Variational approximation of molecular kinetics from short off-equilibrium simulations

Hao Wu,[*] Feliks Nüske,[†] Fabian Paul,[‡] Stefan Klus,[§] Péter Koltai,[¶] and Frank Noé[**]

Department of Mathematics and Computer Science,
Freie Universität Berlin, Arnimallee 6, 14195 Berlin, Germany

Abstract

Markov state models (MSMs) and Master equation models are popular approaches to approximate molecular kinetics, equilibria, metastable states, and reaction coordinates in terms of a state space discretization usually obtained by clustering. Recently, a powerful generalization of MSMs has been introduced, the variational approach of conformation dynamics (VAC) and its special case the time-lagged independent component analysis (TICA), which allow us to approximate molecular kinetics and reaction coordinates by linear combinations of smooth basis functions or order parameters. While MSMs can be learned from trajectories whose starting points are not sampled from an equilibrium ensemble, TICA and VAC have as yet not enjoyed this property, and thus previous TICA/VAC estimates have been strongly biased when used with ensembles of short trajectories. Here, we employ Koopman operator theory and ideas from dynamic mode decomposition (DMD) to show how TICA/VAC can be used to estimate the unbiased equilibrium distribution from short-trajectory data and further this result in order to construct unbiased estimators for expectations, covariance matrices, TICA/VAC eigenvectors, relaxation timescales, and reaction coordinates.

[*] hao.wu@fu-berlin.de

[†] feliks.nueske@fu-berlin.de; H. Wu and F. Nüske contributed equally to this work.

[‡] fab@zedat.fu-berlin.de

[§] stefan.klus@fu-berlin.de

[¶] peter.koltai@fu-berlin.de

[**] frank.noe@fu-berlin.de

I. INTRODUCTION

With the ability to generate extensive and high-throughput molecular dynamics (MD) simulations [1–9], the spontaneous sampling of rare-events such as protein folding, conformational changes and protein-ligand association have become accessible [10–17]. Markov state models (MSMs) [18–25], Master-equation models [26–28] and closely related approaches [29–33] have emerged as powerful frameworks for the analysis of extensive MD simulation data, as they approximate the true kinetics without requiring strong prior definition of relevant reaction coordinates [23, 34], allow a large variety of mechanistic information to be extracted [10, 35, 36], experimental observables to be computed and structurally interpreted [12, 28, 37–40]. They provide a direct approximation of the dynamic modes describing the slow conformational changes that are identical or closely related to the so-called reaction coordinates, depending on which notion of that term is employed [41–45]. An especially powerful feature of MSMs and similar approaches is that they can be estimated from non-equilibrium data – more specifically, the MSM transition probabilities $p_{ij}(\tau)$, i.e. the probability that the trajectory is found in a set A_j a time lag τ after it has been found in a set A_i,

$$p_{ij}(\tau) = \mathrm{Prob}\left[x(t+\tau) \in A_j \mid x(t) \in A_i\right],$$

is a conditional transition probability. $p_{ij}(\tau)$ can be estimated without bias even if the trajectory is not initiated from a global, but only a local equilibrium distribution [23]. Consequently, given $c_{ij}(\tau)$ transition events between states i and j at lag time τ, the maximum likelihood estimator of the transition probability can be easily shown to be

$$p_{ij}(\tau) = \frac{c_{ij}(\tau)}{\sum_k c_{ik}(\tau)}, \tag{1}$$

i.e. the fraction of the number of transitions to j conditioned on starting in i. This conditionality is a key reason why MSMs have become popular to analyze short distributed simulations that are started from arbitrary configurations whose relationship to the equilibrium distribution is initially unknown.

However, when estimating (1) from simulation data, one does not generally obtain a time-reversible estimate, i.e. the stationary probabilities of the transition matrix, π_i, will usually not fulfill the detailed balance equations $\pi_i p_{ij} = \pi_j p_{ji}$, even if the underlying dynamics are microscopically time-reversible. Compared to a reversible transition matrix, a transition matrix with independent estimates of p_{ij} and p_{ji} has more free parameters, resulting in larger statistical uncertainties, and moreover may possess complex-valued eigenvalues and eigenvectors, which exclude or exacerbate various analyses [46]. Since most molecular dynamics simulations are in thermal equilibrium and thus fulfill at least a generalized microscopic reversibility (Appendix B in [47]), it is desirable to force p_{ij} to fulfill detailed balance, which both reduces statistical uncertainty and enforces a real-valued spectrum [46, 48]. In old studies, the pragmatic solution to this problem was often to symmetrize the count matrix, i.e. to simply set $c_{ij}^{\mathrm{sym}} = c_{ij} + c_{ji}$, which is equivalent to evaluating the simulation trajectory forward and backward, and which leads to a transition matrix with detailed balance when inserted into (1). However, it has been known since at least 2008 that this estimator is strongly biased, and therefore reversible maximum likelihood and Bayesian estimators have been developed [22, 23, 28, 46, 48, 49]. These algorithms formulate the estimation problem as an optimization or sampling problem of the transition matrix constrained to fulfill detailed balance. The idea of these algorithms becomes clear when writing the reversible maximum likelihood estimator in two subsequent steps, as demonstrated in [46]:

1. *Reweighting*: Estimate the stationary distribution π_i given all transition counts c_{ij} and a reversible Markov model.

2. *Estimation*: Insert π_i and c_{ij} into an equation for the transition matrix to obtain a maximum likelihood estimate of p_{ij} with detailed balance.

Recently, a powerful extension to the Markov modeling framework has been introduced: the variational approach of conformation dynamics (VAC) [50–52]. It has been known for many years that Markov state models are good approximations to molecular kinetics if their largest eigenvalues and eigenvectors approximate the eigenvalues and eigenfunctions of the Markov operator governing the full-phase space dynamics [18, 34, 53], moreover the first few eigenvalues and eigenvectors are sufficient to compute almost all stationary and kinetic quantities of interest [37, 38, 54–56]. The VAC has generalized this idea beyond discrete states and formulated the approximation problem of molecular kinetics in terms of an approach that is similar to the variational approach in quantum mechanics [50–52]. It is based on the following variational principle: If we are given a set of n orthogonal functions of state space, and evaluate the autocorrelation of the molecular dynamics in these functions at lag time τ, these will give us lower bounds to the true eigenvalues $\lambda_1(\tau)$, ..., $\lambda_n(\tau)$ of the Markov operator, equivalent to an (under)estimate of relaxation timescales and an (over)estimate of relaxation rates. Only if the n functions used are the eigenfunctions themselves, then their autocorrelations will be maximal and identical to the true eigenvalues $\lambda_1(\tau)$, ..., $\lambda_n(\tau)$. This principle allows to formulate variational optimization algorithms to approximate the eigenvalues and eigenfunctions of the Markov operator. The linear variational approach proceeds as follows:

1. Fix an arbitrary basis set $\boldsymbol{\chi} = [\chi_1(\mathbf{x}), ..., \chi_n(\mathbf{x})]$ and evaluate the values of all basis functions for all sampled MD configurations $\mathbf{x}$.

2. Estimate two covariance matrices, the instantaneous (PCA) covariance matrix $\mathbf{C}(0)$, and the time-lagged covariance matrix $\mathbf{C}(\tau)$ from the basis-set-transformed data.

3. Solve a generalized eigenvalue problem involving both $\mathbf{C}(0)$ and $\mathbf{C}(\tau)$, and obtain estimates for the eigenvalues $\lambda_i(\tau)$ and expansion coefficients $\mathbf{b}_i$. The estimate for the i-th eigenfunction is then given by

$$\psi_i(\mathbf{x}) = \sum_j b_{ij}\chi_j(\mathbf{x}). \tag{2}$$

This approach provides the optimal linear representation (2). Note that the functions $\boldsymbol{\chi}$ can be arbitrary nonlinear functions in the original coordinates $\mathbf{x}$, which allows complex nonlinear dynamics to be encoded even within this linear optimization framework. The variational approach has spawned a variety of follow-up works, for example it has been shown that the algorithm called blind source separation, time-lagged or time-structure based independent component analysis (TICA) established in signal processing and machine learning [57–59] is a special case of the VAC [51]. TICA is now widely used in order to reduce the dimensionality of MD data sets to a few slow collective coordinates, in which MSMs and other kinetic models can be built efficiently [51, 60, 61]. The VAC has been used to generate and improve guesses of collective reaction coordinates [62, 63]. A VAC-based metric has been defined which transforms the VAC estimates into a space in which Euclidean distance corresponds to kinetic distance [64, 65]. A kernel version of TICA/VAC has been proposed in [66], and it has been suggested to use the VAC eigenvalues in order to perform kinetic model selection by means of cross-validation [67]. A tensor-based approach to find the representation of eigenfunctions in terms of products of simple one-coordinate functions has been formulated [68], and basis sets for peptide dynamics have been proposed [69].

Despite the popularity of VAC and TICA, their estimation from MD data is still in the stage that MSMs had been about a decade ago: A direct estimation of covariance matrices will generally provide a non-symmetric $\mathbf{C}(\tau)$ matrix and complex eigenvalues/eigenfunction estimates that are not consistent with reversible molecular dynamics. In order to avoid this problem, the current state of the art is to enforce the symmetrization of covariance matrices directly [51, 60, 66]. This approach – which is analogous to symmetrizing count matrices in MSM estimation – introduces a strong bias when the simulation data are not in equilibrium. In lack of a better estimator, this approach is currently used also with short distribution MD simulations despite the fact that the resulting timescales and eigenfunctions may be biased or even misleading. This problem is addressed in the present paper.

For reversible dynamics, TICA and VAC are identical to dynamic mode decomposition (DMD) [70–73] and extended dynamic mode decomposition (EDMD) [74], respectively. However DMD and EDMD have been developed in the context of dynamical systems and fluid mechanics where data is often nonreversible and non-stationary. Thus, the theory upon which DMD/EDMD are based [75] can be used in order to formulate estimators for TICA and VAC that are also unbiased in the presence of short non-equilibrium simulations. First it is shown that the direct estimate of covariance matrices provides an unbiased TICA/VAC estimator for nonreversible dynamics. Then an unbiased estimator for reversible dynamics is derived, which involves two steps analogously to optimal reversible MSM estimation:

1. *Reweighting*: Estimate a reweighting vector u_i with an entry for each basis function given the empirical covariance matrices $\mathbf{C}(0)$ and $\mathbf{C}(\tau)$.

2. *Estimation*: Insert u_i and $\mathbf{C}(0)$, $\mathbf{C}(\tau)$ into an equation for the equilibrium estimates of $\mathbf{C}(0)$ and $\mathbf{C}(\tau)$ in order to obtain an unbiased reversible estimate for computing eigenvalues and eigenfunctions.

In addition to this result, the reweighting vector u_i allows us to approximate *any* equilibrium estimate in terms of a linear combination of our basis functions from off-equilibrium data. Thus, we obtain a generalized estimator for equilibrium stationary and kinetic quantities without the need to compute clusters and to construct a Markov state model. The new methods are illustrated on toy examples with stochastic dynamics and a benchmark protein-ligand binding problem. All analyses in this paper were made using the PyEMMA program version 2.2 (www.pyemma.org) [76].

II. VARIATIONAL APPROACH OF CONFORMATION DYNAMICS (VAC)

A. Variational principle of conformation dynamics

For simulations of molecular dynamics (MD), it is natural to model simulation trajectories of a molecular system as an ergodic and time-reversible Markov process $\{\mathbf{x}_t\}$ living in a phase space Ω by defining $\mathbf{x}_t$ as a collection of all variables that can determine the conformational progression after time t (e.g., positions and velocities of all atoms). Ergodicity implies that the probability density p_t of system state $\mathbf{x}_t$ at time t tends to a unique stationary density $\mu(\mathbf{x})$ as $t \to \infty$, and reversibility can be described by the detailed balance condition

$$p(\mathbf{x},\mathbf{y};\tau)\,\mu(\mathbf{x}) = p(\mathbf{y},\mathbf{x};\tau)\,\mu(\mathbf{y}), \quad \forall \mathbf{x},\mathbf{y} \in \Omega, \tag{3}$$

where $p(\mathbf{x},\mathbf{y};\tau)$ denotes the transition density from $\mathbf{x}$ to $\mathbf{y}$ with lag time τ, i.e., the conditional probability density of $\mathbf{x}_{t+\tau} = \mathbf{y}$ given $\mathbf{x}_t = \mathbf{x}$. Under these conditions, the time evolution of the ensemble of the molecular system can be decomposed into a set of relaxation processes as

$$p_{t+\tau}(\mathbf{x}) = \sum_{i=1}^{\infty} \mathrm{e}^{-\frac{\tau}{t_i}}\,\mu(\mathbf{x})\,\psi_i(\mathbf{x})\,\langle \psi_i, p_t\rangle, \tag{4}$$

where t_i are relaxation timescales sorted in decreasing order, ψ_i are eigenfunctions of the transfer operator or Koopman operator of $\{\mathbf{x}_t\}$ with eigenvalues $\lambda_i(\tau) = \mathrm{e}^{-\frac{\tau}{t_i}}$ (see Section III A). Inner products $\langle \psi_i, p_t\rangle = \int \mathrm{d}\mathbf{x}\ \psi_i(\mathbf{x})\,p_t(\mathbf{x})$ measure projections of p_t onto the corresponding eigenspace. The first spectral component is given by the constant eigenfunction $\psi_1(\mathbf{x}) = \mathbb{1}(\mathbf{x}) \equiv 1$ and and infinite timescale $t_1 = \infty > t_2$ corresponding to the stationary state of the system. Obviously, the long-term temporal behavior of a molecular system can be modeled by only a few dominant spectral components of the system dynamics associated with leading eigenvalues since the remaining part decays quickly with τ.

The eigenvalues and eigenfunctions can also be formulated by the following variational principle [50, 52]: *For any $m \geq 1$, the first m eigenfunctions $\psi_1, \ldots, \psi_m$ are the solution of the following optimization problem*

$$\begin{aligned} &\max_{f_1,\ldots,f_m} \sum_{i=1}^{m} \mathbb{E}_\mu\left[f_i(\mathbf{x}_t)\,f_i(\mathbf{x}_{t+\tau})\right], \\ &\text{s.t.}\quad \mathbb{E}_\mu\left[f_i(\mathbf{x}_t)^2\right] = 1, \\ &\qquad\ \ \mathbb{E}_\mu\left[f_i(\mathbf{x}_t)\,f_j(\mathbf{x}_{t+\tau})\right] = 0,\ \textit{for } i \neq j, \end{aligned} \tag{5}$$

and the maximum value is the sum of $\lambda_1, \ldots, \lambda_m$, where $\mathbb{E}_\mu[\cdot]$ denotes the expected value with $\mathbf{x}_t$ sampled from the stationary density μ. Notice that each term $\mathbb{E}_\mu\left[f_i\left(\mathbf{x}_t\right) f_i\left(\mathbf{x}_{t+\tau}\right)\right]$ in the objective function can be interpreted as a Rayleigh quotient of the transfer operator or Koopman operator and the conclusion is identical to the Rayleigh-Ritz principle [50]. Therefore, for every other set of functions that aims at approximating the true eigenfunctions, the eigenvalues will be underestimated, and we can use this variational principle in order to search for the best approximation of eigenfunctions and eigenvalues.

According to this formulation, the eigenfunctions $\psi_1, \ldots, \psi_m$ can be interpreted as m slow coordinates, that are related or possibly equivalent to what are commonly called "reaction coordinates" that satisfy the following properties:

- They are uncorrelated.
- They describe the directions of the slow kinetics with the maximal autocorrelations $\mathbb{E}_\mu\left[\psi_i\left(\mathbf{x}_t\right)\psi_i\left(\mathbf{x}_{t+\tau}\right)\right] = \lambda_i(\tau)$.
- Population changes along these coordinates decay exponentially with $\lambda_i(\tau) = \mathrm{e}^{-\frac{\tau}{t_i}}$.

Thus, the dominant spectral components are key to the analysis and understanding of conformation dynamics of molecular systems, where eigenvalues characterize timescales of conformation dynamics and eigenfunctions are an ideal choice of reaction coordinates. In what follows, we will investigate how to approximate the spectral components from MD simulation data.

B. Linear variational approach

In this paper, we focus on the finite-dimensional approximation of spectral components of conformation dynamics, which approximates each eigenfunction by a linear combination of real-valued conformational basis functions $\boldsymbol{\chi} = \left(\chi_1, \ldots, \chi_m\right)^\top$

$$\hat{\psi}_i(\mathbf{x}) = \sum_j b_{ij}\chi_j(\mathbf{x}) = \mathbf{b}_i^\top \boldsymbol{\chi}(\mathbf{x}) \tag{6}$$

with expansion coefficients b_{ij}.

A common way to get such approximations for analysis of MD simulations is the variational approach of conformation dynamics (VAC) [50, 52], which is based on the variational formulation (5) of spectral components.

Within the linear expansion (6), the optimal approximation of eigenvalues λ_i and eigenfunctions ψ_i according to (5) are the solutions of the generalized eigenvalue problem [50]

$$\mathbf{C}\left(\tau\right)\mathbf{B}=\mathbf{C}\left(0\right)\mathbf{B}\boldsymbol{\Lambda}, \tag{7}$$

with $\boldsymbol{\Lambda}=\mathrm{diag}\left(\lambda_1,\ldots,\lambda_m\right)$ and $\mathbf{B}=\left(\mathbf{b}_1,\ldots,\mathbf{b}_m\right)$. Here,

$$\mathbf{C}\left(0\right)=\mathbb{E}_\mu\left[\boldsymbol{\chi}\left(\mathbf{x}_t\right)\boldsymbol{\chi}\left(\mathbf{x}_t\right)^\top\right], \tag{8}$$

$$\mathbf{C}\left(\tau\right)=\mathbb{E}_\mu\left[\boldsymbol{\chi}\left(\mathbf{x}_t\right)\boldsymbol{\chi}\left(\mathbf{x}_{t+\tau}\right)^\top\right] \tag{9}$$

are correlation matrix and time-lagged correlation matrix of the basis functions in the equilibrium ensemble. This conclusion suggests the following approximation procedure:

1. Estimate correlation matrices $\mathbf{C}\left(0\right)$ and $\mathbf{C}\left(\tau\right)$ from data.

2. Solve the generalized eigenvalue problem (7).

3. Output estimated eigenvalues $\hat{\lambda}_i$ and eigenfunctions $\hat{\psi}_i$. The latter can be used in order to define reaction coordinates between the metastable states of the system, or as essential kinetic coordinates in order to reduce the dimensionality of the problem and interpret the molecular events occurring with slow rates [51, 60].

The VAC provides a general framework for the finite-dimensional approximation of spectral components of conformation dynamics, and two widely used analysis methods, time-lagged independent component analysis (TICA) [51, 57, 60] and Markov state models (MSMs) [23], are both special cases of VAC.

TICA: In TICA, basis functions are mean-free molecular coordinates (internal or Cartesian) or order parameters (e.g. contact maps), $\boldsymbol{\chi}=\mathbf{r}-\mathbb{E}_\mu[\mathbf{r}]$, where $\mathbf{r}$ contains the selected coordinates and $\mathbb{E}_\mu[\mathbf{r}]$ are the means. Then the resulting estimates $\boldsymbol{\psi}$ of eigenfunctions can be viewed as a set of linearly independent components (ICs) with autocorrelations $\lambda_i(\tau)$. The dominant ICs can be used to reduce the dimension of the molecular system.

Notice that using mean free coordinates is equivalent to removing the stationary spectral component $(\lambda_1,\psi_1)\equiv(1,\mathbb{1})$, thus TICA will only contain the dynamical components, starting from (λ_2,ψ_2).

MSM: The MSM is a special case of the VAC while using the indicator functions as basis set:

$$\chi_i(\mathbf{x}) = \begin{cases} 1, & \text{for } \mathbf{x} \in A_i, \\ 0, & \text{for } \mathbf{x} \notin A_i, \end{cases} \tag{10}$$

where $A_1, \ldots, A_m$ form a partition of the phase space Ω. With such basis functions, the correlation matrix $\mathbf{C}(0)$ is a diagonal matrix with $[\mathbf{C}(0)]_{ii} = \Pr(\mathbf{x}_t \in A_i)$ being the equilibrium probability of A_i, and the (i,j)-th element $[\mathbf{C}(\tau)]_{ij} = \Pr(\mathbf{x}_t \in A_i, \mathbf{x}_{t+\tau} \in A_j)$ of the time-lagged correlation matrix $\mathbf{C}(\tau)$ is equal to the equilibrium frequency of the transition from A_i to A_j. Thus, a piecewise-constant approximation of eigenfunctions

$$\psi_j(\mathbf{x}) = [\mathbf{B}]_{ij}, \text{ for } \mathbf{x} \in A_i, \tag{11}$$

and the corresponding eigenvalues are given by the generalized eigenvalue problem (7), which can be equivalently transferred into an eigenvalue problem as

$$\mathbf{C}(\tau)\mathbf{B} = \mathbf{C}(0)\mathbf{B}\mathbf{\Lambda} \quad \Rightarrow \quad \mathbf{P}(\tau)\mathbf{B} = \mathbf{B}\mathbf{\Lambda} \tag{12}$$

if the equilibrium probability of each A_i is positive, where $\mathbf{P}(\tau) = \mathbf{C}(0)^{-1}\mathbf{C}(\tau)$ is the transition matrix of the MSM with $[\mathbf{P}(\tau)]_{ij} = \Pr(\mathbf{x}_{t+\tau} \in A_j | \mathbf{x}_t \in A_i)$. This is consistent with the conclusion obtained in the literature on MSMs [34].

The choice of more general basis functions for the VAC is beyond the scope of this paper, and some related work can be found in [52, 68, 69].

C. Estimation of correlation matrices

The remaining problem is how to obtain estimates of $\mathbf{C}(0)$ and $\mathbf{C}(\tau)$. For convenience, we introduce the following notation: we take all sampled coordinates $\mathbf{x}_t$ of a trajectory, evaluate their basis function values $\chi_1(\mathbf{x}_t)$, ..., $\chi_m(\mathbf{x}_t)$, and define the following two matrices:

$$\mathbf{X} = \begin{pmatrix} \chi_1(\mathbf{x}_1) & \cdots & \chi_m(\mathbf{x}_1) \\ \vdots & & \vdots \\ \chi_1(\mathbf{x}_{T-\tau}) & \cdots & \chi_m(\mathbf{x}_{T-\tau}) \end{pmatrix} \in \mathbb{R}^{N\times m}, \tag{13}$$

$$\mathbf{Y} = \begin{pmatrix} \chi_1(\mathbf{x}_{\tau+1}) & \cdots & \chi_m(\mathbf{x}_{\tau+1}) \\ \vdots & & \vdots \\ \chi_1(\mathbf{x}_T) & \cdots & \chi_m(\mathbf{x}_T) \end{pmatrix} \in \mathbb{R}^{N\times m}. \tag{14}$$

where each row corresponds to one stored timestep. Thus, $\mathbf{X}$ contains the first $N = T - \tau$ time steps and $\mathbf{Y}$ contains the last $N = T-\tau$ time steps. Assuming that $\{\mathbf{x}_t\}$ is ergodic, $\mathbf{C}(0)$ and $\mathbf{C}(\tau)$ can be directly estimated by time averages of $\boldsymbol{\chi}(\mathbf{x}_t)\,\boldsymbol{\chi}(\mathbf{x}_t)^\top$ and $\boldsymbol{\chi}(\mathbf{x}_t)\,\boldsymbol{\chi}(\mathbf{x}_{t+\tau})^\top$ over the trajectory:

$$\hat{\mathbf{C}}(0) = \frac{1}{N}\mathbf{X}^\top\mathbf{X}, \tag{15}$$

$$\hat{\mathbf{C}}(\tau) = \frac{1}{N}\mathbf{X}^\top\mathbf{Y}. \tag{16}$$

For the purpose of solving the eigenvalue problem (7), the factor $1/N$ may be ignored. Furthermore, multiple trajectories $k = 1, ..., K$ are trivially handled by adding up their contributions, e.g. $\hat{\mathbf{C}}(0) = \frac{1}{\sum_k N_k}\sum_k \mathbf{X}_k^\top\mathbf{X}_k$, etc.

Due to statistical noise or non-equilibrium starting points, the time-lagged correlation matrix $\hat{\mathbf{C}}(\tau)$ estimated by this method is generally not symmetric, even if the underlying dynamics are time-reversible. Thus, the eigenvalue problem (7) may yield complex eigenvalues and eigenvectors, which are undesirable in analysis of statistically reversible MD simulations. The relaxation timescales t_i can be computed from complex eigenvalues as $t_i = -\tau/\ln|\lambda_i(\tau)|$ by using the norm of eigenvalues, but it is *a priori* unclear how to perform component analysis and dimension reduction as in TICA based on complex eigenfunctions.

In order to avoid the problem of complex estimates, a symmetric estimator is often used in applications, which approximates $\mathbf{C}(0)$ and $\mathbf{C}(\tau)$ by empirically averaging over all transition pairs $(\mathbf{x}_t, \mathbf{x}_{t+\tau})$ and their reverses $(\mathbf{x}_{t+\tau}, \mathbf{x}_t)$, which is equivalent to averaging the time-forward and the time-inverted trajectory:

$$\hat{\mathbf{C}}_{\mathrm{sym}}(0) \approx \frac{1}{2N}\left(\mathbf{X}^\top\mathbf{X} + \mathbf{Y}^\top\mathbf{Y}\right), \tag{17}$$

$$\hat{\mathbf{C}}_{\mathrm{sym}}(\tau) \approx \frac{1}{2N}\left(\mathbf{X}^\top\mathbf{Y} + \mathbf{Y}^\top\mathbf{X}\right), \tag{18}$$

so that the estimate of $\mathbf{C}(\tau)$ is always symmetric and the generalized eigenvalue problem (7) has real-valued solutions.

For equilibrium simulations, i.e. if the simulation starting points are sampled from the global equilibrium, or the simulations are much longer than the slowest relaxation times, Eqs. (17) and (18) are unbiased estimates of $\mathbf{C}_\mu(0)$ and $\mathbf{C}_\mu(\tau)$ and can also be derived from the maximum likelihood estimation by assuming a multivariate normal distribution of $(\mathbf{x}_t, \mathbf{x}_{t+\tau})$ [66]. The major difficulty of this approach arises from non-equilibrium data, i.e. simulations whose starting points are not drawn from the equilibrium distribution and are not long enough to reach that equilibrium during the simulation. In this situation, (17) and (18) are biased estimates, i.e. they do not converge to the true covariance matrices and provide biased VAC/TICA results even in the limit of infinitely many trajectories.

The difference between the direct estimation and symmetric estimation methods of correlation matrices becomes clear when considering the MSM special case. Since the transition matrix is $\mathbf{P} = \mathbf{C}(0)^{-1}\mathbf{C}(\tau)$, as shown in Section II B, transition matrices of MSMs given by the two estimators are

$$[\mathbf{P}]_{ij} = \frac{c_{ij}(\tau)}{\sum_{k=1}^{m} c_{ik}(\tau)}, \qquad \text{(direct estimation)} \qquad (19)$$

$$[\mathbf{P}]_{ij} = \frac{c_{ij}(\tau) + c_{ji}(\tau)}{\sum_{k=1}^{m} c_{ik}(\tau) + c_{kj}(\tau)}, \qquad \text{(symmetric estimation)} \qquad (20)$$

respectively. If the transition dynamics between discrete states $A_1, \ldots, A_m$ are exactly Markovian, the direct estimator converges to the true transition matrix in the large-data limit for non-equilibrium or even nonreversible, whereas the symmetric estimator does not. However, the direct estimator may give a nonreversible transition matrix with complex eigenvalues, which is why the symmetric estimator has been frequently used before 2008 until it has been replaced by reversible maximum likelihood and Bayesian estimators [22, 23, 28, 46, 48, 49]. How do we resolve this problem in the more general case of VAC (or more specifically, TICA) estimation? Below, we will introduce a solution based on Koopman operator theory and dynamic mode decomposition (DMD).

III. DYNAMIC MODE DECOMPOSITION (DMD)

A. Koopman operator description of conformation dynamics

According to the Koopman operator theory [75], the dynamics of a dynamical system that is Markovian in phase space can be fully described by an integral operator $\mathcal{K}_\tau$, called *Koopman operator*, which maps an observable quantity $f(\mathbf{x}_t)$ at time t, to its expectation at time $t + \tau$ as

$$\begin{aligned} \mathcal{K}_\tau f(\mathbf{x}) &= \mathbb{E}\left[f(\mathbf{x}_{t+\tau}) \,|\mathbf{x}_t = \mathbf{x}\right] \\ &= \int \mathrm{d}\mathbf{y}\ p(\mathbf{x}, \mathbf{y}; \tau)\, f(\mathbf{y}). \end{aligned} \qquad (21)$$

If the dynamics fulfill detailed balance, the spectral components $\{(\lambda_i, \psi_i)\}$ discussed above are in fact the eigenvalues and eigenfunctions of the Koopman operator:

$$\mathcal{K}_\tau \psi_i = \lambda_i \psi_i \tag{22}$$

under the detailed balance condition. Notice that the operator description and decomposition of molecular kinetics can also be equivalently provided by the transfer operator, or backward propagator and the forward propagator [23], which propagate ensemble densities instead of observables. We exploit the Koopman operator in this paper because it is the only one of these operators that can be reliably approximated from non-equilibrium data in general. See Section III B and Appendix A for a more detailed analysis.

Eq. (22) suggests the following way for spectral estimation: We can first approximate the Koopman operator from data, and then extract the spectral components from the estimated operator.

B. (Extended) dynamic mode decomposition

Like in the VAC, we can also approximate the Koopman operator $\mathcal{K}_\tau$ by its projection $\mathcal{K}_\tau^{\mathrm{proj}}$ onto the subspace spanned by basis function $\boldsymbol{\chi}$ which satisfies

$$\mathcal{K}_\tau f \approx \mathcal{K}_\tau^{\mathrm{proj}} f \in \mathrm{span}\{\chi_1, \ldots, \chi_m\} \tag{23}$$

for any function f in the space spanned by $\boldsymbol{\chi}$. As the Koopman operator is linear, even if the dynamics are nonlinear, it can be approximated by a matrix $\mathbf{K} = (\mathbf{k}_1, \ldots, \mathbf{k}_m) \in \mathbb{R}^{m\times m}$ as

$$\mathcal{K}_\tau^{\mathrm{proj}} \left(\sum_{i=1}^{m} c_i \chi_i \right) = \sum_{i=1}^{m} c_i \mathbf{k}_i^\top \boldsymbol{\chi}, \tag{24}$$

with

$$\mathbf{k}_i^\top \boldsymbol{\chi} = \mathcal{K}_\tau^{\mathrm{proj}} \chi_i \approx \mathcal{K}_\tau \chi_i \tag{25}$$

representing a finite-dimensional approximation of $\mathcal{K}_\tau \chi_i$. After a few algebraic steps [74], it can be shown that eigenfunctions of $\mathcal{K}_\tau^{\mathrm{proj}}$ also have the form $\psi_i = \mathbf{b}_i^\top \boldsymbol{\chi}$, and eigenvalues and eigenfunctions of $\mathcal{K}_\tau^{\mathrm{proj}}$ can also be calculated by the eigenvalue problem

$$\mathbf{KB} = \mathbf{B\Lambda}, \tag{26}$$

where definitions of $\mathbf{\Lambda}, \mathbf{B}$ are the same as in (7).

A mathematically equivalent formulation of this approach was introduced in the fluid mechanics field as Dynamic Mode Decomposition (DMD) in [70, 72], and it was found to approximate the Koopman operator in [71]. Further extensions are discussed e.g. in [73]. An Extended Dynamic Mode Decomposition (EDMD) using general basis functions was described in [74].

Considering that

$$\mathbb{E}\left[\chi_i\left(\mathbf{x}_{t+\tau}\right)|\mathbf{x}_t\right]=\mathcal{K}_\tau\chi_i\left(\mathbf{x}_t\right)\approx\mathbf{k}_i^\top\boldsymbol{\chi}\left(\mathbf{x}_t\right) \tag{27}$$

for each transition pair $\left(\mathbf{x}_t,\mathbf{x}_{t+\tau}\right)$ in simulations, the matrix $\mathbf{K}$ can be determined via minimizing the mean square error between $\mathbf{k}_i^\top\boldsymbol{\chi}\left(\mathbf{x}_t\right)$ and $\chi_i\left(\mathbf{x}_{t+\tau}\right)$ as

$$\begin{aligned}\mathbf{K}&=\arg\min_{\mathbf{K}}\frac{1}{N}\sum_{t=1}^{T-\tau}\sum_{i=1}^{m}\left\|\mathbf{k}_i^\top\chi_i\left(\mathbf{x}_t\right)-\chi_i\left(\mathbf{x}_{t+\tau}\right)\right\|^2\\&=\arg\min_{\mathbf{K}}\frac{1}{N}\left\|\mathbf{XK}-\mathbf{Y}\right\|^2\\&=\hat{\mathbf{C}}\left(0\right)^{-1}\hat{\mathbf{C}}\left(\tau\right),\end{aligned} \tag{28}$$

With covariance matrices given by their direct estimates (15-16). Here $\|\cdot\|$ denotes the Frobenius norm of matrices, and the basis functions are assumed to be linearly independent on the simulation data so that $\hat{\mathbf{C}}\left(0\right)$ is invertible. In applications, the linear indpendence can be achieved by decorrelation of basis functions (see Section IV A). Thus, EDMD is algorithmically equivalent to the linear variational approach (7) with a direct estimation of the covariance matrix (15-16).

If the simulation is reversible and in equilibrium, and statistics are such that the estimate of $\hat{\mathbf{C}}\left(\tau\right)$ is symmetric, then this is also equal to the symmetrized estimation. However, for off-equilibrium data the difference of the empirical covariance matrices $\hat{\mathbf{C}}_{\mathrm{sym}}\left(0\right)$ and $\hat{\mathbf{C}}_{\mathrm{sym}}\left(\tau\right)$ to the true expectations is large and in this case the symmetric estimator involves a large bias. In contrast, suppose that the ensemble of $\left\{\mathbf{x}_1,\ldots,\mathbf{x}_{T-\tau}\right\}$ follows a probability distribution $\rho\left(\mathbf{x}\right)$, then $\hat{\mathbf{C}}\left(0\right)$ and $\hat{\mathbf{C}}\left(\tau\right)$ are unbiased estimates of non-equilibrium correlation matrices $\mathbb{E}_\rho\left[\boldsymbol{\chi}\left(\mathbf{x}_t\right)\boldsymbol{\chi}\left(\mathbf{x}_t\right)^\top\right]$ and $\mathbb{E}_\rho\left[\boldsymbol{\chi}\left(\mathbf{x}_t\right)\boldsymbol{\chi}\left(\mathbf{x}_{t+\tau}\right)^\top\right]$ instead of $\mathbf{C}\left(0\right)$ and $\mathbf{C}\left(\tau\right)$, and the matrix $\mathbf{K}$ given by (28) minimizes the error (see Appendix B)

$$\sum_i\left\langle\mathbf{k}_i^\top\boldsymbol{\chi}-\mathcal{K}_\tau\chi_i,\mathbf{k}_i^\top\boldsymbol{\chi}-\mathcal{K}_\tau\chi_i\right\rangle_\rho, \tag{29}$$

where $\left\langle\cdot,\cdot\right\rangle_\rho$ denotes the the inner product defined by $\left\langle f,g\right\rangle_\rho=\int\mathrm{d}\mathbf{x}\,\rho\left(\mathbf{x}\right)f\left(\mathbf{x}\right)g\left(\mathbf{x}\right)$. Therefore, $\mathbf{K}$ is still a finite-dimensional approximation of $\mathcal{K}_\tau$ even if $\rho\neq\mu$ because of the non-equilibrium of simulation data, which implies that EDMD is applicable to non-equilibrium data without the assumption of reversibility and is equivalent to the direct VAC estimate.

At this point, EDMD and the VAC with nonreversible covariance matrix estimate are algorithmically identical and only differ by the way they were derived - see [77] for a mathematical analysis. However, we can use DMD theory in order to go further and formulate an unbiased reversible estimator.

C. Estimation of the equilibrium distribution

Not only is EDMD robust when using non-equilibrium data, we can also utilize the Koopman matrix $\mathbf{K}$ to recover the equilibrium properties of the molecular system. The principle of importance sampling [78] states that the equilibrium ensemble average of an observable $f\left(\mathbf{x}_t\right)$ can be unbiasedly estimated by the weighted mean

$$\mathbb{E}_\mu\left[f\left(\mathbf{x}_t\right)\right] \approx \frac{1}{N}\sum_{t=1}^{T-\tau}\frac{\mu\left(\mathbf{x}_t\right)}{\rho\left(\mathbf{x}_t\right)}f\left(\mathbf{x}_t\right), \tag{30}$$

As analytical expressions of μ and ρ are generally unavailable, we approximate the ratio between them by a linear combination of basis functions $\boldsymbol{\chi}$ as

$$\frac{\mu\left(\mathbf{x}\right)}{\rho\left(\mathbf{x}\right)} \approx \mathbf{u}^\top\boldsymbol{\chi}\left(\mathbf{x}\right) \tag{31}$$

From the invariance condition $\mathbb{E}_\mu\left[\boldsymbol{\chi}\left(\mathbf{x}_{t+\tau}\right)\right] = \mathbb{E}_\mu\left[\boldsymbol{\chi}\left(\mathbf{x}_t\right)\right]$ and the normalization condition $\int \mathrm{d}\mathbf{x}\ \mu\left(\mathbf{x}\right) = 1$ on the stationary distribution μ, we can show following algebraic constraints on $\mathbf{u}$:

$$\mathbf{u}^\top\hat{\mathbf{C}}\left(0\right)\mathbf{K}\hat{\mathbf{C}}\left(0\right)^{-1} = \mathbf{u}^\top, \tag{32}$$

$$\mathbf{u}^\top\hat{\mathbf{C}}\left(0\right)\mathbf{v} = 1 \tag{33}$$

in the limit of large statistics (see Appendix C for proof). Here, $\mathbf{v}$ is the vector that combines the basis functions to represent the constant $\mathbb{1}$ function, i.e.

$$\mathbf{v}^\top\boldsymbol{\chi} = \mathbb{1}, \tag{34}$$

and as shown in the algorithms below, we can use the following trick to end up with a known $\mathbf{v}$: (1) “Whiten” the data by orthogonalizing it, and then normalizing each data column to have a variance of 1, (2) add the constant function to the basis set by adding a column of 1's to the input data matrices $\mathbf{X}$ and $\mathbf{Y}$.

Thus, we can compute a vector proportional to $\mathbf{u}$ as the left eigenvector of $\hat{\mathbf{C}}(0)\mathbf{K}\hat{\mathbf{C}}(0)^{-1}$ with eigenvalue 1 and normalize it by dividing by $\mathbf{u}^\top\hat{\mathbf{C}}(0)\mathbf{v}$ in order to satisfy (33).

Besides equilibrium ensemble averages in the form of (30), we can also approximate time-lagged cross correlations between observable quantities at equilibrium. For two observables $f_1 = \mathbf{c}_1^\top\boldsymbol{\chi}$ and $f_2 = \mathbf{c}_2^\top\boldsymbol{\chi}$ in $\mathrm{span}\{\chi_1, \ldots, \chi_m\}$, we have

$$\mathbb{E}_\mu\left[f_1(\mathbf{x}_t) f_2(\mathbf{x}_{t+k\tau})\right] = \mathbb{E}_\mu\left[f_1(\mathbf{x}_t)\cdot\mathcal{K}_\tau^k f_2(\mathbf{x}_t)\right] \approx \mathbf{c}_1^\top\hat{\mathbf{C}}_{\mathrm{eq}}(0)\mathbf{K}^k\mathbf{c}_2. \tag{35}$$

Here,

$$\hat{\mathbf{C}}_{\mathrm{eq}}(0) = \frac{1}{N}\sum_{t=1}^{T-\tau}\left(\mathbf{u}^\top\boldsymbol{\chi}(\mathbf{x}_t)\right)\boldsymbol{\chi}(\mathbf{x}_t)\boldsymbol{\chi}(\mathbf{x}_t)^\top \tag{36}$$

$$= \frac{1}{N}\mathbf{X}^\top\mathrm{diag}(\mathbf{X}\mathbf{u})\,\mathbf{X} \tag{37}$$

is the estimate of $\mathbf{C}(0) = \mathbb{E}_\mu\left[\boldsymbol{\chi}(\mathbf{x}_t)\boldsymbol{\chi}(\mathbf{x}_t)^\top\right]$ given by the reweighting and $\hat{\mathbf{C}}_{\mathrm{eq}}(0)\mathbf{K}^k$ is the corresponding estimate of time-lagged correlation matrix $\mathbf{C}(k\tau)$.

D. Reversible EDMD

If $\{\mathbf{x}_t\}$ satisfies the detailed balance condition (3), the time-lagged cross correlation between two arbitrary observable quantities $f_1(\mathbf{x}_t)$ and $f_2(\mathbf{x}_t)$ at equilibrium is symmetric in the sense of $\mathbb{E}_\mu\left[f_1(\mathbf{x}_t) f_2(\mathbf{x}_{t+k\tau})\right] = \mathbb{E}_\mu\left[f_2(\mathbf{x}_t) f_1(\mathbf{x}_{t+k\tau})\right]$ and $\mathbf{C}(k\tau)$ is a symmetric matrix. Therefore, we can symmetrize the EDMD estimate of $\mathbf{C}(\tau)$ as

$$\hat{\mathbf{C}}_{\mathrm{eq}}(\tau) \approx \frac{1}{2}\left(\hat{\mathbf{C}}_{\mathrm{eq}}(0)\mathbf{K} + \mathbf{K}^\top\hat{\mathbf{C}}_{\mathrm{eq}}(0)\right) \tag{38}$$

and modify the matrix $\mathbf{K}$ as

$$\tilde{\mathbf{K}} = \frac{1}{2}\hat{\mathbf{C}}_{\mathrm{eq}}(0)^{-1}\left(\hat{\mathbf{C}}_{\mathrm{eq}}(0)\mathbf{K} + \mathbf{K}^\top\hat{\mathbf{C}}_{\mathrm{eq}}(0)\right) \approx \mathbf{C}(0)^{-1}\mathbf{C}(\tau). \tag{39}$$

In the case of reversible dynamics, the reversible EDMD given by (39) may be desirable because it yields real-valued spectral components even in the existence of statistical noise and modeling error. In addition, it can be shown that (32) holds after replacing $\mathbf{K}$ by $\tilde{\mathbf{K}}$, i.e., the reweighting vector $\mathbf{u}$ remains fixed for the reversible EDMD. (See Appendix D for more detailed analysis.) Unlike the symmetric estimation for VAC, the symmetrization in (38) does not affect the unbiasedness of the estimate.

IV. ALGORITHMS

A. Decorrelation of basis functions

In Section III B, the basis functions $\boldsymbol{\chi}$ are assumed to be linearly independent on the sampled data so that $\hat{\mathbf{C}}(0)$ is invertible and the matrix $\mathbf{K}$ given in (28) is well defined. In some publications, e.g. [74], $\mathbf{K}$ is calculated as $\mathbf{K} = \hat{\mathbf{C}}(0)^{\dagger}\hat{\mathbf{C}}(\tau)$ by using the pseudoinverse $\hat{\mathbf{C}}(0)^{\dagger}$ of $\hat{\mathbf{C}}(0)$, however this approach cannot completely avoid numerical instabilities. In this paper, we utilize principal component analysis (PCA) [79] to explicitly reduce correlations between basis functions as

$$\boldsymbol{\chi} = \begin{pmatrix} \mathrm{PCA}\left[\boldsymbol{\chi}_o | \rho\right] \\ \mathbb{1} \end{pmatrix}. \tag{40}$$

Here, $\boldsymbol{\chi}_o$ denotes the original basis functions which may be linearly dependent, $\mathrm{PCA}\left[\boldsymbol{\chi}_o|\rho\right]$ denotes the PCA whitening transformation of the original basis functions $\boldsymbol{\chi}_o$ according to the empirical distribution ρ of $(\mathbf{x}_1, ..., \mathbf{x}_{T-\tau})$. Whitening means: (i) transform the data into all available principal components and (ii) scale coordinates to have a variance of 1. The dimension of $\mathrm{PCA}\left[\boldsymbol{\chi}_o|\rho\right]$ is equal to the number of positive eigenvalues of the covariance matrix of $\boldsymbol{\chi}_o$ which is larger than a small numerical cutoff $\epsilon_0 > 0$ (see Appendix E for the implementation details of the PCA transformation). The last basis function is set to be $\mathbb{1}$ in (40) so that $\mathbf{v}^{\top}\boldsymbol{\chi} = \mathbb{1}$ with $\mathbf{v} = (0, \dots, 0, 1)^{\top}$.

Similarly, the estimate $\hat{\mathbf{C}}_{\mathrm{eq}}(0)$ of the equilibrium correlation matrix $\mathbf{C}(0)$ given by (36) may yield numerical singularities for reversible EDMD estimation if it is not positive definite (see (39) and Appendix D). In order to overcome this problem, we can further decorrelate basis functions $\boldsymbol{\chi}$ according to the estimated equilibrium distribution $\mu(\mathbf{x}) = \mathbf{u}^{\top}\boldsymbol{\chi}(\mathbf{x}) \cdot \rho(\mathbf{x})$ to get a set of new basis functions $\boldsymbol{\chi}_s$ as

$$\boldsymbol{\chi}_s = \begin{pmatrix} \mathrm{PCA}\left[\boldsymbol{\chi} | \mu\right] \\ \mathbb{1} \end{pmatrix}. \tag{41}$$

It can be easily verified that the equilibrium correlation matrix of $\boldsymbol{\chi}_s$ at lag time zero is an identity matrix, so the Koopman matrix within the subspace of $\boldsymbol{\chi}_s$ is

$$\begin{aligned}\mathbf{K}_s &= \mathbb{E}_{\mu}\left[\boldsymbol{\chi}_s(\mathbf{x}_t)\,\boldsymbol{\chi}_s(\mathbf{x}_{t+\tau})^{\top}\right] \\ &= \frac{1}{N}\sum_{t=1}^{T-\tau}\left(\mathbf{u}^{\top}\boldsymbol{\chi}(\mathbf{x}_t)\right)\boldsymbol{\chi}_s(\mathbf{x}_t)\,\boldsymbol{\chi}_s(\mathbf{x}_t)^{\top}\end{aligned} \tag{42}$$

and the corresponding reversible estimation is given by

$$\tilde{\mathbf{K}}_s = \frac{1}{2}\left(\mathbf{K}_s + \mathbf{K}_s^\top\right). \tag{43}$$

The relationships between $\boldsymbol{\chi}_o$, $\boldsymbol{\chi}$, $\boldsymbol{\chi}_s$ and $\mathbf{K}$, $\mathbf{K}_s$ can be briefly summarized as follows: $\boldsymbol{\chi}$ is a linearly independent basis of $\boldsymbol{\chi}_o$ based on the empirical distribution ρ and $\boldsymbol{\chi}_s$ forms a basis of $\boldsymbol{\chi}$ based on the equilibrium distribution μ. $\mathbf{K}$ and $\mathbf{K}_s$ are approximations of the Koopman operator with respect to $\boldsymbol{\chi}$ and $\boldsymbol{\chi}_s$, and they yield the equivalent approximations if the matrix $\hat{\mathbf{C}}_{\mathrm{eq}}(0)$ is positive definite. In practice, $\mathbf{K}$ can be used for estimation of spectral components and equilibrium distributions without the constraint of reversibility, whereas $\mathbf{K}_s$ can achieve reversible estimates in a numerically stable way.

B. Algorithms

Based on all the above discussions, a general analysis procedure for MD data with given conformational basis functions $\boldsymbol{\chi}_o$ can be summarized by the following algorithms:

Algorithm 1: Nonreversible VAC / TICA

1. Perform the decorrelation (40) to obtain a set of linearly independent basis functions $\boldsymbol{\chi}$.

2. Compute the matrix $\mathbf{K}$ by (28) and solve the eigenvalue problem $\mathbf{KB} = \mathbf{B\Lambda}$.

3. Output spectral components: Eigenvalues $\hat{\lambda}_i$ and eigenfunctions $\hat{\psi}_i$. Both may have imaginary components that are either due to statistical noise or due to real nonreversible processes if the true dynamics are nonreversible.

Algorithm 2: Estimation of equilibrium properties

1. Compute $\mathbf{K}$ as in Algorithm 1.

2. Compute $\mathbf{u}$ as a left eigenvector of $\mathbf{K}$ satisfying $\mathbf{u}^\top\mathbf{K} = \mathbf{u}^\top$ and $\mathbf{u}^\top\mathbf{v} = 1$, where $\mathbf{v} = (0, \ldots, 0, 1)^\top$. (Note that $\hat{\mathbf{C}}(0)$ is an identity matrix after the decorrelation in Step 1.)

3. Compute the matrix $\hat{\mathbf{C}}_{\mathrm{eq}}(0)$ by (36) as the unbiased estimate of $\mathbf{C}(0)$.

4. Output:

(a) Equilibrium expectations: $\mathbb{E}_{\mu}\left[f\left(\mathbf{x}_{t}\right)\right]=\frac{1}{N}\sum_{t=1}^{T-\tau}\left(\mathbf{u}^{\top}\boldsymbol{\chi}\left(\mathbf{x}_{t}\right)\right)f\left(\mathbf{x}_{t}\right)$ for a given observable f.

(b) Equilibrium time-lagged correlations: $\mathbb{E}_{\mu}\left[f_{1}\left(\mathbf{x}_{t}\right)f_{2}\left(\mathbf{x}_{t+k\tau}\right)\right]=\mathbf{c}_{1}^{\top}\hat{\mathbf{C}}_{\mathrm{eq}}\left(0\right)\mathbf{K}^{k}\mathbf{c}_{2}$ for $f_{1}=\mathbf{c}_{1}^{\top}\boldsymbol{\chi}$ and $f_{2}=\mathbf{c}_{2}^{\top}\boldsymbol{\chi}$.

Algorithm 3: Reversible VAC / TICA

1. Compute $\mathbf{K}$ and $\hat{\mathbf{C}}_{\mathrm{eq}}\left(0\right)$ as in Algorithms 1 and 2.

2. Perform the decorrelation of $\boldsymbol{\chi}$ by (41) according to the equilibrium distribution to get basis functions $\boldsymbol{\chi}_{s}$.

3. Compute $\mathbf{K}_{s}$ by (42).

4. Perform the reversibility modification $\tilde{\mathbf{K}}_{s}=\frac{1}{2}\left(\mathbf{K}_{s}+\mathbf{K}_{s}^{\top}\right)$ and solve the eigenvalue problem $\tilde{\mathbf{K}}_{s}\mathbf{B}_{s}=\mathbf{B}_{s}\mathbf{\Lambda}_{s}$ of $\tilde{\mathbf{K}}_{s}$.

5. Output spectral components: Eigenvalues $\hat{\lambda}_{i}$ and eigenfunctions $\hat{\psi}_{i}$. These eigenvalues and eigenfunctions are real-valued. The dimensionality of the data can be trivially reduced by discarding the eigenfunctions with small eigenvalues.

V. APPLICATIONS

In this section, we apply three different estimators of VAC (or TICA) for spectral estimation to the same data sets: the symmetric estimator with symmetrization of time-lagged correlation matrices, the direct estimator which is also equivalent to the estimator derived by EDMD, and the reversible estimator proposed in Section 39. In addition, we compare the estimated equilibrium distribution provided by the direct estimator and that calculated by histogram counting in order to demonstrate the validity of the proposed reweighting method.

A. One-dimensional diffusion process

As a first example, we consider a one-dimensional diffusion process $\{x_t\}$ in a double-well potential with phase space $[0,2]$ as shown in Fig. 1A. In order to validate the robustness of different estimators, we start all simulations far from equilibrium, in the region $[0,0.2]$

(Fig. 1C). The set of basis functions for estimators is constructed by using 100 Gaussian functions with random parameters. For more details on the simulation model and experimental setup, see Appendix F 1.

Fig. 1B shows estimates of the slowest relaxation timescale ITS_2 based on 500 independent short simulation trajectories with length 0.2 time units. The largest relaxation timescale t_2 is computed from λ_2 as $t_2 = -\tau/\ln|\lambda_2(\tau)|$ and is a constant independent of lag time according to (4). For such non-equilibrium data, the symmetric estimator significantly underestimates the relaxation timescale for such non-equilibrium data and gives even worse results with longer lag times. The direct and reversible estimators, on the other hand, converge quickly to the true timescale before $\tau = 0.01$. The equilibrium distribution density of $\{x_t\}$ computed from Algorithm 2 with lag time 0.01 is shown in Fig. 1C. In contrast to the empirical histogram density given by direct counting, the direct estimator effectively recovers the equilibrium property of the process from non-equilibrium data.

Fig. 1D summarizes the empirical probability of the potential well I and the estimate given by the direct estimator with different simulation trajectory lengths, where the lag time for EDMD is still 0.01 and the accumulated simulation time is kept fixed to be 100. Due to the ergodicity of the process, the empirical probability converges to the true value as the trajectory length increases. The convergence rate, however, is very slow as shown in Fig. 1D, and empirical probability is close to the true value only for trajectories longer than 2. When using the reweighting method proposed here, the estimated probability is robust with respect to changes in trajectory length, and unbiased even for very short trajectories.

B. Two-dimensional diffusion process

We now discuss an example of a two-dimensional diffusion process $\{(x_t, y_t)\}$ which has three potential wells as shown in Fig. 2A, where all simulations are initialized with $(x_0, y_0) \in [-2, -1.5] \times [-1.5, 2.5]$, and the set of basis functions for spectral estimation consists of 100 Gaussian functions with random parameters (see Appendix F 2 for details).

We generate 8000 short simulation trajectories with length 1.25 and show the empirical free energy of the simulation data in Fig. 2B. Comparing Fig. 2B and Fig. 2A, it can be seen that most of the simulation data are distributed in the area $x \leq 0$ and simulations are very far away from the equilibrium state. Therefore, the symmetric estimator cannot capture

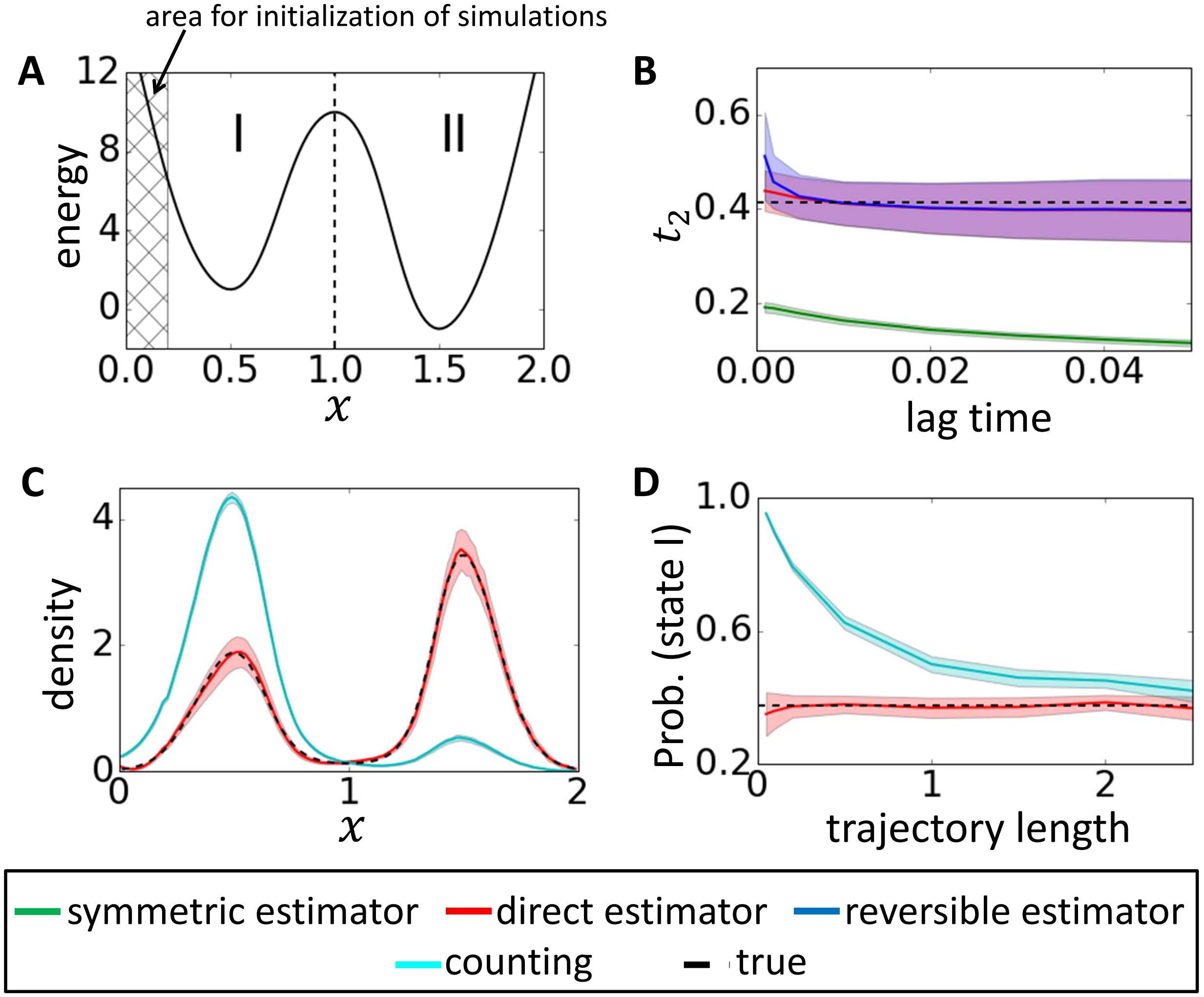

Figure 1. Estimation results of a one-dimensional diffusion process. (A) Dimensionless energy $U(x)$, where the dashed line represents the border of the two potential wells I and II. The shaded area denotes the region where initial states are drawn for simulations. (B) The slowest relaxation timescale estimated by the previously used symmetric estimator, the direct estimator and the present reversible estimator with different lag times. (C) Stationary density of states obtained from equilibrium probabilities of 100 uniform bins, where the probabilities are estimated from the direct estimator and direct counting. (D) Estimates of the equilibrium probability of the potential well I given by the direct estimator and direct counting with different simulation trajectory lengths. In (B-D), solid lines and shaded regions indicate mean values and one standard derivation error intervals obtained from 30 independent experiments.

the spectral components of the process as illustrated in Fig. 2D, whereas the direct and the present reversible estimator can still provide accurate eigenvalues and the equilibrium density (see Figs. 2C and 2D). Note that the two slowest relaxation timescales plotted in Fig. 2D are computed from λ_2 and λ_3, respectively.

For such a two-dimensional process, it is also interesting to investigate the slowest modes predicted by TICA. Fig. 2A displays the slowest modes computed from the exact equilibrium distribution with lag time $\tau = 0.01$. Notice that the slowest mode is parallel to x-axis, which is related to transitions between potential wells I and II, and the second IC is parallel to the y-axis, which is related to transitions between {I,II} and III. However, if we extract ICs from simulation data by using the previous symmetric estimator, the result is significantly different as shown in Fig. 2B, where the first IC characterizes transitions between I and III. The ICs given by the direct and reversible estimators suggested in this work can be seen in Fig. 2C. They are still different from those in Fig. 2A because the equilibrium distribution is difficult to approximate with only linear basis functions, but much more accurate than the estimates obtained by the previously used symmetric estimator in Fig. 2B.

Fig. 2E summarizes the estimation errors of estimated equilibrium distribution obtained by using simulations with different trajectory lengths, where the accumulated simulation time is kept fixed to be 10^4, the lag time for estimators is $\tau = 0.005$, and the error is evaluated as the total variation distance between the estimated probability distributions of the three potential wells and the true reference. Fig. 2F shows angles of linear ICs approximated from the same simulation data with lag time $\tau = 0.01$. Both of the figures clearly demonstrate the superiority of the direct and reversible estimators suggested here.

C. Protein-Ligand Binding

We revisit the the binding process of benzamidine to trypsin which was studied previously in Refs. [11, 76]. The data set consists of 52 trajectories of 2μs and four trajectories of 1 μs simulation time, resulting in a total simulation time of 108 μs. From the simulations, we extract a feature set of 223 nearest neighbor heavy-atom contacts between all trypsin residues and the ligand. We then perform TICA using the symmetrized estimate (previous standard), the direct estimate and the reversible estimate discussed in Sec. III D. In order to obtain uncertainties, we compute 100 bootstrapping estimates. In Figure 3 A-C, we show the three

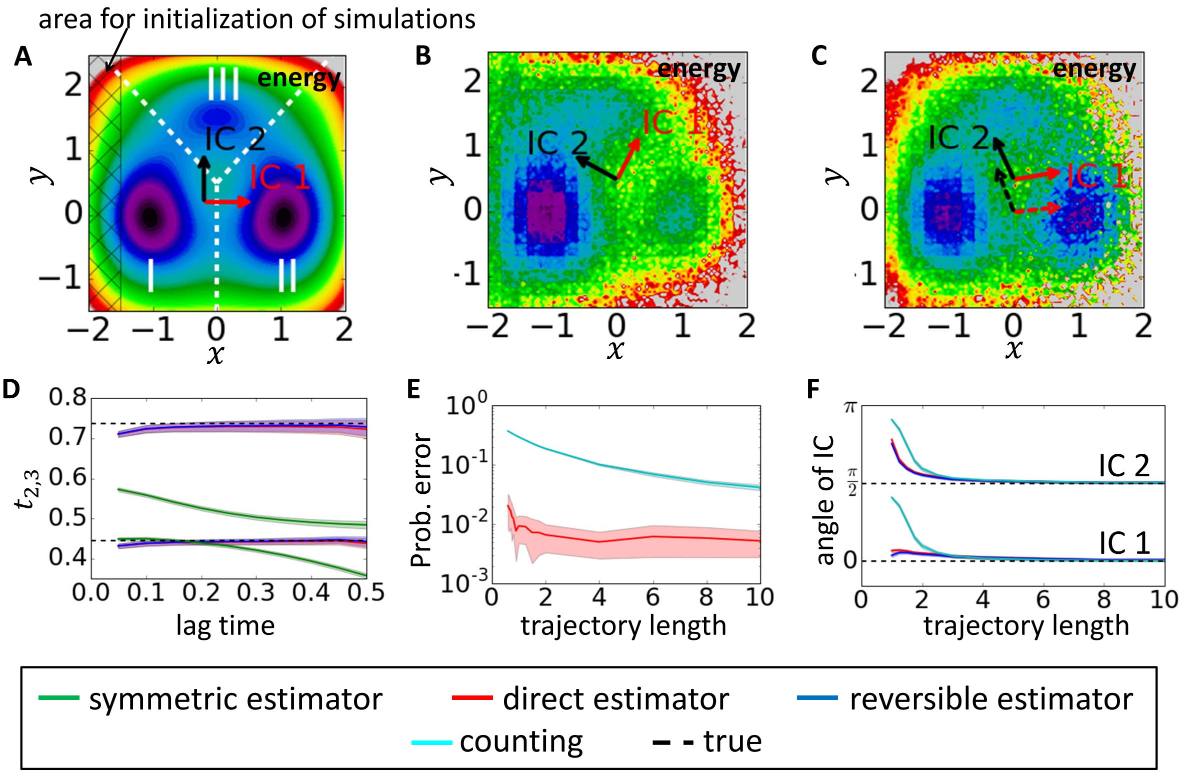

Figure 2. Estimation results of a two-dimensional diffusion process. (A) Free energy of the process, where the dashed line represents the border of potential wells I, II, and III. The shaded area denotes the region where initial states are drawn for simulations, and the two linear ICs obtained from TICA with exact statistics. (B, C) Free energies computed from the equilibrium density estimated via direct counting and the direct estimator. Solid arrows in C and D indicate directions of estimated ICs given by the symmetric and direct estimators respectively, and dashed arrows indicate that given by the reversible estimator. (D) Estimates of the two slowest relaxation timescales. (E) Estimation errors of equilibrium distributions. (F) Angles of estimated ICs. In (D–F), solid lines and shaded regions indicate mean values and one standard derivation error intervals obtained from 30 independent experiments.

slowest implied timescales as estimated by the three approaches discussed above. We observe that both the direct and the reversible estimator provide a larger slowest implied timescale than the symmetric estimator. The slowest timescale estimated by the reversible estimator only converges with increasing lag time if extremely high estimates from the bootstrapping are discarded. This instability is likely due to the simple choice of the basis function used here – it is known that the trypsin-benzamidin binding kinetics involves internal conformational changes of trypsin [16]. In Fig. 3D–F, we display the projection of the data onto the first two TICA components for all three estimates (the first TICA components of the direct estimate are coincidentally purely real here). The eigenvectors used for the dimensionality reduction were estimated at lag time $\tau = 100\,\mathrm{ns}$. The projections are qualitatively similar, revealing three minima of the landscape, labeled 1, 2, and 3. In all three cases, these centers correspond to the same macro-states of the system, shown underneath in Figure 3 G-H. Center 1 corresponds to the ligand being either unbound or loosely attached to the protein. The other two states are different conformational arrangements of the bound state of the ligand.

VI. CONCLUSION

Using dynamic mode decomposition theory, we have shown that the variational approach of conformation dynamics and the time-lagged independent component analysis can be made without bias even if just empirical out-of-equilibrium estimates of the covariance matrices are available, i.e. they can be applied to ensembles of short MD simulations starting from arbitrary starting point. The crucial point is that the forceful symmetrization of the empirical covariances practiced in previous studies must be avoided.

In order to facilitate an unbiased symmetric estimate of covariance matrices, we have proposed a reweighting technique in which the weights of sampled configurations can be estimated using a first pass of VAC/TICA, and be applied in order to turn the empirical (out-of-equilibrium) estimates of covariance matrices into estimates of the equilibrium covariance matrices. These matrices can then be symmetrized without introducing a bias from the empirical distribution, resulting in real-valued eigenvalue and eigenfunction estimates.

With these algorithms, VAC and TICA inherit the same benefits that MSMs have enjoyed since nearly a decade: we can generate optimal and unbiased reversible and nonreversible estimate from either long equilibrium trajectories or swarms of short trajectories not started from equilibrium.

An additional result shown in this paper is the computation of the reweighting factors of sampled configurations that turn the biased empirical distribution into an unbiased estimate of the equilibrium distribution. This provides a conceptual novelty: We can compute variationally optimal estimates of equilibrium properties (expectation values, distributions) from

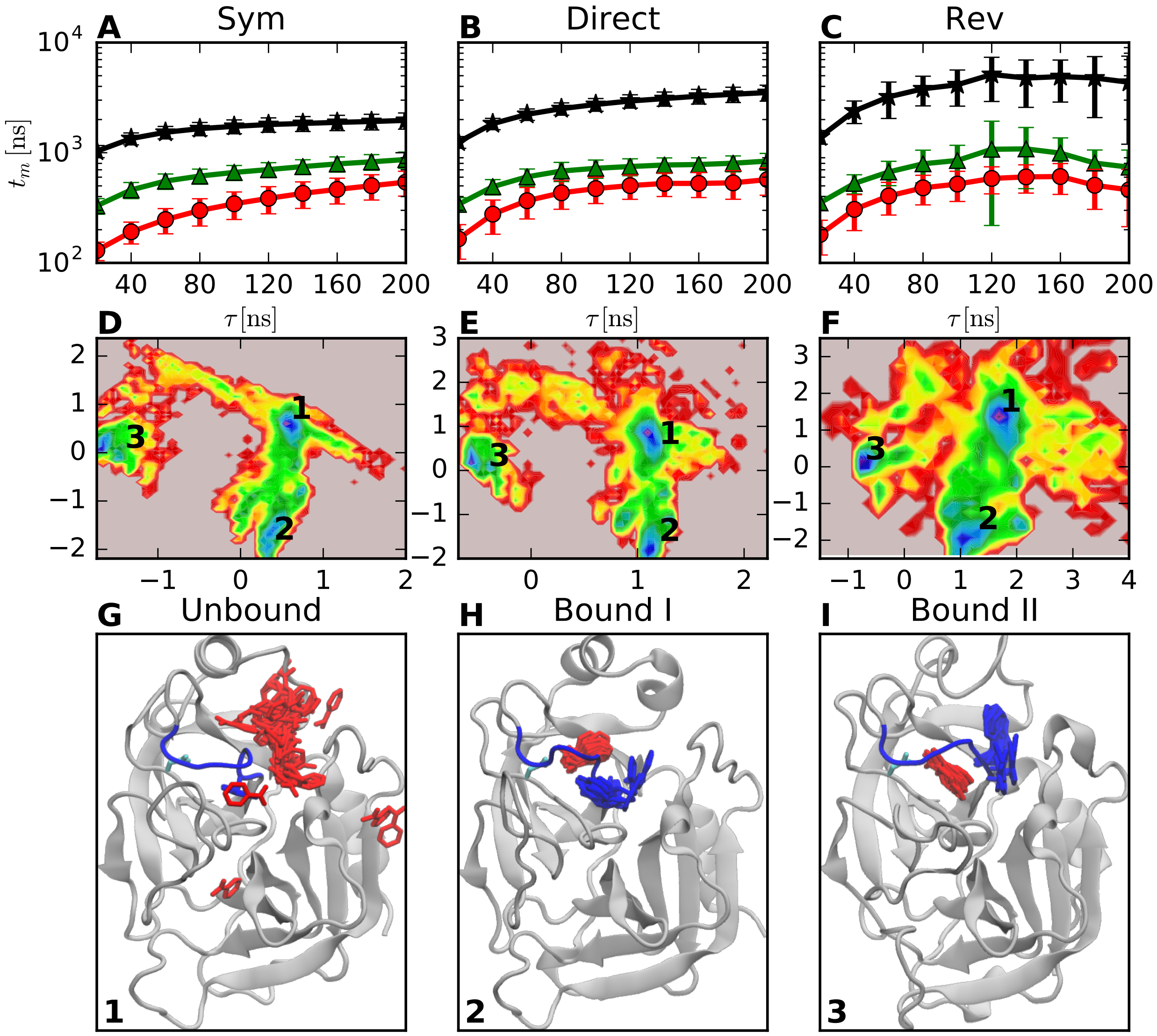

Figure 3. Results for MD simulations of the trypsin-benzamidine binding process. A–C: Three slowest implied timescales as a function of the lag time estimated by the previously used symmetric estimator (A), the direct estimator (B) and the reversible estimator suggested here (C). D–F: Projections of the data into the two-dimensional space of slowest dynamical eigenvectors for the three estimation methods. In all cases, we can discern three minima of the landscape, labeled 1–3. For all three methods, minima 1-3 correspond to the same macro-states of the system. Representative structures of these states are shown in G-I. State 1 represents the ligand being unbound or loosely attached to the protein. States 2 and 3 are different conformational arrangements of the bound state, in particular of the binding loop including Trp 215 [16].

out-of-equilibrium data using an approach that involves arbitrary sets of basis functions and computing covariance matrices between them. However, the viability of this approach critically depends on the suitability of the basis functions employed, and this aspect will be investigated in future studies.

Appendix A: Dynamical operators

Besides the Koopman operator $\mathcal{K}_\tau$, the conformation dynamics of a molecular system can also be described by the forward operator $\mathcal{P}_\tau$ and transfer operator, or called backward operator, $\mathcal{T}_\tau$ [23], which describe the evolution of ensemble densities as

$$\begin{aligned} p_{t+\tau}\left(\mathbf{x}\right) &= \mathcal{P}_\tau p_t\left(\mathbf{x}\right) \\ &= \int \mathrm{d}\mathbf{y}\ p\left(\mathbf{y},\mathbf{x};\tau\right) p_t\left(\mathbf{y}\right) \end{aligned} \tag{A1}$$

and

$$\begin{aligned} u_{t+\tau}\left(\mathbf{x}\right) &= \mathcal{T}_\tau u_t\left(\mathbf{x}\right) \\ &= \int \mathrm{d}\mathbf{y}\ \frac{\mu\left(\mathbf{y}\right)}{\mu\left(\mathbf{x}\right)} p\left(\mathbf{y},\mathbf{x};\tau\right) u_t\left(\mathbf{y}\right), \end{aligned} \tag{A2}$$

where $p_t\left(\mathbf{x}\right)$ denotes the probability density of $\mathbf{x}_t$ and $u_t\left(\mathbf{x}\right) = \mu\left(\mathbf{x}\right)^{-1} p_t\left(\mathbf{x}\right)$ denotes the density weighted by the inverse of the stationary density. The relationship between the three operators can be summarized as follows:

1. $\mathcal{K}_\tau$ is adjoint to $\mathcal{T}_\tau$ in the sense of

$$\left\langle \mathcal{K}_\tau f_1, f_2 \right\rangle_\mu = \left\langle f_1, \mathcal{T}_\tau f_2 \right\rangle_\mu \tag{A3}$$

for any $f_1, f_2 \in L^2_\mu$. If $\{\mathbf{x}_t\}$ is reversible, $\mathcal{K}_\tau$ and $\mathcal{T}_\tau$ are self-adjoint with respect to $\left\langle \cdot, \cdot \right\rangle_\mu$, i.e., $\mathcal{K}_\tau = \mathcal{T}_\tau$.

2. Defining the multiplication operator $\mathcal{M}_\mu : L^2_\mu \mapsto L^2_{\mu^{-1}}$ as $\mathcal{M}_\mu f\left(\mathbf{x}\right) = \mu\left(\mathbf{x}\right) \cdot f\left(\mathbf{x}\right)$, the Markov propagator $\mathcal{P}_\tau$ can be expressed as

$$\mathcal{P}_\tau = \mathcal{M}_\mu \mathcal{T}_\tau \mathcal{M}_\mu^{-1}. \tag{A4}$$

Under the detailed balance condition, $\mathcal{P}_\tau$ is self-adjoint with respect to $\left\langle \cdot, \cdot \right\rangle_{\mu^{-1}}$.

We can also find the finite-dimensional approximation $\mathcal{P}_\tau\chi_i \approx \mathbf{p}_i^\top \boldsymbol{\chi}$ and $\mathcal{T}_\tau\chi_i \approx \mathbf{t}_i^\top \boldsymbol{\chi}$ of $\mathcal{P}_\tau$ and $\mathcal{T}_\tau$ by minimizing errors $\sum_i \left\langle \mathbf{p}_i^\top \boldsymbol{\chi} - \mathcal{P}_\tau\chi_i, \mathbf{p}_i^\top \boldsymbol{\chi} - \mathcal{P}_\tau\chi_i \right\rangle_w$ and $\sum_i \left\langle \mathbf{t}_i^\top \boldsymbol{\chi} - \mathcal{T}_\tau\chi_i, \mathbf{t}_i^\top \boldsymbol{\chi} - \mathcal{T}_\tau\chi_i \right\rangle_w$ for some weight function $w(\mathbf{x})$. However, it is still unknown how to select the weight functions so that the approximation errors can be easily computed from simulation data as in the approximation of $\mathcal{K}_\tau$. For example, if we select $w(\mathbf{x}) = \rho(\mathbf{x})^{-1}$, the approximation error of $\mathcal{P}_\tau$ is

$$\begin{aligned}\sum_i \left\langle \mathbf{p}_i^\top \boldsymbol{\chi} - \mathcal{P}_\tau\chi_i, \mathbf{p}_i^\top \boldsymbol{\chi} - \mathcal{P}_\tau\chi_i \right\rangle_{\rho^{-1}} &= \sum_i \left\langle \mathbf{p}_i^\top \boldsymbol{\chi}, \mathbf{p}_i^\top \boldsymbol{\chi} \right\rangle_{\rho^{-1}} - 2\sum_i \left\langle \mathbf{p}_i^\top \boldsymbol{\chi}, \mathcal{P}_\tau\chi_i \right\rangle_{\rho^{-1}} \\ &\quad + \sum_i \left\langle \mathcal{P}_\tau\chi_i, \mathcal{P}_\tau\chi_i \right\rangle_{\rho^{-1}} \\ &= \sum_i \mathbb{E}_\rho \left[\frac{\mathbf{p}_i^\top \boldsymbol{\chi}(\mathbf{x}_t)\boldsymbol{\chi}(\mathbf{x}_t)^\top \mathbf{p}_i}{\rho(\mathbf{x}_t)^2} \right] \\ &\quad -2\sum_i \mathbb{E}_\rho \left[\frac{\mathbf{p}_i^\top \boldsymbol{\chi}(\mathbf{x}_{t+\tau})\chi_i(\mathbf{x}_t)}{\rho(\mathbf{x}_{t+\tau})\rho(\mathbf{x}_t)} \right] \\ &\quad + \sum_i \left\langle \mathcal{P}_\tau\chi_i, \mathcal{P}_\tau\chi_i \right\rangle_{\rho^{-1}} \end{aligned} \quad \text{(A5)}$$

where the last term on the right hand side is a constant independent of $\mathbf{p}_i$, and the computation of the first two terms is infeasible for unknown ρ. For $\mathcal{T}_\tau$, the weight function is generally set to be $w = \rho$, and the corresponding approximation error is then

$$\begin{aligned}\sum_i \left\langle \mathbf{t}_i^\top \boldsymbol{\chi} - \mathcal{T}_\tau\chi_i, \mathbf{t}_i^\top \boldsymbol{\chi} - \mathcal{T}_\tau\chi_i \right\rangle_{\rho} &= \sum_i \left\langle \mathbf{t}_i^\top \boldsymbol{\chi}, \mathbf{t}_i^\top \boldsymbol{\chi} \right\rangle_{\rho} - 2\sum_i \left\langle \mathbf{t}_i^\top \boldsymbol{\chi}, \mathcal{T}_\tau\chi_i \right\rangle_{\rho} \\ &\quad + \sum_i \left\langle \mathcal{T}_\tau\chi_i, \mathcal{T}_\tau\chi_i \right\rangle_{\rho} \\ &= \sum_i \mathbb{E}_\rho \left[\mathbf{t}_i^\top \boldsymbol{\chi}(\mathbf{x}_t)\boldsymbol{\chi}(\mathbf{x}_t)^\top \mathbf{t}_i \right] \\ &\quad -2\sum_i \mathbb{E}_\rho \left[\frac{\rho(\mathbf{x}_{t+\tau})\mu(\mathbf{x}_t)}{\mu(\mathbf{x}_{t+\tau})\rho(\mathbf{x}_t)} \cdot \mathbf{t}_i^\top \boldsymbol{\chi}(\mathbf{x}_{t+\tau})\chi_i(\mathbf{x}_t) \right] \\ &\quad + \sum_i \left\langle \mathcal{T}_\tau\chi_i, \mathcal{T}_\tau\chi_i \right\rangle_{\rho} \end{aligned} \quad \text{(A6)}$$

which is difficult to estimate unless the empirical distribution ρ is consistent with μ or the system is reversible. (For reversible systems, $\mathcal{K}_\tau = \mathcal{T}_\tau$ and the finite-dimensional approximation of $\mathcal{K}_\tau$ is therefore also that of $\mathcal{T}_\tau$.) In general cases, only the Koopman operator can be reliably estimated from the non-equilibrium data.

Appendix B: Limit of the EDMD approximation error

The mean square error of the EDMD approximation is

$$\mathrm{MSE} = \frac{1}{N} \sum_{t=1}^{T-\tau} \sum_{i=1}^{m} \left\| \mathbf{k}_i^\top \chi_i \left(\mathbf{x}_t \right) - \chi_i \left(\mathbf{x}_{t+\tau} \right) \right\|^2 \tag{B1}$$

Under the condition $N \to \infty$, we have

$$\begin{aligned} \mathrm{MSE} &= \sum_{i=1}^{m} \int \mathrm{d}\mathbf{x}\ \rho\left(\mathbf{x}\right) \left(\mathbf{k}_i^\top \boldsymbol{\chi} - \mathcal{K}_\tau \chi_i \right)^\top \left(\mathbf{k}_i^\top \boldsymbol{\chi} - \mathcal{K}_\tau \chi_i \right) \\ &= \sum_{i=1}^{m} \left\langle \mathbf{k}_i^\top \boldsymbol{\chi} - \mathcal{K}_\tau \chi_i, \mathbf{k}_i^\top \boldsymbol{\chi} - \mathcal{K}_\tau \chi_i \right\rangle_\rho \end{aligned}$$

Appendix C: Proof of (32) and (33)

Here, we define

$$\mathbf{C}_\rho\left(0\right) = \mathbb{E}_\rho \left[\boldsymbol{\chi}\left(\mathbf{x}_t\right) \boldsymbol{\chi}\left(\mathbf{x}_t\right)^\top \right]. \tag{C1}$$

Obviously, $\hat{\mathbf{C}}\left(0\right)$ is an unbiased estimate of $\mathbf{C}_\rho\left(0\right)$ with $\hat{\mathbf{C}}\left(0\right) \to \mathbf{C}_\rho\left(0\right)$ as $N \to \infty$.

Since

$$\begin{aligned} \mathbb{E}_\mu \left[\boldsymbol{\chi}\left(\mathbf{x}_{t+\tau}\right) \right] &= \mathbb{E}_\mu \left[\mathcal{K}_\tau \boldsymbol{\chi}\left(\mathbf{x}_t\right) \right] \\ &= \mathbb{E}_\mu \left[\mathbf{K}^\top \boldsymbol{\chi}\left(\mathbf{x}_t\right) \right] \\ &= \int \mathrm{d}\mathbf{x}\ \mathbf{u}^\top \boldsymbol{\chi}\left(\mathbf{x}\right) \cdot \rho\left(\mathbf{x}\right) \cdot \mathbf{K}^\top \boldsymbol{\chi}\left(\mathbf{x}\right) \\ &= \mathbf{K}^\top \left(\int \mathrm{d}\mathbf{x}\ \rho\left(\mathbf{x}\right) \boldsymbol{\chi}\left(\mathbf{x}\right) \boldsymbol{\chi}\left(\mathbf{x}\right)^\top \right) \mathbf{u} \\ &= \mathbf{K}^\top \mathbf{C}_\rho\left(0\right) \mathbf{u} \end{aligned} \tag{C2}$$

and

$$\begin{aligned} \mathbb{E}_\mu \left[\boldsymbol{\chi}\left(\mathbf{x}_t\right) \right] &= \int \mathrm{d}\mathbf{x}\ \mathbf{u}^\top \boldsymbol{\chi}\left(\mathbf{x}\right) \cdot \rho\left(\mathbf{x}\right) \cdot \boldsymbol{\chi}\left(\mathbf{x}\right) \\ &= \mathbf{C}_\rho\left(0\right) \mathbf{u}, \end{aligned} \tag{C3}$$

we can obtain from $\mathbb{E}_\mu \left[\boldsymbol{\chi}\left(\mathbf{x}_{t+\tau}\right) \right] = \mathbb{E}_\mu \left[\boldsymbol{\chi}\left(\mathbf{x}_t\right) \right]$ that

$$\mathbf{u}^\top \mathbf{C}_\rho\left(0\right) \mathbf{K} = \mathbf{u}^\top \mathbf{C}_\rho\left(0\right). \tag{C4}$$

In addition, the integral of $\mu(\mathbf{x})$ over all phase space can be expressed as

$$\begin{aligned}\int \mathrm{d}\mathbf{x}\ \mu(\mathbf{x}) &= \int \mathrm{d}\mathbf{x}\ \mu(\mathbf{x})\,\boldsymbol{\chi}(\mathbf{x})^\top \mathbf{v} \\ &= \int \mathrm{d}\mathbf{x}\ \mathbf{u}^\top \boldsymbol{\chi}(\mathbf{x})\,\boldsymbol{\chi}(\mathbf{x})^\top \mathbf{v} \\ &= \mathbf{u}^\top \mathbf{C}_\rho(0)\,\mathbf{v} \qquad \text{(C5)}\end{aligned}$$

Therefore,

$$\mathbf{u}^\top \mathbf{C}_\rho(0)\,\mathbf{v} = 1. \qquad \text{(C6)}$$

Appendix D: Analysis of the reversible estimator

Considering that

$$\begin{aligned}\hat{\mathbf{C}}_{\mathrm{eq}}(0)\,\mathbf{v} &= \frac{1}{N}\mathbf{X}^\top \mathrm{diag}(\mathbf{X}\mathbf{u})\,\mathbf{X}\mathbf{v} \\ &= \frac{1}{N}\mathbf{X}^\top \mathrm{diag}(\mathbf{X}\mathbf{u})\,\mathbf{1} \\ &= \frac{1}{N}\mathbf{X}^\top \mathbf{X}\mathbf{u} \\ &= \hat{\mathbf{C}}(0)\,\mathbf{u} \qquad \text{(D1)}\end{aligned}$$

and

$$\begin{aligned}\mathbf{K}\mathbf{v} &= \hat{\mathbf{C}}(0)^{-1}\,\hat{\mathbf{C}}(\tau)\,\mathbf{v} \\ &= \hat{\mathbf{C}}(0)^{-1}\left(\frac{1}{N}\mathbf{X}^\top \mathbf{Y}\mathbf{v}\right) \\ &= \hat{\mathbf{C}}(0)^{-1}\left(\frac{1}{N}\mathbf{X}^\top \mathbf{X}\mathbf{v}\right) \\ &= \hat{\mathbf{C}}(0)^{-1}\,\hat{\mathbf{C}}(0)\,\mathbf{v} \\ &= \mathbf{v}\end{aligned}$$

where $\mathbf{1}$ denotes a column vector of ones of appropriate size, thenthe modified matrix $\tilde{\mathbf{K}}$ given by (39) satisfies

$$\begin{aligned}
\mathbf{u}^\top \hat{\mathbf{C}}(0)\tilde{\mathbf{K}} &= \frac{1}{2}\left(\mathbf{u}^\top \hat{\mathbf{C}}(0)\mathbf{K} + \mathbf{u}^\top \hat{\mathbf{C}}(0)\hat{\mathbf{C}}_{\mathrm{eq}}(0)^{-1}\mathbf{K}^\top \hat{\mathbf{C}}_{\mathrm{eq}}(0)\right) \\
&= \frac{1}{2}\left(\mathbf{u}^\top \hat{\mathbf{C}}(0) + \mathbf{v}^\top \hat{\mathbf{C}}_{\mathrm{eq}}(0)\hat{\mathbf{C}}_{\mathrm{eq}}(0)^{-1}\mathbf{K}^\top \hat{\mathbf{C}}_{\mathrm{eq}}(0)\right) \\
&= \frac{1}{2}\left(\mathbf{u}^\top \hat{\mathbf{C}}(0) + \hat{\mathbf{C}}_{\mathrm{eq}}(0)\right) \\
&= \mathbf{u}^\top \hat{\mathbf{C}}(0) = \mathbf{u}^\top \hat{\mathbf{C}}(0)\mathbf{K} \qquad \text{(D2)} \\
\tilde{\mathbf{K}}\mathbf{v} &= \frac{1}{2}\left(\mathbf{K}\mathbf{v} + \hat{\mathbf{C}}_{\mathrm{eq}}(0)^{-1}\mathbf{K}^\top \hat{\mathbf{C}}_{\mathrm{eq}}(0)\mathbf{v}\right) \\
&= \frac{1}{2}\left(\mathbf{v} + \hat{\mathbf{C}}_{\mathrm{eq}}(0)^{-1}\mathbf{K}^\top \hat{\mathbf{C}}(0)\mathbf{u}\right) \\
&= \frac{1}{2}\left(\mathbf{v} + \hat{\mathbf{C}}_{\mathrm{eq}}(0)^{-1}\hat{\mathbf{C}}(0)\mathbf{u}\right) \\
&= \frac{1}{2}\left(\mathbf{v} + \hat{\mathbf{C}}_{\mathrm{eq}}(0)^{-1}\hat{\mathbf{C}}_{\mathrm{eq}}(0)\mathbf{v}\right) \\
&= \mathbf{v} = \mathbf{K}\mathbf{v}. \qquad \text{(D3)}
\end{aligned}$$

So the reweighting vector $\mathbf{u}$ remains fixed after the modification, and the estimated eigenfunction with eigenvalue 1 is still $\mathbf{v}^\top \boldsymbol{\chi} = \mathbb{1}$.

In addition, if $\hat{\mathbf{C}}_{\mathrm{eq}}(0)$ is positive-definite, the matrix

$$\hat{\mathbf{C}}_{\mathrm{eq}}(0)^{\frac{1}{2}}\tilde{\mathbf{K}}\hat{\mathbf{C}}_{\mathrm{eq}}(0)^{-\frac{1}{2}} = \frac{1}{2}\left(\hat{\mathbf{C}}_{\mathrm{eq}}(0)^{\frac{1}{2}}\mathbf{K}\hat{\mathbf{C}}_{\mathrm{eq}}(0)^{-\frac{1}{2}} + \hat{\mathbf{C}}_{\mathrm{eq}}(0)^{-\frac{1}{2}}\mathbf{K}^\top \hat{\mathbf{C}}_{\mathrm{eq}}(0)^{\frac{1}{2}}\right) \qquad \text{(D4)}$$

is symmetric, which implies that the eigenvalues of $\tilde{\mathbf{K}}$ are real.

Appendix E: Detailed decorrelation procedure of basis functions

Suppose that $\mathbf{v}_o^\top \boldsymbol{\chi}_o = \mathbb{1}$, then the mean value and covariance matrix of $\{\boldsymbol{\chi}_o(\mathbf{x}_1), \ldots, \boldsymbol{\chi}_o(\mathbf{x}_{T-\tau})\}$ can be computed as

$$\frac{1}{N}\sum_{t=1}^{T-\tau}\boldsymbol{\chi}_o(\mathbf{x}_t) = \hat{\mathbf{C}}_o(0)\mathbf{v}_o \qquad \text{(E1)}$$

$$\frac{1}{N}\sum_{t=1}^{T-\tau}\left(\boldsymbol{\chi}_o(\mathbf{x}_t) - \hat{\mathbf{C}}_o(0)\mathbf{v}_o\right)\left(\boldsymbol{\chi}_o(\mathbf{x}_t) - \hat{\mathbf{C}}_o(0)\mathbf{v}_o\right)^\top = \hat{\mathbf{C}}_o(0) - \hat{\mathbf{C}}_o(0)\mathbf{v}_o\mathbf{v}_o^\top \hat{\mathbf{C}}_o(0), \text{(E2)}$$

where

$$\hat{\mathbf{C}}_o(0) = \frac{1}{N}\sum_{t=1}^{T-\tau}\boldsymbol{\chi}_o(\mathbf{x}_t)\boldsymbol{\chi}_o(\mathbf{x}_t)^\top. \qquad \text{(E3)}$$

Suppose the truncated eigendecomposition of the covariance matrix is

$$\hat{\mathbf{C}}_o(0) - \hat{\mathbf{C}}_o(0)\,\mathbf{v}_o\mathbf{v}_o^\top\hat{\mathbf{C}}_o(0) \approx \mathbf{Q}_d^\top\mathbf{S}_d\mathbf{Q}_d, \tag{E4}$$

where the diagonal of matrix $\mathbf{S}_d$ contains all positive eigenvalues that are larger than ϵ_0 and absolute values of all negative eigenvalues ($\epsilon_0 = 10^{-10}$ in our applications). Then the decorrelation can be implemented as

$$\begin{aligned}\boldsymbol{\chi} &= \begin{bmatrix}\mathbf{Q}_d^\top\mathbf{S}_d^{\frac{1}{2}}\left(\boldsymbol{\chi} - \hat{\mathbf{C}}_o(0)\,\mathbf{v}_o\right)\\ \mathbb{1}\end{bmatrix}\\ &= \begin{bmatrix}\mathbf{Q}_d^\top\mathbf{S}_d^{\frac{1}{2}}\left(\mathbf{I} - \hat{\mathbf{C}}_o(0)\,\mathbf{v}_o\mathbf{v}_o^\top\right)\\ \mathbf{v}_o^\top\end{bmatrix}\boldsymbol{\chi}_o\\ &= \mathbf{R}_{\hat{\mathbf{C}}_o(0),\mathbf{v}_o}^\top\boldsymbol{\chi}_o, \end{aligned}\tag{E5}$$

with

$$\mathbf{R}_{\hat{\mathbf{C}}_o(0),\mathbf{v}_o}^\top = \begin{bmatrix}\mathbf{Q}_d^\top\mathbf{S}_d^{\frac{1}{2}}\left(\mathbf{I} - \hat{\mathbf{C}}_o(0)\,\mathbf{v}_o\mathbf{v}_o^\top\right)\\ \mathbf{v}_o^\top\end{bmatrix}. \tag{E6}$$

Similarly, the decorrelation of $\boldsymbol{\chi}$ according to the equilibrium distribution can also be implemented as

$$\boldsymbol{\chi}_s = \mathbf{R}_{\hat{\mathbf{C}}_{\mathrm{eq}}(0),\mathbf{v}}^\top\boldsymbol{\chi}. \tag{E7}$$

Appendix F: Simulation models and experimental setups

1. One-dimensional diffusion process

The diffusion processes in Section V A is driven by the Brownian dynamics

$$\mathrm{d}x_t = -\nabla U(x_t)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}W_t \tag{F1}$$

where $\beta = 0.3$, sample interval is 0.002, x_0 is uniformly drawn in $[0, 0.2]$, and the potential function is given by

$$U(x) = \frac{\sum_{i=1}^{5}\left(|x - c_i| + 0.001\right)^{-2}u_i}{\sum_{i=1}^{5}\left(|x - c_i| + 0.001\right)^{-2}} \tag{F2}$$

with $c_{1:5} = (-0.3, 0.5, 1, 1.5, 2.3)$. Simulations are implemented by a reversibility preserving numerical discretization scheme proposed in [**?**] with bin size 0.02. The basis functions for estimators are chosen to be

$$\chi_i(x) = \exp\left(-\left(w_i x + b_i\right)^2\right), \tag{F3}$$

where w_i and b_i are randomly drawn in $[-1,1]$ and $[0,1]$.

2. Two-dimensional diffusion process

The dynamics of the two-dimensional diffusion process in Section V B has the same form as (F1), where $\beta = 0.5$, sample interval is 0.05, $\mathbf{x}_0 = (x_0, y_0)$ is uniformly drawn in $[-2,-1.5] \times [-1.5, 2.5]$, and the potential function is chosen as in [80] by

$$\begin{aligned} U(x,y) = 3\exp&\left(-x^2 - \left(y-\frac{1}{3}\right)^2\right) \\ &-3\exp\left(-x^2 - \left(y-\frac{5}{3}\right)^2\right) \\ &-5\exp\left(-(x-1)^2 - y^2\right) \\ &-5\exp\left(-(x+1)^2 - y^2\right) \\ &+\frac{1}{5}x^4 + \frac{1}{5}\left(y-\frac{1}{3}\right)^4 . \end{aligned} \tag{F4}$$

Simulations are implemented by the same algorithm as in Appendix F 1 with bin size 0.2×0.2. The basis functions for estimators are also Gaussian functions

$$\chi_i(\mathbf{x}) = \exp\left(-\left(\mathbf{w}_i^\top \mathbf{x} + b_i\right)^2\right), \tag{F5}$$

with random weights $\mathbf{w}_i \in [-1,1] \times [-1,1]$ and $b_i \in [0,1]$.

[1] M. Shirts and V. S. Pande, Science **290**, 1903 (2000).

[2] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten, J. Comput. Chem. **26**, 1781 (2005).

[3] M. Harvey, G. Giupponi, and G. D. Fabritiis, J. Chem. Theory Comput. **5**, 1632 (2009).

[4] I. Buch, M. J. Harvey, T. Giorgino, D. P. Anderson, and G. De Fabritiis, J. Chem. Inf. Model. **50**, 397 (2010).

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

[6] P. Eastman, M. S. Friedrichs, J. D. Chodera, R. J. Radmer, C. M. Bruns, J. P. Ku, K. A. Beauchamp, T. J. Lane, L.-P. Wang, D. Shukla, T. Tye, M. Houston, T. Stich, C. Klein, M. R. Shirts, and V. S. Pande, J. Chem. Theory Comput. **9**, 461 (2013).

[7] S. L. Grand, A. W. Goetz, and R. C. Walker, Chem. Phys. Comm. **184**, 374 (2013).

[8] S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl, Bioinformatics **29**, 845 (2013).

[9] S. Doerr, M. J. Harvey, F. Noé, and G. D. Fabritiis, J. Chem. Theory Comput. **12**, 1845 (2016).

[10] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, Proc. Natl. Acad. Sci. USA **106**, 19011 (2009).

[11] I. Buch, T. Giorgino, and G. De Fabritiis, Proc. Natl. Acad. Sci. USA **108**, 10184 (2011).

[12] G. R. Bowman, V. A. Voelz, and V. S. Pande, J. Am. Chem. Soc. **133**, 664 (2011).

[13] S. K. Sadiq, F. Noé, and G. De Fabritiis, Proc. Natl. Acad. Sci. USA **109**, 20449 (2012).

[14] D.-A. Silva, D. R. Weiss, F. P. Avila, L.-T. Da, M. Levitt, D. Wang, and X. Huanga, Proc. Natl. Acad. Sci. USA **111**, 7665 (2014).

[15] D. Shukla, Y. Meng, B. Roux, and V. S. Pande, Nat. Commun. **5**, 3397 (2014).

[16] N. Plattner and F. Noé, Nature Commun. **6**, 7653 (2015).

[17] T. F. Reubold, K. Faelber, N. Plattner, Y. Posor, K. Branz, U. Curth, J. Schlegel, R. Anand, D. Manstein, F. Noé, V. Haucke, O. Daumke, and S. Eschenburg, Nature **525**, 404 (2015).

[18] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, J. Comput. Phys. **151**, 146 (1999).

[19] W. C. Swope, J. W. Pitera, and F. Suits, J. Phys. Chem. B **108**, 6571 (2004).

[20] F. Noé, I. Horenko, C. Schütte, and J. C. Smith, J. Chem. Phys. **126**, 155102 (2007).

[21] J. D. Chodera, K. A. Dill, N. Singhal, V. S. Pande, W. C. Swope, and J. W. Pitera, J. Chem. Phys. **126**, 155101 (2007).

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

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

[24] C. Schütte, F. Noé, J. Lu, M. Sarich, and E. Vanden-Eijnden, J. Chem. Phys. **134**, 204105 (2011).

[25] G. R. Bowman, V. S. Pande, and F. Noé, eds., *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation.*, Advances in Experimental Medicine and Biology, Vol. 797 (Springer Heidelberg, 2014).

[26] D. S. Chekmarev, T. Ishida, and R. M. Levy, J. Phys. Chem. B **108**, 19487 (2004).

[27] S. Sriraman, I. G. Kevrekidis, and G. Hummer, J. Phys. Chem. B **109**, 6479 (2005).

[28] N. V. Buchete and G. Hummer, J. Phys. Chem. B **112**, 6057 (2008).

[29] F. Noé, H. Wu, J.-H. Prinz, and N. Plattner, J. Chem. Phys. **139**, 184114 (2013).

[30] H. Wu and F. Noé, Multiscale Model. Simul. **12**, 25 (2014).

[31] E. Rosta and G. Hummer, J. Chem. Theory Comput. **11**, 276 (2015).

[32] H. Wu and F. Noé, J. Chem. Phys. **142**, 084104 (2015).

[33] H. Wu, F. Paul, C. Wehmeyer, and F. Noé, Proc. Natl. Acad. Sci. USA **113**, E3221 (2016).

[34] M. Sarich, F. Noé, and C. Schütte, Multiscale Model. Simul. **8**, 1154 (2010).

[35] P. Metzner, C. Schütte, and E. Vanden-Eijnden, Multiscale Model. Simul. **7**, 1192 (2009).

[36] A. Berezhkovskii, G. Hummer, and A. Szabo, J. Chem. Phys. **130**, 205102 (2009).

[37] F. Noé, S. Doose, I. Daidone, M. Löllmann, J. D. Chodera, M. Sauer, and J. C. Smith, Proc. Natl. Acad. Sci. USA **108**, 4822 (2011).

[38] B. G. Keller, J.-H. Prinz, and F. Noé, Chem. Phys. **396**, 92 (2012).

[39] W. Zhuang, R. Z. Cui, D.-A. Silva, and X. Huang, *J. Phys. Chem. B*, J. Phys. Chem. B **115**, 5415 (2011).

[40] B. Lindner, Z. Yi, J.-H. Prinz, J. C. Smith, and F. Noé, J. Chem. Phys. **139**, 175101 (2013).

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

[42] B. Peters, J. Chem. Phys. **125**, 241101 (2006).

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

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

[45] M. A. Rohrdanz, W. Zheng, and C. Clementi, Ann. Rev. Phys. Chem. **64**, 295 (2013).

[46] B. Trendelkamp-Schroer, H. Wu, F. Paul, and F. Noé, J. Chem. Phys. **143**, 174101 (2015).

[47] A. Bittracher, P. Koltai, and O. Junge, J. Appl. Dyn. Syst. **14**, 1478 (2015).

[48] F. Noé, J. Chem. Phys. **128**, 244103 (2008).

[49] B. Trendelkamp-Schroer and F. Noé, J. Phys. Chem. **138**, 164113. (2013).

[50] F. Noé and F. Nüske, Multiscale Model. Simul. **11**, 635 (2013).

[51] G. Perez-Hernandez, F. Paul, T. Giorgino, G. D Fabritiis, and F. Noé, J. Chem. Phys. **139**, 015102 (2013).

[52] F. Nüske, B. G. Keller, G. Pérez-Hernández, A. S. J. S. Mey, and F. Noé, J. Chem. Theory Comput. **10**, 1739 (2014).

[53] S. M. W. Huisinga and C. Schuette, Ann. Appl. Probab. **14**, 419458 (2004).

[54] P. Deuflhard and M. Weber, ZIB Report **03-09** (2003).

[55] S. Kube and M. Weber, J. Chem. Phys. **126**, 024103 (2007).

[56] M. Cameron and E. Vanden-Eijnden, J. Stat. Phys. **156**, 427 (2014).

[57] L. Molgedey and H. G. Schuster, Phys. Rev. Lett. **72**, 3634 (1994).

[58] A. Ziehe and K.-R. Müller, in *ICANN 98* (Springer Science and Business Media, 1998) pp. 675–680.

[59] E. O. Aapo Hyvärinen, Juha Karhunen, *Independent Component Analysis* (John Wiley & Sons, 2001).

[60] C. R. Schwantes and V. S. Pande, J. Chem. Theory Comput. **9**, 2000 (2013).

[61] Y. Naritomi and S. Fuchigami, J. Chem. Phys. **134**, 065101 (2011).

[62] L. Boninsegna, G. Gobbo, F. Noé, and C. Clementi, J. Chem. Theory Comput. **11**, 5947 (2015).

[63] R. T. McGibbon and V. S. Pande, arXiv:1602.08776 (2016).

[64] F. Noé and C. Clementi, J. Chem. Theory Comput. **22**, 5002 (2015).

[65] F. Noé and C. Clementi, J. Chem. Theory Comput. ((submitted)).

[66] C. R. Schwantes and V. S. Pande, J. Chem. Theory Comput. **11**, 600 (2015).

[67] R. T. McGibbon and V. S. Pande, J. Chem. Phys. **142**, 124105 (2015).

[68] F. Nüske, R. Schneider, F. Vitalini, and F. Noé, J. Chem. Phys. **144**, 054105 (2016).

[69] F. Vitalini, F. Noé, and B. G. Keller, J. Chem. Theory Comput. **11**, 3992 (2015).

[70] P. J. Schmid and J. Sesterhenn, in *61st Annual Meeting of the APS Division of Fluid Dynamics. American Physical Society* (2008).

[71] C. W. Rowley, I. Mezić, S. Bagheri, P. Schlatter, and D. S. Henningson, J. Fluid Mech. **641**, 115 (2009).

[72] P. J. Schmid, J. Fluid Mech. **656**, 5 (2010).

[73] J. H. Tu, C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. N. Kutz, J. Comput. Dyn. **1**, 391 (2014).

[74] M. O. Williams, I. G. Kevrekidis, and C. W. Rowley, J. Nonlinear Sci. **25**, 1307 (2015).

[75] I. Mezić, Nonlinear Dyn. **41**, 309 (2005).

[76] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Perez-Hernandez, M. Hoffmann, N. Plattner, J.-H. Prinz, and F. Noé, J. Chem. Theory Comput. **11**, 5525 (2015).

[77] S. Klus, P. Koltai, and C. Schütte, arXiv:1512.05997 (2015).

[78] P. W. Glynn and D. L. Iglehart, Manag. Sci. **35**, 1367 (1989).

[79] K. Pearson, Phil. Mag. **2**, 559 (1901).

[80] P. Metzner, C. Schütte, and E. Vanden-Eijnden, J. Chem. Phys. **125** (2006), 10.1063/1.2335447.