Combine umbrella sampling with integrated tempering method for efficient and accurate calculation of free energy changes of complex energy surface

Mingjun Yang,[1] Lijiang Yang,[2] Yiqin Gao,[2] and Hao Hu[1,a)]
[1]*Department of Chemistry, The University of Hong Kong, Pokfulam Road, Hong Kong, China*
[2]*College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China*

(Received 21 March 2014; accepted 25 June 2014; published online 24 July 2014)

Umbrella sampling is an efficient method for the calculation of free energy changes of a system along well-defined reaction coordinates. However, when there exist multiple parallel channels along the reaction coordinate or hidden barriers in directions perpendicular to the reaction coordinate, it is difficult for conventional umbrella sampling to reach convergent sampling within limited simulation time. Here, we propose an approach to combine umbrella sampling with the integrated tempering sampling method. The umbrella sampling method is applied to chemically more relevant degrees of freedom that possess significant barriers. The integrated tempering sampling method is used to facilitate the sampling of other degrees of freedom which may possess statistically non-negligible barriers. The combined method is applied to two model systems, butane and ACE-NME molecules, and shows significantly improved sampling efficiencies as compared to standalone conventional umbrella sampling or integrated tempering sampling approaches. Further analyses suggest that the enhanced performance of the new method come from the complemented advantages of umbrella sampling with a well-defined reaction coordinate and integrated tempering sampling in orthogonal space. Therefore, the combined approach could be useful in the simulation of biomolecular processes, which often involves sampling of complex rugged energy landscapes. © *2014 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4887340]

I. INTRODUCTION

Efficient calculation of accurate free energy change along a well-defined reaction coordinate (RC) is a central question in the physical chemistry of many important chemical and biomolecular processes. Among many developed simulation methods, [1–8] umbrella sampling (US) is one of the most widely employed.[9] There are two key ingredients in the US method: the identification of one or multiple reaction coordinates and the design of biasing potentials as a function of the RC.

Given the complexity of the energy landscape of multi-atom systems, the identification of the RC itself becomes a challenging issue.[10,11] In typical US simulations, RC is defined as a combination of simple geometric terms, such as bond lengths, bond angles, and dihedral angles on the basis of chemical intuition. It is assumed that the combination of the small number of selected geometric properties can characterize the major part of the reaction progress. In simple systems this type of definition is often sufficient for capturing the essence of the reaction.

Once the RC is determined, a discrete set of RC values is selected to cover the whole range of the reaction process. For each RC value, a biasing potential is applied to simulations to generate appropriate sampling of the system in the vicinity of the given value of RC. The biasing potential is expected to change the relative Boltzmann weight of different conformations along the RC. As a result, sufficient sampling can be obtained for those conformations that originally have small statistical weights and are difficult to sample in regular simulations. Once the biasing potential is decided, US simulations can be carried out embarrassingly parallel. The trajectories from all simulations can be combined together with posterior analysis methods such as the weighted histogram analysis method (WHAM)[12–15] or the maximum likelihood method.[16–18] The free energy change along the RC, or the potential of mean force (PMF) of the reaction process, can be reconstructed too.

After applying the biasing potential, the US simulation is technically identical to ordinary MD simulation. Therefore, it will suffer any technical difficulties that normal MD might experience. Putting the issue of correctness of RC aside, the success of the US simulations then depends on the (approximately) converged sampling in all the degrees of freedom orthogonal to the reaction coordinate in conformational space. Even though in many chemical reactions this requirement is likely to be satisfied, it would become a serious issue if there are multiple reaction channels along the reaction coordinate and the transitions between different channels are inadequately sampled in US simulations.

An example 2D potential energy landscape for this scenario is illustrated in Fig. 1. Obviously, the correctness of the 1D US simulations depends on the converged sampling in the vertical direction. As there are two parallel paths connecting stationary states A and B, the 1D PMF along the RC must reflect the proper statistical weights of the two paths, in particular, the correct sampling of the two transition states C_1

a)Author to whom correspondence should be addressed. Electronic mail: haohu@hku.hk

 © 2014 AIP Publishing LLC

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 147.8.204.164 On: Mon, 26 Jan 2015 07:00:03

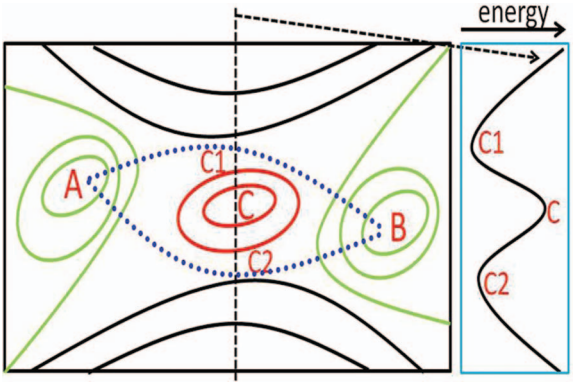

FIG. 1. An illustrative potential energy landscape for the parallel reaction paths. A and B are two energy minima. There are two reaction paths connecting A and B (dashed thick blue line). X-axis is regarded as the reaction coordinate for the transition between A and B. C is a high-energy region often regarded as a hidden barrier in the direction perpendicular to the reaction coordinate. C_1 and C_2 are the transition states of the two transition paths, respectively.

and C_2. If a biasing potential is applied to RC which limits the sampling of RC to regions close to C_1 and C_2, the corresponding US simulation must be able to sample both C_1 and C_2 in the same trajectory. This suggests there must be sufficient conformational transitions to cross the barrier of C relative to C_1 and C_2. In that sense the region C becomes a *hidden barrier* for conformational transition between C_1 and C_2 along the vertical direction. One must also be reminded that the importance of sufficient sampling of C only appears when localized biased sampling methods like US are employed. In ordinary sampling, if the length of the simulation is not a concern, direct sampling of C_1 and C_2, rather than C, is the most critical issue. The latter is determined by the free energy difference between C_1 and A or B, and between C_2 and A or B. Usually, it is assumed that the free energy difference between C_1 and A or B, or between C_2 and A or B, is much larger than that between C_1 or C_2 and C. That is, the target process is still dominated by the conformational transition along the RC.

Given this example, hidden barriers and parallel reaction paths appear to be two terms reflecting the same issue in conformational sampling. As shown, the existence of parallel paths along the RC already suggests a barrier in the perpendicular direction. Otherwise, the two paths would merge into one statistically significant path. For regions close to the transition states of the two paths, the correct simulation results would require converged sampling of two energy minimal regions as suggested by the energy curve of the intersecting plane along the vertical direction. If the height of the hidden barrier is low and, thus, can be sufficiently sampled within the length of US simulations, the 1D US results would be correct. As the length of individual US simulation is often significantly shorter than that of the normal MD, the evolution of the system along the reaction process will likely limit to small regions near one reaction path if the hidden barrier is high. Therefore, 1D US simulations could not provide correct results.

In principle, one could perform multi-dimensional US for the situation illustrated here, if one knows that there are hidden barriers in directions other than the reaction coordinate. There are, however, several technical difficulties for doing so. First, the existence of hidden barrier is not always known a prior. Second, the number of independent US simulations would grow exponentially with the dimensionality of US. It is often already difficult to define properly one RC, identifying additional RC would be even more challenging. Furthermore, multi-dimensional US sampling, e.g., for a potential energy landscape depicted in Fig. 1, would unavoidably waste many simulations in the corner regions which are statistically very insignificant.

In addition to umbrella sampling, a large group of enhanced sampling methods has been developed which often does not require an explicit definition of reaction coordinate. Great attention has been paid in recent years to the class of generalized ensemble methods. This includes the methods of replica-exchange (RE),[19,20] simulated tempering (ST),[21,22] multicanonical ensemble,[23–25] and the integrated tempering sampling (ITS).[26] Each of these approaches has shown distinct advantages and efficiency in tackling different problems. Of them, ITS is performed on an effective potential constructed by averaging over multiple Boltzmann distributions at different temperatures. Subsequent ITS simulations can be conducted only at temperature of interests, instead of at multiple different temperatures as in RE or ST methods. Thus, ITS has relatively lower demands for computational resources. Applications of ITS have shown that the sampling of high-energy region can be remarkably enhanced [27–31] and the efficiency is at least as good as other enhanced sampling method.[27,32] ITS needs not to use RCs, thus it can be applied to complex processes with rugged energy landscape such as protein folding.[30,31] This is one of the advantages of ITS and many other enhanced sampling methods. On the other hand, statistical weights for high-energy regions are uniformly increased regardless of their relevance to the target process, unless a small subset of degrees of freedom is selected to separate from the rest as in the selective ITS (SITS) method.[32] This would become a disadvantage if ITS was applied to processes with practically well-known RCs, because significant amount of the simulation efforts would be spent in sampling regions both statistically and chemically unimportant.

The unique features of ITS method, i.e., single-copy simulation and non-specifically enhanced sampling in phase space, make it a convenient method to combine with RC-guided multiple-window methods like US in the simulation of conformational transition and reaction processes. Here, we report our development and application of a simulation approach combining both ITS and US methods. We show that this combined new approach can achieve significantly improved sampling efficiency for complex landscape like the one depicted in Fig. 1. The theory will be provided in Sec. II, followed by computational details of the application to two model systems, namely, the butane and the ACE-NME molecules. Simulation results of the conformational dynamics of butane and ACE-NME molecules using different sampling strategies will be discussed and compared in Sec. III. The mechanisms of the sampling enhancement, possible extensions, and applications of the new method are discussed afterwards.

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 147.8.204.164 On: Mon, 26 Jan 2015 07:00:03

II. METHODS

In this section, we first outline the ITS method, which is followed by the theory of combined ITS and US method. Later, the computational details for the application of the methods to two molecular systems are provided.

A. Integrated tempering sampling

ITS was originally developed by Gao and co-workers.[26,27,33,34] In this method, a generalized non-Boltzmann ensemble was constructed by summing canonical distributions over a set of different temperatures. The generalized distribution can be written as

$$P(U(\mathbf{R})) = \sum_{k=1}^{N} n_k e^{-\beta_k U(\mathbf{R})}, \tag{1}$$

where $U(\mathbf{R})$ is the potential energy as a function of the coordinates $\mathbf{R}$ of the molecule and n_k is a weighting factor for the temperature T_k with

$$\beta_k = 1/k_B T_k. \tag{2}$$

The configurational partition function of the system at a given temperature is

$$Z_k = \int e^{-\beta_k U(\mathbf{R})} d\mathbf{R}. \tag{3}$$

The value of n_k can be determined by applying the condition of $n_1 Z_1 = n_2 Z_2 = \cdots = n_N Z_N$, which ideally will produce a uniform distribution in the temperature space within the range $[T_1, T_N]$. In practice, n_k will have to be determined with short trial simulations prior to the production run. An effective potential energy can then be defined for the production simulation temperature T_0 as

$$e^{-\beta_0 \tilde{U}(\mathbf{R})} = P(U(\mathbf{R})) = \sum_{k=1}^{N} n_k e^{-\beta_k U(\mathbf{R})} \tag{4}$$

or

$$\tilde{U}(\mathbf{R}) = \frac{-1}{\beta_0} \ln \sum_{k=1}^{N} n_k e^{-\beta_k U(\mathbf{R})}. \tag{5}$$

According to this equation, one practical advantage of the ITS scheme is that the value of $\tilde{U}(\mathbf{R})$ and $U(\mathbf{R})$ can be uniquely mapped once the coefficients $\{n_k\}$ are known. The corresponding force for propagating coordinates in MD simulations can be computed with

$$\tilde{\mathbf{F}}_i = -\frac{\partial \tilde{U}(\mathbf{R})}{\partial \mathbf{R}_i} = \frac{\sum_{k=1}^{N} n_k \beta_k e^{-\beta_k U(\mathbf{R})}}{\beta_0 \sum_{k=1}^{N} n_k e^{-\beta_k U(\mathbf{R})}} \mathbf{F}_i, \tag{6}$$

where

$$\mathbf{F}_i = -\frac{\partial U(\mathbf{R})}{\partial \mathbf{R}_i} \tag{7}$$

is the force of the original potential energy of the system.[26,33] As a result, the sampling by ITS simulation can be interpreted as an averaged sampling from multiple MD simulations each at a temperature T_k and with the potential energy $U(\mathbf{R})$, and with a weighting factor of

$$p_k = \frac{\int n_k e^{-\beta_k U(\mathbf{R})} d\mathbf{R}}{\int e^{-\beta_0 \tilde{U}(\mathbf{R})} d\mathbf{R}} = \frac{n_k Z_k}{\sum_{j=1}^{N} n_j Z_j}. \tag{8}$$

The sampling enhancement of ITS can be attributed to two factors. Firstly, the ITS MD simulation under effective potential $\tilde{U}(\mathbf{R})$ at β_0 includes contribution of samples from original potential $U(\mathbf{R})$ at higher temperature $T_k (T_k > T_0)$ as seen from Eq. (8). This factor ensures that the ITS simulation can overcome barriers more efficiently than that of normal MD simulations at T_0. Secondly, the low- and high-energy regions of effective potential $\tilde{U}(\mathbf{R})$ match with these of the original potential $U(\mathbf{R})$ since $\tilde{U}(\mathbf{R})$ is a monotonic function of $U(\mathbf{R})$ from Eq. (5). This implies the statistically favored low-energy regions of $U(\mathbf{R})$ still possess larger statistical weights in the samples of $\tilde{U}(\mathbf{R})$ in ITS simulations. This feature helps to maintain good balance of samples between low- and high-energy regions and, thus, can provide accurate unbiased canonical distribution $\rho(\mathbf{R})$ through reweighting from $\tilde{U}(\mathbf{R})$ to $U(\mathbf{R})$. Therefore, by taking the two factors together, ITS simulation is equivalent to performing a trajectory on an *attenuated* energy surface with lower barriers in comparison to that of the original potential, which can result in much more frequent barrier transition of the system.

We like to comment here on some differences between the simulated tempering (ST) and ITS methods. Both methods are constructed on the framework of multiple canonical ensembles. Practically, all canonical ensembles will be sampled in ST simulations, even though it appears there is only one trajectory. As the target system periodically changes its temperature/Hamiltonian, the trajectory, in fact, is pieced together from many fragments each corresponding to a short trajectory in a canonical ensemble of specific temperature or Hamiltonian. In other words, each sample obtained in ST simulations belongs to one of the real canonical ensembles. The need to switch between different temperatures or Hamiltonians, in fact, causes a difficult practical issue on the frequency of change. On the other hand, with the construction of an effective potential energy (Eq. (5)), the sample in the ITS simulation cannot be directly assigned to any real canonical ensemble. Instead, each ITS sample contains information of all canonical ensembles covered in the simulations, mixed together in a mean-field fashion. Therefore, the ITS method does not suffer the well-known problem of switching frequency in many other multi-canonical ensemble methods. We note that it would be premature to simply presume one method is superior to other methods, as in molecular simulations the choice of method depends on the detailed properties of the system. This is also one of the important viewpoints the current work is trying to stress.

B. Combining ITS with US

In typical umbrella sampling simulations,[14,15,35,36] the target molecular event is characterized by a preselected reaction coordinate which is thought to best characterize the

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 147.8.204.164 On: Mon, 26 Jan 2015 07:00:03

reaction progress. A biasing potential, often in the form of a quadratic function,

$$V_i^b(\Omega(\mathbf{R})) = \frac{1}{2}K_i(\Omega(\mathbf{R}) - \omega_i)^2, \tag{9}$$

is applied to each different parallel simulations. Here, K_i and ω_i are the preset force constant and center of the distribution, $\Omega(\mathbf{R})$ is the reaction coordinate, also often referred to as the order parameter or collective variable/degree of freedoms, defined as a function of atomic positions. The potential energy in each MD simulation of US, often termed as a "sampling window," is

$$U_i(\mathbf{R}) = U(\mathbf{R}) + V_i^b(\mathbf{R}). \tag{10}$$

By varying ω_i and the correspondingly K_i, one can force the system to sample through the whole conformational space of the reaction coordinate $\Omega(\mathbf{R})$, even for regions with high free energy and hardly being observed in normal MD with unbiased energy function.

As proposed in the Introduction, enhanced sampling technique such as ITS can be combined with umbrella sampling to improve the sampling in the degrees of freedom orthogonal to RC. When ITS is used, the biased total potential energy for the ith window should be $\tilde{U}_i^b(\mathbf{R})$,

$$\tilde{U}_i^b(\mathbf{R}) = \tilde{U}(\mathbf{R}) + V_i^b(\Omega(\mathbf{R})), \tag{11}$$

where $\tilde{U}(\mathbf{R})$ is the effective ITS potential defined in Eq. (5). Here, we assume that the same effective ITS potential is applied to different US sampling windows, i.e., the same set of weighting factors is used for the effective ITS potential in all umbrella windows. More generally, one may define the biased potential energy of each US window as

$$\tilde{U}_i^b(\mathbf{R}) = \tilde{U}_i(\mathbf{R}) + V_i^b(\Omega(\mathbf{R})), \tag{12}$$

where $\tilde{U}_i(\mathbf{R})$ now is the effective ITS potential determined specifically for the sampling window i. In contrast to $\tilde{U}(\mathbf{R})$ in Eq. (11), different $\{n_k\}$ values are used for each $\tilde{U}_i(\mathbf{R})$ of a given umbrella window. Thus, Eq. (12) is a general case of Eq. (11) and is used in the remaining discussion of Sec. II.

By employing the biased potential defined in Eq. (12), the unbiased canonical distribution $\rho_i(\mathbf{R})$ of the original potential energy can be recovered from the biased distribution $\rho_i^b(\mathbf{R})$ through

$$\rho_i(\mathbf{R}) = \rho_i^b(\mathbf{R})e^{\beta_0(\tilde{U}_i(\mathbf{R})-U(\mathbf{R})+V_i^b(\Omega(\mathbf{R})))}e^{-\beta_0 f_i}, \tag{13}$$

where the term f_i accounting for the free energy correction to the sampling window can be computed with

$$\begin{aligned} f_i &= -\frac{1}{\beta_0}\ln\frac{Z_i^b}{Z_i} \\ &= -\frac{1}{\beta_0}\ln\int e^{-\beta_0\left(\tilde{U}_i(\mathbf{R})-U(\mathbf{R})+V_i^b(\Omega(\mathbf{R}))\right)}\rho_i(\mathbf{R})d\mathbf{R}. \end{aligned} \tag{14}$$

Here, Z_i^b and Z_i are the canonical partition function for the biased and unbiased system, respectively.

WHAM can be used to analyze the simulation data if one treats the combined term $\tilde{U}_i(\mathbf{R}) - U(\mathbf{R}) + V_i^b(\Omega(\mathbf{R}))$ as the biasing potential. Following the original derivation,[15] the values of $\{f_i\}$ can be determined iteratively through

$$\rho(\mathbf{R}) = \sum_{i=1}^{N_W}\frac{m_i\rho_i^b(\mathbf{R})}{\sum_{j=1}^{N_W}m_j e^{-\beta_0\left(\left(V_j^b(\mathbf{R})-f_j\right)+\left(\tilde{U}_j(\mathbf{R})-U(\mathbf{R})\right)\right)}} \tag{15}$$

and

$$\begin{aligned} e^{-\beta_0 f_k} &= \int e^{-\beta_0\left(\tilde{U}_k(\mathbf{R})-U(\mathbf{R})+V_k^b(\mathbf{R})\right)}\rho(\mathbf{R})d\mathbf{R} \\ &= \sum_{i=1}^{N_W}\sum_{l=1}^{m_i}\frac{e^{-\beta_0\left(\tilde{U}_k\left(\mathbf{R}_{i,l}\right)-U\left(\mathbf{R}_{i,l}\right)+V_k^b\left(\mathbf{R}_{i,l}\right)\right)}}{\sum_{j=1}^{N_W}m_j e^{-\beta_0\left(\left(V_j^b\left(\mathbf{R}_{i,l}\right)-f_j\right)+\left(\tilde{U}_j\left(\mathbf{R}_{i,l}\right)-U\left(\mathbf{R}_{i,l}\right)\right)\right)}}. \end{aligned} \tag{16}$$

Here, m_i is the number of samples recorded in the ith simulation window.

After converged values of $\{f_i\}$ have been obtained, the unbiased probability distribution $\rho(\mathbf{R})$ can be computed from Eq. (15). Then the distribution of other quantities can be derived from $\rho(\mathbf{R})$, e.g., the reweighted probability distribution along a collective variable $\Theta(\mathbf{R})$:

$$\rho(\theta) = \int \rho(\mathbf{R})\,\delta(\Theta(\mathbf{R}) - \theta)\,d\mathbf{R}, \tag{17}$$

which can be conveniently evaluated from the recorded snapshots. The associated PMF along $\Theta(\mathbf{R})$ can be computed as

$$A(\theta) = -\frac{1}{\beta_0}\ln\rho(\theta). \tag{18}$$

C. Computational details

The ITS method and the combined ITS-US method were implemented in an in-house program QM^{4}D.[37] The methods were applied to the simulation of two systems, namely, the butane molecule and the ACE-NME molecule, in gas phase (Fig. 2). In all simulations, MD time step was 1 fs. The temperature of the simulation system was maintained at 300 K by Langevin dynamics.[38] $\{n_k\}$ values in ITS and ITS-US simulations were determined following the optimization procedure suggested previously.[33]

The butane molecule was treated by the classical CHARMM force field.[39] This molecule was used to examine the correctness of our method since standard MD simulation can provide accurate sampling of its conformational states, as such the results can serve as a reference for other simulations. To this end, several different sampling strategies, including standard MD, ITS, 1D US, and the combined ITS-US methods, were carried out to compute the potential of mean force along the rotation of the C1-C2-C3-C4 dihedral angle (Fig. 2(a)). The weighting factors $\{n_k\}$ were optimized under the original potential and then used in both ITS and ITS-US production runs.

The ACE-NME molecule in gas phase was used to investigate the isomerization of the peptide bond (Fig. 2(b)). This molecule was treated by the SCCDFTB method which is necessary for providing a quantum mechanical description

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 147.8.204.164 On: Mon, 26 Jan 2015 07:00:03

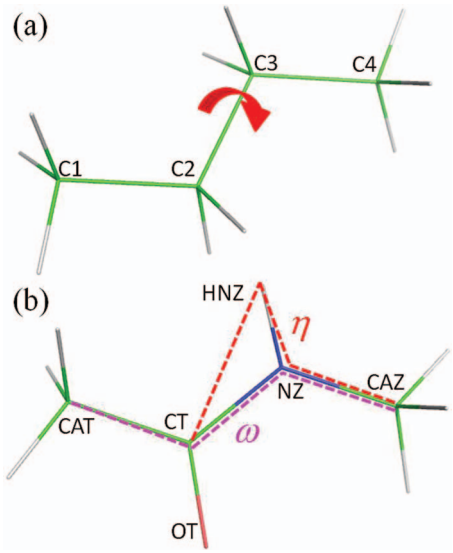

FIG. 2. Structural model of (a) butane and (b) ACE-NME. The isomerization of ACE-NME is characterized by two reaction coordinates, the backbone dihedral ω: CAT-CT-NZ-CAZ and the improper dihedral η: CT-HNZ-NZ-CAZ.

for the peptide bond isomerization.[40,41] Simulations with different sampling strategies were carried out, including 1D US along ω, 1D ITS-US along ω, and 2D US along ω and η. For 2D US, 600 ps simulations were performed for each window. Simulation time of each 1D US window is 6000 ps. Simulation time of each 1D ITS-US window is 1500 ps. The force constants and window spacing were chosen to maintain sufficient overlap between neighboring umbrella windows, with smaller intervals and larger force constants around the barrier regions and larger intervals and smaller force constants near the free energy valleys. For the dihedral ω, 54 unevenly distributed windows within $(-180^\circ, 180^\circ]$ were selected with force constants of 70–160 kcal/mol/rad^2 used (Table SI).[42] For the improper dihedral η, 19 unevenly distributed windows in the range of $[99^\circ, 180^\circ]$ and $[-180^\circ, -99^\circ]$ were employed with a force constant of 90 kcal/mol/rad^2. In order to compare the sampling efficiency, the same set of restraining forces and window locations along ω was employed in 2D US, 1D US, and 1D ITS-US simulations. In ITS-US simulations, 60 intermediate temperatures in the range of 273 K to 700 K were used.

To examine the influence of the weighting factors on the sampling efficiency of ITS-US, we carried out 5 different set of ITS-US production runs using, respectively, $\{n_k\}$ values optimized from (1) the original potential for all windows, (2) the biased umbrella potential at 0°, (3) the biased umbrella potential at 45°, (4) the biased umbrella potential at 90°, and (5) the biased umbrella potential for each window. Note that in the last case each individual ITS-US simulation used a different set of $\{n_k\}$.

Also to compare the sampling efficiency of the current ITS-US method with other widely used methods, the replica exchange MD (REMD) simulations were performed for the ACE-NME system with the SCCDFTB method. Two different variants of REMD simulations were performed, one uses replicas in the temperature space (T-REMD) and the other in the space of biasing potential (RE-US). For T-REMD, a total of 6 replicas with an exponentially distributed temperature from 300 K to 700 K were adopted, resulting in an average acceptance ratio of ~50%. For the RE-US simulations, the same set of biasing potentials as in 1D US was used and the temperatures were maintained at 300 K for all replicas. In contrast to T-REMD, only the biasing potentials were exchanged in the RE-US simulations according to the Metropolis criterion. The average acceptance ratio for RE-US is around 30%. In both simulations, exchange was attempted between the neighboring replicas every 1000 MD steps and each replica was simulated with 12 ns and 6 ns for T-REMD and RE-US, respectively.

III. RESULTS

A. Internal rotation of butane

For the butane molecule in gas phase, the conformational dynamics investigated here is the rotation of the dihedral C1-C2-C3-C4. As shown, the barrier height is ~5.0 kcal/mol. A barrier of this height can be well sampled in MD simulations at room temperature with 8 × 32 ns. All methods tested gave results in excellent agreement with each other, suggesting the correctness of the ITS-US method (Fig. S1).

B. Peptide bond *cis-trans* isomerization in ACE-NME

The ACE-NME molecule contains a peptide bond which is the fundamental linkage between amino acids in proteins. The peptide bond isomerization is often assumed to proceed mainly through the rotation of the backbone dihedral ω. However, the configuration of the backbone nitrogen atom may become non-planar during the isomerization process. Another quantity is, thus, required to identify different states of the nitrogen configuration. Therefore, *cis-trans* isomerization of the ACE-NME could serve a model system to provide useful information for the peptide bond isomerization in proteins and polypeptides. One straightforward reaction coordinate for the isomerization is the dihedral ω(CAT-CT-NZ-CAZ), while a second *improper* dihedral η(CT-HNZ-NZ-CAZ) defines the chiral configuration of the nitrogen atom which is also important in the isomerization process (Fig. 2(b)).

In the ITS-US simulations, the production run requires a set of weighting factors $\{n_k\}$ to be determined in prior. The weighting factors can improve the sampling at high-energy regions to accelerate barrier transitions in the remaining conformational space. In ITS-US, the weighting factors can be optimized in the presence or absence of the biasing potentials. To examine if the results are sensitive to the weighting factors, several independent ITS-US production runs were carried out using $\{n_k\}$ values optimized under different biasing potentials, i.e., localized in different phases regions. The 1D PMF along ω can converge to each other within 150 ps simulation of each window and the transition regions are also sampled comparably in the 2D PMF maps (Fig. S2). This result suggests that $\{n_k\}$ values optimized from all these schemes can attenuate the original energy landscape in similar fashion. Despite the difference in the phase spaces sampled for

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 147.8.204.164 On: Mon, 26 Jan 2015 07:00:03

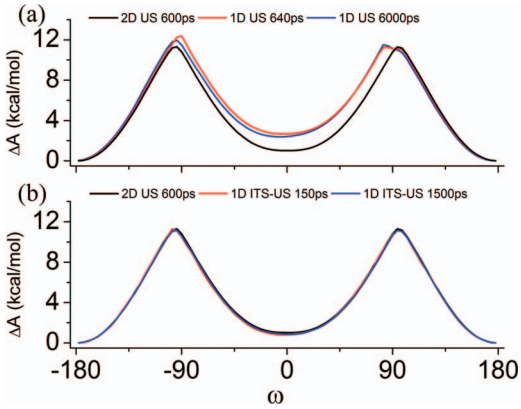

FIG. 3. 1D PMF profile along ω in ACE-NME. (a) Comparison of results of conventional 1D US and 2D US; (b) comparison of results of 1D ITS-US and 2D US.

optimizing the values, different sets of weighting factors apparently improve the sampling in approximately equal efficiency. Therefore, unless stated otherwise, we will only discuss the ITS-US simulations using $\{n_k\}$ values derived from the original potential in Sec. II C. We first compare the 1D PMF along ω from different sampling strategies. The 1D PMF results of 2D US simulations were generated from Eqs. (17) and (18) by integrating out the η degree of freedom in the joint distribution $\rho(\omega, \eta)$. Because of the extensive sampling in 2D US simulations, we regard this result to be valid and use it as reference for comparison. Compared to the 2D US simulations, 1D US simulations show diverged results of 1D PMF for simulations of different lengths (Fig. 3(a)). Even for 1D US simulations with 6000 ps per window, the results did not converge to the reference results of 2D US simulations. On the contrary, the results of 1D ITS-US show excellent agreement to the 2D US results, even for simulations with only 150 ps per window (Fig. 3(b)).

To determine the reason why 1D US simulations failed to generate correct PMF for the assumed isomerization process along ω, 2D PMF was generated for each different set of simulations. It is evident that depending on the value of improper dihedral η (or the equivalent η'), isomerization along ω can proceed with two parallel paths through the transition states around $\pm 130^\circ$ ($\eta' \approx \pm 50^\circ$) in either clockwise ($\omega$: 0° to 180°) or anti-clockwise (ω: 0° to -180°) direction (Fig. 4(a)). For the correct calculation of 1D PMF along ω, the degree of freedom of η must be sufficiently sampled to reflect the correct statistical weights for conformations along the two paths. Once there are non-negligible barriers in the motions along η, conventional 1D US sampling would not be able to efficiently cross barrier to provide converged sampling. This problem was clearly demonstrated by the 2D PMF generated from the trajectories of 1D US along ω (Fig. 4(b)). Due to the barrier in the direction of η, the conformations sampled in different windows in 1D US do not overlap in the transition regions along the direction of η around $\pm 130^\circ$ ($\eta' \approx \pm 50^\circ$). In fact, neither of the transition state of the two paths was sampled in 1D US simulations. The poor sampling in these regions is the reason why both the height and position of the transition states in 1D US samplings are incorrect (Fig. 3(a)).

In contrast, 1D ITS-US, even for data extracted from simulation fragments of only 150 ps, provided significantly improved samplings (Figs. 4(c) and S3).[42] A significant amount of conformations in the transition region have been sampled in 150 ps of 1D ITS-US simulations. For the parallel isomerization paths, one of the two transition states with lower free energy barrier was sampled properly, thus providing correct results for 1D PMF of ω. When the simulation time was increased to 1500 ps per window, the 2D PMF well reproduces

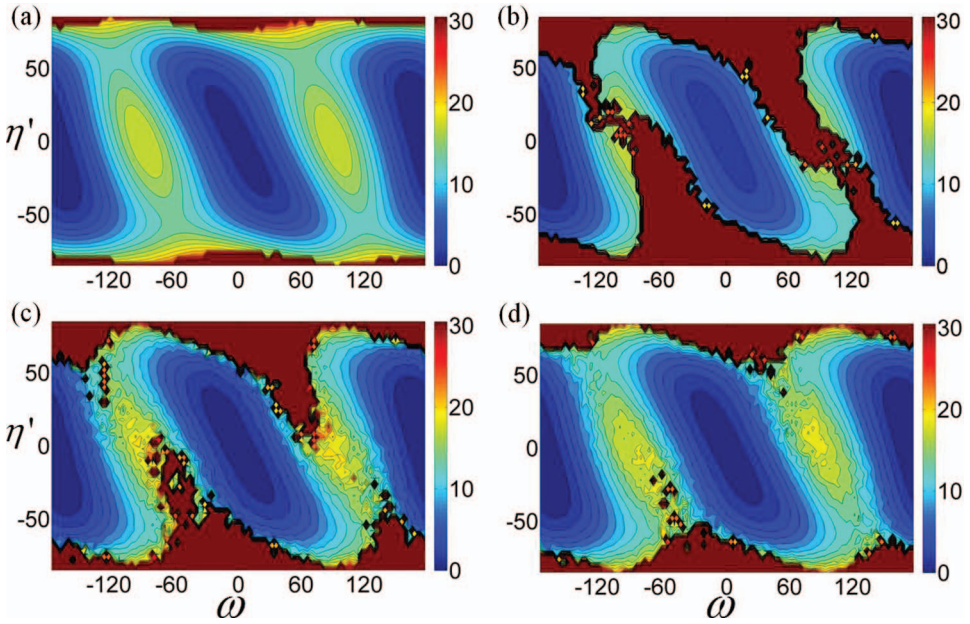

FIG. 4. 2D PMF maps of the ACE-NME molecule. (a) Results reconstructed directly from 2D US simulations of 600 ps per window; (b) results projected from 1D US simulations of 6000 ps per window; (c) results projected from 1D ITS-US simulations of 150 ps per sampling window; (d) results projected from 1D ITS-US simulations of 1500 ps per sampling window. Note for clarity, the Y-axis is defined as η'. $\eta' = \eta - \pi$ when $\eta > 0$ and $\eta' = \eta + \pi$ when $\eta < 0$. The unit for both axes is degree.

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 147.8.204.164 On: Mon, 26 Jan 2015 07:00:03

the results of 2D US illustrated by the clear appearance of two parallel isomerization paths (Fig. 4(d)).

IV. DISCUSSION

In this study, we show a combination of ITS and US can efficiently generate satisfactory sampling for system with complex energy landscape and possibly multiple parallel transformation paths. The success of this ITS-US approach is due to the proper combination of the advantages of both methods. On one hand, the degree of freedom with significant energy barrier is more efficiently sampled by RC guided US approach. In the ACE-NME system, the height of the barrier along ω is about 12 kcal/mol. In the current case, without RC, general enhanced sampling methods might still be able to sample the barrier but with very low efficiency (Figs. S4 and S5).[42] The reason for their low efficiency is that the target process involves a much more localized conformational motion. It would be the best scenario if enhanced sampling can be directly applied to the specific motions instead of spreading among all degrees of freedom.

On the other hand, with an increase in the number of degrees of freedom, the energy landscape of large molecule becomes more and more complicated. Numerous minima, maxima, and transition saddle points emerge even for those degrees of freedom regarded as chemically less relevant. This rugged energy landscape is the ultimate reason for the hierarchy of conformational motions in biomolecules. Even under the assumption that there is a single reduced coordinate in which the barrier dominates the target processes, sampling the remaining conformation degrees of freedom with lower but non-negligible barriers remains important. In limited examples high-dimensional US could be employed, but with significantly increased technical difficulty and computational costs. General enhanced sampling techniques, such as ITS, become an effective approach to sample these degrees of freedoms.

A simple comparison with the simulation time can be used to provide a rough picture for the level of improvement the ITS-US approach made with respect to normal US. For the ACE-NME system tested using the same set of simulation parameters, 150 ps simulations of each window of 1D ITS-US generate PMF in good agreement with reference results, while plain 1D US simulation up to 6000 ps per window cannot provide results of the same quality. The ratio of computational cost for 1D ITS-US vs US is ~1:40. On the other hand, even without optimizing the number of US windows, the total simulation length of 1D ITS-US is $0.6 \times 54 = 32.4$ ns, which provided results comparable to or even better than 8 parallel ITS simulations with a total length of $8 \times 96 = 768$ ns (Figs. S4 and S5).[42] Both ratios demonstrate again the much-improved efficiency in the ITS-US approach.

To show the effectiveness of the ITS-US approach, we also plotted the distribution of potential energy in different simulations. As shown, the fluctuation of potential energy in conventional US simulations is small, thus crossing a barrier of even moderate height is very difficult (Fig. 5). In contrast, ITS samples the potential energy in a range significantly broader than normal MD, thus providing sufficient sampling for crossing barriers of low and medium heights.

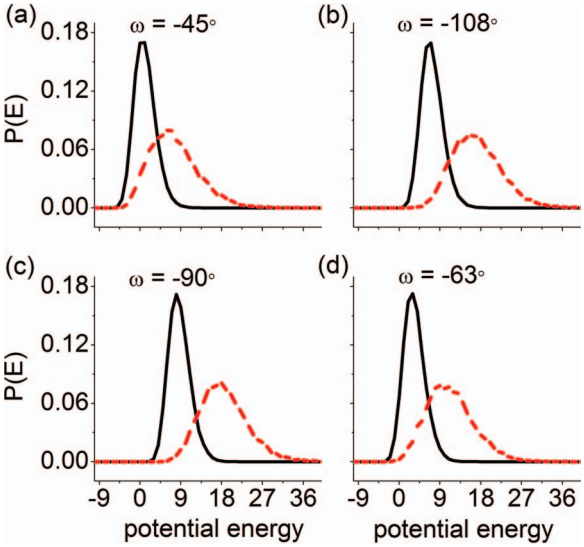

FIG. 5. Distribution of potential energy in simulations with different restraints. Black solid line: 1D US; red dashed: 1D ITS-US. The energy was shifted by 8453.78 kcal/mol in all panels to plot the distribution in the vicinity of zero.

We note that the current approach still possesses great potential to further improve the efficiency. The first improvement could be made to use different set of $\{n_k\}$ parameters for the simulation of different US windows. Second, as shown in the theory of ITS, the relative conformational weight along RC will unavoidably be adjusted by ITS. Therefore, the restraining potential and the corresponding number of US windows in US simulations could be re-optimized in this combined ITS-US approaches. Therefore, a smaller number of US windows might be utilized in ITS-US simulations.

Even though the current study demonstrates the high efficiency of the combined ITS-US approach, this new method to some extent still depends on the identification of an appropriate RC. In systems where a RC is difficult to identify, e.g., protein folding, simple ITS might be practically more advantageous. Moreover, the results of ITS simulations can be analyzed to identify proper order parameters of the target process, which can be used subsequently in the combined enhanced sampling simulations to improve the accuracy of the results. It is noted that the optimization of weighting factors would become challenging with the increase of system size. The contribution of the improved sampling efficiency will also deteriorate with the increase of system size. In this sense, the application of ITS-US will face the common "curse of dimensionality" shared by many, if not all, generalized ensemble methods, e.g., REMD and ST.[51–56] One way to tackle this problem is to reduce the number of atoms included into the ITS potential, e.g., the development of selective ITS to mainly enhance the sampling of solute and maintain the solvent environment nearly unperturbed.[32] This will be explored in our future researches.

Several extensions have been proposed to the US methods to improve the efficiency of free energy calculations in the past several decades. Instead of performing multiple-window simulations with local biasing potential at each US window, the adaptive US method, originally proposed by

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 147.8.204.164 On: Mon, 26 Jan 2015 07:00:03

Mezei[43] and Hooft *et al.*[44] and further developed by Karplus and co-workers,[45,46] updates the biasing potential iteratively over the whole range of RC in a single simulation. The samples generated in all iterations can be combined to reconstruct the PMF profiles in posterior analysis. Similarly, some other extensions of US can also carry out the simulation in one trajectory, e.g., self-healing US and local evaluation US. In self-healing US, a biasing potential is progressively constructed using the time-dependent probability density distribution and slow dynamics can be accelerated along the RC.[47] The local evaluation US combines the searching power of local evaluation to build up an optimized biasing potential at the first stage and the sampling ability of US to explore relevant conformational space at the second stage.[48] In this method, the biasing potential can be constructed by fragments and then used for larger molecules.[49] Metadynamics and US have also been combined and executed in a sequential fashion to study the ion permeation in an ion channel.[50] Nevertheless, the efficiency of most methods strongly depends on a predefined RC, while the sampling of the remaining conformational space was hardly improved. This would generate poorly converged results when the RC cannot be accurately defined.

Enhanced sampling methods such as simulated tempering (ST) and replica-exchange (RE) have also been combined with US methods. [51–56] In the ST-US and RE-US methods, the transition or exchange attempts are made in the temperature and/or parameter space of the umbrella potentials. If implemented in a straightforward way where transitions or exchanges are made simultaneously in both the temperature and the parameter space, the total number of simulated canonical components will equal to the product between the number of temperatures and umbrella windows. The larger number of replicas will consume more computational resources and take longer simulation time for a roundtrip from the first replica to the last. If transitions or exchanges are only allowed in the US parameter spaces, the sampling along the degrees of freedom orthogonal to the RC is hardly improved for RE-US. To give a rough comparison of the sampling efficiency for the complex energy landscape we studied here, the T-REMD and RE-US simulations were also carried out for the ACE-NME system. Even though T-REMD produced a correct relative free energy for different local minima (around 0° and ±180°), the high-energy transition regions are hardly sampled at all. The 1D PMF (Fig. 6(a)) of T-REMD and RE-US showed clear deviation from converged results from 2D US simulations. Inspection of the 2D PMF (Figs. 6(b) and 6(c)) suggested that the barrier regions were poorly sampled for both methods. For RE-US, the sampling of transition regions is only slightly improved compared to the plain 1D US simulation with the same simulation length for each umbrella window (Fig. 6(c)). We believe the results are due to the relatively fast, focused, and significant energy change of the target system when making conformational change along the reaction coordinate, which certainly is not an ideal situation for sampling methods that do not employ a reaction coordinate. Therefore, the ITS-US method performs with a distinct advantage for sampling the multiple parallel paths if a major reaction coordinates can be easily identified.

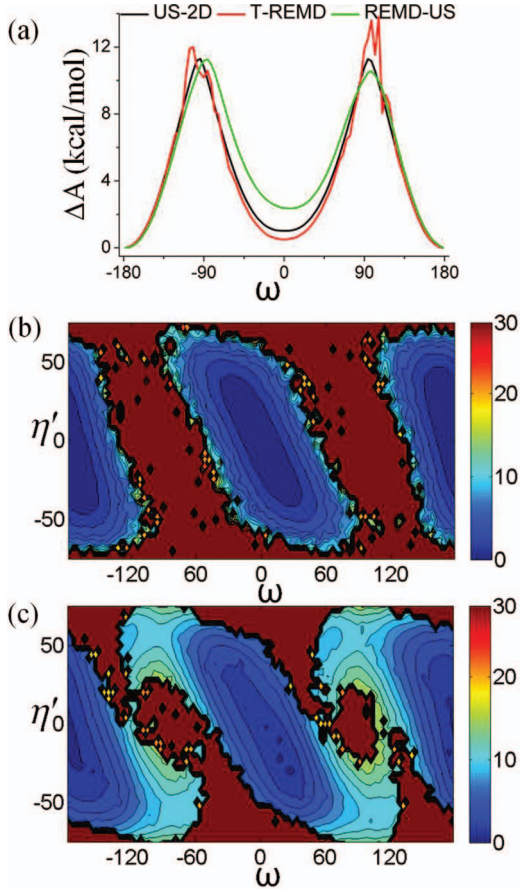

FIG. 6. PMF profiles from T-REMD and RE-US simulations. (a) 1D PMF along the dihedral ω. (b) 2D PMF along ω and η' from T-REMD. (c) 2D PMF along ω and η' from RE-US. The simulation length for T-REMD is 12 ns per replica and that for RE-US is 6 ns per replica. WHAM was used to compute the PMF profiles from all replicas of T-REMD and RE-US simulations, respectively. Note for clarity, the Y-axis is defined as η' in 2D PMF maps. $\eta' = \eta - \pi$ when $\eta > 0$ and $\eta' = \eta + \pi$ when $\eta < 0$.

V. CONCLUSIONS

In conclusion, the ITS-US method developed here can produce remarkably improved sampling efficiency for complex energy surface such as the multiple parallel reaction paths we tested in this study. The combined approach provides great application potential for a broad range of interesting questions, such as the reaction in solution, enzymatic catalysis, and conformational transition in molecular recognition, etc.

ACKNOWLEDGMENTS

We thank the Research Grants Council of Hong Kong, the University Development Fund on Fast Algorithms, Strategic Research Themes in Clean Energy and in Computation and Information, and Seed Funding for Basic Research at the University of Hong Kong for providing financial supports, the high-performance computing facility of the computer center at HKU for providing computing resources.

[1] Y. Hu, W. Hong, Y. Shi, and H. Liu, J. Chem. Theory Comput. **8**, 3777 (2012).
[2] L. Zheng, M. Chen, and W. Yang, Proc. Natl. Acad. Sci. U.S.A. **105**, 20227 (2008).

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 147.8.204.164 On: Mon, 26 Jan 2015 07:00:03

[3]H. Hu and W. Yang, Annu. Rev. Phys. Chem. **59**, 573 (2008).
[4]E. Darve, D. Rodriguez-Gomez, and A. Pohorille, J. Chem. Phys. **128**, 144120 (2008).
[5]*Free Energy Calculations: Theory and Applications in Chemistry and Biology*, edited by C. Chipot and A. Pohorille (Springer, 2007), p. 1.
[6]C. Micheletti, A. Laio, and M. Parrinello, Phys. Rev. Lett. **92**, 170601 (2004).
[7]H. Hu, R. H. Yun, and J. Hermans, Mol. Simul. **28**, 67 (2002).
[8]J. Hermans, J. Phys. Chem. **95**, 9029 (1991).
[9]J. Kaestner, Wiley Interdiscip. Rev. Comput. Mol. Sci. **1**, 932 (2011).
[10]A. Singharoy, S. Cheluvaraja, and P. Ortoleva, J. Chem. Phys. **134**, 044104 (2011).
[11]P. G. Bolhuis, C. Dellago, and D. Chandler, Proc. Natl. Acad. Sci. U.S.A. **97**, 5877 (2000).
[12]A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **63**, 1195 (1989).
[13]S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, J. Comput. Chem. **13**, 1011 (1992).
[14]B. Roux, Comput. Phys. Comm. **91**, 275 (1995).
[15]M. Souaille and B. Roux, Comput. Phys. Commun. **135**, 40 (2001).
[16]T. S. Lee, B. K. Radak, A. Pabis, and D. M. York, J. Chem. Theory Comput. **9**, 153 (2013).
[17]Z. Q. Tan, J. Am. Stat. Assoc. **99**, 1027 (2004).
[18]C. Bartels, Chem. Phys. Lett. **331**, 446 (2000).
[19]U. H. E. Hansmann, Chem. Phys. Lett. **281**, 140 (1997).
[20]Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **314**, 141 (1999).
[21]A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov, and P. N. Vorontsovvelyaminov, J. Chem. Phys. **96**, 1776 (1992).
[22]E. Marinari and G. Parisi, Europhys. Lett. **19**, 451 (1992).
[23]B. A. Berg and T. Neuhaus, Phys. Lett. B **267**, 249 (1991).
[24]B. A. Berg and T. Neuhaus, Phys. Rev. Lett. **68**, 9 (1992).
[25]N. Nakajima, H. Nakamura, and A. Kidera, J. Phys. Chem. B **101**, 817 (1997).
[26]Y. Q. Gao, J. Chem. Phys. **128**, 064105 (2008).
[27]L. Yang, Q. Shao, and Y. Q. Gao, J. Chem. Phys. **130**, 124111 (2009).
[28]Q. Shao and Y. Q. Gao, J. Chem. Phys. **135**, 135102 (2011).
[29]Q. Shao, L. Yang, and Y. Q. Gao, J. Chem. Phys. **135**, 235104 (2011).
[30]Q. Shao, J. Shi, and W. Zhu, J. Chem. Phys. **137**, 125103 (2012).
[31]Q. Shao, W. Zhu, and Y. Q. Gao, J. Phys. Chem. B **116**, 13848 (2012).
[32]L. Yang and Y. Q. Gao, J. Chem. Phys. **131**, 214109 (2009).
[33]Y. Q. Gao, J. Chem. Phys. **128**, 134111 (2008).
[34]Y. Q. Gao, L. Yang, Y. Fan, and Q. Shao, Int. Rev. Phys. Chem. **27**, 201 (2008).
[35]G. M. Torrie and J. P. Valleau, Chem. Phys. Lett. **28**, 578 (1974).
[36]G. M. Torrie and J. P. Valleau, J. Comput. Phys. **23**, 187 (1977).
[37]X. Q. Hu, H. Hu, and W. T. Yang, QM4D: An integrated and versatile quantum mechanical/molecular mechanical simulation package (2013).
[38]W. F. Van Gunsteren and H. J. C. Berendsen, Mol. Simul. **1**, 173 (1988).
[39]A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, 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, 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).
[40]M. Elstner, T. Frauenheim, E. Kaxiras, G. Seifert, and S. Suhai, Phys. Status Solidi B **217**, 357 (2000).
[41]M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, and G. Seifert, Phys. Rev. B **58**, 7260 (1998).
[42]See supplementary material at http://dx.doi.org/10.1063/1.4887340 for the force constants and window locations used in US simulations, the PMF along ω computed by 150 ps segment of standalone umbrella sampling and ITS simulations.
[43]M. Mezei, J. Comput. Phys. **68**, 237 (1987).
[44]R. W. W. Hooft, B. P. Vaneijck, and J. Kroon, J. Chem. Phys. **97**, 6690 (1992).
[45]C. Bartels and M. Karplus, J. Phys. Chem. B **102**, 865 (1998).
[46]C. Bartels and M. Karplus, J. Comput. Chem. **18**, 1450 (1997).
[47]S. Marsili, A. Barducci, R. Chelli, P. Procacci, and V. Schettino, J. Phys. Chem. B **110**, 14011 (2006).
[48]H. S. Hansen and P. H. Huenenberger, J. Comput. Chem. **31**, 1 (2010).
[49]H. S. Hansen, X. Daura, and P. H. Huenenberger, J. Chem. Theory Comput. **6**, 2598 (2010).
[50]Y. Zhang and G. A. Voth, J. Chem. Theory Comput. **7**, 2277 (2011).
[51]Y. Mori and Y. Okamoto, Phys. Rev. E **87**, 023301 (2013).
[52]S. Park, T. Kim, and W. Im, Phys. Rev. Lett. **108**, 108102 (2012).
[53]A. Mitsutake and Y. Okamoto, Phys. Rev. E **79**, 047701 (2009).
[54]A. Mitsutake and Y. Okamoto, J. Chem. Phys. **130**, 214105 (2009).
[55]J. Curuksu and M. Zacharias, J. Chem. Phys. **130**, 104110 (2009).
[56]Y. Sugita, A. Kitao, and Y. Okamoto, J. Chem. Phys. **113**, 6042 (2000).

This article is copyrighted as indicated in the article. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 147.8.204.164 On: Mon, 26 Jan 2015 07:00:03