SGOOP-d: Estimating kinetic distances and reaction coordinate dimensionality for rare event systems from biased/unbiased simulations

Sun-Ting Tsai,[1] Zachary Smith,[2] and Pratyush Tiwary*[3, a]

[1] *Department of Physics and Institute for Physical Science and Technology, University of Maryland, College Park 20742, USA.*

[2] *Biophysics Program and Institute for Physical Science and Technology, University of Maryland, College Park 20742, USA.*

[3] *Department of Chemistry and Biochemistry and Institute for Physical Science and Technology, University of Maryland, College Park 20742, USA.*

Understanding kinetics including reaction pathways and associated transition rates is an important yet difficult problem in numerous chemical and biological systems especially in situations with multiple competing pathways. When these high-dimensional systems are projected on low-dimensional coordinates often needed for enhanced sampling or for interpretation of simulations and experiments, one often ends up losing the kinetic connectivity of the underlying high-dimensional landscape. Thus in the low-dimensional projection metastable states might appear closer or further than they actually are. To deal with this issue, in this work we develop a formalism that learns a multi-dimensional yet minimally complex reaction coordinate (RC) for generic high-dimensional systems, such that when projected along this RC, all possible kinetically relevant pathways can be demarcated and the true high-dimensional connectivity is maintained. One of the defining attributes of our method lies in that it can work on long unbiased simulations as well as biased simulations often needed for rare event systems. We demonstrate the utility of the method by studying a range of model systems including conformational transitions in a small peptide Ace-Ala₃-Nme, where we show how SGOOP-derived two-dimensional and three-dimensional reaction coordinate can capture the kinetics for 23 and all 28 out of the 28 dominant state-to-state-transitions respectively.

I INTRODUCTION

It has been a problem of longstanding theoretical and practical interest to model reaction pathways and transition mechanisms in generic chemical and biological systems.[1–7] Due to recent progress in high performance computing, brute-force Molecular Dynamics (MD) simulations with all-atom resolution have enabled a possible way to do such analysis in femtosecond temporal and all-atom spatial precision, making it a useful tool for studying diverse phenomena. However this leads to a deluge of data resulting from explicit enumeration of all atomic coordinates over a very large number of MD timesteps. To make sense of such high-dimensional trajectories resulting from MD, it is a common practice to project them along low-dimensional coordinates identified with one of many dimensionality reduction schemes.[8–11] However, more often that not in such schemes one ends up losing the kinetic connectivity of the high-dimensional landscape. This can thus lead to incorrect interpretation of MD trajectories, for example making molecular conformations appear closer to each other than they are and obfuscating interconversion pathways between them.[12]

In this work, we develop a formalism that learns a multi-dimensional yet minimally complex reaction coordinate (RC), such that when projected along this RC, all possible kinetically relevant pathways can be demarcated and the true high-dimensional connectivity is maintained. The central idea is to calculate the interconversion times between different pairs of metastable states, which can be defined *a priori* or learnt on-the-fly,[13] and monitor how these distances change by adding additional dimensions to the RC. The procedure is stopped when the interconversion times do not vary with additional RC components. The interconversion times are calculated using the commute distance framework proposed by Noé, Clementi and co-workers.[14,15] While such a kinetic or commute distance based procedure is indeed already recommended best practice in the construction of Markov State Models (MSMs),[16] it is not directly amenable to rare event systems that might be undersampled, or accessible only through biased simulations.

To deal with this issue, in this work we combine the the commute distance[14,15] with the Maximum Caliber based "Spectral Gap Optimization of Order Parameters (SGOOP)" approach.[17] This amounts to inducing a distance metric, which we call "SGOOP-d" that preserves kinetic truthfulness, and can be calculated from long unbiased simulations as well as biased simulations. Such biased simulations are often unavoidable in the study of rare events in chemical and biological physics. Here we use metadynamics[18] as an example of the biasing method to illustrate the usefulness of SGOOP-d, while anticipating that the method directly applies to other biasing pro-
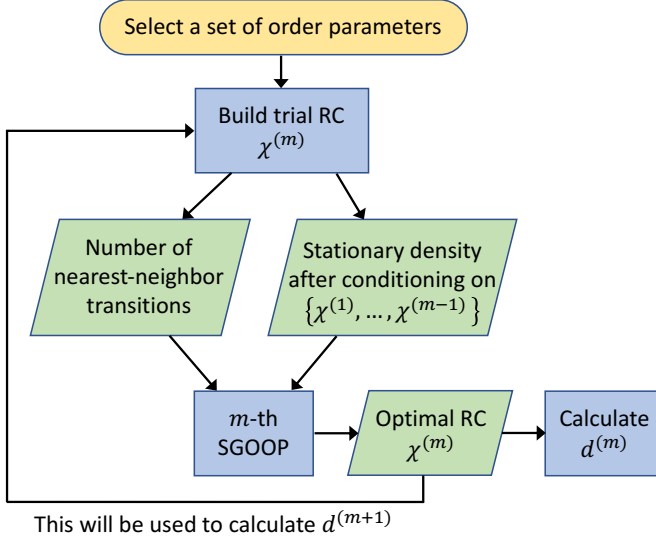
---
a) Electronic mail: ptiwary@umd.edu

Figure 1. This flowchart describes the calculation of the m-th RC component $\chi^{(m)}, m \geq 1$ through multi-dimensional spectral gap optimization. For each m we calculate $d^{(m)}$ in Eq. 10, which represents the contribution to commute distance on the basis of this m-th component. The optimal RC $\chi^{(m)}$ will be fed to the next SGOOP calculation for finding $d^{(m+1)}$. This iteration will stop when we obtain convergence in state-to-state d_{comm}^2 values with addition of RC components. The commute distance d_{comm}^2 will be the sum of all the $d^{(m)}$ obtained in the iteration.

tocols as well.[19] We demonstrate the utility of the method by studying a range of model systems including conformational transitions in a small peptide Ace-Ala$_3$-Nme. In this system for instance, one has a total of at least 28 inter-state transitions. As we show here, with only two component-RC learnt from SGOOP-d we do accurately capture most of the 28 pairs of distances, with minimal improvement achieved by adding a 3rd component to the RC. Similar results are obtained on the basis of input trajectories coming from metadynamics simulations biased along pre-selected biasing variables. Open-source software detailing the method has also been released.

II THEORY

II.1 Commute Distance and Commute Map

Our work builds upon the powerful advances first introduced by Clementi, Noé and co-workers that allow quantifying a kinetically truthful distance metric between generic molecular configurations.[14,15] One such notion of "kinetic distance" was introduced in Ref. 14, which was then generalized in Ref. 15 as the "commute distance". Both of these distances amount to transformations of the input coordinate space into a new space wherein Euclidean distances directly correspond to interconversion times. Here we summarize the basic ideas which origi-

nated from diffusion maps[20,21] but were later generalized to Markovian dynamics.[14,15]

We consider a generic dynamical system undergoing Markovian dynamics in a finite state space Ω. The local density $\rho_t(\mathbf{x})$, $\forall \mathbf{x} \in \Omega$ can be propagated in time t through

$$\rho_{t+\tau}(\mathbf{y}) = \int_{\mathbf{x} \in \Omega} \rho_t(\mathbf{x}) p_\tau(\mathbf{y}|\mathbf{x}) d\mathbf{x} \equiv \mathcal{P} \circ \rho_t(\mathbf{x}) \quad (1)$$

where $p_\tau(\mathbf{y}|\mathbf{x})$ is the transition density of finding the system at state $\mathbf{y}$ at time $t+\tau$ given that we have started it at state $\mathbf{x}$ at time t. Equivalently, Eq. 1 defines a Markov operator $\mathcal{P}$ and describes how an initial distribution $\rho_t(\mathbf{x})$ at time t propagates to the distribution $\rho_{t+\tau}(\mathbf{y})$ at a later time $t+\tau$. One usual assumption made here is that there exists a unique equilibrium distribution $\pi(\mathbf{x})$ which satisfies

$$\pi(\mathbf{x}) = \mathcal{P} \circ \pi(\mathbf{x}) \quad (2)$$

At the same time, we can write an equivalent equation for the weighted density $\nu_t(\mathbf{x}) = \rho_t(\mathbf{x})/\pi(\mathbf{x})$

$$\pi(\mathbf{y})\nu_{t+\tau}(\mathbf{y}) = \int p_\tau(\mathbf{y}|\mathbf{x})\pi(\mathbf{x})\nu_t(\mathbf{x}) d\mathbf{x} = \mathcal{T} \circ \nu_t(\mathbf{x}) \quad (3)$$

where $\mathcal{T}$ is the corresponding backward operator, also called the transfer operator. With this formalism, following the literature on diffusion maps[21] one defines a distance measure $D_\tau^2(\mathbf{x}_1, \mathbf{x}_2)$ between two points $\mathbf{x}_1, \mathbf{x}_2$ in the state space of a random walk as

$$D_\tau^2(\mathbf{x}_1, \mathbf{x}_2) = \int_{\mathbf{y} \in \Omega} \frac{|p_\tau(\mathbf{y}|\mathbf{x}_1) - p_\tau(\mathbf{y}|\mathbf{x}_2)|^2}{\pi(\mathbf{y})} d\mathbf{y} \quad (4)$$

This definition can be seen[21] as equivalent to (a) preparing two ensembles initially located at $\mathbf{x}_1$ and $\mathbf{x}_2$, (b) letting them evolve by a lag time τ, and then (c) computing the difference between the subsequently resulting probability distributions. In order to make use of Eq. 4, one needs the transition density $p_\tau(\mathbf{y}|\mathbf{x})$. To facilitate its computation,[14] we assume that the transfer operator $\mathcal{T}$ has N discrete eigenpairs and assume reversible dynamics/detailed balance $\pi(\mathbf{x}) p_\tau(\mathbf{y}|\mathbf{x}) = \pi(\mathbf{y}) p_\tau(\mathbf{x}|\mathbf{y})$:

$$p_\tau(\mathbf{y}|\mathbf{x}) = \sum_{j=0}^{N-1} \lambda_j(\tau) \psi_j(\mathbf{x}) \pi(\mathbf{y}) \psi_j(\mathbf{y}) \quad (5)$$

where λ_j and ψ_j are the corresponding eigenvalues and eigenvectors of the transfer operator $\mathcal{T}$. This allows calculation of Eq. 4 directly as:

$$D_\tau^2(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^{N-1} (\lambda_j \psi_j(\mathbf{x}_1) - \lambda_j \psi_j(\mathbf{x}_2))^2 \quad (6)$$

In Eq. 6 the summation starts at $j = 1$ since the $j = 0$ eigenvector for the transfer operator $\mathcal{T}$ is a constant in $\mathbf{x}$−space. By further integrating out the lag time τ in Eq. 6, we can make Eq. 6 insensitive to the choice of the lag time, and in this way we arrive at the definition of

the commute distance d_{comm}^2:

$$d_{\text{comm}}^2(\mathbf{x}_1, \mathbf{x}_2) = \int_0^\infty D_\tau^2(\mathbf{x}_1, \mathbf{x}_2) d\tau$$

$$= \sum_{j=1}^{N-1} \left(\sqrt{\frac{t_j}{2}} \psi_j(\mathbf{x}_1) - \sqrt{\frac{t_j}{2}} \psi_j(\mathbf{x}_2) \right)^2 \quad (7)$$

where $t_j = -\frac{\tau}{\ln \lambda_j}$ is the relaxation timescale associated with jth eigenvector. Often one uses the rate $k_j^{-1} = t_j$ instead of the timescale.[22] Eq. 7 now has a Euclidean distance form and a direct physical meaning: it is approximately the averaged time the system spends to commute between two states.[15] The distance d_{comm} is thus called the "commute distance", and the associated mapping

$$\mathbf{x} \mapsto (\sqrt{\frac{t_1}{2}} \psi_1, ..., \sqrt{\frac{t_{N-1}}{2}} \psi_{N-1}) \quad (8)$$

is called the "commute map".

Assuming that the dynamics in the $\mathbf{x}$−space is Markovian and fully sampled giving access to eigenvalues and eigenvectors of $\mathcal{T}$, we can then use Eq. 7 to calculate a Euclidean distance which approximates the commute time in the $\mathbf{x}$−space. It is also worth pointing out that in Eq. 7 the timescales follow $t_1 \geq t_2 \geq ...0$, which implies that the commute distance increases monotonically with consideration of further eigenvectors of $\mathcal{T}$, and that there is an increasingly vanishing contribution from every additional eigenvector that we consider. If such a distance can be obtained through Eq. 7, it is very useful for analyzing high-dimensional trajectories arising from well-sampled simulations as shown for instance in Ref. 14 and 15. However many if not most real-world applications are characterized by rare events, wherein the system stays trapped in the part of the configuration space it was initiated from and rarely visits other regions. Adequate and reliable sampling of the underlying configuration space thus remains a longstanding challenge in computational chemistry and physics. This implies that the eigenvectors and eigenvalues needed to evaluate the various terms in Eq. 7 are simply not available or far from reliable. In fact, the dominant first few components of the commute map could even serve as biasing coordinates along which the sampling could be enhanced through methods such as umbrella sampling, metadynamics or others. This brings out the inverse nature of the problem wherein constructing an accurate commute distance depends on sufficient sampling of the eigenvalues and eigenvectors of the transfer operator, but the sampling itself could benefit greatly from the knowledge of the commute map.

II.2 Calculating commute distances for rare events

In this section we develop a formalism for obtaining commute distances in poorly sampled rare-event systems where access to $\mathcal{T}$ and its eigenvectors/eigenvalues is not straightforward. The central idea is to perform biased sampling to accelerate exploration of the configuration space. Here we use metadynamics as the biased sampling method, but the developed formalism should be more generically applicable. While this basic idea is simple, there are however, at least two major, immediate difficulties when applying Eq. 7 with metadynamics or other similar enhanced sampling methods. First, the use of any sort of biasing corrupts the kinetics of the system, critical to calculating accurate eigenvalues and eigenvectors of the transfer operator $\mathcal{T}$. Second, the biasing itself needs access to the slow modes of the system, which are the dominant components of the commute map in Eq. 8. In SGOOP, we find these slow modes from such transition matrix but only look at its dynamics along 1-d coordinate, where we refer to these slow modes as the reaction coordinate (RC) for the system.[23,24] As mentioned in Sec. II 1 the different components ψ_i of the commute map have a vanishing relevance to the calculation of the commute distance as $i \gg 1$, and thus one can stop after the first few dominant components and bias these components in any biasing method of choice. However, without knowing the commute map, it is hard to calculate the dimensionality and components of the RC which would then be biased.

II.2.1 Commute distance calculation for rare events with SGOOP

In this sub-section we develop an approach that circumvents both of the above described challenges. Our work builds on the RC discovery method "Spectral Gap Optimization of Order Parameters (SGOOP)".[13,17,25] Summarily, SGOOP in its original form is a method for obtaining a one-dimensional RC given static and dynamic information about a multi-dimensional system by combining these information in a maximum Caliber or path entropy framework.[26,27] Here we use SGOOP to induce a commute distance metric for complex high-dimensional systems that can be calculated from a combination of biased simulations and short unbiased trajectories. SGOOP constructs the RC as a combination of pre-selected candidate order parameters $(s_1, ..., s_d)$, which can be thought of as a set of basis functions using which we are trying to describe our problem. The dimensionality d is kept high enough so that dynamics in the high-dimensional s−space is likely Markovian, needed for the formalism described in Sec. II 1. The central ideas behind SGOOP[17] in its original form can be summarized as the following three points:
(i) It uses a reweighting protocol[28] to estimate the equilibrium distribution $P_0(s_1, ..., s_d)$ from an initial metadynamics simulation performed by biasing some trial RC.
(ii) In addition, it uses short unbiased MD simulations to obtain dynamical observables pertaining to the system. These observables could be the position-dependent diffusivity or more typically, number of nearest-neighbor transitions along some binned trial RCs.
(iii) By combining (i) and (ii) SGOOP constructs transi-

tion rate matrices along different putative RCs.

Different candidate one-dimensional RCs are then ranked in terms of the number of metastable states they demarcate, and then in terms of the timescale separation (or the spectral gap) between the slow and fast modes as projected on any RC. The optimal RC maximizes both of these. We recently introduced a multi-dimensional version of SGOOP[13] which makes it possible to extend the dimensionality of the RC in SGOOP. Each additional RC component $\chi^{(i)}, i \geq 2$ is constructed in a way that it captures features indiscernible in the previous components through a conditional probability factorization described in Sec. II 2 2. This de-emphasizes the features already captured by the components so identified. With multiple iterations of the SGOOP protocol one can identify a multi-dimensional RC $\chi = \{\chi^{(1)}, \chi^{(2)}, ...\}$. Assuming that a satisfactorily large number of components have been included in χ, any two points $\{\mathbf{x}, \mathbf{x}'\} \in \Omega$ can then be mapped without substantial loss of information to its values in the χ space as $\{\chi, \chi'\}$. Whether the dimensionality of the RC χ is indeed sufficient or not is a non-trivial question to answer, which we will address later in this section and in Sec. III. We can then reformulate Eq. 7 as

$$d_{\text{comm}}^2(\mathbf{x}, \mathbf{x}') = d_{\text{comm}}^2(\chi, \chi')$$
$$= \sum_{j=1}^{N-1} \frac{1}{2k_j} \left[\psi_j(\mathbf{x}) - \psi_j(\mathbf{x}')\right]^2$$
$$= \sum_{j=1}^{N-1} \frac{1}{2k_j^{(1)}} \left[\psi_j^{(1)}(\mathbf{x}) - \psi_j^{(1)}(\mathbf{x}')\right]^2 \quad (9)$$

In the above equation we have made use of the mapping $\mathbf{x} \to \chi$ learnt from SGOOP, but otherwise it still needs the eigenvalues and eigenvectors of the transfer operator $\mathcal{T}$. In the final line we have introduced a superscript (1) to indicate the case where the first RC $\chi^{(1)}$ learnt from SGOOP is indeed sufficient for the system at hand. In such a case, SGOOP yields a Maximum Caliber based rate matrix $K^{(1)}$ for transitions between grid points along suitably discretized $\chi^{(1)}$. Details of the construction of this rate matrix are described in Sec. II 2 2 while illustrative examples are provided in Sec. III. By diagonalizing the rate matrix $K^{(1)}$ we obtain the eigenvalues $k_1^{(1)}, k_2^{(1)}, ...$ and corresponding eigenvectors $\psi_1^{(1)}, \psi_2^{(1)}, ...$ to use in Eq. 9.

The above commute distance so obtained can be understood as an estimate of true commute distance using the 1-dimensional projected RC $\chi^{(1)}$. However, as shown in Sec. III and also emphasized in the literature on numerous occasions,[12] a 1-dimensional projection is often not kinetically truthful and does not reflect the connectivity of underlying high-dimensional space. We thus consider additional RC components $\chi^{(m)}$ from the SGOOP protocol, with eigenvalues $k_1^{(m)}, k_2^{(m)}, ...$ and corresponding eigenvectors $\psi_1^{(m)}, \psi_2^{(m)}, ...$, where $m \geq 1$ denotes which RC component we are looking at. Each such component induces its own contribution to the commute distance

which we add to the contribution of the 1st component $\chi^{(1)}$ in Eq. 9 yielding the central equation of this work for a $M-$component RC:

$$d_{\text{comm}}^2(\mathbf{x}, \mathbf{x}')$$
$$= \sum_{m=1}^{M} \sum_{j=1}^{N-1} \frac{1}{2k_j^{(m)}} \left[\psi_j^{(m)}(\mathbf{x}) - \psi_j^{(m)}(\mathbf{x}')\right]^2 \equiv \sum_{m=1}^{M} d^{(m)}$$
$$(10)$$

Here $d^{(m)}$ is the contribution to the commute distance arising from the mth RC component, while $k_j^{(m)}$ and $\psi_j^{(m)}$ are the jth eigenvalue and eigenvector of the Maximum Caliber-based transition matrix $K^{(m)}$ calculated along along RC-component $\chi^{(m)}$ (Sec. II 2 2).

We want to mention two important points here. Firstly, for any RC component $\chi^{(m)}$ for $m \geq 1$, the construction of the rate matrix $K^{(m)}$ as detailed in Sec. II 2 2 ensures that the rates are ordered as per $0 < k_1^{(m)} \leq k_2^{(m)} \leq ...$. This leads to a useful property that the commute distance is a strictly monotonically increasing function of adding further RC components as well as further eigenvectors along any RC component. By monitoring how $d_{\text{comm}}^2 = d^{(1)} + d^{(2)} + ...$ converges with addition of RC components, we can quantify the dimensionality of the RC needed for a given system at hand. Secondly, the intuitive idea behind going from Eq. 9 to Eq. 10 is that different eigenvectors are orthogonal to each other allowing for a Euclidean distance measure. This is strictly true for the SGOOP-derived eigenvectors along a given RC component, i.e. the dot product of $\psi_j^{(m)}$ and $\psi_k^{(m)}$ is 0 $\forall j, k, m \geq 0$ as mentioned in Sec. II 2 2. However when comparing $\psi_j^{(m)}(\chi_m)$ and $\psi_k^{(n)}(\chi_n)$ for $m \neq n$ i.e. for different RC components through multiple rounds of SGOOP[13] this is not strictly true, and thus we expect Eq. 10 to be an upper bound for the commute distance.

II.2.2 Rate matrix estimates from Maximum Caliber based SGOOP

In this sub-section we discuss further details of the construction of the rate matrix $K^{(m)}$ through SGOOP, which is central to calculating commute distance in Eq. 10. Mathematically this can be written as follows. Once the first RC component $\chi^{(1)}$ has been learnt by SGOOP, we focus our attention on the probability distribution P_1 conditional on the knowledge of χ defined as:

$$P_1(s_1, ..., s_d) \equiv P_0(s_1, ..., s_d | \chi^{(1)})$$
$$= \frac{P_0(s_1, ..., s_d)}{P_0(\chi^{(1)})} \quad (11)$$

where we have used that the equilibrium probability $P_0(s_1, ..., s_d, \chi^{(1)}) = P_0(s_1, ..., s_d)$ as $\chi^{(1)}$ is a deterministic function of $(s_1, ..., s_d)$. The next round of SGOOP is then performed on data sampled from P_1 instead of P_0, which yields the second RC component $\chi^{(2)}$ that captures features missed by $\chi^{(1)}$. The procedure can be repeated

for further RC components and can be performed using any enhanced sampling method.[13] Here we illustrate it using metadynamics. By performing well-tempered metadynamics simulation along $\chi^{(1)}$ where one builds a bias $V_b(\chi^{(1)})$, it can be shown that

$$P_0(\chi^{(1)}) \propto e^{-\beta F(\chi^{(1)})} \propto e^{+\beta[\frac{\gamma}{\gamma-1}V_b(\chi^{(1)})]}$$

$$P_1 \equiv \frac{P_0(s_1,...,s_d)}{P_0(\chi^{(1)})} \propto e^{-\beta[F(s_1,...,s_d)+V_b(\chi^{(1)})]} \qquad (12)$$

where $\beta = 1/k_B T$, γ is the bias factor for well-tempered metadynamics,[18] and F is the free energy of the system. Therefore, P_1 is simply the unreweighted/biased probability density obtained by sampling in the presence of bias potential $V_b(\chi^{(1)})$. Following the strategy proposed in Ref. 13, the rate matrix along any putative RC χ can be built as follows:

$$K^{(1)}_{mn} = \begin{cases} -\frac{\langle N \rangle}{\sum \sqrt{\pi_n \pi_m}}\sqrt{\frac{\pi_n}{\pi_m}}, & \text{if } n \neq m \\ -\sum_{k \neq m} K^{(1)}_{mk}, & \text{if } n = m \end{cases} \qquad (13)$$

where $\langle N \rangle$ is the total number of nearest-neighbor transitions per unit time, counted along a suitably discretized RC $\chi = \{\chi_n\}$ with n indicating grid index and $\pi \equiv P_0$ is the corresponding stationary density. For the first round of SGOOP to learn $\chi^{(1)}$, $\langle N \rangle$ is calculated from short unbiased MD simulations. The $K^{(1)}$ matrices are then constructed for different putative RCs and its eigenvalues used to screen for the best RC $\chi^{(1)}$ with highest spectral gap, as well as the eigenvalues and eigenvectors along this RC to be used in Eq. 9.

For learning the second component $\chi^{(2)}$ and other higher order components, we generalize Eq. 13 as follows:[13]

$$K^{(2)}_{mn} = \begin{cases} -\frac{\langle N \rangle^{(1)}}{\sum \sqrt{\pi_n^{(1)} \pi_m^{(1)}}}\sqrt{\frac{\pi_n^{(1)}}{\pi_m^{(1)}}}, & \text{if } n \neq m \\ -\sum_{k \neq m} K^{(2)}_{mk}, & \text{if } n = m \end{cases} \qquad (14)$$

In Eq. 14, $\pi^{(1)} \equiv P_1$ is defined in Eq. 12. $\langle N \rangle^{(1)}$ denotes the average number of first-nearest neighbor transitions along a putative RC observed per unit time, but now measured in the biased simulation performed by sampling from this conditional probability density P_1. The procedure can then be easily generalized for constructing rate matrices $K^{(3)}, K^{(4)},...$ for learning further RC components. As mentioned previously in Sec. II 2 1, by diagonalizing the rate matrix $K^{(i)}, i \geq 1$ we obtain the eigenvalues $k_1^{(i)}, k_2^{(i)},...$ and corresponding eigenvectors $\psi_1^{(i)}, \psi_2^{(i)},...$ to use in Eq. 9.

III RESULTS

In this section we demonstrate the usefulness and reliability of the SGOOP[13,17] based commute distance[14,15] protocol developed in Sec. II, which we label "SGOOP-d" for convenience, by applying it to a range of analytical potentials, as well as to small molecules with rare conformational transitions between different metastable states. Low-dimensional projections of these high-dimensional potentials can in general lead to a spurious number of barriers and inter-basin connectivity.[12,29] Here we show how to use SGOOP-d to ascertain the minimal dimensionality of the RC that preserves the kinetic aspects of the underlying high-dimensional landscape. To do so we calculate the state-to-state commute distances, and monitoring how these change and eventually converge with increase in RC dimensionality. This is done using either biased or long unbiased simulations. We can also use the RC so learnt to perform further efficient and reliable biased simulations. We consider different types of unbiased and biased trajectories to demonstrate the general applicability of our proposed framework. Numerical and computational details of these systems have been provided in the Supplementary Information (SI).

III.1 Analytical potentials

The analytical potentials used here are originally inspired from Ref. 12. These are built with two degrees of freedom x and y, but with varying number of metastable states and barriers separating them. Thus a 1-d projection is not always guaranteed to be kinetically truthful. Specifically we consider a 3-state potential and three 4-state potentials labeled 4A and 4B (Figs. 2 (a)-(c)). For each of these we build inter-state commute distances using one-dimensional and two-dimensional RCs, with different components expressed as linear combinations of x and y. Since the underlying dimensionality is two, here we will demonstrate the results with up to two-dimensional RC. In such a case we can simplify Eq. 10 by introducing

$$\hat{d}^{(m)} = k_1^{(m)} d^{(m)} \qquad (15)$$

and then writing

$$d_{\text{comm}}(\mathbf{x}_1, \mathbf{x}_2) = d^{(1)} + d^{(2)}$$
$$= \frac{1}{k_1^{(1)}}\hat{d}^{(1)} + \frac{1}{k_1^{(2)}}\hat{d}^{(2)} \qquad (16)$$

To see how good a job the RC components do at reconstructing the state-to-state connectivity, we further parameterize Eq. 16 by using a tunable parameter K for the ratio of eigenvalues $\frac{k_1^{(1)}}{k_1^{(2)}}$, yielding

$$k_1^{(1)} d_{\text{comm}}(\mathbf{x}_1, \mathbf{x}_2) \equiv k_1^{(1)} d_K \equiv \hat{d}^{(1)} + K\hat{d}^{(2)} \qquad (17)$$

We highlight here that in our framework K is not a free parameter that needs to be tuned. Instead, it can be approximated on the basis of Maximum Caliber based rate matrices (Sec. II 2 2) as:

$$K^* \equiv \frac{k_1^{(1)}}{k_1^{(2)}} \qquad (18)$$

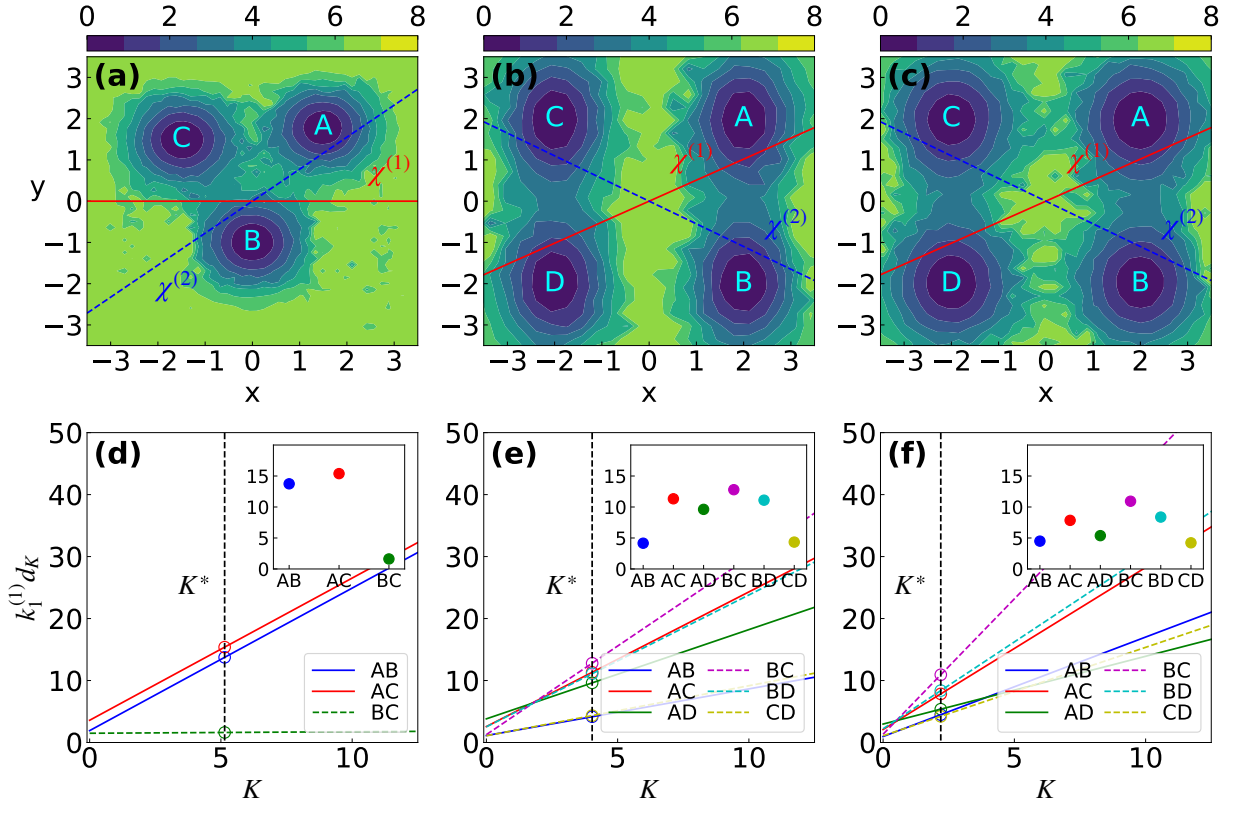However, as the Maximum Caliber based rate estimates

Figure 2. (a)-(c) show the analytical 3-state and 4-state potentials 4A, 4B respectively. In (a)-(c) we have also provided the two RC components $\chi^{(1)}$ (solid red lines) and $\chi^{(2)}$ (dashed blue lines) evaluated using Eq. 17. Contours in all plots are separated by $k_B T$. In (d)-(f) the estimated commute distances d_K between different pairs of states (in arbitrary units) are plotted as function of K, where the benchmark parameter K^* in each case is specified as the vertical black dashed line. The insets in (d)-(f) show the commute distance between each metastable states at $K = K^*$. As explained in Sec. III 1, using $K = K^*$ gives the right kinetic connectivity between different metastable states for each of the model potentials.
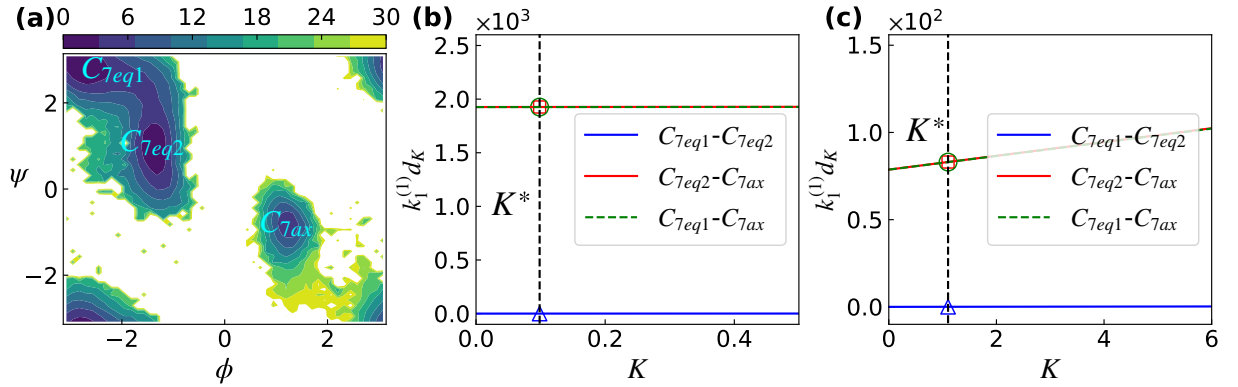

Figure 3. (a) Free energy surface as function of ϕ and ψ obtained by reweighting metadynamics simulation biasing along 1-d RC $\chi^{(1)}$ specified in Table. II . The positions of three metastable states are specified. (b) shows the corresponding d_K as a function of K for each pair of metastable states obtained from the biased simulation. (c) shows the same information generated from a long unbiased simulation. In both (b) and (c) the benchmark parameter K^* is shown through a vertical black dashed line, while empty circles correspond to the expected best estimates of commute distances.

are approximate and might depend on choice of the dynamical constraints and quality of sampling,[27] we can vary K around K^* and ascertain the sensitivity of the calculation of commute distance with respect to the accuracy of K.
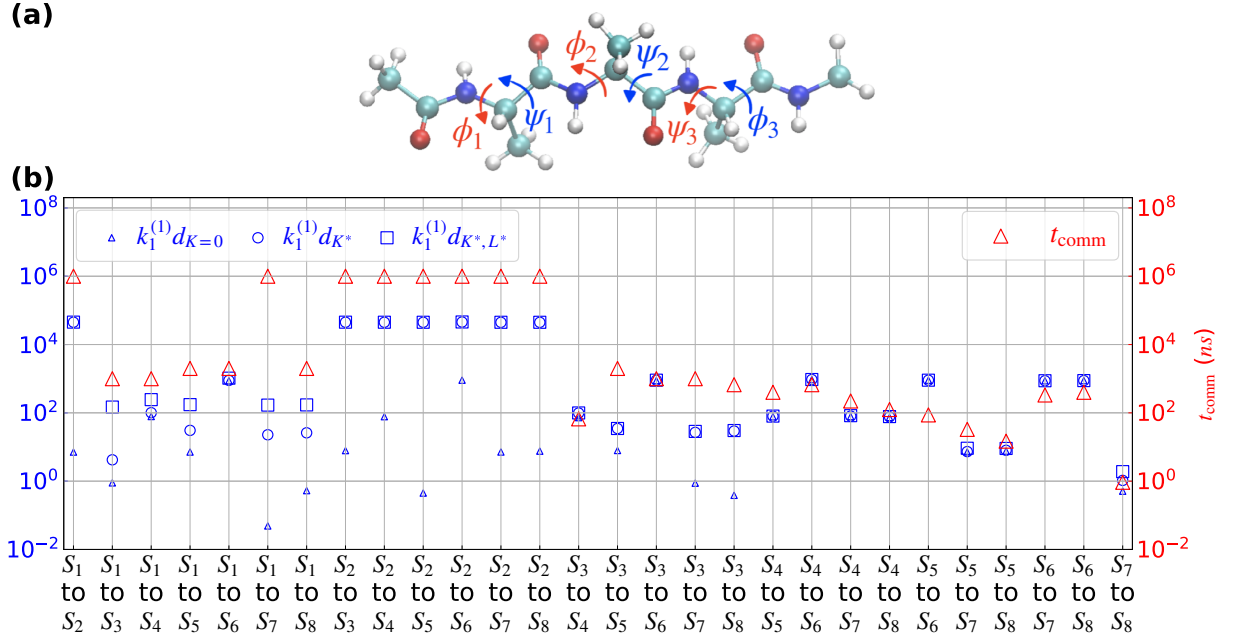
Fig. 2 and Table. I detail the two RC-components

(a)

(b)

Figure 4. In this figure, (a) provides the molecular structure of Ace-Ala$_3$-Nme with the corresponding dihedral angles. The corresponding metastable states are detailed in SI. (b) provides the estimated commute distances using one-dimensional, two-dimensional and three-dimensional RC respectively (blue triangles, blue circles and blue squares, left axis). Corresponding to their calculation, these are labelled respectively $k_1^{(1)} d_{K=0}$, $k_1^{(1)} d_{K^*}$ and $k_1^{(1)} d_{K^*,L^*}$ (in arbitrary units). (b) also provides the estimated commute time (t_{comm} (red triangles, right axis) calculated from long unbiased simulation of Ace-Ala$_3$-Nme.

$\chi^{(1)}$ and $\chi^{(2)}$ so obtained for the different model potentials. Here using $K = 0$ is equivalent to using only the first component $\chi^{(1)}$ to determine the commute distance, while increasing non-zero values of K captures increasing contributions from the second component $\chi^{(2)}$ through Eq. 17. As can be seen for the 3-state system (Fig. 2 (d)), considering only the first component $\chi^{(1)}$ would lead to an erroneous conclusion that the pairs of states AB, AC and BC are all kinetically equidistant. This is not consistent with the high-dimensional data sampled shown in Fig. 2 (a), where the barrier experienced between the states BC is much lower than for AB and AC. By adding the second component $\chi^{(2)}$ to the kinetic distance in Eq. 17 using $K = K^*$, we recover this correct picture. Furthermore the picture is relatively insensitive to the exact choice of K^* thereby showing the robustness of the procedure to possible inaccuracies in the Maximum Caliber estimate. Similar conclusions regarding kinetically truthful picture consistent with the data can be drawn for the remaining three 4-state potentials shown in Fig. 2. In both Fig. 2 (e) and (f), using only the 1-d RC $\chi^{(1)}$, AB, BC, and CD are equally short, while AD is the slowest transition. This erroneous connectivity has been corrected after adding a second component of RC $\chi^{(2)}$, where AB and CD are equally shortest at $K = K^*$. Note that in both insets AD is slightly lower which shows the noisy nature in the Maximum Caliber based estimation of transition rates.

Systems		$\theta^{(1)}/\pi$	$\theta^{(2)}/\pi$
3-state		0.00	0.21
4-state	4A (Fig. 2(b))	0.15	0.84
	4B (Fig. 2(c))	0.15	0.84

Table I. In this table, we have shown the first and second components of the reaction coordinate $\chi^{(1)}$ and $\chi^{(2)}$ found for each model analytical potential. The angles $\theta^{(1)}$ and $\theta^{(2)}$ in the table define $\chi^{(i)} = \cos(\theta^{(i)})x + \sin(\theta^{(i)})y$.

III.2 Alanine dipeptide

The next system we use to illustrate our method is the well-studied alanine dipeptide. Here we consider the molecule as characterized by three dihedral angles ϕ, ψ and θ. This molecule has three metastable configurations (Fig. 3(a)) which can be characterized by using only ϕ and ψ, while θ plays a role in characterizing the transition between the metastable states.[30] Here we express the different RC components as linear combinations of 6 order parameters, namely cosines and sines of the 3 aforementioned dihedrals, with the final optimized coefficients listed in Table. II. The spectral gap in SGOOP is optimized using a basin-hopping algorithm.[31–33] These RC components and associated information are then plugged into Eq. 17 to estimate the commute distance d_K. In Figs. 3(b)-(c) we show the commute distance so calculated using an input biased trajectory and a benchmark

Systems	RCs	Coefficients
Alanine dipeptide	$\chi^{(1)}$	$(0.643, 0.778, -0.133, -0.088, -0.221, -0.165)$
	$\chi^{(2)}$	$(0.827, 1.166, -0.120, 0.578, 0.013, 0.240)$
Ace-Ala$_3$-Nme	$\chi^{(1)}$	$(0.187, -1.127, -0.228, -2.362, 0.230, 1.176)$
	$\chi^{(2)}$	$(1.174, 0.738, 0.132, 0.716, 0.356, 2.827)$
	$\chi^{(3)}$	$(-0.037, -0.839, 0.557, 1.454, 1.693, 1.624)$

Table II. This table shows the reaction coordinates found for alanine dipeptide and Ace-Ala$_3$-Nme. For alanine dipeptide, two RC components both expressed as $\chi = a\cos\phi + b\sin\phi + c\cos\psi + d\sin\psi + e\cos\theta + f\sin\theta$ with their 6 respective coefficients are listed. For Ace-Ala$_3$-Nme, three RC components all expressed as $\chi = a\cos\phi_1 + b\sin\phi_1 + c\cos\phi_2 + d\sin\phi_2 + e\cos\phi_3 + f\sin\phi_3$ with their 6 respective coefficients are listed.

long unbiased trajectory respectively. The biased trajectory was generated by doing well-tempered metadynamics along 1-d RC $\chi^{(1)}$ defined in Table. II. See SI for further details of both the biased and unbiased simulations.

For this simple system, the commute distances d_K show similar connectivities for $K = 0$ and $K = K^*$, which shows that one RC is indeed sufficient to describe the system in terms of recovering state-to-state connectivity between all 3 metastable states. Both types of input trajectories show a near degenerate structure with two pairs of states kinetically separated from each other, while one pair very close.

III.3 Ace-Ala$_3$-Nme

In this final section, we demonstrate our method on a more complicated molecular system, namely the peptide Ace-Ala$_3$-Nme with a much larger number of metastable states, and an even larger number of state-to-state transitions.[34] Simulation details are provided in SI. As discussed in Ref. 34 the three dihedral angles ϕ_1, ϕ_2, ϕ_3 are sufficient to characterize the $2^3 = 8$ dominant metastable states corresponding to positive and negative parts of the Ramachandran diagram for the 3 central Alanine residues. The RC components used in computing SGOOP-d distances are calculated as linear combination of cosines and sines of these 3 dihedral angles, thereby amounting to a total of 6 order parameters. We consider the 8 most dominant metastable states labelled S_1,..., S_8 and the associated $\binom{8}{2} = 28$ inter-state transitions. The corresponding dihedral angles for these 8 states are tabulated in the SI. Here we consider up to three RC components, and demonstrate that after considering 3 components the commute distances converge especially for the slower state-to-state transitions. They are also in agreement with the benchmark calculations on this system through counting transitions in the higher dimensional underlying space from a long unbiased trajectory. The final optimized solutions for all three RC components are shown in Table. II. Here in order to add a third RC component we generalize Eq. 17 by introduc-

ing an additional parameter L:

$$k_1^{(1)} d_{K,L} \equiv \hat{d}^{(1)} + K\hat{d}^{(2)} + L\hat{d}^{(3)} \qquad (19)$$

Similar to what was done for K^* in Eq. 18 we can approximate L^* as

$$L^* \equiv \frac{k_1^{(1)}}{k_1^{(3)}} \qquad (20)$$

With a long enough unbiased MD trajectory, we can also calculate the commute time t_{comm} between two metastable states through a simple counting protocol (see SI and Ref. 29). In Fig. 4, we show SGOOP-d distances calculated using Eq. 19 with 1, 2 and 3 RC components, and compare them with the corresponding 28 t_{comm} values between these 8 metastable states in the same plot. It can be seen from plot that with only the use of two RC components SGOOP-d already provides a good estimate of relative inter-state connectivity and commute distances between 23 of the 28 pairs of states. However, in order to get the correct connectivity for the remaining 5 pairs of states as well, we have to include the third RC component. We emphasize that in Fig. 4 the slowest 8 transitions have the same reference commute time as we were to observe any such transition events even in the 1 μs long unbiased simulation. Thus the reference commute times for these states serve as approximate lower bounds to the true values.

IV CONCLUSION

In summary, in this work we have developed a computationally efficient formalism labeled "SGOOP-d" and summarized in the flowchart in Fig. 1, that can help towards solving a longstanding important problem in chemical physics and physical chemistry. Namely, how many dimensions should a projection from high-dimensions into low-dimensional reaction coordinates (RC) have, so that (1) the projection is kinetically and thermodynamically truthful to the underlying landscape, and (2) these minimal number of components can then be used to perform biasing simulations without fear of missing slow degrees of freedom. The formalism here makes the best of two different approaches, namely commute map[15] and

SGOOP[17]. This way it induces a distance metric which we call SGOOP-d that is applicable to biased rare event systems as well as unbiased trajectories with arbitrary quality of sampling. The kinetically truthful RC learnt here can then also be used to improve the sampling quality of the biased simulation itself[35] or as a progress coordinate in path-based sampling methods.[36–40] We thus believe that going forward our work represents a useful tool in the study of kinetics in rare event systems with multiple states and interconnecting pathways.

V ACKNOWLEDGMENTS

REFERENCES

[1] J Juraszek, G Saladino, TS Van Erp, and FL Gervasio. Efficient numerical reconstruction of protein folding kinetics with partial path sampling and pathlike variables. *Physical Review Letters*, 110(10):108106, 2013.

[2] Pratyush Tiwary, Jagannath Mondal, and Bruce J Berne. How and when does an anticancer drug leave its binding site? *Science advances*, 3(5):e1700014, 2017.

[3] Haiyang Niu, Pablo M Piaggi, Michele Invernizzi, and Michele Parrinello. Molecular dynamics simulations of liquid silica crystallization. *Proceedings of the National Academy of Sciences*, 115(21):5348–5352, 2018.

[4] Sun-Ting Tsai, Zachary Smith, and Pratyush Tiwary. Reaction coordinates and rate constants for liquid droplet nucleation: Quantifying the interplay between driving force and memory. *The Journal of chemical physics*, 151(15):154106, 2019.

[5] Sooheyong Lee, Haeng Sub Wi, Wonhyuk Jo, Yong Chan Cho, Hyun Hwi Lee, Se-Young Jeong, Yong-Il Kim, and Geun Woo Lee. Multiple pathways of crystal nucleation in an extremely supersaturated aqueous potassium dihydrogen phosphate (kdp) solution droplet. *Proceedings of the National Academy of Sciences*, 113(48):13618–13623, 2016.

[6] Jorjethe Roca, Naoto Hori, Saroj Baral, Yogambigai Velmurugu, Ranjani Narayanan, Prasanth Narayanan, D Thirumalai, and Anjum Ansari. Monovalent ions modulate the flux through multiple folding pathways of an rna pseudoknot. *Proceedings of the National Academy of Sciences*, 115(31):E7313–E7322, 2018.

[7] Jan-Hendrik Prinz, Hao Wu, Marco Sarich, Bettina Keller, Martin Senne, Martin Held, John D Chodera, Christof Schütte, and Frank Noé. Markov models of molecular kinetics: Generation and validation. *The Journal of chemical physics*, 134(17):174105, 2011.

[8] Stefano Piana, Kresten Lindorff-Larsen, and David E Shaw. Protein folding kinetics and thermodynamics from atomistic simulation. *Proceedings of the National Academy of Sciences*, 109(44): 17845–17850, 2012.

[9] Robert B Best and Gerhard Hummer. Reaction coordinates and rates from transition paths. *Proceedings of the National Academy of Sciences*, 102(19):6732–6737, 2005.

[10] Gerhard Hummer and Attila Szabo. Optimal dimensionality reduction of multistate kinetic and markov-state models. *The Journal of Physical Chemistry B*, 119(29):9029–9037, 2015.

[11] Gareth A Tribello and Piero Gasparotto. Using dimensionality reduction to analyze protein trajectories. *Frontiers in molecular biosciences*, 6:46, 2019.

[12] Alexandros Altis, Moritz Otten, Phuong H Nguyen, Rainer Hegger, and Gerhard Stock. Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis. *The Journal of chemical physics*, 128(24):06B620, 2008.

[13] Zachary Smith, Debabrata Pramanik, Sun-Ting Tsai, and Pratyush Tiwary. Multi-dimensional spectral gap optimization of order parameters (sgoop) through conditional probability factorization. *J. Chem. Phys.*, 149(23):234105, 2018.

[14] Frank Noé and Cecilia Clementi. Kinetic distance and kinetic maps from molecular dynamics simulation. *J. Chem. Theor. Comp.*, 11(10):5002–5011, 2015.

[15] Frank Noéé, Ralf Banisch, and Cecilia Clementi. Commute maps: Separating slowly mixing molecular configurations for kinetic modeling. *J. Chem. Theor. Comp.*, 12(11):5620–5630, 2016.

[16] Brooke E Husic and Vijay S Pande. Markov state models: From an art to a science. *Journal of the American Chemical Society*, 140(7):2386–2396, 2018.

[17] Pratyush Tiwary and BJ Berne. Spectral gap optimization of order parameters for sampling complex molecular systems. *Proceedings of the National Academy of Sciences*, 113(11):2839–2844, 2016.

[18] Omar Valsson, Pratyush Tiwary, and Michele Parrinello. Enhancing important fluctuations: Rare events and metadynamics from a conceptual viewpoint. *Annual review of physical chemistry*, 67:159–184, 2016.

[19] Pratyush Tiwary and Axel van de Walle. A review of enhanced sampling approaches for accelerated molecular dynamics. *Multiscale Materials Modeling for Nanomechanics*, pages 195–221, 2016.

[20] Boaz Nadler, Stephane Lafon, Ioannis Kevrekidis, and Ronald R Coifman. Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators. In *Advances in neural information processing systems*, pages 955–962, 2006.

[21] Ronald R Coifman, Stephane Lafon, Ann B Lee, Mauro Maggioni, Boaz Nadler, Frederick Warner, and Steven W Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the national academy of sciences*, 102(21):7426–7431, 2005.

[22] Guillermo Pérez-Hernández, Fabian Paul, Toni Giorgino, Gianni De Fabritiis, and Frank Noé. Identification of slow molecular order parameters for markov model construction. *The Journal of chemical physics*, 139(1):07B604_1, 2013.

[23] Ao Ma and Aaron R Dinner. Automatic method for identifying reaction coordinates in complex systems. *The Journal of Physical Chemistry B*, 109(14):6769–6779, 2005.

[24] Andreas Bittracher, Ralf Banisch, and Christof Schütte. Data-driven computation of molecular reaction coordinates. *The Journal of Chemical Physics*, 149(15):154103, 2018.

[25] Pratyush Tiwary and BJ Berne. Predicting reaction coordinates in energy landscapes with diffusion anisotropy. *The Journal of chemical physics*, 147(15):152701, 2017.

[26] Steve Pressé, Kingshuk Ghosh, Julian Lee, and Ken A Dill. Principles of maximum entropy and maximum caliber in statistical physics. *Rev. Mod. Phys.*, 85(3):1115, 2013.

[27] Kingshuk Ghosh, Purushottam D Dixit, Luca Agozzino, and Ken A Dill. The maximum caliber variational principle for nonequilibria. *Annual review of physical chemistry*, 71:213–238, 2020.

[28] Pratyush Tiwary and Michele Parrinello. A time-independent free energy estimator for metadynamics. *The Journal of Physical Chemistry B*, 119(3):736–742, 2015.

[29] Sun-Ting Tsai, En-Jui Kuo, and Pratyush Tiwary. Learning molecular dynamics with simple language model built upon long short-term memory neural network. *Nature communications*, 11(1):1–11, 2020.

[30] Pratyush Tiwary and Michele Parrinello. From metadynamics to dynamics. *Physical review letters*, 111(23):230602, 2013.

[31] David Wales et al. *Energy landscapes: Applications to clusters, biomolecules and glasses.* Cambridge University Press, 2003.

[32] David J Wales and Jonathan PK Doye. Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms. *The Journal of Physical Chemistry A*, 101(28):5111–5116, 1997.

[33] Zhenqin Li and Harold A Scheraga. Monte carlo-minimization approach to the multiple-minima problem in protein folding. *Proceedings of the National Academy of Sciences*, 84(19):6611–6615, 1987.

[34] Jakub Rydzewski and Omar Valsson. Multiscale reweighted stochastic embedding (mrse): Deep learning of collective variables for enhanced sampling. *arXiv preprint arXiv:2007.06377*, 2020.

[35] Giovanni Bussi and Alessandro Laio. Using metadynamics to explore complex free-energy landscapes. *Nat. Rev. Phys.*, pages 1–13, 2020.

[36] Daniel M Zuckerman and Lillian T Chong. Weighted ensemble simulation: review of methodology, applications, and software. *Annual review of biophysics*, 46:43–57, 2017.

[37] Ron Elber. Milestoning: An efficient approach for atomically detailed simulations of kinetics in biophysics. *Annual review of biophysics*, 49:69–85, 2020.

[38] Lane W Votapka, Benjamin R Jagger, Alexandra L Heyneman, and Rommie E Amaro. Seekr: simulation enabled estimation of kinetic rates, a computational tool to estimate molecular kinetics and its application to trypsin–benzamidine binding. *The Journal of Physical Chemistry B*, 121(15):3597–3606, 2017.

[39] Hao Jiang, Amir Haji-Akbari, Pablo G Debenedetti, and Athanassios Z Panagiotopoulos. Forward flux sampling calculation of homogeneous nucleation rates from aqueous nacl solutions. *The Journal of chemical physics*, 148(4):044505, 2018.

[40] Ryan S DeFever and Sapna Sarupria. Contour forward flux sampling: Sampling rare events along multiple collective variables. *The Journal of chemical physics*, 150(2):024103, 2019.