Variational cross-validation of slow dynamical modes in molecular kinetics

Robert T. McGibbon and Vijay S. Pande

Department of Chemistry, Stanford University, Stanford, California 94305, USA

Markov state models are a widely used method for approximating the eigenspectrum of the molecular dynamics propagator, yielding insight into the long-timescale statistical kinetics and slow dynamical modes of biomolecular systems. However, 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. Here, we consider cross-validation with a new objective function for estimators of these slow dynamical modes, a generalized matrix Rayleigh quotient (GMRQ), which measures the ability of a rank-m projection operator to capture the slow subspace of the system. It is shown that a variational theorem bounds the GMRQ from above by the sum of the first m eigenvalues of the system's propagator, but that this bound can be violated when the requisite matrix elements are estimated subject to statistical uncertainty. This overfitting can be detected and avoided through cross-validation. These result make it possible to construct Markov state models for protein dynamics in a way that appropriately captures the tradeoff between systematic and statistical errors. © *2015 AIP Publishing LLC.* [http://dx.doi.org/10.1063/1.4916292]

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 involves not only just the structure of a unique single conformation but also 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 nuclear magnetic resonance spectroscopy—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 has 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, a MD trajectory, is a time series of Cartesian positions (and perhaps momenta) of dimension $3N$ ($6N$ if momenta 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 high-dimensional. 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.[18–24]

In this work, we combine two central ideas from machine learning and chemical physics—hyperparameter selection via cross-validation and variational approaches for linear operator eigenproblems—to create a new method for discriminating between alternative simplified models for molecular kinetics constructed from MD simulations. Towards this end, we prove a new variational theorem concerning the simultaneous approximation of the first m eigenfunctions of the propagator of a high-dimensional reversible stochastic dynamical system, which mathematically formalize the slow collective dynamical motions we wish to identify in molecular systems.

II. CROSS VALIDATION

In seeking to estimate the slowest collective dynamical modes of a molecular system from a finite set of stochastic trajectories, statistical noise is unavoidable. These dynamical

modes, which we formally identify as the first m eigenfunctions, ϕ_i, of the propagator, an integral operator associated with the dynamics of a molecular system (vide infra), are functions on $\mathbb{R}^{3N}$ to $\mathbb{R}$. Like the ground state wave function in quantum chemistry, these eigenfunctions can only be approximately represented in any finite basis set. Reducing this approximation error, a statistical bias, motivates the use of larger and more flexible basis sets. Unfortunately, in an effect known as the *bias-variance tradeoff*,[25,26] larger basis sets tend to exacerbated a competing source of error, the model variance: with a fixed simulation data set but additional adjustable parameters due to a larger basis set, the model estimates of these eigenfunctions become more unstable and uncertain.

As has been known since at least the early 1930s, training a statistical algorithm and evaluating its performance on the same data set generally yield overly optimistic results.[27] For this reason, standard practice in supervised machine learning is to divide a data set into separate training and test sets. The model parameters are estimated using the training data set, but performance is evaluated separately by scoring the now-trained model on the separate test set, consisting of data points that were left out during the training phase. To avoid overfitting, the choice between alternative models is made using test set performance not training set performance.

However, because researchers may expend significant effort to collect data sets, the exclusion of a large fraction of the data set from the training phase can be a costly proposition. k-fold cross-validation is one alternative that can be more data-economical, where the data are split into k disjoint subsets, each of which is cycled as both the training and test sets.

Let X be a collection of molecular dynamics trajectories (the data set), which we assume for simplicity to consist of a multiple of k independent MD trajectories of equal length. In k-fold cross validation, the trajectories are split into k equally sized disjoint subsets, called folds, denoted $X^{(i)}$, for $i \in \{1, 2, \ldots, k\}$. These will serve as the test sets. Let $X^{(-i)} = X \setminus X^{(i)}$ denote the set of all trajectories excluded from fold i; these will serve as the training sets.

Consider an algorithm to estimate the m slowest dynamical modes of the system, g. Examples of such estimators include Markov state models (MSMs)[28] and time-structured independent components analysis (tICA).[29,30] The result of this calculation, the estimated eigenfunctions, $\hat{\phi}_{1:m}$, is taken to be a function of both an input dataset, X, and a set of hyperparameters, θ, which may include settings such as the number of states or clustering algorithm in a MSM or the basis set used in tICA,

$$\hat{\phi}_{1:m} = (\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_m) = g(X, \theta). \qquad (1)$$

Furthermore, consider an objective function, $O(\hat{\phi}_{1:m}, X')$, which evaluates a set of proposed eigenfunctions, $\hat{\phi}_{1:m}$, and a (possibly new) dataset X', returning a single scalar measuring the performance or accuracy of these eigenfunctions. The mean cross validation performance of a set of hyperparameters is defined by the following expression, which builds k models on each of the training sets and scores them on the corresponding test sets,

$$MCV(\theta) = \frac{1}{k} \sum_{i=1}^{k} O(g(X^{(-i)}, \theta), X^{(i)}). \qquad (2)$$

Model selection can be performed by finding the hyperparameters, $\theta^* = \arg\max_\theta MVC(\theta)$, which maximize the cross validation performance. Many variants of this protocol with different procedures for splitting the training and test sets, such as repeated random subsampling cross-validation and leave-one-out cross validation, are also possible.[31]

The remainder of this work seeks to develop a suitable objective function, O, for estimates of the slow dynamical modes in molecular kinetics that can be used as shown above in a cross-validation protocol. This task is complicated by the fact that no ground-truth values of true eigenfunctions, ϕ_i, are available either in the training or test sets. Nevertheless, our goal is to construct a *consistent* objective function, such that as the size of a molecular dynamics data set, X, grows larger, the maximizer of $O(\cdot, X)$ converges in probability to the true propagator eigenfunctions, $\phi_{1:m}$,

$$\arg\max_{\hat{\phi}_{1:m}} O(\hat{\phi}_{1:m}, X) \xrightarrow{p} \phi_{1:m}. \qquad (3)$$

III. THEORY BACKGROUND

We begin by introducing the notion of the propagator and its eigenfunctions 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.*[28] which contains a longer and more thorough discussion.

Consider a time-homogeneous, ergodic, continuous-time Markov process $\mathbf{x}(t) \in \Omega$ which is reversible with respect to a stationary distribution $\mu(\mathbf{x}) : \Omega \to \mathbb{R}^+$. Where necessary for concreteness, we take the phase space, Ω, to be $\mathbb{R}^{3N}$, where N is the number of atoms of a molecular system. The system's stochastic evolution over an interval $\tau > 0$ is described by a transition probability density

$$p(\mathbf{x}, \mathbf{y}; \tau)d\mathbf{y} = \mathbb{P}[\mathbf{x}(t + \tau) \in B_\epsilon(\mathbf{y}) \mid \mathbf{x}(t) = \mathbf{x}], \qquad (4)$$

where $B_\epsilon(\mathbf{y})$ is the open ϵ-ball centered at $\mathbf{y}$ with infinitesimal measure $d\mathbf{y}$.

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

$$p_{t+\tau}(\mathbf{y}) = \int_\Omega d\mathbf{x}\, p(\mathbf{x}, \mathbf{y}; \tau)\, p_t(\mathbf{x}) = \mathcal{P}(\tau) \circ p_t(\mathbf{y}), \qquad (5)$$

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

$$\mathcal{P}(\tau) \circ \phi_i = \lambda_i \phi_i. \qquad (6)$$

Furthermore, due to the reversibility of the underlying dynamics, $\mathcal{P}(\tau)$ is compact and self-adjoint with respect to

a μ^{-1} weighted scalar product,[32]

$$\langle f, g \rangle_{\mu^{-1}} = \int_\Omega d\mathbf{x} \, f(\mathbf{x}) g(\mathbf{x}) \mu^{-1}(\mathbf{x}), \tag{7}$$

where f and g are arbitrary scalar functions on Ω. The propagator has a unique largest eigenvalue $\lambda_1 = 1$ with corresponding eigenfunction $\phi_1(\mathbf{x}) = \mu(\mathbf{x})$. The remaining eigenvalues are real and positive, can be sorted in descending order, and can be normalized to be μ^{-1}-orthonormal. Using the spectral decomposition of $\mathcal{P}(\tau)$, the conformational distribution of an ensemble at arbitrary multiples of τ can be written as a sum of exponentially decaying relaxation processes,

$$p_{t+k\tau}(\mathbf{x}) = \sum_{i=1}^{\infty} \lambda_i^k \langle p_t, \phi_i \rangle_{\mu^{-1}} \phi_i \tag{8}$$

$$= \mu(\mathbf{x}) + \sum_{i=2}^{\infty} \exp\left(-\frac{k\tau}{t_i}\right) \langle p_t, \phi_i \rangle_{\mu^{-1}} \phi_i, \tag{9}$$

where $t_i = -\dfrac{\tau}{\ln \lambda_i}$. The eigenfunctions $\phi_i(\mathbf{x})$ for $i = 2, \dots$ can thus be interpreted as dynamical modes, along each of which the system relaxes towards equilibrium with a characteristic timescale, τ_i. Many molecular systems are characterized by m individual *slow* timescales with eigenvalues close to one, separated from the remaining eigenvalues by a *spectral gap*. These slowest collective degrees of freedom, such as protein folding coordinates and pathways associated with enzyme activation/deactivation, are often identified with key functions in biological systems. The remaining small eigenvalues correspond to faster dynamical processes that rapidly decay to equilibrium. 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 to the propagator.

Furthermore, not only do these slow eigenfunctions form a convenient basis but also 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. (8) at its first m terms is the closest possible rank-m approximation to $\mathcal{P}(\tau)$ 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 λ_i and associated eigenfunctions ϕ_i are sorted in descending order by eigenvalue. Define the rank-m operator $\mathcal{P}_m$ such that $\mathcal{P}_m \circ f = \sum_{i=1}^{m} \lambda_i \langle f, \phi_i \rangle_{\mu^{-1}} \phi_i$ and let $\mathcal{A}_m$ be an arbitrary rank m operator. Then,*

$$\mathcal{P}_m = \operatorname*{argmin}_{\text{rank}(\mathcal{A}_m) \leq m} \| \mathcal{A}_m - \mathcal{P} \|_{\mu^{-1}}. \tag{10}$$

Proof. This is the extension of the familiar Eckart-Young theorem to self-adjoint linear operators. The original result is by Schmidt.[33] See Courant and Hilbert (pp. 161)[34] and Micchelli and Pinkus[35] for further details.   □

While Theorem 1 is a statement about operator approximation, it can also be viewed as a statement about optimal dimensionality reduction for description of slow dynamics.

Over all m-dimensional dimensionality reductions, the one which projects the dynamics onto its first m propagator eigenfunctions preserves to the largest degree information about the long-timescale evolution of the original system.

Note however that rank-constrained propagator, $\mathcal{P}_m$, while optimal by spectral norm is not generally positivity-preserving, as proved in Appendix B, which is an important property of the propagator necessary for its probabilistic interpretation in Eq. (5).

IV. OBJECTIVE FUNCTION AND SUBSPACE VARIATIONAL PRINCIPLE

In this section, we introduce the objective function discussed abstractly in Sec. II. We show that both the existing tICA[29,30] and MSM[28,36–39] methods can be interpreted as procedures which directly optimize this criteria using different restricted families of basis functions. Furthermore, we show that in the infinite-data limit, when the requisite matrix elements can be computed without error, a variational bound governs this objective function: *ansatz* eigenfunctions, $\hat{\phi}_{1:m}$, which differ from the propagator's true eigenfunctions, $\phi_{1:m}$, are always assigned a score which is less than the score of the true eigenfunctions.

Unfortunately, in the more typical finite-data regime, this variational bound can be violated in a pernicious manner: as the size of the basis set increases past some threshold, models can give continually increasing training set performance (even breaking the variational bound), even as they get *less* accurate when measured on independent test data sets. This observation underscores the practical value of cross-validation in estimators for the slow dynamical processes in molecular kinetics.

Our results build on the important contributions of Noé and Nüske[40] and Nüske *et al.*,[41] who introduced a closely related variational approach for characterizing the slow dynamics in molecular systems. Our novel contribution stems from an approach to the problem through the lens of cross-validation, with its need for a single scalar objective function. While previous work focuses on estimators of each of the propagator eigenfunctions, ϕ_i, one at time, we focus instead on measures related to the simultaneous estimation of all of the first m eigenfunctions collectively.

Theorem 2. *Let $\mathcal{P}$ be compact linear operator whose eigenvalues $\lambda_1 > \lambda_2 \geq \lambda_3, \dots$ 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 $\Omega \to \mathbb{R}$, $f = \{f_i(\cdot)\}_{i=1}^{m}$. Let $\mathbb{S}^m$ and $\mathbb{S}_{++}^m$ be the space of $m \times m$ real symmetric matrices and positive definite matrices, respectively. Define a matrix $P \in \mathbb{S}^m$ with $P_{ij} = \langle f_i, \mathcal{P} \circ f_j \rangle_{\mu^{-1}}$ and a matrix $Q \in \mathbb{S}_{++}^m$ with $Q_{ij} = \langle f_i, f_j \rangle_{\mu^{-1}}$. Define $\mathcal{R}_{\mathcal{P}}[f]$ as*

$$\mathcal{R}_{\mathcal{P}}[f] = \operatorname{Tr}(P Q^{-1}). \tag{11}$$

Then,

$$\mathcal{R}_{\mathcal{P}}[f] \leq \sum_{i=1}^{m} \lambda_i. \tag{12}$$

Lemma 3. The equality in Eq. (12) holds for $f = \{\phi_1, \phi_2, \ldots, \phi_m\}$ and any set of m functions, f, such that span(f) = span$(\{\phi_1, \phi_2, \ldots, \phi_m\})$.

The proof of Theorem 2 follows from the Ky Fan theorem.[42,43] Its proof, as well as the proof of Lemma 3, can be found in Appendix A.

This result implies that the slow eigenspace of the molecular propagator can be numerically determined by simultaneously varying a set of *ansatz* functions f to maximize $\mathcal{R}_{\mathcal{P}}[f]$. If the maxima is found, then f are the desired eigenfunctions up to a rotation. The matrix P has the form of a time-lagged covariance matrix between the *ansatz* functions, describing the tendency of the system to move from regions of phase space strongly associated one *ansatz* function to another in time τ. The matrix Q acts like a normalization, giving the equilibrium overlap between *ansatz* functions. Note that when the trial functions, f, are μ^{-1}-orthonormal, Q is simply the identity. Under these conditions, $\mathcal{R}_{\mathcal{P}}[f]$ then assumes a simple form as the sum of the individual Ritz values of the trial functions.

Physically, $\mathcal{R}_{\mathcal{P}}[f]$ can be interpreted as the "slowness" of the lower-dimensional dynamical process formed by projecting a high-dimensional process through the *m ansatz* functions. The maximization of $\mathcal{R}_{\mathcal{P}}[f]$ corresponds to a search for the coordinates along which the system decorrelates as slowly as possible.

Because it is bounded by the sum of the first *m* true eigenfunctions of the propagator, $\mathcal{R}_{\mathcal{P}}[f]$ is the foundation of the sought objective function for cross-validation of estimators for the slow dynamical modes in molecular kinetics. Unfortunately, it cannot be calculated exactly from a molecular dynamics simulation. Next, we show how the requisite matrix elements, P_{ij} and Q_{ij}, can be approximated from MD. The noise in these approximations will be a function of both the amount of available simulation data and the size and flexibility of a basis set, leading to the bias variance tradeoff discussed earlier. By the continuous mapping theorem and Theorem 2, consistency of the objective function (in the sense of Eq. (3)) is established if these estimators for P and Q are consistent.

A. Basis function expansion

Equipped with this variational theorem, 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. This reduces the problem of searching over the space of sets of *m* functions to a problem of finding a weight matrix that linearly mixes the basis functions.

Let $\{\varphi_a\}_{a=1}^n$ be a set of *n* functions on $\Omega \to \mathbb{R}$, which will be used as basis functions in which to expand the slowest *m* propagator eigenfunctions. Physically motived basis functions for protein dynamics might include measurements of protein backbone dihedral angles, the distances between particular atoms, or similar structural coordinates. The basis can also be indicator functions for specific regions of phase space—the "Markov states" in a MSM.

Following Nüske *et al.*,[41] we expand the *m ansatz* eigenfunctions as μ-weighted linear combinations of the basis

functions, $f_i(\mathbf{x}) = \sum_a A_{ia}\mu(\mathbf{x})\varphi_a(\mathbf{x})$, where $A \in \mathbb{R}^{n\times m}$ is a weight matrix of arbitrary expansion coefficients. From the basis functions, we define the time-lagged covariance and overlap matrices $C \in \mathbb{S}^n$ and $S \in \mathbb{S}_{++}^n$, respectively, such that $C_{ij} = \langle \mu\varphi_i, \mathcal{P} \circ \mu\varphi_j \rangle_{\mu^{-1}}$ and $S_{ij} = \langle \mu\varphi_i, \mu\varphi_j \rangle_{\mu^{-1}}$.

Then, by exploiting the linearity of the basis function expansion, the matrices P and Q can be written as matrix products involving the expansion coefficients, correlation, and overlap matrices,

$$P = A^T C A, \tag{13}$$

$$Q = A^T S A. \tag{14}$$

These equations can be interpreted in a simple way: the time-lagged correlation and overlap of the *ansatz* functions with respect to one another can be formed from two similar matrices involving only the basis functions, C and S, and the expansion coefficients, A. When the *ansatz* functions, f, are constructed this way, $\mathcal{R}_{\mathcal{P}}[f]$ reduces to the generalized matrix Rayleigh quotient (GMRQ), $\mathcal{R}_{\mathcal{P}}[f] = \mathcal{R}(A; C, S) = \mathcal{R}(A)$,

$$\mathcal{R}(A) \equiv \text{Tr}\big(A^T C A (A^T S A)^{-1}\big). \tag{15}$$

Following Lemma 3, we note 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 Ω. This space is referred to as a *Grassmann manifold*.[44] Note that when $m = 1$, P and Q are scalars and $\mathcal{R}(A)$ reduces to a standard generalized Rayleigh quotient.

Furthermore, with fixed basis functions, 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[29,30] and the Ritz method.[40]

B. Estimation of matrix elements from MD

Equipped with a collection of basis functions, $\{\varphi\}$, how can C and S be estimated from a MD dataset? As previously shown by Nüske *et al.*,[41] the matrix elements C and S can be estimated from an equilibrium molecular dynamics simulations, $X = \{\mathbf{x}_t\}_{t=1}^T$, by exploiting the ergodic theorem and measuring the correlation between the basis functions, with or without a time lag,

$$C_{ij} = \langle \mu\varphi_i, \mathcal{P}(\tau) \circ \mu\varphi_j \rangle_{\mu^{-1}} \tag{16}$$

$$= \int_{x\in\Omega} \int_{y\in\Omega} d\mathbf{x}\, d\mathbf{y}\, \mu(\mathbf{y})\, \varphi_i(\mathbf{y})\, p(\mathbf{x}, \mathbf{y}; \tau)\, \varphi_j(\mathbf{x})$$
$$\tag{17}$$

$$\approx \frac{1}{T - \tau} \sum_{t=1}^{T-\tau} \varphi_i(\mathbf{x}_t)\varphi_j(\mathbf{x}_{t+\tau}), \tag{18}$$

$$S_{ij} = \langle \mu\varphi_i, \mu\varphi_j \rangle_{\mu^{-1}} \tag{19}$$

$$= \int_{x\in\Omega} \varphi_i(\mathbf{x})\varphi_j(\mathbf{x})\mu(\mathbf{x}) \tag{20}$$

$$\approx \frac{1}{T} \sum_{t=1}^T \varphi_i(\mathbf{x}_t)\varphi_j(\mathbf{x}_t). \tag{21}$$

Note that for Theorem 2 to be applicable, C is required to be symmetric, a property which is likely to be violated by the estimator Eq. (18). For this reason, in practice, we use an estimator that averages the matrix computed in Eq. (18) with its transpose. We call this method as transpose symmetrization, and it amounts to including each trajectory twice in the dataset, once in the forward and once in the reversed direction, as discussed by Schwantes and Pande.[29]

MSMs[28,36–39] are particular case of the proposed method, which have been widely applied to the analysis of biomolecular simulations,[45–53] where the basis functions are chosen to indicator functions on a collection of non-overlapping subsets of the conformation space. Given a set of discrete non-overlapping states which partition Ω, $\mathcal{S} = \{s_i\}_{i=1}^n$, such that $s_i \subseteq \Omega$, $\bigcup_{i=1}^n s_i = \Omega$ and $s_i \cap s_j = \emptyset$, and define

$$\varphi_i^{\text{MSM}}(\mathbf{x}_t) = \begin{cases} 1, & \text{if } \mathbf{x}_t \in s_i. \\ 0, & \text{otherwise.} \end{cases} \tag{22}$$

Using this basis set, as previously shown by Nüske et al.,[41] estimates of the correlation matrix elements C_{ij} can be obtained following Eq. (18) by counting the number of observed transitions between sets s_i and s_j. The overlap matrix S is diagonal with entries, S_{ii}, that estimate the stationary probabilities of the sets, $S_{ii} \approx \int_{\mathbf{x} \in s_i} d\mathbf{x}\, \mu(\mathbf{x})$.

For the particular case of MSM basis sets, in contrast to the somewhat crude transpose symmetrization method, a more elegant enforcement of symmetry of C can be accomplished via a maximum likelihood estimator following Algorithm 1 of Prinz et al.[28]

Equipped with these estimators for C and S from MD data, Eq. (15) now has a form which is suitable for use as a cross-validation objective function, $O(\hat{\phi}_{1:m}, X')$. The proposed eigenfunctions, which may have been trained on a *different* dataset, are numerically represented by expansion coefficients, $\hat{A}$, and C and S act as sufficient statistics from the test dataset X'; the GMRQ objective function is $\mathcal{R}(\hat{A}; C(X'), S(X'))$.

V. ALGORITHMIC REALIZATION

The central practical purpose of cross-validation with GMRQ is, given a MD dataset, to select a set of appropriate basis functions with which to construct MSMs for system's kinetics. Note that Eq. (22) leaves substantial flexibility in the definition of the basis set, since the partitioning of phase space into states is left unspecified.

Methodologies for constructing these states include clustering the conformations in the dataset using a variety of distance metrics, clustering algorithms, and dimensionality reduction techniques. Let θ be a variable containing the settings for these procedures, which parameterizes a function, $g_\theta^{\text{MSM}}(X)$, that, given a collection of MD trajectories, returns a set of n states, $\mathcal{S}$.

Procedurally, GMRQ-based cross-validation for MSMs is a protocol for assigning a scalar score, $MCV(\theta)$, to the MSM hyperparameters, θ, with the following steps.

1. Separate the full data set into k disjoint folds, as described in Sec. II.

2. For each fold, i, use the training data set, $X^{(-i)}$, to construct a set of states, $\mathcal{S}^{(-i)} = g_\theta(X^{(-i)})$.

3. Use the states $\mathcal{S}^{(-i)}$ and the training data set $X^{(-i)}$ to build a Markov state model. This entails clustering the dataset to obtain the basis functions (states), $\{\varphi\}$, estimating the training set correlation and overlap matrices $C^{(-i)}$ and $S^{(-i)}$ from the trajectories, and computing their first m generalized eigenvectors, $A = \arg\max_A \mathcal{R}(A; C^{(-i)}, S^{(-i)})$, with a standard generalized symmetric eigensolver (e.g., LAPACK's ᴅsʏɢv subroutine).[54]

4. These eigenvectors maximize the GMRQ on the training set, but how do they perform when tested on new data? Using the test set data, $X^{(i)}$, and the states, $\mathcal{S}^{(-i)}$, *as determined from the training set*, compute the test set correlation and overlap matrices, $C^{(i)}$ and $S^{(i)}$. These trained eigenvectors, A, are scored on the test set by $\mathcal{R}(A; C^{(i)}, S^{(i)})$. The key metric for model selection, the cross-validation mean test set generalized matrix Rayleigh quotient is

$$MVC(\theta) = k^{-1} \sum_{i=1}^k \mathcal{R}(A;\, C^{(i)}, S^{(i)}). \tag{23}$$

As an overfitting diagnostic, we also calculate the cross-validation mean training set GMRQ,

$$MVC'(\theta) = k^{-1} \sum_{i=1}^k \mathcal{R}(A; C^{(-i)}, S^{(-i)}). \tag{24}$$

5. Finally, the entire procedure is repeated for many choices of θ, and the hyperparameter set that maximizes the mean cross validation score is chosen as the best model, $\theta^\star = \arg\max_\theta MVC(\theta)$.

For this approach, one symptom of overfitting—the construction of models that describe the statistical noise in X rather than the underlying slow dynamical processes—is an overestimation of the eigenvalues of the propagator and overestimation of the GMRQ. Related statistical methods, such as kernel principal components analysis which also involve spectral analysis of integral operators under non-negligible statistical error suffer from the same effect, which has been termed variance inflation.[55–57]

Left unspecified in this protocol are three important parameters: the degree of cross validation, k, the number of desired eigenfunctions, m, and the correlation lag time, τ. In our experiments, following common practice in the machine learning community, we use $k = 5$. Especially in the data-limited regime, the tradeoffs involving the choice of k are not entirely clear, as the objective lacks the form of an empirical risk minimization problem.[26,58] For MSMs, substantial attention in the literature has been paid to the selection of the lag time, τ.[28,38,59] With fixed basis function, it has been shown that the eigenfunction approximation error is a decreasing function of the τ, which motivates the use of larger values.[60] On the other hand, larger values of τ limit the temporal resolution of the model. For MSMs of protein folding, the authors' experience suggests that appropriate values for τ are often in the range between 1 and 100 ns. Finally, we suggest that m, the rank of GMRQ, be selected based on the number of slow dynamical processes in the system, as determined

by an apparent gap in the eigenvalue spectrum of $\mathcal{P}(\tau)$, or heuristically to a value between 2 and ~10.

VI. SIMULATIONS

A. Double well potential

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

$$\frac{dx_t}{dt} = -\nabla V(x_t) + \sqrt{2D}R(t), \qquad (25)$$

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

$$V(x) = 1 + \cos(2x) \qquad (26)$$

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 are shown in Fig. 1(b). We computed the true eigenvalues of the system's propagator to machine precision by discretizing the propagator matrix elements on a dense grid (see Appendix C). The timescale of the slowest relaxation process in this system is $t_2 \approx 7115.3$ steps and the dataset contains approximately 94 transition events.

We now consider the construction of Markov state models for this system, and in particular, the selection of the number of states, n, using states, $\mathcal{S} = \{s_i\}_{i=1}^n$, which evenly partition the region between $x = -\pi$ and $x = \pi$ into n equally spaced intervals,

$$s_i = \left[-\pi + \frac{2\pi}{n}(i-1), \ -\pi + \frac{2\pi}{n}i \right). \qquad (27)$$

When n 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 estimated by the MSM is proportional to n^2, we expect that for n too large, our models will be overfit. We therefore use 5-fold cross validation with the GMRQ to select the appropriate number of states, balancing these competing effects. The cross-validation GMRQ for the first two eigenvectors ($m = 2$, $\tau = 100$ steps) of the MSMs is shown in Fig. 1(a), 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 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. (12)— 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. The training set eigenvectors, the maximizers of the training set GMRQ, are actually exploiting noise in the dataset more so than modeling the propagator eigenfunctions. 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: Octaalanine

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 octaalanine, 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.

For clustering, we considered three distance metrics. The first was the backbone ϕ and ψ 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 distribution of reciprocal interatomic distances (DRID) distance metric introduced by Zhou and Caflisch,[61] 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.[62] We also considered three clustering algorithms; k-centers,[63] a landmark version of UPGMA hierarchical clustering (see Appendix D), and k-means.[64] For each pair of distance metric and clustering algorithm (excluding k-means and RMSD which is incompatible),[65] we
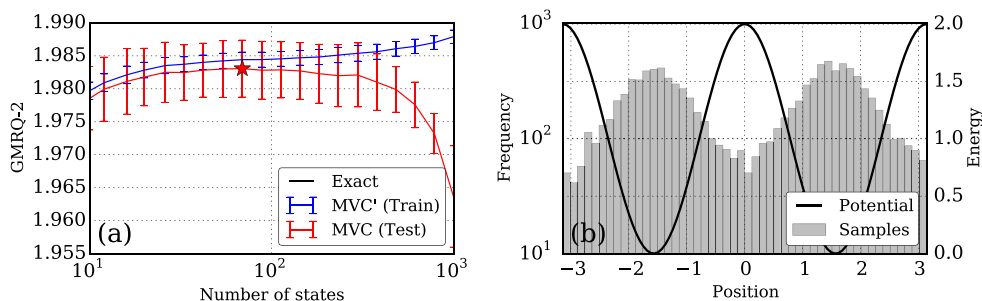

FIG. 1. Model selection for MSMs of a double well potential. Error bars indicate standard deviations over the 5 folds of cross validation. See text for details.

performed 5-fold cross validation using between 10 and 500 states for the clustering. 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 reëvaluated 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 processes. 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.[63]

VII. 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 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 theorem 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 2 is analogous to what has
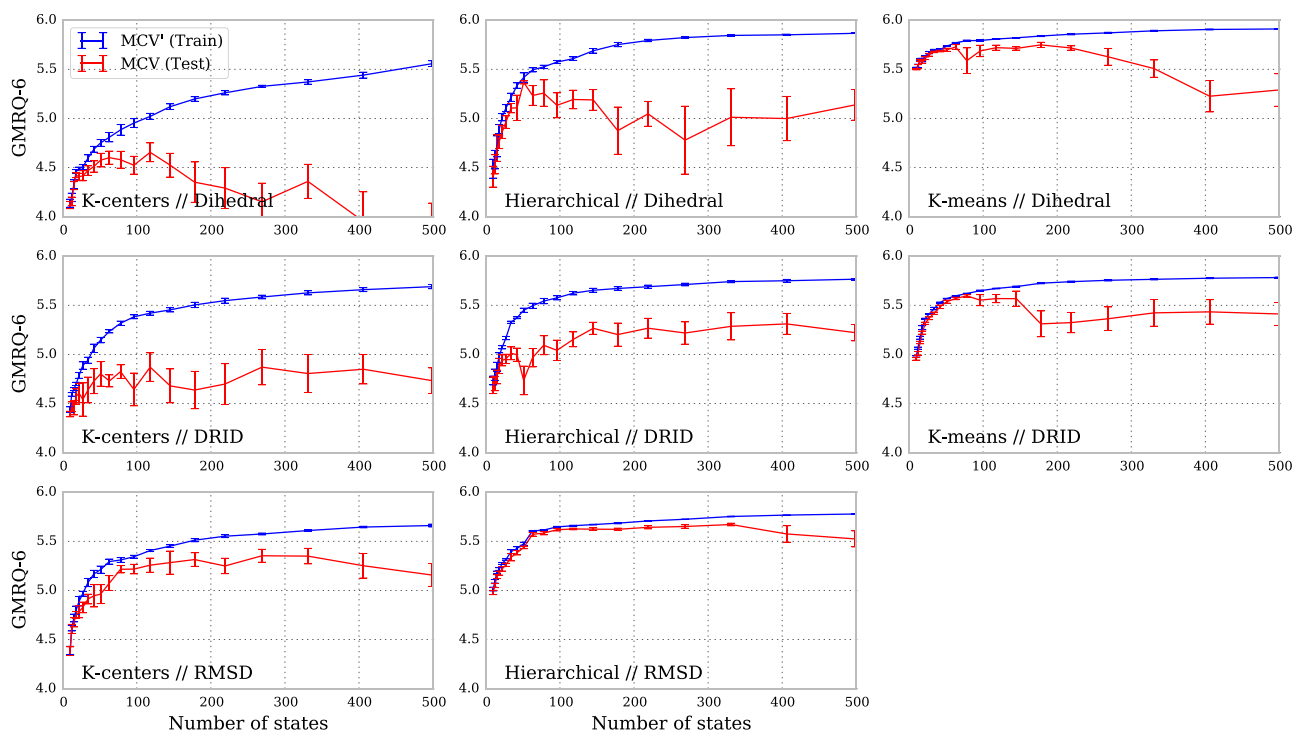

FIG. 2. Comparison of 8 methods for building MSMs under 5-fold cross validation, evaluated using the rank-6 GMRQ. We used the k-centers, k-means, and landmark-based ($n_{\text{landmarks}} = 5000$) UPGMA hierarchical clustering algorithms, with the DRID[61] and backbone dihedral angle featurizations. Error bars indicate the standard error in the mean over the cross validation folds.

been called the ensemble or trace variational principle in that field,[66–69] 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.[70] 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. (15) in this work.[71] 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.[72,73]

B. Comparison to likelihood maximization

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

As we show in Sec. IV B, and Nüske et al.[41] have shown earlier, MSMs arise naturally from a maximization of Eq. (12) 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,[25,74] parameterization,[75] and coarse-graining.[76,77] 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 positivity preserving, which is essential to form a probabilistic likelihood function in the sense of Kellogg, Lange, and Baker[74] or McGibbon, Schwantes, and Pande.[25] On the other hand, the two views *are* tightly coupled; their connection is given by the error bounds proved by Sarich, Noé and Schütte.[60] When the model gives a good approximation to the slow propagator eigenspace (low eigenfunction approximation error, high GMRQ), a good approximation to the long-timescale transition probabilities is obtained.

Cross validation with the log likelihood requires either a generative model for the high dimensional data, such as a hidden Markov model (HMM)[78] or dimensionality reduction before model comparison. This is a major disadvantage, because accurate and tractable generative models for time

series with tens or hundreds of thousands dimensions are not generally available. However, treating dimensionality reduction as a preprocessing and applying probabilistic models afterwards, as done by McGibbon, Schwantes, and Pande,[25] does not enable quantitative comparison between alternative competing dimensionality reduction protocols. With the GMRQ, on the other hand, the need for a reference state decomposition or high-dimensional generative model is eliminated,[76] and different dimensionality reduction procedures can easily be compared in a quantitative manner, as shown in Fig. 2.

VIII. 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. During the training step, tICA and MSMs are specific instance of this method with different types basis functions. This method extends these tools, making it possible to score trained models on new datasets and to perform hyperparameter selection.

We have applied this approach to compare eight different protocols for Markov state model construction on a set of MD simulations of the octaalanine 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. 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. Free and open source software fully implementing these methods is available in the MDTraj and MSMBuilder3 packages from http://mdtraj.org and http://msmbuilder.org, along with example and tutorials.[79] While the lag time, τ 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 a MSM, can be chosen to 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 automated.[80]

APPENDIX A: PROOFS OF THEOREM 2 AND LEMMA 3

Proof of Theorem 2. The eigenfunctions, ϕ_i, of $\mathcal{P}(\tau)$ form a complete basis. Expand each $f_i = \sum_a w_{ai}\phi_a$ with coefficients $W \in \mathbb{R}^{\infty \times m}$ with $W_{ni} = \langle f_i, \phi_n \rangle_{\mu^{-1}}$,

$$P_{ij} = \langle f_i, \mathcal{P} \circ f_j \rangle_{\mu^{-1}} \tag{A1}$$

$$= \left\langle \sum_a W_{ai}\phi_a, \mathcal{P} \circ \sum_b W_{bj}\phi_b \right\rangle_{\mu^{-1}} \tag{A2}$$

$$= \sum_a W_{ai}W_{aj}\lambda_a, \tag{A3}$$

$$Q_{ij} = \langle f_i, f_j \rangle_{\mu^{-1}} \tag{A4}$$

$$= \left\langle \sum_a W_{ai}\phi_a, \sum_b W_{bj}\phi_b \right\rangle_{\mu^{-1}} \tag{A5}$$

$$= \sum_a W_{ai}W_{aj}. \tag{A6}$$

We define the diagonal matrix $D(\lambda)$ with $D_{ii} = \lambda_i$. Then, Eqs. (A3) and (A6) can be rewritten in matrix form,

$$P = W^T D(\lambda) W, \tag{A7}$$

$$Q = W^T W. \tag{A8}$$

Let $F = Q^{1/2} \in \mathbb{S}_{++}^m$ be the (unique) positive definite square root of Q, which is guaranteed to exist because Q is positive definite and $B = WF^{-1}$. Then, rearrange the objective function using the cyclic property of the trace,

$$\mathcal{R}_\mathcal{P}[f] = \mathrm{Tr}\big(\underbrace{W^T D(\lambda) W}_{P} \underbrace{(FF)^{-1}}_{Q^{-1}}\big) \tag{A9}$$

$$= \mathrm{Tr}\big(F^{-1}W^T D(\lambda) W F^{-1}\big) \tag{A10}$$

$$= \mathrm{Tr}\big(B^T D(\lambda) B\big). \tag{A11}$$

Note that $B^T B = F^{-1}W^T W F^{-1} = I_m$. Therefore, by application of the Ky Fan theorem,[42,43]

$$\mathcal{R}_\mathcal{P}[f] \le \sum_i^m \lambda_i, \tag{A12}$$

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

Proof of Lemma 3. Following Absil *et al.*,[44] let $f = \{f_1, f_2, \ldots, f_m\}$ and $M \in \mathbb{R}^{m \times m}$ be an arbitrary invertible matrix. Define a new collection of functions $g = \{g_1, g_2, \ldots, g_m\}$, such that $g_j = \sum_{i=1}^m M_{ij}f_i$ and a matrix $W' \in R^{\infty \times m}$ such that $W'_{ni} = \langle g_i, \phi_n \rangle_{\mu^{-1}}$. Expanding the matrix elements of W', we note that

$$W' = WM. \tag{A13}$$

Then, using Eqs. (A7) and (A8), $\mathcal{R}_\mathcal{P}[g]$ can be written as a matrix expression involving W' and subsequently rewritten involving W and M. Expansion of the matrix products and application of the cyclic property of the trace confirms that

$$\mathcal{R}_\mathcal{P}[g] = \mathcal{R}_\mathcal{P}[f],$$

$$\mathcal{R}_\mathcal{P}[g] = \mathrm{Tr}\big(W'^T D(\lambda) W'(W'^T W')^{-1}\big) \tag{A14}$$

$$= \mathrm{Tr}\big((WM)^T D(\lambda)(WM)((WM)^T(WM))^{-1}\big) \tag{A15}$$

$$= \mathrm{Tr}\big(M^T W^T D(\lambda) W M^{-1}(W^T W)^{-1}M^{-T}\big) \tag{A16}$$

$$= \mathrm{Tr}\big(W^T D(\lambda) W(W^T W)^{-1}\big) \tag{A17}$$

$$= \mathcal{R}_\mathcal{P}[f]. \tag{A18}$$

The significance of this result is that it demonstrates $\mathcal{R}_\mathcal{P}$ to be invariant to linear transformations of f which preserve the space spanned by the functions. Much like the Ritz value of an trial vector is invariant to rescaling, or the angle between two planes is invariant to linear transformations of the basis vectors defining the planes, $\mathcal{R}_\mathcal{P}[f]$ is only a functional of the space spanned by f. This space—the set of all m-dimensional linear subspaces of a vector or Hilbert space—is referred to as a *Grassmann manifold*.[44]

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, $\mathcal{P}_m(\tau)$, built from the first m eigenpairs of $\mathcal{P}(\tau)$ is not in general a nonnegativity-preserving operator. That is, for some valid initial distributions, $p_t(\mathbf{x})$, the propagated distribution, $\tilde{p}_{t+\tau}^{(m)}(\mathbf{x}) = \mathcal{P}_m(\tau) \circ p_t(\mathbf{x})$, fails to be non-negative throughout Ω and thus does not represent a valid probability distribution,

$$p_{t+\tau}^{(m)}(\mathbf{x}) \not\ge 0 \ \forall \ \mathbf{x} \in \Omega. \tag{B1}$$

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

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

with infinitesimal generator $\mathcal{L}$ given by

$$\mathcal{L} = \frac{\partial^2}{\partial x^2} + 2\frac{\partial}{\partial x}x \tag{B3}$$

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, ξ_n, defined by

$$\mathcal{L} \circ \phi_n(x) = \xi_n\phi_n(x), \tag{B4}$$

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

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

$$\xi_n = -2n, \tag{B6}$$

$$c_n^2 = (2^n n! \pi)^{-1}, \tag{B7}$$

where the normalizing constants, c_n, are chosen such that $\langle \phi_n, \phi_m \rangle_{\mu^{-1}} = \delta_{nm}$.

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

$$\mathcal{P}(\tau) = e^{\tau \mathcal{L}}. \tag{B8}$$

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

$$\lambda_n = e^{-\tau \xi_n}. \tag{B9}$$

We now define the rank-m truncated propagator, $\mathcal{P}_m(\tau)$, such that

$$\mathcal{P}_m(\tau) \circ p_t = \sum_{n=0}^{m-1} \lambda_n \langle p_t, \phi_n \rangle_{\mu^{-1}} \phi_n \tag{B10}$$

$$= \sum_{n=0}^{m-1} e^{-2n\tau} c_n e^{-x^2} H_n(x)$$

$$\times \left[\int_{-\infty}^{\infty} dx' \, c_n \sqrt{\pi} \, p_t(x') H_n(x') \right]. \tag{B11}$$

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

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

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

$$\tilde{p}_\tau^{(2)}(x) = \frac{1}{\sqrt{\pi}} e^{-x^2} \left(1 + 2x x_0 e^{-2\tau} \right). \tag{B13}$$

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

$$\tilde{p}_\tau^{(2)}(x) < 0, \quad \text{when} \begin{cases} x < -e^{2\tau}/2x_0 & \text{if } x_0 > 0 \\ x > -e^{2\tau}/2x_0 & \text{if } x_0 < 0 \end{cases}. \tag{B14}$$

Because of this non-positivity, $\tilde{p}_\tau^{(2)}(x)$ 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: DOUBLE-WELL POTENTIAL INTEGRATOR AND EIGENFUNCTIONS

To discretize the Brownian dynamics stochastic differential equation in Eq. (25) with reflecting boundary conditions

at $-\pi$ and π, we used the Euler integrator,

$$x_{t+1} = bc \left(x_t + \left(\nabla V(x_t) + \sqrt{2D} R(t) \right) \Delta t \right), \tag{C1}$$

where steps that went outside the boundaries by a given distance were reflected back into the interval with a matching distance to the boundary

$$bc(x) = \begin{cases} 2\pi - x, & \text{if } x > \pi, \\ -2\pi - x, & \text{if } x < -\pi, \\ x, & \text{otherwise.} \end{cases} \tag{C2}$$

We computed the propagator eigenvalues by discretizing the interval into n MSM states $\{s_i\}_{i=1}^n$, following Eq. (27) and computing the matrix elements without stochastic sampling. This calculation is more straightforward by working with the transition matrix $T \in \mathbb{R}^{n \times n}$,

$$T_{ij} = \mathbb{P} \left[x_{t+\tau} \in s_j | x_t \in s_i \right], \tag{C3}$$

instead of the correlation and overlap matrices, C and S, directly. Note that as shown by Nüske *et al.*[41] and Prinz *et al.*,[28] $T = S^{-1}C$. Thus, the eigenvalues of T are identical to the generalized eigenvalues of (C, S).

To calculate the matrix elements of T, we consider each state, s_i, represented by its left endpoint, x_i. For each pair of states (i, j), we calculate the probability of the random force required to transition between them in one step, from Eq. (C1), taking into account the fact that because of the reflecting boundary conditions, the transition could have taken place via a transition outside the interval followed by a reflection.

Let δx be the width of the states, $\delta x = n^{-1} 2\pi$. For each $i \in \{1, \ldots n\}$ and $k \in \{-n, \ldots, n\}$, we calculate the action for a step from x_i to $x_i + k\delta x$,

$$a_{ik} = \frac{1}{\sqrt{2\pi}} \exp \left(\frac{((x_i + k\delta x) - x_i + \nabla V(x_i)\Delta t)}{\sqrt{2D}} \right). \tag{C4}$$

Let $j'_{ik} \in \{1, \ldots, n\}$ be the index of state which contains $bc(x_i + k\delta x)$. We calculate T_{ij} by summing the appropriate action terms which, because of the boundary conditions, get reflected into the same ending state,

$$T_{ij} = \sum_{k=-n}^{n} a_{ik} \, \delta_{j'_{ik}, j}, \tag{C5}$$

where $\delta_{i,j}$ is the Kronecker delta. This calculation is implemented in the file `brownian1d.py` in the MSMBuilder3 package. The eigenvalues of T converge rapidly as n increases. Our results in Fig. 1 use $n = 500$.

APPENDIX D: 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.[81] 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 are on average closest to

$$s(x^*) = \underset{n}{\arg\min} \frac{1}{|S_n|} \sum_{x \in S_n} d(x^*, x). \tag{D1}$$

APPENDIX E: SIMULATION SETUP

We performed all-atom molecular dynamics simulations of terminally blocked octaalanine (Ace-(Ala)$_8$-NHMe) in explicit solvent using the GROMACS 4 simulation package,[82] the AMBER ff99SB-ILDN-NMR forcefield,[83] and the TIP3P water model.[84] The system was energy minimized, followed by 1 ns of equilibration using the velocity rescaling thermostat (reference temperature of 298 K, time constant of 0.1 ps),[85] Parrinello-Rahman barostat (reference pressure of 1 bar, time constant of 1 ps, isotropic compressibility of 5×10^{-5} bar),[86] and Verlet integrator (time step 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.[87] Twenty six such simulations were performed, with production lengths between 20 and 150 ns each. The total aggregate sampling was 1.74 μs.

[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).

[12]P. E. Lopes, J. Huang, J. Shim, Y. Luo, H. Li, B. Roux, and A. D. MacKerell, Jr., J. Chem. Theory Comput. **9**, 5430 (2013).

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

[14]D. 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 (ACM, New York, 2009), pp. 39:1–39:11.

[15]M. Shirts and V. S. Pande, Science **290**, 5498 (2000).

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

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

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

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

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

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

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

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

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

[25]R. T. McGibbon, C. R. Schwantes, and V. S. Pande, J. Phys. Chem. B **118**, 6475 (2014).

[26]V. N. Vapnik and V. Vapnik, Statistical Learning Theory (Wiley, New York, 1998), Vol. 2.

[27]S. C. Larson, J. Educ. Psychol. **22**, 45 (1931).

[28]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).

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

[30]G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, J. Chem. Phys. **139**, 015102 (2013).

[31]O. Z. Maimon and L. Rokach, Data Mining and Knowledge Discovery Handbook (Springer, 2005), Vol. 1.

[32]C. Schütte, "Conformational dynamics: Modelling, theory, algorithm, and application to biomolecules," Habilitation thesis (Free University Berlin, 1998).

[33]E. Schmidt, Math. Ann. **63**, 433 (1907).

[34]R. Courant and D. Hilbert, Methods of Mathematical Physics, Methods of Mathematical Physics Vol. 1 (Wiley, 2008).

[35]C. A. Micchelli and A. Pinkus, J. Approximation Theory **24**, 51 (1978).

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

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

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

[39]J. D. Chodera and F. F. Noé, Curr. Opin. Stuct. Biol. **25**, 135 (2014).

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

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

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

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

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

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

[46]S. Muff and A. Caflisch, J. Chem. Phys. **130**, 125104 (2009).

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

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

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

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

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

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

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

[54]E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, LAPACK Users' Guide, 3rd ed. (Society for Industrial and Applied Mathematics, Philadelphia, PA, 1999).

[55]B. Schölkopf, A. Smola, and K.-R. Müller, Neural Comput. **10**, 1299 (1998).

[56]U. Kjems, L. K. Hansen, and S. C. Strother, in Advances in Neural Information Processing Systems 13 (MIT Press, 2000), pp. 549–555.

[57]T. J. Abrahamsen and L. K. Hansen, J. Mach. Learn. Res. **12**, 2027 (2011).

[58]M. Cornec, "Concentration inequalities of the cross-validation estimator for Empirical Risk Minimiser," preprint arXiv:1011.0096 (2010).

[59]S. Park and V. S. Pande, J. Chem. Phys. **124**, 054118 (2006).

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

[61]T. Zhou and A. Caflisch, J. Chem. Theory Comput. **8**, 2930 (2012).

[62]D. L. Theobald, Acta Crystallogr., Sect. A: Found. Crystallogr. **61**, 478 (2005).

[63]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).

[64]S. Lloyd, IEEE Trans. Inf. Theory **28**, 129–137 (1982).

[65]M. Senne, B. Trendelkamp-Schroer, A. S. Mey, C. Schütte, and F. Noé, J. Chem. Theory Comput. **8**, 2223 (2012).

[66]A. K. Theophilou, J. Phys. C: Solid State Phys. **12**, 5419 (1979).

[67]E. K. U. Gross, L. N. Oliveira, and W. Kohn, Phys. Rev. A **37**, 2805 (1988).

[68]N. Gidopoulos, P. Papaconstantinou, and E. Gross, Physica B **318**, 328 (2002).

[69]R. Lai, J. Lu, and S. Osher, "Density matrix minimization with $\ell 1$ regularization," e-prints arXiv (2014).

[70]C. R. Rao, J. R. Stat. Soc. Ser. B Stat. Methodol. **10**, 159 (1948).

[71]G. Baudat and F. Anouar, Neural Comput. **12**, 2385 (2000).

[72]L. Clemmensen, T. Hastie, D. Witten, and B. Ersbøll, Technometrics **53** (2011).

[73]B. K. Sriperumbudur, D. Torres, and G. Lanckriet, Mach. Learn. **85**, 3 (2011).

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

[75]E. K. Rains and H. C. Andersen, J. Chem. Phys. **133**, 144113 (2010).

[76]S. Bacallado, J. D. Chodera, and V. Pande, J. Chem. Phys. **131**, 045106 (2009).

[77]G. R. Bowman, J. Chem. Phys. **137**, 134111 (2012).

[78]R. T. McGibbon, B. Ramsundar, M. M. Sultan, G. Kiss, and V. S. Pande, in Proceedings of the 31st International Conference on Machine Learning, Beijing, China, Vol. 32 (Omnipress, 2014), pp. 1197–1205.

[79]R. T. McGibbon, K. A. Beauchamp, C. R. Schwantes, L.-P. Wang, C. X. Hernández, M. P. Harrigan, T. J. Lane, J. M. Swails, and V. S. Pande, "MDTraj: A modern, open library for the analysis of molecular dynamics trajectories," bioRxiv, 2014.

[80]J. Snoek, H. Larochelle, and R. P. Adams, in Advances in Neural Information Processing Systems, Vol. 25 (Lake Tahoe, USA, 2012) pp. 2951–2959.

[81]D. Müllner, J. Stat. Softw. **53**, 1 (2013).

[82]B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, J. Chem. Theory Comput. **4**, 435 (2008).

[83]D.-W. Li and R. Brüschweiler, Angew. Chem. **122**, 6930 (2010).

[84]W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).

[85]G. Bussi, D. Donadio, and M. Parrinello, J. Chem. Phys. **126**, 014101 (2007).

[86]M. Parrinello and A. Rahman, J. Appl. Phys. **52**, 7182 (1981).

[87]T. Darden, D. York, and L. Pedersen, J. Chem. Phys. **98**, 10089 (1993).