Advances in Chemical Physics, Volume 150

ARTICLE *in* ADVANCES IN CHEMICAL PHYSICS · JANUARY 2012

Impact Factor: 1.49 · DOI: 10.1002/9781118197714.ch6 · Source: PubMed

Xiongwu Wu
National Heart, Lung, and B…

EFFICIENT AND UNBIASED SAMPLING OF BIOMOLECULAR SYSTEMS IN THE CANONICAL ENSEMBLE: A REVIEW OF SELF-GUIDED LANGEVIN DYNAMICS

XIONGWU WU[1], ANA DAMJANOVIC[1,2], AND BERNARD R. BROOKS[1]

[1] *Laboratory of Computational Biology, National Heart, Lung, and Blood Institute (NHLBI), National Institutes of Health (NIH), 5635 Fishers Lane, Bethesda, MD 20892-9314, USA*

[2] *Department of Biophysics, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218, USA*

CONTENTS

Advances in Chemical Physics, Volume 150, First Edition. Edited by Stuart A. Rice and Aaron R. Dinner.

I. THE CONFORMATIONAL SEARCH PROBLEM

Conformational search is problematic for simulation systems where populated states are either separated by energy barriers or kinetic bottlenecks or are spread across a distance that corresponds to significant conformational changes. In biological systems, conformational search is very challenging because biological molecules such as proteins or DNA have a huge conformational space and numerous energy barriers. Biological relevant events, such as protein folding [1], ligand binding, conformational signal transduction, and so on, occur in a timescale far exceeding that accessible by current realistic simulations [2].

The conformation search problem for macromolecules has been the subject of intense efforts for many decades. There are numerous methods and approaches, each with various strengths and weaknesses, and there are several review articles that survey these methods rather well [3–9].

Among the many methods for efficient conformational search, the self-guided molecular dynamic (SGMD) [10, 11] and the self-guided Langevin dynamics (SGLD) [12–14] simulation methods are somewhat unique. The term "self-guided" refers to the manner in which the information learned during a simulation is used to enhance the conformational search of the very same simulation. The core of these methods is the use of local averages of force and momentum as a guiding force that accelerates barrier crossing in a manner that can also preserve the canonical ensemble. Even though these methods have been discussed in studies by Norberg and Nilsson [8], Tai [9], and Christen and Van Gunsteren [3], this chapter presents a more complete description that includes recent developments.

To better understand how SGLD relates to the many other sampling and search methods, it is worthwhile to categorize sampling methods by considering the following nine questions for which we contrast SGLD with alternative methods:

1. Are structures found by iterative sampling, or are structures found with a construction/library/buildup/genetic procedure?

The standard techniques of Metropolis Monte Carlo (MC), molecular dynamics (MD), and Langevin dynamics (LD) are typical iterative sampling methods that are designed to sample the canonical ensemble by default. By contrast, there is a wide variety of buildup [15] and construction methods that make use of libraries [16–19], genetic algorithms [20–24], or exhaustive enumeration [25–28]. There are also combined methods, such as the conformation space annealing (CSA) [29, 30], which uses multiple techniques to generate an extensive variety of widely separated conformations. The SGLD methods are extensions of LD and the SGMD methods are extensions of MD and both involve direct iterative sampling.

2. Is the method efficient relative to standard MD? How much so?

The use of MD or LD yields excellent results in the long-time limit, but for macromolecular systems, they are often too pedestrian to be optimal for efficient conformational searching. Barrier height much larger than $10kT$ becomes rare events that are best explored with other methods. Methods that rely on MD or LD for some degrees of freedom (e.g., typical free energy perturbation simulations) converge well only if those degrees of freedom do not have important barriers in the 10–$20kT$ range. Accurate calculation of free energy depends on efficient conformational sampling. If SGLD, or one of its variants, is used as a replacement for MD, then the problematic barriers can be found in the 20–$30kT$ range, a range that simply is not explored with standard MD. Whether free energy convergence is improved depends on the macromolecular system and specifically what important states are separated by those higher barriers that are inaccessible by MD. Better sampling does not simply equate with better convergence behavior.

3. Is the canonical ensemble directly generated, or via reweighting, or is a nonensemble collection of structures generated?

There are three categories here. Methods that search only space without producing a canonical ensemble are useful in many ways but cannot be used for calculating free energies or potentials of mean force. Methods that can generate a canonical ensemble do so directly or via a reweighting procedure that corrects bias introduced by the sampling procedure. For example, if an MD simulation is done at an elevated temperature, an ensemble average at a lower temperature can be obtained by reweighting the contribution of each frame by using

$$\langle P\rangle_T = \frac{\left\langle P \exp\left(-\frac{T'-T}{kTT'}E_{\mathrm{p}}\right)\right\rangle_{T'}}{\left\langle \exp\left(-\frac{T'-T}{kTT'}E_{\mathrm{p}}\right)\right\rangle_{T'}} \tag{1}$$

The reweighting factors changes exponentially with the temperature difference and the fluctuation of the total potential energy. In practice, the reweighting procedure works well only when the temperature differences are small and for smaller systems. For large systems, the fluctuation of the potential energy is also large and the averaging converges very poorly, or not at all. This procedure is thus not considered to be size extensive. Simulation methods that directly generate the canonical ensemble are preferred and can be used with very large systems. Variants of SGLD preserve the ensemble via reweighting or directly. The SGLD variants that preserve the ensemble directly do not sample as efficiently as the original SGLD method, but still much more efficiently than LD.

4. Is the trajectory continuous?

Methods with continuous trajectories are suitable for pathway studies, while the others must focus only on conformational sampling and ensemble generation. There are several methods that generate the desired ensemble, but the trajectory may not be continuous. Temperature replica exchange (TREx) [31] is such a method. The ensemble at the specific target temperature is discontinuous whenever that temperature is involved in an accepted exchange. Thus, replica exchange approaches cannot be used to measure correlations for events that occur on a timescale longer than the mean time between accepted exchanges.

5. Is the timescale preserved, or is the timescale lost via acceleration?

Classical MD does preserve the timescale, and if the sampling is sufficiently good, then rates can be calculated directly from time correlation functions. For SGLD, the continuous trajectories can be analyzed using time correlation functions, but the connection between simulation time and real time is not straightforward. The rate acceleration of the crossing of any given potential energy barrier depends on a number of factors. It is not safe to assume that rank order of rates in the accelerated SGLD system is the same as the rank order of rates found with MD. Additional development work is needed before SGLD can be used to accurately estimate transition rates. It may be reasonable to assume that the order of events seen with SGLD simulations may reflect the same order of events that would be observed with very long MD simulations, but there is no formal justification for this assumption.

6. Is the sampling method direct, or indirect via exchanges or couplings?

TREx achieves accelerated sampling via exchanges to simulations at other temperatures [31, 32] or simulations on a modified Hamiltonians [33]. This is also referred to as parallel tempering. Such approaches have both benefits and weaknesses. One practical weakness is that it is more difficult to combine TREx with other sampling methods. For example, combining TREx with metadynamics (MetaD) has been developed and used to good effect [34]; however, it would have been considerably easier to simply replace the LD integrator with an SGLD type of integrator, and the overall results would have likely been improved with lower simulation costs, especially for larger systems where a large number of replicas with small temperature differences are needed to obtain converged ensemble averages.

7. Does there need to be a predetermination of enhanced degrees of freedom, or are all degrees of freedom enhanced?

Many sampling methods require a predetermination of important degrees of freedom. With the targeted molecular dynamics method (TMD) [35], the target degree of freedom must be predetermined. With metadynamics [36], the bias potential degrees of freedom must be predetermined. If one knows exactly what degrees of freedom need to be enhanced, then this is fine. But in many cases, such knowledge is not known to sufficient detail. In such cases, an unbiased method may be preferred. SGLD enhances all motion without the need to predetermine which degrees of freedom to enhance. It enhances all degrees of freedom that are coupled to the local averages of force and momentum.

8. Is there an effective maximum barrier height, or is all space explored?

Another concern is when a sampling method is too efficient. For example, with the CHARMM22 protein force field [37], both L- and D-amino acids have low-energy conformations. When structures are generated with CSA [30], all combinations of chirality are found in peptides. If one wants only L-amino acids, then restraints are required. However, for simulations methods, such chirality transitions are never observed due to the large energy barrier. Another often unwanted protein conformation is a *cis*-peptide. Except for prolines and adjacent cystines, such conformations are generally unwanted in simulations. With MD or LD, the *cis–trans* isomerization barrier of a peptide bond is insurmountable, however, with SGLD, such barrier height of the *cis–trans* can be crossed. To avoid this problem, the force field can be modified, restraints can be added, or less aggressive SGLD parameters can be employed. But when only aggressive SGLD parameters are used, one should carefully monitor such dihedral orientations.

9. Is the method size extensive?

An ideal sampling method is size extensive, meaning that it can be run on increasingly large simulation systems without breaking down. Molecular dynamics is size extensive, in that the various degrees of freedom are essentially propagated independently. Replica exchange methods are not size extensive since the number of replicas required also increases with system size. Methods where the enhanced degrees of freedom must be predetermined are generally not size extensive. Methods that connect to a desired ensemble via an exponential reweighting formula are not size extensive because energy differences are contained in the exponent. This limits the method to small systems or larger systems where only a small region is accelerated. The SGLDfp method is unique in that it is apparently the only direct accelerated sampling method that is both size extensive and preserves the correct ensemble.

The SGMD [10, 11] and SGLD [12–14] simulation methods were developed for an efficient conformational search and have found many applications to study rare events, such as protein folding [32, 38–45], ligand binding [46–48], docking [49], conformational transitions [50–53], crystallization [11, 54–57], and surface absorption [58, 59].

Despite many applications of SGLD and SGMD methods, the lack of understanding of the guiding effect on conformational distribution and conformational search hindered the acceptance of this method in simulation studies. Recently, a quantitative understanding of the perturbation on conformational distribution by the local average momentum-based guiding force has been achieved [14]. The partition function of an SGLD ensemble is quantitatively related to the so-called low-frequency properties defined based on the local averaging scheme. Through the SGLD partition function, the conformational distribution obtained in SGLD simulations can be converted to a canonical ensemble distribution and ensemble average properties can be calculated from SGLD simulations through reweighting.

On the basis of the understanding of SGLD conformational distribution, and combined with SGMD simulation method [10, 11], we developed the force–momentum-based self-guided Langevin dynamics or SGLDfp [13]. This method adds a force-based guiding force to cancel any conformational bias in the momentum-based guiding effect [13]. Through this combined guiding force, SGLDfp achieves an unbiased canonical conformational distribution without the need for reweighting.

In this chapter, we provide a comprehensive picture of the SGLD method to explain why and how it works. We first present the history of the development of SGMD and SGLD methods. The theoretical basis as well as the simulation methods in a variety of forms is provided in a comprehensive way in Section III. In Section IV through several simple systems, we explain the reason why SGLD can enhance conformational search. In Section V, we review the applications of SGMD and SGLD in computational studies. Finally, in Section VI, we present the development direction and guidance in applying SGMD and SGLD methods.

II. HISTORY OF THE SGMD AND SGLD METHODS

The idea of a self-guided simulation is to promote conformational transitions according to the information extracted during the same simulation in order to achieve faster convergence in conformational sampling. The information extracted during a simulation is called a local average property. The average is taken over the conformational space near the current conformation and can be approximately estimated by the following function:

$$\langle P\rangle_{\mathrm{L}}[n] = \frac{L-1}{L}\langle P\rangle_{\mathrm{L}}[n-1] + \frac{1}{L}P[n] \tag{2}$$

Here, L is the number of local conformations used for the averaging and $P[n]$ is a conformational property at conformation n. The symbol $\langle\ \rangle_{\mathrm{L}}$ denotes a local average. The contribution of any conformation to a local average decays exponentially with a decay factor of L.

The local averaging was first utilized to estimate the mean solvation force in protein folding simulations with explicit solvent [41]. Explicit solvent molecules dampen protein motion and the conformational transition is slow. Furthermore, the noise from the solvent collisions is overwhelming; as a result, much of protein motion appears as a random walk. The mean force of the interaction with the solvent represents the solvation free energy force, which excludes the noise of solvent interaction and guides the protein to conformations favored by solvation. In this method, the solvent environment is simulated with a Monte Carlo method, while the protein was simulated with molecular dynamics, and the mean solvation force was replaced by the local average solvent interaction force and calculated using Eq. (2) with a local average size of $L = 10$. The equation of motion for the protein can be written as

$$\dot{\mathbf{p}}_i = \mathbf{f}_i + \tilde{\mathbf{s}}_i$$

where $\dot{\mathbf{p}}_i$ and $\mathbf{f}_i$ are the time derivative of momentum and the interaction force of atom i in a protein, respectively. $\tilde{\mathbf{s}}_i$ is the local average force of solvent on protein atom i.

The SGMD [10, 11] simulation method was developed by extending the local average force to all atoms and by including all nonbonded forces.

$$\dot{\mathbf{p}}_i = \mathbf{f}_i + \mathbf{g}_i$$

Here, $\mathbf{g}_i$ is the guiding force, which is calculated as a local average of the nonbonded force:

$$\mathbf{g}_i(t) = \lambda\langle\mathbf{f}_i(t) + \lambda\mathbf{g}_i(t-\delta t)\rangle_{\mathrm{L}} = \left(1 - \frac{\delta t}{t_{\mathrm{L}}}\right)\mathbf{g}_i(t-\delta t) + \frac{\delta t}{t_{\mathrm{L}}}\lambda\left(\mathbf{f}_i(t) + \lambda\mathbf{g}_i(t-\delta t)\right) \tag{3}$$

The parameter, λ, is the guiding factor, δt is the time step, and $t_{\mathrm{L}} = L\delta t$ is the local averaging time. In an SGMD simulation, the system undergoes an accelerated systematic motion, which is defined by the local averaging time, t_{L}, while maintaining a desired temperature. Many applications have demonstrated that SGMD simulations have enhanced conformational search ability [43–47, 55, 56, 58, 59]. Shinoda and Mikami extended SGMD to the *NPT* ensemble [60] and later combined it with the rigid body dynamics [61].

There are several drawbacks when applying the SGMD method. First, the guiding force calculated by Eq. (3) is correlated with the force field and results in an unwanted alteration of the conformational distribution. Second, for molecular systems, high-frequency bonded interactions need to be excluded in the guiding force calculation to avoid excessive noise. Third, as pointed by Lahiri et al. [62], the guiding force derived from the local average of actual forces may not be sufficient to enhance conformational searching in stochastic dynamics simulations.

Andricioaei et al. extended the self-guiding idea to a hybrid Monte Carlo simulation method (MHMC) to enhance conformational sampling efficiency [63]. They used the local average momentum as a guide to bias the initial choice of momenta at each step. They demonstrated that their self-guided enhanced sampling method enhances conformational sampling efficiency while producing, theoretically, correct thermodynamic average properties in the weak perturbation limit.

The local average momentum has some advantages over the local average force to be used as the guiding force. However, due to the correlation between the local average momentum and the instantaneous momentum, directly applying the guiding force of this type could make fast objects move faster and cause an uneven distribution of kinetic energy throughout the simulation system. This problem can be solved with Langevin dynamics where every degree of freedom is constrained independently with a thermal bath.

LD simulation has been a very useful tool in macromolecule studies [64]. It is also used as a temperature control scheme to maintain constant temperature [65]. Obviously, introducing a guiding force to accelerate the systematic motion can enhance conformational search efficiency of an LD simulation. On the basis of the position Langevin equation, we found that the guiding force can be represented by the local average of friction forces, which is proportional to the local average of momenta. Therefore, the guiding force takes the form of local average momentum in the SGLD method

$$\dot{\mathbf{p}}_i = \mathbf{f}_i + \lambda_i \gamma_i (\langle \mathbf{p}_i \rangle_{\mathrm{L}} - \xi \mathbf{p}_i) - \gamma_i \mathbf{p}_i + \mathbf{R}_i \tag{4}$$

where γ_i is the collision frequency and $\mathbf{R}_i$ is a random force for particle i. The parameter, ξ, is an energy conservation parameter, which is set to cancel the extra energy input from the guiding forces.

The enhanced conformational searching ability of SGMD and SGLD is demonstrated by their many applications in protein folding [32, 38, 43, 44, 45], ligand binding [46–48], conformational transitions [50–53], phase transitions [11, 54, 57, 66, 67], and surface adsorption [58, 59]. There are several method developments along the same concept of SGLD. For example, Yang and Gao presented an approximate method to use a relatively short normal dynamics simulation to obtain slow motion information to propagate structural changes in the slow degrees of freedom [5, 68]. Similarly, MacFadyen et al. proposed a method that uses a directional negative friction force to enhance sampling efficiency for rare events [69]. In earlier applications, there was a lack of understanding about why conformational search is accelerated and how the guiding forces affect conformational distribution. The most practical question is how to obtain canonical conformational distribution with the accelerated conformational search techniques. These questions hindered the application of SGMD and SGLD in quantitative studies, such as free energy calculations.

Recently, a quantitative understanding of the perturbation on conformational distribution by the local average momentum-based guiding force has been achieved [14]. The partition function in an SGLD ensemble is quantitatively related to the so-called low-frequency properties defined based on the local averaging scheme. Through the SGLD partition function, the conformational distribution obtained in SGLD simulations can be converted to a canonical ensemble distribution, and ensemble average properties can be calculated from SGLD simulations through reweighting. Because the energy distribution of an SGLD simulation is very close to the canonical distribution, the reweighting can be calculated with high accuracy for a large range of guiding factors. Another convenience is that the reweighting factor can be computed efficiently either on-the-fly or during postprocessing, which means that SGLD can be used to compute free energies in a direct manner without the need for postprocessing.

Based on the understanding of the SGLD conformational distribution, an SGLD combined with the SGMD simulation method [10, 11] called force–momentum-based self-guided Langevin dynamics has been developed. This method uses the force-based guiding force to cancel any conformational bias due to the momentum-based guiding effect [13]. Through this combined guiding force, conformational search can be accelerated while preserving canonical conformational distribution without the need for reweighting. In other words, the method is explicitly designed so that every sampled conformation has the same reweighting coefficient. One drawback of SGLDfp is that it is not as efficient as SGLD with reweighting. As a rough rule of thumb, SGLD will cross barriers of $20kT$ at the rate that LD or MD will cross barriers of $10kT$ (an effective doubling of temperature), but SGLDfp only crosses barriers of $15kT$ at the same rate. Details depend on systems and parameters used, so this is only a rough guide.

This progress in the understanding of the SGLD conformational distribution and conformational search and the development of the SGLDfp method open the door for numerous types of quantitative simulation studies.

III. THERMODYNAMICS OF SGMD AND SGLD

A. Low-Frequency and High-Frequency Properties

Thermal motion in a molecular system has a wide distribution of frequencies. Chemical bonds vibrate and bend at high frequencies, while ion transport and protein folding events occur on a relatively long time. High-frequency events repeat on a short timescale and are often the easiest to study in molecular simulations. However, it is the low-frequency events that are important for many macroscopic behaviors, such as protein folding, binding, and conformational rearrangements, but are often beyond the timescale accessible by molecular simulations with available computing resources.

Low-frequency properties are related to low-frequency events. For example, dimerization of a pair of water molecules depends on the relative position between the water molecules. This dimer energy means the energy when the two water molecules are at the dimer state, that is, the average among all bond vibration and bending states. This dimer energy represents the energy at a frequency of dimerization, which is a slow event compared to bond vibration and bending. At each given moment, bond vibration and bending, and even electron density fluctuations, produce an instantaneous energy deviation that depends on the high-frequency motions. The energy associated with such high-frequency motions is called high-frequency energy. Compared to the bond vibration and bending, dimer energy is low-frequency energy, which is an average over all the vibration and bending states. For slow events, low-frequency properties give a more accurate picture, while for fast events, high-frequency properties are needed to describe them.

We propose to define a low-frequency property by the so-called local average property. A local averaging procedure involving an exponential decay average [10, 11, 12, 41]), typically on force or momentum, is performed by the following equation:

$$\langle P\rangle_{\mathrm{L}} = \frac{1}{L}\sum_{i=n-L+1}^{n} P[i] = \frac{1}{t_{\mathrm{L}}}\int_{t-t_{\mathrm{L}}}^{t} P(\tau)\,d\tau \approx \left(1-\frac{1}{L}\right)\tilde{P}[n-1] + \frac{1}{L}P[n]$$
$$= \left(1-\frac{\delta t}{t_{\mathrm{L}}}\right)\tilde{P}(t-\delta t) + \frac{\delta t}{t_{\mathrm{L}}}P(t) = \tilde{P} \tag{5}$$

Here, $P[i]$ represents property P at the ith data point and $P(t)$ represents the one at time t. As can be seen from Eq. (5), a local average, denoted as "$\langle\ \rangle_{\mathrm{L}}$," is calculated by averaging over the most recent L points or the most recent $t_{\mathrm{L}} = L\delta t$ time period. Here, δt is the time interval between data points. We call L the local averaging size and t_{L} the local averaging time. This average can be approximately calculated as an evolving average with a constant updating of current value as shown in the right-hand side of Eq. (5). This evolving average is denoted by a "$\sim$" cap: $\tilde{P}$. Because all local averages in this work are calculated as evolving averages, we also use "$\langle P\rangle_{\mathrm{L}}$" to represent evolving averages when the cap "$\sim$" is not easy to print. Corresponding to the low-frequency properties, we define high-frequency properties as the difference between instantaneous properties and their low-frequency ones: $P-\tilde{P}$. Both the low- and high-frequency properties are conformational dependent or time dependent and can be expressed as functions of time $\tilde{P}(t)$ and $P(t)-\tilde{P}(t)$, in molecular dynamics simulation.

The local averaging shown in Eq. (5) suppresses high-frequency effects and emphasizes low-frequency contributions. From Eq. (5), we can see that the local averaging time, t_{L}, determines the range of contributing frequencies. To better understand the evolving averaging, we can rearrange Eq. (5) to the following form:

$$\frac{\tilde{P}(t)-\tilde{P}(t-\delta t)}{\delta t} = \frac{P(t)-\tilde{P}(t-\delta t)}{t_{\mathrm{L}}}$$

When $\delta t \to 0$, we have

$$\frac{d\tilde{P}(t)}{dt} = \frac{P(t)-\tilde{P}(t)}{t_{\mathrm{L}}}$$

This differential equation can be solved by

$$\tilde{P}(t) = \frac{1}{t_{\mathrm{L}}} \int_0^t P(\tau)\, e^{-(t-\tau)/t_{\mathrm{L}}}\, d\tau \tag{6}$$

Therefore, a property at any moment provides an exponentially decaying contribution to the evolving average as a function of time. The decaying rate depends on the local averaging time t_{L}.

The separation of the low-frequency properties and the high-frequency properties is at the heart of the SGLD simulation method. The low-frequency properties are calculated through the evolving averaging shown in Eq. (5). To explain the behavior of the evolving averages, we use $q(t) = \sin(2\pi\varpi t)$ as an example function of frequency ϖ to show how frequency and local averaging time affect the evolving average.

Substituting $q(t) = \sin(2\pi\varpi t)$ into Eq. (6), we get its evolving average

$$\tilde{q}(t) = \frac{2\pi\varpi t_{\mathrm{L}}(e^{-t/t_{\mathrm{L}}} - \cos(2\pi\varpi t)) + \sin(2\pi\varpi t)}{1 + 4\pi^2 t_{\mathrm{L}}^2 \varpi^2} \tag{7}$$

As can be seen from Eq. (7), for high frequency, $2\pi\varpi t_{\mathrm{L}} \gg 1$, the amplitude of $\tilde{q}(t)$ is inversely proportional to ϖ, while for low frequency, $2\pi\varpi t_{\mathrm{L}} \ll 1$, $\tilde{q}(t) \approx q(t)$. The local averaging time, t_{L}, defines the separation of what is high frequency and what is low frequency compared to a local averaging frequency of $\varpi_{\mathrm{L}} = 1/t_{\mathrm{L}}$. This example shows that the evolving averaging suppresses the high-frequency contribution, while it has less effect on low-frequency components. The high-frequency portion can be expressed as

$$q(t) - \tilde{q}(t) = \frac{-2\pi\varpi t_{\mathrm{L}}(e^{-t/t_{\mathrm{L}}} - \cos(2\pi\varpi t)) + 4\pi^2 t_{\mathrm{L}}^2 \varpi^2 \sin(2\pi\varpi t)}{1 + 4\pi^2 t_{\mathrm{L}}^2 \varpi^2} \tag{8}$$

As can be seen from Eq. (8), when $2\pi\varpi t_{\mathrm{L}} \gg 1$, $q(t) - \tilde{q}(t) \approx \sin(2\pi\varpi t) = q(t)$, and when $2\pi\varpi t_{\mathrm{L}} \ll 1$, $q(t) - \tilde{q}(t) \approx -2\pi\varpi t_{\mathrm{L}}(e^{-t/t_{\mathrm{L}}} - \cos(2\pi\varpi t)) \to 0$. That is, the high-frequency portion keeps the high-frequency contributions while suppressing the low-frequency components.

Figure 1a shows the example function and its evolving averages at different local averaging times. Clearly, one can see that the frequencies of the averaging results remain the same as the example function, but the amplitudes and phases are very different from each other. When $\varpi t_{\mathrm{L}} = 0.1$, this function represents a low-frequency motion and its evolving average has a magnitude similar to the function. When $\varpi t_{\mathrm{L}} = 10$, this function represents a high-frequency motion and the magnitude of its evolving average is very small compared to the function. Figure 1b shows an averaging result as a function of ϖt_{L}. The envelope function represents the amplitude of the averages. Clearly, one can see that, with a small ϖt_{L}, the amplitude of the average is similar to the example function, while with a

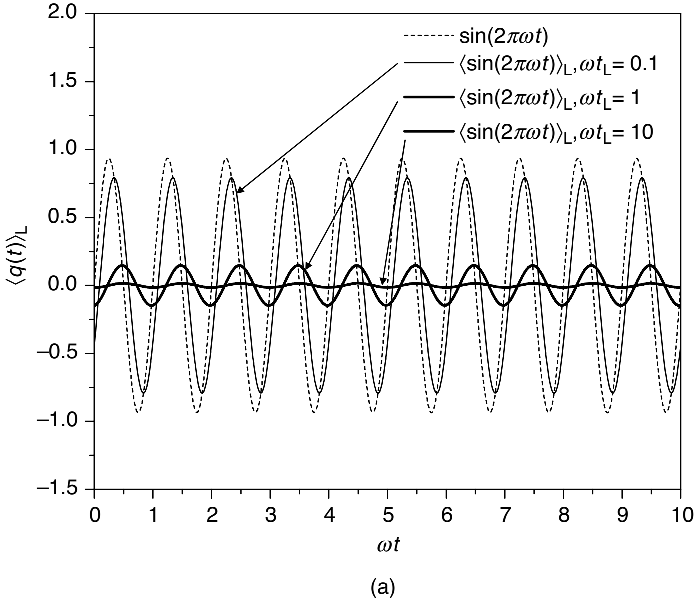


Figure 1. (**a**) The example function, $q(t) = \sin(2\pi\varpi t)$, and its evolving averages at three local averaging times: $\varpi t_{\mathrm{L}} = 0.1$, 1, and 10. (**b**) The evolving average of the example function as a function of the frequency. The envelope curves show the amplitude as a function of ϖt_{L}. At small ϖt_{L}, which corresponds to a low frequency, the amplitude is approaching 1, very similar to that of the example function, while at a large ϖt_{L}, which corresponds to a high frequency, the amplitude approaches 0.

large ϖt_{L}, the amplitude of the average approaches zero, indicating that the low-frequency function will remain in the evolving average and the high-frequency function will be suppressed.

In summary, conformational properties can be separated into high-frequency and low-frequency properties based on t_{L}. Through the local averaging, many low-frequency properties can be obtained in molecular simulation. For example, low-frequency forces

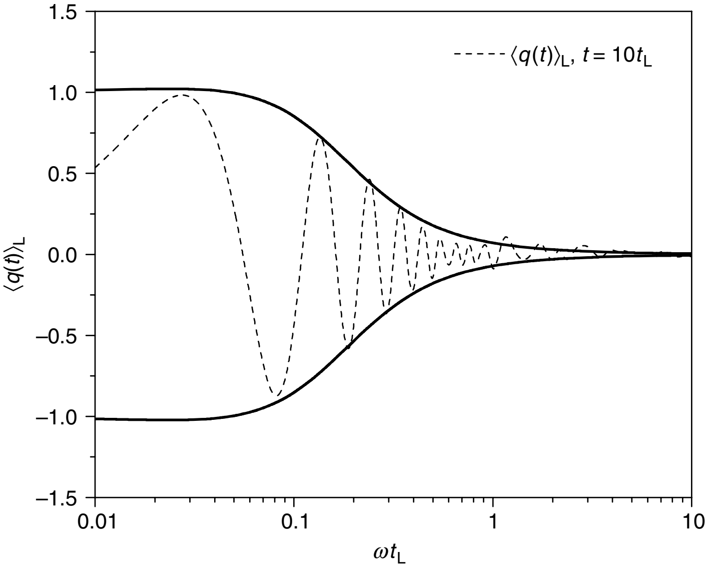
$$\tilde{\mathbf{f}}_i(t) = \left(1 - \frac{\delta t}{t_{\mathrm{L}}}\right) \tilde{\mathbf{f}}_i(t - \delta t) + \frac{\delta t}{t_{\mathrm{L}}} \mathbf{f}_i(t)$$

low-frequency momenta

$$\tilde{\mathbf{p}}_i(t) = \left(1 - \frac{\delta t}{t_{\mathrm{L}}}\right) \tilde{\mathbf{p}}_i(t - \delta t) + \frac{\delta t}{t_{\mathrm{L}}} \mathbf{p}_i(t)$$

and low-frequency potential energies

$$\tilde{E}_{\mathrm{p}}(t) = \left(1 - \frac{\delta t}{t_{\mathrm{L}}}\right) \tilde{E}_{\mathrm{p}}(t - \delta t) + \frac{\delta t}{t_{\mathrm{L}}} E_{\mathrm{p}}(t) \tag{9}$$



(b)

Figure 1. (*Continued*)

We can calculate some derived low-frequency quantities from these low-frequency properties, such as the low-frequency temperature:

$$\tilde{T} = \frac{1}{N_{DF}k}\left\langle \sum_i \frac{\tilde{\mathbf{p}}_i^2}{m_i} \right\rangle \tag{10}$$

Here, N_{DF} is the number of degrees of freedom and k is the Boltzmann constant. m_i is the mass of particle i and the summation runs over all atoms in a system. The bracket, $\langle\rangle$, represents an ensemble average.

B. SGMD and SGLD Simulation Methods

Because molecular dynamics can be regarded as a special case of Langevin dynamics, to be general, we give the following description and explanation based on the self-guided Langevin dynamics. The equation of the self-guided motion can be written in the following general form:

$$\dot{\mathbf{p}}_i = \mathbf{f}_i + \mathbf{g}_i - \gamma_i \mathbf{p}_i + \mathbf{R}_i \tag{11}$$

where $\dot{\mathbf{p}}_i$ and $\mathbf{f}_i$ are the time derivative of momentum and the interaction force of particle i, respectively. $\mathbf{R}_i$ is a random force, which is related to mass, m_i, the

TABLE I
The Guiding Forces Used in Self-Guiding Molecular Dynamics and Self-Guided Langevin Dynamics in a Variety of Derivative Forms

Name	Parameters	Guiding force[a]
SGMD [10, 11][b]	$\lambda^{\mathrm{f}}, t_{\mathrm{L}}$	$\mathbf{g}_i(t) = \lambda_i^{\mathrm{f}} \left(\tilde{\mathbf{f}}_i^{(\mathrm{nb})}(t) + \mathbf{g}_i(t-\delta t) - \xi^{\mathrm{p}}\mathbf{p}_i(t)\right)$
SGMDf	$\lambda^{\mathrm{f}}, t_{\mathrm{L}}$	$\mathbf{g}_i(t) = \lambda_i^{\mathrm{f}} \left(\tilde{\mathbf{f}}_i(t) - \xi^{\mathrm{p}}\mathbf{p}_i(t)\right)$
SGMDp[c]	$\lambda_i^{\mathrm{p}}, t_{\mathrm{L}}$	$\mathbf{g}_i(t) = \lambda_i^{\mathrm{p}}\gamma^0 (\tilde{\mathbf{p}}_i(t) - \xi^{\mathrm{p}}\mathbf{p}_i(t))$
SGMDfp[d]	$\lambda_i^{\mathrm{p}}, t_{\mathrm{L}}$	$\mathbf{g}_i(t) = \lambda^{\mathrm{f}}\tilde{\mathbf{f}}_i(t) - \xi^{\mathrm{f}} f_i(t) + \lambda_i^{\mathrm{p}}\gamma^0 (\tilde{\mathbf{p}}_i(t) - \xi^{\mathrm{p}}\mathbf{p}_i(t))$
SGLD or SGLDp [12, 14]	$\lambda_i^{\mathrm{p}}, t_{\mathrm{L}}, \gamma_i$	$\mathbf{g}_i(t) = \lambda_i^{\mathrm{p}}\gamma_i (\tilde{\mathbf{p}}_i(t) - \xi^{\mathrm{p}}\mathbf{p}_i(t))$
SGLDf	$\lambda^{\mathrm{f}}, t_{\mathrm{L}}, \gamma_i$	$\mathbf{g}_i(t) = \lambda_i^{\mathrm{f}} \left(\tilde{\mathbf{f}}_i(t) - \xi^{\mathrm{p}}\gamma_i\mathbf{p}_i(t)\right)$
SGLDfp [13]	$\lambda_i^{\mathrm{p}}, t_{\mathrm{L}}, \gamma_i$	$\mathbf{g}_i(t) = \lambda^{\mathrm{f}}\tilde{\mathbf{f}}_i(t) - \xi^{\mathrm{f}}\mathbf{f}_i(t) + \lambda_i^{\mathrm{p}}\gamma_i (\tilde{\mathbf{p}}_i(t) - \xi^{\mathrm{p}}\mathbf{p}_i(t))$

[a] The parameter, ξ^{p}, is an energy conservation factor to cancel the energy input from the guiding force and can be calculated by $\sum_i \mathbf{g}_i(t) \cdot \dot{\mathbf{r}}_i(t) = 0$. $\gamma^0 = 1\,\mathrm{ps}^{-1}$ is a force converting factor.

[b] In SGMD, only nonbonded forces, $f_i^{(\mathrm{nb})}$, are used to calculate the guiding force.

[c] SGMDp is SGLD with $\gamma_i = 0$.

[d] SGMDfp is SGLDfp with $\gamma_i = 0$.

collision frequency, γ_i, and simulation temperature, T, by the following equation:

$$\langle \mathbf{R}_i(0)\mathbf{R}_i(t)\rangle = 2m_i kT\gamma_i\delta(t) \tag{12}$$

$\mathbf{g}_i$ is called the guiding force and is calculated based on the low-frequency momentum, low-frequency force, or both. Even though Eq. (11) is in the form of the self-guided Langevin dynamics, it can represent an SGMD motion when the collision frequency, γ_i, and the random force, $\mathbf{R}_i$, are zero. From Eq. (11), we can see that molecular dynamics and Langevin dynamics are special cases of SGLD when the guiding force is zero and the collision frequency is zero. Depending on how the guiding force is calculated, Eq. (11) can represent different kinds of self-guided dynamics motion. For example, SGMD calculates the guiding force with nonbonded forces [10, 11], SGLD uses momenta [12] and is also referred to as SGLDp, and SGLDfp uses both forces and momenta [13]. As a summary, Table I lists the derivative forms of SGLD and their guiding forces. So far, SGMD, SGLD, and SGLDfp have been well documented, while SGMDf, SGMDp, SGLDf, and SGMDfp have not been studied.

C. Conformational Distribution in SGLD

The guiding force in an SGLD simulation is designed to accelerate the low-frequency motion so that the conformational search efficiency can be enhanced. It has two types of effects on a simulation system. First, the guiding force enhances low-frequency motion as measured by the increase in the low-frequency temperature, and it also reduces the high-frequency motion due to the energy conservation force that comes with the guiding force. Second, the guiding force produces a bias in the energy surface. To understand the conformational distribution in SGLD simulation, we separately examine the low-frequency motion and high-frequency motion.

In the low-frequency conformational space, the equation of motion can be expressed as a low-frequency portion of Eq. (11):

$$\dot{\tilde{\mathbf{p}}}_i = \tilde{\mathbf{f}}_i + \tilde{\mathbf{g}}_i - \gamma_i \tilde{\mathbf{p}}_i + \tilde{\mathbf{R}}_i \tag{13}$$

The low-frequency motion is on the low-frequency potential energy surface, $\tilde{E}_{\mathrm{p}}$, under the low-frequency interaction force, $\tilde{\mathbf{f}}_i$, and the low-frequency guiding force, $\tilde{\mathbf{g}}_i$, the low-frequency friction force, $\gamma_i \tilde{\mathbf{p}}_i$, and the low-frequency random force, $\tilde{\mathbf{R}}_i$. Based on the "position Langevin equation," the momentum and the momentum-based guiding force are correlated with the interaction force [12]. Therefore, the total low-frequency force, $\tilde{\mathbf{f}}_i + \tilde{\mathbf{g}}_i - \gamma_i \tilde{\mathbf{p}}_i + \tilde{\mathbf{R}}_i$, acts as a force from a scaled low-frequency potential energy surface, $E_{\mathrm{lf}} = \lambda_{\mathrm{lf}} \tilde{E}_{\mathrm{p}}$. The low-frequency energy factor, λ_{lf}, can be approximated according to the average projection of the total low-frequency force in the direction of the low-frequency interaction forces:

$$\lambda_{\mathrm{lf}} = \frac{\left\langle \sum_i (\tilde{\mathbf{f}}_i + \tilde{\mathbf{g}}_i - \gamma_i \tilde{\mathbf{p}}_i) \tilde{\mathbf{f}}_i \right\rangle}{\left\langle \sum_i \tilde{\mathbf{f}}_i \tilde{\mathbf{f}}_i \right\rangle} \tag{14}$$

Beside the scaling effect in the low-frequency potential energy, the guiding force also enhances the low-frequency motion. This enhanced low-frequency motion corresponds to an elevated thermal temperature in the low-frequency conformational space. We define this thermal temperature in the low-frequency conformational space as T_{lf}. It is reasonable to assume that T_{lf} is proportional to the low-frequency temperature $\tilde{T}$:

$$\frac{T_{\mathrm{lf}}}{T} = \frac{\tilde{T}}{\tilde{T}_0} \tag{15}$$

Here, $\tilde{T}_0$ is the low-frequency temperature when $\lambda = 0$ and is called the reference low-frequency temperature. On the basis of the definition, we know that $\tilde{T}_0$ depends on the simulation condition and the local averaging time t_{L}.

To understand the relationship between $\tilde{T}$ and $\tilde{T}_0$, we can rewrite the low-frequency motion, Eq. (11), to a Langevin dynamics form:

$$\dot{\tilde{\mathbf{p}}}_i = \tilde{\mathbf{f}}_i - \chi_{\mathrm{lf}} \gamma_i \tilde{\mathbf{p}}_i + \tilde{\mathbf{R}}_i \tag{16}$$

Equation (16) corresponds to a Langevin dynamics with a collision frequency of $\chi_{\mathrm{lf}} \gamma_i$. The factor, χ_{lf}, is called the low-frequency collision factor and can be calculated according to the projection of the low-frequency guiding force in the direction of the low-frequency friction force:

$$\chi_{\mathrm{lf}} = \frac{\left\langle \sum_i (\gamma_i \tilde{\mathbf{p}}_i - \tilde{\mathbf{g}}_i)\gamma_i \tilde{\mathbf{p}}_i \right\rangle}{\left\langle \sum_i \gamma_i^2 \tilde{\mathbf{p}}_i \tilde{\mathbf{p}}_i \right\rangle} \tag{17}$$

Based on the Langevin dynamics relation [Eq. (12)], with a given distribution of random forces, the product of temperature and collision frequency is a constant:

$$T\gamma_i = \frac{\langle \mathbf{R}_i(0)\mathbf{R}_i(t)\rangle}{2m_i k\delta(t)} \tag{18}$$

The reference low-frequency temperature, $\tilde{T}_0$, corresponds to the low-frequency temperature at a collision frequency of γ_i, while the low-frequency temperature in an SGLD simulation, $\tilde{T}$, corresponds to that at the collision frequency of $\chi_{\mathrm{lf}}\gamma_i$. Because the guiding force does not affect the random force, we have from Eq. (18)

$$\tilde{T}_0 = \tilde{T}\chi_{\mathrm{lf}} \tag{19}$$

Equation (19) provides a relationship between χ_{lf}, $\tilde{T}_0$, and $\tilde{T}$. We can calculate χ_{lf} from Eq. (16) or Eq. (19) with$\chi_{\mathrm{lf}} = \tilde{T}_0/\tilde{T}$, which is more accurate if $\tilde{T}_0$ is obtained from a previous SGLD simulation with $\lambda = 0$.

Combining the scaling in the low-frequency potential energy surface and the acceleration in the low-frequency motion, we have the partition function in the low-frequency conformational space:

$$\Theta_{\mathrm{lf}}(N, V, T_{\mathrm{lf}}) = \sum_{\Omega_{\mathrm{lf}}} \exp\left(-\frac{E_{\mathrm{lf}}}{kT_{\mathrm{lf}}}\right) = \sum_{\Omega_{\mathrm{lf}}} \exp\left(-\frac{\lambda_{\mathrm{lf}}\chi_{\mathrm{lf}}E_{\mathrm{p}}}{kT}\right) \tag{20}$$

Here, we use Ω_{lf} to represent the low frequency conformational space, and $\Omega_{\mathrm{hf}}$to represent the high frequency conformational space.

Similarly, in the high-frequency conformational space, the equation of motion can be expressed as the difference between the instantaneous motion [Eq. (11)] and the low-frequency motion [Eq. (13)]:

$$\dot{\mathbf{p}}_i - \dot{\tilde{\mathbf{p}}}_i = \mathbf{f}_i - \tilde{\mathbf{f}}_i + \mathbf{g}_i - \tilde{\mathbf{g}}_i - \gamma_i(\mathbf{p}_i - \tilde{\mathbf{p}}_i) + \mathbf{R}_i - \tilde{\mathbf{R}}_i \tag{21}$$

The potential energy surface is approximated as the scaled high-frequency potential energy surface, $E_{\mathrm{hf}} = \lambda_{\mathrm{hf}}\left(E_{\mathrm{p}} - \tilde{E}_{\mathrm{p}}\right)$. The high-frequency energy factor, λ_{hf}, is calculated as the average projection of the total high-frequency force in the direction of the high-frequency interaction force:

$$\lambda_{\mathrm{hf}} = \frac{\left\langle \sum_i \left(\mathbf{f}_i - \tilde{\mathbf{f}}_i + \mathbf{g}_i - \tilde{\mathbf{g}}_i - \gamma_i(\mathbf{p}_i - \tilde{\mathbf{p}}_i)\right)(\mathbf{f}_i - \tilde{\mathbf{f}}_i)\right\rangle}{\left\langle \sum_i (\mathbf{f}_i - \tilde{\mathbf{f}}_i)(\mathbf{f}_i - \tilde{\mathbf{f}}_i)\right\rangle} \tag{22}$$

Again, we define the effective thermal temperature in the high-frequency conformational space as T_{hf} and assume that T_{hf} is proportional to the high-frequency temperature, $T - \tilde{T}$:

$$\frac{T_{\mathrm{hf}}}{T} = \frac{T - \tilde{T}}{T - \tilde{T}_0} \tag{23}$$

Similarly, we can calculate the high-frequency collision factor:

$$\chi_{\mathrm{hf}} = \frac{T - \tilde{T}_0}{T - \tilde{T}} = \frac{T - \chi_{\mathrm{lf}}\tilde{T}}{T - \tilde{T}} = 1 - \frac{\left\langle \sum_i \gamma_i(\mathbf{g}_i - \tilde{\mathbf{g}}_i) \cdot (\mathbf{p}_i - \tilde{\mathbf{p}}_i) \right\rangle}{\left\langle \sum_i \gamma_i^2(\mathbf{p}_i - \tilde{\mathbf{p}}_i) \cdot (\mathbf{p}_i - \tilde{\mathbf{p}}_i) \right\rangle} \tag{24}$$

Combining the scaling in the high-frequency potential energy surface and the repression of the high-frequency motion, we have the partition function in the high-frequency conformational space:

$$\Theta_{\mathrm{hf}}(N, V, T_{\mathrm{lf}}) = \sum_{\Omega_{\mathrm{hf}}} \exp\left(-\frac{E_{\mathrm{hf}}}{kT_{\mathrm{hf}}}\right) = \sum_{\Omega_{\mathrm{hf}}} \exp\left(-\frac{\lambda_{\mathrm{hf}}\chi_{\mathrm{hf}}E_{\mathrm{p}}}{kT}\right) \tag{25}$$

The overall partition function of an SGLD ensemble is the product of that in the low- and high-frequency conformational spaces:

$$\begin{aligned}\Theta(N, V, T) &= \Theta_{\mathrm{lf}}(N, V, T_{\mathrm{lf}})\Theta_{\mathrm{hf}}(N, V, T_{\mathrm{hf}}) \\ &= \sum_{\Omega_{\mathrm{lf}}} \exp\left(-\frac{E_{\mathrm{lf}}}{kT_{\mathrm{lf}}}\right) \sum_{\Omega_{\mathrm{hf}}} \exp\left(-\frac{E_{\mathrm{hf}}}{kT_{\mathrm{hf}}}\right) \\ &= \sum_{\Omega} \exp\left(-\frac{\lambda_{\mathrm{lf}}\chi_{\mathrm{lf}}E_{\mathrm{lf}} + \lambda_{\mathrm{hf}}\chi_{\mathrm{hf}}E_{\mathrm{hf}}}{kT}\right)\end{aligned} \tag{26}$$

The total conformational space is a combination of the two: $\Omega = \Omega_{\mathrm{lf}} \cdot \Omega_{\mathrm{hf}}$.

In summary, at a given temperature, T, the guiding force produces the following effects in both low- and high-frequency conformational spaces:

(a) In the low-frequency conformational space, the low-frequency energy surface, $\tilde{E}_{\mathrm{p}}$, is modified by a factor of λ_{lf}. The effective temperature is changed from T to

$$T_{\mathrm{lf}} = \frac{\tilde{T}}{\tilde{T}_0}T = \frac{T}{\chi_{\mathrm{lf}}}$$

(b) In the high-frequency conformational space, the high-frequency energy surface, $E_{\mathrm{p}} - \tilde{E}_{\mathrm{p}}$, is modified by a factor of λ_{hf}. The effective temperature is changed from T to

$$T_{\mathrm{hf}} = \frac{T - \tilde{T}}{T - \tilde{T}_0}T = \frac{T}{\chi_{\mathrm{hf}}}$$

The partition function of a canonical ensemble from an LD simulation can be related to that of an SGLD ensemble by the following equation:

$$\begin{aligned}\Theta_{\mathrm{LD}} &= \sum \exp\left(-\frac{\tilde{E}_{\mathrm{p}}}{kT} - \frac{E_{\mathrm{p}} - \tilde{E}_{\mathrm{p}}}{kT}\right) \\ &= \sum \exp\left(-\lambda_{\mathrm{lf}}\chi_{\mathrm{lf}}\frac{\tilde{E}_{\mathrm{p}}}{kT} - \lambda_{\mathrm{hf}}\chi_{\mathrm{hf}}\frac{E_{\mathrm{p}} - \tilde{E}_{\mathrm{p}}}{kT}\right) \\ &\quad \times \exp\left((\lambda_{\mathrm{lf}}\chi_{\mathrm{lf}} - 1)\frac{\tilde{E}_{\mathrm{p}}}{kT} + (\lambda_{\mathrm{hf}}\chi_{\mathrm{hf}} - 1)\frac{E_{\mathrm{p}} - \tilde{E}_{\mathrm{p}}}{kT}\right) \\ &= \Theta_{\mathrm{SGLD}}\langle w_{\mathrm{SGLD}}\rangle_{\mathrm{SGLD}}\end{aligned} \tag{27}$$

Here, w_{SGLD} is called the SGLD reweighting factor:

$$w_{\mathrm{SGLD}} = \exp\left((\lambda_{\mathrm{lf}}\chi_{\mathrm{lf}} - 1)\frac{\tilde{E}_{\mathrm{p}}}{kT} + (\lambda_{\mathrm{hf}}\chi_{\mathrm{hf}} - 1)\frac{E_{\mathrm{p}} - \tilde{E}_{\mathrm{p}}}{kT}\right) \tag{28}$$

Any ensemble average, $\langle P\rangle$, can be calculated in an SGLD simulation as

$$\langle P\rangle = \frac{\langle P w_{\mathrm{SGLD}}\rangle_{\mathrm{SGLD}}}{\langle w_{\mathrm{SGLD}}\rangle_{\mathrm{SGLD}}} \tag{29}$$

Because SGLD simulation does not change temperature, the average energy contribution to the reweighting factor can be removed:

$$\begin{aligned}w_{\mathrm{SGLD}} &= \exp\left((\lambda_{\mathrm{lf}}\chi_{\mathrm{lf}} - 1)\frac{\langle E\rangle}{kT}\right) \\ &\quad \times \exp\left((\lambda_{\mathrm{lf}}\chi_{\mathrm{lf}} - 1)\frac{\tilde{E}_{\mathrm{p}} - \langle E\rangle}{kT} + (\lambda_{\mathrm{hf}}\chi_{\mathrm{hf}} - 1)\frac{E_{\mathrm{p}} - \tilde{E}_{\mathrm{p}}}{kT}\right) \\ &= C w'_{\mathrm{SGLD}}\end{aligned} \tag{28$'$}$$

As can be seen from Eq. (28′), the reweighting factor, w'_{SGLD}, depends on the energy change, instead of the total energy, and can be calculated much more easily numerically than w_{SGLD}. The factors, λ_{lf}, λ_{hf}, χ_{lf}, and χ_{hf}, are all close to 1. Therefore, w'_{SGLD} is actually used in place of w_{SGLD} for reweighting calculation [Eq. (29)]. The reweighting factor of SGLD simulations has relatively narrower value range than other approaches such as with high-temperature simulations, which enable accurate reweighting calculation in SGLD simulations. The SGLD reweighting factor can be calculated on-the-fly during an SGLD simulation to simplify a postprocessing of a simulation trajectory.

D. Conformational Search in SGLD

In SGLD simulations, the guiding factor, λ, is a unitless input parameter whose value is often hard to decide for its lack of physical meaning. For convenience in describing the conformational search ability of an SGLD simulation, we define a self-guiding temperature, T_{sg}, based on the effective temperatures in the low- and high-frequency conformational spaces:

$$T_{\mathrm{sg}} = \frac{T_{\mathrm{lf}}}{T_{\mathrm{hf}}} T = \frac{\tilde{T}(T - \tilde{T}_0)}{\tilde{T}_0(T - \tilde{T})} T \tag{30}$$

The self-guiding temperature, T_{sg}, provides a rough measure of the conformational searching ability in the unit of temperature. An SGLD simulation with a self-guiding temperature of T_{sg} has a conformational search ability comparable to that of a high-temperature simulation at the temperature of T_{sg}. As can be seen from Eq. (30), for an LD simulation, $\tilde{T} = \tilde{T}_0$, we have $T_{\mathrm{sg}} = T$. For an SGLD simulation with $\lambda > 0$, we have $\tilde{T} > \tilde{T}_0$ and $T_{\mathrm{sg}} > T$, and with $\lambda < 0$, we have $\tilde{T} < \tilde{T}_0$ and $T_{\mathrm{sg}} < T$. T_{sg} can be used as a guidance for the choice of λ. For example, it is reasonable to choose a λ that produces $T_{\mathrm{sg}} = 2T$. However, when λ is large and T_{sg} is too large compared to T, it is difficult to obtain accurate canonical ensemble through reweighting with Eqs. (25) and (26). Therefore, λ should be chosen to balance the acceleration of conformational search and the accuracy in converting the conformational distribution.

E. Force–Momentum-Based Self-Guided Langevin Dynamics Simulation Method

SGMD utilizes the local average forces while SGLD uses the local average momentum to calculate the guiding force to achieve accelerated conformational search. These two types of guiding forces have opposite bias effect on the low-frequency energy surface. The low-frequency force, $\tilde{\mathbf{f}}_i$, favors low $\tilde{E}_{\mathrm{p}}$ states, just as normal forces do, while the low-frequency momentum, $\tilde{\mathbf{p}}_i$, favors high $\tilde{E}_{\mathrm{p}}$ states, just as high temperature does. These two types of low-frequency properties can be combined in such a way that the bias effects are cancelled.

Let us define a guiding force, $\mathbf{g}_i$, as a linear combination of $\tilde{\mathbf{f}}_i$ and $\tilde{\mathbf{p}}_i$ in the following form:

$$\mathbf{g}_i(t) = \lambda^{\mathrm{f}} \tilde{\mathbf{f}}_i(t) - \xi^{\mathrm{f}} \mathbf{f}_i(t) + \lambda_i^{\mathrm{p}} \gamma_i \left(\tilde{\mathbf{p}}_i(t) - \xi^{\mathrm{p}} \mathbf{p}_i(t) \right) \tag{31}$$

Here, λ^{f} is the force guiding factor and ξ^{f} is the force damping factor. The energy conservation factor, ξ^{p}, is calculated by the following equation to cancel the energy input from the guiding force at every time step:

$$\xi^{\mathrm{p}} = \frac{\sum_i \left(\lambda^{\mathrm{f}} \tilde{\mathbf{f}}_i - \xi^{\mathrm{f}} \mathbf{f}_i + \lambda_i^{\mathrm{p}} \gamma_i \tilde{\mathbf{p}} \right)_i \cdot \dot{\mathbf{r}}_i}{\sum_i \lambda_i^{\mathrm{p}} \gamma_i \mathbf{p}_i \cdot \dot{\mathbf{r}}_i} \tag{29}$$

The low-frequency energy factor now becomes

$$\begin{aligned}
\lambda_{\mathrm{lf}} &= 1 + \frac{\left\langle \sum_i (\tilde{\mathbf{g}}_i - \gamma_i \tilde{\mathbf{p}}_i)\, \tilde{\mathbf{f}}_i \right\rangle}{\left\langle \sum_i \tilde{\mathbf{f}}_i \tilde{\mathbf{f}}_i \right\rangle} \\
&= 1 + \frac{\left\langle \sum_i \left\langle \lambda^{\mathrm{f}} \tilde{\mathbf{f}} - \xi^{\mathrm{f}} \mathbf{f}_i + \lambda_i^{\mathrm{p}} \gamma_i \, (\tilde{\mathbf{p}}_i - \xi^{\mathrm{p}} \mathbf{p}_i) \right\rangle_{\mathrm{L}} - \gamma_i \tilde{\mathbf{p}}_i) \tilde{\mathbf{f}}_i \right\rangle}{\left\langle \sum_i \tilde{\mathbf{f}}_i \tilde{\mathbf{f}}_i \right\rangle} \\
&\approx 1 + \lambda^{\mathrm{f}} - \xi^{\mathrm{f}} + \frac{\left\langle \sum_i \left(\tilde{\mathbf{g}}_i^{\mathrm{p}} - \gamma_i \tilde{\mathbf{p}}_i \right) \tilde{\mathbf{f}}_i \right\rangle}{\left\langle \sum_i \tilde{\mathbf{f}}_i \tilde{\mathbf{f}}_i \right\rangle} \\
&= \lambda_{\mathrm{lf}}^{\mathrm{p}} + \lambda^{\mathrm{f}} - \xi^{\mathrm{f}}
\end{aligned} \tag{32}$$

The high-frequency energy factor is

$$\begin{aligned}
\lambda_{\mathrm{hf}} &= 1 + \frac{\left\langle \sum_i (\mathbf{g}_i - \tilde{\mathbf{g}}_i - \gamma_i (\mathbf{p}_i - \tilde{\mathbf{p}}_i))\, (\mathbf{f}_i - \tilde{\mathbf{f}}_i) \right\rangle}{\left\langle \sum_i (\mathbf{f}_i - \tilde{\mathbf{f}}_i)(\mathbf{f}_i - \tilde{\mathbf{f}}_i) \right\rangle} \\
&= 1 + \frac{\left\langle \sum_i \left(\lambda^{\mathrm{f}} \langle \mathbf{f}_i - \tilde{\mathbf{f}}_i \rangle_{\mathrm{L}} - \xi^{\mathrm{f}} (\mathbf{f}_i - \tilde{\mathbf{f}}_i) + (\mathbf{g}_i^{\mathrm{p}} - \tilde{\mathbf{g}}_i^{\mathrm{p}}) - \gamma_i (\mathbf{p}_i - \tilde{\mathbf{p}}_i) \right) (\mathbf{f}_i - \tilde{\mathbf{f}}_i)_i \right\rangle}{\left\langle \sum_i (\mathbf{f}_i - \tilde{\mathbf{f}}_i)(\mathbf{f}_i - \tilde{\mathbf{f}}_i) \right\rangle} \\
&\approx 1 - \xi^{\mathrm{f}} + \frac{\left\langle \sum_i \left(\mathbf{g}_i^{\mathrm{p}} - \tilde{\mathbf{g}}_i^{\mathrm{p}} - \gamma_i (\mathbf{p}_i - \tilde{\mathbf{p}}_i) \right) \left(\mathbf{f}_i - \tilde{\mathbf{f}}_i \right) \right\rangle}{\left\langle \sum_i (\mathbf{f}_i - \tilde{\mathbf{f}}_i)(\mathbf{f}_i - \tilde{\mathbf{f}}_i) \right\rangle} \\
&= \lambda_{\mathrm{hf}}^{\mathrm{p}} - \xi^{\mathrm{f}}
\end{aligned} \tag{33}$$

Here, we use $\lambda_{\mathrm{lf}}^{\mathrm{p}}$ and $\lambda_{\mathrm{hf}}^{\mathrm{p}}$ to represent the momentum-based energy factors calculated using Eqs. (13) and (22). $\lambda_{\mathrm{lf}}^{\mathrm{p}}$ and $\lambda_{\mathrm{hf}}^{\mathrm{p}}$, as well as χ_{lf}, and χ_{hf}, are calculated during simulations as long-time evolving averages [13]. We can set λ^{f} and ξ^{f} during a simulation in such a way

$$\xi^{\mathrm{f}} = \lambda_{\mathrm{hf}}^{\mathrm{p}} - \frac{1}{\chi_{\mathrm{hf}}} \tag{34}$$

and

$$\lambda^{\mathrm{f}} = \frac{1}{\chi_{\mathrm{lf}}} - \frac{1}{\chi_{\mathrm{hf}}} - \lambda_{\mathrm{lf}}^{\mathrm{p}} + \lambda_{\mathrm{hf}}^{\mathrm{p}} \tag{35}$$

so that

$$\lambda_{\mathrm{lf}} \chi_{\mathrm{lf}} = 1 \tag{36}$$

$$\lambda_{\mathrm{hf}} \chi_{\mathrm{hf}} = 1 \tag{37}$$

and we have

$$
\begin{aligned}
\Theta_{\mathrm{SGLDfp}} &= \sum \exp\left(-\lambda_{\mathrm{lf}}\chi_{\mathrm{lf}}\frac{\tilde{E}_{\mathrm{p}}}{kT} - \lambda_{\mathrm{hf}}\chi_{\mathrm{hf}}\frac{E_{\mathrm{p}} - \tilde{E}_{\mathrm{p}}}{kT}\right) \\
&= \sum \exp\left(-\frac{\tilde{E}_{\mathrm{p}}}{kT} - \frac{E_{\mathrm{p}} - \tilde{E}_{\mathrm{p}}}{kT}\right) \\
&= \Theta_{\mathrm{LD}}
\end{aligned}
\tag{38}
$$

From above equations, we can see that by using a guiding force with balanced local average force components as shown in Eqs. (34) and (35), we can directly obtain an unbiased conformational distribution. Therefore, an ensemble average property can be directly calculated from an SGLDfp simulation:

$$\langle P \rangle = \langle P \rangle_{\mathrm{SGLDfp}}$$

With such a direct approach, the sampled conformation can be directly used for computing ensemble averages, such as free energy. As such, SGLDfp equation of motion can directly replace MD or LD for any nondriven degree of freedom in a rather unbiased manner. For example, the generalized ensemble (GE) methods [70–72] are enabled via the free energy flattening (or effectively flattening) treatment. Therefore, these methods intrinsically suffer from the diffusion sampling problem [70, 71]. Complementary to the GE strategy, SGLD or SGLDfp improves the sampling by improving local diffusion. One can naturally expect that the combination of the SGLD or SGLDfp method and the efficient GE method such as the orthogonal space random walk method will lead to significant sampling improvement; this expectation should be especially true when the collective variables associated with a large number of degrees of freedom such as essential energy [70, 73] or generalized force [72, 74] are employed.

Details of the simulation algorithms of SGMD [10, 11, 60, 61], SGLD [12, 14], and SGLDfp [13] have been reported previously. SGLD is available in CHARMM [75, 76] version 32 and later, as well as Amber 9 or later [77a (Case et al. 2006)]. SGLD reweighting and SGLDfp have been implemented into CHARMM version 36 and will be available in Amber 12. Because SGLD and SGLDfp simulations involve extra calculation only in the propagation of the equation of motion compared to normal LD simulation, the cost of SGLD and SGLDfp simulation is almost identical to an LD simulation for the same number of time steps. SGLD and SGLDfp simulations do keep more arrays in memory because of the need to store the guiding forces, as well as some arrays for the weighting factor calculation.

To run an SGLD or SGLDfp simulation, one can either set λ (or λ_i^{p} for SGLDfp) or set a target self-guiding temperature, T_{sg}^0, defined by Eq. (30). When T_{sg}^0 is set,

λ_i^{p} is adjusted in such a way

$$\lambda_i^{\mathrm{p}}(t) = \lambda_i^{\mathrm{p}}(t - \delta t) + \frac{\delta t}{t_{\mathrm{est}}} \frac{T_{\mathrm{sg}}^0 - T_{\mathrm{sg}}}{T} \tag{39}$$

so that T_{sg} will approach T_{sg}^0. Here, t_{est} is a response time to mantain the convergence of the estimation. Typically, we set test $> 10t_{\mathrm{L}}$. λ^{f} and ξ^{f} will be calculated according to Eqs. (34) and (35) in the same way as when λ_i^{p} is set. Because T_{sg} is a derived quantity, its value range is limited by the simulation temperature, system size, and other SGLD parameters, T_{sg}^0 must be set close to the simulation temperature to produce a converged λ_i^{p}. For example, one may set $T_{\mathrm{sg}}^0 = 1.2T$ for an SGLDfp simulation. To achieve an optimal performance, it may be necessary to briefly explore various SGLD parameters to find an optimal set of parameters for a particular system.

IV. CHARACTERISTICS OF THE SELF-GUIDED LANGEVIN DYNAMICS

We use several model systems to demonstrate the nature and explain the characteristics of SGLD simulations. The model systems we choose are a skewed double well system, argon liquid, an alanine dipeptide, and a pentamer peptide. Through these model systems, we demonstrate the effect of the guiding force on kinetic energy and potential energy distributions, low- and high-frequency motion, and energy barrier crossing ability. In addition, we examine how the low- and high-frequency properties change with the guiding factor λ, the local averaging time t_{L}, and the collision frequency γ. Because only λ_i^{p} is set and is the same for all particles, in the following description, the guiding factor, λ, refers to the momentum guiding factor λ_i^{p}.

A. The Skewed Double Well System

A skewed double well system represents the simplest system with an energy barrier to cross. This system has only one particle and the particle moves on a fixed energy surface. The skewed double well potential energy (in kcal mol^{-1}) has the following form:

$$\varepsilon_{\mathrm{p}}(x, y, z) = \varepsilon_{xz}(x, z) + \varepsilon_y(y) = \left(500(x^2 + z^2)\right) + \left(y^2(y-2)^2 + 0.25y\right) \tag{40}$$

Figure 2 shows the energy surface of this double well potential. This energy surface is designed in such a way that it restricts the particle to move near the y-axis with two energy minima of different depths, -0.0038 and 0.4960 kcal mol^{-1}, along the y-axis at (0, -0.0299, 0) and (0, 1.9672, 0), respectively. The potential is symmetric around y-axis with a strong dependence on the distance from y-axis,

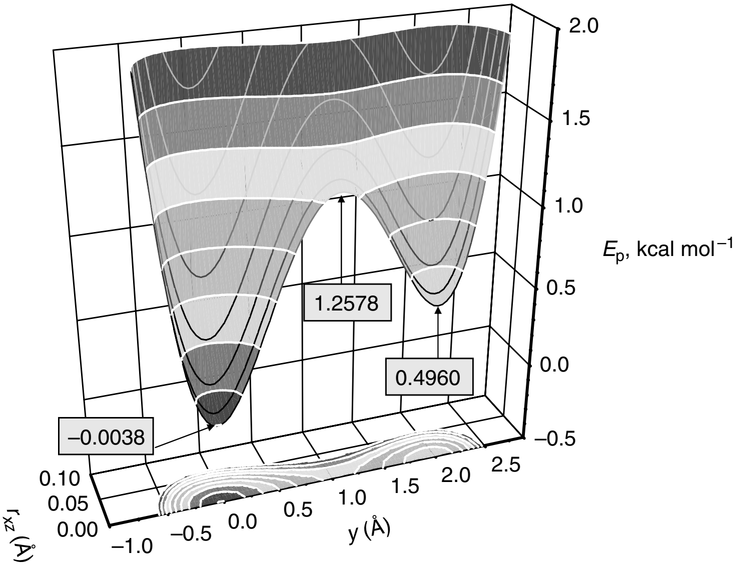


Figure 2. The energy surface of the skewed double well potential. The potential is symmetric around the y-axis and $r_{xz} = \sqrt{x^2 + z^2}$ is the distance to the y-axis.

$r_{xz} = \sqrt{x^2 + z^2}$. The minimum transition energy from one well to the other well is 1.2578 kcal mol^{-1} at (0, 1.0627, 0) between the two wells. Such a design forces the particle to have a high-frequency motion in the x–z direction and a low-frequency motion in the y direction.

An argon atom was used to represent the particle. Simulations were carried out with a local averaging time, $t_L = 0.2$ ps. A time step of 1 fs was used and the simulation length was 1 μs for each simulation. The collision frequency was 10 ps^{-1} except noted otherwise. To help illustrate the guiding force effect, we used a large range of the guiding factor, up to $\lambda = 2$.

Kinetic energy is transferred from high- to low-frequency degrees. This double well system has only three degrees of freedom in x, y, and z directions. In the y direction, the atom has low-frequency motion, while in the x and z directions, it has high-frequency motion. We calculate the temperature components based on its velocity components to examine the kinetic energy changes with the guiding force. Figure 3a and b shows the kinetic energies in the low- (along the y-axis) and high-frequency directions (perpendicular to the y-axis or along the x–z plane) as function of y and r_{xz}, respectively. The top and bottom panels of Fig. 3a show the y and x–z components of temperature as functions of the y coordinate. At $\lambda = 0$, T_y and T_{xz} are almost constant throughout the accessible y coordinate range. Large fluctuations are observed around the energy barrier $y = 1$ Å due to the poor

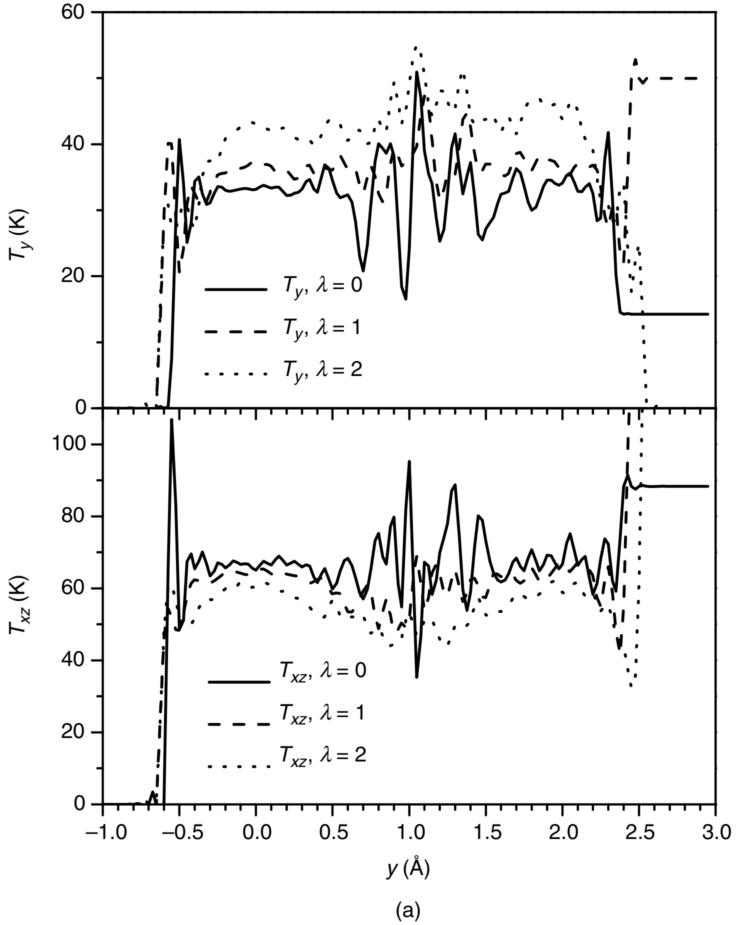
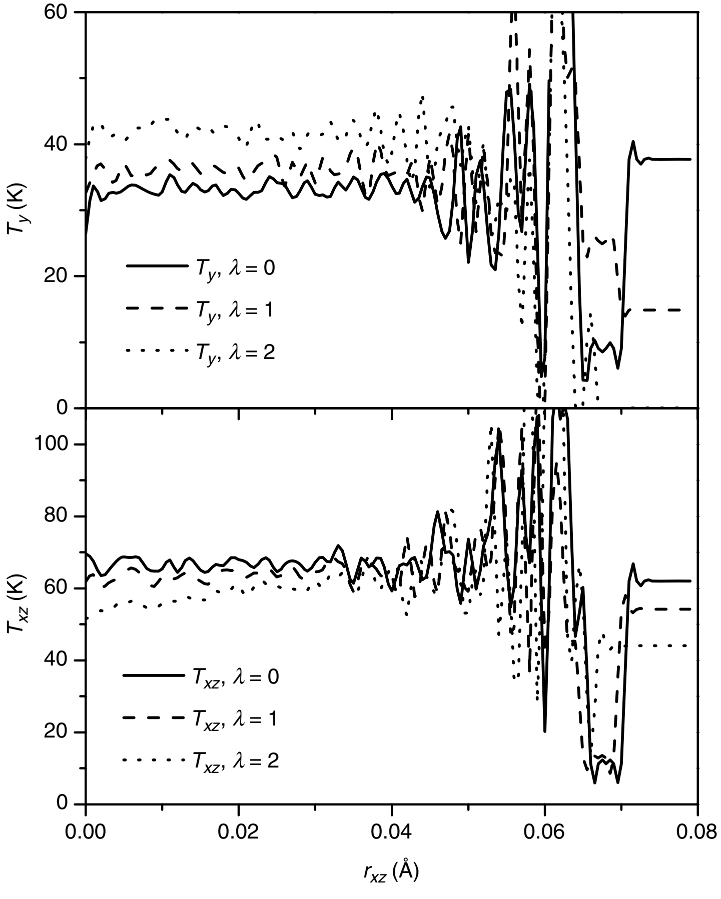


Figure 3. Temperature distribution of the skewed double well system. (**a**) Temperature in y-coordinates. (**b**) Temperature in r_{xz}.

sampling in this region. As λ increases, T_y increases, while T_{xz} decreases. The changes in T_y and T_{xz} are not uniform. Larger changes can be seen in the energy barrier region than in the well regions. This result explains why the guiding force helps energy barrier crossing. The guiding force pumps kinetic energy from high-frequency degrees of freedom to low-frequency degrees of freedom to overcome energy barriers, and the higher the energy barrier, the more the kinetic energies transferred. The kinetic energy transfer can also be seen in Fig. 3b in the x–z coordinate range, but in this high-frequency coordinate range, more kinetic energy transfer is observed in low-energy region (smaller r_{xz}). Once the barrier is crossed,

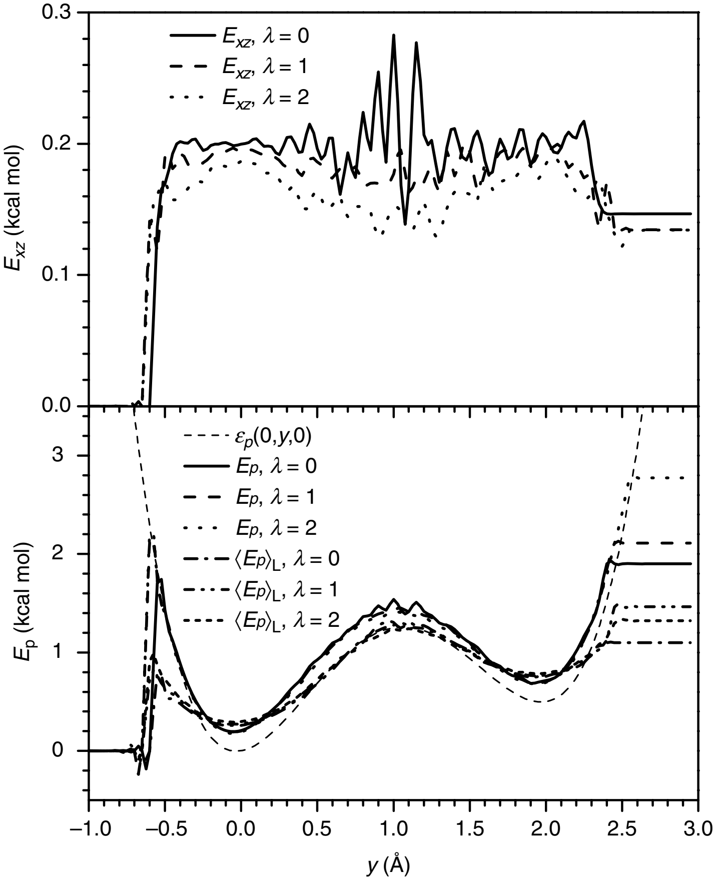


(b)

Figure 3. (*Continued*)

the excess kinetic energy in the low-frequency motion is returned to the high-frequency degrees of freedom in a nonthermostat manner. The overall effect can be seen as "energy borrowing."

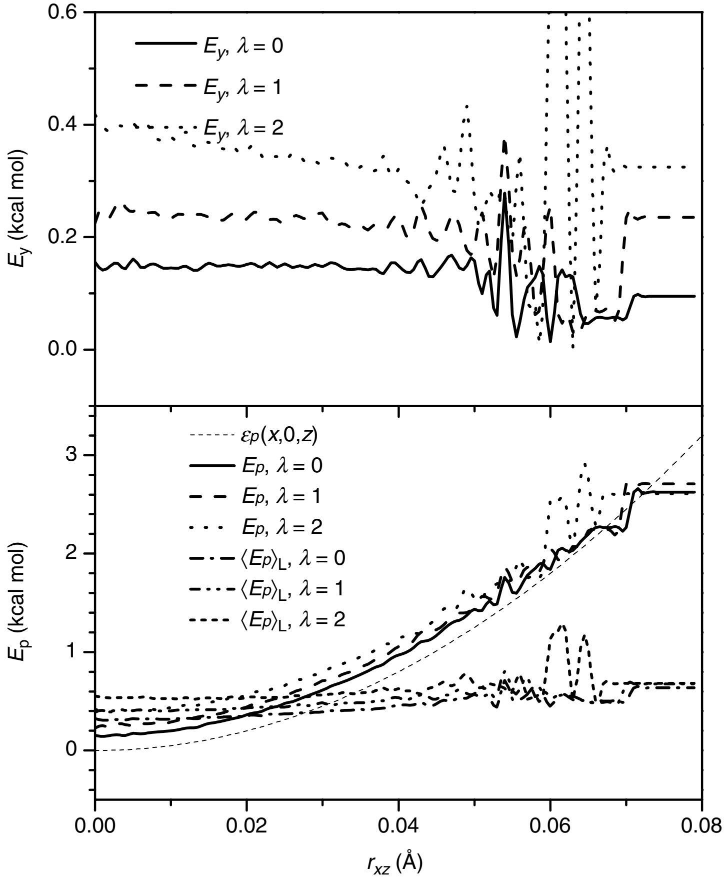
The guiding force favors low-potential energy region in the high-frequency degrees of freedom and high potential energy in the low-frequency degrees of freedom. Figure 4 shows the average potential energy and its components as functions of the coordinates. Figure 4a shows the energies along the y coordinates. The top panel of Fig. 4a shows the x–z component of the total energy, Exz, which represents the high-frequency portion. In LD simulation ($\lambda = 0$), Exz is almost flat throughout the accessible y coordinate range except a large fluctuation around



(a)

Figure 4. Potential energies of the skewed double well system. (**a**) Average potential energies in y-coordinates. (**b**) Average potential energies in x–z direction.

the barrier region (around $y = 1$ Å) due to poor sampling. When λ increases, Exz decreases and it decreases more at the energy barrier region. The fluctuation in the energy barrier region becomes much smaller because of the improved sampling in the SGLD simulation. The bottom panel of Fig. 4a shows the y component, $E_y = \varepsilon_y(y) = \varepsilon_p(0, y, 0)$, and the total potential energy, E_p, as well as the low-frequency potential energy, $\tilde{E}_p$. The total potential energy is the sum of the y component and the x–z component shown in the top panel: $E_p = E_{xz} + E_y$. E_y depends only on the y coordinate and will not change with λ. Even though, we can see from Fig. 4a that the accessible y range increases with λ, indicating that higher energy states are reached with larger λ.



(b)

Figure 4. (*Continued*)

Comparing E_p and $\tilde{E}_p$ in Fig. 4a, we can see that $\tilde{E}_p$ has smaller energy barrier than E_p. Low-frequency energy surface tends to have lower energy barriers. In other words, the low-frequency energy surface is smoother than the original energy surface. Enhanced motion in the low-frequency energy surface can be more efficient to cross energy barrier than that in the original energy surface.

Figure 4b shows the average potential energy and its components at different r_{xz}. From the top panel of Fig. 4b, we can see that in LD simulation ($\lambda = 0$), E_y is almost flat and in SGLD simulations, E_y increases with λ and increases more for smaller r_{xz}. The lower panel of Fig. 4b shows that $\tilde{E}_p$ is almost flat, indicating that high-frequency energies are averaged out in the local averaging process.

Overall, the guiding force accelerates the low-frequency motion while it slows down the high-frequency motion. As a result, the simulation has enhanced ability to overcome energy barriers in the low-frequency conformational space while making high-frequency states more stable. These features contrast SGLD against high-temperature simulations. SGLD can preserve high-frequency structures while enhancing conformational search in the low-frequency conformational space. High-temperature simulation will destabilize all structures.

Another important parameter for SGLD simulations is t_L. It is used to define the low-frequency property and the high-frequency property through the evolving averaging [Eq. (5)]. The choice of t_L will affect which motions will be enhanced and which motions will be suppressed. A larger t_L will result in more motion falling into the high-frequency motion category and less into the low-frequency motion category, which is demonstrated in Fig. 5. The low-frequency temperature, $\tilde{T}$, accounts

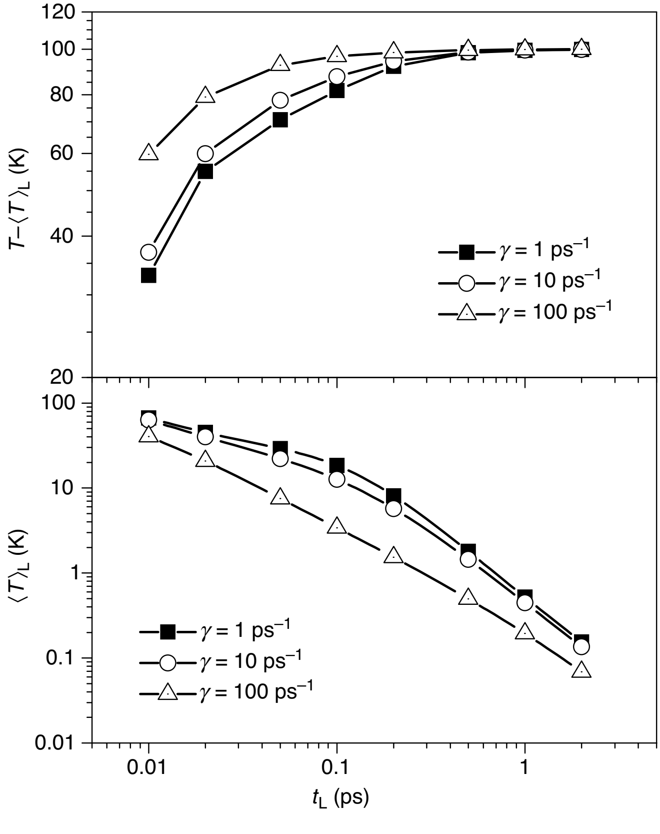


Figure 5. The low- and high-frequency temperatures of the skewed double well system at different local averaging times.

for the kinetic energy of the low-frequency motion. As t_{L} increases, $\tilde{T}$ decreases (lower panel of Fig. 5), while the high-frequency temperature, $T - \tilde{T}$, increases (top panel of Fig. 5). Low-frequency temperature decreases with the collision frequency. Figure 5 also shows the effect of collision frequency. As γ increases, $\tilde{T}$ decreases. This is because an increase in γ will increase the friction force, which will suppress more low-frequency motion than high-frequency motion.

Now let us examine the conformational search ability of SGLD and SGLDfp simulations. Figure 6 shows the trajectories of the particle in the LD, SGLD, and

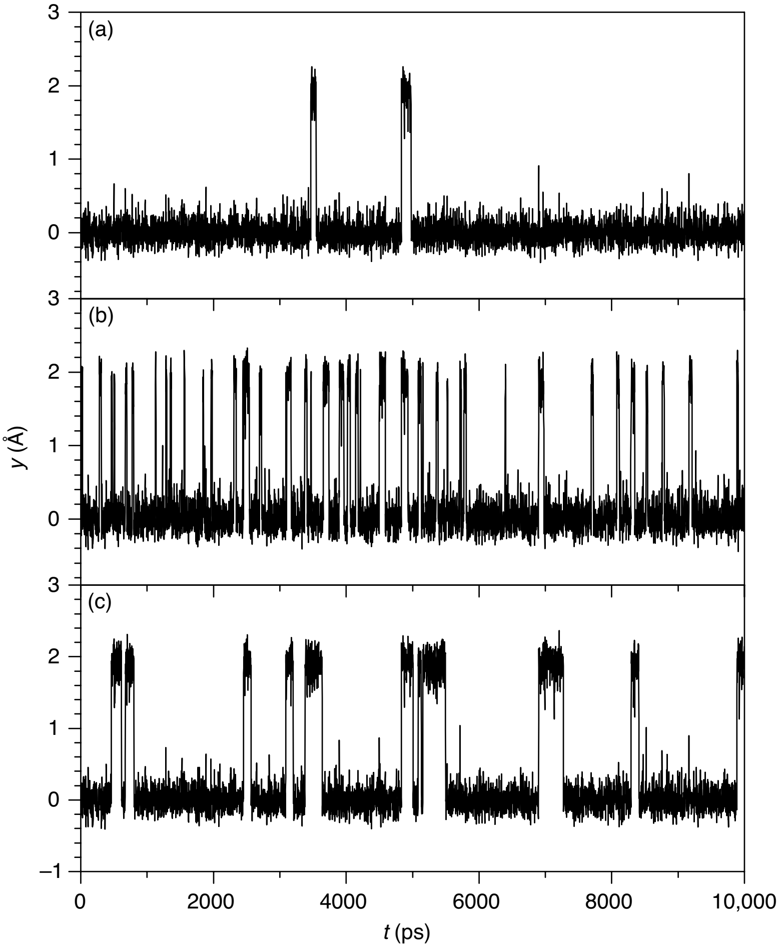


Figure 6. Trajectories of the particle on the double well potential. (**a**) LD simulation. (**b**) SGLD simulation with $\lambda = 1$. (**c**) SGLDfp simulation with $\lambda = 1$. The collision frequency is 10 ps^{-1} and temperature is 80K.

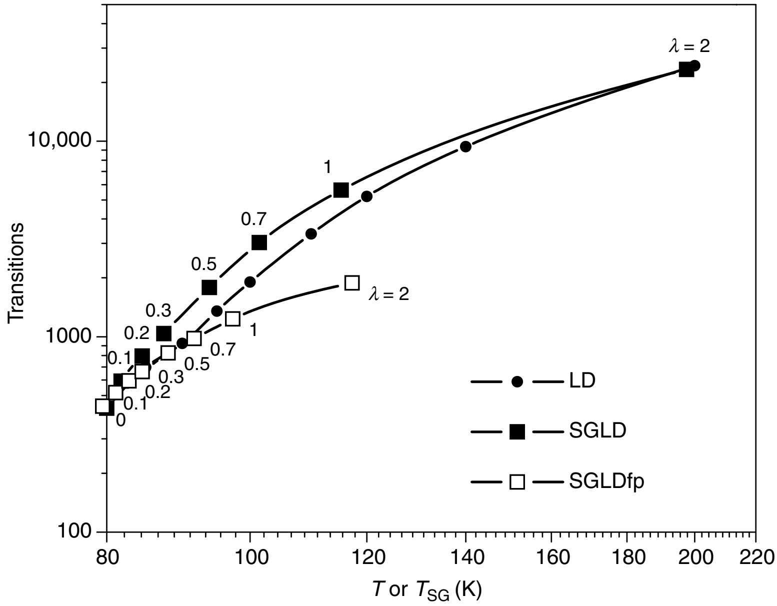


Figure 7. Transitions of the particle on the double well potential in high-temperature LD simulations and in SGLD or SGLDfp simulations. x-axis is T for the LD simulations and is the self-guiding temperature, T_{sg}, for the SGLD or SGLDfp simulations. The guiding factors are labeled in the plot and the temperature is 80K for the SGLD or SGLDfp simulations. The collision frequency is 10 ps^{-1} and the transition count starts with 1. The simulation length is 1000 ns.

SGLDfp simulations. Both SGLD and SGLDfp simulations were run with $\lambda = 1$. Clearly, both SGLD and SGLDfp simulations increased transition rates compared to the LD simulation. However, the SGLDfp simulation shows fewer transitions than the SGLD simulation due to the inclusion of a force-based guiding force to preserve the canonical ensemble.

The self-guiding temperature, T_{sg}, is introduced to describe the conformational search ability [14]. Figure 7 compares the transition rate in high-temperature LD simulations as a function of temperature and in SGLD or SGLDfp simulations as a function of T_{sg}. The guiding factor, λ, is labeled for each data point of the SGLD and SGLDfp simulations. The transition rate increases with T in the LD simulations and increases with T_{sg} in the SGLD or SGLDfp simulations. Even though the curves show different change rates with T or T_{sg}, they demonstrate that T_{sg} in the SGLD or SGLDfp simulations roughly reflect the transition rate of the LD simulations with $T \approx T_{sg}$, especially when λ is small. The purpose of introducing T_{sg} is to provide a measurement of conformational search ability with certain physical meaning. It should be noted that an LD simulation at $T \approx T_{sg}$ is very different from an SGLD or SGLDfp simulation with a self-guiding temperature of T_{sg}. The major difference is that SGLD and SGLDfp simulations are performed at a temperature of interest that normally is lower than T_{sg}. The conformational distribution and energy distribution of SGLD simulations are very much closer to that of LD simulations at the same temperature, rather than a high temperature where LD at $T \approx T_{\mathrm{sg}}$, while the distributions of SGLDfp simulations are the same as that of LD simulations at the lower temperature.

From Fig. 7, the difference between SGLD and SGLDfp simulations can be seen. In these simulations, the transition rates in both SGLD and SGLDfp simulations increase with T_{sg}, which depends on λ. At the same λ, an SGLD simulation has higher T_{sg} than an SGLDfp simulation. Even at the same T_{sg}, SGLD simulation has a higher transition rate. In the SGLD simulation with $\lambda = 1$, the transition rate is about 13 times that of the LD simulation (i.e., $\lambda = 0$). However, in the SGLDfp simulation with $\lambda = 1$, the transition rate is only 2.9 times the LD rate. SGLDfp shows a reduced enhancement in energy barrier crossing compared to the SGLD simulations, especially when λ is large. Therefore, the preservation of conformational distribution without reweighting comes at a cost of the reduced enhancement in conformational searching.

The collision frequency, γ, in Langevin dynamics plays an important role in representing a thermostatic environment. Through this skewed double well system, we can examine its effect on SGLD and SGLDfp simulations.

We performed a series of SGLD and SGLDfp simulations with $\lambda = 1$ at various γ and T, and the transition rates are shown in Fig. 8. The collision frequency controls the diffusion and the temperature corresponds to relative energy barrier heights. At $T = 100$, 60, and 40K, the average y energies are 0.152, 0.0793, and 0.0561 kcal mol^{-1}, respectively. In kT scale, the energy differences between the global minimum and the transition barrier are $6.35kT$, $10.58kT$, and $15.87kT$, and the relative barrier heights from the average y energies to the transition barrier are $5.56kT$, $9.89kT$, and $15.1kT$ at $T = 100$, 60, and 40K, respectively.

In Fig. 8, we can see that the transition rates of LD simulations decrease with γ at all temperatures. For the convenience of plotting, the transition count starts with 1. A transition value of 1 means the particle has never crossed the energy barrier. As can be seen in Fig. 8, at 40K, LD cannot overcome the energy barrier during the simulation time with a collision frequency larger than 10 ps^{-1}. Higher γ reduces diffusion and slows down all events in LD simulations, regardless of their energy barriers. The transition rates of both SGLD and SGLDfp simulations are higher than those in the LD simulations, demonstrating that SGLD and SGLDfp can enhance the barrier crossing and diffusion. The difference between SGLD and LD or between SGLDfp and LD increases as γ increases, indicating that the larger the friction force, the more acceleration the SGLD and SGLDfp will have. Comparing the SGLD and SGLDfp simulations, we can see that the SGLDfp simulations have much fewer transitions than the SGLD simulations. This result

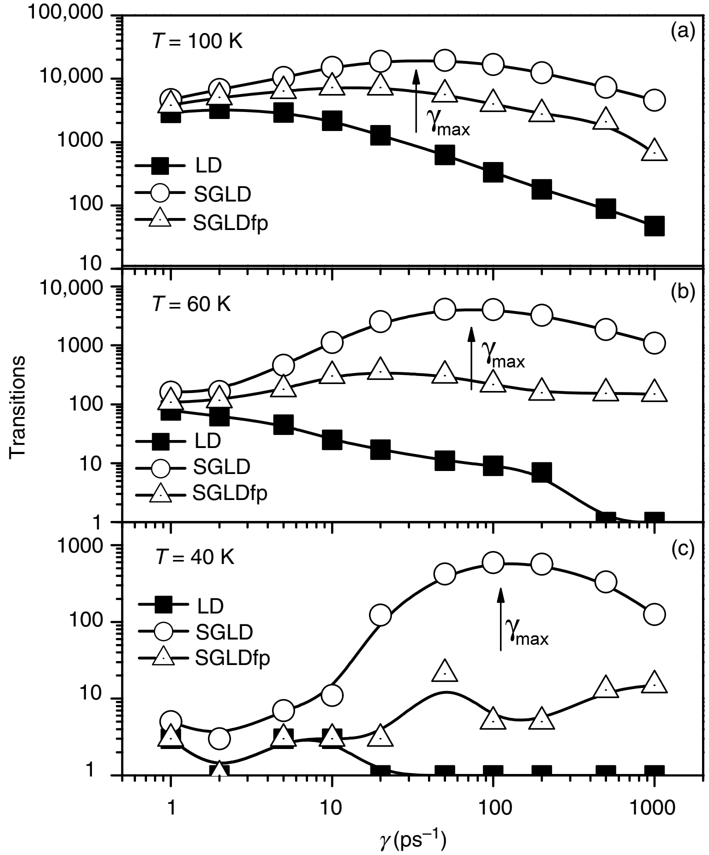


Figure 8. Transitions of the particle on the skewed double well potential at different collision frequencies and temperatures. A guiding factor of $\lambda = 1$ was used for all simulations. The transition count starts with 1. The simulation length is 1000 ns. The energy barrier heights from the y average energies are (**a**) $5.56kT$ at $T = 100$K; (**b**) $9.89kT$ at $T = 60$K; (**c**) $15.1kT$ at $T = 40$K.

indicates that SGLDfp sacrifices the enhancement in conformational search to maintain correct conformational distribution. When γ approaches zero, the low-frequency force and the low-frequency momentum become highly correlated and the guiding effect approaches zero in the SGLDfp simulations. As can be seen from Fig. 8, the SGLDfp simulations have similar transition rates as the LD simulations when γ is small. This result means that SGLDfp performs better at larger γ.

In Fig. 8, there is a maximum in the transition rate at each temperature in the SGLD and SGLDfp simulations. Before the maximum collision frequency, $\gamma_{\max}$, the transition rate increases with γ and after that the transition rate decreases with γ. This is because as γ increases, the guiding force increases and the low-frequency motion is enhanced. An increase in low-frequency motion, combined with the increase in γ, will lead to an increase in friction forces. At $\gamma_{\max}$, the guiding effect is balanced by the friction effect. When $\gamma > \gamma_{\max}$, the slow down effect by the friction force surpasses the guiding effect and brings down the transition rate.

The value of $\gamma_{\max}$ depends on the energy barrier. A higher energy barrier will result in a larger $\gamma_{\max}$. Comparing the transitions at different temperatures, $\gamma_{\max}$ shifts up when temperature decreases. For SGLD simulations, at 100K, $\gamma_{\max}$ is between 20 and 50 ps^{-1}, while at 60K, $\gamma_{\max}$ is between 50 and 100 ps^{-1}, and at 40K, $\gamma_{\max}$ is between 100 and 200 ps^{-1}. At a lower temperature, energy barrier becomes a more dominant factor for the transition and the low-frequency motion is slower, making $\gamma_{\max}$ larger before the guiding effect is balanced by the friction force. Also, we can see in Fig. 8 that in the SGLDfp simulations, $\gamma_{\max}$ values are always higher than those in the SGLD simulations. This is because SGLDfp has less energy barrier crossing ability than SGLD, which delays the maximum collision frequency where the guiding effect is balanced by the friction effect. Figure 8 also demonstrates that SGLD and SGLDfp can overcome energy barriers as high as $15kT$ (at 40K) with reasonable transition rates where no LD transition is observed. Even at 30K (corresponding to an energy barrier of $20kT$), we observed up to 100 transitions in the SGLD simulations (data not shown).

For macromolecular systems with a wide variety of barrier heights, a consensus value of γ needs to be used. Within CHARMM, different γ values can be applied to each atom, so that each part of a macromolecular system can be optimally enhanced. For example, the γ parameters that maximize the diffusion constant of water are different from those that maximally enhance protein side-chain transitions.

Figure 9 shows the conformational search ability as measured by the self-guiding temperature, T_{sg} (top panel), and by the transition rate (lower panel) as functions of t_{L}. All simulations were performed at 100K and are 1 ms in length. As can be seen in both panels, there is an optimal t_{L} at each γ. The optimal t_{L} increases as γ increases. The optimal t_{L} depends on the frequency of the barrier crossing motion. A large γ will slow down the crossing motion, making the optimal t_{L} larger. Based on the transition rates, the optimal t_{L} is 0.03, 0.1, and 0.2 ps for $\gamma = 1$, 10, and 100 ps^{-1}, respectively. Comparing T_{sg} and the transition rate in Fig. 9, we can see that T_{sg} correlates with the transition rate fairly well, again validating the use of T_{sg} to measure conformational search ability.

We examine the ensemble distributions from the SGLD and SGLDfp simulations at 80K with different guiding factors (Fig. 10). The average y energy of the system at 80K is 0.107 kcal mol^{-1}. The energy barrier height from the average y energy to the transition energy is $7.24kT$ and the energy difference between the two wells is $3.14kT$. Figure 10 compares the potential energy distributions in the

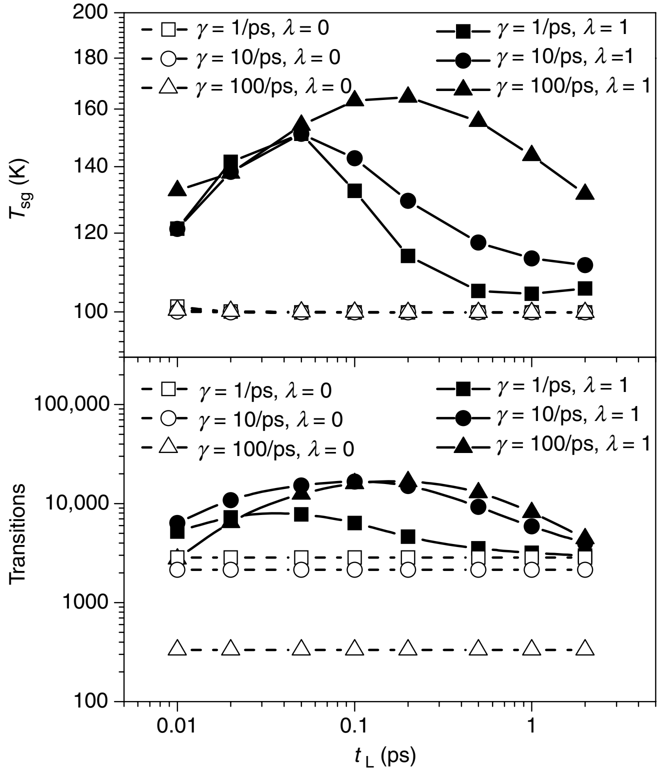


Figure 9. SGLD simulations of the double well system at different local averaging times t_L. Upper panel shows the self-guiding temperatures and lower panel shows the transitions crossing the energy barrier. The simulations are performed at $T = 100$K and are 1000 ns in length.

SGLD simulations and the SGLDfp simulations. In SGLD simulations, as can be seen in Fig. 10a, as λ increases, the distribution decreases in the low-energy region and increases in the high-energy region. Figure 10b shows the reweighted energy distributions [14]. Clearly, all curves converge fairly well to the one with $\lambda = 0$, except when the guiding factor is very large, $\lambda = 2$, indicating the weighting scheme can convert the SGLD distributions to the canonical distribution. Figure 10c shows the results from the SGLDfp simulations. The densities at different guiding factors converge together, even with $\lambda = 2$, proving that the SGLDfp simulations preserve the energy distribution to a reasonable accuracy.

To further demonstrate the preservation in conformational distribution in SGLDfp simulations, we plot the conformational density as a function of the y coordinate in Fig. 11. Figure 11a shows the distributions from SGLD simulations

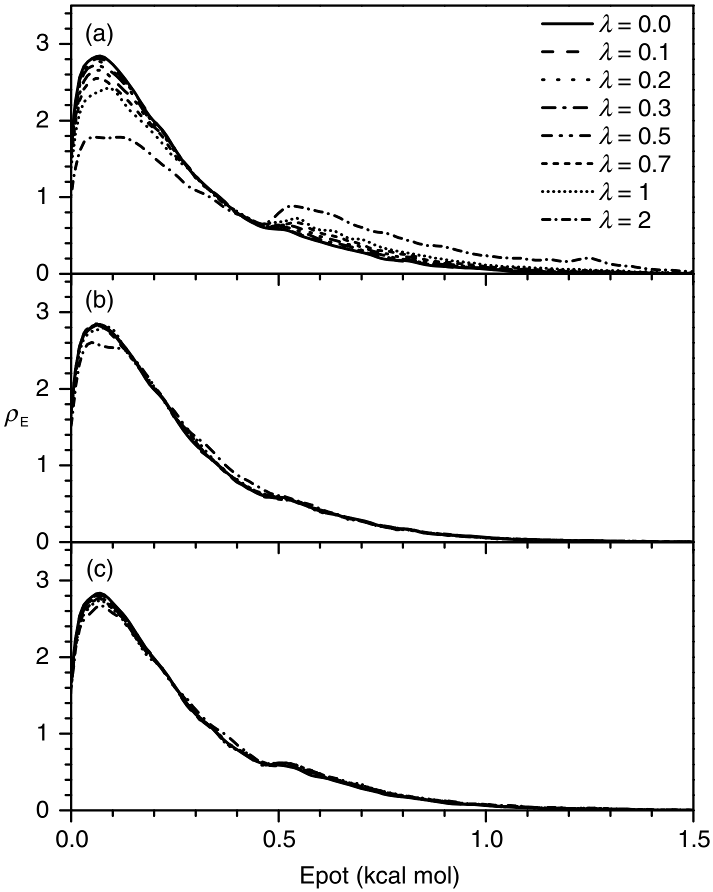


Figure 10. The potential energy distributions of the double well system. (**a**) SGLD unweighted. (**b**) SGLD reweighted. (**c**) SGLDfp. The collision frequency is 10 ps^{-1} and temperature is 80K.

at different guiding factors. There are two peaks with different heights, corresponding to the two skewed double wells. Examining the peak heights at different λ values, we can see that as λ increases, the left peak (the higher peak) decreases, while the right peak (the lower peak) grows. Figure 11b shows the reweighted conformational distributions of the SGLD simulations. All distributions converge fairly well to the one with $\lambda = 0$, except when $\lambda = 2$, validating the weighting scheme. The SGLDfp results are shown in Fig. 11c. The densities at different guiding factors almost overlap with each other, except when $\lambda = 2$, proving that the SGLDfp simulation preserves the conformational distribution well. When the

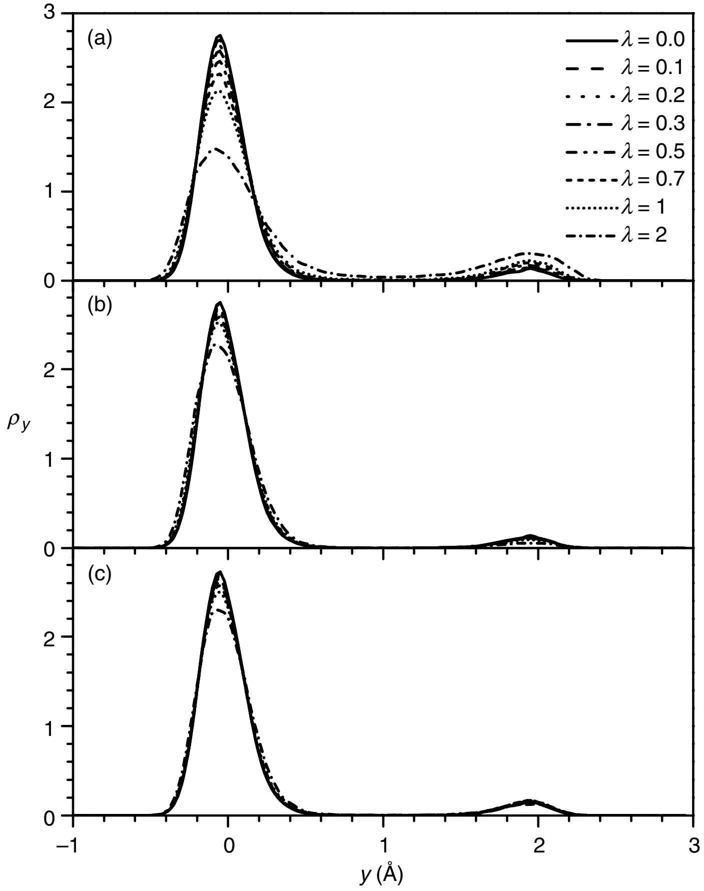


Figure 11. The y-coordinate distributions of the double well system. (**a**) SGLD unweighted; (**b**) SGLD reweighted; (**c**) SGLDfp. The collision frequency is 10 ps^{-1} and temperature is 80K.

guiding factor is too large, here, $\lambda = 2$, the perturbation of the momentum-based guiding force is too large to be described by the reweighting factor or be compensated by the force-based guiding force. These results indicate that $\lambda \leq 1$ is the recommended guiding factor range for SGLD reweighting or SGLDfp simulation. This finding is independent of the integration time step of a simulation.

To quantitatively compare the LD result and SGLD and SGLDfp results, we plot the root mean square deviations (RMSD) of the SGLD and SGLDfp distributions from the LD result in Fig. 12. The upper panel and lower panels of Fig. 12 show the RMSDs of the energy distributions, $\delta\rho_{\mathrm{E}}$, and the RMSDs of the y distributions,

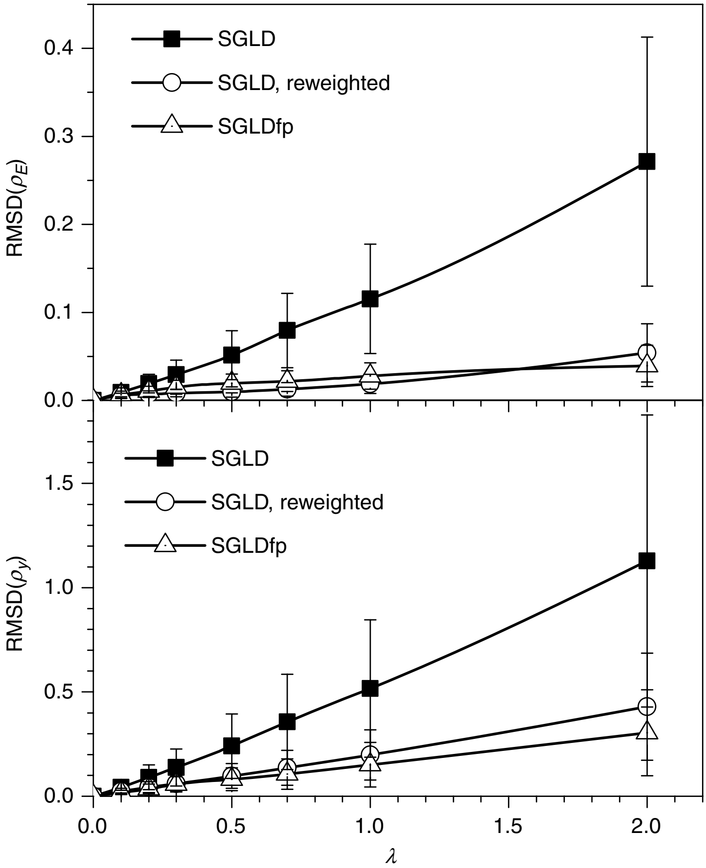


Figure 12. Root mean square deviations of the SGLD and SGLDfp distributions from the LD distributions. The upper panel shows the deviations in the potential energy distributions (kcal mol^{-1}) from Fig. 10 and the lower panel shows the deviations in the y distributions from Fig. 11.

$\delta\rho_y$, for the SGLD simulations before and after reweighting, as well as for the SGLDfp simulations. The SGLDfp distributions, as well as the reweighted SGLD distributions, show much reduced deviations from the LD distribution than the SGLD distributions. For this system, the SGLDfp distributions and the reweighted SGLD distributions have similar deviations from the LD distributions. The RMSD increase with λ in both the reweighted SGLD result and the SGLDfp result is likely due to statistical noise that increases with the guiding force and the approximation made in separating high- and low-frequency motion. A more detailed discussion of reweighting accuracy in simulation can be found elsewhere [77]. The end result is that both SGLD with reweighting and SGLDfp are sufficiently accurate, when used properly, to both enhance sampling and preserve the ensemble.

B. Argon Fluid

Argon liquid represents a typical homogeneous system. It is a convenient system to examine ensemble average properties. Argon atoms were described by the Lennard-Jones 6-12 potentials with $\varepsilon = 119.8$K and $\sigma = 3.405$ Å. In this example system, 500 argon atoms were placed in a cubic periodic box ($28.53 \times 28.53 \times 28.53$ Å^3). A time step of 1 fs was used for all simulations. The simulation length was 10 ns for each simulation. The temperature was set to 100K except otherwise noted. Nonbonded interactions were calculated using the isotropic periodic sum (IPS) method [52, 75, 78]. The following rationalized polynomial 3D IPS potentials are used for Lennard-Jones potential calculation.

Lennard-Jones IPS potentials:

$$\varepsilon_{\text{disp}}^{\text{IPS}}(r, R) = \begin{cases} -\frac{C_{ij}}{r^6} - \frac{C_{ij}}{R^6}\left(\frac{1341}{3064} + \frac{77}{141}\left(\frac{r}{R}\right)^2 + \frac{61}{141}\left(\frac{r}{R}\right)^4 + \frac{56}{141}\left(\frac{r}{R}\right)^8\right) & r \le R \\ 0 & r > R \end{cases} \tag{41}$$

$$\varepsilon_{\text{rep}}^{\text{IPS}}(r, R) = \begin{cases} \frac{A_{ij}}{r^{12}} + \frac{A_{ij}}{R^{12}}\left(\frac{23}{3620} + \frac{8}{151}\left(\frac{r}{R}\right)^2 + \frac{66}{151}\left(\frac{r}{R}\right)^6 + \frac{100}{151}\left(\frac{r}{R}\right)^{10}\right) & r \le R \\ 0 & r > R \end{cases} \tag{42}$$

Here, R is the radius of the IPS local region, or the cutoff distance. To quantitatively compare the SGLD and high-temperature LD simulations, we plot the average potential energies against diffusion constants in Fig. 13. Diffusion constants measure the conformational change in the slowest frequencies and can be a good measurement of conformational search efficiency. The diffusion constants were calculated with a fixed center of mass to avoid any exaggeration due to the enhanced motion of the center of mass. As can be seen from Fig. 13, SGLD increases diffusion constants with much smaller energy deviations than LD simulations at elevated temperatures. This plot tells us that SGLD can speed up conformational searches with little change in conformational distribution, while high-temperature LD simulation speeds up conformational search, but searches a conformational space far away from that of the temperature of interest.

The weighted average potential energies are also plotted against diffusion constants in Fig. 13. For SGLD, the weighted potential energy is very flat against diffusion constant. In other words, through the on-the-fly weighting procedure,

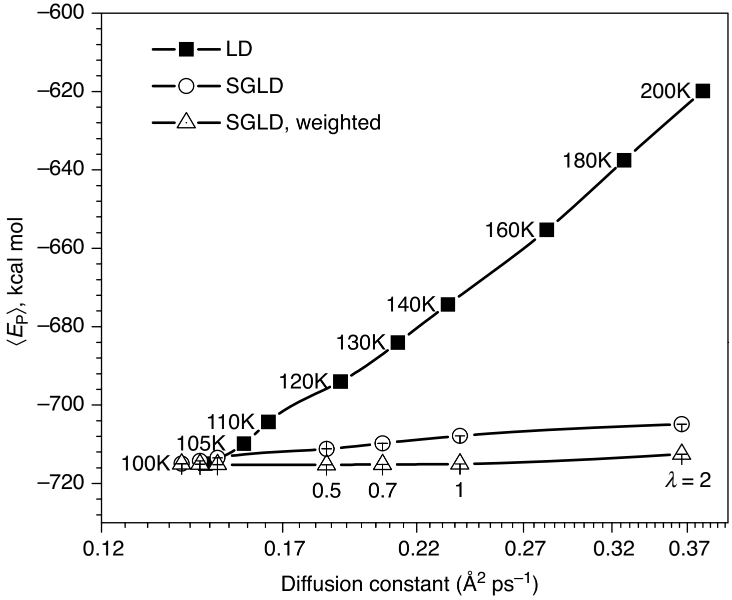


Figure 13. Average potential energies versus diffusion constants for the argon liquid in the LD simulations at different temperatures (as labeled) and in the SGLD simulations at different guiding factors (as labeled). The collision frequency is 1 ps^{-1}. The SGLD simulations were performed at 100K.

SGLD can speed up conformational searches and produce an accurate conformational distribution.

This result also serves as an example that SGLD not only increases the energy barrier crossing rate, but also accelerates the diffusion process. The speedup in conformational search by SGLD is not only through overcoming energy barriers, but also through enhancing damped low-frequency motion.

C. Alanine Dipeptide

Alanine dipeptide is the simplest molecule that is relevant to proteins. The conformation of this molecule is mainly characterized by two dihedral angles, ϕ: CT-N-Cα-C and ψ: N-Cα-C-NT (Fig. 14). The CHARMM all-atom force field [37] was used to describe the interactions. Here, we used a distance-dependent dielectric constant of 4r without cutoffs to represent solvent screening effect to simplify the example.

All simulations were performed with a time step of 2 fs and SHAKE algorithm [79] was employed to fix the bond lengths. Each simulation lasted 200 ns and conformations of every 2 ps were saved for postanalysis. The SGLD and SGLDfp

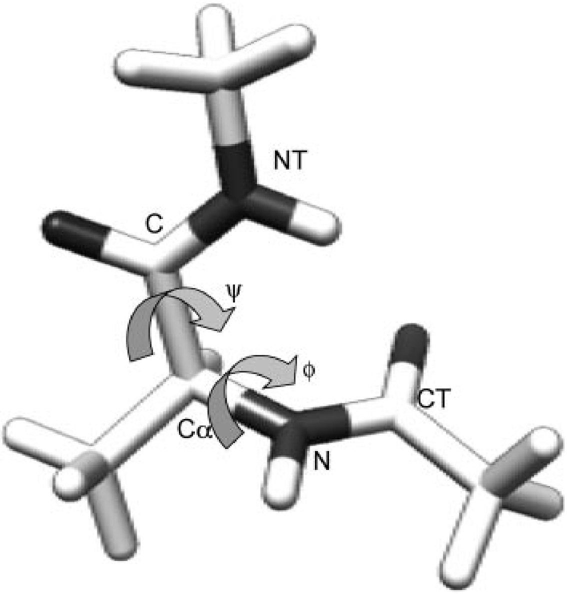


Figure 14. A conformation of an alanine dipeptide. Chemical bonds are shown as sticks. Two backbone dihedral angles, ϕ and ψ, are marked by arrows.

simulations were performed with a local averaging time of $t_L = 0.2$ ps and a temperature of 300 K. A collision frequency of 10 ps^{-1} was used for all the simulations.

Figure 15 compares the $\phi - \psi$ dihedral angle distributions of the alanine dipeptide in LD, SGLD before and after reweighting, and SGLDfp simulations. For this small molecule at the simulation conditions, LD can sample the conformational space fairly well. Comparing the distribution from the LD simulation with that of the SGLD simulation, we can see that the one from the SGLD simulation has a lower peak at (−90°, 170°) and a broader baseline near (−50°, 30°), indicating the changing in the $\phi - \psi$ distribution by the guiding effect in the SGLD simulation. After reweighting, the $\phi - \psi$ distribution from the SGLD simulation becomes similar to that of the LD simulation, demonstrating that the SGLD distribution can be converted to the LD distribution through reweighting. Comparing the $\phi - \psi$ distributions from the SGLDfp simulation and the LD simulation, one can clearly see that they agree with each other fairly well. The root mean square differences from the normalized LD distribution are 1.08, 0.574, and 0.380 for the SGLD distributions before and after reweighting and the SGLDfp distribution, respectively. These are not fully converged values, and we expect that they would get better with longer simulation time.

To demonstrate the conformational search ability, we compare the SGLD and SGLDfp simulations with high-temperature LD simulations. To quantitatively

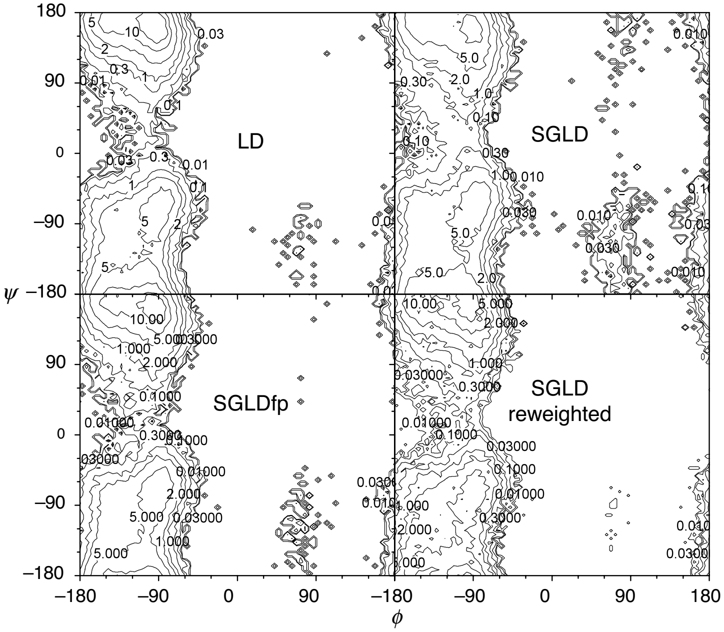


Figure 15. $\phi - \psi$ dihedral angle distribution of the alanine dipeptide in the LD, SGLD, and SGLDfp simulations. The collision frequency is $\gamma = 10$ ps^{-1} and simulation temperature is 300K. A guiding factor of $\lambda = 1$ was used for both SGLD and SGLDfp simulations.

compare the conformational search ability, we calculated the transition rate for the dihedral angles (ϕ, ψ) to transfer from one local minimum at (−90°, −70°) to another local minimum at (−90°, 170°). One transfer is counted when (ϕ, ψ) is changing from within 40° of one local minimum to within 40° of the other local minimum.

Figure 16 shows average potential energy as a function of the transition rate in the high-temperature LD simulations as well as in the SGLD and SGLDfp simulations. The average potential energy reflects the conformational distribution to a certain degree. A change in the average energy indicates a change in conformational distribution. As can be seen from Fig. 16, the high-temperature simulation increases the transition rate, but it also significantly increases the average potential energy. While in the SGLD and SGLDfp simulations, the average potential energy has little change except for the SGLD simulations with $\lambda = 1$. The transition rate increases significantly with λ in both SGLD and SGLDfp simulations, even though the SGLDfp simulations have fewer transitions compared to the SGLD simulations with the same λ. It is also clear from Fig. 16 that the SGLDfp simulation preserves the average energy better. This figure indicates that while the

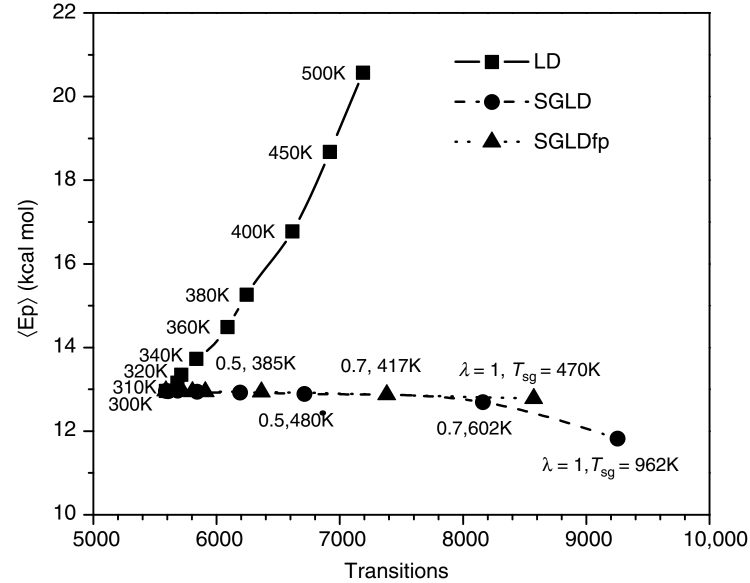


Figure 16. Average energies against transition rates in high-temperature LD, SGLD, and SGLDfp simulations. The simulation temperatures of the LD simulations, as well as the guiding factors and the self-guiding temperatures of the SGLD and SGLDfp simulations, are labeled beside their data points. The collision frequency is $\gamma = 10\ \text{ps}^{-1}$ for all simulations. The simulation temperature is 300K for the SGLD and SGLDfp simulations.

conformational search is accelerated in the high-temperature simulation, the simulation is searching a conformational space that is little relevant to the conformational space at the temperature of interest. In other words, the search is enhanced, but the probability of finding the folded conformation may not be enhanced. SGLD or SGLDfp accelerates the conformational search with little change in ensemble distribution, increasing the chance to reach the folded state.

D. Folding of a Pentamer Peptide

Protein folding is a major challenge for conformational search. Owing to many degrees of freedom of proteins, the conformational space of a protein is huge and exhaustive conformational search is often impossible. We believe that a reasonable hypothesis of protein folding is that the conformational space accessible for a protein is limited, and that the protein can find its folded state quickly by moving through this accessible conformational space. Methods such as high-temperature simulations can accelerate conformational search, but they greatly increase the accessible conformational space. This may reduce the probability to reach the folded state. Because of the many degrees of freedom, conformational space increases exponentially with the accessibility. An increase in the accessible conformational space not only makes the conformational search problem worse, but also may alter the folding pathway or inhibit folding altogether. The temperature replica exchange can enhance sampling while preserving the proper ensemble, but significant difficulties are encountered if the sampled temperatures cross a phase transition at the melting temperature. The SGLD approach avoids the need to generate an ensemble at the many temperatures. The ability to preserve the conformational distribution makes SGLDfp a suitable method to study problems where conformational distribution preservation is critical.

To demonstrate the application of the SGLDfp method in protein folding study, we performed folding simulations for a pentamer peptide [43, 80], which forms a type II turn according to experimental observation. The sequence of the pentamer peptide is Tyr-Pro-Gly-Asp-Val. To simplify demonstration, all simulation conditions were the same as that for the alanine dipeptide simulations described above. A temperature of 300K and a collision frequency of 1 ps^{-1} were set for all the simulations. The guiding factor was $\lambda = 0.5$ for the SGLD simulation and $\lambda = 1$ for the SGLDfp simulation, so that both the simulations have similar conformational search ability. All simulations were started from an extended conformation and were 200 ns in length.

Because a large number of conformations were visited during these simulations, to simplify the description, we clustered the conformations to six major clusters using the local maximum clustering method [42]. The distances between conformations are calculated as the sum of the difference square of the backbone dihedral angles. Figure 17 shows the representative structures of these six major clusters. Clusters 1 and 4 have a broad turn involving Pro-Gly-Asp with the proline carbonyl oxygen pointing up and down, respectively. Clusters 2 and 3 have a tight

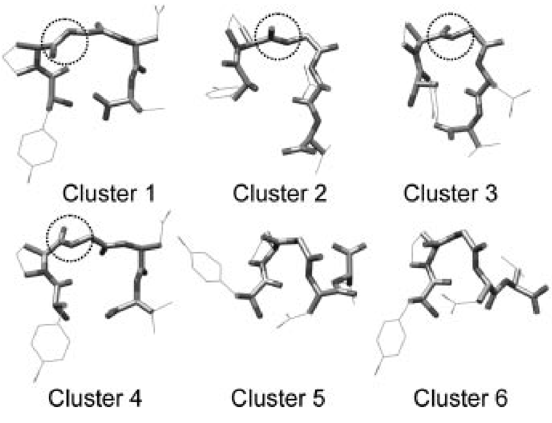


Figure 17. The representative conformations of the six major clusters of the pentamer peptide. Backbone atoms are shown as thick sticks and side-chain heavy atoms are shown as thin sticks. Hydrogen atoms are not shown for clarity.

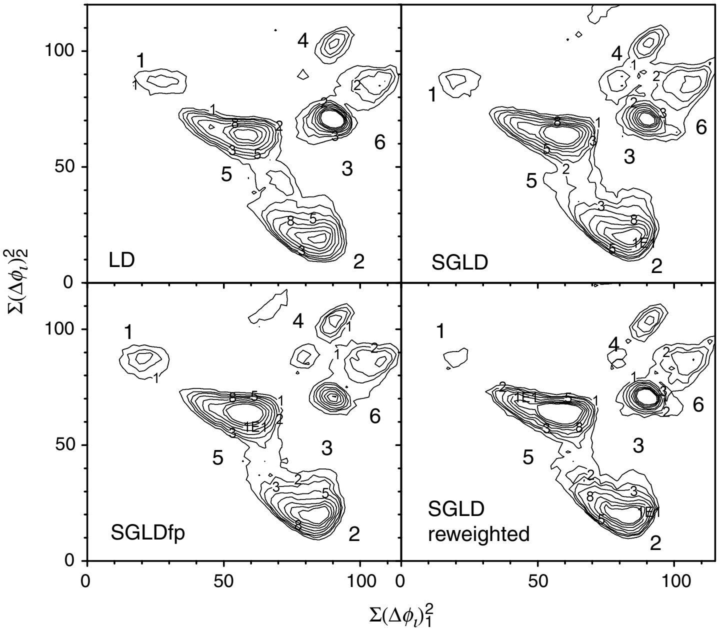


Figure 18. Conformational distributions in the LD, SGLD, and SGLDfp simulations. Conformational distances from the center conformations of clusters 1 and 2 are used as x and y coordinates to show the distributions in two dimensions. All simulations are done with $\gamma = 1$ ps^{-1} and $T = 300$K. The guiding factor is $\lambda = 0.5$ for SGLD and $\lambda = 1$ for SGLDfp.

turn involving Pro-Gly with the proline carbonyl oxygen pointing up and down, respectively. Clusters 5 and 6 form a helical coil with a C-terminus pointing up and down, respectively.

Figure 18 compares the conformational distributions obtained from the LD, SGLD, and SGLDfp simulations. The conformational distributions are shown in two-dimensional contour plots with the distances to the center conformations of clusters 1 and 2 as x and y coordinates, respectively. Even though the peptide has only five residues, the conformational space is large and the LD simulation of 200 ns may not necessarily properly sample the whole conformational space. All six major clusters can be clearly identified in these simulations, even though the SGLD and SGLDfp simulation results have some trace amounts of other clusters. The density from the SGLD simulation shows broader peaks than those in LD and SGLDfp results. After reweighting, the SGLD result has peaks as sharp as the other results. The SGLDfp result resembles the LD result fairly well, again demonstrating that the SGLDfp method is excellent in preserving the conformational distribution. The RMSDs from the LD result are 1.44, 1.59, and 0.81 for the SGLD results before

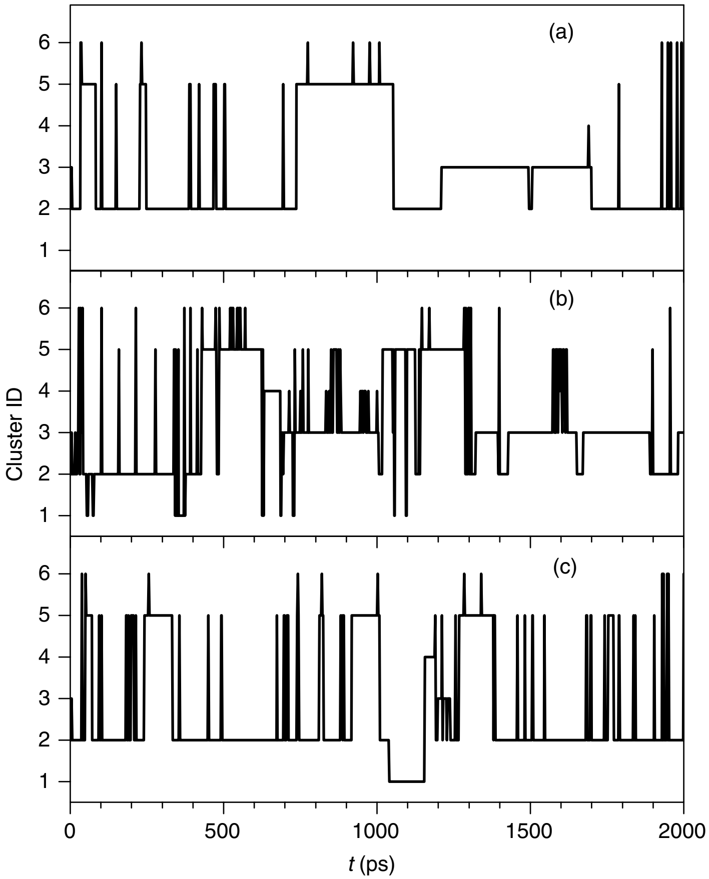


Figure 19. Conformational transitions between the clusters in the first 2000 ps of the LD, SGLD, and SGLDfp simulations. All simulations are done with $\gamma = 1$ ps^{-1} and $T = 300$K. The guiding factor is $\lambda = 0.5$ for SGLD and $\lambda = 1$ for SGLDfp.

and after reweighting and the SGLDfp result, respectively. The large RMSD for the reweighted SGLD result is because the reweighting introduces significant noise in the not fully converged distribution plots.

Figure 19 plots the cluster transitions during the first 2 ns of simulations. As can be seen, the LD simulation (Fig. 19a) did not reach cluster 1 during the first 2 ns, and the transitions between clusters are not as frequent as in the SGLD (Fig. 19b) and SGLDfp (Fig. 19c) simulations. The most frequent transitions were between cluster 2 and cluster 5. This agrees with Fig. 18 which shows that there are the two major clusters and that they are not separated by a significant barrier. There are also significant transitions between cluster 2 and cluster 3, but not between cluster

2 and cluster 4, agreeing with Fig. 18 which shows that cluster 2 and cluster 4 are separated by clusters 3, 5, and 6. This example demonstrates that SGLDfp is an excellent approach for protein folding study, accelerating conformational search while maintaining reasonable conformational distribution.

V. APPLICATIONS

Here, we review the applications of SGMD and SGLD methods in several scientific areas, including protein folding, modeling of protein structures and complexes, protein conformational rearrangements, water penetration, surface adsorption, crystallization, and phase transitions.

A. Protein Folding

Protein folding is one of the most active areas that utilize molecular simulations. However, studies of protein folding have been hindered by the timescale issues. Protein folding occurs on timescales of microseconds and longer. While several groups have reported MD simulations on a timescale of microseconds and longer, such simulation timescales are still not accessible to a majority of MD simulators. The benchmarking and fine-tuning of existing force fields is another problem in the field and is expected to improve as more and more structures folded through the simulations can be compared with experimental structures. SGMD/SGLD will aid the field of protein folding by easing the conformational search limitations.

The earliest application of SGMD in protein folding was in the study of reversible folding of a linear pentamer peptide YPGDV. NMR studies have shown that this peptide has a significant population (50%) of a type II turn conformation in aqueous solution [80]. This peptide was simulated in water with atomically detailed representation for both peptide and solvent molecules at 300K using the SGMD. During a 2 ns SGMD simulation started from a fully extended conformation, the peptide folded into a type II turn-like conformation and then undergoes unfolding and refolding several times [43]. Simulations with regular MD [43, 81] failed to reach the experimentally observed turn structure in 2 ns. Five major conformational clusters were obtained from the 2 ns SGMD simulation and the most populated conformational cluster was a type II reverse turn-like conformation. The structure of the most populated conformational cluster identified through the SGMD simulations was consistent with the NMR data, and the estimated relative NMR NOE strengths of proton pairs based on the SGMD trajectory are in good agreement with the experimental data. Figure 20 shows typical conformational clusters observed during the folding simulation.

SGMD simulations were further employed to study helix folding in explicit water [44]. A 16-residue alanine-based helical peptide [82], Ace-(AAQAA)-Y-NH2, was simulated for 10 ns. The reversible folding (folding, unfolding, and

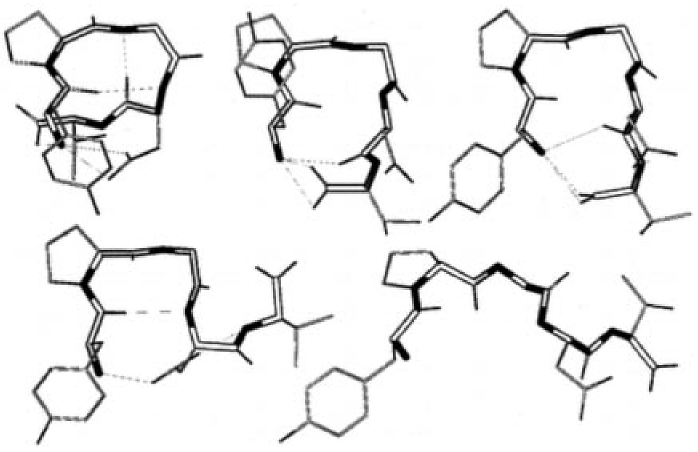


Figure 20. Five major clusters identified from the trajectory obtained from the 2 ns SGMD simulation [43] (upper row: clusters I, II, and III; lower row: clusters IV and V).

refolding) of this peptide in explicit water at 274K was successfully accomplished. Consistent with experimental results, the helix was found to be the major secondary structural element in aqueous solution, and among different helix forms, the α-helix was the dominant form. Conformational analysis of our simulation results showed that turns and 3_{10}-helices play an essential role in the folding of the α-helix. Conventional MD simulations for the same system failed to explore the conformational space in a 10 ns period. An MD simulation started with an extended conformation remained in a random coil structure throughout the 10 ns period as shown in Fig. 21a, and an MD simulation started with a complete helix remained a complete helix as shown in Fig. 21b. In the SGMD simulation, a variety of conformational states were observed, and their populations are shown in Fig. 22.

β-hairpin folding is a challenge for molecular dynamics simulations due to its long folding time. Using SGMD method, for the first time, β-hairpin folding was directly observed in explicit water simulation [42, 45]. The sequence of the peptide is Tyr-Gln-Asn-Pro-Asp-Gly-Ser-Gln-Ala. Strong NMR NOE evidence indicates that this peptide folds into a β-hairpin structure in aqueous solution [83]. Reversible folding process of this β-hairpin was simulated with the SGMD method, and details of the folding process were analyzed. Figure 23a shows a typical β-hairpin structure observed in the simulations. This structure was first reached in about 20 ns. Figure 23b shows the excellent agreement between the experimental NOEs and the average distances of the corresponding atom pairs.

Recently, Lee and Olson combined SGLD with temperature-based replica exchange to perform protein folding simulation [32]. They tested the performance

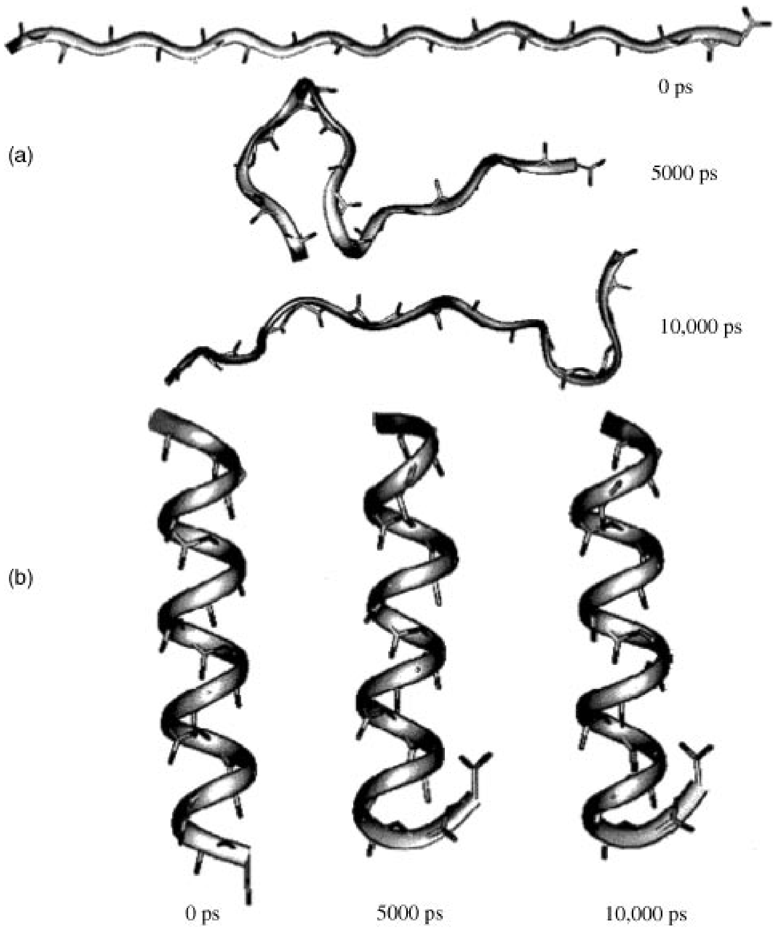


Figure 21. Snapshots of the peptide conformations obtained in the two 10 ns MD simulations [44]: (**a**) simulation started from a fully extended conformation; (**b**) simulation started from a complete helix conformation.

and accuracy of the MD-ReX, LD-ReX, and SGLD-ReX simulations on the prediction of thermodynamic folding observables of the Trp-cage miniprotein. The PARAM22 + CMAP force field was used together with the generalized Born molecular volume implicit solvent model. They found that the SGLD-ReX folds up the protein somewhat faster than the two conventional ReX approaches, in contrast to the 65-fold speedup of helix formation reported in the original SGLD study [12]. The likely explanation is that ReX already provides sufficient sampling enhancement for MD and LD to overcome the unfolded/folded transition barrier

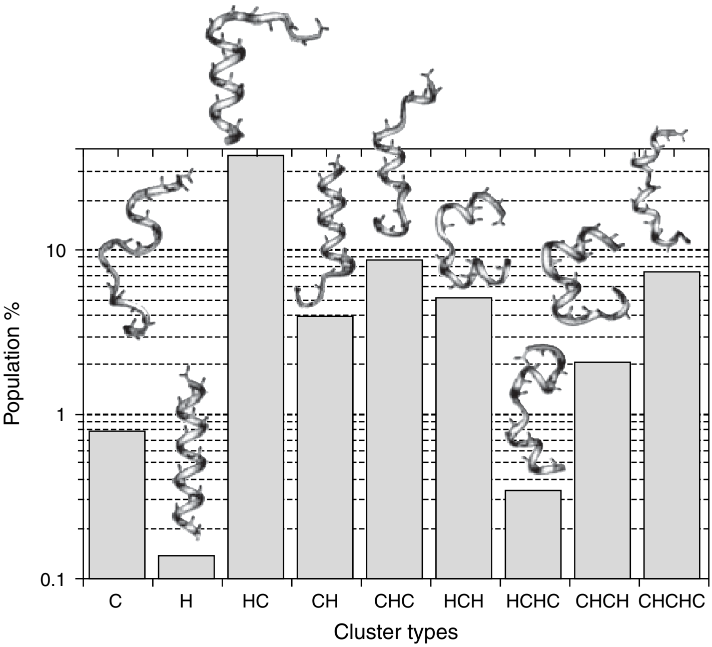


Figure 22. Major secondary structure clusters observed during the 10 ns SGMD simulation with $\lambda = 0.4$ [44]. Conformations are clustered on the basis of the number and location of helix segments in the peptide.

to fold up the Trp-cage. Their result suggests that SGLD-ReX improves sampling convergence by reducing topological folding barriers between energetically similar near-native states. Also, they found that SGLD-ReX predicted the melting temperatures, heat capacity curves, and folding free energies that are closer in agreement to the experimental observations. Figure 24 shows the energy and RMSD distribution in MD-ReX, LD-ReX, and SGLD-ReX. All three methods sample the nearest to native basin (1 Å) at their respective transition temperatures, with SGLD-ReX having the most density there. Since the nearest to native basin does not appear to be the lowest in free energy, this could be due to the fact that SGLD-ReX performs the most excursions among basins in a given simulation time. Another positive feature of SGLD-ReX shown in Fig. 24g–i is how similar the PMFs are among the different starting conformations and data collection times. This suggests that, of the three methods, SGLD-ReX is the most self-consistent and arguably the most converged, at least in the conformational space of compact folds. The 150–200 ns data windows of MD-ReX and LD-ReX do have qualitative agreement with

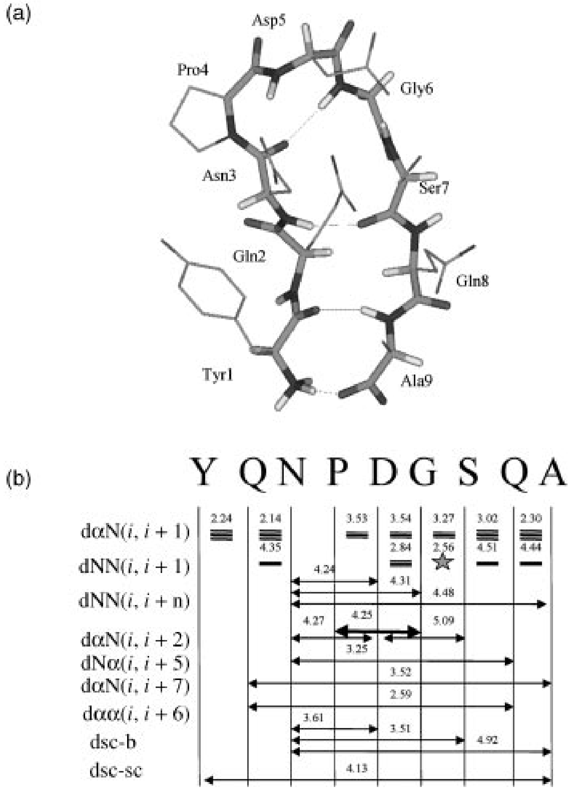


Figure 23. Reversible β-hairpin folding simulation with SGMD [42, 45]. (**a**) A typical folded structure of the peptide obtained in the simulation (at 21,000 ps). For clarity, side-chain hydrogens are not shown. The backbone atoms are shown as thick sticks and side-chain atoms as thin sticks. Interstrand hydrogen bonds are marked by dashed lines. (**b**) NMR NOEs observed in the peptide aqueous solution 2 (arrow bars between residues) and the average hydrogen pair distances (numbers in Å above NOE bars) in the β-hairpin structure obtained in our simulation. α, N, sc, and b represent the hydrogen atoms on α-carbon, amide nitrogen, side chain (β-carbon in our calculation), and backbone (amide nitrogen in our calculation). The thickness of the NOE bars represents the strength of the NOEs reported. Generally, NOEs are strong for hydrogen pair distances within 3 Å, medium between 3 and 4 Å, and weak between 4 and 5 Å.

SGLD-ReX, suggesting that longer equilibration times could bring these three methods into better agreement.

Lee and Chang characterized the denatured state of the human prion protein (huPrP) 121–230 through SGLD simulations [38]. Misfolding and aggregation of the prion protein (PrP) are responsible for the development of fatal transmissible

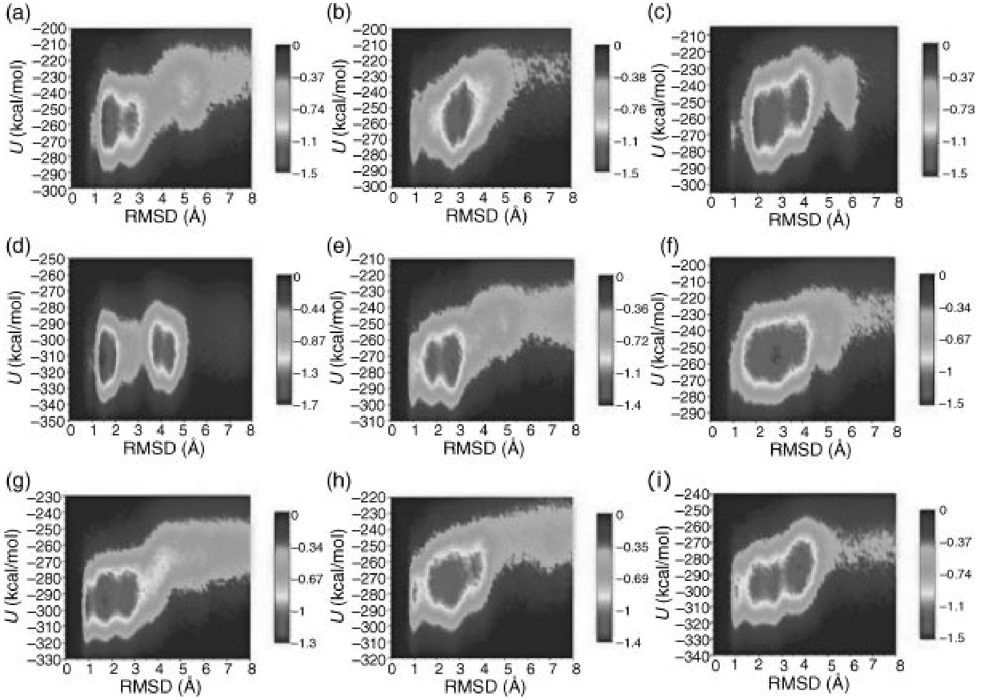


Figure 24. Combined SGLD with replica exchange in protein folding simulation [32]. Free energy landscapes at respective melting temperatures ($\Delta G_{\text{fold}} = 0$) of individual simulations (method/starting structure/simulation data) in the coordinates of potential energy, U, and C_α RMSD to native: (**a**) MD-ReX/*trans*/50–100 ns ($T = 351.3$K), (**b**) MD-ReX/native/50–100 ns ($T = 354.6$K), (**c**) MD-ReX/native/150–200 ns ($T = 348.2$K), (**d**) LD-ReX/*trans*/50–100 ns ($T = 290.1$K), (**e**) LD-ReX/native/50–100 ns ($T = 335.1$K), (**f**) LD-ReX/native/150–200 ns ($T = 353.9$K), (**g**) SGLD-ReX/*trans*/50–100 ns ($T = 311.2$K), (**h**) SGLD-ReX/native/50–100 ns ($T = 331.5$K), and (**i**) SGLD-ReX/native/150–200 ns ($T = 306.4$K).

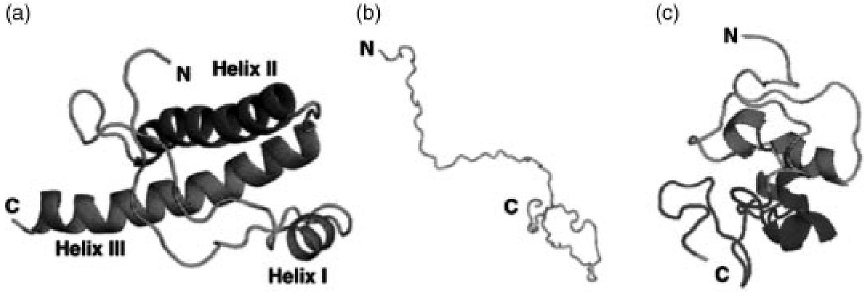


Figure 25. Characterization of prion denatured state with SGLD simulations [38]. (**a**) The NMR structure of huPrP 121–231 (PDB 1hjn [19]), (**b**) fully unfolded huPrP 121–230 at 600K, and (**c**) the simulated denatured structure of huPrP from the most populated cluster.

neurodegenerative diseases. To gain insight into possible aggregation-prone denatured states, multiple SGLD simulations starting from the extended conformation of the huPrP 121–230 were performed. The simulations were performed with an implicit solvent and were 50 ns long. The structural analysis indicated that the most populated denatured state of huPrP is partially folded with helical content. Experimental observation indicated that PrP fibril is rich in β-sheet structure. Lack of β-structure suggests that β-sheets in amyloid fibrils may be formed from intermolecular interactions rather than intramolecular forces. Figure 25 shows the partially unfolded structure of huPrP 121–230.

Wen and coworkers have studied protein folding with Poisson–Boltzmann molecular dynamics with self-guiding forces (SG-PBMD) [39]. They investigated the sampling efficiency with SG-PBMD in molecular dynamics with the PB implicit solvent when self-guiding forces are added. They found an impressive efficiency as measured by fluctuations of potential energy, radius of gyration, backbone RMSD, the number of unique clusters, and distribution of low RMSD structures over time compared to a high-temperature dynamics simulation. They performed *ab initio* folding simulations of BBA1 and villin headpiece and discussed folding pathways for the two small proteins. They found topological agreement between the folded state observed in their simulation and the theoretical native states (Fig. 26). The denatured state of the BBA1 miniprotein was discussed in more detail in a subsequent publication [40].

B. Molecular Modeling and Docking

Characterization of the solution structure of peptides has been the goal of many simulation studies. Yang et al. used SGMD to study solution conformations of wild-type and mutated Bak BH3 peptides via dynamical conformational sampling

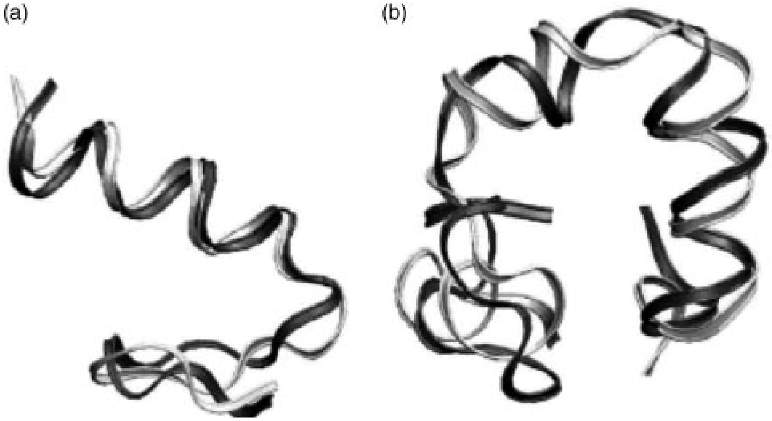


Figure 26. Poisson–Boltzmann molecular dynamics simulation with the self-guiding force [39, 40]. Superposition of theoretical native states (gray) and native-like structures found in folding simulations (black): (**a**) BBA1; (**b**) villin headpiece.

[48]. The BH3 domain of the Bcl-2 family of proteins plays a critical role in the regulation of apoptosis. Their SGMD simulations showed that the Bak peptide exhibits a partially formed helical structure with a fairly stable six-residue helical segment at the N-terminus and a less stable approximately four-residue helical segment at the C-terminus. Additional SGMD simulations of two mutated Bak peptides found that the R5G mutation greatly affects the solution conformations of the peptide and that the overall helix ratio decreases by a factor of 2 compared to the wild-type Bak peptide, but the R5A mutation does not affect significantly the peptide solution conformations observed in the wild type. To quantitatively examine the effects of mutations on each residue, they calculated the helical propensity for each residue from the 10 ns simulations for these three peptides (Fig. 27). Analysis of representative conformations of the R5A mutant suggested that the relatively stable helical segment close to the N-terminus may greatly facilitate its binding to Bcl-xL.

Chandrasekaran et al. utilized SGMD in modeling of a protein complex between the protein Z-dependent protease inhibitor (ZPI) and the factor Xa (FXa), a serine protease that plays a key role in the blood coagulation cascade [49]. The Michaelis complex of human ZPI/FXa was built using homology modeling, protein–protein docking, and molecular dynamics simulation methods. The ZPI/FXa complex built through the docking method was subjected to SGMD simulation to enhance conformational sampling efficiency. The aim was to examine whether the conformation of ZPI/FXa obtained through docking moved toward the conformation obtained through homology modeling or if it explored a different conformational path.

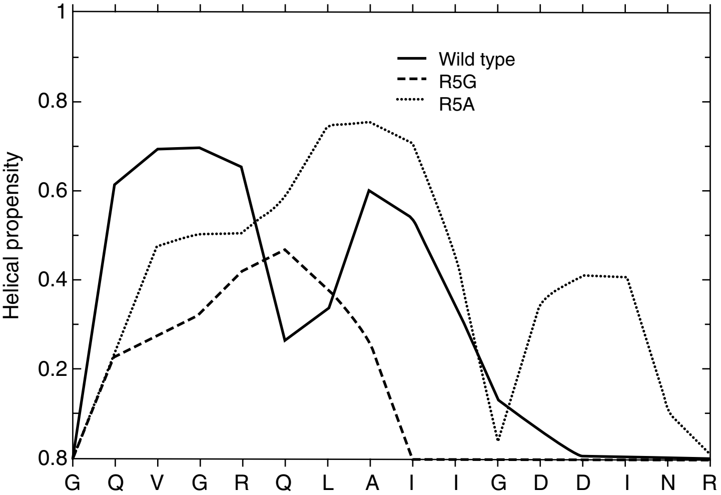


Figure 27. Bak BH3 simulation results with SGMD [48]. Effects of mutation on the helical propensity from the 10 ns wild-type and R5G and R5A sampling simulations.

The motivation behind using accelerated conformational sampling was to obtain a qualitative idea, in a reasonable timescale, regarding the direction of movement of ZPI in the ZPI/FXa complex obtained through the docking method. Figure 28 shows the complex model they obtained.

Understanding the fundamental principles that govern the binding of a guest molecule to its host and accurate prediction of the binding mode of the guest/host complex are important goals in guest–host chemistry and have implications in structure-based drug design. Varady et al. performed a computational investigation of benzyl alcohol (the guest) binding to β-cyclodextrin (the host) in the presence of explicit water molecules (Fig. 29) using both SGMD and conventional MD simulations [47]. In their SGMD simulations, competitive and reversible binding of the guest molecules to the host was observed. Analysis of the simulation trajectories (Fig. 30) showed that one major complexed conformational cluster is in good agreement with the complex structure determined using the X-ray diffraction method. In addition, several other major binding modes were also identified in aqueous solution. Investigation of the binding forces showed that the burial of the phenyl group in the cavity of β-cyclodextrin, but not the hydrogen bonding interaction between the guest and the host, is the major change for binding, suggesting that hydrophobic interaction may be responsible for the formation of the complex. To verify the predictions made by the SGMD method, two 12.5 ns conventional MD simulations with the same initial setup and same conditions as for the two SGMD simulation runs were performed. In addition, a 10 ns long conventional

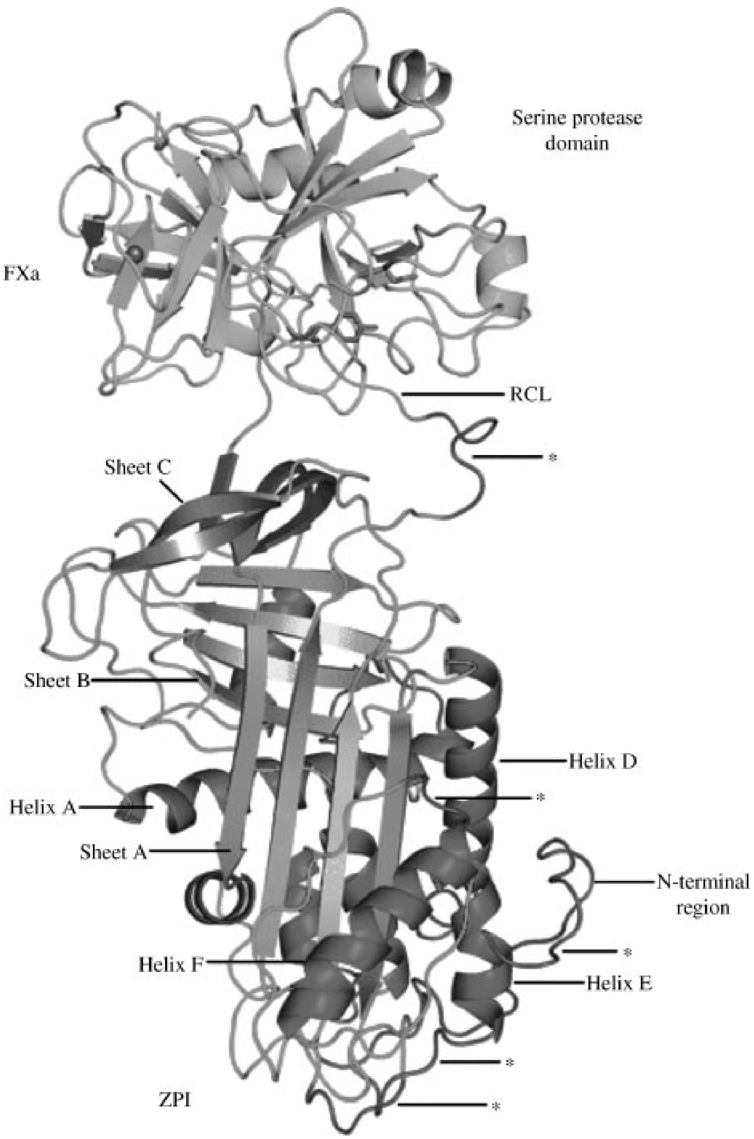


Figure 28. Solvent equilibrated models for protein Z-dependent protease inhibitor and its initial reactive complex with coagulation factor Xa [49].

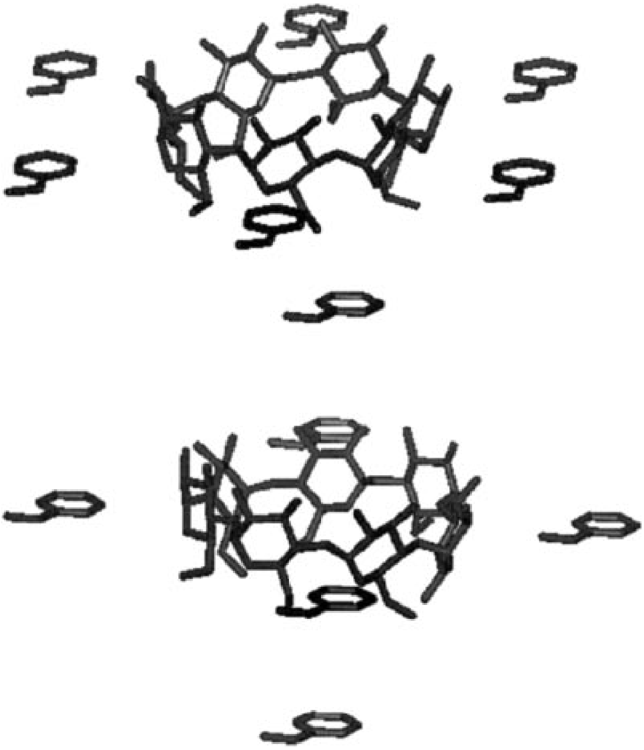


Figure 29. Guest–host binding simulation with SGMD [47]. Starting conformations for SGMD (SGMD #1 and #2) and corresponding MD simulations (MD #3 and #4). For clarity, only heavy atoms are shown. The six benzyl alcohol molecules are either all around the "rim" of the host (top) or four around the rim, one "below," and one "above" β-cyclodextrin (bottom).

MD simulation starting from the crystal structure of the complex was performed. The MD simulations predicted major solution binding modes similar to those identified through the SGMD simulations, including the conformational cluster that is essentially the same as that found in the X-ray structure. The studies showed that the SGMD method is an efficient way to study competitive and reversible binding of guest molecules to their hosts in aqueous solution. This study result indicates that SGMD may also be useful to study the binding of drug molecules to their macromolecular targets.

Lung et al. used SGMD to study conformations of a small peptide (called G1) that binds to the Grb2-SH2 domain but not the src SH2 domain [46]. G1 is a candidate to be an inhibitor for the function of the Grb2-SH2 domain that binds to specific tyrosine phosphorylated motifs on activated GF receptors. Overexpression

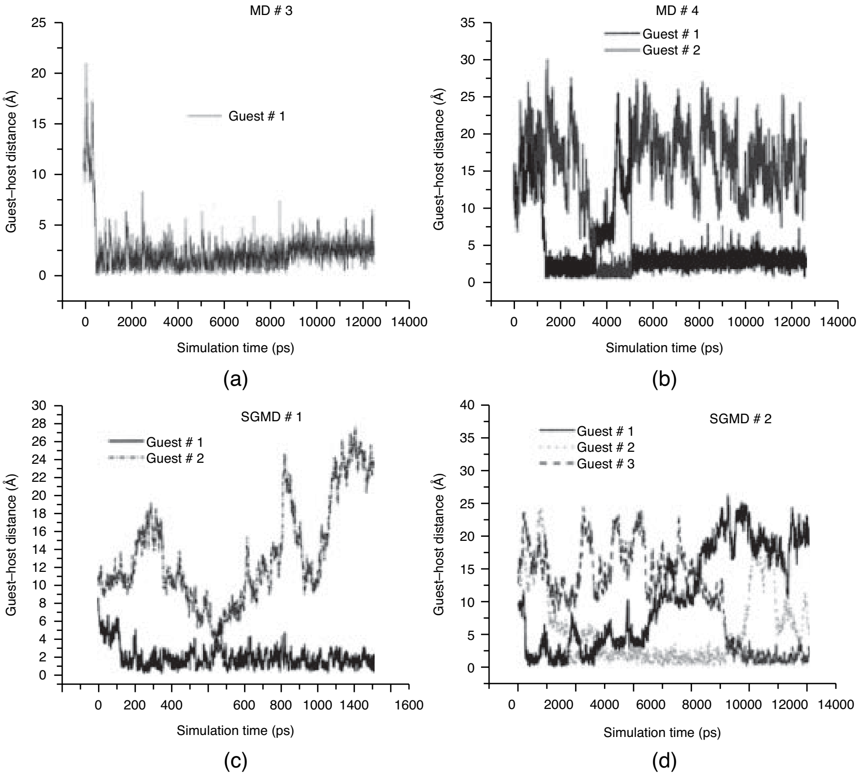


Figure 30. Distance between the centers of the mass of guest and host molecules during MD and SGMD simulations [47]. For clarity, we plotted only the distances for guest molecules, which become tightly bound (within 3 Å) to the host molecule for longer than 30 ps during simulations.

of these receptors, or constitutive activation of this pathway, is highly relevant to a number of diseases, including breast cancer. Thus, blocking Grb2-SH2 function provides a promising therapeutic target for the development of new antitumor agents. Conformations of the G1 peptide in explicit solvent were generated with the SGMD simulations. For the SGMD simulations, the local averaging time t_1 was set at 2 ps and the guiding factor was set at 0.5. The four major conformational clusters of G1 identified from an SGMD simulations are shown in Fig. 31. Molecular modeling studies suggest that the G1 peptide can adopt low-energy solution conformations, which allow its Tyr3 and Asn5 to mimic the corresponding pTyr and Asn residues in the natural phosphopeptide ligand. Moreover, its Glu1 residue can interact with the positively charged binding site in Grb2-SH2, thus partially compensating for the absence of a phosphate group in G1.

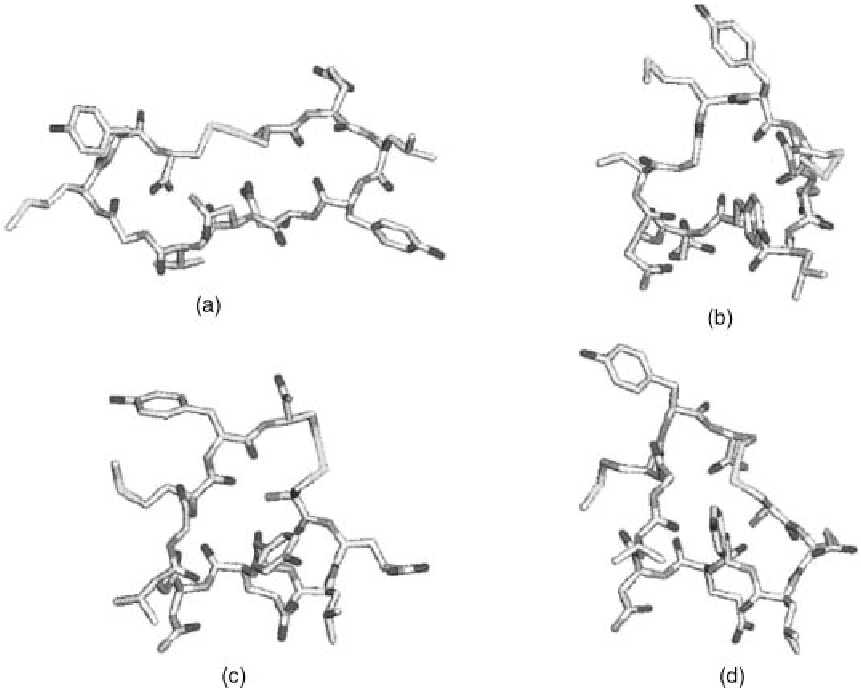


Figure 31. Four major conformational clusters of G1 identified from an SGMD simulation in explicit water [46]. These clusters were identified using all 500 conformations recorded during the 500 ps SGMD simulation. The conformations in conformational cluster (A) were similar to the starting conformation, whereas the conformations in the other three clusters were substantially different from the starting conformation and exhibited circular open-chain backbone conformations lacking evidence of intramolecular interactions.

Owing to the enhanced conformational search ability of SGMD/SGLD, this method is often used as an efficient way to explore conformational space. Shao et al. [84] used MD and SGMD as a tool to generate conformational library to test different clustering algorithms. In case of a 10-mer polyadenine single strand of DNA, standard 5 ns long MD simulations yielded conformations that were fully stacked and helical on a 5 ns timescale. SGMD was used to generate single-strand structures more representative of the true ensemble and to generate a set of diverse conformations for clustering. SGMD parameters that were utilized were significantly greater than those routinely applied. When used in this manner, the SGMD rapidly moves the DNA and effectively samples a wide range of "unfolded" conformations in short (1 ns) runs. Configurations generated with SGMD were then used as starting structures for a standard MD run. The structures generated with MD were further used for clustering.

C. Protein Conformational Transitions

Protein conformational transitions play a central role in key cellular processes such as signal transduction. Owing to the large size of protein systems and long timescale of transition events, describing such events has been a challenge for simulation studies. The SGLD method has been used successfully to provide qualitative insights into the mechanisms and types of conformational relaxation that occur upon ligand binding.

Damjanovic et al. applied SGLD to study protein conformational reorganization triggered by charging of internal ionizable residue in three variants of protein staphylococcal nuclease (SNase) [52]. SGLD simulations with five different sets of guiding parameters (including $\lambda = 0$, that is, no guiding) were performed and compared to each other, as well as to the structural information available through CD, steady state Trp fluorescence, and NMR spectroscopy. Simulations of the wild-type protein, which does not contain internal ionizable residues and does not undergo conformational transitions, served to calibrate and benchmark the simulations. Comparison of the amount of backbone relaxation in the wild-type protein as measured through the average secondary structure content showed only small amounts of secondary structure loss, exclusively localized to the termini of β-strands and α-helices. The observations were consistent between SGLD and LD simulations, with the SGLD simulations with $\lambda = 1$ exhibiting slightly smoother transitions at helical termini. In contrast to the wild-type protein, the three variants that contain internal ionizable residues exhibit experimental evidence of structural relaxation triggered by charging of internal groups. Figure 32 shows the secondary structure changes observed during the SGLD simulations. The structural trends observed in the simulations are in general agreement with experimental observations. The I92D variant, which unfolds globally upon ionization of Asp-92, in simulations often exhibits extensive hydration of the protein core and sometimes also significant perturbations of the β-barrel. In the crystal structure of the V66R variant, the $\beta 1$ strand from the β-barrel is domain swapped; in the simulations, the $\beta 1$ strand is sometimes partially released. The V66K variant, which in solutions shows reorganization of six residues at the C-terminus of helix $\alpha 1$ and perturbations in the β-barrel structure, exhibits fraying of three residues of helix $\alpha 1$ in one simulation and perturbations and partial unfolding of three β strands in a few other simulations. Overall, the use of SGLD simulations was shown to facilitate observation of conformational transitions in proteins where such conformational relaxation is believed to exist.

In another study of variants of the V66E form of SNase [51], SGLD was benchmarked against LD in its ability to reproduce hydration state and rotameric substates of internal Glu-66 side chain, when the side chain is in a neutral state (Fig. 33). Because of the intricate coupling between the hydration state and the rotameric states of this internal side chain, the correct sampling of side-chain

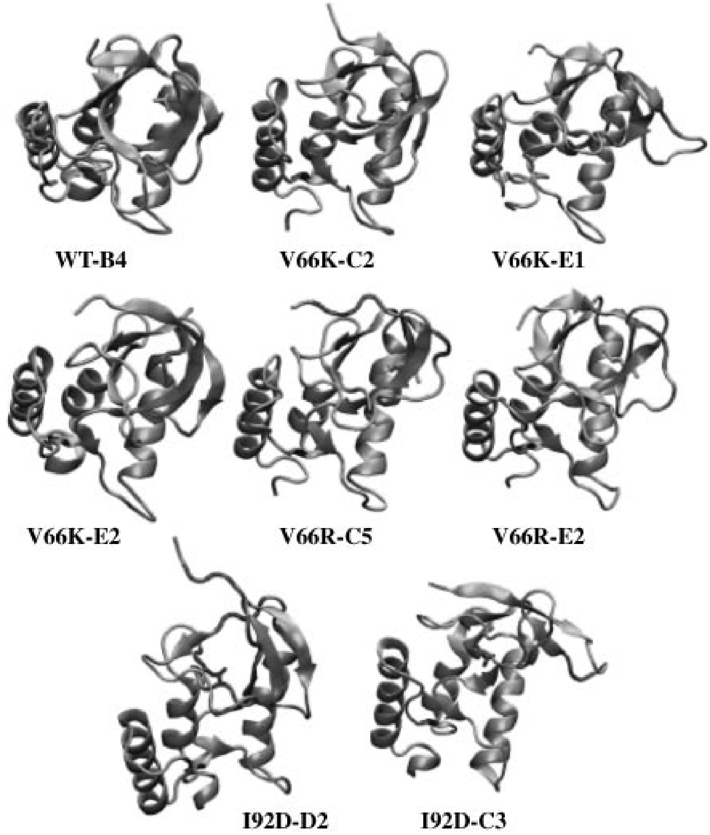


Figure 32. SNase conformational reorganization from SGLD simulations [52]. The residues undergoing change in the secondary structure are shown in red. The substituted ionizable groups are shown in stick representation.

conformations may require very long simulation times. Alternatively, multiple simulations started with different initial velocities can achieve more effective sampling of side-chain conformations. In this study, populations of two side-chain conformations were studied based on 40 short LD and SGLD simulations. The results of simulations with LD and SGLD methods yield side chain and water populations that agree up to 8%. In contrast, the results of simulations started with and without the crystallographic water molecules differed by as much as 20%. However, the simulations were not fully converged and with additional simulation time the simulations with different initial hydration states could have converged to the same value. Similarly, the differences in populations observed in simulations with SGLD

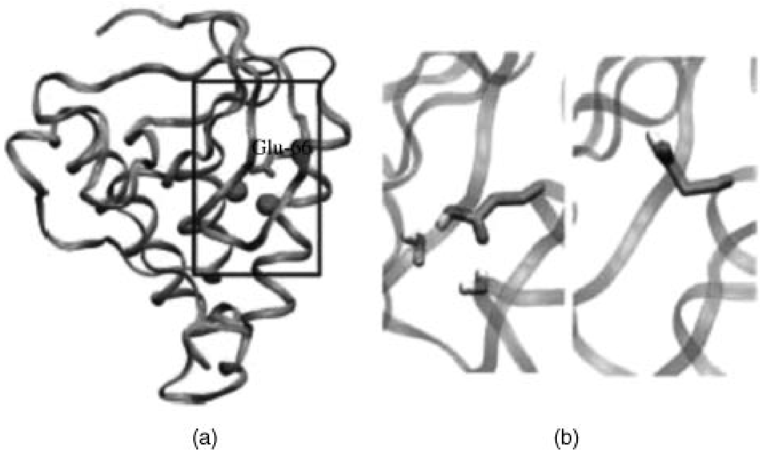


Figure 33. (**a**) Crystal structure of the V66E variant of SNase [51]. (**b**)Snapshots from MD simulations representative of the straight (left) and the twisted (right) conformation of the Glu-66 side chain.

and LD could likely be attributed to the difference in sampling efficiency of the two methods during the same simulation time. Surprisingly, when performance of SGLD in sampling of conformational transitions was benchmarked, it was found that the number of hops between the two conformations of the side chain was only slightly larger in SGLD than in LD (228 versus 191). We believe that this is because the conformational transitions in this case were heavily influenced by the fluctuations in the hydration state of the side chain. The hydration state of the side chain was dependent on the penetration and exiting of water molecules from the protein interior and the guiding parameters used in the study most likely do not enhance such motions.

Conformational transitions induced by dephosphorylation in the NtrC protein were studied through multiple SGLD simulations [50] (Fig. 34). SGLD simulations provided a way to examine structural and dynamical properties of the receiver domain of nitrogen regulatory protein C (NtrCr) and study pathways of conformational transitions induced by dephosphorylation. NtrC is a signaling protein regulated by phosphorylation of an Asp-54 residue in NtrCr. It is believed that the protein undergoes conformational transitions between inactive and active forms on a microsecond timescale. Phosphorylation of NtrC^r stabilizes the active form of the protein. The major structural difference between the two forms is in the orientation of the regulatory helix $\alpha 4$. SGLD and MD simulations of the phosphorylated active form structure suggest a mostly stable but broad structural ensemble of this protein. The finite difference Poisson–Boltzmann calculations of the pK_a values of the active site residues suggest an increase in the pK_a of His-84 on

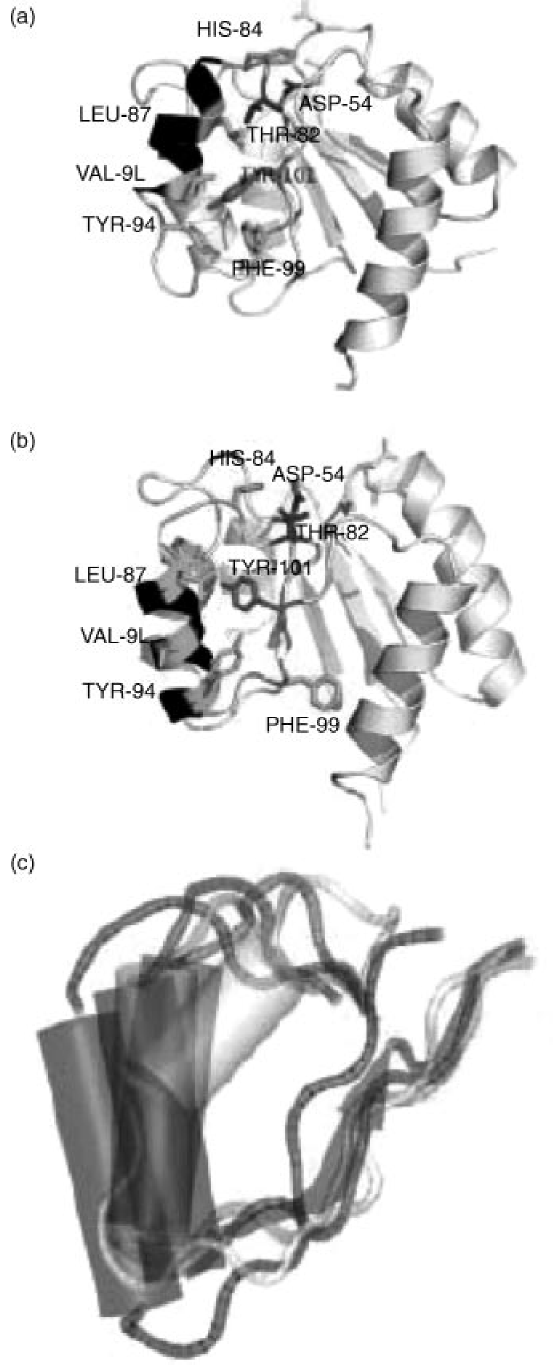


Figure 34. SGLD simulation of conformational transitions induced by dephosphorylation in the NtrC protein [50]. (**a**) NMR structure of the inactive form of NtrCr. The key helix 4 is shown in black. (**b**) NMR structure of the active form of NtrCr. (**c**) Conformations of the key helix 4.

phosphorylation of Asp-54. In SGLD simulations of the phosphorylated active form with charged His-84, the average position of the regulatory helix $\alpha4$ is found closer to the starting structure than in simulations with the neutral His-84. To model the transition pathway, the phosphate group was removed from the simulations. After 7 ns of simulations, the regulatory helix $\alpha4$ was found approximately halfway between positions in the NMR structures of the active and inactive forms. Even though the simulations were too short to observe the full range of conformational transitions between the active and inactive forms of the protein, the study illustrates the potential utility of the SGLD method in providing the atomic-level details about the pathways of conformational transitions and role of particular residues in conformational transitions induced by ligand binding/unbinding.

SGLD was recently used to study conformational changes in a membrane transporter protein lactose permease (LacY) [53]. LacY undergoes a conformational change from a state that is open to the cytoplasm to the state that is open to the periplasm in response to sugar binding and protonation of Glu-269 residue (Fig. 35). SGLD simulations were used to enhance conformational sampling in simulations of LacY with implicit description of the membrane. SGLD simulations were followed by MD simulations with an explicit description of a fully hydrated bilayer. Control simulations without the sugar bound and without the protonated Glu-269 were performed to verify that in this case there are no conformational changes in the periplasmic half. Indeed, only simulations with the sugar bound and with the protonated Glu-269 resulted in conformational changes in the periplasmic half. In those simulations, the pore radius of the lumen increased by

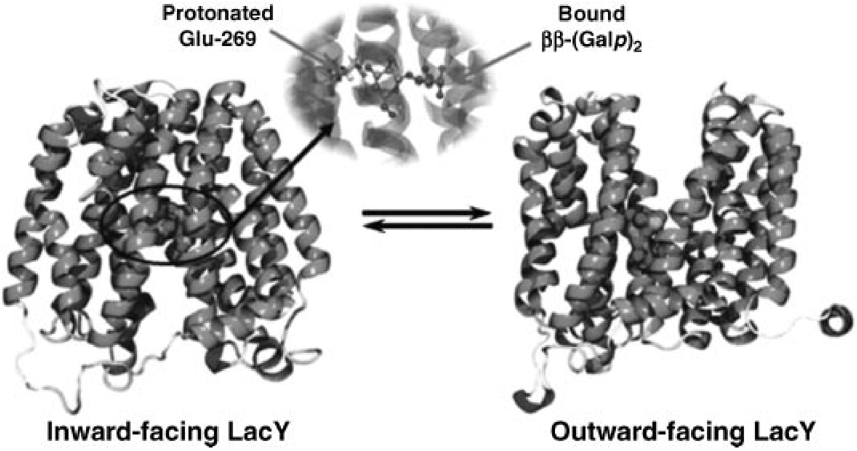


Figure 35. SGLD study of conformational changes in a membrane transporter protein lactose permease (LacY) [53]. Proton translocation to Glu-269 and sugar binding trigger LacY conformational change from the inward-facing to the outward-facing state.

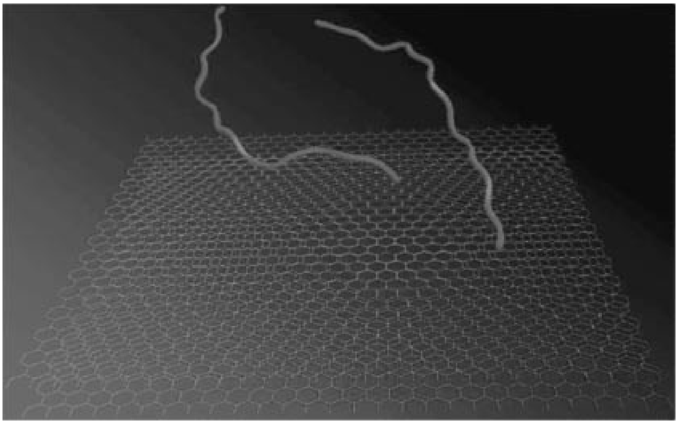


Figure 36. Molecular dynamics simulations of the ionic complementary peptide EAK16-II on the hydrophobic HOPG surface [58, 59]. The adsorption and initial assembly process of EAK16-II on the surface were revealed. Hydrophobic alanine residues are found to be energetically favorable when in contact with the HOPG surface. It is the hydrophobic interaction that drives the adsorption of the first peptide molecule.

3.5 Å on the periplasmic side, while the pore radius decreased by 2.5 Å on the cytoplasmic side. SGLD simulations were found to enhance observations of structural changes. The periplasmic open conformations were found to agree with experimental data. The comparison with the experiments suggests a possible incomplete closure of the cytoplasmic side; however, the closure is large enough to prevent the sugar from being transported to the cytoplasm [53].

D. Surface Adsorption

SGMD simulations were used to study adsorption of the ionic complementary peptide EAK16-II on the hydrophobic HOPG surface (Fig. 36) [58, 59]. Protein adsorption plays an important role in bioactive implant devices and drug delivery materials design. Ionic complementary peptides are novel nanobiomaterials with many biomedical applications, and understanding of the fundamentals of peptide adsorption on the surface is important for peptide applications in biotechnology and nanotechnology. The studies examine the roles of the hydrophobic interaction, electrostatic interactions, and hydrogen bonding interactions on the adsorption of the peptide molecules under neutral, acidic, and basic conditions. Figure 37 shows the snapshots of the peptide EAK16-II on the HOPG surface.

E. Crystallization and Phase Transitions

Argon crystallization was studied with SGMD [11]. A system of 500 argon atoms was used in the simulations (Fig. 38). The starting structure was created by melting

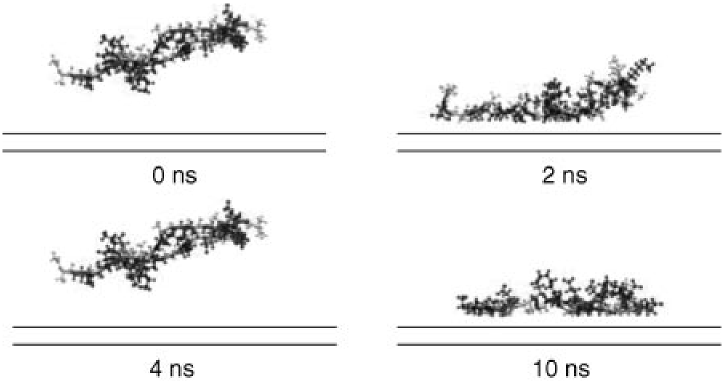


Figure 37. Snapshots of the peptide EAK16-II on the HOPG surface [58, 59]. The two peptide molecules are shown with van der Waals spheres.

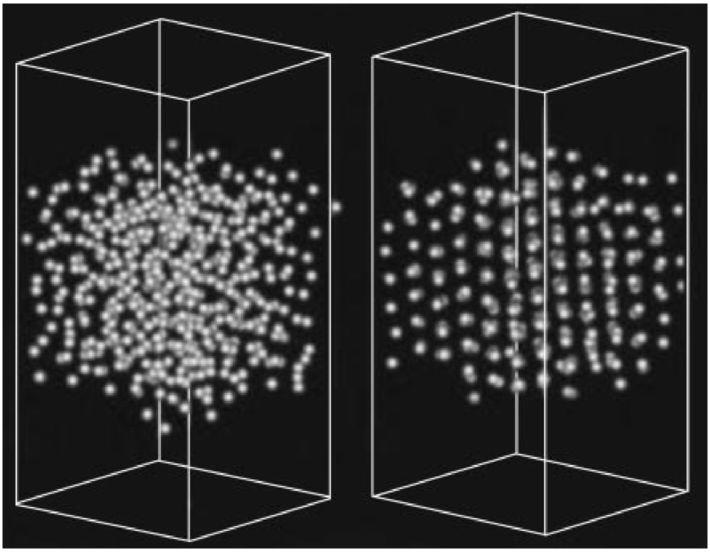


Figure 38. Crystallization of argon liquid observed in SGMD simulations [11]. Snapshots of the argon film system at $T^* = 0.501$ (60K). (**a**) Initial liquid structure and (**b**) crystallized structure. A tetragonal periodic boundary condition with sides $a = b = 28.53$ Å along x and y axes and $c = 57.06$ Å along z-axis is applied to the system.

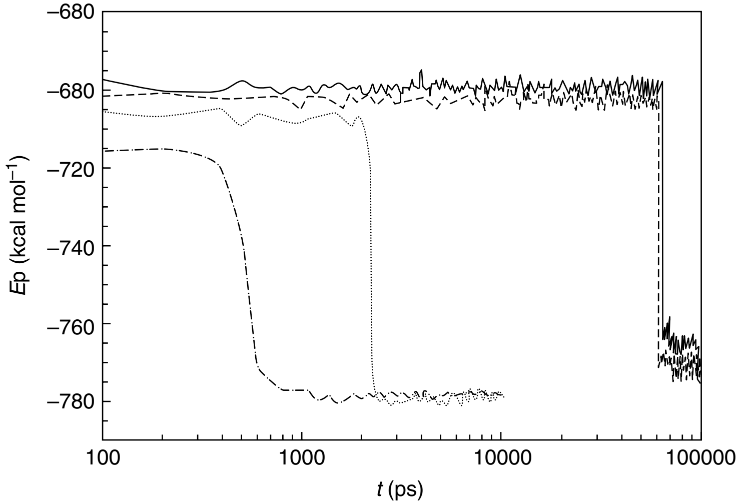


Figure 39. The potential energies of the supercooled argon film system ($T^* = 0.501$) during the crystallization simulations [11]. Simulations were $T^* = 0.501$ and with $t_L = 0.2$ ps and different values. The solid line represents the results from a conventional MD simulation. The dashed line, dotted line, and centered lines represent the results in the SGMD simulations with $t_L = 0.2$ ps and $\lambda = 0.02$, 0.05, and 0.1, respectively.

a fcc crystal at 120K, cooling down, and equilibrating at 60K. The equilibrated argon liquid film was simulated using the conventional MD method and the SGMD method at $t_L = 0.2$ ps and $\lambda = 0.02$, 0.05, and 0.1. In the conventional MD simulation, it took 65 ns before crystallization occurred. In the SGMD simulations, the crystallization occurred at 63, 2, and 0.5 ns with $\lambda = 0.02$, 0.05, and 0.1, respectively. Figure 39 shows the potential energy changes during these simulations. Phase transitions are evident by the sharp decline in potential energy.

Sinoda and Mikami extended the SGMD method to the isothermal–isobaric ensemble and applied it to study crystallization of an argon fluid in a supercooled state [60]. They found that the pressure- and temperature-induced crystallization was considerably accelerated with the use of a suitable parameter set in the SGMD method, as long as the system is not in a glass state.

Production of amorphous silicon has been simulated with SGMD [55, 56, 66]. Choudhary and Clancy used the SGMD method to study evolution of a quenched sample of liquid silicon. The validity of the results using SGMD was provided by comparison to a conventional MD algorithm simulated under constant temperature conditions for more than 100 ns. They found that it was important to perform a sensitivity analysis of the effect of the SGMD parameters before applying the self-guided MD scheme. They demonstrated that using a suitable set of parameters in

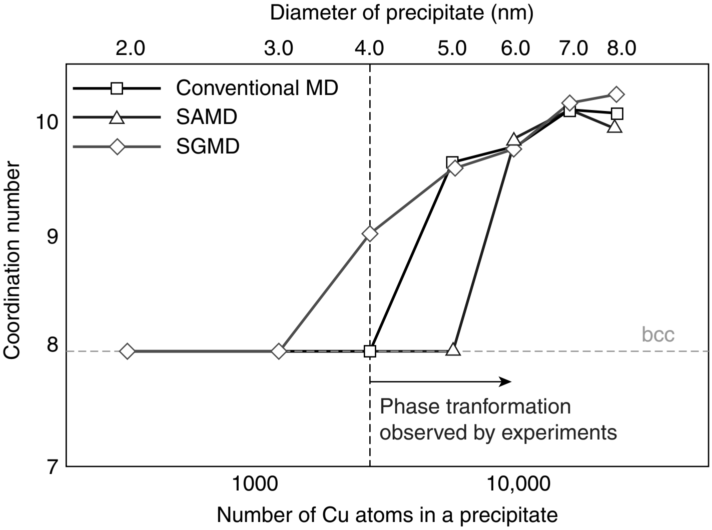


Figure 40. Comparison of phase transition process in different simulations [54]. Coordination number with respect to the size of Cu precipitates in conventional MD, simulated annealing MD, and SGMD ($\lambda = 0.1$, $t_L = 0.2$ ps).

the SGMD method improved the structural evolution compared to a conventional MD scheme, even in the glass state. They concluded that SGMD provides an important tool for observing the evolution of slowly changing processes.

SGMD was used to study phase transformation of Cu precipitate in Fe-Cu alloy [54, 57]. It was shown that the SGMD method can accelerate calculating the bcc to 9R structure transformation of a small precipitate, enabling the transformation without introducing any excess vacancies. Figure 40 compares the size of the Cu precipitate at which the phase transformation occurs in conventional MD, simulated annealing MD (SAMD), and SGMD. In conventional MD and SAMD, phase transformation occurred when the precipitate was larger than 5.0 and 6.0 nm, respectively. However, in the SGMD simulation, the size of the Cu precipitate needed to change the coordination number was 4.0 nm, which is in good agreement with the lower bound of what was experimentally observed.

VI. SUMMARY

Since their development, the SGMD [10, 11] and SGLD [12] methods have been employed to study many slow processes and events. A theoretical understanding of the methods was achieved only after recent progress in quantitative description of SGLD ensembles [13, 14]. The low-frequency motion defined by the local averaging time, t_L, determines the efficiency of conformational search. Energy barriers and diffusion limits are among the causes of slow low-frequency motion. The enhancement in conformational search efficiency in SGLD simulations is achieved through transferring kinetic energy from high-frequency degrees of freedom to low-frequency degrees of freedom so that the low-frequency motion is accelerated while the high-frequency motion is suppressed. Once a barrier is crossed, the excess energy is returned to the high-frequency degrees of freedom. We refer to this effect as "energy borrowing," and it can occur with minimal effect on the overall conformational distribution.

The guiding force can have various compositions that will lead to various forms of SGLD simulation method (Table I) depending on how the guiding force is calculated. When the friction constant is reduced to zero, an LD simulation is reduced to an MD simulation and the SGLD simulation method is transferred to the SGMD simulation method. Depending on how the guiding force is calculated, SGLD can be transformed to SGLDf and SGLDfp. And when the friction constant reduces to zero, SGLD is transformed to SGMDf, SGMDp, or SGMDfp. When only nonbonded forces are used for the guiding force calculation, it is transformed to the original SGMD simulation method.

The partition function of an SGLD ensemble can be expressed with the low- and high-frequency properties. From the SGLD partition function, we can convert SGLD conformational distribution to a canonical conformational distribution, and canonical ensemble averages can be calculated in SGLD simulations through reweighting on-the-fly or during postprocessing. It should be noted that the reweighting approach becomes intractable for large systems where the range of the reweighting factors can be large. In this case, the convergence is poor because the reweighting approach is not size extensive.

The SGLDfp method incorporates both the local average momentum and the local average force in such a way that a canonical ensemble conformational distribution is directly sampled. Therefore, the SGLDfp approach can directly sample the canonical conformational space while accelerating conformational search and can be used in conjunction with many other techniques, such as umbrella sampling or free energy perturbation, to improve convergence. The SGLDfp approach is seen to be size extensive. Doubling the size of the system does not seem to impact the quality of the distribution.

The enhanced conformational search ability can be measured by the self-guiding temperature, T_{sg}, which is calculated with the low- and high-frequency temperature of an SGLD simulation. An SGLD simulation with a self-guiding temperature of T_{sg} will have a conformational searching ability comparable to a high-temperature simulation at $T = T_{sg}$. In a typical SGLD simulation, one can set λ to make $T_{sg} = 2T$, while in a typical SGLDfp simulation one can set λ to achieve $T_{sg} = 1.2T$. In other words, a typical SGLD simulation has an enhanced conformational searching ability comparable to a high-temperature simulation with its temperature doubled.

The performance of an SGLD simulation can be turned with three parameters, the guiding factor, λ, the local averaging time, t_L, and the collision frequency, γ. The parameter λ determines the strength of the guiding effect and is recommended to take values between 0 and 1. When $\lambda = 0$, an SGLD simulation reduces to an LD simulation and if $\gamma = 0$, an SGLD simulation reduces to an MD simulation. The parameter t_L determines which low-frequency motions will be enhanced and which high-frequency motions will be suppressed. $t_L = 0.2$ ps is default for an SGLD simulation that has been used for secondary structure folding simulations. For lower frequency motions such as protein domain motion, larger t_L, say 1 ps, would be more suitable. This value range for various types of molecular motions will be the topic of future studies. γ is related to the diffusion in a simulation system. Also, γ is a factor in the guiding force calculation. Therefore, increasing γ will slow down thermal diffusion and increase the guiding effect. Considering these competing two effects, there is an optimal γ value that maximizes the conformational search ability.

Temperature-based replica exchange method has been widely used in conformational search and sampling. However, for large systems, many replicas with small temperature difference are needed to have reasonable transition rates. The quantitative understanding of the SGLD partition function makes it possible to perform guiding factor-based replica exchange simulation at a constant temperature.

The SGLDfp is unique in that it greatly enhances sampling while directly preserving the canonical ensemble. It is an ideal approach to problems where ensemble distribution preservation is critical, such as protein folding and pathway studies, or when computing free energies.

SGLD and SGLDfp can also be used in conjunction with many other sampling techniques [70–73] that currently rely on MD or LD to sample conformational space. As an efficient and accurate simulation approach, we believe SGLD will play an important role in molecular simulation studies of processes such as protein folding, structure prediction, conformational arrangements, free energy calculations, binding mode prediction, and protein function studies.

Acknowledgment

This research was supported by the Intramural Research Program of the NIH, NHLBI. A.D. was partially supported by NIH Grant RO1 GM073838 to Bertrand Garcia-Moreno at Johns Hopkins University.

References

1. C. M. Dobson and M. Karplus, *Curr. Opin. Struct. Biol.*, **9**, 92–101 (1999).
2. S. A. Adcock and J. A. McCammon, *Chem. Rev.*, **106**, 1589–615 (2006).
3. M. Christen and W. F. Van Gunsteren, *J. Comput. Chem.*, **29**, 157–166 (2008).
4. N. Foloppe and I. J. Chen, *Curr. Med. Chem.*, **16**, 3381–3413 (2009).
5. Y.Q. Gao, L. Yang, Y. Fan, and Q. Shao, *Int. Rev. Phys. Chem.*, **27**, 201–227 (2008).

6. K. Klenin, B. Strodel, D.J. Wales, and W. Wenzel, *Biochim. Biophys. Acta*, **1814**(8), 977–1000 (2011).
7. A. Liwo, C. Czaplewski, S. Oldziej, and H. A. Scheraga, *Curr. Opin. Struct. Biol.* **18**, 134–139 (2008).
8. J. Norberg and L. Nilsson *Q. Rev. Biophys.*, **36**, 257–306 (2003).
9. K. Tai, *Biophys. Chem.*, **107**, 213–220 (2004)
10. X. Wu and S. Wang, *J. Phys. Chem. B*, **102**, 7238–7250 (1998).
11. X. Wu and S. Wang, *J. Chem. Phys.*, **110**, 9401–9410 (1999).
12. X. Wu and B. R. Brooks, *Chem. Phys. Lett.*, **381**, 512–518 (2003).
13. X. Wu and B. R. Brooks, *J. Chem. Phys.,* **135**, 204101 (2011a).
14. X. Wu and B.R. Brooks, *J. Chem. Phys.*, **134**, 134108 (2011b).
15. N. Budin, S. Ahmed, N. Majeux, and A. Caflisch, *Comb. Chem. High Throughput Screen.*, **4**, 661–673 (2001).
16. I. Kolossvary and W. C. Guida, *J. Comput. Chem.*, **20**, 1671–1684 (1999).
17. M. J. Loferer, I. Kolossvary, and A. Aszodi, *J. Mol. Graph. Model.*, **25**, 700–710 (2007).
18. C. McMartin and R. S. Bohacek, *J. Comput. Aided Mol. Des.*, **11**, 333–344 (1997).
19. D. C. Spellmeyer, A. K. Wong, M. J. Bower, and J. M. Blaney, *J. Mol. Graph. Model.* **15**, 18–36 (1997).
20. M. L. Beckers, L. M. Buydens, J. A. Pikkemaat, and C. Altona *J. Biomol. NMR*, **9**, 25–34 (1997).
21. T. Dandekar and P. Argos, *Protein Eng.*, **5**, 637–645 (1992).
22. G. Jones, P. Willet, R. C. Glen, and A. R. Leach, *J. Mol. Biol.*, **267**(3), 727–748 (1997).
23. S. M. Le Grand and K. M. Merz, Jr., The genetic algorithm and protein tertiary structure prediction, in *The Protein Folding Problem and Ternary Structure Prediction*, K. M. Merz, ed., Birkhauser, Boston, 1994, 109–124.
24. H. Ogata, Y. Akiyama, and M. Kanehisa, *Nucleic Acids Res.* **23**, 419–426 (1995).
25. R. C. Brower, G. Vasmatzis, M. Silverman, and C. Delisi, *Biopolymers*, **33**, 329–334 (1993).
26. W. M. Brown, J. L. Faulon, and K. Sale, *Comput. Biol. Chem.*, **29**, 143–150 (2005).
27. Y. Duan and P. A. Kollman, *IBM Syst. J.*, **40**, 297–309 (2001).
28. J. L. Faulon, K. Sale, and M. Young, *Protein Sci.*, **12**, 1750–1761 (2003).
29. J. Lee, A. Liwo, D. R. Ripoll, *Proteins Suppl*, **3**, 204–208 (1999a).
30. J. Lee, A. Liwo, and H. A. Scheraga, *Proc. Natl. Acad. Sci. USA*, **96**, 2025–2030 (1999b).
31. Y. Sugita and Y. Okamoto, *Chem. Phys. Lett.*, **314**, 141–151 (1999).
32. M. S. Lee and M. A. Olson, *J. Chem. Theory Comput.*, **6**, 2477–2487 (2010).
33. H. Fukunishi, O. Watanabe, and S. Takada, *J Chem. Phys.*, **116**, 9058–9067 (2002).
34. G. Bussi, F. L. Gervasio, A. Laio, and M. Parrinello, *J. Am. Chem. Soc.*, **128**, 13435–13441 (2006).
35. J. Schlitter, M. Engels, and P. Kruger, *J. Mol. Graph.*, **12**, 84–89 (1994).
36. A. Stirling, M. Iannuzzi, A Laio, and M. Parrinello, *ChemPhysChem*, **5**, 1558–1568 (2004).
37. A. D. MacKerell Jr., D. Bashford, M. Bellott, R. L. Dunbrack Jr., J. D. Evanseck, et al., *J. Phys. Chem. B*, **102**, 3586–3616 (1998).
38. C. I. Lee and N. Y. Chang, *Biophys. Chem.*, **151**, 86–90 (2010).
39. E. Z. Wen, M. J. Hsieh, P. A. Kollman, and R. Luo, *J. Mol. Graph. Model.*, **22**, 415–424 (2004).
40. E. Z. Wen and R. Luo *J. Chem. Phys.*, **121**, 2412–2421 (2004).
41. X.-W. Wu and S.-S. Sung, *Proteins*, **34**, 295–302 (1999).

42. X. Wu and B. R. Brooks, *Biophys. J.*, **86**, 1946–1958 (2004).
43. X. Wu and S. Wang, *J. Phys. Chem. B*, **104**, 8023–8034 (2000).
44. X. Wu and S. Wang, *J. Phys. Chem. B*, **105**, 2227–2235 (2001).
45. X. Wu, S. Wang, and B. R. Brooks, *J. Am. Chem. Soc.*, **124**, 5282–5283 (2002).
46. F. D. T Lung, Y. Q. Long, P. P. Roller, C. R. King, and J. Varady, et al., *J. Pept. Res.* **57**, 447–454 (2001).
47. J. Varady, X. Wu, and S. Wang, *J. Phys. Chem. B*, **106**, 4863–4872 (2002).
48. C. Y. Yang, Z. Nikolovska-Coleska, P. Li, P. Roller, and S. Wang, *J. Phys. Chem. B*, **108**, 1467–1477 (2004).
49. V. Chandrasekaran, C. J. Lee, P. Lin, R. E. Duke, and L. G. Pedersen, *J. Mol. Model.*, **15**, 897–911 (2009).
50. A. Damjanovic, E. B. García-Moreno, and B. R. Brooks, *Proteins*, **76**, 1007–1019 (2009).
51. A. Damjanovic, B. T. Miller, T. J. Wenaus, P. Maksimovic, E. Bertrand García-Moreno, and B. R. Brooks, *J. Chem. Inf. Model.*, **48**, 2021–2029 (2008a).
52. A. Damjanovic, X. Wu, E. B. García-Moreno, and B. R. Brooks, *Biophys. J.*, **95**, 4091–4101 (2008b).
53. P. Y. Pendse, B. R. Brooks, and J. B. Klauda, *J. Mol. Biol.*, **404**, 506–521 (2010).
54. Y. Abe and S. Jitsukawa, *Philos. Mag. Lett.*, **89**, 535–543 (2009).
55. D. Choudhary and P. Clancy, *J. Chem. Phys.* **122**, 154509 (2005a).
56. D. Choudhary, P. Clancy, *J. Chem. Phys.*, **122**, 1–10 (2005b).
57. T. Tsuru, A. B. E Yosuke, Y. Kaji, T. Tsukada, and S. Jitsukawa, *Zairyo/J. Soc. Mater. Sci.*, **59**, 583–588 (2010).
58. Y. Sheng, W. Wang, and P. Chen, *J. Phys. Chem. C*, **114**, 454–459 (2010a).
59. Y. Sheng, W. Wang, and P. Chen, *Protein Sci.*, **19**, 1639–1648 (2010b).
60. W. Shinoda and M. Mikami, *Chem. Phys. Lett.*, **335**, 265–272 (2001).
61. W. Shinoda and M. Mikami, *J. Comput. Chem.*, **24**, 920–930 (2003).
62. A. Lahiri, L. Nilsson, and A. Laaksonen, *J. Chem. Phys.*, **114**, 5993–5999 (2001).
63. I. Andricioaei, A. R. Dinner, and M. Karplus, *J. Chem. Phys.*, **118**, 1074–1084 (2003).
64. M. P. Allen and D. J. Tildesley, Computer Simulations of Liquids, Clarendon Press, Oxford, 1987.
65. R. W. Pastor, B. R. Brooks, and A. Szabo, *Mol. Phys.*, **65**, 1409–1419 (1988).
66. D. Choudhary and P. Clancy, Proceedings of the 2002 International Conference on Computational Nanoscience and Nanotechnology (ICCN 2002) 2002, pp. 159–162.
67. S. Chowdhury, W. Zhang, C. Wu, G. Xiong, and Y. Duan, *Biopolymers*, **68**, 63–75 (2003).
68. L. Yang and Y. Q. Gao, *J. Phys. Chem. B*, **111**, 2969–2975 (2007).
69. J. MacFadyen, J. Wereszczynski, and I. Andricioaei, *J. Chem. Phys.*, **128**, 114112 (2008).
70. H. Li, D. Min, Y. Liu, and W. Yang, *J. Chem. Phys.*, **127**, 094101 (2007).
71. D. Min and W. Yang, *J. Chem. Phys.*, **128**, 094106 (2008).
72. L. Zheng, M. Chen, and W. Yang, *J. Chem. Phys.*, **130**, 234105 (2009).
73. L. Zheng and W. Yang, *J. Chem. Phys.*, **129**, 014105 (2008).
74. D. Min, M. Chen, L. Zheng, Y. Jin, M. A. Schwartz, et al. *J Phys. Chem. B*, **115**(14), 3924–3935 (2011).
75. B. R. Brooks, C. L. Brooks, III, A. D. Mackerell, Jr., L. Nilsson, R. J. Petrella, et al., *J. Comput. Chem.*, **30**, 1545–1614 (2009).

76. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, et al., *J. Comput. Chem.*, **4**, 187–217 (1983).

77. T. Shen and D. Hamelberg, *J. Chem. Phys.*, **129**, 034103 (2008).

77a. D. A. Case, T. A. Darden, T. E. Cheatham Iii, C. L. Simmerling, J. Wang, R. Duke, R. Luo, K. M. Merz, D. A. Pearman, M. Crowley, B. Walker, B. Wang, S. Hayik, A. Roitberg, G. Seabra, X. Wu, S. Brozell, V. Tsui, H. Gohlke, J. Mongan, V. Hornak, G. Cui, P. Beroza, D.H. Mathews, C. Schafmeister, W. S. Ross, and P. A. Kollman, *AMBER* 9. (University of California, San Francisco, 2006).

78. X. Wu and B. R. Brooks. *J. Chem. Phys.*, **122**, 44107 (2005).

79. J. P. Ryckaert, G. Ciccotti, and H. J. C Berendsen, *J. Comput. Phys.*, **23**, 327–341 (1977).

80. H. J. Dyson, M. Rance, R. A. Houghten, P. E. Wright, and R. A. Lerner, *J. Mol. Biol.* **201**, 201–217 (1988).

81. D. J. Tobias, J. E. Mertz, C. L. Brooks, 3rd., *Biochemistry*, **30**, 6054–6058 (1991).

82. J. M. Scholtz, E. J. York, J. M. Stewart, and R. L. Baldwin, *J. Am. Chem. Soc.*, **113**, 5104 (1991).

83. F. J. Blanco, M. A. Jimenez, J. Herranz, M. Rico, J. Santoro, J. L. Nieto. *J. Am. Chem. Soc.*, **115**, 5887–5888 (1993).

84. J. Shao, S. W. Tanner, N. Thompson, and T. E. Cheatham, III, *J. Chem. Theory Comput.*, **3**, 2312–2334 (2007).