2 Simulation of Biomolecules

Hamiltonian Systems and *Statistical Physics* forms the basis of mathematical modeling of biomolecules. After introducing Hamiltonian mechanics, we describe commonly used force fields to set up a classical Hamiltonian of a molecular system, and introduce two major approaches to extract statistical and dynamical information: *Molecular Dynamics* and *Monte Carlo methods*. In Sect. 2.6 we define *metastable conformations* of a biomolecule in terms of a canonical ensemble.

2.1 Hamiltonian Mechanics

Biomolecular systems are typically described in terms of classical mechanics [6], which is a good compromise between a detailed and reliable description of the dynamics of a (bio-)molecule and its computational feasibility [2, 60].

A broad class of molecular systems with n atoms (or center of forces) is described via *generalized coordinates* $x \in \mathbb{R}^{3n}$ and *generalized momenta* $p \in \mathbb{R}^{3n}$ by a *separable Hamiltonian*

$$\mathcal{H}(x,p) = \mathcal{T}(p) + \mathcal{V}(x), \tag{2}$$

where $\mathcal{H}(x,p)$ defines the total energy of the system, which further splits into a kinetic part $\mathcal{T}(p)$ and a potential part $\mathcal{V}(x)$, respectively.

Hamiltonians that are used to model biomolecules are of a more specific form: x refers to the Cartesian coordinates of the n atoms and $\mathcal{H}$ is of the form

$$\mathcal{H}(x,p) = \frac{1}{2}p^T M^{-1} p + \sum_{i=1}^{k} \mathcal{V}_{(i)}(x), \qquad x,p \in \mathbb{R}^{3n} \tag{3}$$

where the $3n \times 3n$ diagonal matrix M is a mass matrix that contains the masses of the respective atoms on its diagonal. In $\mathcal{V}$ interaction between atoms are modeled as sums of various potential parts $\mathcal{V}_{(i)}$.

From $\mathcal{H}$ one can derive the *canonical equations of motion*

$$\dot{x} = \frac{\partial \mathcal{T}}{\partial p} \qquad \text{and} \qquad \dot{p} = -\frac{\partial \mathcal{V}}{\partial x}, \tag{4}$$

which forms an autonomous system of $6n$ first order ordinary differential equations. Since (4) is only dependent on $\mathcal{H}$, the potential $\mathcal{V}$ and the mass matrix M totally determine the dynamics of the molecular system.

Equations (4) together with initial values $x_0 = x(0)$ and $p_0 = p(0)$ for the coordinates and momenta, respectively, form an initial value problem. If the right hand side of (4) is locally Lipschitz-continuous, a unique solution does exist. In that case the phase flow

$$\Phi^t(x_0,p_0) := ((x(t),p(t));\, (x_0,p_0)) \tag{5}$$

denotes the state of the system at time t for an initial value (x_0, p_0).

The special structure of the canonical equations implies at least three remarkable properties of Hamiltonian systems [6]:

1. *Conservation of the total energy* $E \equiv H(x(t), p(t))$, which follows from

$$\frac{dH}{dt} = \sum_{k=1}^{3n} \left(\frac{\partial H}{\partial x_k} \dot{x_k} + \frac{\partial H}{\partial p_k} \dot{p_k} \right) = \sum_{k=1}^{3n} \left(\frac{\partial H}{\partial x_k} \frac{\partial H}{\partial p_k} - \frac{\partial H}{\partial p_k} \frac{\partial H}{\partial x_k} \right) = 0.$$

2. *Time-reversibility* of the phase flow Φ^t, i.e.,

$$\Phi^t(x', -p') = (x, -p) \qquad \text{for} \quad \Phi^t(x, p) = (x', p').$$

   Time-reversibility follows from the special structure of (4) and the general property of a phase flow Φ to be symmetric (i.e., $\Phi^{-t}\Phi^t(x, p) = (x, p)$).

3. *Conservation of the phase space volume* is known as Liouville's theorem: a global property of Φ is to leave the volume $V(A) = \int_A dx\, dp$ of a subsets A of the phase space invariant, i.e.,

$$V(\Phi^t A) = V(A),$$

   which is a direct consequence of the symplecticness of a Hamiltonian phase flow.

These three properties will become of special importance in the context of the Hybrid Monte Carlo method in Sect. 4.2.2.

2.2 Molecular Force Fields

The first step before a simulation can be started for a specific molecular system is to set up a Hamiltonian that reflects reasonable well all properties of interest. Then, in a next step, numerical and stochastic simulations provide insight into diverse aspects such as the behavior of a single long time trajectory, thermodynamical quantities, or structural information. That way, computer simulations help to fill the gap between theory and experiment.

The quantities of interest depend largely on the kind of molecular system under consideration (e.g., gases, liquids, molecules, DNA-segments, small peptides, large proteins). Gases and liquids are often simulated in order to investigate thermodynamic quantities when the system undergoes a phase transition, whereas a primary question concerning biomolecules is to find out typical three-dimensional structures at a fixed temperature.

Different kind of interactions act between atoms as forces. In quantum mechanics the motion of atoms is primarily governed by electromagnetic interactions. These interactions are associated to the general form of a

classical biomolecule's potential function, which is often described by the following parts:

$$\begin{aligned}\mathcal{V}(x) &= \sum_{\text{bonds}} \mathcal{V}_{\text{bonds}} + \sum_{\text{angles}} \mathcal{V}_{\text{angles}} + \sum_{\text{torsions}} \mathcal{V}_{\text{torsions}} \\ &\quad + \sum_{\text{atom pairs}} (\mathcal{V}_{\text{Lennard-Jones}} + \mathcal{V}_{\text{Coulomb}})\end{aligned}$$

The potential function $\mathcal{V}$ can roughly be classified into three parts:

1. *bonded interactions:* The bond structure of a molecule determines a number of short-range interactions. A harmonic representation is used for bond and bond-angle oscillations, whereas a typically three-minima potential is used for torsion angles.

2. *non-bonded interactions:* Non-bonded interactions due to electronic and nuclear charges are modeled via pairwise interactions of Lennard-Jones- and Coulomb-type potentials. They have a long-range effect and make up the biggest part in the computation of larger systems.

3. *interactions with an environment:* A biomolecule is typically surrounded by water molecules, which additionally lead to long-range Lennard-Jones- and Coulomb-type interactions between the biomolecule and its environment. In many biomolecular simulations, however, the Hamiltonian solely consists of intra-molecular forces (i.e., simulation is carried out in vacuum).

To choose a suitable Hamiltonian including all of its parameters for a given molecular system is a formidable task. Parameters for potential parts are derived from a mixture of experimental sources, physical insight, or quantum mechanical simulations of small subsystems [2, 60]. We next have a closer look on some commonly used force field, which allows to set up a Hamiltonian for a wide range of molecules. We start with a simple model for n-alkanes, which will serve us as an algorithmic test environment in the following.

United Atom Representation. Ryckaert and Bellemans [108, 109] proposed in 1967 a simple model for *n-alkanes* in order to investigate by simulating a gas or liquid of similar n-alkanes of how the internal structure of n-alkanes affects thermodynamical quantities and vice versa. An n-alkane

$$\mathrm{CH_3} - \mathrm{CH_2} - \cdots - \mathrm{CH_2} - \mathrm{CH_3}$$

is a linear chain of single bonded carbon atoms, with remaining valencies are bonded to hydrogens. In this model, the term "united atom" refers to

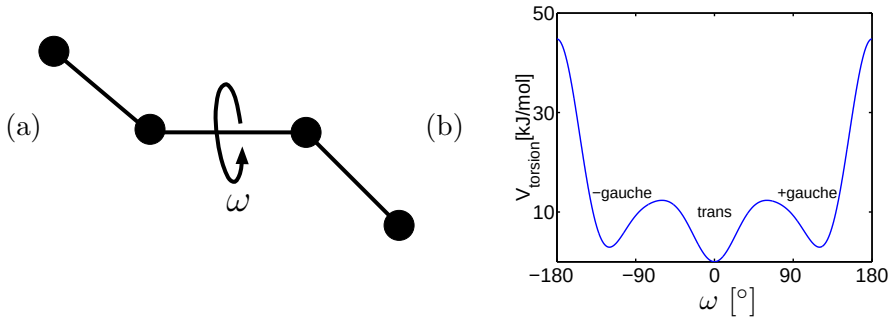

Figure 4: (a) United atom model of n-butane with torsion angle ω. (b) Torsion angle potential $\mathcal{V}_{\text{tor}}$. The main minimum corresponds to the *trans* orientation of the angle, the two side minima to the $\pm$*gauche* orientations.

the approach to model CH_3 and CH_2 groups as single center of forces, which leads to drastically reduced computational cost for simulations.

To set up a Hamiltonian, let $\mathcal{H}(x,p)$ with $x,p \in \mathbb{R}^{3d}$ be the Hamiltonian for d united atoms, $x_k, p_k \in \mathbb{R}^3$ for $k = 1,\dots,d$ the respective coordinates and momenta of one united atom, and M the mass matrix. Then, for a single n-alkane its Hamiltonian is given by

$$\begin{aligned}
\mathcal{H}(x,p) \;=\; & \frac{1}{2}\,p^T M^{-1} p && \text{kinetic part } \mathcal{T}(p)\\
& + \textstyle\sum_{i=1}^{k-1} \mathcal{V}_{\text{bonds}}(x_i, x_{i+1}) && \text{bond terms}\\
& + \textstyle\sum_{i=1}^{k-2} \mathcal{V}_{\text{angle}}(x_i, x_{i+1}, x_{i+2}) && \text{bond angle terms}\\
& + \textstyle\sum_{i=1}^{k-3} \mathcal{V}_{\text{tor}}(x_i, x_{i+1}, x_{i+2}, x_{i+3}) && \text{torsion angles}\\
& + \textstyle\sum_{\substack{i,j\\ i<j-3}} \mathcal{V}_{\text{LJ}}(x_i, x_j) && \text{Lennard-Jones terms}
\end{aligned} \tag{6}$$

The Hamiltonian consists of the kinetic part, and the potential part splits into bond- and angle oscillations, torsion angle rotations, and Lennard-Jones terms for non-bonded interactions (for details see [40, 109]).

The Hamiltonian of a single n-butane (see Fig. 4) will serve us as an illustrative example throughout this thesis. It consists of four united-atom groups only (two CH_3 and two CH_2 compounds), yet it reveals two typical problems of larger biomolecules: a multiscale dynamics due to fast oscillations in the bond and bond-angle potentials compared to slow overall structural changes, which are essentially effected by torsion angle rotations; and

the formation of metastable conformations, which can also be described in terms of the torsion angle by its *trans* and $\pm$-*gauche orientations*.

GROMOS96. For larger biomolecules additional and refined atomic interactions can be taken into account by the GROMOS96 force field [127]. Comparable to the united atoms of the n-alkane model the term "extended atom" is used to denote that some hydrogens are covered by corresponding heavy atoms. Moreover, GROMOS96 contains an extra covalent energy term for out-of-plane oscillations.

As an example, Fig. 5 shows the triribonucleotide adenylyl*(3'-5')*cytidylyl *(3'-5')*cytidin [*r(ACC)*]. The global structure of r(ACC) can be roughly described by eight parameters per nucleotide. From a biochemical point of view, the torsion angles χ and the the pseudorotation angle P and its phase θ [4] are of special interest for conformational analysis.

MMFF. In the all-atom Merck Molecular Force Field (MMFF) [62] all H-atoms are modeled explicitly. This gives a more detailed description of atomic interactions but leads in contrast to the other force fields to an increase in computational cost. One of its distinctive features is to allow the setup of Hamiltonians for a wide class of molecular system (e.g., Fig. 1 on page 6 shows the HIV inhibitor VX-478, a small biomolecule that can inhibit the function of HIV protease). Also, we analyze n-pentane in Sect. 6.2 as well as a larger biomolecule in Sect. 6.3 by means of MMFF.

Remarks. As we can see from these three different force fields, the choice of the Hamiltonian for a model system is not a priori given and is dependent on many factors as detail of description, range of applicability, or computational cost.

A big challenge in modeling biomolecules is the inclusion of a solvent in the model, which makes the introduction of boundary conditions necessary. Yet, a direct approach (e.g., by explicitly modeling interactions between the biomolecule and a huge amount of water molecules) leads to a drastic increase in computational cost. These and further important aspects of modeling and simulation are discussed in detail in standard text books about molecular dynamics [2, 46, 60].

2.3 Molecular Dynamics

Now assume that we already have set up a separated Hamiltonian $\mathcal{H} = \mathcal{T} + \mathcal{V}$ by means of one of the force fields described in Sect. 2.2. Then, the canonical equations are derived as in (4) and the deterministic dynamics of the system is given by the trajectory $\Phi^t(x_0, p_0)$ for some initial values (x_0, p_0). In practice, it is for all but the simplest Hamiltonians impossible to determine the exact solution $\Phi^t(x_0, p_0)$; instead, we have to use some integrator (i.e., a

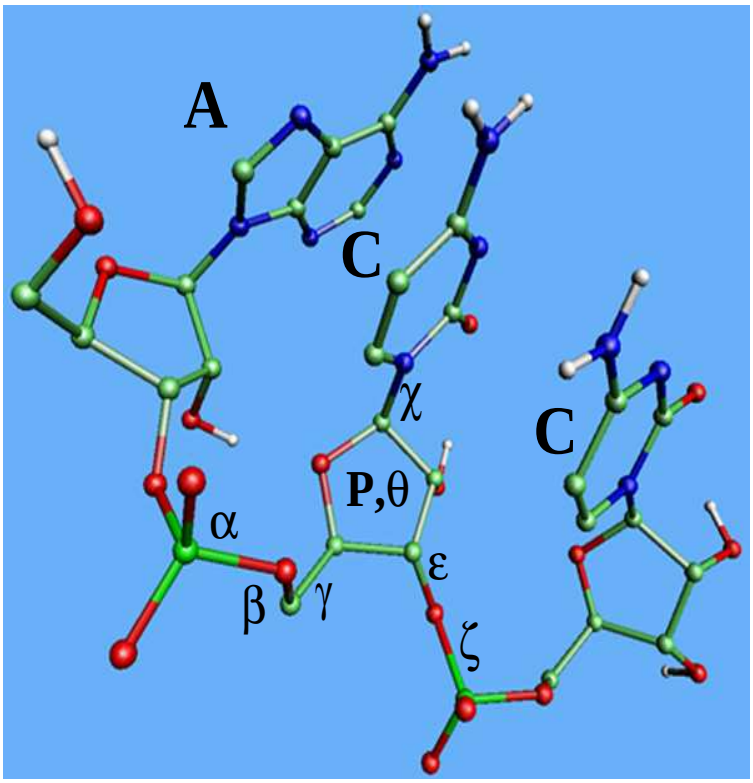

Figure 5: The triribonucleotide adenylyl*(3'-5')*cytidylyl*(3'-5')*cytidin [*r(ACC)*] in the extended atom representation of GROMOS96 [127]. A and C denote the bases adenine and cytosine. Small Greek letters refer to torsion angles, which are necessary for a rough reconstruction of the molecule's configuration. The torsion angles of the ribose (i.e., the five atoms forming a ring structure) can be approximated by the pseudorotation angle P and the phase θ [4].

numerical integration scheme), and therefore inevitably introduce some error by computing a discretized numerical solution $(\Psi^\tau)^k(x_0, p_0)$ at discrete time steps $k\tau$ for $k = 1, \ldots, n$ [26].

Leapfrog. A well-known numerical method for separable Hamiltonians is the *Leapfrog* (or *Verlet*) integrator [128]. Its integration scheme for one time step τ is given by

$$(x_{k+1}, p_{k+1}) = \Psi^\tau(x_k, p_k),$$

where

$$\begin{aligned} x_{k+\frac{1}{2}} &= x_k + \frac{\tau}{2}\frac{\partial \mathcal{T}}{\partial p}(p_k), \\ p_{k+1} &= p_k - \tau \frac{\partial \mathcal{V}}{\partial x}(x_{k+\frac{1}{2}}), \\ x_{k+1} &= x_{k+\frac{1}{2}} + \frac{\tau}{2}\frac{\partial \mathcal{T}}{\partial p}(p_1). \end{aligned}$$

Apparently, this scheme consists of a half-step in the coordinates x, a full step in the momenta p, and another half-step in x; all in all we need only one evaluation of the force field per time step. For an iterative computation of $\Psi^{\tau k}$ the last half-step of x_{k+1} can be further combined with the first half-step of $x_{k+1+\frac{1}{2}}$, resulting in two shifted series of full steps in the coordinates and momenta, respectively. The shift of $\tau/2$, which is due to the initial and final half-step in x, guarantees the Leapfrog scheme to be time-reversible. Moreover, the discrete phase flow $\Psi : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ is symplectic, and therefore inherits two important properties of the phase flow Φ.

Other sophisticated integrators could be used instead, many of which are also reversible and symplectic [110]. Reversible and symplectic integrators are also known to reproduce conservation of energy reasonable well. The Verlet scheme is the simplest representative in the family of so-called partitioned Runge-Kutta methods [61], with an order of approximation of 2. Higher order integrators produce much better results for the price of a higher computational cost, at least if one is interested in good discrete approximations wrt. the exact solution. We will stick to the Leapfrog scheme in the following, since in the context of MCMC it is sufficient to use a computational inexpensive scheme that is reversible and symplectic.

Interpretation of Simulation Outcome. To start a numerical long-term simulation we have to determine reasonable initial values for the coordinates x_0 and momenta p_0. It is impossible to get exact values from experimental measurements, and even if one agrees on x_0 the remaining energy has to be assigned rather randomly on p_0. Long-term dynamics is known to be chaotic, and even worse, unpredictable errors in long-term simulations can cause the approximated discrete solution to end up in a totally different part of the phase space than the exact solution; additionally, preservation of energy is not guaranteed for long time spans and large time steps. Nevertheless, short-time simulations do not suffer from numerical problems, and sometimes even long-term simulations can provide valuable insight, especially for non-equilibrium situations.

2.4 Canonical Ensemble

Statistical physics provides an elegant way to get rid of the problem of initial values and erroneous long-term simulations. The focus shifts towards a global view of the molecular system involving the entire phase space rather than the part sampled by a single energy preserving deterministic trajectory. By probabilistic modeling it is possible to obtain averages of observables or to identify typical three-dimensional structures defined in a statistical setting.

The *canonical ensemble* (other frequently used terms are: *canonical distribution*, *Gibbs-*, *Boltzmann-*, or *Gibbs-Boltzmann distribution*) is given on a continuous state space Ω by its density

$$f_{\mathcal{H}}(x,p) = \frac{h_{\mathcal{H}}(x,p)}{Z_{h_{\mathcal{H}}}} = \frac{\exp[-\beta\,\mathcal{H}(x,p)]}{\int_\Omega \exp[-\beta\,\mathcal{H}(x,p)]\,dx\,dp}, \tag{7}$$

where $h_{\mathcal{H}}(x,p) = \exp[-\beta\,\mathcal{H}(x,p)]$ is the unnormalized density, $Z_{h_{\mathcal{H}}}$ its normalizing constant, and $\beta = 1/(k_{\mathrm{B}}T)$ the inverse temperature depending on the temperature T and Boltzmann's constant k_{B}. From a physical point of view the canonical ensemble is associated with a simulation in a heat bath (i.e., a simulation with constant volume, temperature, and number of particles).

Since $\mathcal{H}$ is separable we can split f into its potential and momenta part:

$$f_{\mathcal{V}}(x) = \frac{\exp[-\beta\,\mathcal{V}(x)]}{Z_{h_{\mathcal{V}}}(x)} \qquad \text{and} \qquad f_{\mathcal{T}}(p) = \frac{\exp[-\beta\,\mathcal{T}(x)]}{Z_{h_{\mathcal{T}}}(x)} \tag{8}$$

The momenta part $f_{\mathcal{T}}$ is a multivariate Gaussian distribution and can therefore be treated analytically. The real challenge is to extract information from the potential part $f_{\mathcal{V}}$.

Typical observables are the energy or the energy of some potential part. The probability of a biomolecule to have some special structural property can be stated as

$$\int_\Omega \mathbf{1}_A(x) f_{\mathcal{V}}(x)\,dx = \int_A f_{\mathcal{V}}(x)\,dx$$

where the subset $A \subseteq \Omega$ could for example express restrictions on some internal degrees of freedom.

Example of a Canonical Ensemble. Let us again consider the united atom representation of n-butane. In order to illustrate two characteristics typical for the canonical ensemble we have a closer look at the 12-dimensional probability density function $f_{\mathcal{V}}$ for temperatures at 300 K and 1000 K.

1. In Fig. 6 (a) the two distributions are plotted versus the torsion angle (see Fig. 4). The probability clearly concentrates in low energy regions; an effect that becomes more apparent for decreased temperature (in fact, in the limit $\beta \to \infty$ all probability will be concentrated around the set of global minima).

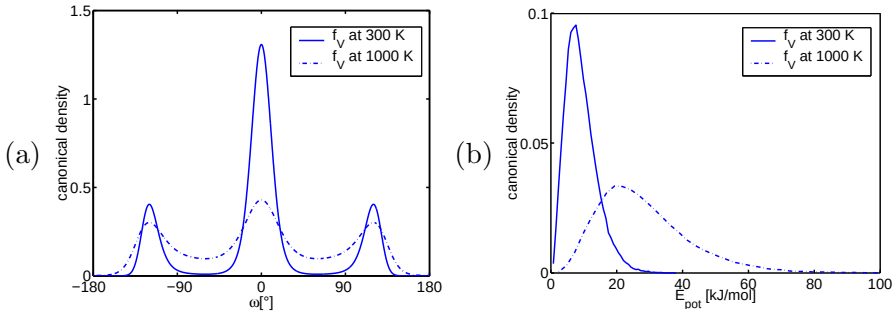

Figure 6: Canonical distribution at $T = 300\,\mathrm{K}$ and $T = 1000\,\mathrm{K}$. (a) Plotted versus the torsion angle. (b) Energy distribution of $f_{\mathcal{V}}$.

2. In Fig. 6 (b) the energy distribution of $f_{\mathcal{V}}$ is plotted. We observe a relatively small overlap between the two distributions. The overlap decreases for higher dimensions and larger temperature differences.

In summary, by decreasing the temperature T the density concentrates more and more in the local minima of $\mathcal{V}$ and therefore in lower energy regions. Although the canonical densities gives the impression of a big overlap in Fig. 6 (a), one should have in mind that in fact the actual overlap can be very small.

Remark. Depending on the modelled experimental situation and the application in mind, other ensembles than the canonical one are used in statistical physics (e.g., the *microcanonical* or *grand-canonical ensemble*). Yet, for simulations in our context (an ensemble of a system consisting of a single molecule with or without embedding in a solvent) where one aims at an understanding of internal structures rather than interactions between different molecules the canonical ensemble is the method of choice.

2.5 Sampling Schemes

In order to extract (thermo-)dynamical quantities one needs to gather information about the canonical distribution, which is usually done by drawing samples from $f_{\mathcal{V}}$. Two main approaches for this task are *Molecular Dynamics* and *Markov chain Monte Carlo*.

2.5.1 Canonical Molecular Dynamics

A simple application of a discrete phase flow Ψ to a molecular system would at best draw samples from a microcanonical ensemble (provided that discretization errors can be neglected and the physical ergodic hypotheses is valid for the molecular system under consideration). Yet, by appropriate

(stochastic) remodeling it is possible to draw samples from a canonical ensemble by means of molecular dynamics. However, one should keep in mind that pure molecular dynamics approaches are always prone to errors due to numerical integration.

Nosé-Hoover Thermostats. In Nosé-Hoover thermostats [70, 100] additional degrees of freedom are introduced which act as an external system on the physical system. Central to this approach is that the microcanonical distribution in the augmented set of variables is equivalent to a canonical distribution, which is obtained by projecting on the original coordinates and appropriate scaled momenta.

Langevin Dynamics. Another possibility to extend the canonical equations is to introduce a stochastic term that aims at resembling a stochastic interaction with a heat bath [67]. A stochastic differential equation of that kind is the Langevin equation

$$\begin{pmatrix} \dot{x} \\ \dot{p} \end{pmatrix} = \begin{pmatrix} p \\ -\nabla\mathcal{V}(x) - \gamma\, p + \sigma\dot{W} \end{pmatrix}, \tag{9}$$

where $\gamma > 0$ is some friction constant and $\sigma\dot{W}$ some white noise given in terms of a standard Brownian motion W. One can show that (9) defines a continuous time Markov process on the phase space with the canonical distribution $f_{\mathcal{H}}$ as its invariant distribution; the temperature β given by $\beta = 2\gamma/\sigma^2$ is the result of an equilibration between the friction constant γ and stochastic excitation regulated by σ [104].

In the high friction case ($\gamma \gg 1$) the Langevin equation can be approximated by the Smoluchowski equation

$$\dot{\mathbf{x}} = -\frac{1}{\gamma}\,\nabla\mathcal{V}(x) + \frac{\sigma}{\gamma}\,\dot{W}\,, \tag{10}$$

which similar to the Langevin equation defines a continuous time Markov process with the canonical distribution $f_{\mathcal{V}}$ as its invariant distribution, this time restricted on the coordinate space.

2.5.2 Monte Carlo Schemes

In a Markov chain Monte Carlo simulation one aims at drawing samples from $f_{\mathcal{V}}$ irrespective of any dynamical information. In short, one needs to perform the following two steps:

1. *Modeling*
   Construct a transition kernel K with $f_{\mathcal{V}}$ being its unique stationary distribution. The crucial point is to find a K that leads to a rapidly mixing Markov chain.

2. *Simulation*
   Realize a Markov chain $\mathcal{X}$ associated with K. Markov chain theory guarantees that for $n \to \infty$ (where n is the number of update steps) we obtain samples from $f_{\mathcal{V}}$.

Due to its generality and simplicity Markov chain Monte Carlo is a powerful approach to draw samples from high dimensional probability distributions like $f_{\mathcal{V}}$. On the other hand, only few of them (e.g., HMC) are accurate to cope with problems arising from high-dimensionality and rugged energy landscape of biomolecules. Like in our UC approach, often a well established MCMC method serves as the basis for more sophisticated approaches which draw samples from generalized ensembles. With the notable exception of HMC, Markov chain Monte Carlo does not provide dynamical information wrt. the Hamiltonian, but rather makes use of the freedom to propose non-physical updates to improve mixing properties.

2.6 Metastable Conformations

Given a Hamiltonian $\mathcal{H}$ and an inverse target temperature β we can start a simulation in the canonical ensemble by either a molecular dynamics or Monte Carlo method. If metastability is present in the system a simulation will remain for a rather long time in some subset A of the state space Ω before a sudden change moves the system on to another subset B.

Let us denote by $\kappa(A, B)$ the transition probability from A to B wrt. the canonical density $f_{\mathcal{V}}$. With K being the transition kernel of the Markov chain under consideration we have

$$\kappa(A, B) = \frac{\int_A K(x, B) f_{\mathcal{V}}(x)\, dx}{\int_A f_{\mathcal{V}}(x)\, dx}, \tag{11}$$

where $K(x, B)$ denotes the probability to move from the point $x \in \Omega$ to the set $B \subseteq \Omega$ in one step of the Markov chain.

Thus, *metastable conformations* are sets $A \subseteq \Omega$ with

$$\kappa(A, A) \approx 1. \tag{12}$$

By definition, the term "metastable conformation" refers to a subset A of the state space Ω within the molecule can move around freely and from which it will exit only with low probability. Typically, in a metastable conformation the overall structure of a biomolecule is well preserved whereas other parts (e.g., bonds and bond angles) can oscillate freely.

Metastable conformations are dependent not only on the potential $\mathcal{V}$ of the molecular system, but also on the temperature β of the canonical density and the underlying dynamics. In contrast to that, we use the term "metastable set" to refer to non-physical dynamics or to emphasize the Monte Carlo viewpoint. For example, Hybrid Monte Carlo (HMC) can be regarded

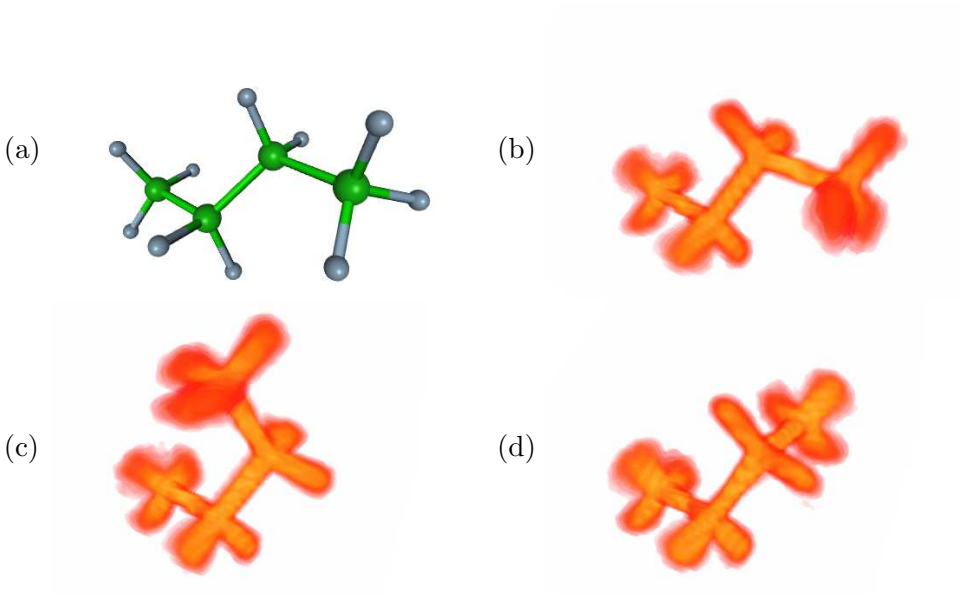

Figure 7: (a) A n-butane configuration with an explicit representation of H-atoms as in MMFF (torsion angle $\omega \approx 0°$). (b)–(d) Foggy representation of three metastable conformations (by alignment of three of the C-atoms) induced by the torsion angle potential $\mathcal{V}_{\text{tor}}$ (cf. Fig. 4). The metastable sets (b), (c), and (d) correspond to the −gauche-, trans-, and +gauche-conformations of n-butane, respectively.

as a sophisticated Metropolis algorithm, where metastability is interpreted due to non-physical sampling in the usual Monte Carlo context; yet it is also possible to use the dynamical information of a HMC sample to directly identify metastable conformations [113].

We already outlined in the introduction that the most probable three-dimensional structures of a molecule determine the functionality of a biomolecule in an organic environment. Although the probability to be within a metastable conformation is not part of its characterization, metastable conformations of high probability represent the biggest portion of the canonical distribution and are therefore the most important ones.

Figure 7 illustrates n-butane in three different metastable conformations. Conformations are separated by its torsion angle, which in fact is the only internal degree of freedom of structural importance for this simple molecule. In Fig. 1 we already illustrated two conformations of a much larger biomolecule, the inhibitor VX-478 of the enzyme HIV-protease.

Metastable sets are understood in terms of dynamical fluctuations within the canonical ensemble; a purely geometric approach to identify metastable sets by clustering the sample points according to geometric similarity is therefore insufficient to describe the dynamical situation. In Sect. 3.2.3 we present an identification strategy which is based on a *dynamical cluster algorithm*.