

Biomolecular Sampling: Algorithms, Test Molecules, and Metrics

Scott S. Hampton, Paul Brenner, Aaron Wenger, Santanu Chatterjee, and
Jesús A. Izaguirre

Department of Computer Science and Engineering, University of Notre Dame,
Notre Dame, Indiana 46556, USA

Summary. We compare the effectiveness of different simulation sampling techniques by illustrating their application to united atom butane, alanine dipeptide, and a small solvated protein, BPTI. We introduce an optimization of the Shadow Hybrid Monte Carlo (SHMC) algorithm, a rigorous method that removes the bias of molecular dynamics (MD). We also evaluate the ability of constant-temperature MD methods (based on Langevin and Nosé-Hoover dynamics) to achieve uniform thermal equilibrium. Our results show the the superiority of Langevin dynamics over Nosé-Hoover dynamics to achieve thermal equilibrium. They also illustrate the inherent limitation of sampling protocols that rely on sampling the microcanonical and canonical ensemble at only one temperature, and the importance of generalized ensemble approaches such as replica exchange that sample using different temperatures. Finally, we show how SHMC is able to remove bias and scale with timestep and system size. Presented, herein, are a set of sampling algorithms, test molecules,

2 S. S. Hampton, P. Brenner, A. Wenger, S. Chatterjee, and J. A. Izaguirre
and metrics. We make the sampling methods discussed available via the open source
and freely distributed PROTOMOL molecular simulation framework.

Key words: Monte Carlo, Molecular Dynamics, Replica Exchange, Shadow
Hybrid Monte Carlo, Langevin Impulse, Nosé-Hoover dynamics.

1.1 Introduction

For a system of N atoms, define configuration space as a dN dimensional vector of positions, where d is the number of degrees of freedom per atom. Similarly, phase space is defined by both configuration and momenta and is $2dN$ dimensional. The problem of sampling the configuration space of biomolecules such as proteins is an inherently difficult problem. The difficulties are due to the high dimensionality of phase space ($6N$), the ruggedness of the energy functions, the presence of multiple time scales, high energy barriers, and the introduction of systematic error (bias) by sampling methods.

We compare the effectiveness of different simulation sampling techniques by illustrating their application to two simple molecules, united atom butane and alanine dipeptide, and in a few cases to a small solvated protein, BPTI. The methods we test are based on molecular dynamics (MD) or Monte Carlo (MC) sampling techniques. We introduce an optimization of our Shadow Hybrid Monte Carlo algorithm, a rigorous method that removes the bias of MD. We also evaluate the ability of constant-temperature MD methods (based

on Langevin and Nosé-Hoover dynamics) to achieve uniform thermal equilibrium. Our results show the inherent limitation of sampling protocols that rely on sampling the microcanonical and canonical ensemble at only one temperature, and the importance of generalized ensemble approaches such as the replica exchange method (REM). REM is based on making several simulations at different temperatures, and combining the results to get better sampling at the target temperature. In REM, a random walk is performed on temperature space, which in turn induces a walk on configuration space.

At the same time, we present a set of sampling algorithms, test molecules, and metrics. We make the sampling methods tested available via the open source and freely distributed PROTOMOL [26,27] molecular simulation framework.

The problem of sampling can be thought of as estimating expectation values for a function $A(\Gamma)$ with respect to a probability density function (p.d.f.) $\rho(\Gamma)$, where $\Gamma = \begin{pmatrix} X \\ P \end{pmatrix}$, and X and P are the vectors of collective positions and momenta. Then,

$$\langle A(\Gamma) \rangle_\rho = \int A(\Gamma) \rho(\Gamma) d\Gamma. \quad (1.1)$$

Examples of observables A are potential energy, pressure, free energy, and distribution of solvent molecules in vacancies [11,21,34]. For the sampling of configuration space of biological molecules, the space spanned by the positions X , ρ typically corresponds to a constant system size N , constant temperature T , and constant volume V ensemble (canonical ensemble),

$$\rho_{\text{NVT}}(\Gamma) \propto \frac{\exp(-\mathcal{H}(\Gamma)/(k_{\text{B}}T))}{\int \exp(-\mathcal{H}(\Gamma)/(k_{\text{B}}T))d\Gamma}, \quad (1.2)$$

such that \mathcal{H} is the Hamiltonian or total energy of the system, k_{B} is Boltzmann’s constant, and T is temperature. Integration is performed over all of phase space. Another common ensemble is the microcanonical, constant N , V , and E , where E is the total energy of the system. The p.d.f. then becomes [1]:

$$\rho_{\text{NVE}}(\Gamma) \propto \frac{\delta(\mathcal{H}(\Gamma) - E)}{\int \delta(\mathcal{H}(\Gamma) - E)d\Gamma}, \quad (1.3)$$

where $\delta()$ chooses only those energy states equal to E .

Biomolecular sampling is a general term which should be qualified based on the sampling quantity of interest. We subgroup the quantities into geometric conformations and thermodynamic properties (free energy, entropy, chemical potential, kinetic and potential energy, temperature, and pressure). We consider both conformational sampling and computation of observables in our tests. In Section 1.2, we refer the reader to multiple papers which further describe the fundamental hurdles of sampling.

We first present the sampling algorithms tested. Second, we introduce the test molecules, simulation parameters, and applicable sampling metrics. Third, we present the results from applying the sampling algorithms to the test molecules. We conclude with a discussion of results and overview of promising methods.

1.2 Related Work

There are several reviews of classical biomolecular sampling methods and newer methods [2, 5, 9, 22, 29, 39]. Leach [22] provides a good review of conformational analysis techniques but simulation details are unavailable. Berne and Straub [5] provide a fairly thorough discussion of sampling methods and introduce several sampling metrics. Nardi and Wade [29] reviewed the application of several sampling techniques to simple biomolecules, showing extensive results. These authors do not include a discussion of SHMC or REM.

Our presentation of molecular simulation parameters and test metrics is based on the work by Barth, Leimkuhler, and Reich [2]. We have extended the metrics to consider the problem of sampling and quantification of bias in more depth.

1.3 Sampling Algorithms

We study simulation-based methods for conformational sampling. The key technique for sampling of configuration space of a biomolecule is constant-temperature MD, which is widely applicable and flexible. We compare Nosé dynamics to Langevin dynamics in their ability to achieve thermal equilibrium (needed to do sampling). MD methods have a timestep dependent bias, which we quantify for sample metrics in Section 1.5. MC methods, in contrast, are rigorous sampling methods. A rigorous method removes the effect of systematic errors such as discretization error, although the effect of rounding errors

is neglected. We study two MC methods: the Hybrid Monte Carlo (HMC) method, that uses MD as a global MC move, and SHMC, which improves the scaling of HMC with system size and MD timestep. Finally, we consider replica exchange, which can be used with either MD or MC methods, and which greatly improves their sampling ability.

1.3.1 Molecular Dynamics

Molecular dynamics solves Newton’s equations of motion to sample from the microcanonical ensemble. In the thermodynamic limit of infinite system size, the microcanonical and canonical ensembles are equivalent for ergodic systems. In practice, it is more convenient to sample from the canonical ensemble.

Propagators for MD, such as Leapfrog, can be written as

$$\Gamma' = \Psi(\Gamma), \quad (1.4)$$

where Ψ is the MD integrator, and $\Gamma = \begin{pmatrix} X \\ P \end{pmatrix}$, the vector of collective positions and momenta. For example, the velocity-form of the Verlet method can be expressed as:

$$X^{n+1} = X^n + \Delta t M^{-1} P^n - \frac{1}{2} \Delta t^2 M^{-1} F(X^n), \quad (1.5)$$

$$P^{n+1} = P^n - \frac{1}{2} \Delta t (F(X^n) + F(X^{n+1})), \quad (1.6)$$

where M is the matrix of atomic masses, and $F(X) = -\nabla U(X)$ is the force, assumed to be the negative gradient of a potential energy function.

The temperature of the system is related to the time-average of the kinetic energy of the system. A popular way of controlling temperature can be achieved by coupling the system to an external heat bath, such as in Nosé-Hoover dynamics [14, 30]. Nosé proposed a Hamiltonian of the form:

$$\mathcal{H}^{\text{Nosé}} = \mathcal{H}(x, p/s) + \frac{p_s^2}{2Q} + gkT \ln s, \quad (1.7)$$

where Q is the Nosé mass, s is the thermostating variable, p_s is its conjugate momentum and x and p are positions and momenta respectively.

Langevin dynamics is also a way of generating the canonical ensemble. Here, the equations of motion are modified to include a damping factor γ in the velocity and a random forcing term from a Gaussian distribution with zero mean and variance $\sqrt{2\gamma k_B T} M^{-1/2}$, cf. [7].

1.3.2 Hybrid Monte Carlo

HMC combines an MD trajectory with an MC rejection step. It takes advantage of the long steps in phase space that can be achieved through MD while eliminating the inaccuracies due to a finite timestep and other numerical artifacts through the MC step. One step of HMC is shown in Algorithm 1, where $\beta = (k_B T)^{-1}$. The output of Algorithm 1 is either a new $\Gamma' = \Psi(\Gamma)$, or simply the original input point in phase space, $\Gamma' = \Gamma$.

1.3.3 Shadow Hybrid Monte Carlo

SHMC is a new method recently published by Izaguirre and Hampton [17] for sampling the phase space of large molecules, particularly biological molecules.

Algorithm 1 Hybrid Monte Carlo

Hybrid Monte Carlo:

1. Given $X := x$
 2. Generate P with Gaussian p.d.f.
 3. $\Gamma := \begin{pmatrix} X \\ P \end{pmatrix}$
 4. $\Gamma' := \Psi(\Gamma)$ (run NVE MD for a small number of steps
using, for example, the Verlet MD integrator Ψ
given by Eqs. (1.5) and (1.6))
 5. Accept Γ' with probability
$$\min \left\{ 1, \frac{\exp(-\beta(\mathcal{H}(\Gamma')))}{\exp(-\beta(\mathcal{H}(\Gamma)))} \right\}$$
 6. If rejected, $\Gamma' := \Gamma$
-

It improves sampling of HMC by allowing larger timesteps and system sizes in the MD step. The acceptance rate of HMC decreases exponentially with increasing system size N or timestep δt . For a proof see [17, p. 588] and [8, 13, 28]. This poor performance is due to the effect of discretization errors introduced by the numerical integrator on the acceptance function.

SHMC achieves an asymptotic $O(N^{1/4})$ speedup over HMC by sampling from all of phase space using high order approximations to a shadow or modified Hamiltonian exactly integrated by a symplectic MD integrator. SHMC satisfies microscopic reversibility and is a rigorous sampling method. SHMC requires extra storage, modest computational overhead, and a reweighting step to obtain averages from the canonical ensemble. This method has been validated by numerical experiments that compute observables for different

molecules, ranging from a small n -alkane butane with 4 united atoms to a larger solvated protein with 14,281 atoms. In these experiments, SHMC achieves an order of magnitude speedup over HMC in sampling efficiency for medium sized proteins.

Description of SHMC:

SHMC samples the target density $\tilde{\rho}(\mathbf{x}, \mathbf{p})$, where

$$\tilde{\rho}(\mathbf{x}, \mathbf{p}) \propto \exp\left(-\beta\tilde{\mathcal{H}}(\mathbf{x}, \mathbf{p})\right), \quad (1.8)$$

$$\tilde{\mathcal{H}}(\mathbf{x}, \mathbf{p}) = \max\{\mathcal{H}(\mathbf{x}, \mathbf{p}), \mathcal{H}_{[2k]}(\mathbf{x}, \mathbf{p}) - c\}. \quad (1.9)$$

Here, $\mathcal{H}_{[2k]}(\mathbf{x}, \mathbf{p})$ is the more accurate shadow Hamiltonian, defined by Skeel and Hardy (2001), and c is an arbitrary constant that limits the amount by which $\mathcal{H}_{[2k]}$ is allowed to depart from $\mathcal{H}(\mathbf{x}, \mathbf{p})$. In Figure 1.1, we plot the 4th and 8th order shadow Hamiltonians, as well as the true Hamiltonian for a typical system. As can be seen in the figure, the shadow Hamiltonian has much smaller fluctuations.

Algorithm 2 lists one iteration of SHMC. The first step is to generate a set of momenta, P , often chosen via a Gaussian distribution. P is accepted based on a Metropolis criterion step proportional to the difference of the total and shadow energies. This step is repeated until a set of momenta are accepted. Next, the system is integrated using MD and accepted with probability proportional to Eq. (1.8).

After all iterations of SHMC have been completed, in order to calculate unbiased values, the observables are reweighted using the following formula:

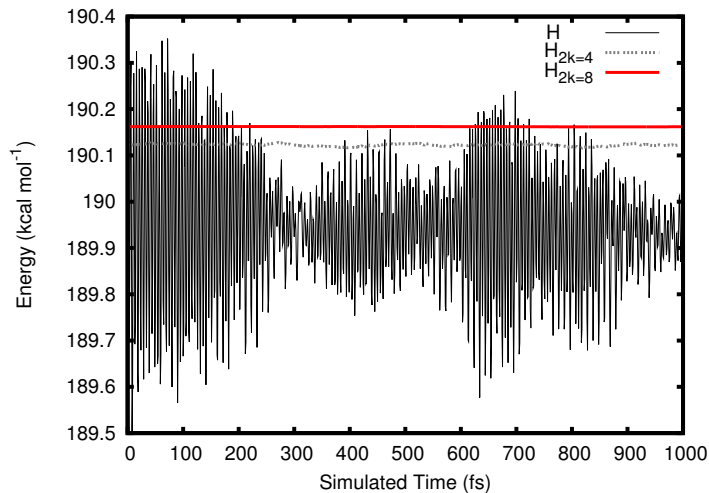


Fig. 1.1. Energy and Shadow energy (4^{th} and 8^{th} order) for a Decalanine molecule.

$$\langle A \rangle_{\rho_{\text{NVT}}} = \frac{\sum_{i=1}^m w_i A_i}{\sum_{i=1}^m w_i}, \quad (1.10)$$

where

$$w_i = \frac{\exp(-\beta\mathcal{H}(\Gamma_i))}{\exp(-\beta\tilde{\mathcal{H}}(\Gamma_i))}. \quad (1.11)$$

There are two criteria in the optimization of c . The first is to minimize the difference in the shadow and total energies so that the reweighted observables of $\mathcal{H}_{[2k]}$ have a reasonable variance. If the difference is left unbounded, the mean of the observable will be consistent with the correct ensemble average, but the standard deviation grows with system size and timestep. In particular, without introducing c in SHMC, it is not possible to define the ensemble from which the method samples and thus prove microscopic reversibility, see [17]. The second criterion is to maximize the computational efficiency of SHMC.

Algorithm 2 Shadow Hybrid Monte Carlo .

1. **MC Step:** Given $X := x$, generate P with p.d.f. $\tilde{\rho}(X, p)$, using the acceptance-rejection method:
 - a) Generate P having Gaussian p.d.f.
 - b) Accept with probability $\min \left\{ 1, \frac{\exp(-\beta(\mathcal{H}_{[2k]}(X, P) - c))}{\exp(-\beta\mathcal{H}(X, P))} \right\}$
 - c) Repeat (1a) - (1b) until P is accepted.
 2. **MD Step:** Given $\Gamma := \begin{pmatrix} X \\ P \end{pmatrix}$, and $R := \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}$,
 - a) $\Gamma' := R\Psi(\Gamma)$, (Run NVE MD for a small number of steps, where the MD integrator Ψ nearly conserves $\mathcal{H}_{[2k]}$)
 - b) Accept Γ' with probability $\min \left\{ 1, \frac{\tilde{\rho}(\Gamma')}{\tilde{\rho}(\Gamma)} \right\}$
 - c) If rejected, $\Gamma' = \Gamma$.
-

Figures 1.2a and 1.2b demonstrate how the choice of c affects the runtime of a simulation. First, we plot the number of momenta rejections per SHMC step, followed by the number of position rejections per SHMC step. It is important to note that rejecting new positions is much more costly than rejecting the momenta as the forces must be calculated for an MD trajectory to produce new positions but not for new momenta. For large c , our algorithm reduces to HMC. Thus, the HMC-regime has many rejections of positions since c increases with system size and timestep, whereas large negative c SHMC has a high rejection of momenta instead.

In general, we are not able to predict c *a priori*, but we have devised a heuristic that can choose an optimal c at runtime. In order to automatically

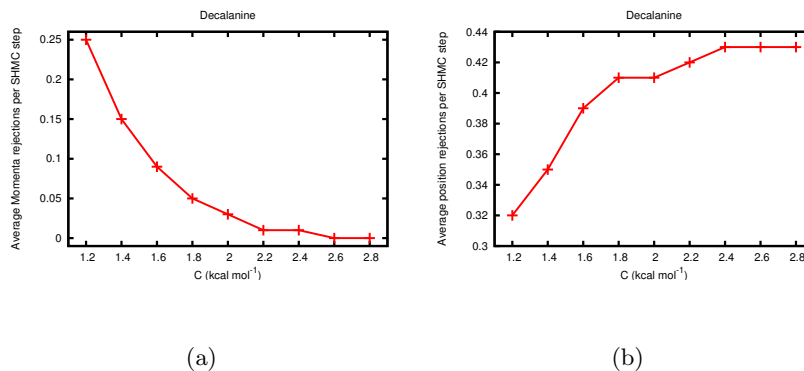


Fig. 1.2. (a) Average number of rejected momenta per step; and (b) Average number of position rejections per step as a function of the choice of c for a 66-atom decalanine. Simulations run for 100 ps with 25 MD steps per SHMC step.

choose c , we first run initialization steps (20) of our algorithm using a default value for c . During each warm-up step, we record the difference between the total and shadow energies two times: at the beginning of the MD run, immediately after the momenta are accepted, and at the end of the MD run. In a well equilibrated system, the energies at these two points should be fluctuating about some average value. This gives us forty samples and the mean value of this difference is chosen as an initial guess for c , denoted c_0 .

Up to this point, this heuristic tends to choose large c . Therefore, we do additional relaxation steps (20) where the value of c is allowed to change slightly. Even when the initial value of c_0 is large, we noticed that big changes during relaxation are not effective. For this reason, each relaxation step is allowed to change c by no more than $\pm \log_{10} c_0$. We also want c to converge

during the relaxation instead of jumping between predetermined values. In order to accomplish this, we multiply any suggested change by a decreasing factor, $1 - \frac{i}{N_R}$, where N_R is the number of relaxation steps and i is the current step, $0 \leq i < N_R$. During this phase, we may or may not change c depending on a predetermined criteria. In [17], we suggest that rejecting a few momenta during each step of SHMC is beneficial to the algorithm. We use an additional rejection factor to influence c to give us the desired number of momenta rejections. Based on our tests, we change c at each step of the relaxation according to Table 1.1. Thus, if after the third step of the relaxation

Table 1.1. Change in c for next relaxation step based on the number of rejected momenta for the current step.

Rejected momenta	0	1-5	6-10	11-15	16-20	21-25	> 25
Rejection factor	-1	0	0.5	1	1.5	2	3

there were 17 momenta rejections, the new value of c would be:

$$c_4 = c_3 + \left(1 - \frac{3}{20}\right) \times 1.5 \times \log_{10} c_0.$$

In this way, we lower the value of c if no rejections occur and raise the value if we reject too many. We have found that this method will, on average, give us the desired 5-10 momenta rejections per timestep for most values of the timestep Δt . As Δt grows so does the discretization error and the number of rejections of the momenta will also grow.

1.3.4 Replica Exchange Method

The method of replica exchange or “exchange Monte Carlo method”, as introduced by Hukushima and Nemoto [15], uses a number of non-interacting MD or MC simulations running in parallel each at a different temperature. Replica coordinates are swapped between temperature neighbors according to a probability function, with an intent to enable a replica at low temperature to surpass barriers between its local energy minima. The method has found favor because the probability weighting factors are known *a priori*, unlike other generalized-ensemble algorithms which require detailed determination. The method is shown in Algorithm 3 following Okamoto [31].

Algorithm 3 Replica Exchange

1. Each replica in the canonical ensemble of fixed temperature is simulated simultaneously and independently for a certain number of MD or MC steps
2. A pair of replicas at neighboring temperatures, say $\gamma_m^{[i]}$ and $\gamma_{m+1}^{[j]}$, are exchanged with the probability

$$w(\gamma_m^{[i]}|\gamma_{m+1}^{[j]}) = \min(1, \exp(-\beta_m - \beta_{m+1})(U(x^{[j]}) - U(x^{[i]}))), \quad (1.12)$$

where $U(x)$ is the potential energy.

Quantities of interest such as conformational distributions and thermodynamic properties can be derived from the individual replica output data sets through utilization of single histogram or multiple-histogram reweighting techniques such as the weighted histogram analysis method WHAM [20].

1.4 Test Systems, Methods, and Metrics

1.4.1 United Atom Butane

The 4-atom, united atom model butane, recently summarized by Barth *et al.* [2], is useful for elementary evaluation of sampling algorithms. Analytical calculations, relative to the dihedral energy, are easily generated for verification and validation purposes. Our model uses a two component term for the dihedral energy equation which closely approximates the original model [33]. Baseline analytical values are reported for subsequent comparison with the sampling algorithms.

Molecule Parameters:

Parameters for the united atom butane which we utilized in our testing are listed in Table 1.2. Note the absence of Coulombic forces. The non-bonded Lennard-Jones forces were evaluated using all-pairs with no cutoffs.

Metric - Average Dihedral Energy:

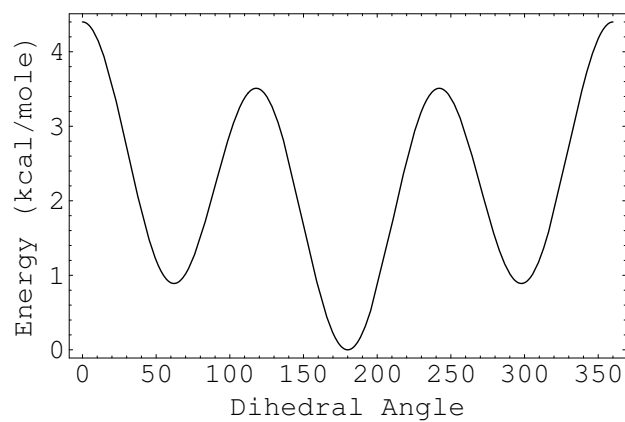
The average dihedral energy for united-atom butane can be computed analytically according to the following equation reported by Fischer *et al.* [10]:

$$\langle U_D \rangle = \frac{\int U_D(\omega) \exp[-\beta U_D(\omega)] d\omega}{\int \exp[-\beta U_D(\omega)] d\omega}, \quad (1.13)$$

where U_D is the potential energy for the dihedral angle and ω is over the interval $[-\pi, \pi]$. The value at 300 K is given in Table 1.3.

Table 1.2. Parameters for United Atom Butane (Based on CHARMM 19 [24, 25])

Bonds	
CH2-CH2	$\kappa_b = 225.0 \text{ kcal/mol}$, $b_{eq} = 1.52 \text{ \AA}$
CH2-CH3	$\kappa_b = 225.0 \text{ kcal/mol}$, $b_{eq} = 1.54 \text{ \AA}$
Angle	
CH3-CH2-CH2	$\kappa_\theta = 45.0 \text{ kcal/mol}$, $\theta_{eq} = 110^\circ$
Dihedral	
CH3-CH2-CH2-CH3	$\kappa_\phi = 1.6 \text{ kcal/mol}$, $\phi_{eq} = 0^\circ$, $n_\phi = 3$
	$\kappa_\phi = 0.6 \text{ kcal/mol}$, $\phi_{eq} = 0^\circ$, $n_\phi = 1$
Lennard-Jones	
CH2	$\epsilon = 0.1142 \text{ kcal/mol}$, $\sigma = 2.235 \text{ \AA}$
CH3	$\epsilon = 0.1811 \text{ kcal/mol}$, $\sigma = 2.165 \text{ \AA}$

**Fig. 1.3.** United atom butane energy for dihedral angle.

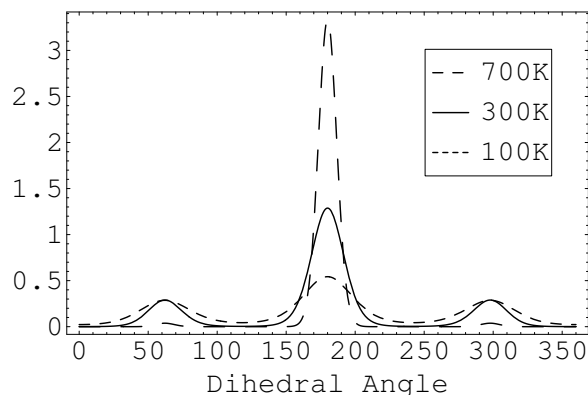


Fig. 1.4. United atom butane dihedral distribution.

Metric - Dihedral Distribution:

The dihedral energy of butane is shown in Figure 1.3. The probability density of dihedral values for united atom butane at 300 K is plotted in Figure 1.4 as calculated by

$$\frac{\exp[-\beta U_D(\omega)]d\omega}{\int \exp[-\beta U_D(\omega)]d\omega}. \quad (1.14)$$

1.4.2 Alanine Dipeptide

Alanine dipeptide is a frequently used test problem in MD simulations. It has only 23 atoms and it exhibits conformational flexibility in the two-dimensional ϕ - ψ plane. Also, it has multiple *CO* and *NH* units for *H*-bonding and pendant side chain (methyl group) off of the main backbone (Figure 1.5).

The change in conformation seen through the dynamics of the backbone dihedral and its size make it a suitable molecule for studying conformational sampling. In alanine dipeptide, using CHARMM22 force field, the lowest free

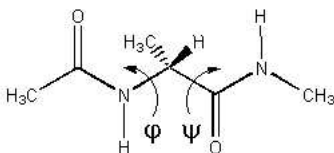


Fig. 1.5. Diagram of Alanine Dipeptide (source: <http://www.chem.umass.edu/nermmw/alanine.html>)

energy structure is the extended β conformation, $(\phi, \psi) = (-81, 81)$, while in solution, the extended β , $(\phi, \psi) = (-81, 153)$ and right-handed α -helical, $(\phi, \psi) = (-81, 63)$, conformations are nearly isoenergetic. In solution, a secondary minimum at $(\phi, \psi) = (63, -81)$ occurs at approximately 2.3 kcal/mol at global free-energy minimum [32]. The ψ dihedral angle corresponds to $N-C_{\alpha}-C-N$ and the ϕ to $C-N-C_{\alpha}-C$.

1.4.3 Bovine Pancreatic Trypsin Inhibitor (BPTI)

BPTI, protein data bank entry 6PTI [4], is commonly studied because of its small size. We use a solvated version with 73 additional water molecules for a total of 1,101 atoms. We also solvate BPTI for a larger system size of 14,281 atoms. The systems were prepared for simulation using NAMD 2.5 [18]. First, we ran minimization for 1,000 steps. Next, 20,000 steps of constant temperature, constant pressure (NPT) MD were run to equilibrate the system at a temperature of 300 K and pressure of 1 atm.

Metric - Conformation Strings

In order to identify unique conformations, we use a “dihedral string” method similar to that reported in [17] and [19]. A set of key dihedrals from the molecule are observed and their values resolved into an alphanumeric tag representing which energetic well the dihedral currently resides in. The energetic maximas for each dihedral in the input evaluation set are determined during initialization using a modified Brent’s method and dihedral parameter knowledge from the topology (non zero phase shifts are normalized). By concatenating the tags together, we form a conformation string for each timestep. Comparing strings, we can classify and analyze the number of unique conformations.

1.5 Simulation Results

1.5.1 United Atom Butane

In order to compare and contrast the ability of conformational sampling algorithms to surpass energy barriers we ran two simulation sets. The first simulation set included the HMC and SHMC algorithms. We also ran the same test set at longer times steps (4 and 8 fs). No significant differentiation from the 1 fs data was observed as one is able to take longer steps with united atom butane than would be possible for larger all-atom proteins. Simulations were run using 10 MD steps per MC step, periodic boundary conditions, the Leapfrog

20 S. S. Hampton, P. Brenner, A. Wenger, S. Chatterjee, and J. A. Izaguirre
integrator, all bonded forces, and full Lennard-Jones nonbonded forces. Performance at three temperatures (100, 300, and 700 K) was evaluated over 1 ns of simulation time. Average dihedral energies and the dihedral distributions were calculated. The performance at 300 K can be found in Table 1.3. For this simple molecule, HMC and SHMC performed similarly so SHMC data is omitted. The results relative to BPTI more appropriately contrast SHMC performance over HMC. We observe that HMC is unable to surpass the energy barrier sufficiently to sample the correct distribution; it reports an expected average dihedral energy approximately half of the analytical expectation.

For the second simulation set, we performed three 1 ns tests using NVE Leapfrog, NVT Langevin Impulse, and a replica exchange with NVT Langevin Impulse. All tests utilized a 1 fs timestep, periodic boundary conditions, and the same force field used in the HMC simulations. Langevin simulations were run with a γ value of 5000 ps^{-1} . The replica exchange simulation consisted of eight replicas at temperatures: 200 K, 239 K, 286 K, 342 K, 409 K, 489 K, 585 K, and 700 K [37]. Each replica was simulated using the Langevin impulse NVT method for 125 ps with a timestep of 1 fs. A replica exchange was attempted every 400 fs. It should be noted that all three tests represent the same amount of simulation time (a total of 1 ns), but the REM approach offers the additional benefit of fully parallel execution.

The results of the REM were analyzed using the “temperature” variant of the weighted histogram analysis method [12]. As is evident from Figures 1.6, 1.7, and 1.8, the REM simulation achieved wider and more accurate sampling

of the dihedral space than the standard MD simulations, suggesting that REM is superior for investigating the conformation space. In turn, NVT Langevin Impulse was far superior to NVE Leapfrog. We include these results in Table 1.3.

Table 1.3. Analytical, NVE Leapfrog, Hybrid Monte Carlo (HMC), Langevin Impulse, and Replica Exchange using Langevin Impulse (REM) values for united atom butane at 300 K.

	$\langle U_D \rangle$ in kcal/mol	gauche- in %	trans in %	gauche+ in %
Analytical	0.628	16.0	68.0	16.0
Leapfrog	0.096	0.0	100.0	0.0
HMC	0.335	0.2	99.6	0.2
Langevin	0.71	5.8	55.3	38.9
REM	0.619	18.4	66.5	15.1

1.5.2 Alanine Dipeptide

A brief analysis of two different MD implementations of the NVT ensemble, the Nosé-Hoover method using single thermostat and the Langevin impulse method, was performed to determine which achieved thermal equilibration more quickly. Langevin Impulse is exact for constant forces and has been shown to be a good discretization of the Langevin equation [16, 35]. Thermal equilibrium is important in the context of sampling because one typically

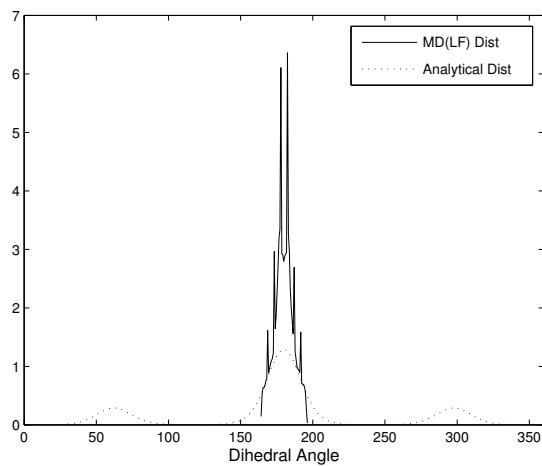


Fig. 1.6. United atom butane dihedral distributions: Leapfrog NVE MD (1 ns). The dashed lines show the correct analytical distributions

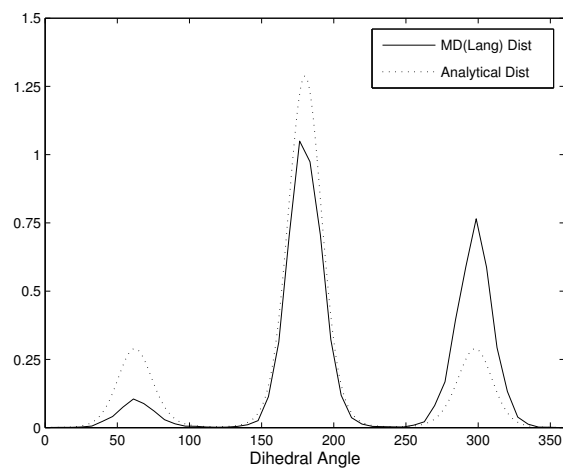


Fig. 1.7. United atom butane dihedral distributions: Langevin Impulse NVT MD (1 ns). The dashed lines show the correct analytical distributions

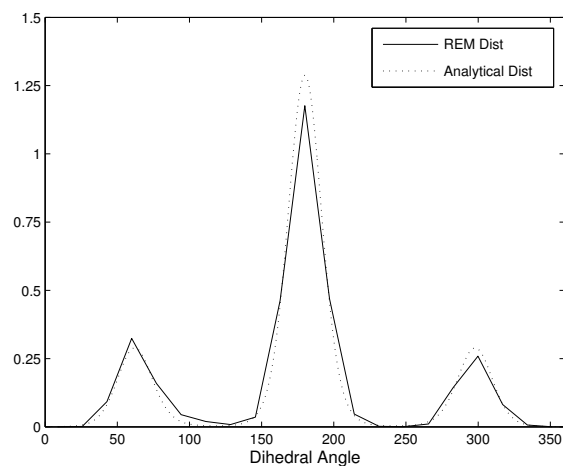


Fig. 1.8. United atom butane dihedral distributions: Replica Exchange Langevin NVT (1 ns total simulation time for 8 replicas - each replica 125 ps only). The dashed lines show the correct analytical distributions

starts sampling only when the system has achieved equilibrium. In addition, some methods such as umbrella sampling [6, 38] use biasing potentials to explore different conformations, and quick equilibration is necessary for efficient sampling.

We equilibrated alanine dipeptide for 500 ps using a 1 fs timestep in the NPT ensemble. Additionally, we examined the temperature of subsystems within the larger system. We expect an equipartition of the kinetic energy. Thus, we use the following criteria to judge the effectiveness of the NVT ensembles: 1) Does the ensemble achieve the desired system temperature? In other words, is the average system temperature equal to the target and is

the standard deviation small? 2) Does the ensemble maintain a consistent temperature among system components?

An alanine dipeptide molecule solvated with 607 water molecules was used in the simulations. For each method, a 10 ps simulation was performed with a timestep of 0.1 fs. All bonded forces were utilized, including bond, bond angle, dihedral, and improper. For non-bonded forces, both Lennard-Jones and Coulombic forces were calculated using an all-pairs algorithm. Additionally, an artificial harmonic weighting potential was added to the system to force the ϕ dihedral to a desired value. The weighting potential was a function of the dihedral angle: $U(\phi) = k(\phi_{cur} - \phi_{goal})^2$ where ϕ_{cur} is the current value of the dihedral, ϕ_{goal} is the desired value, and k is a force constant. We want to simulate using MD techniques around the point in phase space represented by the reaction coordinate ϕ_{goal} . This is similar to umbrella sampling where a weighting function is used to bias a reaction coordinate to a particular configuration. The weighting potential was utilized to create biasing in the system to adequately test the NVT methods. We used a γ value of 5000 ps^{-1} for the simulation using the Langevin Impulse method. The ϕ_{goal} has been set to 60° . The k was set to 100 kcal/mol .

As is clear from Figures 1.9a and 1.9b and Table 1.4, the Langevin impulse method was superior in our test simulations. It maintained all subsystems near the target temperature throughout the simulation. Conversely, the Nosé-Hoover method, despite maintaining the total system temperature near the target value, produced a significant temperature differential between the

alanine dipeptide molecule and the water. Even after 10 ps, the Nosé-Hoover method did not achieve the desired subsystem temperatures. These results are in agreement with the contribution of Barth, Leimkuhler, and Sweet in this volume.

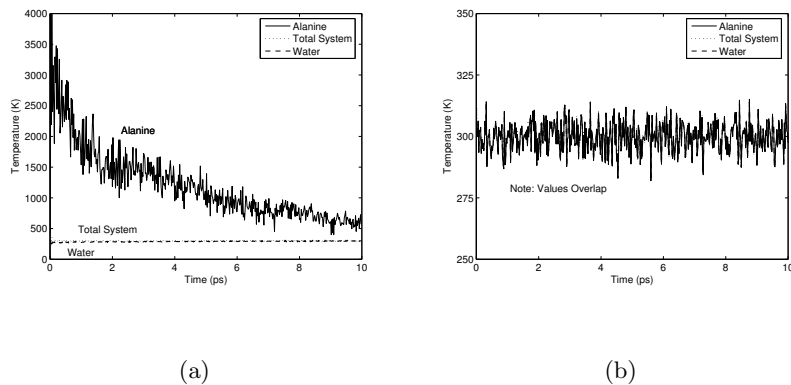


Fig. 1.9. Thermal equilibration for water and solute of solvated alanine dipeptide (a) Nosé-Hoover NVT at 300 K; (b) Langevin impulse NVT at 300 K

Table 1.4. Subsystem temperatures for NVT ensembles with a target temperature of 300 K

Subsystem	Nosé-Hoover NVT		Langevin impulse NVT	
	Average (K)	Std Dev	Average (K)	Std Dev
Alanine	1215.50	596.55	302.10	84.27
Water	288.88	9.63	300.05	5.79
Total System	300.11	7.38	300.23	5.78

1.5.3 BPTI

Figures 1.10 and 1.11 show the cost for HMC and SHMC of producing a different conformation (using the conformation string) as a function of the timestep for 1,101-atom and 14,281-atom solvated BPTI respectively. These charts were selected from our recently published SHMC work [17] to demonstrate how SHMC allows much longer timesteps than HMC. The 1,101 atom uses an MD trajectory of 42 fs for each HMC or SHMC step, whereas the 14,281 uses an MD trajectory of 15 fs for each HMC or SHMC step. The speedup in sampling efficiency, given by the ratio of computational cost of HMC to SHMC, ranges from 2 to 10 as the timestep is increased.

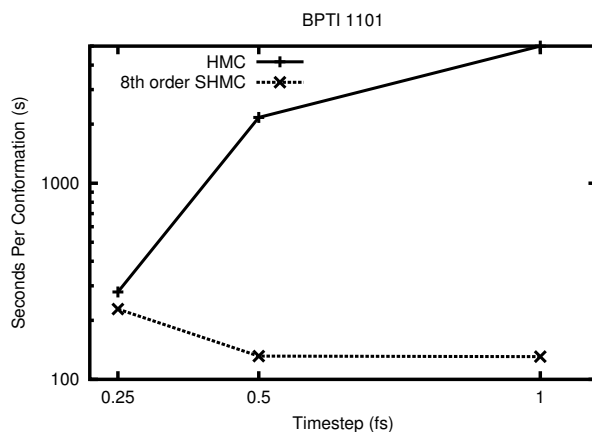


Fig. 1.10. Average computer time per discovered conformation for 1,101 atom BPTI.

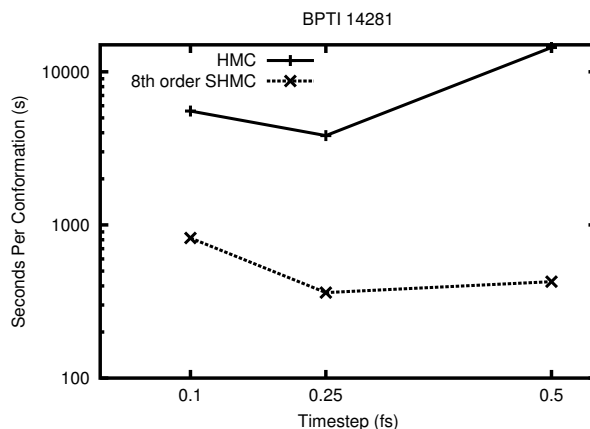


Fig. 1.11. Average computer time per discovered conformation for the 14,281 atom BPTI.

1.5.4 Bias

In Figure 1.12, we show the effect of bias on an MD simulation, as well as how MC methods remove the bias. We simulated BPTI with 1,101 atoms for 20 ps, using Leapfrog, HMC, and SHMC for 6 different timesteps. HMC and SHMC use 25 MD steps per MC step. Error bars are 1 standard deviation and are averaged over 3 runs. From the figure, it can be seen that as the timestep increases, there is an obvious upward drift in the potential energy that is greater than the statistical error. A least squares fit is plotted through the points. SHMC, on the other hand, is fluctuating about an average value that falls within the statistical error. Thus, the bias is removed. For the larger timesteps, SHMC is beginning to show more deviation due to a decrease in accepted moves. HMC does not show removal of bias because its poor scala-

bility does not allow it to accept enough moves beyond a timestep of 0.5 fs in this test.

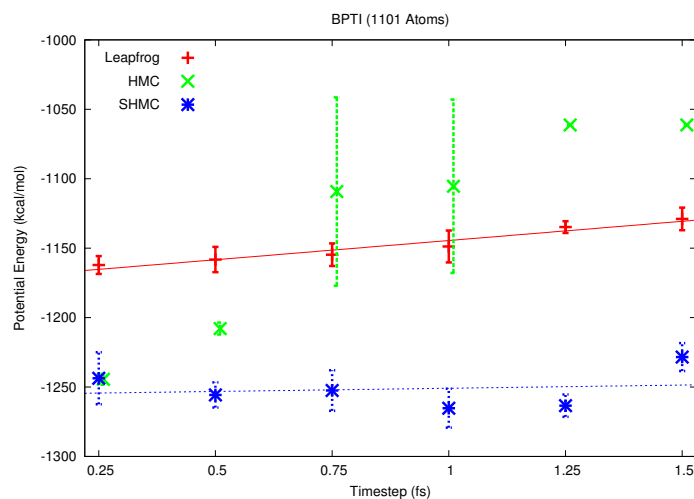


Fig. 1.12. Bias in the average potential energy of MD, HMC, and SHMC as a function of timestep

1.6 Discussion

From our results with butane, it is evident that REM could be considered the best method for conformational sampling of biomolecules. For optimal performance, one needs to choose the right distribution of the temperature. But, REM has the advantage that the weighting factor is known in advance. Also, REM is a suitable candidate for parallel implementation. Indeed, this method allows us to sample around all energy wells under low temperature.

REM can work with either MD or MC methods. Thus, if one needs to be exact, one can use HMC or SHMC for each replica simulation.

There are several different optimizations proposed based on general REM. One of them is multidimensional replica exchange method (MREM). In this method, a pair of replicas with different temperature or different parameters of the potential energy are interchanged [36]. This method provides greater flexibility over the general REM and it is based on the following observation: For the M noninteracting replicas, the Hamiltonian \mathcal{H} of the system could be different for different replicas depending on some parameter(s). One particularly interesting instance of MREM is replica-exchange umbrella-sampling.

SHMC is exact since it eliminates the bias associated with MD sampling of molecular systems. SHMC can use much higher timesteps than classical HMC, particularly as the system size increases. In practice, one needs to choose an extra parameter c . We have presented an efficient method for automatically determining this additional parameter and shown its effectiveness in computing potential energy of solvated BPTI. We show that SHMC is more than ten times more efficient than HMC for a medium size solvated BPTI with 14,281 atoms.

Nosé dynamics is a popular scheme to simulate molecular systems under constant temperature. But it is based on the ergodic hypotheses, which does not necessarily hold for small or stiff systems. In our experiment with alanine dipeptide, we have found a significant gap in subsystem temperature from the target value. Several extensions of the original Nosé-Hoover method exist

30 S. S. Hampton, P. Brenner, A. Wenger, S. Chatterjee, and J. A. Izaguirre
which address this issue. There are extensions which use generalized heat
baths to control the temperature of the subsystem. Auxiliary heat baths are
coupled into thermostating variables. But the design of auxiliary heat bath
could slow down the simulation significantly. One of the interesting extensions
along this line is due to Leimkuhler and Sweet. Their work is based on the
idea of adding multiple thermostats to a Hamiltonian which has been modified
by Nosé’s method, retaining sampling from the canonical ensemble. They
employ a regularizing term in the Hamiltonian to bound the integrals over the
thermostats. A detailed discussion on this topic and related methods can be
found in [23]. Recursive multiple thermostat (RMT) is a recent development
by which equilibration is obtained from a more complicated interaction of
thermostat variables with the physical variables. The work by Berth *et al.* [3]
have shown that the RMT is most promising of all methods based on Nosé-
Hoover chains for correct thermostating.

From our experiment with NVT integrators on solvated alanine dipeptide
it is evident that Langevin Impulse can be used to achieve thermal equi-
libration quickly and accurately compared to the Nosé-Hoover method. The
former method kept the temperature of the biomolecule and the solvent close
to the target temperature for the entire course of the MD run. On the other
hand, the Nosé-Hoover formulation is more flexible and can be used to derive
other equations of motion.

A complex but powerful solution to the problems associated with poten-
tial barriers is due to Rosso *et al.* [32]. The idea is to separate out a reactive

subspace using a coordinate transformation. Then an adiabatic separation is imposed over this subspace and from the adiabatic probability distribution free energy surface is computed. Roughness of the free-energy manifolds have been explored using a higher temperature for a certain subset of reaction coordinates. This technique, known as adiabatic free-Energy dynamics (AFED), can also be applied to efficiently sample over the entire conformational space. However, this method has a high overhead for larger molecules since the coordinate transformations have to be derived on-the-fly. This is an instance of a broader class of algorithms based on smoothing the potential energy surface. A user of sampling methods will generally need to have access to a whole range of tools to solve their problems. Our hope is that this paper serves as an introduction to some of the very useful and new methods.

References

1. M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Clarendon Press, Oxford, New York, 1987. Reprinted in paperback in 1989 with corrections.
2. E. Barth, B. Leimkuhler, and S. Reich. A test set for molecular dynamics algorithms. In T. Barth, M. Griebel, D. Keyes, R. Nieminen, D. Roose, and T. Schlick, editors, *Computational Methods for Macromolecules: Challenges and Applications*, volume 24 of *Lecture Notes in Computational Science and Engineering*, pages 73–103. Springer, 2002.
3. E. Barth, B. Leimkuhler, and C. Sweet. Approach to thermal equilibrium in biomolecular simulation. 2005. To appear in *Proc. of Algorithms for Macro-*

- 32 S. S. Hampton, P. Brenner, A. Wenger, S. Chatterjee, and J. A. Izaguirre
molecular Modeling IV. Lecture Notes in Computational Science and Engineering (LNCSE). Springer Verlag. New York and Berlin.
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*, pages 235–242, 2000.
 5. B. J. Berne and J. E. Straub. Novel methods of sampling phase space in the simulation of biological systems. *Curr. Opin. Struct. Biol.*, 7:181–189, 1997.
 6. D. L. Beveridge and F. M. DiCapua. Free energy via molecular simulation: Applications to chemical and biological systems. *Ann. Rev. Biophys. Biophys. Chem.*, 18:431–492, 1989.
 7. A. Brünger, C. B. Brooks, and M. Karplus. Stochastic boundary conditions for molecular dynamics simulations of ST2 water. *Chem. Phys. Lett.*, 105:495–500, 1982.
 8. M. Creutz. Global Monte Carlo algorithms for many-fermion systems. *Phys. Rev. D*, 38(4):1228–1238, 1988.
 9. R. Elber. Reaction path studies of biological molecules. In R. Elber, editor, *Recent Developments in Theoretical Studies of Proteins (Advanced Series in Physical Chemistry, Vol. 7)*, Singapore, 1996. World Scientific.
 10. A. Fischer, F. Cordes, and C. Schütte. Hybrid Monte Carlo with adaptive temperature in mixed-canonical ensemble: Efficient conformational analysis of RNA. *J. Comp. Chem.*, 19(15):1689–1697, 1998.
 11. D. Frenkel and B. Smit. *Understanding Molecular Simulation*. Academic Press, San Diego, 2nd edition, 2002.
 12. E. Gallicchio, M. Andrec, A. K. Felts, and R. M. Levy. Temperature weighted histogram analysis method, replica exchange, and transition. *J. Chem. Phys.*,

- 109:6722–6731, 2005.
13. S. Gupta, A. Irbäck, F. Karsch, and B. Petersson. The acceptance probability in the hybrid Monte-Carlo method. *Phys. Lett. B*, 242:437–443, 1990.
 14. W. G. Hoover. Canonical dynamics: Equilibrium phase-space distribution. *Phys. Rev. A*, 31(3):1695–1697, 1985.
 15. K. Hukushima and K. Nemoto. Exchange Monte Carlo method and application to spin glass simulations. *J. Phys. Soc. of Japan*, 65(6):1604–1608, 1996.
 16. J. A. Izaguirre, D. P. Catarello, J. M. Wozniak, and R. D. Skeel. Langevin stabilization of molecular dynamics. *J. Chem. Phys.*, 114(5):2090–2098, 2001.
 17. J. A. Izaguirre and S. S. Hampton. Shadow hybrid Monte Carlo: An efficient propagator in phase space of macromolecules. *J. Comput. Phys.*, 200(2):581–604, 2004.
 18. L. Kalé, R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan, and K. Schulten. NAMD2: Greater scalability for parallel molecular dynamics. *J. Comput. Phys.*, 151:283–312, 1999.
 19. P. D. Kirchhoff, M. B. Bass, B. A. Hanks, J. Briggs, A. Collet, and J. A. McCammon. Structural fluctuations of a cryptophane host: A molecular dynamics simulation. *J. Am. Chem. Soc.*, 118:3237–3246, 1996.
 20. S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg. The weighted histogram analysis method for free-energy calculations on biomolecules. i: The method. *J. Comput. Chem.*, 13(8):1011–1021, 1992.
 21. A. R. Leach. *Molecular Modelling: Principles and Applications*. Addison-Wesley, Reading, Massachusetts, July 1996.

- 34 S. S. Hampton, P. Brenner, A. Wenger, S. Chatterjee, and J. A. Izaguirre
22. A. R. Leach. *Molecular Modelling: Principles and Applications*. Prentice Hall, 2nd edition, Mar. 2001.
23. B. J. Leimkuhler and C. R. Sweet. The canonical ensemble via symplectic integrators using Nosé and Nosé-Poincaré chains. *J. Chem. Phys.*, 121(1):108–116, 2004.
24. A. D. MacKerell Jr., D. Bashford, M. Bellott, R. L. Dunbrack Jr., J. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, I. W. E. Reiher, B. Roux, M. Schlenkrich, J. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-hydrogen empirical potential for molecular modeling and dynamics studies of proteins using the CHARMM22 force field. *J. Phys. Chem. B*, 102:3586–3616, 1998.
25. A. D. MacKerell Jr., D. Bashford, M. Bellott, R. L. Dunbrack Jr., J. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, B. Roux, M. Schlenkrich, J. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. Self-consistent parameterization of biomolecules for molecular modeling and condensed phase simulations. *FASEB J.*, A143:6, 1992.
26. T. Matthey. *Framework Design, Parallelization and Force Computation in Molecular Dynamics*. PhD thesis, University of Bergen, Bergen, Norway, 2002.
27. T. Matthey, T. Cickovski, S. S. Hampton, A. Ko, Q. Ma, M. Nyerges, T. Raeder, T. Slabach, and J. A. Izaguirre. PROTOMOL: An object-oriented framework for prototyping novel algorithms for molecular dynamics. *ACM Trans. Math. Softw.*, 30(3):237–265, 2004.

28. B. Mehlig, D. W. Heermann, and B. M. Forrest. Hybrid Monte Carlo method for condensed-matter systems. *Phys. Rev. B*, 45(2):679–685, 1992.
29. F. Nardi and R. Wade. *Molecular Dynamics. From Classical to Quantum Methods*, chapter 21, pages 859–898. Elsevier Science B.V., first edition, 1999.
30. S. Nosé. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.*, 81(1):511–519, 1984.
31. Y. Okamoto. Generalized-ensemble algorithms: enhanced sampling techniques for Monte Carlo and molecular dynamics simulations. *J. Molecular Graphics and Modelling*, 22:425–439, 2004.
32. L. Rosso, J. B. Abrams, and M. E. Tuckerman. Mapping the backbone dihedral free-energy surfaces in small peptides in solution using adiabatic free-energy dynamics. *J. Phys. Chem. B*, 109:4162–4167, 2005.
33. J.-P. Ryckaert and A. Bellemans. Molecular dynamics of liquid alkanes. *Faraday Discussions*, 66:95–106, 1978.
34. T. Schlick. *Molecular Modeling and Simulation - An Interdisciplinary Guide*. Springer-Verlag, New York, NY, 2002.
35. R. D. Skeel and J. A. Izaguirre. An impulse integrator for Langevin dynamics. *Mol. Phys.*, 100(24):3885–3891, 2002.
36. Y. Sugita, A. Kitao, and Y. Okamoto. Multidimensional replica-exchange method for free-energy calculations. *J. Chem. Phys.*, 113(15):6042–6051, 2000.
37. Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314:141–151, 1999.
38. Torrie and Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.*, 23:187–199, 1977.

- 36 S. S. Hampton, P. Brenner, A. Wenger, S. Chatterjee, and J. A. Izaguirre
39. W. F. van Gunsteren, T. Huber, and A. E. Torda. Biomolecular modelling: Overview of types of methods to search and sample conformational space. volume 330, pages 253–268. AIP, 1995.