Freie Universität Berlin
Fachbereich Mathematik und Informatik
Studiengang Bioinformatik

Master's thesis

Accuracy, stability, convergence of rigorous thermodynamic sampling methods

Alexander Riemer*

2006/08/18

supervised by Dr. Frank Cordes† and Prof. Dr. Paul Wrede‡

*Freie Universität Berlin
†Konrad-Zuse-Zentrum für Informationstechnik, Computational Drug Design Group
‡Charité Universitätsmedizin Berlin, Institute for Molecular Biology and Bioinformatics

Contents

1. Introduction

Ever since the advent of computers, the behavior of microscopic systems of particles has been a primary subject to be studied in computer simulations. The basic methodology for such simulations, most notably Monte Carlo methods [42, 43, 44] and molecular dynamics [36], had already been developed by the 1950s. Computer simulations are hoped to give an understanding of structural or dynamic properties of molecular systems which cannot be observed directly. Of special interest are biochemical molecular systems, where large molecules or clusters of molecules, primarily proteins, act like molecular machines which perform a multitude of different functions in metabolism, transport processes, coordinated movement, immune defense, and signal transduction [38]. With the computational power available in massively parallel computer systems today, it has become possible to explore structure, function and dynamics of ever larger and more complex biochemical systems in a mathematically rigorous way, which has led to the emergence of the discipline of *computational drug design*, the aim of which is to identify novel drug molecules which bind to a given receptor molecule, thus providing new impulses for pharmaceutical research.

The biochemical function of a molecule is basically determined by its 3-dimensional structure. The interaction between two biomolecules, e.g. of a small molecule called a *ligand* with a protein, its *target*, is only possible when the ligand sterically "fits" into the target's binding site. Typical ligands are highly flexible biomolecules that can switch between several metastable conformations, each of which has a different 3-dimensional shape. In order to predict *in silico* whether a ligand binds to a target or not, the ligand's main conformations have to be known. So-called "3D structure generators" such as CONCORD [50] and CORINA [24] use different heuristics on databases of known structures in order to quickly generate representatives of molecular conformations. However, there is no way to estimate the error of such methods, and they return no information at all about the statistical distribution of the generated representatives. Information about the statistical weights of the different conformations, steric variance within a conformation, and transition probabilities between conformations allow the prediction of the dynamics of the intermolecular interaction that is being studied and can only be obtained from thermodynamic simulations. A transition from one conformation to the other can be brought about by (possibly simultaneous and correlated) rotations around single bonds within the molecule. Additionally, the interaction between two biomolecules, which is a dynamic process, induces conformational changes in each of the reactants. This allows them to recruit further interaction partners, which can lead to cascades of interactions, as they are found in the signalling pathways of all eukaryote organisms [3]. In

this thesis, however, the focus will be on exploring the static thermodynamic distribution of biomolecules, and transition processes will be of minor concern.

While biomolecules can be very flexible, they do not assume any state in conformational space with equal probability. The probability of a transition from one molecule configuration to another one is determined by the difference in total energy between the two configurations. Thus, the energy landscape associated with the conformational space defines a statistical distribution which favors low-energy states while disallowing physically forbidden states (e.g. two atoms can neither overlap nor move too far away from each other while connected by a chemical bond).
In order to determine the metastable conformations of a biomolecule, a cluster analysis is performed on a large sample of molecule configurations which is generated in a sampling phase according to the thermodynamically correct distribution at the desired temperature.

The goal of this work is to compare three methods for exploring a molecule's statistical distribution in conformational space with respect to the following questions.

- How fast does a method converge against the "true" distribution?
- How sensitive is it to the choice of initial configurations?
- How closely does a method approximate the "true" distribution in a given time?
- What is the computational cost of each method?

The three methods under consideration are all based on the hybrid Monte Carlo method but use a variety of approaches to accelerate convergence compared to a simple hybrid Monte Carlo approach.
ZIBgridfree uses a meshless partitioning of the conformational space. It was originally implemented as "HuMFree" by Holger Meyer from April 2004 to February 2005 in the course of his master's thesis [45]. The method was developed by Marcus Weber in his doctoral thesis [73]. The Replica Exchange method was added by Alexander Riemer from August to November 2005. It uses independent sampling runs at different temperatures which are allowed to exchange positions at certain intervals. The sampling strategy "ConfJump" [71] uses known minima of the potential energy surface to accelerate sampling by randomly introducing jumps from the proximity of one minimum to the proximity of another one, thus effectively escaping "trapping" within the basin of attraction of one local minimum. It has been developed by Lionel Walter and Marcus Weber and was implemented by Lionel Walter from October 2005 to June 2006. The three techniques have been implemented within a common framework that allows them to be combined easily [47].

This thesis aims at more than just a comparison of different sampling methods. Methods for monitoring convergence of a Markov chain Monte Carlo sampling and for

comparing the quality of different sampling runs needed to be developed in the first place. Almost as a byproduct of this thesis, a graph-theoretic recursive algorithm has been developed to find all rotationally symmetric functional groups in arbitrary biomolecules.
A metric is proposed for measuring the difference between two sampling results (cf. 4.2) in order to be able to compare sampling methods. Further, a variant of this metric is used to define a new semi-empirical convergence indicator based on molecule symmetries (cf. 4.3). The performance of each of the three sampling techniques under consideration is assessed in a series of numerical experiments conducted on three increasingly complex biomolecules (see chapter 5).

1.1. Outline

The following chapter gives an overview of the basics of statistical mechanics and conformation analysis. Both molecular dynamics and Markov chain Monte Carlo methods are presented in chapter 2 along with the hybrid Monte Carlo approach which combines the two.

The three sampling techniques under consideration, ZIBgridfree, Replica Exchange, and ConfJump, will be presented in chapter 3. Theoretical considerations concerning the efficiency of each method compared to pure hybrid Monte Carlo will be given as well.

Chapter 4 deals with the issue of convergence diagnostics, i.e. algorithms for estimating whether the thermodynamic distribution sampled by a molecular simulation is sufficiently close to the molecule's "true" distribution. In addition to that, methods for comparing the different sampling strategies are developed in the same chapter: In section 4.2, a metric for measuring the difference between two sampling results is developed, which is based on histograms over 1-dimensional sampled distributions. A semi empirical convergence criterion that employs knowledge about rotational symmetries in the molecule under consideration is developed based on this metric in section 4.3. In connection with this symmetry criterion, a graph-theoretic algorithm for automatic detection of molecule symmetries is developed in section 4.3.2.

The numerical simulations that were performed for assessing the performance of the different sampling methods are described in chapter 5. A measure for the performance of a sampling technique is developed in section 5.1 based on the metric developed in section 4.2.

Chapter 6 presents the results of the numerical experiments.

Finally, a conclusion is given in chapter 7 along with an outlook, especially regarding the future goal of simulating large molecular systems.

2. Basics

The goal of a conformation analysis is to divide a molecule's conformational space into metastable regions, i.e. to find a partition of the conformation space with the following property: For a given period of time, the transition probability from one region to itself is high, while transition probabilities between any two different metastable regions are minimal. Transition here means physically feasible transitions within that period of time according to the system's dynamics as usually simulated by molecular dynamics (cf. 2.3).

A conformation analysis that divides conformations based on differences in free energy consists of two phases, *sampling* and *clustering*. During the sampling phase, molecule configurations are generated according to the correct thermodynamic distribution of the molecule at a given sampling temperature. Afterwards, these configurations are clustered into metastable regions.

The molecule to be analyzed is given by its N atoms and the bonds between them as well as atom and bond types. A specific 3-dimensional molecule configuration is described as a position state q of the system which is a $3N$-dimensional vector of atom coordinates $q \in \mathbb{R}^{3N} = \Omega$. Let $p \in \mathbb{R}^{3N}$ analogously to q be the collective momentum vector and $\mathcal{M}$, a $3N \times 3N$-matrix with

$$\mathcal{M}_{ij} = \begin{cases} m_{\lfloor (i-1)/3 \rfloor + 1}, & i = j \\ 0, & \text{else} \end{cases}, \tag{2.1}$$

the mass matrix of the molecule, where m_k is the mass of atom k.

In classical mechanics, the total energy of a microstate (q, p) of the system is described by a separable Hamiltonian

$$\begin{aligned} H(q,p) &= V(q) + K(p) \\ &= V(q) + \frac{1}{2} p^\top \mathcal{M} p, \end{aligned} \tag{2.2}$$

which is the sum of the potential energy V and the kinetic energy K. K depends only on the momenta p, while V can be calculated from the positions q alone. While the kinetic energy can be calculated directly from atom masses and momenta, the potential V is approximated by a molecular force field which describes V as the sum of energy terms for binding and non-binding interactions between atoms. All methods discussed in this thesis have been implemented using the Merck molecular force field (MMFF) [31].

2.1. Canonical Ensemble

Rather than simulating the behavior of an individual molecule over time, molecules are simulated within a statistical ensemble, which assigns a probability measure to any point (q, p) in the molecule's phase space $\Gamma = \Omega \times \mathbb{R}^{3N}$. It can be thought of as a large (possibly infinite) number of realizations of a random experiment – in this case observing the microstate of the molecule at an arbitrary point in time [21]. Sampling then consists in generating the ensemble according to the underlying probability distribution, i.e. drawing samples from that distribution. In conformation dynamics, we are interested in ensembles that are in thermodynamic equilibrium, i.e. stationary ensembles, in which the underlying probability density function is time-independent.

The quantities of interest are the expected values of observables over the statistical ensemble. An observable is any function $A : \Gamma \to \mathbb{R}$ that assigns a real number to every point (q, p) in phase space. Examples for observables are total energy H, potential energy V, kinetic energy K, geometric properties such as the value of a particular torsion angle in the molecule (cf. 2.5), but also the degree of membership for a certain metastable conformation (see section 3.2.1).

In the *canonical* or *NVT ensemble*, a molecule is considered a subsystem of fixed volume V that is embedded in an infinitely large thermal bath with a constant temperature T, with which it continuosly exchanges energy, while its mean kinetic energy remains constant in the limit. Since chemical reactions are not allowed in the simulation, the number of particles in the system N is also constant [23].

Phase space microstates (q, p), which describe the system's positions q and momenta p, are distributed according to a Boltzmann distribution:

$$\pi(q, p) = \frac{1}{Q} \exp(-\beta H(q, p)), \tag{2.3}$$

where

$$Q = \int_\Gamma \exp(-\beta H(q, p)) \, \mathrm{d}q \, \mathrm{d}p \tag{2.4}$$

is a normalization factor used to make π a probability distribution. Since it is an integral over all possible states of a $6N$-dimensional system, this *partition function* can only be calculated analytically for the most simple systems.
The temperature enters the equation in the form of $\beta = 1/k_B T$, the inverse temperature. $k_B = 1.38065 \cdot 10^{-23}\mathrm{J/K}$ is Boltzmann's constant.

Since in the canonical ensemble, the system is in constant thermal contact with the environment, there is no limitation on the total energy of an actual state of the system. Thus, any microstate (q, p) is in principle reachable, and every open subset of Γ has a non-zero probability.

Substituting equation 2.2 in equation 2.3 yields

$$\begin{aligned} \pi(q, p) &= \frac{1}{Q_q \cdot Q_p} \exp(-\beta V(q)) \exp(-\beta K(p)) \\ &= \rho(q) \cdot \eta(p). \end{aligned} \tag{2.5}$$

The thermodynamic distribution can be split into independent distributions ρ of positions and η of momenta. In fact, in a conformation analysis, one is only interested in observables in q. Therefore, it is sufficient to sample from the position distribution ρ. The expected value of an observable $A : \Omega \to \mathbb{R}$ is the integral

$$\begin{aligned} \langle A \rangle_\rho &= \int_\Omega A(q)\rho(q)\,\mathrm{d}q \\ &= \frac{1}{Q_q}\int_\Omega A(q)\exp(-\beta V(q))\,\mathrm{d}q \end{aligned} \tag{2.6}$$

over the whole position space Ω.
Let $(q_1, \ldots, q_n)$ be an independent sequence of molecule configurations distributed according to ρ. Then it follows from the law of large numbers that the sample means

$$\bar{A} = \frac{1}{n}\sum_{i=1}^{n} A(q_i) \tag{2.7}$$

converges to the expected value $\langle A \rangle_\rho$ for $n \to \infty$ [55]. In addition to that, it follows from the central limit theorem that with increasing n the sampling error, i.e. the difference between the sampled distribution and ρ, decreases asymptotically in $\mathcal{O}(\sqrt{n})$ almost surely.

However, as the partition function Q_q is unknown and hard to compute, it is not possible to directly draw samples from ρ. The Markov chain Monte Carlo approach (MCMC) generates samples from a probability distribution ρ by constructing an ergodic Markov chain that has ρ as its unique stationary distribution.

2.2. Markov Chain Monte Carlo

The MCMC method was developed in the late 1940s and early 1950s by Metropolis, Ulam, Fermi, von Neumann, Teller et al. for studying the diffusion of neutrons in fissible material and also already for molecular simulations. This work led to the Metropolis algorithm which was published in 1953 [43]. As stated above, the idea is to generate a *dependent* sequence $\left(q^{(n)}\right)$ of random vectors $q^{(n)} \in \Omega$ that are distributed according to ρ for $n \to \infty$.

This Markov chain must be ergodic and meet the criterion of *detailed balance*,

$$\rho(q)P(q \to \tilde{q}) = \rho(\tilde{q})P(\tilde{q} \to q), \tag{2.8}$$

in order for its unique stationary distribution to be the thermodynamically correct equilibrium distribution ρ. The convergence rate is $\mathcal{O}(\sqrt{n})$ just as for independent random vectors [21]. Detailed balance is a restatement of the constraint of thermodynamic equilibrium, i.e. in the limit there is no net flow between any two open subsets A and B of the position space Ω. It is also called *microscopic reversibility*.

Substituting ρ from equation 2.5 into equation 2.8 yields

$$\begin{aligned} \frac{P(q \to \tilde{q})}{P(\tilde{q} \to q)} &= \frac{\exp(-\beta V(\tilde{q}))}{\exp(-\beta V(q))} \\ &= \exp(-\beta \Delta V). \end{aligned} \tag{2.9}$$

The ratio of the probabilities of a transition and its reversal depends only on the potential at the positions q and $\tilde{q}$ and is thus directly computable with only two evaluations of the force field (which, in the case of MMFF and many other force field models, has a computational cost of $\mathcal{O}(N^2)$).

The Metropolis algorithm is one of the most popular MCMC strategies in use today. It splits the transition from a state q to a state $\tilde{q}$ into two steps: A trial step and an acceptance step, in which the new state $\tilde{q}$ is accepted with a probability of P_{acc} and rejected in favor of resting in q with $1 - P_{\text{acc}}$:

$$P(q \to \tilde{q}) = P_{\text{gen}}(q \to \tilde{q}) \cdot P_{\text{acc}}(q \to \tilde{q}). \tag{2.10}$$

The trial step must ensure that every state $q \in \Omega$ is in principle reachable, i.e. any open subset of Ω must have a non-zero probability. If, in addition to that, the trial step is chosen symmetrically, i.e. $P_{\text{gen}}(q \to \tilde{q}) = P_{\text{gen}}(\tilde{q} \to q)$, equation 2.9 becomes

$$\frac{P_{\text{acc}}(q \to \tilde{q})}{P_{\text{acc}}(\tilde{q} \to q)} = \exp(-\beta \Delta V). \tag{2.11}$$

By choosing

$$\begin{aligned} P_{\text{acc}}(q \to \tilde{q}) &= \min\left\{1, \frac{\rho(\tilde{q})}{\rho(q)}\right\} \\ &= \min\left\{1, \exp(-\beta \Delta V)\right\} \end{aligned} \tag{2.12}$$

this equation is easily satisfied. This choice of acceptance probability is called the *Metropolis criterion*.

The resulting Markov chain is *irreducible* because in the trial generation algorithm, any position state is in principle reachable from any other, so that every open subset of Ω has a non-zero probability, i.e. all the states communicate. Due to the possibility of rejecting a trial, the Markov chain is also *aperiodic*. A Markov chain that is both irreducible and aperiodic is ergodic [2]. Therefore, the unique stationary distribution exists. Detailed balance, the algorithm used for generating trials, and the acceptance criterion ensure that it is the Boltzmann distribution at the sampling temperature T.

Constructing an ergodic Markov chain whose transitions are split into trial and acceptance steps with the additional constraint of a symmetric trial step and the Metropolis acceptance criterion allows generating random variables distributed according to ρ without knowledge of the partition function Q_q.

The Metropolis Algorithm

Starting from an initial configuration $q^{(0)}$, repeat the following:

1. From the current state $q^{(i)} = q$, generate a trial $\tilde{q}$ by a perturbation technique that satisfies detailed balance and has a symmetric proposal probability.
2. Calculate the acceptance probability
$$P_{\text{acc}}(q \to \tilde{q}) = \min\left\{1, \exp(-\beta \Delta V)\right\}.$$
3. Generate a uniformly distributed random number $\zeta \in [0, 1)$.
4. Set the new configuration
$$q^{(i+1)} := \begin{cases} \tilde{q}, & \zeta < P_{\text{acc}} \\ q, & \text{else} \end{cases}. \tag{2.13}$$

This is done either for a fixed number of times n or until some error measure indicates convergence. See chapter 4 for a discussion of convergence monitors.

The main problem of this algorithm is that in order to be efficient, it has to propose a new configuration $\tilde{q}$ that is substantially different from q but also has a high probability of being accepted, i.e. $\tilde{q}$ must be of similar or lower potential energy than q but must be as far away from q as possible so that the algorithm will cover a large amount of space in a short time. The efficiency of any sampling strategy that is based on the Metropolis algorithm or its generalization, the Metropolis-Hastings algorithm [33], is dependent on two quantities:

- the computational cost of the trial step and
- the average acceptance probability or alternatively the
$$\text{acceptance ratio} = \frac{\#\text{ accepted steps}}{n}. \tag{2.14}$$

The hybrid Monte Carlo approach employs molecular dynamics to generate trials with a high acceptance ratio at an acceptable computational cost.

2.3. Molecular Dynamics

Molecular dynamics (MD) [21, 23, 37, 55] simulates the behavior of a molecular system over time as a many-body system in terms of classical mechanics, i.e. it solves the Newtonian or Hamiltonian equations of motion, respectively, for the given system by numerical integration. Quantum effects such as induced changes in the

electronic density of a molecule are ignored. In contrast to Monte Carlo approaches, molecular dynamics is a deterministic method.

MD simulates the motion of the system under the influence of a specified force field (the potential energy function V). Given an ideal integrator, MD reproduces the correct physical behavior of the system over time, within the limitations of classical mechanics and the force field used to describe interactions between atoms.

For a system of N atoms the equations of motion can be written as a set of two differential equations:

$$\begin{aligned} v(t) &= \dot{q}(t), \\ F(q) &= \mathcal{M}\dot{v}(t) = -\nabla V\left(q(t)\right). \end{aligned} \tag{2.15}$$

The velocity v of a particle, the product of its mass m and momentum p, is the derivative of that particle's position with respect to time. The force F acting on a particle is the negative gradient of the potential at the particle's position. Additional terms are sometimes added to the force F to simulate interactions with the environment.

Since analytic solutions for this system of differential equations are known only for very simple systems, it is necessary to employ numerical integrators. A very popular integrator used in molecular dynamics is the velocity Verlet integrator [23, 65]. Like the Euler integrator and other Verlet-type integrators it is derived from a Taylor expansion of the trajectory $q(t)$. Verlet integrators are based on a second-order Taylor approximation in which the third-order terms cancel thus leaving a local error in position of $\mathcal{O}(\tau^4)$, where τ is the length of the integration step [69]. The velocity Verlet integrator updates position q and velocity $\dot{q} = v$ according to the following equations:

$$\begin{aligned} q(t+\tau) &= q(t) + \tau\dot{q}(t) + \frac{\tau^2}{2}\mathcal{M}^{-1}F(t), \\ \dot{q}(t+\tau) &= \dot{q}(t) + \frac{\tau}{2}\mathcal{M}^{-1}\left(F(t) + F(t+\tau)\right). \end{aligned} \tag{2.16}$$

The time step length τ is typically on the order of 1fs $= 10^{-15}$s so as to be able to correctly simulate high-frequency processes such as bond vibrations or bond-angle oscillations. By repeatedly applying equations 2.16, starting from some initial configuration $(q^{(0)}, \dot{q}^{(0)})$, a trajectory is generated which describes the change of the dynamic variables with time.

In contrast to the MCMC approach, which samples the canonical ensemble, a molecular dynamics trajectory samples a part of the *microcanonical* or NVE-ensemble, in which number of particles N, volume V and total energy $E = H$ is constant. Since states of constant energy are not necessarily connected, an MD trajectory is not an ergodic Markov chain. Moreover, an integrator that exactly conserves energy is theoretically impossible [21]. However, symplectic integrators such as varieties of the Verlet integrator conserve the total energy of the system on average. A variety of approaches exists for molecular dynamics in different ensembles

such as the NVT ensemble, e.g. by rescaling momenta or adding correcting terms to the force F in equation 2.16 [23, 55].

The average of an observable $A : \Omega \to \mathbb{R}$ on an MD trajectory of n time steps starting at time $t_0 = 0$ is calculated as the time average over the trajectory:

$$\bar{A} = \frac{1}{n} \sum_{i=0}^{n-1} A\left(q(i\tau)\right). \tag{2.17}$$

The ergodic hypothesis [21, 49, 55], which is one of the fundamental axioms of statistical mechanics, posits that a molecular system will assume all possible microstates (q, p) within some *ergodic component* $\hat{\Omega} \subseteq \Omega$ (which contains all points that are compatible with the constraint of conservation of energy (or conservation of energy on average)) for $t \to \infty$ $(n \to \infty)$. Therefore, the unique time average of an observable A exists in $\hat{\Omega}$. The ergodic hypothesis states further that $\bar{A}$, as calculated by equation 2.17, converges towards the expected value of A over the microcanonical ensemble,

$$\bar{A}_\infty = \langle A \rangle_{\rho_{NVE}}. \tag{2.18}$$

In practice, the ergodic hypothesis can usually not be proven and may even be false for special cases.

While molecular dynamics has a number of advantages, such as simulating "true" dynamics, which allows estimating kinetic properties of the system, it also has severe disadvantages, especially when applied to conformation dynamics:

- It has a high error amplification due to numerical errors, effectively disallowing simulations over a long period of time,
- a very low time step length τ, because of which an MD trajectory can only cover a small region of phase space in a given period of time, and
- since MD simulations model the system's true dynamics, they tend to get trapped within basins of attraction of local minima (metastabilities) for long times.

2.4. Hybrid Monte Carlo

The hybrid Monte Carlo strategy (HMC) [7, 12, 19, 21] combines Markov chain Monte Carlo and molecular dynamics in order to efficiently generate samples from the canonical ensemble of the molecule at the given temperature T. HMC is a Markov chain Monte Carlo method which is based on the Metropolis-Hastings algorithm [33], which, in contrast to the Metropolis algorithm, does not require a symmetric trial step. The Metropolis-Hastings algorithm satisfies equation 2.9 and thus detailed

balance by choosing the following acceptance criterion:

$$\begin{aligned} P_{\text{acc}}(q \to \tilde{q}) &= \min\left\{1, \frac{\rho(\tilde{q})P_{\text{gen}}(\tilde{q} \to q)}{\rho(q)P_{\text{gen}}(q \to \tilde{q})}\right\} \\ &= \min\left\{1, \exp(-\beta\Delta V)\frac{P_{\text{gen}}(\tilde{q} \to q)}{P_{\text{gen}}(q \to \tilde{q})}\right\}. \end{aligned} \tag{2.19}$$

In hybrid Monte Carlo, the trial step consists in a short MD trajectory. Trial generation in this way has a moderate computational effort but also a high probability of being accepted in the subsequent acceptance step since MD on average conserves the system's total energy thus only generating physically meaningful configurations. The method requires that MD simulations be performed with an integrator that is both time-reversible and preserves phase space volume [21]. The symplectic velocity Verlet integrator (given in equations 2.16) has both properties.

For every step in the Markov chain, a short MD trajectory is computed, starting from the current position state q and a randomly generated momentum state p, which is distributed according to the Boltzmann distribution η (see equation 2.5). Since molecular dynamics is deterministic, the outcome $(\tilde{q}, \tilde{p})$ of the MD simulation depends only on the initial state (q, p). As the initial position state q is given, the trial probability in equation 2.19 depends only on the distribution of the initial momenta p:

$$P_{\text{gen}}(q \to \tilde{q}) = \eta(p) = \frac{1}{Q_p}\exp(-\beta K(p)). \tag{2.20}$$

The integrator used in the MD simulation is reversible, i.e. if the state $(\tilde{q}, \tilde{p})$ is generated from (q, p) in l iterations, then l integration steps starting from $(\tilde{q}, -\tilde{p})$ will generate the state $(q, -p)$. Therefore, the probability of generating q from $\tilde{q}$ depends only on the distribution of the start momenta $-\tilde{p}$:

$$P_{\text{gen}}(\tilde{q} \to q) = \eta(-\tilde{p}) = \frac{1}{Q_p}\exp(-\beta K(-\tilde{p})). \tag{2.21}$$

The kinetic energy K is a quadratic function in the momenta p (see equation 2.2). Therefore, $K(-\tilde{p}) = K(\tilde{p})$. Thus, equation 2.19 becomes

$$\begin{aligned} P_{\text{acc}}(q \to \tilde{q}) &= \min\left\{1, \exp(-\beta\Delta V)\frac{\exp(-\beta K(\tilde{p}))}{\exp(-\beta K(p))}\right\} \\ &= \min\left\{1, \exp(-\beta\Delta H)\right\}. \end{aligned} \tag{2.22}$$

The acceptance probability of a hybrid Monte Carlo step depends on the total energy H. However, the trial momentum $\tilde{p}$ is discarded (as is the whole MD trajectory), and only the next position (q or $\tilde{q}$) needs to be stored. It is worth noting that if the system's total energy H is exactly conserved in the trajectory, the trial is accepted with probability 1.

The HMC Algorithm

Starting from an initial configuration $q^{(0)}$, repeat the following:

1. Draw a random collective momentum vector p from the Boltzmann distribution η for the simulation temperature T.
2. Let $q = q^{(i)}$ denote the current position state. Run a short MD simulation of a fixed length l starting from (q, p). Let $(\tilde{q}, \tilde{p})$ denote the microstate after l iterations.
3. Calculate the acceptance probability
$$P_{\text{acc}}(q \to \tilde{q}) = \min\left\{1, \exp(-\beta \Delta H)\right\}.$$
4. Generate a uniformly distributed random number $\zeta \in [0, 1)$.
5. Set the new configuration
$$q^{(i+1)} := \begin{cases} \tilde{q}, & \zeta < P_{\text{acc}} \\ q, & \text{else} \end{cases}. \tag{2.23}$$

Again, this is done either for a fixed number of times n or until convergence is detected.

2.5. Conformational space

After generating a sufficient number of samples from the canonical ensemble of a molecule, the metastabilities in the molecule's position space have to be identified. Generally, metastabilities are *almost invariant* subsets of the state space, i.e. non-equilibrium states which are stable for longer periods of time. When considering the dynamics of the system under consideration for some given period of time, the transition probability from any metastable region to itself is high while transitions between two different metastable regions occur with low probability; for a formal definition of almost invariant subsets see [57]. In order to facilitate the metastability analysis, the conformation space has to be defined in a meaningful way.

A molecular system of $3N$ particles has $3N - 6$ degrees of freedom. However, a molecule's metastabilities can usually be described in terms of very few degrees of freedom. Since in metastability analysis, one is interested in slow transition processes, bond-angle and bond-length oscillations can be neglected due to the fact that their frequencies are very high. Thus, it is sufficient to define a molecule's conformational space in terms of a selection of its dihedral angles, i.e. in terms of rotations around chemical bonds.

Let a_1, a_2, a_3, and a_4 be four atoms in the molecule under consideration which are connected by the chemical bonds (a_1, a_2), (a_2, a_3) and (a_3, a_4) (and possibly others).

The *dihedral angle* defined by the dihedral (a_1, a_2, a_3, a_4) is the angle between the planes spanned by the triangles (a_1, a_2, a_3) and (a_2, a_3, a_4) (see fig. 2.1). Rotations around the bond (a_2, a_3) lead to different values for the dihedral angle. It is also called a torsion angle.

Dihedral coordinates are invariant to rotation and translation of the whole system, which is good for conformation analysis, since absolute atom positions are irrelevant. The conformational space can be further reduced by omitting those dihedral angles that have no potential to define metastabilities, i.e. dihedrals that are either completely rigid or extremely flexible. Consequently, a dihedral (a_1, a_2, a_3, a_4) is excluded if

- a_1 or a_4 is hydrogen (such a dihedral's flexibility is almost unrestricted), or
- the bond $(a2, a3)$ is not a single bond (only single bonds are rotatable).

In addition to that, no two dihedrals are used for defining the conformational space that describe the same single bond. The set of dihedrals obtained by removing the ultra-flexible and inflexible dihedrals restricted to one dihedral per single bond will be called the set of important or "heavy" dihedrals throughout this thesis. Figure 2.1 illustrates the concept on the butane molecule.

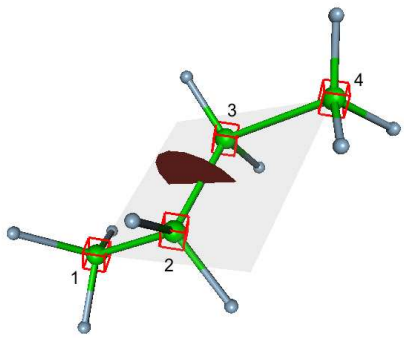

Figure 2.1.: The only "heavy" dihedral angle in the butane molecule. Note that the dihedral angle is defined as the angle between the planes spanned by atoms 1,2,3 and atoms 2,3,4, respectively.

Since torsion angles are a cyclic measure, the Euclidian distance is not a metric in torsion angle space [45]. Let $\varphi, \psi \in [0, 2\pi)^d$ be two configurations given as points in the conformational space defined by the molecule's d heavy dihedrals. The Euclidian distance on the torus between φ and ψ is

$$\delta(\varphi, \psi) = \sqrt{\sum_{i=1}^{d} (\varphi_i \ominus \psi_i)^2}, \tag{2.24}$$

with

$$\varphi_i \ominus \psi_i = \begin{cases} 2\pi - (\varphi_i - \psi_i), & \varphi_i - \psi_i > \pi \\ 2\pi + (\varphi_i - \psi_i), & \varphi_i - \psi_i < -\pi \\ \varphi_i - \psi_i, & \text{else} \end{cases} . \tag{2.25}$$

With this intuitive metric, two angles can differ by no more than π (180°). Consequently, the difference between two points in conformational space can be no greater than $\sqrt{\pi^2 d}$.

3. Sampling strategies

3.1. Overview

The potential energy surface V of a biomolecule is usually very rough, i.e. regions with a low potential energy are separated by high energy barriers. In a Markov chain Monte Carlo sampling of the canonical ensemble at physiologically relevant temperatures (around 300K) this hinders transitions between different low-energy regions due to the fact that the acceptance criterion for an MCMC step depends on the potential energy V (or the total energy $H = V + K$ in the case of HMC). Metastabilities in configuration space, the very phenomenon that is examined by conformation dynamics, make the sampling process slow by causing a "trapping" effect, where the sampling generates configurations from within the basin of attraction of one local minimum for a long time, while the interesting transitions between different local minima, which correspond to conformational changes, are observed very rarely. This effect is known as "broken ergodicity" and can lead to a very slow convergence of the hybrid Monte Carlo method.

In order to overcome an energy barrier between two adjacent metastable regions, the HMC algorithm must by chance generate a vector of initial momenta p which both

- "points towards" an energy barrier that is not too far away for a short MD trajectory to pass and
- infuses the system with enough energy so as to allow the MD trajectory to actually pass through the region of high potential energy rather than being diverted.

Finally, the end point of the MD trajectory has to be accepted.

Generating random momenta that carry the system in one HMC step (or very few steps with some accepted high-energy states on the path) from the basin of attraction of one local minimum of the potential energy surface to that of another becomes more and more unlikely with increasing size and complexity of the molecule. In fact, a molecule's complexity (in terms of containing certain "complex" structures) is far more important than its size as is illustrated by cyclohexane, which has only 18 atoms and whose configurations can be described in terms of only three torsion angles (disregarding high-frequency oscillations as well as translation and rotation of the whole molecule). However, cyclohexane's metastabilities are separated by very high energy barriers, and it is extremely difficult to accurately sample its Boltzmann

distribution at 300K. Figure 3.1 shows the major conformations of the molecule and their thermodynamic distribution in conformational space.

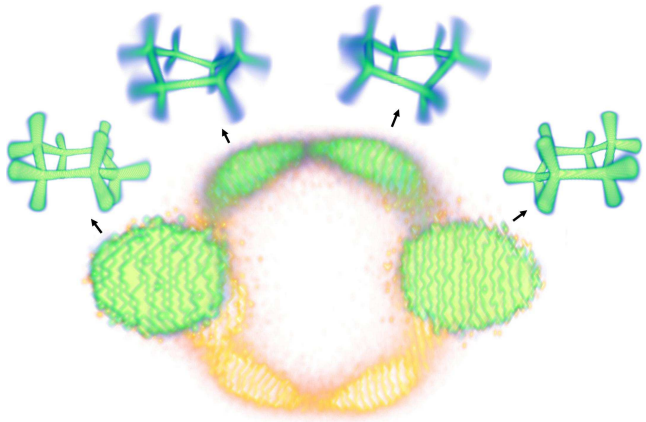

Figure 3.1.: Thermodynamic distribution of cyclohexane in its conformational space at 300K. Also shown are four conformations (left-most and right-most: 'chair' conformations, center: two 'twist' conformations) corresponding to different metastable regions in conformational space. The two chair conformations have a combined thermodynamic weight of more than 99%.

Of course, a sampling should need as few simulation steps as possible, i.e. generate a minimum amount of redundant data. On the other hand, all major local minima of the potential energy surface (metastabilities) have to be found and assigned thermodynamically correct weights.

In order to accelerate the sampling, different so-called *umbrella strategies* [66, 67, 68] can be employed to systematically modify the probability distribution to be sampled (e.g. by lowering energy barriers) so as to make the Markov chains "mix" faster. These modifications are designed in such a way that their effect can be eliminated from the resulting trajectories by reweighting, which allows estimating the original, unmodified probability distribution.
Replica Exchange and ConfJump use different systematic potential modifications that facilitate sampling by "flattening" or "smoothing" the thermodynamic distribution, while ZIBgridfree uses a soft meshless partitioning of the conformation space, where ideally each subset of the conformation space does not assign a high weight to more than one major local minimum. Using a soft partitioning, i.e. restricting the sampling to certain subsets of the conformational space by erecting artificial energy barriers, allows estimating transition probabilities between the partitions of conformational space, which, in turn, allow correct reweighting of the different subsamplings to form a combined estimate of the Boltzmann distribution at the sampling temperature.

3.2. ZIBgridfree

In large molecular systems with high-dimensional conformational spaces, the potential energy surface is very rough. It is desirable to be able to discretize space in a way that does not entail an exponential computational cost and sample the thermodynamic density on each partition, separately. ZIBgridfree [45, 73, 75] uses a meshless discretization of the conformational space. By sampling different subsets of the conformational space separately, less metastabilities occur within each subset, and thus, the HMC sampling converges fast. Instead of a crisp partitioning of the conformational space (as e.g. a Voronoi tessellation), ZIBgridfree employs a partitioning that is function-based ("fuzzy") rather than set-based. This is achieved by adding softly limiting functions to the potential energy V. These potential modifications do not have the effect of smoothing the potential, but rather, they (softly) restrict the sampling to certain regions in conformational space so that it is easier to sample all physically relevant regions of the conformational space, i.e. all regions with a high statistical weight. The potential modifications are defined adaptively with respect to covering a large amount of the physically relevant regions, which are identified by a presampling.

Rather than using one Markov chain to sample the unmodified potential, ZIBgridfree subdivides the conformational space by defining a potential modification for each partition and then launches one Markov chain for each modified potential energy function. ZIBgridfree pursues an *uncoupling-coupling* strategy [22]. In an *uncoupling* step the conformational space is partitioned, and subsequently, for each partition of the space a distribution is sampled that has a lower variance than the original distribution because it contains fewer local minima. Due to the lowered variance, each sampling converges fast. Afterwards, the samplings of the different partitions of the conformational space are reweighted and combined in the *coupling* step so that the resulting linear combination of the sampled partial densities is an approximation of the target distribution.

3.2.1. Soft-characteristic molecular conformations

ZIBgridfree is based on the concept of *conformation dynamics* as described by Deuflhard, Schütte et al. [16, 17, 59]. Conformations are defined in terms of almost-characteristic membership functions rather than classical sets in conformational space. The goal then is to identify a set of C conformations defined by membership functions $\chi_1, \ldots, \chi_C : \Omega \to [0, 1]$ (see [18]). The functions χ_i are non-negative, i.e. for all $q \in \Omega$:

$$\chi_i(q) \geq 0, \quad i = 1, \ldots, C, \tag{3.1}$$

and form a partition of unity,

$$\forall q \in \Omega : \sum_{i=1}^{C} \chi_i(q) = 1. \tag{3.2}$$

Conformations defined on the basis of membership functions χ_i have overlapping partial density functions $\tilde{\rho}_i$ associated with them:

$$\tilde{\rho}_i(q) = \frac{\chi_i(q)\rho(q)}{\tilde{w}_i}, \tag{3.3}$$

where the partition functions $\tilde{w}_i = \int_\Omega \chi_i(q)\rho(q)\,\mathrm{d}q$ are the thermodynamic weights, and ρ is the spatial Boltzmann distribution (see equation 2.5). Note that the membership functions χ_i are by their definition observables over the canonical ensemble under consideration, as each function χ_i assigns a real number to every point $q \in \Omega$. The thermodynamic weight $\tilde{w}_i$ is then the expected value of χ_i under the distribution ρ (see equation 2.6). While the integral

$$Z_i = \int_\Omega \chi_i(q)\,\mathrm{d}q \tag{3.4}$$

is the fraction of the conformational space "covered" by conformation i, $\tilde{w}_i$ is the fraction of the thermodynamic density over Ω that conformation i accounts for. Note that in the case of a set-based approach with conformations $S_1, \ldots, S_C \subset \Omega$ with characteristic functions

$$\xi_i(q) = \begin{cases} 1, & q \in S_i \\ 0, & \text{else} \end{cases} \tag{3.5}$$

replacing the membership functions χ_i, the "coverage" integral Z_i is equal to the volume of S_i, and $\tilde{w}_i$ is the same as the integral of ρ over S_i. Figure 3.2 illustrates the difference between a crisp and a soft discretization of conformation space on a 1-dimensional example. Figure 3.3 shows the decomposition of a Boltzmann distribution over Ω into partial density functions $\tilde{\rho}_i$ by the soft partitioning functions shown in figure 3.2.

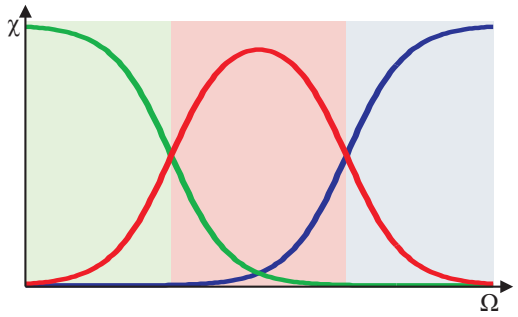

Figure 3.2.: Partitioning of a set Ω into three "subsets" either via soft-characteristic functions (lines) or a crisp partitioning into classical sets (boxes).

The expected value of a spatial observable $A : \Omega \to \mathbb{R}$ can be calculated separately for each function-based conformation χ_i under the partial density $\tilde{\rho}_i$ of that

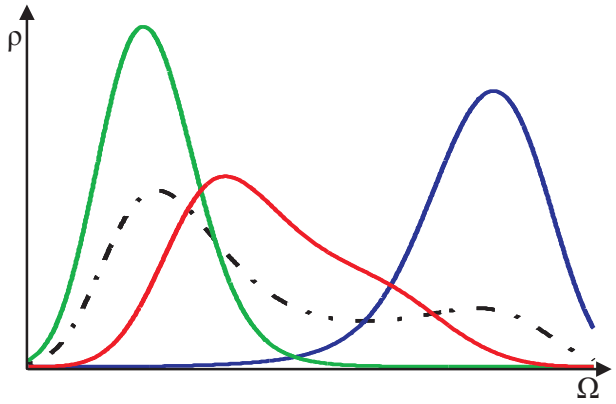

Figure 3.3.: A Boltzmann distribution (dashed black line) over Ω and the partial density functions derived from it via the three soft partitioning functions from figure 3.2.

conformation (see equation 3.3):

$$\begin{aligned} \langle A \rangle_{\rho,\chi_i} &= \frac{1}{\tilde{w}_i} \langle A\chi_i \rangle_\rho \\ &= \frac{1}{\tilde{w}_i} \int_\Omega A(q)\chi_i(q)\rho(q)\,\mathrm{d}q. \end{aligned} \tag{3.6}$$

The soft-characteristic conformations χ_i can be interpreted as *macrostates* in configuration space that are fully described by modified potential energy functions $\tilde{V}_i$ with

$$\tilde{V}_i(q) = V(q) - \frac{1}{\beta} \ln\left(\chi_i(q)\right). \tag{3.7}$$

This follows from an interpretation of the partial density functions substituting ρ from equation 2.5 in equation 3.3 (see also [73]):

$$\begin{aligned} \frac{1}{\tilde{w}_i}\chi_i(q)\rho(q) &= \frac{1}{\tilde{w}_i Q_q}\chi_i(q)\exp\left(-\beta V(q)\right) \\ &= \frac{1}{\tilde{w}_i Q_q}\exp\left(-\beta(V(q) - \frac{1}{\beta}\ln\left(\chi_i(q)\right))\right). \end{aligned} \tag{3.8}$$

3.2.2. Partitioning by membership basis functions

A central concept in ZIBgridfree is the approximation of the unknown conformation membership functions χ_i from a function basis $\phi_1, \ldots, \phi_s : \Omega \to [0,1]$. If this function basis has the same properties as the membership functions $\chi_1, \ldots, \chi_C$, namely non-negativity (equation 3.1) and partition of unity (equation 3.2), then each conformation χ_l is a convex combination of the basis functions ϕ_i (see [73]).

$$\chi_l = \sum_{i=1}^{s} \chi_{\mathrm{disc}}(i,l)\phi_i, \quad i = 1, \ldots, C, \tag{3.9}$$

where χ_{disc} is the matrix of linear combination factors which is row-stoachastic, i.e.

$$\sum_{l=1}^{s} \chi_{\text{disc}}(i,l) = 1, \quad i = 1, \ldots, C. \tag{3.10}$$

The number of basis functions s is chosen sufficiently greater than the anticipated number of conformations C.

The basis functions form a soft partitioning of Ω as well but are not necessarily metastable. Consequently, the concepts of thermodynamic weights (defined analogously to equation 3.3) and potential modifications as defined in equation 3.7 apply to the basis functions as well. From here onward, V_i will denote the modified potential corresponding to the basis function ϕ_i, and w_i will denote the thermodynamic weight of ϕ_i. ρ_i will denote the partial density function corresponding to ϕ_i.

The goal of cluster analysis will be to identify both the correct number of clusters C and the matrix χ_{disc} of linear combination factors from samplings of the partial densities ρ_i associated with the basis functions ϕ_i so as to obtain the set of membership functions χ_l by applying equation 3.9. The membership basis functions are also referred to as *shape functions* in meshless methods.

ZIBgridfree defines the membership basis functions ϕ_i by means of a set of defining *nodes* $\{k_1, \ldots, k_s\} \subset \Omega$. Nodes are placed equidistantly in the relevant part of configuration space which is identified beforehand in a presampling at a high temperature (cf. 3.2.4). As most of configuration space is physically "forbidden" due to extremely high potential energy in regions where atoms either overlap or are too far away from each other to maintain chemical bonds, the amount of "relevant" (i.e. physically allowed) space that has to be covered is hoped not to grow exponentially with the number of atoms N [45, 73]. The definition of basis functions as

$$\phi_i := \frac{W_i}{\sum_{j=1}^{s} W_j}, \quad i = 1, \ldots, C, \tag{3.11}$$

follows the partition of unity method of Shepard [62]. With *radial basis functions* W_i with

$$W_i(q) = \exp\left(-\alpha\, \delta^2(q, k_i)\right), \quad i = 1, \ldots, C, \tag{3.12}$$

the basis functions ϕ_i are unimodal, non-negative, and continuously differentiable and form a partition of unity [73]. $\delta^2(q, k_i)$ is the squared distance of the projections of q and k_i into the space of heavy dihedrals as defined in section 2.5. The shape parameter α is chosen in dependence on the number of nodes s and the given node distance θ. The meshfree discretization using soft-characteristic basis functions ϕ_i is a generalized Voronoi tessellation which converges towards a Voronoi tessellation for $\alpha \to \infty$. The basis functions ϕ_i have their maximum at the defining node k_i and decrease exponentially with growing distance from k_i. Consequently, the modified potential V_i is identical to V at position k_i while the difference between V and V_i increases exponentially in the distance from k_i.

ZIBgridfree samples the Boltzmann distributions corresponding to the modified potentials V_i separately, which can even be done parallelly as each V_i can be evaluated at every position $q \in \Omega$ independently of all V_j with $j \neq i$. The current implementation of ZIBgridfree [47] supports both serial and (massively) parallel sampling. The algorithm is described in more detail in the following sections.

3.2.3. The algorithm (outline)

1. Perform a (relatively short) presampling at a high temperature on the original potential V (cf. section 3.2.4). Let Q denote the set of generated configurations.

2. Place nodes $k_1, \ldots, k_s \in Q$ approximately equidistantly within relevant regions of Ω only (cf. 3.2.5).

3. Define a meshless soft discretization by constructing basis functions $\phi_1, \ldots \phi_s : \Omega \to [0, 1]$ from $k_1, \ldots, k_s$ as described in section 3.2.2.

4. Perform HMC sampling of each partial density ρ_i which is induced by the modified potential V_i corresponding to the basis function ϕ_i (cf. 3.2.6).

5. Accumulate the transition matrix $\bar{P}$ and the overlap matrix $\bar{S}$ from the trajectories generated in step 4.

6. Calculate thermodynamic weights of the partial densities ρ_i (cf. 3.2.7).

7. Determine the number of conformations C and the matrix of linear combination factors χ_{disc} by Robust Perron Cluster Analysis in order to obtain conformation membership functions $\chi_1, \ldots, \chi_C$ from the membership basis $\phi_1, \ldots, \phi_s$ (see 3.2.8.

3.2.4. Presampling

In the first step of the algorithm, a presampling at a high temperature is performed on the unmodified potential. The sampled distribution, which differs from the Boltzmann distribution at temperature T only in the parameter β (see section 2.1), shares all important minima and maxima with the target distribution while being generally more variable. Therefore, the Markov chains exploring this distribution in HMC sampling are expected to mix better. The increased variability stems from the increased temperature which causes the random initial momenta for the HMC steps to be higher on average than when sampling at temperature T. The presampling effectively yields a rough overview of the potential energy landscape which allows identification of the low-energy regions. Convergence of the presampling is monitored by a Gelman-Rubin criterion (see section 4.1) which is more tolerant than for the regular sampling. After all, the goal is not estimation of the high-temperature distribution but merely finding all relevant regions in conformation space.

3.2.5. Choice of nodes

In order to reflect only physically relevant parts of the conformational space, nodes are chosen from the presampling trajectory. As energetically forbidden and very unlikely states are not assumed during presampling, nodes are only generated within the relevant regions of conformational space. This results in a problem-adaptive discretization of the conformational space as the modified potentials V_i defined by the nodes k_i will assign very high values to all regions that have never been visited during presampling. Therefore, it is crucial that the presampling discovers all low-energy regions. Meyer [45] proposed an iterative strategy for finding the optimal presampling temperature. In practice, it suffices to choose the presampling temperature high enough, as no error results from the inclusion of regions that have a low statistical weight at temperature T but are assumed easily at the presampling temperature. The computational overhead from sampling a few low-weight partial densities is probably not very great overall, especially when compared to the overhead resulting from multiple presamplings at different temperatures. The node selection algorithm of ZIBgridfree (see [45, 73]) chooses nodes from the presampling trajectory that are spaced approximately equidistantly and no closer to each other than the given minimum node distance θ:
Let Q denote the set of molecule configurations generated in presampling and Q^* the list of nodes which is initially empty. Further, let L be another list of molecule configurations which is also initially empty.

1. Pick an arbitrary configuration $k_1 \in Q$ and add k_1 to Q^*.
2. Calculate the distances of all geometries $q \in Q$ to k_1 (in heavy dihedral space; cf. section 2.5).
3. Add all configurations $q \in Q$ to L, sorted by their distance from k_1.
4. Repeat:
   a) Let k denote the configuration that was added to Q^* most recently. Remove from L all geometries q with $\delta(q, k) < \theta$.
   b) If L is not empty, add the first element of L to Q^* and remove it from L.

   until L is empty.

Note that step 4a does not require testing all elements of L for their distance to the node k that was added in the previous step of the algorithm. The reason for this is as follows:
Let $a = \delta(k, k_1)$ be the distance of the node k that was added to Q^* in the previous step to the initial node k_1. For all molecule configuration $q \in Q$ with $\delta(q, k_1) > \theta + a$, it follows that $\delta(q, k) > \theta$, because δ is a metric and the triangle inequality

$$\delta(k_1, k) + \delta(k, q) = a + \delta(k, q) > \delta(k_1, q) \tag{3.13}$$

holds for all $k, k_1, q \in \Omega$. Consequently, in step 4a, the search in the list L can be stopped when the first element with $\delta(q, k_1) > a$ is encountered. Therefore, the total computational cost of node selection is the sum of $\mathcal{O}(|Q| \log |Q|)$ for sorting all $|Q|$ configurations and $\mathcal{O}(|Q|)$ for the loop in step 4 which per iteration removes at least one element from L and examines each element at most twice. This yields a total computational cost of $\mathcal{O}(|Q| \log |Q|)$ for node generation.

This moderately low computational cost (as the presampling is kept fairly short) allows controlling the overall computational cost of the sampling phase by limiting the number of nodes from the outset. This is done by repeating the node selection algorithm for a different values of θ which are generated by a binary search until the target number of nodes is (approximately) reached. Compared to the sampling or even the presampling, the cost of node selection is negligible.

3.2.6. Sampling of partial densities

ZMFree uses different kinds of sampling in order to obtain information about

a) the overlap between two basis functions ϕ_i and ϕ_j,

$$\bar{S}(i,j) = \phi_i(q)\phi_j(q)\rho(q)\,\mathrm{d}q, \tag{3.14}$$

and

b) the transition probabilities

$$P(i,j) = \frac{\int_\Omega \phi_i(q) P^\tau \phi_j(q)\,\mathrm{d}q}{\int_\Omega \phi_i(q)\rho(q)\,\mathrm{d}q}, \tag{3.15}$$

where P^τ is the Markov operator that describes the propagation of the system in the canonical ensemble in time span τ; see [58]. As the partition membership functions ϕ_i are soft-characteristic functions, equation 3.15 describes a "fuzzy" concept of transition probabilities as well.

Both the overlap and the transition matrix can be used for reweighting the partial densities (cf. 3.2.8).

Figure 3.4 shows the sampling scheme used by ZIBgridfree for each partial density ρ_i. A regular HMC sampling is performed to generate a sequence of configurations $q_1, \ldots, q_n$, and from every state q_j, a short MD trajectory is launched on the original, unmodified potential to obtain a new position state q'_j. The sequence $(q_1, \ldots, q_n)$ is a realization of a Markov chain, as it is generated by HMC sampling. The sequence $(q'_1, \ldots, q'_n)$, however, is also a realization of a Markov chain because molecular dynamics is deterministic. Thus, the transition from q'_j to q'_{j+1} is determined solely by the transition from q_j to q_{j+1} (and the random initial momenta drawn in steps j and $j+1$), as it can be realized by (deterministically) moving from q'_j to q_j, then from q_j to q_{j+1}, which is the HMC step, and finally from q_{j+1} to q'_{j+1}, deterministically once more. The "horizontal" chain (states q_j) can be used to

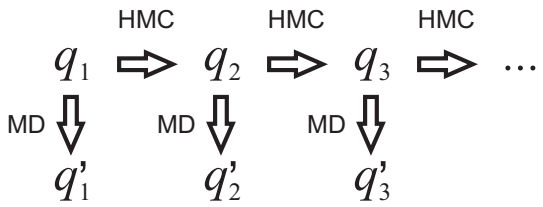

Figure 3.4.: The sampling process employed by ZIBgridfree for every modified distribution. Configurations $q_1, \ldots, q_n$ are generated by HMC sampling of a modified potential. The sequence (q'_j) is generated from the first sequence "on the fly" by drawing random momenta $\tilde{p}$ distributed according to the Boltzmann distribution η for every state q_j and launching a short MD trajectory from $(q_j, \tilde{p})$. These MD simulations are performed on the unmodified potential.

calculate the overlap between partial densities, while the "vertical chain" (states q'_j) is used to estimate the Markov operator P^τ and accumulate transition probabilities between partial densities. This sampling approach is presented in more detail in [73] and [45].

3.2.7. Computation of thermodynamic weights

The first step in conformation analysis based on trajectories generated by the sampling approach outlined in section 3.2.6 is the computation of a matrix M which contains information about the degree of membership of configurations sampled from the partial density ρ_i in all partitions ϕ_j of the conformational space:

$$M_{ij} = \frac{1}{n_i} \sum_{k=1}^{n_i} \phi_j \left(q_k^{(i)} \right). \tag{3.16}$$

If the basis functions ϕ_i are interpreted as abstract states in a Markov chain, then the stochastic matrix M is an estimate of the transition matrix of that Markov chain, i.e. M_{ij} is the probability of moving from the fuzzy set ϕ_i to the fuzzy set ϕ_j,

$$M_{ij} = \langle \phi_j \rangle_{\rho_i}. \tag{3.17}$$

Alternatively, the degrees of membership of $q_k^{\prime(i)}$ in ϕ_j can be used in equation 3.17.

This Markov chain is ergodic (see [45]), as detailed balance holds in the canonical ensemble. Therefore, the unique stationary distribution π of M exists. Because of detailed balance, the basis functions ϕ_i must be waited against each other in such a way that the net flow between them is zero:

$$\pi_i M_{ij} = \pi_j M_{ji}. \tag{3.18}$$

Therefore, the components of the stationary distribution π_i are, in fact, the thermodynamic weights w_i (see theorem 4.11 in [73]) and can be computed by eigenvalue

iteration, as

$$w = \pi = \lim_{n\to\infty} M^n \alpha \tag{3.19}$$

with an arbitrary initial distribution α.

The weights w_i are the linear combination factors that are used to reconstruct the Boltzmann distribution ρ on Ω from the the partial densities ρ_i:

$$\rho = \sum_{i=1}^{s} w_i \rho_i. \tag{3.20}$$

Note that weights are computed by evaluating the degree of membership of every sampling point from the sampling of each partial density ρ_i in every soft partition ϕ_i. This means that if the sampling of one partial density ρ_i does not converge, all weights will be flawed. ZIBgridfree must sample accurately the distribution over every soft partition ϕ_i, even those whose thermodynamic weights are intrinsically low due to a generally high level of potential energy.

3.2.8. Transition and overlap matrix and conformation analysis

Afterwards, the estimated thermodynamical weights w_i are used to compute the overlap integral matrix $\bar{S}$ and the transition matrix P (see equations 3.14 and 3.15). $\bar{S}$ is constructed as

$$\bar{S}(i,k) = \langle \phi_i, \phi_k \rangle_\rho = \sum_{j=1}^{s} w_j \langle \phi_i, \phi_k \rangle_{\rho_j} = \sum_{j=1}^{s} w_j \bar{S}_j(i,k), \tag{3.21}$$

approximated from individual summands

$$\bar{S}_j(i,k) \approx \frac{1}{n_j} \sum_{l=1}^{n_j} \phi_i(q_l^{(j)}) \phi_k(q_l^{(j)}), \tag{3.22}$$

each of which is built from a trajectory $\left(q_1^{(j)}, \ldots, q_{n_j}^{(j)}\right)$ that is a sampling of the partial density ρ_j.
Analogously, the matrix $\bar{P}$ is constructed as

$$\bar{P}(i,k) = \langle \phi_i, P^\tau \phi_k \rangle_\rho = \sum_{j=1}^{s} w_j \langle \phi_i, P^\tau \phi_k \rangle_{\rho_j} = \sum_{j=1}^{s} w_j \bar{P}_j(i,k), \tag{3.23}$$

where the individual summands for each trajectory $\left(q_1^{(j)}, \ldots q_{n_j}^{(j)}\right)$ are

$$\bar{P}_j(i,k) \approx \frac{1}{n_j} \sum_{l=1}^{n_j} \phi_i(q_l^{(j)}) \phi_k(q_l'^{(j)}), \tag{3.24}$$

where $q_l'^{(j)}$ are the configurations generated by 'vertical' sampling (cf. 3.2.6). P is then obtained by making $\bar{P}$ stochastic. For more details, see [45, 73].

If metastable conformations exist, the overlap integral matrix $\bar{S}$ is almost block-structured (see figure 3.5) after a suitable permutation. The same holds true for the transition matrix P. Robust Perron Cluster Analysis [18, 72] is used to find this permutation and thus the matrix of linear combination factors χ_{disc} that transforms the vector of basis functions $(\phi_1, \ldots, \phi_s)$ into a vector of conformation membership functions $(\chi_1, \ldots, \chi_C)$.

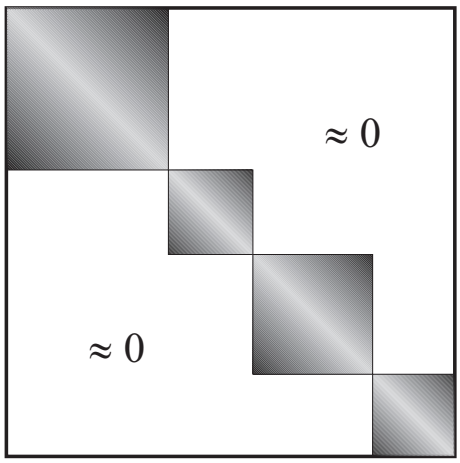

Figure 3.5.: After suitable permutation, the overlap matrix $\bar{S}$ or the transition matrix P of a metastable system has a block structure, where basis functions within each block communicate, while transitions to (or overlap with) basis functions outside a block are seldom (or weak).

3.2.9. Convergence criterion

ZMFree allows more than any other sampling technique to control the sampling error, by directly estimating that error. Theoretically, only with an infinite number of points n one could be sure that the transition matrix M has been estimated correctly. Weber, Kube et al. [74] pick up the idea of Weber [73] to estimate the sampling error $||E||_\infty = ||M - M_{tr}||$, the difference between the true matrix M and the estimation M_{tr} obtained from generating finitely many sampling points. The case that $||E||_\infty \leq \epsilon$ is equivalent to the statement that the $||\cdot||_1$-norms of row vectors of E are small,

$$||E(i,:)||_1 \leq \epsilon, \quad i = 1, \ldots, s. \tag{3.25}$$

The rows i of E correspond to different subsamplings $\left(q_1^{(j)}, \ldots q_{n_j}^{(j)}\right)$. The *normE* convergence indicator computes the ith row of M_{tr} for each subsampling according to equation 3.17. The difference between these rows in vector-$||\cdot||_1$-norm is measured, and the maximum distance is compared to a given ϵ.

3.2.10. Efficiency of ZIBgridfree

The ZIBgridfree approach has to produce $\mathcal{O}(ns)$ sampling points by the HMC method, which is only better than pure hybrid Monte Carlo if the sampling of each partial density ρ_i converges s times as fast as that of HMC on the unmodified potential. However, there is no theoretical limit to the mean time that a molecular system spends in one metastable conformation – consider diamond, which is very hard to change experimentally into graphite despite the latter conformation's lower potential energy [30]. It is expected that from some threshold for the size or complexity of the molecule onward, no sampling strategy can hope to sample the unmodified Boltzmann density in a reasonable time, and ZIBgridfree becomes more efficient and also more reliable than other strategies. $\mathcal{O}(ns)$ is also the computational cost of ZIBgridfree sampling in terms of memory usage, as every point in every subsampling has to be stored. By making use of parallelization, at least the time cost can be lowered considerably. It is also worth noting that the computational cost of sampling analysis is on the order of $\mathcal{O}(ns^3)$ if $\bar{S}$ and P are calculated from summands S_j and P_j as given by equations 3.21–3.24, respectively.

3.3. Replica Exchange

Ideally, the sampling would consist in a random walk in energy space rather than in position space. This would allow a fast discovery of all local minima of the potential energy surface. Unfortunately, there is no direct way to construct a random walk in energy space so that usually it cannot be done efficiently. The Replica Exchange method is a generalized-ensemble approach that consists in a random walk in temperature space which in turn induces a random walk in energy space [64] thus allowing the simulation to jump out of local minima more easily. Replica Exchange simulations have successfully been applied to macromolecules [10, 39, 51, 54].

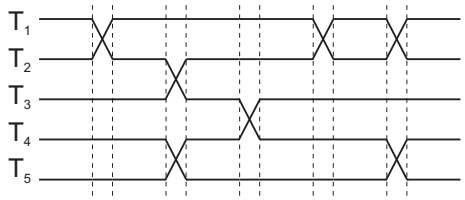

Figure 3.6.: Replica Exchange method for 5 replicas. Non-interacting copies of the system at different temperatures T_i are allowed to exchange positions (or temperatures) at regular intervals.

The principle of the Replica Exchange method is shown in figure 3.6. The basic idea is to consider M independent copies or *replicas* of the system to be simulated on which non-interacting simulations at M different temperatures are performed. At periodic intervals positions q are exchanged between replicas according to a Monte Carlo acceptance criterion. At high simulation temperatures it is easier to pass

energy barriers since, due to higher momenta, the effectively sampled probability density function is generally flatter, including the barriers. However, sampling at a higher temperature means generating samples from a different thermodynamic distribution. While the sampling points created in this way can be reweighted to the Boltzmann distribution at temperature T, this is only a heuristic and in no way equivalent to drawing samples from the target distribution in a mathematically rigorous way. For this reason, a single simulation at a high temperature is not sufficient. In a Replica Exchange simulation the replicas at high temperatures provide new start positions for the HMC chain at the relevant (low) sampling temperature T, which allows jumps out of the basin of attraction of a local minimum. This is illustrated for a 1-dimensional potential energy function and two HMC chains at different temperatures in figure 3.7. The subsequent conformation analysis is done based only on the chain at T, all sampling data at higher temperatures are discarded.

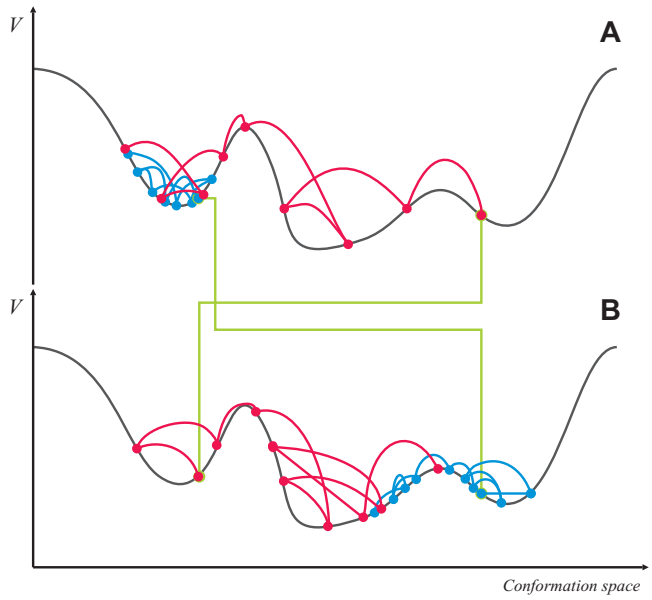

Figure 3.7.: Replica exchange. The potential energy surface is sampled by a high-temperature (red) and a low-temperature (blue) HMC chain.
(A) before and (B) after a replica exchange step. Replica exchange avoids trapping of the low-temperature chain in local minima.

In practice, M non-interacting hybrid Monte Carlo chains are started at M different temperatures where for every time step i there exists a one-to-one mapping between chains and temperatures. All chains are propagated simultaneously using hybrid Monte Carlo sampling as described in section 2.4. A replica exchange step is performed after every t_{RE} simulation steps. The replica exchange is realized technically as an exchange of temperatures rather than positions between two chains. This reduces the amount of data that has to actually be moved in memory, which is especially useful when chains are to be propagated truly parallelly on different CPUs. Let prior to an exchange step HMC chain i be at temperature m and replica j be at

temperature n. The exchange step then corresponds to the state transition

$$x = (\ldots, q_m^{[i]}, \ldots, q_n^{[j]}, \ldots) \rightarrow x' = (\ldots, q_n^{[i]}, \ldots, q_m^{[j]}, \ldots). \tag{3.26}$$

For transitions between states in this generalized ensemble [32], detailed balance is assumed as well, which means

$$\pi_{\mathrm{GE}}(x) P_{\mathrm{xchg}}(x \rightarrow x') = \pi_{\mathrm{GE}}(x') P_{\mathrm{xchg}}(x' \rightarrow x) \text{ with} \tag{3.27}$$

$$\pi_{\mathrm{GE}}(x) = \frac{1}{Q_{\mathrm{GE}}} \exp\left(- \sum_{i=1}^{M} \beta_{m(i)} V(q^{[i]}) \right), \tag{3.28}$$

where $m(i)$ is the index of the temperature of chain i and Q_{GE} is again a normalization factor that is used to obtain a probability distribution. The assumption of detailed balance is necessary for the Boltzmann distribution at each temperature to be an invariant measure of the Markov operator associated with the generalized ensemble [32].

Substituting equation 3.28 into equation 3.27 yields

$$\begin{aligned} \frac{P_{\mathrm{xchg}}(x \rightarrow x')}{P_{\mathrm{xchg}}(x' \rightarrow x)} &= \frac{\pi_{\mathrm{GE}}(x')}{\pi_{\mathrm{GE}}(x)} \\ &= \exp\left[-\beta_m V(q^{[j]}) - \beta_n V(q^{[i]}) + \beta_m V(q^{[i]}) + \beta_n V(q^{[j]}) \right] \\ &= \exp\left[(\beta_n - \beta_m) \left(V(q^{[j]}) - V(q^{[i]}) \right) \right] \\ &=: \exp(-\Delta). \end{aligned} \tag{3.29}$$

The detailed balance constraint can thus easily be met by choosing the acceptance criterion as

$$P_{\mathrm{xchg}}(x \rightarrow x') = \min\left\{1, \exp(-\Delta)\right\}. \tag{3.30}$$

3.3.1. Efficiency of the Replica Exchange method

The resulting acceptance ratio decreases exponentially as the distance $|\beta_n - \beta_m|$ of inverse temperatures increases. Therefore, replica exchange is only attempted between chains at adjacent temperatures. In [64] Sugita and Okamoto formulate the following criteria for evaluating the efficiency of the Replica Exchange method:

(a) The simulation temperatures should be chosen from the interval $[T, T_{\max}]$ in such a way that the acceptance ratios are approximately equal for all pairs of temperatures under consideration.

(b) The number of temperatures (and chains) M should be chosen so that the acceptance ratios for all pairs of temperatures under consideration are higher than 10%.

(c) The maximum simulation temperature $T_{\max}$ should be chosen high enough to avoid trapping in local minima, i.e. all major local minima have to be found within an acceptable period of time.

The first two criteria are easy to test, and criterion (b) can be ensured simply by starting short test runs with different numbers of chains and measuring the acceptance ratios. Different algorithms exist to calculate optimal choices for the simulation temperatures with respect to obtaining equal acceptance ratios for all pairs of temperatures. For small to medium-sized molecules, however, choosing temperatures with exponentially increasing distance already yields very good results in terms of criterion (a). In this approximation the sampling temperatures are closer to one another near the relevant temperature T than in the high-temperature region. A set of temperatures with this property is generated by

$$T_i = T \cdot a^i \text{ with } a = \left(\frac{T_{\max}}{T}\right)^{\frac{1}{M-1}}, \quad i = 0, \ldots, M-1. \tag{3.31}$$

In contrast to the first two efficiency indicators, criterion (c) can only be evaluated empirically or estimated indirectly. The former requires knowledge of the "true" conformations. As usually experimental data on the molecule's conformations are not available, statistical methods have to be employed. In fact, the question whether the maximum temperature in Replica Exchange has been chosen high enough is related to the question whether an MCMC simulation has been run long enough. It must also be noted that for systems that are stabilized primarily by hydrogen bonds or van der Waals forces, it is not allowable to use arbitrarily high simulation temperatures, as that would destabilize the system under consideration.

The computational overhead associated with the Replica Exchange compared to pure hybrid Monte Carlo is thus $\mathcal{O}(Mn)$. In order to be efficient, a Replica Exchange sampling has to be M times as fast as an HMC sampling at temperature T only.

3.4. ConfJump

The ConfJump strategy [71] employs *a priori* knowledge of the shape of the potential energy surface and thus of the Boltzmann distribution to be sampled. It facilitates transitions between different low-energy regions by introducing artificial jumps from one low-energy region to another into the sampling process. Thus, while still using HMC sampling to obtain physically correct transition probabilities, the average time the simulation spends within the basin of attraction of one local minimum of the potential energy is considerably shortened by occasional "jump steps" so that trapping is actively avoided. The combined Markov process which uses two different transition operators is ergodic and satisfies detailed balance (cf. 2.2), thus sampling the thermodynamically correct distribution. The ConfJump method is closely related to Smart Darting Monte Carlo [1] and the Jump Between Wells approach [60, 61].

ConfJump needs a preprocessing step in which a minimization algorithm is used to generate representatives of all important low-energy regions of the potential energy surface. Let $M = \{m_1, \ldots, m_C\}$ denote the set of these representatives. In the

current implementation this is done using the ConFlow algorithm by Holger Meyer, which is based on the RPROP algorithm [52]. This method is very fast and has been found empirically to be able to identify all important conformations of a wide variety of small to medium-sized biomolecules [46].

The information about low-energy regions can then be used in a standard Metropolis Monte Carlo approach as described in section 2.2 to propose jumps from the proximity of one local minimum of the potential energy to a point in the proximity of another. More precisely, ConfJump determines the configuration $m_j \in M$ that is closest to the current position state $q \in \Omega$ and then randomly chooses another configuration $m_k \in M$ and proposes a new configuration $\tilde{q} \in \Omega$ whose relative position to m_k is determined by the relative position of q to m_j. Throughout this work the following intuitive algorithm is used to obtain $\tilde{q}$ from q, m_j, and m_k:
Let x be the Z-matrix representation of q. Then $\tilde{x}$ is obtained from x by adding the difference vector $(m_k - m_j)$ to x. Transforming $\tilde{x}$ back to Cartesian coordinates yields $\tilde{q}$. Z-matrix coordinates are a very popular form of internal coordinates which are invariant to translation and rotation of the molecule and otherwise describe a molecule's position state accurately [37].

Trials generated in this way are subsequently accepted with a probability of $P_{\text{acc}}(q \to \tilde{q}) = \min\{1, \exp(-\beta\Delta V)\}$. This is the usual Metropolis acceptance criterion which requires a symmetric trial step, i.e. the jump from a point $q \in \Omega$ in the proximity of a low-energy configuration m_j to a point $\tilde{q}$ whose nearest neighbor from M is a configuration m_k must be proposed with the same probability as the reverse jump. Due to the constraint of symmetric trial steps, the trial $\tilde{q}$ must also be rejected if its nearest neighbor in M is not m_k. As the Metropolis acceptance criterion depends on the difference in potential energy between q and $\tilde{q}$, it can be expected that the acceptance probability improves if the probability to propose m_k given a point q whose nearest neighbor from M is m_j is based on the potential energy difference between m_j and m_k instead of proposing all low-energy configurations $m_k \in M$ with the same probability.

3.4.1. Jump Proposition Matrix

Therefore, in a second preprocessing step a jump proposition matrix $\mathcal{A}$ is calculated whose entries $\mathcal{A}_{jk}$ are the probabilities to propose a configuration m_k from a point whose nearest neighbor from M is m_j. Consequently, $\mathcal{A}$ must be a stochastic matrix, i.e.

$$\sum_{k=1}^{C} \mathcal{A}_{jk} = 1, \quad j = 1, \ldots, C. \tag{3.32}$$

In order for the Metropolis algorithm to be applicable, the detailed balance condition given by equation 2.8 must be satisfied. Choosing $\mathcal{A}$ symmetric, i.e.

$$\mathcal{A}_{jk} = \mathcal{A}_{kj}, \quad j, k = 1, \ldots, C, \tag{3.33}$$

ensures detailed balance as the proposed new position $\tilde{q}$ depends only on the position of m_k and the relative position of q to m_j.

Let $\hat{\mathcal{A}}$ be a $C \times C$-matrix with

$$\hat{\mathcal{A}}_{jk} := \begin{cases} \exp\left(-\beta \left|V(m_k) - V(m_j)\right|\right), & j \neq k \\ 0, & j = k \end{cases}. \tag{3.34}$$

The doubly-stochastic symmetric matrix $\mathcal{A}$ is computed by scaling the symmetric non-stochastic matrix $\hat{\mathcal{A}}$ using Ruiz's algorithm [53]. Using the jump proposition matrix $\mathcal{A}$ for trial generation is hoped to yield a high acceptance ratio as

- the acceptance probability depends on $V(\tilde{q}) - V(q)$ and due to spatial proximity q and $\tilde{q}$ are expected to be close in potential energy to m_j and m_k, respectively, and
- it is hoped that regions of similar potential energy have similar shapes as well.

The second point is very important as a trial $\tilde{q}$ whose nearest neighbor in M is not m_k has to be rejected as well. Figure 3.8 illustrates on a 2-dimensional conformation space how the acceptance ratio can decrease when the given low-energy regions have different geometric shapes. As $\mathcal{A}_{jk}$ is proportional to $\exp\left(-\beta \left|V(m_k) - V(m_j)\right|\right)$ for $j \neq k$, the transition from m_j to m_k and the reverse transition have an equally high probability of being accepted. Setting $\hat{\mathcal{A}}_{jj} = \mathcal{A}_{jj} = 0$ skips some unnecessary computations since the effect of accepting $m_k = m_j$ is the same as rejecting m_j (and staying in m_j).

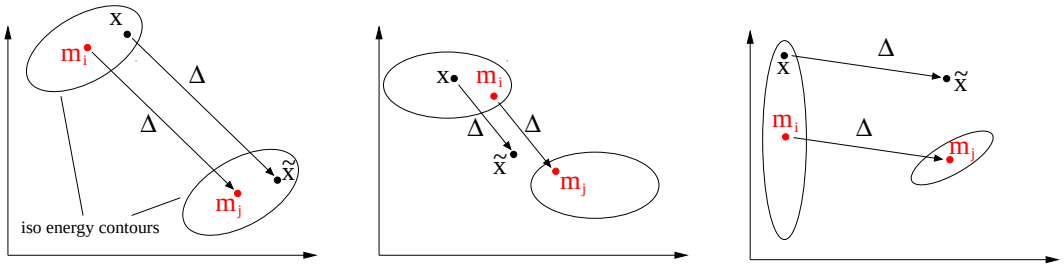

Figure 3.8.: Jump steps proposed by ConfJump in three different scenarios. Coordinate axes represent two different internal coordinates. The left panel shows a jump step that is accepted. The low-energy regions represented by m_i and m_j have a very similar shape. If m_i and m_j differ strongly in their relative positions to the regions they represent as shown in the central panel, the acceptance ratio is low. The same holds true if the two regions have very different shapes as shown on the right.

3.4.2. ConfJump as a rigorous sampling method

A Metropolis Monte Carlo sampling using only the jump method would not be ergodic as from a starting point $q^{(0)}$ only a small part of the configuration space Ω

is reachable by a series of jump steps. However, the method can be combined with a regular HMC approach which results in a Markov chain Monte Carlo method in which the configuration $q^{(i+1)}$ is determined from the current configuration $q^{(i)}$ by attempting a jump step with a fixed low probability P_{jump} and an HMC step with probability $1 - P_{\text{jump}}$. While the jumping Metropolis Monte Carlo is not ergodic and thus does not have a unique stationary distribution it satisfies detailed balance with respect to the thermodynamically correct distribution by construction. Mathematically, the Boltzmann distribution at temperature T is one possible invariant measure of the non-ergodic Markov operator of the jumping Metropolis Monte Carlo. In contrast to that, hybrid Monte Carlo is an ergodic Markov process, and the underlying Markov operator has the target distribution as its unique invariant measure, i.e. it is the unique stationary distribution [21]. If the two are combined by making a jump step with probability P_{jump} and an HMC step with $1 - P_{\text{jump}}$, an ergodic Markov process results whose unique stationary distribution is the Boltzmann distribution at temperature T.

3.4.3. The ConfJump Algorithm

Let $M = \{m_1, \ldots, m_C\}$ be a set of C low-energy configurations given in internal coordinates (Z-matrix representation) obtained from some minimization algorithm on the potential energy function V. Further, let P_{jump} denote the constant fixed probability of making a jump step rather than an HMC step.

Preprocessing: Compute the jump proposition matrix $\mathcal{A}$ (cf. 3.4.1).

Starting from an initial configuration $q^{(0)} \in \Omega$, repeat the following:

1. Generate a uniformly distributed random number $\zeta \in [0, 1)$.
2. If $\zeta > P_{\text{jump}}$, perform an HMC step (cf. 2.4).
3. Else, perform a jump step:

   Let $q = q^{(i)}$ denote the current state, and let x be the Z-matrix representation of q.

   (a) Find the nearest low-energy configuration $m_j \in M$ to x. This is done based on the cyclic Euclidian distance in the space of important dihedral angles (cf. 2.5).
   (b) Select a second low-energy configuration m_k with probability $\mathcal{A}_{jk}$.
   (c) Compute $\tilde{x} = x + (m_k - m_j)$. Let $\tilde{q}$ be $\tilde{x}$ transformed into Cartesian coordinates.
   (d) Find the nearest low-energy configuration $X \in M$ to $\tilde{x}$.
   (e) If $X \neq m_k$, set $q^{(i+1)} := q$.

(f) Else, accept $\tilde{q}$ according to the Metropolis acceptance criterion (cf. 2.2), i.e. generate a uniformly distributed random number $\xi \in [0, 1)$ and set the new configuration

$$q^{(i+1)} := \begin{cases} \tilde{q}, & \xi < \min\{1, \exp(-\beta \Delta V)\} \\ q, & \text{else} \end{cases} . \tag{3.35}$$

This is done either for a fixed number of times n or until convergence is detected.

Afterwards, conformations can be identified from the trajectory $\left(q^{(i)}\right)$ by successive Perron Cluster Analysis as described by Cordes et al. in [13].

3.4.4. Efficiency of the ConfJump strategy

A point that is only briefly discussed in [71] is the efficiency of the ConfJump strategy. In fact, only a "proof of concept" using numerical examples is provided. Some considerations regarding the theoretical efficiency of ConfJump will be presented here.

When compared to pure HMC, the ConfJump strategy has very little computational overhead as long as the acceptance ratio for jump steps is reasonably high. Its use of a jump proposition matrix for trial generation is hoped to improve the efficiency over the similar Jump Between Wells method [60, 61] and has a low computational overhead of $\mathcal{O}(C)$ as Ruiz's algorithm usually converges within a few iterations.

However, ConfJump relies fundamentally on precomputed information about low-energy regions of the potential energy V which is a very rough high-dimensional function. A global search strategy has to be employed in order to find all minima of the conformational space. Any such algorithm is necessarily affected by the "curse of dimensionality", i.e. even if only a projection of V into some lower-dimensional space (e.g. the space of heavy dihedral angles) is explored, the search algorithm will still have a computational cost that is exponential in the dimension of the search space. In fact, the number of local minima tends to increase exponentially with increasing size of the system under consideration [29]. Additionally, little *a priori* information about V can be employed as it is a multimodal nonconvex function [5]. Therefore, the only available option when searching for all local minima is a systematic search. Furthermore, in [76] Wille and Vennik proved theoretically that searching for all minima of the Lennard-Jones part of the potential[1] alone is already an *NP-hard* problem.

It must also be stressed that low-energy regions in high-dimensional very rough potential energy landscapes can hardly be expected to be all of a similar shape as required for a good acceptance rate of jump steps. Rather, we would expect the shape of low-energy regions to become more and more irregular. This poses a strong

[1] The Lennard-Jones potential is the additive part of V that describes the non-covalent, non-electrostatic interactions between pairs of atoms, i.e. repulsion between atoms whose electron orbitals overlap and van-der-Waals attraction [23, 37, 55].

problem for ConfJump as the direction of jumps is determined exclusively by the relative positions of q, m_j, and m_k to each other (see step 3c of the algorithm).

For these reasons, the applicability of ConfJump is limited to small to medium-sized molecules from the outset. In practice, the ConFlow algorithm is able to identify the low-energy regions of a wide-variety of drug-sized molecules.

4. Convergence diagnostics

When discussing Markov chain Monte Carlo algorithms in chapter 2, one very important question has been left open: For how many steps should the algorithm be iterated until the sampled distribution is a reasonably good approximation of the "true" distribution? Some criterion is needed to determine when improvements in the quality of the approximation of the target distribution can no longer be expected from continuing the simulation. A related question that is no less important is: How do we differentiate between "good" and "bad" sampling runs? Given two or more sampling results, which one approximates the physically correct distribution best? This requires a distance measure, preferably a metric, on a suitably defined space of sampling results.

While any rigorously conducted MCMC sampling converges towards the thermodynamically correct distribution of the system under consideration in $\mathcal{O}(\sqrt{n})$ [21] with the number of simulation steps n going to infinity almost surely, this is only statistical convergence, and it is very hard to tell in practice when a given upper bound for the sampling error has been reached. Therefore, heuristics must be used which, while not generally able to detect true convergence, can at least give a necessary condition for convergence. All convergence criteria are necessarily unreliable for slowly mixing Markov chains [15], i.e. for chains whose state space is divided into subsets between which transitions are rare. This is an intrinsic property of MCMC algorithms applied to molecular systems in some physically meaningful statistical ensemble where the sampled high-dimensional probability density functions are very rough [70]. No generally applicable convergence criterion for ergodic Markov processes is able to reliably distinguish between true convergence and local convergence within some metastable region in conformational space [9]. If a convergence monitor signals convergence, it is always possible that yet undiscovered regions with high statistical weight exist which are separated from the sampled subset by high energy barriers [45]. Obviously, using more than one convergence diagnostic can give a stricter criterion of (global) convergence and thus increase the probability to detect local convergence.

A wide variety of approaches is in use for convergence diagnostics; for an overview see [9] or [15]. Most methods are based on analyses of the properties of Markov processes and applicable to a wide field of problems. The most commonly used of these is presented in section 4.1.
In addition to that, knowledge of the properties of the systems under consideration can (and should) be employed. In section 4.3 a semi-empirical convergence criterion is developed which can give an independent necessary condition for convergence of the MCMC method for molecules that contain rotational symmetries.

Generalization of this criterion leads to a histogram-based method for comparing the results of two different sampling runs which is presented in section 4.2.

4.1. The Gelman-Rubin Criterion

The Gelman-Rubin statistic [8, 26] is one of the most widely used convergence indicators in practice. Gelman and Rubin's algorithm has a low computational cost, it is applicable to any type of ergodic Markov process and very easy to implement. Gelman and Rubin's approach requires multiple independent Markov chains which are launched from different starting points from an overdispersed distribution. It differentiates between true convergence and a trapping in some subset of the conformational space on the basis of a comparison of the variance within each chain with the variance between chains for some set of one-dimensional real-valued observables.

For m independent Markov chains with a length of n steps each the average of the m within-chain variances for an observable θ is given by

$$W = \frac{1}{m(n-1)} \sum_{j=1}^{m} \sum_{i=1}^{n} (\theta_j^i - \bar{\theta}_j)^2, \tag{4.1}$$

where θ_j^i denotes the value of θ at step i in chain j. The variance between the m chain means $\bar{\theta}_j$ is

$$\frac{1}{n} B = \frac{1}{m-1} \sum_{j=1}^{m} (\bar{\theta}_j - \bar{\theta})^2. \tag{4.2}$$

From W and B the total variance of the observable θ can be estimated:

$$\hat{\sigma}^2 = \left(1 - \frac{1}{n}\right) W + \frac{1}{n} B. \tag{4.3}$$

If the MCMC simulation has converged, W and $\hat{\sigma}^2$ are almost equal since W and B converge (statistically) to the same value. If, however, one chain gets trapped within one local subset of the conformational space which it never leaves while other chains generate a significant number of samples from other regions as well, then W will be lower than B. $\hat{\sigma}^2$ is an overly strict estimate of the total variance. Taking the sampling variance of both $\hat{\mu} = \bar{\theta}$ and $\hat{\sigma}^2$ into account yields a *pooled posterior variance estimate*

$$\hat{V} = \hat{\sigma}^2 + \frac{B}{mn}. \tag{4.4}$$

The Gelman-Rubin statistic, also called *potential scale reduction*,

$$\sqrt{\hat{R}} = \sqrt{\frac{\hat{V}}{W}}, \tag{4.5}$$

the square root of the ratio between the pooled and within-chain variance estimate, is then used as a measure for how much closer the sampled distribution might become

to the stationary distribution if the simulation were run longer. $\sqrt{\hat{R}}$ is always greater than 1 and converges to $1 + \frac{1}{mn} \overset{n\to\infty}{\longrightarrow} 1$ almost surely. Gelman and Rubin suggest running the simulation at least until $\sqrt{\hat{R}}$ is less than 1.1 or 1.2 [25, 27].

In order to generate starting points from an "overdispersed" distribution, a short disperse sampling is performed in which m independent Markov chains are launched from the same arbitrary starting point at a high temperature T_{disperse}. This leads to a Boltzmann distribution that covers the distribution at temperature T in the sense that it shares all important minima and maxima with it while being generally more variable. A subsequent very short burn-in sampling at a temperature $T_{\text{burn-in}} \leq T$ ensures that the points from which the m Markov chains used for the actual sampling start lie within important regions under the distribution at temperature T, i.e. usually near local minima of the potential energy surface.

For all experiments described in this thesis the implementation by Holger Meyer [47] is used which monitors convergence in all d important dihedral angles (cf. 2.5) by using both the sine and the cosine of each torsion angle φ_i as linearized observables which results in $2d$ observables for each of which a potential scale reduction factor is computed at regular intervals. Approximate convergence is assumed when the maximum of these factors drops below a fixed threshold (usually 1.01).

Monitoring convergence by the Gelman-Rubin method has a low additional computational cost since the samples generated in all m chains are equally used for the subsequent cluster analysis. This is possible because due to ergodicity running multiple moderately long Markov chains is equivalent to running one very long chain in the limit [28]. Only a small computational overhead of $\mathcal{O}(m)$ arises from the m short disperse and burn-in samplings.

The Replica Exchange method (cf. 3.3) which inherently uses multiple independent Markov chains is directly accessible to Gelman and Rubin's method. With m chains at different temperatures which are allowed to exchange temperatures in frequent intervals the within-chain variance and between-chain variance as calculated by equations 4.1 and 4.2 also converge to the same value. However, the combined distribution sampled by these "switching" chains on a generalized ensemble (see figure 3.6) is different from the target distribution (which, in fact, is only sampled by the Markov chain that is obtained by piecing together the segments at the sampling temperature T). Therefore, a greater total variance is expected from the combined distribution, but the Gelman-Rubin convergence monitor is still applicable and no less reliable than on m Markov chains sampling the Boltzmann distribution at temperature T. This were not the case if the RE method were implemented using m chains at different temperatures which exchange positions rather than temperatures since in that setting each chain would sample a different distribution.

4.2. Comparing Sampling Results

The primary goal of this thesis, comparing the performance of the three HMC-based sampling methods presented in chapter 3 on a certain set of molecules, requires com-

parisons of different sampling runs. Different ideas were considered with the aim that comparisons can be performed easily and independently of the generating sampling technique while incorporating as much as possible of the information generated in the sampling process. The idea to compare clustering results was quickly discarded as there is no way to define an informed distance measure on sets of clusters in a high-dimensional space. Further, it was felt that the sampling results should be compared in a more direct way. Thus, a metric on approximated statistical distributions, which are directly computable from any sampling result, was developed, which is presented in this section. A variant of this metric has been developed which is able to monitor convergence during sampling. The resulting symmetry criterion is presented in the following section.

The result of an HMC sampling run with a total length of n steps is a time series (trajectory) of molecule configurations $(q^1, \ldots, q^n)$. Projected into the conformational space defined by "heavy" dihedrals (cf. 2.5), it becomes a time series $(\Phi^1, \ldots, \Phi^n)$ whose data points are vectors of dihedral angles $\Phi = (\varphi_1, \ldots, \varphi_d)$, where d is the number of dihedral angles used to define the conformational space. This projection discards information from those degrees of freedom which do not define metastabilities.

When comparing two sampling results it makes no sense to look at individual molecule configurations. Rather, we want to compare the distributions sampled by the two simulation runs. It should be noted that in doing so all information about the order in which the sampling points where generated (and thus information about transition probabilities) is discarded. Instead of trying to compare two d-dimensional sampled probability density functions, comparisons are performed on the basis of the projections of the sampled distribution into each of the d 1-dimensional subspaces of the conformational space, i.e. we look at the sampled distributions in each dihedral angle separately. A sampling result $\mathcal{S}$ is thus interpreted as a tuple of approximated 1-dimensional statistical distributions $\mathcal{S} = (\rho_1, \ldots, \rho_d)$ for each dihedral angle, each defined on the interval $[0, 2\pi)$. Being density functions, the functions ρ_i are non-negative with

$$\int_0^{2\pi} \rho_i(\varphi)\,\mathrm{d}\varphi = 1.$$

These distributions can easily be discretely approximated as histograms for each torsion angle by binning all configurations Φ^j according to their value of φ_i using z bins of equal width.

1-dimensional projections are used instead of the original sampled distribution because comparing two d-dimensional is simply not practicable. Comparing two d-dimensional functions using a discretization of z bins in every dimension has a computational cost of z^d (and requires the same number of memory cells for storing the resulting histogram) which can only be done in a reasonable time for very low values of d. The determining factor for the cost of comparing sampling results should be the number of sampling points n which has a linear influence on the computational

cost as every point has to be processed exactly once when accumulating a histogram of the density of the sampling points. Comparing sets of d 1-dimensional histograms instead has a cost of $z \cdot d$ (both in time and memory) which is clearly preferable. The same idea of looking at the d degrees of freedom separately rather than discretizing the original d-dimensional space is used in [13] for cluster analysis.

If a metric defined on the space of d 1-dimensional projections of the d-dimensional sampled distribution indicates a distance of zero between two sampling results, this is only a necessary condition for the original two approximated d-dimensional distributions being identical. However, since both distributions are the results of samplings of the same molecule exploring the same potential energy landscape, correlations between dihedral angles are expected to have the same effect in both sampling results so that the difference in the original sampled distributions is not expected to be significantly greater than the difference measured between the sets of their 1-dimensional projections. Moreover, if that difference lies below some threshold so that the sampling results would be considered "similar", the distance between the original d-dimensional distributions is expected to be "low" as well.

Let $\mathcal{S}_1 = (\rho_1, \ldots, \rho_d)$ and $\mathcal{S}_2 = (\sigma_1, \ldots, \sigma_d)$ be two sampling results given by the approximated distributions ρ_i and σ_i, respectively, for each dihedral angle φ_i. An informative measure for the difference between $\mathcal{S}_1$ and $\mathcal{S}_2$ within one torsion angle φ_i is the L_1-metric in function space:

$$\delta_i(\mathcal{S}_1, \mathcal{S}_2) = \frac{1}{2} \int_0^{2\pi} |\rho_i(\varphi) - \sigma_i(\varphi)| \, \mathrm{d}\varphi. \tag{4.6}$$

The factor $\frac{1}{2}$ ensures that values of δ_i are always in the interval $[0, 1]$:

$$\int_0^{2\pi} |\rho_i(\varphi) - \sigma_i(\varphi)| \, \mathrm{d}\varphi \leq \int_0^{2\pi} \rho_i(\varphi) + \sigma_i(\varphi) \, \mathrm{d}\varphi = 2. \tag{4.7}$$

The value 2 is not only an upper bound for the difference integral, but the integral can actually take on this value, namely if $\rho_i(\varphi) = 0$ for every point φ with $\sigma_i(\varphi) > 0$ and vice versa.

The metric defined in equation 4.6 can be extended to a metric over tuples of dihedral distributions by simply averaging over all dihedrals:

$$\begin{aligned} \delta(\mathcal{S}_1, \mathcal{S}_2) &= \frac{1}{d} \sum_{i=1}^{d} \delta_i(\mathcal{S}_1, \mathcal{S}_2) \\ &= \frac{1}{2d} \sum_{i=1}^{d} \left(\int_0^{2\pi} |\rho_i(\varphi) - \sigma_i(\varphi)| \, \mathrm{d}\varphi \right). \end{aligned} \tag{4.8}$$

This metric is highly informative since it uses information from every point in every histogram and is therefore well-suited for measuring distances between sampling results.

Let $W = (w_1, \ldots, w_d) \in [0,1]^d$ be a weight vector with $\sum_{i=1}^{d} w_i = 1$. Then the weighted average over the metric's values for each dihedral,

$$\tilde{\delta}(\mathcal{S}_1, \mathcal{S}_2) = \frac{1}{2} \sum_{i=1}^{d} w_i \cdot \left(\int_0^{2\pi} |\rho_i(\varphi) - \sigma_i(\varphi)| \, \mathrm{d}\varphi \right), \tag{4.9}$$

is a metric as well. Note that δ is a special case of $\tilde{\delta}$ for $w_i = 1/d$, $i = 1, \ldots, d$. $\tilde{\delta}$ allows weighting the histograms for the different dihedral angles against each other, which can be used to put more emphasis on major metastabilities separated from the rest of conformational space by high potential energy barriers than on minor metastabilities which correspond to shallower local minima in conformational space. As an example, $w = 1/\lambda_2$ can be used, where the second eigenvalues of each dihedral's transition matrix T discovered by successive Perron Cluster Analysis [13]. This means weighting the dihedrals by their degree of metastability.

In practice the density functions ρ_i are approximated by histograms H_i each of which consists of z bins $H_i^1, \ldots, H_i^z$ of equal width which form a discretization of the interval $[0, 2\pi)$. The histograms are normalized, i.e. for all $i = 1, \ldots, d$:

$$\sum_{j=1}^{z} H_i^j = 1.$$

The difference between two sampling results given as sets of histograms H and J is then calculated as

$$\delta(H, J) = \frac{1}{2d} \sum_{i=1}^{d} \sum_{j=1}^{z} \left| H_i^j - J_i^j \right| \tag{4.10}$$

which is the average bin-wise difference of all d pairs of histograms. Again, histograms for different dihedrals can be weighted against each other in a way analogous to equation 4.9.

The histogram-based metric thus defined is a very widely applicable method for comparing different sampling results as it only depends on the sampling points themselves. Even the results of ZIBgridfree (cf. 3.2), which consist of a set of s sampling results with different weights $w_1, \ldots, w_s \in [0,1]$ each, can be compared to each other and to sampling results from other techniques. The total histogram H for one ZIBgridfree sampling run is computed from the normalized histograms $\tilde{H}_1, \ldots, \tilde{H}_s$ of the results from the samplings of the s partial densities (see 3.2.2) which are weighted by the thermodynamic weights calculated in the sampling analysis:

$$H = \sum_{i=1}^{s} w_i \tilde{H}_i. \tag{4.11}$$

Similarly, the method can be applied to results of sampling techniques that assign individual weights to all sampling points. These point weights can easily be taken into account when accumulating the histograms: When a sampling point with weight w

is determined to fall into a bin b, the counter for b is not increased by 1 but by w.

With this distance measure the quality of different sampling runs can be judged by comparing the different sampling results to a reference. Unfortunately, there exists no general method for assessing the quality of a sampling run, and thus it is in general impossible to create a reliable reference run. However, one can at least use a sampling run as reference in which one has great confidence, e.g. because of

- having run the simulation for a very long time,
- obtaining very similar results (as measured by the metric presented in this section) from very long simulations with different sampling methods, e.g. Replica Exchange and ConfJump, from different starting points, and/or
- finding many features of the sampling result in accordance with chemical intuition and possibly expert knowledge.

As mentioned in section 3.4 the ConfJump approach has been found to be reliable for typical drug-like molecules of small to medium size for which the ConFlow algorithm can reliably identify representatives of all important low-energy regions. Therefore, for the numerical experiments conducted for this thesis reference runs were created by running several long ConfJump simulations (5 HMC chains at 200000 steps each), verifying that all pair-wise distances were below a threshold of 0.03 and taking the simulation result as reference that had the lowest distance to all others. This reference was then verified by performing long simulations using the Replica Exchange methods and comparing the results to the reference run.

4.3. Symmetry criterion for convergence

When assessing the convergence behavior of different HMC-based sampling methods, it is clearly not advisable to rely on Gelman and Rubin's statistic alone, as

- no convergence indicator is able to reliably discern true convergence on the whole conformational space from local convergence within some metastable region (see [9] and also page 43 in this thesis), and
- when dealing with very rough high-dimensional functions such as the Boltzmann distributions of biomolecules, it is quite probable that some region in conformational space with a high statistical weight is never reached by any of the Markov chains, a case in which the Gelman-Rubin statistic falsely indicates convergence.

Therefore, it has been one of the goals of this thesis from the outset to develop a new convergence criterion to be used in addition to Gelman and Rubin's method which would incorporate knowledge about the system to be simulated. The idea for the criterion presented in this section stems from the observation that many

biomolecules contain rotational symmetries which should, of course, be reproduced in sampling. If e.g. the molecule under consideration contains a symmetric planar ring that is connected to the rest of the molecule by one single bond[1], the distribution of the torsion angle corresponding to that single bond over all molecule configurations generated should be periodic with a period of π (see fig. 4.1). A configuration with the torsion angle at a value of ψ and the configuration that has the same torsion angle set to $\pi + \psi$ but is otherwise identical to the first one behave physically and chemically in the same way and are therefore generated with equal probability in sampling.
A measure for the sampling error in a rotationally symmetric dihedral is defined on the basis of the metric for comparing histograms developed in section 4.2. This criterion proposed here is only applicable to molecules containing rotational symmetries. However, a cursory look at the ligand structures stored in the Protein Data Bank [4] reveals such symmetries, particularly symmetric planar ring structures to be an abundant feature of drug-like molecules. It is worth noting that the symmetry criterion is applicable to a large fraction of the class of peptide ligands (see e.g. [14, 20, 34]), as the amino acids phenylalanine and tyrosine each contain a symmetric aromatic ring.

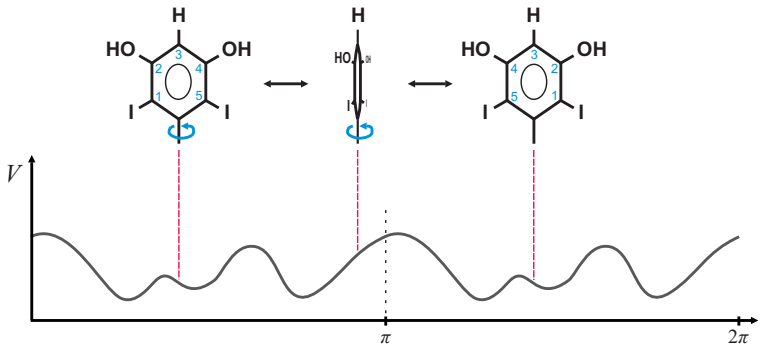

Figure 4.1.: Rotation of a symmetric planar ring and its effect on the potential energy.

In order to define a measure of sampling error based on this, a histogram is created for each rotationally symmetric torsion angle of the molecule. This is done by binning the configurations generated by sampling according to their value for the symmetric torsion angle. Throughout this work a fixed bin width of $5°(\frac{\pi}{36})$ was used. Then the sections of the histogram that are expected to be identical due to molecule symmetries are compared to each other. This is done by applying the error measure for comparing histograms presented in section 4.2 to all pairs of symmetric histogram sections (see fig. 4.2). The symmetry error measure is derived from equation 4.6. It is defined as the mean bin-wise difference between all pairs of symmetric sections of the (normalized) histogram H (as defined on page 48) for a symmetric torsion angle φ.

[1]This connecting bond must lie on a symmetry axis of the ring.

For 180° rotational symmetry the symmetry error is calculated as

$$E_{\text{sym}}(\varphi) = \sum_{i=1}^{z/2} \left| H_i - H_{\frac{z}{2}+i} \right| . \tag{4.12}$$

Using a bin width of $\frac{\pi}{36}$ yields a number of bins $z = 72$.
In the case of 120° rotational symmetry the average difference between three pairs of histogram sections is used:

$$E_{\text{sym}}(\varphi) = \frac{1}{2} \sum_{i=1}^{z/3} \left(\left| H_i - H_{\frac{z}{3}+i} \right| + \left| H_i - H_{\frac{2z}{3}+i} \right| + \left| H_{\frac{z}{3}+i} - H_{\frac{2z}{3}+i} \right| \right) . \tag{4.13}$$

The factor $\frac{1}{2}$ again ensures that the error is in the interval $[0, 1]$ and is calculated as $\frac{1}{3}$, for averaging between 3 pairs of histogram sections, divided by $\frac{2}{3}$ which is the maximum average difference between these pairs.

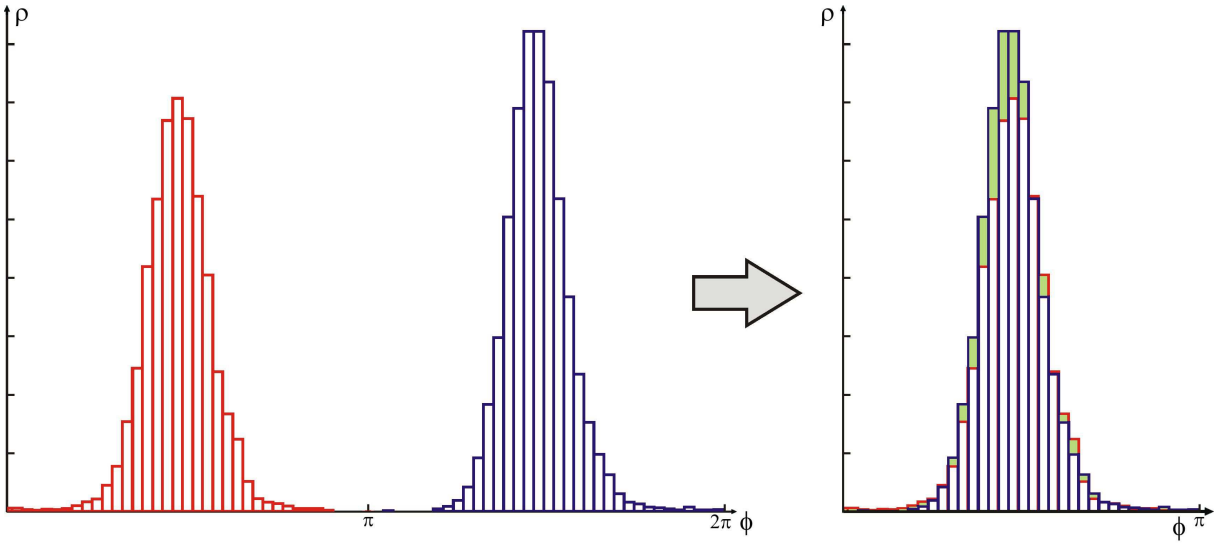

Figure 4.2.: The symmetry error for a single bond with 180° rotational symmetry is measured as average bin-wise distance between periodic sections of the corresponding histogram.

Thus, we obtain an informative measure for the sampling error which yields a necessary condition for convergence: If the sampling error is still above some fixed threshold, the MCMC sampling has not converged, yet. The convergence criterion gets stricter with every rotationally symmetric single bond in the molecule as, like with the Gelman-Rubin indicator, the maximum of all symmetry errors is used as convergence monitor.

4.3.1. Applicability of the symmetry criterion

The symmetry criterion is applicable to all MCMC methods that sample a Boltzmann distribution, such as ConfJump and Replica Exchange. When performing Replica

Exchange, either the combined distribution of all chains (which is not a Boltzmann distribution but preserves symmetric behavior) or a Markov chain that is composed of all segments that are at the sampling temperature T can be used. The latter approach was chosen for the simulations performed for this thesis as it was felt that this would give a stricter criterion of convergence due to the fact that only the low-temperature data are used in the cluster analysis.

ZIBgridfree samples a series of different distributions based on modified potential energy functions which do not necessarily assign the same energy value to two symmetric configurations. Reconstructing the overall sampled distribution requires reweighting of the sampling results under each modified potential against each other which has a computational cost of $\mathcal{O}(ns^3)$, where s is the number of modified potential energy functions and n is the number of sampling steps per individual sampling run (cf. 3.2.6). Moreover, the weights change as the sampling progresses. Thus, building the histograms for the symmetric torsion angles requires looking at all time steps of the samplings under each potential modification. Therefore, a convergence monitor based on symmetry errors should not be used for ZIBgridfree due to its prohibitive computational cost. However, it is easy to calculate symmetry errors during cluster analysis after the correct weights are calculated. The histogram for a symmetric torsion angle is built by adding histograms for the sampling runs in each potential modification which are multiplied by the respective weight of the corresponding partial density function.

In an RE or ConfJump simulation with an interval of convergence tests t_{test} the histogram at a time t can be reused when estimating the distribution at time $t+t_{\text{test}}$. Therefore, each convergence test based on the symmetry criterion only needs to look at the last t_{test} sampling steps. This is, in fact, less than the computational cost of the Gelman-Rubin convergence monitor which has to look at all $t + t_{\text{test}}$ steps. Figure 4.3 shows the symmetry error decreasing in a typical sampling run using the Replica Exchange method for the molecule L-benzylsuccinic acid (BZS) shown in figure 5.1, which contains one 180° rotationally symmetric bond.

The fact that multiple chains are sampled as a requirement of Gelman and Rubin's method is useful for the symmetry criterion as well. By using 5 chains which is neither divisible by 2 nor by 3, we know that at least one chain must sample the transition between the 2 (or 3) symmetric parts of a monitored dihedral's distribution in order for an approximately equal number of points being generated from all symmetric parts. If e.g. all 5 chains never crossed the barrier between the two periodic regions in the distribution of a 180° rotationally symmetric torsion angle, there would be at best $\frac{3}{5}$ of the sampling points in one region and $\frac{2}{5}$ in the other. The symmetry criterion is thus able to recognize this type of local convergence.

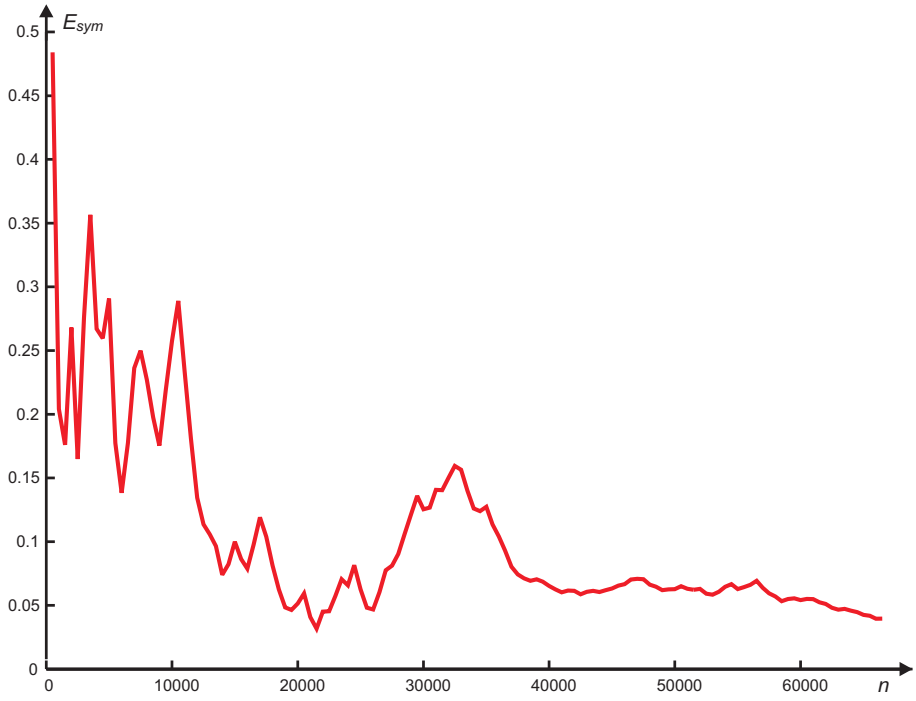

Figure 4.3.: The symmetry error decreases with growing number of simulation steps.

4.3.2. Automatic detection of molecule symmetries

Of course, it is desirable that the user of conformation analysis software such as the ZIBgridfree program [47] need not input the information about rotational symmetries. Rather, it should be possible to identify such symmetries automatically without any input from the user. Therefore, the following algorithm has been developed for examining symmetric properties of a molecule based on its topology.

Rotational symmetry of single bonds in a molecule is a property of the molecule's topology rather than its geometry. If two "branches" of a molecule are considered symmetric based on a topological analysis, the only geometric property that remains to be tested is whether there exist chirality centers in the branches. Therefore, rotational symmetries in a molecule are mainly determined based on a graph representation of the molecule.

Let a graph $G = (V, E)$ with a set of nodes V and a set of undirected edges $E \subseteq V \times V$ be a graph representation of the given molecule, i.e. each node $v \in V$ represents one atom by storing the atom's unique index and its atomic number, and each edge $e = (u, v) = (v, u) \in E$ represents a bond between atom u and atom v. Then a node v with degree $g = \deg(v) \geq 3$ is a symmetry center if after removing one edge (u, v) from the graph, the component of the remaining graph that contains v and its remaining neighbors $v_1, \ldots, v_{g-1}$ can be split into $g - 1$ non-overlapping isomorphic subgraphs so that no two edges (v, v_i) and (v, v_j), $i \neq j$, are part of the same subgraph. Such a graph partitioning is shown schematically in figure 4.4. Then the edge (u, v) is rotationally symmetric if it is a single bond. Since we are

interested in biomolecules, it is sufficient to consider nodes with a degree of 3 or 4, i.e. 2 symmetric branches resulting in a 180° rotational symmetry or 3 symmetric branches which gives a 120° rotational symmetry.

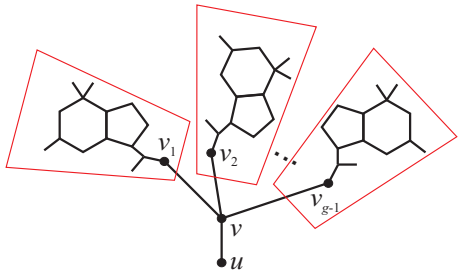

Figure 4.4.: An edge is rotationally symmetric if the connected component bordering on one node v of the edge can be split into $\deg(v) - 1$ isomorphic subgraphs.

All symmetry centers (and consequently all rotationally symmetric single bonds) for 180° rotational symmetry can be found efficiently by the following recursive algorithm. It is assumed implicitly that no atom has more than four binding partners. Full pseudocode can be found in appendix A.

Recursive algorithm for identifying symmetry centers

Start from an empty list of symmetric dihedrals.

For every node $v \in V$ with $\deg(v) = 3$ that is adjacent to a single bond described by a "heavy" dihedral (cf. 2.5), repeat the following:

- For each neighbor u of v do:

  1. Mark u and v as visited, all other nodes as unvisited.
  2. Let l, r be the other two neighbors of v.
     Call function compareSubgraphs(v, l, v, r) to determine whether the branches starting with the directed edges $v \to l$ and $v \to r$ are isomorphic.
  3. If the result of step 2 is `True`, the bond described by the edge (u, v) is rotationally symmetric.
     Identify the dihedral that describes the bond (u, v) and add it to the list of symmetric dihedrals if each of the isomorphic branches contains more than one atom.

compareSubgraphs(*from*1, *to*1, *from*2, *to*2)
Test the following cases in the order given:

Case 1: *to*1 = *to*2, i.e. a ring closes.

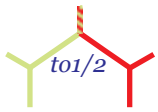

Figure 4.5.: Recursion ends if two branches meet.

a) $to1$ has ≤ 3 neighbors (at most one unvisited neighbor):

Return `True`.

b) $to1$ has four neighbors (two unvisited neighbors, l and r):

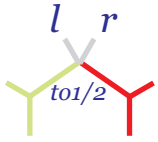

Figure 4.6.: Two branches meet at a new branching point.

Mark $to1 = to2$ as visited.
If COMPARESUBGRAPHS$(to1, l, to1, r)$, then return `True`.
Else, unmark $to1$ and return `False`.

Case 2: Nodes $to1$ and $to2$ are of different types of atoms or have a different number of neighbors.
▷ Atoms are incompatible ⇒ backtrack.
Return `False`.

Case 3: $to1$ has exactly one neighbor (the one we came from).

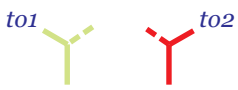

Figure 4.7.: Recursion ends at matching terminal atoms.

Return `True`.

Case 4: $to1$ and $to2$ have a different number of unvisited neighbors.
▷ One branch is growing into the other ⇒ backtrack.
Return `False`.

Case 5: $to1$ has no unvisited neighbors.

Return `True`.

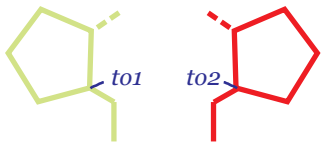

Figure 4.8.: Recursion ends with a ring closing on each branch.

Else: Recurse into branches.
Set *result* ← `False`.
Mark *to*1 and *to*2 as visited.

a) *to*1 has 1 unvisited neighbor:

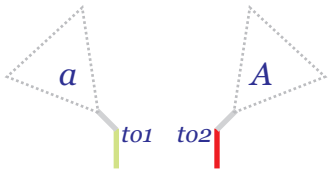

Figure 4.9.: Only one path to pursue on each branch.

Set *result* ← compareSubgraphs($to1, a, to2, A$).

b) *to*1 has 2 unvisited neighbors:

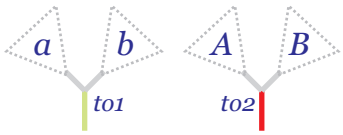

Figure 4.10.: Branching point with 2 branches on each side.

▷ The pairs of isomorphic subbranches are either (a, A) and (b, B) or (a, B) and (b, A).
Call compareSubgraphs recursively to identify isomorphic pairs of subbranches.
Set *result* accordingly.

c) *to*1 has 3 unvisited neighbors:

▷ Try to find a permutation of (A, B, C) that is isomorphic to (a, b, c).
Call compareSubgraphs recursively to identify isomorphic pairs of subbranches.
Set *result* accordingly.

Figure 4.11.: Branching point with 3 branches on each side.

If *result* = `True`, check for chirality.

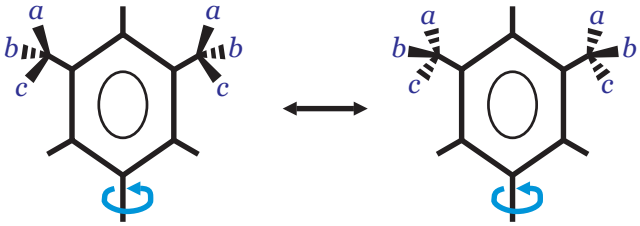

Figure 4.12.: Two branches are not symmetric if they contain chiral atoms.

If the atom *to*1 is a chirality center, set *result* ← `False`
▷ See "A note on chirality" on page 57.

If *result* = `False`, unmark *to*1 and *to*2.
Return *result*.

The algorithm is very similar for 120° rotational symmetry. This variation is omitted here.

A note on chirality

Chirality (or "handedness") is a form of stereoisomerism which in organic chemistry occurs with certain carbon atoms [48]. When a carbon atom is connected to 4 different functional groups, these can be arranged in two different ways that represent nonsuperimposable mirror images of each other. Two molecules that differ only in the configuration at one chiral center from each other have different physical and chemical properties.

When looking for 120° symmetry, three branches that contain chirality centers can only be accepted as symmetric if the chirality is the same on all branches. Otherwise the resulting structure is not rotationally symmetric.

When trying to identify 180° rotational symmetry, however, two branches of a molecule that are identified as symmetric based on topology must not contain any chiral carbon atoms (see figure 4.12 for an illustration). Therefore, whenever in both branches a node is discovered that has three undiscovered neighbors, the branches are considered not symmetric if the functional groups to which the four neighbors belong are all different. This property is tested in case 5c of the algorithm by applying a variant of compareSubgraphs to pairs of the branches that start with *from*1, *a*,

b, and c, respectively, until an isomorphic pair is found or all pairs have found to be not isomorphic.

5. Numerical Experiments

The three sampling methods presented in chapter 3, ZIBgridfree, Replica Exchange, and ConfJump were compared by trying to estimate the thermodynamically correct distribution at 300K for the three model systems presented in section 5.2.

5.1. Performance measure for sampling runs

The most important quantities for comparing different sampling methods are the time requirement, the mean accuracy, and possibly the memory requirement of each strategy. It was decided to deal with the latter only theoretically and only incorporate the former two quantities into a measure of performance. The reciprocal of the product of a time measure and a measure of sampling error is well suited as performance measure as both time and sampling error will be close to zero in the optimal case and large for "bad" sampling runs. It is impossible to normalize the two measures against each other as no general statement is possible about the ratio between one unit of time and one unit of error (in fact, this ratio is being examined in this study). This leads to the problem that it is impossible to tell which case is worse, that of a high sampling error after a short sampling time or that of a long sampling yielding a low sampling error, where the product of the two is the same in both cases. However, it is not the goal of this thesis to compare arbitrary sampling runs, and samplings are always run either until all convergence criteria signal convergence or until a fixed number of time steps which is chosen relatively high. Therefore, the sampling error can be expected to be approximately on the same order of magnitude for all samplings while the actual number of iterations needed by each sampling run can differ strongly. At least, the constellation of a high sampling error at a low sampling time is actively prevented. Thus, the performance measure proposed here can be thought of as a measure of time until convergence augmented by a punishment factor for sampling error.

As all sampling techniques under consideration are based on the hybrid Monte Carlo approach (cf. section 2.4), it is sufficient to measure the time of each sampling run in terms of total HMC steps. Corrections are necessary only for the preprocessing step of ConfJump in which representatives from all low-energy regions are generated (see 3.4). No correction was used for the presampling phase of ZIBgridfree, as that value is low compared to the total number of time steps in 100 subsamplings (cf. 5.3).

As the identification of all low-energy regions in conformational space is not based on HMC, a correction factor ν is introduced, which is the average time needed by a simulation per HMC step on the same computer on which the preprocessing

was performed. Thus, the time needed for detecting all low-energy regions in the preprocessing can be expressed in HMC steps by division by ν.

The sampling error is measured as the distance between a sampling result and the result of a reference run by the metric developed in section 4.2. Reference runs were generated for each molecule by performing several very long runs ($1.2 \cdot 10^6$ steps overall) of ConfJump and Replica Exchange, respectively, and choosing the one which had the least distance to all others after removing obvious outliers. The sampling runs used for creating the reference were discarded afterwards. Visual inspection of the distributions in all heavy dihedrals also shows almost no flaws for all the reference runs for every model system used (see chapter 6). The sampling error must, of course, be considered zero if it lies below some fixed threshold due to the inherent uncertainty inherent in the generation of reference runs which is explained in more detail in section 4.2. The unweighted form of the sampling error (calculated by equation 4.10) is used for evaluating the quality of all simulations. It has been found empirically that when comparing pairs of random histograms, the average bin-wise difference is 15.0%. When evaluating sampling error alone, samplings that have an average bin-wise difference from the reference of less than 1%, are considered equal. Values above 1% are considered very low up to 3%, low between 3% and 6%, medium between 6% and 11% and high if they lie above 11%. The same scale applies to the symmetry error.

The performance of a sampling run $\mathcal{S}$ is measured against a reference run $\mathcal{S}_{\text{ref}}$ in all practical experiments as

$$G(\mathcal{S}) = \frac{1}{n \cdot \delta(\mathcal{S}, \mathcal{S}_{\text{ref}})}, \tag{5.1}$$

where n is the total number of HMC steps performed during sampling which is corrected as described above.

5.2. Molecules used for this study

The performance of the sampling methods was measured for three different ligand molecules which were extracted from the Protein Data Bank (PDB) [4]. The particular choice of ligand molecules used here was inspired by Boström [6]. Only molecules were chosen that contain 180° rotational symmetries so as to be able to use the semi-empirical convergence criterion developed in section 4.3. Table 5.1 shows that the three molecules chosen from the PDB differ considerably with respect to their size and complexity. The structural formulas of the molecules are shown in figures 5.1, 5.2, and 5.3. The single bonds that correspond to heavy dihedrals are labeled with numbers.

L-Benzylsuccinate (found in the PDB under its "HET ID" BZS) is an inhibitor of carboxypeptidase A [40]. It consists of 25 atoms of which 10 are hydrogen. The molecule is shown in figure 5.1 and has one rotationally symmetric single bond which connects the aromatic ring to the rest of the molecule. Its conformations

HET ID	atoms	H atoms	d	d_{sym}
BZS	25	10	5	1
TOP	39	18	5	2
BSI	46	18	7	2

Table 5.1.: The molecules used for this study. d is the number of heavy dihedrals, and d_{sym} is the number of rotationally symmetric dihedrals (180° rotational symmetry).

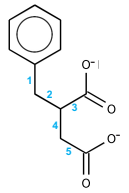

Figure 5.1.: L-Benzylsuccinate (BZS). Numbers 1–5 indicate heavy dihedrals.

are described in terms of 5 "heavy" dihedral angles. BZS is the smallest and least complex system considered in this work.

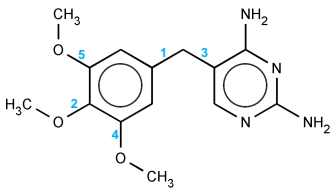

Figure 5.2.: Trimethoprim (TOP), an antibiotic.

Trimethoprim (HET ID: TOP) is an antibiotic that works by inhibiting bacterial dihydrofolate reductases [11]. The molecule consists of 39 atoms of which 18 are hydrogen and is more complex than L-benzylsuccinate. Of the 5 heavy dihedrals in the Trimethoprim molecule two are rotationally symmetric, namely the two bonds that lie in the symmetry axis of the aromatic ring shown on the left hand side in figure 5.2, one facing the greater part of the molecule, the other facing the central (-OCH3)-group.

BSI (2-(Biphenyl-4-sulfonyl)-1,2,3,4-tetrahydro-isoquinoline-3-carboxylate) is an inhibitor of the enzyme neutrophil collagenase which is also called matrix metalloproteinase 8 (MMP-8) [41]. At 46 atoms of which 18 are hydrogen, BSI is not only the largest but also the most complex molecule under consideration in this work. As clearly visible in figure 5.3 the molecule contains a non-aromatic ring (top left), which is why a similar behavior to that of cyclohexane (see figure 3.1) can be expected from BSI. It is expected that this ring can assume two very different

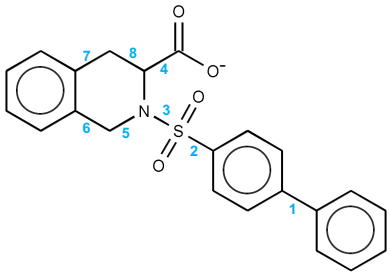

Figure 5.3.: 2-(Biphenyl-4-sulfonyl)-1,2,3,4-tetrahydro-isoquinoline-3-carboxylate (BSI).

configurations that are separated by extremely high energy barriers which makes the sampling very difficult. Of the 8 heavy dihedrals spanning the molecule's conformational space 2 describe rotationally symmetric single bonds, namely that connecting the two aromatic rings on the right hand side and the bond that connects one of these rings to the sulfur atom.

5.3. Simulation details and choice of parameters

5 experiments were conducted for each molecule and each sampling strategy under consideration which yields a total of 45 simulation runs. The 5 simulations for one molecule using the same technique were run with the same set of parameters except for the initial state of the random number generator which was chosen differently for each simulation run. All computer simulations were performed using the ZIBgridfree framework [47]. All methods developed for this thesis, most importantly the symmetry criterion (cf. 4.3) and the algorithm for automatic detection of molecule symmetries (cf. 4.3.2) have been implemented within this framework which already contained all three simulation techniques compared in this thesis. The ZIBgridfree program is written in C++ and uses libraries from amira [63] and amiraMol [56].

All experiments were run at a temperature of 300K which is near to typical physiological temperatures. Every individual HMC sampling run (including the presampling phase of ZIBgridfree) was started with a disperse phase at a temperature of 2000K for 300 HMC steps and a burn-in phase at 300K for 10 steps in order to ensure that the Markov chains start in different regions of Ω. All disperse and burn-in sampling steps are discarded (cf. section 4.1). Each HMC proposal was generated by 60 integration steps of molecular dynamics. The length of an MD step was chosen as 1.3fs. All three methods were used with the parameters set to values that were found to be suitable in earlier experiments, see e.g. [45, 71].

ZIBgridfree simulations were restricted to 100 nodes resulting in 100 partial distributions to be sampled. The maximum number of HMC steps within for the sampling of each partial density was set to 20000 per chain. 60 MD steps were performed for trial generation for HMC in the "horizontal" sampling (see section ref-

sec:zmfsampling), while 30 steps of MD were used for generating the configurations q'_j in the "vertical" sampling. In presampling the maximum number of HMC steps allowed was 18000. 5 Markov chains were launched per simulation both in presampling and in sampling resulting in a total upper bound of approximately $11 \cdot 10^6$ HMC steps depending on the actual number of nodes used. The presampling was performed at a temperature of 2500K, and convergence was detected by a Gelman-Rubin statistic (see section 4.1) using a threshold of 1.05. The convergence of the sampling was monitored by ... Convergence checks were performed every 500 HMC steps. In order to accelerate calculations a cutoff value of 10^{-6} was used below which the value of a basis function ϕ_i was set to zero.

Replica Exchange simulations for all three molecules were performed with 10 chains at temperatures of 300K, 387.46K, 500.43K, 646.33K, 834.77K, 1078.14K, 1392.48K, 1798.45K, 2322.79K, and 3000K which were determined by equation 3.31. The maximum number of steps allowed per chain was set to 100000 which amounts to a total upper bound of 10^6 HMC steps. Convergence was monitored by a combination of the Gelman-Rubin statistic with a threshold of 1.01 and the symmetry criterion (cf. section 4.3) using a threshold of 0.04. Convergence tests were performed at intervals of 500 HMC steps.

For the ConfJump simulations the same convergence criteria were used as for Replica Exchange (although the interpretation of the value of the Gelman-Rubin statistic changes slightly for Replica Exchange, see section 4.1). The same total upper bound of 10^6 HMC steps was chosen (mainly due to memory limitations) which corresponds to 200000 maximally allowed steps in each of 5 Markov chains. The same sets of precomputed representatives of the low-energy regions of the potential energy as in [71] were used for all three molecules. As ConfJump simulations are expected to converge fast from the simulations performed in [71], convergence was checked every 200 steps. The probability of jump steps was set to $P_{\text{jump}} = 0.2$ for all simulations (see 3.4).
Correction factors ν (cf. 5.1), which are used to express the time for generating representatives of all local minima of the molecule, were calculated for all three molecules from ConfJump simulations of 1000000 steps in length (similar to the actual simulation runs in setup) as shown in table 5.2.

HET ID	t_{ConFlow}	ν	c
BZS	470s	83.1	39000
TOP	1700s	47.0	80000
BSI	2700s	34.5	93000

Table 5.2.: Time of generating representatives of low-energy regions t_{ConFlow} (in s), number of HMC steps per second ν (in a ConfJump run), and c, the product of the two, for the three model systems (approximate values).

6. Results

In the following tables 'error' is the sampling error, 'steps' is the total number of HMC steps performed, the column 'corrected' contains the corrected number of steps in the case of ConfJump, 'performance' is the performance as calculated by equation 5.1 while 'E_{sym}' and 'G-R' contain the final values of the symmetry and Gelman-Rubin criterion, respectively. $\hat{\mu}$ denotes the means over the values for the 5 respective sampling results while $\hat{\sigma}$ is the estimated standard deviation.

6.0.1. L-Benzylsuccinate

Figure 6.1 shows the 1-dimensional projections of the Boltzmann distribution of L-benzylsuccinate at 300K sampled by the reference run, a ConfJump simulation with a length of 120000 steps per chain performed with overly strict convergence criteria but with the other parameters set to the values given in section 5.3. The three dihedral angles corresponding to rotationally symmetric single bonds, namely the bond next to the aromatic ring (top left panel in figure 6.1) and the two bonds that are adjacent to the carboxyl groups (bottom panels), show nearly perfect symmetry at visual inspection. The symmetry error for the monitored dihedral (1) is very low at 2.06%. The distribution of dihedral 2 (see figure 5.1; top center in figure 6.1) shows two peaks with different weight. Remarkably, dihedral 3 (top right panel) shows only a single peak which is probably due to the strong repulsion between the two negatively charged carboxyl groups.

The ZIBgridfree strategy achieves a very low to medium sampling error except for one outlier which is highly different at 20.5% average bin-wise difference from the reference (see table 6.1). Excluding this outlier (line 4 in table 6.1), the average is low at 5.83% with a standard deviation of 3.02%. Therefore, it can be concluded that the ZIBgridfree method can reproduce the Boltzmann distribution at $T = 300$K with a fairly low sampling error when using a meshless discretization into $s = 100$ partial densities. The symmetry error is very low to medium, again except for simulation run 4, with values between 2.9 and 8.6%. It must be noted that the results produced by ZIBgridfree have a high standard deviation, which is almost as high as the means with respect to sampling error, symmetry error and overall performance relative to the reference run.
The method needs many time steps, and frequently, the sampling of a partial density function does not converge (according to the criterion that was used (see 5.3)). Therefore, the average simulation time (measured in HMC steps) is high at about $11.5 \cdot 10^6$ with a standard deviation of 322000. Consequently, the performance is low at values between $4 \cdot 10^{-7}$ and $4 \cdot 10^{-6}$.

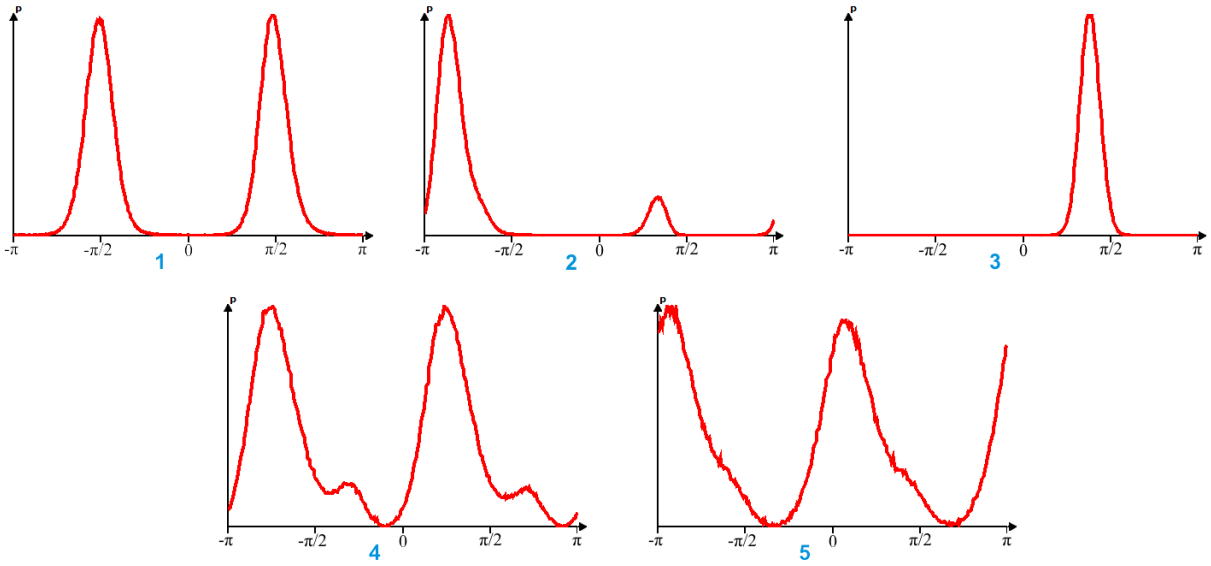

Figure 6.1.: The sampled distributions of the five heavy dihedrals of L-benzylsuccinate. Numbers below the diagrams refer to the dihedrals marked in figure 5.1.

	error	steps	performance	E_{sym}
1	0.0897	$11.47 \cdot 10^6$	$9.721 \cdot 10^{-7}$	0.0857
2	0.0458	$11.75 \cdot 10^6$	$1.857 \cdot 10^{-6}$	0.0364
3	0.0754	$11.72 \cdot 10^6$	$1.132 \cdot 10^{-6}$	0.0288
4	0.2054	$10.96 \cdot 10^6$	$4.444 \cdot 10^{-7}$	0.1592
5	0.0223	$11.60 \cdot 10^6$	$3.873 \cdot 10^{-6}$	0.0428
$\hat{\mu}$	0.0877	$11.50 \cdot 10^6$	$1.656 \cdot 10^{-6}$	0.0706
$\hat{\sigma}$	0.0708	322000	$1.339 \cdot 10^{-6}$	0.0542

Table 6.1.: Results for BZS using ZIBgridfree.

Replica Exchange is able to reproduce the sampling result from the reference run very well (see table 6.2). The mean sampling error is low at 3.6% average bin-wise difference to the reference. However, all experiments except for the fourth in table 6.2, where the sampling obviously has not converged (see the right-most column, 'G-R'), have yielded values that are below this average. Excluding the outlier gives a very low mean error of 2.79% and a standard deviation of 0.42%. The mean symmetry error is low at a value of 5.8% with a standard deviation of 2.35%. Replica exchange converges fairly well according to Gelman and Rubin's convergence monitor within the maximally allowed number of HMC steps. This upper bound is reached in 3 of the 5 sampling runs. The average simulation time is 946000 HMC steps with a standard deviation of 358000. The average performance of the RE method on BZS is $4.238 \cdot 10^{-5}$, $5.002 \cdot 10^{-5}$ after removing the outlier, with a standard deviation of $2.8795 \cdot 10^{-5}$ or $2.677 \cdot 10^{-5}$, respectively. This is about 26 times the

performance of ZIBgridfree.

	error	steps	performance	E_{sym}	G-R
1	0.0293	$1.2 \cdot 10^6$	$2.848 \cdot 10^{-5}$	0.0642	1.02
2	0.0319	$1.2 \cdot 10^6$	$2.611 \cdot 10^{-5}$	0.0948	1.05
3	0.0284	445000	$7.912 \cdot 10^{-5}$	0.0363	1.008
4	0.0704	$1.2 \cdot 10^6$	$1.183 \cdot 10^{-5}$	0.0530	1.36
5	0.0220	685000	$6.636 \cdot 10^{-5}$	0.0398	1.008
$\hat{\mu}$	0.03640	946000	$4.238 \cdot 10^{-5}$	0.0576	1.09
$\hat{\sigma}$	0.01937	358000	$2.880 \cdot 10^{-5}$	0.0235	0.152

Table 6.2.: Results for BZS using Replica Exchange.

In the simulations of L-benzylsuccinate, ConfJump yielded the best results on average, with a mean sampling error of 2.17% which is very low (see table 6.3. Highly accurate results are obtained with a high reliability as the standard deviation of the sampling error is only 0.97%. The ConfJump sampling converged in every case, i.e. the sampling error dropped below 0.04, and the Gelman-Rubin indicator went below a threshold of 1.01 within the limit set for the number of HMC steps. In fact, the sampling converged after less than 100000 steps (20000 steps per chain) in 4 of 5 runs.
The number of sampling steps was corrected by adding the correction value $c = 39000$ from table 5.2 for BZS. The resulting approximate total simulation time was 128800 with a very high standard deviation of 192200 due to the outlier in line 5 of table 6.3. The performance calculated on the basis of these values was $4.639 \cdot 10^{-4}$ on average and had a standard variance of $1.369 \cdot 10^{-4}$. Thus, for L-benzylsuccinate the performance of ConfJump was 32 times that of Replica Exchange and 827 times as high as that of ZIBgridfree.

	error	steps	corrected	performance	E_{sym}	G-R
1	0.0351	16000	55000	$5.181 \cdot 10^{-4}$	0.0393	1.007
2	0.0237	34000	73000	$5.791 \cdot 10^{-4}$	0.0393	1.009
3	0.0207	80000	119000	$4.059 \cdot 10^{-4}$	0.0389	1.007
4	0.0213	44000	83000	$5.651 \cdot 10^{-4}$	0.0397	1.005
5	0.0078	470000	509000	$2.513 \cdot 10^{-4}$	0.0400	1.0007
$\hat{\mu}$	0.0217	128800	167800	$4.639 \cdot 10^{-4}$	0.03945	1.006
$\hat{\sigma}$	0.0097	192200	192200	$1.369 \cdot 10^{-4}$	0.0004	0.003

Table 6.3.: Results for BZS using ConfJump.

6.0.2. Trimethoprim

The sampled distributions over the 5 heavy dihedrals of Trimethoprim are shown in figure 6.2. Again, the reference run was created by the ConfJump strategy running

for 120000 steps per chain. The rotational symmetry of dihedral 1 is somewhat imperfectly reflected by the reference run (see top left panel), but still within acceptable limits at a symmetry error of approximately 4.65%. Use of this particular run as reference is justified by the fact that of all Replica Exchange and ConfJump runs it is the one with the lowest mean distance to all others. The rotational symmetry of dihedral 2 situated on the opposite side of the ring on the left hand side in figure 5.2 is considerably better reproduced by the result of the reference run, as a visual examination of the top center panel reveals. The distributions of the dihedrals of the two methoxy groups at the sides of the symmetric ring (bottom panels) are nearly identical (except for a shift by π), which is expected because the ring is symmetric and the functional groups are equal. Therefore, they act chemically and physically in the same way.

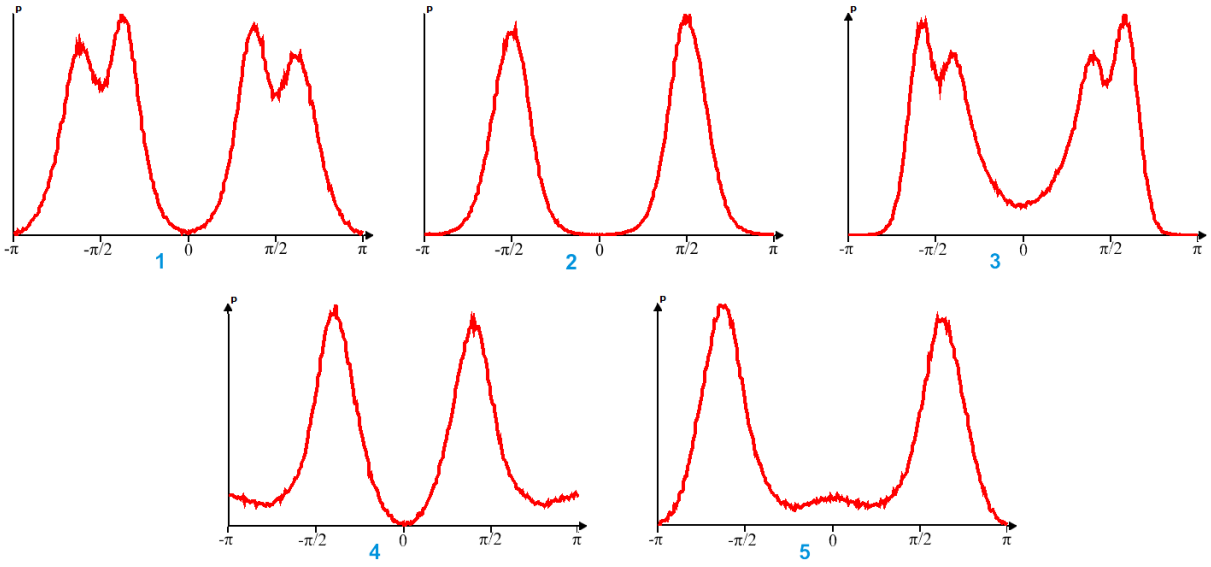

Figure 6.2.: The sampled distributions of the five heavy dihedrals of Trimethoprim. Numbers under the panels refer to the numbering of heavy dihedrals in figure 5.2.

ZIBgridfree produces consistently very low sampling errors with respect to the reference for Trimethoprim at an average value of 2.66% with a standard deviation of 0.6% (see table 6.4. Unfortunately, the symmetry error of the simulations was not measured in the simulations. However, at visual inspection the dihedral distributions look very similar to those in figure 6.2 for all 5 simulation runs (not shown).
ZIBgridfree's overall performance on Trimethoprim is considerably better than for L-benzylsuccinate. The sampling needs on average $9.5 \cdot 10^6$ steps at a standard deviation of only 17000, which results in a mean performance of $4.2 \cdot 10^{-6}$ with a standard deviation of $1.25 \cdot 10^{-6}$.

The Replica Exchange technique produces sampling results for Trimethoprim with a low average sampling error of 3.88% at a standard deviation of 1.16% (see table 6.5,

	error	steps	performance
1	0.0290	$9.715 \cdot 10^6$	$3.548 \cdot 10^{-6}$
2	0.0345	$9.510 \cdot 10^6$	$3.045 \cdot 10^{-6}$
3	0.0244	$9.445 \cdot 10^6$	$4.332 \cdot 10^{-6}$
4	0.0276	$9.545 \cdot 10^6$	$3.802 \cdot 10^{-6}$
5	0.0172	$9.260 \cdot 10^6$	$6.280 \cdot 10^{-6}$
$\hat{\mu}$	0.0266	$9.495 \cdot 10^6$	$4.201 \cdot 10^{-6}$
$\hat{\sigma}$	0.00638	165000	$1.2515 \cdot 10^{-6}$

Table 6.4.: Results for Trimethoprim using ZIBgridfree.

i.e. the results are reliably good. The symmetry error (which is actually the maximum of two symmetry errors, one for each symmetric dihedral) is, however, in the medium range at 10% with a standard deviation of 3.31%. All simulation runs are considered to have converged by the Gelman-Rubin criterion after the maximally allowed number of 10^6 HMC steps.
As all 5 simulations have been run for the same time, the average performance depends solely on the sampling error, which is very low in 4 of 5 cases. The mean performance is thus $27.5 \cdot 10^{-6}$ with a standard deviation of $7.48 \cdot 10^{-6}$. This is 6.5 times as high as that of ZIBgridfree.

	error	steps	performance	E_{sym}	G-R
1	0.0274	10^6	$3.654 \cdot 10^{-5}$	0.1085	1.002
2	0.0347	10^6	$2.882 \cdot 10^{-5}$	0.0749	1.005
3	0.0558	10^6	$1.791 \cdot 10^{-5}$	0.1109	1.007
4	0.0450	10^6	$2.222 \cdot 10^{-5}$	0.1456	1.004
5	0.0312	10^6	$3.206 \cdot 10^{-5}$	0.0615	1.003
$\hat{\mu}$	0.0388	10^6	$2.751 \cdot 10^{-5}$	0.1003	1.004
$\hat{\sigma}$	0.01155	0	$7.484 \cdot 10^{-6}$	0.03309	0.0017

Table 6.5.: Results for Trimethoprim using Replica Exchange.

The ConfJump strategy was more successful in reproducing the reference result than Replica Exchange but less so than ZIBgridfree (see table 6.6. The mean sampling error is 3.46% which is low. The standard deviation is 1.65%. As with the RE simulations, the symmetry errors differ strongly between different sampling runs, which gives rise to the conjecture that rotation around the single bond corresponding to dihedral 1 of the molecule is sterically hindered to a high degree, possibly due to electrostatic attraction between partial charges of different sign in the two rings. The mean symmetry error is 9.7% with a standard deviation of 3.89%.
All simulations have been run for 10^6 steps, 120000 in each chain. Thus, the performance only depends on the sampling error. A correction value of $c = 80000$ was added to the simulation time for identifying representatives of all low-energy regions in conformational space (see table 5.2. The mean performance was $3.332 \cdot 10^{-5}$ with a

standard deviation of $1.812 \cdot 10^{-5}$. For Trimethoprim, ConfJump has a mean performance that is 1.2 times that of Replica Exchange and 7.9 times that of ZIBgridfree.

	error	steps	corrected	performance	E_{sym}	G-R
1	0.0152998	10^6	$1.08 \cdot 10^6$	$6.052 \cdot 10^{-5}$	0.066784	1.0244
2	0.0219898	10^6	$1.08 \cdot 10^6$	$4.211 \cdot 10^{-5}$	0.0724	1.06272
3	0.0332182	10^6	$1.08 \cdot 10^6$	$2.787 \cdot 10^{-5}$	0.078386	1.02529
4	0.0521558	10^6	$1.08 \cdot 10^6$	$1.775 \cdot 10^{-5}$	0.105084	1.02876
5	0.0505242	10^6	$1.08 \cdot 10^6$	$1.833 \cdot 10^{-5}$	0.161262	1.02625
$\hat{\mu}$	0.03463756	10^6	$1.08 \cdot 10^6$	$3.332 \cdot 10^{-5}$	0.0967832	1.033484
$\hat{\sigma}$	0.016546929	0	0	$1.812 \cdot 10^{-5}$	0.038920972	0.016424458

Table 6.6.: Results for Trimethoprim using ConfJump.

6.0.3. BSI

Figure 6.3 illustrates the Boltzmann distribution of BSI sampled by the reference run at 300K, a simulation run of 10^6 steps using the ConfJump method, projected into 5 of its 8 heavy dihedral angles. The top left panel and the central panel at the top show the distributions of the two rotationally symmetric dihedrals that correspond to the bond between the two aromatic 6-rings on the right hand side in figure 5.3 (left) and between the sulfonyl group and the adjacent planar ring (center). Both distributions show only minor flaws at visual inspection. The single bond adjacent to the carboxyl group, which corresponds to dihedral 4 is also rotationally symmetric which is reproduced well by the reference run, as can be seen in the bottom left panel in figure 6.3. The distribution of dihedral 3 which is situated between the S- and the N-atom shows two peaks with very different statistical weights (top right panel), and a similar behavior can be seen in the distribution of dihedral 8 which lies inside the non-aromatic ring of the molecule. The three heavy dihedrals that are not shown have only one visible peak each. Therefore, if BSI should show a similar behavior to cyclohexane with respect to a large conformational change induced by a "flip" of the non-aromatic ring, this behavior is at least not reproduced well by the simulation. However, it is also conceivable that the large functional groups that surround that ring force it into one of the two possible conformations most of the time. This speculation is supported by the fact that none of the 15 sampling runs evaluated below assigned a higher statistical weight to the small peak seen in dihedral 8, and, in fact, most sampling runs failed to reproduce it at all.

The Replica Exchange simulations of BSI all produced results with a very low sampling error except for one outlier with a low sampling error (see table 6.7). The mean sampling error is 3.15% with a standard deviation of 1.45%. However, the symmetry error is high at 24 to 28.5%.
As none of the simulations has converged within the maximally allowed time steps, all simulations ran for 10^6 time steps. The performance thus depends only on the

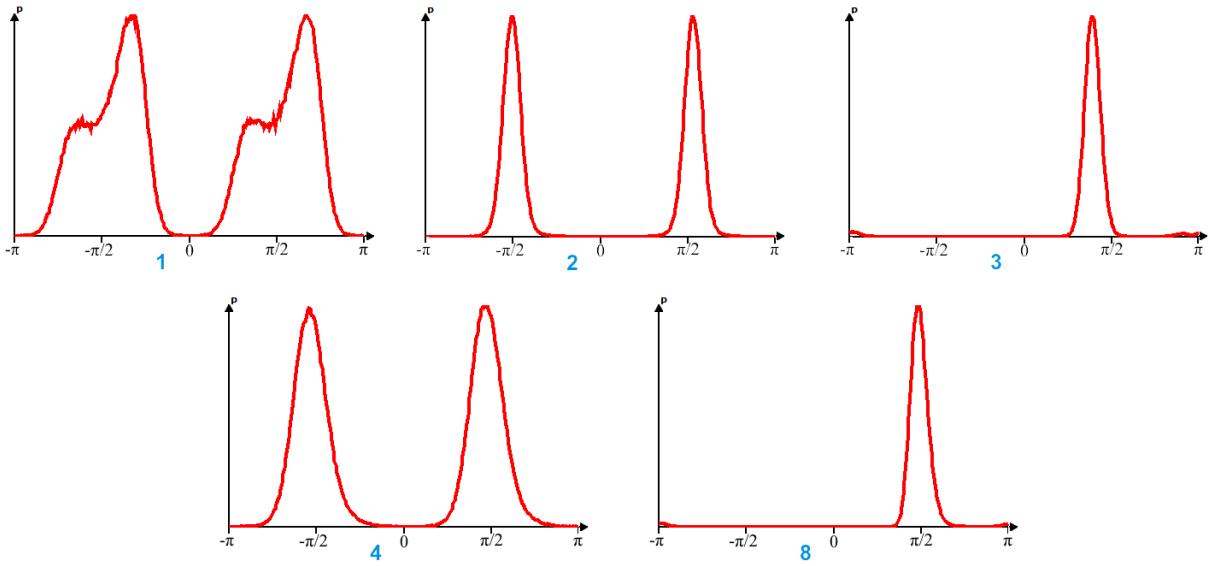

Figure 6.3.: The sampled distributions of five of the eight heavy dihedrals of BSI. The numbers under each diagram correspond to the dihedrals marked in figure 5.3.

sampling error. The mean performance is $3.58 \cdot 10^{-5}$ while the estimated standard deviation is $1.15 \cdot 10^{-5}$.

	error	steps	performance	E_{sym}	G-R
1	0.0203	10^6	$4.927 \cdot 10^{-5}$	0.2723	1.04
2	0.0257	10^6	$3.887 \cdot 10^{-5}$	0.285	1.03
3	0.0274	10^6	$3.645 \cdot 10^{-5}$	0.2429	1.02
4	0.0272	10^6	$3.678 \cdot 10^{-5}$	0.2476	1.03
5	0.0570	10^6	$1.755 \cdot 10^{-5}$	0.2410	1.19
$\hat{\mu}$	0.0315	10^6	$3.578 \cdot 10^{-5}$	0.2578	1.06
$\hat{\sigma}$	0.01452	0	$1.1460 \cdot 10^{-5}$	0.01973	0.074

Table 6.7.: Results for BSI using Replica Exchange.

For BSI, the ConfJump approach also produced results with a low sampling error (see table 6.8). The means was 0.8%, while the standard deviation was 0.32%. These exremely low results compared to the other methods (especially Replica Exchange) combined with the high symmetry error with a very low variance gives some reason to doubt the validity of the reference run.
None of the simulations has converged, like in the case of RE, within the maximally allowed time steps. All simulations ran for 10^6 time steps. Therefore, the performance is dependent only on the sampling error. 93000 steps have been added to the sampling time, corresponding to the time for preprocessing. The mean performance is $1.23 \cdot 10^{-4}$ while the estimated standard deviation is $3.978 \cdot 10^{-5}$.

	error	steps	corrected	performance	E_{sym}	G-R
	0.00534	10^6	$1.093 \cdot 10^6$	$1.714 \cdot 10^{-4}$	0.2516	1.08
	0.00634	10^6	$1.093 \cdot 10^6$	$1.443 \cdot 10^{-4}$	0.2525	1.10
	0.00714	10^6	$1.093 \cdot 10^6$	$1.282 \cdot 10^{-4}$	0.2521	1.04
	0.01369	10^6	$1.093 \cdot 10^6$	$6.684 \cdot 10^{-5}$	0.2487	1.04
	0.00875	10^6	$1.093 \cdot 10^6$	$1.046 \cdot 10^{-4}$	0.2528	1.12
$\hat{\mu}$	0.00825	10^6	$1.093 \cdot 10^6$	$1.231 \cdot 10^{-4}$	0.2516	1.07
$\hat{\sigma}$	0.00329	0	0	$3.975 \cdot 10^{-5}$	0.0017	0.036

Table 6.8.: Results for BSI using ConfJump.

Unfortunately, due to technical difficulties, the results obtained from the ZIBgridfree could not be evaluated for this thesis.

6.0.4. Performance comparison

Figure 6.4 shows a plot of accuracy (1−sampling error) vs. time in HMC steps for the simulations of the three model systems with all ConfJump and Replica Exchange. The corrected times are used for ConfJump.
ConfJump produces a lower sampling error (higher accuracy) than Replica Exchange for all three molecules. Overall, very few simulations converged, due to the strict choice of convergence criteria. In the case of BZS, however, ConfJump was able to beat Replica Exchange on both accounts by producing a better result in a much shorter time. Surprisingly both methods fare best on BSI, the most complex molecule. However, there are reasons to doubt the validity of the reference run in that case.

The mean results of ZMFree are shown in comparison for BZS and Trimethoprim in figure 6.5. ZMFree has a large computational overhead compared to ConfJump and Replica Exchange. In figure 6.5, its results appear far to the right because of this. ZMFree was able to sample the distributions of Trimethoprim and BZS sufficiently well and in the case of Trimethoprim even produced the lowest sampling error of all three methods.

While ZMFree needs a more thorough sampling of the conformational space than the other two methods, it also gains more information than ConfJump and RE. ZIBgridfree is the only method able to compute transition probabilities between the conformations. The high number of basis functions needed for an accurate sampling can in part be dealt with by parallelization. As mentioned in section 3.2, the current implementation already uses a parallelization, and in fact, in every sampling, three partial density functions were sampled at the same time.

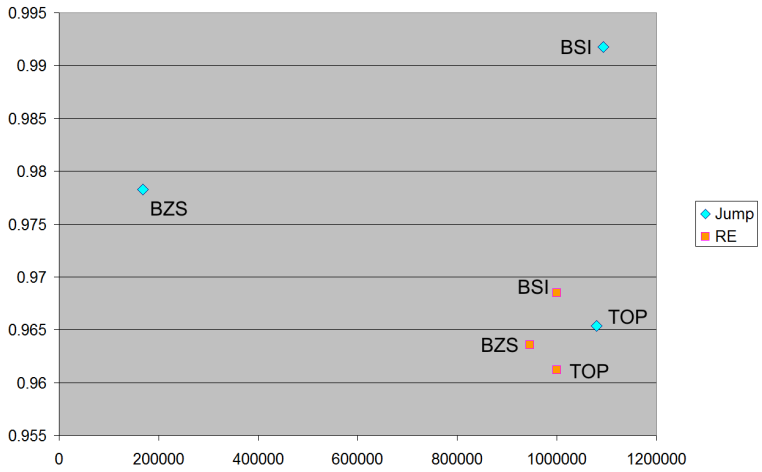

Figure 6.4.: Mean accuracy (y-axis) vs. time (x-axis) for the ConfJump and Replica Exchange simulations.

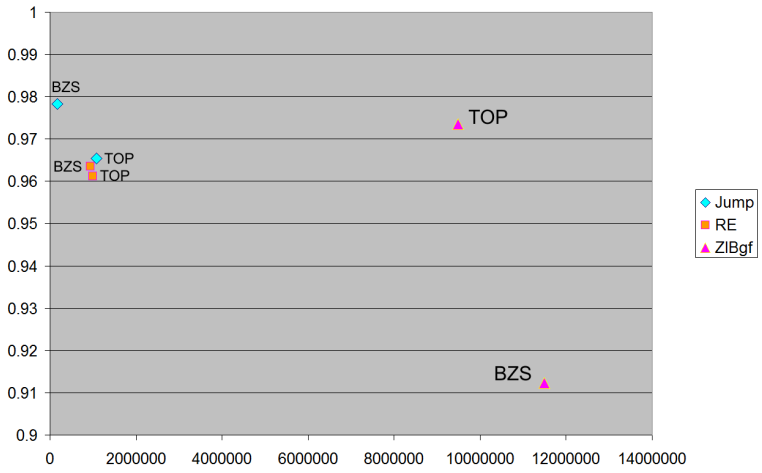

Figure 6.5.: Mean accuracy (y-axis) vs. time (x-axis) for the ZIBgridfree simulations in comparison to the values from the ConfJump and Replica Exchange simulations.

7. Conclusion

In this thesis a method for comparing sampling results from different Markov chain Monte Carlo methods was developed and applied to samplings of three typical drug-like molecules using three different sampling methods, ZIBgridfree, Replica Exchange and ConfJump.

It has been shown that a method is generally more stable the less information it needs about high-energy transition regions in conformational space.

- ConfJump employs knowledge about the shape of the potential and is thus able to bypass high-energy regions altogether.
- In Replica Exchange, high-temperature chains must pass through high-energy regions in sampling in order to be able to discover different low-energy regions which are then sampled accurately by the chain at the sampling temperature. However, Replica Exchange does not need to sample high-energy regions accurately.
- ZIBgridfree, on the other hand, requires accurate sampling of transition regions in order to correctly weight different low-energy regions against each other.

For small to medium-sized molecules where it is affordable to generate representatives of all low-energy regions in conformational space, the ConfJump approach seems to be both the most accurate and the most stable method. By switching between HMC steps and jump steps that carry the system swiftly from one metastable region to the next and thus actively avoiding the problem of broken ergodicity, the ConfJump approach can greatly accelerate the sampling. However, as the dimension of the conformational space grows, ConfJump will invariably become less efficient, less stable and ultimately also less accurate than other methods. On the one hand, low-energy regions in a high-dimensional, rough potential energy landscape will be more irregularly shaped than in lower dimensions, which is a critical problem for the efficiency of ConfJump. This is due to the fact that the jump vector is determined independent of the shape of the target region (solely on the basis of one representative of that region) but is only accepted if it "hits" the target. On the other hand, identifying all low-energy regions in the d-dimensional conformational space has a computational cost that is exponential in d. This soon leads to a prohibitive computational cost as d grows. By relying on precomputed representatives of low-energy regions, ConfJump gives up the crucial advantage of Monte Carlo methods over e.g. numerical integration, namely being able to approximate high-dimensional statistical distributions at a computational cost that does not depend on the dimension of the problem but only on the number of samples generated.

Replica Exchange has been found in numerical experiments to be able to approximate the Boltzmann distributions of drug-sized molecules almost as well as ConfJump, with a slightly higher average sampling error with a somewhat higher variance. This means that it is less stable numerically than ConfJump. Exchanging temperatures between replicas at certain intervals cannot avoid trapping in basins of attraction of local minima as well as the ConfJump approach because no information is available on the destination of the "jump" associated with a replica exchange. One problem of the Replica Exchange method is the enormous amount of redundant data that is generated. When sampling at ten temperatures only one tenth of the data generated during sampling can be used for conformation analysis. In order to be able to reach all regions of conformational space in an acceptable time on average, the maximum temperature must be chosen high enough. However, the hybrid Monte Carlo method and especially the molecular dynamics integration are bound to encounter numerical difficulties when working with very high temperatures. Worse, systems such as DNA or clusters of lipids or proteins that are stabilized by weak molecular interaction forces such as van der Waals forces and hydrogen bonds can impossibly be simulated at high temperatures as high temperatures would break the stabilizing interactions, and the system being studied would simply fall apart. Another problem for Replica Exchange that is independent of this consideration is that the acceptance probability of two replicas exchanging temperatures depends on the potential energy difference between the position states of the two chains. In large systems the interesting low-energy regions will likely be far away from each other which leads to a decrease in the acceptance ratio as it is less likely that two chains are at similar energy levels, and the probability of a high-temperature chain being in a high-energy region is high.

ZIBgridfree has also been found to be able to get very close to a given reference run in most cases. However, the method is not very robust with respect to initial conditions. ZIBgridfree will occasionally generate samplings with a large sampling error. The reason for this is that this method relies on being able to weight all pairs of "adjacent" sampled partial densities correctly against each other, which requires a high accuracy of sampling also in high-energy transition regions which are seldom visited in sampling. Nevertheless, ZIBgridfree must be considered the most promising strategy when the goal is to be able to simulate large systems as no other technique discussed here is, in principle, able to deal with very rough potential energy surfaces on high-dimensional conformational spaces. It seems inevitable to discretize very complex Boltzmann distributions and look at uncoupled partial densities separately. Very likely, ZIBgridfree's approach for weighting partial densities against each other, which relies on accurate sampling of high-energy transition regions, is bound to fail on very rough potential energy surfaces. However, better methods for weighting the different partitions against each other are being discussed already and will be the subject of further research. It might be possible to use the ConfJump method to quickly and accurately explore transition probabilities between partial densities that contain low-energy regions without needing detailed information about the transition regions in between. Additionally, ZIBgridfree is the only method that yields

transition probabilities between metastable regions thus allowing examination of the dynamics of the system under consideration.

The semi-empirical convergence indicator for Markov chain Monte Carlo methods that was developed in this thesis can be used to supplement convergence monitors that are based solely on properties of Markov chains. This convergence indicator is widely applicable as symmetric planar rings and other rotationally symmetric groups are abundant in biomolecules and occur particularly frequently in the class of peptide ligands. In the numerical experiments conducted for this thesis cases were observed where the Gelman-Rubin statistic indicated convergence while the symmetry error was still high as well as the reversed situation. Therefore, the convergence criterion that uses the Gelman-Rubin statistic and the symmetry error in combination is a more powerful criterion than either method alone. The computational cost of the combined method is lower than twice the cost of Gelman-Rubin due to the reusing of histograms by the symmetry monitor. The method owes some of its easy applicability to the graph-theoretic algorithm for finding rotationally symmetric groups in molecules that was developed in this thesis.

Acknowledgments

First, I would like to deeply thank Dr. Marcus Weber for helpful suggestions and fruitful discussions throughout all stages of this work. Marcus Weber is also acknowledged for bringing to my attention the problem of automatic detection of molecule symmetries.

I would like to thank Lionel Walter and Dr. Frank Cordes for interesting discussions and suggestions, Susanna Kube and Marcus Weber for proofreading parts of this thesis, and Prof. Dr. Paul Wrede for pointing out 3D structure generators.

I would further like to thank Johannes Schmidt-Ehrenberg for help with amira and amiraMol and Wolgang Pyszkalski for swift technical assistance in the final stages of this work.

Image Credits

- Figure 2.1, created using `amira` [78].
- Figure 3.1, simulation and image courtesy M. Weber and H. Meyer, taken with permission from [75].
- Figure 3.2 and fig:partialdens, based on plots created with MATLAB [35].
- Figure 3.8, courtesy L. Walter, taken with permission from [71].
- Figure 4.2, based on a diagram created with `gnuplot` [77].
- Figure 4.3, based on a diagram created with `gnuplot` [77].

- Figures 5.1, 5.2, and 5.3, extracted from visualizations created by the Protein Data Bank's "ligand summary" viewer [4].
- Figures 6.4 and 6.5, created with Microsoft Excel.

A. Algorithm for automatic detection of molecule symmetries

The following recursive algorithm is used to detect all single bonds with 180° rotational symmetry. It operates on a graph representation $G = (V, E)$ of the molecule whose nodes $v \in V$ represent atoms (by storing index and atomic number) and whose undirected edges $e = (u, v) = (v, u) \in E \subseteq V \times V$ represent bonds between atoms. This has been implemented as an adjacency list in which each node stores the indices of its neighbors. It is assumed that no atom has more than four binding partners (an extremely rare phenomenon in biomolecules).

An array *discovered* of $N = |V|$ binary flags is used to mark which atoms have already been visited. *discovered*$[i]$ = `True` means that node i lies on one of the two branches that are being compared at that moment. The *discovered* flags are used to prevent branches from growing into themselves or each other – each atom can only belong to one branch in which it also cannot occur twice.

The core of the algorithm is the function compareSubgraphs which determines whether two branches starting with the directed edges *from*1 → *to*1 and *from*2 → *to*2 are isomorphic. This is done by removing the edges (*from*1, *to*1) and (*from*2, *to*2) from the graph and recursively trying to split the component of the remaining graph that contains *to*1 and *to*2 into two isomorphic subgraphs (where *to*1 is in one branch and *to*2 in the other).

```
1:  Initialize list of symmetric dihedrals symList ← [].
2:  for all nodes v with 3 neighbors do
3:      if v is adjacent to a single bond in a 'heavy' dihedral then
                                  ▷ See section 2.5 for definition of 'heavy' dihedrals.
4:          for all neighbors u of v do
5:              for i ← 1, N do
6:                  discovered[i] ← False
7:              end for
8:              discovered[u] ← discovered[v] ← True
9:              Let l, r be the other 2 neighbors of v.
10:             if compareSubgraphs(v, l, v, r) then
11:                 Identify the dihedral D that describes the bond (u, v)
12:                 symList.append(D)
13:             end if
14:         end for
15:     end if
```

```
16: end for

17: function compareSubgraphs(from1, to1, from2, to2)
    ▷ Checks whether the two branches starting with the directed edges from1 → to1
    and from2 → to2 are isomorphic.
18:     if to1 = to2 then                                          ▷ a ring closes
19:         if nNeighbors(to1) ≤ 3 then
20:             ▷ singular ring appendage is part of both branches ⇒ skip
21:             return True
22:         else          ▷ 4 neighbors ⇒ recurse into non-ring neighbors of ring link
23:             Let l, r be the non-ring neighbors of to1 = to2.
24:             discovered[to1] ← True
25:             if compareSubgraphs(to1, l, to1, r) then
26:                 return True
27:             else                                                ▷ backtrack
28:                 discovered[to1] ← False
29:                 return False
30:             end if
31:         end if
32:     else if (atomType[to1] ≠ atomType[to2]) or (nNeighbors(to1) ≠nNeighbors(to2))
    then
33:         ▷ atoms to1 and to2 are incompatible ⇒ backtrack
34:         return False
35:     else if nNeighbors(to1) = 1 then
36:         ▷ reached (compatible) terminal atoms
37:         return True
38:     end if

39:     ▷ Build lists of undiscovered neighbors
40:     neighbors1 ← neighbors2 ← []
41:     for i ← 0, nNeighbors(to1) do
42:         if not discovered[neighbor[to1][i]] then
43:             neighbors1.append(neighbor[to1][i])
44:         end if
45:     end for
46:     for i ← 0, nNeighbors(to2) do
47:         if not discovered[neighbor[to2][i]] then
48:             neighbors2.append(neighbor[to2][i])
49:         end if
50:     end for

51:     if neighbors1.size() ≠ neighbors2.size() then
52:         ▷ one branch grows into the other ⇒ backtrack
53:         return False
```

```
54:     else if neighbors1.size() = 0 then
55:         ▷ a side ring closes on each branch; already checked
56:         return True
57:     end if

58:     ▷ Recurse through branches
59:     a ← neighbors1[0], b ← neighbors1[1], c ← neighbors1[2]
60:     A ← neighbors2[0], B ← neighbors2[1], C ← neighbors2[2]
61:     ▷ (provided these exist)
62:     discovered[to1] ← discovered[to2] ← True
63:     result ← False
64:     if neighbors1.size() = 1 then
65:         ▷ only one path to pursue on each branch
66:         result ← compareSubgraphs(to1, a, to2, A)
67:     else if neighbors1.size() = 2 then
68:         ▷ Either A corresponds to a and B to b or A corresponds to b and B to a:
69:         result ← (compareSubgraphs(to1, a, to2, A)
70:           and compareSubgraphs(to1, b, to2, B))
71:           or (compareSubgraphs(to1, a, to2, B)
72:           and compareSubgraphs(to1, b, to2, A))
73:     else if neighbors1.size() = 3 then
74:         ▷ Check all possible combinations
75:         if compareSubgraphs(to1, a, to2, A) then
76:             result ← (compareSubgraphs(to1, b, to2, B)
77:               and compareSubgraphs(to1, c, to2, C))
78:               or (compareSubgraphs(to1, b, to2, C)
79:               and compareSubgraphs(to1, c, to2, B))
80:         else if (not result) and compareSubgraphs(to1, a, to2, B) then
81:             result ← (compareSubgraphs(to1, b, to2, A)
82:               and compareSubgraphs(to1, c, to2, C))
83:               or (compareSubgraphs(to1, b, to2, C)
84:               and compareSubgraphs(to1, c, to2, A))
85:         else if (not result) and compareSubgraphs(to1, a, to2, C) then
86:             result ← (compareSubgraphs(to1, b, to2, A)
87:               and compareSubgraphs(to1, c, to2, B))
88:               or (compareSubgraphs(to1, b, to2, B)
89:               and compareSubgraphs(to1, c, to2, A))
90:         end if
91:         ▷ Check chirality of corresponding triples of atoms
92:         if result then
93:             result ← checkChirality(to1, from1, neighbors1[0], neighbors1[1], neighbors1[2])
94:             ▷ (using a variant of compareSubgraphs on a second discovered
    array)
95:         end if
```

```
96:     end if
97:     if not result then
98:         discovered[to1] ← discovered[to2] ← False
99:     end if
100:    return result
101: end function
```

The algorithm can easily be adapted for 120° symmetry (not shown).

Bibliography

[1] I. Andricioaiei, J. Straub, and A. Voter. Smart Darting Monte Carlo. *J. Chem. Phys.*, 114(16):6994–7000, 2001.

[2] Ehrhard Behrends. *Introduction to Markov Chains*. Vieweg Verlagsgesellschaft, 1st edition, 2000.

[3] Jeremy M. Berg, John L. Timoczko, and Lubert Stryer. *Biochemistry*. Palgrave Macmillan, 5th edition, 2002.

[4] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235–242, 2000. `http://www.rcsb.org/pdb`.

[5] J.F. Bonnans, J.Ch. Gilbert, C. Lemaréchal, and C.A. Sagastizábal. *Numerical Optimization – Theoretical and Practical Aspects*. Universitext. Springer-Verlag, Berlin, 2003.

[6] Jonas Boström. Reproducing the conformations of protein-bound ligands: A critical evaluation of several popular conformational searching tools. *J. Comput.-Aided Mol. Des.*, 15(12):1137–1152, 2001.

[7] A. Brass, B.J. Pendleton, Y. Chen, and B. Robson. Hybrid Monte Carlo simulations theory and initial comparison with molecular dynamics. *Biopolymers*, 33(8):1307–1315, 1993.

[8] S.P. Brooks and A. Gelman. General Methods for Monitoring Convergence of Iterative Simulations. *J. Comput. Graph. Stat.*, 7(4):434–455, 1998.

[9] S.P. Brooks and G.O. Roberts. Assessing Convergence of Markov Chain Monte Carlo Algorithms. Technical report, University of Cambridge, 1997.

[10] M. Cecchini, F. Rao, M. Seeber, and A. Caflisch. Replica exchange molecular dynamics simulations of amyloid peptide aggregation. *J. Chem. Phys*, 121:10748–10756, 2004.

[11] J.N. Champness, A. Achari, S.P. Ballantine, P.K. Bryant, C.J. Delves, and D.K. Stammers. The structure of Pneumocystis carinii dihydrofolate reductase to 1.9 A resolution. *Structure*, 2(10):915–924, 1994.

[12] M.E. Clamp, P.G. Baker, C.J. Stirling, and A. Brass. Hybrid Monte Carlo: An efficient algorithm for condensed matter simulation. *J. Comput. Chem.*, 15(8):838–846, 1994.

[13] Frank Cordes, Marcus Weber, and Johannes Schmidt-Ehrenberg. Metastable Conformations via successive Perron-Cluster Cluster Analysis of dihedrals. Technical report 02-40, Zuse Institute Berlin, 2002.

[14] J. Couet, S. Li, T. Okamoto, T. Ikezu, and M.P. Lisanti. Identification of Peptide and Protein Ligands for the Caveolin-scaffolding Domain. Implications for the Interaction of Caveolin with Caveolae-Associated Proteins. *J. Biol. Chem.*, 272(10):6525–6533, 1997.

[15] M.K. Cowles and B.P. Carlin. Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *J. Am. Stat. Assoc.*, 91(434):883–904, 1996.

[16] Peter Deuflhard. From Molecular Dynamics to Conformational Dynamics in Drug Design. In M. Kirkilionis, S. Krömker, R. Rannacher, and F. Tomi, editors, *Trends in Nonlinear Analysis*, pages 269–288. Springer-Verlag, Berlin, 2003.

[17] Peter Deuflhard and Christof Schütte. Molecular Conformation Dynamics and Computational Drug Design. In J.M. Hill and R. Moore, editors, *Applied Mathematics Entering the 21st Century. Invited Talks from the ICIAM 2003 Congress, Sydney, Australia*, 2004.

[18] Peter Deuflhard and Marcus Weber. Robust Perron Cluster Analysis in Conformation Dynamics. In M. Dellnitz, S. Kirkland, Neumann M., and C. Schütte, editors, *Lin. Alg. Appl. – Special Issue on Matrices and Methematical Biology*, volume 398C, pages 161–184. 2005.

[19] S. Duane, A.D. Kennedy, B.J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Phys. Lett. B*, 195:216–222, 1987.

[20] M. Filter, M. Eichler-Mertens, A. Bredenbeck, F.O. Losch, T. Sharav, A. Givehchi, P. Walden, and P. Wrede. A Strategy for the Identification of Canonical and Non-canonical MHCI-binding Epitopes Using an ANN-based Epitope Prediction Algorithm. *QSAR & Comb. Sci.*, 25(4):350–358, 2006.

[21] Alexander Fischer. Die Hybride Monte–Carlo Methode in der Molekülphysik. Master's thesis, Freie Universität Berlin, 1997. In German.

[22] Alexander Fischer. An Uncoupling-Coupling Technique for Markov Chain Monte Carlo Methods. Technical report 00-04, Zuse Institute Berlin, 2006.

[23] Daan Frenkel and Berend Smit. *Understanding Molecular Simulation.* Academic Press, 2nd edition, 2002.

[24] Johann Gasteiger and Jens Sadowski. From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev.*, 93:2567–2581, 1993.

[25] A. Gelman. Inference and monitoring convergence. In W. Gilks, S. Richardson, and D.J. Spiegelhalter, editors, *Practical Markov Chain Monte Carlo*, pages 131–143. Chapman & Hall, London, UK, 1996.

[26] A. Gelman and D. Rubin. Inference from Iterative Simulation using Multiple Sequences. *Statist. Sci.*, 7:457–511, 1992.

[27] A. Gelman and D. Rubin. Markov chain Monte Carlo Methods in Biostatistics. *Stat. Meth. Med. Res.*, 5:339–355, 1996.

[28] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London, UK, 1996.

[29] Stefan Goedecker. Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems. *J. Chem. Phys.*, 120(21):9911–9917, 2004.

[30] Y.G. Gogotsi, A. Kailer, and K.G. Nickel. Transformation of diamond to graphite. *Nature*, 401:663–664, 1999.

[31] Thomas A. Halgren. Merck molecular force field. I–V. *J. Comput. Chem.*, 17(5–6):490–641, 1996.

[32] Okamoto Y. Hansmann, U.H.E. Generalized-ensemble Monte Carlo method for systems with rough energy landscape. *Phys. Rev. E*, 56(2):2228–2233, 1997.

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

[34] K. Inaba, S. Turley, T. Iyoda, F. Yamaide, S. Shimoyama, C.R. e Sousa, R.N. Germain, I. Mellman, and R.M. Steinman. The Formation of Immunogenic Major Histocompatibility Complex Class II-Peptide Ligands in Lysosomal Compartments of Dendritic Cells Is Regulated by Inflammatory Stimuli. *J. Exp. Med.*, 191(6):927–936, 2000.

[35] The MathWorks Inc. MATLAB(R) 6.5.0, 1984-2002.

[36] Martin Karplus. Molecular dynamics of biological macromolecules: A brief history and perspective. *Biopolymers*, 68(3):350–358, 2002.

[37] Andrew R. Leach. *Molecular Modelling: Principles and Applications*. Prentice Hall, 2002.

[38] Georg Löffler and Petro E. Petrides. *Biochemie und Pathobiochemie*. Springer-Verlag, 5th edition, 1997. In German.

[39] Z. Lu, H. Hu, W. Yang, and P.E. Marszalek. Simulating Force-Induced Conformational Transitions in Polysaccharides with the SMD Replica Exchange Method. *Biophys. J*, 2006.

[40] S. Mangani, P. Carloni, and P. Orioli. Crystal structure of the complex between carboxypeptidase A and the biproduct analog inhibitor L-benzylsuccinate at 2.0 A resolution. *J. Mol. Biol.*, 223(2):573–578, 1992.

[41] H. Matter, W. Schwab, D. Barbier, G. Billen, B. Haase, B. Neises, M. Schudok, W. Thorwart, H. Schreuder, V. Brachvogel, P. Lonze, and K.U. Weithmann. Quantitative structure-activity relationship of human neutrophil collagenase (MMP-8) inhibitors using comparative molecular field analysis and X-ray structure analysis. *J. Med. Chem.*, 42(11):1908–1920, 1999.

[42] N. Metropolis. The Beginning of the Monte Carlo Method. *Los Alamos Science, Special Issue*, pages 125–130, 1987.

[43] 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(6):1087–1092, 1953.

[44] N. Metropolis and S. Ulam. The Monte Carlo method. *J. Am. Stat. Assoc.*, 44:335–341, 1949.

[45] Holger Meyer. Die Implementierung und Analyse von HuMFree – einer gitterfreien Methode zur Konformationsanalyse von Wirkstoffmolekülen. Master's thesis, Freie Universität Berlin, 2005. In German.

[46] Holger Meyer, Frank Cordes, and Marcus Weber. ConFlow: A new space-based Application for complete Conformational Analysis of Molecules. Technical report 06-31, Zuse Institute Berlin, 2006. in preparation.

[47] Holger Meyer, Marcus Weber, Alexander Riemer, and Lionel Walter. ZIB-gridfree, 2004–2006. Software package for HMC simulation and conformation analysis based upon C++ classes of amiraMol [56] using the Merck Molecular Force Field [31] implemented by T. Baumeister and parametrized by F. Cordes. Robust Perron Cluster Analysis implemented by M. Weber and J. Schmidt-Ehrenberg. Status: August 2006. Software owned by Zuse Institute Berlin.

[48] David L. Nelson and Michael M. Cox. *Lehninger Principles of Biochemistry*, chapter 1, pages 16–21. W.H. Freeman, New York, NY, USA, 4th edition, 2004.

[49] Adrian Patrascioiu. The Ergodic Hypothesis: A Complicated Problem in Mathematics and Physics. *Los Alamos Science, Special Issue*, pages 263–279, 1987.

[50] RS Pearlman. Rapid generation of high quality approximate 3D molecular structures. *Chem. Des. Auto. News*, 2:1–7, 1987.

[51] J.W. Pitera and W. Swope. Understanding folding and design: Replica-exchange simulations of "trp-cage" miniproteins. *PNAS*, 100(13):7587–7592, 2003.

[52] Martin Riedmiller and Heinrich Braun. A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm. In *Proceedings of the IEEE International Conference on Neural Networks*, pages 586–591. IEEE Press, 1993.

[53] Daniel Ruiz. A Scaling Algorithm to Equilibrate both Rows and Columns in Matrices. Technical report RAL-TR-2001-034, Rutherford Appleton Laboratory, 2001.

[54] K.Y. Sanbonmatsu and A.E. Garcia. Structure of met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics. *Proteins*, 46:225–234, 2002.

[55] Tamar Schlick. *Molecular Modeling and Simulation: An Interdisciplinary Guide*. Springer-Verlag, New York, NY, USA, 2002.

[56] Johannes Schmidt-Ehrenberg, Daniel Baum, and Hans-Christian Hege. Visualizing dynamic molecular conformations. In *IEEE Visualization 2002*, pages 235–242. IEEE Computer Society Press, 2002.

[57] Ch. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. A Direct Approach to Conformational Dynamics based on Hybrid Monte Carlo. *J. Comput. Phys., Special Issue on Computational Biophysics*, 151:146–168, 1999.

[58] Christof Schütte. *Conformational Dynamics: Modelling, Theory, Algorithm and Application to Biomolecules*. Habilitation thesis, Freie Universität Berlin, 1998.

[59] Christof Schütte and Wilhelm Huisinga. *Biomolecular Conformations can be Identified as Metastable Sets of Molecular Dynamics*, volume X, pages 699–744. North-Holland, 2003.

[60] H. Senderowitz, F. Guarnieri, and W.C. Still. A Smart Monte Carlo Technique for Free Energy Simulations of Multiconformal Molecules. Direct Calculation of the Conformational Population of Organic Molecules. *J. Am. Chem. Society*, 117:8211–8219, 1995.

[61] H. Senderowitz and W.C. Still. Simple but smart monte carlo algorithm for free energy simulations of multiconformational molecules. *J. Comput. Chem.*, 19(15):1736–1745, 1998.

[62] D. Shepard. A two-dimensional interpolation function for irregularly spaced data. In *Proc. 23rd ACM Nat. Conf.*, pages 517–524, 1968.

[63] D. Stalling, M. Westerhoff, and H.C. Hege. Amira: A Highly Interactive System for Visual Data Analysis. In C.D. Hansen and C.R. Johnson, editors, *The Visualization Handbook*, chapter 38, pages 749–767. Elsevier, 2005.

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

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

[66] G.M Torrie and J.P. Valleau. Monte Carlo free energy estimates using non-Boltzmann sampling: Application to the sub-critical Lennard-Jones fluid. *Chem. Phys. Lett.*, 28(4):578–581, 1974.

[67] G.M Torrie and J.P. Valleau. Monte Carlo study of a phase-separating liquid mixture by umbrella sampling. *J. Chem. Phys.*, 66(4):1402–1408, 1977.

[68] G.M Torrie and J.P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.*, 23:187–199, 1977.

[69] L. Verlet. Computer "Experiments" on Classical Fluids I. thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.*, 159:98–103, 1967.

[70] David Wales. *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses (Cambridge Molecular Science)*. Cambridge University Press, 2004.

[71] Lionel Walter and Marcus Weber. ConfJump: A fast biomolecular sampling method which drills tunnels through high mountains. Technical report 06-26, Zuse Institute Berlin, 2006.

[72] Marcus Weber. Clustering by using a simplex structure. Technical report 04-03, Zuse Institute Berlin, 2004.

[73] Marcus Weber. *Meshless Methods in Conformation Dynamics*. PhD thesis, Freie Universität Berlin, 2006.

[74] Marcus Weber, Susanna Kube, Lionel Walter, and Peter Deuflhard. Well-conditioned computation of probability densities for metastable conformations. Technical report 06-39, Zuse Institute Berlin, 2006. in preparation.

[75] Marcus Weber and Holger Meyer. ZIBgridfree – Adaptive Conformation Analysis with qualified Support of Transition States and Thermodynamic Weights. Technical report 05-17, Zuse Institute Berlin, 2006.

[76] L.T. Wille and J. Vennik. Computational complexity of the ground-state determination of atomic clusters. *J. Phys. A*, 18(8):L419–L422, 1985.

[77] Thomas Williams and Colin Kelley. Gnuplot, version 4.0, 1986–1993, 1998, 2004. `http://www.gnuplot.info`.

[78] ZIB and Mercury Computer Systems, Berlin. amira and amiraMol, 1999–2004.