
Article:
Krivov, SV (2021) Blind Analysis of Molecular Dynamics. Journal of Chemical Theory and Computation, 17 (5). pp. 2725-2736. ISSN: 1549-9618

https://doi.org/10.1021/acs.jctc.0c01277

Blind analysis of molecular dynamics.

Sergei V. Krivov*

Astbury Center for Structural Molecular Biology, Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, United Kingdom

E-mail: s.krivov@leeds.ac.uk

Abstract

We describe a non-parametric approach for accurate determination of the slowest relaxation eigenvectors of molecular dynamics. The approach is blind as it uses no system specific information. In particular, it does not require a functional form with many parameters to closely approximate eigenvectors, e.g., a linear combinations of molecular descriptors or a deep neural network, and thus no extensive expertise with the system. We suggest a rigorous and sensitive validation/optimality criterion for an eigenvector. The criterion uses only eigenvector timeseries and can be used to validate eignevectors computed by other approaches. The power of the approach is illustrated on long atomistic protein folding trajectories. The determined eigenvectors pass the validation test at timescale of 0.2 ns, much shorter than alternative approaches.

1 Introduction

Molecular dynamics simulations increasingly produce massive trajectories.[1,2] Accurate analysis and interpretation of such data are widely recognized as fundamental bottlenecks that could limit their applications, especially in the forthcoming era of exascale computing.[3–9] A rigorous way to analyze dynamics in such data is to describe/approximate it by diffusion

on the free energy landscape, free energy as a function of reaction coordinates (RCs). For such a description to be quantitatively accurate, the RCs should be chosen in an optimal way.[5,10] The committor function is an example of such RCs, that can be used to compute some important properties of the dynamics exactly.[10] The eigenvectors (EVs) of the transfer operator are another example.[11,12] They are often used to decrease the dimensionality of the dynamics during the construction of Markov state models (MSM).[13,14] Incidentally, one embarrassingly parallel strategy to exascale simulations consists of running a very large number of short trajectories independently, which are later combined using MSMs in order to obtain a long time behavior.[7,15]

The minimal lag time when a MSM becomes approximately Markovian, which can be estimated by the convergence of implied timescales or by Chapman-Kolmogorov criterion, is a good indicator of the accuracy of the constructed model. State of the art approaches have lag times in the range of tens of nanoseconds.[13,14,16,17] Shorter lag times mean more accurate putative EVs and MSMs, as well as shorter trajectories and higher efficiency for the simple strategy of exascale simulations. Here we present an approach, which determines EVs for protein folding trajectories, which pass a stringent EV validation test at much shorter lag time of trajectory sampling interval of 0.2 ns.

A major difficulty of parametric approaches is that they require a functional form with many parameters to approximate RCs, e.g., linear combinations of molecular descriptors/features[13,14] or deep neural networks.[16,17] While, e.g., it was argued that "the expressive power of neural networks provides a natural solution to the choice-of-basis problem",[16] finding the optimal architecture of a neural network and input variables are difficult tasks. The suggested approach is non-parametric and can approximate any RC with high accuracy without system specific information. Instead of optimizing the parameters of the approximating function, the approach directly optimizes RC time-series. Such approaches, which use no system specific information and operate in generic, system agnostic terms, such as EVs and eigenvalues, RCs, committors,[10] optimality criteria, free energy landscapes, we propose to call blind, in

analogy with the blind source separation approaches.

The blind approaches should especially be useful in the following cases: i) the initial analysis of the systems dynamics, when the knowledge of the system is very limited; ii) analyses, where one does not want to introduce any bias, e.g., due to the employed function approximation, or one does not have a satisfactory function approximation; iii) they can be used aposteriory, to check if possible bias in the analysis has altered the results.

The initial framework of non-parametric RC optimization was described in Ref. 18. Ref. 10 introduced adaptive version of the approach for the committor RC to treat realistic systems with relatively limited sampling, e.g., state-of-the-art atomistic protein folding trajectories.[1,2,19] To avoid overfitting in such a system, the approach performs RC optimization in an adaptive manner by focusing on less optimized spatiotemproal regions of RC. The latter are identified by using the committor optimality criterion.[20] The current paper makes the following contributions to the nonparametric framework. First, we suggest a rigorous and sensitive validation/optimality criterion for EVs. Second, as we discuss below, the optimization of EVs is inherently unstable.[18] It is not a drawback of the non-parametric approach *per se*, but is rather due to the unsupervised nature of the problem itself. One seeks EVs with the smallest eigenvalues, which describe the slowest relaxation dynamics, however, not all of such EVs are of interest. Here, we describe a few heuristics to suppress the instability. Third, we describe an adaptive approach, which avoids overfitting for realistically sampled systems. Fourth, we illustrate the power of the approach by determining accurate EVs for realistic protein folding trajectories: HP35 double mutant[19] and FIP35.[1]

The paper is as follows. The Methods section starts by reviewing the conventional, parametric approach of EVs approximation. Then, the non-parametric framework of RC optimization is introduced. An iterative, non-parametric approach of EV optimization is described. A stringent EV validation/optimality criterion is suggested. A protein folding trajectory is used to illustrate that iterative EV optimization has an inherently instability. It can converge to EVs with smaller eigenvalues, but of no interest, which we denote as spurious

EVs. An approach with heuristics to suppress the instability is described. Application of the EV criterion shows that during EV optimization some regions of the putative EV are underfitted (suboptimal), while other are overfitted. The criterion is adopted to perform optimization in an adaptive, more uniform way. We conclude by discussing the obtained protein folding free energy landscapes.

2 Method

2.1 Variational optimization of eigenvectors.

Assume that system dynamics is described/approximated by a Markov chain with transition probability matrix $P(i|j,\Delta t)$ for transition from state j to state i after time interval Δt. Note that this assumption is used only for the derivation of equations. One does not need to know the actual Markov chain, meaning that this assumption does not restrict the applicability of the algorithm.

Given a very long equilibrium trajectory $\mathbf{X}(k\Delta t_0)$, where Δt_0 is the trajectory sampling interval, and using a *very* fine-grained clustering of the configuration space of the system, one can, *in principle*, estimate the transition matrix $P(i|j,\Delta t) = n(i|j,\Delta t)/n(j,\Delta t)$, where time interval (lag time) Δt equals Δt_0 or its multiple, $n(i|j,\Delta t)$ is the number of transitions from cluster j to cluster i after time interval Δt, observed in the trajectory and $n(j,\Delta t) = \sum_i n(i|j,\Delta t)$ is the total number of transitions out of cluster i, which is proportional to the equilibrium probability. Knowing $P(i|j,\Delta t)$ one can estimate the left eigenvectors

$$\sum_i u'_\gamma(i) P(i|j,\Delta t) = e^{-\mu_\gamma \Delta t} u'_\gamma(j), \tag{1}$$

where index γ numbers eigenvectors, $u'_\gamma(i)$ is γ-th eigenvector as a function of cluster node (i), μ_γ is the corrsponding γ-th eigenvalue. For equilibrium dynamics with the detailed balance, $n(i|j,\Delta t) = n(j|i,\Delta t)$, which we assume here, the smallest eigenvalue, $\mu_0 = 0$,

and the corresponding eigenvector is constant $u'_0(j) = 1$, all other eigenvalues are real and positive $\mu_{\gamma>0} > 0$.

To simplify the description of system's dynamics one can project its high-dimensional trajectory on a few EVs with lowest eigenvalues. These EVs describe slowest relaxation modes of the dynamics, and a free energy landscape as a function of these EVs can provide a simplified model of the relaxation dynamics. To project trajectory on EV u_γ one computes EV time-series as $u_\gamma(k\Delta t_0) = u'_\gamma(i(k\Delta t_0))$; where primed variable, $u'(i)$, denotes EV as a function of cluster index i, $u(k\Delta t_0)$ denotes EV as a function of trajectory (trajectory snapshot number or trajectory time), while $i(k\Delta t_0)$ denotes cluster index as a function of trajectory.

In practice, very long trajectories are rarely available, which makes this approach with accurate very fine-grained clustering non viable. Number of clusters grows exponentially with the dimensionality of the configuration space, which also limits the approach to low-dimensional configuration space. The proposed approach determines rather accurate approximations to the time-series of a few lowest eigenvectors, $u_\gamma(k\Delta t_0)$, without performing clustering at all.

Variational approaches are a promising alternative to the clustering approach.[5,13,14,18] A functional form (FF) with many parameters $R(\boldsymbol{X}, \alpha_i)$ (usually a weighted sum) is suggested as an approximation to EVs. One numerically optimizes the parameters by e.g., maximizing the auto-correlation function[13,14] or minimizing the total squared displacement.[18]

Namely, given a long equilibrium multidimensional trajectory $\boldsymbol{X}(k\Delta t_0)$, one computes the reaction coordinate time-series $r(k\Delta t_0) = R(\boldsymbol{X}(k\Delta t_0), \alpha_i)$. Here and below r denotes any reaction coordinate, while u is reserved for putative EVs. The functional form R approximates the first left EV, if it provides the minimum to the total squared displacement $\Delta r^2(\Delta t) = \sum_{k=1}^{N-\Delta t/\Delta t_0} [r(k\Delta t_0+\Delta t) - r(k\Delta t_0)]^2$, under the constraint $\sum_{k=1}^{N} r(k\Delta t_0)^2 = 1$. Note that, due to the constraint, the minimization of $\Delta r^2(\Delta t)$ is equivalent to the maximization of the auto-correlation function $C(r, \Delta t) = \sum_{k=1}^{N-\Delta t/\Delta t_0} r(k\Delta t_0+\Delta t) r(k\Delta t_0)$; we neglect here small differ-

ence between constrains $\sum_{k=1}^{N} r^2(k\Delta t_0) = 1$ and $\sum_{k=1}^{N-\Delta t/\Delta t_0} r^2(k\Delta t_0 + \Delta t) + r^2(k\Delta t_0) = 2$. The functional form R approximates the γ-th left EV if it provides the minimum to the $\Delta r^2(\Delta t)$ under constraint $\sum_k r(k\Delta t_0)^2 = 1$ and is orthogonal to the previous $\gamma - 1$ EVs $\sum_k r(k\Delta t_0) u_j(k\Delta t_0) = 0$, $j = 1, ..., \gamma - 1$.

It is straightforward to prove this principle. Consider Markov chain, describing the dynamics. Let indexes i and j denote the states of the chain and $r'(i)$ is an RC as a function of state i. Consider trajectory, i.e., a sequence of states $i(k\Delta t_0)$ of length N, which define the RC time-series as $r(k\Delta t_0) = r'(i(k\Delta t_0))$. The total squared displacement equals $\Delta r^2(\Delta t) = N \sum_{ij} [r'(i) - r'(j)]^2 P(i|j, \Delta t) P(j)$, while the constraint is $N \sum_j r'^2(j) P(j) = 1$, where $P(j)$ denotes equilibrium probability. Using 2λ as the Lagrange multiplier, differentiating with respect to $r'(j)$ and assuming the detailed balance one obtains Eq. 1 with $\lambda = e^{-\mu \Delta t}$.

Consider EV time-series approximation by a linear combination of basis functions $r(k\Delta t_0) = \sum_j \alpha_j f_j(k\Delta t_0)$. Using λ as the Lagrange multiplier the optimal values of parameters, $\alpha_j^\star$, that provide minimum to $\Delta r^2(\Delta t)$ under constraint $\sum_k r^2(k\Delta t_0) = 1$ can be found as a solution of the generalized eigenvalue problem

$$\sum_j A_{ij}(\Delta t) \alpha_j^\star = \lambda \sum_j B_{ij} \alpha_j^\star \tag{2a}$$

$$A_{ij}(\Delta t) = \sum_{k=1}^{N-\Delta t/\Delta t_0} \Delta f_i(k\Delta t_0) \Delta f_j(k\Delta t_0) \tag{2b}$$

$$B_{ij} = \sum_{k=1}^{N} f_i(k\Delta t_0) f_j(k\Delta t_0), \tag{2c}$$

where $\Delta f_i(t) = f_i(t + \Delta t) - f_i(t)$ denotes the forward time difference. The solutions of the eigenvalue problem are found numerically by standard linear algebra methods. Since both matrices are symmetric, the eigenvalues are real. Assume that eigenvalues are sorted as $\lambda_0 = 0 < \lambda_1 < \lambda_2 ....$ Then, the γ-th solution of Eq. 2a, denoted as $\alpha_j^{\star\gamma}$, corresponds to putative RC time-series $r(k\Delta t_0) = \sum_j \alpha_j^{\star\gamma} f_j(k\Delta t_0)$, which approximates γ-th EV time-series $u_\gamma(k\Delta t_0)$

2.2 Estimation of eigenvalues and implied timescales

The minimal value of the $\Delta r^2(\Delta t)$ functional, attained when r approximates EV u, equals $\Delta u^2(\Delta t) = 2(1 - e^{-\mu \Delta t})$, which, for small Δt, gives $\Delta u^2(\Delta t) \approx 2\mu \Delta t$; it is assumed here that EV is normalized as $\sum_k u^2(k\Delta t_0) = 1$. Correspondingly, the maximum value of the auto-correlation term equals $C(u, \Delta t) = e^{-\mu \Delta t}$. They can be used to estimate the eigenvalues μ, or the so called implied timescales $\hat{\tau} = 1/\mu$ as

$$\mu = -\ln[1 - \Delta u^2(\Delta t)/2]/\Delta t \tag{3a}$$

$$\mu = -\ln[C(u, \Delta t)]/\Delta t \tag{3b}$$

as functions of lag time Δt. Large lag times mask suboptimality of the putative EV and lead to a more accurate estimates of μ and $\hat{\tau}$. However at very large lag times it becomes difficult to accurately estimate an exponentially decreasing value of $C(u, \Delta t) = e^{-\mu \Delta t}$, since its statistical accuracy is limited by the number of transitions between regions where u is positive and negative, i.e., different free energy minima. An accurate and robust estimate should have statistical errors much smaller than the estimated value. A characteristic lag time $\Delta t^\star$, where the two are comparable could be roughly estimated as $(\mu T)^{-1/2} = e^{-\mu \Delta t^\star}$, where T is the total duration of the trajectory. The lag time chosen to accurately estimate the eigenvalues and the implied timescales, which we denote as Δt_∞, should be chosen much smaller than $\Delta t^\star$. An EV optimization is considered to be converged when eigenvalue estimated with lag time of interest Δt is close to the accurate eigenvalue, i.e., $\mu(\Delta t) \approx \mu(\Delta t_\infty)$.

In application to the HP35 protein, considered here, $\Delta t^\star \sim 10^4 \Delta t_0$ and we took Δt_∞ ten times smaller, $\Delta t_\infty = 1024 \Delta t_0 = 204.8$ ns as $\Delta t_0 = 0.2$ ns. The described approach determines EVs with eigenvalues (and implied timescales) accurate at the lag time of trajectory sampling interval of 0.2 ns, i.e., $\mu(\Delta t_0) \approx \mu(\Delta t_\infty)$.

2.3 Non-parametric optimization of eigenvectors

A major weakness of parametric approaches that approximate RCs by using a functional form (FF) with many parameters, e.g., a linear combination of collective variables or a neural network, is that it is difficult to suggest a good FF approximating EVs. The difficulty becomes apparent if one remembers that such a FF should be able to accurately project a few million snapshots of a very high-dimensional trajectory. In particular, it implies an extensive knowledge of the system, and that such a FF is likely to be system specific.

Recently we have suggested a non-parametric approach for the determination of the committor function, which bypasses the difficult problem of finding an appropriate FF.[10,18] The power of the approach was demonstrated by applying it to the equilibrium folding trajectory of the HP35 double mutant. The determined RC closely approximates the committor as was validated by the optimality criterion - $Z_{C,1}$ (defined below) is constant up to the expected statistical noise.[20] The approach performs optimization of the RC in a uniform manner by focusing optimization on the time scales and the regions of the putative RC which are most suboptimal.

The general idea of iterative non-parametric RC optimization is as follows.[10,18] We start with a seed RC time-series $r(k\Delta t_0)$. During each iteration we consider a variation of RC as $r(k\Delta t_0) + \delta r(k\Delta t_0)$, where $\delta r(k\Delta t_0)$ can be a time-series of any function of configuration space, collective variables and, hence, the RC itself. For example, one can take $\delta r(k\Delta t_0) = f(r(k\Delta t_0), y(k\Delta t_0))$, where $y(k\Delta t_0)$ is time-series of a randomly chosen collective variable or a coordinate of the configuration space and $f(r, y) = \sum_{ij} \alpha_{ij} r^i y^j$ is a low degree polynomial. The coefficients/parameters of the variation are chosen such that $r(k\Delta t_0) + \delta r(k\Delta t_0)$ provides the best approximation to the target optimal RC (e.g., the committor or EVs). Specifically, they deliver optimum to the corresponding target functional. For the optimization of EVs, considered here, they can be found as solutions of Eq. 2. The RC time-series is updated $r(k\Delta t_0) \leftarrow r(k\Delta t_0) + \delta r(k\Delta t_0)$ and the process is repeated. Iterating the process one repeatedly improves the putative RC time-series by incorporat-

ing information contained in different coordinates or collective variables. The process stops when, e.g., the target functional is close to its optimal value, meaning that the putative RC is a close approximation of the target RC.

Importantly, while the result of each iteration may depend on the exact choice of the family of collective variables or the functional form of the variation, the final RC does not, since it provides the optimum to a (non-parametric) target functional when the optimization converges, which makes this approach non-parametric. It is assumed that the family of collective variables contains all the important information about the dynamics of interest. If the system obeys some symmetry (e.g., the rotational and translational symmetries for biomolecules), then the RCs should obey the same symmetry. A simple way to ensure this is to use collective variables that respect the symmetry. For example, the distances between randomly chosen pairs of atoms or sin and cos of dihedral angles can be suggested as standard sets of collective variables.

Here we extend the approach to non-parametric determination of eigenvectors. Specifically, given a multidimensional trajectory $\boldsymbol{X}(k\Delta t_0)$ and the number of the slowest eigenvectors required n_{ev}, the approach determines time-series of the required eigenvectors $u_\gamma(k\Delta t_0)$ and corresponding eigenvalues μ_γ, where $k = [1, N]$ and N is the trajectory length and $1 \leq \gamma \leq n_{\text{ev}}$.

We start with seed EVs time-series, $u_\gamma(k\Delta t_0)$, $1 \leq \gamma \leq n_{\text{ev}}$, for example the distance time-series between randomly chosen pairs of atoms. Then, the EVs time-series are improved iteratively. To simultaneously update all EVs during each iteration, we consider a variation of EVs time-series as

$$r(k\Delta t_0) = \sum_\gamma \alpha_\gamma u_\gamma(k\Delta t_0) + f(u_\beta(k\Delta t_0), y(k\Delta t_0)), \tag{4}$$

here, $y(k\Delta t_0)$ is the time-series of a randomly chosen collective variable of the original multidimensional space $\boldsymbol{X}$, β denotes index of an active EV, whose contribution to the

variation is higher then linear, and $f(u,y) = \sum a_{ij} u^i y^j$ is a low degree polynomial.

All the time-series in the variation (Eq. 4) are denoted as basis functions $f_j(k\Delta t_0)$; the variation can be written as $r(k\Delta t_0) = \sum_j \alpha_j f_j(k\Delta t_0)$, where vector α_j now contains both parameters α_γ and coefficients of the polynomial a_{ij}. The optimal values of the parameters, $\alpha_j^\star$, are chosen such that the variation provides the best approximation to an EV time-series. They are determined by numerically solving Eq. 2. The first $n_{\rm ev}$ solutions, denoted as $\alpha_j^{\star\gamma}$, are used to update the putative time-series of γ-th EV as $u_\gamma(k\Delta t_0) \leftarrow \sum_j \alpha_j^{\star\gamma} f_j(k\Delta t_0)$, and the iterative process is repeated.

The generalized eigenvalue problem, Eq. 2, does not have a solution if the basis functions contain the same time-series twice. Here, the time-series of the active EV, $u_\beta(k\Delta t_0)$, is included in both the first sum and the polynomial. The same is true for EV u_0, corresponding to $\mu_0 = 0$, whose time-series is a constant. To have these time-series only once we assume that the constant and u terms are removed from the polynomial.

Inclusion of the linear combination of all EVs into the RC variation (Eq. 4) means that this variation can be considered as a variation $u_\gamma + \delta u$ of every EV in turn. It ensures, in particular, that every updated EV has the corresponding EV at the previous iteration as a baseline. Active EVs can be selected randomly, or one may select the least optimal EV, i.e., the one having the largest ratio $\mu(\Delta t)/\mu(\Delta t_\infty)$. The iterative optimization is considered to be converged, when eigenvalues of all eigenvectors of interest estimated with the lag time of interest Δt are close to the accurate values, i.e., $\mu_\gamma(\Delta t) \approx \mu_\gamma(\Delta t_\infty)$.

Thus, a minimal algorithm of non-parametric EV optimization is as follows. **Initialization:** Set seed EVs time-series. $u_0(k\Delta t_0) = 1$. For $1 \le \gamma \le n_{\rm ev}$ select randomly a collective variable y and set $u_\gamma(k\Delta t_0) = y(k\Delta t_0)$. Select the lag time of interest Δt and the lag time Δt_∞ to test convergence. For example, $\Delta t = \Delta t_0$ and $\Delta t_\infty = 1024\Delta t_0$. **Iterations:** Select active EV, u_β, as the most suboptimal one, i.e., the one with the largest ratio $\beta = \arg\max_\gamma \mu_\gamma(\Delta t)/\mu_\gamma(\Delta t_\infty)$, or just randomly. Select collective variable time-series $y(k\Delta t_0)$. Compute basis functions of Eq. 4, solve Eq. 2 and updates the EVs time-

series $u_\gamma(k\Delta t_0)$. **Stopping:** Stop if the optimization has converged: $\mu_\gamma(\Delta t) < \mu_\gamma(\Delta t_\infty)$ for $1 \leq \gamma \leq n_{\rm ev}$.

To explicitly illustrate the iterative character of the optimization, the algorithm can be written as $u_1^{n+1}, ..., u_{n_{ev}}^{n+1} = F(u_1^n, ..., u_{n_{ev}}^n, \beta^n, y^n)$, where superscript n denotes values of variables at n-th iteration, and $F(u_1^n, ..., u_{n_{ev}}^n, \beta, y)$ denotes a function/procedure that takes a set of n_{ev} EVs time-series u_γ, the index of active eigenvector β and time-series of collective variable y, computes basis functions of Eq. 4, solves Eq. 2 and returns a set of updated n_{ev} time-series v_γ, that better approximate the EVs.

Selecting collective variable time-series $y(k\Delta t_0)$ means random selection from the provided set of collective variables. For example, if one takes a standard set of collective variables - the inter-atom distances, then every time a collective variable is requested, one selects a random pair of atoms i and j, and returns the distance time-series between the atoms $r_{ij}(k\Delta t_0)$ computed from the trajectory.

Selection of Δt_∞. Lag time Δt_∞ is used to test the convergence of EV optimization as $\mu_\gamma(\Delta t) \approx \mu_\gamma(\Delta t_\infty)$. From one side, it should be chosen as long as possible, to mask the deficiencies of putative EV time-series and have a more accurate estimate of eigenvalue μ_γ. From the other side, very long Δt_∞ lead to large statistical uncertainties in the estimation of $\mu_\gamma(\Delta t_\infty)$, as discussed in Sect. 2.2. One strategy of selection of Δt_∞ is to, first, perform optimization with Δt_∞ conservatively selected just a few times longer than the lag time of interest Δt, determine the statistical uncertainties as a function of lag time using bootstrapping and use that for an informed selection of Δt_∞.

Selection of the polynomial. Generally, the higher is the degree of the polynomial $f(u, y)$, the faster is the optimization, though more computationally demanding. However a very high degree may lead to numerical instabilities and strong overfitting. The following strategy was found useful: use a polynomial $f(u, y)$ with a relatively small degree (3-6) for updates involving u_β and y followed by a polynomial $f(u)$ of a high degree (e.g., 10-16) for updates involving only u_β, where u_β is the active EV.

2.4 Eigenvector validation/optimality criterion

An accurate eigenvalue or the corresponding implied timescales can serve as an indicator that the putative RC time-series closely approximates an EV. However, these metrics provide a rather crude, cumulative estimate of the accuracy of putative EVs. It is possible that while an eigenvalue is accurate, some parts of EV are overfitted/overoptimized, while other underfitted. To check for that, we describe a more stringent EV optimality/validation criterion $\Theta(x, \Delta t)$.

The criterion is an extension of the $Z_{C,1}$ criterion for the committor reaction coordinate. $Z_{C,1}$ can be straightforwardly computed from time-series $r(k\Delta t_0)$: each transition of trajectory from $x_1 = r(i\Delta t)$ to $x_2 = r(i\Delta t + \Delta t)$ adds $1/2|x_1 - x_2|$ to $Z_{C,1}(x, \Delta t)$ for all points x between x_1 and x_2.[10,20] Jupyter notebooks illustrating usage of $Z_{C,\alpha}$ profiles and the committor and eigenvector criteria are available at https://github.com/krivovsv/CFEPs.[21]

$Z_{C,1}$ has a number of useful properties.[10,20] If reaction coordinate q closely approximates the committor function, then $Z_{C,1}(q, \Delta t) \approx N_{AB}$, where N_{AB} is the number of transitions between boundary states, i.e., from A to B, or from B to A. For a suboptimal reaction coordinate r, $Z_{C,1}(r, \Delta t)$ values generally decrease to the limiting value of N_{AB}, as Δt increases. The larger the difference between $Z_{C,1}(r, \Delta t_1)$ and $Z_{C,1}(r, \Delta t_2)$ the less optimal the reaction coordinate around r. This property is used to find suboptimal spatio-temporal regions and focus optimization on them to make it more uniform.

The constancy of $Z_{C,1}(q, \Delta t)$ along the committor q follows from the following. Consider Markov chain, describing the dynamics. Let indexes i and j denote the states of the chain and $x(i)$ their position on an RC. Value of $Z_{C,1}(x, \Delta t)$ can change, in a step-wise fashion, only when position x goes through a particular state j, i.e., x goes from $x(j) - 0$ to $x(j) + 0$ and equals[20]

$$\Delta Z_{C,1}(x(j), \Delta t) = \sum_i [x(i) - x(j)] n(i|j, \Delta t). \tag{5}$$

It is zero for the committor function (if j is not a boundary state) since committor is defined

by the following equation

$$\sum_i [q(i) - q(j)] P(i|j, \Delta t) = 0 \quad \text{for } j \neq A, B \tag{6a}$$

$$q(A) = 0, \quad q(B) = 1 \tag{6b}$$

and $n(i|j, \Delta t) = P(i|j, \Delta t)P(j)$. Eq. 1 is different from 6a which means that $Z_{C,1}$ along an eigenvector is not constant. However Eq. 1 can be rewritten as

$$\sum_i [u'(i) - u'(j)] n(i|j, \Delta t) = (1 - e^{-\mu \Delta t})[0 - u'(j)] n(j), \tag{7}$$

and interpreted in the following way. On the left hand side we have change in $Z_{C,1}$ around $u'(j)$ computed in the standard way. It is proportional to the change of $Z_{C,1}$ computed for a virtual trajectory consisting of collection of transitions 0 to $u'(j)$ and back to 0 made $n(j)$ times for every j. We denote the second profile as $Z^0_{C,1}$. Since both profiles are 0 at large negative x and have proportional changes, they are proportional themselves $Z_{C,1}(x, \Delta t) = (1 - e^{-\mu \Delta t}) Z^0_{C,1}(x, \Delta t)$. Note that $Z^0_{C,1}(x, \Delta t = m\Delta t_0) = Z^0_{C,1}(x, \Delta t_0)/m$. Consider the following variable

$$\Theta(x, \Delta t) = -\ln \frac{Z_{C,1}(x, \Delta t)}{(1 - e^{-\mu \Delta t}) Z^0_{C,1}(x, \Delta t)}. \tag{8}$$

Validation: If putative time-series $u(i\Delta t_0)$ and μ closely approximates an EV and the corresponding eigenvalue, then $\Theta(x, \Delta t) \approx 0$ for all Δt and all x along u. An accurate estimate of μ is obtained from the EV time-series using Eq. 3 at large lag times.

$Z_{C,1}(x, \Delta t)$ can be interpreted as a local density of the total squared displacement $\Delta r^2(\Delta t)/2$, since $\int Z_{C,1}(x, \Delta t) dx = \Delta r^2(\Delta t)/2$.[20] Analogously, $Z^0_{C,1}(x)$ can be considered as a local density of $\sum_k r^2(k\Delta t_0)$. The constraint optimization problem is equivalent to finding minimum of an integral of $Z_{C,1}(x, \Delta t)$ under constraint that an integral of $Z^0_{C,1}(x)$ is 1. When a putative coordinate closely approximates an eigenvector, the local densities are proportional.

Optimality: For a suboptimal coordinate $\Theta(x, \Delta t) < 0$, because $Z_{C,1}(x, \Delta t)$ is larger

than that for the optimal coordinate. The bigger the difference between $\Theta(x, \Delta t_1)$ and $\Theta(x, \Delta t_2)$ for $t_1 > t_2$ the less optimal is u around x. $\Theta(x, \Delta t) \to 0$ as Δt increases.

2.5 Inherent instability of iterative EV optimization

We illustrate the minimal algorithm of non-parametric optimization by determining the first eigenvector u_1 for a long equilibrium trajectory of double mutant of HP35 protein consisting of 1509392 snapshots at 380 K and the sampling interval of $\Delta t_0 = 0.2$ ns.[19] We used $\Delta t = \Delta t_0$, $\Delta t_\infty = 1024 \Delta t_0$, and polynomials $f(u_1, y)$ of degree 4 and $f(u_1)$ of degree 12. Inter-atom distance were used as a set of collective variables.

Fig. 1a shows $\mu(\Delta t_0)$ as a function of iteration number for ten representative optimization runs started with different random seed numbers. For most of the runs $\mu(\Delta t_0)$ steadily converges to the same eigenvalue of $\mu(\Delta t_0) \sim 2.68 \cdot 10^{-4}$ in units of Δt_0^{-1}. It indicates robustness and reproducibility of the non-parametric optimization. The putative time-series, after a few thousands iterations, can provide a rather good approximation to an EV with the corresponding eigenvalue within a small factor from the exact value.

However, two of the runs, showed by red and blue colors, converged to different EVs with different eigenvalues. Optimization run, showed by red color on Fig. 1a is rather short. It started as other runs but quickly converged to a spurious EV. The EV has a peculiar free energy profile (FEP), $F(u_1)$, shown on Fig. 1b; the FEP is estimated from a histogram. The EV has a rather large amplitude $A(u_1) = \max(u_1) - \min(u_1) \approx 175$ and describes a transition to a low populated, shallow minimum. Inspection of the EV time-series shows that it has made only one transition from the main minimum around $u_1 \sim 0$ to the shallow minimum around $u_1 \sim 175$ and back. Optimization run, showed by blue color, initially followed the gray lines, however around 720-th iteration in deviated abruptly, which can be seen by the abrupt change of the eigenvalue on Fig. 1a. FEP of the putative EV just before the iteration is shown by black line on Fig. 1c, and is very close to the FEPs of EVs the runs, colored gray, converged to. The blue line on Fig. 1c shows the FEP of the putative EV just after the

abrupt transition, which has a much higher barrier and more structure. Correspondingly, the EV has a much smaller eigenvalue of $\mu(\Delta t_0) \sim 8.58 \cdot 10^{-6}$ in units of Δt_0^{-1}. However, the FEP does not describe the folding dynamics. The two minima of the FEP, $u_1 < -1$ and $u_1 > -1$, have identical FEPs when projected on the root-mean-square-deviation from the native structure RC. Closer inspection shows that the main barrier describes a rotation of a dihedral angle corresponding to a transition between two permutational isomers of GLN 67 residue. Collective variable $y = r_{ij}$ that contributed to this abrupt deviation is the distance between atoms 209 and 491, which correspond to OE1 and HA4 in GLN 67. The permutational isomers correspond to exchange of hydrogen atoms HA4 and HE41. Thus, while this EV has a smaller eigenvalue, it has no connection to folding and is of very limited interest.

To summarize, the two deviated runs illustrate that the problem of determining the slowest EVs has the following inherent "instability".[18] The algorithm seeks EVs with the smallest eigenvalues, which describe the slowest dynamics. However, some of such EVs, which we denote as spurious EVs, are not of interest. For example, in protein folding, such an EV could describe a much slower torsion angle isomerization process.[12,18] The spurious EV shown by blue color on Fig. 1, describes a permutational isomerization, that happened 7 times in the course of the entire trajectory that contains about 140 folding-unfolding events. Another, more frequent possibility, is due to limited sampling. There are many parts of the configuration space that were visited very few times or even just once (the shallow basin on Fig. 1b), and EVs describing those transitions have small eigenvalues. Thus, starting with an EV of interest, the algorithm may eventually converge to a spurious EV, with smaller eigenvalue, but of no interest. In general terms, this peculiarity of EV optimization is due to its unsupervised nature: we seek any EV with smallest eigenvalue. Optimization of the committor function, which is a variant of supervised learning, as the function interpolates between two given boundary states of interest, is free of such a problem.[10,18]

For systems, where the likelihood of switching to spurious EVs is not large, the simplest

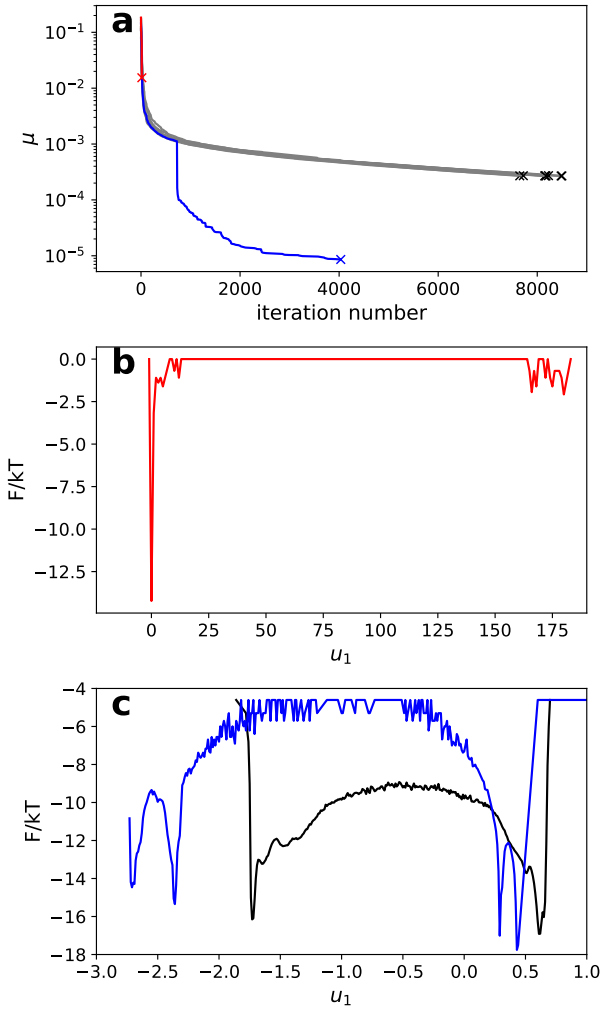


Figure 1: Application of the minimal algorithm of non-parametric EV optimization to the determination of u_1 of HP35. Ten representative optimization runs started with different seed numbers. **a**) Estimate of the eigenvalue $\mu(\Delta t_0)$ in units of Δt_0^{-1} as a function of iteration number; cross at the end of line indicates where optimization has converged, i.e., $\mu_\gamma(\Delta t_0) \approx \mu_\gamma(\Delta t_\infty)$. Most of the lines (gray) fall on the same curve, indicating the robustness and reproducibility of the non-parametric approach. For them the eigenvalues steadily decrease toward the target value. However, two optimization runs converged to different, spurious eigenvectors (see text). Run, colored run, converged to a EV, which FEP, $F(u_1)$, is shown on **b**). Optimization run, colored blue, initially followed the common trajectory and deviated abruptly around 720-th iteration and converged to an EV with lower eigenvalue. Free energy profiles of putative EVs just before the iteration and straight after are shown on **c**) by black and blue colors, respectively.

is to just discard the optimization runs that have converged to spurious EVs, and keep those where EVs of interest are found. Such systems can be analyzed with the minimal algorithm described above. For other systems, where the likelihood of switching to spurious EVs is large, one needs a more systematic approach of suppressing the instability. We describe a few heuristics to suppress the instability, which were sufficient to determine the lowest EVs of realistic protein folding trajectories.

As illustrated on Fig. 1, a shift to a spurious EV happens usually in an abrupt manner and results in significant changes in the EV time-series. Thus, allowing only gradual changes of the putative EV time-series. should help suppress the instability. A main idea is to keep a fraction of trajectory points, selected with probability p_{fix} (0.5 here), fixed during each iteration. It penalizes large changes in the EV time-series, since during optimization the distance between consecutive points is minimized. Allowing, an overall shift and change of scale, it means that fixed points are transformed according to Eq. 4, with contributions from the polynomial set to zero, i.e., all eigenvectors contribute linearly. Increasing p_{fix} enforces a more gradual change of eigenvectors during optimization.

The eigenvalue of an EV, estimated at large lag time Δt_∞, changes rather little after an initial settling phase. Hence, a relatively large change (5 % here), is an indication that an EV has changed significantly. Iterations with such changes are not accepted.

Collective variables that promote transitions to spurious EVs, (e.g., like that on Fig. 1c) can be filtered out. A simple collective variable y that depends on a few coordinates only, e.g., the inter-atom distance, is first transformed to the first EV as its function $y \to u_1(y)$. If the corresponding eigenvalue is very small it means that y does not describe a collective process, such as protein folding, and is likely to describe a spurious EV. Such a variable is discarded.

In the infrequent cases, when, in spite of the heuristics employed, the algorithm switches to a spurious EV, the optimization is restarted. Such events are detected by the following heuristics: one monitors the amplitude of an EV $A(u)$. When the amplitude reaches a

relatively large value, it indicates of a spurious EV analogous to that on Fig. 1b; e.g., compare the amplitudes of EVs on Fig. 1b and Fig. 1c.

Note that, usually, the likelihood of switching to a spurious EV from the very start is rather small. Thus with a large likelihood a randomly selected collective variable will naturally lead to the slowest EVs describing a collective process, like protein folding, i.e., the EVs of interest. There is no need to specifically select an EV of interest which keeps the analysis unbiased and blind.

The optimization algorithm with heuristics to suppressed instability is as follows. **Initialization:** Set seed EVs time-series. $u_0(k\Delta t_0) = 1$. For $1 \leq \gamma \leq n_{\text{ev}}$ select randomly a collective variable y and set $u_\gamma(k\Delta t_0) = y(k\Delta t_0)$. Set the starting lag time $\Delta t > \Delta t_0$ and the lag time Δt_∞ to test convergence. For example, $\Delta t = 256\Delta t_0$ and $\Delta t_\infty = 1024\Delta t_0$. Set the p_{fix} probability, e.g., $p_{\text{fix}} = 0.5$. **Iterations:** Select the set of fixed points with probability p_{fix}. Select active EV, u_β, as the most suboptimal one, i.e., the one with the largest ratio $\beta = \arg\max_\gamma \mu_\gamma(\Delta t)/\mu_\gamma(\Delta t_\infty)$, or just randomly. Select randomly collective variable y. Compute basis functions of Eq. 4. Set polynomial basis functions to 0 for fixed points/frames. Solve Eq. 2 and compute updates for the EVs. If the update passes the safety checks for the suppression of the instability, update the EVs. If optimization has diverged: an EV amplitude $A(u)$ has crossed the threshold (30 here), restart the optimization. **Stopping:** Optimization with current lag time stops when the eigenvalue estimate is close to the accurate value $\mu_\gamma(\Delta t) < \mu_\gamma(\Delta t_\infty)$. If $\Delta t > \Delta t_0$, then Δt is halved and optimization with smaller Δt is continued. If $\Delta t = \Delta t_0$ optimization stops.

Selection of collective variables. To filter out collective variables that promote transitions to spurious EVs one proceeds as follows. Select a random pair of atoms i and j, and compute the distance time-series between the atoms $y(k\Delta t_0) = r_{ij}(k\Delta t_0)$ from the trajectory. Compute $u_1(y)$ and the corresponding eigenvalue, using the Eq. 2 with basis functions $f_j(k\Delta t_0) = y^j(k\Delta t_0)$ for $0 \leq j \leq 12$. If eigenvalue is smaller than a threshold (e.g., $\mu(\Delta t) < 10^{-4}$ here), reject y and repeat the process with another pair of atoms.

Selection of p_{fix}. The larger is p_{fix} the more robust, but slower is optimization. One is advised to start with $p_{\text{fix}} = 0.5$ and adjust according to the performance of the algorithms.

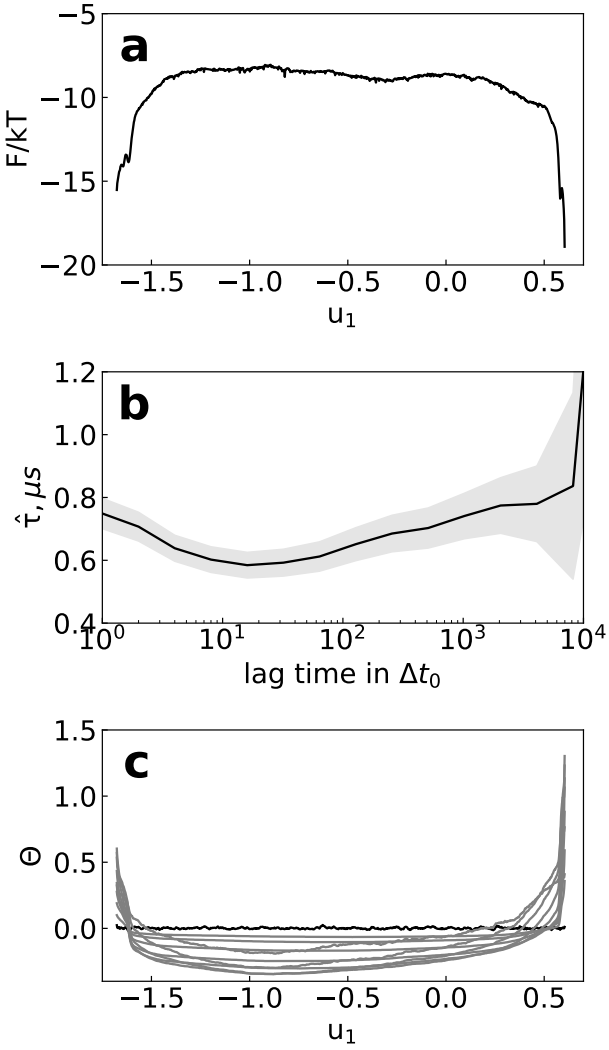


Figure 2: Non-parametric optimization of first eigenvector u_1 of HP35 with heuristics to suppress instability. **a**) Free energy as a function of u_1. **b**) Implied timescale $\hat{\tau}$ as a function of lag time Δt. Uncertainties (shaded areas) were computed with bootstrap. Uncertainties rapidly increase as the lag time approaches $\Delta t^{\star}$. **c**) Optimality criterion for an eigenvector $\Theta(u_1, \Delta t)$ for different lag times $\Delta t/\Delta t_0 = [1, 2, 4, ..., 2^{10}]$. $\Theta(u_1, \Delta t_0)$ is shown by solid black line.

Fig. 2 shows the application of the algorithm with the suppressed instability to the determination of u_1 of HP35. Fig. 2a shows FEP as a function of the first EV $F(u_1)$. It has a simple shape of one free energy (FE) barrier and two minima.

Fig. 2b shows that an accurate estimate of implied timescales is possible with lag time

of the trajectory sampling interval of $\Delta t_0 = 0.2$ ns. The figure confirms the choice of $\Delta t_\infty = 1024\Delta t_0$. It is sufficiently long, so that the estimates of the eigenvalue or implied timescale at this lag time agrees with those at longer lag times. At the same time it is sufficiently short so that the estimates have small uncertainty.

The uncertainties of the estimate of implied timescales rapidly increase as Δt approaches $\Delta t^\star$. As explained in Sect. 2.2 it is difficult to accurately estimate the exponentially decreasing auto-correlation function $C(r, \Delta t) \approx e^{-\mu \Delta t}$. $C(r, \Delta t)$ decreases as lag time increases because a larger fraction of points transit to the other basin with the opposite sign of EV. The statistical error of the estimate of $C(r, \Delta t)$ is determined by the total number of transitions between the basins. And when, with increasing lag time, the estimate of $C(r, \Delta t)$ becomes close to its statistical error, the uncertainty of μ_1 estimate rapidly increases.

Fig. 2c shows the EV optimality/validation criterion $\Theta(x, \Delta t)$. In particular, it shows that $\Theta(x, \Delta t_0) > \Theta(x, \Delta t)$ around the barrier and $\Theta(x, \Delta t_0) < \Theta(x, \Delta t)$ around minima for large $\Delta t > \Delta t_0$. It means that the putative u_1 time-series does not approximate the EV uniformly. It overfits the EV around the barrier region and underfits around the minima.

To conclude, while the accurate eigenvalue suggest that the putative time-series closely approximates u_1, the more stringent EV optimality/validation criterion shows that the time-series overfits u_1 in some parts and underfits in other.

2.6 Adaptive optimization

Our aim is to determine such an EV time-series $u(i\Delta t_0)$ that it passes the validation test, i.e., $\Theta(u, \Delta t) \approx 0$ up to statistical uncertainty. A way to do this is to perform optimization more uniformly, so that all regions of the putative EVs become underfitted to the same degree and stop optimization just before overfitting. Such an adaptive optimization is performed by focusing on less optimized parts of putative EVs. Before every iteration one scans $\Theta(x, \Delta t)$ profiles to find most suboptimal/underfitted regions. Position dependent $p_{\text{fix}}(x)$ is introduced in such a way as to be smaller for more underfitted regions. Smaller $p_{\text{fix}}(x)$ means less

constraints and thus faster optimization. The obtained results are robust with respect to specific form of $p_{\text{fix}}(x)$ employed. More details are given in the Appendix.

The generic adaptive non-parametric EV optimization algorithm is as follows. **Initialization:** Set seed EVs time-series. $u_0(k\Delta t_0) = 1$. For $1 \leq \gamma \leq n_{\text{ev}}$ set $u_\gamma(k\Delta t_0) = y(k\Delta t_0)$, where y is a randomly selected collective variable, e.g., from the standard set. Set the initial lag time to a large value, e.g., $\Delta t = 256\Delta t_0$. Set Δt_∞, for example, $\Delta t_\infty = 1024\Delta t_0$ **Iterations:**

1. Select active EV, u_β, as the most suboptimal one, i.e., the one with the largest ratio $\beta = \arg\max_\gamma \mu_\gamma(\Delta t)/\mu_\gamma(\Delta t_\infty)$, or just randomly.

2. Scan $\Theta(x, \Delta t)$ profiles for the active EV to find most suboptimal/underfitted regions and compute the position dependent $p_{\text{fix}}(x)$. Determine fixed points/frames: for frame k take the position of the frame along the active EV, $x = u_\beta(k\Delta t_0)$, and choose the frame to be fixed with probability $p_{\text{fix}}(x)$.

3. Select randomly collective variable y. Compute basis functions of Eq. 4. Set polynomial basis functions to 0 for fixed points/frames. Solve Eq. 2 and compute updates for the EVs.

4. Perform safety checks. If optimization has diverged, an eigenvector amplitude $A(u) = \max(u) - \min(u)$ has crossed the threshold (30 here), restart the optimization by going to **Initialization**. If safety checks are passed, update the EVs.

Stopping:

1. If $\Delta t > \Delta t_0$ and optimization has converged for current lag time: $\mu_\gamma(\Delta t) < 1.2\mu_\gamma(\Delta t_\infty)$ for $1 \leq \gamma \leq n_{\text{ev}}$, continue optimization with halved lag time $\Delta t \leftarrow \Delta t/2$.

2. Stop if $\Delta t = \Delta t_0$ and optimization has converged: $\mu_\gamma(\Delta t_0) < \mu_\gamma(\Delta t_\infty)$ for $1 \leq \gamma \leq n_{\text{ev}}$.

It is advantageous to stop optimization at larger lag times $\Delta t > \Delta t_0$ a bit earlier, i.e., when $\mu_\gamma(\Delta t) < 1.2\mu_\gamma(\Delta t_\infty)$. It, first, speeds up the overall optimization and, second, optimization with smaller lag times continues to improve $\mu_\gamma(\Delta t)$.

Fig. 3 shows application of the adaptive approach to determine the first two EVs for the HP35 trajectory. Fig. 3a shows that $\Theta(x, \Delta t)$ is much closer to zero (bounded by ± 0.2) compared to Fig. 2c, indicating that u_1 is now better approximates the EV. The FEP $F(u_1)$ also shows more structure in the minima. This additional structure disappeared on Fig. 2a because the regions were not sufficiently optimized. The second EV similarly has $\Theta(x, \Delta t)$ close to zero (Fig. 3b). The implied timescales are accurate starting from the shortest lag time of 0.2 ns (Fig. 3c).

Note, that it is difficult to compare free energy barriers along different EVs u_1 and u_2 directly. First, the correspondence between the barriers can be elucidated only by considering the free energy surface as a function of both EVs (see below); for example barrier around $u_1 \sim -1$ corresponds to that around $u_2 \sim 2$. Second, different EVs provide different, highly nonlinear projections of the configuration space; regions separated on one EV can overlap on another.

How accurately do the FEPs on Fig. 3 describe the kinetics? For example, the FEP along the committor can be used to compute *exactly* such important properties of kinetics as the equilibrium flux, the mean first passage times, and the mean transition path times between any two regions on the committor.[10] Exactly here means that these quantities computed from the one-dimensional diffusion model are equal to that computed directly from the multidimensional trajectory. It, thus, can be used to obtain *direct* accurate estimates of, e.g., free energy barriers and pre-exponential factors.[10] The accuracy is limited only by the accuracy of the determined committor. An EV, while being different, could be quite close to the committor between the boundary minima, especially around the transition state (TS) region.[22] It can be used to compute the properties approximately. The relative error could be roughly estimated by applying the committor optimality/validation criterion[20] $Z_{C,1}$ to

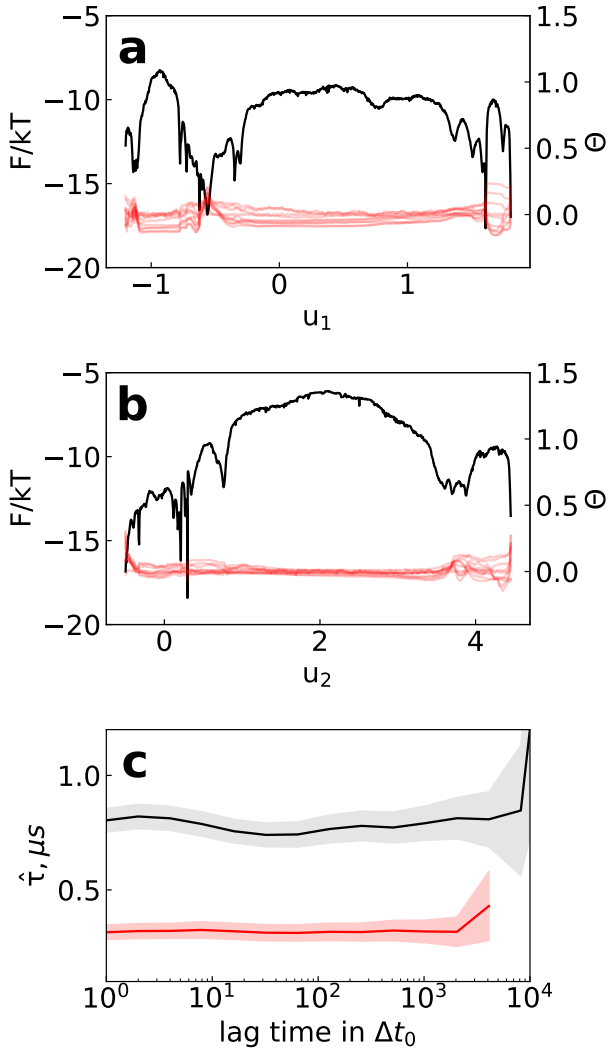


Figure 3: Adaptive non-parametric optimization of first two eigenvectors u_1 and u_2 of HP35. **a**) Free energy (black) and optimality criterion (red) as functions of u_1. **b**) Those as as functions of u_2; **c**) Implied timescale $\hat{\tau}$ for u_1 (black) and u_2 (red) as functions of lag time Δt; uncertainties (shaded areas) were computed with bootstrap.

the EV time-series (Fig. 4) and for the first EV the error is around 30%. For example, taking boundaries along u_1 at $A = -0.565$ and $B = 1.8$ (at local minima on Fig 3a) one obtains the following estimates with the diffusive model:[10] $N_{AB} = 58$, $\text{mfpt}_{AB} = 3974$ ns, $\text{mfpt}_{BA} = 1194$ ns, $\text{mtpt}_{AB} = 227$ ns, and directly from trajectory: $N_{AB} = 49$, $\text{mfpt}_{AB} = 4787$ ns, $\text{mfpt}_{BA} = 1330$ ns, $\text{mtpt}_{AB} = 323$ ns; here N_{AB} is the number of transition from A to B, or B to A, mfpt_{AB} is the mean first passage time from A to B, mtpt_{AB} is the mean transition path time between A and B. For boundaries at $A = -0.565$ and $B = 1.607$ estimates from diffusive model are $N_{AB} = 75$, $\text{mfpt}_{AB} = 2915$ ns, $\text{mfpt}_{AB} = 1111$ ns, $\text{mtpt}_{AB} = 61$ ns and directly from trajectory $N_{AB} = 77$, $\text{mfpt}_{AB} = 2790$ ns, $\text{mfpt}_{AB} = 1102$ ns, $\text{mtpt}_{AB} = 93$ ns.

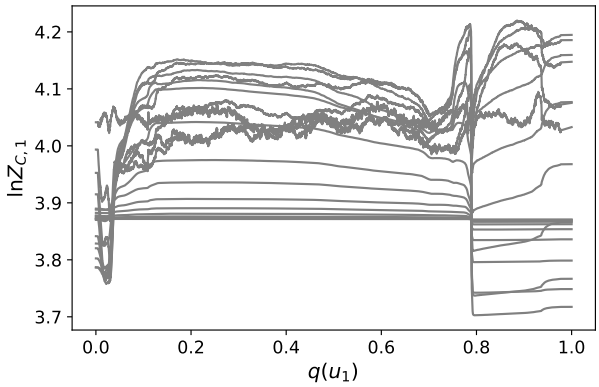


Figure 4: Committor optimality/validation criterion[20] applied to u_1. u_1 is first transformed to $q(u_1)$, committor as a function of u_1. $Z_{C,1}(x, \Delta t)$ along $q(u_1)$ are computed for $\Delta t = 1, 2, ...2^{20}$. Deviations of $\ln Z_{C,1}(x, \Delta t)$ from a constant a bounded by ± 0.3, which translates to relative error around 30% in estimation of kinetic properties.

3 Protein folding landscapes and dynamics.

Using $F(u_i)$ (Fig. 3) for the analysis and description of the dynamics is not very convenient as the diffusion coefficient varies significantly along the EVs.[10] It is more convenient to use a "natural" coordinate, which we denote as $\tilde{u}_i$, where the diffusion coefficient is constant $D(\tilde{u}_i) = 1$. It is related to u_i by the following monotonous transformation $d\tilde{u}_i/du_i = D(u_i)^{-1/2}$.[23]

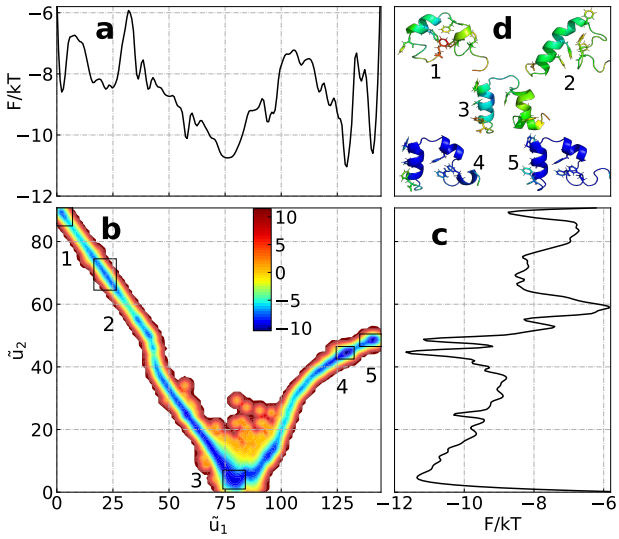


Figure 5: Free energy landscapes of HP35 double mutant: **a**) $F(\tilde{u}_1)$, **c**) $F(\tilde{u}_2)$, and **b**) $F(\tilde{u}_1, \tilde{u}_2)$; the color bar shows F/kT. **d**) shows representative structures for the rectangular regions around the free energy minima on **b**); colors code the root-mean-square fluctuations of atomic positions around the average structure from 0.5 Å (blue) to 13 Å (red).

The FEP along the first EV $F(\tilde{u}_1)$ (Fig. 5a) is in agreement with $F(\tilde{q})$, the FEP along the optimal folding coordinate - the committor between the denatured and the native states; here $\tilde{q}$ denotes the committor monotonously transformed to a natural coordinate.[10] $F(\tilde{u}_1)$ and $F(\tilde{q})$, in particular, both have minima 3, 4 and 5 and the folding barrier of ~ 3.5 kT, confirming that the approach works. There are however also important differences: $F(\tilde{q})$ does not show minima 1 and 2 and minima 4 and 5 are in the opposite order. This is due to the employed definition of the boundary denatured and native states for the committor, defined as structures that have the C_α root-mean-squared-deviation (rmsd) from the native structure smaller than 0.5 Å and greater than 10.5 Å, respectively. Using the native minimum (4) with the smallest rmsd as the boundary state forces it to be the rightmost minimum on $F(\tilde{q})$, while $F(\tilde{u}_1)$ reveals that kinetically 5 is the rightmost minimum. Minima 1, 2 and 3 all have very similar projections on the rmsd, and the boundary state with large rmsd is equally connected to all of them, preventing their separation along $\tilde{q}$. This illustrates that proper definition of boundary states for committor is a difficult problem. As even such a natural approach as using the rmsd leads to inaccuracies. The problem is likely to be more

severe for more complex cases, e.g., intrinsically disordered proteins, allosteric transitions, etc, which could be treated by the proposed approach.

Once constructed, the landscapes (Fig. 5) can be postprocessed to obtain descriptions of minima, TSs, pathways in terms of easy-interpretable coordinates, e.g., dihedral angles, distances,[24] or secondary and tertiary structures. For example, since we can easily identify structures that belong to every region of the FES, a supervised machine learning model can be trained to assign these structures to these regions. It will make the model to learn to identify the most important molecular coordinates, e.g., inter-atom distances or dihedral angles, that discriminate these states.[24] Or, more generally, one can consider a standard machine learning regression problem of approximating the determined EVs coordinates u_1 and u_2, by a function of e.g., selected collective variables or inter-atom distances or dihedral angles. The regression problem is simpler then the original problem of accurate determination of the slowest EVs. Note, that it is, probably, easier to approximate $\tilde{u}_1$ and $\tilde{u}_2$, where TS and minima have similar scales.

Here we analyze the FES in terms of tertiary structures. For every free energy minimum we find the geometric average of all the structures in the minimum. Each structure is optimally superposed on the first trajectory structure of the minimum. A structure from the trajectory closest to the geometric average is found and is considered as a representative structure for the minimum. The process is repeated a few iterations with all the structures superposing on the representative structure, until the latter stops changing. The root-mean-square fluctuations (rmsf) for every residue is computed as the square root of the mean squared distances of all the atoms of the residue between the representative structure and all the superposed structures. Cartoon pictures of representative structures, colored according to the rmsf, from 0.5 Å (blue) to 13 Å (red) are shown on Fig. 5d.

In minimum 3 the protein is almost folded: all three helices are formed with a relatively high propensity and are all at the right positions. The hydrophobic core is not formed and the structure is rather flexible. In the native minimum (4) the folding is completed

by forming the hydrophobic core and making the structure stable. Near-native minimum 5 has first and third helices partially unraveled.[25] In minimum 3 residues 18-24 form a turn, connecting second and third helices, whereas in minima 1 and 2, they form a helix with > 90 % propensity. It leads to the possibility of the second and third helices forming a single long helix in minimum 2 and a longer second helix in 1.

The two-dimensional FES $F(\tilde{u}_1, \tilde{u}_2)$ can be used to find the correspondence among the minima on the FEPs and see the evidence of parallel pathways. In particular, the FES for HP35 has an L-like shape and shows no evidence of parallel pathways. The one-dimensional FEPs, i.e., $F(\tilde{u}_1)$, on the other hand, are better suited for the quantitative analysis of the dynamics, like determining free energies of TSs and minima, free energy barriers and pre-exponential factors, computing rates, mean first passage times, etc.

We have also applied the approach the FIP35 protein trajectory (Fig. 6).[1] The EV validation test $\Theta(x, \Delta t)$ was bounded by ± 0.2 for both EVs. This trajectory has only 15 folding-unfolding events, which illustrates that the approach can analyze systems with very limited sampling. $F(\tilde{u}_1)$ shows two minima with an intermediate state in agreement with other studies.[26,27] The two-dimensional FES has an A-like shape and shows the evidence of two parallel pathways, i.e., protein folds from 1 to 4 via 2 or 3. The representative structure of 2 has the first hairpin formed, while that of 3 has the second hairpin formed to a much larger degree. Surprisingly, region 3 is a TS rather than an intermediate state. It probably explains why this pathway is much less populated. It might be difficult to detect this pathway using MSMs. The intermediate TS is much less populated, thus, a rather large clustering size could be required to have a representative statistics. However, a large clustering size will make it more likely that points from the TS are assigned to free energy minima, which are much more populated.

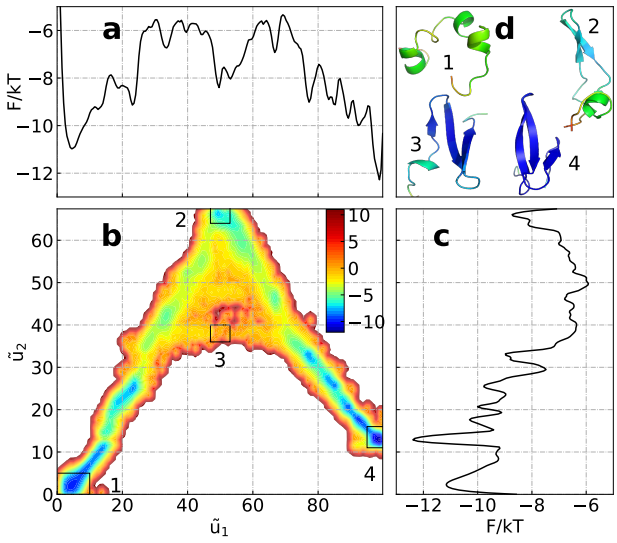


Figure 6: Free energy landscapes of FIP35: notation as in Fig. 5.

4 Concluding Discussion

We have described a blind approach for the determination of the slowest relaxation eigenvectors from an equilibrium trajectory. The approach determined the first and second slowest eigenvectors for the HP35 and FIP35 proteins with high spatio-temporal accuracy, as validated by the stringent criterion at the shortest lag time of 0.2 ns. In contrast to alternative (parametric) approaches, which require approximating functions with many parameters and extensive expertise with the system, the approach directly determines eigenvectors time-series and uses no system specific information. The optimality criterion is another important ingredient of the approach, which makes the uniform optimization possible. The approach can be used in cases when one does not want to introduce any bias in the analysis, e.g., due to employed approximating functions, or one does not have good approximating functions. It can also be used aposteriory to check if possible bias in the analysis has altered the results. As the HP35 example illustrates, even a seemingly innocent and natural choice of boundary states can hide the inherent complexity of the landscape.

The approach was illustrated by analyzing long equilibrium trajectories, i.e., state-of-the-art atomistic protein folding simulations.[1,2] However, generating such trajectories by

brute-force molecular dynamics is very computationally demanding. A number of advanced sampling methods have been suggested to alleviate the sampling problem, e.g, umbrella sampling,[28,29] steered-MD,[30] replica-exchange,[31,32] meta-dynamics.[33] If an enhanced sampling method generates unbiased, equilibrium sampling, possibly consisting of many trajectories, e.g., the trajectories of the base replica of the recently suggested REHT method,[34] they can be analyzed by the approach directly. One just needs to extend the summation to all the trajectories in Eq. 2. For other approaches, which can be used to determine equilibrium probabilities, however perturb the natural dynamics of the system, the equilibrium sampling needs to be generated first. It can be done, e.g., by starting many trajectories with natural, unbiased dynamics from the obtained configurations with equilibrium probabilities. It is possible to extend the developed non-parametric approaches to non-equilibrium sampling, which is generated, for example, by adaptive sampling methods. Mainly, it requires the change of the optimization functional and correspondingly the equations for optimal parameters of the variation (Eq. 2 here) and is discussed elsewhere.[35]

Here, we consider EV as a function of a trajectory, rather than a function of configuration space. In principle, it is possible to record all the parameters of the transformations and collective variables during EV optimization (training) and apply them later, in the same order, to new (test) data. That would make the determined EV a function of the configuration space. Here, however, we did not record the transformations, and computed the EV only for the configurations along a trajectory.

It is instructive to compare the proposed approach with alternative approaches. Diffusion maps,[36] Laplacian eigenmaps[37] and Isomap[38] are non-parametric generic dimensionality reduction methods. The main difference between these approaches and the proposed approach is that the former analyse a model of the dynamics, while the later - the actual dynamics. For example, the diffusion maps and Laplacian eigenmaps effectively define transition matrix between the configurations as the heat (diffusion) kernel according to the distance between them. It can be said, loosely, that these methods perform dimensionality reduction with a

focus on preserving the properties (proximity) of a given configuration space. However, it is well known that the geometric proximity is a poor indicator of kinetics proximity. Configurations which are close geometrically can be separated by high barriers, while motions along the low energy normal modes (i.e., low barriers) are generally associated with large conformational changes.

A large collection of parametric approaches, e.g., tICA,[13,14] VAMP,[39] EDMD[39,40] aims to approximate the slowest eigenvectors by multy-parametric functions, for example, a linear combination of collective variables or a neural network. Their major weakness is that their performance is limited by the choice of the employed functional forms and the input/collective variables. Since finding, e.g., an optimal architecture of a neural network or informative collective variables are difficult tasks. While intuition can help to solve these problems for low-dimensional model systems, the difficulty in the case of complex realistic systems becomes apparent, when one realizes that such a function should be able to accurately project a few million snapshots of a very high-dimensional trajectory. In particular, it implies an extensive knowledge of the system, and that an acceptable solution is likely to be system specific.

The proposed method is non-parametric and can approximate any EV with high accuracy. While each iteration may depend on the exact choice of the family of collective variables/molecular descriptors/features, the final EVs do not, since they provide optimum to a (non-parametric) target functional, when the optimization converges. We assume that the employed input variables contain all the information about the dynamics of interest. For the analysis of biomolecular simulations one can suggest the inter-atom distances as the standard sets of input variables. The iterative optimization of EVs, using these standard input variables, is a more generic and a more efficient approach than custom design of multi-parametric functions. As Fig. 1a shows, a few thousands iterations can provide a rather good approximation to an EV with the corresponding eigenvalue within a small factor from the exact value. The determined EVs pass a stringent validation test at a very short

time-scales of the trajectory sampling interval. It means that the obtained EVs time-series are more accurate than those obtained with alternative approaches. They provide a higher temporal resolution in the description of the dynamics. Much shorter lag time also means that much shorter trajectories are required for the simple strategy of exascale simulations and thus a much larger possible speedup over direct, brute-force simulations.

Acknowledgments

I am grateful to David Shaw and his coworkers for making the folding trajectories available.

5 Appendix

5.1 Adaptive non-parametric optimization of eigenvectors

The simple, non-adaptive algorithm, optimizes EVs in a non-uniform way analogous to the committor case. It is easier to optimize free energy barriers than minima. To perform optimization in a uniform way one needs first to detect sub-optimal regions of EVs and focus optimization on them. To detect the most suboptimal regions for current lag time Δt, we first find a longer lag time Δt_1, which exhibits the most nonuniformness in the distance between $\Theta(x, \Delta t_1)$ and $\Theta(x, \Delta t)$:

$$\Delta t_1 = \arg\sup_{t_i}[\max_x \Delta\Theta(x, \Delta t_i, \Delta t) - \min_x \Delta\Theta(x, \Delta t_i, \Delta t)], \tag{9}$$

here $\Delta\Theta(x, \Delta t_i, \Delta t) = \Theta(x, \Delta t_i) - \Theta(x, \Delta t)$. Then, the relative degree of suboptimality of region around x is defined as

$$s(x) = \exp[\Delta\Theta(x, \Delta t_1, \Delta t) - \max_x \Delta\Theta(x, \Delta t_1, \Delta t)] \tag{10}$$

It takes maximal value of 1 for the most suboptimal part where the difference between $\Theta(x, \Delta t_1)$ and $\Theta(x, \Delta t)$ is maximal. To focus optimization on such suboptimal regions we make p_{fix} position dependent, large for optimal regions and small for suboptimal regions. Consequently, the optimization is more focused on less optimized regions, because they have a smaller number of fixed points and are less constraint. For example, an extremely over-optimized region might have $p_{\text{fix}} = 1$, i.e., all the points fixed and thus it will not be optimized at all. Here we used

$$p_{\text{fix}}(x) = \min[1, p_{\text{fix}} \times s(x)^{-10}] \tag{11}$$

Before every iteration, the $p_{\text{fix}}(x)$ values are computed for active (k-th) eigenvector, and are used to select fixed points. Namely, a point at time moment $i\Delta t$, that has eigenvector coordinate $u_\beta(i\Delta t)$ is selected to be fixed with probability $p_{\text{fix}}(u_\beta(i\Delta t))$.

References

(1) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **2010**, *330*, 341–346.

(2) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How Fast-Folding Proteins Fold. *Science* **2011**, *334*, 517–520.

(3) Freddolino, P. L.; Harrison, C. B.; Liu, Y.; Schulten, K. Challenges in protein-folding simulations. *Nat Phys* **2010**, *6*, 751–758.

(4) Schwantes, C. R.; Pande, V. S. Modeling Molecular Kinetics with tICA and the Kernel Trick. *J. Chem. Theory Comput.* **2015**, *11*, 600–608.

(5) Banushkina, P. V.; Krivov, S. V. Optimal reaction coordinates. *WIREs Comput Mol Sci* **2016**, *6*, 748–763.

(6) Noé, F.; Clementi, C. Collective variables for the study of long-time kinetics from molecular trajectories: theory and methods. *Curr. Opin. Struct. Biol.* **2017**, *43*, 141–147.

(7) Jung, H.; Covino, R.; Hummer, G. Artificial Intelligence Assists Discovery of Reaction Coordinates and Mechanisms from Molecular Dynamics Simulations. **2019**, arXiv: 1901.04595 [physics:chem–ph].

(8) Peters, B. Common Features of Extraordinary Rate Theories. *J. Phys. Chem. B* **2015**, *119*, 6349–6356.

(9) Peters, B. Reaction Coordinates and Mechanistic Hypothesis Tests. *Ann. Rev. Phys. Chem.* **2016**, *67*, 669–690.

(10) Krivov, S. V. Protein Folding Free Energy Landscape along the Committor - the Optimal Folding Coordinate. *J. Chem. Theory Comput.* **2018**, *14*, 3418–3427.

(11) Shuler, K. E. Relaxation Processes in Multistate Systems. *The Physics of Fluids* **1959**, *2*, 442–448.

(12) McGibbon, R. T.; Husic, B. E.; Pande, V. S. Identification of simple reaction coordinates from complex dynamics. *J Chem Phys* **2017**, *146*, 044109.

(13) Schwantes, C. R.; Pande, V. S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J. Chem. Theory Comput.* **2013**, *9*, 2000–2009.

(14) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **2013**, *139*, 015102.

(15) Wan, H.; Voelz, V. A. Adaptive Markov state model estimation using short reseeding trajectories. *J. Chem. Phys.* **2020**, *152*, 024103.

(16) Hernández, C. X.; Wayment-Steele, H. K.; Sultan, M. M.; Husic, B. E.; Pande, V. S. Variational encoding of complex dynamics. *Phys. Rev. E* **2018**, *97*, 062412.

(17) Mardt, A.; Pasquali, L.; Wu, H.; Noé, F. VAMPnets for deep learning of molecular kinetics. *Nature Communications* **2018**, *9*, 5.

(18) Banushkina, P. V.; Krivov, S. V. Nonparametric variational optimization of reaction coordinates. *J. Chem. Phys.* **2015**, *143*, 184108.

(19) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. Protein folding kinetics and thermodynamics from atomistic simulation. *PNAS* **2012**, *109*, 17845–17850.

(20) Krivov, S. V. On Reaction Coordinate Optimality. *J. Chem. Theory Comput.* **2013**, *9*, 135–146.

(21) Krivov, S. CFEPs. `https://github.com/krivovsv/CFEPs`, 2020.

(22) Berezhkovskii, A.; Szabo, A. Ensemble of transition states for two-state protein folding from the eigenvectors of rate matrices. *J. Chem. Phys.* **2004**, *121*, 9186–9187.

(23) Krivov, S. V.; Karplus, M. Diffusive reaction dynamics on invariant free energy profiles. *PNAS* **2008**, *105*, 13841–13846.

(24) Brandt, S.; Sittel, F.; Ernst, M.; Stock, G. Machine Learning of Biomolecular Reaction Coordinates. *J. Phys. Chem. Lett.* **2018**, *9*, 2144–2150.

(25) Beauchamp, K. A.; McGibbon, R.; Lin, Y.-S.; Pande, V. S. Simple few-state models reveal hidden complexity in protein folding. *PNAS* **2012**, *109*, 17807–17813.

(26) Krivov, S. V. The Free Energy Landscape Analysis of Protein (FIP35) Folding Dynamics. *J. Phys. Chem. B* **2011**, *115*, 12315–12324.

(27) Boninsegna, L.; Gobbo, G.; Noe, F.; Clementi, C. Investigating Molecular Kinetics by Variationally Optimized Diffusion Maps. *J. Chem. Theory Comput.* **2015**, *11*, 5947–5960.

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

(29) Souaille, M.; Roux, B. Extension to the Weighted Histogram Analysis Method: Combining Umbrella Sampling with Free Energy Calculations. *Comput. Phys. Commun.* **2001**, *135*, 40.

(30) Isralewitz, B.; Baudry, J.; Gullingsrud, J.; Kosztin, D.; Schulten, K. Steered Molecular Dynamics Investigations of Protein Function. *J. Mol. Graphics Modell.* **2001**, *19*, 13.

(31) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141.

(32) Fukunishi, H.; Watanabe, O.; Takada, S. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *J. Chem. Phys.* **2002**, *116*, 9058.

(33) Barducci, A.; Bonomi, M.; Parrinello, M. Metadynamics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 826.

(34) Appadurai, R.; Nagesh, J.; Srivastava, A. High resolution ensemble description of metamorphic and intrinsically disordered proteins using an efficient hybrid parallel tempering scheme. *Nature Communications* **2021**, *12*, 958.

(35) Krivov, S. Non-Parametric Analysis of Non-Equilibrium Simulations. **2021**, arXiv: 2102.03950 [physics:chem–ph].

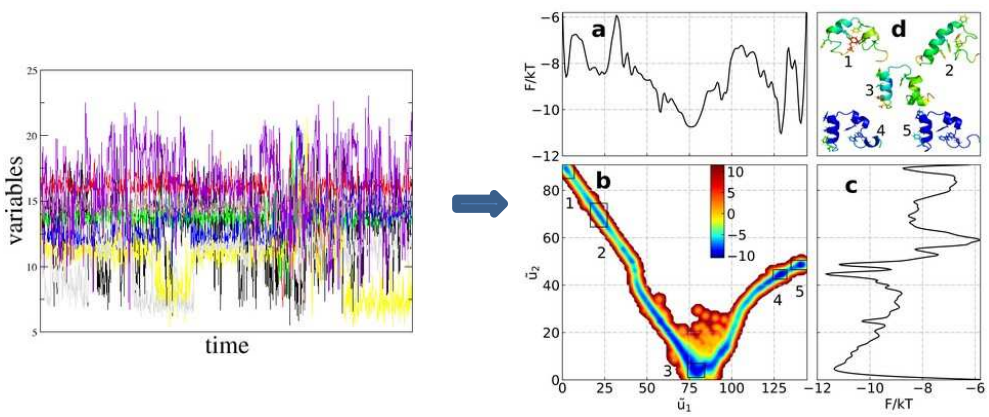
(36) Coifman, R. R.; Lafon, S. Diffusion maps. *Appl. Comput. Harmon. Anal.* **2006**, *21*, 5–30.

(37) Belkin, M.; Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **2003**, *15*, 1373–1396.

(38) Tenenbaum, J. B.; Silva, V. d.; Langford, J. C. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* **2000**, *290*, 2319–2323.

(39) Wu, H.; Nüske, F.; Paul, F.; Klus, S.; Koltai, P.; Noé, F. Variational Koopman models: Slow collective variables and molecular kinetics from short off-equilibrium simulations. *J. Chem. Phys.* **2017**, *146*, 154104.

(40) Williams, M. O.; Kevrekidis, I. G.; Rowley, C. W. A Data-Driven Approximation of the Koopman Operator: Extending Dynamic Mode Decomposition. *J. Nonlinear Sci.* **2015**, *25*, 1307–1346.



For Table of Contents Only