Bezier curve string method for the study of rare events in complex chemical systems

Citation: Bellucci, Michael A., and Bernhardt L. Trout. "Bezier Curve String Method for the Study of Rare Events in Complex Chemical Systems." The Journal of Chemical Physics 141, no. 7 (August 21, 2014): 074110. © 2014 AIP Publishing LLC

As Published: http://dx.doi.org/10.1063/1.4893216

Publisher: American Institute of Physics (AIP)

Persistent URL: http://hdl.handle.net/1721.1/92363

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Bézier curve string method for the study of rare events in complex chemical systems

Michael A. Bellucci and Bernhardt L. Trout

Citation: The Journal of Chemical Physics **141**, 074110 (2014); doi: 10.1063/1.4893216
View online: http://dx.doi.org/10.1063/1.4893216
View Table of Contents: http://scitation.aip.org/content/aip/journal/jcp/141/7?ver=pdfcov
Published by the AIP Publishing

Bézier curve string method for the study of rare events in complex chemical systems

Michael A. Bellucci and Bernhardt L. Trout[a)]
Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

(Received 19 April 2014; accepted 5 August 2014; published online 20 August 2014)

We present a new string method for finding the most probable transition pathway and optimal reaction coordinate in complex chemical systems. Our approach evolves an analytic parametric curve, known as a Bézier curve, to the most probable transition path between metastable regions in configuration space. In addition, we demonstrate that the geometric properties of the Bézier curve can be used to construct the optimal reaction coordinate near the most probable reaction path, and can further be used to devise a ranking vector capable of identifying precisely which collective variables are most important for governing the transition between metastable states. We discuss the algorithmic details of the Bézier curve string method, analyze its stability, accuracy and efficiency, and illustrate its capabilities using model potential energy functions. In particular, we use the degree elevation property of Bézier curves to develop an algorithm that adaptively learns the degree polynomial necessary to accurately represent the most probable transition path. Subsequently, we apply our method to the isomerization of alanine dipeptide, and demonstrate that the reaction coordinate obtained from the Bézier curve string method is in excellent agreement with the optimal reaction coordinate constructed from an aimless shooting and maximum likelihood procedure. Finally, we apply our method to a large complex system and study the homogenous nucleation of benzene from the melt. In these two examples, we illustrate that the ranking vector correctly identifies which collective variables govern these chemical transitions. © *2014 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4893216]

I. INTRODUCTION

The evolution of a complex molecular system often exhibits motion on widely separated time scales since different physical interactions in the molecular potential induce different characteristic timescales. Some examples of systems exhibiting this behavior include conformational changes of large biomolecules, molecular assembly of macromolecules, ligand binding, and nucleation events in phase transitions. The appearance of long time scales is a manifestation of large free energy barriers or entropic bottlenecks that confine the system in metastable basins in regions of phase space. Transitions out of these metastable basins only occur when sufficiently large energy fluctuations localized in these regions allow the system to overcome the free energy barriers. These types of transitions are referred to as rare events because these large energy fluctuations occur rarely with respect to the typical molecular time scales present in the system. In addition, these transitions generally occur on time scales that are much longer than those accessible by direct molecular dynamics simulations, making it impractical to study these systems with traditional methods. Consequently, more sophisticated computational methodologies are required to overcome this time scale problem.

In addition to overcoming the time scale problem, identifying the underlying mechanisms that cause these transitions is a fundamental challenge by itself. In general, the system evolves in high dimensional phase space, but because of the separation of time scales, the transitions between metastable states is often governed only by a subset of variables that evolve on the slower time scales present in the system. Therefore, to gain insight into the mechanism of the reaction, it is more useful to reduce the dimensionality of the system and describe the system as a function of important collective motions. These collective motions are known as collective variables, and ideally there exists some combination of them that can be formed into a single variable that completely describes the reaction between metastable states. This single variable description of the system in terms of collective variables is known as the reaction coordinate. The major challenge in gaining insight into the mechanism is therefore identifying which collective variables correctly describe the reaction and how each of them participates in the reaction coordinate. In addition, it is of paramount importance to choose the correct collective variables to describe the system because the exclusion of important degrees of freedom can lead to artificially low or high free energy barriers, incorrect reaction rates and mechanistic interpretation. Consequently, when modeling any large complex chemical system, a systematic approach to test reaction coordinates is needed.

To develop a systematic approach for testing reaction coordinates, one necessarily needs a criterion for what the reaction coordinate should be. For a phase space model of any chemical reaction, the unique reaction coordinate is the solution to the backward Kolmogorov equation,[1–6] which in the chemical physics literature, is known as the committor probability. The committor probability, $q(x)$, is a function that

[a)]Electronic mail: trout@mit.edu

specifies at each point in configuration space or phase space, the probability that a trajectory initiated at that point will succeed in making a transition to a product state. Furthermore, it has been shown[1,2] that the isosurfaces of the committor probability completely characterize the progress of the reaction. Unfortunately, for all but the simplest of systems, obtaining the exact committor function is impractical. However, it can often be approximated to varying levels of accuracy. Ideally, for a computational method to be successful at modeling large complex systems, at the very least, it must be capable of overcoming the time scale issue and approximating the committor function in some region of configuration space. A number of methods have been developed to address these issues. Of these methods, we focus our discussion here on two of the most successful methods: path sampling methods and string methods.

Path sampling methods are generally based on the transition path sampling (TPS) framework, which has been utilized to obtain insight into the mechanism of rare events in a variety of chemical and biological systems.[3,4,7,8] Under this framework, an ensemble of unbiased reactive trajectories is obtained between metastable basins by performing a Monte Carlo procedure in path space. This methodology has been used to study melt crystallization in sodium halides,[9] pressure induced polymorphic transformations,[10] evaporation coefficients of water,[11] and to generate nucleation pathways to hexagonal ice in water.[12,13] Recently, a variant of TPS known as the aimless shooting algorithm was developed in our group.[14,15] In this method, the TPS algorithm is modified and combined with a likelihood maximization method capable of finding the optimal reaction coordinate. It has been successfully used to obtain insight into the nucleation mechanism in the Ising Model[14] and benzene from the melt,[16] as well as polymorph transformations in organic compounds.[15,17,18] Despite the success of these methods, the diffusive nature of the system and ruggedness of the free energy landscape can limit their applicability due to long simulation times needed to gather adequate statistics.

Alternatively, the string method,[19,20] and its variants,[21–24] can circumvent long simulation times even for diffusive systems or systems with rugged free energy landscapes. The string method is a general approach that provides a reliable way of calculating minimum energy paths (MEP) or minimum free energy paths (MFEP) for barrier-crossing events. The method proceeds by evolving strings, i.e., smooth curves, to the most probable transition path between metastable regions in configuration space. It has been used to study hydrophobic collapse of hydrated chains,[25] membrane adhesion,[26] capillary condensation,[27] and to find nucleation pathways in block copolymers.[28] However, in contrast to path sampling methods, the string method using collective variables is a biased method, requiring one to choose suitable collective variables *a priori*. Moreover, the string method converges to a single path in collective variable space, whereas path sampling methods build an ensemble of unbiased transition paths, which in the case of aimless shooting, can be used to systematically construct the optimal reaction coordinate. Therefore, the string method inherently provides different information than path sampling methods, making it cumbersome to test reaction coordinate models using simulation data *a posteriori*.

The objective of this work is to develop an approach for determining important collective variables and the optimal reaction coordinate, comparable in accuracy to a maximum likelihood approach, from information available from the string method in collective variables. In this work, we show that provided that the number of collective variables is large enough, the mechanism of the reaction can be adequately captured by finding the most probable reaction path in this large set of collective variables. In addition, it has been shown that isocommittor surfaces can be locally approximated near the MFEP by the normal planes of the curve. The normal plane near the $q(x) = 1/2$ point on the string gives us an estimate for configurations that comprise the transition state ensemble, and by definition, motion in this direction does not lead toward the reactant or product region. In contrast, motion in the direction of the tangent plane leads to the most significant progress in the reaction. Moreover, since the tangent plane is orthogonal to the committor isosurfaces, it must be proportional to the change in the reaction coordinate, $\nabla q(x)$, since the gradient of any function is orthogonal to its level sets. Consequently, to determine what collective variables are important in governing the transition from reactants to products, one can analyze the components of the unit velocity vector and unit acceleration vector tangent to the MFEP at the $q(x) = 1/2$ point on the string. Together, these components impart a natural rank on the collective variables in terms of importance; the larger the magnitude of the components associated with a particular collective variable, the more important the collective variable is in governing the transition. Unfortunately, in the string method in collective variables, the string is represented by a piecewise cubic spline, and therefore, there is little information regarding the components of these vectors. One could always approximate these vectors numerically as was done in the original string method,[19] but it was shown[20] that evolving the string using the numerical derivatives was numerically unstable and less accurate than the current approach, which uses a Lagrange multipliers method to evolve the string.

In the spirit of the original string method, we have developed a numerically stable and accurate string method using Bézier curves to find MEPs and MFEPs on potential or free energy surfaces. The Bézier curve is an analytic parametric curve equipped with a Bernstein polynomial basis set and has been used in a variety of applications.[29–33] A natural advantage of using a Bézier curve in the representation of the string is that any geometric property of the curve can be computed analytically, including the velocity and acceleration vectors. This allows us to perform the reaction coordinate analysis described above. In addition, we demonstrate that the natural properties of Bézier curves, such as degree elevation, imparts a level of flexibility in the algorithm proposed here, and allows us to compute the MEP or MFEP to a high level of accuracy. The discussion of our method proposed herein is as follows: we begin with an overview of the string method, Bézier curves, their properties, and the Bézier curve string method algorithm. We then develop a reaction coordinate analysis for quantifying how important collective

variables are in governing rare event transitions relative to one another. Subsequently, we use model potentials to test the robustness of our method and its convergence properties. We further apply our method to study the isomerization of the alanine dipeptide molecule, and show that the reaction coordinate obtained from the Bézier curve string method is equivalent to results obtained from a likelihood maximization reaction coordinate analysis performed on data obtained from aimless shooting. In addition, we demonstrate that the reaction coordinate analysis we develop here allows us to determine precisely which collective variables are governing the conformational transition. Finally, we apply our method to a large complex chemical system, and study the mechanism of homogenous nucleation of benzene from the melt.

II. OVERVIEW OF THE STRING METHOD

The Bézier curve string method is based on the string method in collective variables algorithm of Maragliano *et al.*, and so, we present a general overview of the string method and refer the reader to Ref. 21 for a detailed description of the algorithm. The main objective of the string method is to find the MEP or MFEP for chemical systems. Denote by $V(x)$ the potential energy of the system of interest and assume that $V(x)$ has at least two minima, corresponding to a reactant state and a product state. By definition, a MEP is a curve, φ, connecting the reactant and product basins that satisfies

$$\nabla V(\varphi)^{\perp} = 0, \tag{1}$$

$$\nabla V(\varphi)^{\perp} = \nabla V(\varphi) - (\nabla V(\varphi) \cdot \hat{\tau})\hat{\tau}, \tag{2}$$

where $\hat{\tau}$ is the unit tangent vector. Therefore, the MEP is a curve where the normal component of the gradient of the potential is zero, or equivalently, the gradient of the potential is parallel to the tangent vector at every point along the curve. The underlying idea of the string method is to find MEPs by evolving a curve by

$$v^{\perp} = -\nabla V(\varphi)^{\perp}, \tag{3}$$

where $v^{\perp}$ is the normal velocity of the curve, since stationary solutions of (3) satisfy (1). Similarly, given a set of N collective variables, $\theta(x) = (\theta_1(x), \ldots, \theta_N(x))$, where N is less than the dimensionality of the full system, then the free energy of the system can be written as

$$F(z) = -k_B T \ln\left(Z^{-1} \int_{\mathbb{R}^n} e^{-\beta V(x)} \prod_{j=1}^{N} \delta(z_j - \theta_j(x)) dx \right), \tag{4}$$

where Z is the canonical partition function. We can see that if N is less than the dimensionality of the full system, then only the degrees of freedom used to define the collective variables are explicitly taken into account in the free energy, whereas all other degrees of freedom contribute to the free energy in an average sense. Consequently, to adequately capture the barrier heights and shape of the full dimensional free energy surface along the MFEP, one must choose collective variables that are relevant to the underlying molecular transition.

Let $z(\alpha)$ be a path in collective variable space, then the MFEP connecting a reactant state and a product state is a curve that satisfies

$$M(z)\nabla_z F(z)^{\perp} = 0, \tag{5}$$

$$M(z)\nabla_z F(z)^{\perp} = M(z)\nabla_z F(z) - (M(z)\nabla_z F(z) \cdot \hat{\tau})\hat{\tau}, \tag{6}$$

where $F(z)$ is the free energy of the system and M(z) is a tensor that accounts for the curvilinear nature of the collective variables. The MFEP is found by evolving the curve by the corresponding normal velocity equation until the curve is stationary. Once the string has converged to the MFEP, the free energy along the string can be determined using the following:

$$\frac{dF(z(\alpha))}{d\alpha} = \nabla_z F(z(\alpha)) \cdot \frac{dz(\alpha)}{d\alpha}, \tag{7}$$

$$F(z(\alpha)) - F(z(0)) = \int_0^{\alpha} \nabla_z F(z(\alpha')) \cdot \frac{dz(\alpha')}{d\alpha'} d\alpha'. \tag{8}$$

III. BÉZIER CURVES

A Bézier curve is a parametric curve originally developed for use in computer aided geometric design,[29] but has also been used in approximation theory,[30,31] optimal control theory,[32] and in applied mathematics to solve partial differential equations.[33] Moreover, Bézier curves are very useful in curve fitting; since every real-valued continuous function on the interval $[a, b]$ can be uniformly approximated by a nth degree polynomial, every continuous function has a Bézier curve representation. A Bézier curve is defined as a linear combination of Bernstein polynomials,

$$z_i(\alpha) = \sum_{j=0}^{n} P_{i,j} B_{n,j}(\alpha), \quad \alpha \in [0, 1], \tag{9}$$

$$B_{n,j} = \binom{n}{j} \alpha^j (1-\alpha)^{n-j}, \tag{10}$$

where $z_i(\alpha)$ represents the ith component of the parametric curve $z(\alpha)$, which corresponds to the ith collective variable, α is a parameter, n is the degree of the polynomial, $B_{n,j}(\alpha)$ are Bernstein polynomials, and $P_{i,j}$ is a set of control points that together form a control polygon (see Fig. 1). The velocity and acceleration of the curve have the following analytic expressions:

$$v_i(\alpha) = \sum_{j=0}^{n-1} n(P_{i,j+1} - P_{i,j}) B_{n-1,j}(\alpha), \tag{11}$$

$$a_i(\alpha) = \sum_{j=0}^{n-2} n(n-1)(P_{i,j+2} - 2P_{i,j+1} + P_{i,j}) B_{n-2,j}(\alpha). \tag{12}$$

One of the most useful properties of Bézier curves is the degree elevation property. Any nth degree Bézier curve can be exactly represented as a Bézier curve of degree $n+1$, and hence, as a curve of any degree $n' > n$. When raising the

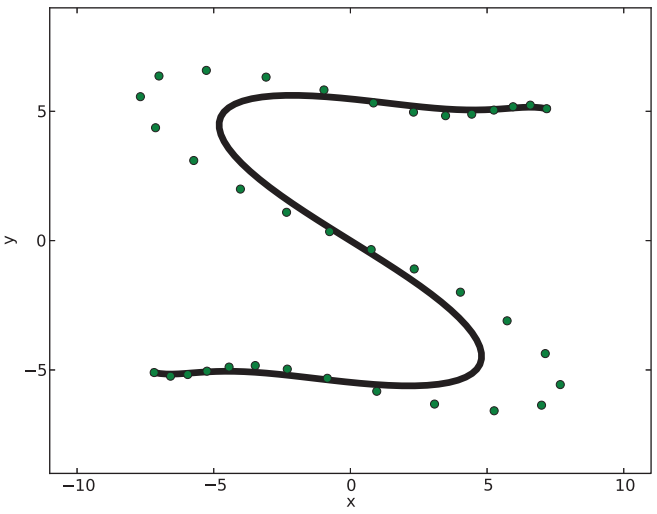


FIG. 1. Example of a two dimensional Bézier curve with control points overlaid in green. Aside from the endpoints of the curve, the remaining $n-2$ control points do not necessarily lie on the curve itself.

degree of the Bézier curve, one necessarily needs to determine the control points for the equivalent $n+1$ Bézier curve. The $n+1$ control points can be computed using the following equation:

$$P_{i,j}^{n+1} = \frac{j}{n+1} P_{i,j-1}^{n} + \frac{n+1-j}{n+1} P_{i,j}^{n}. \tag{13}$$

The degree elevation property of Bézier curves is especially useful because it allows us to adaptively increase the degree of the polynomials in the Bézier curve to achieve a high level of accuracy when evolving the string. In general, if the collective variables are sufficiently well-behaved when describing a barrier-crossing event, as is the case for most collective variables, then a Bézier curve can describe the transition to a very high level of accuracy using a small Bernstein polynomial basis.

IV. BÉZIER CURVE STRING METHOD ALGORITHM

The Bézier curve string method algorithm can be broken down into a series of simple steps:

1. Generate Initial String
2. Compute Mean Force and Metric Tensor
3. Evolve String
4. Reparameterization
5. Evaluate Error
6. Degree Elevation

These steps are described in greater detail below.

A. Initial Bézier string

In practice, the Bézier string, $z(\alpha, t)$, is discretized into a set of m images by discretizing the parameter α in (9) into a set of m points. Since the string method is a local optimization method, it is advantageous to have an initial string close to, in the Fréchet distance sense, the most probable transition path. For example, to generate an initial string to study crystallization, one would typically take the crystal of interest, slowly melt it in a molecular dynamics simulation, and choose configurations along the melting trajectory as the discretized points along the initial string. Let $z0$ represent the initial string in collective variable space determined from a molecular dynamics simulation, let $B = (B_{n,0} \ldots, B_{n,n})$ be the $m \times n$ Bernstein polynomial matrix, and let $P = (P_0, \ldots, P_N)$ be the $n \times N$ control point matrix, then we necessarily have the following condition:

$$BP = z0. \tag{14}$$

Therefore, finding an initial Bézier string amounts to finding the control points, P, that make the Bézier string best approximate some initial string $z0$. For a detailed description of how to determine the optimal control points for the initial string, we refer the reader to Appendix A.

B. Mean force and metric tensor

The evolution of the string requires the evaluation of the mean force, $\nabla_z F(z)$, and the tensor $M(z)$ locally around the string. Since this aspect of our algorithm is performed in the same way as in the original string method in collective variables algorithm, we refer the reader to Ref. 21 for a detailed description of how the mean force and metric tensor are determined from simulation. Once the mean force and metric tensor are determined, we compute the normal component of the mean force using (6). The unit tangent vector of the curve can be computed analytically with (11) using the relation $\hat{\tau} = v/|v|$.

C. String dynamics

To evolve the Bézier curve toward the MFEP, one evolves the m images on the string in the direction of the normal velocity of the curve. In practice, this amounts to finding the control points that reproduce the correct string after evolution by the normal velocity. This evolution can be viewed as a curve-fitting problem that involves first evolving the string by the normal velocity and then finding the control points that best reproduce the time-evolved string. However, a much more simple and stable approach is to allow the control points to be a time-dependent dynamical system. Using a forward Euler integration scheme, we can determine the time evolved control points using the following:

$$P_k^{h+1} = P_k^h - \frac{B_{n,k} \cdot M(z) \nabla_z F(z)^{\perp}}{B_{n,k} \cdot B_{n,k}} \Delta t, \quad k = 1, \ldots, n-1. \tag{15}$$

For a derivation of this equation, please refer to Appendix B. Our algorithm performs this update on all the control points except the control points on the endpoints of the string. For the control points on the endpoints, we evolve them toward the minima of the reactant and product basins by evolving them by the full force instead of the normal force, i.e., we evolve them by

$$P_k^{h+1} = P_k^h - M(z(\alpha_k)) \nabla_z F(z(\alpha_k)) \Delta t, \quad k = 0, n, \tag{16}$$

where $-M(z(\alpha_k))\nabla_z F(z(\alpha_k))$ is the force at the corresponding endpoints. Note that this step must be performed before evolving the $n-2$ control points by the normal force to maintain stability in the equations of motion.

D. Reparameterization

The reparameterization step is performed to ensure that the images along the string are equidistant from one another, or equivalently, it is the realization of an arc length parameterization of the curve. In the original string method algorithm, this step was necessary to enforce a Lagrange multiplier constraint throughout the evolution of the string. However, in our algorithm, this step is not explicitly necessary, but is often useful to ensure adequate sampling across the string. The most straightforward parameterization of a Bézier curve is to simply discretize α into m points, which in turn, discretizes the string into m images. This parameterization is usually sufficient, but in regions of high curvature, i.e., where the string is bending, this parameterization can lead to images being very close together. As a result, in regions of low curvature, the images can be widely separated, and thus, the sampling in these regions will be sparse. Fortunately, a Bézier curve has an arbitrary parameterization, and therefore, any parameterization can be realized by performing a parameter substitution, $\alpha = f(u)$. If in particular, we let $u = s$, then our curve will be parameterized with respect to arc length s provided that $f(s)$ is the inverse function of the arc length. The function $f(s)$ cannot be computed analytically, but it can be determined numerically. To find $f(s)$ we first compute the arc length function of the curve with numerical integration

$$s(\alpha) = \int_0^{\alpha} |v(\alpha')| d\alpha', \tag{17}$$

where $|v(\alpha')|$ is the speed of the curve. After determining the arc length function of the curve, the inverse function $f(s)$ can then be computed numerically with the following:

$$f(s) = \int_0^{s} |v(s')|^{-1} ds'. \tag{18}$$

Once we know the function $f(s)$, we determine what values of α correspond to an equal arc length parameterization. To do this, we interpolate $f(s)$ using a cubic spline interpolation as a function of the computed arc length $s(\alpha)$, then we discretize s into m equidistant points, and subsequently calculate the new values of α that correspond to the equal arc length parameterization from the interpolated function $f(s)$. Using the calculated values of α in the Bézier curve yield an equal arc length parameterization.

E. Error evaluation

As a simple measure of error, we can define our error function to be

$$E_{MFEP} = \max_i |M(z(\alpha_i))\nabla_z F(z(\alpha_i))^{\perp}|, \tag{19}$$

where *max* represents the maximum value of $|M(z(\alpha_i))\nabla_z F(z(\alpha_i))^{\perp}|$ over the string. To judge convergence of the string, we evolve the string until E_{MFEP} drops below some tolerance level, *TOL*. E_{MFEP} is a fairly stringent measure of error, since it ensures that the maximum magnitude of $M(z(\alpha_i))\nabla_z F(z(\alpha_i))^{\perp}$ never exceeds *TOL*, which can arbitrarily be set to some desired accuracy level. However, this definition of error can lead to slow convergence. To see why this is the case, we first note that (19) may be rewritten as

$$E_{MFEP} = \max_i [|M(z(\alpha_i))\nabla_z F(z(\alpha_i))|^2 (1 - \cos(\phi)^2)]^{1/2}, \tag{20}$$

where ϕ is the angle between $M(z(\alpha_i))\nabla_z F(z(\alpha_i))$ and $\hat{\tau}$, the unit tangent vector of the string. We see from (20), that the convergence of the string is governed by $|M(z(\alpha_i))\nabla_z F(z(\alpha_i))|^2$ and ϕ, and if in particular, $|M(z(\alpha_i))\nabla_z F(z(\alpha_i))|^2$ happens to be very large, then $\phi \ll 1^{\circ}$ in order to satisfy $E_{MFEP} \approx 0$. In practice, the string is stationary long before $\phi \ll 1^{\circ}$, and so, further evolution of the string beyond the point it is stationary makes little difference in the MFEP. Therefore, (20) is not a very efficient error function for measuring convergence. Furthermore, since the magnitude of $|M(z(\alpha_i))\nabla_z F(z(\alpha_i))|^2$ depends on the system and is not known *a priori*, it is difficult to set an appropriate value for *TOL* that coincides with the string being stationary. However, we see from (20) that the defining criterion for a MFEP is such that $M(z)\nabla_z F(z)$ is parallel to $\hat{\tau}$, or equivalently, that $(1 - \cos(\phi)^2) = 0$. Therefore, we can define an equivalent error function based solely on geometrical considerations as

$$E_{MFEP} = \max_i (1 - \cos(\phi)^2), \tag{21}$$

where *max* is defined with respect to the maximum $|M(z(\alpha_i))\nabla_z F(z(\alpha_i))^{\perp}|$ along the string. Moreover, we define $TOL = 1 - \cos(\phi)^2$, for some small value of ϕ, typically $\phi < 0.5^{\circ}$. In this way, we are still minimizing (20), but our criterion for judging convergence is system independent and better defined. Consequently, the algorithm spends less time trying to achieve very small values of ϕ, as it would if we used (20) as an error function, and therefore, it is much more efficient.

F. Degree elevation

When evolving the Bézier string, we generally begin with a low degree Bézier curve, usually the lowest degree Bézier curve that fits the data for the initial string below some threshold, and subsequently make use of the degree elevation property throughout the evolution of the string. This approach allows the algorithm to adaptively find the degree of the polynomial that best represents the MFEP. Moreover, it prevents the Bézier string from becoming kinked, which can occur if the degree of the polynomial is much greater than is needed to represent the MFEP. In practice, a $n+1$ degree elevation is achieved by increasing the Bernstein basis set to B_{n+1} and recursively solving for the $n+1$ control points using (13).

The simplest approach to determine when to perform a degree elevation is to increase the degree of the basis set every T time steps. However, since we can explicitly measure $M(z)\nabla_z F(z)^{\perp}$ when searching for the MFEP, we can therefore

compute the exact error on the string in our method. Consequently, a much better approach at determining when to perform a degree elevation is to monitor the error of the string and perform a degree elevation when the change in error drops below a tolerance level, Δ. In addition, as the string converges toward the MFEP, the change in error necessarily gets smaller, and therefore, the tolerance level, Δ, that governs whether or not to perform a degree elevation should get smaller as well. Consequently, every time we perform a degree elevation, we rescale Δ by some number C, which is a number in the range $0 \leq C \leq 1$. This leads to an exponential decay in the value of Δ, i.e., $\Delta = \Delta_0 C^r$, where Δ_0 is the initial tolerance, and r is the number of times the change in error drops below Δ throughout the evolution of the string.

In general, the convergence time depends on the parameters Δ_0 and C. However, there exists a set of parameters for which the algorithm performs optimally. To find this set of parameters, we adopt a simple and reliable approach for approximating the optimal parameters from simulation.

The change in error as a function of time, denoted $\Delta E(t)$, to a good approximation can be thought of as following some unknown exponential decay model, $\Delta E(t) = \Delta E_0 (C_E)^t$, where $0 \leq C_E \leq 1$. The function Δ in our algorithm represents a model for this exponential decay. Therefore, we can use the first few data points of the simulation to estimate good values for Δ_0 and C in our model. A suitable value for Δ_0 is the initial change in error between the first and second times steps from simulation, i.e., $\Delta_0 = \Delta E_0$. Similarly, the parameter C can be thought of as a rate of convergence, μ, which can be measured as

$$\mu = \frac{\Delta E(t+1)}{\Delta E(t)}. \tag{22}$$

Therefore, we can estimate the parameter C by measuring the change in error in the simulation during the first few time steps and then calculating the rate of convergence. Once we have done this, we set the parameter C to be slightly less than that of the measured rate of convergence so that the algorithm does not perform an excessive number of degree elevations throughout the simulation. In general, we find that values of C in the range $0.90 \leq C \leq 0.95$ work well.

V. COMMITTOR FUNCTION, REACTION COORDINATE, AND RANKING VECTOR

Consider a system governed by overdamped Langevin dynamics,

$$dx = -\frac{\partial V(x)}{\partial x} dt + \sqrt{2k_B T}\, dw, \quad x \in \Omega \subset \mathbb{R}^n, \tag{23}$$

where $V(x)$ is the potential energy of the system, k_B is the Boltzmann constant, T is the temperature, and dw is the standard Brownian motion. The system described by (23) samples configurations from the canonical ensemble and has the standard equilibrium distribution

$$\rho(x) = \frac{\exp(-\beta V(x))}{Z} \tag{24}$$

in the configuration space Ω, where $Z = \int_\Omega \exp(-\beta V(x)) dx$ is the partition function. Suppose that one is interested in understanding the mechanism of a reaction between two metastable states described by two subsets, A and B, of the configuration space Ω. Then the unique reaction coordinate, $q(x)$, that characterizes the transition between A and B is defined as the solution of the backward Kolmogorov equation,

$$\beta^{-1} \nabla \cdot \nabla q(x) - \nabla V(x) \cdot \nabla q(x) = 0, \tag{25}$$

$$q|_{x \in A} = 0, \quad q|_{x \in B} = 1.$$

The committor probability, $q(x)$, is a function that specifies at each point in configuration space, the probability that a trajectory initiated at that point will succeed in making a transition to a product state. With this definition, we can see that the committor probability leads to a natural generalization of the concept of a transition state. For simple systems with few degrees of freedom and sufficiently smooth free energy landscapes, transition states are often identified by finding saddle points on the free energy landscape. However, for large systems with many degrees of freedom, the free energy landscape can often be very rugged with numerous saddle points, giving this definition little significance. In contrast, true statistical transition states can be found by looking at configurations on the $q(x)$=1/2 isosurface, since any trajectory initiated from these configurations have a 50% chance to proceed toward the product state or return to the reactant state. Furthermore, it has been shown that the MFEP proceeds in the orthogonal direction to the committor isosurfaces.[1,2,21,22] Consequently, the committor isosurfaces completely characterize the progress of the reaction, and therefore, the committor function can be viewed as the ideal reaction coordinate.

Unfortunately, Eq. (25) is a multidimensional partial differential equation, and as such, it has no analytic solution. Furthermore, the large dimensionality of the system makes it impossible to find a solution with traditional numerical methods. Although, it is often not necessary to know the committor probability in all of configuration space, but rather, only locally around the MFEP. Since the MFEP represents the most probable transition path, it follows that the vast majority of transition pathways between metastable states occur near or on the MFEP. To a good approximation,[1,2,21] the probability to observe a transition pathway away from the MFEP decreases with increasing distance from the MFEP. Consequently, to gain insight into the reaction mechanism, it is sufficient to approximate the committor function in some finite neighborhood of the MFEP.

The most direct approach at approximating the committor probability is to choose points in configuration space, sample random initial velocities from the Boltzmann distribution, integrate trajectories from these points, and record the number of times each trajectory ended in the product basin versus the reactant basin. By computing the histogram of these outcomes, one can approximate the committor probability at any point. For systems exhibiting widely separated time scales, this approach is generally only feasible near the transition state isosurface. The aimless shooting algorithm developed in our group utilizes this fundamental idea within a TPS framework to build estimates of the $q(x) = 1/2$ isosurface. However, unlike TPS where a new trajectory is obtained by modifying

velocities from a previous trajectory, a new trajectory in aimless shooting is obtained by choosing a configuration along the previous trajectory (near the putative $q(x) = 1/2$ isosurface), and drawing new velocities from the Boltzmann distribution. By drawing new velocities from the Boltzmann distribution, the outcomes of the path sampling are independent, and thus, these outcomes can be used to estimate committor probabilities at each shooting point. Therefore, during the course of the aimless shooting, one keeps track of the outcomes of each trajectory, i.e., whether or not the sampled path connects the reactant and product basins. To find the optimal reaction coordinate, one uses a model function, which is a function of some trial reaction coordinate, $r(z)$, and optimizes the parameters in the reaction coordinate using a maximum likelihood procedure such that the model function best explains the outcomes from the aimless shooting. Often, the reaction coordinate is simply represented as a linear combination of collective variables, such as

$$r(z) = \sum_{j=1} c_j z_j - c_0. \tag{26}$$

The method can find higher order reaction coordinates, but this planar reaction coordinate is usually sufficient.

In contrast, the string method uses a variational approach to locally approximate the isocommittor surfaces near the MEP or MFEP. As the solution of (25), $q(x)$ has an equivalent characterization as being the minimizer of

$$I = \int_{\Omega/A\cup B} |\nabla q(x)|^2 \exp(-\beta V(x)) dx. \tag{27}$$

Using a planar approximation for the committor isosurfaces, it was shown in Ref. 22 that (27) can be simplified as

$$I = \int_0^1 (q'(\alpha))^2 \exp(-\beta F(\alpha)) |\hat{n}_{iso} \cdot \varphi'|^{-1} d\alpha, \tag{28}$$

where $\hat{n}_{iso}$ is the unit normal to the isocommittor surface and φ' is the tangent vector of the string. The minimization condition for this functional is such that $\hat{n}_{iso}$ is parallel to φ', and given this minimization condition, the minimizer of (28) subject to the boundary conditions in (25) is given by

$$q(\alpha) = \frac{\int_0^\alpha \exp(\beta F(\alpha')) d\alpha'}{\int_0^1 \exp(\beta F(\alpha')) d\alpha'}. \tag{29}$$

Since $\hat{n}_{iso} \parallel \varphi'$, we necessarily have the condition that the unit normal of the MEP or MFEP is a first order approximation to the isocommittor surfaces and that φ' is a first order approximation to the reaction coordinate. Provided that the string has been converged in a large set of collective variables, the reaction coordinate obtained from the string method should be equivalent to the one obtained from the aimless shooting and maximum likelihood approach under the planar approximation. We demonstrate this below in Sec. VI.

Using the analytic properties of the Bézier curve, we can go a step further and assess which collective variables are participating the most in the transition from reactant to product. In general, this information is more elucidating than the reaction coordinate itself because it provides information about which collective variables are likely to be the governing variables in the system. Since $\hat{n}_{iso} = \nabla q(x)/|\nabla q(x)|$ and $\hat{n}_{iso} \parallel \varphi'$, when the string is converged, we must have the condition that

$$\hat{\tau} = \frac{\nabla q(x)}{|\nabla q(x)|}, \tag{30}$$

where $\hat{\tau}$ is the unit tangent vector of the string, or equivalently, the unit velocity vector tangent to the string. Therefore, $\hat{\tau}$ gives information about the magnitude and direction of the change in reaction coordinate as a function of the collective variables of the system. This is particularly important near the transition state isosurface since changes in the reaction coordinate are responsible for inducing the transition. Therefore, the relative level of "importance" of a collective variable can be assessed by looking at its corresponding magnitude in the unit velocity and unit acceleration vectors tangent to the string near the transition state isosurface, which can be determined using (29). To devise a simple ranking system for the importance of the collective variables, we use the following:

$$\frac{\hat{\tau} \circ \hat{\tau} + \hat{a} \circ \hat{a} \Delta t}{|\hat{\tau} \circ \hat{\tau} + \hat{a} \circ \hat{a} \Delta t|}, \tag{31}$$

where $\hat{\tau}$ and $\hat{a}$ are the unit velocity and unit acceleration vectors tangent to the string, respectively, and $\circ$ denotes the Hadamard product, or equivalently, component-wise multiplication. The appearance of Δt in (31) is due to the fact that the acceleration is proportional to Δt^2, while the velocity is proportional to Δt with respect to motion in the system. Therefore, when determining the importance of a variable, more weight should be allocated to its velocity component than to its acceleration component. To get (31), we simply divided by Δt. In addition, since the signs of the components do not matter for our purposes, we use the Hadamard product to ensure each component in the two vectors is positive before adding them together. Once we compute (31), we sort the resulting vector from largest to smallest to determine the relative rank of importance of the collective variables.

VI. MODEL EXAMPLES

To illustrate the efficiency of the Bézier curve string method in finding MEPs, we first applied it to the Mueller potential[20] and the circle potential,[20] which are simple two-dimensional test systems commonly used in testing the performance of algorithms. We first studied the convergence properties of the Bézier curve string method without degree elevation and converged a series of Bézier strings as a function of increasing number of Bernstein polynomial basis functions, or equivalently, increasing number of control points since the number of control points is always equal to the number of basis functions. In addition, the number of control points is also equal to the degree of the Bernstein polynomial basis set.

For the Mueller potential, we used a linear interpolation between the basins of attraction as the initial string, and for the circle potential, the initial string was generated from the following:

$$x = \cos(\pi t), \quad y = -0.5 \sin(\pi t), \tag{32}$$

Note that the locations of the minima are not required *a priori*. As long as the end points of the initial string lie in the

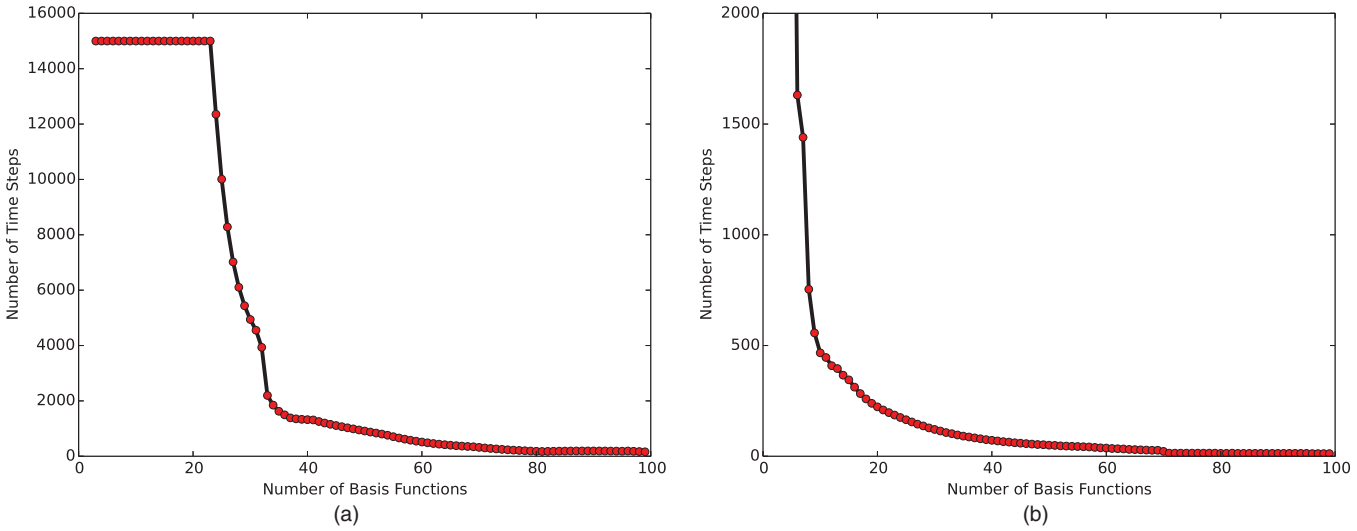


FIG. 2. Number of time steps to converge the Bézier strings below a given tolerance *TOL* as a function of the number of basis functions for (a) Mueller potential and (b) circle potential. The tolerance was defined as $TOL = 1 - \cos(\phi)^2$, with $\phi = 0.5^\circ$, which ensures that the angle between $\nabla V(x)$ and $\hat{\tau}$ is less than 0.5° at the maximum error point on the string. The points at 15 000 in (a) show that the Bézier string did not converge below *TOL* in the 15 000 steps of the simulation. Similarly, the first three data points in (b) did not converge below *TOL* as well, but are omitted for clarity.

neighborhood of the minima, they are identified automatically since the end points of the string evolve by $-\nabla V(x)$ and their positions are not affected by the reparameterization step. For our studies here, the Bézier strings, $\varphi(\alpha)$, were discretized into $m = 30$ images and the reparameterization of the strings was carried out every time step. The normal force along the string was computed using (2) and $\hat{\tau}$ was computed by normalizing (11). The control points were evolved with

$$P_k^{h+1} = P_k^h - \frac{B_{n,k} \cdot \nabla V(\varphi)^\perp}{B_{n,k} \cdot B_{n,k}} \Delta t, \quad k = 1, \ldots, n-1 \quad (33)$$

$$P_k^{h+1} = P_k^h - \nabla V(\varphi(\alpha_k)) \Delta t, \quad k = 0, n,$$

which corresponds to forward Euler integration. For each string, we fixed the number Bernstein polynomial basis functions, and evolved each string for 15 000 time steps using a time step of $\Delta t = 0.5 \times 10^{-4}$. The number of basis functions used in the convergence study ranged from $n = 3$ to $n = 99$ in increments of 1.

The resulting convergence profiles of the Bézier strings on the Mueller potential and on the circle potential are shown in Figs. 2(a) and 2(b), respectively. From Fig. 2, we see the number of time steps it took to converge each Bézier string below a given tolerance, *TOL*, as a function of increasing number of basis functions. For this study, we used (21) as the error function, and defined the tolerance as $TOL = 1 - \cos(\phi)^2$, with $\phi = 0.5^\circ$, which ensures that the angle between $\nabla V(x)$ and $\hat{\tau}$ is less than 0.5° at the maximum error point on the string. In Fig. 2(a), we see that Bézier strings with $n < 24$ do not converge below *TOL* in the 15 000 time steps of the simulation, highlighting the stringent requirements of the error function. Similarly, for the circle potential, Bézier strings with $n < 5$ do not converge below *TOL* in the 15 000 time steps of the simulation. However, we also see from Fig. 2 that the convergence time decreases exponentially with increasing number of basis functions. For example, for the circle potential with $n = 5$, it took 4073 steps to converge, whereas with $n = 15$, it took 291 steps to converge, and with $n = 99$, it took only 13 steps to converge. In addition, for the Mueller potential with $n = 99$, it took only 162 steps to converge. These results demonstrate that for appropriately sized Bernstein polynomial basis sets, the algorithm can be remarkable efficient. Examples of converged strings for the Mueller and circle potentials are shown in Fig. 3.

In addition to studying the convergence properties using fixed sized Bernstein polynomial basis sets, we used the degree elevation algorithm, described in Sec. IV F, to find MEPs on both the Mueller and circle potentials. We performed these studies to determine the effect of the degree elevation algorithm on the convergence time, as compared to the previous studies with fixed sized Bernstein polynomial basis sets. For these studies, we chose the same initial strings as before, but we also ran separate simulations using initial strings where the endpoints were not located at the minima in their corresponding basins (see Fig. 4). In addition, nearly all of the simulation parameters that were used in the control point convergence studies were adopted in these studies. The only difference is that we used an initial Bernstein polynomial basis set of $n = 3$ for each string, and the reparameterization of the strings was carried out every 50 time steps, demonstrating that reparameterization need not be carried out every time step. For both the Mueller potential and circle potential, the Δ_0 and C parameters were estimated from the first few time steps of the simulation by measuring the initial change in error, ΔE_0, and the rate of convergence, μ, respectively. We used values of Δ_0=0.1 and C=0.91 for both the Mueller and circle potentials for the simulations that used an initial string with endpoints at their corresponding minima. For the simulations using the initial string with endpoints not at their corresponding

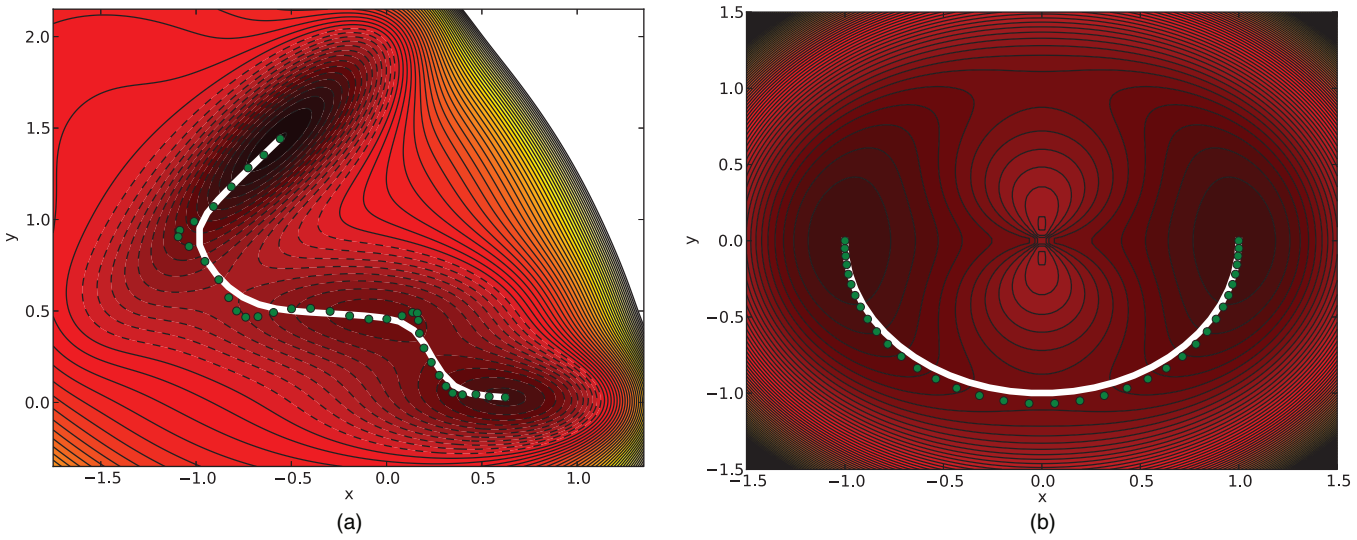


FIG. 3. Minimum energy paths for the (a) Mueller potential and (b) circle potential with control points overlaid in green. The number of control points is 35 for both (a) and (b).

minima, we kept the same value for C but changed the values of Δ_0 to $\Delta_0 = 0.01$ and $\Delta_0 = 0.0001$ for the Mueller and circle potentials, respectively. Using a smaller value for Δ_0 has little effect on the convergence properties of the algorithm since the convergence properties are more strongly influenced by the rate of convergence parameter C, but it does help increase the stability of the string when the endpoints are not located at their corresponding minima. The smaller values for Δ_0 delays the start of the degree elevation algorithm, and so, it allows time for the endpoints of the string to evolve toward the minima before the degree elevations begin. Note that this step is only necessary if the endpoints of the string are initially located very far away from the basins of attraction, as we have chosen to highlight here. If the endpoints of the string are initially in the basins of attraction, but not at the minima, this step is not necessary. In any case, a very simple and effective solution aside from the one we have adopted here would be to first evolve the string for a finite period of time with a small Bernstein polynomial basis set, and subsequently use the degree elevation algorithm to enhance the rate of convergence of the string once the endpoints of the string are near their corresponding minima.

In comparison to the previous study where we utilized a fixed number of basis polynomials, the results from these studies demonstrate that using the degree elevation property of Bézier curves speeds up convergence. For the simulations using an initial string with endpoints at their corresponding minima, the resulting convergence properties are shown in

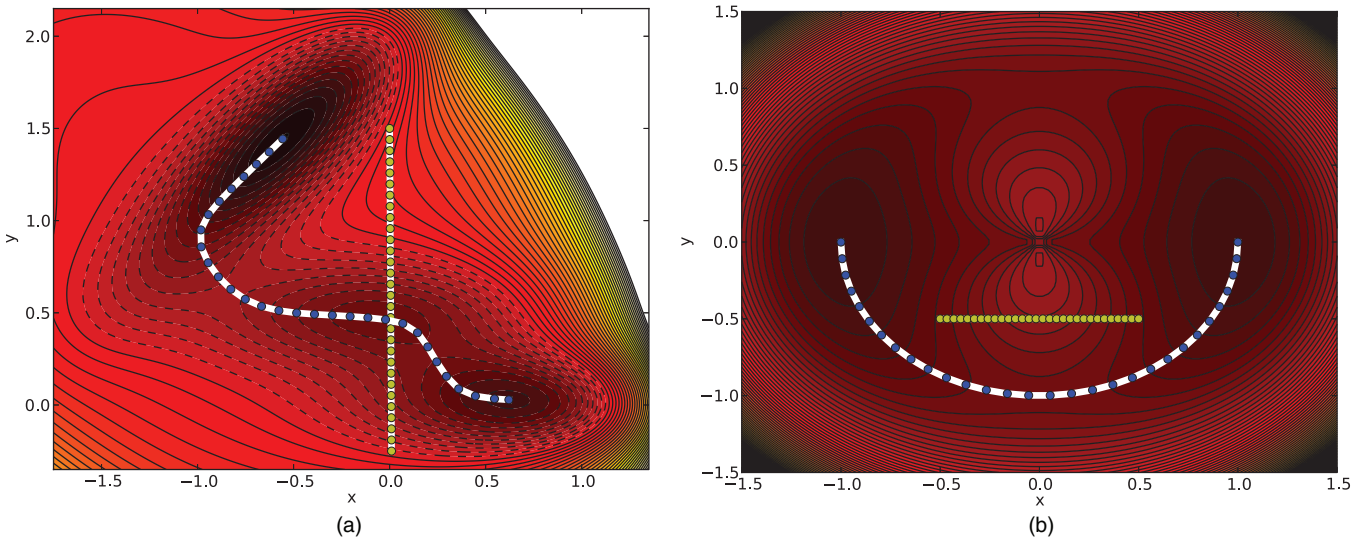


FIG. 4. Minimum energy paths for the (a) Mueller potential and (b) circle potential obtained from the Bézier curve string method using degree elevations with images (blue points). The initial string is overlaid (yellow points) for both (a) and (b).

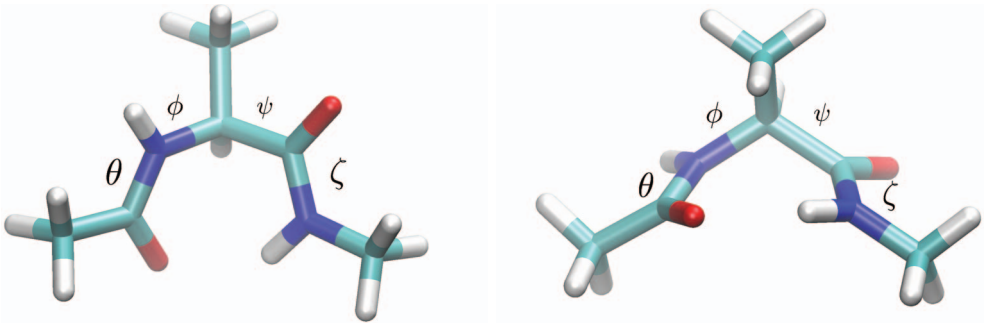
TABLE I. Comparison of the convergence properties of the degree elevation algorithm to that of the fixed basis set algorithm. The convergence is defined as the number of time steps to converge the Bézier string below a given tolerance *TOL*. For the degree elevation algorithm, each string initially had a Bernstein polynomial basis set of $n = 3$ and the number of basis functions reported above is the size of the Bernstein polynomial basis set at convergence.

	Mueller potential		Circle potential	
	Fixed basis set algorithm	Degree elevation algorithm	Fixed basis set algorithm	Degree elevation algorithm
Basis functions	81	81	39	39
Convergence	174	130	77	46

Table I. These results indicate that performing degree elevations at appropriate times throughout the evolution of the string can lead to faster convergence times than if one were to use a fixed sized basis set throughout. For the simulations using an initial string with endpoints far away from their corresponding minima, the Bézier string converged in 249 steps and resulted in a Bernstein polynomial basis set of $n = 71$ on the Mueller potential, and for the circle potential, the Bézier string converged in 744 steps and resulted in a Bernstein polynomial basis set $n = 38$. The longer convergence times are a result of the initial strings being much farther away from their corresponding MEPs, and so, the strings naturally must evolve for a longer period of time. Moreover, the endpoints of the string evolve with a steepest descent dynamics, which can become inefficient in basins with low curvature. The much longer convergence time in the circle potential is a reflection of this fact. However, the broader implication of these results is that the algorithm is capable of adaptively learning the size of the polynomial basis set that best represents the MEP, and therefore it alleviates the need for a knowledge of a suitable sized basis set *a priori*. Furthermore, since the bottleneck in any string method algorithm is the evaluation of the potential force, increasing the number of basis functions does not slow down the performance of the algorithm. This is particularly true for large systems, where the evaluation of the mean force and metric tensor are overwhelmingly the bottleneck in the performance of the algorithm.

VII. ALANINE DIPEPTIDE

In order to demonstrate that the Bézier curve string method is capable of finding MFEPs in chemical systems, we have analyzed the isomerization of the alanine dipeptide molecule at 298 K in vacuum. For this simple system, we have studied the transition between the two metastable conformers, C_{7eq} and C_{7ax}, which has been extensively studied in the literature with a variety of different methods.[6,8,21,22,34–37] The two metastable states of alanine dipeptide can be defined as local minima in the space of the two dihedral angles ϕ and ψ. The two conformers of alanine dipeptide are shown in Fig. 5, as well as the dihedral angles used in our study as collective variables. Of the two conformers in this system, the C_{7eq} conformer has the deepest minimum and is approximately located at $(\phi,\psi) = (-83.2,74.5)$, while the C_{7ax} conformer is approximately located at $(\phi,\psi) = (70,-70)$ in the two dihedral angle collective variable space. In our study, we have used the Bézier curve string method to find MFEPs in the space of two collective variables, (ϕ,ψ), and in the space of four collective variables $(\theta, \phi, \psi, \zeta)$. We have chosen to perform these studies using these collective variables for two reasons: to benchmark our method, since there are previous string method studies of this system using these collective variables,[21] and because it has been shown that the set of four dihedral angles, $(\theta, \phi, \psi, \zeta)$, sufficiently characterize the isomerization, whereas the set of two dihedral angles, (ϕ, ψ), does not.[21]



(a) (b)

FIG. 5. The C_{7eq} and C_{7ax} conformers of alanine dipeptide are shown in (a) and (b), respectively. The central carbon atom is referred to as C_α. All four dihedral angles used as collective variables are shown in both (a) and (b), and are defined as the dihedral angles between the following groups of atoms: (O, C, N, C_α) for θ, (C, N, C_α, C) for ϕ, (N, C_α, C, N) for ψ, and (C_α, C, N, H) for ζ.

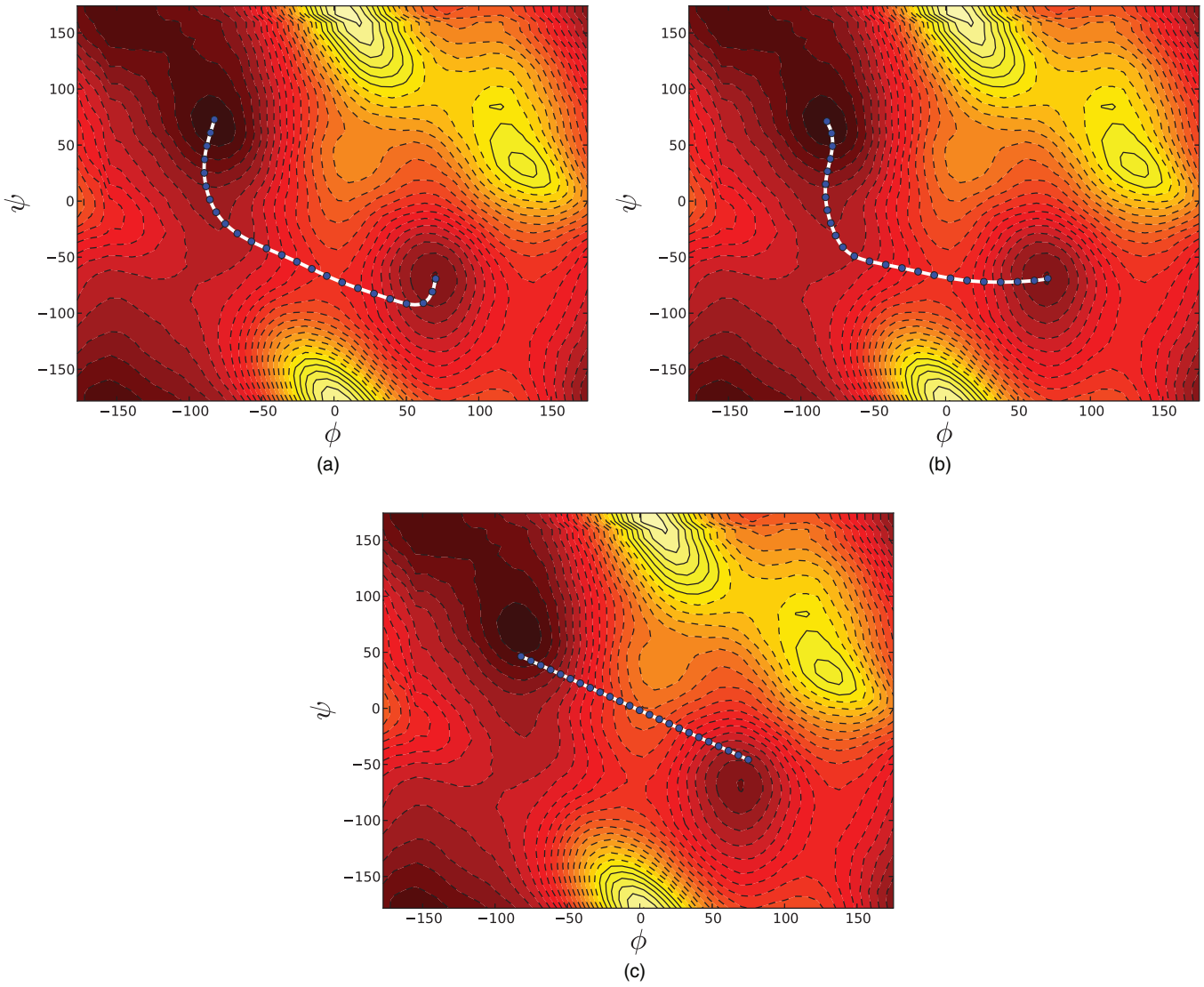


FIG. 6. Minimum free energy paths obtained from the Bézier curve string method using the (a) two dihedral angles (ϕ, ψ) and (b) four dihedral angles (θ, ϕ, ψ, ζ) as collective variables. Note that for graphical purposes, (b) is the projection of the string from the four dimensional space (θ, ϕ, ψ, ζ) into the two dimensional space (ϕ, ψ). (c) Initial string used in the search for a MFEP in both the two dihedral angle and four dihedral angle collective variable space. Note that the endpoints are not at the minima in their corresponding basins.

The simulations were performed within the framework of the DL_POLY molecular dynamics package.[38] In all simulations, we used an all-atom representation of alanine dipeptide in the CHARMM force field.[39,40] We used the DL_FIELD program to ensure that the correct CHARMM force field parameters were used within the DL_POLY molecular dynamics package. For the dynamics, we used a time step of 0.1 fs and a Nosé-Hoover thermostat to maintain the temperature at 298 K. The force field was extended to include harmonic potentials involving the collective variables, with force constants $k = 1000$ kcal/(mol rad^2). These potentials were used to perform restrained dynamics around each image of the string and compute the mean force. The DL_POLY package was modified to compute the collective variables as well as the quantities needed for the computation of the mean force and tensor $M(z)$. For each image on the string, both the mean force and tensor were computed as the ensemble average over 250 ps of molecular dynamics simulation.

The initial string used in the simulations for the MFEP is shown in Fig. 6(c). This initial string was used in the search for a MFEP in both the two dihedral angle and four dihedral angle collective variable space. The initial string was constructed by taking the C_{7eq} and C_{7ax} structures and performing a linear interpolation in the corresponding collective variable space, resulting in 24 images on the string. The ϕ and ψ dihedral angles were purposely perturbed by a random amount before performing the linear interpolation so that the initial string did not have endpoints exactly at the minima. For the initial string, we used an initial Bernstein polynomial basis set of $n = 10$, and we used values of $\Delta_0 = 0.25$ and $C = 0.91$ in the degree elevation algorithm. Once the initial string was set up, the procedure in Sec. IV was applied until the string converged on the MFEP. For the evolution of the string, we used a time step of $\Delta t = 0.5 \times 10^{-4}$ and performed the reparameterization of the string every time step. In order for the string to converge to the MFEP, about 100 updates were needed in two

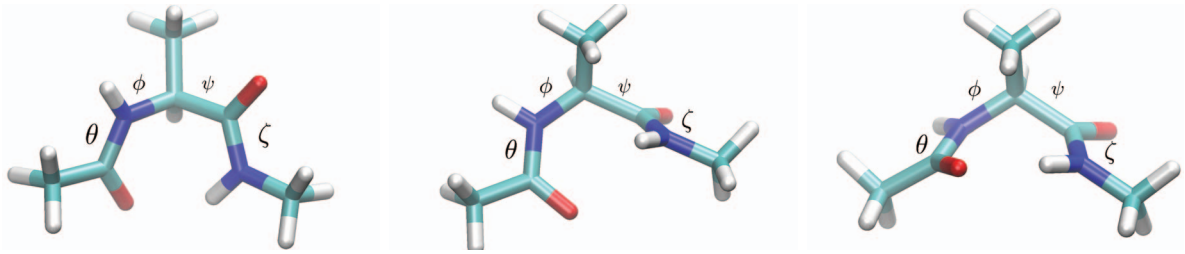


FIG. 7. The C_{7eq} and C_{7ax} conformers of alanine dipeptide are shown in (a) and (c), respectively, whereas the alanine dipeptide geometry at the $q(s) = 1/2$ point along the MFEP is shown in (b). Comparing these configurations, it is evident that the ψ dihedral angle at the transition state in (b) more closely resembles the ψ dihedral angle of the product state in (c), whereas the ϕ dihedral angle does not. This suggests that the ψ dihedral angle evolves on a faster time scale relative to the ϕ dihedral angle.

angles and about 150 in four angles, and in both cases, the resulting Bézier string had a Bernstein polynomial basis set of $n = 40$ at convergence. At the last update, 1 ns simulations were performed for each image on the string to minimize statistical error in $\nabla_z F(z)$ and $M(z)$.

The resulting MFEPs obtained from the Bézier curve string method in the two dihedral angle and four dihedral angle collective variable space are shown in Figs. 6(a) and 6(b), respectively. Since the MFEP in Fig. 6(b) represents a curve in the four dimensional space, $(\theta, \phi, \psi, \zeta)$, we have projected this curve into the two dimensional space (ϕ, ψ) for graphical purposes. Furthermore, the MFEPs in Fig. 6 are superimposed on the adiabatic potential energy landscape of alanine dipeptide for the collective variables, ϕ and ψ, since these collective variables should participate the most in the transition. This surface is defined as the minimum potential energy of the full system for fixed values of ϕ and ψ, and for this simple system, this surface governs the transition. For the MFEP defined in two dihedral angle collective variable space, we see that the path goes through the major saddle point on the surface and that our MFEP is in excellent agreement with previous strings and paths obtained for this system in two dimensional collective variable space.[21] However, we also note that the MFEP does not proceed orthogonally to the equipotential contour lines of the adiabatic potential energy landscape as should be the case if the (ϕ, ψ) collective variables completely described the transition. It has been shown that the true MFEP of the full dimensional system can be thought of as the centerline of a transition tube that carries the vast majority of probability current from one metastable state to another over the free energy landscape.[1,2] In the limit that the temperature of the system goes to zero, the transition tube should shrink to the centerline of this tube and become the MEP. Therefore, the true MFEP should coincide with the MEP on the adiabatic potential landscape, which in turn, should proceed in the orthogonal direction to the equipotential contour lines of the surface. This suggests that by defining the MFEP in two dihedral angles alone, we have integrated out important degrees of freedom in the free energy necessary to describe the transition. Consequently, the free energy surface defined solely by (ϕ, ψ) does not adequately capture the free energy of the full dimensional system. In contrast, from Fig. 6(b), we see that the MFEP in the space of four collective variables does proceed in the orthogonal direction of the equipotential contour lines of the adiabatic potential energy surface. This suggests that the free energy surface defined by $(\theta, \phi, \psi, \zeta)$ sufficiently captures the free energy of the full dimensional system. Furthermore, the MFEP in the space of the four collective variables is in very good agreement with the one calculated in Ref. 22, where the string was converged in the full dimensional Cartesian space. This further suggests that the four dihedral angles provide an accurate description of the transition between the metastable states. This also demonstrates how converging a string in a larger set of collective variables can lead to adequate recovery of the features of the full dimensional system necessary to describe the transition.

The results presented in Fig. 6 clearly demonstrate that the four dihedral angle collective variable space provides a better description of the transition between the C_{7eq} and C_{7ax} conformers, but does not give information about which of the four collective variables is most important in governing the transition. To determine which of the four collective variables is most responsible for inducing the transition, we first computed the committor probability along the MFEP from (29). In Fig. 7(b), the molecular geometry of alanine dipeptide corresponding to the $q(s) = 1/2$ point along the MFEP is shown. Comparing this structure to the C_{7eq} and C_{7ax} structures in Fig. 7, we see that the transition must take place by a faster rotation of the ψ dihedral angle relative to the ϕ dihedral angle, which intuitively indicates that the collective variable corresponding to the ϕ dihedral angle moves on a much slower time scale, and thus is the collective variable that governs the isomerization near the transition state. This behavior is evident when one analyzes values of ϕ and ψ along the MFEP in Fig. 6(b). However, what is not immediately evident is how the θ and ζ dihedral angles participate in the transition, or how important each collective variable is relative to one another. To gain insight into the latter, we have ranked the importance of the collective variables relative to one another using (31). From (31), we get (0.017,0.998,0.057,0.004) as the ranking vector, where each number in this vector corresponds to the collective variables in the order $(\theta, \phi, \psi, \zeta)$. As expected, the ranking vector correctly identifies the ϕ collective variable as the variable overwhelmingly responsible for inducing

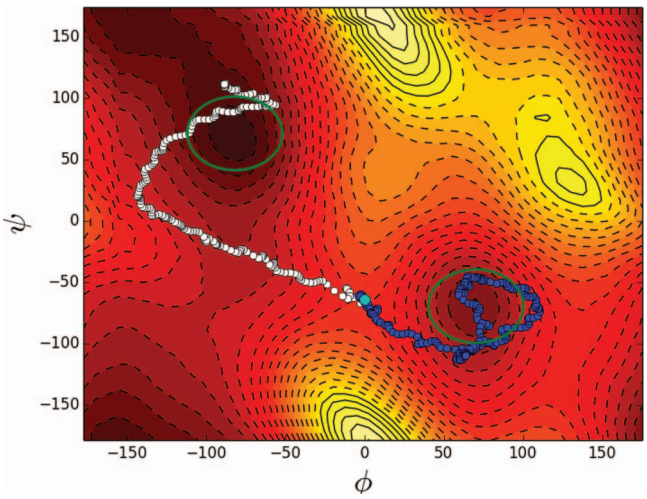


FIG. 8. Example of an aimless shooting trajectory for the alanine dipeptide system. The reactant and product regions on the potential are enclosed within the green circles, the blue path designates the forward path, the white path designates the backward path, and the larger turquoise point near the saddle point of the surface designates the shooting point.

the transition from C_{7eq} to C_{7ax}. This result can also be interpreted as meaning that near the $q(s) = 1/2$ transition state isosurface, only motion in the ϕ collective variable leads to any significant progress in the reaction. This can also be seen in Fig. 6(b), where it is apparent that the tangent vector (not shown) to the string near the saddle point is nearly parallel to the ϕ axis. It is important to note however that these results do not mean that the other dihedral angles play no role in the overall mechanism at different stages of the transition. For example, the faster rotation of ψ relative to ϕ demonstrates that the ψ dihedral angle is important in the early stages of the isomerization. Moreover, both the θ and ζ dihedral angles are important to adequately describe the overall molecular transition but their importance near the transition state is negligible in comparison to the ϕ dihedral angle. Due to the analytic representation of the Bézier curve, the ranking vector can be used to quantify the relative importance of collective variables at any stage of the molecular transition since it can be computed at any point along the MFEP. However, our discussion here is focused on the ranking vector near the transition state since we are interested in elucidating the collective variables that govern the isomerization.

To determine the accuracy of our ranking vector and reaction coordinate analysis proposed here, we have used an aimless shooting and maximum likelihood analysis for comparison. The aimless shooting and maximum likelihood analysis should provide a high quality test of our method since the aimless shooting sampling is based on unbiased trajectories in the full dimensional Cartesian space and the maximum likelihood approach utilizes the outcomes of these trajectories to construct the optimal reaction coordinate. The procedure we followed to implement the aimless shooting and maximum likelihood analysis can be found in Ref. 15. To begin, we have used our MFEP in four collective variables as the initial reactive trajectory needed for the aimless shooting algorithm. More specifically, we have used the configuration at the $q(s) = 1/2$ point along the string as our initial shooting point. We have defined the reactant and product basins as shown in Fig. 8, and we have integrated the forward and backward paths from each shooting point for 10 ps using a time step of 0.1 fs. From this sampling, we keep track of the outcomes of the forward and backward paths for each shooting point, i.e., whether or not the forward and backward paths passed through the designated regions on the potential landscape. For the maximum likelihood optimization, we used a linear reaction coordinate, as in (26), and optimized the parameters of the reaction coordinate with respect to the likelihood function such that the model function, $q(r) = 0.5(1 + \tanh(r))$, which is a function of the reaction coordinate, best fits the outcomes from aimless shooting. The aimless shooting procedure was repeated until the parameters and location of the optimized reaction coordinate determined from the likelihood maximization no longer changed with respect to increasing the size of the shooting point ensemble. The results of this procedure led to a shooting point ensemble with about 6000 configurations, and is shown in Fig. 9. In Fig. 9, we have overlaid the shooting point ensemble with the MFEP in the four dihedral angle collective variable space, as well as the isocommittor surfaces and reaction coordinates determined from the maximum likelihood method and Bézier curve string method. From Fig. 9, we see that the isocommittor surface and reaction coordinate determined from the Bézier curve string method is in excellent agreement with the isocommittor surface and reaction coordinate determined from the aimless shooting and maximum likelihood procedure. The reaction coordinate we have obtained from the maximum likelihood analysis has components of (−0.053, 0.966, −0.245, 0.057)

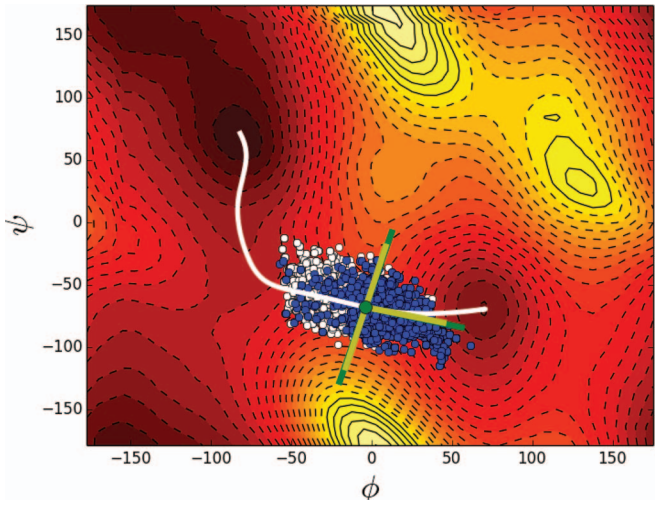


FIG. 9. Shooting point ensemble from the aimless shooting procedure overlaid with the MFEP (white path) in four collective variables (θ, ϕ, ψ, ζ) projected onto the two collective variable space (ϕ, ψ). The shooting points colored in blue resulted in forward trajectories that led to the C_{7ax} product basin, whereas the shooting points colored in white resulted in backward trajectories that led to the C_{7eq} reactant basin. The linear isocommittor surface (normal to the path) and reaction coordinate (tangent to the path) determined from the maximum likelihood procedure is colored in yellow, while the linear isocommittor surface (normal to the path) and reaction coordinate (tangent to the path) determined from the Bézier curve string method is colored in green. The $q(s) = 1/2$ point on the string is represented by the green point along the string. Note that the isocommittor surface and reaction coordinate in green is elongated for clarity.

after transforming the reaction coordinate hyperplane (26) to parametric form. By comparison, the reaction coordinate obtained from the Bézier curve string method has components of (−0.127, 0.954, −0.259, 0.066). In addition, the shooting point ensemble and forward trajectory outcomes as well as the control points that define the Bézier curve in the space of four collective variables have been made available in the supplementary material.[41] This agreement is significant because it demonstrates that the reaction coordinate analysis we have obtained from the Bézier curve string method, which is a biased method in a reduced-dimension collective variable space, provides equivalent information to that of the aimless shooting and maximum likelihood method, which is an unbiased method in the full Cartesian space. In addition, since the reaction coordinate we have obtained is accurate, the information provided by the ranking vector must also be reliable. This gives us confidence that the collective variable associated with the ϕ dihedral angle is the variable that governs the transition. The broader implication of this result is that the method we propose here is capable of overcoming the time scale issue associated with systems exhibiting rare events, and simultaneously determining the reaction coordinate and importance of collective variables relative to one another. However, it is hard to imagine a more simple system than the alanine dipeptide system, and so, to verify that our method is successful for large complex systems, we have applied it to the much more challenging problem of homogenous nucleation of benzene.

VIII. HOMOGENOUS NUCLEATION OF BENZENE

We have used our method to study crystallization, since this problem should provide a rigorous test of the effectiveness of our method at overcoming challenges inherent in large complex systems. For this study, we have chosen to use our method to model the homogenous nucleation of the Form I crystal of benzene from the melt since there are previous studies of this system using both aimless shooting and the string method in collective variables.[16,42]

In order to describe the liquid to solid phase transition, one necessarily needs a suitable set of collective variables that are capable of distinguishing liquid states from solid states. In general, this is a highly non-trivial problem, but has been extensively addressed in the paper by Santiso and Trout,[43] where it was shown that a set of collective variables suitable for describing liquid-solid phase transitions can be systematically defined for any system. In addition, it was shown that these collective variables are not only capable of distinguishing between liquid and solid states, but also between different polymorphs of crystals. Here we provide a general overview of these collective variables, and refer the reader to Ref. 43 for a detailed description.

The set of collective variables utilized in this study are based on a generalized pair distribution function, which for any given crystalline system at zero K, gives sharp peaks in specific regions of configuration space that can be used to uniquely describe a crystal. At finite temperature however, these signature peaks spread due to thermal motion in the system. In the method presented by Santiso and Trout, this peak spreading is approximated by a set of model probability density functions, and the pair distribution function at finite temperature is represented as a linear combination of a product of these model probability density functions. The probability density functions themselves are functions of internal coordinates such as distances, bond orientations, and relative orientations, and the parameters appearing in these models are optimized to reproduce information easily obtained from straightforward molecular dynamics simulations. In addition, the collective variables can be further refined to provide information about local crystalline order by discretizing the simulation box into a set of cells, and defining a set of collective variables within these cells. In doing so, the collective variables are sensitive enough to detect seeds forming in any region of space, and therefore, are extremely useful in determining where and when nucleation events take place.

To begin, we obtained the structure of the benzene Form I crystal (refcode: BENZEN) from the Cambridge Structural Database,[44] which corresponds to the experimental study of Bacon *et al.*[45] and is the stable polymorph of benzene at low temperatures and pressures.[46] The Form I crystal of benzene belongs to the space group *Pbca*, where the benzene molecules are arranged on an orthorhombic lattice in the unit cell. The values of the lattice parameters have been taken from the experimental crystal structure.[45] For our study, we used a system size of 720 molecules, which corresponds to 8640 atoms. We used an all-atom representation of benzene in the CHARMM22 force field[39,40] and all simulations were performed with the NAMD software package.[47] In addition, the NAMD software package was modified to compute the collective variables as well as the quantities needed for the computation of the mean force and tensor $M(z)$. The parameters for the model probability density functions used in the collective variables were taken from Ref. 43, and the simulation box was divided into a $5 \times 5 \times 5$ grid, resulting in 125 cells. The collective variables used in our simulation were pair distribution functions defined between pairs of benzene molecules, and were a function of bond orientation and of relative orientation variables. The bond orientation variable was defined as the angle between the center of mass vector between two benzene molecules and the normal vector to the molecular plane of the reference benzene molecule. The relative orientation variable was defined as the angle between the normal vectors of the molecular planes of both benzene molecules. Initial equilibration and energy minimization on the benzene system was carried out at 200 K and 1 bar using the NAMD software package. After equilibration, the system size was 45.13 Å $\times$ 48.32 Å $\times$ 41.84 Å. For each image on the string, the mean force and tensor were computed as the ensemble average over 350 ps in the NPT ensemble using a time step of 1 fs. A Langevin thermostat with a damping coefficient of 5 ps^{-1} was employed to maintain the temperature at 200 K, and a Langevin piston with a piston period of 100 fs and damping coefficient of 50 fs was utilized to maintain the pressure at 1 bar. Periodic boundary conditions were used, and long-range electrostatics were treated using a particle mesh Ewald (PME) summation.

In our study, we first converged the MFEP in the space of 125 collective variables, which were functions solely of the bond orientation collective variables. Upon convergence,

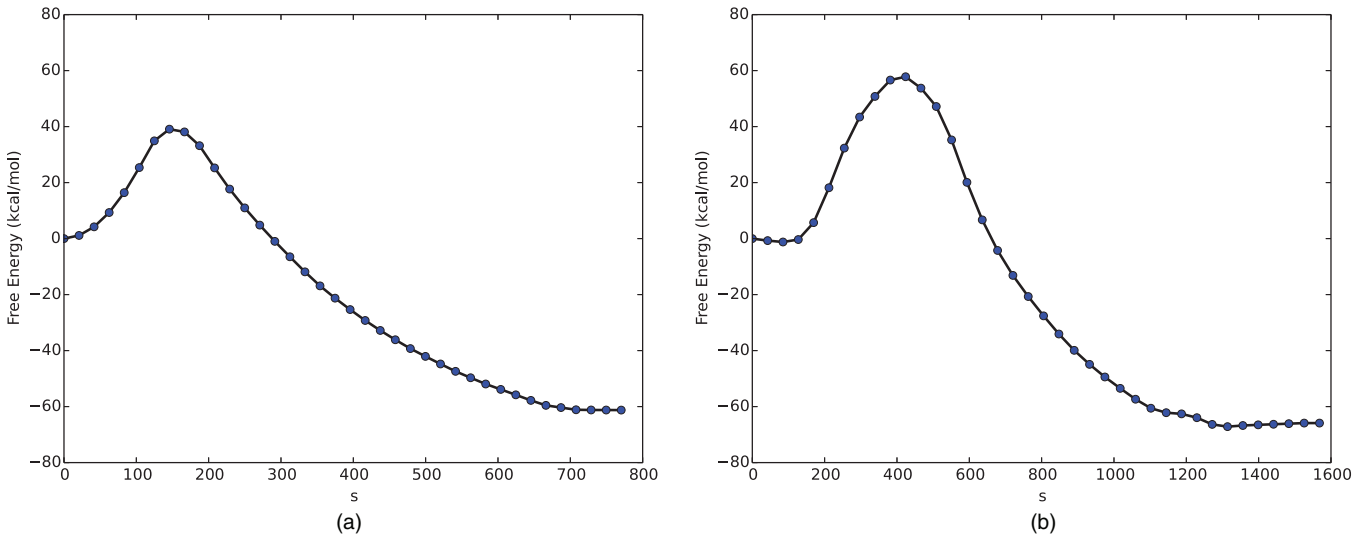


FIG. 10. Free energy profile for the MFEP in (a) bond orientation collective variable space and (b) bond orientation and relative orientation collective variable space as a function of arc length. The liquid state corresponds to the leftmost point, while the crystal state corresponds to the rightmost point in both (a) and (b).

we then increased the size of the collective variable space to 250; 125 collective variables that were a function of bond orientation and 125 collective variables that were a function of relative orientation. Each of the different types of collective variables, either bond orientation or relative orientation, occupied one of the 125 cells that resulted from discretization of the simulation box. The initial string in our study was obtain by first slowly melting the Form I crystal over a very long molecular dynamics trajectory, and then subsequently fitting each component of the Bézier curve to the corresponding collective variable time series using a Bernstein polynomial basis set of $n = 35$. Once the initial string was fit, we discretized the string into $m = 38$ equal arc length images and chose configurations from the melting trajectory that best matched the values of the collective variables at these images. Once the initial string was set up, the procedure in Sec. IV was applied until the string converged on the MFEP. For the evolution of the string, we used a time step of $\Delta t = 0.01$, we performed the reparameterization of the string every time step, and we used values of $\Delta_0 = 0.35$ and $C = 0.91$ in the degree elevation algorithm. In order for the string to converge to the MFEP, about 150 updates were needed using the bond orientation collective variables and about 75 in both the bond orientation and relative orientation collective variables. The much shorter convergence time in the bond orientation and relative orientation collective variable space indicates that the converged MFEP in the bond orientation collective variable space is close to the MFEP in the bond orientation and relative orientation collective variable space. In both cases, the resulting Bézier string had a Bernstein polynomial basis set of n=110 at convergence. At the last update, 1 ns simulations were performed for each image on the string to minimize statistical error in $\nabla_z F(z)$ and $M(z)$.

The resulting MFEPs clearly cannot be plotted due to the high dimensionality of the system, but we can analyze the resulting free energy profiles associated with each path, which are shown in Fig. 10. From Fig. 10, we see that in both cases, the crystal structure has a lower free energy than the liquid state at 200 K as expected since the crystal structure is the stable state at 200 K. We also note that our free energy profile in the bond orientation collective variable space is in excellent agreement with a previous study of this system utilizing the string method and the bond orientation collective variables.[42] However, it is quite evident that the free energy profiles are different from one another. In each case, the free energy difference between the liquid and solid states are nearly identical, but the free energy maximum along each MFEP is different. For the MFEP in bond orientation collective variable space, the free energy barrier to crystallization is about 40 kcal/mol, whereas for the MFEP in bond orientation and relative orientation collective variable space, the free energy maximum is about 60 kcal/mol. This indicates that the MFEP in bond orientation collective variable space has an underestimated free energy barrier relative to the MFEP in bond orientation and relative orientation collective variable space. Therefore, the use of bond orientation collective variables alone is not a large enough set of collective variables to describe the transition. Evidence for this is also apparent when we computed the ranking vector for each MFEP near the transition state isosurface to determine which collective variables were most important for the transition. If the relative orientation collective variables were not participating in the transition, then the majority of components in the ranking vector corresponding to relative orientation collective variables should have magnitude much less than that of the components of the bond orientation collective variables. However, this is not what we observe when we perform this analysis. In Fig. 11, a histogram of the magnitudes of the components in the ranking vector is shown as a function of the cell number. In this figure, the magnitudes associated with bond orientation collective variables are colored in blue, whereas the magnitudes of the relative orientation variables are colored in red.

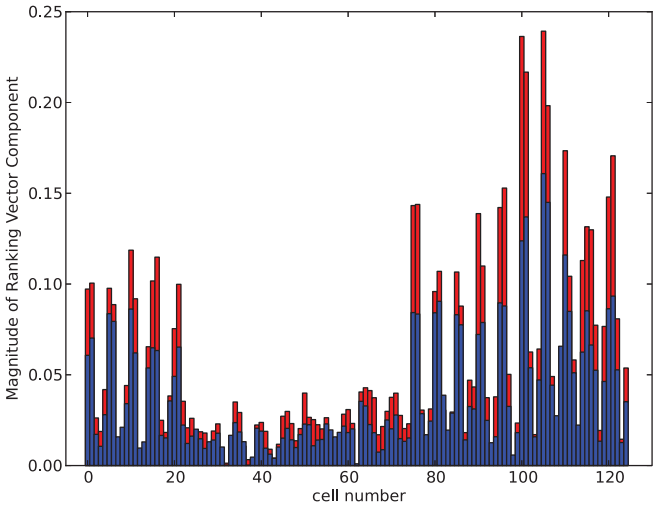


FIG. 11. Magnitude of components of ranking vector as a function of cell number. Bond orientation values are colored in blue, whereas relative orientation values are colored in red.

In each case, we see that the components associated with relative orientation collective variables are larger than the components associated with the bond orientation collective variables in nearly every cell within the simulation box, indicating that these collective variables are more important for describing the transition. This is in accord with results reported by Shah *et al.*,[16] where an aimless shooting and maximum likelihood analysis was performed on the homogenous nucleation of benzene from the melt. In this study, the maximum likelihood analysis demonstrated that from a very large set of candidate collective variables, reaction coordinate models with the relative orientation collective variable consistently yielded much higher likelihood scores.

After computing the ranking vector, we have analyzed the configurations near the $q(s) = 1/2$ transition state isosurface for each MFEP. Figure 12 shows the configurations near the $q(s) = 1/2$ transition state isosurface along each MFEP. To demonstrate the effectiveness of the ranking vector at finding important collective variables, we have colored the spatial cells in red in Fig. 12 where the corresponding collective variables have large components in the ranking vector. In each case, we define a large component to be greater than 0.1. Taking into consideration the periodic boundary conditions in the system, it is very clear that all these cells lie adjacent to one another in space. It is a good sign that the collective variables identified as important belong to spatially localized crystalline clusters rather than random cells spread across the system. In addition, the fact that the ranking vector identified regions where crystalline clusters have formed instead of regions with liquid-like structure gives us confidence that it correctly identifies collective variables and regions of space that cause the liquid to solid phase transition. We also see from Fig. 12 that the transition state configuration in both bond orientation and relative orientation collective variable space has clusters with a much higher degree of local crystalline order than the clusters present in the configuration obtained from bond orientation collective variables alone. This indicates that the inclusion of relative orientation collective variables helps the system organize into to well formed crystalline clusters near the transition state, which again illustrates their importance in governing the transition. Taking our analysis a step further, we have extracted the structures of the nucleation seeds by locating the cells with the highest components in the ranking vector that are adjacent to one another. In both cases, this procedure identified five cells that have components in the ranking vector that are well separated in terms of magnitude from all other components. For both collective variables, the resulting seed structures were found to be in cells 100, 101, 105, 106, and 110. For the identified nucleation seeds, the clusters were composed of 16 molecules and 28 molecules. We note that the latter of these cluster sizes is very similar in size to the average critical cluster size of 32 reported by Shah *et al.*[16]

Despite the fact that the relative orientation collective variables govern the nucleation near the transition state, the

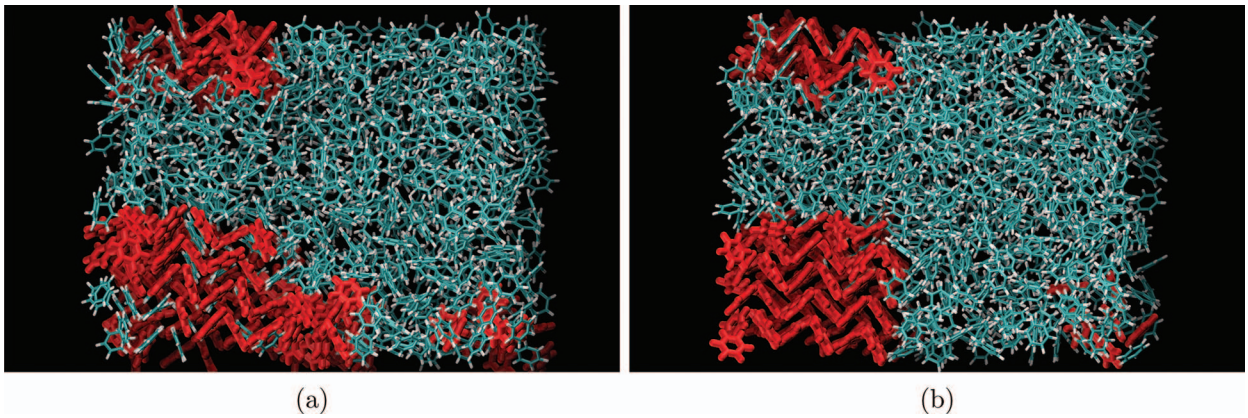


FIG. 12. (a) Configuration corresponding to the $q(s) = 1/2$ point on the MFEP in bond orientation collective variable space. (b) Configuration corresponding to the $q(s) = 1/2$ point on the MFEP in bond orientation and relative orientation collective variable space. The larger benzene molecules colored in red indicate cells where the corresponding collective variable has a large component in the ranking vector. These cells indicate where the reaction coordinate is changing the most and are the regions in space where the benzene molecules are organizing themselves into the Form I crystal. Note that the configuration in (b) has better local crystalline order at the transition state than the configuration in (a).

optimal reaction coordinate in this study was determined to be a combination of bond orientation and relative orientation collective variables. Indeed, both of these collective variables are important for describing the mechanism of the phase change, which is consistent with the results presented in Ref. 16. In particular, analysis of the ranking vector in the early stages of the transition demonstrates that the bond orientation collective variables are more important than the relative orientation collective variables. This trend continues until the system approaches the transition state and the magnitudes of the components of the relative orientation collective variables in the ranking vector suddenly become much larger than those of the bond orientation collective variables and continue to dominate until the crystal is fully formed. This behavior could signify that the bond orientation collective variables are better capable of describing a semi-ordered dense phase in comparison to the relative orientation collective variables. However, the relative orientation collective variables are much more sensitive to crystalline order, and so, once the semi-ordered crystalline phase is formed, the relative orientation collective variables provide a strong driving force for crystallization. This is consistent with the large sudden change in the ranking vector near the transition state that persists until the crystal is fully formed.

IX. CONCLUSION

In this research, we have developed an approach for finding MEPs and MFEPs in complex chemical systems exhibiting rare event transitions. Our approach utilizes Bézier curves, which are analytic parametric curves defined as a linear combination of Bernstein basis polynomials, and are capable of accurately approximating any continuous algebraic curve. We have shown that by using the degree elevation property of Bézier curves, our algorithm is capable of adaptively learning the size and degree of the Bernstein basis set necessary to reproduce MEPs or MFEPs, circumventing the need for estimating these quantities *a priori*. In addition, the use of the degree elevation property increases the stability of the string as it evolves toward the most probable transition path, and simultaneously increases the convergence speed of the algorithm. The algorithm proposed here evolves the string by a forward Euler integration, but we note that more sophisticated optimization algorithms can be used to accelerate the convergence of the string.

In addition to developing the Bézier curve string method algorithm, we have devised a ranking vector for the purpose of determining which collective variables are most important for governing chemical transitions. In general, this information is more elucidating than finding the reaction coordinate itself because it provides information about which collective variables are most likely to initiate the molecular transition. To illustrate the utility of the ranking vector, and to demonstrate that the reaction coordinate obtained from the Bézier curve string method is consistent with the reaction coordinate obtain from an aimless shooting and maximum likelihood procedure, we applied our method to the isomerization of the alanine dipeptide molecule. Our analysis indicates that the use of two dihedral angles, (ϕ, ψ), as collective variables does not completely capture the mechanism of the transition, whereas the use of four dihedral angles, $(\theta, \phi, \psi, \zeta)$, does adequately capture the mechanism of the transition. Moreover, the MFEPs obtained using our method in both two dihedral angle collective variable space and four dihedral angle collective variable space are consistent with previous studies[21,22] using the string method. After convergence of the MFEPs, we have used the ranking vector to identify the ϕ dihedral angle as the collective variable most responsible for governing the conformational transition. This is consistent with the fact that we observe the ϕ dihedral angle evolving on the slowest of timescales along the MFEP. Since the collective variable corresponding to the ϕ dihedral angle evolves on the slowest of timescales, it is the rate-limiting variable, and therefore governs the transition. In addition, we have also demonstrated that the reaction coordinate and isocommittor surface obtained from our method are equivalent to the reaction coordinate and isocommittor surface obtained from an aimless shooting and maximum likelihood analysis. This is significant because it demonstrates that the reaction coordinate we have obtained from the Bézier curve string method, which is a biased method in a reduced-dimension collective variable space, provides equivalent information to that of the aimless shooting and maximum likelihood method, which is an unbiased method in the full Cartesian space.

Finally, to test our method in a large complex chemical system, we have applied it to study the homogenous nucleation of benzene from the melt. The study of crystallization using all atom simulations is challenging for a variety of reasons, but the main challenges are due to the diffusive nature of the crystallization process, the prevalence of large free energy barriers and rugged free energy landscapes, and the difficulty associated with defining suitable collective variables to describe the transition. Consequently, testing our method for such a demanding problem provides evidence for its effectiveness and applicability. For our simulations, we adopted the collective variables described in Ref. 43. In our study, we converged two MFEPs for this system using a large set of collective variables: the first set consisted of 125 bond orientation collective variables and the second set consisted of 250 collective variables, of which 125 were bond orientation collective variables and 125 were relative orientation collective variables. The free energy profile obtained from our method is consistent with the free energy profile reported in Ref. 42, where the string method in collective variables was applied to the benzene system using the bond orientation collective variables, giving us confidence in our converged MFEPs. Our analysis of these paths indicates that the use of both bond orientation and relative orientation collective variables was necessary to sufficiently capture the mechanism of the phase transition. However, we also demonstrated that the relative orientation collective variables were more important for characterizing the mechanism. The importance of the relative orientation collective variable in describing the transition is consistent with a previous study[16] on the benzene system using an aimless shooting and maximum likelihood analysis, which demonstrated that reaction coordinate models with the relative orientation collective variable consistently yielded much higher likelihood scores. After convergence of the MFEPs,

we used the ranking vector to identify a subset of collective variables that govern the phase transition. It was shown that all of these collective variables lie adjacent to one another in space, and are part of spatially localized crystalline clusters where the nucleation event takes place. The fact that the ranking vector identified regions in configuration space where crystalline clusters formed instead of regions with liquid-like structure signifies that it correctly identified collective variables and regions of space that govern the liquid to solid phase transition.

The method we propose here is capable of overcoming the time scale issue associated with systems exhibiting rare events, and simultaneously determining the reaction coordinate and importance of collective variables relative to one another. We are currently applying it to more complicated nucleation problems in an effort to understand the underlying mechanisms inherent in these systems. This will be the subject of our future work.

ACKNOWLEDGMENTS

The authors thank Erik Santiso and Geoff Wood for useful discussions. In addition, we thank Erik Santiso for the C++ code that implements the crystallization collective variables into NAMD. The authors would like to kindly acknowledge support from Novartis through the Novartis-MIT Center for continuous manufacturing.

APPENDIX A: OPTIMAL CONTROL POINTS FOR INITIAL STRING

Although the Bernstein polynomials are not orthonormal, we can find the optimal control points to approximate the initial string, in the least squares sense, with the following equation:

$$P = (B^T B)^{-1} B^T z0. \tag{A1}$$

However, as the number of basis functions increase the $B^T B$ matrix can become ill-conditioned, and so, numerical instabilities associated with finding its inverse can lead to highly oscillatory, or sporadic control points. Highly oscillatory control points should be avoided because they can affect the accuracy of the tangent vectors. As a simple visual check of the accuracy of the tangent vectors, one can plot each component of the Bézier curve fit and the associated tangent vectors at each point as a function of α.

A much more numerically stable way of solving for the control points is to first find the QR decomposition of the matrix B, where Q and R are defined as

$$Q = (e_0, \ldots, e_n), \tag{A2}$$

$$R = \begin{pmatrix} e_0 \cdot B_{n,0} & e_0 \cdot B_{n,0} & e_0 \cdot B_{n,0} & \cdots \\ 0 & e_1 \cdot B_{n,1} & e_1 \cdot B_{n,1} & \cdots \\ 0 & 0 & e_2 \cdot B_{n,2} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{A3}$$

The matrix Q is an orthonormal matrix and the matrix R is an upper triangular matrix where the components are the dot products between the orthonormal basis functions and the Bernstein polynomial basis functions. Substituting $B = QR$ in (14) and using the fact that $Q^T Q = I$, we get

$$RP = Q^T z0. \tag{A4}$$

Since R is an upper triangular matrix, we can find the control points by recursively back solving from P_n to P_0 with the following equation:

$$P_i = \frac{1}{R_{i,i}} \left(\sum_{j=0}^{n} Q_{j,i} z0_j - \sum_{j=i+1}^{n} R_{i,j} P_j \right), \quad i = n, \ldots, 1. \tag{A5}$$

Even for large basis sets, this procedure can be used to find the initial Bézier string without numerical instability. This procedure should be repeated, slowly increasing the degree of the polynomial, n, until the Bézier string best approximates the initial string.

APPENDIX B: TIME-DEPENDENT CONTROL POINTS

The MFEP is the stationary solution of the following dynamics:

$$\dot{z}(\alpha, t) = -M(z) \nabla_z F(z)^{\perp}. \tag{B1}$$

To derive Eq. (15), we first make the following change of variables for the control points:

$$U_k = P_k + \sum_{j \neq k}^{n} \frac{B_{n,k} \cdot B_{n,j}}{B_{n,k} \cdot B_{n,k}} P_j, \tag{B2}$$

where U_k is the new kth control point. The reason for this change of variables will become clear in the discussion below. Letting the control points be time-dependent, from (B1), we get

$$B\dot{U} = -M(z) \nabla_z F(z)^{\perp}, \tag{B3}$$

which in the case of a forward Euler integration yields

$$BD = -M(z) \nabla_z F(z)^{\perp} \Delta t, \tag{B4}$$

$$D = U^{h+1} - U^h. \tag{B5}$$

To find U^{h+1}, we need to minimize the sum of squares error

$$E = \sum_{i=1}^{m} \left(\sum_{j=0}^{n} B_{n,j}(\alpha_i) D_j + M(z(\alpha_i)) \nabla_z F(z(\alpha_i))^{\perp} \Delta t \right)^2, \tag{B6}$$

where the first summation is over the m images on the string and the second summation is over the n Bernstein polynomials. For clarity we drop the first sum and α parameter with the understanding that the procedure below holds for every image on the string. To minimize E, we take the derivative and set it equal to zero,

$$\frac{\partial E}{\partial U_k^{h+1}} = 2 B_{n,k} \left(\sum_{j=0}^{n} B_{n,j} D_j + M(z) \nabla_z F(z)^{\perp} \Delta t \right) = 0. \tag{B7}$$

Substituting in (B5) and simplifying, we get

$$\sum_{j=0}^{n} B_{n,k} B_{n,j} U_j^{h+1} = \sum_{j=0}^{n} B_{n,k} B_{n,j} U_j^{h} - B_{n,k} M(z) \nabla_z F(z)^{\perp} \Delta t. \quad \text{(B8)}$$

This equation is seemingly problematic because in order to find the time-evolved control point U_k^{h+1}, one necessarily needs to know all other time-evolved control points due to the presence of the cross terms in the above expression. The cross terms arise in the above expression because the Bernstein polynomials are not orthonormal. However, making use of the change of variables in (B2), we can greatly simplify this equation. Using the change of variables (B2) in the first term of (B8), we get

$$\sum_{j=0}^{n} B_{n,k} B_{n,j} \left(P_j^{h+1} + \sum_{l \neq j}^{n} \frac{B_{n,j} B_{n,l}}{B_{n,j} B_{n,j}} P_l^{h+1} \right). \quad \text{(B9)}$$

Expanding the summation and simplifying, we get

$$B_{n,k} B_{n,k} \left(P_k^{h+1} + \sum_{j \neq k}^{n} \frac{B_{n,k} B_{n,j}}{B_{n,k} B_{n,k}} P_j^{h+1} \right) + \sum_{j \neq k}^{n} B_{n,k} B_{n,j} \left(P_j^{h+1} + \sum_{l \neq j}^{n} \frac{B_{n,j} B_{n,l}}{B_{n,j} B_{n,j}} P_l^{h+1} \right), \quad \text{(B10)}$$

$$B_{n,k} B_{n,k} P_k^{h+1} + 2 \sum_{j \neq k}^{n} B_{n,k} B_{n,j} P_j^{h+1} + \sum_{j \neq k}^{n} \sum_{l \neq j}^{n} B_{n,k} B_{n,l} P_l^{h+1}. \quad \text{(B11)}$$

The last term in (B11) can be further simplified as

$$\sum_{j \neq k}^{n} \sum_{l \neq j}^{n} B_{n,k} B_{n,l} P_l^{h+1} = (n-1) B_{n,k} B_{n,k} P_k + (n-2) \sum_{j \neq k}^{n} B_{n,k} B_{n,j} P_j^{h+1}. \quad \text{(B12)}$$

Substituting (B12) into (B11) and simplifying, we get

$$n \left(B_{n,k} B_{n,k} P_k^{h+1} + \sum_{j \neq k}^{n} B_{n,k} B_{n,j} P_j^{h+1} \right) = n B_{n,k} B_{n,k} U_k^{h+1}. \quad \text{(B13)}$$

Since the same arguments above apply to the second term in (B8), this equation simplifies to the following:

$$n U_k^{h+1} = n U_k^{h} - \frac{B_{n,k} M(z) \nabla_z F(z)^{\perp} \Delta t}{B_{n,k} B_{n,k}}. \quad \text{(B14)}$$

Dividing both sides by n, letting $\Delta t' = \Delta t / n$, and including contributions from all images over the string, we get the following update rule:

$$U_k^{h+1} = U_k^{h} - \frac{B_{n,k} \cdot M(z) \nabla_z F(z)^{\perp} \Delta t'}{B_{n,k} \cdot B_{n,k}}. \quad \text{(B15)}$$

This update rule for the control points is remarkably simple, computationally inexpensive, and it completely avoids any kind of numerical instability associated with inverting matrices. In addition, in regard to the change of variables in (B2), since we could have just as easily let

$$P_k = W_k + \sum_{j \neq k}^{n} \frac{B_{n,k} \cdot B_{n,j}}{B_{n,k} \cdot B_{n,k}} W_j, \quad \text{(B16)}$$

then the change of variables itself is arbitrary since we can in principle always find some set W that satisfies (B16). Consequently, we can revert back to our original notation, which leads to (15) as the update rule for the control points.

[1] W. E, W. Ren, and E. Vanden-Eijnden, Chem. Phys. Lett. **413**, 242 (2005).
[2] W. E and E. Vanden-Eijnden, Annu. Rev. Phys. Chem. **61**, 391 (2010).
[3] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, Annu. Rev. Phys. Chem. **53**, 291 (2002).
[4] C. Dellago, P. G. Bolhuis, and P. L. Geissler, Adv. Chem. Phys. **123**, 1 (2002).
[5] G. Hummer, J. Chem. Phys. **120**, 516 (2004).
[6] A. Ma and A. R. Dinner, J. Phys. Chem. B **109**, 6769 (2005).
[7] C. Dellago, P. G. Bolhuis, and D. Chandler, J. Chem. Phys. **108**, 9236 (1998).
[8] P. G. Bolhuis, C. Dellago, and D. Chandler, Proc. Natl. Acad. Sci. U.S.A. **97**, 5877 (2000).
[9] D. Zahn, J. Phys. Chem. B **111**, 5249 (2007).
[10] D. Zahn and S. Leoni, Phys. Rev. Lett. **92**, 250201 (2004).
[11] P. Varilly and D. Chandler, J. Phys. Chem. B **117**, 1419 (2013).
[12] R. Radhakrishnan and B. L. Trout, J. Chem. Phys. **117**, 1786 (2002).
[13] R. Radhakrishnan and B. L. Trout, J. Am. Chem. Soc. **125**, 7743 (2003).
[14] B. Peters and B. L. Trout, J. Chem. Phys. **125**, 054108 (2006).
[15] B. Peters, G. T. Beckham, and B. L. Trout, J. Chem. Phys. **127**, 034109 (2007).
[16] M. Shah, E. E. Santiso, and B. L. Trout, J. Phys. Chem. B **115**, 10400 (2011).
[17] G. Beckham, B. Peters, C. Starbuck, N. Variankaval, and B. L. Trout, J. Am. Chem. Soc. **129**, 4714 (2007).
[18] G. Beckham, B. Peters, and B. L. Trout, J. Phys. Chem. B **112**, 7460 (2008).
[19] W. E, W. Ren, and E. Vanden-Eijnden, Phys. Rev. B **66**, 052301 (2002).
[20] W. E, W. Ren, and E. Vanden-Eijnden, J. Chem. Phys. **126**, 164103 (2007).
[21] L. Maragliano, A. Fischer, E. Vanden-Eijnden, and G. Ciccotti, J. Chem. Phys. **125**, 024106 (2006).
[22] W. Ren, E. Vanden-Eijnden, P. Maragakis, and W. E, J. Chem. Phys. **123**, 134109 (2005).
[23] W. E, W. Ren, and E. Vanden-Eijnden, J. Phys. Chem. B **109**, 6688 (2005).
[24] E. Vanden-Eijnden and M. Venturoli, J. Chem. Phys. **130**, 194103 (2009).
[25] T. F. Miller III, E. Vanden-Eijnden, and D. Chandler, Proc. Natl. Acad. Sci. U.S.A. **104**, 14559 (2007).
[26] C. Z. Zhang and Z. G. Wang, Phys. Rev. E **77**, 021906 (2008).
[27] C. Qiu, T. Qian, and W. Ren, J. Chem. Phys. **129**, 154711 (2008).
[28] X. Cheng, L. Lin, W. E, P. Zhang, and A. C. Shi, Phys. Rev. Lett. **104**, 148301 (2010).
[29] R. Farouki, Comput. Aided. Geom. D **29**, 379 (2012).
[30] S. Pal, P. Ganguly, and P. K. Biswas, Pattern Recogn. **40**, 2730 (2007).
[31] Y. J. Ahn and H. O. Kim, J. Comput. Appl. Math. **81**, 145 (1997).
[32] M. B. Egerstedt and C. F. Martin, IEEE Trans. Automat. Contr. **49**, 1728 (2004).
[33] J. Wu, Appl. Math. Comput. **219**, 3655 (2012).
[34] B. M. Pettitt and M. Karplus, Chem. Phys. Lett. **121**, 194 (1985).
[35] R. Czerminski and R. Elber, J. Chem. Phys. **92**, 5580 (1990).
[36] T. Lazaridis, D. J. Tobias, C. L. Brooks III, and M. E. Paulaitis, J. Chem. Phys. **95**, 7612 (1991).
[37] D. J. Tobias and C. L. Brooks III, J. Phys. Chem. **96**, 3864 (1992).
[38] I. T. Todorov, W. Smith, K. Trachenko, and M. T. Dove, J. Mater. Chem. **16**, 1911 (2006).
[39] A. MacKerell, J. Wiorkiewicz-Kuczera, and M. Karplus, J. Am. Chem. Soc. **117**, 11946 (1995).
[40] A. MacKerell, D. Bashford, M. Bellott, R. Dunbrack, J. Evanseck, M. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. Lau, C. Mattos, S. Michnick, T. Ngo, D. Nguyen, B.

Prodhom, W. Reiher, B. Roux, M. Schlenkrich, J. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus, J. Phys. Chem. B **102**, 3586 (1998).
[41]See supplementary material at http://dx.doi.org/10.1063/1.4893216 for the shooting point ensemble, forward trajectory outcomes, and Bézier string control points.
[42]E. E. Santiso and B. L. Trout, "A general method for molecular modeling of nucleation from the melt," J. Chem. Phys. (submitted).
[43]E. E. Santiso and B. L. Trout, J. Chem. Phys. **134**, 064109 (2011).
[44]F. H. Allen, Acta Crystallogr. B: Struct. Sci. **58**, 380 (2002).
[45]G. Bacon, N. Curry, and S. Wilson, Proc. R. Soc. London A **279**, 98 (1964).
[46]P. Raiteri, R. Martonak, and M. Parrinello, Angew. Chem., Int. Ed. Engl. **44**, 3769 (2005).
[47]J. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. Skeel, L. Kale, and K. Schulten, J. Comput. Chem. **26**, 1781 (2005).