Variational cross-validation of slow dynamical modes in molecular kinetics

Robert T. McGibbon\textsuperscript{1} and Vijay S. Pande\textsuperscript{1}

\textit{Department of Chemistry, Stanford University, Stanford CA 94305, USA}

(Dated: 31 July 2014)

We consider the problem of robustly determining the \( m \) slowest dynamical modes of a reversible dynamical system, with a particular focus on the analysis of equilibrium molecular dynamics simulations. We show that the problem can be formulated as the variational optimization of a single scalar functional, a generalized matrix Rayleigh quotient (GMRQ), which measures the ability of a rank-\( m \) projection operator to capture the slow dynamics of the system. While a variational theorem bounds the GMRQ from above by the sum of the first \( m \) eigenvalues of the system’s propagator, we show that this bound can be violated when the requisite matrix elements are estimated subject to statistical uncertainty. Furthermore, this overfitting can be detected and avoided through cross-validation in which the GMRQ is evaluated for the purpose of model selection on data that was held out during training. These result make it possible to, for the first time, construct a unified, consistent objective function for the parameterization of Markov state models for protein dynamics which captures the tradeoff between systematic and statistical errors.

Keywords: molecular dynamics, propagator, cross-validation, Rayleigh quotient, variational principle
I. INTRODUCTION

Conformational dynamics are central to the biological function of macromolecular systems such as signaling proteins, enzymes, and channels. The molecular description of processes as diverse as protein folding, kinase activation, voltage-gating of ion channels, and ubiquitin signaling involve not just the structure of a unique single conformation, but on the conformational dynamics between a multitude of states accessible on the potential energy surface.\[1,4\] These dynamics occur on a range of timescales and have varying degrees of structural complexity: localized vibrations may occur on the 0.1 ps timescale, while large-scale structural changes like protein folding can take seconds or longer.\[5\] Although many experimental techniques – most notably X-ray crystallography and NMR – can yield detailed structural information on functional conformations, the experimental characterization of the dynamical processes, intermediate conformations and transition pathways in macromolecular systems remains exceptionally challenging.\[6,7\]

Atomistic molecular dynamics (MD) simulations can complement experiment and provide a powerful tool for probing conformational dynamics, allowing researchers to directly visualize and analyze the time evolution of macromolecular systems in atomic detail. Three major challenges for MD simulation of complex systems are the accuracy of the potential energy functions, adequate sampling of conformational space, and quantitative analysis of simulation results. The state-of-the-art on all three fronts has advanced rapidly in recent years. A new generation of increasingly accurate forcefields have recently emerged, such as those which include explicit polarizability and have been parameterized more systematically.\[8-12\] On the sampling problem, the introduction of graphical processing units (GPUs) has dramatically expanded the timescales accessible with MD simulation at modest cost, and specialized MD-specific hardware and distributed computing networks have yielded further gains.\[13-17\] In this work, we focus on the remaining challenge, the quantitative analysis of MD simulations.

Despite, or perhaps because of their detail, MD simulations require further analysis in order to yield insight into macromolecular dynamics or quantitative predictions capable of being tested experimentally. The direct result of a simulation, an MD trajectory, is a time series of Cartesian coordinates (and perhaps velocities) of dimension $3N$ ($6N$ if velocities are retained), where $N$ is the number of atoms in the system. Because routine MD simulations may contain tens or hundreds of thousands of atoms, these time series are extremely...
A multitude of methods have been proposed for reducing the dimensionality or complexity of MD trajectories and enabling the analysis of the system’s key long-lived conformational states, dynamical modes, transition pathways, and essential degrees of freedom. A central challenge in this enterprise, however, has been the lack of a single unambiguous method by which to evaluate and compare different proposed dimensionality reductions, or to use as a guide to find optimally informative projections.

To illustrate this challenge, consider the application of principal components analysis (PCA), a classical linear dimensionality reduction technique, to a protein MD simulation. Should PCA be applied directly to the time series of atomic coordinates, or instead to some nonlinear transformation thereof, such as the coordinates after rotational alignment to a reference structure or the backbone $\phi$ and $\psi$ dihedral angles of the protein chain? Or, when clustering a set of MD trajectories (which can be viewed as a dimensionality reduction projecting the conformations in a trajectory from $\mathbb{R}^{3N}$ to $\{0,1\}^p$, where $p$ is the number of clusters), should the distance metric between conformations used to define the clustering measure the root-mean-square deviation of atomic positions, the change in residue-residue contact distances, the deviation in the reciprocal inter-atomic distances, or something else? Various arguments favoring particular methodologies have appeared in the literature, but we are aware of no unified theory governing such a choice in a computable manner. When the proposed dimensionality reduction reduces each conformation to a single real number, $Q(x) : \mathbb{R}^{3N} \rightarrow \mathbb{R}$, and the system is characterized by two metastable macrostates $A$ and $B$, the transmission coefficient or committor – the probability of arriving at $A$ before arriving at $B$ – can be viewed as an ideal dimensionality reduction, and proposed models can be evaluated by their correlation with the committor. However, there is no natural extension of this concept to a situation in which many metastable states exist and the rank of the proposed dimensionality reduction, $m$, is greater than 1.

Here, we propose a method backed by a variational theorem for comparing alternative rank-$m$ dimensionality reductions of a high-dimensional dynamical system. The method is based on a unified approximation of the first $m$ eigenfunctions of the propagator for an equilibrium stochastic dynamical system. Unlike committor values, the method does not rely on the existence or identification of particular metastable states. We show that both the existing time-structure independent components analysis (tICA) and Markov state model (MSM) methods can be interpreted as directly optimizing this criteria using different
restricted families of basis functions. Furthermore, we show that the variational bounds governing the approximation of these slow modes can be violated when the requisite matrix elements are estimated subject to statistical uncertainty, but that the use of cross-validation enables both the systematic and statistical error (bias and variance) in the construction of these reduced-rank representations to be balanced systematically. Finally, these tools make it possible to quantitatively select hyperparameters, such as the clustering method and number of states in an MSM. These results are demonstrated on three systems: simulations of a double well potential, terminally-block octalanine, and ubiquitin.

Our results build on the work of Noé and Nüske and Nüske et al., who introduced a closely related variational approach to characterizing the slow dynamics in molecular systems. While their formulation involves the stepwise optimization individual \( ansatz \) eigenfunctions with mounting orthogonality constraints, our approach arrives at the same result during training via the optimization of a single scalar functional of a collection of \( m \) \( ansatz \) eigenfunctions simultaneously. This formulation uniquely enables the evaluation of the proposed eigenfunctions on new data which was held out during the fitting step, which we show to be essential to avoid overfitting.

II. THEORY

A. Preliminaries

We begin by introducing the problem setting from a mathematical perspective, introducing the key variables and notation that will be essential for the remainder of this work. We largely follow the order of presentation in Prinz et al. which contains a longer and more thorough discussion.

Consider a time-homogeneous, ergodic, continuous-time Markov process \( x(t) \in \Omega \) which is reversible with respect to a stationary distribution stationary distribution \( \mu(x) : \Omega \to \mathbb{R}^+ \). The system’s evolution over an interval \( \tau > 0 \) is described by a transition probability density

\[
p(x, y; \tau)dy = \mathbb{P}[x(t + \tau) \in y + dy | x(t) = x].
\]

Consider an ensemble of such systems at time \( t \), distributed according to some probability
distribution \( p_t(x) \). After waiting for a duration \( \tau \), the distribution evolves to

\[
p_{t+\tau}(y) = \int_{\Omega} dx \, p(x, y; \tau) \, p_t(x) = \mathcal{P}(\tau) \circ p_t(y),
\]

(2)

which defines the continuous integral operator \( \mathcal{P}(\tau) \) called the propagator with lag time \( \tau \). The propagator \( \mathcal{P}(\tau) \) admits a natural decomposition in terms of its eigenfunctions and eigenvalues

\[
\mathcal{P}(\tau) \circ \phi_i = \lambda_i \phi_i.
\]

(3)

The propagator \( \mathcal{P}(\tau) \) is self-adjoint with respect to the \( \mu^{-1} \) weighted scalar product,

\[
\langle f, g \rangle_{\mu^{-1}} = \int_{\Omega} dx \, f(x)g(x)\mu^{-1}(x),
\]

(4)

and has a unique largest eigenvalue \( \lambda_1 = 1 \) with corresponding eigenfunction \( \phi_1(x) = \mu(x) \). The remaining eigenvalues are real and positive, and can be sorted in descending order. The eigenfunctions are taken to be \( \mu^{-1} \)-orthonormal. Using the spectral decomposition of \( \mathcal{P}(\tau) \), an ensemble’s distribution at arbitrary multiples of \( \tau \) can be written as a sum of exponentially decaying relaxation processes

\[
p_{n\tau}(x) = \sum_{i=1}^{\infty} \lambda_i^k \langle p_t, \phi_i \rangle_{\mu^{-1}} \phi_i,
\]

(5)

\[
= \mu(x) + \sum_{i=2}^{\infty} \exp \left( -\frac{k\tau}{t_i} \right) \langle p_t, \phi_i \rangle_{\mu^{-1}} \phi_i,
\]

(6)

where \( t_i = -\frac{\tau}{\ln \lambda_i} \). The eigenfunctions \( \phi_i(x) \) for \( i = 2, ... \) can thus be interpreted as dynamical modes of the system, each of which relax towards the equilibrium distribution with a characteristic time-scale \( \tau_i \). Many molecular systems are characterized by \( m \) individual slow time-scales with eigenvalues close to one, separated from the remaining eigenvalues by a spectral gap. These remaining small eigenvalues correspond to faster dynamical processes that rapidly decay. Under these conditions, the long-time dynamics induced by the propagator can be well described by consideration of only these slow eigenfunctions – that is, a rank-\( m \) low-rank approximation.

Furthermore, not only do these eigenfunctions form a convenient basis, in fact they lead to an optimal reduced-rank description of the dynamics. That is, each of the partial sums formed by truncating the expansion in Eq. (5) at its first \( m \) terms is the closest possible rank-\( m \) approximation to \( \mathcal{P} \) in spectral norm. This statement is made precise by the following theorem.
Theorem 1. Let \( \mathcal{P} \) be compact linear operator which is self-adjoint with respect to an inner product \( \langle \cdot, \cdot \rangle_{\mu^{-1}} \). Assume that the eigenvalues \( \lambda_i \) and associated eigenfunctions \( \phi_i \) are sorted in descending order by eigenvalue. Define the operator \( \mathcal{P}_m \) such that \( \mathcal{P}_m \circ f = \sum_{i=1}^{m} \lambda_i \langle f, \phi_i \rangle_{\mu^{-1}} \phi_i \). Then,

\[
\mathcal{P}_m = \arg\min_{\text{rank}(A_m) \leq m} ||A_m - \mathcal{P}||_{\mu^{-1}}.
\]

Proof. This is the extension of the familiar Eckart-Young theorem to self-adjoint linear operators. The original result is by Schmidt.\(^{40}\) See Courant and Hilbert (pp 161)\(^{41}\) and Micchelli and Pinkus\(^{42}\) for further details.

Because of both this optimality property and their natural physical interpretation as the system’s essential degrees of freedom, we view the numerical identification of these eigenfunctions as the central mathematical goal in dimensionality reduction of complex stochastic dynamical systems.

B. Variational Principle for Eigenspaces

As is the case with the Hamiltonian in quantum mechanics, the analytical solution to the eigenvalue problem in Eq. (3) is only possible for certain toy systems. We therefore proceed by reformulating these eigenfunctions as the collective maximizer of a certain variational optimization problem, which provides an objective function by which various approximate solutions can be obtained and compared.

Because \( \mathcal{P}(\tau) \) is bounded and self-adjoint, variational principles governs the approximation of its eigenvalues and eigenfunctions. The following statement is a generalization of the variational principle introduced for \( \mathcal{P}(\tau) \) by Noé and Nüske\(^{38}\) which applies jointly to the estimation of the eigenspace formed by the span of multiple eigenfunctions.

Theorem 2. Let \( \mathcal{P} \) be compact linear operator whose eigenvalues \( \lambda_1 > \lambda_2 \geq \lambda_3, \ldots \) are bounded from above and which is self-adjoint with respect to an inner product \( \langle \cdot, \cdot \rangle_{\mu^{-1}} \). Furthermore, let \( f \) be an arbitrary set of \( m \) linearly independent functions on \( \mathbb{R}^N \to \mathbb{R} \), \( f = \{f_i(\cdot)\}_{i=1}^{m} \). Define a matrix \( P(f) \in \mathbb{S}^m \) with \( P_{ij} = \langle f_i, \mathcal{P} \circ f_j \rangle_{\mu^{-1}} \), and a matrix \( Q(f) \in \mathbb{S}^m_{++} \) with \( Q_{ij} = \langle f_i, f_j \rangle_{\mu^{-1}} \). Define \( \mathcal{R}_\mathcal{P}[f] \) as

\[
\mathcal{R}_\mathcal{P}[f] = \text{Tr} \left( P(f) Q(f)^{-1} \right).
\]
Then,
\[ R_P[f] \leq \sum_{i=1}^{m} \lambda_i. \] (9)

**Proof.** The eigenfunctions \( \phi_i \) of \( P \) form a complete basis. Expand each \( f_i = \sum a_i \phi_a \) with coefficients \( W \in \mathbb{R}^{\infty \times m} \) with 
\[ W_{ni} = \langle f_i, \phi_n \rangle_{\mu^{-1}}. \]

\[ P_{ij} = \langle f_i, P \circ f_j \rangle_{\mu^{-1}} \] (10)
\[ = \left\langle \sum_a W_{ai} \phi_a, P \circ \sum_b W_{bj} \phi_b \right\rangle_{\mu^{-1}} \] (11)
\[ = \sum_a W_{ai} W_{aj} \lambda_a \] (12)
\[ P = W^T D(\lambda) W \] (13)
\[ Q_{ij} = \langle f_i, f_j \rangle_{\mu^{-1}} \] (14)
\[ = \left\langle \sum_a W_{ai} \phi_a, \sum_b W_{bj} \phi_b \right\rangle_{\mu^{-1}} \] (15)
\[ = \sum_a W_{ai} W_{aj} \] (16)
\[ Q = W^T W \] (17)

Let \( F = Q^{1/2} \in S_{++}^m \) be the square root of \( Q \). Then, rearrange the objective function

\[ R_P = \text{Tr} \left( W^T D(\lambda) W (FF)^{-1} \right) \] (18)
\[ = \text{Tr} \left( F^{-1} W^T D(\lambda) W F^{-1} \right) \] (19)
\[ = \text{Tr} \left( B^T D(\lambda) B \right), \] (20)

where \( B = WF^{-1} \). Note that \( B^T B = F^{-1} W^T W F^{-1} = I_m \). Therefore, by application of the Ky Fan theorem,\(^{4344}\)

\[ R_P[f] \leq \sum_{i} \lambda_i, \] (21)

and the equality holds when \( f = \{ \phi_1, \phi_2, \ldots, \phi_m \} \).

**Lemma 3.** The equality in Eq. (9) holds for any set of \( m \) functions, \( f \), such that \( \text{span}(f) = \text{span}(\{ \phi_1, \phi_2, \ldots, \phi_m \}) \).
Proof (Absil et al\textsuperscript{[19]}). For any invertible $p$-by-$p$ matrix $M$

\[
\mathcal{R}_p(WM) = \text{Tr} \left( (WM)^T D(\lambda) (WM) ((WM)^T (WM))^{-1} \right) \quad (22)
\]
\[
= \text{Tr} \left( M^T W^T D(\lambda) W M^{-1} (W^T W)^{-1} M^{-T} \right) \quad (23)
\]
\[
= \text{Tr} \left( W^T D(\lambda) W (W^T W)^{-1} \right) \quad (24)
\]
\[
= \mathcal{R}_\rho(W) \quad (25)
\]

This result implies that the slow eigenspace of $\mathcal{P}(\tau)$ can be numerically determined by varying a set of ansatz functions $f$ to maximize $\mathcal{R}_p[f]$. Note that when the trial functions $f$ are $\mu^{-1}$-orthonormal, $Q = I$. $Q$ can thus understood to simply ensure normalization. Under these conditions, $\mathcal{R}_p[f]$ then assumes a simple form as the sum of the individual Ritz values of the trial functions.

C. Linear Combination of Fixed Basis Functions

Equipped with this variational principle, we now consider the construction of an approximation to the dominant eigenspace of $\mathcal{P}(\tau)$ using linear combinations of functions from a finite basis set.

Let $\{ \varphi_i \}_{i=1}^n$ be a set of $n$ basis functions on $\mathbb{R}^N \rightarrow \mathbb{R}$. Using this basis set, we expand a vector of $m$ trial functions, $g = \{ g_i \}_{i=1}^m$ as linear combinations of the basis functions, with
\( g_j(\cdot) = \sum_{i=1}^{n} A_{ij} \varphi_i, \) where \( A = \{a_{ij}\} \in \mathbb{R}^{n \times m} \) is the matrix of expansion coefficients.

\[
P_{ij} = \langle g_i, P \circ g_j \rangle_{\mu^{-1}} \tag{26}
\]

\[
= \langle \sum_{a=1}^{n} A_{ai} \varphi_i, P \sum_{b=1}^{n} A_{bj} \varphi_j \rangle_{\mu^{-1}} \tag{27}
\]

\[
= \sum_{a,b=1}^{n} A_{ai} A_{bj} \langle \varphi_i, \varphi_j \rangle_{\mu^{-1}} \tag{28}
\]

\[
P = A^T C A \tag{29}
\]

\[
Q_{ij} = \langle g_i, g_j \rangle_{\mu^{-1}} \tag{30}
\]

\[
= \langle \sum_{a=1}^{n} A_{ai} \varphi_i, \sum_{b=1}^{n} A_{bj} \varphi_j \rangle_{\mu^{-1}} \tag{31}
\]

\[
= \sum_{a,b=1}^{n} A_{ai} A_{bj} \langle \varphi_i, \varphi_j \rangle_{\mu^{-1}} \tag{32}
\]

\[
Q = A^T S A \tag{33}
\]

where \( C \in S^n \) is a matrix with \( C_{ij} = \langle \varphi_i, P \circ \varphi_j \rangle_{\mu^{-1}} \) and \( S \in S_{++}^n \) is a matrix with \( S_{ij} = \langle \varphi_i, \varphi_j \rangle_{\mu^{-1}} \).

Thus, when \( f \) is expanded in a finite basis set, \( \mathcal{R}_P[f] \) reduces to the generalized matrix Rayleigh quotient (GMRQ), \( \mathcal{R}(A) = \mathcal{R}(A; C, S) \)

\[
\mathcal{R}(A; C, S) \equiv \text{Tr} \left( A^T C A (A^T S A)^{-1} \right) \tag{34}
\]

From the proof of Lemma 3, we can see that \( \mathcal{R}(A) \) is a function only of column span of \( A \), and is not affected by rescaling, or the application of any invertible transformation of the columns. Therefore, the optimization of \( \mathcal{R}(A) \) can be seen as a single optimization problem over the set of all \( m \)-dimensional linear subspaces of \( \mathbb{R}^n \). This space is referred to as a Grassmann manifold. Note that when \( m = 1 \), \( \mathcal{R}(A) \) reduces to the standard generalized Rayleigh quotient.

Furthermore, by the Ky Fan theorem, the training problem, \( A^* = \arg \max_A \mathcal{R}(A; C, S) \), is solved directly by a matrix \( A^* \) with columns that are the \( m \) generalized eigenvectors of \( C \) and \( S \) with the largest eigenvalues, and this eigenproblem is identical to the one introduced for the tICA method. Therefore, the GMRQ method herein and the stepwise optimization of the ansatz eigenfunctions subject to mutual orthogonality constraints, as specified by Schwantes and Pande and Pérez-Hernández et al. are functional identical during training when the basis set consists solely of linear functions.
D. Estimation of Matrix Elements

For simulation studies of molecular kinetics, $\mathcal{P}(\tau)$ is not directly available, and is only visible by way of one or more sampled MD trajectories, $\{x_t\}_{t=1}^T$, of the dynamical system it describes. First, introduce an auxiliary $\mu^{-1}$-weighted basis set $\chi_i(x) = \mu^{-1}(x)\varphi(x)$. The matrix elements $\{C_{ij}\}$ and $\{S_{ij}\}$ are expectation values, which can be estimated from $\{x_t\}_{t=1}^T$.

\[
C_{ij} = \langle \varphi_i, \mathcal{P}(\tau) \varphi_j \rangle_{\mu^{-1}}
\]
\[
= \int_{x \in \Omega} \int_{y \in \Omega} dx \, dy \, \mu^{-1}(y) \, \varphi_i(y) \, p(x, y; \tau) \, \varphi_j(x)
\]
\[
= \int_{x \in \Omega} \int_{y \in \Omega} dx \, dy \, \chi_i(y) \, \mu(x) p(x, y; \tau) \, \chi_j(x)
\]
\[
= \mathbb{E} [\chi_i(x_{t+\tau}) \chi_j(x_t)]
\]

\[
S_{ij} = \langle \varphi_i, \varphi_j \rangle_{\mu^{-1}}
\]
\[
= \int_{x \in \Omega} \chi_i(x) \chi_j(x) \mu(x)
\]
\[
= \mathbb{E} [\chi_i(x_t) \chi_j(x_t)]
\]

E. Markov State Models

Markov state models (MSMs)\cite{33,34} are special case of the proposed method that have been widely applied to the analysis of biomolecular simulations\cite{46,47}, where the basis functions are restricted to indicator functions on a collection of non-overlapping subsets of the conformation space. That is, given a set of discrete non-overlapping states which partition $\Omega$, $S = \{s_i\}_{i=1}^m$, such that $s_i \subseteq \Omega$, $\bigcup_{i=1}^m s_i = \Omega$, and $s_i \cap s_j = \emptyset$, define

\[
\chi_i^{\text{MSM}}(x_t) = \begin{cases} 
1, & \text{if } x_t \in s_i, \\
0, & \text{otherwise.} 
\end{cases}
\]

Then, estimates of the correlation matrix elements $C_{ij}$ can be trivially obtained following Eq. (38) by counting the number of observed transitions between sets $s_i$ and $s_j$, and the overlap matrix $S$ is diagonal with $S_{ii}$ equal to the stationary probabilities of the sets, $\pi_i$\cite{39}.
F. Cross Validation

As has been known since at least the early 1930s, training a statistical algorithm and evaluating its performance on the same data generally yields overly optimistic results.\textsuperscript{55} For the approach considered herein, this manifests generally as the overestimation of the eigenvalues of the propagator. Related statistical methods, such as kernel principle components analysis (kPCA)\textsuperscript{56} which also involve the approximate partial diagonalization of integral operators given a finite number of samples suffer from the same effect, which has been termed variance inflation\textsuperscript{57,58}

The problem stems from the fact that the true values of the correlation matrix $C$ and overlap matrix $S$ are not available. While we may instead calculate their maximum likelihood estimators $\hat{C}$ and $\hat{S}$ by replacing the expectation values in Eq. (38) and Eq. (41) with averages over a dataset, this retains no guarantee that maximization of $\mathcal{R} (A; \hat{C}, \hat{S})$ also maximizes $\mathcal{R} (A; C, S)$, or that $\mathcal{R} (A; \hat{C}, \hat{S}) < \sum_{i=1}^{k} \lambda_i$.

In particular, we find empirically that as the number of basis functions grows, $\max_A \mathcal{R} (A; \hat{C}, \hat{S})$ continues to increase. However, when a set of ansantz eigenvectors, $\hat{A}$, are trained by maximizing the GMRQ on a set of training data, $\hat{A} = \max_A \mathcal{R} (A; \hat{C}_{\text{train}}, \hat{S}_{\text{train}})$, and then evaluated on data which was held out during the training, the test set value of the GMRQ, $\mathcal{R} (\hat{A}; \hat{C}_{\text{test}}, \hat{S}_{\text{test}})$, tends to decrease beyond a critical number of basis functions, indicating that our proposed dimensionality reduction, $\hat{A}$, fails to generalize to new data and is overfitting the noise in the training set.

With this situation in mind, we propose the selection of a suitable set of basis functions and any appropriate regularization parameters by way of $k$-fold cross-validation\textsuperscript{59} In this procedure, the available MD trajectories are split into $k$ disjoint groups (folds). For each family of basis functions to be evaluated and each fold, a model is trained with the specified fold held out from the training set, and then scored on the held out fold. The chosen parameters are then selected to be those which maximize the mean test set generalized matrix Rayleigh quotient.
III. EXPERIMENTS

A. Double Well Potential

In order to gain intuition about the method, we begin considering one of simplest possible systems: Brownian dynamics on a double well potential. We consider a one dimensional diffusion in which a single particle evolves according to the stochastic differential equation

\[
\frac{dx_t}{dt} = -\nabla V(x_t) + \sqrt{2D}R(t)
\]  

(43)

where \( V \) is the reduced potential energy, \( D \) is the diffusion constant, and \( R(t) \) is a zero-mean delta-correlated stationary Gaussian process. For simplicity, we consider the potential

\[
V(x) = 1 + \cos(2x)
\]  

(44)

with reflecting boundary conditions at \( x = -\pi \) and \( x = \pi \). Using an Euler integrator, a time step of \( \Delta t = 10^{-3} \), and diffusion constant \( D = 10^3 \), we simulated 10 trajectories starting from \( x = 0 \) of length \( 10^5 \) steps, and saved the position every 100 steps. The potential and histogram of the resulting data points is shown in the right panel of Fig. 1. We computed the true eigenvalues of the system’s propagator to machine precision by discretizing the Fokker-Planck equation on a dense grid.

We now consider the construction of Markov state models, and in particular the selection of the number of states, \( k \). When \( k \) is too low, we expect that the discretization error in the MSM will dominate, and our basis will not be flexible enough to capture the first eigenfunction of the propagator. On the other hand, because the number of parameters in the MSM is proportional to \( k^2 \), we expect that for \( k \) too large, our models will be overfit.

We therefore split the 10 trajectories into 5 folds of two to perform 5-fold cross validation. For each trial value of \( k \), we build five different MSMs, each of which was fit using 4 of the 5 folds and then tested on the left-out fold.

Using states which divided the region between \( -\pi \) and \( \pi \) into \( N \) equal width bins, we built Markov state models for the system for a series of \( N \) between 5 and 1000. The mean GMRQ for the first two eigenvectors (stationary distribution and slowest dynamical process) of the MSMs is shown in the left panel of Fig. 1 along with the exact value of the GMRQ. The blue training curve gives the average GMRQ over the folds when scoring the models on the same trajectories that they were fit with, and is simply equal to the mean sum of the
FIG. 1. Model selection for MSMs of a double well potential. Error bars indicate standard deviations over the 5 folds of cross validation.

first two eigenvalues of the MSMs, whereas the red curve shows the mean GMRQ evaluated on the left-out test trajectories.

The training GMRQ increases monotonically, and we note with particular emphasis that it increases past the exact value when using a large number of states. This indicates that the models built with more than 200 states predict slower dynamics than the true propagator. This effect is impossible in the limit of infinite data as demonstrated by Eq. [9] – it is a direct manifestation of overfitting, and indicates why straightforward variational optimization without testing on held-out data or consideration of statistical error fails in a data-limited regime. On the other hand, the test GMRQ displays an inverted U-shaped behavior and achieves a maximum at $k = 61$. These models thus achieve the best predictive accuracy in capturing the systems slow dynamics, given the finite data available.

### B. Comparison of Clustering Procedures: Octalanine

What methods of MSM construction are most robustly able to capture the long-timescale dynamics of protein systems? To address this question, we performed a series of analyses of 27 molecular dynamics trajectories of terminally-blocked octalanine, a small helix forming peptide. We used 8 different methods to construct the state discretization using clustering with three distance metrics and three clustering algorithms.

We considered three distance metrics. The first was the backbone $\phi$ and $\psi$ dihedral angles. Each conformation was represented by the sine and cosine of these torsions for a
total of 32 features per frame, and distances for clustering were computed using a Euclidean metric. Second, we considered the DRID distance metric introduced by Zhou and Caflisch, using the \( C_\alpha \), \( C_\beta \), \( C \), \( N \), and \( O \) atoms in each residue. Finally, we considered the Cartesian minimal root mean square deviation (RMSD) using the same set of atoms per residue. We also considered three clustering algorithms, \( k \)-centers, a landmark version of UPGMA hierarchical clustering (see Appendix C), and \( k \)-means.

For each pair of distance metric and clustering algorithm (excluding \( k \)-means & RMSD which are incompatible), we performed 5-fold cross validation using between 10 and 500 states. Evaluation with the GMRQ requires prior specification of the number of eigenfunctions to monitor – the rank, \( m \) – and the lag time, \( \tau \); for this experiment we heuristically chose a lag time of \( \tau = 10 \) ps, and \( m = 6 \), to capture the first five dynamical processes in addition to the stationary distribution. The results are shown in Fig. 2, with blue curves indicating the mean GMRQ on the training set, and red curves indicating the mean performance on the held-out sets. We find that in all cases, the performance on the training set is optimistic, in the sense that the ansatz eigenvectors fit during training score more poorly when re-evaluated on held out data. Furthermore, although the training curves all continue to increase with respect to the number of states within the parameter range studied – which might be interpreted from a variational perspective as the quality of the models continually increasing – the performance on the test sets tends to peak at a moderate number of states and then decrease. We interpret this as a sign of overfitting when the number of states is too large, with models fitting the statistical noise in the dataset rather than the underlying slow dynamical modes. Of the parameters studied, the best performance appears to be using the combination of \( k \)-means clustering with the dihedral distance metric, using between 50 and 200 states. We also note that \( k \)-centers appears to yield particularly poor models for all distance metrics, which may be rationalized on the basis that, by design, the algorithm selects outlier conformations to serve as cluster centers.

C. Value of tICA preprocessing: Ubiquitin

To what extent can dimensionality reduction with tICA be used to improve the predictive quality of MSMs? While tICA itself is an estimator for the slow eigenspace of the propagator, the restriction to linear basis functions is relatively severe. Instead of simply replacing the
FIG. 2. Comparison of 8 methods for building MSMs under 5-fold cross validation, evaluated using the rank-6 GMRQ. We used the $k$-centers, $k$-means, and landmark-based ($n_{\text{landmarks}} = 5000$) UPGMA hierarchical clustering algorithms, with the DRID and backbone dihedral angle featurizations. Error bars indicate the standard error in the mean over the cross validation folds.

MSM estimator with tICA, the approach advocated by Schwantes and Pande and Pérez-Hernández et al. was instead to chain the estimators, using tICA as a preprocessing tool to coarsely identify a smaller number slow degrees of freedom, with MSMs built in this lower-dimensional metric space.

Here, we test this stepwise approach using the GMRQ under 5-fold cross-validation on simulations of the near-native state dynamics of ubiquitin. Our simulation dataset consists of $N$ independent trajectories (see Appendix A for details) each started from the 1XD3 crystal structure. We began by extracting the cosine and sine of each of the backbone $\phi$ and $\psi$ dihedral angles, and consider three types of analysis, each of which is scored by a cross-validated rank-3 GMRQ:

(a) Apply tICA directly to the extracted dihedral angles, building the optimal linear rank-3 approximation to the propagator using these coordinates.

(b) Apply MSMs directly to the torsions, clustering with $k$-means into between 10 and 1000 states.
FIG. 3. Three approximations to the rank-3 approximate propagator under 5-fold cross-validation. (a) tICA applied to the backbone dihedral angles. (b) MSMs built directly on the dihedral data using $k$-means with varying numbers of states. (c) MSMs built after tICA dimensionality-reduction. The MSMs built with tICA display significantly better performance on data that was left-out used during parameterization.

(c) Construct MSMs using the tICA dimensionality reduction, using an initial projection into a 4-dimensional subspace which is then clustered via $k$-means into between 10 and 1000 states.

The training and test performance of each of these three procedures is shown in Fig. 3. Whereas MSMs built by directly clustering the dihedral angles appear to show convergence on training performance, they display decreased and highly variable performance on held-out data. In contrast, across the range of parameters studied the MSMs built using tICA display similar performance on training data and significantly improved test-set performance. These results demonstrate that tICA can significantly aid the robust detection of slow dynamical modes in molecular kinetics.
IV. DISCUSSION

Some amount of summarization, coarse-graining or dimensionality reduction of molecular dynamics data sets is a necessary part of their use to answer questions in biological physics. In this work, we argue that the goal of this effort should essentially be to find the dominant eigenfunctions of the system’s propagator, an unknown integral operator controlling the system’s dynamics. We show that this goal can be formulated as the variational optimization of a single scalar functional, which can be approximated using trajectories obtained from simulation and a parametric basis set. Although overfitting is a concern with finite simulation data, this risk can be mitigated by the use of separate training and test sets or cross-validation.

When the basis sets are restricted to mutually-orthogonal indicator functions or linear functions of the input coordinates, this method corresponds to the existing MSM and tICA methods. Unlike previous formulations, it provides a method by which MSM and tICA solutions can be “scored” on new data sets that were not used during parameterization, making it possible to measure the generalization performance of these methods and choose the various hyperparameters required for each method, such as the number of MSM states or clustering method. Furthermore, the extension to other families of basis functions (e.g Gaussians) is straightforward, and GMRQ provides a natural quantitative basis on which to conclude whether these new methods are superior to existing basis sets.

A. Connections to quantum mechanics and machine learning

The variational principle for eigenspaces in this work has strong connections to work in two other related fields: excited state electronic structure theory in quantum mechanics and multi-class Fisher discriminant analysis in machine learning. In quantum mechanics, Theorem 1 is analogous to what has been called the ensemble or trace variational principle in that field, which bounds the sum of the energy of the first $m$ eigenstates of the Hamiltonian by the trace of a matrix of Ritz values. While the goal of finding just the ground-state eigenfunction ($m = 1$) is more common in computational quantum chemistry, the simultaneous optimization of many eigenstates is critical for many applications including band-structure calculations for materials in solid state physics.
Furthermore, in machine learning, this work has an analog in the theory multi-class Fisher
discriminant analysis. Here, the goal is to find a low-rank projection of a labeled multi-
class dataset which maximizes the between-class variance of the dataset while controlling
the within-class variances. The optimal discriminant vectors are shown to be the first \( k \)
generalized eigenvectors of an eigenproblem involving these two variance matrices – the
problem shares the same structure as Eq. (34) in this work. We anticipate that this parallel
will aid the development of improved algorithms for the identification of slow molecular
eigenfunctions, especially with respect to regularization and sparse formulations.

B. Comparison to likelihood maximization

While we focus on the identification of the dominant eigenfunctions of the system’s prop-
agator, a different viewpoint is that analysis of MD should essential entail the construction
of probabilistic, generative models over trajectories, fit for example by maximum likelihood
or Bayesian methods.

As we show in Section II E and Nüske et al. have shown earlier, MSMs arise naturally
from a maximization of Eq. (9) when the ansatz eigenfunctions are constrained to be linear
combinations of a set of mutually orthogonal indicator functions. However, MSMs can also
be viewed directly as probabilistic models, constructed by maximizing a likelihood function
of the trajectories with respect to the model parameters. This probabilistic view has, in
fact, been central to the field, driving the development of improved methods for example in
model selection, parameterization, and coarse-graining. To what extent does this imply that the variational and probabilistic views are equivalent?

In Appendix B we show that while these two views may coincide for the particular choice
of basis set with MSMs, they need not be equivalent in general. In fact, the GMRQ-optimal
model formed by the first \( m \) eigenfunctions of the propagator need not be non-negativity
preserving, which is essential to form a probabilistic likelihood function in the sense of
Kellogg, Lange, and Baker or McGibbon, Schwantes and Pande.

For MD datasets, the use probabilistic models often require preprocessing the simulation
simulation, for example by extracting a set of backbone dihedral angles or inter-atomic
distances. The models then use an emission distribution over this preprocessed data. Because the support of the likelihood function is then on the preprocessed data instead of
the the original trajectory data, none of the machinery of probability theory is applicable for central question of deciding which preprocessing protocol to employ.\textsuperscript{[71]} The procedure described herein has no such limitation, as shown in Fig. 2.

V. CONCLUSIONS

The proliferation of new and improved methods for constructing low-dimensional models of molecular kinetics given a set of high-resolution MD trajectories has been a boon to the field, but the lack of a unified theoretical framework for choosing between alternative models has hampered progress, especially for non-experts applying these methods to novel biological systems. In this work we have presented a new variational theorem governing the estimation of the space formed by the span of multiple eigenfunctions of the molecular dynamics propagator. With this method, a single scalar-valued functional scores a proposed model on a supplied data set, and the use of separate testing and training data sets makes it possible to quantify and avoid statistical overfitting. Both time-structure independent components analysis (tICA) and Markov state models (MSMs) are shown to be a specific instance of this method with different types basis functions.

We have applied this approach to compare eight different protocols for Markov state model construction on a set of MD simulations of the octalanine peptide. We find that of the methods tested, $k$-means clustering with the dihedral angles using between 50 and 200 states appears to outperform the other methods, and that the $k$-centers cluster method can be particularly prone to poor generalization performance. We also find, in a separate experiment on simulations of Ubiquitin, that preprocessing with tICA can substantially improve the generalization performance of MSMs. To our knowledge, this work is the first to enable such quantitative and theoretically well-founded comparisons of alternative parameterization strategies for MSMs.

We anticipate that this work will open the door to more complete automation and optimization of MSM construction. While the lag time, $\tau$ and rank, $m$, of the desired model must be manually specified, other key hyperparameters that control difficult-to-judge statistical tradeoffs, such as the number of states in an MSM, can be chosen be optimizing the cross-validation performance. Furthermore, given recent advances in automated hyperparameter optimization in machine learning, we anticipate that this search itself can be fully
ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation and National Institutes of Health under NIH R01-GM62868 and NSF MCB-0954714. We thank Christian R. Schwantes and Krikamol Muandet for helpful discussions, as well as Jeffrey K. Weber for access to the octalanine simulation.

Appendix A: Simulation Setups

**Octalanine** We performed all-atom molecular dynamics simulations of terminally-blocked octalanine (Ace-(Ala)$_8$-NHMe) in explicit solvent using the GROMACS 4 simulation package, the AMBER ff99SB-ILDN-NMR forcefield, and the TIP3P water model. The system was energy minimized, followed by 1 ns of equilibration using the velocity rescaling thermostat (reference temperature of 298K, time constant of 0.1 ps), Parrinello-Rahman barostat (reference pressure of 1 bar, time constant of 1 ps, isotropic compressibility of $5 \text{ bar}$), and verlet integrator (timestep of 2 fs). Production simulations were performed in the canonical ensemble using the same integrator and thermostat. Nonbonded interactions in all cases were treated with the particle mesh Ewald method, using a real space cutoff distance for Ewald summation as well as for van der Waals interactions of 10.0 Å. Twenty six such simulations were performed, with production lengths between 20 and 150 ns each. The total aggregate sampling was 1.74 µs.

**Ubiquitin** We performed all-atom molecular dynamics simulations of ubiquitin carboxyl-terminal esterase L3 (referred to in the text simply as ubiquitin) in implicit solvent using the OpenMM 6.0 simulation package, and the AMBER ff99SB-ILDDB forcefield. The system was energy-minimized starting from the protein data bank 1XD3 crystal coordinates. Simulations were performed using Langevin dynamics with a target temperature of 300 K, a friction coefficient of 91 ps$^{-1}$, and timestep of 2 fs. Bonds to hydrogen atoms were constrained, and nonbonded interactions were cut off at 10.0 Å using the reaction field method. Ten such simulations were performed, each of length 100 ns. The total aggregate sampling was 1 µs.
Appendix B: Tension between spectral and probabilistic approaches

Here we show, by way of a simple analytical example, the extent to which the variational and probabilistic approaches to the analysis of molecular dynamics data are indeed distinct. By explicitly constructing the propagator eigenfunctions for a Brownian harmonic oscillator, we show that the rank-$m$ truncated propagator, $P_m(\tau)$, built from the first $m$ eigenpairs of $P(\tau)$ is not in general a nonnegativity-preserving operator. That is, for some valid initial distributions, $p_t(x)$, the propagated distribution, $\tilde{p}_{t+\tau}^{(m)}(x) = P_m(\tau) \circ p_t(x)$, fails to be non-negative throughout $\Omega$ and thus does not represent a valid probability distribution.

$$p_{t+\tau}^{(m)}(x) \not\geq 0 \forall \, x \in \Omega \quad (B1)$$

This indicates that variational and probabilistic approaches have the potential to be almost contradictory in what they judge to be “good” models of molecular kinetics.

Consider the diffusion of a Brownian particle in the potential $U(x) = x^2$. For simplicity, we take the temperature and diffusion constant to be unity. This is an Ornstein-Uhlenbeck process, and the dynamics are described by the Smoluchowski equation,

$$\frac{\partial}{\partial t} p_t(x) = L \circ p_t(x), \quad (B2)$$

with infinitesimal generator $L$ given by

$$L = \frac{\partial^2}{\partial x^2} + 2 \frac{\partial}{\partial x} x, \quad (B3)$$

and stationary distribution $\mu(x) = \pi^{-1/2} e^{-x^2}$.

We can expand the generator in terms of its eigenfunctions, $\phi_n(x)$, and eigenvalues, $\xi_n$, defined by,

$$L \circ \phi_n(x) = \xi_n \phi_n(x), \quad (B4)$$

which can be recognized as the Hermite equation whose solutions are related to the Hermite polynomials, $H_n$. For $n = \{0, 1, \ldots\}$ the solutions are

$$\phi_n(x) = c_n e^{-x^2} H_n(x), \quad (B5)$$

$$\xi_n = -2n, \quad (B6)$$

$$c_n^2 = \left(2^n n! \pi \right)^{-1}, \quad (B7)$$
where the normalizing constants, $c_n$, are chosen such that $\langle \phi_n, \phi_m \rangle_{\mu^{-1}} = \delta_{nm}$.

The propagator $P(\tau)$ can be formed by integrating Eq. (B1) with respect to $t$, giving

$$P(\tau) = e^{\tau L}.$$  \hspace{1cm} (B8)

$P(\tau)$ shares the same eigenfunctions as $L$. Its eigenvalues, $\lambda_n$, are related to the eigenvalues of $L$ by

$$\lambda_n = e^{-\tau \xi_n}.$$  \hspace{1cm} (B9)

We now define the rank-$m$ truncated propagator, $P_m(\tau)$, such that

$$P_m(\tau) \circ p_t = \sum_{n=0}^{m-1} \lambda_n \langle p_t, \phi_n \rangle_{\mu^{-1}} \phi_n$$

$$= \sum_{n=0}^{m-1} e^{-2n \tau} c_n e^{-x^2} \left[ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} dx' c_n \sqrt{\pi} p_t(x') H_n(x') \right]$$  \hspace{1cm} (B10)

Consider an initial distribution, $p_t(x) = \delta(x - x_0)$, propagated forward in time by $P_m$. Let $\tilde{p}_r^{(m)} = P_m(\tau) \circ \delta(x - x_0)$. Then, Eq. (B11) simplifies to

$$\tilde{p}_r^{(m)}(x) = \sum_{n=0}^{m-1} \frac{1}{2^n n! \sqrt{\pi}} e^{-2n \tau} e^{-x^2} H_n(x) H_n(x_0).$$  \hspace{1cm} (B12)

Consider now the specific case of $m = 2$. Using the explicit expansion $H_0(x) = 1$, and $H_1(x) = 2x$, we have

$$\tilde{p}_r^{(2)}(x) = \frac{1}{\sqrt{\pi}} e^{-x^2} \left( 1 + 2xx_0 e^{-2r} \right).$$  \hspace{1cm} (B13)

Note that Eq. (B13) has a zero when $x = -e^{2r}/2x_0$, and that

$$\tilde{p}_r^{(2)}(x) < 0 \quad \text{when} \quad \begin{cases} x < -e^{2r}/2x_0 & \text{if } x_0 > 0 \\ x > -e^{2r}/2x_0 & \text{if } x_0 < 0. \end{cases}$$  \hspace{1cm} (B14)

Because of this non-positivity, $\tilde{p}_r^{(2)}(x)$ is not a valid probability distribution.

This example demonstrates that the rank-$m$ truncated propagator need not, in general, preserve the positivity of distributions it acts on. Therefore, if such a model of the dynamics are fit or assessed via maximum-likelihood methods on datasets consisting of observed transitions, despite being optimal by spectral norm, the true rank-$m$ truncated propagator may appear to give a log likelihood of $-\infty$. The variational and probabilistic approaches to modeling molecular kinetics can indeed be very different.
Appendix C: Landmark UPGMA Clustering

Landmark-based UPGMA (Unweighted Pair Group Method with Arithmetic Mean) agglomerative clustering is a simple scalable hierarchical clustering which does not require computing the full matrix of pairwise distances between all data points. The procedure first subsamples \( l \) “landmark” data points at regular intervals from the input data. These data points are then clustered using the standard algorithm, resulting in \( n \) clusters. Let \( S_n \) be the set of landmark data points assigned by the algorithm to the cluster \( n \), and \( d(x, x') \) be the distance metric employed. Then, each remaining data point in the training set as well as new data points from the test set, \( x^* \), are assigned to cluster, \( s(x^*) \in \{1, \ldots, n\} \), whose landmarks they are on average closest to:

\[
    s(x^*) = \arg\min_n \frac{1}{|S_n|} \sum_{x \in S_n} d(x^*, x).
\]  

(C1)

REFERENCES

1. K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl, Annu. Rev. Biophys. 37, 289 (2008).
2. M. Huse and J. Kuriyan, Cell 109, 275 (2002).
3. E. Vargas, V. Yarov-Yarovoy, F. Khalili-Araghi, W. A. Catterall, M. L. Klein, M. Tarek, E. Lindahl, K. Schulten, E. Perozo, F. Bezanilla, et al., J. Gen. Physiol. 140, 587 (2012).
4. A. H. Phillips, Y. Zhang, C. N. Cunningham, L. Zhou, W. F. Forrest, P. S. Liu, M. Steffek, J. Lee, C. Tam, E. Helgason, et al., Proc. Natl. Acad. Sci. U.S.A. 110, 11379 (2013).
5. G. Careri, P. Fasella, E. Gratton, and W. Jencks, Crit. Rev. Biochem. Mol. Biol. 3, 141 (1975).
6. H. B. Buergi and J. D. Dunitz, Acc. Chem. Res. 16, 153 (1983).
7. A. Mittermaier and L. E. Kay, Science 312, 224 (2006).
8. L.-P. Wang, J. Chen, and T. Van Voorhis, J. Chem. Theory Comput. 9, 452 (2012).
9. L. Huang and B. Roux, J. Chem. Theory Comput. 9, 3543 (2013).
10. J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio Jr, et al., J. Phys. Chem. B 114, 2549 (2010).
11. R. B. Best, X. Zhu, J. Shim, P. E. Lopes, J. Mittal, M. Feig, and A. D. MacKerell Jr, J. Chem. Theory Comput. 8, 3257 (2012).
12P. E. Lopes, J. Huang, J. Shim, Y. Luo, H. Li, B. Roux, and A. D. MacKerell Jr, J. Chem. Theory Comput. 9, 5430 (2013).

13J. E. Stone, D. J. Hardy, I. S. Ufimtsev, and K. Schulten, J. Mol. Graphics Modell. 29, 116 (2010).

14D. E. Shaw, R. O. Dror, J. K. Salmon, J. P. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, M. M. Deneroff, B. Batson, K. J. Bowers, E. Chow, M. P. Eastwood, D. J. Ierardi, J. L. Klepeis, J. S. Kuskin, R. H. Larson, K. Lindorff-Larsen, P. Maragakis, M. A. Moraes, S. Piana, Y. Shan, and B. Towles, in Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis (2009) pp. 39:1–39:11.

15M. Shirts and V. S. Pande, Science 10, 43 (2006).

16I. Buch, M. J. Harvey, T. Giorgino, D. Anderson, and G. De Fabritiis, J. Chem. Inf. Model. 50, 397 (2010).

17K. J. Kohlhoff, D. Shukla, M. Lawrenz, G. R. Bowman, D. E. Konerding, D. Belov, R. B. Altman, and V. S. Pande, Nature Chem. 6, 15 (2014).

18J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, and W. C. Swope, J. Chem. Phys. 126, 155101 (2007).

19P. Deuflhard and M. Weber, Linear Algebra Appl. 398, 161 (2005).

20M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi, J. Chem. Phys. 134, 124116 (2011).

21A. Altis, P. H. Nguyen, R. Hegger, and G. Stock, J. Chem. Phys. 126, 244111 (2007).

22P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi, Proc. Natl. Acad. Sci. U.S.A. 103, 9885 (2006).

23S. V. Krivov and M. Karplus, Proc. Natl. Acad. Sci. U.S.A. 101, 14766 (2004).

24E. Weinan and E. Vanden-Eijnden, J. Stat. Phys. 123, 503 (2006).

25I. Jolliffe, Principal component analysis (Wiley Online Library, 2005).

26D. L. Theobald, Acta Crystallogr., Sect. A: Found. Crystallogr. 61, 478 (2005).

27E. H. Kellogg, O. F. Lange, and D. Baker, J. Phys. Chem. B 116, 11405 (2012).

28T. Zhou and A. Caflisch, J. Chem. Theory Comput. 8, 2930 (2012).

29R. T. McGibbon and V. S. Pande, J. Chem. Theory Comput. 9, 2900 (2013).

30R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. S. Shakhnovich, J. Chem. Phys. 108, 334 (1998).

31C. R. Schwantes and V. S. Pande, J. Chem. Theory Comput. 9, 2000 (2013).
G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, J. Chem. Phys. 139, 015102 (2013).

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

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

V. S. Pande, K. Beauchamp, and G. R. Bowman, Methods 52, 99 (2010).

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

J. D. Chodera and F. F. Noé, Curr. Opin. Struct. Biol. 25, 135 (2014).

F. Noé and F. Nüske, Multiscale Model. Simul. 11, 635 (2013).

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

E. Schmidt, Math. Ann. 63, 433 (1907).

R. Courant and D. Hilbert, Methods of Mathematical Physics Methods of Mathematical Physics No. v. 1 (Wiley, 2008).

C. A. Micchelli and A. Pinkus, J. Approx. Theory 24, 51 (1978).

K. Fan, Proc. Natl. Acad. Sci. U.S.A. 35, 652 (1949).

M. L. Overton and R. S. Womersley, SIAM J. Matrix Anal. Appl. 13, 41 (1992).

P.-A. Absil, R. Mahony, R. Sepulchre, and P. Van Dooren, SIAM Rev. 44, 57 (2002).

D. Sezer, J. H. Freed, and B. Roux, J. Phys. Chem. B 112, 11014 (2008).

S. Muff and A. Caffisch, J. Chem. Phys. 130, 125104 (2009).

I. Buch, T. Giorgino, and G. De Fabritiis, Proc. Natl. Acad. Sci. U.S.A. 108, 10184 (2011).

V. A. Voelz, G. R. Bowman, K. Beauchamp, and V. S. Pande, J. Am. Chem. Soc. 132, 1526 (2010).

K. A. Beauchamp, R. T. McGibbon, Y.-S. Lin, and V. S. Pande, Proc. Natl. Acad. Sci. U.S.A. 109, 17807 (2012).

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

S. K. Sadiq, F. Noé, and G. De Fabritiis, Proc. Natl. Acad. Sci. U.S.A. 109, 20449 (2012).

O. P. Choudhary, A. Paz, J. L. Adelman, J.-P. Colletier, J. Abramson, and M. Grabe, Nat. Struct. Mol. Biol. 21, 626 (2014).

D. Shukla, Y. Meng, B. Roux, and V. S. Pande, Nat. Commun. 5, (2014).

S. C. Larson, J. Educ. Psychol. 22, 45 (1931).
56 B. Schölkopf, A. Smola, and K.-R. Müller, Neural Comput. 10, 1299 (1998).
57 U. Kjems, L. K. Hansen, and S. C. Strother, in Advances in Neural Information Processing Systems 13 (2000) pp. 549–555.
58 T. J. Abrahamsen and L. K. Hansen, J. Mach. Learn. Res. 12, 2027 (2011).
59 S. Arlot and A. Celisse, Statist. Surv. 4, 40 (2010).
60 K. A. Beauchamp, G. R. Bowman, T. J. Lane, L. Maibaum, I. S. Haque, and V. S. Pande, J. Chem. Theory Comput. 7, 3412 (2011).
61 S. Lloyd, IEEE Trans. Inform. Theory 28, 129–137 (1982).
62 M. Senne, B. Trendelkamp-Schroer, A. S. Mey, C. Schütte, and F. Noé, J. Chem. Theory Comput. 8, 2223 (2012).
63 A. K. Theophilou, J. Phys. C 12, 5419 (1979).
64 E. K. U. Gross, L. N. Oliveira, and W. Kohn, Phys. Rev. A 37, 2805 (1988).
65 N. Gidopoulos, P. Papaconstantinou, and E. Gross, Physica B 318, 328 (2002).
66 R. Lai, J. Lu, and S. Osher, ArXiv e-prints (2014).
67 C. R. Rao, J. R. Stat. Soc. Ser. B Stat. Methodol. 10, pp. 159 (1948).
68 G. Baudat and F. Anouar, Neural Comput. 12, 2385 (2000).
69 L. Clemmensen, T. Hastie, D. Witten, and B. Ersbøll, Technometrics 53, (2011).
70 B. K. Sriperumbudur, D. Torres, and G. Lanckriet, Mach. Learn. 85, 3 (2011).
71 R. T. McGibbon, C. R. Schwantes, and V. S. Pande, J. Phys. Chem. B 118, 6475 (2014).
72 E. K. Rains and H. C. Andersen, J. Chem. Phys. 133, 144113 (2010).
73 S. Bacallado, J. D. Chodera, and V. Pande, J. Chem. Phys. 131, 045106 (2009).
74 G. R. Bowman, J. Chem. Phys. 137, 134111 (2012).
75 R. T. McGibbon, B. Ramsundar, M. M. Sultan, G. Kiss, and V. S. Pande, in Proceedings of the 31st International Conference on Machine Learning, Vol. 32 (Beijing, China, 2014) pp. 1197–1205.
76 J. Snoek, H. Larochelle, and R. P. Adams, in Advances in Neural Information Processing Systems, Vol. 25 (Lake Tahoe, USA, 2012) pp. 2951–2959.
77 B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, J. Chem. Theory Comput. 4, 435 (2008).
78 D.-W. Li and R. Brüschweiler, Angew. Chem. 122, 6930 (2010).
79 W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. 79, 926 (1983).
80 G. Bussi, D. Donadio, and M. Parrinello, J. Chem. Phys. 126, 014101 (2007).
81 M. Parrinello and A. Rahman, J. Appl. Phys. 52, 7182 (1981).
82 T. Darden, D. York, and L. Pedersen, J. Chem. Phys. 98, 10089 (1993).
83 P. Eastman, M. S. Friedrichs, J. D. Chodera, R. J. Radmer, C. M. Bruns, J. P. Ku, K. A. Beaucamp, 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).
84 K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, Proteins: Struct., Funct., Bioinf. 78, 1950 (2010).
85 S. Misaghi, P. J. Galardy, W. J. N. Meester, H. Ovaa, H. L. Ploegh, and R. Gaudet, J. Biol. Chem. 280, 1512 (2005).
86 I. G. Tironi, R. Sperb, P. E. Smith, and W. F. van Gunsteren, J. Chem. Phys. 102, 5451 (1995).
87 D. Müllner, J. Stat. Softw. 53, 1 (2013).