JCTC Journal of Chemical Theory and Computation

Article

Cite This: J. Chem. Theory Comput. 2018, 14, 14−29

pubs.acs.org/JCTC

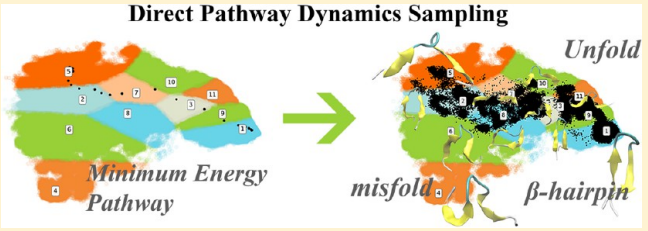
Dynamics Sampling in Transition Pathway Space

Hongyu Zhou and Peng Tao*

Department of Chemistry, Center for Drug Discovery, Design, and Delivery (CD4), Center for Scientific Computation, Southern Methodist University, Dallas, Texas 75275, United States of America

Supporting Information

ABSTRACT: The minimum energy pathway contains important information describing the transition between two states on a potential energy surface (PES). Chain-of-states methods were developed to efficiently calculate minimum energy pathways connecting two stable states. In the chain-of-states framework, a series of structures are generated and optimized to represent the minimum energy pathway connecting two states. However, multiple pathways may exist connecting two existing states and should be identified to obtain a full view of the transitions. Therefore, we developed an enhanced sampling method, named as the direct pathway dynamics sampling (DPDS) method, to facilitate exploration of a PES for multiple pathways connecting two stable states as well as addition minima and their associated transition pathways. In the DPDS method, molecular dynamics simulations are carried out on the targeting PES within a chain-of-states framework to directly sample the transition pathway space. The simulations of DPDS could be regulated by two parameters controlling distance among states along the pathway and smoothness of the pathway. One advantage of the chain-of-states framework is that no specific reaction coordinates are necessary to generate the reaction pathway, because such information is implicitly represented by the structures along the pathway. The chain-of-states setup in a DPDS method greatly enhances the sufficient sampling in high-energy space between two end states, such as transition states. By removing the constraint on the end states of the pathway, DPDS will also sample pathways connecting minima on a PES in addition to the end points of the starting pathway. This feature makes DPDS an ideal method to directly explore transition pathway space. Three examples demonstrate the efficiency of DPDS methods in sampling the high-energy area important for reactions on the PES.



1. INTRODUCTION

Sampling transition pathways that connect two states for a given system are an important and active area of methodology development in computational chemistry. Chain-of-states methods were developed to obtain a minimum energy pathway connecting two predefined states. Many chain-of-states methods were developed to obtain minimum energy pathways.[1] Elber and Karplus first used a line integral representation of a discretized path for optimization.[2] Following the same line of thinking, nudged elastic band (NEB) methods were developed by projecting out perpendicular components of elastic forces and parallel components of the true force with respect to the path under minimization.[3,4] In zero temperature string (ZTS) methods, the states along the transition pathway are evenly distributed and re-distributed along the fitted path after each step of minimization or simulation.[5−8] It should be noted that the ZTS method could automatically optimize end structures to local minima, regardless of whether the starting pathway connects two minima or not. Many variations of chain-of-states methods with improved efficiency and accuracy have been developed subsequently, but will not be covered in detail.[9−12]

Due to the high number of degrees of freedom, the convergence of chain-of-states methods can be slow in some applications, especially for macromolecules. Therefore, specialized chain-of-states methods were developed to generate reference reaction paths as a good approximation to the true minimum energy pathway (MEP) with fast convergence rates. For example, the replica path (RPATH) method implemented in the CHARMM program package utilizes harmonic restraints of both best-fit root-mean-square distances (RMSD) and meta-angles defined by RMSD to control the distribution of replicas along the pathway and the smoothness of the pathway.[13] In addition to harmonic restraints, equal distance holonomic constraints have also been implemented in CHARMM to maintain an even distribution of replicas along the pathway during the pathway optimization.[14] Based on a recent benchmark study, the MEP calculated by these reference pathway methods are comparable with NEB results, but with a superior convergence rate.[15] Because of low computational cost, such reference pathway methods could be used for extensive sampling of pathways.

In addition to determining the MEP, sampling along the pathway to estimate the transition free energy is also an active area of methodology development. To obtain a sufficient sampling around high-energy states for the accurate free energy estimation, enhanced sampling methods were developed. While not typically considered as an enhanced sampling method,

Received: June 9, 2017
Published: November 30, 2017

DOI: 10.1021/acs.jctc.7b00606
J. Chem. Theory Comput. 2018, 14, 14−29

umbrella sampling methods are arguably the most widely applied methods to gain sufficient sampling in high-energy regions to calculate a free energy profile with regard to the predefined order parameters.[16−19] Metadynamics methods do not sample toward a specific transition pathway but, instead, discourage revisiting already sampled states to promote exploration of other parts of a potential energy surface (PES), including high-energy states.[20−23] Yang and co-workers proposed a minimum free energy pathway method combined with hybrid quantum mechanical and molecular mechanical free energy (QM/MM-FE) methods to couple the optimization of the MEP of an enzymatic reaction pathway using QM/MM methods with sampling of the environment to obtain the most probable reaction pathway.[24,25] Transition path sampling (TPS) methods were first proposed by Pratt[26] and much further developed by Chandler and his collaborators.[27−31] Different from other enhanced sampling methods, no specific reaction pathway defined by a certain reaction coordinate is required in TPS methods. In TPS methods, a large number of simulations are carried out to sample the transitions between two target stable states. To enhance the success rate of transitions in the simulations, the TPS simulations often start from high-energy states, which are preferably close to the transition state regions between two states. Roux and co-workers combined string methods with a large number of short simulations, which were referred to as a swarm of trajectories, to identify the most probable transition pathways between two stable states.[32] Vanden-Eijnden and his co-workers further advanced the reaction pathway sampling by proposing transition path theory (TPT).[33−40] In TPT, all the transition pathways connecting two states are viewed as a part of a long trajectory, which samples both states numerous times. The probability density function and probability current function were proposed to measure the transition between two end states in the TPT. Dominant reaction path (DRP) methods were developed by Faccioli and co-workers based on the Fokker−Planck equation to search for the transition pathway with a minimum action potential.[41−45]

In addition to the development of the transition pathway optimization and enhanced sampling methods either targeting a specific transition pathway or between two predefined states, further methodology development is still necessary to explore potential multiple transition pathways connecting two states and detect additional minima on the PES of systems of interest and their associated transition pathways. The chain-of-states framework has been applied to sample the potentials of the mean force along a reference pathway showing certain advantages over related methods.[46] In a similar vein, we proposed a new method to combine the chain-of-states framework with molecular dynamics (MD) simulations to directly sample the target PES in the transition pathway space. We refer to this approach as the direct pathway dynamics sampling (DPDS) method. The remaining part of this work is organized as the following. Computational framework and methods are described in section 2. Test of the DPDS method using three model systems is presented in section 3. Both advantages and limitations of this method are discussed in section 4. The contribution is concluded in section 5.

2. COMPUTATIONAL METHOD

2.1. Chain-of-States Representation of Reaction Pathway. A transition pathway represented by a series of structures on a PES is the basic representation of a simulation system. The original chain-of-states algorithm was developed for energy optimization on a PES to characterize the MEP. The final MEP from any given representation of a transition pathway is likely a local MEP instead of the global one. Therefore, by combining the chain-of-states framework with molecular dynamics simulations, the sampling efficiency on a PES could be greatly enhanced. In a chain-of-states framework, a series of structures described by certain coordinates are constructed to represent a transition pathway from one state to another. Cartesian coordinates are a convenient option to generate and maintain replicas within a chain-of-states framework. The distance between two states (replicas) i and j, d^{ij}, can be defined as the following:[14,15]

$$d^{i,j} = \sqrt{\frac{\sum_{l=1}^{N} w_l (r_l^i - U^{ij} r_l^j)^2}{\sum_{l=1}^{N} w_l}} \tag{1}$$

where N is the total number of atoms in the system, r_l^i is the Cartesian coordinate of atom l in the structure i, U^{ij} is the rotation matrix to superimpose the structure j over the structure i for least-squares fitting between two structures, w_l is the weight factor for atom l, which includes atomic mass and additional factors.

In addition, any other appropriate reaction coordinates (RCs) or collective variables (CVs) can also be employed to maintain these replicas. The initial pathway can start with any state, and the MEP is a reasonable starting point. The end replicas are preferred to be two stable states as minima on the PES.

2.2. Restraints To Control Chain-of-States Pathways. Restraints are used to maintain the distance between adjacent states and prevent collapse of the pathway into a local minimum. The harmonic potential using best-fit RMSD has been shown as an effective metric to control the interstructure distance in the optimization of the MEP within the chain-of-states framework. The total potential energy associated with these added harmonic potentials to control RMSD between adjacent replicas is defined as the following[13,15]

$$E_{\text{rms}} = \sum_{i=1}^{n-1} \frac{1}{2} k_{\text{rms}} (d^{i,i+1} - \bar{d})^2 \tag{2}$$

where k_{rms} is the harmonic force constant to restrain RMSD distances between adjacent replicas along the reaction pathway, $d^{i,i+1}$ is the best-fit RMSD between replica i and $i + 1$, $\bar{d}$ is the average RMSD between adjacent replicas, and n is the total number of replicas of the given pathway.

To control the smoothness of the sampling pathway, an additional potential could be added to restrain the pseudoangle between structures along the pathway:[13,15]

$$E_{\text{angle}} = \sum_{i=1}^{n} \frac{1}{2} k_{\text{ang}} (\text{COSMAX} - \cos(\theta_i))^2 \quad \text{COSMAX} > \cos(\theta_i)$$

$$E_{\text{angle}} = 0 \quad \text{COSMAX} \le \cos(\theta_i) \tag{3}$$

The pseudoangle θ_i, illustrated in Figure 1, describes the smoothness of the pathway. The force constant k_{ang} controls the smoothness of the pathway by keeping θ_i from getting too small. COSMAX is a cutoff value determining the value of $\cos(\theta_i)$ subjected to the control of pathway smoothness. Both pathway RMSD and pseudoangle θ_i controlling are implemented in the RPATH module of the CHARMM program package.[13]

2.3. Molecular Dynamics Simulation within Chain-of-States Formalism. The conventional application of chain-of-states is to minimize the MEP, with the change of structures along the pathway only to reduce either the total energy of the

DOI: 10.1021/acs.jctc.7b00606
J. Chem. Theory Comput. 2018, 14, 14−29

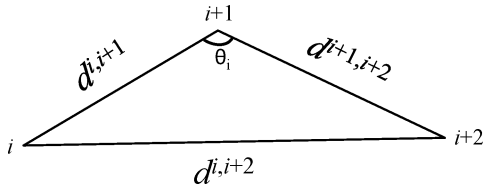


Figure 1. Illustration of angle θ_i around replica $i + 1$ in pathway control. $d^{i,i+1}$ is the distance between replica i and $i + 1$. It is similar to $d^{i+1,i+2}$ and $d^{i,i+2}$.

pathway when using restraints or forces perpendicular to the pathway. To achieve the goal of enhancing sampling efficiency on a PES, the dynamical propagation needs to be integrated with the chain-of-states control. This integration is implemented through adding pathway control forces, including distance potential and angle potential, into the potential functions for dynamical propagation. The force originating from the harmonic potential based on the RMSD restraints applied on atom l in structure i with Cartesian coordinates as r_l^i is defined as

$$f^i_{\mathrm{rmsd},l} = \frac{\partial E_{\mathrm{rms}}}{\partial r_l^i} \tag{4}$$

Similarly, the force originating from the harmonic potential applied to pseudoangle θ associated with structures, i, $i + 1$, and $i + 2$, is calculated as

$$f^i_{\mathrm{ang},l} = \frac{\partial E_{\mathrm{ang}}}{\partial r_l^i} \tag{5}$$

Both $f^i_{\mathrm{rmsd},l}$ and $f^i_{\mathrm{ang},l}$ are added to the force used for the MD simulations integration scheme for atom l in structure i. This integration scheme was implemented in CHARMM associated with the RPATH module.[47]

2.4. Nudged Elastic Band. Although NEB methods were originally developed to optimize the transition pathway on a PES to obtain the MEP, it is also suitable for the current framework. In the NEB methods, the forces along the pathway are projected using the pathway tangent vector (τ_i).

$$\begin{aligned}
\mathbf{F}_i &= \mathbf{F}_i^{\perp} + \mathbf{F}_i^{\parallel} \\
\mathbf{F}_i^{\perp} &= -\nabla V(\mathbf{r}_i)\cdot(1 - \boldsymbol{\tau}_i\boldsymbol{\tau}_i) \\
\mathbf{F}_i^{\parallel} &= -\nabla_i\left(\frac{1}{2}k\sum_{i=1}^{N}(\Delta l^i - \overline{\Delta l})^2\right)\cdot(\boldsymbol{\tau}_i\boldsymbol{\tau}_i)
\end{aligned} \tag{6}$$

where $\mathbf{F}_i^{\perp}$ and $\mathbf{F}_i^{\parallel}$ are the components of force that are perpendicular and parallel to the tangent vector of replica i ($\boldsymbol{\tau}_i$), respectively. V is the potential energy function, and Δl is the distance between adjacent replicas.

2.5. Control of the Two End Replicas. The main goal of conventional chain-of-states methods is to obtain an MEP connecting two minima. In most cases, two minima are characterized before obtaining the MEP connecting them. In such situations, two predefined minima are set as two end replicas on the initial chain-of-states setup and are constrained by removing all its degrees of freedom during the optimization process. Although restraint as harmonic potential can also be employed to control the end replicas, constraint is used for evaluating sampling efficiency of DPDS under the strict condition.

In DPDS applications, the end replicas could be handled differently to enhance the sampling efficiency. Three different ways to control the end replicas of the reaction pathway in DPDS methods were tested. (1) Controlling two ends: two end replicas in a chain-of-states setup are constrained similarly to the conventional application of the chain-of-states optimization toward the MEP. Therefore, the DPDS method will sample the reaction pathway space specifically connecting two predefined minima. (2) Controlling one end: only one end replica in a chain-of-states setup is constrained during the DPDS. In this way, the other end replica can potentially sample different space other than the initial structure. This can provide sufficient flexibility to the sampling system and allow additional minima to be identified during the simulations. (3) Releasing two ends: none of the end replicas in a chain-of-states setup is controlled by constraint during the DPDS, giving the most flexibility in terms of sampling the PES. In this way, the DPDS method can sample many stationary structures on a PES and identify transition pathways connecting these structures. In general, the DPDS method can be considered as a four-dimensional (4D) sampling method with the fourth dimension as transition pathway.

2.6. Clustering Analysis of Pathway Simulations. Similar to the clustering analysis of different conformers of flexible molecules, the transition pathways can also be subjected to clustering analysis to identify distinctive pathways. Generalized coordinates depending on the systems can be used for such clustering analysis. For a nonlinear system with m atoms, there can be l ($l \in [1, 3m - 6]$) generalized coordinates $\{g_1(\mathbf{R}), g_2(\mathbf{R}), ..., g_l(\mathbf{R})\}$ to be utilized to describe the transitions associated with conformational and structural changes. $\mathbf{R}$ is the set of $3m$ Cartesian coordinates ($\mathbf{x}_i$, $i = 1, 3m$) of all m atoms in the system. For meaningful comparison, the translational and rotational degrees of freedom should be projected out in most cases. For pathways comprising n structures with each structure described as l generalized coordinates, each structure i can be described by its l generalized coordinates $\{g_{i1}(\mathbf{R}_i), g_{i2}(\mathbf{R}_i), g_{i3}(\mathbf{R}_i), ..., g_{il}(\mathbf{R}_i)\}$; $\mathbf{R}_i$ is the set of $3m$ Cartesian coordinates of structure i. Therefore, a given pathway can be described by its pathway generalized coordinates matrix:

$$\begin{pmatrix} g_{11}(\mathbf{R}_1) & \cdots & g_{1l}(\mathbf{R}_1) \\ \vdots & \ddots & \vdots \\ g_{n1}(\mathbf{R}_n) & \cdots & g_{nl}(\mathbf{R}_n) \end{pmatrix}$$

For multiple pathways with the same number of structures, these pathway generalized coordinate matrices can be collected and subjected to vector quantization methods, such as k-means clustering analysis. For the three testing models in this study, different generalized coordinates were used to describe their potential energy surfaces and transitions. The clustering analyses were carried out using pathway generalized coordinate matrices for each testing system.

2.7. Implementation of DPDS Method. The basic steps of the DPDS method are outlined as a flowchart in Figure 2, comprising three parts: initial pathway preparation, dynamics pathway simulation, and final analysis. The goal of the preparation part is to generate an appropriate transition path representation as a starting point for the pathway dynamics simulations. To achieve this goal, one can adapt a conventional chain-of-states scheme to generate an MEP. Following this approach, as illustrated in the flowchart, two structures representing the start and end (for example reactant and

DOI: 10.1021/acs.jctc.7b00606
J. Chem. Theory Comput. 2018, 14, 14−29

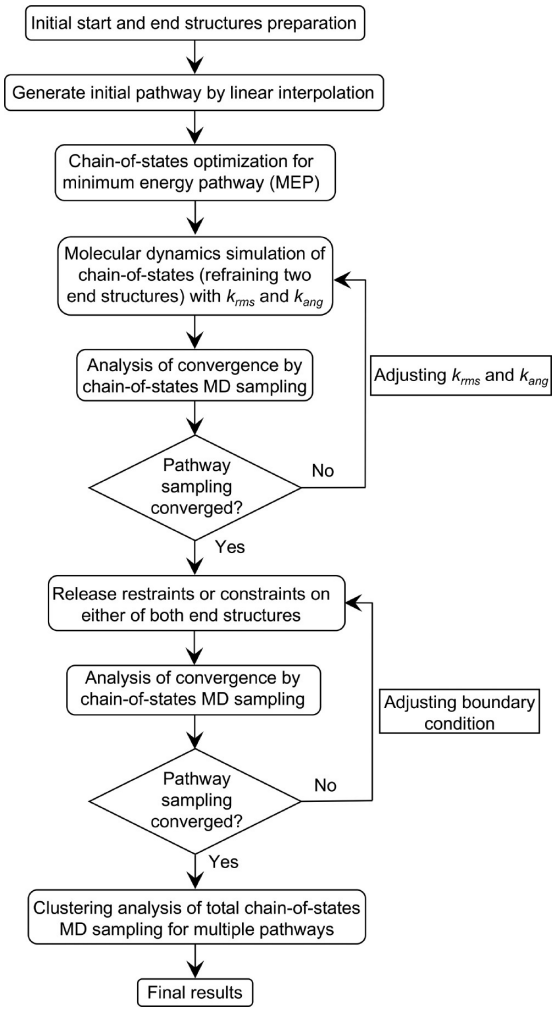


Figure 2. Flowchart for direct pathway dynamics sampling (DPDS) method.

product) states of a transition process should be identified initially. Then a series of structures connecting these two states can be constructed through linear interpolation or other means and be subjected to the chain-of-states pathway optimization to obtain an approximate MEP.

The optimized pathway will then be subjected to the dynamics simulations outlined in part two. If there is special interest in the transition between two end replicas of the initial pathway, either restraints (such as harmonic potential) or constraints (with these two structures being treated as rigid bodies) could be applied on these two end replicas during the simulations to prevent the sampling from escaping these states. Using k_{rms} and k_{ang} as controlling factors, molecular dynamics simulations can be applied to the initial pathway to sample the transitions connecting the initial two end states. Various analyses, including using order parameters, *k*-means clustering, or advanced time–structure independent component analysis, can be applied to monitor the coverage of simulations and evaluate the convergence of the simulations. If exploration of the potential energy surface is desired to identify additional stationary states and their associated transition pathways, the restraints or constraints on either one or both end replicas can be removed during the simulations to allow the sampling of the different area on the potential energy surface. In addition, some adjustment of k_{rms} and k_{ang} can also be applied during this part of the simulations. The coverage of simulations should also be monitored to estimate convergence of the simulations.

As the final analysis in part three, the clustering analysis should be carried out on all the pathway simulations or some selected part. *k*-means clustering algorithms can be applied for this purpose. The clustering analysis will reveal all the distinctive transition pathways sampled in the simulations that connect either predefined or newly discovered stationary structures on the target potential energy surface. If necessary, further exploration or sampling can be carried out for any newly identified transition pathway.

3. RESULTS

Three systems with different complexity and levels of theory were employed to test the DPDS method.

3.1. Isomerization of Alanine Dipeptide. The first test case is the isomerization of the alanine dipeptide (*N*-acetylalanyl-*N*-methyl-amide). Two backbone dihedral angles (ϕ and ψ) can be used to describe the isomerization process of this molecule (Figure 3A). An MEP was generated connecting two conformers C_{7eq} and C_{ax} corresponding to two minima on the PES (Figure 3B).[15] The CHARMM 22 force field[48] with CMAP backbone dihedral angle corrections[49] was used for the calculation. No solvent molecule is present in the model system. The chain-of-states calculations of the alanine dipeptide were carried out using 25 replicas. The barrier for the isomerization with reference to conformer C_{ax} is 8.74 kcal/mol. This MEP was used as the start pathway in the following DPDS simulations.

The distribution of a 1 μs conventional MD simulation of the alanine dipeptide is projected on the two-dimensional (2D) surface defined by ϕ and ψ (Figure 3C). The sampling covers two main attraction basins *a* and *b* circled by dashed lines in Figure 3C (basin *b* spreads to four corners of the plot as one single basin). The term attraction basin refers to the region with minimum free energy compared to adjacent regions. Apparently, the area between two basins was not sampled efficiently at all,

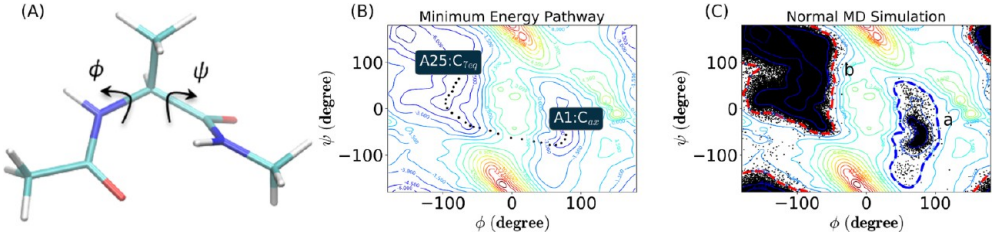


Figure 3. Alanine dipeptide as test case: (A) alanine dipeptide structure and two dihedral angles (ϕ and ψ) as reaction coordinates; (B) initial minimum energy pathway for alanine dipeptide isomerization; (C) distribution of conventional molecular dynamics simulations of alanine dipeptide on its potential energy surface with reference to ϕ and ψ. Two main attraction basins are labeled as *a* and *b*.

DOI: 10.1021/acs.jctc.7b00606
J. Chem. Theory Comput. 2018, 14, 14−29

because every time the sampling changes from one basin to the other basin, the transition process is transient without significant sampling of the pathway connecting two basins.

3.1.1. RPATH/Restraints Using k_{rms} Only. To explore the effect of k_{rms} on the sampling efficiency, k_{ang} was set to zero in the simulations of this section. Using a chain-of-states pathway with 25 structures, with restraint of distance between adjacent structures, the sampling was enhanced in various ways. With a rather small k_{rms} at 0.1 kcal/(mol·Å^2), the sampling of the smaller attraction basin was significantly enhanced (Figure 4A). To save

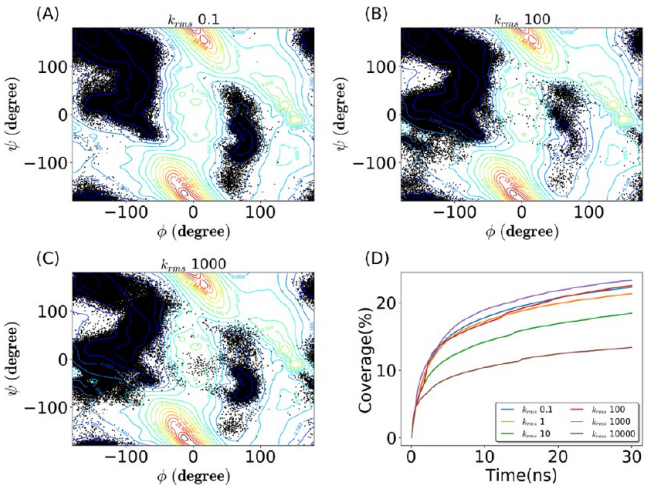


Figure 4. DPDS simulations of alanine dipeptide with different k_{rms} and PES coverage of each simulation: (A) k_{rms} = 0.1; (B) k_{rms} = 100; (C) k_{rms} = 1000; (D) plot of PES coverage along the simulation time.

space, the unit for k_{rms} (kcal/(mol·Å^2)) may not be presented in the remaining of the text. Various transition regions were sampled. However, the transition region between two basins was not sampled sufficiently due to the small k_{rms}. With larger k_{rms} of 100 and 1000, the sampling between basins including the transition from the larger basin to itself was much enhanced (Figure 4B,C). For the comparison of sampling efficiency, the PES of alanine dipeptide was divided into 40,000 grid squares by dividing both ϕ and ψ into 200 bins. The coverage of the surface was estimated based on the distribution of sampling trajectories among grid squares and plotted in Figure 4D. With k_{rms} ranging between 0.1 and 10000, the coverage of the PES varies significantly with k_{rms} = 10 showing the most efficient sampling (magenta line in Figure 4D). The samplings of k_{rms} as 1, 10, and 10000 did not show significant improvement of coverage on the PES and are listed in Table S1 in the Supporting Information. It should be noted that the two end replicas are constrained during the direct pathway simulations described here and in sections 3.1.2 and 3.1.3.

3.1.2. RPATH/Restraints Using k_{ang}. Another controlling factor, k_{ang}, is also applied. To test effectiveness of k_{ang}, a rather small k_{rms} of 0.1 kcal/(mol·Å^2) is used. The most effective k_{ang} values are 10, 100, and 1000 kcal/mol (Figure 5). To save space, the unit for k_{ang} (kcal/mol) may not be presented in the remainder of the text. Interestingly, different k_{ang} values led to the significant enhancement of sampling in different transition regions. With a k_{ang} value of 10 (Figure 5A), the transition path region connecting the same attraction basin *b* was extensively sampled (circled by blue dashed line), and a novel transition path region connecting two main attraction basins *a* and *b* was also revealed (circled by red dashed line). With a k_{ang} value of 100, the

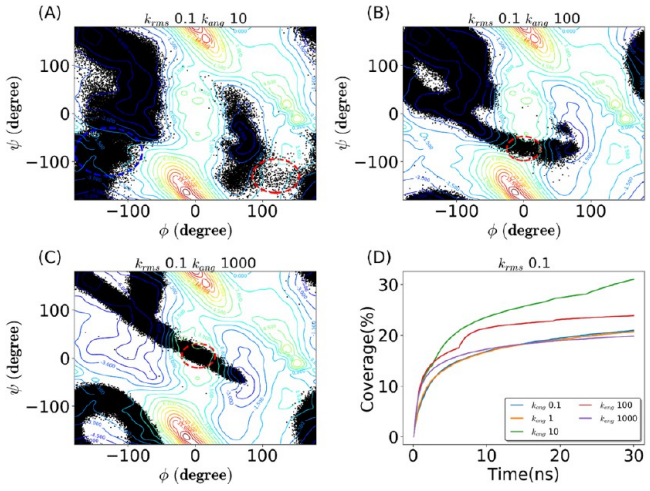


Figure 5. DPDS simulations of alanine dipeptide with different k_{ang} (while k_{rms} = 0.1 kcal/(mol·Å^2)) and PES coverage of each simulation: (A) k_{ang} = 10 kcal/mol; (B) k_{ang} = 100; (C) k_{ang} = 1000; (D) plot of PES coverage along the simulation time.

transition path region covering the original MEP of alanine dipeptide as the starting pathway of DPDS was extensively sampled (Figure 5B, circled by red dashed line). With a larger k_{ang} value of 1000, a second transition path region parallel to the starting pathway for DPDS (Figure 3B) was reached and extensively sampled (Figure 5C, circled by a red dashed line). Using k_{ang} as the other controlling factor, different valleys for transition pathways on the PES could be reached and sampled extensively using the DPDS method. The samplings of k_{ang} as 0.1 and 1 did not show significant improvement of coverage on the PES and are listed in Table S1 in the Supporting Information.

3.1.3. RPATH/Restraints Using k_{rms} and k_{ang}. Based on the outcome of simulations testing k_{rms} and k_{ang} individually, different combinations of k_{rms} and k_{ang} were used for alanine dipeptide simulations to gain a better understanding of the efficiency of the DPDS method. With k_{rms} as 10 and k_{ang} as 10, both transition path regions connecting attraction basin *b* to itself and the one connecting basins *a* and *b* were extensively sampled (Figure 6A,

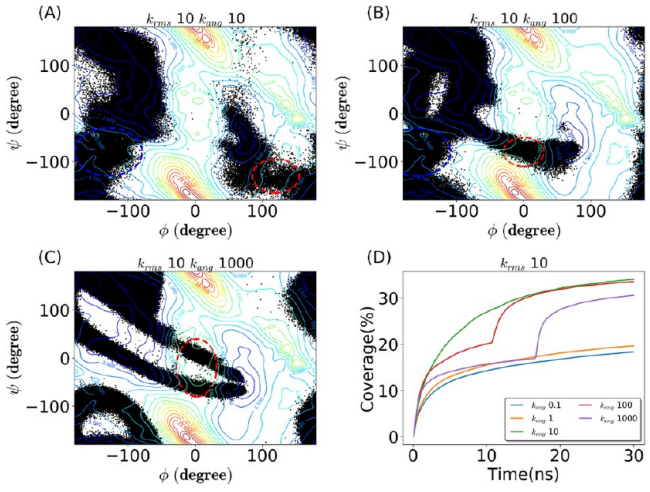


Figure 6. DPDS simulations of alanine dipeptide with different k_{ang} (while k_{rms} = 10) and PES coverage of each simulation to demonstrate high efficiency of sampling: (A) k_{ang} = 10; (B) k_{ang} = 100; (C) k_{ang} = 1000; (D) plot of PES coverage along the simulation time.

DOI: 10.1021/acs.jctc.7b00606
J. Chem. Theory Comput. 2018, 14, 14−29

circled by dashed lines). With k_{rms} as 10 and k_{ang} as 100, the transition path region covering the original MEP as the starting point of the simulation was extensively sampled (Figure 6B, circled by a red dashed line), similar to the previous simulation with the same k_{ang} but k_{rms} as 0.1 (Figure 5B). However, different from that case, the transition path region connecting the attraction basin b to itself was also extensively sampled with k_{rms} as 10 (Figure 6B, circled by a blue dashed line). Strikingly, with k_{rms} as 10 and k_{ang} as 1000, two transition pathway regions parallel to each other and connecting attraction basins a and b were both extensively sampled (Figure 6C, circled by dashed line). This observation demonstrates DPDS as a powerful method to explore and sample multiple transition pathway regions connecting two states. Another interesting observation of the sampling efficiency is that the PES coverage of the simulation with k_{ang} as 100 (cyan line in Figure 6D) or 1000 (magenta line in Figure 6D) have dramatic increases around 10 and 16 ns, respectively. These sudden increases of coverage occurred because pathway sampling switched to a different region on the PES. To illustrate these jumps in the DPDS simulation from one transition pathway region to a different region, the distributions before and after the jumping are illustrated in Figures 7 and 8 for these two simulations. The samplings with

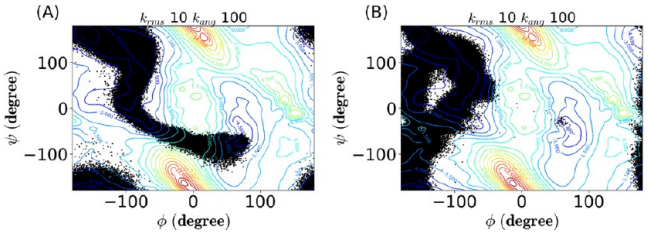


Figure 7. Switch of sampling between different transition pathway regions during the DPDS simulations with k_{rms} = 10 and k_{ang} = 100: (A) sampling during the first 10 ns; (B) sampling during the remaining 20 ns.

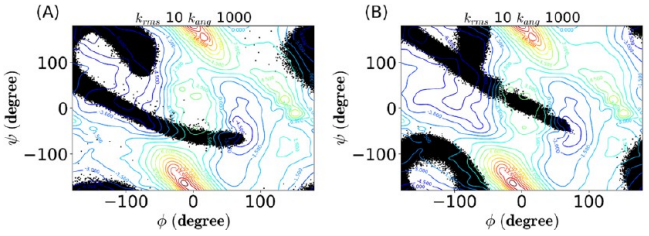


Figure 8. Switch of sampling between different transition pathway regions during the DPDS simulations with k_{rms} = 10 and k_{ang} = 1000: (A) sampling during the first 16 ns; (B) sampling during the remaining 14 ns.

k_{ang} as 0.1 and 1 did not show significant improvement of the coverage on the PES and are listed in Table S1 in the Supporting Information.

3.1.4. RPATH/Restraints with Removing Constraints on Either One or Both Ends. For many systems, not only are transition pathways unknown but also the minimum on the PES or stable states are also unknown. It will be beneficial if DPDS could help to search new minima while directly sampling the transition pathway space. To use the DPDS method as a potential tool to explore the PES, the constraints on either one or both of the end replicas could be removed to allow the sampling of the different states on the PES.

First, the constraint on replica number 25 (A25) was removed for alanine dipeptide DPDS. Therefore, only replica number 1 (A1) was constrained. Replica A25 is a minimum energy structure in basin b and A1 is a minimum energy structure in basin a (Figure 3B). The coverage and sampling efficiency with k_{rms} = 0.1 and k_{ang} = 10 are similar to the simulation with constraints on both end replicas but with higher coverage (34.5%; Figure 9A). However, with k_{rms} = 10 and k_{ang} = 10, DPDS

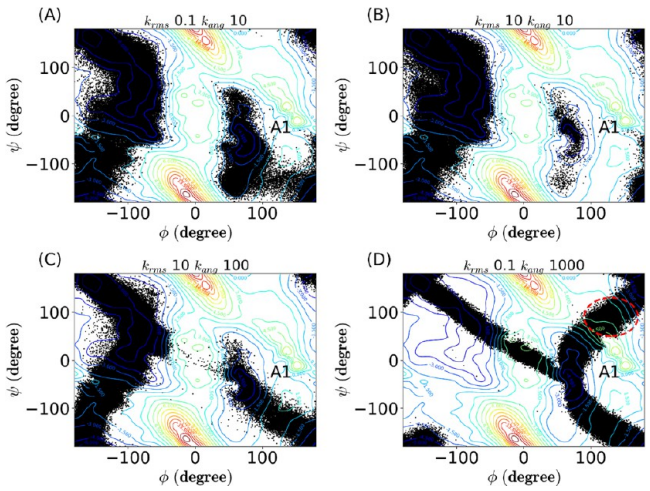


Figure 9. DPDS simulations of alanine dipeptide with constraint only on replica A1 with different k_{rms} and k_{ang} combinations: (A) k_{rms} = 0.1, k_{ang} = 10; (B) k_{rms} = 10, k_{ang} = 10; (C) k_{rms} = 10, k_{ang} = 100; (D) k_{rms} = 0.1, k_{ang} = 1000.

did not show improvement in either minimum sampling or PES coverage (Figure 9B). With k_{rms} = 10 and k_{ang} = 100, transition pathway regions that connect basin b to itself or basins a and b were sampled extensively (Figure 9C). Interestingly, a new transition pathway region connecting attraction basins a and b through the top right corner of the PES was sampled with k_{rms} = 0.1 and k_{ang} = 1000 simulation (Figure 9D, circled by red dashed line). Various combinations of k_{rms} and k_{ang} were also tested with constraint only on A25 with similar sampling coverage, and are listed in Table S1 in the Supporting Information.

Second, the constraint on A1 was removed for alanine dipeptide DPDS. Therefore, only A25 was constrained. The coverage and sampling efficiency with constrained A25 combination of k_{rms} = 0.1 and k_{ang} = 10 (Figure 10A) and k_{rms} = 10 and k_{ang} = 10 (Figure 10B) are similar to the simulation with constraint only on A1 (Figure 9A,B). With k_{rms} = 10 and k_{ang} = 100, a new sampling pattern on the PES was observed (Figure 10C). However, with strong k_{rms} (1000), the DPDS simulation could be trapped in a local attraction basin when the constraint on the A1 replica is removed, ultimately leading to less efficient sampling (Figure 10D). Various combinations of k_{rms} and k_{ang} were also tested with constraint only on A1 with similar sampling coverage, and are listed in Table S1 of the Supporting Information.

Third, the constraints on both end replicas were removed for alanine dipeptide DPDS. The coverage and sampling efficiency with a combination of k_{rms} = 0.1 and k_{ang} = 10 (Figure 11A) are similar to the previous simulations without constraint on either end replica, indicating that a weak k_{rms} does not pull either end replica out of attraction basins after removing the constraint forces. With a combination of k_{rms} = 10 and k_{ang} = 10, both minimum states and PES coverage are improved (Figure 11B). Interestingly, with k_{rms} = 10 and k_{ang} = 100, basin a is much less sampled compared to other samplings (Figure 11C). This is

DOI: 10.1021/acs.jctc.7b00606
J. Chem. Theory Comput. 2018, 14, 14−29

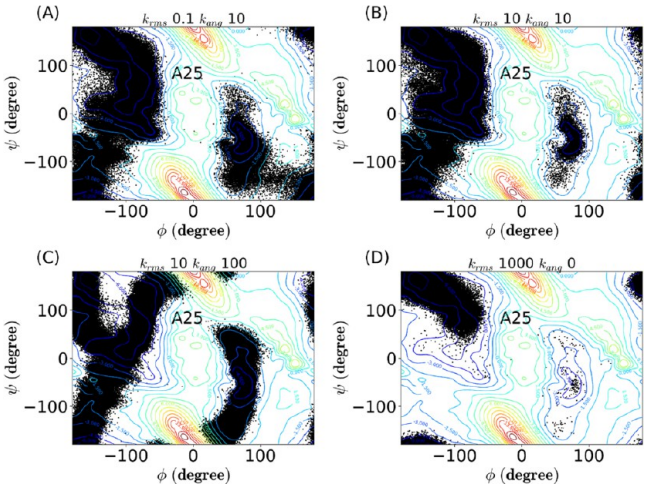


Figure 10. DPDS simulations of alanine dipeptide with constraint only on replica A25 with different k_{rms} and k_{ang} combinations: (A) k_{rms} = 0.1, k_{ang} = 10; (B) k_{rms} = 10, k_{ang} = 10; (C) k_{rms} = 10, k_{ang} = 100; (D) k_{rms} = 1000, k_{ang} = 0.

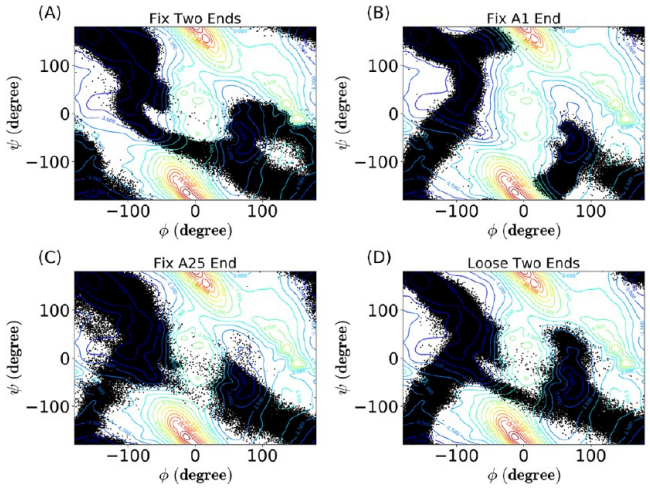


Figure 12. DPDS simulations of alanine dipeptide within nudged elastic band (NEB) framework: (A) with constraints on both ends; (B) with constraints only on replica A1; (C) with constraint only on replica A25; (D) without constraint on either end replica.

because the pathway was trapped within basin *b* during this simulation.

The above simulations show that, with a proper setup, the sampling efficiency could be much improved. New transition pathways could be discovered and sampled for further analysis.

3.1.5. NEB Results. Another popular chain-of-states method, the NEB method, was also applied within the DPDS framework. Both k_{rms} and k_{ang} were applied to control the sampling efficiency. With k_{rms} = 0.1 and k_{ang} = 100, the sampling showed the most effective results. The sampling with constraints on both ends was the most effective in finding the minimum and covering the PES (Figure 12A). Two different transition pathway regions connecting two attraction basins were detected and extensively sampled. With constraint on either end replicas, the simulations are similar to each other. The transition pathway region connecting the attraction basin *b* to itself was detected and extensively sampled in both simulations with constraints on either end replica (Figure 12B,C). In the simulations without constraints on either end replica, all the transition pathway regions detected in the above three simulations were also detected and extensively sampled (Figure 12D). All other DPDS results using the NEB method are illustrated in Table S2 in the Supporting Information.

3.1.6. Distinguished Pathways Identified through DPDS. All the DPDS simulations carried out for the alanine dipeptide were combined and subjected to clustering analysis to identify distinguished transition pathways. The clustering analysis was carried out using a density-based spatial clustering of application with noise (DBSCAN) algorithm.[50] For each pathway snapshot containing 25 replicas sampled using the DPDS method, two dihedral angles of alanine dipeptide in each replica leading to a total of 50 dihedral angles for each pathway were used for the clustering analysis. A cutoff of 200° was chosen for the ideal results, in which a total of 17 clusters were generated. Among these 17 clusters, nine clusters representing unique transition pathways were identified and are listed in Figure 13. All 17 clusters are presented in Table S3 in the Supporting Information. Pathway 1 represents transitions from attraction basin *a* to *b*. It also has a loop structure within basin *b* (Figure 13A). Pathway 2 has a shape similar to that of pathway 1, but with significant shifting toward attraction basin *a* (Figure 13B). In addition to the transition pathways connecting attraction basins *a* and *b*, pathway 3 presents another transition pathway that comes from basin *b* going through basin *a* before going back to basin *b* (Figure 13C). Pathway 4 represents a straight transition pathway going through both attraction basins, with a loop within attraction basin *b* (Figure 13D). Pathway 5 represents two new transition pathways; one connects two attraction basins, and the other one connects attraction basin *b* to itself (Figure 13E). The major part of pathway 6 resides in attraction basin *b* with the remaining part going through attraction basin *a* (Figure 13F). However, the transition between the two attraction basins is actually a false pathway, because it goes through an energy peak by having two adjacent replicas residing on the two sides of the peak with no replica in the high-energy area. This resulted from the simulations with small k_{rms}. Pathway 7 resides in attraction

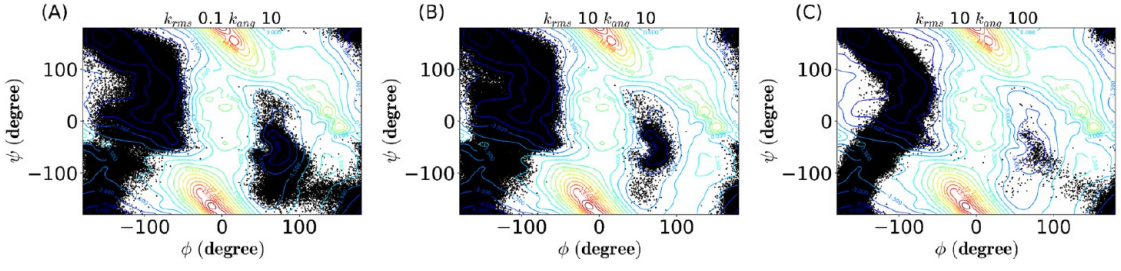


Figure 11. DPDS simulations of alanine dipeptide without constraint on either replica A1 or A25 with different k_{rms} and k_{ang} combinations: (A) k_{rms} = 0.1, k_{ang} = 10; (B) k_{rms} = 10, k_{ang} = 10; (C) k_{rms} = 10, k_{ang} = 100.

DOI: 10.1021/acs.jctc.7b00606
J. Chem. Theory Comput. 2018, 14, 14−29

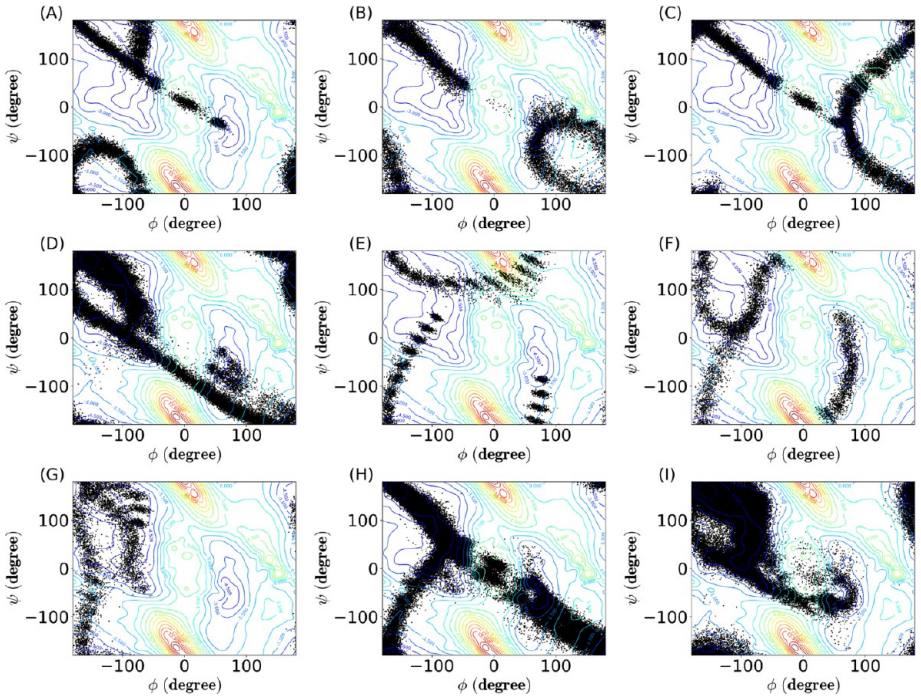
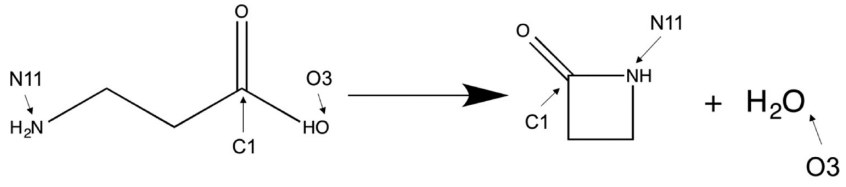


Figure 13. Unique transition pathways identified through clustering analysis of alanine dipeptide DPDS simulations. Nine among a total of 17 pathways are listed. All 17 pathways are illustrated in Table S3 in the Supporting Information.

Scheme 1



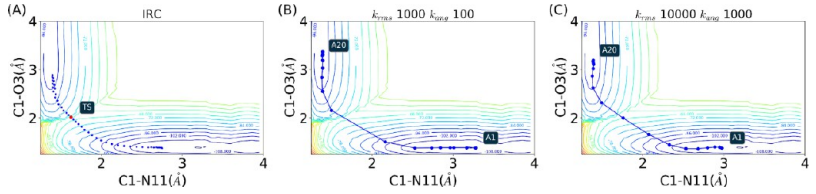


Figure 14. Minimum energy pathways of intramolecular condensation of β-alanine reaction: (A) using intrinsic reaction coordinates; (B) using RPATH method in CHARMM (k_{rms} = 1000, k_{ang} = 100); (C) using RPATH method in CHARMM (k_{rms} = 10000, k_{ang} = 1000).

basin *b* and with the transition pathway connecting basin *b* itself (Figure 13G). Pathway 8 (Figure 13H) represents a transition pathway connecting basin *b*, which is also represented in pathways 5, 6, and 7, and the transition pathway connecting two attraction basins, also seen in pathways 1 and 3. Pathway 9 (Figure 13I) represents an alternative pathway connecting attraction basins *a* and *b*, which is similar to the one represented in pathway 4. The aggregation of replicas was due to the small k_{rms} value used to carry out simulations that contribute significantly to this cluster.

The clustering analyses demonstrate the effectiveness of using the DPDS method as a tool to sample multiple pathways on the PES. With different combinations of controlling factors k_{rms} and k_{ang} as well as constraints on the end replicas, the DPDS could automatically sample multiple pathways, which provides valuable information for complicated processes. Careful analyses are also essential to obtain information about unique reaction pathways sampled in the simulations.

3.2. SCC-DFTB Reaction Pathway Sampling. The second test case is the intramolecular condensation reaction of β-alanine, a simple organic reaction (Scheme 1). Despite its apparent chemical simplicity, this reaction presents a real challenge for pathway sampling with two chemical bonds forming and two chemical bonds breaking simultaneously. The MEP connecting the reactant and product was constructed using the intrinsic reaction coordinate (IRC) method.[51] Its distribution on the PES is illustrated in Figure 14A. The lengths of two key bonds, carbon−oxygen (C1−O3) and carbon−nitrogen (C1−N11), are used as reaction coordinates to construct the PES. The actual transition state identified from a separate quantum mechanical calculation was also plotted as a red dot along the MEP. The MEPs connecting reactant and product were also constructed using RPATH, the chain-of-states method implemented in

DOI: 10.1021/acs.jctc.7b00606
J. Chem. Theory Comput. 2018, 14, 14−29

CHARMM (Figure 14B,C). The energy barrier from the IRC calculation is around 47.268 kcal/mol for this reaction.

3.2.1. DPDS with Combining k_{rms} and k_{ang}. Using chain-of-states with 20 structures along the pathway, different combinations of k_{rms} and k_{ang} were used to carry out DPDS for this reaction. Self-consistent charge density functional tight binding (SCC-DFTB) was shown to reproduce energetic results similar to those obtained from B3LYP/6-31+G(d,p) level of theory.[52] Therefore, the SCC-DFTB method[53,54] was applied for all DPDS simulations of this reaction. DFTB3 mio parameters with third order correction were used for the calculation.[55] No dispersion correction was employed in the current study. The Anderson mixing scheme was used in SCF iterations. The convergence criterion for the SCF cycle was set to be 10^{-7}. For each combination of k_{rms} and k_{ang} values, DPDS were carried out at 300 K for 1 ns using a 1 fs time step.

First, the constraints were applied on both end replicas during the simulations. The simulations demonstrated that large k_{rms} and particularly large k_{ang} values were necessary to enhance the sampling along the pathway, especially the region close to the transition state region (Figure 15). With k_{ang} smaller than 100, no apparent enhanced sampling was observed. With larger k_{rms} and k_{ang}, the sampling along the pathway was much enhanced (Figure 15A–C). The most effective combination to sample the transition pathways of this system is k_{rms} = 10000 and k_{ang} = 1000 (Figure 15D). The DPDS is evenly distributed along the pathway with extensive sampling throughout the pathway connecting two minimum states.

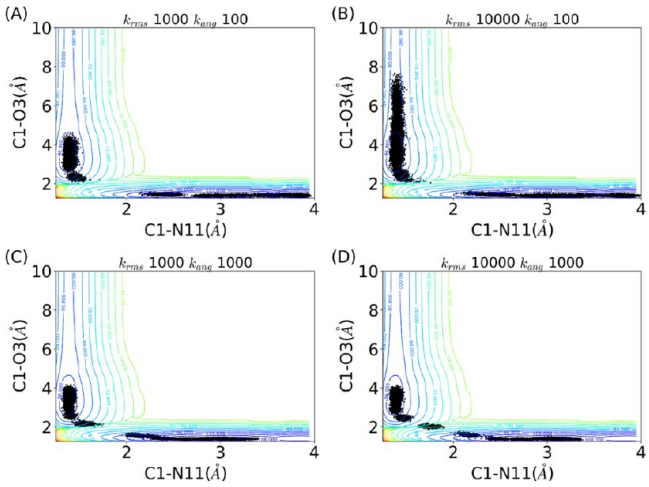


Figure 15. Simulation of intramolecular condensation of β-alanine reaction using DPDS method. Constraints were applied on both end replicas: (A) k_{rms} = 1000, k_{ang} = 100; (B) k_{rms} = 10000, k_{ang} = 100; (C) k_{rms} = 1000, k_{ang} = 1000; (D) k_{rms} = 10000, k_{ang} = 1000.

3.2.2. DPDS with Removing Constraint on Either One or Both Ends. When the constraint on A20 (product; see Figure 14) was removed leaving constraint only on A1, with k_{rms} = 100 and k_{ang} = 1000, the sampling in the product region was extended toward large C1–O3 distances significantly (Figure 16A). The reason for this is due to the escaping of the water molecule as a product away from the β-lactam ring product. With k_{rms} = 1000 and k_{ang} = 1000, the sampling is better controlled with enhanced sampling along the pathway (Figure 16B).

When the constraint on A1 (reactant; see Figure 14) was removed leaving constraint only on A20, the sampling in the reactant region was much extended (Figure 16C). With combinations of large k_{rms} and k_{ang}, (k_{rms} = 1000 and k_{ang} = 1000) and (k_{rms} = 10000 and k_{ang} = 1000), the sampling along the pathway was also enhanced (Figure 16C,D). When the constraints on both A1 and A20 were removed, the sampling in the reactant region was also much extended with a combination of large k_{rms} = 1000 and k_{ang} = 1000 (Figure 16E). With k_{rms} = 10000 and k_{ang} = 1000, the sampling along the pathway is much enhanced without constraints on either end replicas (Figure 16F).

3.3. β-Hairpin. As a third example, a more complex system of β-hairpin peptide folding was used to test sampling efficiency of DPDS method.[56,57] The sequence of the peptide is GEWTYD-DATKTFTVTE. The fold structure was obtained from a crystallographic structure available from the protein data bank[58] (PDB code 3GB1, residues 41–56). An extended structure of this peptide was generated using CHARMM.[47] To better evaluate and compare the sampling efficiency, the overall distribution of peptide conformation was generated based on all the obtained sampling and used as a background for distribution plots in this section. To build an overall distribution, a Markov state method, MSMbuilder,[59] was used to calculate dihedral angles along the backbone of the peptide for each structure. Then, time–structure independent component analysis (tICA) implemented in MSMbuilder was applied on all simulations of the peptide. The first two dominant components tICA1 and tICA2 were used to plot the overall distribution of β-hairpin simulations. A k-means clustering method was used to cluster the overall distribution into 11 clusters (Figure 17). Based on the nature of these clusters, we divided those clusters into three groups: β-hairpin (fold region), unfold region, and misfold region (Figure 17).

Two conventional MD simulations starting from either the fold structure or an extended structure were carried out for 1 μs. The simulation starting from the extended structure was trapped in cluster 4 with presence in clusters 2 and 6 in the misfold region (Figure 18A). On the other hand, the simulation starting from the fold structure was mainly distributed in clusters 3 and 9 in both the fold and unfold regions (Figure 18B).

An MEP was generated with 25 replicas connecting the extended and fold conformations (Figure 18C). The DPDS was carried out starting from this MEP with different k_{rms} and k_{ang} combinations and two end replicas being constrained. The distribution of each simulation is illustrated in Table S4 in the Supporting Information. Clustering analysis using k-means clustering algorithm[60] was carried out, and divided simulations with the same control of end replicas together into 10 clusters. For the DPDS with two end replicas being constrained, four out of 10 pathways (PA1 to PA4) are illustrated in Figure 19. Pathway PA1 demonstrates a smooth transition through a series of unfold clusters and fold clusters (5, 7, 10, 11, 9, and 1) (Figure 19A). Another slightly different pathway PA2 goes through clusters 2 and 8 instead of cluster 7 before cluster 3 (5, 2, 8, 3, 11, 9, and 1) (Figure 19B). Pathway PA3 actually starts in cluster 2 with extensive sampling and goes through a misfold cluster 6 (2, 6, 8, 3, 9, and 1; Figure 19C). These three clusters represent smooth transition channels between two end states through different clusters. Pathway PA4 represents a stochastic transition between two end states. With switching back and forth, this pathway samples clusters 5, 2, 6, 8, 7, 10, 11, 3, 9, and 1 (Figure 19D). Several other pathways are similar to pathway PA4 (Table S5 in Supporting Information). The uneven distribution shown in pathway PA4 could result from a weak k_{rms} that contributes to this cluster.

DOI: 10.1021/acs.jctc.7b00606
J. Chem. Theory Comput. 2018, 14, 14–29

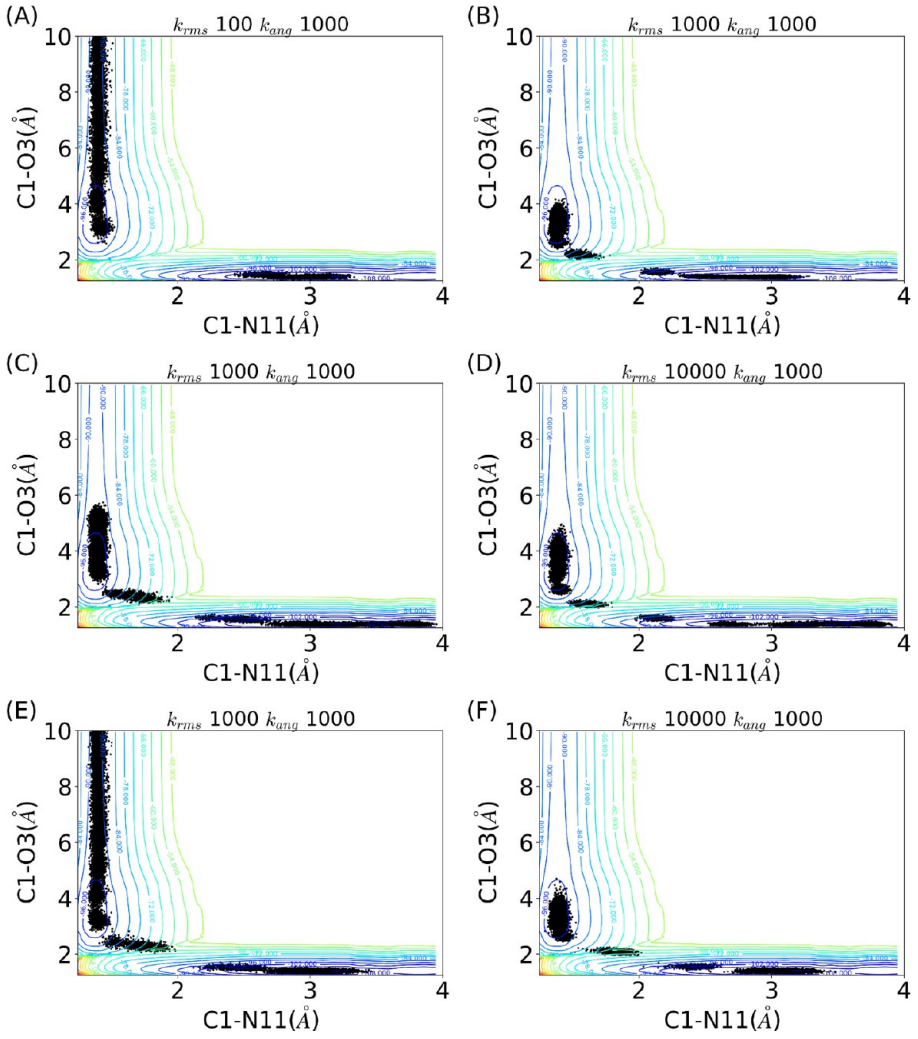


Figure 16. DPDS simulations of intramolecular condensation of β-alanine reaction with removed constraint on end replicas and different combinations of k_{rms} and k_{ang}: (A) constraint only on replica A1 k_{rms} = 100, k_{ang} = 1000; (B) constraint only on replica A1, k_{rms} = 1000, k_{ang} = 1000; (C) constraint only on replica A20, k_{rms} = 1000, k_{ang} = 1000; (D) constraint only on replica A20, k_{rms} = 10000, k_{ang} = 1000; (E) no constraints, k_{rms} = 1000, k_{ang} = 1000; (F) no constraint, k_{rms} = 10000, k_{ang} = 1000.

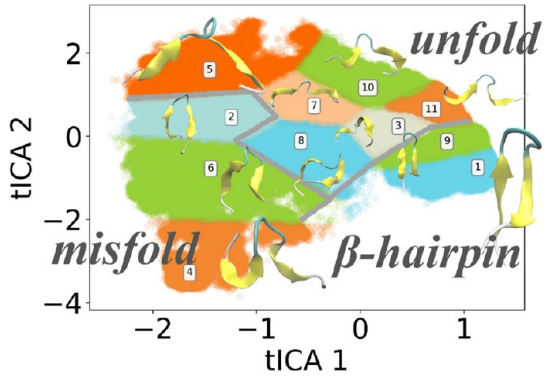


Figure 17. Overall conformational distribution and cluster analysis of β-hairpin peptide based on all DPDS and 1 μs conventional molecular dynamics (MD) simulations of fold/unfold states. A total of 11 clusters are identified with representative conformers illustrated.

Four representative pathways (PB1 to PB4) are illustrated in Figure 20 among 10 clusters generated from simulations without constraining the fold end replica (labeled as A25). In pathway PB1, end replica A25 remains in cluster 1. This pathway represents transition through clusters 5, 2, 6, 8, 3, 9, and 1 (Figure 20A). Pathway PB2 (Figure 20B) is similar to pathway PA2 (Figure 19B) with end replica A25 remaining in cluster 1. In pathway PB3, end replica 25 migrates to cluster 9. Pathway PB3 has uneven distribution and goes through clusters 5, 2, 7,10, 11, 9, and 1 and ends within cluster 9 (Figure 20C). End replica 25 also migrates to cluster 9 in pathway PB4, which goes through clusters 5, 2, 10, 7, 8, 6, 3, 11, 1, and 9 in a zigzag pattern (Figure 20D).

When the constraint on replica 1 is removed, DPDS displays much more diversity than the above samplings. Pathway PC1 (Figure 21A) starts with replica 1 in cluster 2 as a misfold structure and goes through clusters 8, 3, 11, 9, and 1, similar to part of pathway PA2 (Figure 19B). Pathway PC2 (Figure 21B) has replica 1 remaining in cluster 5 and half of it similar to pathway PC1. Pathway PC3 (Figure 21C) starts with replica 1 in cluster 2, goes through clusters 5 and 6, and leads to a second half similar to pathway PC1. In pathways PC4, PC5, and PC6, replica 1 samples different part of clusters 6 and 4 as misfold regions (Figure 21D−F). All three pathways have zigzag patterns, and go through the unfold region before reaching the fold region.

When constraints on both end replicas were removed, DPDS did not present much more diverse pathways (Figure 22).

DOI: 10.1021/acs.jctc.7b00606
J. Chem. Theory Comput. 2018, 14, 14−29

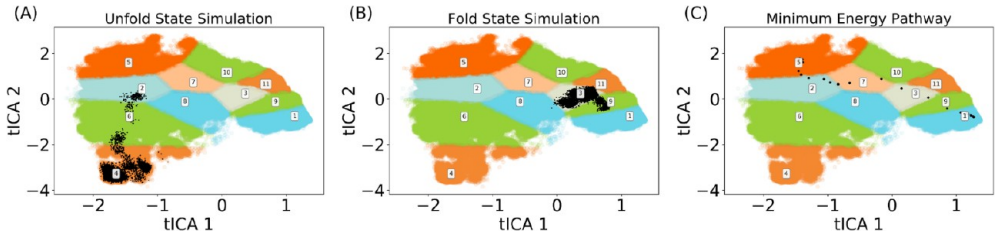


Figure 18. Distributions of conventional MD simulations staring from either an extended or fold conformation and initial minimum energy pathway: (A) MD simulations staring from extended conformation; (B) MD simulations staring from a folded conformation; (C) minimum energy pathway.

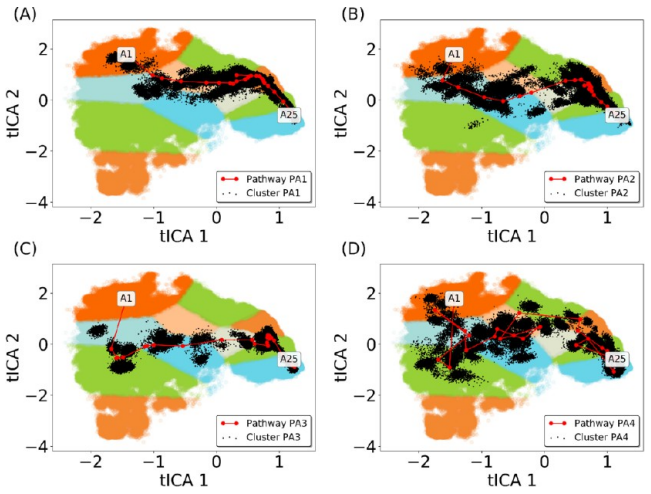


Figure 19. Clusters and four representative pathways (PA1 through PA4) of β-hairpin based on DPDS simulations with constraints on both end replicas (A1 and A25).

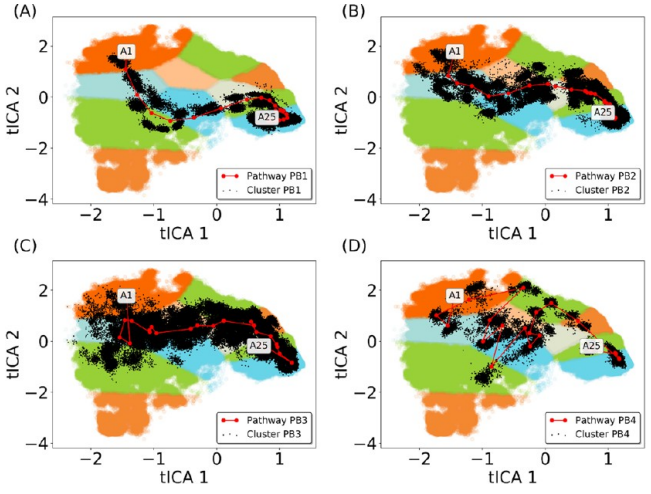


Figure 20. Clusters and their representative pathways of β-hairpin based on DPDS with constraint only on replica 1.

Pathway PD1 (Figure 22A) is similar to previous pathways, such as PB2. Pathway PD2 (Figure 22B) is very similar to pathway PC1. Pathway PD3 (Figure 22C) starts in cluster 2 and goes through cluster 6, both in the misfold region, before going through the unfold region and reaching the fold region. The end replica 25 in pathway PD3 is rather close to cluster 3 in the unfold region. Pathways PD4, PD5, and PD6 are zigzag pathways with replica 1 heavily sampling the misfold region (Figure 22D–F).

From the sampling, it seems to be extremely unlikely that misfold states of this β-hairpin can directly change to the fold states. All the simulations presented in this study indicate that the misfold states need to unfold first before reaching the fold states.

3.4. Comparison with the Transition Path Sampling Method. For comparison, another pathway sampling method, the TPS method,[61] implemented in CHARMM was applied on alanine dipeptide isomerization and β-hairpin peptide folding. To apply the TPS method, two attraction basins and an appropriate start geometry for successful shooting trajectories are required. In TPS method, new transition pathways are generated from old pathways through shooting and shifting algorithms.[62] In the shooting algorithm, a random frame in the old pathway is selected and modified by adding a random momentum perturbation. MD simulations are carried out from the selected frame along the new velocities in both forward and backward directions in time to generate a new transition pathway. In the shifting algorithm, a random frame and its atomic velocities in the old pathway are selected. MD simulations are carried out from the selected frame and velocities in either forward or backward directions in time to generate a new transition pathway with the same length as the old one, effectively shifting the old pathway either forward or backward in time. The shifting algorithm is complementary to the shooting algorithm. Although the new pathways overlap with the old pathway, shifting moves will improve the convergence of transition path sampling.[62] Successful shooting trajectories are those simulations starting from the start geometry that connects two attraction basins. Normally structures similar to the transition state could serve as the start geometry leading to successful shooting trajectories.

For alanine dipeptide, two attraction basins *a* and *b* are defined based on the dihedral angles used as reaction coordinates. To ensure sufficient TPS sampling, these two basins need to be defined relatively large (for *a*, $\phi(40°,100°)$ and $\psi(-150°,25°)$, blue rectangle in Figure 23A; for *b*, $\phi(-170°,-50°)$ and $\psi(-50°,170°)$, red rectangle in Figure 23A). After some initial tests, an appropriate start geometry of alanine dipeptide was identified. Starting from this initial geometry, a total of 1,000,000 shooting pathways each as 2 ps long were carried out. The shooting and shifting movement ratio was chosen as 9:1 for optimal sampling efficiency. The acceptance rate is about 9% for shooting movement and 90% for pathway shifting movement. The TPS simulations of alanine dipeptide lead to good coverage of the potential energy surface with 72% coverage (Figure 23A). Although most major transition pathways connecting basins *a* and *b* were sampled, one transition pathway between two basins was not detected by these otherwise comprehensive simulations.

For the β-hairpin, the attraction basins were defined using two structural order parameters, one as distance *d* between α carbons of two terminal residues (Gly1 and Glu16) and the other one as angle θ defined by Gly1, Ala8, and Glu16 α carbons (with Ala8 as the vertex). The two attraction basins were defined as fold

DOI: 10.1021/acs.jctc.7b00606
J. Chem. Theory Comput. 2018, 14, 14–29

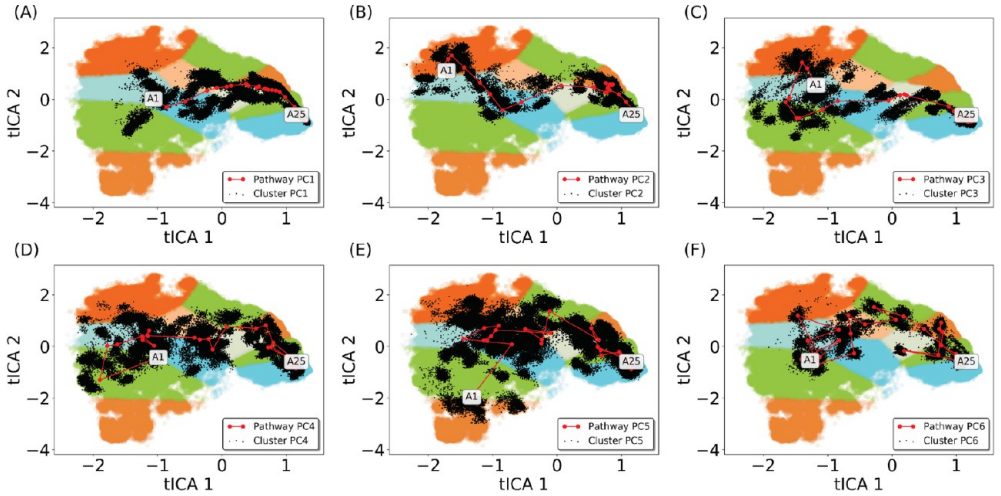


Figure 21. Clusters and their representative pathways of β-hairpin based on DPDS with constraint only on replica A25.

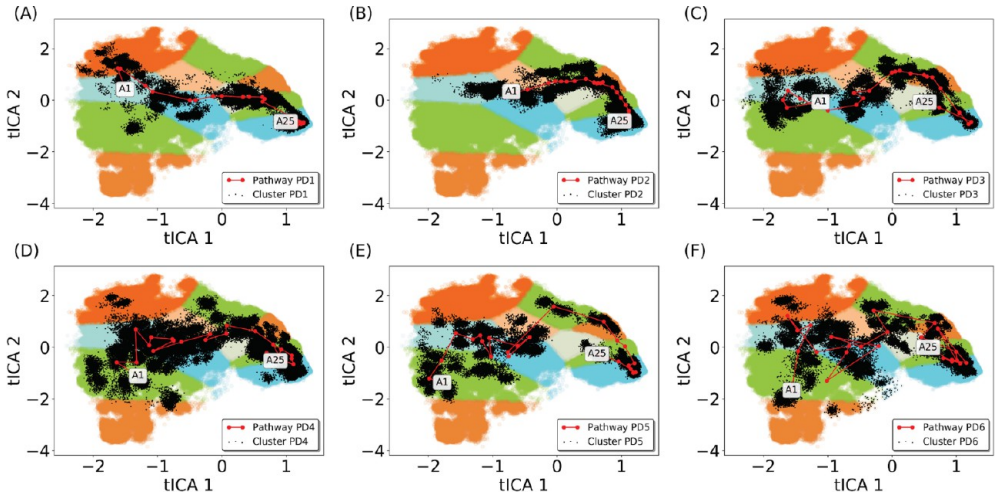


Figure 22. Clusters and their representative pathways of β-hairpin based on DPDS without constraint on end replicas.

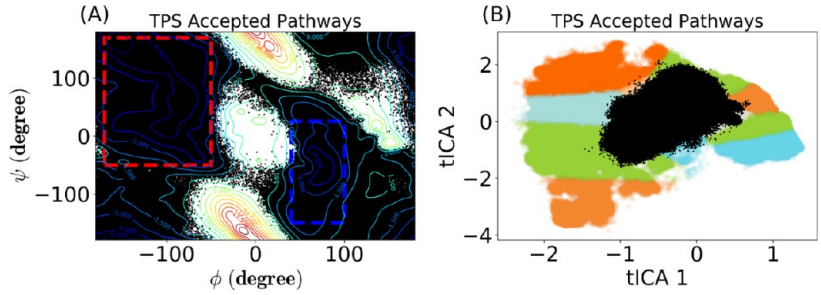


Figure 23. Transition path sampling (TPS) results: (A) alanine dipeptide isomerization (two attraction basins defined for the setup of TPS method illustrated in rectangles); (B) β-hairpin folding. The attracting basins for the TPS method are defined using order parameters different from tICA coordinates used in this plot and, thus, could not be illustrated.

(d(0.0Å,8.0Å), θ (0°,45°)) and unfold (d(22.0Å,40.0Å), θ (70°,180°)). Although there are different means to define attraction basins more relevant to folding, the above order parameters were chosen for the purpose of interpretation and implementation. The replica number 13 from the minimum energy pathway of β-hairpin folding was chosen as the initial geometry for TPS simulations. Starting from this geometry, a total of 1,000,000 shooting pathways each as 10 ps long were carried out. Longer trajectories are necessary for shooting simulations to reach both attraction basins. The shooting and pathway shifting movement ratio was chosen as 9:1 for optimal sampling efficiency. Similar to the alanine dipeptide case, the acceptance rate is about 9% for shooting movement and 90% for pathway shifting movement. The TPS simulations mainly cover unfold regions, part of the misfold region, and a very small part of the fold region (Figure 23B). The lack of coverage on the folded structures is due to the quick termination of the shooting trajectories when reaching an attraction basin representing folded structures.

DOI: 10.1021/acs.jctc.7b00606
J. Chem. Theory Comput. 2018, 14, 14−29

In the above TPS simulations, the shooting trajectories can be up to a certain length (2 ps for alanine dipeptide and 10 ps for β-hairpin), but the length of each trajectory varies significantly. Therefore, it is not convenient to carry out the pathway clustering analysis. The plots of the TPS simulations on the potential energy surfaces of these two systems could be compared with the DPDS. For the simple alanine dipeptide case, the TPS simulations provided sufficient coverage on the PES, but missed a key transition pathway. For the β-hairpin, the TPS simulations only cover a portion of the regions sampled by the DPDS simulations.

4. DISCUSSION

In this study, we developed and tested the DPDS method as a new approach for direct sampling in the transition pathway space. Through the analyses of three test cases, the DPDS method demonstrates certain advantages. The chain-of-states setup of the method ensures the sampling of transition pathway on the PES of the target systems. This could greatly enhance the sampling efficiency when the main goal of sampling is exploring feasible transition pathways connecting multiple minimum states. The DPDS method is compatible with most chain-of-states methods, such as nudged elastic band and string methods. Another advantage of the DPDS method is that minimum knowledge about the pathways on the target PES is required a priori. In addition, the dynamical sampling increases the probability of the sampling to reach high-energy barrier transition pathways even starting with the global minimum energy pathway. By removing constraints or restraints on either or both end replicas, the DPDS could reach any transition pathway on a PES.

However, it should be emphasized that, like all other enhanced sampling methods, it is not guaranteed that DPDS can exhaust all the feasible pathways even with long simulation time. Therefore, the use of parameters k_{rms} and k_{ang} provides additional control of DPDS to reach different transition pathways. With small k_{rms} and k_{ang}, the sampling will favor low-energy space. By increasing k_{rms}, one could obtain the sampling more faithful to the actual transition pathways. On the other hand, by increasing k_{ang}, the DPDS is more likely to switch to and detect multiple transition pathways. When testing the different combinations of k_{rms} and k_{ang}, the DPDS of alanine dipeptide isomerization exhausts all the major transition pathways identified on its PES.[63] Analyzing DPDS results to obtain transition pathway information is relatively easy, because each snapshot of the sampling is a transition pathway itself. Clustering analysis of DPDS results will directly lead to multiple transition pathways connecting multiple minima on the targeting PES. However, one should be cautious about applying high k_{ang} in DPDS, because high k_{ang} will force smoothing of the transition pathway and lead to high-energy barriers.

One may be concerned about the dependence of the DPDS sampling on the initial pathway. Therefore, an MEP (Figure 24A) different from the one used in the DPDS simulation of alanine dipeptide was employed as the initial pathway and subjected to three DPDS simulations using optimal combinations of k_{rms} and k_{ang} parameters. The simulation with k_{rms} = 10 and k_{ang} = 10 (Figure 24B) starting from this MEP is similar to the simulation starting from the original MEP with the same k_{rms} and k_{ang} (Figure 6A). The simulation with k_{rms} = 10 and k_{ang} = 100 does sample the transition region between attraction basins *a* and *b* (circled region in Figure 24C) where the original MEP is located. Although the simulation with k_{rms} = 10 and k_{ang} = 1000 presents unique coverage on the PES (Figure 24D), the pathways

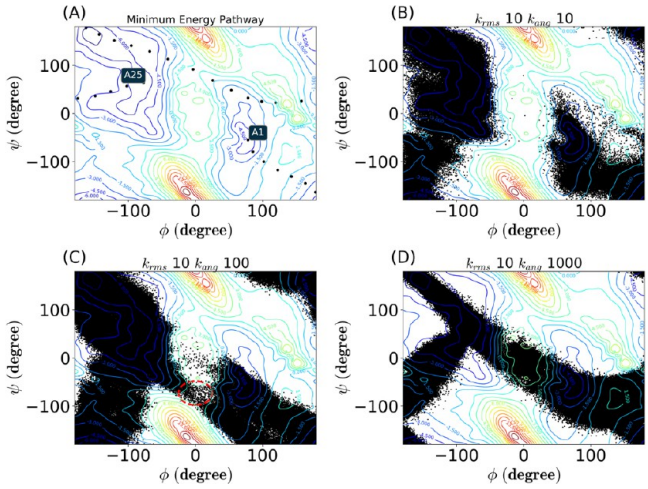


Figure 24. DPDS simulations of alanine dipeptide with an alternative start pathway: (A) alternative initial pathway (both A1 and A25 as two end replicas constrained during the simulation; different k_{rms} and k_{ang} combinations); (B) k_{rms} = 10, k_{ang} = 10; (C) k_{rms} = 10, k_{ang} = 100; (D) k_{rms} = 10, k_{ang} = 1000.

sampled in this simulation were also sampled in the DPDS with constraint on replica A1 (Figure 8C,D). This suggests that the dependence of DPDS on the start pathway is not significant.

Because of the convenience of the setup, DPDS can be easily applied with QM/MM methods to sample chemical reaction pathways. The sampling of intramolecular condensation reaction of β-alanine using the DPDS method indicates that high k_{rms} and k_{ang} are necessary for sufficient sampling of chemical reactions. This is due to the inherent high-energy barrier of chemical reactions, which requires large force to maintain even distribution along the pathway. Another popular sampling method along the chemical reaction pathway is the umbrella sampling. In an umbrella sampling method, harmonic potentials using order parameters or collective variables are implemented for sampling windows to force the simulations of certain regions. Using DPDS method, one has the flexibility through different weighting factors to control the RMSD distance between replicas. The umbrella sampling is often combined with the weighted histogram analysis method (WHAM) to obtain free energy information along the sampling windows. By obtaining enhanced sampling along multiple pathways, the DPDS method can also be used to estimate free energy information on the targeting PES. This is under development for future publication.

The DPDS method could be the most effective to search for transition pathways of a complicated system as demonstrated by the β-hairpin peptide in this study. Many complicated biomolecular processes, such as protein folding, ligand binding, and protein−protein binding, do not have convenient order parameters or collective variables to describe the processes as transition pathways. This could be resolved by the DPDS method through sampling transition pathways connecting two end states, which can be unfold/fold proteins or protein unbound/bound with its binding partners. Without defining specific order parameters, DPDS could potentially detect the most probable transition pathways connecting two end states and explore multiple transition pathways. Again, different combinations of k_{rms} and k_{ang} can be an effective means to drive sampling among different transition pathways.

Based on the flexibility from the combinations of k_{rms} and k_{ang}, DPDS simulations could be carried out with different emphases.

DOI: 10.1021/acs.jctc.7b00606
J. Chem. Theory Comput. 2018, 14, 14−29

On one hand, with low k_{rms} and k_{ang} values, the simulations will sample the attraction basins with moderate increasing coverage of the transition region as shown in Figure 4. These simulations could be utilized to detect and sample attraction basins. On the other hand, larger k_{rms} and k_{ang} values will enhance the detection and sampling of transition pathway regions as shown in Figures 5 and 6. For complex systems, as demonstrated by the β-hairpin in this study, the DPDS simulations with different combinations of k_{rms} and k_{ang} are useful to explore the potential energy surface closely related to the transition processes of interest. These simulations will provide not only the information about potential attraction basins but the transition pathways connecting these basins as well. Although transition pathways sampled in DPDS simulations are not minimum energy pathways, the representative pathways generated from clustering analyses can be subjected to further optimization to obtain minimum energy pathways.

Although the TPS method could be powerful to sample the transition pathway space for a simple system as demonstrated by alanine dipeptide, the efficient application of this method for complex systems is challenging and requires the careful setup of the simulations. In the β-hairpin peptide case, the TPS simulations only sampled a portion of the overall sampling space from the DPDS simulations. This is because TPS is limited to two predefined attraction basins, which are used as criteria to terminate the simulations. Although the shifting movement of TPS could help to enhance the sampling efficiency, these termination criteria still limit the sampling efficiency of new minima or attraction basins. The DPDS simulations, especially when removing restraints or constraints on both or either end replicas, could greatly enhance the detection and the sampling of new minimum or attraction basins for complex systems.

There is also a benefit for the computational cost of DPDS method in terms of sampling the high-energy regions along transition pathways. Comparing to the TPS method, it is not necessary to choose an appropriate initial geometry a priori for shooting simulations in DPDS method. Because of the chain-of-states framework, the "acceptance" rate for transition pathways in DPDS simulations is 100% compared to the 9% acceptance rate of shooting trajectories in the TPS simulations for the two test cases in this study, even after our best effort to fine-tune the TPS simulations. It should be noted that our experience of using the TPS method is limited, and it is likely that a better setup could lead to a higher acceptance rate. Nevertheless, the setup for high sampling efficiency of transition pathways could prevent effective application of the TPS method especially for complex systems. In general, the convenience of the setup within the chain-of-states framework in DPDS will help to generate sufficient coverage of transition pathway regions and to explore the potential energy surface for additional attraction basins.

Similar to many other enhanced simulation methods, there is no foolproof way to guarantee the convergence of the DPDS simulations. However, one still could have a good idea about the converging trend of the simulations. To evaluate the convergence of DPDS simulations, one of the key factors is selecting appropriate generalized coordinates for distribution plots. This would be relatively easy for simple cases, such as the alanine dipeptide isomerization and the β-alanine intramolecular condensation reaction tested in this study. For complex systems with high degrees of freedom, appropriate dimension reduction will be necessary and critical. It was shown in this and several other studies[64−66] that time−structure independent component analysis can be very effective to construct generalized coordinates to be used to plot the distribution of dynamics simulations. The distribution of accumulated DPDS simulations projected on the suitable generalized coordinates will be an effective tool for not only monitoring the convergence of the simulations but also identifying unique transition pathways through clustering analysis.

5. CONCLUSION

In this study, we developed a direct pathway dynamics sampling (DPDS) method for efficient sampling of a potential energy surface and exploring transition pathways. Two parameters k_{rms} and k_{ang} could be implemented for effective controlling of pathway sampling. Sampling with small k_{rms} and k_{ang} will favor a low-energy space, similar to conventional molecular dynamics sampling. Higher k_{rms} will lead to enhanced sampling along the pathway with decreasing distances among replicas. Higher k_{ang} will smoothen the pathway and increase the likelihood for the simulation to switch to new transition pathways. Using different combinations of k_{rms} and k_{ang}, the DPDS method can sample the majority of the PES important for transitions and detect multiple pathways, which would not be easily obtained and analyzed otherwise. The convenience of setup and analysis of DPDS results for transition pathway information makes this method an ideal option for transition pathway sampling related to biomolecules, such as proteins. Using RMSD combining with weight factors, complex processes such as β-hairpin peptide folding were efficiently sampled and multiple folding pathways were identified. In summary, the DPDS method provides a simple and effective means to directly sample transition pathways for complex systems and can be easily combined with various levels of theory.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.7b00606.

Distributions of DPDS simulations and clustering analysis of alanine dipeptide isomerization and β-hairpin peptide folding (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Email: ptao@smu.edu.

ORCID

Peng Tao: 0000-0002-2488-0239

Funding

This work was partially supported by the Edward R. Biehl Graduate Fellowship (H.Z.), SMU Dean's Research Council, and American Chemical Society Petroleum Research Fund (Grant 57521-DNI6).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Computational time was provided by Southern Methodist University's Center for Scientific Computation and Texas Advanced Computing Center (TACC) at the University of Texas at Austin. We thank Dr. Shuanghong (Sharon) Hou for critical reading of the manuscript and fruitful discussions and Dr. Alex Lippert for help with the manuscript preparation.

DOI: 10.1021/acs.jctc.7b00606
J. Chem. Theory Comput. 2018, 14, 14−29

REFERENCES

(1) Tao, P.; Larkin, J. D.; Brooks, B. R., Reaction Path Optimization and Sampling Methods and Their Applications for Rare Events. In *Some Applications of Quantum Mechanics*; Pahlavani, M. R., Ed.; InTech: Rijeka, Croatia, 2012; pp 27−66, DOI: 10.5772/35351.

(2) Elber, R.; Karplus, M. A method for determining reaction paths in large molecules: Application to myoglobin. *Chem. Phys. Lett.* **1987**, *139* (5), 375−380.

(3) Jónsson, H.; Mills, G.; Jacobsen, K. W., Nudged Elastic Band Method for Finding Minimum Energy Paths of Transitions. In *Classical and Quantum Dynamics in Condensed Phase Simulations*; Berne, B. J., Ciccotti, G., Coker, D. F., Eds.; World Scientific: Singapore, 1998; pp 385−404.

(4) Henkelman, G.; Jóhannesson, G.; Jónsson, H. Methods for Finding Saddle Points and Minimum Energy Paths. In *Theoretical Methods in Condensed Phase Chemistry*; Schwatz, S. S., Ed.; Kluwer Academic: New York, 2000; pp 269−302.

(5) E, W.; Ren, W.; Vanden-Eijnden, E. String method for the study of rare events. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2002**, *66* (5), 052301.

(6) Ren, W. Higher Order String Method for Finding Minimum Energy Paths. *Commun. Math. Sci.* **2003**, *1* (2), 377−384.

(7) E, W.; Ren, W.; Vanden-Eijnden, E. Simplified and improved string method for computing the minimum energy paths in barrier-crossing events. *J. Chem. Phys.* **2007**, *126* (16), 164103.

(8) Cameron, M.; Kohn, R. V.; Vanden-Eijnden, E. The String Method as a Dynamical System. *J. Nonlinear Sci.* **2011**, *21* (2), 193−230.

(9) Bergonzo, C.; Simmerling, C. An Overview of String-Based Path Sampling Methods. *Annu. Rep. Comput. Chem.* **2011**, *7*, 89−97.

(10) Birkholz, A. B.; Schlegel, H. B. Path optimization by a variational reaction coordinate method. I. Development of formalism and algorithms. *J. Chem. Phys.* **2015**, *143* (24), 244101.

(11) Hernández, E. R.; Herrero, C. P.; Soler, J. M. A chain-of-states acceleration method for the efficient location of minimum energy paths. *J. Chem. Phys.* **2015**, *143* (18), 184104.

(12) Birkholz, A. B.; Schlegel, H. B. Path optimization by a variational reaction coordinate method. II. Improved computational efficiency through internal coordinates and surface interpolation. *J. Chem. Phys.* **2016**, *144* (18), 184101.

(13) Woodcock, H. L.; Hodoscek, M.; Sherwood, P.; Lee, Y. S.; Schaefer, H. F., III; Brooks, B. R. Exploring the quantum mechanical/molecular mechanical replica path method: a pathway optimization of the chorismate to prephenate Claisen rearrangement catalyzed by chorismate mutase. *Theor. Chem. Acc.* **2003**, *109* (3), 140−148.

(14) Brokaw, J. B.; Haas, K. R.; Chu, J. W. Reaction Path Optimization with Holonomic Constraints and Kinetic Energy Potentials. *J. Chem. Theory Comput.* **2009**, *5* (8), 2050−2061.

(15) Tao, P.; Hodoscek, M.; Larkin, J. D.; Shao, Y.; Brooks, B. R. Comparison of Three Chain-of-States Methods: Nudged Elastic Band and Replica Path with Restraints or Constraints. *J. Chem. Theory Comput.* **2012**, *8* (12), 5035−5051.

(16) Torrie, G. M.; Valleau, J. P. Monte Carlo free energy estimates using non-Boltzmann sampling: Application to the sub-critical Lennard-Jones fluid. *Chem. Phys. Lett.* **1974**, *28* (4), 578−581.

(17) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23* (2), 187−199.

(18) Bartels, C.; Karplus, M. Multidimensional adaptive umbrella sampling: Applications to main chain and side chain peptide conformations. *J. Comput. Chem.* **1997**, *18* (12), 1450−1462.

(19) Kästner, J. Umbrella sampling. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1* (6), 932−942.

(20) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (20), 12562−12566.

(21) Bussi, G.; Gervasio, F. L.; Laio, A.; Parrinello, M. Free-Energy Landscape for β Hairpin Folding from Combined Parallel Tempering and Metadynamics. *J. Am. Chem. Soc.* **2006**, *128* (41), 13435−13441.

(22) Raiteri, P.; Laio, A.; Gervasio, F. L.; Micheletti, C.; Parrinello, M. Efficient Reconstruction of Complex Free Energy Landscapes by Multiple Walkers Metadynamics. *J. Phys. Chem. B* **2006**, *110* (8), 3533−3539.

(23) Laio, A.; Gervasio, F. L. Metadynamics: A method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Rep. Prog. Phys.* **2008**, *71* (12), 126601.

(24) Zhang, Y.; Liu, H.; Yang, W. Free energy calculation on enzyme reactions with an efficient iterative procedure to determine minimum energy paths on a combined ab initio QM/MM potential energy surface. *J. Chem. Phys.* **2000**, *112* (8), 3483−3492.

(25) Hu, H.; Lu, Z. Y.; Parks, J. M.; Burger, S. K.; Yang, W. T. Quantum mechanics/molecular mechanics minimum free-energy path for accurate reaction energetics in solution and enzymes: Sequential sampling and optimization on the potential of mean force surface. *J. Chem. Phys.* **2008**, *128* (3), 034105.

(26) Pratt, L. R. A statistical method for identifying transition states in high dimensional problems. *J. Chem. Phys.* **1986**, *85* (9), 5045−5048.

(27) Bolhuis, P. G.; Dellago, C.; Geissler, P. L.; Chandler, D. Transition path sampling: Throwing ropes over mountains in the dark. *J. Phys.: Condens. Matter* **2000**, *12* (8A), A147.

(28) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. TRANSITION PATH SAMPLING: Throwing Ropes Over Rough Mountain Passes, in the Dark. *Annu. Rev. Phys. Chem.* **2002**, *53* (1), 291−318.

(29) Dellago, C.; Bolhuis, P. G.; Geissler, P. L. Transition Path Sampling. In *Advances in Chemical Physics*, Vol. *123*; Prigogine, I., Rice, S. A., Eds.; John Wiley & Sons: Hoboken, NJ, USA, 2002; pp 1−78, DOI: 10.1002/0471231509.ch1.

(30) Dellago, C.; Bolhuis, P. G. Activation Energies from Transition Path Sampling Simulations. *Mol. Simul.* **2004**, *30* (11−12), 795−799.

(31) Dellago, C.; Bolhuis, P. G. Transition Path Sampling and Other Advanced Simulation Techniques for Rare Events. In *Advanced Computer Simulation Approaches for Soft Matter Sciences III*; Holm, C.; Kremer, K., Eds.; Advances in Polymer Science, Vol. *221*; Springer: Berlin, Heidelberg, 2009; pp 167−233, DOI: 10.1007/978-3-540-87706-6_3.

(32) Pan, A. C.; Sezer, D.; Roux, B. Finding Transition Pathways Using the String Method with Swarms of Trajectories. *J. Phys. Chem. B* **2008**, *112* (11), 3432−3440.

(33) E, W.; Ren, W.; Vanden-Eijnden, E. Finite temperature string method for the study of rare events. *J. Phys. Chem. B* **2005**, *109* (14), 6688−6693.

(34) Ren, W.; Vanden-Eijnden, E.; Maragakis, P.; E, W. Transition pathways in complex systems: Application of the finite-temperature string method to the alanine dipeptide. *J. Chem. Phys.* **2005**, *123* (13), 134109.

(35) E, W.; Ren, W.; Vanden-Eijnden, E. Transition pathways in complex systems: Reaction coordinates, isocommittor surfaces, and transition tubes. *Chem. Phys. Lett.* **2005**, *413* (1), 242−247.

(36) Maragliano, L.; Fischer, A.; Vanden-Eijnden, E.; Ciccotti, G. String method in collective variables: minimum free energy paths and isocommittor surfaces. *J. Chem. Phys.* **2006**, *125* (2), 024106.

(37) E, W.; Vanden-Eijnden, E. Towards a Theory of Transition Paths. *J. Stat. Phys.* **2006**, *123* (3), 503.

(38) Metzner, P.; Schütte, C.; Vanden-Eijnden, E. Illustration of transition path theory on a collection of simple examples. *J. Chem. Phys.* **2006**, *125* (8), 084110.

(39) Vanden-Eijnden, E. Transition Path Theory. In *Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology*, Vol. *1*; Ferrario, M., Ciccotti, G., Binder, K., Eds.; Springer: Berlin, Heidelberg, 2006; pp 453−493.

(40) Vanden-Eijnden, E.; Venturoli, M. Revisiting the finite temperature string method for the calculation of reaction tubes and free energies. *J. Chem. Phys.* **2009**, *130* (19), 194103.

(41) Faccioli, P.; Sega, M.; Pederiva, F.; Orland, H. Dominant Pathways in Protein Folding. *Phys. Rev. Lett.* **2006**, *97* (10), 108101.

(42) Sega, M.; Faccioli, P.; Pederiva, F.; Garberoglio, G.; Orland, H. Quantitative Protein Dynamics from Dominant Folding Pathways. *Phys. Rev. Lett.* **2007**, *99* (11), 118102.

DOI: 10.1021/acs.jctc.7b00606
J. Chem. Theory Comput. 2018, 14, 14−29

(43) Faccioli, P. Characterization of Protein Folding by Dominant Reaction Pathways. *J. Phys. Chem. B* **2008**, *112* (44), 13756−13764.
(44) Autieri, E.; Faccioli, P.; Sega, M.; Pederiva, F.; Orland, H. Dominant reaction pathways in high-dimensional systems. *J. Chem. Phys.* **2009**, *130* (6), 064106.
(45) Mazzola, G.; Beccara, S. a.; Faccioli, P.; Orland, H. Fluctuations in the ensemble of reaction pathways. *J. Chem. Phys.* **2011**, *134* (16), 164109.
(46) Hudson, P. S.; White, J. K.; Kearns, F. L.; Hodoscek, M.; Boresch, S.; Lee Woodcock, H. Efficiently computing pathway free energies: New approaches based on chain-of-replica and Non-Boltzmann Bennett reweighting schemes. *Biochim. Biophys. Acta, Gen. Subj.* **2015**, *1850* (5), 944−953.
(47) Brooks, B. R.; Brooks, C. L., 3rd; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: the biomolecular simulation program. *J. Comput. Chem.* **2009**, *30* (10), 1545−1614.
(48) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102* (18), 3586−3616.
(49) Mackerell, A. D.; Feig, M.; Brooks, C. L. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* **2004**, *25* (11), 1400−1415.
(50) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X., A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*; AAAI Press: Portland, OR, USA, 1996; pp 226−231.
(51) Schlegel, H. B. Exploring potential energy surfaces for chemical reactions: An overview of some practical methods. *J. Comput. Chem.* **2003**, *24* (12), 1514−1527.
(52) Cui, Q.; Elstner, M.; Kaxiras, E.; Frauenheim, T.; Karplus, M. A QM/MM Implementation of the Self-Consistent Charge Density Functional Tight Binding (SCC-DFTB) Method. *J. Phys. Chem. B* **2001**, *105* (2), 569−585.
(53) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1998**, *58* (11), 7260−7268.
(54) Elstner, M.; Frauenheim, T.; Kaxiras, E.; Seifert, G.; Suhai, S. A Self-Consistent Charge Density-Functional Based Tight-Binding Scheme for Large Biomolecules. *Phys. Status Solidi B* **2000**, *217* (1), 357−376.
(55) Gaus, M.; Cui, Q.; Elstner, M. DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB). *J. Chem. Theory Comput.* **2011**, *7* (4), 931−948.
(56) Yang, H.; Wu, H.; Li, D.; Han, L.; Huo, S. Temperature-Dependent Probabilistic Roadmap Algorithm for Calculating Variationally Optimized Conformational Transition Pathways. *J. Chem. Theory Comput.* **2007**, *3* (1), 17−25.
(57) Duan, M.; Fan, J.; Li, M.; Han, L.; Huo, S. Evaluation of Dimensionality-Reduction Methods from Peptide Folding−Unfolding Simulations. *J. Chem. Theory Comput.* **2013**, *9* (5), 2490−2497.
(58) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235−242.
(59) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. MSMBuilder2: Modeling Conformational Dynamics on the Picosecond to Millisecond Scale. *J. Chem. Theory Comput.* **2011**, *7* (10), 3412−3419.
(60) MacQueen, J. In Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1: Statistics*, Berkeley, CA, USA, 1967; Le Cam, L. M., Neyman, J., Eds.; University of California Press: Berkeley, CA, USA, 1967; pp 281−297.
(61) Hagan, M. F.; Dinner, A. R.; Chandler, D.; Chakraborty, A. K. Atomistic understanding of kinetic pathways for single base-pair binding and unbinding in DNA. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100* (24), 13922−13927.
(62) Dellago, C., Bolhuis, P. G., Geissler, P. L., Transition Path Sampling Methods. In *Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology*, Vol. *1*; Ferrario, M., Ciccotti, G., Binder, K., Eds.; Springer: Berlin, Heidelberg, 2006; pp 349−391.
(63) Jang, H.; Woolf, T. B. Multiple pathways in conformational transitions of the alanine dipeptide: an application of dynamic importance sampling. *J. Comput. Chem.* **2006**, *27* (11), 1136−41.
(64) Naritomi, Y.; Fuchigami, S. Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: The case of domain motions. *J. Chem. Phys.* **2011**, *134* (6), 065101.
(65) Schwantes, C. R.; Pande, V. S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J. Chem. Theory Comput.* **2013**, *9* (4), 2000−2009.
(66) Schwantes, C. R.; Shukla, D.; Pande, V. S. Markov State Models and tICA Reveal a Nonnative Folding Nucleus in Simulations of NuG2. *Biophys. J.* **2016**, *110* (8), 1716−1719.

DOI: 10.1021/acs.jctc.7b00606
J. Chem. Theory Comput. 2018, 14, 14−29