[i]/2.0;
    }
  }
}
```

The variables are explained in Table A.I, as far as they are not self-explanatory. In delta_ekin the difference of the old and new squared velocity is accumulated. This makes sense, since a velocity update happens quite seldom, and so the kinetic energy does not have to be calculated from scratch every time step after applying the `thermostating()` function. Since the Box-Müller method is used for generating random numbers from a Gaussian distribution with squared width `sigma_sq`, always *two* such random numbers are produced. Therefore, for the first $2n$ dimensions ($n \in \mathbb{N}$) two velocity components are randomly chosen at once (lines `6-14`). Afterwards, for an odd number of dimensions in the system (e.g. for one- or three-dimensional systems), one component is left (lines `15-20`). Although, this code replication in lines `8-10`, `11-13` and `17-19` is unaesthetic, it provides a reasonable speed-up, since calculating two new random numbers for *each* velocity component would be twice as costly. For the sake of completeness here is an implementation of the Box-Müller method:

```
void gaussian(sigma_sq,*x,*y) {
  r=sqrt(-2*sigma_sq*log(1-RAN01()));
  theta=2*M_PI*RAN01();
  x=r*cos(theta);
  y=r*sin(theta);
}
```

There is still some room for small optimisation, which has not been included here for readability. E.g. not using a function call when calculating `x` and `y` would be faster. Also doing the check about the number of dimensions (line `15`) only once in the beginning instead for every colliding particle could (slightly) speed up the performance.

The implementation is simple – the function `thermostating()` has to be called once per MD time step.

A.2 Nosé-Hoover-Chain Thermostat

As already mentioned while explaining the principle of the Nosé-Hoover-Chain thermostat, both Ref. [21] and [22] give pseudo source code. Unfortunately, both include small spelling mistakes, which can be confusing. Additionally, each implies one optimisation, which can be

Table A.I: Variables used in the pseudo source code of the Andersen thermostat.

`T`	desired equilibrium temperature	T
`dt`	time step	δt
`ekin`	kinetic energy	E_{kin}
`Np`	number of particles in the system	
`Nd`	number of dimensions of the system	
`nu`	collision frequency	ν
`m[i]`	mass of particle `i`	m
`v[i,j]`	`j`th velocity component of particle `i`	$v_{i,j}$
`RAN01()`	random number $\in [0,1)$ (used generator [53])	

combined. Therefore, it seems to be expedient to give a third version of pseudo code, which is similar but not equal to the two variants in the given references:

```
 1 void thermostating() {
 2   E1set=Nf*k_B*T;
 3   E2set=k_B*T;
 4   scale=1.0;                        // no initial particle vel. scaling
 5   for (k=1; k<=nc; k=k+1) {
 6     for (j=1; j<=m; j=j+1) {
 7       dts2=w[j]*dt/((double)nc*2.0);// precalculate fractional time steps
 8       dts4=dts2/2.0;
 9       dts8=dts4/2.0;
10       axi[1]=(ekin*2.0-E1set)/Q[1]; // calculate thermostat acceleration
11       for (i=2; i<=M; i=i+1) {
12         axi[i]=(Q[i-1]*vxi[i-1]*vxi[i-1]-E2set)/Q[i];
13       }
14       vxi[M]=vxi[M]+axi[M]*dts4;    // update thermostat velocities
15       for (i=M-1; i>=1; i=i-1) {    // attention: process "backwards"
16         s=exp(-vxi[i+1]*dts8);
17         vxi[i]=(vxi[i]*s+(axi[i]*dts4))*s;
18       }
19       s=exp(-vxi[1]*dts2);
20       scale=scale*s;                // accumulate particle vel. scaling
21       ekin=ekin*s*s;                // scale particle kinetic energy
22       for (i=1; i<=M; i=i+1) {      // update thermostat positions
23         xi[i]=xi[i]+vxi[i]*dts2;
24       }
25       axi[1]=(ekin*2.0-E1set)/Q[1]; // update thermostat acc. and vel.
26       for (i=1; i<=M-1; i++) {
27         s=exp(-vxi[i+1]*dts8);
28         vxi[i]=(vxi[i]*s+(axi[i]*dts4))*s;
29         axi[i+1]=(Q[i]*vxi[i]*vxi[i]-E2set)/Q[i+1];
30       }
31       vxi[M]=vxi[M]+axi[M]*dts4;
32     }
33   }
34   for (i=0; i<Nf; i=i+1) {          // scale particle velocities
35     v[i]=v[i]*scale;
36   }
37 }
```

The whole chapter 2 and especially the therein given references should help to get a deeper understanding of the algorithm itself. The two main `for`-loops in line `5` and `6` handle the smaller time steps and the higher order integration with respect to the Nosé-Hoover degrees of freedom. The main effect of the whole function is the aggregation of the velocity scaling factor `scale`. The higher order integration is coupled to certain prefactors `w[j]` for each iteration of the `j`-loop, which are collected in Table A.III up to order $m = 7$, according to [22]. Here also is still left some room for optimisations. E.g. line `2` and `3` must be carried out only once for each temperature, and if simulating with a fixed time step size `dt` the values `dts2`, `dts4` and `dts8` could be precalculated for each value `w[j]`.

The implementation of the Nosé-Hoover-Chain thermostat as given here is done as follows (for a full example see e.g. Ref. [20], appendix E.2):

```
setup();
for (i=0; i<no_steps; i=i+1) {
  ...
  thermostating();
  velocity_stoermer_verlet();
  thermostating();
  ...
  measurement();
  ...
}
```

I.e., because of the separation of the Hamiltonian into the Nosé-Hoover part and the stand-alone "classic" system by a Trotter factorisation (compare eq. (2.76)), the Nosé-Hoover propagation is calculated twice: in the beginning and in the end of each MD time step, each time with a step size $\delta t/2 =$`dt/2`. The function `velocity_stoermer_verlet()` stands for the integration of the stand-alone system with an algorithm analogous to eq. (2.16). In appendix E.2 of Ref. [20] this function is denoted with `pos_vel()`.

Table A.II: Variables used in the pseudo source code of the Nosé-Hoover-Chain thermostat.

`T`	desired equilibrium temperature	T
`dt`	time step	δt
`ekin`	kinetic energy	E_{kin}
`Nf`	number of degrees of freedom	
`M`	number of Nosé-Hoover particles (NH-Chain length)	M
`nc`	separation of one time step for NHC	n_c
`m`	order of factorisation	m
`w[i]`	factorisation prefactors (see Table A.III)	w_j
`m[i]`	mass of particle `i`	m_i
`v[i]`	`i`th velocity component (with respect to all particles)	v
`Q[i]`	virtual mass of `i`th Nosé-Hoover particle	Q_i
`xi[i]`	position of `i`th Nosé-Hoover particle	ξ_i
`vxi[i]`	velocity of `i`th Nosé-Hoover particle	v_{ξ_i}
`axi[i]`	acceleration of `i`th Nosé-Hoover particle	$\dot{p}_{\xi_i}/Q_i$

Table A.III: Prefactors for higher order Trotter schemes as given in [22].

m	1	3	5	7
w_1	1	1.3512071919596576340	0.41449077179437573714	−1.17767998417887
w_2		−1.7024143839193152681	0.41449077179437573714	0.235573213359357
w_3		1.3512071919596576340	−0.65796308717750294857	0.78451361047756
w_4			0.41449077179437573714	1.3151863206839
w_5			0.41449077179437573714	0.78451361047756
w_6				0.235573213359357
w_7				−1.17767998417887

Bibliography

[1] T. E. Creighton, *Proteins: structures and molecular properties.* Freeman, New York, 2nd. ed., 1993.

[2] C. B. Anfinsen, *Principles that Govern the Folding of Protein Chains*, *Science* **181** (1973) 223–230.

[3] J. C. Venter, *The Sequence of the Human Genome*, *Science* **291** (2001) 1304–1351.

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

[5] F. H. Stillinger, T. Head-Gordon, and C. L. Hirshfeld, *Toy model for protein folding*, *Phys. Rev. E* **48** (1993) 1469–1477.

[6] F. H. Stillinger and T. Head-Gordon, *Collective aspects of protein folding illustrated by a toy model*, *Phys. Rev. E* **52** (1995) 2872–2877.

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

[8] S. Nosé, *A molecular dynamics method for simulations in the canonical ensemble*, *Mol. Phys.* **100** (2002) 191–198. Reprint of the original paper from 1983.

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

[10] G. J. Martyna and M. L. Klein, *Nosé-hoover chains: The canonical ensemble via continuous dynamics*, *J. Chem. Phys.* **97** (1992) 2635–2643.

[11] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *Equation of State Calculations by Fast Computing Machines*, *J. Chem. Phys.* **21** (1953) 1087–1092.

[12] K. Hukushima and K. Nemoto, *Exchange Monte Carlo Method and Application to Spin Glass Simulations*, *Phys. Soc. J.* **65** (1996) 1604–1608.

[13] K. F. Lau and K. A. Dill, *A Lattice Statistical Mechanics Model of the Conformational and Sequence Spaces of Proteins*, *Macromolecules* **22** (1989) 3986–3997.

[14] A. Irbäck, C. Peterson, and F. Potthast, *Identification of amino acid sequences with good folding properties in an off-lattice model*, *Phys. Rev. E* **55** (1997) 860–867.

[15] A. Irbäck, C. Peterson, F. Potthast, and O. Sommelius, *Local interactions and protein folding: A three-dimensional off-lattice approach*, *J. Chem. Phys.* **107** (1997) 273–282.

[16] H.-P. Hsu, V. Mehra, and P. Grassberger, *Structure optimization in an off-lattice protein model*, *Phys. Rev. E* **68** (2003) 037703-1–4.

[17] M. Bachmann, H. Arkin, and W. Janke, *Multicanonical study of coarse-grained off-lattice models for folding heteropolymers*, *Phys. Rev. E* **71** (2005) 031906-1–11.

[18] W. F. van Gunsteren and H. J. C. Berendsen, *Algorithms for macromolecular dynamics and constraint dynamics*, *Mol. Phys.* **34** (1977) 1311–1327.

[19] H. C. Andersen, *Rattle: A Velocity Version of the Shake Algorithm for Molecular Dynamics Calculations*, *J. Comput. Phys.* **52** (1983) 24–34.

[20] D. Frenkel and B. Smit, *Understanding Molecular Simulation*, vol. 1 of *Computational Science Series*. Academic Press, London, 2nd. ed., 2002.

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

[22] R. Windiks, "Thermostatting in Molecular Dynamics Simulations." http://homepage.swissonline.ch/windiks/documents/thermostats.pdf.

[23] L. Verlet, *Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules*, *Phys. Rev.* **159** (1967) 98–103.

[24] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 1. Wiley, New York, 1957.

[25] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 2. Wiley, New York, 1966.

[26] W. Janke, *Pseudo Random Numbers: Generation and Quality Checks*, in *Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms* (J. Grotendorst, D. Marx, and A. Muramatsu, eds.), vol. 10, pp. 423–445, John von Neumann Institute for Computing, Jülich, 2002.

[27] M. E. Tuckerman, C. J. Mundy, and G. J. Martyna, *On the classical statistical mechanics of non-Hamiltonian systems*, *Europhys. Lett.* **45** (1999) 149–155.

[28] K. Cho, J. D. Joannopoulos, and L. Kleinman, *Constant-temperature molecular dynamics with momentum conservation*, *Phys. Rev. E* **47** (1993) 3145–3151.

[29] G. J. Martyna, *Remarks on "Constant-temperature molecular dynamics with momentum conservation" [28]*, *Phys. Rev. E* **50** (1994) 3234–3236.

[30] Y. Liu and M. E. Tuckerman, *Generalized Gaussian moment thermostatting: A new continuous dynamical approach to the canonical ensemble*, *J. Chem. Phys.* **112** (2000) 1685–1700.

[31] M. E. Tuckerman and M. Parrinello, *Integrating the Car-Parrinello equations. I. Basic integration techniques*, *J. Chem. Phys.* **101** (1994) 1302–1315.

[32] M. Suzuki, *General theory of fractal path integrals with applications to many-body theories and statistical physics*, *J. Math. Phys.* **32** (1991) 400–407.

[33] H. Yoshida, *Construction of higher order symplectic integrators*, *Phys. Lett. A* **150** (1990) 262–268.

[34] W. Janke, *Statistical Analysis of Simulations: Data Correlation and Error Estimation*, in *Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms* (J. Grotendorst, D. Marx, and A. Muramatsu, eds.), vol. 10, pp. 423–445, John von Neumann Institute for Computing, Jülich, 2002.

[35] W. Janke, *Histograms and All That*, in *Computer Simulations of Surfaces and Interfaces, NATO Science Series, II. Mathematics, Physics and Chemistry* (B. Dünweg, D. P. Landau, and A. I. Milchev, eds.), vol. 114, pp. 137–157, NATO Advanced Study Institute. Kluwer, Dordrecht, 2003.

[36] A. M. Ferrenberg and R. H. Swendsen, *Optimized Monte Carlo Data Analysis*, *Phys. Rev. Lett.* **63** (1989) 1195–1198.

[37] M. Newman and G. T. Barkema, *Monte Carlo Methods in Statistical Physics.* Oxford University Press, 1999.

[38] B. A. Berg, *Multicanonical Simulations Step by Step*, *Comp. Phys. Commun.* **153** (2003) 397–406.

[39] C. Clementi, H. Nymeyer, and J. N. Onuchic, *Topological and Energetic Factors: What Determines the Structural Details of the Transition State Ensemble and "En-route" Intermediates for Protein Folding? An Investigation for Small Globular Proteins*, *J. Mol. Biol.* **298** (2000) 937–953.

[40] W. Kerler and P. Rehberg, *Simulated-tempering procedure for spin-glass simulations*, *Phys. Rev. E* **50** (1994) 4220–4225.

[41] M. Bachmann and W. Janke, *Multicanonical Chain-Growth Algorithm*, *Phys. Rev. Lett.* **91** (2003) 208105-1–4.

[42] S. Schnabel, *Thermodynamische Eigenschaften und Faltungskanäle von Coarse-Grained Heteropolymeren.* Diploma thesis, Institut für Theoretische Physik, Universität Leipzig, 2005.

[43] H. S. Chan and K. A. Dill, *Protein Folding in the Landscape Perspective: Chevron Plots and Non-Arrhenius Kinetics*, *Proteins: Struct. Funct. Genet.* **30** (1998) 2–33.

[44] H. Kaya and H. S. Chan, *Solvation Effects and Driving Forces for Protein Thermodynamic and Kinetic Cooperativity: How Adequate is Native-centric Topological Modeling?*, *J. Mol. Biol.* **326** (2003) 911–931.

[45] A. Kallias, *Thermodynamics and Folding Kinetics of Coarse-Grained Protein Models.* Diploma thesis, Institut für Theoretische Physik, Universität Leipzig, 2005.

[46] U. H. E. Hansmann and L. T. Wille, *Global Optimization by Energy Landscape Paving*, *Phys. Rev. Lett.* **88** (2002) 068105.

[47] S. E. Jackson and A. R. Fersht, *Folding of Chymotrypsin Inhibitor 2. First Evidence for a Two-State Transition*, *Biochemistry* **30** (1991) 10428–10435.

[48] N. D. Socci and J. N. Onuchic, *Kinetic and thermodynamic analysis of proteinlike heteropolymers: Monte Carlo histogram technique*, *J. Chem. Phys.* **103** (1995) 4732–4744.

[49] A. Irbäck, F. Sjunnesson, and S. Wallin, *Three-helix-bundle protein in a Ramachandran model*, *Proc. Natl. Acad. Sci. USA* **97** (2000) 13614–13618.

[50] M. Griebel, S. Kanpek, G. Zumbusch, and A. Caglar, *Numerische Simulationen in der Moleküldynamik.* Springer, Berlin, 2004.

[51] S. B. Ozkan, I. Bahar, and K. A. Dill, *Transition states and the meaning of* Φ*-values in protein folding kinetics*, *Nature Struct. Bio.* **8** (2001) 765–770.

[52] A. R. Fersht and V. Daggett, *Protein Folding and Unfolding at Atomic Resolution*, *Cell* **108** (2002) 573–582.

[53] G. Marsaglia, A. Zaman, and W. W. Tsang, *Toward a universal random number generator*, *Stat. Prob. Lett.* **9** (1990) 35–39.

Danksagung

Zunächst danke ich meinen Betreuern Prof. Dr. Wolfhard Janke und Dr. Michael Bachmann, die mir nicht nur jederzeit mit großem Einsatz bei auftretenden Problemen behilflich waren, sondern mir auch durch ein gut durchdachtes Thema die Möglichkeit gegeben haben viel zu lernen und einen Einblick in ein hochinteressantes Forschungsgebiet zu erhalten.

Ein besonderer Dank gilt meinen lieben Eltern, die mir durch ihre stetige Unterstützung in jeder Form das Studium ermöglicht haben. Ebenfalls danken möchte ich meiner Freundin Isabel, die mir in vielen Situationen wichtigen Beistand geleistet hat (und leistet).

Für die Arbeit war ich stets auf ein zuverlässiges Computersystem angewiesen. Dies wurde von Dr. Elmar Bittner und Andreas Nussbaumer gewährleistet, die neben ihrer eigenen wissenschaftlichen Arbeit (nicht nur) den Linux Cluster im Institut für theoretische Physik perfekt verwalteten und pflegten, und sich zu jeder Zeit um etwaige Wünsche und Probleme sofort kümmerten. Dafür möchte ich mich herzlich bedanken.

Für zahlreiche Hilfestellungen bei der Bewältigung von computerspezifischen und programmiertechnischen Schwierigkeiten danke ich insbesondere Andreas Nussbaumer, aber auch Erik Lorenz, Anna Kallias und Stefan Schnabel.

Aufschlussreiche Diskussionen über Proteinmodelle und Proteinfaltung wie auch über diverse Methoden der computerorientierten statistischen Physik konnte ich mit Anna Kallias, Stefan Schnabel, Thomas Vogel, Andreas Nussbaumer, Erik Lorenz und Rainer Bischof führen.

Nicht zuletzt verdanke ich einer langen Reihe netter Kommilitonen und Freunde, dass ich mein Studium und auch die Diplomarbeit stets mit Freude angehen konnte und für den nötigen Ausgleich gesorgt war. Ich denke dabei besonders an Andreas C., Andreas N., Anna K., Erik L., Franziska “Franzi I” L., Friedmar S., Johannes “Inju” P., Rainer B., Volker K. und an den Verein “Taekwondo YomChi”.

Selbständigkeitserklärung

Hiermit erkläre ich, die vorliegende Arbeit selbständig und ohne unerlaubte fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die in der Literatur angegebenen Quellen benutzt und sämtliche der Literatur entnommenen Textstellen und Bilder als solche kenntlich gemacht.

Leipzig, den 28.10.2005

Einverständniserklärung

Hiermit erkläre ich mich einverstanden, dass meine Diplomarbeit nach positiver Begutachtung zur Benutzung in der Zweigstelle Physik der Universitätsbibliothek zur Verfügung gestellt wird.

Leipzig, den 28.10.2005