JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

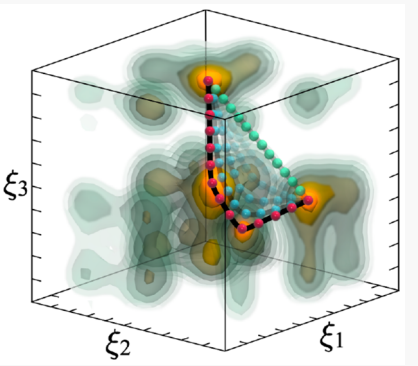
A Companion Guide to the String Method with Swarms of Trajectories: Characterization, Performance, and Pitfalls

Haochuan Chen, Dylan Ogden, Shashank Pant, Wensheng Cai, Emad Tajkhorshid, Mahmoud Moradi, Benoît Roux, and Christophe Chipot*

Cite This: https://doi.org/10.1021/acs.jctc.1c01049 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The string method with swarms of trajectories (SMwST) is an algorithm that identifies a physically meaningful transition pathway—a one-dimensional curve, embedded within a high-dimensional space of selected collective variables. The SMwST algorithm leans on a series of short, unbiased molecular dynamics simulations spawned at different locations of the discretized path, from whence an average dynamic drift is determined to evolve the string toward an optimal pathway. However conceptually simple in both its theoretical formulation and practical implementation, the SMwST algorithm is computationally intensive and requires a careful choice of parameters for optimal cost-effectiveness in applications to challenging problems in chemistry and biology. In this contribution, the SMwST algorithm is presented in a self-contained manner, discussing with a critical eye its theoretical underpinnings, applicability, inherent limitations, and use in the context of path-following free-energy calculations and their possible extension to kinetics modeling. Through multiple simulations of a prototypical polypeptide, combining the search of the transition pathway and the computation of the potential of mean force along it, several practical aspects of the methodology are examined with the objective of optimizing the computational effort, yet without sacrificing accuracy. In light of the results reported here, we propose some general guidelines aimed at improving the efficiency and reliability of the computed pathways and free-energy profiles underlying the conformational transitions at hand.



1. INTRODUCTION

Many important phenomena of physical, chemical, and biological relevance occur on time scales that largely exceed milliseconds (ms), e.g., large conformational transitions in complex biological objects. Observing by means of all-atom simulations broad movements between protein domains connected allosterically could be valuable to complement experiment, and help address important biological challenges. In spite of significant advances on the hardware and the software fronts, the time scales amenable to molecular dynamics (MD)—commonly from tens of microseconds (μs) on commodity clusters to a few milliseconds (ms) on special-purpose computers[1] and large arrays of computing nodes equipped with graphics processing units (GPUs)—remain orders of magnitude less than those spanned by biological processes. Computational investigation of such rare events, over ms and beyond, requires more than brute-force simulations on faster computers to dilate both time and length scales. The limitations that brute-force MD impose require other avenues to be examined to sample the underlying slow degrees of freedom.

The string method goes directly to the heart of this challenge by trying to determine the dominant transition pathway connecting two metastable states constructed as a one-dimensional curve within a reduced subspace of selected degrees of freedom, for example, a set of atomic Cartesian coordinates (CCs) or some mathematical transformations of CCs that are already known (also commonly referred to as "collective variables" or CVs).[2,3] In practice, the string method represents the curvilinear pathway as a chain of M discrete "images" or "nodes" in the subspace of the CVs or CCs, each node corresponding to a full copy of the entire system. One of its variants, the string method with swarms-of-trajectories (SMwST),[4] consists of progressively refining a trial transition pathway on the basis of the average dynamic local drift at each node. Toward this end, multiple short, unbiased trajectories are spawned at each node, forming a swarm, from which the mean dynamic drift is estimated on the fly. The string connecting two metastable basins is then refined through successive iterations, until the local drift at each node is zero in directions orthogonal to the tangent of the path, hence the name "zero-drift pathway" (ZDP). To put the significance of the string method into proper context, it is helpful to first

Received: October 18, 2021

review the computational approaches that are commonly used to study slow molecular processes and rare events.

Enhanced-sampling approaches, notably free-energy calculations,[5] have long been an essential tool to characterize the dynamics of slow degrees of freedom. Exploring with all-atom computer simulations biological events involving entangled, slow movements of large amplitude remains, however, thwarted by conceptual and methodological hurdles rooted in (i) the definition of the CVs capable of describing the conformational transitions associated with these events and (ii) the algorithm that encourages sampling along these CVs. Enhancing sampling to investigate phenomena spanning long time scales is commonly achieved by defining a general-extent parameter, $\xi(\mathbf{x})$, i.e., a model of the reaction coordinate (RC), and encouraging exploration of important regions of configurational space through (i) the application of an external force along $\xi(\mathbf{x})$, as is commonly done with free-energy-oriented approaches, like the adaptive biasing force (ABF),[6,7] metadynamics (MtD),[8] and umbrella sampling (US)[9] algorithms; (ii) the definition of interfaces along $\xi(\mathbf{x})$ and determination of transition probabilities between them, which forms the basis of schemes like milestoning[10] and simulation-enabled estimation of kinetic rates (SEEKR);[11] (iii) the definition of regions in the direction of $\xi(\mathbf{x})$, and preferential sampling of these regions, as is done in weighted-ensembles simulations;[12] (iv) spawning short trajectories, beginning and ending at the reference state, and pushing further along $\xi(\mathbf{x})$, which is the central idea of the adaptive multilevel splitting (AMS) algorithm.[13]

Still, each method is plagued by slow degrees of freedom in orthogonal space, which hampers efficient sampling. This limitation, due to a misrepresentation of the RC, results in sampling nonuniformity and quasi non-ergodicity in the direction of $\xi(\mathbf{x})$. To address these shortcomings, common to nearly all importance-sampling algorithms,[5,14] a number of options should be considered, namely, (i) turning to ergodic-sampling algorithms,[15] like bias-exchange US,[16] or multiple-walker ABF (MW-ABF),[17,18] hoping to populate important regions in orthogonal space; (ii) modifying the definition of $\xi(\mathbf{x})$, notably through a dimensionality reduction scheme,[19−21] $\{x_1, ..., x_N\} \rightarrow \{z_1, ..., z_n\}$, where $\{x_1, ..., x_N\}$ represents the complete set of Cartesian coordinates (CCs) of the molecular object of interest, and $\{z_1, ..., z_n\}$, the set of CVs, with $n \ll N$; (iii) turn to specialized importance-sampling algorithms,[5,14] like the most recent well-tempered MtD-extended adaptive biasing force (WTM−eABF) scheme,[22,23] which introduces ergodicity in the sampling. There is a strong connection between the choice of the importance-sampling algorithm and the representation of $\xi(\mathbf{x})$, which necessarily impacts not only sampling efficacy, but also the underlying dynamics. This connection naturally raises the question—what constitutes a suitable set of CVs, or, said differently, what are the important degrees of freedom that contribute predominantly to $\boldsymbol{\xi}(\mathbf{x})$, which has been tackled by many research groups in the past 20 years with a host of approaches that cannot all be cited here.[2,4,10,21,24−30]

Failure to model the RC in complex biological objects and erroneous description of the underlying dynamics stem from our reductionist view of the CV space—e.g., describing the permeation of a substrate across the cell membrane in terms of a Euclidian distance projected onto the normal to the interface results in non-Markovian dynamics.[31] A number of methods, in particular, Markov state models (MSMs),[25,32] have been devised to identify from multiple short MD simulations the relevant CVs associated with conformational transitions. Tools like time-structure independent components analysis (tICA)[27] can be utilized in conjunction with MSMs to extract the slowest movements from time series of linear combinations of input CVs. In stark contrast with principal component analysis (PCA),[33] which identifies motions corresponding to the largest variance, tICA targets through time-correlation analysis the slowest degrees of freedom. Variants of PCA, like relative PCA (RPCA),[34] have recently emerged, and offer a more satisfactory approach for singling out the relevant degrees of freedom in geometric transitions. MSMs and tICA are, however, costly, requiring usually long simulations and knowledge of the metastabilities to infer meaningful information.[35] It might be argued that the necessity of ms-simulations to address ms-biological phenomena defies the purpose of time scale bridging. Furthermore, assumption that the Markov chain yields perforce the correct kinetics constitutes a conceptual leap of faith.[36] This methodological ceiling in our ability to define a proper RC from scratch has limited our ability to investigate slow and complex processes that are underlying the function of large protein assemblies.

It is in this broad context that the string method,[2−4] which is built upon transition-path theory (TPT),[37,38] helps deepen our understanding of the concept of RC. In TPT, the RC is fundamentally associated with the concept of committor,[24,39] defined as the commitment probability that a trajectory starting from a given initial condition will reach target state B before crossing state A. In the vicinity of states A and B, which are metastabilities of the free-energy landscape, the committor approaches 0 and 1, respectively. The transition region is foliated into isocommittor surfaces crossed by reactive trajectories. One statistically important pathway connects state A to state B by following the probability flux densities (see Figure 1).[40] The string method aims at determining such a pathway, thereby helping better define an optimal RC for a slow process of interest.

Up until now, the string method has been applied to capture conformational transition pathways in various molecular processes, such as the hydrophobic collapse of a hydrated chain,[41] activation of c-Src kinase,[42−44] mechanical coupling in myosin,[45,46] substrate selection in DNA polymerase,[47] amyloidogenic isomerization of β2-microglobulin,[48] ligand-induced transition in adenylate kinase,[49] DFG-flip in insulin receptor kinase,[50,51] ion-channels gating,[52,53] V_0-[54] and V_1-ATPase,[55] ATP-driven calcium pumping in SERCA,[56] drug permeation through porins,[57] and membrane transport proteins.[58−61] For illustrative purposes, results of the SMwST algorithm applied to the activation of c-Src kinase[42] and the multidrug transporter MsbA[58] will be briefly described.

In the first application of the SMwST algorithm, Gan and Roux investigated the molecular basis of the activation process in human tyrosine kinases.[42] The calculated transition pathway between two crystallographic structures of human tyrosine kinases[62,63] revealed a two-step activation process in which opening of the activation loop is followed by rotation of the αC helix. More extensive SMwST calculations later confirmed the main conclusions.[44] Subsequent MSM-based computations exploited the information from the string pathway to provide additional insights into the activation/deactivation mechanism of the Src kinase.[64] It is noteworthy that alternative approaches to the SMwST have also been employed to study conformational transition in kinases, including the adaptively biased path

https://doi.org/10.1021/acs.jctc.1c01049
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

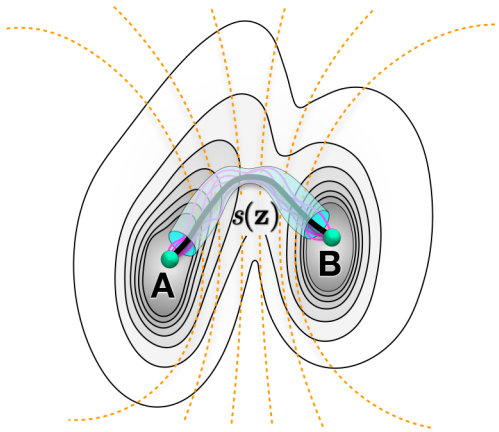


Figure 1. Localized tube connecting the end states of a geometric transition. Shaded contours represent the different free-energy levels. In the vicinity of states A and B, which are metastabilities of the free-energy landscape, the committor approaches 0 and 1, respectively. A commitment probability, $p_B(\mathbf{x})$, of 1/2 corresponds to the ensemble of transition states from which random trajectories have an equal probability to commit to state A or state B. The committor foliates the transition region between A and B into a set of so-called isocommittor surfaces characterized by $p_B(x) = \alpha$, for $\alpha \in (0, 1)$ (dashed lines), and is a convenient representation of the RC. It is generally accepted that most paths connecting states A and B lie in a tube, where isocommittors are planar and crossed by a number of trajectories, which, per unit area, corresponds to successive transition flux densities. The tube embraces a concentration of reactive trajectories (magenta), and the thick curve at its center is a zero-drift pathway (ZDP), denoted here as $s(\mathbf{z})$.

optimization (ABPO) approach for the determination of the principal curve defining a conformational transition between two known end states. The adaptively biased path optimization strategy utilizes unrestricted, enhanced sampling in the region of a path in the subspace of the CVs to identify a broad path between two stable end-states.[65] For instance, Post and co-workers turned to ABPO to examine the transitions within the catalytic domain of c-Src and CDK2 kinase.[66−68]

In a different application, the SMwST algorithm was employed to gain mechanistic insights behind the complex, large-scale conformational changes in a class of ATP-binding cassette (ABC) transporter called MsbA.[58] MsbA is involved in exporting a variety of substrates across the lipid bilayer. In order to fulfill its function this transporter undergoes large-scale structural transitions between inward- and outward-facing states. However, the static snapshots of the protein in different conformational states suggest that ATP-mediated dimerization/dissociation of nucleotide-binding domains (NBDs) is coupled to the structural changes in the transmembrane-binding domains (TMDs). The most relevant transition pathway captured from the SMwST algorithm highlighted a highly coupled NBD−TMD interaction (which follows alternating access model) and engages orientational change in the NBDs to facilitate cytoplasmic opening.

Once the optimal path is obtained through the SMwST algorithm, the associated thermodynamic properties, such as the free-energy difference between the initial and the final states, and the free-energy barrier along the string, can be computed by using enhanced-sampling methods[14,15,69] with path collective variables (PCVs)[70] or path-metadynamics variables (PMVs).[71] These thermodynamic properties are essential in practical biophysical and chemical applications, e.g., determining the binding affinity of protein−ligand complexes and the spontaneity of a chemical reaction. In addition, kinetic properties, such as transition rate and permeation coefficients, can be computed by employing models of diffusive motions.[72,73] Alternatively, nonequilibrium pulling with Jarzynski equality and Weiertrass transform[74,75] along the string also provides insight into the thermodynamic and kinetic properties and computes the system free energy. Our work presented herein examines the effectiveness of different PCVs and parameters in the computation of potential of mean force (PMF) along the string, and also discusses the determination of kinetic properties.

By exploiting developments for asynchronous communication between copies,[16] the virtual absence of overhead for threads running in parallel makes the SMwST algorithm ideally suited for distribution over large arrays of computing units. Still, the search for the optimal path by means of the iterative SMwST algorithm is computationally intensive and, depending on the complexity of the molecular objects at play and the implementation of the code, may require substantial resources, which explains why this methodology has not been hitherto applied in a more routine fashion. Furthermore, cost-effectiveness of the SMwST algorithm is subservient to the optimal choice of their different parameters—i.e., from the discretization scheme of the string to the simulation times guaranteeing a suitable estimation of the average dynamic drift, as well as the initial guess of the pathway connecting the basins of free-energy landscape, which, for the intricate string underlying large conformational transitions, can be critical.[58] Defining a set of guidelines to optimize the computational investment of SMwST calculations by choosing the most suitable parameters is, therefore, eminently desirable. In the present contribution, after outlining the theoretical underpinnings of the SMwST algorithm, we examine in a series of simulations of the well-documented prototypical alanine tripeptide[28,76] how the choice of the string parameters impacts the optimization process and ultimately the resulting path. Next, beyond SMwST calculations, we review how PMFs can be determined along the string, focusing on PCVs[70,71] and their applicability. We close on suggestions of an optimum strategy for the identification of minimum free-energy pathways, comparing alternate schemes, whereby SMwST calculations are followed by importance-sampling simulations, and multidimensional free-energy calculations are associated with a post-hoc path search. We also show how kinetic properties like diffusivities can be readily inferred from the PMF computed along the string.

2. METHODS

2.1. Theoretical Underpinnings. Following pioneering work from Pratt,[77] and Elber and Karplus,[78] the nudged elastic band[79] method was introduced nearly three decades ago as a path-finding algorithm to determine the minimum energy path (MEP) between two stable states, or metastabilities, as a chain of copies of the system of interest. The zero-temperature string method formalized these ideas by considering a chain of discrete images evolving along the potential energy gradient while keeping the length of each link identical through a reparametrization procedure.[2] It is worth noting that keeping the length of all the links along the chain identical is a necessary condition to prevent images from pooling together into the metastabilities. This procedure also allows the reaction path to remain well-resolved, especially in high-energy

https://doi.org/10.1021/acs.jctc.1c01049
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

transition regions. Maragliano et al.[3] expanded this framework to define the string on a free-energy surface in CV space, where each node evolves according to the mean force and a metric tensor that can be determined as conditional averages. To avoid calculating these average quantities, the SMwST,[4] also called the "drift method",[80] relies on a large number of short, unbiased trajectories, launched from the positions of the images distributed along the curvilinear abscissa to determine the optimal pathway. The drift method is formally equivalent to the mean force method of Maragliano et al.,[3] in the limit where the trajectories of the swarms are extremely short.[81] As discussed below, the SMwST algorithm is, however, a much simpler and scalable approach from a computational point of view.

Let us consider a molecular system of N atoms described by the atomic coordinates $\mathbf{x} \in \mathcal{R}^{3N}$ and the potential energy $U(\mathbf{x})$. The equilibrium distribution of the molecular configuration at a temperature T is given by the Boltzmann distribution:

$$p(\mathbf{x}) = \frac{e^{-\beta U(\mathbf{x})}}{\int d\mathbf{x}\, e^{-\beta U(\mathbf{x})}} \tag{1}$$

where $\beta = 1/k_B T$, in which k_B is the Boltzmann's constant. We are interested in characterizing the slow transitions between two basins corresponding to stable states defined by a set of n CVs $\mathbf{z} = \{z_1, z_2, ..., z_n\}$, such that $n \ll N$. We define $W(\mathbf{z})$, the PMF in terms of the CV $\mathbf{z}$, such that

$$e^{-\beta W(\mathbf{z})} = \frac{\int d\mathbf{x}\, \delta(\mathbf{z} - \mathbf{z}(\mathbf{x}))\, e^{-\beta U(\mathbf{x})}}{\int d\mathbf{x}\, e^{-\beta U(\mathbf{x})}} \tag{2}$$

We assume that at some temporal resolution characterized by time scale $\delta\tau$, the CVs evolve according to an overdamped Langevin dynamics:

$$z_i(\delta\tau) = z_i(0) + \sum_j \left(-\beta D_{ij}(\mathbf{z}(0)) \frac{\partial}{\partial z_j} W(\mathbf{z}(0)) + \frac{\partial}{\partial z_j} D_{ij}(\mathbf{z}(0)) \right) \delta\tau + \zeta_i(0) \tag{3}$$

where D_{ij} is the diffusion tensor and $\xi_i(0)$ is a Gaussian white noise, with $\langle \zeta_i(0) \rangle = 0$, and $\langle \zeta_i(0) \zeta_j(0) \rangle = 2 D_{ij} \delta\tau$. It should be emphasized, however, that no assumptions are made here about the underlying microscopic dynamics that govern the evolution of the atomic coordinates $\mathbf{x}(t)$, giving rise to this effective overdamped Langevin dynamics of the collective variables $\mathbf{z}$. The atomic coordinates $\mathbf{x}(t)$ may evolve according to any typical stochastic—e.g., Langevin, or deterministic—e.g., Newtonian, dynamics. Taking an average over eq 3 and using the vector notation, we have

$$\frac{\langle \delta\mathbf{z}(\delta\tau) \rangle}{\delta\tau} = -\beta \mathbf{D}(\mathbf{z}(0)) \cdot \nabla W(\mathbf{z}(0)) + \nabla \cdot \mathbf{D}(\mathbf{z}(0)) \tag{4}$$

where $\delta\mathbf{z}(\delta\tau) = \mathbf{z}(\delta\tau) - \mathbf{z}(0)$ represents the evolution of the CVs. For $\langle \delta\mathbf{z}(\delta\tau) \rangle$ to stay along a path for any $\mathbf{z}(0)$ on the path, the right-hand side of eq 4 needs to be parallel to the path, or, in brief, $(-\beta \mathbf{D} \cdot \nabla W + \nabla \cdot \mathbf{D})^{\perp} = 0$. At convergence, the images along the optimal string display a zero drift orthogonal to the path. For this reason, we call this object the zero-drift path (ZDP). The ZDP has the property that a system initiated anywhere on the path, and let free to evolve dynamically for a short period of time ($O(\delta\tau)$), will only evolve on average along the path, and not orthogonal to the path (i.e., when repeated multiple times from the same initial point). This is the most general condition for the ZDP within the overdamped Langevin assumption above. If the diffusion tensor is independent of $\mathbf{z}$, then the condition is reduced to $(\mathbf{D} \cdot \nabla W)^{\perp} = 0$. Finally, if $\mathbf{D}$ is proportional to the identity matrix (i.e., D is a scalar and not a tensor), then the optimal path from SMwST is the MFEP, i.e., $(\nabla W)^{\perp} = 0$.

Let us assume that an arbitrary path is represented by a string of M equidistant images $\{\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, ..., \mathbf{z}^{(M)}\}$, where $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(M)}$ represent states A and B, respectively, and $|\mathbf{z}^{(i)} - \mathbf{z}^{(i-1)}| = |\mathbf{z}^{(i)} - \mathbf{z}^{(i+1)}|$. An algorithm for converging an arbitrary initial string would evolve path until $\langle \delta\mathbf{z}(\delta\tau) \rangle^{\perp} \approx 0$ for any $\mathbf{z}(0) = \mathbf{z}^{(i)}$. To evolve an initial path toward the ZDP, the SMwST algorithm uses the average drift evaluated from multiple unbiased trajectories of simulation time $\delta\tau$ initiated from each image $\mathbf{z}^{(i)}$, with or without white noise, as an approximation of $\langle \delta\mathbf{z}^{(i)}(\delta\tau) \rangle$ above. For each image i, the system $\mathbf{x}$ is first restrained to keep the CVs close to the image center $\mathbf{z}^{(i)}$. The restraint is then released to generate an unbiased trajectory of length $\delta\tau$. For each image i, the above procedure can be repeated multiple times to generate multiple unbiased trajectories of length $\delta\tau$, all starting around the same image center $\mathbf{z}^{(i)}$, hence the idea of a swarm of trajectories. The restraining part could be done either in one shot for all copies of the same image as in the serial version of SMwST[4] or independently for each copy as in the parallel version of the SMwST algorithm.[16,60] In the latter, each copy is an independent simulation including a restraining and a release part, while in the former, a single biased simulation is performed first to generate a number of initial conformations to initiate multiple independent unbiased simulations. After each iteration, the new image centers are moved, either deterministically or in a randomized fashion, from $\mathbf{z}^{(i)}$ to $\mathbf{z}^{(i)} + \langle \delta\mathbf{z}^{(i)}(\delta\tau) \rangle \approx \overline{\mathbf{z}(\delta\tau)}$, where $\overline{\mathbf{z}(\delta\tau)}$ is the average collective variable position of all copies of image i at the end of the unbiased simulation of length $\delta\tau$. The image centers are then modified to satisfy the reparameterization condition to keep them equidistant. This procedure can be repeated until the image centers converge within a desired accuracy.

2.2. Potential-of-Mean-Force Calculations and Path-Collective Variables. Now that we have the ZDP, i.e., a parametrized curvilinear abscissa in the subspace of the CVs, we need to associate the latter to a progress variable, which will allow a free-energy change between the reference and the target state—i.e., metastabilities of the multidimensional free-energy landscape, to be calculated. The concept of progress variable provides a convenient framework for dimensionality reduction and the compact description of the conformational transition by means of a one-dimensional free-energy profile, or PMF, reflecting, in principle, the correct dynamics of the molecular objects. The introduction of a progress variable necessarily implies a projection of the CV or CC space onto the string that represents the average transition path, resulting in differentiable geometric expressions, along which a free-energy change can be calculated. An example of such expressions is furnished by the PCVs,[70] which are presented here in a variant of their original formulation

https://doi.org/10.1021/acs.jctc.1c01049
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

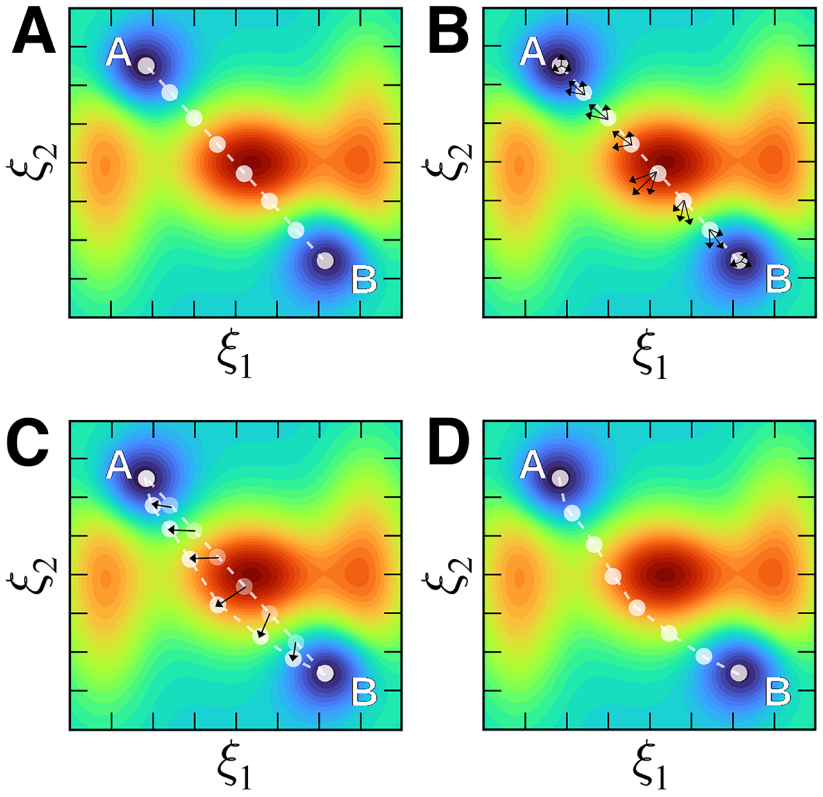


Figure 2. Schematic representation of an iteration in the SMwST. The string shown as a dashed, white line with its eight images is overlaid on top of the free-energy landscape, $\Delta G(\xi_1, \xi_2)$, determined along two CVs, ξ_1 and ξ_2. (A) Initial string generated by SMD and equilibrium simulations. (B) Forming a swarm of short and unbiased trajectories for each image. (C) Measuring the average displacements and moving the images accordingly. (D) Reparametrizing the string to make the images equidistant.

$$s(\mathbf{z}) = \lim_{\lambda\to\infty} \frac{\int_0^1 \mathrm{d}t\ t\mathrm{e}^{-\lambda f[\mathbf{z}-\mathbf{z}(t)]}}{\int_0^1 \mathrm{d}t\ \mathrm{e}^{-\lambda f[\mathbf{z}-\mathbf{z}(t)]}} \quad \text{and}$$

$$z(\mathbf{z}) = \lim_{\lambda\to\infty}\left\{-\frac{1}{\lambda}\ln\int_0^1 \mathrm{d}t\ \mathrm{e}^{-\lambda f[\mathbf{z}-\mathbf{z}(t)]}\right\} \tag{5}$$

where $\mathbf{z}$ is a vector modeling the RC, discretized with the positions of the images of the string. In the actual numerical implementation, a discretized version of eq 5 is employed.[70] The function f is a mean-square displacement (MSD), and λ is a smoothing parameter, related to the inverse of the MSD between two consecutive images. While $s(\mathbf{z})$ is a progress variable along the string, varying between 0 (state A) and 1 (state B), the ancillary variable $z(\mathbf{z})$ may be construed as the radius of a tube that embraces the string[82] and confines sampling in its vicinity. A convenient framework for the determination of the PMF, notably from a coarse approximation of the ZDP, consists of mapping the free energy in two dimensions, $s(\mathbf{z})$ and $z(\mathbf{z})$, and recovering the one-dimensional free-energy profile from the marginal distribution of $s(\mathbf{z})$.

The proposed metric, leaning on MSDs between consecutive images, proved, however, unsatisfactory in the event of conformational degeneracies. In addition, while the free energy is expected to be resilient to λ for simple curvilinear abscissas, like that of toy models,[70] complexity of the CV or CC space and crookedness of the string impede the choice of λ, inducing severe instabilities in the trajectories due to singularity of the MSD as it approaches 0. Slight deviations from the ideal value of λ results in not just sampling inefficiency, but clear signs of artifacts that render sampling outright nonphysical. Moreover, due to distinct relaxation times, finding a common value of λ for both $s(\mathbf{z})$ and $z(\mathbf{z})$ has proven challenging. These limitations have motivated the design of alternate geometric expressions,[71,76] more robust to the choice of their parameters, e.g., path-metadynamics variables (PMVs),[71] which, unlike the original PCV formulation, produce unique values of $s(\mathbf{z})$ and $z(\mathbf{z})$, and are defined as

$$\begin{cases} s(\mathbf{z}) = \dfrac{m}{M} \pm \dfrac{1}{2M}\left\{\dfrac{[(\mathbf{v}_1\cdot\mathbf{v}_3)^2 - |\mathbf{v}_3|^2(|\mathbf{v}_1|^2 - |\mathbf{v}_2|^2)]^{1/2} - (\mathbf{v}_1\cdot\mathbf{v}_3)}{|\mathbf{v}_3|^2} - 1\right\} \\ z(\mathbf{z}) = \left|\mathbf{v}_1 + \dfrac{1}{2}\left\{\dfrac{[(\mathbf{v}_1\cdot\mathbf{v}_3)^2 - |\mathbf{v}_3|^2(|\mathbf{v}_1|^2 - |\mathbf{v}_2|^2)]^{1/2} - (\mathbf{v}_1\cdot\mathbf{v}_3)}{|\mathbf{v}_3|^2} - 1\right\}\mathbf{v}_4\right| \end{cases} \tag{6}$$

where $\mathbf{v}_1 = \mathbf{s}_m - \mathbf{z}$ is a vector connecting the current position to the closest image, $\mathbf{v}_2 = \mathbf{z} - \mathbf{s}_{m-1}$ is a vector connecting the second closest image to the current position, $\mathbf{v}_3 = \mathbf{s}_{m+1} - \mathbf{s}_m$ is a vector connecting the closest image to the third closest one, and $\mathbf{v}_4 = \mathbf{s}_m - \mathbf{s}_{m-1}$ is a vector connecting the second closest image to the closest one. m and M are, respectively, the current index of the closest image and the total number of images of the string. If the current position is on the left of the closest reference image, the $\pm$ in the expression of $s(\mathbf{z})$ is a positive sign—otherwise, it is a negative sign.

https://doi.org/10.1021/acs.jctc.1c01049
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

A limitation of both sets of PCVs discussed above and those from similar approaches is the reliance on the position of images, which needs to be known prior to calculating the PCVs. If the images are ill-defined (e.g., parts of the path are cluttered with multiple images), the PCVs will be ill-defined as well. In the case of cluttered regions of the path, the numbering of images could also be another source of inaccuracies, as the image numbers may not represent the progress along the path accurately.

In the context of SMwST and other variations of the string method, the image centers are being updated iteratively. A reparametrization procedure is often used to ensure the images are equidistant along the path, and sometimes the curvature of the path is also adjusted to avoid cluttering of images, which is prone to occur in the case of high curvature. However, a more robust approach would be the use of Voronoi tessellation, where each image represents a Voronoi cell. The approach may not be ideal as an on-the-fly reparametrization for SMwST due to its computational cost. Note that in a parallel SMwST, all simulations need to stop during the reparametrization process, which is typically done in a single processor and thus imposes a considerable cost. On the other hand, the Voronoi tessellation or other parametrization methods can be used at the end of the SMwST simulations to redefine the image centers for the purpose of having well-defined PCVs for free energy calculations. An example of a centroidal Voronoi tessellation approach for finding optimized image centers from existing MD data along a given path is the so-called post-hoc string method.[60] One may for instance use the SMwST generated conformations from the post-convergence iterations to build centroidal Voronoi cells and then place the image centers at the centroids of the cells that consecutively connect state A to B.

2.3. Computational Details. *String Protocol and Convergence Assessment.* In SMwST, the path from state A to B is represented by M discretized images in either the CV space or the CC space, and these images undergo a number of iterative optimization steps. The iterative protocol has been discussed in detail elsewhere.[3,4] Here, for the sake of completeness and consistency, we summarize the iterative process into five steps as follows:

1. An initial path discretized with M images from state A to B is generated by steered molecular dynamics (SMD). Alternatively, the positions of images on the initial path can be obtained from a direct linear interpolation in CV or CC space, as shown in Figure 2A.

2. For each image on the string, a swarm of N trajectories are formed by running $\delta\tau$ unbiased simulations, as shown in Figure 2B. The displacement of the jth trajectory of ith image, $\Delta\mathbf{z}_{i,j}$, is measured and collected.

3. The average displacement of ith image and the corresponding standard deviation are computed as $\overline{\Delta\mathbf{z}_i} = \frac{1}{N}\sum_{j=1}^{N}\Delta\mathbf{z}_{i,j}$ and $\sigma(\Delta\mathbf{z}_i) = \sqrt{\overline{\Delta\mathbf{z}_i^2} - \overline{\Delta\mathbf{z}_i}^2}$, respectively. The ith image is then moved toward a random displacement that is drawn from a Gaussian distribution with a mean $\overline{\Delta\mathbf{z}_i}$, and a standard deviation $\sigma(\Delta\mathbf{z}_i)$. To be more specific, the new position of image i, $\mathbf{z}_i'$, is updated as

$$\mathbf{z}_i' = \mathbf{z}_i + \overline{\Delta\mathbf{z}_i} + \alpha\sigma(\Delta\mathbf{z}_i)X \tag{7}$$

where X is a random number drawn from the standard normal distribution $\mathcal{N}(0, 1)$, and α is a factor regulating the random fluctuations. It ought to be noted that eq 7, termed randomized updating, is different from the deterministic updating employed in the original SMwST algorithm,[4] which uses only the average displacement. Akin to simulated annealing, the introduction of α enables to search within a broader configurational space, and as the iterative process goes on, α linearly decreases to zero. This step is illustrated in Figure 2C.

4. The images are reparametrized to ensure that they are equidistant on the string, as shown in Figure 2D.

5. Each image undergoes an equilibrium simulation of time τ with its position restrained at $\mathbf{z}_i'$. The resultant string is then treated as a new initial string and goes over this iterative process from step 2 again.

The above iterative process should be continued until the convergence is reached. Assuming that a total number of K iterations are performed, and $\mathbf{z}_i^k$ refers to the ith image of the kth iteration, then the convergence can be measured by the root-mean-square deviation (RMSD) of $\mathbf{z}^k$ and a reference path $\mathbf{z}^{\mathrm{ref}}$:

$$\mathrm{RMSD}(\mathbf{z}^k, \mathbf{z}^{\mathrm{ref}}) = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(\mathbf{z}_i^k - \mathbf{z}_i^{\mathrm{ref}})^2} \tag{8}$$

If $\mathrm{RMSD}(\mathbf{z}^k, \mathbf{z}^{\mathrm{ref}})$ remains constant as k grows, or $\mathrm{RMSD}(\mathbf{z}^k, \mathbf{z}^{k+1})$ is near zero at the kth iteration, then the images on the path can be considered unchanged in the iterative process, which also indicates that there are no perpendicular movements of the images, and thus, the convergence is reached. After the convergence, the ZDP is obtained, and the PMF along it can be estimated by employing CV-based enhanced sampling methods,[5,14] such as meta-eABF,[22] with PCVs like eq 5 and eq 6. For a string optimized in CC space, the initial end points must be selected very carefully to get rid of conformations lying in an physically unrealistic high-energy space. In other words, the initial end points ought to be optimized in all degrees of freedom that the CCs involve. If this is not possible, or computationally prohibitive, it is highly recommended to transform the resultant pathway in CC space into one in CV space, and then constitute the PCVs using the corresponding CVs toward PMF calculations.

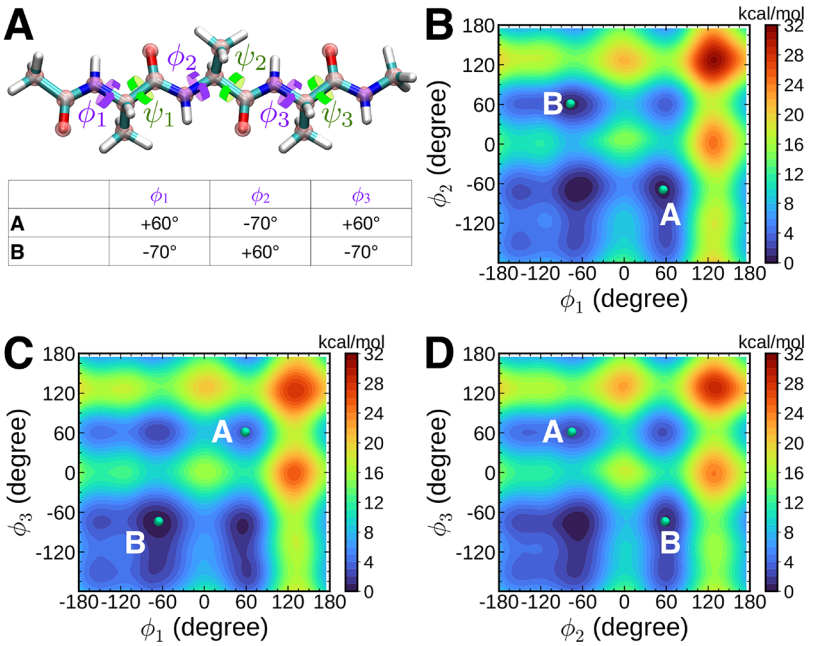
It should be noted that in theory, ZDPs are not unique. If possible, one should try to generate different ZDPs by varying the parameters of the optimization, the initial pathway, or the CVs utilized and then identify amid the PMFs determined with these ZDPs which one possesses the lowest free-energy barrier. A more accurate approach consists of computing rate constants, which, in turn, requires determining first the diffusivities, $D(s(\mathbf{z}))$. Therefore, in this work, we employed a Bayesian scheme to determine the position-dependent diffusivities from a cubic interpolated $D(s(\mathbf{z}))$,[73] and computed the rate constant, $k_{\mathrm{A}\to\mathrm{B}}$,[83] as follows:

$$k_{\mathrm{A}\to\mathrm{B}} = \left\{\int_{s_{\mathrm{A}}}^{s_{\mathrm{B}}} \mathrm{d}s \frac{\mathrm{e}^{\beta\Delta G(s)}}{D(s)} \times \int_{s_{\mathrm{A}}}^{s^*} \mathrm{d}s\ \mathrm{e}^{-\beta\Delta G(s)}\right\}^{-1} \tag{9}$$

where s_{A} and s_{B} are the PCVs of end points, and s^* is a point between s_{A} and the nearest free-energy barrier, chosen so that $\Delta G(s^*)$ is several $k_{\mathrm{B}}T$ above $\Delta G(s_{\mathrm{A}})$. The most probable transition path (MPTP) can then be determined as the ZDP that has the highest rate constant.

Application and Simulation Details. In order to examine the sensitivity of SMwST calculations on the choice of different

https://doi.org/10.1021/acs.jctc.1c01049
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX



	ϕ_1	ϕ_2	ϕ_3
A	+60°	-70°	+60°
B	-70°	+60°	-70°

Figure 3. (A) Illustration of the prechosen CVs in the CV-space string calculations, namely, the three dihedral angles (ϕ_1, ϕ_2, and ϕ_3) of the trialanine example. The heavy atoms involved in the CC-space string calculations are marked by pink bubbles. States A and B represent the starting and the ending configurations of the path, respectively. Panels (B), (C), and (D) depict the two-dimensional maps inferred from the marginal distributions of the three-dimensional free-energy landscape determined along ϕ_1, ϕ_2, and ϕ_3. The green dots denote the positions of the metastabilities A and B of the free-energy landscape.

parameters, namely, the number of images (M), unrestrained simulation time ($\delta\tau$), restrained equilibration time (τ), number of copies in a swarm (N), initial pathway, updating strategy—deterministic or randomized, and the selection of CVs, we systematically performed these simulations on a toy model, alanine tripeptide (Figure 3A). The three-dimensional PMF along the dihedral angles ϕ_1, ϕ_2, and ϕ_3, previously shown to be an important low-dimensional phase space for trialanine,[28] is mapped by the extended generalized ABF (egABF) method,[84] and its projections on (ϕ_1, ϕ_2), (ϕ_1, ϕ_3), and (ϕ_2, ϕ_3) are shown in Figure 3B,C,D, respectively. In the three-dimensional PMF, we selected two notable local minima, namely, $A(60°, -70°, 60°)$ and $B(-70°, 60°, -70°)$, and attempted to find the MPTPs from A to B via SMwST with different combinations of M, N, τ, $\delta\tau$, and updating strategy—deterministic or randomized, as shown in Table 1. The initial pathways from A to B were generated by SMD, and in CC-space simulations, different initial pathways (detailed in the Supporting Information) were explored, while the other parameters were kept identical. In order to avoid undesired translations and rotations of the capped trialanine that may break the string reparametrization, the orientation and the center-of-mass of the peptide are fixed in CC-space SMwST simulations.

The SMwST simulations were carried out by NAMD 2.14[85] while running all copies in parallel. The trialanine system was modeled by the AMBER ff14SB force field,[86] and the temperature was kept at 300 K by a Langevin thermostat. A time step of 0.5 fs was used to integrate the equations of motion for both short-range and long-range interactions.

After running the SMwST simulations, it was observed that by varying the parameters, the simulations result in different pathways, namely, A-M_1-M_3-B, A-M_2-M_3-B, A-M_1-M_4-B, A-M_2-M_6-B, and A-M_5-M_6-B as listed in Table 1. For reference purposes, the five pathways are also plotted on top of the three-dimensional free-energy landscape in Figure 4A, and the PMFs along them (shown in Figure 4B) in Figure 4A are extracted by a post-hoc path search method.[87] To examine whether the resultant pathways are exactly the MPTPs in configurational space, we also compute the PMFs with eABF[88−90] with two different sets of variables, namely, the PMVs proposed by Ensing co-workers[71] (eq 6) and the PCVs proposed by Parrinello and co-workers[70] (eq 5), and subsequently determine the rate constants defined in eq 9. To better visualize and differentiate the captured pathways, we employed a dihedral-angle principal component analysis (dPCA)[91] to the path ensembles, and projected all paths in the (ϕ_1, ϕ_2, ϕ_3) space onto the two leading principle components (PCs), as depicted in Figure 5.

3. RESULTS

3.1. Miscellaneous Aspects of the SMwST Algorithm.

Coarse-Graining of the String. To address the key question of the minimum number of images required for the successful application of SMwST, we performed independent simulations on the capped alanine tripeptide for four different discretization schemes of the string, namely, 5, 15, 25, and 35 images, with varying equilibration time and number of copies per image (Table 1). First, we gauged the convergence of these simulations by monitoring the propagation of the images, on a low-dimensional space, at each iteration of the string optimization. After 100 iterations, we do not observe any significant movement of the images perpendicular to the path, suggesting that convergence is achieved in all the simulations (see Figures S1−S3). Furthermore, to dissect the effect of adding images to the string, we calculated the PMF along the pathway obtained from the converged SMwST optimizations, using two different approaches, namely, PCVs and PMVs. As depicted in Figures 6 and 7, by adding images to the string, we observe significant improvement in the energy basins and

https://doi.org/10.1021/acs.jctc.1c01049
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Table 1. Summary of Simulation Setups and Parameters

String Type	Index	M	N	τ (ps)	$\delta\tau$ (fs)	Updating	ZDP
CV space	1	5	20	10	5	Deterministic	None
	2	5	20	50	5	Deterministic	None
	3	5	20	100	5	Deterministic	None
	4	5	100	10	5	Deterministic	None
	5	5	200	10	5	Deterministic	None
	6	15	20	10	5	Deterministic	A-M_1-M_3-B
	7	15	20	50	5	Deterministic	A-M_1-M_3-B
	8	15	20	100	5	Deterministic	A-M_2-M_3-B
	9	15	100	10	5	Deterministic	A-M_1-M_3-B
	10	15	200	10	5	Deterministic	A-M_1-M_3-B
	11	25	20	10	5	Deterministic	A-M_2-M_3-B
	12	25	20	50	5	Deterministic	A-M_2-M_3-B
	13	25	20	100	5	Deterministic	A-M_2-M_3-B
	14	35	10	10	5	Deterministic	A-M_2-M_3-B
	15	35	20	10	5	Deterministic	A-M_2-M_3-B
	16	35	20	50	5	Deterministic	A-M_2-M_3-B
	17	35	20	100	5	Deterministic	A-M_2-M_3-B
	18	35	10	10	5	Randomized	A-M_1-M_3-B
	19	35	20	10	1	Randomized	A-M_1-M_4-B
	20	35	20	10	5	Randomized	A-M_1-M_3-B
	21	35	20	10	10	Randomized	A-M_1-M_3-B
	22	35	20	10	50	Randomized	Loop
	23	35	20	10	100	Randomized	Loop
	24	35	20	50	5	Randomized	A-M_1-M_3-B
	25	35	20	100	5	Randomized	A-M_1-M_3-B
	26	35	50	10	5	Randomized	A-M_1-M_3-B
CC space	27	25	20	10	10	Deterministic	A-M_2-M_3-B
	28	25	20	10	10	Deterministic	A-M_2-M_3-B
	29	25	20	10	10	Deterministic	A-M_2-M_3-B
	30	35	20	10	10	Deterministic	A-M_2-M_3-B
	31	35	20	10	10	Deterministic	A-M_2-M_3-B
	32	35	20	10	10	Deterministic	A-M_2-M_3-B
	33	25	20	10	10	Randomized	A-M_2-M_3-B
	34	25	20	10	10	Randomized	A-M_5-M_6-B
	35	25	20	10	10	Randomized	A-M_2-M_6-B
	36	35	20	10	10	Randomized	A-M_2-M_3-B
	37	35	20	10	10	Randomized	A-M_5-M_6-B
	38	35	20	10	10	Randomized	A-M_2-M_6-B

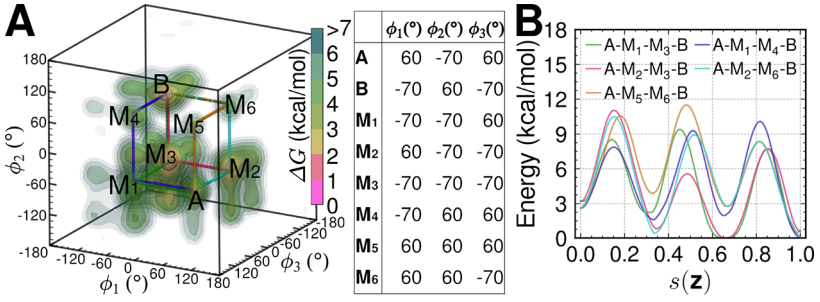


Figure 4. (A) Three-dimensional free-energy landscape of trialanine determined along torsional angles ϕ_1, ϕ_2, and ϕ_3. The local minima are marked as A, B, and M_1 through M_6, respectively. The positions of the local minima in the (ϕ_1, ϕ_2, ϕ_3) space are listed on the right-hand side of the three-dimensional free-energy landscape. (B) Free-energies along the five pathways extracted from the three-dimensional PMF in (A): A-M_1-M_3-B (green), A-M_2-M_3-B (pink), A-M_1-M_4-B (blue), A-M_2-M_6-B (cyan), and A-M_5-M_6-B (orange).

barriers (simulations 1, 6, 11, and 15 in Figure 6). More specifically, 5 images resulted in a very coarse PMF, precluding clean demarcation of the intermediate conformations of the peptide chain. Moreover, although the PMFs of simulations 6–10 in Figure 6 correctly feature the relevant free-energy minima, marked deviations between the PCVs and PMVs imply that using 15 images may still be suboptimal. We conclude that for the capped alanine tripeptide, both 25 and 35 images constitute reasonable choices. In addition, as may be observed from Table 1, it is found that varying the number of

https://doi.org/10.1021/acs.jctc.1c01049
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

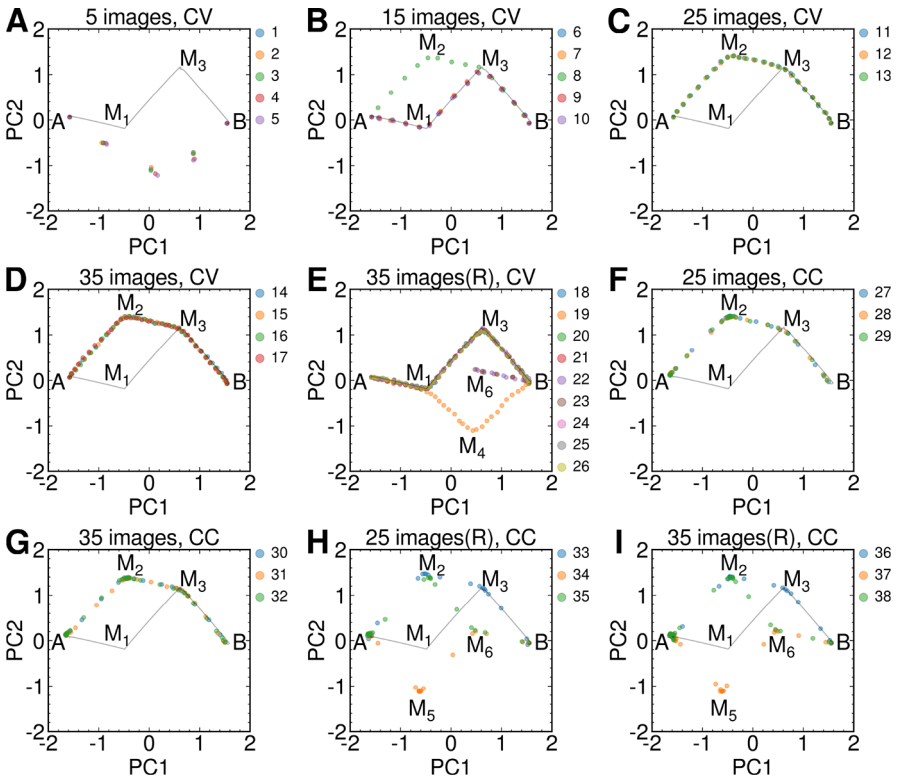


Figure 5. Images of the strings of the last iteration projected to the leading principle components (PC1 and PC2). The annotations A, B, and M_1−M_6 in the subfigures correspond to the local minima in Figure 4. The gray line and the numbered dots shown with different colors represent, respectively, the MFEP, as determined from a post-hoc path search of the three-dimensional free-energy landscape,[87] and the pathways identified from the simulations detailed in Table 1.

images may result in different ZDPs. Simulations 6 and 7, which rest on 15 images, lead to pathway A-M_1-M_3-B, while simulations 11 and 12, which rest on 25 images, lead to pathway A-M_2-M_3-B. Interestingly enough, the different PMFs generated from the SMwST optimizations suggest that the conformational transition of the capped alanine tripeptide obeys a three-step mechanism with an energetic difference of ∼3 kcal/mol between the two end states. This difference can be ascribed, at least in part, to the formation of main-chain hydrogen bonds (see Figure S4A,B). We also observe that the PMFs determined using eq 5 are somewhat more sensitive to the coarseness of the string, specifically near the two end states, A and B. This observation might be rooted in the singularity of the MSD in these regions.

Number of Copies. We also examined the effect of varying the number of copies per image by performing SMwST calculations with 20, 100, and 200 copies per image. These simulations were performed for both 5- and 15-image strings, with τ = 10 ps per image (i.e., simulations 3−5 and 8−10 in Table 1). Although the simulations using 5 images only result in very coarse pathways, convergence is, indeed, faster when using more copies, as shown in Figure S1C,D. Furthermore, comparing simulations 6 (20 copies), 9 (100 copies), and 10 (200 copies) in Table 1, increased convergence rate is also observed for the 15-image SMwST simulations, as shown in Figure S1F, I, and J. Considering that the SMwST is actually a path-optimizing process in a free-energy landscape, a possible reason for the improvement of convergence is that when more copies per image are used, the average displacement of each image is closer to the optimal direction toward the ZDP.

Effects of the Unrestrained MD Simulation Time to Form the Swarm ($\delta\tau$). The length of the short, unbiased trajectories forming the swarm, $\delta\tau$, is one of the most distinct parameters of the SMwST algorithm. It has been shown recently that $\delta\tau$ should be chosen to ensure that the CVs evolve according to Markovian dynamics.[92] While the present work does not focus on this fundamental aspect of $\delta\tau$, we have performed a series of 35-image, 20-copy, and 10-ps SMwST simulations in CV space, and the value of $\delta\tau$ was varied from 1 to 100 fs, as detailed in Table 1 (simulations 15−19) to evaluate the effects of this parameter in practice. Simulation 19, which uses a shorter $\delta\tau$ than simulations 20−23, lead to a final pathway different from A-M_1-M_4-B. Its rate of convergence is also slower than that of simulations 20−23, as shown in Figure S2C. In addition, we find that simulations 22 and 23 actually result into a folded pathway with loops like A-M_1-M_3-B-M_3-B-M_6-B, as depicted in Figure 5E. A possible reason for this artifact is that for a large enough $\delta\tau$, the images at the middle of the string evolve quickly and aggregate in local minima such as B, M_3, and M_6. Moreover, we observe that this process can occur in a nonsequential fashion, resulting in the formation of loops. For instance, images $\mathbf{z}_i$ and $\mathbf{z}_{i+k}$ share the same local minimum. Consequently, if eq 7 is used to update the positions of the new images, $\delta\tau$ ought to be as short as a few femtoseconds. Theoretically, the images lying on the pathway and their neighborhood can be regarded as vertices on an undirected graph, and the artifactual loops or cycles can be detected and removed by forming a tree using various path-searching algorithms.[93] Another strategy to detect and remove loops consists of finding the nearest image of image i and remove the intermediate images between the nearest image

https://doi.org/10.1021/acs.jctc.1c01049
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

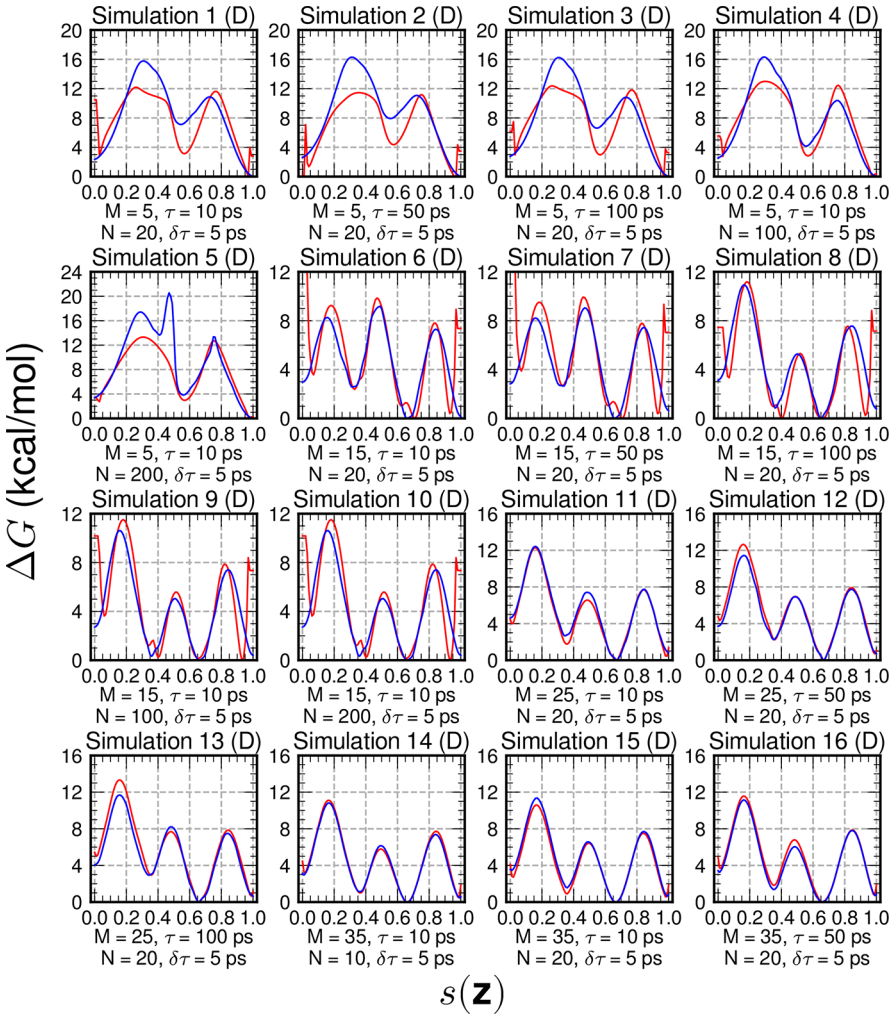


Figure 6. PMFs along the PCV and PMV of simulations 1−16 in Table 1. The red and blue lines correspond the PCV defined in eq 5 and the PMV defined in eq 6, respectively.

and image $i + 1$ or image $i - 1$. Alternatively, one may consider scaling the drifts by a small factor, k, between 0 and 1, to slow down the evolution of the drifts, thereby avoiding spurious loops. In other words, the positions of the different images could be updated as

$$\mathbf{z}_i' = \mathbf{z}_i + k\overline{\Delta \mathbf{z}_i} + \alpha\sigma(k\Delta \mathbf{z}_i)X \tag{10}$$

This strategy, however, may greatly slow down convergence. To summarize, convergence of finding a ZDP by the SMwST can be accelerated using a large $\delta\tau$, but in doing so, we also increase the probability to encounter loops in the final pathway, especially for rugged free-energy landscapes featuring multiple minima. Nevertheless, this issue could be safely avoided by turning to loop-removing algorithms, which are detailed in the Supporting Information. After running the simulations, the standard deviations of the drifts for all images were averaged over all iterations and plotted in Figure S1. To further analyze the drifts, their variances associating different $\delta\tau$ were also averaged over all iterations and images. As shown in Figure S6, the average variances grow linearly with respect to $\delta\tau$.

Effects of Randomization during Position Update. Traditionally, the SMwST algorithm only updates in a deterministic manner the positions of the different images of the path on the basis of the average drifts, that is

$$\mathbf{z}_i' = \mathbf{z}_i + \overline{\Delta \mathbf{z}_i} \tag{11}$$

which is precisely how we proceeded for simulations 1−17 and 27−32 (see Table 1). In this contribution, we explored a randomized updating strategy embodied in eq 7, which takes into account the standard deviations of the copies of the different images. This randomization strategy was employed in 35 image SMwST optimizations in CV space, as well as 25-image and 35-image SMwST optimizations in CC space, corresponding to simulations 18−26 and 33−38 (see Table 1). The PMFs along the pathways from these simulations are shown in Figures 7 and 8. As shown in the calculation of the rate constants below, the new ZDP, A-M_1-M_3-B, found by means of the randomized updating is, indeed, the MPTP. Comparing the group of simulations without (simulations 14−17, 27−32) and with randomization (simulations 18, 20, 24, 25, 33−38), we observe that in either CC-space or CV-space SMwST simulations, the introduction of white noise as the positions of the images are updated broadens the search in configurational space, induces diversity in the generated ZDPs, and, thus, increases the probability of identifying the MPTP.

Effects of the Restrained MD Equilibration Time. A potentially important parameter of the SMwST is the simulation time, τ, needed to equilibrate each image after the reference positions of the CVs have been updated and the

https://doi.org/10.1021/acs.jctc.1c01049
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

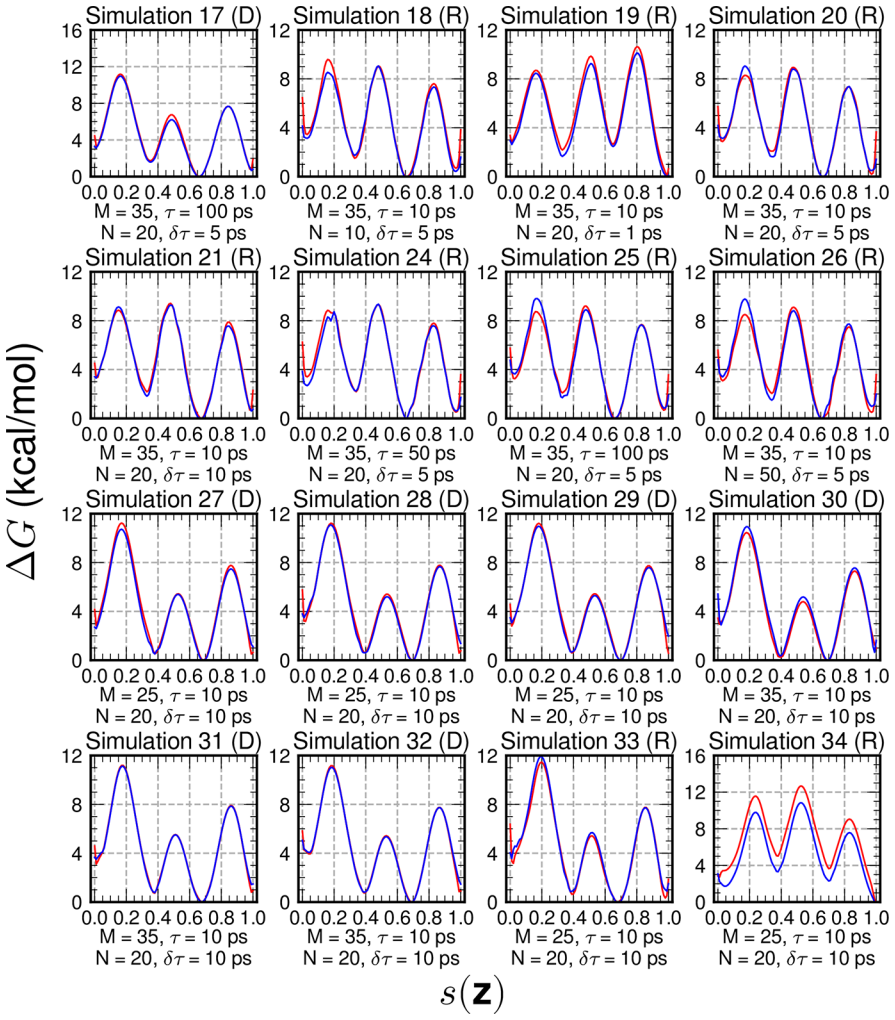


Figure 7. PMFs along the PCV and PMV of simulations 17−21 and 24−34 in Table 1. The red and blue lines correspond the PCV defined in eq 5 and the PMV defined in eq 6, respectively.

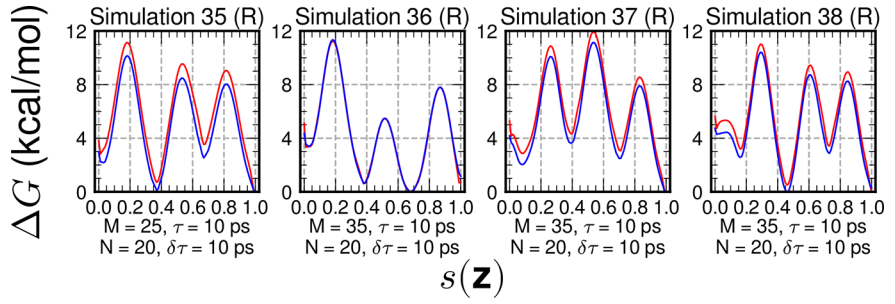


Figure 8. PMFs along the PCV and PMV of simulations 35−38 in Table 1. The red and blue lines correspond the PCV defined in eq 5 and the PMV defined in eq 6, respectively.

string has been reparametrized. To address this point, we performed additional SMwST optimizations, varying τ, namely, 10, 50, and 100 ps, with a fixed number of copies per image, referred to as simulations 1−3 (5 images), 6−8 (15 images), 11−13 (25 images), and 15−17 (35 images). Although no significant difference is observed between the different PMFs upon increase of the equilibration time, we, nonetheless, identify two pathways, namely, A-M_1-M_3-B and A-M_2-M_3-B, when the string is discretized with 15 images (see simulations 6−10 in Figure 6). Conversely, when the discretization scheme involves 25 images, all the ZDPs go from A to B via the A-M_2-M_3-B pathway. It is worth noting that our test system, trialanine, is small and can be equilibrated within a short period of time, so that varying τ has little impact on the generated pathways. In SMwST simulations of larger, more realistic biomolecular objects, a longer τ may, however, be required to update the images to the correct new positions.

3.2. String Optimizations in CC and CV Spaces. We have observed that the images of the pathways resulting from CC-space SMwST optimizations tend to gather in the local minima of the three-dimensional CV space (see Figure S7 and Figure 5F−I). Yet, in CC space, the images remain

https://doi.org/10.1021/acs.jctc.1c01049
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

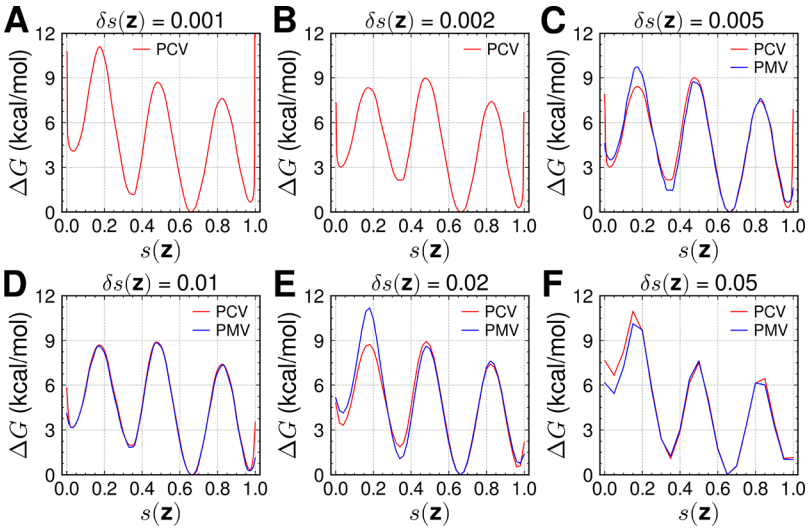


Figure 9. Free-energy landscapes along $s(\mathbf{z})$ with different choices of bin width $\delta s(\mathbf{z})$, which determines how fine-grained in eABF calculations. In (A) and (B), the eABF simulations using the bin widths of 0.001 and 0.002 with the PMV resulted in severe instabilities of the trajectory, and the corresponding PMFs are, therefore, not plotted.

equidistantly separated from each other. Such is not the case in three-dimensional space, owing to the higher degeneracy around the minima. Poor description of the free-energy landscape by means of the PMF calculation stems from the exaggerated number of images occupying the same region of the "intrinsic manifold"—i.e., the intrinsic CV space containing the slow degrees of freedom, (ϕ_1, ϕ_2, ϕ_3) being an approximation of it. In other words, the CC space is a bad approximation of the latter. Consequently, to determine the PMF along a pathway resulting from a CC-space SMwST optimization, it may still be necessary to find a proper set of CVs to transform the pathway in CV space prior to calculating the PMF along the PCV or PMV. In this contribution, we transformed the pathways from CC-space SMwST simulations to the (ϕ_1, ϕ_2, ϕ_3) space, and reparametrized them again as a preamble to the PMF calculations.

Furthermore, it ought to be noted that in CC-space SMwST optimization, string reparametrization requires measuring the optimal RMSDs between consecutive images. In theory, this step can be achieved following two different strategies, namely, (i) aligning structures on-the-fly during reparametrization, or (ii) fixing the orientation and center-of-mass of the system to avoid undesired rotations and translations. These strategies, however, cannot guarantee a perfect reparametrization, that is, the best fits, or optimal RMSDs, between images being approximately equivalent. Consequently, for small, paradigmatic biological objects like trialanine, reparametrization is markedly affected by roto-translational invariance, which may also introduce artifacts into the resultant pathways, inducing ruggedness. This limitation of the reparametrization step performed in CC-space is detailed in the Supporting Information. Conversely, for larger, more realistic biological objects, such as ATPase,[54,55] this may not be the case. As a result, it is recommended to employ SMwST optimization in CV space whenever a suitable set of CVs can be identified. If the molecular object at hand is, however, excessively complex, and the CVs describing the conformational transition of interest cannot be readily identified, SMwST optimization in CC space may be the only viable choice available.

3.3. Determining the PMF along the String.

A key difference between eq 6 and eq 5 is that the former is constructed in a geometric way, without the need of the auxiliary variable λ, thereby simplifying the setup of the simulations. Nevertheless, one may argue that there is a useful rule of thumb for choosing λ, for example, the inverse of the mean squared displacement between successive images,[70] namely,

$$\frac{1}{\lambda} \simeq \frac{1}{M-1}\sum_{i=1}^{M-1}(\mathbf{z}_{i+1}-\mathbf{z}_i)^2 \tag{12}$$

In the case of our toy model of a capped alanine tripeptide, setting λ according to eq 12 is perfectly appropriate. However, for biological objects involving complex conformational transition, whereby the path is defined in a CV space of higher dimensionality, the use of eq 12 may not be the optimal choice for setting λ. Said differently, although in our application, both the PCV and PMV definitions work well, robustness of λ in PCV-based free-energy calculations remains a matter of concern, most notably when the biological process at hand corresponds to very intricate pathways.

Aside from λ, another parameter that affects the PMF calculation along PCVs[70] or PMVs[71] is the discretization of the CV $s(\mathbf{z})$, which is determined by $\delta s(\mathbf{z})$, the width of the bins, wherein the instantaneous collective forces are accrued along $s(\mathbf{z})$. While lowering the value of $\delta s(\mathbf{z})$ increases the resolution of the PMF, it also reduces the efficiency of its computation, since longer simulation time is required to collect the instantaneous forces in every bin along the pathway. Moreover, to keep in an eABF free-energy calculation the extended CV synchronized with $s(\mathbf{z})$, in approximately the same bin, a stiffer spring connecting the extended CV to $s(\mathbf{z})$ is required, which, in turn, may render the simulation unstable. Based on the pathway obtained from simulation 20 in Table 1, we ran PMF calculations along PMV and PCV, by varying $\delta s(\mathbf{z})$ from 0.001 to 0.05. As shown in Figure 9A, the simulation crashes when using the PMV in eq 6 with low $\delta s(\mathbf{z})$. Conversely, in Figure 9E and F, the free-energy landscapes obtained with a high $\delta s(\mathbf{z})$ are coarse due to an inadequate discretization of $s(\mathbf{z})$. In summary, as shown in Figure 9C, balancing computational

https://doi.org/10.1021/acs.jctc.1c01049
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

efficiency and accuracy, a value of 0.01 should constitute a reasonable choice for $\delta s(\mathbf{z})$, which was then used for all PMF calculations in Figures 6−8.

3.4. Determining the MPTP amid Candidate ZDPs. Although Figures 6 and 7 indicate that the PMFs along the pathway A-M_1-M_3-B possess the lowest free-energy barrier, which can also be verified in Figure 4, the possibilities of the five pathways happening in the transformation from A to B remain unknown. To solve this issue, following PCV- or PMV-based free-energy calculations, the CV-dependent diffusivity along $s(\mathbf{z})$ was determined by means of a Bayesian inference scheme (see Figure S8).[73] In a nutshell, knowing the probability of the trajectory, given the force acting along $s(\mathbf{z})$ and the diffusivity, we sought the probability of these quantities given the trajectory generated in the course of the PMF calculation. Next, the rate constants for the different ZDPs were computed with eq 9, and listed in Table 2. It can be

Table 2. Rate Constants Determined from the Different ZDPs Identified by the SMwST for Trialanine, and Using Two Metrics for the Computation of the PMFs along the Pathways

	$k_{A\to B}$ (ns^{-1})	
ZDP	PCV[70]	PMV[71]
A-M_1-M_3-B	3.62×10^{-2}	4.10×10^{-2}
A-M_2-M_3-B	2.04×10^{-3}	1.55×10^{-3}
A-M_1-M_4-B	2.52×10^{-3}	1.54×10^{-3}
A-M_2-M_6-B	7.40×10^{-4}	8.71×10^{-4}
A-M_5-M_6-B	1.01×10^{-4}	9.59×10^{-5}

observed that the ZDP going from A to B via the free-energy minima M_1 and M_3 has the largest rate constant and, therefore, corresponds to the shortest mean first-passage time (MFPT),[83,94] which is equal to the inverse of $k_{A\to B}$. Although the rate constants computed in light of PCV- and PMV-based PMF calculations are somewhat quantitatively different, the ZDP A-M_1-M_3-B is likely the MPTP that we want.

4. CONCLUSION

In this contribution we have examined with a critical eye the merits and the limitations of the SMwST, a method designed for identifying likely transition pathways between distinct conformational states in a broad range of molecular objects undergoing complex structural transitions. While SMwST undoubtedly represents a powerful tool for the investigation of conformational equilibria by finding a probable pathway—the string—connecting well-characterized basins of high-dimensional free-energy landscapes, it must be emphasized that this pathway is not necessarily the MPTP, but rather one plausible candidate amid a number of other likely pathways sharing the common property of being devoid of an average drift. Said differently, what a suitably converged SMwST optimization supplies is a ZDP, which, for a given conformational transition, is not unique. Among the different ZDPs associated to this conformational transition, one is the MPTP, identifiable from the corresponding rate constants describing the diffusion between the two free-energy minima of interest. We further note that the MFEP is a ZDP parallel to the gradient of the PMF, i.e., literally the gradient of the PMF with respect to the CVs, and may not necessarily always coincide with the MPTP.

Though conceptually simple in its formulation and practical implementation, the SMwST rests on a number of assumptions and parameters generally underappreciated, or overlooked, and that ought to be considered with great care. Most important, a meaningful optimization of the transition pathway is subservient to the appropriate choice of three highly critical parameters:

a. Definition of the CV space. This is the most crucial aspect to obtain a physically meaningful transition pathway in any application of the string method. The set of CVs ought to be chosen with care, so that they can describe the conformational transition of interest with suitable accuracy. On the other hand, if the dimensionality of the CV is excessive, this can burden convergence of the string optimization, and potentially lead to the wrong ZDP, as will be discussed in greater detail hereafter.
b. The number of images, M, along the string. If M is too small, the pathway will be oversimplified—or ill-defined. Conversely, if M is too large, the string may be prone to generate loops and spurious fluctuations.
c. The unrestrained simulation time, $\delta\tau$. This is one of the key parameters underlying the assumptions of the SMwST algorithm. Fundamentally, $\delta\tau$ should be chosen such that the dynamics of the CVs is truly Markovian.[92] In practical applications of the SMwST algorithm, this parameter can also impact the string optimization in various ways. On the one hand, in a free-energy landscape featuring multiple minima, loops may form if $\delta\tau$ is too large, and in this case, turning to a loop-removing algorithm employed at the reparametrization stage becomes necessary. On the other hand, if $\delta\tau$ is too small, progress at each iteration may be small, and, as a result, convergence would be slowed down.

In addition, the computed transition pathway depends also on a number of parameters that are important in practice

d. The restraining force constant, k, to tether the images of the discretized pathway. As depicted in Figure S9, if k is too small, the different images may not coincide with the desired position prior to the drifting process, resulting in a lack of accuracy. Conversely, too large a value of k may entail trajectory instabilities, and even possible distortions of the molecular object at hand.
e. The restrained simulation time, τ, of the swarms of trajectories. If τ is too small, here again, an offset of the different images from the desired position may occur. In addition, equilibration prefacing the drifting process may be suboptimal, and likely to propagate throughout the string optimization.
f. The number of trajectories, N, generated in the swarms to determine the drift along the string. The accuracy and the computational cost increase as N gets larger. On the other hand, if N is too small, convergence will be slowed or impeded due to noise.
g. Choice of the initial—or guess string. Ideally, the choice of the initial string should have minimal bearing on the SMwST algorithm. However, in particular for large molecular objects, should the guess pathway be too far from the actual ZDP, optimization could be very costly, as will be also discussed further hereafter.
h. Randomized updating. Randomizing the update of the new positions of images can help optimize the string in a

https://doi.org/10.1021/acs.jctc.1c01049
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

larger configurational space, and, thus, lead to different resultant ZDPs, which increases the probability of finding the correct MPTP.

The above parameters are also inter-related with some deep ramifications as to which ZDP the SMwST optimization might converge to—and how fast convergence will be reached. For instance, the selection of coarse variables, and, hence, the definition of the CV space necessarily impact which ZDP is identified by the algorithm, as illustrated in the comparison of the pathways generated by the CC-based and dihedral-based path optimizations detailed in Table 1. However rudimentary, the toy model used herein, consisting of a terminally capped alanine tripeptide, provides a cogent demonstration of how different parameters can change the outcome of a SMwST optimization. Although the choice of the initial pathway may be of lesser relevance for a tripeptide, and was, thus, only marginally discussed here, it is particularly important for larger biological objects undergoing intricate conformational transitions. It is reasonable to assume that under these premises, the SMwST algorithm will in most cases find the closest ZDP to the initial string.

An assessment of convergence constitutes a critical aspect of SMwST simulations. Once a pathway has been optimized, its characterization remains incomplete without the associated PMF calculation along the identified ZDP. As has been discussed in detail here, this free-energy calculation can be performed, employing a well-suited importance-sampling algorithm in association with PCVs or PMVs. The choice of the CVs that define the PCV is not without consequences. We may choose the same CVs that were utilized in the SMwST optimization, or other CVs to define PCVs. For instance, as we indicate it here, it is possible, and sometimes even advantageous, to turn to dihedral-based PCVs to perform the PMF calculation, even though the pathway has been optimized in CC space. Moreover, although both PCVs and PMVs proved satisfactory for the determination of the PMF along the pathways identified in our study, the PCVs require an additional parameter, namely, λ, the value of which may need to be tuned for intricate pathways associated to complex biological processes.

The PMF per se is, however, insufficient to not only ascertain whether the SMwST optimization has converged appropriately, but also characterize the generated ZDP. As a function of the topology of the computed one-dimensional free-energy profile, a committor analysis may not be always feasible. As an alternative, employing Bayesian inferences, it is possible to determine the CV-dependent diffusivity, from whence the MFPT and, hence, the rate constant can be obtained. The ZDP that corresponds to the largest rate constant, or the shortest MFPT, is the MPTP. Ideally, it would be desirable to replicate the SMwST optimization, using, for instance, distinct guess strings and distinct noises, as a way to identify different ZDPs in high-dimensional CV or CC space, and pinpoint which, amid the ensemble of plausible pathways, is the actual MPTP. While this strategy represents the best practice for the discovery of transition pathways using SMwST simulations, it is most unfortunately not amenable to the investigation of complex conformational changes in large biological objects, owing to the prohibitive cost of one string optimization, which ordinarily marshals significant computational resources. Under these premises, one is limited to a single SMwST optimization and left to speculate that the generated ZDP constitutes a reasonable approximation of the MPTP.

In the arsenal of algorithms developed in recent years to treat processes involving very slow conformational changes in molecular objects, the SMwST algorithm constitutes an appealing option owing to its conceptual simplicity and amenability to very high-dimensional CV or CC space. Because the algorithm is embarrassingly parallelizable and scalable, the search of transition pathways connecting metastable states on a complex free-energy landscape can be distributed onto large arrays of computing units, resulting in rapid optimization of the trial string. The strength of the algorithm lies not only in its ability to successfully yield physically plausible pathways in very large and complex biomolecular systems,[42−61] but also in the possibility to combine it with other methods such as milestoning,[10,47,95] weighted-ensemble simulations,[96] or MSMs,[64,97] thereby helping advance our mechanistic insights into these systems. Our hope is that the set of guidelines compiled in the present contribution will help deepen our understanding of the algorithm to gain efficiency in the preparation, practical application, and post-hoc analysis of SMwST simulations of intricate conformational transitions.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.1c01049.

Generation of initial pathways in the CC-space SMwST, details of elimination of loops, detailed discussion of implementing reparametrization in CC-space with roto-translational invariance, projections of pathways in principle components, average variances of drifts in simulations of different $\delta\tau$, standard deviations of drifts averaged over iterations, PMFs and CV-dependent diffusivity profiles along ZDPs, and pathways from the CC-space SMwST simulations of different restraining force constants (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Christophe Chipot − *Laboratoire International Associé Centre National de la Recherche Scientifique et University of Illinois at Urbana−Champaign, Unité Mixte de Recherche no 7019, Université de Lorraine, 54506 Vandœuvre-lès-Nancy, France; Theoretical and Computational Biophysics Group, NIH Center for Macromolecular Modeling and Bioinformatics, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana−Champaign, Urbana, Illinois 61801, United States; Department of Physics, University of Illinois at Urbana−Champaign, Urbana, Illinois 61801, United States;* orcid.org/0000-0002-9122-1698; Email: chipot@illinois.edu

Authors

Haochuan Chen − *Research Center for Analytical Sciences, College of Chemistry, Tianjin Key Laboratory of Biosensing and Molecular Recognition, Nankai University, Tianjin 300071, China; Theoretical and Computational Biophysics Group, NIH Center for Macromolecular Modeling and Bioinformatics, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana−Champaign,*

https://doi.org/10.1021/acs.jctc.1c01049
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Urbana, Illinois 61801, United States; Laboratoire International Associé Centre National de la Recherche Scientifique et University of Illinois at Urbana−Champaign, Unité Mixte de Recherche no 7019, Université de Lorraine, 54506 Vandœuvre-lès-Nancy, France; orcid.org/0000-0001-6447-1096

Dylan Ogden − *Department of Chemistry and Biochemistry, University of Arkansas, Fayetteville, Arkansas 72701, United States*

Shashank Pant − *Theoretical and Computational Biophysics Group, NIH Center for Macromolecular Modeling and Bioinformatics, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana−Champaign, Urbana, Illinois 61801, United States;* Present Address: Loxo Oncology @ Lilly, Louisville, CO 80027, USA; orcid.org/0000-0002-8222-3616

Wensheng Cai − *Research Center for Analytical Sciences, College of Chemistry, Tianjin Key Laboratory of Biosensing and Molecular Recognition, Nankai University, Tianjin 300071, China;* orcid.org/0000-0002-6457-7058

Emad Tajkhorshid − *Theoretical and Computational Biophysics Group, NIH Center for Macromolecular Modeling and Bioinformatics, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana−Champaign, Urbana, Illinois 61801, United States; Department of Biochemistry and Center for Biophysics and Quantitative Biology, University of Illinois at Urbana−Champaign, Urbana, Illinois 61801, United States;* orcid.org/0000-0001-8434-1010

Mahmoud Moradi − *Department of Chemistry and Biochemistry, University of Arkansas, Fayetteville, Arkansas 72701, United States;* orcid.org/0000-0002-0601-402X

Benoît Roux − *Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, Illinois 60637, United States;* orcid.org/0000-0002-5254-2712

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.1c01049

Notes

Shashank Pant is currently an employee of Loxo Oncology @ Lilly and is a shareholder of stock in Eli Lilly and Co. The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This study was supported by the National Natural Science Foundation of China (22073050), the US National Institutes of Health (R15-GM139140 and P41-GM104601), the National Science Foundation (MCB-1517221 and CHE-1945465), the France and Chicago Collaborating in The Sciences (FACCTS) program, and the Agence Nationale de la Recherche (ProteaseInAction). S.P. would like to thank Beckman Institute for Graduate Fellowship.

■ REFERENCES

(1) Shaw, D.; Deneroff, M.; Dror, R.; Kuskin, J.; Larson, R.; Salmon, J.; Young, C.; Batson, B.; Bowers, K.; Chao, J.; Eastwood, M.; Gagliardo, J.; Grossman, J.; Ho, C.; Ierardi, D.; Kolossváry, I.; Klepeis, J.; Layman, T.; McLeavey, C.; Moraes, M.; Mueller, R.; Priest, E.; Shan, Y.; Spengler, J.; Theobald, M.; Towles, B.; Wang, S. Anton, a Special-Purpose Machine for Molecular Dynamics Simulation. *SIGARCH Comput. Archit. News* **2007**, *35*, 1−12.

(2) E, W.; Ren, W.; Vanden-Eijnden, E. String Method for the Study of Rare Events. *Phys. Rev. B* **2002**, *66*, 052301.

(3) 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*, 024106.

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

(5) Chipot, C.; Pohorille, A., Eds. *Free Energy Calculations. Theory and Applications in Chemistry and Biology*; Springer Verlag: Berlin, Heidelberg, NY, 2007.

(6) Darve, E.; Pohorille, A. Calculating Free Energies Using Average Force. *J. Chem. Phys.* **2001**, *115*, 9169−9183.

(7) Comer, J.; Gumbart, J. C.; Hénin, J.; Lelièvre, T.; Pohorille, A.; Chipot, C. The Adaptive Biasing Force Method: Everything You Always Wanted to Know, But Were Afraid to Ask. *J. Phys. Chem. B* **2015**, *119*, 1129−1151.

(8) Laio, A.; Parrinello, M. Escaping Free Energy Minima. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562−12565.

(9) Torrie, G. M.; Valleau, J. P. Monte Carlo Study of Phase Separating Liquid Mixture by Umbrella Sampling. *J. Chem. Phys.* **1977**, *66*, 1402−1408.

(10) Faradjian, A. K.; Elber, R. Computing Time Scales from Reaction Coordinates by Milestoning. *J. Chem. Phys.* **2004**, *120*, 10880−10889.

(11) Votapka, L. W.; Jagger, B. R.; Heyneman, A. L.; Amaro, R. E. SEEKR: Simulation Enabled Estimation of Kinetic Rates. A Computational Tool to Estimate Molecular Kinetics and Its Application to Trypsin-Benzamidine Binding. *J. Phys. Chem. B* **2017**, *121*, 3597−3606.

(12) Huber, G. A.; Kim, S. Weighted-Ensemble Brownian Dynamics Simulations for Protein Association Reactions. *Biophys. J.* **1996**, *70*, 97−110.

(13) Teo, I.; Mayne, C. G.; Schulten, K.; Leliévre, T. Adaptive Multilevel Splitting Method for Molecular Dynamics Calculation of Benzamidine-Trypsin Dissociation Time. *J. Chem. Theory Comput.* **2016**, *12*, 2983−2989.

(14) Lelièvre, T.; Stoltz, G.; Rousset, M. *Free Energy Computations: A Mathematical Perspective*; Imperial College Press, 2010.

(15) Abrams, C.; Bussi, G. Enhanced Sampling in Molecular Dynamics Using Metadynamics, Replica-Exchange, and Temperature-Acceleration. *Entropy* **2014**, *16*, 163−199.

(16) Jiang, W.; Phillips, J.; Huang, L.; Fajer, M.; Meng, Y.; Gumbart, J. C.; Luo, Y.; Schulten, K.; Roux, B. Generalized Scalable Multiple Copy Algorithms for Molecular Dynamics Simulations in NAMD. *Comput. Phys. Commun.* **2014**, *185*, 908−916.

(17) Minoukadeh, K.; Chipot, C.; Lelièvre, T. Potential of Mean Force Calculations: A Multiple-Walker Adaptive Biasing Force Approach. *J. Chem. Theory Comput.h* **2010**, *6*, 1008−1017.

(18) Comer, J.; Phillips, J.; Schulten, K.; Chipot, C. Multiple-Replica Strategies for Free-Energy Calculations in NAMD: Multiple-Walker Adaptive Biasing Force and Walker Selection Rules. *J. Chem. Theory Comput.* **2014**, *10*, 5276−5285.

(19) Wang, Y.; Ribeiro, J. M. L.; Tiwari, P. Past−Future Information Bottleneck Framework for Sampling Molecular Reaction Coordinate, Thermodynamics and Kinetics. *Nat. Commun.* **2019**, *10*, 3573.

(20) Smith, Z.; Pramanik, D.; Tsai, S.-T.; Tiwary, P. Multi-Dimensional Spectral Gap Optimization of Order Parameters (SGOOP) through Conditional Probability Factorization. *J. Chem. Phys.* **2018**, *149*, 234105.

(21) Pant, S.; Smith, Z.; Wang, Y.; Tajkhorshid, E.; Tiwary, P. Confronting Pitfalls of AI-Augmented Molecular Dynamics Using Statistical Physics. *J. Chem. Phys.* **2020**, *153*, 234118.

(22) Fu, H.; Zhang, H.; Chen, H.; Shao, X.; Chipot, C.; Cai, W. Zooming Across the Free-Energy Landscape: Shaving Barriers, and Flooding Valleys. *J. Phys. Chem. Lett.* **2018**, *9*, 4738−4745.

(23) Fu, H.; Shao, X.; Cai, W.; Chipot, C. Taming Rugged Free-Energy Landscapes Using an Average Force. *Acc. Chem. Res.* **2019**, *52*, 3254−3264.

(24) Bolhuis, P. G.; Dellago, C.; Chandler, D. Reaction Coordinates of Biomolecular Isomerization. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 5877−5882.
(25) Bowman, G. R.; Pande, V. S.; Noé, F. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*; Springer Science & Business Media, 2013; Vol. *797*.
(26) Moradi, M.; Tajkhorshid, E. Mechanistic Picture for Conformational Transition of a Membrane Transporter at Atomic Resolution. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 18916−18921.
(27) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of Slow Molecular Order Parameters for Markov Model Construction. *J. Chem. Phys.* **2013**, *139*, 015102.
(28) Tiwary, P.; Berne, B. J. Spectral Gap Optimization of Order Parameters for Sampling Complex Molecular Systems. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, 2839−2844.
(29) Chen, W.; Tan, A. R.; Ferguson, A. L. Collective Variable Discovery and Enhanced Sampling Using Autoencoders: Innovations in Network Architecture and Error Function Design. *J. Chem. Phys.* **2018**, *149*, 072312.
(30) Wehmeyer, C.; Noé, F. Time-Lagged Autoencoders: Deep Learning of Slow Collective Variables for Molecular Kinetics. *J. Chem. Phys.* **2018**, *148*, 241703.
(31) Chipot, C.; Comer, J. Subdiffusion in Membrane Permeation of Small Molecules. *Sci. Rep.* **2016**, *6*, 35913.
(32) Chodera, J. D.; Noé, F. Markov State Models of Biomolecular Conformational Dynamics. *Curr. Opin. Struct. Biol.* **2014**, *25*, 135−144.
(33) Amadei, A.; Linssen, A. B.; Berendsen, H. J. Essential Dynamics of Proteins. *Proteins* **1993**, *17*, 412−425.
(34) Ahmad, M.; Helms, V.; Kalinina, O. V.; Lengauer, T. Relative Principal Components Analysis: Application to Analyzing Biomolecular Conformational Changes. *J. Chem. Theory Comput.* **2019**, *15*, 2166−2178.
(35) Plattner, N.; Doerr, S.; de Fabritiis, G.; Noé, F. Complete Protein-Protein Association Kinetics in Atomic Detail Revealed by Molecular Dynamics Simulations and Markov Modelling. *Nat. Chem.* **2017**, *9*, 1005−1011.
(36) Noé, F.; Wu, H.; Prinz, J.-H.; Plattner, N. Projected and Hidden Markov Models for Calculating Kinetics and Metastable States of Complex Molecules. *J. Chem. Phys.* **2013**, *139*, 184114.
(37) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. Transition Path Sampling: Throwing Ropes Over Mountain Passes, in the Dark. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291−318.
(38) E, W.; Vanden-Eijnden, E. Transition-Path Theory and Path-Finding Algorithms for the Study of Rare Events. *Annu. Rev. Phys. Chem.* **2010**, *61*, 391−420.
(39) Onsager, L. Initial Recombination of Ions. *Phys. Rev.* **1938**, *54*, 554−557.
(40) E, W.; Ren, W.; Vanden-Eijnden, E. Transition Pathways in Complex Systems: Reaction Coordinates, Isocommittor Surfaces, and Transition Tubes. *Chem. Phys. Lett.* **2005**, *413*, 242−247.
(41) Miller, T.; Vanden-Eijnden, E.; Chandler, D. Solvent Coarse-Graining and the String Method Applied to the Hydrophobic Collapse of a Hydrated Chain. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 14559−14564.
(42) Gan, W.; Yang, S.; Roux, B. Atomistic View of the Conformational Activation of Src Kinase Using the String Method with Swarms-of-Trajectories. *Biophys. J.* **2009**, *97*, L8−L10.
(43) Meng, Y.; Lin, Y. L.; Roux, B. Computational Study of the "DFG-flip" Conformational Transition in c-Abl and c-Src Tyrosine Kinases. *J. Phys. Chem. B* **2015**, *119*, 1443−1456.
(44) Fajer, M.; Meng, Y.; Roux, B. The Activation of c-Src Tyrosine Kinase: Conformational Transition Pathway and Free Energy Landscape. *J. Phys. Chem. B* **2017**, *121*, 3352−3363.
(45) Ovchinnikov, V.; Trout, B. L.; Karplus, M. Mechanical Coupling in Myosin V: A Simulation Study. *J. Mol. Biol.* **2010**, *395*, 815−833.
(46) Ovchinnikov, V.; Karplus, M.; Vanden-Eijnden, E. Free Energy of Conformational Transition Paths in Biomolecules: The String Method and Its Application to Myosin VI. *J. Chem. Phys.* **2011**, *134*, 085103.
(47) Kirmizialtin, S.; Nguyen, V.; Johnson, K. A.; Elber, R. How Conformational Dynamics of DNA Polymerase Select Correct Substrates: Experiments and Simulations. *Structure* **2012**, *20*, 618−627.
(48) Stober, S. T.; Abrams, C. F. Energetics and Mechanism of the Normal-to-Amyloidogenic Isomerization of β2-Microglobulin: On-the-Fly String Method Calculations. *J. Phys. Chem. B* **2012**, *116*, 9371−9375.
(49) Matsunaga, Y.; Fujisaki, H.; Terada, T.; Furuta, T.; Moritsugu, K.; Kidera, A. Minimum Free Energy Path of Ligand-Induced Transition in Adenylate Kinase. *PLoS Comp. Biol.* **2012**, *8*, No. e1002555.
(50) Vashisth, H.; Maragliano, L.; Abrams, C. F. "DFG-flip" in the Insulin Receptor Kinase Is Facilitated by a Helical Intermediate State of the Activation Loop. *Biophys. J.* **2012**, *102*, 1979−1987.
(51) Vashisth, H.; Abrams, C. F. All-atom Structural Models of Insulin Binding to the Insulin Receptor in the Presence of a Tandem Hormone-binding Element. *Proteins: Struct. Funct. Bioinf.* **2013**, *81*, 1017−1030.
(52) Zhu, F.; Hummer, G. Pore Opening and Closing of a Pentameric Ligand-gated Ion Channel. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 19814−19819.
(53) Lev, B.; Murail, S.; Poitevin, F.; Cromer, B. A.; Baaden, M.; Delarue, M.; Allen, T. W. String Method Solution of the Gating Pathways for a Pentameric Ligand-Gated Ion Channel. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, E4158−E4167.
(54) Roh, S.-H.; Shekhar, M.; Pintilie, G.; Chipot, C.; Wilkens, S.; Singharoy, A.; Chiu, W. CryoEM and MD Infer Water−mediated Proton Transport and Autoinhibition Mechanisms of V_O Complex. *Science Adv.* **2020**, *6*, No. eabb9605.
(55) Singharoy, A.; Chipot, C.; Moradi, M.; Schulten, K. Chemomechanical Coupling in Hexameric Protein−protein Interfaces Harnesses Energy within V−Type ATPases. *J. Am. Chem. Soc.* **2017**, *139*, 293−310.
(56) Das, A.; Rui, H.; Nakamoto, R.; Roux, B. Conformational Transitions and Alternating-Access Mechanism in the Sarcoplasmic Reticulum Calcium Pump. *J. Mol. Biol.* **2017**, *429*, 647−666.
(57) Prajapati, J. D.; Fernańdez Solano, C. J.; Winterhalter, M.; Kleinekathöfer, U. Characterization of Ciprofloxacin Permeation Pathways Across the Porin OmpC Using Metadynamics and a String Method. *J. Chem. Theory Comput.* **2017**, *13*, 4553−4566.
(58) Moradi, M.; Tajkhorshid, E. Mechanistic Picture for Conformational Transition of a Membrane Transporter at Atomic Resolution. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 18916−18921.
(59) Moradi, M.; Tajkhorshid, E. Computational Recipe for Efficient Description of Large-Scale Conformational Changes in Biomolecular Systems. *J. Chem. Theory Comput.* **2014**, *10*, 2866−2880.
(60) Moradi, M.; Enkavi, G.; Tajkhorshid, E. Atomic-Level Characterization of Transport Cycle Thermodynamics in the Glycerol-3-Phosphate:Phosphate Transporter. *Nat. Commun.* **2015**, *6*, 8393.
(61) Ke, M.; Yuan, Y.; Jiang, X.; Yan, N.; Gong, H. Molecular Determinants for the Thermodynamic and Functional Divergence of Uniporter GLUT1 and Proton Symporter XylE. *PLoS Comput. Biol.* **2017**, *13*, No. e1005603.
(62) Xu, W.; Doshi, A.; Lei, M.; Eck, M. J.; Harrison, S. C. Crystal Structures of c-Src Reveal Features of Its Autoinhibitory Mechanism. *Mol. Cell* **1999**, *3*, 629−638.
(63) Cowan-Jacob, S. W.; Fendrich, G.; Manley, P. W.; Jahnke, W.; Fabbro, D.; Liebetanz, J.; Meyer, T. The Crystal Structure of a c-Src Complex in an Active Conformation Suggests Possible Steps in c-Src Activation. *Structure* **2005**, *13*, 861−871.
(64) Shukla, D.; Meng, Y.; Roux, B.; Pande, V. S. Activation Pathway of Src Kinase Reveals Intermediate States as Targets for Drug Design. *Nat. Commun.* **2014**, *5*, 1−11.

https://doi.org/10.1021/acs.jctc.1c01049
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

(65) Dickson, B. M.; Huang, H.; Post, C. B. Unrestrained Computation of Free Energy along a Path. *J. Phys. Chem. B* **2012**, *116*, 11046−11055.
(66) Huang, H.; Zhao, R.; Dickson, B. M.; Skeel, R. D.; Post, C. B. αC Helix as a Switch in the Conformational Transition of Src/CDK-like Kinase Domains. *J. Phys. Chem. B* **2012**, *116*, 4465−4475.
(67) Wu, H.; Post, C. B. Protein Conformational Transitions from All-Atom Adaptively Biased Path Optimization. *J. Chem. Theory Comput* **2018**, *14*, 5372−5382.
(68) Wu, H.; Huang, H.; Post, C. B. All-atom Adaptively Biased Path Optimization of Src Kinase Conformational Inactivation: Switched Electrostatic Network in the Concerted Motion of αC Helix and the Activation Loop. *J. Chem. Phys.* **2020**, *153*, 175101.
(69) Chipot, C.; Pohorille, A., Eds. *Free Energy Calculations: Theory and Applications in Chemistry and Biology*; Springer Series in Chemical Physics; Springer-Verlag: Berlin Heidelberg, 2007.
(70) Branduardi, D.; Gervasio, F. L.; Parrinello, M. From A to B in Free Energy Space. *J. Chem. Phys.* **2007**, *126*, 054103.
(71) Díaz Leines, G.; Ensing, B. Path Finding on High-Dimensional Free Energy Landscapes. *Phys. Rev. Lett.* **2012**, *109*, 020601.
(72) Hummer, G. Position-Dependent Diffusion Coefficients and Free Energies from Bayesian Analysis of Equilibrium and Replica Molecular Dynamics Simulations. *New J. Phys.* **2005**, *7*, 34.
(73) Comer, J.; Chipot, C.; González-Nilo, F. D. Calculating Position-Dependent Diffusivity in Biased Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2013**, *9*, 876−882.
(74) Hummer, G.; Szabo, A. Free Energy Reconstruction from Nonequilibrium Single-Molecule Pulling Experiments. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 3658−3661.
(75) Hummer, G.; Szabo, A. Free Energy Profiles from Single-Molecule Pulling Experiments. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 21441−21446.
(76) Hovan, L.; Comitani, F.; Gervasio, F. L. Defining an Optimal Metric for the Path Collective Variables. *J. Chem. Theory Comput.* **2019**, *15*, 25−32.
(77) Pratt, L. A Statistical-Method for Identifying Transition-States in High Dimensional Problems. *J. Chem. Phys.* **1986**, *85*, 5045−5048.
(78) Elber, R.; Karplus, M. A Method for Determining Reaction Paths in Large Molecules: Application to Myoglobin. *Chem. Phys. Lett.* **1987**, *139*, 375−380.
(79) Mills, G.; Jónsson, H. Quantum and Thermal Effects in H 2 Dissociative Adsorption: Evaluation of Free Energy Barriers in Multidimensional Quantum Systems. *Phys. Rev. Lett.* **1994**, *72*, 1124.
(80) Johnson, M. E.; Hummer, G. Characterization of a Dynamic String Method for the Construction of Transition Pathways in Molecular Reactions. *J. Phys. Chem. B* **2012**, *116*, 8573−8583.
(81) Maragliano, L.; Roux, B.; Vanden-Eijnden, E. Comparison Between Mean Forces and Swarms-of-Trajectories String Methods. *J. Chem. Theor. Comp.* **2014**, *10*, 524−533.
(82) Zhao, R.; Shen, J.; Skeel, R. D. Maximum Flux Transition Paths of Conformational Change. *J. Chem. Theory Comput.* **2010**, *6*, 2411−2423.
(83) Berezhkovskii, A. M.; Szabo, A. Committors, First-Passage Times, Fluxes, Markov States, Milestones, and All That. *J. Chem. Phys.* **2019**, *150*, 054106.
(84) Zhao, T.; Fu, H.; Lelièvre, T.; Shao, X.; Chipot, C.; Cai, W. The Extended Generalized Adaptive Biasing Force Algorithm for Multidimensional Free-Energy Calculations. *J. Chem. Theory Comput.* **2017**, *13*, 1566−1576.
(85) Phillips, J. C.; Hardy, D. J.; Maia, J. D. C.; Stone, J. E.; Ribeiro, J. V.; Bernardi, R. C.; Buch, R.; Fiorin, G.; Hénin, J.; Jiang, W.; McGreevy, R.; Melo, M. C. R.; Radak, B. K.; Skeel, R. D.; Singharoy, A.; Wang, Y.; Roux, B.; Aksimentiev, A.; Luthey-Schulten, Z.; Kalé, L. V.; Schulten, K.; Chipot, C.; Tajkhorshid, E. Scalable Molecular Dynamics on CPU and GPU Architectures with NAMD. *J. Chem. Phys.* **2020**, *153*, 044130.
(86) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696−3713.
(87) Fu, H.; Chen, H.; Wang, X.; Chai, H.; Shao, X.; Cai, W.; Chipot, C. Finding an Optimal Pathway on a Multidimensional Free-Energy Landscape. *J. Chem. Inf. Model.* **2020**, *60*, 5366−5374.
(88) Lelièvre, T.; Rousset, M.; Stoltz, G. Computation of Free Energy Profiles with Parallel Adaptive Dynamics. *J. Chem. Phys.* **2007**, *126*, 134111.
(89) Fu, H.; Shao, X.; Chipot, C.; Cai, W. Extended Adaptive Biasing Force Algorithm. An On−the−Fly Implementation for Accurate Free−Energy Calculations. *J. Chem. Theory Comput.* **2016**, *12*, 3506−3513.
(90) Lesage, A.; Lelièvre, T.; Stoltz, G.; Hénin, J. Smoothed Biasing Forces Yield Unbiased Free Energies with the Extended-System Adaptive Biasing Force Method. *J. Phys. Chem. B* **2017**, *121*, 3676−3685.
(91) Altis, A.; Nguyen, P. H.; Hegger, R.; Stock, G. Dihedral Angle Principal Component Analysis of Molecular Dynamics Simulations. *J. Chem. Phys.* **2007**, *126*, 244111.
(92) Roux, B. String Method with Swarms-of-Trajectories, Mean Drifts, Lag Time, and Committor. *J. Phys. Chem. A* **2021**, *125*, 7558−7571.
(93) Sedgewick, R.; Wayne, K. *Algorithms*, 4th ed.; Addison-Wesley, 2011; Chapter 4, pp 514−556.
(94) Szabo, A.; Schulten, K.; Schulten, Z. First Passage Time Approach to Diffusion Controlled Reactions. *J. Chem. Phys.* **1980**, *72*, 4350−4357.
(95) Vanden-Eijnden, E.; Venturoli, M. Markovian Milestoning with Voronoi Tessellations. *J. Chem. Phys.* **2009**, *130*, 194101.
(96) Adelman, J. L.; Grabe, M. Simulating Rare Events Using a Weighted Ensemble-Based String Method. *J. Chem. Phys.* **2013**, *138*, 044105.
(97) Pan, A. C.; Roux, B. Building Markov State Models along Pathways to Determine Free Energies and Rates of Transitions. *J. Chem. Phys.* **2008**, *129*, 064107.

https://doi.org/10.1021/acs.jctc.1c01049
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX