1 Introduction

In this thesis we present the *Uncoupling-Coupling* Method (UC), a new approach for drawing samples from high-dimensional probability distributions based on the *Markov Chain Monte Carlo* (MCMC) methodology. Applications in the fields of Statistical Physics and Bayesian Statistics naturally lead to the investigation of a narrow distribution, located in separated parts of high probability on a high-dimensional state space, from which it is not feasible to directly generate independent samples. In such cases, the MCMC method provides a powerful and flexible framework for computer simulations.

Structural properties of biomolecules are often investigated by means of MCMC, a challenging application which attracts the attention from researches coming from biology, chemistry, physics, bioinformatics, and mathematics. Typical three-dimensional structures (so called *metastable conformations* or *metastable sets*) determine the functionality of a biomolecule in an organic environment. Knowledge of metastable conformations can be used for example in pharmaceutical research to determine the likelihood of a biomolecule of being developed into a drug. To support such tasks, analyzing properties of biomolecules by means of computer simulation has been dramatically enhanced over the last decade. This problem cannot be solved by sheer computer power; a proper modeling of the physical and chemical situation together with the development of robust algorithms, which adapt to the structure of the problem, are of great importance to obtain reliable results.

We can reformulate the problem of detecting and describing metastable conformations of a biomolecule by means of statistical physics in terms of the canonical ensemble, a high-dimensional probability distribution that typically consists of separated parts of high probability. Metastable conformations are understood as fluctuations within the canonical ensemble wrt. some Markov operator, and MCMC provides the basis to actually identify them. Yet, drawing samples from this distribution by MCMC is hampered by the trapping problem—the sample path of a Monte Carlo Markov chain jumps rarely if at all within a finite simulation time between *metastable sets*, which causes a slow mixing and thus slow convergence of the chain.

The UC algorithm directly addresses the trapping problem by hierarchically decomposing the state space into metastable sets. Its characteristic features are:

- hierarchical and adaptive construction of a patchwork of bridge distributions, which embeds the target distribution in an auxiliary distribution;
- identification of metastable sets based on dominant eigenvectors of associated Markov operators;

- combined resampling and annealing via bridge distributions restricted to metastable sets;
- independent parallel sampling of rapidly mixing Markov chain;
- and a proper reweighting of all samples to the target distribution.

In addition, we proof that Monte Carlo Markov chains restricted by our method are indeed rapidly mixing on identified metastable sets. From a MCMC viewpoint UC distinguishes itself from existing auxiliary distributions or parallel sampling techniques by actually decomposing the state space into metastable sets. UC can also be looked at from a dynamical systems' viewpoint: in this context, analyzing the spectrum of Markov operators by means of associated Markov chains leads to an adaptive domain decomposition of the state space. In this thesis, we combine the MCMC methodology with an algorithmic exploitation of spectral properties.

Markov Chain Monte Carlo. The original MCMC method was developed in physics by Metropolis et al. [89], and later generalized and put into a statistical framework by Hastings [66]. The Metropolis (or Metropolis-Hastings) algorithm [34, 63, 84, 105] is the most common form of MCMC and essentially builds upon Markov chain theory [16, 19, 93].

Suppose that we are interested in a distribution given by a density function f with values in $\Omega \subseteq \mathbb{R}^d$, from which it is practically impossible to draw independent samples (e.g., this could be the canonical distribution of a physical system or the posterior distribution in Bayesian statistics). Our goal is to obtain expectations of some function g with respect to f, i.e., computing the integral

$$I_f(g) = \int g(x) f(x)\, dx.$$

The Metropolis algorithm realizes a Markov chain $\mathcal{X} = X^{(1)}, X^{(2)}, X^{(3)}, \ldots$ having f as its invariant density. A sample $\mathbf{x} = (x^{(1)}, \ldots, x^{(n)})$ of $\mathcal{X}$ is obtained by accepting a proposal step $x_{\text{prop}}^{(k+1)}$ with a probability that only depends on the ratio of $f(x_{\text{prop}}^{(k+1)})/f(x^{(k)})$, thereby avoiding a computation of the unknown normalizing constant (which in its integral representation is typically hard to evaluate). The generated (dependent) random sample $\mathbf{x}$ then enables us to estimate the integral I_f as

$$\hat{I}_f(g) = \frac{1}{n} \sum_{k=1}^{n} g(x^{(k)}). \tag{1}$$

A special variant of the Metropolis algorithm is the Gibbs' sampler [52] (also known as the heatbath method in statistical physics [120]), which is

build upon iterative sampling of *conditional distributions* (i.e., distributions that are restricted to a subspace of the target distribution). Its popularity mainly stems from its easy and often efficient applicability to statistical inference problems [56].

In practice, it is hard to construct a Metropolis algorithm for a specific application such that the Markov chain has good mixing properties, which is essential for a good convergence rate. At least, it has been shown that a basic convergence property (which can be expressed via geometric bounds for eigenvalues of the Metropolis Markov chain [31]) hold for specific problem classes. For a geometrically ergodic Markov chain its mixing rate depends on the 2nd largest eigenvalue of its associated reversible Markov operator, which can serve as an indicator for the sampling lenght in order to estimate integrals as (1). The bigger the spectral gap in the spectrum of the Markov operator between $\lambda_1 = 1$ and the 2nd largest eigenvalue λ_2, the better is the mixing property of the Markov chain. If λ_2 is bounded far away from 1, we speak of a *rapidly mixing Markov chain*.

In real applications, MCMC often suffers from an extremely slow mixing as a result of getting trapped in metastable sets, making it virtually impossible to obtain a reliable sampling. To overcome this notoriously difficult problem, many researchers from diverse fields contributed over the last two decades a variety of advanced techniques and extensions to the standard MCMC scheme (for an overview, see [18, 36, 56, 83, 84]).

Many advanced techniques can be formulated as *data augmentation schemes* [69, 122], where new variables are introduced artificially in the system in order to improve mixing. For example, in Statistical Physics, Swendsen and Wang invented a cluster algorithm which prevents critical slowing down for the Ising and Potts model [121], which was further modified and generalized in [98, 132]. Another important method is *Hybrid Monte Carlo* (HMC) [15, 33] by Duane et al., which combines *molecular dynamics* with the Metropolis algorithm. HMC enables large moves in the state space by using short molecular dynamics trajectories of the underlying Hamiltonian system as proposal steps. Interestingly, HMC is even used in the statistics community [94, 96]. Both, the Swendsen-Wang algorithm and HMC, can be regarded as data augmentation schemes.

Instead of drawing samplings directly from the target distribution the set up of an *auxiliary distribution* which is smoother and enables better mixing is seen as a powerful approach to attack the trapping problem (e.g., by introducing a "temperature" parameter). After drawing samples from the auxiliary distribution, the samples are reweighted to the target distribution [37]. *Umbrella Sampling* [126] by Torrie and Valleau is one of the first methods that followed this strategy based on modifications in the Hamiltonian describing the target distribution. The same idea lies behind generalized ensembles: without introducing expert knowledge in the system under consideration, Berg and Neuhaus proposed the *Multicanonical Monte*

Carlo method [9, 133], which seeks to sample in a flat energy distribution over a broad temperature range; whereas the *1/k ensemble method* [68] by Hesselbo and Stinchcombe aims at a flat entropy distribution. Connected to these approaches is Geyer's *Parallel Tempering* [53], where Metropolis Markov chains, which can exchange their actual states, run in parallel at different temperatures. These and many others popular MCMC methods have been combined and extended in various ways [36, 65, 75, 84].

To perform bridge sampling in a small temperature range, Fischer et al. proposed *Adaptive Temperature HMC* (ATHMC), combining ideas from umbrella sampling and HMC [42]. We will use ATHMC (or alternatively, the related *Potential Scaling HMC* (PSHMC) method, which we introduce in this thesis) as one of the building blocks for UC.

Metastability. The phenomenon of *metastability* arises naturally in a variety of systems with complex dynamical behavior (e.g., biomolecules [45], climate models [48], or computer networks [22]). Mathematical modeling leads to a (possibly stochastic) dynamical system with a huge number of degrees of freedom. Suppose, we would observe a single trajectory of such a dynamical system for a long period of time: the trajectory would remain for a certain time in one metastable set (i.e., in one part of the state space), followed by a sudden and apparently random transition into another metastable set, and so on. Hence, a reduced description, which essentially characterizes the system's behavior, would consist of (a) an identification of the metastable sets, (b) the probability to stay within these metastable sets, and (c) transition rates between them.

To describe metastable behavior precisely in mathematical terms, Dellnitz and Junge suggested recently to analyze the global behavior of a dynamical system via its associated *Perron-Frobenius operator* rather than by a (possibly ill-conditioned) single long-term trajectory [24]. Investigation of the associated Perron-Frobenius operator revealed intrinsic connections between dominant eigenvalues of its spectrum and *almost invariant sets* (which turn out to be metastable sets in our setting). The application of this approach to dynamical systems, where the essential part takes place on a low-dimensional subspace, led to a discretized eigenvalue problem for the Perron-Frobenius operator, which is efficiently solved by a *multilevel subdivision technique*.

By identifying almost invariant sets of the associated Hamiltonian of a molecule with its metastable conformations, Deuflhard et al. applied the approach of Dellnitz and Junge to small molecular systems [27]. Yet, the curse of dimension prevented the application of subdivision techniques for higher dimensional systems.

A thorough reformulation in terms of statistical physics and Hybrid Monte Carlo (HMC) sampling by Schütte et al. in [112, 113] extended

these algorithmic approaches to be applicable for high-dimensional systems. Therein, metastable conformations are understood as fluctuations within the canonical ensemble with respect to the Hamiltonian dynamics. This is achieved by introducing a *transfer operator*, which (like in the approach of Dellnitz and Junge) reflects metastable conformations in its spectral structure. The mathematical justification of this approach is based on the transfer operator to be a self-adjoint Markov operator. Since in the context of MCMC we simulate by means of *reversible* Markov chains associated with self-adjoint Markov operators, transfer operator techniques can be directly applied to Metropolis Markov chains. Recently, the transfer operator approach has also been investigated for a broader class of dynamical situations [71, 114].

The central theme of the approaches by Dellnitz and Junge, and Schütte et al. is to exploit information contained in eigenfunctions corresponding to eigenvalues which are close to the unit circle. In connection to the latter approach, Deuflhard et al. described a *dynamical cluster algorithm* for the identification of metastable sets in nearly uncoupled reversible Markov chains [30]. Identification of metastable sets is also one of the central pillars in the UC algorithm, and we will base it in this thesis on the strategies described in [30, 112, 113].

Biomolecules. Biomolecules are the building blocks of life. The aim of analyzing biomolecules (by experiment or by computer simulation) is to reveal structural, chemical, and biological information in order to understand its function in an organic environment and its physiological impact on a living system.

The structure of a biomolecule has a great influence on its chemical and biological properties. For example, the function of a large protein depends on certain active sites, which only show up after folding from its primary to its tertiary structure (the actual three-dimensional structure) has taken place [17, 97]. In the same way, the structure of a ligand (i.e., a small biomolecule) determines to a large amount if binding to a receptor (e.g., to the active site of a large protein) is possible or not.

One main difficulty for understanding such processes is that biomolecules do not exist in a unique structure. They rather fluctuate and oscillate for a long time within a metastable conformation, and occasionally perform a transition to another one (see Fig. 1). Metastable conformations are generally assumed to exist in a hierarchical order [45]. Getting knowledge about typical metastable conformations and their hierarchical structure is an important aspect to reveal the function of a biomolecule.

At this point, computer simulations of (bio-)molecular systems can help to fill the gap between theory and experiment [2, 46, 60]; they provide insight into the dynamics of molecules not accessible otherwise. A typical

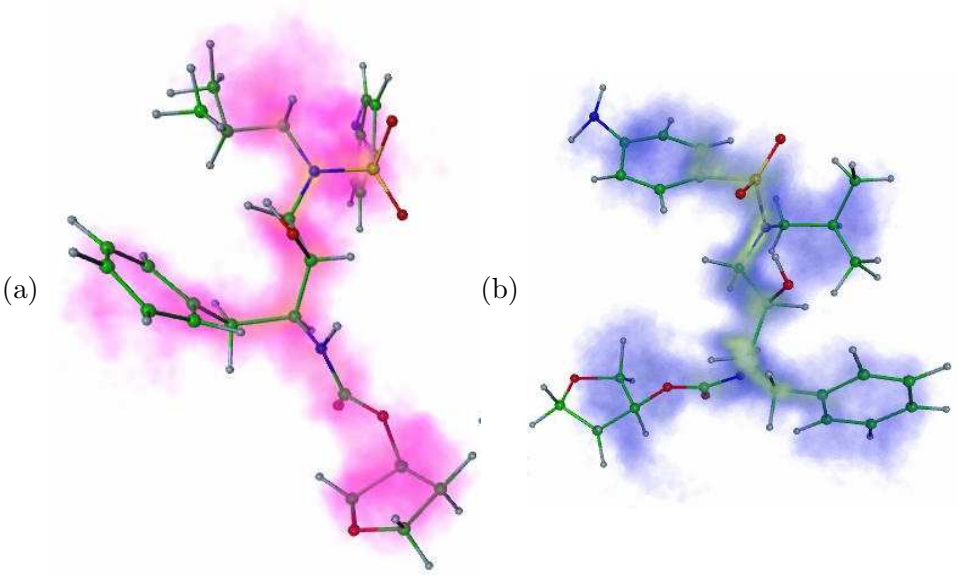

Figure 1: A small biomolecule (inhibitor VX-478 of the enzyme HIV-protease) is shown in a ball and stick representation modeled by the Merck molecular force field. (a) and (b) show two distinct metastable conformations. Thereby, the foggy parts indicate flexibility within the respective metastable conformation.

framework for simulations is a statistical description in terms of the canonical distribution (which reflects the experimental situation) by setting up a classical Hamiltonian that is then analyzed by either *Molecular Dynamics* or *Markov Chain Monte Carlo* (MCMC) methods. Yet, the construction of appropriate force fields, its highly nonlinear dynamics which causes long term trajectories to be ill-conditioned, the multi-timescale nature of biomolecular processes (ranging from 1 femtosecond up to several seconds for proteins [59, 97]), and last but not least the existence of metastable conformations make computer simulations a formidable task.

A wide range of MCMC methods, which are aiming to draw samples from the canonical distribution (and therefore trying to tackle the trapping problem), have been constructed for biomolecular systems [11, 36]. For biomolecules modeled in terms of a Hamiltonian system, methods based on HMC [33, 15] seems to be a natural choice; and in fact, they have been proved to be efficient (e.g., by a combination with Multicanonical Monte Carlo as described in [65]; or by ATHMC bridge sampling [42] which we use as part of UC). Another reason why we choose a HMC-based Metropolis algorithm is its intriguing connection to the transfer operator approach, which aims at identifying metastable conformations. That way, metastable sets of a HMC-based Markov chain are strongly connected to metastable conformations of the biomolecule [112].

The computational task we address could be treated in pure mathemat-

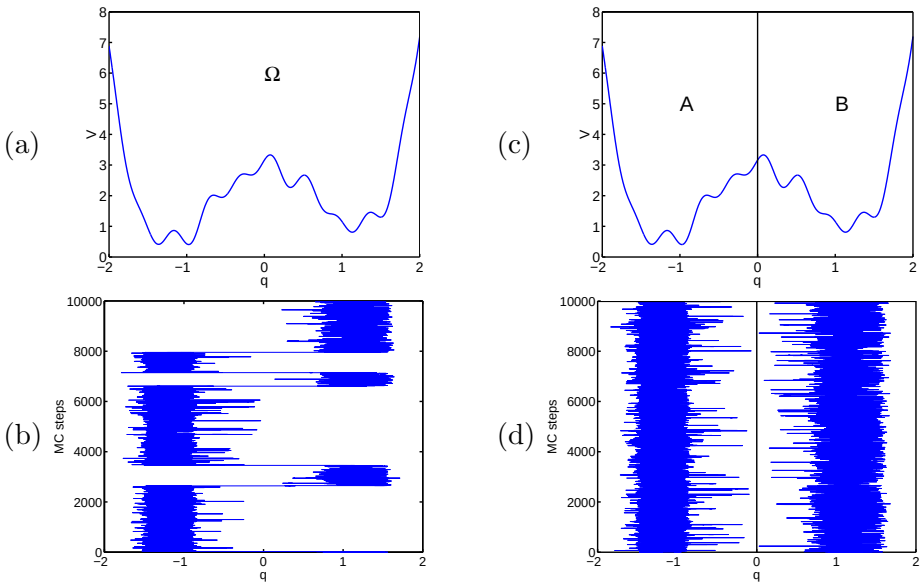

Figure 2: (a) A one-dimensional potential $\mathcal{V}$ with two distinct wells. We can think of them as the two metastable conformations shown in Fig. 1. (b) Sample path of a *slowly mixing* Markov chain on the whole state space that draws samples from a canonical distribution; the 2nd largest eigenvalue 0.9988 of the associated Markov operator is close to 1. (c) The potential $\mathcal{V}$ is separated in two metastable sets A and B. (d) Sample paths of restricted *rapidly mixing* Markov chains on A and B, respectively. The 2nd largest eigenvalues (0.8341 and 0.8129, respectively) of the associated restricted Markov operators are both bounded away from 1.

ical terms. Yet it should be clear, that knowledge about the underlying physical model are vital to understand the output of computer simulations when applied to real applications. Therefore, this thesis starts with some background information about biomolecular modeling, which illustrates the role of our algorithmic approach in this challenging application field.

Uncoupling-Coupling. *Uncoupling-Coupling* (UC), which was first presented in [41, 43], is based on the MCMC methodology and integrates aspects from stochastic complementation [91], simulated annealing [78, 81], macrostate dissection [20], bridge and path sampling techniques [42, 50], high-dimensional discretization [47, 72], transfer operators [112, 113], and dynamical cluster algorithms [30]; it provides a general framework to automatically build up and draw samples from a patchwork of distributions by means of rapidly mixing Markov chains.

The trapping problem we are confronted with is illustrated in Fig. 2 (a) and (b) for the situation of two metastable sets. In contrast, Fig. 2 (c) and (d) demonstrates the rapidly mixing of Markov chains restricted on the metastable sets. In UC, we exploit algorithmically these rapidly mixing properties. The basic idea of UC for this non-hierarchical situation con-

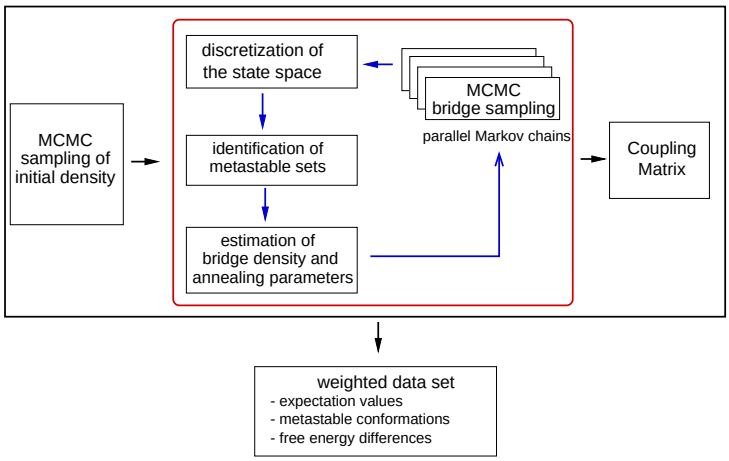

Figure 3: Scheme of the UC algorithm.

sists of the following steps: (i) drawing samples from a high-temperature distribution, (ii) identifying metastable sets from the sample, (iii) setting up bridge distributions on the identified metastable sets (which allow for reliable reweighting to the target distribution at low temperature), and (iv) computing a coupling factor between these two metastable sets by building a "bridge" between them with the help of the initial high-temperature simulation.

In general, we have to deal with a hierarchy of metastable sets. Figure 3 shows the UC scheme for a hierarchical extension of this basic idea. The *uncoupling step* starts with an initial sampling at sufficiently high temperature. Next, the state space is decomposed into metastable sets, and restricted Markov chains (which draw samples from bridge distributions that are annealed towards the low temperature target distribution) are restarted in parallel. This procedure is recursively applied until all annealed bridge distributions have reached the target distribution. That way, we obtain a patchwork of overlapping bridge distributions (for details see Fig. 22 on page 77), which is then analyzed in the *coupling step*. All samples can be reweighted to the target distribution by means of a global auxiliary distribution, which then allows for the computation of expectation values or free energy differences.

By UC we provide a general framework which is not dependent on a particular method for one of its constituent parts (e.g., high-dimensional discretization, identification of metastable sets, choice of bridge distribution, MCMC method,...). For example, one could think of replacing ATHMC bridge sampling by restricted versions of more sophisticated (but also more complex) MCMC methods like multicanonical sampling, parallel tempering, or whichever method one prefers. At the end of Sect. 5, we give a detailed summary of the algorithmic realization used in this thesis for numerical

investigation.

Outline. In Sect. 2 some methodological background about biomolecular computer simulations is introduced. Our framework is a statistical description of a biomolecule in a canonical ensemble. By setting up a Hamiltonian of the biomolecule via a generic molecular force field various Monte Carlo as well as molecular dynamics based methods can be employed to extract expectation values, dynamical information, or structural properties. A phenomenological description of metastable conformations illustrates the trapping problem associated with a direct application of these sampling methods.

Section 3 starts with the development of the mathematical theory. At first, we introduce the basic notation from Markov chain theory for a finite state space together with the classical limit theorems and their rate of convergences, which form the basis of the Monte Carlo method. When generalized to a continuous state space the transition matrix of a Markov chain is replaced by a transition kernel and some associated Markov operator. We describe our concept of metastability, which leads us to a characterization of metastable sets via the eigenvalues and eigenvectors of the associated Markov operator. We combine discretization techniques for high-dimensional continuous state spaces with an identification strategy for finite state spaces, which eventually enables us to identify metastable sets.

In Sect. 4, we introduce *Markov Chain Monte Carlo* (MCMC). We concentrate on the Metropolis-Hastings algorithm, and we have a closer look on convergence rates and convergence estimators. As a MCMC variant, which is a combination of molecular dynamics with the Metropolis-Hastings algorithm, we describe the *Hybrid Monte Carlo* (HMC) method; HMC turns out to be especially suitable for simulation of molecular systems. Since we need an easily manageable bridge sampling method later on as part of UC we also introduce two bridge sampling methods at this point: *Adaptive Temperature* HMC (ATHMC) and *Potential Scaling* HMC (PSHMC). Additionally, we give an overview of popular extensions to MCMC algorithms like *Multicanonical Monte Carlo* or *Parallel Tempering*.

After these intensive preparations we eventually come in Sect. 5 to the core of the UC algorithm. First, the principal idea of UC and the use of bridge distributions is illustrated by a simple example. Then, we describe the hierarchical *Uncoupling* procedure, where identification of metastable sets plays a decisive role. To provide a better understanding of the uncoupling strategy, we investigate the spectra of restricted Markov operators in Sect. 5.2. Our aim in the *Coupling* step is to compute coupling factors between identified metastable sets. To that end we define a coupling matrix such that the desired coupling factors form its invariant distribution. We show how to estimate entries of the coupling matrix via quotients of normal-

izing constants, which in turn are derived from bridge distribution samples. Altogether, this leads us to an overall convergence of expectation values wrt. the target distribution. At the end of this section, we present an overview of the UC scheme.

We apply UC to biomolecules in Sect. 6. First, we give a detailed illustration of the algorithm by analyzing n-butane and n-pentane, which allow to demonstrate the potential advantages and the accuracy of the method. Next, a small alteration in the uncoupling step results in an improved robustness by decomposing the state space into an *essential hierarchy*. Finally, a constituent of green tea, epigallocatechin gallate, leads our investigations towards the goal of analyzing biomolecules of pharmacological interest.

Acknowledgment. First and foremost I thank all the people from the Biocomputing Group at the Free University and the Molecular Dynamics Group at the Konrad-Zuse-Zentrum for their support, encouragement, and company.

Among all these people it is my greatest pleasure to thank one of my advisors, Professor Christof Schütte, for his support and wise advise, as well as his incredible patience. I owe him special gratitude. I also have to express my very special thanks to my advisor Professor Peter Deuflhard, who guided me towards the fields of applied mathematics and interdisciplinary research. As president of the Konrad-Zuse-Zentrum he gave me constant support. The many fruitful discussions with both of my advisors during my time at the Konrad-Zuse-Zentrum were the starting point for this thesis. My special thanks goes to my room-mate Wilhelm Huisinga for helpful comments, questioning theory, and contributing to the pleasant working atmosphere. Also, it is my pleasure to thank Illia Horenko for his encouragement and friendship.

I am indebted to Frank Cordes for collaborative work that enabled the theory presented here to be applied towards real-world problems. I also want to express my thank to Johannes Schmidt-Ehrenberg for producing pictures with Amira, Christian Salzmann for fixing computer problems, Eike Meerbach for pointing out literature, and Ralf Forster and Sonja Waldhausen for proofreading.

Last but not least I want to thank my parents for always supporting me over the years.