Replica exchanging self-guided Langevin dynamics for efficient and accurate conformational sampling

Xiongwu Wu,[1,a)] Milan Hodoscek,[2] and Bernard R. Brooks[1]
[1]*Laboratory of Computational Biology, National Heart, Lung, and Blood Institute (NHLBI), National Institutes of Health (NIH), Bethesda, Maryland 20892, USA*
[2]*Center for Molecular Modeling, National Institute of Chemistry, Hajdrihova 19, SI-1000, Ljubljana, Slovenia*

(Received 19 April 2012; accepted 29 June 2012; published online 24 July 2012)

This work presents a replica exchanging self-guided Langevin dynamics (RXSGLD) simulation method for efficient conformational searching and sampling. Unlike temperature-based replica exchanging simulations, which use high temperatures to accelerate conformational motion, this method uses self-guided Langevin dynamics (SGLD) to enhance conformational searching without the need to elevate temperatures. A RXSGLD simulation includes a series of SGLD simulations, with simulation conditions differing in the guiding effect and/or temperature. These simulation conditions are called stages and the base stage is one with no guiding effect. Replicas of a simulation system are simulated at the stages and are exchanged according to the replica exchanging probability derived from the SGLD partition function. Because SGLD causes less perturbation on conformational distribution than high temperatures, exchanges between SGLD stages have much higher probabilities than those between different temperatures. Therefore, RXSGLD simulations have higher conformational searching ability than temperature based replica exchange simulations. Through three example systems, we demonstrate that RXSGLD can generate target canonical ensemble distribution at the base stage and achieve accelerated conformational searching. Especially for large systems, RXSGLD has remarkable advantages in terms of replica exchange efficiency, conformational searching ability, and system size extensiveness. © *2012 American Institute of Physics*. [http://dx.doi.org/10.1063/1.4737094]

I. INTRODUCTION

Conformational searching and sampling is a fundamental process of molecular systems. In the real world, molecular thermal motions are the driving force for conformational searching. Raising temperatures can accelerate thermal motions and is often very effective for speeding up physical and chemical processes. For computational studies, it is often difficult to properly search and sample the conformational space of large molecular systems with current computing resources. There are many ways to accelerate conformational searching and sampling, such as modifying energy surfaces and raising temperatures. Raising temperatures is often more convenient and has been widely used. For example, simulated annealing and temperature-based replica exchange have found applications in many computational studies. However, raising temperature causes changes in conformational distribution, and often leads to complications such as protein unfolding and phase transition.

Unlike high temperature simulations that accelerate all thermal motions, the self-guided Langevin dynamics (SGLD) (Refs. 1–3) enhances only the low frequency motion that is the most important for conformational searching and sampling. SGLD is unique in that with a simple local averaging scheme, it selectively enhances molecular motions based on their frequencies without modifying energy surfaces or raising temperatures. SGLD simulations have been applied to many studies of long time scale events, such as the conformational reorganization of protein staphylococcal nuclease (SNase),[4] hydration state and rotameric substates of SNase,[5] conformational transitions induced by dephosphorylation in nitrogen regulatory protein C (NtrC) protein,[6] conformational transitions in a membrane transporter protein lactose permease (LacY),[7] and characteristics of the denatured state of the human prion (huPrP).[8]

The concept of SGLD can be illustrated by Fig. 1. In normal dynamics, such as in Langevin dynamics, kinetic energy distributes evenly among all degrees of freedom, i.e., $kT/2$ per degree of freedom. The alanine dipeptide shown in Fig. 1 has high frequency motions like bond vibration and bond bending, and low frequency motions, such as the bond rotations about the ϕ, ψ dihedral angles. All these motions have the same kinetic energy, or temperature, T, in the canonical ensemble. The low frequency motions, the changes of the ϕ, ψ dihedral angels in this case, are the limiting steps for conformational searching. These low frequency motions can be enhanced by elevating temperatures. However, at high temperatures, all motions, most of them high frequency motions, are more energetic, which dramatically enlarge the accessible conformational space and shift major conformational distribution to otherwise unpopulated regions. In SGLD, the guiding forces enhance the low frequency motions, as defined by the local averaging time, t_L, while suppressing high frequency

a)Author to whom correspondence should be addressed. Electronic mail: wuxw@nhlbi.nih.gov. Telephone: 301-451-6251. Fax: 301-480-6496.

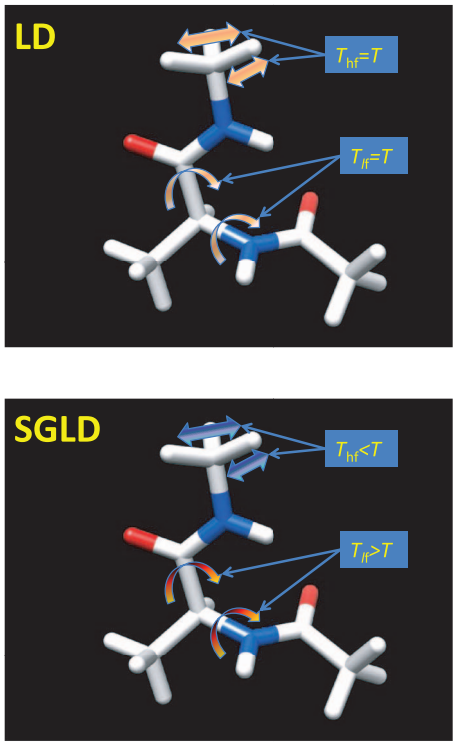

FIG. 1. Thermal motions of an alanine dipeptide in Langevin dynamics and in self-guided Langevin dynamics simulations. Atoms are drawn as sticks. Carbon, oxygen, nitrogen, and hydrogen atoms are colored as grey, red, blue, and white, respectively. In LD, all motions have kinetic energies equivalent to a temperature of T, while in SGLD, low frequency motion gain more kinetic energy, becoming hotter, and high frequency motions lose some kinetic energy, becoming cooler. It is the low frequency motion that controls the conformational searching, therefore, SGLD achieves enhanced conformational searching ability without raising temperature.

motions to maintain the overall temperature. In other words, the low frequency motions will gain more kinetic energy and become hotter. At the same time, the high frequency motions will lose kinetic energy and become cooler. The overall temperature remains unchanged. The enhancement in the low frequency motions will increase the transition of the backbone dihedral angles, which leads to an accelerated conformational searching. Because most of degrees of freedom are high frequent in nature, enhancing only the low frequency motions has a much smaller disturbance on the conformational distribution than enhancing all motions due to elevated temperatures.

SGLD relies on the guiding forces to enhance the low frequency motions. Under the effect of the guiding forces, SGLD has its own conformational distribution, defined as the SGLD ensemble. The partition function of the SGLD ensemble has been derived recently[2] which allows a conversion between the SGLD ensemble and the canonical ensemble so that canonical ensemble properties can be calculated from SGLD simulations through reweighting. Based on the SGLD partition function, we further developed the force-momentum based self-guided Langevin dynamics (SGLDfp)[9] method, to directly sample the canonical ensemble without the need of reweighting.

Temperature-based replica exchange methods utilize elevated temperatures to accelerate conformational search.[10–14] Through replica exchanging, a system can overcome energy barriers at high temperatures and maintain conformational distributions at different temperatures. For large systems, elevating temperatures causes a significant change in energy distributions, which requires a large number of replicas to achieve reasonable exchange probabilities, which in turn increases simulation cost accordingly. This size dependence is termed as not size extensive. Many efforts have been dedicated to address this difficulty for large systems.[15–26]

Because SGLD accelerates conformational searching and sampling, it can also be used for replica exchange simulations. Lee and Olson applied SGLD directly to their temperature-based replica-exchange (SGLD-ReX) simulations in their study of protein folding.[27] Because the SGLD partition function was not available at that time, they did not include the guiding force effects in their calculation of the exchange probabilities. Therefore, it is not surprising to observe certain deviations in their simulation results. To apply SGLD properly in replica exchange simulations, the SGLD partition function must be incorporated into the exchange probability to maintain the conformational distributions at different simulation conditions.

This work presents the replica exchanging self-guiding Langevin dynamics (RXSGLD) method. The goal of this work are as follows: (1) to describe the details of the RXSGLD method and to show how to derive the exchange probability from the SGLD partition function, (2) to examine how accurately RXSGLD samples the canonical ensemble, and (3) to evaluate the conformational searching ability of RXSGLD as compared to a temperature-based replica exchange simulation method.

II. THEORY AND METHODS

The RXSGLD simulation method is a combination of the replica exchange approach and the SGLD method. By incorporating the SGLD partition function into the exchange probability, RXSGLD can maintain a canonical ensemble at its base stage while achieve enhanced conformational searching and sampling. Because SGLD plays a central role in this method, we first briefly introduce the SGLD method[1–3,9] and its conformational distribution, followed by a detailed description of the RXSGLD method.

A. SGLD simulation method

For any particle, i, the equation of the self-guided motion has the following general form:

$$\dot{\mathbf{p}}_i = \mathbf{f}_i + \mathbf{g}_i - \gamma_i \mathbf{p}_i + \mathbf{R}_i, \tag{1}$$

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

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

Equation (1) contains a guiding force, $\mathbf{g}_i$, which is calculated based on the momentum, $\mathbf{p}_i$, and the low frequency momentum, $\tilde{\mathbf{p}}_i$,

$$\mathbf{g}_i(t) = \lambda_i \gamma_i \left(\tilde{\mathbf{p}}_i(t) - \xi \mathbf{p}_i(t)\right). \tag{3}$$

Here, λ_i is the guiding factor, which defines the strength of the guiding force. When $\lambda_i = 0$, Eq. (1) reduces to that of Langevin dynamics. The parameter, ξ, is an energy conservation factor to eliminate any net energy input from the guiding force,

$$\sum_i \mathbf{g}_i \cdot \dot{\mathbf{r}}_i = \sum_i \lambda_i \gamma_i \tilde{\mathbf{p}}_i \cdot \dot{\mathbf{r}}_i - \xi \sum_i \lambda_i \gamma_i \mathbf{p}_i \cdot \dot{\mathbf{r}}_i = 0. \tag{4}$$

The summation in Eq. (4) runs over all particles in a simulation system. Eq. (4) means that the guiding force will not cause energy flow between the system and its environment. Instead, the guiding force will cause energy flow between different motion modes within a simulation system. The energy conservation factor on each time step is determined from the following equation:

$$\xi = \frac{\sum_i \lambda_i \gamma_i \tilde{\mathbf{p}}_i \cdot \dot{\mathbf{r}}_i}{\sum_i \lambda_i \gamma_i \mathbf{p}_i \cdot \dot{\mathbf{r}}_i}. \tag{5}$$

The low frequency portion of any property, P, as denoted by a "$\sim$" cap, $\tilde{P}$, is calculated as a local average in the following progressive way:

$$\tilde{P}(t) = \left(1 - \frac{\delta t}{t_L}\right) \tilde{P}(t - \delta t) + \frac{\delta t}{t_L} P(t). \tag{6}$$

This calculation is very memory efficient and is simply done by an update with current instantaneous values. This local averaging acts like a low frequency filter that reduces high frequency components while keeps low frequency contributions.[2] Therefore, the results from Eq. (6) are called the low frequency properties. Correspondingly, we call $P - \tilde{P}$ the high frequency property. According to Eq. (3), the guiding force has the characteristics of the low frequency momentum. As a result, it resonates and enhances low frequency motions, which in turn, accelerates conformational searching and sampling.

The low frequency and high frequency properties defined by Eq. (6) should not be understood as a separation of properties based on the motion modes. They both contain contributions from all motion modes, but with different proportions. The low frequency properties contain more contributions from low frequency motion modes, while the high frequency properties contain more from high frequency ones. In other words, every motion mode contributes to both the low frequency and the high frequency properties, but the proportion depends on its frequency.

The effects of the guiding force on the low frequency motions and high frequency motions depend on simulation systems, simulation conditions, and many other factors, such as the coupling between motion modes. Therefore, it is difficult to characterize quantitatively the effects of the guiding forces theoretically. Nonetheless, we can evaluate the effects from SGLD simulations. To summarize, we use $\lambda_{\mathrm{lf}}$and λ_{hf} to represent the bias effects on the low frequency and high frequency energy surfaces, and use χ_{lf} and χ_{hf} to represent the effects on the low frequency and the high frequency motions, respectively. Taken together, an SGLD ensemble has a configurational partition function of the following form:[2]

$$\Theta_{\mathrm{SGLD}} \approx \sum \exp\left(-\frac{\lambda_{\mathrm{lf}}\chi_{\mathrm{lf}}\tilde{E}_{\mathrm{p}}}{kT} - \frac{\lambda_{\mathrm{hf}}\chi_{\mathrm{hf}}(E_{\mathrm{p}} - \tilde{E}_{\mathrm{p}})}{kT}\right). \tag{7}$$

Here, the summation runs over all microscopic states. The factors, $\lambda_{\mathrm{lf}}$and λ_{hf}, which are called the low frequency energy factor and the high frequency energy factor, respectively, are calculated as the average projections of the total forces in the direction of the interaction forces as follows:

$$\lambda_{\mathrm{lf}} = \frac{\left\langle \sum_i (\tilde{\mathbf{f}}_i + \tilde{\mathbf{g}}_i - \gamma_i \tilde{\mathbf{p}}_i)\tilde{\mathbf{f}}_i \right\rangle}{\left\langle \sum_i \tilde{\mathbf{f}}_i \tilde{\mathbf{f}}_i \right\rangle}, \tag{8}$$

$$\lambda_{\mathrm{hf}} = \frac{\left\langle \sum_i (\mathbf{f}_i - \tilde{\mathbf{f}}_i + \mathbf{g}_i - \tilde{\mathbf{g}}_i - \gamma_i(\mathbf{p}_i - \tilde{\mathbf{p}}_i))(\mathbf{f}_i - \tilde{\mathbf{f}}_i) \right\rangle}{\left\langle \sum_i (\mathbf{f}_i - \tilde{\mathbf{f}}_i)(\mathbf{f}_i - \tilde{\mathbf{f}}_i) \right\rangle}. \tag{9}$$

The factors, χ_{lf} and χ_{hf}, which are called the low frequency collision factor and the high frequency collision factor, respectively, are calculated according to the projections of the guiding forces in the direction of the friction forces as follows:

$$\chi_{\mathrm{lf}} = \frac{\tilde{T}_0}{\tilde{T}} = 1 - \frac{\left\langle \sum_i \tilde{\mathbf{g}}_i \gamma_i \tilde{\mathbf{p}}_i \right\rangle}{\left\langle \sum_i \gamma_i^2 \tilde{\mathbf{p}}_i \tilde{\mathbf{p}}_i \right\rangle}, \tag{10}$$

$$\begin{aligned} \chi_{\mathrm{hf}} &= \frac{T - \tilde{T}_0}{T - \tilde{T}} = \frac{T - \chi_{\mathrm{lf}}\tilde{T}}{T - \tilde{T}} \\ &= 1 - \frac{\left\langle \sum_i \gamma_i (\mathbf{g}_i - \tilde{\mathbf{g}}_i) \cdot (\mathbf{p}_i - \tilde{\mathbf{p}}_i) \right\rangle}{\left\langle \sum_i \gamma_i^2 (\mathbf{p}_i - \tilde{\mathbf{p}}_i) \cdot (\mathbf{p}_i - \tilde{\mathbf{p}}_i) \right\rangle}. \end{aligned} \tag{11}$$

$\tilde{T}$ is called the low frequency temperature, which is calculated from the low frequency momentum as

$$\tilde{T} = \frac{1}{N_{\mathrm{DF}} k} \left\langle \sum_i \frac{\tilde{\mathbf{p}}_i^2}{m_i} \right\rangle. \tag{12}$$

$\tilde{T}_0$ is the reference low frequency temperature, which corresponds to the low frequency temperature when the guiding factors, $\{\lambda_i\}$, are zero.

From Eqs. (8)–(11), we can see that in a Langevin dynamics (LD) simulation, $\lambda_{\mathrm{lf}} = 1$, $\lambda_{\mathrm{hf}} = 1$, $\chi_{\mathrm{lf}} = 1$, and $\chi_{\mathrm{hf}} = 1$. Therefore, from Eq. (7) we have

$$\Theta_{\mathrm{LD}} = \sum \exp\left(-\frac{E_{\mathrm{p}}}{kT}\right). \tag{13}$$

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

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

$$\times \exp\left((\lambda_{\rm lf}\chi_{\rm lf}-1)\frac{\tilde{E}_{\rm p}}{kT}+(\lambda_{\rm hf}\chi_{\rm hf}-1)\frac{E_{\rm p}-\tilde{E}_{\rm p}}{kT}\right)\Bigg)$$

$$=\Theta_{\rm SGLD}\langle w_{\rm SGLD}\rangle_{\rm SGLD}.$$

Here, $w_{\rm SGLD}$ is called the SGLD reweighting factor,

$$w_{\rm SGLD}=\exp\left((\lambda_{\rm lf}\chi_{\rm lf}-1)\frac{\tilde{E}_{\rm p}}{kT}+(\lambda_{\rm hf}\chi_{\rm hf}-1)\frac{E_{\rm p}-\tilde{E}_{\rm p}}{kT}\right). \tag{14}$$

Any ensemble average, $\langle P\rangle_{\rm LD}$, can be calculated in an SGLD simulation as

$$\langle P\rangle_{\rm LD}=\frac{\langle P w_{\rm SGLD}\rangle_{\rm SGLD}}{\langle w_{\rm SGLD}\rangle_{\rm SGLD}}. \tag{15}$$

To quantitatively describe the conformational search ability of an SGLD simulation, we define the self-guiding temperature as

$$T_{\rm SG}=\frac{\chi_{\rm hf}}{\chi_{\rm lf}}T=\frac{\tilde{T}(T-\tilde{T}_0)}{\tilde{T}_0(T-\tilde{T})}T. \tag{16}$$

The self-guiding temperature, $T_{\rm SG}$, provides a rough measure of the conformational searching ability in the unit of temperature. An SGLD simulation with a self-guiding temperature of $T_{\rm SG}$ has conformational search ability comparable to that of a high temperature simulation at $T=T_{\rm SG}$. In SGLD simulations, one can either set the guiding factors, $\{\lambda_i\}$, or set a target self-guiding temperature, $T^0_{\rm SG}$, and adjust the guiding factors, $\{\lambda_i\}$, so that $T_{\rm SG}$ approaches $T^0_{\rm SG}$.

B. Replica exchanging self-guiding Langevin dynamics simulation

A replica exchange simulation consists of a number of parallel simulations of identical systems, which are called replicas, at different simulation conditions, which are called stages. For a typical temperature based replica exchange simulation, replicas represent different conformational states, and stages represent different temperatures. A replica exchange means either two replicas exchanging their stages, or two stages exchanging their replicas. Both schemes have the same results but different exchanging information, which only affects communication efficiency and post processing procedures.

Figure 2 illustrates the basic scheme of a RXSGLD simulation. There are $k+1$ stages with different simulation conditions. The stage with the conditions of interest is designated as stage 0 ($T^{(0)}_{\rm SG}=T$ and $T^{(0)}=T$), and is called the base stage. The other k stages have different guiding temperatures, $T^{(i)}_{\rm SG}$, and the same or different temperatures, $T^{(i)}$. The stage with the maximum condition is called the top stage with $T^{(k)}$ and $T^{(k)}_{\rm SG}$. We use the base and top conditions to denote a replica exchange simulation, such as $T_{\rm SG}=T^{(0)}_{\rm SG}/T^{(k)}_{\rm SG}$ and $T=T^{(0)}/T^{(k)}$. To make all exchanges between neighboring stages to have similar acceptance ratios, it is suggested to have temperatures exponentially distributed.[12] We use the following formula to set temperatures and the self-guiding temperatures for each

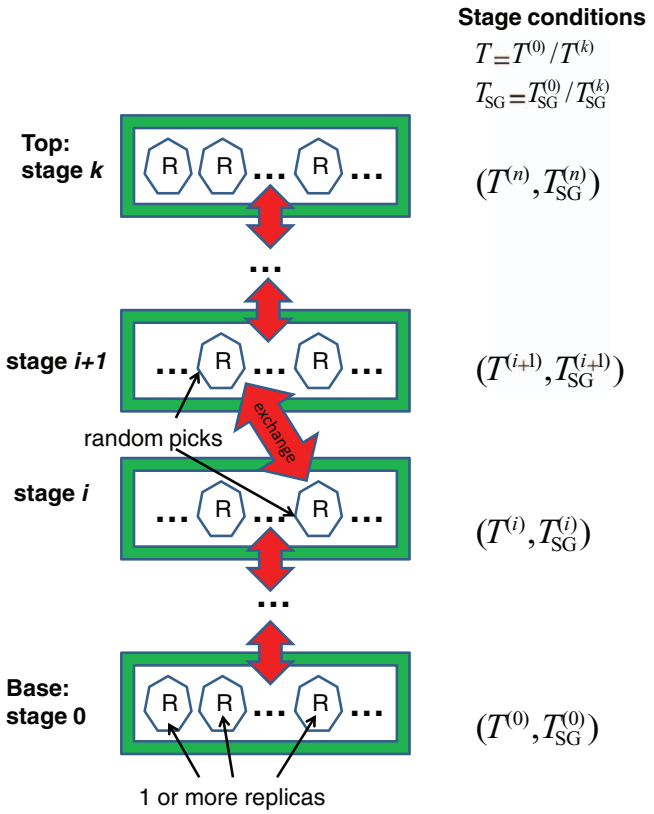

FIG. 2. A basic scheme describing the replica exchanging self-guided Langevin dynamics simulation. There are multiple stages with different simulation conditions, $T^{(i)}$and $T^{(i)}_{\rm SG}$. The base stage has the simulation condition of interest, $T^{(0)}$and $T^{(0)}_{\rm SG}=T^{(0)}$. A simulation system is replicated to many copies, called replicas. On each stage there are one or more replicas. Between stages, a pair of randomly chosen replicas are exchanged according to the exchanging probability. A TRXLD simulation has different stage temperatures, $T^{(i)}\geq T^{(0)}$, but no guiding force, $T^{(i)}_{\rm SG}=T^{(0)}$, and a RXSGLD simulation has difference self-guiding temperatures, $T^{(i)}_{\rm SG}\geq T^{(0)}$.

stage:

$$T^{(i)}=T^{(0)}\left(\frac{T^{(k)}}{T^{(0)}}\right)^{\frac{i}{k}}, \tag{17a}$$

$$T^{(i)}_{\rm SG}=T^{(0)}_{\rm SG}\left(\frac{T^{(k)}_{\rm SG}}{T^{(0)}_{\rm SG}}\right)^{\frac{i}{k}}=T^{(0)}\left(\frac{T^{(k)}_{\rm SG}}{T^{(0)}}\right)^{\frac{i}{k}}. \tag{17b}$$

The right expression of Eq. (17b) is due to the fact that $T^{(0)}_{SG}=T^{(0)}$. At different stages the temperature-based replica exchanging Langevin dynamics (TRXLD) has only different temperatures, while RXSGLD has different guiding temperatures, $\{T^{(i)}_{\rm sg}\}$, and/or different temperatures. In this work, we keep temperatures constant, $T^{(i)}=T^{(0)}$, in all the RXSGLD simulations.

The simulation system is replicated to many replicas. As shown in Fig. 2, one or more replicas are assigned to each stage. The number of replicas in each stage can be different. Between stages, a pair of randomly chosen replicas is exchanged according to the exchange probability, which will be described below. For simplicity, all simulations in this work use only one replica per stage.

The key quantity for a successful replica exchange simulation is the exchange probability. Based on the SGLD

partition function, Eq. (7), it is straightforward to derive the exchange probability between different stages. We use $\mathbf{X}_m^{(i)}$ and T_m to represent conformation i and temperature of stage m. According to the SGLD partition function, Eq. (7), the distribution probability of $\mathbf{X}_m^{(i)}$ is

$$\begin{aligned}&\rho_{\mathrm{SGLD}}\left(X_m^{(i)}\right)\\&= \frac{1}{\Theta_{\mathrm{SGLD}}^{(m)}} \exp\left(-\frac{\lambda_{\mathrm{lf}}^{(\mathrm{m})}\chi_{\mathrm{lf}}^{(\mathrm{m})}\tilde{E}_{\mathrm{p}}^{(i)}}{kT_m} - \frac{\lambda_{\mathrm{hf}}^{(\mathrm{m})}\chi_{\mathrm{hf}}^{(\mathrm{m})}(E_{\mathrm{p}}^{(i)} - \tilde{E}_{\mathrm{p}}^{(i)})}{kT_m}\right)\\&= \frac{1}{\Theta_{\mathrm{SGLD}}^{(m)}} \exp\left(-\tilde{\mu}_{\mathrm{m}}\tilde{E}_{\mathrm{p}}^{(i)} - \mu_{\mathrm{m}}E_{\mathrm{p}}^{(i)}\right). \qquad (18)\end{aligned}$$

Here, the parameters are defined as

$$\tilde{\mu}_m = \frac{\lambda_{\mathrm{lf}}^{(m)}\chi_{\mathrm{lf}}^{(m)} - \lambda_{\mathrm{hf}}^{(m)}\chi_{\mathrm{hf}}^{(m)}}{kT_m} \qquad (19a)$$

and

$$\mu_m = \frac{\lambda_{\mathrm{hf}}^{(m)}\chi_{\mathrm{hf}}^{(m)}}{kT_m}. \qquad (19b)$$

When two replicas exchange, the state of the replica system changes from $\{\ldots,\mathbf{X}_m^{(i)}, \ldots,\mathbf{X}_n^{(j)}, \ldots\}$ to $\{\ldots,\mathbf{X}_m^{(j)},\ldots, \mathbf{X}_n^{(i)},\ldots\}$, and the exchange probability, π_{RX}, can be expressed in the following form:

$$\begin{aligned}\pi_{\mathrm{RX}}\left(\left\{\mathbf{X}_m^{[i]}, \mathbf{X}_n^{[j]}\right\} \to \left\{\mathbf{X}_m^{[j]}, \mathbf{X}_n^{[i]}\right\}\right) &= \frac{\rho_{\mathrm{SGLD}}\left(\mathbf{X}_m^{[j]}\right)\rho_{\mathrm{SGLD}}\left(\mathbf{X}_n^{[i]}\right)}{\rho_{\mathrm{SGLD}}\left(\mathbf{X}_m^{[i]}\right)\rho_{\mathrm{SGLD}}\left(\mathbf{X}_n^{[j]}\right)}\\&= \exp\left(-\tilde{\mu}_m\left(\tilde{E}_{\mathrm{p}}\left(\mathbf{X}_m^{[j]}\right)-\tilde{E}_{\mathrm{p}}\left(\mathbf{X}_m^{[i]}\right)\right) - \mu_m\left(E_{\mathrm{p}}\left(\mathbf{X}_m^{[j]}\right) - E_{\mathrm{p}}\left(\mathbf{X}_m^{[i]}\right)\right) - \tilde{\mu}_n\left(\tilde{E}_{\mathrm{p}}\left(\mathbf{X}_n^{[i]}\right)-\tilde{E}_{\mathrm{p}}\left(\mathbf{X}_n^{[j]}\right)\right)\right.\\&\quad\left. - \mu_n\left(E_{\mathrm{p}}\left(\mathbf{X}_n^{[i]}\right) - E_{\mathrm{p}}\left(\mathbf{X}_n^{[j]}\right)\right)\right)\\&\approx \exp\left(-\left(\tilde{\mu}_m - \tilde{\mu}_n\right)\left(\tilde{E}_{\mathrm{p}}\left(\mathbf{X}_n^{[j]}\right) - \tilde{E}_{\mathrm{p}}\left(\mathbf{X}_m^{[i]}\right)\right) - \left(\mu_m - \mu_n\right)\left(E_{\mathrm{p}}\left(\mathbf{X}_n^{[j]}\right) - E_{\mathrm{p}}\left(\mathbf{X}_m^{[i]}\right)\right)\right). \qquad (20)\end{aligned}$$

Here, we approximate that the low frequency energies at different stages are the same for the same conformation: $\tilde{E}_{\mathrm{p}}(\mathbf{X}_m^{[j]}) \approx \tilde{E}_{\mathrm{p}}(\mathbf{X}_n^{[j]})$, and $\tilde{E}_{\mathrm{p}}(\mathbf{X}_n^{[i]}) \approx \tilde{E}_{\mathrm{p}}(\mathbf{X}_m^{[i]})$. This approximation is accurate if $T_m = T_n$, which is recommended for RXSGLD simulations.

For TRXLD, $\lambda_{\mathrm{lf}}^{(m)} = 1$, $\lambda_{\mathrm{hf}}^{(m)} = 1$, $\chi_{\mathrm{lf}}^{(m)} = 1$, and $\chi_{\mathrm{hf}}^{(m)} = 1$, we have $\tilde{\mu}_m = \tilde{\mu}_n = 0$, and $\mu_m = \beta_m = \frac{1}{kT_m}$, $\mu_n = \beta_n = \frac{1}{kT_n}$, and the exchange probability takes the well known form:[12]

$$\begin{aligned}&\pi_{\mathrm{TRXLD}}\left(\left\{\mathbf{X}_m^{[i]}, \mathbf{X}_n^{[j]}\right\} \to \left\{\mathbf{X}_m^{[j]}, \mathbf{X}_n^{[i]}\right\}\right)\\&\quad= \exp\left(-\left(\beta_m - \beta_n\right)\left(E_{\mathrm{p}}\left(\mathbf{X}_n^{[j]}\right) - E_{\mathrm{p}}\left(\mathbf{X}_m^{[i]}\right)\right)\right). \qquad (21)\end{aligned}$$

To evaluate the exchange probability in a RXSGLD simulation, one needs the low frequency exchange coefficient, $\tilde{\mu}_m = \beta_m(\lambda_{\mathrm{lf}}^{(m)}\chi_{\mathrm{lf}}^{(m)} - \lambda_{\mathrm{hf}}^{(m)}\chi_{\mathrm{hf}}^{(m)})$, and the high frequency exchange coefficient, $\mu_m = \beta_m\lambda_{\mathrm{hf}}^{(m)}\chi_{\mathrm{hf}}^{(m)}$, which in turn, need parameters $\lambda_{\mathrm{lf}}^{(m)}$, $\lambda_{\mathrm{hf}}^{(m)}$, $\chi_{\mathrm{lf}}^{(m)}$, and $\chi_{\mathrm{hf}}^{(m)}$ at each stage. One way to calculate these parameters is from individual SGLD presimulations at all the stage conditions. A more convenient alternative is like in the SGLDfp method,[9] to estimate these parameters during the simulations. The ensemble averages for the calculation of λ_{lf}, λ_{hf}, χ_{lf}, and χ_{hf} in Eqs. (8)–(11) are estimated during simulations as evolving averages,

$$\bar{\alpha}(t) = \left(1 - \frac{\delta t}{t_{\mathrm{est}}}\right)\bar{\alpha}(t - \delta t) + \frac{\delta t}{t_{\mathrm{est}}}\alpha(t). \qquad (22)$$

Here, $\alpha(t)$ represents an instantaneous value of any quantity, and $\bar{\alpha}(t)$ represents its estimated average. The estimation time, t_{est}, is set according to the system size and estimation accuracy. Typically, we choose $t_{\mathrm{est}} = 10t_{\mathrm{L}}$.

At each exchange interval, the exchange probability between a pair of neighboring stages, $\pi_{\mathrm{RX}}(\{\mathbf{X}_m^{[i]}, \mathbf{X}_{m+1}^{[j]}\} \to \{\mathbf{X}_m^{[j]}, \mathbf{X}_{m+1}^{[i]}\})$, is calculated according to Eq. (20) and the acceptance is determined by the Metropolis criterion: $\min\{1, \pi_{\mathrm{RX}}(\{\mathbf{X}_m^{[i]}, \mathbf{X}_{m+1}^{[j]}\} \to \{\mathbf{X}_m^{[j]}, \mathbf{X}_{m+1}^{[i]}\})\}$. Here, m is alternately odd and even stage numbers. Once a replica exchange is accepted, the replicas at the mth and the $(m+1)$th stages are exchanged. Due to the differences in either the temperatures or the self-guiding temperatures, or both, the momentum, $\mathbf{p}_i$, is scaled by a temperature-scaling factor,

$$\mathbf{p}_i'^{(n)} = s_{mn}\mathbf{p}_i^{(m)}, \qquad (23a)$$

$$\mathbf{p}_i'^{(m)} = s_{nm}\mathbf{p}_i^{(n)}, \qquad (23b)$$

$$s_{mn} = \sqrt{\frac{T_n}{T_m}} = \frac{1}{s_{nm}}. \qquad (24)$$

Also, the low frequency momentum, $\tilde{\mathbf{p}}_i$, is scaled by a low frequency temperature-scaling factor,

$$\tilde{\mathbf{p}}_i'^{(n)} = \tilde{s}_{mn}\tilde{\mathbf{p}}_i^{(m)}, \qquad (25a)$$

$$\tilde{\mathbf{p}}_i'^{(m)} = \tilde{s}_{nm}\tilde{\mathbf{p}}_i^{(n)}, \qquad (25b)$$

$$\tilde{s}_{mn} = \sqrt{\frac{\tilde{T}_n}{\tilde{T}_m}} = \frac{1}{\tilde{s}_{nm}}. \qquad (26)$$

Between exchanges, standard SGLD simulations are performed at all stages. The SGLD simulation details can be found elsewhere.[1,2]

C. Simulation details

The RXSGLD method has been implemented and is available in CHARMM[28,29] version c36 and AMBER[30] version 12. The simulation results reported here are obtained with CHARMM. Eight stages are used for all replica exchange simulations reported here. Exchanges are attempted every 1000 time steps. Simulation information and trajectories are output by stages separately for easy processing and analysis. All RXSGLD simulations use a local average time of 0.2 ps. All RXSGLD simulations in this work have $T_{\rm SG}^{(i)} > T_{\rm SG}^{(0)} = T^{(0)}$ and $T^{(i)} = T^{(0)}$ and all TRXLD simulations have $T_{\rm SG}^{(i)} = T^{(0)}$ and $T^{(i)} > T^{(0)}$, with $T_{\rm SG}^{(i)}$ and $T^{(i)}$ on stage i are calculated with Eqs. (17a) and (17b).

III. RESULTS AND DISCUSSIONS

In this section, through three example systems we examine the conformational sampling accuracy and conformational search efficiency of the RXSGLD method. There are many replica exchange methods available,[12,15–19,21–26,31–75] it is not practical to have a thorough comparison with all the methods. In this work, we only compare RXSGLD with TRXLD to provide a general understanding of its performance.

A. The skewed double well system

First, we use this thoroughly studied simple system to examine the accuracy and efficiency of the RXSGLD method. 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 skewed double well energy surface of the following form:

$$\varepsilon_{\rm p}(x,y,z) = \frac{a}{w^2}(x^2+z^2) + \frac{b}{w^4}y^2(y-w)^2 + \frac{s}{w}y. \quad (27)$$

The parameter, a, defines the energy surface in the x and z dimensions, and the parameters, b and w, defines a double well potential in the y dimension. There are two wells, one at $y = 0$ and the other at $y = w$. The skew parameter, s, defines the energy difference between the two wells. There are only three degrees of freedom and the partition function can be separated for each degree of freedom,

$$\Theta_{\rm xz} = \frac{\pi kT w^2}{a}, \quad (28a)$$

$$\Theta_{\rm y} = \int_{-\infty}^{\infty} \exp\left(-\frac{\frac{b}{w^4}y^2(w-y)^2 + \frac{s}{w}y}{kT}\right)dy. \quad (28b)$$

The ensemble average properties can be calculated from the following equations:

$$E_{xz} = kT, \quad (29a)$$

$$E_y = \frac{1}{\Theta_{\rm y}}\int_{-\infty}^{\infty}\left(\frac{b}{w^4}y^2(w-y)^2 + \frac{s}{w}y\right)e^{-\frac{\frac{b}{w^4}y^2(w-y)^2+\frac{s}{w}y}{kT}}dy, \quad (29b)$$

$$\rho_{xz}(r_{xz}) = \frac{2a}{kTw^2}e^{-\frac{ar_{xz}^2}{kTw^2}}, \quad (30a)$$

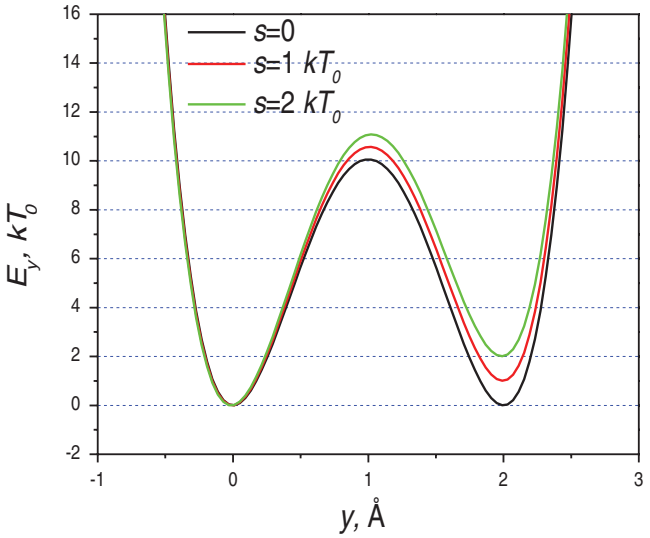

FIG. 3. The skewed double well potentials in the y-dimension. The potential function is shown in Eq. (27). The skew parameter, s, controls the relative depths of the two wells. The energy barrier between the two wells is about $10\,kT_0$.

$$\rho_y(y) = \frac{1}{\Theta_{\rm y}}e^{-\frac{\frac{b}{w^4}y^2(w-y)^2+\frac{s}{w}y}{kT}}. \quad (30b)$$

Here, $r_{xz} = \sqrt{x^2+z^2}$. In this work, we chose $a = 20000kT_0$, $b = 160kT_0$, and $w = 2$Å. Three double well systems with different skew parameters, s $= 0$, kT_0, and $2kT_0$, were examined. Here, $T_0 = 50$ K is the base temperature. Figure 3 shows the three skewed double well potentials in the y dimension. The two wells are at y $= 0$ Å and y $= 2$ Å and have different depths depending on the skew parameter. An argon atom was simulated on the energy surfaces. 8 stages and 8 replicas were used for each replica exchange simulation. The TRXLD simulations were carried out at $T = 50/100$ K and the RXSGLD simulations were at $T = 50$ K, $T_{\rm SG} = 50/100$ K. A collision frequency of 100/ps was used to force the motion of the argon atom into a random walk in nature. A time step of 1 fs was used and the simulation length was 100 ns for each simulation. The local averaging time was set to $t_{\rm L} = 0.2$ps for all RXSGLD simulations.

Ensemble distributions are the first thing to examine to validate a simulation method. Figure 4 compares the average energies of each stage in the TRXLD, RXSGLD, as well as the numerical solutions from Eq. (29a) and Eq. (29b). As can be seen, at the base stages, both methods produce correct ensemble average energies for the three systems. The base stage averages are listed in Table I for a quantitative comparison. These results serve as the first evidence that RXSGLD produces correct ensemble average properties at the base stages. Comparing the average energies of all stages, we can see that RXSGLD has smaller energy differences across the stages than TRXLD, this is because SGLD only enhances the low frequency motions and has less perturbation on conformational distribution than raising temperatures.

To demonstrate the different effects on conformational distributions, we plotted the distributions in the y dimension and in the x-z dimensions from the TRXLD and RXSGLD

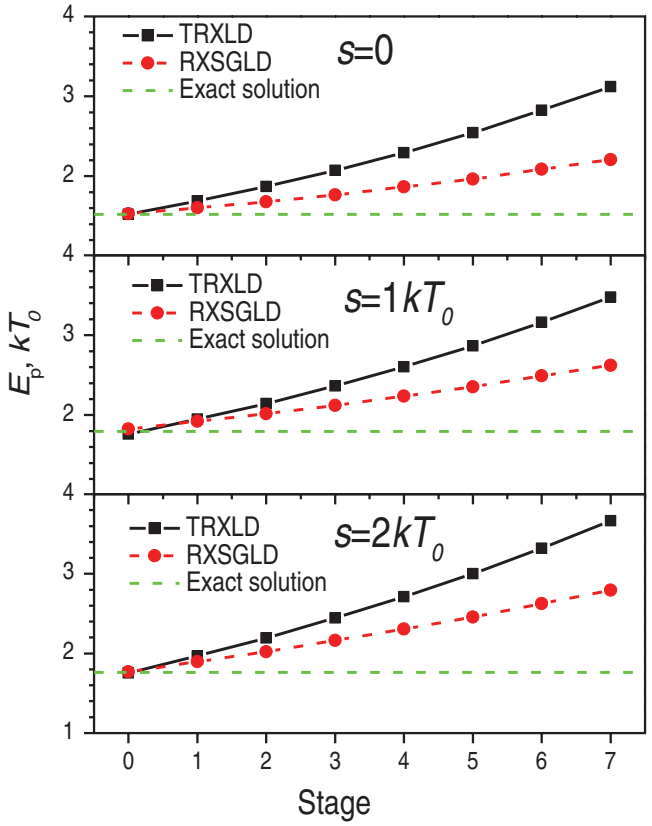

FIG. 4. Average stage energies in the TRXLD and RXSGLD simulations. The solutions from Eq. (29a) are shown as dashed lines. The TRXLD simulations has 8 stages with $T = 50/100$ K, and the RXSGLD simulations have 8 stages with $T_{SG} = 50/100$ K and $T = T_0 = 50$ K.

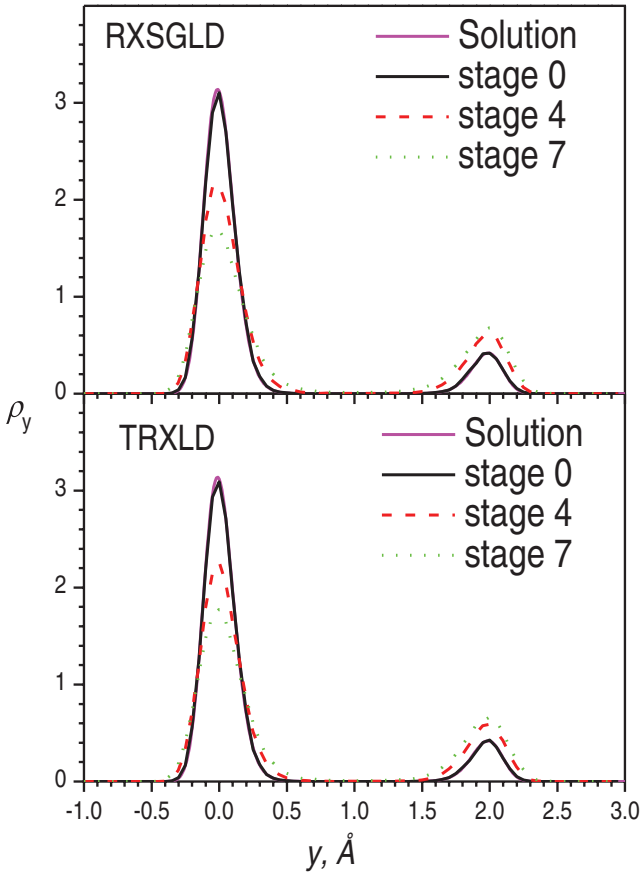

FIG. 5. The y-coordinate distributions at stage 0, 4, and 7 in the TRXLD ($T = 50/100$ K), RXSGLD ($T_{SG} = 50/100$ K and $T = T_0 = 50$ K) simulations. The data is for the skewed double well system with $s = 2kT_0$.

simulations in Fig. 5 and Fig. 6. From Fig. 5, we can see clearly that, at the base stages, i.e., stage 0, both simulations produced distributions that agreed well with the solution of Eq. (30b). These results further demonstrate that RXSGLD preserves canonical ensemble distributions at the base stage. At stage 4 and stage 7, the y-distributions deviate from the numerical solutions due to elevated temperatures in the TRXLD simulations or due to the guiding effects in the RXSGLD simulations. These results demonstrate that in the y dimension, where the atom has a low frequency motion, the guiding forces and raising temperatures have similar effects.

TABLE I. Average properties of the skewed double well systems on the base stage ($T = 50$ K). x_1 is the fraction of distribution in the well near $y = 0$ Å.

Skew parameter	Methods	Epot, kT	x_1
0	solution	1.525	0.5
	TRXLD	1.513 ± 0.002	0.477 ± 0.006
	RXSGLD	1.520 ± 0.002	0.491 ± 0.006
kT	solution	1.796	0.738
	TRXLD	1.760 ± 0.010	0.763 ± 0.005
	RXSGLD	1.827 ± 0.009	0.705 ± 0.006
$2kT$	solution	1.762	0.878
	TRXLD	1.754 ± 0.012	0.881 ± 0.003
	RXSGLD	1.772 ± 0.011	0.876 ± 0.003

Figure 6 shows the distributions in the x-z dimensions where the particle moves in high frequency modes. Clearly, at stage 0, both simulation results agree well with the solution of Eq. (30a). However, at other stages the distributions from TRXLD and from RXSGLD are different. The distributions at stages 4 and 7 from TRXLD deviate significantly from that at the base stage, indicating raising temperatures also change the distributions in the x-z dimensions. However, the distributions on stages 4 and 7 from the RXSGLD simulations are almost the same as on the base stages. That means the guiding forces change little in the x-z dimensions, where the atom has high frequency motions. This difference contrasts SGLD and high temperature simulations, and explains why the energy differences across the stages in RXSGLD are smaller than those in TRXLD. RXSGLD stages differ only in the low frequency motions while TRXLD stages are different in all motions.

The conformational search efficiency can be examined by the convergence in the conformational distribution. We use the distribution fraction in the well near y = 0, denoted as x_1, to examine the convergence. Figure 7 shows x_1 on the base stages during the replica exchange simulations, as well as the solutions using Eqs. (30a) and (30b). By the end of the simulations, all results approach the solutions fairly well (Table I), which again demonstrates that RXSGLD can sample the conformation correctly. As can be seen from Fig. 7, the RXSGLD results approach the solutions faster than the TRXLD results

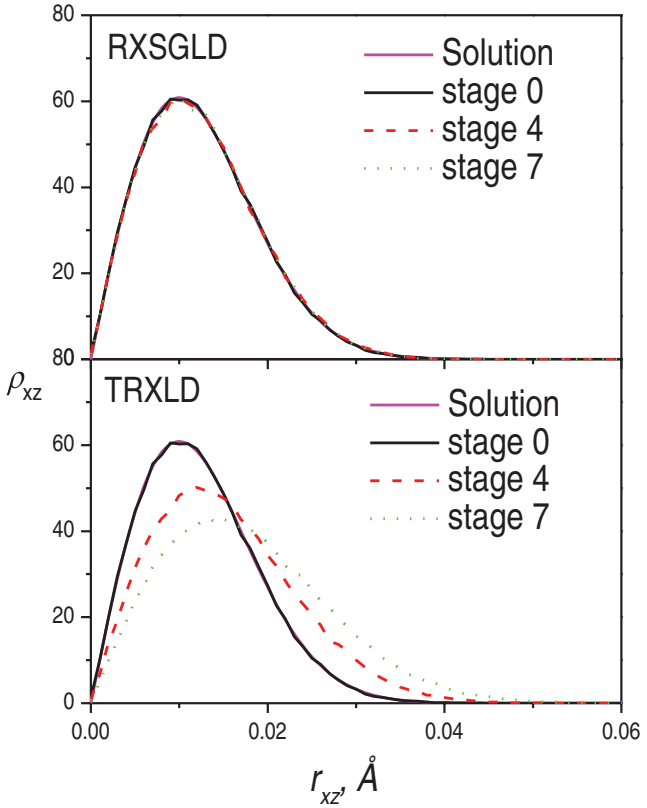

FIG. 6. The x-z distributions at stage 0, 4, and 7 in the TRXLD ($T = 50/100$ K), RXSGLD ($T_{SG} = 50/100$Kand $T = T_0 = 50$K) simulations. The data is for the skewed double well system with $s = 2kT_0$.

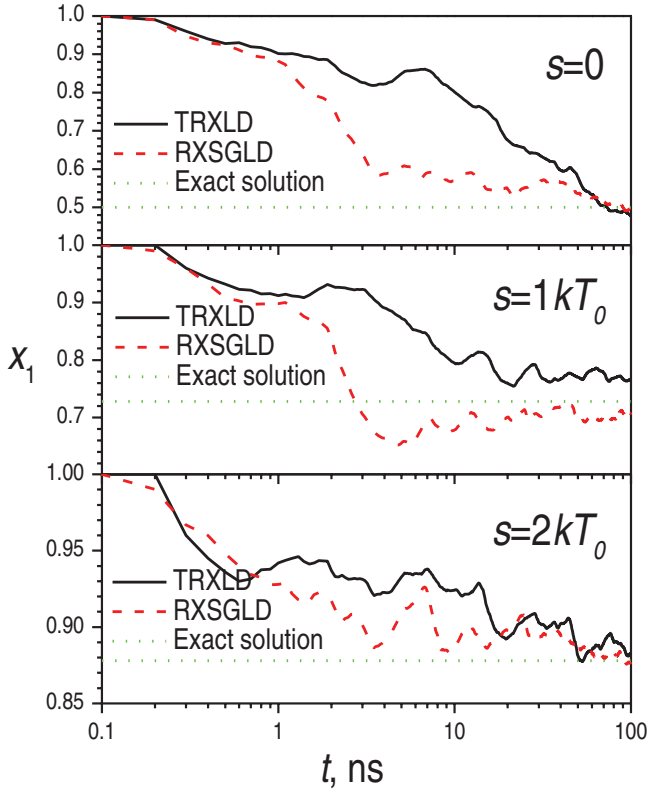

FIG. 7. Distribution fractions of well 1, x_1, on the base stages ($T_0 = 50$ K) from the TRXLD ($T = 50/100$ K) and RXSGLD ($T_{SG} = 50/100$ K and $T = T_0 = 50$ K) simulations. The solutions of Eq. (30) are shown as dashed lines.

for all three cases. These results prove that RXSGLD has better conformational searching efficiency than TRXLD. For this one-particle system, the improvement is not that significant. As will be seen below, differences will be more significant for larger systems.

B. The β-hairpin folding peptide with implicit solvent

Protein folding simulation is a major application of enhanced sampling methods. Because of a large number of degrees of freedom, protein conformational space is huge, making protein folding a challenge for computational studies. Replica exchange methods have been a preferred choice for protein folding studies due to its strong conformational searching ability. Here, we chose a 9-residue β-hairpin folding peptide to examine how the RXSGLD method performs for this realistic system. This 9-residue peptide studied here was designed by Blanco *et al.*[76] and was modified from the β-hairpin of α-amylase inhibitor tendamistat (residues 15-23). The amino acid sequence of this peptide is: Tyr(1)-Gln(2)-Asn(3)-Pro(4)-Asp(5)-Gly(6)-Ser(7)-Gln(8)-Ala(9). The screened Coulomb potential implicit solvent model (SCPISM) was used[77,78] to describe the solvent effects. Using an implicit solvation model in this example allows a fairly thorough sampling of the conformational space so that we can compare the conformational distributions quantitatively.

We performed an 8-stage TRXLD simulation ($T = 274/400$ K) and an 8-stage RXSGLD simulation ($T = 274$ K, $T_{sg} = 274/400$ K). All simulations started from a fully extended conformation and were 200 ns in length. The collision frequency was set to 1/ps.

Because there are many degrees of freedom in proteins, it is often more convenient to cluster protein conformations to provide a simplified description of the conformational distribution. Here, we propose a subset indexing clustering (SIC) method as described below to cluster simulation conformations.

(1) Subsets: separate the conformational variables to subsets:

$$\{x_1, x_2, \ldots, x_n\} = \{(x_{a1}, x_{a2}, \ldots), (x_{b1}, x_{b2}, \ldots), \ldots, (x_{m1}, x_{m2}, \ldots)\} = \{s_a, s_b, \ldots, s_m\}.$$

The conformation variables are chosen based on research interest, such as dihedral angles of amino acids, hydrogen bonds, or secondary structures.

(2) Regions: define regions in the distribution of each subset variables. The distribution is generated from simulation results. For a subset, s_i, there are k_i regions: $\{R_i(1), R_i(2), \ldots, R_i(k_i)\}$, and $s_i \in R_i(1) + R_i(2) + \cdots + R_i(k_i)$. The region index is defined as

$$I_i = \begin{cases} 1 & s_i \in R_i(1) \\ 2 & s_i \in R_i(2) \\ \cdots & \cdots \\ k_i & s_i \in R_i(k_i) \end{cases}. \quad (31)$$

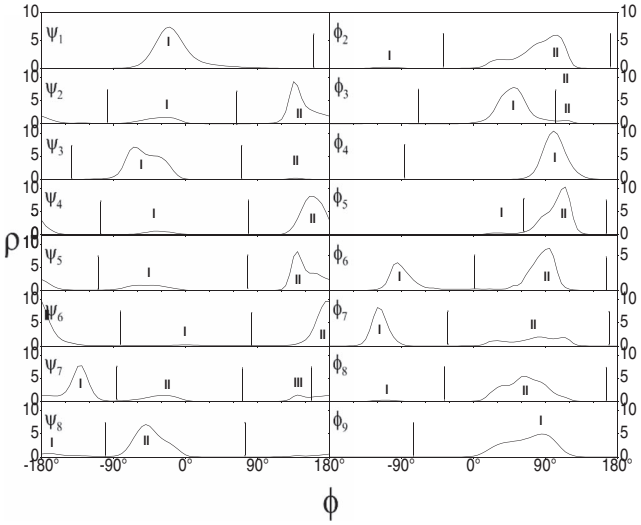

FIG. 8. The distribution of the 16 backbone dihedral angles in the 9-reside β-hairpin folding peptide calculated from the replica exchange simulations. One to three regions are defined as labeled for each dihedral angle for the subset index clustering.

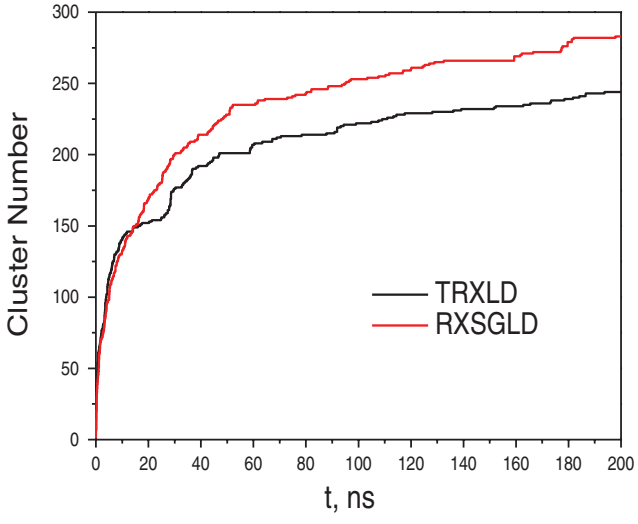

FIG. 9. Total numbers of clusters searched in the TRXLD ($T = 274/500$ K) and RXSGLD ($T_{SG} = 274/400$ K and $T = T_0 = 274$ K) simulations. The 9-residue β-hairpin folding peptide is simulated with an implicit salvation model.

(3) Clusters: A cluster represents a unique list of the region indexes of all subsets, $\{I_1, I_2, \ldots, I_m\}$. Total possible cluster numbers is the product of the numbers of subset regions, $N_c = \prod_{i=1}^{m} k_i$. A conformation is identified as a SIC by indexing its subset regions,

$$\mathrm{SIC}\,(x_1, x_2, \ldots, x_n) = \{I_1, I_2, \ldots, I_m\}.$$

Because the SIC method does not evaluate pairwise properties between conformations, it is very efficient and its computing cost is an order of N. For proteins, it is a natural choice to choose the ϕ, ψ dihedral angles of each amino acid as a subset. The dihedral angle distribution regions for each amino acid in protein structures have been well documented and can be used for the subset index clustering of protein conformations. For this small peptide of 9-amino acids, it is more informational to use each of its 16 ϕ, ψ dihedral angles (tyr(1) does not have ϕ and Ala(9) does not have ψ) as a subset to perform the clustering. Figure 8 plots the population of the 16 dihedral angles from all stages of the replica exchange simulations. Based on the distributions, we define $k_i = 1, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 3, 2, 2, 1$ regions for the 16 dihedral angles, respectively. From the number of regions we can calculate that there are total 12288 possible clusters. The actual number of clusters visited by all the replicas in both the TRXLD and RXSGLD simulations is 1145, among which TRXLD visited 1056 clusters, and RXSGLD visited 730 clusters. However, at the base stage, TRXLD visited 244 clusters and RXSGLD visited 283 clusters. In other words, TRXLD searched many high temperature conformations. It is not surprising that at high temperatures, the replicas searched more conformational space. However, at high temperatures, the replicas very likely climbed into conformations that would have low probabilities at the base temperature and failed to reach the base stage. While in the RXSGLD simulation, the guiding forces accelerate conformational searching but keep the focus on the high probability region. In other words, raising temperatures increases accessible conformational space that lead to a search of low populated regions, while the guiding forces accelerate conformational search without significantly increasing the accessible conformational space so that the conformational search can be more concentrated on high population regions. We can define a conformational searching relevancy (CSR) as the number of clusters on the base stage versus those on all stages. For the TRXLD simulation, the CSR is 244/1056 = 23.1%, while for the RXSGLD simulation, the CSR is 283/730 = 38.8%. These CSR values mean that in the TRXLD simulation only 23.1% of searched conformations are of interest, while in the RXSGLD simulation, about 38.8% searched conformations are relevant.

Figure 9 shows the number of clusters visited at the base stage during the TRXLD and RXSGLD simulations. Clearly, the number of clusters searched in RXSGLD increases faster than that in TRXLD, demonstrating RXSGLD has stronger conformational searching ability.

Because this β-hairpin folding peptide is a more realistic system than the double well system discussed above, it is more interesting to examine how its conformational space is sampled in RXSGLD as compared to TRXLD. Figure 10 plots the cluster populations in the RXSGLD simulation against that in the TRXLD simulation. For easy plotting in a logarithm scale, the cluster populations are counted from 1. In other words, a population of 1 means the cluster has not been visited. As can be seen from Fig. 10, the cluster population in the RXSGLD simulation correlates with that in the TRXLD simulation fairly well, even though significant fluctuation exists in those low population clusters. Figure 10 also shows a linear fit of the data (the red dashed line), which is very close to the ideal equation, $y = x$. This result indicates that RXSGLD can provide correct conformational sampling for this realistic peptide system.

In Fig. 10, we can see there are two high population points, which represent two major clusters. We examined the two structures and found both are hairpin-like conformations.

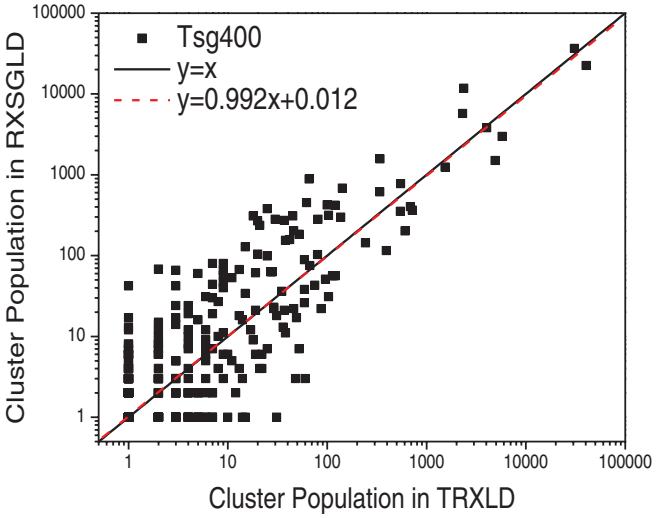

FIG. 10. Comparison of the cluster populations between the TRXLD (T = 274/400 K) result and the RXSGLD (T_{SG} = 274/400 K and $T = T_0$ = 274 K) result. For convenience in plotting in the logarithm scale, the cluster population counts start from 1. The 9-residue β-hairpin folding peptide is simulated with an implicit salvation model. The subset indexing clustering method is described in the text.

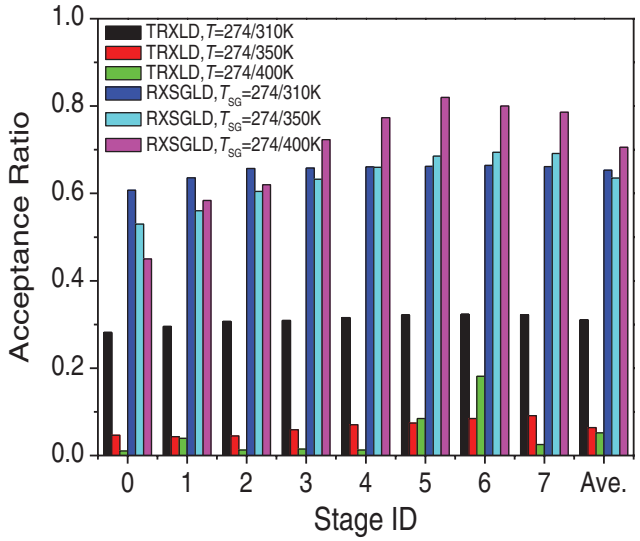

FIG. 11. Acceptance ratios at each stage in the TRXLD and RXSGLD simulations. The 9-residue β-hairpin folding peptide is simulated with explicit water.

A major difference between the two conformations is that Gly(6) changes its dihedral angle, φ_6, from region I to region II (see Fig. 8).

C. The β-hairpin folding peptide in aqueous solutions

A major challenge to temperature-based replica exchange simulation is its difficulties with large systems. When a system is large, the same temperature difference will cause a large energy change, which reduces the exchange probability exponentially according to Eq. (21). This phenomenon is termed as not size extensive. To achieve reasonable replica exchange probability, the number of stages and replicas, must be increased proportionally. When the number of stages is large, the diffusion of replicas from the top stage to the base stage will take a long time, which further reduces the conformational sampling efficiency.

We use an aqueous solution of this β-hairpin folding peptide to examine the application of RXSGLD to large systems. The peptide was dissolved in a box of 829 TIP3P[79] water. A sodium ion was placed in the box to neutralize the system. The box size was 30 × 30 × 30 Å.[3] A collision frequency of 1/ps was used to maintain the temperature. We used the CHARMM 22 force field[80] to calculate energies and used the 3D IPS method with a local region radius of 10 Å for electrostatic and Lennard-Jones energy calculation.[81–83] We performed three 8-stage TRXLD simulations with T = 274/310 K, 274/350 K, and 274/400 K, respectively and three 8-stage RXSGLD simulations at T = 274 K, T_{SG} = 274 K/310 K, 274/350 K, and 274/400 K, respectively. All simulations started from a fully extended conformation and were 20 ns in length.

The efficiency of replica exchange simulations depends on the acceptance ratio of replica exchange. Figure 11 shows the acceptance ratios on each stage in these simulations. When the temperature range is small, T = 274/310K, the average acceptance ratio for TRXLD is 31.1%, which is acceptable. However, for more meaningful temperature

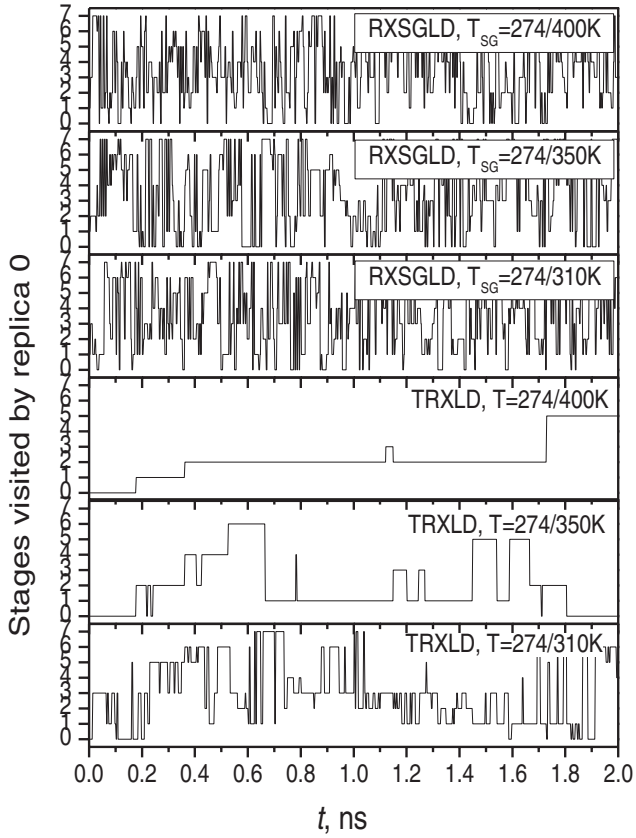

FIG. 12. Stage identities visited by replica 0 in the TRXLD and RXSGLD simulations. The 9-residue β-hairpin folding peptide is simulated with explicit water.

ranges, T = 274/350 K and T = 274/400 K, the average acceptance ratios are 6.4% and 5.2%, which are too low. The smaller the temperature range is, the less conformational searching power a TRXLD simulation can reach. On the other hand, the lower the acceptance ratio is, the less the high temperature stages contribute to the conformational search. Therefore, to accelerate conformational searching for this large system, the number of stages must be increased to obtain a reasonable acceptance ratio and to utilize a reasonable large temperature range. In contrast to TRXLD, all three RXSGLD simulations have high acceptance ratios (65.3%, 63.5%, and 70.2%).

The acceptance ratios may reflect only the transition between neighboring stages. A better picture of a replica exchange simulation is the diffusion of replicas across the stages. Fig. 12 shows the stages visited by replica 0 during the simulations. For TRXLD at T = 274/310K, replica 0 took more than 0.6 ns to reach stage 7, while at T = 274/350K and T = 274/400K, replica 0 did not reach stage 7 until 2.35 ns and 3.28 ns (beyond the plotting range), respectively. While in the three RXSGLD simulations, replica 0 reached stage 7 within 0.1 ns. The fast diffusion across the stages allowed the sampling to take advantage of the maximum enhanced conformational searching at the top stage.

As defined by the exchange probability equations, Eqs. (20) and (21), the reason for such large differences in the exchange ratios is clearly due to the difference in the energy distributions. Fig. 13 shows the potential energy distributions

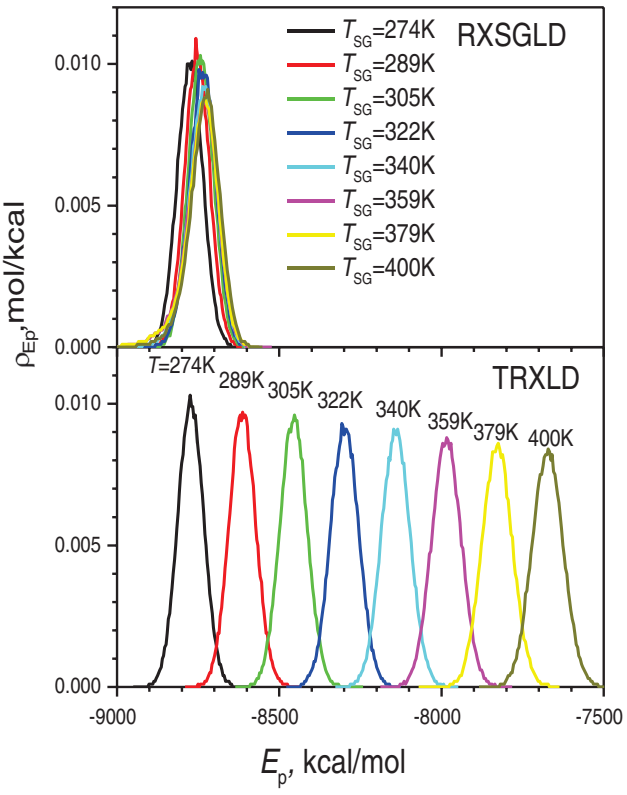

FIG. 13. Potential energy distributions at each stage in the TRXLD (T = 274/400 K) and RXSGLD (T_{SG} = 274/400 K and $T = T_0$ = 274 K) simulations. The 9-residue β-hairpin folding peptide is simulated with explicit water.

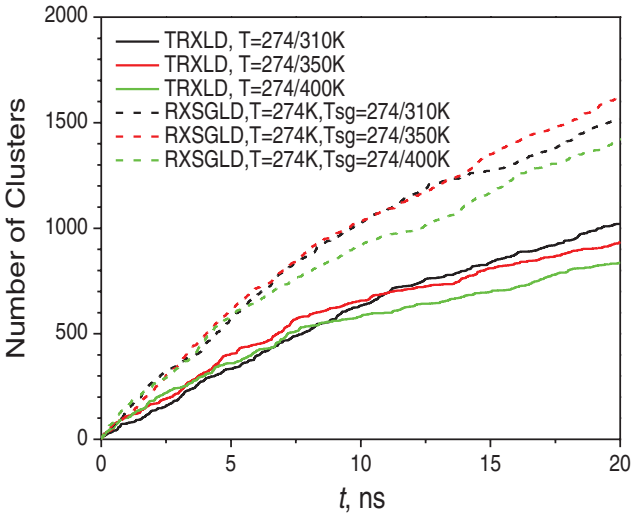

FIG. 14. The total numbers of clusters searched during the TRXLD and RXSGLD simulations. The 9-residue β-hairpin folding peptide is simulated with explicit water.

at different stages in the TRXLD simulation at $T = 274/400$ K and in the RXSGLD simulations at $T_{SG} = 274/400$ K. Clearly, in the TRXLD simulation, the energy distributions are very different from each other and there are very small overlaps between the neighboring stages, which makes exchange probability very low. In the RXSGLD simulation, the stage energy distributions are very close to each other and they overlap with each other significantly. This large overlap in stage energies makes the acceptance ratio high in RXSGLD simulations.

We can compare the conformational searching by examining conformational clusters searched during the simulations. Using the SIC method described above, we clustered the conformations at the base stage of the TRXLD and RXSGLD simulations. Fig. 14 compares the numbers of clusters searched in the simulations. In all RXSGLD simulations, the cluster numbers increased significantly faster than in the TRXLD simulations. These results again prove that RXSGLD has stronger conformational searching ability.

IV. CONCLUSIONS

This work presents the replica exchanging self-guided Langevin dynamics (RXSGLD) simulation method. This method uses SGLD to enhance conformational searching and has high replica exchange efficiency. By avoiding temperature elevation, this method can be applied to large systems with high replica exchange efficiency and can use relatively few replicas to save computing cost. By incorporating the SGLD partition function into the exchanging probability, this method samples the canonical ensemble distribution at the base stage. Therefore, the RXSGLD method can be used as an alternative to the force-momentum based self-guided Langevin dynamics (SGLDfp) to directly sample canonical ensemble without the need of reweighting. Using the skewed double well systems and a β-hairpin folding peptide with implicit salvation model, we demonstrate that RXSGLD produces correct ensemble distributions while improving conformational searching and sampling. Through the β-hairpin folding sim-

ulations in explicit water, we demonstrate that RXSGLD has better size extensiveness than TRXLD.

ACKNOWLEDGMENTS

This research was supported by the Intramural Research Program of the NIH, NHLBI. We thank Eunice Wu for proofreading the manuscript.

[1]X. Wu and B. R. Brooks, Chem. Phys. Lett. **381**, 512 (2003).
[2]X. Wu and B. R. Brooks, J. Chem. Phys. **134**, 134108 (2011).
[3]X. Wu, A. Damjanovic, and B. R. Brooks, "Efficient and unbiased sampling of biomolecular systems in the canonical ensemble: A review of self-guided Langevin dynamics," in *Advances in Chemical Physics*, edited by S. A. Rice, and A. R. Dinner, (Wiley, Hoboken, 2012), Vol. 150, pp. 255.
[4]A. Damjanovic, X. Wu, E. B. Garcia-Moreno, and B. R. Brooks, Biophys. J. **95**, 4091 (2008).
[5]A. Damjanovic, B. T. Miller, T. J. Wenaus, P. Maksimovic, E. Bertrand Garcia-Moreno, and B. R. Brooks, J. Chem. Info. Model. **48**, 2021 (2008).
[6]A. Damjanovi, E. B. Garcia-Moreno, and B. R. Brooks, Proteins: Struct., Funct., Bioinf. **76**, 1007 (2009).
[7]P. Y. Pendse, B. R. Brooks, and J. B. Klauda, J. Mol. Biol. **404**, 506 (2010).
[8]C. I. Lee and N. Y. Chang, Biophys. Chem. **151**, 86 (2010).
[9]X. Wu and B. R. Brooks, J. Chem. Phys. **135**, 204101 (2011).
[10]C. J. Geyer and E. A. Thompson, J. Am. Stat. Assoc. **90**, 909 (1995).
[11]R. H. Swendsen and J.-S. Wang, Phys. Rev. Lett. **57**, 2607 (1986).
[12]Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **314**, 141 (1999).
[13]H. Li, G. Li, B. A. Berg, and W. Yang, J. Chem. Phys. **125**, 144902 (2006).
[14]H. Li and W. Yang, J. Chem. Phys. **126**, 114104 (2007).
[15]A. J. Lee and S. W. Rick, J. Chem. Phys. **131**, 174113 (2009).
[16]S. W. Rick, J. Chem. Phys. **126**, 054102 (2007).
[17]X. Li, C. P. O'Brien, G. Collier, N. A. Vellore, F. Wang, R. A. Latour, D. A. Bruce, and S. J. Stuart, J. Chem. Phys. **127**, 164116 (2007).
[18]X. Li, R. A. Latour, and S. J. Stuart, J. Chem. Phys. **130**, 174106 (2009).
[19]A. Mitsutake, Y. Sugita, Y. Okamoto, R. Faller, Q. Yan, J. J. de Pablo, H. Fukunishi, O. Watanabe, S. Takada, F. Calvo, S. W. Rick, P. Liu, G. A. Voth, H. Kamberaj, A. van der Vaart, P. Brenner, C. R. Sweet, D. VonHandorf, J. A. Izaguirre, S. Trebst, M. Troyer, and U. H. E. Hansmann, J. Chem. Phys. **118**, 6664 (2003).
[20]H. Fukunishi, O. Watanabe, and S. Takada, J. Chem. Phys. **116**, 9058 (2002).
[21]S. Jang, S. Shin, and Y. Pak, Phys. Rev. Lett. **91**, 058305 (2003).
[22]P. Liu, X. Huang, R. Zhou, and B. J. Berne, J. Phys. Chem. B **110**, 19018 (2006).
[23]P. Liu, B. Kim, R. A. Friesner, and B. J. Berne, Proc. Natl. Acad. Sci. U.S.A. **102**, 13749 (2005).
[24]P. Liu and G. A. Voth, J. Chem. Phys. **126**, 045106 (2007).
[25]A. J. Ballard and C. Jarzynski, Proc. Natl. Acad. Sci. U.S.A. **106**, 12224 (2009).
[26]P. Kar, W. Nadler, and U. H. Hansmann, Phys. Rev. E **80**, 056703 (2009).
[27]M. S. Lee and M. A. Olson, J. Chem. Theory Comput. **6**, 2477 (2010).
[28]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**, 1545 (2009).
[29]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).
[30]D. A. Case, T. A. Darden, T. E. Cheatham III, C. L. Simmerling, J. Wang, R. Duke, R. Luo, R. C. Walker, W. Zhang, K. M. Merz, B. P. Roberts, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossvary, 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 11* (University of California, San Francisco, 2010).
[31]J. Kim, T. Keyes, and J. E. Straub, J. Chem. Phys. **132**, 224107 (2010).
[32]L. Su and R. I. Cukier, J. Phys. Chem. B **113**, 9595 (2009).
[33]A. Shmygelska and M. Levitt, Proc. Natl. Acad. Sci. U.S.A. **106**, 1415 (2009).
[34]A. Mitsutake and Y. Okamoto, Phys. Rev. E **79**, 047701 (2009).
[35]A. Mitsutake, J. Chem. Phys. **131**, 094105 (2009).
[36]J. Kim and J. E. Straub, J. Chem. Phys. **130**, 144114 (2009).
[37]J. Kim, T. Keyes, and J. E. Straub, J. Chem. Phys. **130**, 124112 (2009).
[38]S. Kannan and M. Zacharias, J. Struct. Biol. **166**, 288 (2009).
[39]H. Kamberaj and A. van der Vaart, J. Chem. Phys. **130**, 074906 (2009).
[40]M. Fajer, R. V. Swift, and J. A. McCammon, J. Comput. Chem. **30**, 1719 (2009).
[41]C. Czaplewski, S. Kalinowski, A. Liwo, and H. A. Scheraga, J. Chem. Theory Comput. **5**, 627 (2009).
[42]J. Curuksu and M. Zacharias, J. Chem. Phys. **130**, 104110 (2009).
[43]Y. Chebaro, X. Dong, R. Laghaei, P. Derreumaux, and N. Mousseau, J. Phys. Chem. B **113**, 267 (2009).
[44]J. Zhang, W. Li, J. Wang, M. Qin, and W. Wang, Proteins **72**, 1038 (2008).
[45]H. Shen, C. Czaplewski, A. Liwo, and H. A. Scheraga, J. Chem. Theory Comput. **4**, 1386 (2008).
[46]W. Nadler and U. H. Hansmann, J. Phys. Chem. B **112**, 10386 (2008).
[47]M. Kouza, C. K. Hu, and M. S. Li, J. Chem. Phys. **128**, 045103 (2008).
[48]J. Hritz and C. Oostenbrink, J. Chem. Phys. **128**, 144121 (2008).
[49]E. Gallicchio, R. M. Levy, and M. Parashar, J. Comput. Chem. **29**, 788 (2008).
[50]M. Fajer, D. Hamelberg, and J. A. McCammon, J. Chem. Theory Comput. **4**, 1565 (2008).
[51]G. Bellesia, S. Lampoudi, and J. E. Shea, Methods Mol. Biol. **474**, 133 (2008).
[52]M. Athenes and F. Calvo, ChemPhysChem **9**, 2332 (2008).
[53]R. Zhou, Methods Mol. Biol. **350**, 205 (2007).
[54]W. Zheng, M. Andrec, E. Gallicchio, and R. M. Levy, Proc. Natl. Acad. Sci. U.S.A. **104**, 15340 (2007).
[55]C. Thachuk, A. Shmygelska, and H. H. Hoos, BMC Bioinf. **8**, 342 (2007).
[56]L. Su and R. I. Cukier, J. Phys. Chem. B **111**, 12310 (2007).
[57]W. Nadler and U. H. Hansmann, Phys. Rev. E **76**, 057102 (2007).
[58]Y. Mu, Y. Yang, and W. Xu, J. Chem. Phys. **127**, 084119 (2007).
[59]M. B. Kubitzki and B. L. de Groot, Biophys. J. **92**, 4262 (2007).
[60]S. Kannan and M. Zacharias, Proteins **66**, 697 (2007).
[61]H. Kamberaj and A. van der Vaart, J. Chem. Phys. **127**, 234102 (2007).
[62]X. Huang, M. Hagen, B. Kim, R. A. Friesner, R. Zhou, and B. J. Berne, J. Phys. Chem. B **111**, 5405 (2007).
[63]J. Hritz and C. Oostenbrink, J. Chem. Phys. **127**, 204104 (2007).
[64]M. Hagen, B. Kim, P. Liu, R. A. Friesner, and B. J. Berne, J. Phys. Chem. B **111**, 1416 (2007).
[65]P. Brenner, C. R. Sweet, D. VonHandorf, and J. A. Izaguirre, J. Chem. Phys. **126**, 074103 (2007).
[66]D. M. Zuckerman and E. Lyman, J. Chem. Theory Comput. **2**, 12001202 (2006).
[67]W. Zhang, C. Wu, and Y. Duan, J. Chem. Phys. **123**, 154105 (2005).
[68]M. Habeck, M. Nilges, and W. Rieping, Phys. Rev. Lett. **94**, 018105 (2005).
[69]E. Gallicchio, M. Andrec, A. K. Felts, and R. M. Levy, J. Phys. Chem. B **109**, 6722 (2005).
[70]X. Cheng, G. Cui, V. Hornak, and C. Simmerling, J. Phys. Chem. B **109**, 8220 (2005).
[71]A. Mitsutake and Y. Okamoto, J. Chem. Phys. **121**, 2491 (2004).
[72]Y. M. Rhee and V. S. Pande, Biophys. J. **84**, 775 (2003).
[73]P. H. Nguyen, J. Chem. Phys. **132**, 144109 (2010).
[74]S. G. Itoh, H. Okumura, and Y. Okamoto, J. Chem. Phys. **132**, 134105 (2010).
[75]R. I. Cukier, J. Chem. Phys. **134**, 045104 (2011).
[76]F. J. Blanco, M. A. Jimenez, J. Herranz, M. Rico, J. Santoro, and J. L. Nieto, J. Am. Chem. Soc. **115**, 5887 (1993).
[77]S. A. Hassan and E. L. Mehler, Proteins **47**, 45 (2002).
[78]S. A. Hassan, E. L. Mehler, D. Zhang, and H. Weinstein, Proteins **51**, 109 (2003).
[79]W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).

[80]A. D. MacKerell, Jr., D. Bashford, M. Bellott, R. L. Dunbrack, Jr., J. D. Evanseck, M. J. Field, S. Fisher, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, III, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus, J. Phys. Chem. B **102**, 3586 (1998).
[81]X. Wu and B. R. Brooks, J. Chem. Phys. **122**, 44107 (2005).
[82]X. Wu and B. R. Brooks, J. Chem. Phys. **129**, 154115 (2008).
[83]X. Wu and B. R. Brooks, J. Chem. Phys. **131**, 024107 (2009).