AIP | The Journal of Chemical Physics

Nonparametric variational optimization of reaction coordinates

Polina V. Banushkina and Sergei V. Krivov

Citation: The Journal of Chemical Physics **143**, 184108 (2015); doi: 10.1063/1.4935180

View online: http://dx.doi.org/10.1063/1.4935180

View Table of Contents: http://scitation.aip.org/content/aip/journal/jcp/143/18?ver=pdfcov

Published by the AIP Publishing

Nonparametric variational optimization of reaction coordinates

Polina V. Banushkina and Sergei V. Krivov[a)]
Astbury Center for Structural Molecular Biology, Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, United Kingdom

(Received 21 August 2015; accepted 24 October 2015; published online 11 November 2015)

State of the art realistic simulations of complex atomic processes commonly produce trajectories of large size, making the development of automated analysis tools very important. A popular approach aimed at extracting dynamical information consists of projecting these trajectories into optimally selected reaction coordinates or collective variables. For equilibrium dynamics between any two boundary states, the committor function also known as the folding probability in protein folding studies is often considered as the optimal coordinate. To determine it, one selects a functional form with many parameters and trains it on the trajectories using various criteria. A major problem with such an approach is that a poor initial choice of the functional form may lead to sub-optimal results. Here, we describe an approach which allows one to optimize the reaction coordinate without selecting its functional form and thus avoiding this source of error. [http://dx.doi.org/10.1063/1.4935180]

I. INTRODUCTION

Realistic simulations of atomistic dynamics of complex biological processes are becoming increasingly common. These simulations produce a massive amount of data (e.g., trajectories). The trajectories alone without proper analysis, however, do not provide deep insight into the process dynamics. Interpreting the data and constructing reliable and meaningful models of the dynamics, taking into account all important dynamical information from the data, are non-trivial tasks. The large size of trajectories makes the development of automated analysis tools very important. Popular approaches can be roughly divided into three groups: Markov state models (MSMs),[1–5] conformation network analysis,[6–8] and the free energy landscape framework.[9–14] The free energy landscape framework is attractive because it allows one to obtain direct answers to fundamental questions about the dynamics: in particular, the shape of the free energy profile (FEP), the height of the barriers, the structure of the transition states, and kinetic pre-exponential factors.[13,15,16]

The free energy landscape approach consists of projecting the multidimensional trajectories onto so-called reaction coordinates (RCs) or collective variables. The dynamics is then described as diffusion on a low-dimensional free energy landscape, with both diffusion coefficient and free energy being functions of the coordinates. Examples of such dimensionality reductions can be found in a wide range of problems from many different scientific fields: in molecular dynamics simulations,[12,15,17–21] order parameters in physics,[11,22] physically based RCs in single molecular experiments,[23,24] biomarkers in medicine,[25] analysing the game of chess,[26] to name a few.

The reduction of dimensionality from many degrees of freedom to one (or a few) obviously leads to the loss of information and making the choice of the proper RC becomes the critical step. The optimal RC should absorb all the important kinetics information contained in the other degrees of freedom, so that they can be neglected. For example, the optimal RC should correctly partition the whole configuration space into subsets of conformations based on their dynamical and energetic characteristics and preserve all consequential aspects of the dynamics. By contrast, a poor RC does not take the kinetics into account and may inaccurately capture the populations, mixing them together, which is reflected in lower free energy barriers height, masking of intermediate states, or causing the dynamics to be subdiffusive.[11,27,28]

In early applications, RCs were selected based on physical intuition, e.g., number of native contacts, root mean square deviation (rmsd), and radius of gyration, when applied to analysing protein folding trajectories. They were selected like order-parameters, which satisfactory separate the two states — native and denatured. Despite their physical nature and simplicity of calculation, such coordinates are usually not optimal. They satisfactorily partition the configuration space into major free energy minima; however, they do not accurately describe important regions such as transition states and intermediate states.[11] Other popular methods for selection of RC include various variants of principle component analysis (PCA),[21,29,30] diffusion maps,[31–33] Isomap,[34] sketch-map,[20] and full correlation analysis;[35] see also recent reviews.[36–38]

Systematic methods to determine optimal RCs have only started to be developed recently.[36] The folding probability p^{fold} (or committor) is often considered to be an ideal RC to describe equilibrium transition dynamics between any two states.[12,18,28,39–48] The committor equals the probability for the trajectory to reach one state (e.g., the native state in the analysis of protein folding) before it reaches another (e.g., the denatured state) starting from any given configuration. It has been shown that the reaction rate computed based on Kramers

[a)] s.krivov@leeds.ac.uk

equation using the FEP and diffusion coefficient as functions of this coordinate is exact.[28,49,50]

The determination of the p^{fold} coordinate is not straightforward. It can be easily constructed from an accurate MSM; however, the determination of such MSMs, which accurately describe the transition state region, is similarly difficult.[28] A number of variational approaches have been suggested to determine the coordinate, without explicitly constructing the MSM.[12,28,43,48] To this end, a functional form for the RC containing many parameters is suggested. For example, for protein folding, one can take a weighted sum of native and non-native contacts.[12,13,27] Then, one numerically optimizes the weights,[12] or the cutoff distances for contacts[13,27] by optimizing a particular functional, so that in the end, the putative reaction coordinate accurately approximates p^{fold}. The following optimization functionals have been suggested: the probability of being on a transition path,[12,45] the likelihood functional,[43] the cut profiles,[13,27,48] and the total square displacement (TSD).[28]

While these approaches are very promising, they have a common problem that it is difficult to suggest an obviously good functional form for putative RCs. With a poorly chosen functional form, the resulting RC can be sub-optimal. In our previous work, we have been able to construct a RC, which closely approximates p^{fold} only around the transition states.[13,16] While the transition states are the most important parts of the FEP for the dynamics, it would be useful to be able to determine the coordinate that accurately approximates p^{fold} over the entire range.

Usually the functional form is taken as a linear combination of basis functions.[12,13,25,43,48] In particular, this allows one to solve the optimization problem analytically.[25,48] However, such a representation significantly restricts the flexibility of the RC and, as mentioned above, can lead to sub-optimal results. As a way to solve a similar problem in the framework of time-structure based independent components analysis, Schwantes and Pande suggested the use of the kernel trick to arrive at non-linear solutions.[51] Here, we propose another solution to this problem. We describe an approach which allows one to optimize the RC without specifying its functional form at all.

The paper is organized as follows. We start with some theoretical background on the description of the dynamics with a MSM, the p^{fold} coordinate, and eigenvectors. The second left eigenvector can be used to approximate the p^{fold} coordinate in two state systems. Then, we present a variational principle which can be used to determine p^{fold} and left eigenvectors without constructing a MSM. The general idea behind functional form free reaction coordinate optimization is described. It is followed by a detailed description of the practical implementation of the approach for the determination of p^{fold} and the second left eigenvector. The performance of the approach is illustrated on a few model systems.

II. THEORY

We follow the notation and formalism from Ref. 28. We assume that the dynamics of a system of interest can be described as an equilibrium stochastic Markov process in multidimensional configuration space $\vec{X}$, sampled with time interval Δt. In particular, the system trajectory can be represented as $\vec{X}(k\Delta t)$. For molecular dynamics simulations, where the dynamics at small time intervals is deterministic rather than stochastic, we assume that the sampling interval is sufficiently large, so that the system forgets its history and can be approximated as equilibrium Markov process in configuration space. We also assume that the configuration space of the system $\vec{X}$ has been discretized and each state is denoted by integer index i. In this case, the time evolution of $P_i(t)$, the probability distribution for system to be in state i, can be described as

$$P_i(t+\Delta t) = \sum_j P_{ji}(\Delta t) P_j(t), \tag{1}$$

where $P_{ji}(\Delta t)$ is the probability of transition from state i to j after time interval Δt. The equilibrium probability distribution is defined as

$$p_i^{eq} = \sum_j P_{ij}(\Delta t) p_j^{eq}. \tag{2}$$

Given a long equilibrium trajectory generated by the Markov process (Eq. (1)), one can compute the number of transitions $n_{ji}(\Delta t)$ from state i to state j. The number of times state i has been visited,

$$n_i = \sum_j n_{ji}(\Delta t) = \sum_j n_{ij}(\Delta t), \tag{3}$$

which is proportional to p_i^{eq}. Based on n_i and $n_{ji}(\Delta t)$, the elements of the transition probability matrix $\mathbf{P}$ can be estimated as[28]

$$P_{ji}(\Delta t) = \frac{n_{ji}(\Delta t)}{n_i}. \tag{4}$$

Given a MSM, p^{fold} between any two boundary nodes A and B can be determined by solving the following system of linear equations:[52]

$$p_i^{fold} = \sum_j P_{ji} p_j^{fold}, \quad p_A^{fold} = 0, \quad p_B^{fold} = 1. \tag{5}$$

In the case when boundary nodes are not known, approximation of p^{fold} by eigenvectors might be more convenient since one does not have to pick up the nodes A and B which should be on different sides of the transition state. Berezhkovskii and Szabo[53] have shown that for a system with two states and a large free energy barrier, the p^{fold} coordinate can be approximated around the transition state as

$$p_i^{fold} \approx a\left(\frac{v_{2,i}}{v_{1,i}} + b\right), \tag{6}$$

where $v_{1,i}$ and $v_{2,i}$ are the components of the first ($\vec{v}_1$) and the second ($\vec{v}_2$) right eigenvectors of the transition matrix $\mathbf{P}$, respectively. The equation for the right kth eigenvector $\vec{v}_k$ is

$$\lambda_k v_{k,i} = \sum_j P_{ij} v_{k,j}, \tag{7}$$

where λ_k is the kth highest eigenvalue of the transition matrix $\mathbf{P}$. The first eigenvalue λ_1 is equal to 1 and the first eigenvector is equal to the vector of equilibrium probabilities, $v_{1,i} = p_i^{eq}$

(Eq. (2)). The equation for the left kth eigenvector $\vec{u}_k$ can be written as

$$\lambda_k u_{k,i} = \sum_j P_{ji} u_{k,j}. \tag{8}$$

For equilibrium dynamics with detailed balance $P_{ij} p_j^{eq} = P_{ji} p_i^{eq}$, the left and right eigenvectors are related as

$$u_{k,j} = \frac{v_{k,j}}{p_j^{eq}} = \frac{v_{k,j}}{v_{1,j}},$$

which means, in particular, that $u_{1,j} = 1$ and

$$p_j^{fold} \approx a(u_{2,j} + b), \tag{9}$$

where a and b are unimportant constants, i.e., the folding probability is approximated by the second left eigenvector $\vec{u}_2$.

A. Variational approach

Both the left eigenvectors and the p^{fold} coordinate for the Markov process can also be obtained by minimizing the TSD of an equilibrium trajectory projected on a coordinate. Namely, let x_i be a value of a coordinate of state i after projection. Then, the TSD functional computed over an equilibrium trajectory

$$\sum_{i,j} n_{ij}(\Delta t)(x_i - x_j)^2 \tag{10}$$

attains minimum, under specific constraints, when x equals p^{fold} or the left eigenvectors.

Obviously, minimization of the TSD without constraints leads to the trivial solution where all x_i are the same and the total TSD is zero. If one fixes the position of two boundary states, for example, $x_A = 0$ and $x_B = 1$, then the minimum of Eq. (10) is attained when $x_i = p_i^{fold}$.[28] Equating the derivative of the TSD with respect to x_m to 0,

$$\frac{d}{dx_m} \sum_{i,j} n_{ij}(\Delta t)(x_i - x_j)^2 = 0, \tag{11}$$

one obtains

$$n_m x_m = \sum_j n_{jm}(\Delta t) x_j,$$

which after division by n_m and together with Eq. (4) reduces to Eq. (5).

To obtain the left eigenvectors of the transition matrix, one minimizes the TSD, while keeping the total square value of the coordinate normalized

$$\sum_i n_i x_i^2 = 1. \tag{12}$$

It that case, due to the constraint (Eq. (12)), Eq. (10) is simplified to

$$\min(2 - 2\sum_{i,j} n_{ij}(\Delta t) x_i x_j). \tag{13}$$

Using Lagrange multipliers to find the extremum of (half of) Eq. (13),

$$\frac{d}{dx_m}[1 - \sum_{i,j} n_{ij}(\Delta t) x_i x_j + \lambda \sum_i n_i x_i^2] = 0, \tag{14}$$

one obtains

$$\sum_j n_{jm}(\Delta t) x_j = \lambda n_m x_m, \tag{15}$$

which after division by n_m simplifies to Eq. (8). Thus, the constrained optimum is attained when x_m is equal to a left eigenvector of matrix **P**.

The variational principles can be used to determine the p^{fold} and the left eigenvector coordinates without explicitly constructing the MSM, since the TSD and the constraints can be computed directly from a reaction coordinate time-series $x(k\Delta t)$. In that case, the p^{fold} function is a function of configuration space $p^{fold}(\vec{X})$, rather than a function of state p_i^{fold}. Analogously, the left eigenvectors become the left eigenfunction; however, for simplicity, we still refer to them as eigenvectors. From Eq. (10), it follows that to obtain the p^{fold} coordinate, one minimizes

$$\min \sum_{i,j} n_{ij}(\Delta t)(x_i - x_j)^2 = \min \sum_k [x(\Delta t + k\Delta t) - x(k\Delta t)]^2 \tag{16}$$

under the constraints that $x_{k \in A} = 0$ and $x_{k \in B} = 1$, where $k \in A$ and $k \in B$ denote the points $x(k\Delta t)$ that belong to basins A and B, respectively.

From Eq. (13), it follows that to determine the left eigenvectors, one maximizes

$$\max \sum_{i,j} n_{ij}(\Delta t) x_i x_j = \max \sum_k x(\Delta t + k\Delta t) x(k\Delta t) \tag{17}$$

under constraint

$$\sum_i n_i x_i^2 = \sum_k x^2(k\Delta t) = 1. \tag{18}$$

To implement such a variational principle in practice, one needs to explicitly define the RC. To this end, one introduces a functional form $R(\vec{X}, \vec{\alpha})$ for the RC, where $\vec{X}$ are the coordinates of the configuration space and $\vec{\alpha}$ is a vector of parameters. In other words, for every snapshot, the functional $R(\vec{X}(k\Delta t), \vec{\alpha})$ projects the multidimensional trajectory $\vec{X}(k\Delta t)$ onto reaction coordinate $x(k\Delta t)$,

$$x(k\Delta t) = R(\vec{X}(k\Delta t), \vec{\alpha}). \tag{19}$$

Then, the parameters α_i of the functional form are numerically optimized to minimize the functionals (Eq. (16) or (17)) under the corresponding constraints.[13,27,48] In the case when the functional form is a linear combination of basis functions,

$$R(\vec{X}(k\Delta t), \vec{\alpha}) = \sum_i \alpha_i f_i(\vec{X}(k\Delta t)), \tag{20}$$

the optima can be found analytically.[25,48] For the p^{fold} coordinate, one solves the systems of linear equations

$$\mathbf{A}\vec{\alpha} = \mathbf{b}. \tag{21}$$

For the left eigenvectors, one solves the generalized eigenvalue problem

$$\mathbf{C}\vec{\alpha} = \lambda \mathbf{D}\vec{\alpha}. \tag{22}$$

In particular, to approximate the second left eigenvector $\vec{u}_2$, one takes the second generalized eigenvector $\vec{\alpha}$. The derivation of the equations and the expressions for **A**, **C**, **D**, and **b** are presented in the Appendix.

Once the optimal linear coefficients α_i have been found, the multidimensional trajectory $\vec{X}(k\Delta t)$ is projected onto an optimal coordinate by computing $x(k\Delta t)$ for every snapshot (Eqs. (19) and (20)).

Analogous variational approaches for finding the left eigenvectors have been suggested by other groups.[51,54] However, there is an important conceptual difference. In these works, the eigenvectors are evaluated with the aim to approximate the transfer operator, which describes the dynamics without explicitly constructing a MSM. We consider the left eigenvectors (in particular, the second eigenvector) as an approximation to the optimal RC x. Once it has been found, the dynamics of the system is described by the FEP $F(x)$ together with the diffusion coefficient $D(x)$ constructed along x.[13,28,55]

The requirement of the functional form to be a linear combination of basis functions significantly restricts the flexibility of the RC and may lead to suboptimal solutions. We describe a simple idea which allows one to optimize the RC without selecting any specific functional form.

III. METHOD

A. General idea

In principle, any continuous function of coordinates $(X_1,\ldots,X_n)$ can be approximated with arbitrary precision in a compact region of configuration space by a polynomial (e.g., a finite Taylor series),

$$R(X_1,\ldots,X_n) = \sum_{i_1,\ldots,i_n} \alpha_{i_1,\ldots,i_n} X_1^{i_1}\ldots X_n^{i_n}. \qquad (23)$$

Since the function is linear with respect to the Taylor coefficients $\alpha_{i_1,\ldots,i_n}$, the optimal values of the coefficients which provide the extremum to quadratic functionals (16) and (17) under corresponding constraints can be found analytically. However, the dimensionality of the problem (i.e., the number of coefficients required) is huge, which makes such a straightforward approach unrealistic. Instead, we suggest to optimize this function iteratively. At each iteration, one considers only a subset of the Taylor terms and coefficients.

One can imagine different schemes of iterative optimization. We found it useful to take the current putative RC as one of the variables. Namely, at each iteration, we optimize the coordinate in a recursive way $R' = f(R, X_m)$, where R is the RC obtained on the previous step, X_m is a randomly selected coordinate of the configuration space $\vec{X}$, and f is some fixed polynomial of small degree, whose optimum can be found efficiently by solving corresponding Eqs. (21) and (22). Here, we consider the polynomial

$$f(x,y) = \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} a_{ij} x^i y^j, \qquad (24)$$

where (in our case) x corresponds to reaction coordinate R and y corresponds to X_m. After finding the extremum of the quadratic functional, i.e., finding the optimal values for a_{ij}, the RC time series is updated $R(k\Delta t) = R'(k\Delta t)$ and the procedure is repeated with updated $R(k\Delta t)$.

The polynomial $f(x,y)$ has n^2 coefficients and, correspondingly, matrices **A**, **C**, and **D** (Eqs. (21) and (22)) have n^4 elements, which means that the computational cost for each iteration scales with n as $O(n^4)$. An optimal value of n is a compromise between the approximation power, which grows with n and the computational cost. Here, we used $n = 4$.

Such a recursive approach to iterative optimization of RC has a rational explanation. First, a contribution from each basis function to the RC (Eq. (20)) may depend on where the system is between the two free energy basins. Consider, for example, the number of native contacts coordinate for the description of protein folding dynamics. One may argue that formation of some native contacts during the early stages of the folding process may lead to the formation of off-pathway intermediates and actually hinder the folding process. Thus, the contribution from each native contact may depend on the overall progress of the system during folding, and the best way to measure the progress of folding is the RC itself. Second, if the function $f(R, X_m)$ is a low-degree polynomial, then as long as it contains powers of R higher than linear, it is a simple way to implicitly populate R with Taylor monomials $(X_1^{i_1}\ldots X_n^{i_n})$ of arbitrary degree. An alternative is to iteratively introduce different monomials in an explicit way by considering $f(R, X_1^{i_1}\ldots X_n^{i_n})$.

One can also consider recursive transformations with a polynomial of single variable only $R' = g(R)$, where $g(x) = \sum_{i=0}^{n-1} a_i x^i$. Since the polynomial depends on a single variable, one can make it of high degree (20 here) to better approximate the p^{fold} coordinate. Such iterative optimization can be considered as complimentary to the above. It does not use additional information, contained in variables X_m; however, due to the higher degree of the polynomial, it has a higher degree of flexibility in transforming the putative coordinate R. We found alternating use of these two recursive transformations a good optimization strategy.

B. Practical implementation — The p^{fold} coordinate

Here, we describe a particular implementation of the general idea presented above for the determination of the p^{fold} reaction coordinate. Namely, given a long equilibrium multidimensional trajectory $\vec{X}(k\Delta t)$ and two boundary states A and B, we determine the value of p^{fold} for every snapshot of the trajectory, i.e., $R(k\Delta t) = p^{fold}(\vec{X}(k\Delta t))$.

1. *Initialization*

The two states A and B can be specified by providing an initial seed RC time series $rc_0(k\Delta t)$ and two thresholds, so that all points with $rc_0(k\Delta t) < rc_A$ or $rc_0(k\Delta t) > rc_B$ belong to state A or B, respectively, and are assigned $R(k\Delta t \in A) = x_A = 0$ or $R(k\Delta t \in B) = x_B = 1$. These boundary points are not modified during the iterative optimization procedure. The RC time series is initialized by re-scaling the seed RC as

$$R(k\Delta t) = \frac{rc_0(k\Delta t) - rc_A}{rc_B - rc_A}. \qquad (25)$$

2. Iterations

At every iteration, the putative reaction coordinate $R(k\Delta t)$ is optimized by using information contained in a single variable time-series $y(k\Delta t)$ obtained from the multidimensional trajectory $\vec{X}(k\Delta t)$. As $y(k\Delta t)$, one can take a randomly selected component of the multidimensional trajectory $y(k\Delta t) = X_m(k\Delta t)$. For analysis of protein folding dynamics, one would rather take an internal degree of freedom, which is invariant to translations and rotations. In particular, one can use the distance time-series $r_{ml}(k\Delta t)$ between a randomly selected pair of atoms m and l or time-series $\cos(\phi(k\Delta t))$ or $\sin(\phi(k\Delta t))$, where ϕ is a dihedral angle.

The putative coordinate is improved by finding the minimum of the TSD,

$$\min \sum_k [R'(k\Delta t + \Delta t) - R'(k\Delta t)]^2, \tag{26}$$

where $R'(k\Delta t) = f(R(k\Delta t), y(k\Delta t))$ is represented by the polynomial in Eq. (24). The optimal coefficients of the polynomial are found by solving Eq. (21). It may happen that during the initial stage of optimization for some snapshots, one may have $R(k\Delta t) < 0$ or $R(k\Delta t) > 1$. It should not pose a problem as eventually all the values of $R(k\Delta t)$ will be in the range $[0, 1]$.

3. Termination criteria

To stop the iterative improvement, a number of criteria can be suggested. They are based on the observation that if the putative coordinate R is equal to the p^{fold} coordinate, then the partition function $Z_{C,1}(R, \Delta t)$ of the generalized cut profile $F_{C,1}(R, \Delta t)$, computed using transition paths, is position and sampling interval independent and equals the number of transition from A to B,[28]

$$Z_{C,1}(R, \Delta t) = N_{AB}. \tag{27}$$

The partition function $Z_{C,1}(x, \Delta t)$ (at point x) is calculated as half the sum of the distances for those trajectory steps that go through the point x,[28]

$$Z_{C,1}(x, \Delta t) = \frac{1}{2} \sum_k{}' |x(k\Delta t) - x(k\Delta t + \Delta t)|, \tag{28}$$

where the prime sign over the sum indicates here that the sum is taken over such k and x is between $x(k\Delta t)$ and $x(k\Delta t + \Delta t)$.

For example, one can stop optimization, when the mean deviation of $Z_{C,1}(R, \Delta t)$ from N_{AB} reaches a minimum. It is the most stringent criterion and can also be used for testing whether a putative RC is indeed optimal (which is discussed in Subsection III B 4).

To avoid computation of the cut profile, one can integrate $Z_{C,1}(R, \Delta t)$ along R and obtain another criterion[28]

$$\int_0^1 Z_{C,1}(R, \Delta t) dR = \frac{1}{2} \sum_k [R(k\Delta t + \Delta t) - R(k\Delta t)]^2 = N_{AB}, \tag{29}$$

namely, to stop iterations when half the TSD computed along the trajectory is less than or equal to the number N_{AB} of transitions from one state to the other. In the regime of very good sampling, when statistical errors are small and there is no over-fitting, the two criteria are equivalent. The minimum of the TSD is attained when the RC equals the p^{fold} coordinate (Eq. (16)).

The optimization can also be terminated after a predefined number of iterations. That can be useful if the theoretical lower bound of the TSD (Eq. (29)) cannot be reached because of statistical fluctuations.

4. Testing reaction coordinate optimality

Whether a putative coordinate is optimal (i.e., whether it accurately approximates p^{fold}) can be tested by using a recently suggested criterion.[28] It states that for such a coordinate, the cut-based profiles

$$F_{C,1}(R, \Delta t) = -k_B T \ln Z_{C,1}(R, \Delta t) \tag{30}$$

computed from an ensemble of transition path segments are position R and sampling interval Δt independent.[28] Here, the time interval $\Delta t = 1, 2, 4, \ldots$ is measured in integer numbers of the trajectory steps. Constancy of the profiles $F_{C,1}(R, \Delta t)$ implies also that the mean first passage time (mfpt) computed using Kramer's equation for diffusion over FEP $F(R)$ with position dependent diffusion coefficient $D(R)$ is equal to the mfpt computed directly from the trajectories, i.e., it means that the FEP accurately describes the kinetics.[28] If the profile $F_{C,1}(R, \Delta t)$ is getting higher with increasing Δt, it means that consecutive displacements are negatively correlated and the dynamics is sub-diffusive.[28]

The criterion is implemented in the fep1d.py script[56] and works briefly as follows. To analyze an arbitrary coordinate (z), the coordinate is first transformed to $p^{fold}(z)$. To this end, the original coordinate time-series is discretized and a Markov state model is constructed by computing the transition matrix with sampling interval $\Delta t = 1$ (Eq. (4)). Thus, the $p^{fold}(z)$ coordinate is determined (Eq. (5)) and denoted as pfoldMSM(z). Then, profiles $F_{C,1}(pfoldMSM(z), \Delta t)$ are computed along the pfoldMSM(z) coordinate using transition path segments for a sequence of sampling intervals $\Delta t = 1, 2, \ldots, 2^{16}$. $F_{C,1}(pfoldMSM(z), \Delta t)$ computed for $\Delta t = 1$ is position independent by construction. For the optimal coordinate, all constructed $F_{C,1}(pfoldMSM(z), \Delta t)$ should be constant and equal to $F_{C,1}(pfoldMSM(z), \Delta t = 1)$. Below we use the fep1d.py script for transformations of RCs to more convenient ones as well as for construction of profiles and testing RC optimality.

C. Practical implementation — The left eigenvectors

Here, we describe a particular implementation of the approach for the determination of the left eigenvectors. Namely, given a long equilibrium multidimensional trajectory $\vec{X}(k\Delta t)$, we find the projection of the trajectory on the left eigenvector with the second highest eigenvalue $(\vec{u}_2)$, i.e., the value of $u_2(k\Delta t)$ for every snapshot of trajectory.

1. Initialization

As a seed coordinate, one can take a randomly selected component of the multidimensional trajectory $R(k\Delta t) = y(k\Delta t)$.

2. *Iterations*

At every iteration, the putative second left eigenvector is optimized by using information contained in a single variable time-series $y(k\Delta t)$ obtained from the multidimensional trajectory $\vec{X}(k\Delta t)$. The putative eigenvector is updated by finding the maximum of

$$\max \sum_k R'(k\Delta t + \Delta t)R'(k\Delta t) \tag{31}$$

under the constraint

$$\sum_k R'(k\Delta t)^2 = 1, \tag{32}$$

where $R'(k\Delta t) = f(R(k\Delta t), y(k\Delta t))$ is represented by the polynomial in Eq. (24). The optimal parameters α_i of the polynomial are found by solving the generalized eigenvalue problem Eq. (22). Namely, one is interested in the eigenvector α_i, corresponding to the second highest eigenvalue.

3. *Termination criterion*

There is no criterion analogous to that for the p^{fold} coordinate above. The optimization was terminated after a specified number of iterations.

IV. ILLUSTRATIVE EXAMPLES

To illustrate the approach, we apply the method to two model systems — a low dimensional and a high dimensional one. We use them to discuss two types of statistical errors. The first one is independent of the underlying dimensionality of the configuration space and presents even in one-dimensional systems. This is due to limited sampling of the resulting one-dimensional optimal RC. In particular, limited sampling affects the accuracy of the determined FEPs, and consequently, the optimality criterion, i.e., $F_{C,1}(x)$ profile along the coordinate $x = p^{fold}$, is not exactly constant but fluctuates around the true value. These errors should decrease with increasing total number of transition events.

The second type is associated with the infamous curse of dimensionality. With increasing dimensionality, the volume of the configuration space grows exponentially and one may not expect good coverage of all the configuration spaces with a trajectory of realistic length. One outcome of such poor coverage could be over-fitting during optimization.[28] A longer trajectory will cover more parts of the configuration space; however, to cover the entire space, the trajectory should be of unrealistically long length.

A. A low dimensional model system

Consider a system, where two states A and B are connected by two parallel one-dimensional pathways.[28] The configuration space of the system is described by index $i = 1$ or $i = 2$ which denotes the pathway and by continuous variable x, denoting the position along the pathway $0 < x < 1$. The corresponding terminal nodes of the pathways are considered to be identical: $(x = 0, i = 1) = (x = 0, i = 2) = A$ and $(x = 1, i = 1) = (x = 1, i = 2) = B$. Each pathway has a barrier described by the potential

$$U(i,x) = 2\exp[-9(3x - i)^2].$$

The trajectory was generated by simulating Metropolis Monte Carlo (MC) dynamics for 10^8 steps at dimensionless temperature $k_BT = 1.0$ with diffusion coefficient $D(x) = 0.0001$. The coordinates were saved every 10 steps, resulting in the trajectory of 10^7 steps with 1294 transitions in both directions. It is clear that the trajectory covers the entire configuration space, so there are no statistical errors of the second type.

As a seed (sub-optimal) reaction coordinate R, we take position x. To test the optimality of the seed RC, we apply the criterion described in Sec. III; namely, we construct cut profiles $F_{C,1}(x,\Delta t)$ at different sampling intervals $\Delta t = 1, 2, \ldots, 2^{16}$ and check whether they are constant. Figure 1(a) shows that the profile changes significantly with respect to x and Δt. In particular, it changes from $F_{C,1}(x) \approx -6.85$ at $\Delta t = 1$ to $F_{C,1}(x) \approx -6.46$ at $\Delta t = 65\,536$, i.e., it suggests the kinetics which are faster by factor of $\exp(0.4) \approx 1.5$ compared to the actual kinetics and it confirms that the seed RC is sub-optimal.

The seed coordinate is iteratively optimized using the described approach. Briefly, a new R' is sought in the form $R' = f(R, X_m)$ given by the polynomial in Eq. (24). As the coordinate X_m, we randomly select one of the variables of the configuration space: either i or x. States A and B were defined using the seed coordinate with thresholds $rc_A = 0.01$ and $rc_B = 0.99$. The coefficients α_i of the polynomial are found by minimizing the TSD,

$$\min \sum_k (R'(\Delta t + k\Delta t) - R'(k\Delta t))^2,$$

for details see the Appendix (Eqs. (A1) and (A2)). The RC time series is updated $R(k\Delta t) = R'(k\Delta t)$ and the procedure is repeated.

After 20 iterations, the termination criteria (Eq. (29)) were satisfied: the TSD/2 has dropped to 640.331 ($N_{AB} = 647$). Continuing optimization for further 1000 iterations only lowered the TSD/2 to 640.306, indicating that long optimization without a stopping criterion does not lead to over-fitting. Figure 1(b) shows the application of the optimality criterion to the optimized coordinate. As one can see that the profiles are almost constant with respect to x and Δt, meaning that the optimized coordinate closely approximates p^{fold}.

B. A high dimensional model system

Here, we consider a model system with high dimensional configuration space $(\vec{X})$, which cannot be covered by a trajectory of realistic length. In order to be able to verify the final optimized coordinate, we chose a system for which the optimal coordinate is known. Specifically, we consider isotropic diffusion in n-dimensional space (here n = 50) with radially symmetric potential energy $U(r)$. An optimal RC for the diffusion towards the center for such systems is radius $r = \sqrt{\sum_{i=1}^n X_i^2}$ with the FEP constructed along r being

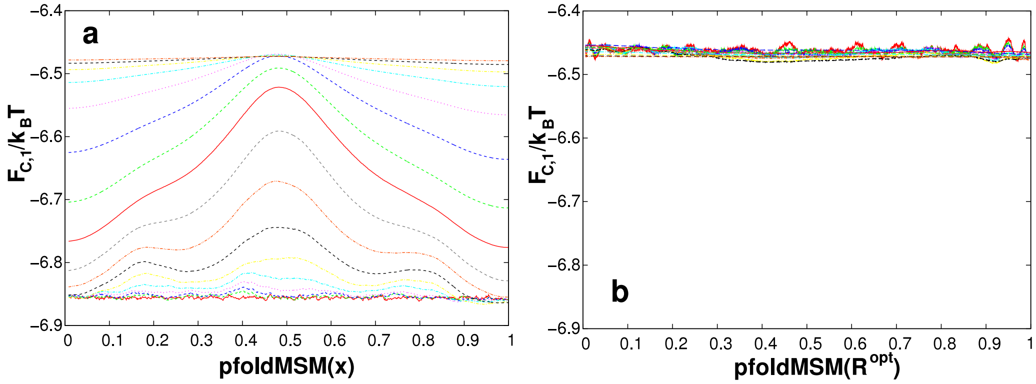

FIG. 1. The optimality criterion applied to the seed (panel (a)) and optimized (panel (b)) RCs. (a) $F_{C,1}$ gets higher with Δt increasing, indicating that the seed coordinate is sub-optimal. (b) $F_{C,1}$ are approximately constant, indicating that the optimized coordinate closely approximates p^{fold}.

$$F(r) = U(r) - (n-1)kT\ln(r).$$

Potential energy along r is given by

$$U(r) = 4\exp[-(r-4)^2] + 4\exp[-(r-6)^2] + 49kT\ln(r),$$

where an infinite wall is placed at $r = 10$. The potential energy is chosen so that the FEP has two large basins with a small intermediate. The trajectory was generated by simulating MC dynamics for 10^6 steps with a diffusion coefficient along each axis $D(X_i) = 1$. It is clear that such a trajectory cannot cover the configuration space densely enough; since just to visit all possible regions with different combinations of coordinate signs, one needs $2^{50} \sim 10^{15}$ points.

As a seed sub-optimal reaction coordinate, $\xi = \sum_i |X_i|$ is taken. To visualize the difference between $F(r)$ and $F(\xi)$ (i.e., the FEPs as functions of the optimal and suboptimal RCs), we proceed as follows. ξ and r are first transformed to $Z_A(\xi)$ and $Z_A(r)$ coordinates, which measure the relative partition function of the coordinate segment between $-\infty$ and ξ or between $-\infty$ and r, respectively,

$$Z_A(x) = \int_{-\infty}^{x} Z_H(w)dw \Big/ \int_{-\infty}^{\infty} Z_H(w)dw.$$

Z_A is invariant to the coordinate transformation, meaning that transition states and intermediate states should have the same positions along Z_A and can be used to compare different coordinates.[27] Then, we construct cut based FEPs $F_C(Z_A(\xi))$ and $F_C(Z_A(r))$ along the transformed RCs. A cut profile $F_C(x)$ along an arbitrary coordinate x is calculated as

$$F_C(x)/k_BT = -\ln(Z_C(x)), \tag{33}$$

where $Z_C(x)$ is the corresponding partition function defined as half the total number of transitions through point x and is also invariant to reaction coordinate transformation.[27] This means that if ξ and r are equivalent (e.g., are connected by a monotonous transformation $\xi = f(r)$), the profiles should coincide $F_C(Z_A(\xi)) = F_C(Z_A(r))$. Figure 2(a) shows that $F_C(Z_A(\xi))$ (green line) is lower than $F_C(Z_A(r))$ (red line). Note that the positions of the transition states along Z_A coincide. Figure 2(c) shows the application of the optimality criterion to the seed RC. The profiles $F_{C,1}(\xi,\Delta t)$ increase with the time interval changing from $\Delta t = 1$ to $\Delta t = 2^{16}$. Though the difference between the first and the last profiles as well as the difference between $F_C(Z_A(\xi))$ and $F_C(Z_A(r))$ (Figure 2(a)) is only around $0.05k_BT$, we show that the ξ coordinate can be further improved.

The seed coordinate is iteratively optimized using the described approach. To this end, a new R' is sought in the form $R' = f(R, X_m)$ given by the polynomial in Eq. (24) by minimizing Eq. (26) (see the Appendix). As the coordinate X_m, we randomly select one of the variables x_i of configuration space. The boundary nodes are defined by $\xi < \xi_A = 10.9$ and $\xi > \xi_B = 48.5$, which are the positions of the minima on $F_C(\xi)$ (Figure 2(b)). The coordinate is iteratively optimized until the TSD-based stopping criterion (Eq. (29)) is satisfied. After 9100 iterations, the termination criterion was satisfied: the TSD/2 has dropped to 988.9 ($N_{AB} = 989$). The optimized reaction coordinate is denoted as R^{opt}. The cut profile $F_C(Z_A(R^{opt}))$ constructed along R^{opt} transformed to the Z_A coordinate approximates $F_C(Z_A(r))$ very closely (Figure 2(a)). Figure 2(d) shows the application of the optimality criterion to R^{opt}. The cut profiles $F_{C,1}(R^{opt},\Delta t)$ just fluctuate around the limiting value, meaning that R^{opt} closely approximates p^{fold} as well as the analytical optimal coordinate r. Continuation of the optimization to 100 000 iterations in total insignificantly decreased the TSD/2 to 986.5. The cut profiles for both optimized coordinates (Figure 2(a) — pink and blue lines) are virtually indistinguishable.

The results allow us to emphasize the following. Even though the difference between the profiles along the optimal r and seed (sub-optimal) ξ coordinates is very small (Figure 2(a)), the proposed approach was able to improve the seed coordinate further and makes it the optimal one (Figures 2(c) and 2(d)), meaning that the approach is very sensitive. Since further optimization to 100 000 iterations in total does not lead to over-optimization or over-fitting, it means that the approach is robust.

To further test the robustness of the algorithm and the obtained results, the following experiments were performed. Five more trajectories of the same length with different random seed values were simulated. The results of their optimization and analysis are very similar to those above, with the only difference being that for some of the trajectories, the TSD/2 could not reach the lower bound of N_{AB} and the optimization had to be terminated after a specified number of iterations. We remind the reader that this is due to statistical fluctuations; the lower bound of TSD/2 is exactly equal to N_{AB} only in the limit

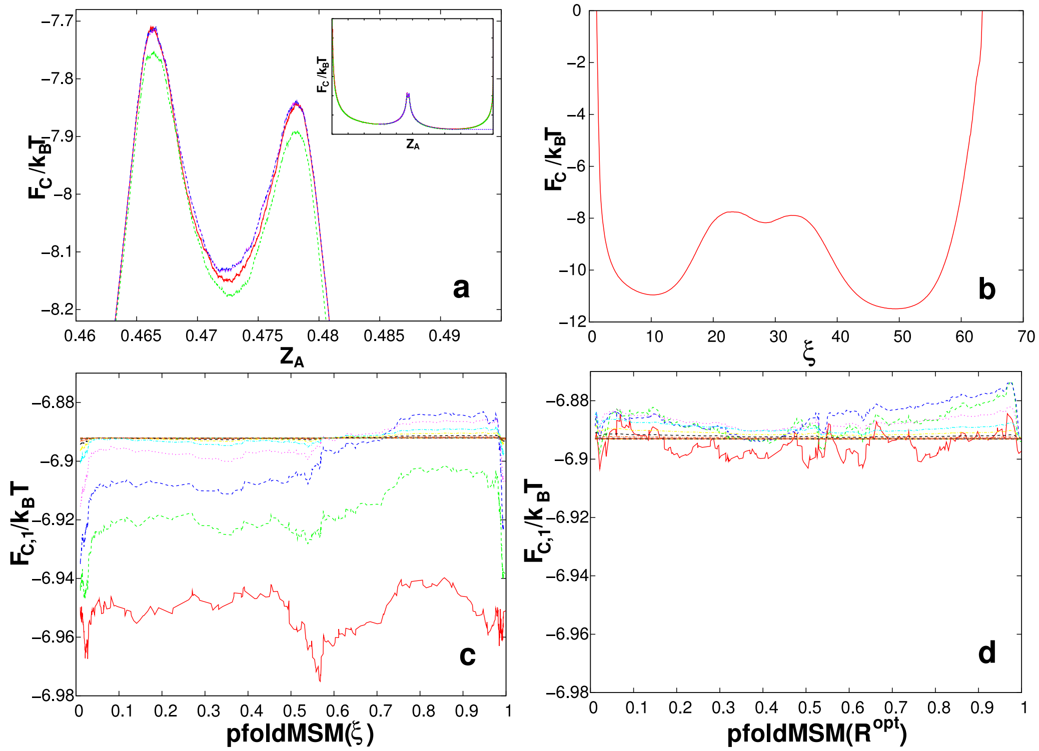

FIG. 2. (a) Cut based FEPs F_C as functions of the following RCs: analytical optimal $r = \sqrt{\sum_i X_i^2}$ (red line); sub-optimal $\xi = \sum_i |X_i|$ (green line); putative optimal coordinate R^{opt} optimized until the TSD based termination criterion has been satisfied (blue line) and that optimized for 100 000 iterations (pink line). The coordinates are rescaled to the Z_A coordinate to facilitate comparison. The plot shows the profiles around the transition states, while the inset shows the entire landscape. (b) Cut based FEP $F_C(\xi)$ along the seed coordinate. (c) Optimality criterion applied to the seed reaction coordinate ξ. (d) Optimality criterion applied to the optimized coordinate R^{opt} (blue line on panel (a)).

of infinite statistics. Additionally, instead of using Eq. (25), the seed coordinate was initialized to a constant value of 0.5 (the boundary nodes have corresponding values of 0 or 1). The results were the same as above.

C. Over-fitting of the under-sampled systems

Here, we illustrate the performance of the algorithm in the case where a system has not been sufficiently sampled. We consider the first 10 000 steps of the trajectory of the high-dimensional system from the previous example. After 20 000 optimization steps, the TSD/2 has converged to 1.294, while the number of transitions is 11, which is an indication of severe over-fitting. The optimality criterion (not shown) confirms this by showing that the cut profiles decrease with increasing sampling interval.

The time-series of the determined optimal coordinate (for the first 10 000 steps) show interesting behavior (Figure 3(a)): it either equals to 0 or 1, or linearly interpolates between the two, so that the snapshots are placed equidistantly. Such an "optimal" coordinate clearly provides an inaccurate description of the dynamics. During each trajectory segment, the system either stays at a boundary or moves with a constant velocity to the opposite boundary. The time-series of the optimal coordinate, determined from the entire trajectory (Figure 3(b)), shows expected behavior, corresponding to stochastic motion back and forth along the coordinate.

This result due to over-fitting can be explained as follows. To find the optimal coordinate, we minimize $\sum_k [x(k\Delta t + \Delta t) - x(k\Delta t)]^2$, under the corresponding constraint. During every iteration, we select the optimal parameters of the polynomial, which update $x(k\Delta t)$, so that the functional yields a smaller value. When a trajectory is relatively short and there are not enough points to densely populate the configuration space, one can assume that all the points of the trajectory $\vec{X}(k\Delta t)$ have very different coordinates, so that a polynomial even with a small degree can be used to separate them significantly. In other words, one can assume that during optimization of under-sampled configuration space, one can change the position of every point $x(k\Delta t)$ independently. In this case, the minimum of the functional can be easily found. Each segment of the trajectory which starts and ends at the boundary points can be considered independently because positions of the boundary points are fixed to 0 and 1. If a trajectory segment starts and ends at 0, then the optimal positions of all its points are $x(k\Delta t) = 0$; if a trajectory segment starts and ends at 1, then all the $x(k\Delta t) = 1$; if a trajectory segment starts and ends at different boundary points, then all the intermediate $x(k\Delta t)$ are positioned equidistantly between 0 and 1.

If, on the contrary, the configuration space has been well sampled, i.e., the neighboring regions of most points are densely populated by other points, which are transformed by the polynomials in a similar way, then one cannot consider these points as independent. In other words, one seeks an optimal position not for every single point, but rather for clusters of points, i.e., one optimizes the positions of an implicitly constructed MSM. The effective size of the clusters

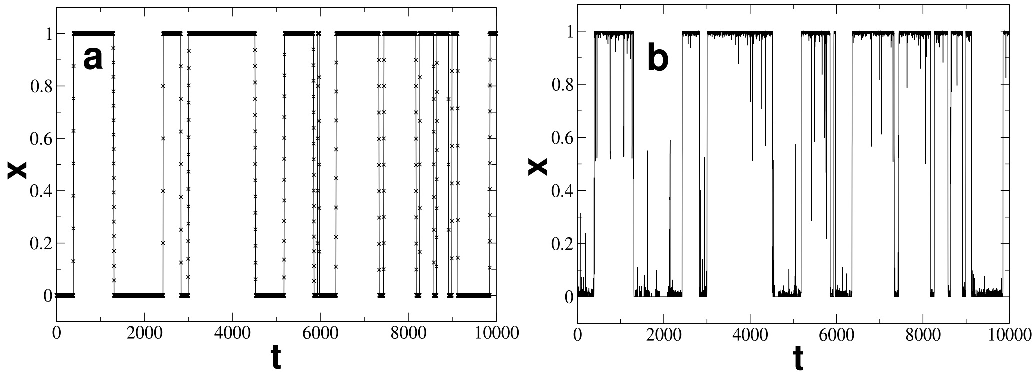

FIG. 3. The time-series of the determined optimal coordinates. (a) Optimization using the first 10 000 steps of the trajectory, which leads to over-fitting. (b) Optimization using the entire trajectory.

is controlled implicitly by the flexibility of the polynomial. The higher degree and more variables the polynomial has, the more independent nearby points shall be. That is why in the proposed iterative optimization method, only a two dimensional subspace of the entire configuration space is considered at every iteration, namely, that consisting of R and y.

D. Iterative optimization of the left eigenvectors

In this section, we illustrate the optimization procedure of finding the second highest left eigenvector. We apply it to the MC trajectory from the high-dimensional model system considered above. We remind the reader that for this type of optimization, there is no need to choose a specific seed coordinate that defines or separates the two boundary states; for example, any variable X_i of configuration space can serve as initial R coordinate.

The optimization iteratively increased the second eigenvalue. The optimized RC (the second eigenvector) is denoted as $\vec{u}_2$. After 600 iterations, the FEP along the corresponding second eigenvector $\vec{u}_2$, transformed to the Z_A coordinate, describes the free energy barrier of interest (Figure 4(a)). However, the barrier of $F_C(Z_A(\vec{u}_2))$ (green line) is slightly lower than that of $F_C(Z_A(r))$ (red line). One may expect that further optimization can make $\vec{u}_2$ closer to r or p^{fold} and correspondingly make the barrier higher. However, optimization for another 200 iterations, while increasing the second eigenvalue, did not improve the free energy barrier and made it even lower (Figure 4(a) — blue line). Fig. 4(b) shows projections of the trajectory on both second eigenvectors. In the time-series projected on $\vec{u}_2$, optimized for 800 iterations, one can see a high peak at around $t = 1.138 \cdot 10^5$, which is absent in the other time-series. It shows that during the additional 200 iterations, the optimization procedure detected a region of configuration space that has been visited very briefly and only once. For the optimization procedure, this means that this region has a high free energy barrier and a high eigenvalue.

It points to the fact that such an eigenvector optimization algorithm possesses quite a generic instability. The optimization procedure tries to find the eigenvector corresponding to the highest eigenvalue. However, the latter is not necessarily the eigenvector of interest. For example, in analyzing a protein folding simulation trajectory with many folding-unfolding transitions and a single *trans-cis* transition of a dihedral angle, the latter has a higher free energy barrier to which the optimization procedure shall eventually converge. The case considered here is that for a high dimensional system, many parts of the configuration space are visited only once, so their effective free energy barrier will be higher than the barrier of interest, which has been sampled many times.

Summarizing the straightforward optimization of the left eigenvectors for multidimensional systems is complicated by an inherent instability, when the optimization at some point runs away from the free energy barrier of interest. In contrast, the optimization of the p^{fold} coordinate is free of such a shortcoming but it needs a seed coordinate for defining boundary nodes A and B. Finding such a coordinate or defining such nodes is not always straightforward. One possibility is to use an initially optimized eigenvector as the seed coordinates to start the p^{fold} optimization.

V. CONCLUDING DISCUSSION

A popular approach to reaction coordinate optimization is to select a functional form with many parameters approximating the coordinate and to train it on the trajectories using various criteria. In the end, one finds optimal values of the parameters such that the functional form provides best approximation to the p^{fold} coordinate. A drawback of such an approach is that it is not trivial to select a functional form that can accurately approximate the optimal coordinate. A poorly chosen functional form can lead to suboptimal results.

We have suggested a nonparametric or functional form free approach which allows one to optimize the reaction coordinate without selecting its functional form. Instead of finding a function that approximates the optimal coordinate, the approach determines the value of the coordinate for every snapshot of the trajectory. In particular, given a long multidimensional equilibrium trajectory $\vec{X}(k\Delta t)$, the algorithm determines the value of the p^{fold} coordinate for every snapshot of the trajectory $p^{fold}(k\Delta t)$, i.e., it projects the multidimensional trajectory onto the optimal coordinate. The approach consists of recursive iterative optimization of the coordinate. At every iteration, the coordinate is improved by considering

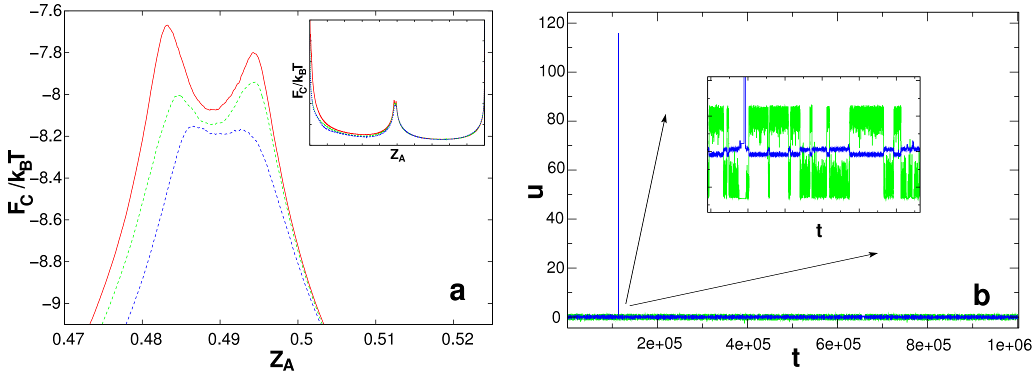

FIG. 4. (a) Cut based FEPs F_C as functions of the following RCs: analytical optimal $r = \sqrt{\sum_i X_i^2}$ (red line); the second left eigenvector $\vec{u}_2$ optimized for 600 iterations (green line) and for 800 iterations (blue). The coordinates were rescaled to the Z_A coordinate to facilitate comparison. The plot shows the profiles around the transition states, while the inset shows the entire landscape. (b) The time-series of the trajectory projected on the left eigenvectors (colored as on panel (a)). The inset enlarges the part of the trajectory that contains the peak.

a small degree polynomial of two variables — the coordinate itself and a randomly selected variable of configuration space.

The approach was successfully tested on a simple one-dimensional system and on a complex multidimensional system, where complete sampling of the configuration space is impractical. In the latter case, the approach demonstrated its sensitivity and robustness. It has improved a seed sub-optimal coordinate to the optimal one, even though the seed coordinate was already quite good; the difference between the free energy profiles constructed along the seed coordinate and the optimal analytical coordinate is of just $0.05kT$. Further optimization for 100 000 iterations in total did not lead to over-fitting. The optimality criterion confirmed that the determined coordinate closely approximates p^{fold} over the entire range, not just around the transition states.

The application of a variant of the approach to find the left eigenvectors to the high dimensional system has shown that the problem of optimizing the left eigenvectors possesses an inherent instability. The approach can find an eigenvector with the highest eigenvalue, but this eigenvector is not necessarily an eigenvector of interest. The left eigenvector coordinate determined by the approach at the early stage of optimization, however, can serve as seed coordinate to start the p^{fold} optimization, especially in cases where the boundary nodes are not straightforward to define.

The next step is to apply the approach for rigorous analysis of the dynamics from state of the art protein folding simulations[15,57] and dynamics in other types of Big Data.[58]

ACKNOWLEDGMENTS

The work has been supported by BBSRC Grant (No. BB/J016055/1).

APPENDIX: DERIVATION OF EQS. 21 AND 22

Here, we derive equations for the vector of coefficients $\vec{\alpha}$ that provide optimum to the optimization functionals (Eqs. (16) and (17)) under different constrains. In the first case, one minimizes (Eq. (16))

$$\min \sum_k [x(\Delta t + k\Delta t) - x(k\Delta t)]^2$$

under the constraint that the positions of the boundary states are fixed, where $x(k\Delta t) = \sum_i \alpha_i f_i(\vec{X}(k\Delta t))$ if the point does not belong to the boundary states and $x(k\Delta t) = 0$ or $x(k\Delta t) = 1$ if the point belongs to state A or B, respectively.

The sum over k can be broken down into four sums: $k \in K_1$ when both points $x(k\Delta t)$ and $x(\Delta t + k\Delta t)$ do not belong to the boundary states

$$\sum_{k\in K_1} [\sum_i \alpha_i f_i(\vec{X}(\Delta t + k\Delta t)) - \sum_i \alpha_i f_i(\vec{X}(k\Delta t))]^2,$$

$k \in K_2$ when the other point belongs to the boundary state A ($x_A = 0$)

$$\sum_{k\in K_2} [\sum_i \alpha_i f_i(\vec{X}(k\Delta t))]^2,$$

$k \in K_3$ when the other point belongs to the boundary state B ($x_B = 1$)

$$\sum_{k\in K_3} [1 - \sum_i \alpha_i f_i(\vec{X}(k\Delta t))]^2,$$

and $k \in K_4$ when both points belong to the different boundary states

$$\sum_{k\in K_4} [1 - 0]^2.$$

Opening the brackets, the entire sum can be written as

$$\sum_{i,j} \alpha_i\alpha_j \sum_{k\in K_1} [f_i(\vec{X}(\Delta t + k\Delta t)) - f_i(\vec{X}(k\Delta t))][f_j(\vec{X}(\Delta t + k\Delta t)) - f_j(\vec{X}(k\Delta t))] + \sum_{i,j} \alpha_i\alpha_j \sum_{k\in K_2} f_i(\vec{X}(k\Delta t))f_j(\vec{X}(k\Delta t))$$
$$+ \sum_{i,j} \alpha_i\alpha_j \sum_{k\in K_3} f_i(\vec{X}(k\Delta t))f_j(\vec{X}(k\Delta t)) - 2\sum_i \alpha_i \sum_{k\in K_3} f_i(\vec{X}(k\Delta t)) + \sum_{k\in K_3} 1 + \sum_{k\in K_4} 1 = \sum_{i,j} \alpha_i\alpha_j A_{ij} - 2\sum_i \alpha_i b_i + c, \qquad \text{(A1)}$$

where the elements of matrix **A** and vector **b** are

$$\begin{aligned} A_{ij} &= \sum_{k\in K_1} [f_i(\vec{X}(\Delta t + k\Delta t)) - f_i(\vec{X}(k\Delta t))][f_j(\vec{X}(\Delta t + k\Delta t)) - f_j(\vec{X}(k\Delta t))] + \sum_{k\in K_2+K_3} f_i(\vec{X}(k\Delta t)) f_j(\vec{X}(k\Delta t)),\\ b_i &= \sum_{k\in K_3} f_i(\vec{X}(k\Delta t)),\\ c &= \sum_{k\in K_3+K_4} 1. \end{aligned} \tag{A2}$$

Taking the derivative of Eq. (A1) with respect to α_m and equating it to 0,

$$\frac{d}{d\alpha_m}\left(\sum_{i,j} \alpha_i \alpha_j A_{ij} - 2\sum_i \alpha_i b_i + c\right) = 0,$$

one obtains the system of linear equations for α coefficients which provide the minimum of the quadratic functional (Eq. (A1))

$$\mathbf{A}\vec{\alpha} = \mathbf{b}.$$

For the eigenvalue problem, one maximizes (Eq. (17))

$$\begin{aligned} \sum_k x(\Delta t + k\Delta t)x(k\Delta t) &= \sum_k \Big[\sum_i \alpha_i f_i(\vec{X}(\Delta t + k\Delta t)) \sum_j \alpha_j f_j(\vec{X}(k\Delta t))\Big]\\ &= \sum_{i,j} \alpha_i \alpha_j \sum_k f_i(\vec{X}(\Delta t + k\Delta t)) f_j(\vec{X}(k\Delta t))\\ &= \sum_{i,j} \alpha_i \alpha_j C_{ij} \end{aligned} \tag{A3}$$

under constraint (Eq. (18))

$$\begin{aligned} \sum_k x(k\Delta t)x(k\Delta t) &= \sum_{i,j} \alpha_i \alpha_j \sum_k f_i(\vec{X}(k\Delta t)) f_j(\vec{X}(k\Delta t))\\ &= \sum_{i,j} \alpha_i \alpha_j D_{ij} = 1, \end{aligned} \tag{A4}$$

where

$$\begin{aligned} C_{ij} &= \sum_k f_i(\vec{X}(\Delta t + k\Delta t)) f_j(\vec{X}(k\Delta t)),\\ D_{ij} &= \sum_k f_i(\vec{X}(k\Delta t)) f_j(\vec{X}(k\Delta t)). \end{aligned}$$

Using the Lagrange multipliers, one finds that the optimal vector $\vec{\alpha}$ is the solution of the generalized eigenvalue problem Eq. (22)

$$\mathbf{C}\vec{\alpha} = \lambda \mathbf{D}\vec{\alpha}.$$

[1] F. Noé and S. Fischer, Curr. Opin. Struct. Biol. **18**, 154 (2008).

[2] G. R. Bowman, K. A. Beauchamp, G. Boxer, and V. S. Pande, J. Chem. Phys. **131**, 124101 (2009).

[3] V. Pande, K. Beauchamp, and G. Bowman, Methods **52**, 99 (2010).

[4] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, J. Chem. Phys. **134**, 174105 (2011).

[5] J.-H. Prinz, B. Keller, and F. Noé, Phys. Chem. Chem. Phys. **13**, 16912 (2011).

[6] F. Rao and A. Caflisch, J. Mol. Biol. **342**, 299 (2004).

[7] D. Prada-Gracia, J. Gómez-Gardeñes, P. Echenique, and F. Falo, PLoS Comput. Biol. **5**, e1000415 (2009).

[8] A. Dickson and C. L. Brooks, J. Am. Chem. Soc. **135**, 4729 (2013).

[9] J. N. Onuchic, N. D. Socci, Z. Luthey-Schulten, and P. G. Wolynes, Folding Des. **1**, 441 (1996).

[10] C. M. Dobson, A. Šali, and M. Karplus, Angew. Chem., Int. Ed. **37**, 868 (1998).

[11] S. V. Krivov and M. Karplus, Proc. Natl. Acad. Sci. U. S. A. **101**, 14766 (2004).

[12] R. B. Best and G. Hummer, Proc. Natl. Acad. Sci. U. S. A. **102**, 6732 (2005).

[13] S. V. Krivov, J. Phys. Chem. B **115**, 12315 (2011).

[14] M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi, J. Chem. Phys. **134**, 124116 (2011).

[15] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, and W. Wriggers, Science **330**, 341 (2010).

[16] P. V. Banushkina and S. V. Krivov, J. Chem. Theory Comput. **9**, 5257 (2013).

[17] S. Jungblut, A. Singraber, and C. Dellago, Mol. Phys. **111**, 3527 (2013).

[18] P. G. Bolhuis, C. Dellago, and D. Chandler, Proc. Natl. Acad. Sci. U. S. A. **97**, 5877 (2000).

[19] G. Berezovska, D. Prada-Gracia, and F. Rao, J. Chem. Phys. **139**, 035102 (2013).

[20] M. Ceriotti, G. A. Tribello, and M. Parrinello, Proc. Natl. Acad. Sci. U. S. A. **108**, 13023 (2011).

[21] Y. Mu, P. H. Nguyen, and G. Stock, Proteins **58**, 45 (2005).

[22] L. O. Hedges, R. L. Jack, J. P. Garrahan, and D. Chandler, Science **323**, 1309 (2009).

[23] H. Yu, A. N. Gupta, X. Liu, K. Neupane, A. M. Brigley, I. Sosova, and M. T. Woodside, Proc. Natl. Acad. Sci. U. S. A. **109**, 14452 (2012).

[24] P. Schuetz, R. Wuttke, B. Schuler, and A. Caflisch, J. Phys. Chem. B **114**, 15227 (2010).

[25] S. V. Krivov, H. Fenton, P. J. Goldsmith, R. K. Prasad, J. Fisher, and E. Paci, PLoS Comput. Biol. **10**, e1003685 (2014).

[26] S. V. Krivov, Phys. Rev. E **84**, 011135 (2011).

[27] S. V. Krivov, PLoS Comput. Biol. **6**, e1000921 (2010).

[28] S. V. Krivov, J. Chem. Theory Comput. **9**, 135 (2013).

[29] A. Altis, P. H. Nguyen, R. Hegger, and G. Stock, J. Chem. Phys. **126**, 244111 (2007).

[30] D. Antoniou and S. D. Schwartz, J. Phys. Chem. B **115**, 2465 (2011).

[31] R. R. Coifman and S. Lafon, Appl. Comput. Harmonic Anal. **21**, 5 (2006).

[32] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis, Appl. Comput. Harmonic Anal. **21**, 113 (2006).

[33] A. L. Ferguson, A. Z. Panagiotopoulos, I. G. Kevrekidis, and P. G. Debenedetti, Chem. Phys. Lett. **509**, 1 (2011).

[34] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, Science **290**, 2319 (2000).

[35] O. F. Lange and H. Grubmüller, Proteins **70**, 1294 (2008).

[36] W. Li and A. Ma, Mol. Simul. **40**, 784 (2014).

[37] D. J. Wales, J. Chem. Phys. **142**, 130901 (2015).

[38] B. Peters, J. Phys. Chem. B **119**, 6349 (2015).

[39] L. Onsager, Phys. Rev. **54**, 554 (1938).

[40] R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. S. Shakhnovich, J. Chem. Phys. **108**, 334 (1998).

[41] P. L. Geissler, C. Dellago, and D. Chandler, J. Phys. Chem. B **103**, 3706 (1999).

[42] A. Ma and A. R. Dinner, J. Phys. Chem. B **109**, 6769 (2005).

[43] B. Peters and B. L. Trout, J. Chem. Phys. **125**, 054108 (2006).

[44] C. D. Snow, Y. M. Rhee, and V. S. Pande, Biophys. J. **91**, 14 (2006).

[45] G. Hummer, J. Chem. Phys. **120**, 516 (2003).

[46] W. E and E. Vanden-Eijnden, J. Stat. Phys. **123**, 503 (2006).

[47] J. D. Chodera and V. S. Pande, Phys. Rev. Lett. **107**, 098102 (2011).

[48] S. V. Krivov, J. Phys. Chem. B **115**, 11382 (2011).

[49] A. M. Berezhkovskii and A. Szabo, J. Phys. Chem. B **117**, 13115 (2013).
[50] J. Lu and E. Vanden-Eijnden, J. Chem. Phys. **141**, 044109 (2014).
[51] C. R. Schwantes and V. S. Pande, J. Chem. Theory Comput. **11**, 600 (2015).
[52] A. Berezhkovskii, G. Hummer, and A. Szabo, J. Chem. Phys. **130**, 205102 (2009).
[53] A. Berezhkovskii and A. Szabo, J. Chem. Phys. **121**, 9186 (2004).
[54] F. Nüske, B. G. Keller, G. Pérez-Hernández, A. S. J. S. Mey, and F. Noé, J. Chem. Theory Comput. **10**, 1739 (2014).
[55] S. V. Krivov and M. Karplus, Proc. Natl. Acad. Sci. U. S. A. **105**, 13841 (2008).
[56] P. V. Banushkina and S. V. Krivov, J. Comput. Chem. **36**, 878 (2015).
[57] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, Science **334**, 517 (2011).
[58] J. Freeman, N. Vladimirov, T. Kawashima, Y. Mu, N. J. Sofroniew, D. V. Bennett, J. Rosen, C.-T. Yang, L. L. Looger, and M. B. Ahrens, Nat. Methods **11**, 941 (2014).