Toward canonical ensemble distribution from self-guided Langevin dynamics simulation

Xiongwu Wu[a] and Bernard R. Brooks
Laboratory of Computational Biology, National Heart, Lung, and Blood Institute (NHLBI), National Institutes of Health (NIH), Bethesda, Maryland 20892, USA

(Received 10 December 2010; accepted 16 March 2011; published online 6 April 2011)

This work derives a quantitative description of the conformational distribution in self-guided Langevin dynamics (SGLD) simulations. SGLD simulations employ guiding forces calculated from local average momentums to enhance low-frequency motion. This enhancement in low-frequency motion dramatically accelerates conformational search efficiency, but also induces certain perturbations in conformational distribution. Through the local averaging, we separate properties of molecular systems into low-frequency and high-frequency portions. The guiding force effect on the conformational distribution is quantitatively described using these low-frequency and high-frequency properties. This quantitative relation provides a way to convert between a canonical ensemble and a self-guided ensemble. Using example systems, we demonstrated how to utilize the relation to obtain canonical ensemble properties and conformational distributions from SGLD simulations. This development makes SGLD not only an efficient approach for conformational searching, but also an accurate means for conformational sampling. [doi:10.1063/1.3574397]

I. INTRODUCTION

The self-guided Langevin dynamics simulation method[1] was developed for efficient conformational searching so that rare events, such as protein folding and ligand binding, can be accessed with much less computing resources. It has been successfully applied to a range of computational studies.[2,3] While it can accelerate slow events to an affordable time scale, the perturbation in conformational distribution from the self-guiding force remained a major concern. For some calculations, such as free energy simulation, conformational search efficiency is a crucial factor to obtain convergent results, while the correct conformational distribution is responsible for accuracy.

Because the guiding force is calculated from the so called local averages, it has been a difficult task to quantitatively understand the effect of the guiding force on ensemble distributions. A common practice for self-guided Langevin dynamics (SGLD) simulation is to limit the guiding factor to a small range so that the effect on conformational distribution is very small and can be neglected.[1] Without a quantitative understanding of the perturbation on conformational distribution, it is difficult to take full advantage of the acceleration that SGLD simulations can achieve.

To obtain correct thermodynamic average properties, Andricioaei *et al.* proposed a Monte Carlo procedure called the momentum-enhanced hybrid Monte Carlo method to include the benefit of the guiding force while preserving the ensemble average properties.[4] In dynamics simulations, the difficulty in characterizing the guiding force effect on ensemble distributions is mainly due to the lack of quantitative definition on the low-frequency motion to be enhanced. To tackle the problem in dynamics simulations, this work proposes a way to separate low-frequency and high-frequency portions of thermodynamic properties through the local averaging procedure. Based on this separation, this work derives a quantitative relation between conformational distribution and guiding parameters. The details of the derivation are described in Sec. II. Examples of applying this relation are provided in Sec. IV.

[a]Author to whom correspondence should be addressed. Electronic mail: wuxw@nhlbi.nih.gov. Present address: Laboratory of Computational Biology, NHLBI, NIH, 5635 Fishers Lane, Room T904, Bethesda, MD 20892. Telephone: 301-451-6251. FAX: 301-480-6496.

II. THEORY AND METHOD

A. The low-frequency and high-frequency properties

Thermal motion in a molecular system has a distribution of frequencies. Chemical bonds vibrate and bend at high frequencies, while ion translation and protein folding events take a relatively long time to happen. High-frequency events repeat on a short time scale and are often most easy to study in molecular simulations. Low-frequency events are important for many macroscopic behaviors, such as protein folding and binding, but often are beyond the time scale accessible by molecular simulations with available computing resources.

Low-frequency properties are related to low-frequency events. For example, interaction between a pair of water molecules depends on the relative position between the water molecules. This interaction energy means the energy at zero frequency, i.e., the average among all bond vibration and bending states. At each given moment, bond vibration and bending, and even electron density fluctuation, produce an instantaneous energy deviation, which depends on the high-frequency motions and is called the high-frequency energy. 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,[1,5] typically on force or momentum, is performed by the following equation:

$$\begin{aligned}\langle \mathbf{p} \rangle_L &= \frac{1}{L}\sum_{i=n-L+1}^{n} \mathbf{p}_i = \frac{1}{t_L}\int_{t-t_L}^{t} \mathbf{p}(\tau)d\tau \\ &\approx \left(1-\frac{1}{L}\right)\tilde{\mathbf{p}}_{n-1} + \frac{1}{L}\mathbf{p}_n \\ &= \left(1-\frac{\delta t}{t_L}\right)\tilde{\mathbf{p}}(t-\delta t) + \frac{\delta t}{t_L}\mathbf{p}(t) = \tilde{\mathbf{p}}. \end{aligned} \tag{1}$$

As can be seen from Eq. (1), a local average, denoted as "$\langle\,\rangle_L$", is calculated by averaging over the most recent L points, or the most recent $t_L = L\delta t$ time period. Here, δt is the time interval between data points. We call L as the local averaging size and t_L as the local average time. This average can be approximately calculated as an evolving average with a constant updating of the current value as shown in the right hand portion of Eq. (1). This evolving average is denoted with a "$\sim$" cap: $\tilde{\mathbf{p}}$. Because all local averages in this work are calculated as evolving averages, we also use "$\langle \mathbf{p} \rangle_L$" 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: $\mathbf{p} - \tilde{\mathbf{p}}$.

The local averaging shown in Eq. (1) suppresses high-frequency effects and emphasizes low-frequency contributions. From Eq. (1) we can see that the local average time, t_L, determines the contribution frequency range. To better understand the evolving averaging, we can rearrange Eq. (1) to the following form:

$$\frac{\tilde{\mathbf{p}}(t) - \tilde{\mathbf{p}}(t-\delta t)}{\delta t} = \frac{\mathbf{p}(t) - \tilde{\mathbf{p}}(t-\delta t)}{t_L}.$$

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

$$\frac{d\tilde{\mathbf{p}}(t)}{dt} = \frac{\mathbf{p}(t) - \tilde{\mathbf{p}}(t)}{t_L}.$$

This differential equation can be solved:

$$\tilde{\mathbf{p}}(t) = \frac{1}{t_L}\int_0^t \mathbf{p}(\tau) e^{-\frac{t-\tau}{t_L}} d\tau. \tag{2}$$

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 average time, t_L.

The separation of the low-frequency properties and the high-frequency properties is at the center of the SGLD simulation method. The low-frequency properties are calculated through the evolving averaging shown in Eq. (1). To explain the behavior of the evolving averaging, we use $q(t) = \sin(2\pi\varpi t)$ as an example function of frequency ϖ to show how frequency and local average time affect the evolving average.

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

$$\tilde{q}(t) = \frac{2\pi\varpi t_L(e^{-t/t_L} - \cos(2\pi\varpi t)) + \sin(2\pi\varpi t)}{1 + 4\pi^2 t_L^2 \varpi^2}. \tag{3}$$

As can be seen from Eq. (3) that for high frequency, $2\pi\varpi t_L \gg 1$, the amplitude of $\tilde{q}(t)$ is inversely proportional to ϖ, while for low frequency, $2\pi\varpi t_L \ll 1$, $\tilde{q}(t) \approx q(t)$. The local average time, t_L, defines the separation of what is high frequency and what is low frequency as compared with a local averaging frequency of $\varpi_L = 1/t_L$. 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

$$\begin{aligned} &q(t) - \tilde{q}(t) \\ &= \frac{-2\pi\varpi t_L(e^{-t/t_L} - \cos(2\pi\varpi t)) + 4\pi^2 t_L^2 \varpi^2 \sin(2\pi\varpi t)}{1 + 4\pi^2 t_L^2 \varpi^2}. \end{aligned} \tag{4}$$

As can be seen from Eq. (4), when $2\pi\varpi t_L \gg 1$, $q(t) - \tilde{q}(t) \approx \sin(2\pi\varpi t) = q(t)$, and when $2\pi\varpi t_L \ll 1$, $q(t) - \tilde{q}(t) \approx -2\pi\varpi t_L(e^{-t/t_L} - \cos(2\pi\varpi t)) \to 0$. That is, the high-frequency portion keeps the high-frequency contributions while suppressing the low-frequency components.

Figure 1(a) shows the example function and its evolving averages at different local average times. Clearly, we 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 $\varpi t_L = 0.1$, this function represents a low-frequency motion and its evolving average has a magnitude similar to the function. When $\varpi t_L = 10$, this function represents a high-frequency motion and the magnitude of its evolving average is very small as compared to the function. Figure 1(b) shows an averaging result as a function of ϖt_L. The envelope function represents the amplitude of the averages. Clearly we can see, with a small ϖt_L, the amplitude of the average has little change from the example function, while with a large ϖt_L, the amplitude of the average approaches zero, indicating that low-frequency function will remain in the evolving average and high-frequency function will be suppressed.

With the evolving averaging, many low-frequency properties can be obtained in molecular simulation. For example, low-frequency forces:

$$\tilde{\mathbf{f}}_i(t) = \left(1-\frac{\delta t}{t_L}\right)\tilde{\mathbf{f}}_i(t-\delta t) + \frac{\delta t}{t_L}\mathbf{f}_i(t);$$

low-frequency momentums:

$$\tilde{\mathbf{p}}_i(t) = \left(1-\frac{\delta t}{t_L}\right)\tilde{\mathbf{p}}_i(t-\delta t) + \frac{\delta t}{t_L}\mathbf{p}_i(t);$$

and low-frequency potential energies:

$$\tilde{E}_p(t) = \left(1-\frac{\delta t}{t_L}\right)\tilde{E}_p(t-\delta t) + \frac{\delta t}{t_L}E_p(t).$$

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

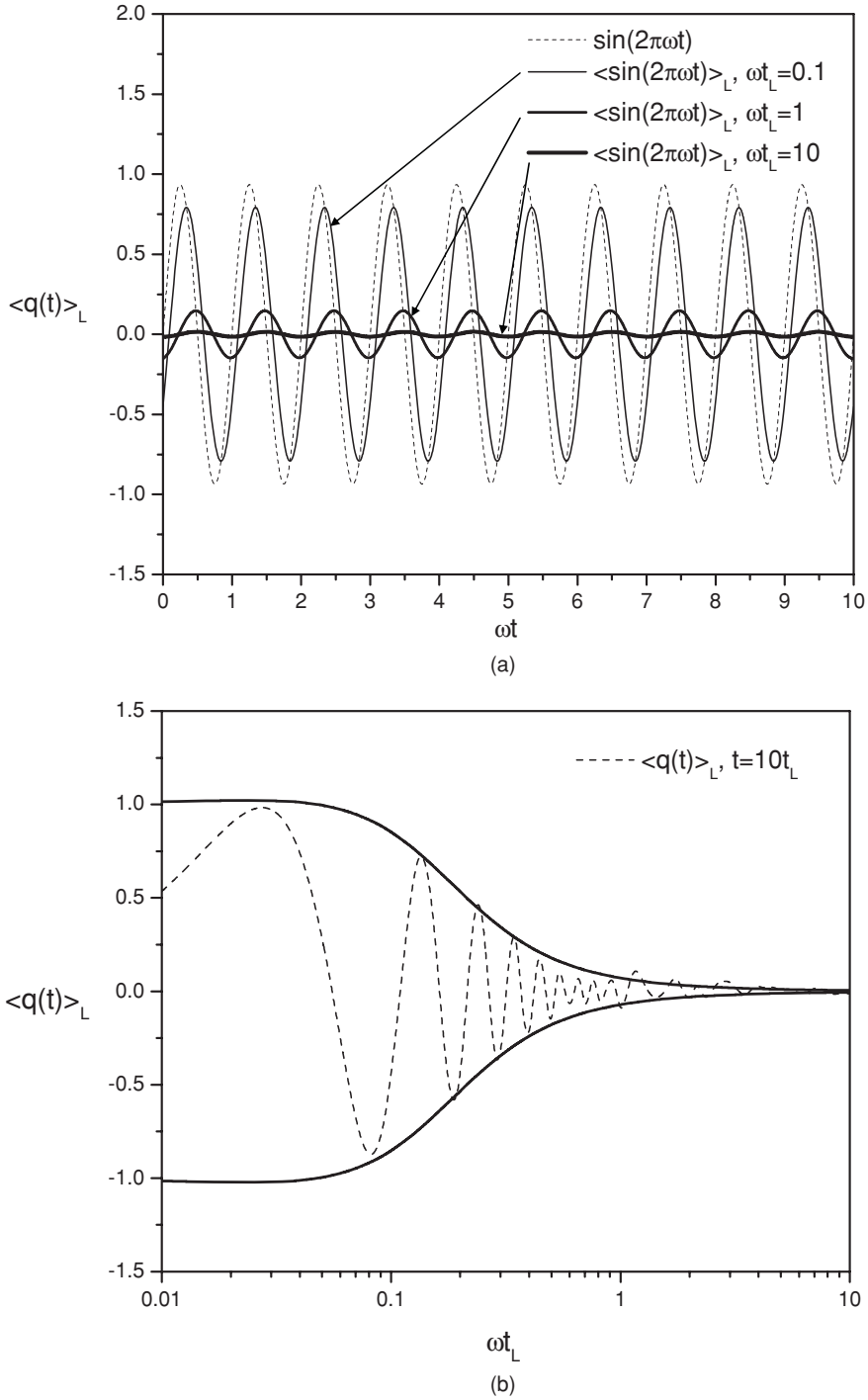

FIG. 1. (a) The example function, $q(t) = \sin(2\pi\varpi t)$, and its evolving averages at three local average times: $\varpi t_L = 0.1$, 1, and 10. (b) The evolving average of the example function as a function of the frequency. The envelope curves show the amplitude as a function of ϖt_L. At small 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.

kinetic energies:

$$\tilde{E}_k = \frac{1}{2}\sum_i \frac{\tilde{p}_i^2}{m_i}, \tag{5}$$

and low-frequency temperature:

$$\tilde{T} = \frac{\tilde{E}_k}{N_{\mathrm{DF}}k}. \tag{6}$$

Here, N_{DF} is the number of degree of freedom and k is the Boltzmann constant.

B. The self-guided Langevin dynamics

Langevin dynamics (LD) is based on the following equation of motion:

$$\dot{\mathbf{p}}_i = \mathbf{f}_i - \gamma_i \mathbf{p}_i + \mathbf{R}_i, \tag{7}$$

where $\dot{\mathbf{p}}_i$ and $\mathbf{f}_i$ are the time derivative of the momentum and the interaction force of particle i, respectively. $\mathbf{R}_i$ is a random force, which is related to mass, m_i, the collision frequency, γ_i, and simulation temperature, T, by the following equation:

$$\langle \mathbf{R}_i(0)\mathbf{R}_i(t)\rangle = 2m_i kT\gamma_i\delta(t). \tag{8}$$

By adding a guiding force, we obtain the equation of motion for SGLD:[1]

$$\dot{\mathbf{p}}_i = \mathbf{f}_i + \mathbf{g}_i - \gamma_i\mathbf{p}_i + \mathbf{R}_i, \tag{9}$$

$\mathbf{g}_i$ is called the guiding force and is calculated based on the low-frequency momentum:

$$\mathbf{g}_i(t) = \lambda_i\gamma_i(\tilde{\mathbf{p}}_i(t) - \xi\mathbf{p}_i(t)). \tag{10}$$

Here, λ_i is the guiding factor. The parameter, ξ, is an energy conservation factor to cancel the energy input from the guiding force,

$$\sum_i \mathbf{g}_i\cdot\dot{\mathbf{r}}_i = \sum_i \lambda_i\gamma_i\tilde{\mathbf{p}}_i\cdot\dot{\mathbf{r}}_i - \xi\sum_i \lambda_i\gamma_i\mathbf{p}_i\cdot\dot{\mathbf{r}}_i = 0. \tag{11}$$

Here, the summation runs over all particles in a simulation system. From Eq. (11) we have

$$\xi = \frac{\sum_i \lambda_i\gamma_i\tilde{\mathbf{p}}_i\cdot\dot{\mathbf{r}}_i}{\sum_i \lambda_i\gamma_i\mathbf{p}_i\cdot\dot{\mathbf{r}}_i}. \tag{12}$$

C. Conformational distribution in SGLD

The guiding force in a SGLD simulation is designed to accelerate the low-frequency motion so the conformational search efficiency can be enhanced. It has two types of effects on a simulation system. First, the guiding force enhances the low-frequency motion as measured by the increase in the low-frequency temperature, and also it reduces the high-frequency motion due to the energy conservation force that comes with the guiding force [see Eq. (10)]. Second, the guiding force produces a bias in the energy surface. Based on these two effects, the partition function of a SGLD ensemble is split into low-frequency and high-frequency parts:

$$\Theta_{\mathrm{SGLD}} = \sum \Omega \exp\left(-\frac{\lambda_{\mathrm{lf}}\tilde{E}_p}{kT_{\mathrm{lf}}} - \frac{\lambda_{\mathrm{hf}}(E_p - \tilde{E}_p)}{kT_{\mathrm{hf}}}\right). \tag{13}$$

Here, λ_{lf} is called the low-frequency energy factor, describing the energy bias in the low-frequency energy surface, $\tilde{E}_p$, and λ_{hf} is the high-frequency energy factor, describing the energy bias in the high-frequency energy surface, $E_p - \tilde{E}_p$. The low-frequency and high-frequency energy surfaces under the guiding effect are $\lambda_{\mathrm{lf}}\tilde{E}_p$ and $\lambda_{\mathrm{hf}}(E_p - \tilde{E}_p)$, respectively. T_{lf} and T_{hf} are the effective temperatures in low-frequency and high-frequency conformational spaces, respectively. In normal conditions without the guiding forces ($\lambda = 0$), $\lambda_{\mathrm{lf}} = \lambda_{\mathrm{hf}} = 1$, and $T_{\mathrm{lf}} = T_{\mathrm{hf}} = T$, we have:

$$\Theta_{\mathrm{SGLD}}(\lambda = 0) = \sum \Omega \exp\left(-\frac{\tilde{E}_p}{kT} - \frac{(E_p - \tilde{E}_p)}{kT}\right)$$
$$= \sum \Omega \exp\left(-\frac{E_p}{kT}\right) = \Theta_{\mathrm{LD}}.$$

In the low-frequency conformational space, the equation of motion can be expressed as an evolving averaging of Eq. (9):

$$\dot{\tilde{\mathbf{p}}}_i = \tilde{\mathbf{f}}_i + \tilde{\mathbf{g}}_i - \gamma_i\tilde{\mathbf{p}}_i + \tilde{\mathbf{R}}_i. \tag{14}$$

The low-frequency energy factor, λ_{lf}, can be calculated according to the projection of the total low-frequency force in the direction of the low-frequency forces:

$$\lambda_{\mathrm{lf}} = \frac{\left\langle \sum_i (\tilde{\mathbf{f}}_i + \tilde{\mathbf{g}}_i - \gamma_i\tilde{\mathbf{p}}_i)\tilde{\mathbf{f}}_i \right\rangle}{\left\langle \sum_i \tilde{\mathbf{f}}_i\tilde{\mathbf{f}}_i \right\rangle}. \tag{15}$$

Similarly, in the high-frequency conformational space, the equation of motion can be expressed as the difference between the instantaneous motion, Eq. (9), and the low-frequency motion Eq. (14):

$$\dot{\mathbf{p}}_i - \dot{\tilde{\mathbf{p}}}_i = \mathbf{f}_i - \tilde{\mathbf{f}}_i + \mathbf{g}_i - \tilde{\mathbf{g}}_i - \gamma_i(\mathbf{p}_i - \tilde{\mathbf{p}}_i) + \mathbf{R}_i - \tilde{\mathbf{R}}_i. \tag{16}$$

The high-frequency energy factor, λ_{hf}, can be calculated according to the projection of the total high-frequency force in the direction of the high-frequency forces:

$$\lambda_{\mathrm{hf}} = \frac{\left\langle \sum_i (\mathbf{f}_i - \tilde{\mathbf{f}}_i + \mathbf{g}_i - \tilde{\mathbf{g}}_i - \gamma_i(\mathbf{p}_i - \tilde{\mathbf{p}}_i))(\mathbf{f}_i - \tilde{\mathbf{f}}_i) \right\rangle}{\left\langle \sum_i (\mathbf{f}_i - \tilde{\mathbf{f}}_i)(\mathbf{f}_i - \tilde{\mathbf{f}}_i) \right\rangle}. \tag{17}$$

Next, let us examine the guiding force effect on low-frequency and high-frequency motions. Under the guiding force, a system experiences an enhanced motion in the low-frequency conformational space. This motion in the low-frequency conformational space can be measured by the low-frequency temperature [see Eq. (6)]. It is reasonable to assume that the effective temperature in the local frequency conformational space is proportional to the low-frequency temperature:

$$T_{\mathrm{lf}} = C_{\mathrm{lf}}\tilde{T}.$$

And the effective temperature in the high-frequency conformational space is proportional to the high-frequency temperature:

$$T_{\mathrm{hf}} = C_{\mathrm{hf}}(T - \tilde{T}).$$

The proportional constants can be estimated from a LD simulation or a SGLD simulation without the guiding forces ($\lambda = 0$), where $T_{\mathrm{lf}} = T_{\mathrm{hf}} = T$:

$$C_{\mathrm{lf}} = \frac{T}{\tilde{T}_0},$$

$$C_{\mathrm{hf}} = \frac{T}{T - \tilde{T}_0}.$$

Here, $\tilde{T}_0$ is the low-frequency temperature when $\lambda = 0$ and is called the reference low-frequency temperature. Based on the definition, we know $\tilde{T}_0$ depends on the simulation condition and the local average time, t_L. Therefore, the partition

function of a SGLD simulation can be written as

$$\Theta_{\mathrm{SGLD}} = \sum \Omega \exp\left(-\frac{\lambda_{\mathrm{lf}}\tilde{T}_0}{\tilde{T}}\frac{\tilde{E}_p}{kT} - \frac{\lambda_{\mathrm{hf}}(T-\tilde{T}_0)}{T-\tilde{T}}\frac{E_p-\tilde{E}_p}{kT}\right). \tag{18}$$

To utilize Eq. (18), in addition to λ_{lf}, λ_{hf}, $\tilde{E}_p$, and $\tilde{T}$ from a SGLD simulation, we also need $\tilde{T}_0$ from a LD simulation or a SGLD simulation with $\lambda = 0$. To avoid this burden, we propose the following way to estimate $\tilde{T}_0$ directly from the same SGLD simulation.

The low-frequency motion, Eq. (14), corresponds to a Langevin dynamics in a low-frequency conformational space and can be rewritten to a Langevin dynamics form:

$$\dot{\tilde{\mathbf{p}}}_i = \tilde{\mathbf{f}}_i - \chi_{\mathrm{lf}}\gamma_i\tilde{\mathbf{p}}_i + \tilde{\mathbf{R}}_i. \tag{19}$$

Equation (19) 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 by the following equation:

$$\chi_{\mathrm{lf}} = \frac{\sum_i(\gamma_i\tilde{\mathbf{p}}_i - \tilde{\mathbf{g}}_i)\gamma_i\tilde{\mathbf{p}}_i}{\sum_i \gamma_i^2\tilde{\mathbf{p}}_i\tilde{\mathbf{p}}_i}. \tag{20}$$

Based on the Langevin dynamics relation, Eq. (8), with a given distribution of random forces, the product of temperature and collision frequency is a constant:

$$T\gamma_i = \frac{\langle \mathbf{R}_i(0)\mathbf{R}_i(t)\rangle}{2m_i k\delta(t)}. \tag{21}$$

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 a 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, from Eq. (21), we have

$$\tilde{T}_0 = \tilde{T}\chi_{\mathrm{lf}}. \tag{22}$$

Equation (22) provides a way to estimate $\tilde{T}_0$ from $\tilde{T}$, which can be calculated directly in a SGLD simulation according to Eq. (6). Therefore, the partition function can be approximated as

$$\Theta_{\mathrm{SGLD}} \approx \sum \Omega \exp\left(-\lambda_{\mathrm{lf}}\chi_{\mathrm{lf}}\frac{\tilde{E}_p}{kT} - \lambda_{\mathrm{hf}}\frac{T-\chi_{\mathrm{lf}}\tilde{T}}{T-\tilde{T}}\frac{E_p-\tilde{E}_p}{kT}\right). \tag{23}$$

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, $\tilde{E}_p$, 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, $E_p - \tilde{E}_p$, is modified by a factor of λ_{hf}. The effective temperature is changed from T to $T_{\mathrm{hf}} = \frac{T-\tilde{T}}{T-\tilde{T}_0}T = \frac{T-\tilde{T}}{T-\chi_{\mathrm{lf}}\tilde{T}}T$.

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

$$\begin{aligned}\Theta_{\mathrm{LD}} &= \sum \Omega \exp\left(-\frac{\tilde{E}_p}{kT} - \frac{E_p-\tilde{E}_p}{kT}\right)\\ &= \sum \Omega \exp\left(-\lambda_{\mathrm{lf}}\frac{\tilde{T}_0}{\tilde{T}}\frac{\tilde{E}_p}{kT} - \lambda_{\mathrm{hf}}\frac{(T-\tilde{T}_0)}{(T-\tilde{T})}\frac{E_p-\tilde{E}_p}{kT}\right)\\ &\quad\times \exp\left(\left(\lambda_{lf}\frac{\tilde{T}_0}{\tilde{T}}-1\right)\frac{\tilde{E}_p}{kT} + \left(\lambda_{\mathrm{hf}}\frac{(T-\tilde{T}_0)}{(T-\tilde{T})}-1\right)\right.\\ &\quad\left.\times\frac{E_p-\tilde{E}_p}{kT}\right).\\ &= \Theta_{\mathrm{SGLD}}\langle w_{\mathrm{SGLD}}\rangle_{\mathrm{SGLD}}\end{aligned} \tag{24}$$

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

$$\begin{aligned}w_{\mathrm{SGLD}} &= \exp\left(\left(\lambda_{\mathrm{lf}}\frac{\tilde{T}_0}{\tilde{T}}-1\right)\frac{\tilde{E}_p}{kT} + \left(\lambda_{\mathrm{hf}}\frac{T-\tilde{T}_0}{T-\tilde{T}}-1\right)\right.\\ &\quad\left.\times\frac{E_p-\tilde{E}_p}{kT}\right)\\ &\approx \exp\left((\lambda_{\mathrm{lf}}\chi_{\mathrm{lf}}-1)\frac{\tilde{E}_p}{kT}\right.\\ &\quad\left.+\left(\lambda_{\mathrm{hf}}\frac{T-\chi_{\mathrm{lf}}\tilde{T}}{T-\tilde{T}}-1\right)\frac{E_p-\tilde{E}_p}{kT}\right).\end{aligned} \tag{25}$$

To calculate the weighting factor according to Eq. (25), we need E_p and $\tilde{E}_p$ for each conformation, as well as $\tilde{T}$, λ_{lf}, λ_{hf}, and χ_{lf} from the SGLD simulation. Once we have the weighting factor, any ensemble average, $\langle A\rangle$, can be calculated in a SGLD simulation as

$$\langle A\rangle = \frac{\langle A w_{\mathrm{SGLD}}\rangle_{\mathrm{SGLD}}}{\langle w_{\mathrm{SGLD}}\rangle_{\mathrm{SGLD}}}. \tag{26}$$

D. The self-guiding temperature

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

$$T_{\mathrm{sg}} = \frac{T_{\mathrm{lf}}}{T_{\mathrm{hf}}}T = \frac{\tilde{T}(T-\tilde{T}_0)}{\tilde{T}_0(T-\tilde{T})}T. \tag{27}$$

The self-guiding temperature, T_{sg}, provides a rough measure of the conformational searching ability in the unit of temperature. A SGLD simulation with a self-guiding temperature of T_{sg} has a conformational search ability comparable to that in a high-temperature simulation at the temperature of T_{sg}. As can be seen from Eq. (27), for a LD simulation, $\tilde{T} = \tilde{T}_0$, we have $T_{\mathrm{sg}} = T$. For a 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 choosing 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 as 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 conformational distribution.

III. SIMULATION DETAILS

To demonstrate the ensemble distribution in SGLD simulations and the conversion to canonical ensembles, we report the results for several simple systems. A leap-frog Verlet algorithm for the SGLD simulation has been implemented into CHARMM (Refs. 6 and 7), version 36 and is described in the Appendix. Because a SGLD simulation involves extra calculation only in the propagation of the equations of motion as compared to a normal LD simulation, the cost of a SGLD simulation is almost identical to a LD simulation for the same number of time steps. SGLD simulations do require additional memory because of the need to store the guiding forces, as well as some arrays for the weighting factor calculation.

IV. RESULTS AND DISCUSSIONS

Through the three model systems presented here we demonstrate three points: (1) effect of guiding forces on conformational search, (2) effect of guiding forces on conformational distribution, and (3) conversion from SGLD conformational distributions to LD conformational distributions.

A. The skewed double well system

A skewed double well system represents the simplest nonsymmetric system with an energy barrier to cross. This system has only one particle and the particle moves on a fixed energy surface. This energy surface is designed in such a way that it restricts the particle to move near the y-axis with two energy minimums of different depths along the y-axis. 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. The potential profiles along the y-axis and across the y-axis are shown in Fig. 2. The potential function (in kcal/mol) is

$$\varepsilon_p = 500(x^2 + z^2) + y^2(y-2)^2 + 0.25y. \quad (28)$$

An argon atom was used to represent the particle. Simulations were carried out at 80 K with a local average time, $t_L = 0.2$ ps. A time step of 1 fs was used and the simulation length was 100 ns for each simulation.

Figure 3 shows two trajectories in y coordinates, one in a LD simulation [Fig. 3(a)] and the other in a SGLD simulation with $\lambda = 1$ [Fig. 3(b)]. Clearly, the transition between the two wells at $y = 0$ Å and $y = 2$ Å are more frequent in the SGLD simulation than that in the LD simulation, demonstrating an enhanced energy barrier overcoming ability in the SGLD simulation. Figure 3(c) shows the number of transitions as a function of T_{sg}. When λ increases, T_{sg} increases, so does the transitions between the wells. At $\lambda = 1$, $T_{sg} = 100.7$ K, the transitions increases by about 10 times more than the transitions in a LD simulation (i.e., $\lambda = 0$ and $T_{sg} = T$

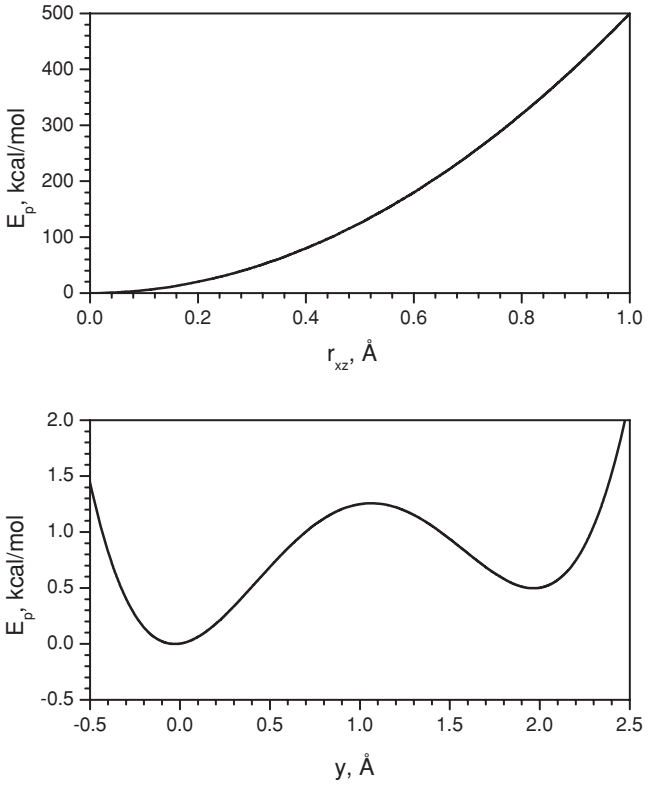

FIG. 2. The skewed double well potential along the y-axis (lower panel) when $r_{xz} = 0$ and perpendicular to the y-axis (upper panel) when $y = 0$.

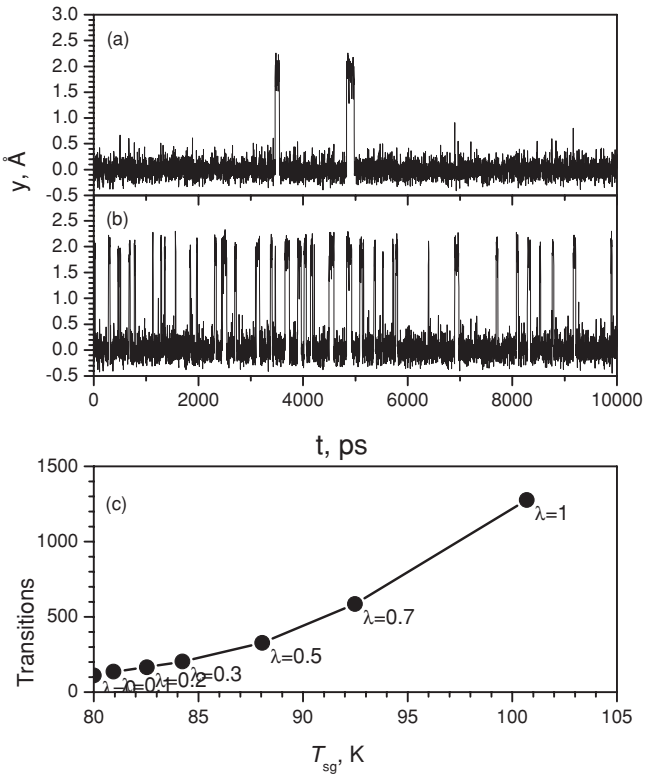

FIG. 3. Transitions of the particle in the double well system. (a) Trajectory in the LD simulation; (b) Trajectory in the SGLD simulation with $\lambda = 1$ where $T_{sg} = 100.7$ K. (c) Transition number as a function of the self-guiding temperature, T_{sg}. The collision frequency is 10/ps and temperature is 80 K.

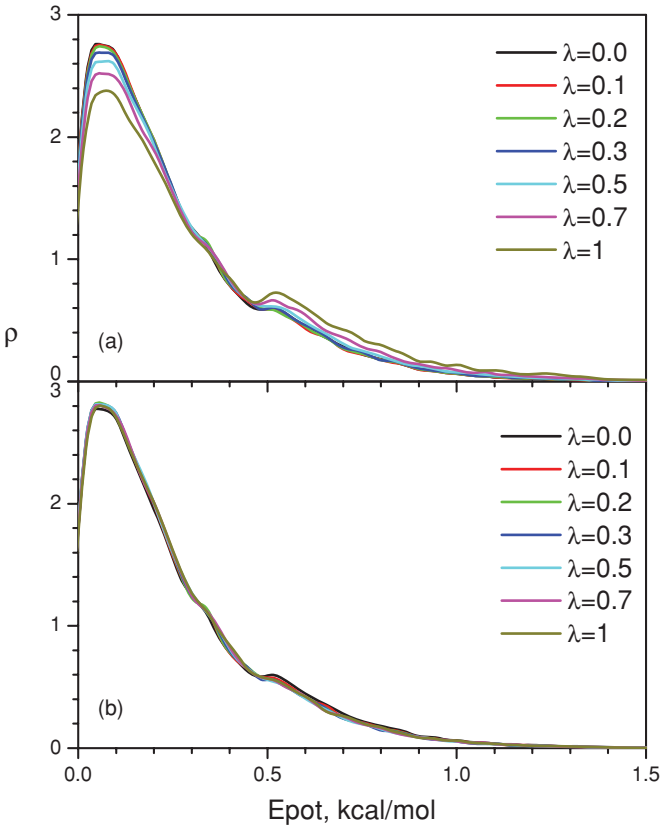

FIG. 4. The energy distributions of the double well system in the SGLD simulations: (a) unweighted; (b) weighted. The collision frequency is 10/ps and temperature is 80 K.

= 80 K). This result demonstrates the dramatic enhancement in the energy barrier crossing ability.

Figure 4(a) shows the potential energy distribution in the SGLD simulations. As λ increases, the distribution decreases in the low-energy region and increases in the high-energy region. Figure 4(b) shows the weighted energy distributions. Clearly, all curves converge fairly well to the one with $\lambda = 0$, demonstrating that the weighting scheme can convert the SGLD distributions to the canonical energy distribution with a reasonable degree of accuracy.

To further examine the guiding effect on the conformational distribution, we plot the conformational density as a function of the y coordinate in Fig. 5. Figure 5(a) shows the distributions at different guiding factors. There are two peaks with different heights, corresponding to the skewed double wells. Examining the peak heights at different λ, we can see that as λ increases, the left peak (the higher peak) decreases, while the right peak (the lower peak) grows. Figure 5(b) shows the weighted conformational distribution. All distributions converge fairly well to the one with $\lambda = 0$. This result again validates the reweighting scheme.

It should be noted that the reweighting scheme, Eqs. (24)–(26), is based on the first order perturbation approximation and is limited to small difference in conformational distribution. As can be seen in Fig. 5(b), when the guiding factor is large, the deviation from the LD distribution increases. Further increasing the guiding factor will make the reweighting hard to converge.

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 100 K except otherwise noted. Nonbonded interactions were calculated using the following rationalized polynomial 3D isotropic periodic sum (IPS) potentials.[6,8]

Lennard-Jones IPS potentials:

$$\varepsilon_{\text{disp}}^{\text{IPS}}(r, R) = \begin{cases} -\dfrac{C_{ij}}{r^6} - \dfrac{C_{ij}}{R^6}\left(\dfrac{1341}{3064} + \dfrac{77}{141}\left(\dfrac{r}{R}\right)^2 + \dfrac{61}{141}\left(\dfrac{r}{R}\right)^4 + \dfrac{56}{141}\left(\dfrac{r}{R}\right)^8\right) & r \le R \\ 0 & r > R \end{cases}, \tag{29}$$

$$\varepsilon_{\text{rep}}^{\text{IPS}}(r, R) = \begin{cases} \dfrac{A_{ij}}{r^{12}} + \dfrac{A_{ij}}{R^{12}}\left(\dfrac{23}{3620} + \dfrac{8}{151}\left(\dfrac{r}{R}\right)^2 + \dfrac{66}{151}\left(\dfrac{r}{R}\right)^6 + \dfrac{100}{151}\left(\dfrac{r}{R}\right)^{10}\right) & r \le R \\ 0 & r > R \end{cases}. \tag{30}$$

Figure 6(a) shows the potential energy distributions in the SGLD simulations at different guiding factors. Clearly, the energy distribution changes with the guiding factor. When applying the weighting scheme, the energy distribution converges together [Fig. 6(b)], except when $\lambda > 1$ where numerical convergence becomes a problem due to the large difference in the conformational distribution. As shown in Eq. (25), the weighting factor varies exponentially with the energies. The weighting scheme will converge poorly if the major distribution to be calculated is not properly sampled

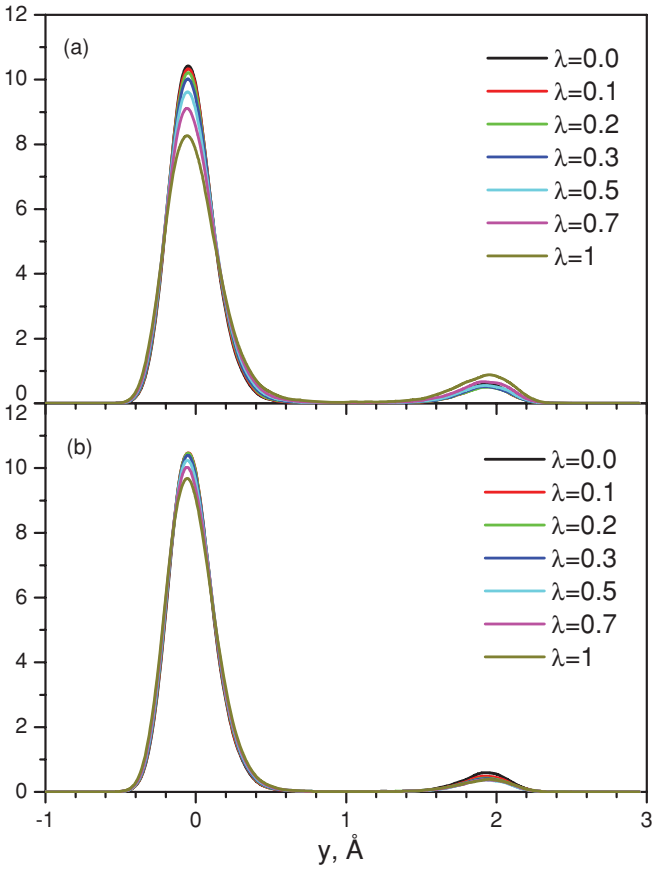

FIG. 5. The *y*-coordinate distributions of the double well system in the SGLD simulations: (a) unweighted; (b) weighted. The collision frequency is 10/ps and temperature is 80 K.

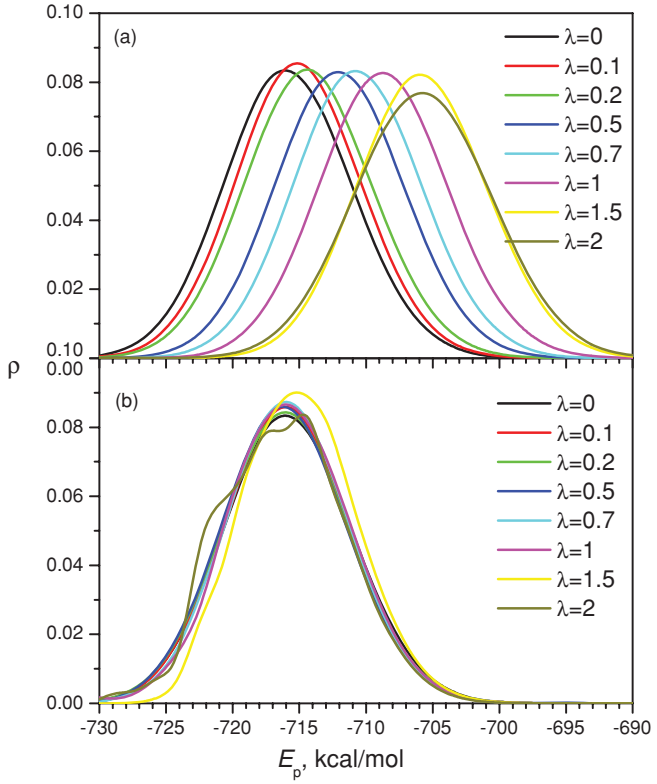

FIG. 6. The energy distributions of the argon liquid in the SGLD simulations at 100 K: (a) unweighted; (b) weighted. The collision frequency is 1/ps.

in the simulation. Note that if the simulation length is significantly increased, convergence would improve for larger λ values. About the precision of reweighting in simulations, Shen and Hamelberg has a more thorough analysis.[9]

Many enhanced sampling techniques come with a certain alteration of the conformational distribution. Increasing temperature is a commonly used approach to speed up a conformational search. However, the change in conformational distribution due to a rise in temperature is significant. Figure 7 shows the potential energy distributions of the argon fluid from LD simulations at different temperatures. Clearly we can see potential energies shift toward high energies when temperature increases. Comparing the distributions at 100 and 140 K, there is little conformation shared by both distributions. In other words, a temperature increase can speed up simulations but most of the conformations searched at 140 K are of little importance to the distribution at 100 K, which makes a reweighting formula to correct for the effects of higher temperature difficult to converge.

Obviously, the potential energy shifts up in a much smaller scale in SGLD simulations [Fig. 6(a)] than that when raising the temperature. Comparing Figs. 6(a) and 7, we can see that the energy deviations due to the guiding effect is much smaller than the deviation due to the temperature increase.

To quantitatively compare the SGLD and high-temperature LD simulations, we plot the average potential energies against diffusion constants in Fig. 8. Diffusion constants measure the conformational change in the slowest frequency and can be a good measurement of the conformational search efficiency. The diffusion constants were calculated with a fixed center of mass to avoid any exaggeration due to the motion of the center of mass. As can be seen from Fig. 8, SGLD increases diffusion constants with much smaller energy deviations than LD simulations at elevated temperatures. This plot tells us that SGLD can speed up

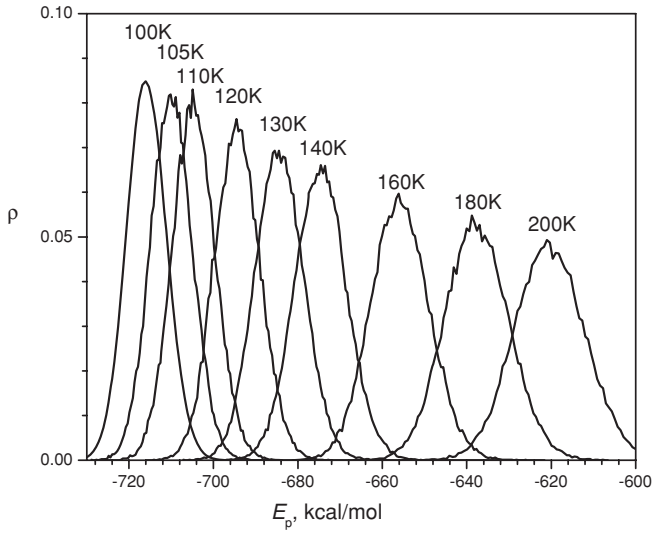

FIG. 7. The energy distributions of the argon liquid in the LD simulations at different temperatures (as labeled). The collision frequency is 1/ps.

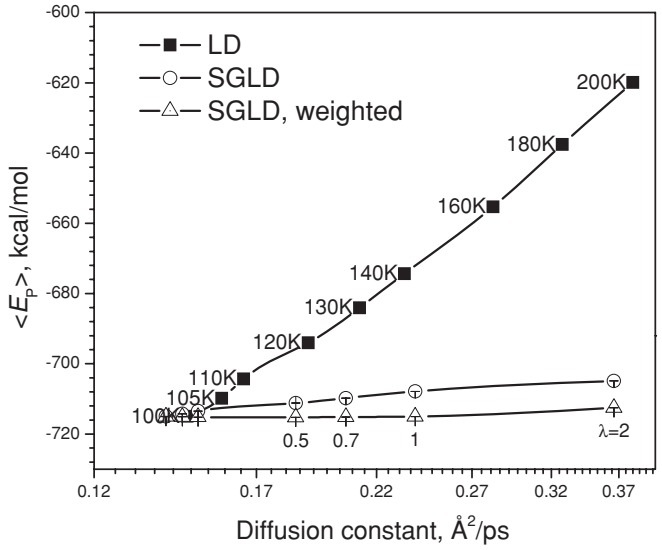

FIG. 8. Average potential energies vs diffusion constants for the argon liquid in the LD simulations at different temperatures (as labeled) and in the SGLD simulations at different guiding factors (as labeled). The collision frequency is 1/ps. The SGLD simulations were performed at 100 K.

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 at the temperature of interest.

The weighted average potential energies are also plotted against diffusion constants in Fig. 8. For SGLD, the weighted potential energy is very flat against diffusion constant. In other words, through the weighting procedure, SGLD can speed up conformational searches and produce accurate conformational distribution.

C. Alanine dipeptide

The Alanine dipeptide is perhaps the simplest and the most well studied molecule that is relevant to proteins.

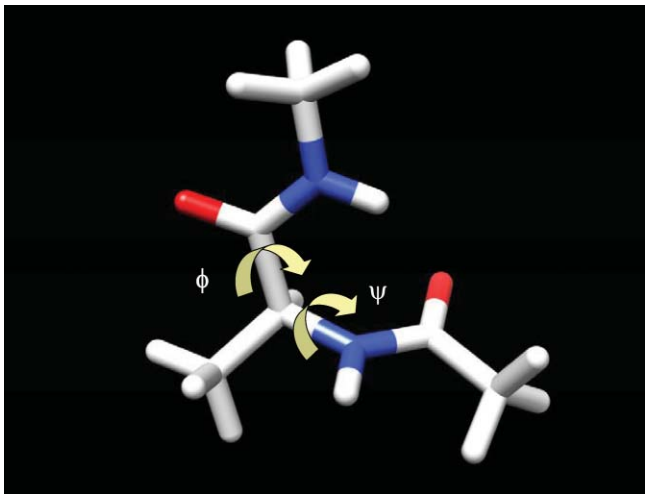

FIG. 9. A conformation of an alanine dipeptide. Chemical bonds are shown as sticks. Oxygen and nitrogen atoms are shown as red and blue, respectively. Two backbone dihedral angles, ϕ and ψ, are marked by arrows.

Figure 9 shows one conformation of an alanine dipeptide. The conformation of this molecule is mainly characterized by two dihedral angles, ϕ: CT–N–Cα–C and ψ: N–Cα–C–NT. The CHARMM all-atom force field[6] was used to describe the interactions. Here, we used a distance-dependent dielectric constant of $4r$ to represent solvent screening effect to simplify the example. The cutoff distance is set to 100 Å to avoid any cutoff effect in the nonbonded interaction calculation within this small molecule.

All simulations were performed with a time step of 2 fs and SHAKE algorithm[10] was employed to fix the bond lengths. Each simulation was 200 ns in length and conformations of every 2 ps were saved for postanalysis. The SGLD simulations were performed with a local average time of $t_L = 0.2$ ps and a temperature of 300 K.

To demonstrate the conformational search of SGLD simulations, we performed high-temperature LD simulations, as well as SGLD simulations with different guiding factors for the alanine dipeptide. To quantitatively describe the conformational search of this peptide, 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 10 shows the transition rate in the LD simulations against the simulation temperature and in the SGLD simulations against the self-guiding temperature. The self-guiding temperature is defined to reflect the conformational searching ability [Eq. (27)] of a SGLD simulation so that users can have a rough idea of how much conformational search ability has been achieved. As can be seen from Fig. 10, the transition rate increases with the temperature or the self-guiding temperature in a similar trend. The transition rates of SGLD simulations is somewhat higher than that of the LD simulations at the

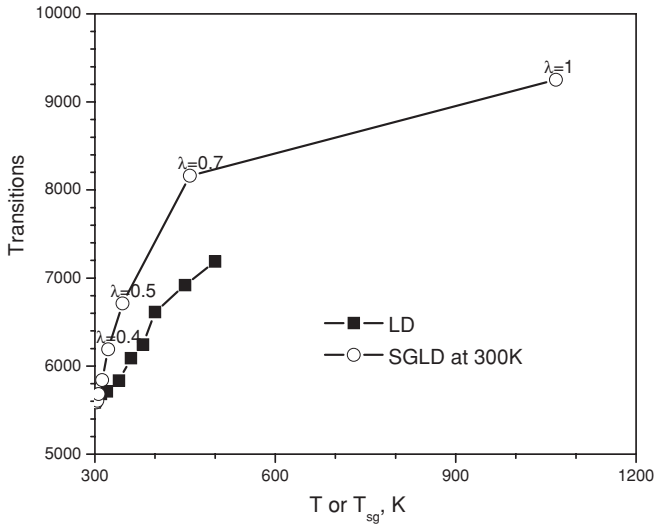

FIG. 10. Conformational transitions of the alanine dipeptide as a function of temperature in the LD simulations and as a function of the self-guiding temperature, T_{sg}, in the SGLD simulations. The self-guiding temperature, T_{sg}, defined by Eq. (27), reflects the conformational searching ability that is comparable to a high-temperature simulation at $T = T_{sg}$. The collision frequency is $\gamma = 10$/ps. The SGLD simulations were performed at 300 K.

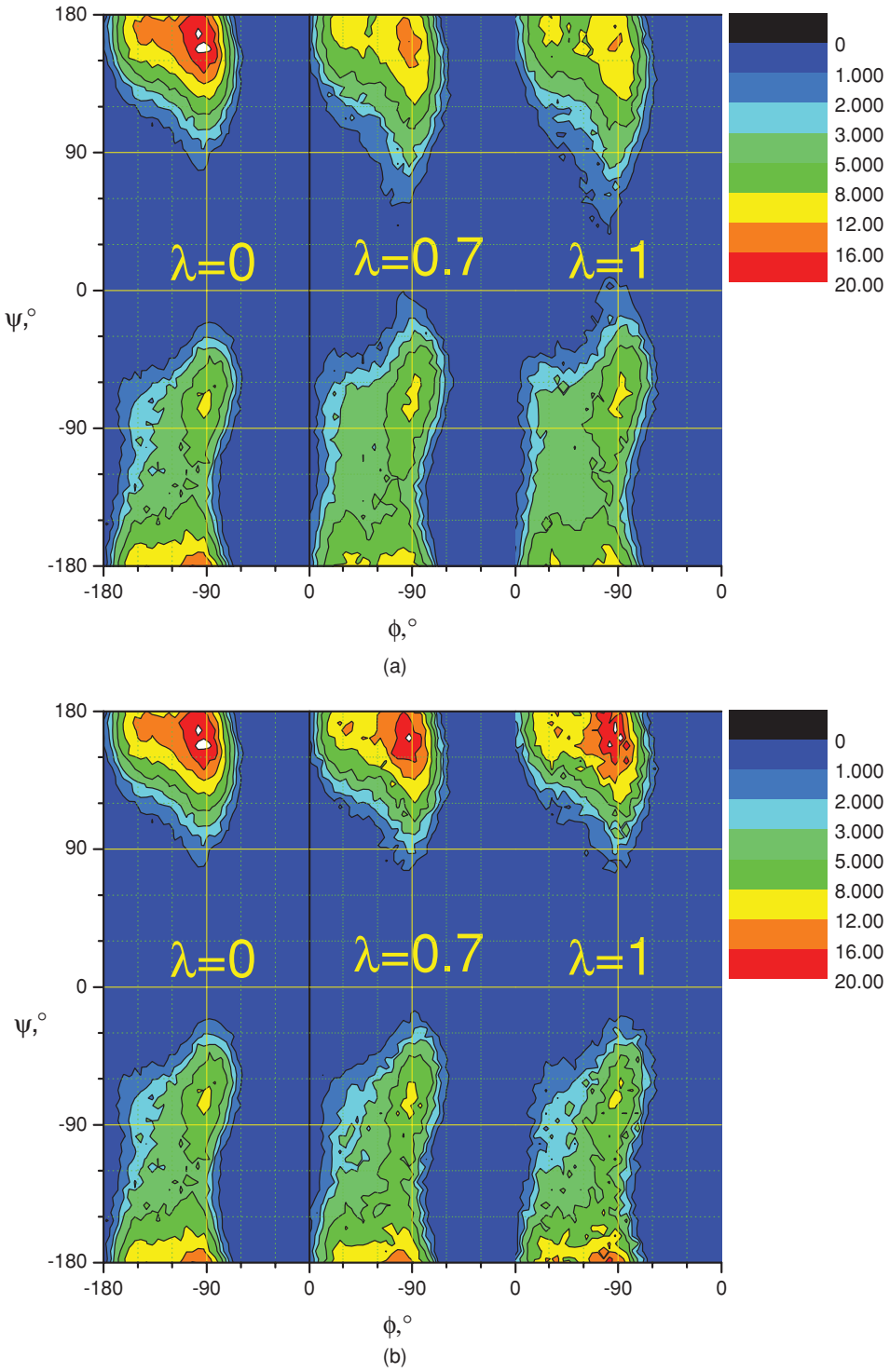

FIG. 11. (a) $\phi-\psi$ distributions of the alanine dipeptide in the LD ($\lambda = 0$) and SGLD simulations at $\lambda = 0.7$ and $\lambda = 1$ before reweighting. The collision frequency is $\gamma = 10$/ps. The SGLD simulations were performed at 300 K. (b) $\phi-\psi$ distributions of the alanine dipeptide in the LD ($\lambda = 0$) and SGLD simulations at $\lambda = 0.7$ and $\lambda = 1$ after reweighting. The collision frequency is $\gamma = 10$/ps. The SGLD simulations were performed at 300 K.

temperature of T_{sg}, indicating T_{sg} somewhat underestimates the conformational searching ability of the SGLD simulations in this case. Figure 10 shows that the SGLD simulations at guiding factors of 0.2, 0.5, and 1 have self-guiding temperatures of 346, 458, and 1067 K, respectively, indicating that the conformational search abilities of these simulations are comparable to that of high-temperature LD simulations at 346, 458, and 1067 K, respectively. It is clear that SGLD simulations have increased the conformational search ability dramatically.

To examine the reweighting scheme in a multidimensional distribution, we plot the $\phi-\psi$ dihedral angle distributions from the LD and SGLD simulations in Fig. 11. Before reweighting, as shown in Fig. 11(a), SGLD simulations have lower peak heights and broader baselines than the LD simulation. This result indicates that the dramatical acceleration

in conformational search accompanies a change in conformational distribution. An increase in the guiding factor, λ, results in an increase in the self-guiding temperatures (Fig. 10), and as a result, the system experiences an enhanced motion in the low-frequency conformational space. This enhanced motion increases conformational search ability, but also flattens the conformational distribution. As the guiding factor increases, the high peaks become lower and the valleys become shallower. After reweighting, as can be seen in Fig. 11(b), the peak heights and the baseline broadness of the SGLD results are quite similar to that of the LD simulation. This result again validates the weighting scheme. Obviously, the reweighting result is noisier at a larger guiding factor. A smaller guiding factor will help reduce the reweighting noise, but have a weaker conformational search ability. Therefore, the guiding factor should be set to have enough conformational search ability while allowing a reweighting of acceptable accuracy.

V. CONCLUSIONS

The conformational distribution from SGLD simulation is quantitatively described through the low-frequency and high-frequency properties. This provides a way to convert conformational distributions from SGLD simulations to canonical ensemble distributions. Through this work, the SGLD simulation method can be used not only to achieve a dramatically enhanced conformational search, but also to produce an accurate conformational distribution. This understanding of the SGLD conformational distribution provides a sound theoretical basis for further development and application of this method.

APPENDIX: SGLD SIMULATION ALGORITHM

To help understand how to calculate ensemble averages from SGLD simulations, we describe a leap-frog Verlet SGLD simulation algorithm below.

(i) Initiate low-frequency variables: $\tilde{E}_p(0) = E_p(0)$, $\tilde{\mathbf{f}}_i(0) = 0$, $\tilde{\mathbf{p}}_i(0) = 0$, and $\tilde{\mathbf{g}}_i(0) = 0$.

(ii) At time step, t, calculate interaction forces, $\mathbf{f}_i(t)$, random forces, $\mathbf{R}_i(t)$, and the uncorrected guiding forces, $\mathbf{g}'_i(t) = \lambda_i \gamma_i \tilde{\mathbf{p}}_i(t)$. The interaction forces, $\mathbf{f}_i(t)$, must include any constraint force as described later. Random forces, $\mathbf{R}_i(t)$, are generated from a Gaussian distribution with zero mean:

$$\rho(\mathbf{R}_i) = \frac{1}{\sqrt{4\pi\gamma_i m_i kT}} e^{-\frac{\mathbf{R}_i^2}{4\gamma_i m_i kT}}. \tag{A1}$$

The low-frequency momentum is calculated using the momentum in the previous half step, $\mathbf{p}_i(t - \frac{\delta t}{2})$:

$$\tilde{\mathbf{p}}_i(t) = \left(1 - \frac{\delta t}{t_L}\right)\tilde{\mathbf{p}}_i(t - \delta t) + \frac{\delta t}{t_L}\mathbf{p}_i\left(t - \frac{\delta t}{2}\right). \tag{A2}$$

(iii) Calculate the energy conservation factor, ξ. The half step velocity, $\dot{\mathbf{r}}_i(t)$, can be expressed in the following form:

$$\dot{\mathbf{r}}_i(t) = \dot{\mathbf{r}}_i\left(t - \frac{\delta t}{2}\right) + \frac{\delta t}{2m_i}(\mathbf{f}_i(t) + \mathbf{g}'_i(t) + \mathbf{R}_i(t)) - \frac{\delta t}{2}(\gamma_i + \xi\lambda_i\gamma_i)\dot{\mathbf{r}}_i(t); \tag{A3}$$

calculate the friction-free velocity at the half step:

$$\dot{\mathbf{r}}'_i(t) = \dot{\mathbf{r}}_i\left(t - \frac{\delta t}{2}\right) + \frac{\delta t}{2m_i}(\mathbf{f}_i(t) + \mathbf{g}'_i(t) + \mathbf{R}_i(t)). \tag{A4}$$

From Eqs. (A3) and (A4), we have

$$\dot{\mathbf{r}}_i(t) = \frac{\dot{\mathbf{r}}'_i(t)}{1 + \frac{(1+\xi\lambda_i)\gamma_i\delta t}{2}} \approx \frac{\dot{\mathbf{r}}'_i(t)}{1 + \frac{\gamma_i\delta t}{2}} - \frac{\dot{\mathbf{r}}'_i(t)}{\left(1 + \frac{\gamma_i\delta t}{2}\right)^2}\frac{\xi\lambda_i\gamma_i\delta t}{2}. \tag{A5}$$

Based on the energy conservation relation, Eq. (11), and neglect the higher power term of ξ, we can solve the energy conservation factor:

$$\xi = \frac{\sum_i^N \lambda_i\gamma_i\tilde{\mathbf{p}}_i(t)\dot{\mathbf{r}}'_i(t)\left(1 + \frac{\gamma_i\delta t}{2}\right)^{-1}}{\sum_i^N \lambda_i\gamma_i m_i\dot{\mathbf{r}}'^2_i(t)\left(1 + \frac{\gamma_i\delta t}{2}\right)^{-2} + \frac{\delta t}{2}\sum_i^N \lambda_i^2\gamma_i^2\tilde{\mathbf{p}}_i(t)\dot{\mathbf{r}}'_i(t)\left(1 + \frac{\gamma_i\delta t}{2}\right)^{-2}}. \tag{A6}$$

The actual guiding force is

$$\mathbf{g}_i(t) = \lambda_i\gamma_i\tilde{\mathbf{p}}_i(t) - \xi\mathbf{p}_i(t) = \lambda_i\gamma_i\tilde{\mathbf{p}}_i(t) - \frac{\xi m_i\dot{\mathbf{r}}'_i(t)}{1 + \frac{(1+\xi\lambda_i)\gamma_i\delta t}{2}}. \tag{A7}$$

(iv) Update low-frequency variables and accumulators for the calculation of the SGLD weighting factor. Low-frequency forces:

$$\tilde{\mathbf{f}}_i(t) = \left(1 - \frac{\delta t}{t_L}\right)\tilde{\mathbf{f}}_i(t - \delta t) + \frac{\delta t}{t_L}\mathbf{f}_i(t).$$

Low-frequency potential energy:

$$\tilde{E}_p(t) = \left(1 - \frac{\delta t}{t_L}\right)\tilde{E}_p(t - \delta t) + \frac{\delta t}{t_L}E_p(t).$$

Low-frequency guiding forces:

$$\tilde{\mathbf{g}}_i(t) = \left(1 - \frac{\delta t}{t_L}\right)\tilde{\mathbf{g}}_i(t - \delta t) + \frac{\delta t}{t_L}\mathbf{g}_i((t).$$

From the low-frequency momentums we can calculate the low-frequency temperature:

$$\tilde{T} = \frac{1}{N_{\mathrm{DF}}}\sum_i \frac{\tilde{\mathbf{p}}_i^2}{m_i}.$$

Update accumulators for the collision factors and energy factors:

$$\mathrm{FLF} = \sum_t \sum_i^N \tilde{\mathbf{f}}_i(t)\cdot\tilde{\mathbf{f}}_i(t),$$

$$\mathrm{FHF} = \sum_t \sum_i^N (\mathbf{f}_i(t) - \tilde{\mathbf{f}}_i(t))\cdot(\mathbf{f}_i(t) - \tilde{\mathbf{f}}_i(t)),$$

$$\mathrm{GLF} = \sum_t \sum_i^N (\tilde{\mathbf{g}}_i(t) - \gamma_i\tilde{\mathbf{p}}_i(t))\cdot\tilde{\mathbf{f}}_i(t),$$

$$\mathrm{GHF} = \sum_t \sum_i^N (\mathbf{g}_i(t) - \tilde{\mathbf{g}}_i(t) - \gamma_i(\mathbf{p}_i(t) - \tilde{\mathbf{p}}_i(t))) \cdot(\mathbf{f}_i(t) - \tilde{\mathbf{f}}_i(t))$$

$$\mathrm{PPLF} = \sum_t \sum_i^N \gamma_i^2\tilde{\mathbf{p}}_i(t)\cdot\tilde{\mathbf{p}}_i(t),$$

$$\mathrm{GPLF} = \sum_t \sum_i^N \tilde{\mathbf{g}}_i(t)\cdot\gamma_i\tilde{\mathbf{p}}_i(t).$$

The collision and energy factors are calculated with the accumulators:

$$\lambda_{\mathrm{lf}} = 1 + \frac{\mathrm{GLF}}{\mathrm{FLF}}, \quad \lambda_{\mathrm{hf}} = 1 + \frac{\mathrm{GHF}}{\mathrm{FHF}}, \quad \chi_{\mathrm{lf}} = \frac{\tilde{T}_0}{\tilde{T}} = 1 - \frac{\mathrm{GPLF}}{\mathrm{PPLF}}. \tag{A8}$$

When $\tilde{T}_0$ is available from a previous SGLD simulation with $\lambda = 0$, $\chi_{\mathrm{lf}} = \frac{\tilde{T}_0}{\tilde{T}}$ is recommended, otherwise, $\chi_{\mathrm{lf}} = 1 - \frac{\mathrm{GPLF}}{\mathrm{PPLF}}$ has to be used. The SGLD weighting factor of each conformation can be calculated as below during a simulation or in a postsimulation processing:

$$w_{\mathrm{SGLD}} = \exp\left((\lambda_{\mathrm{lf}}\chi_{\mathrm{lf}} - 1)\frac{\tilde{E}_p - \bar{E}_p}{kT} + \left(\lambda_{\mathrm{hf}}\frac{T - \chi_{\mathrm{lf}}\tilde{T}}{T - \tilde{T}} - 1\right)\frac{E_p - \tilde{E}_p}{kT}\right). \tag{A9}$$

In Eq. (A9), the average potential energy is subtracted from the low-frequency energy to avoid overflow in calculating the exponential function. With the weighting factor, any ensemble averages can be calculated during a simulation or in a postsimulation process.

(v) Advance velocities to the next half time step:

$$\dot{\mathbf{r}}_i\left(t + \frac{\delta t}{2}\right) = (2\chi_i - 1)\,\dot{\mathbf{r}}_i\left(t - \frac{\delta t}{2}\right) + \chi_i\frac{\delta t}{m_i}(\mathbf{f}_i(t) + \mathbf{g}_i(t) + \mathbf{R}_i(t)). \tag{A10}$$

Here, the scaling parameter, χ_i, is calculated as

$$\chi_i = \left(1 + \frac{(1 + \xi\lambda_i)\gamma_i\delta t}{2}\right)^{-1}. \tag{A11}$$

Then advance positions to the next time step:

$$\mathbf{r}_i(t + \delta t) = \mathbf{r}_i(t) + \dot{\mathbf{r}}_i\left(t + \frac{\delta t}{2}\right)\delta t. \tag{A12}$$

If internal coordinates need to be constrained, apply constraining algorithms, such as SHAKE (Ref. 10) or semiflexible constraint dynamics,[11] to obtain constrained positions, $\mathbf{r}_i^{\mathrm{CON}}(t + \delta t)$, from $\mathbf{r}_i(t + \delta t)$. The constraint forces must be included in the low-frequency force calculation. The constraint forces are calculated by the following equation:

$$\mathbf{f}_i^{\mathrm{CON}}(t + \delta t) = \frac{2m_i}{\delta t^2}\left(\mathbf{r}_i^{\mathrm{CON}}(t + \delta t) - \mathbf{r}_i(t + \delta t)\right). \tag{A13}$$

(vi) Continue to step (ii) with $t = t + \delta t$ until the end of the simulation.

[1] X. Wu and B. R. Brooks, Chem. Phys. Lett. **381**(3–4), 512 (2003).
[2] A. Damjanović, E. B. García-Moreno, and B. R. Brooks, Proteins: Struct., Funct., Bioinf. **76**(4), 1007 (2009); A. Damjanović, B. T. Miller, T. J. Wenaus, P. Maksimović, E. Bertrand García-Moreno, and B. R. Brooks, J. Chem. Inf. Model. **48**(10), 2021 (2008); M. S. Lee and M. A. Olson, J. Chem. Theory Comput. **6**(8), 2477 (2010); C. I. Lee and N. Y. Chang, Biophys. Chem. **151**(1–2), 86 (2010).
[3] A. Damjanović, X. Wu, E. B. García-Moreno, and B. R. Brooks, Biophys. J. **95**(9), 4091 (2008).
[4] I. Andricioaei, A. R. Dinner, and M. Karplus, J. Chem. Phys. **118**(3), 1074 (2003).
[5] X. Wu and S. Wang, J. Chem. Phys. **110**(19), 9401 (1999); J. Phys. Chem. B **102**(37), 7238 (1998); X.-W. Wu and S.-S. Sung, Proteins: Struct. Funct. Genet. **34**(3), 295 (1999).
[6] B. R. Brooks, C. L. Brooks III, A. D. Mackerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, J. Comput. Chem. **30**(10), 1545 (2009).
[7] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, B. Jaun, and M. Karplus, J. Comput. Chem. **4**, 187 (1983).
[8] X. Wu and B. R. Brooks, J. Chem. Phys. **122**(4), 44107 (2005).
[9] T. Shen and D. Hamelberg, J. Chem. Phys. **129**(3), 034103 (2008).
[10] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, J. Comput. Phys. **23**, 327 (1977).
[11] X.-W. Wu and S.-S. Sung, J. Comput. Chem. **19**(14), 1555 (1998).