

RESEARCH ARTICLE | AUGUST 18 2023

## Efficient approximation of molecular kinetics using random Fourier features

Special Collection: [2023 JCP Emerging Investigators Special Collection](#)

Feliks Nüske   ; Stefan Klus 

 Check for updates

*J. Chem. Phys.* 159, 074105 (2023)

<https://doi.org/10.1063/5.0162619>



View  
Online



Export  
Citation

CrossMark

### Articles You May Be Interested In

Identification of kinetic order parameters for non-equilibrium dynamics

*J. Chem. Phys.* (April 2019)

Direct differentiation of human pluripotent stem cells into vascular network along with supporting mural cells

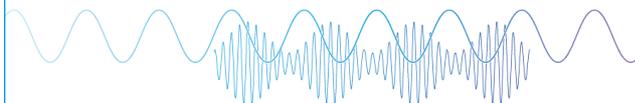
*APL Bioeng* (August 2023)

An enhanced iterative receiver based on vector approximate message passing for deep-sea vertical underwater acoustic communications

*J Acoust Soc Am* (March 2021)

Webinar

Boost Your Signal-to-Noise  
Ratio with Lock-in Detection



Sep. 7th – Register now



Zurich  
Instruments

# Efficient approximation of molecular kinetics using random Fourier features

Cite as: J. Chem. Phys. 159, 074105 (2023); doi: 10.1063/5.0162619

Submitted: 15 June 2023 • Accepted: 24 July 2023 •

Published Online: 18 August 2023



View Online



Export Citation



CrossMark

Feliks Nüske<sup>1,a)</sup>  and Stefan Klus<sup>2,b)</sup> 

## AFFILIATIONS

<sup>1</sup>Max-Planck-Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany

<sup>2</sup>Heriot-Watt University, Edinburgh, United Kingdom

**Note:** This paper is part of the 2023 JCP Emerging Investigators Special Collection.

<sup>a)</sup>Author to whom correspondence should be addressed: [nueske@mpi-magdeburg.mpg.de](mailto:nueske@mpi-magdeburg.mpg.de)

<sup>b)</sup>Email: [S.Klus@hw.ac.uk](mailto:S.Klus@hw.ac.uk)

## ABSTRACT

Slow kinetic processes in molecular systems can be analyzed by computing the dominant eigenpairs of the Koopman operator or its generator. In this context, the Variational Approach to Markov Processes (VAMP) provides a rigorous way of discerning the quality of different approximate models. Kernel methods have been shown to provide accurate and robust estimates for slow kinetic processes, but they are sensitive to hyper-parameter selection and require the solution of large-scale generalized eigenvalue problems, which can easily become computationally demanding for large data sizes. In this contribution, we employ a stochastic approximation of the kernel based on random Fourier features (RFFs) to derive a small-scale dual eigenvalue problem that can be easily solved. We provide an interpretation of this procedure in terms of a finite, randomly generated basis set. By combining the RFF approach and model selection by means of the VAMP score, we show that kernel parameters can be efficiently tuned and accurate estimates of slow molecular kinetics can be obtained for several benchmarking systems, such as deca alanine and the NTL9 protein.

© 2023 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>). <https://doi.org/10.1063/5.0162619>

## I. INTRODUCTION

The automated extraction of essential information about the thermodynamic and kinetic properties of molecular systems from large-scale computer simulations remains one of the major open problems in computational physics and chemistry to this day. The challenge consists of processing very long time series of high-dimensional data in such a way that *relevant* features of the molecular system are determined automatically, with limited user input or intervention. The precise definition of what is regarded as relevant depends on the context, but in many applications, the determination of *slow* relaxation processes and their associated metastable sets is among the most important features to be detected. This applies, for example, to protein folding, ligand binding, or protein–protein association.<sup>1–3</sup>

Let us highlight two milestones that have shaped the field in the last two decades: The first was the inception of Markov State Models (MSMs) in the late 1990s and early 2000s,<sup>4–6</sup> which allowed

for the construction of easily interpretable discrete kinetic models from which a wealth of information could be extracted. The second milestone was the variational approach to conformational dynamics (VAC) for reversible systems<sup>7,8</sup> and its extension, the *variational approach to Markov processes* (VAMP),<sup>9</sup> which provided a rigorously defined scalar objective function (called the VAMP score) to assess the quality of kinetic models. Combined, these developments have led to a standard workflow for the construction of MSMs, usually consisting of a linear dimensionality reduction using *time-lagged independent component analysis* (TICA),<sup>10</sup> geometric clustering, as well as MSM estimation and validation.<sup>11,12</sup> It was also shown that MSMs and the VAC are closely related to the Koopman operator framework<sup>13,14</sup> and, in particular, the *extended dynamic mode decomposition* (EDMD) algorithm,<sup>15,16</sup> which have primarily emerged in the analysis of fluid dynamics and related fields. The Koopman framework is therefore a unifying theme for the construction of kinetic models. Extensions of the EDMD algorithm can also be used to approximate the generator of dynamical systems [e.g., the

Kolmogorov backward operator for stochastic differential equations (SDEs)], as shown in Ref. 17.

Despite the remarkable success of MSMs, the construction process remains somewhat unsatisfactory, at least from a theoretical perspective. The initial TICA step presupposes that there is a kinetically informative linear subspace of the initial set of descriptors used for MSM building. In order to move beyond this requirement, algorithms based on deep neural networks (DNNs) were introduced in recent years,<sup>18,19</sup> allowing one to learn in one step the complete non-linear transformation from raw atomic coordinates to low-dimensional features that optimize the VAMP score. Both the MSM construction process (including state definition) and neural networks require a non-linear optimization step, which is typically solved by means of stochastic algorithms. A significant number of hyper-parameters need to be tuned in both cases, such as network architecture, network depth, and height for DNNs, or number of cluster centers, number of iterations, convergence threshold, etc. for clustering-based MSMs.

A different avenue that has been considered is to employ reproducing kernels, which have long been known in machine learning<sup>20,21</sup> as a powerful non-linear model class with rich theoretical properties. Kernel methods have also been demonstrated in many applications to be fairly robust if the available data are only small- to medium-sized. Kernel-based versions of EDMD were suggested in Refs. 22–24 and applied to high-dimensional molecular simulations in Ref. 25. Extracting slow timescales from kernel representations requires the solution of a large-scale generalized eigenvalue problem, where the size of the matrix depends on the number of snapshots. Even for moderate data sizes, the solution to this problem can be challenging. Moreover, although kernel methods usually require only a few hyper-parameters (such as the bandwidth of the hugely popular Gaussian kernel), these need to be tuned accurately, which requires solving the above-mentioned eigenvalue problem many times, which can be prohibitively time-consuming.

In this work, we show that kernel methods can be made tractable for the analysis of molecular kinetics by employing *Random Fourier Features* (RFFs),<sup>26</sup> a well-known low-rank approximation technique for kernel matrices, to convert the original large-scale eigenvalue problem into a much smaller one that can be easily solved. We derive the associated small-scale matrix problem both for the Koopman operator and its generator and show that this approach can be interpreted as applying EDMD with a specific randomly chosen basis set. We demonstrate that the VAMP score can be effectively used for hyper-parameter tuning of the kernel method. The combination of these ideas leads to an efficient approximation algorithm that is robust with respect to the data size and the stochastic Fourier feature selection. Therefore, our method presents a promising new direction for the automated analysis of molecular kinetics.

The remainder of this paper is structured as follows: In Sec. II, we review elements of the Koopman formalism, the EDMD approximation algorithm, reproducing kernel Hilbert spaces (RKHSs), and Koopman modeling using RKHSs. Then, in Sec. III, we present our main contribution, an approximation algorithm for the Koopman operator or generator based on Random Fourier Features. Further computational details, especially regarding model selection, are discussed in Sec. IV, while applications to benchmarking problems in molecular dynamics (MD) are shown in Sec. V.

## II. THEORETICAL BASICS

We will briefly introduce the stochastic Koopman operator and its generator, as well as numerical methods to estimate these operators and their spectral properties from data.

### A. Stochastic dynamical systems

Computer simulations are routinely used to explore the thermodynamics and kinetics of molecular systems. Many available implementations of molecular dynamics (MD) can be described as a stochastic process  $\{X_t\}_{t \geq 0}$ , where  $t$  is the time and  $X_t \in \mathbb{X} \subset \mathbb{R}^d$  is the state of the system at time  $t$ . In particular, a widely used model in this context is given by *stochastic differential equations* (SDEs) of the form

$$dX_t = F(X_t) dt + G(X_t) dW_t, \quad (1)$$

where  $F: \mathbb{R}^d \rightarrow \mathbb{R}^d$  and  $G: \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$  are the vector- and matrix-valued fields, respectively, called the *drift* and *diffusion* of the SDE, and  $W_t$  is the  $d$ -dimensional Brownian motion. The covariance matrix of diffusion is denoted by

$$a(x) := G(x)G(x)^\top. \quad (2)$$

As far as thermodynamics is concerned, the primary goal of molecular simulations is to sample the *Boltzmann distribution*,

$$d\mu(x) \sim \exp(-\beta V(x)) dx, \quad (3)$$

where  $V$  is the potential energy at position  $x$  and  $\beta^{-1} = k_B T$  is the inverse temperature. Most simulation protocols are built to achieve this goal, i.e., the Boltzmann distribution is invariant, and long time averages converge to spatial averages with respect to  $\mu$  (ergodicity). A standard example is *Brownian or overdamped Langevin dynamics*,

$$dX_t = -\frac{1}{\gamma} \nabla V(X_t) dt + \sqrt{2\gamma^{-1}\beta^{-1}} dW_t \quad (4)$$

with the friction parameter  $\gamma$ .

### B. Koopman operator and generator

When it comes to inferring kinetic information about  $X_t$ , significant progress has been made in recent years by considering an operator-based approach to the statistics of the stochastic process  $X_t$ . For a fixed lag time  $t \geq 0$ , the *Koopman operator*<sup>13</sup>  $\mathcal{K}^t$  acts on a function  $\phi: \mathbb{X} \rightarrow \mathbb{R}$  by taking the conditional expectation,

$$\mathcal{K}^t \phi(x) = \mathbb{E}^x[\phi(X_t)] = \mathbb{E}[\phi(X_t) | X_0 = x]. \quad (5)$$

In other words, evaluation of the function  $\mathcal{K}^t \phi$  at  $x$  yields the expectation of  $\phi$  after starting the dynamics at  $x$  and evolving until time  $t$ . The linear operators  $\mathcal{K}^t$  satisfy the semigroup equation  $\mathcal{K}^{s+t} = \mathcal{K}^s \mathcal{K}^t$ , which implies the existence of another linear operator  $\mathcal{L}$ , called the *Koopman generator*, such that the following linear differential equation holds for the function  $\phi_t = \mathcal{K}^t \phi$ ,

$$\frac{\partial}{\partial t} \mathcal{K}^t \phi(x) = \mathcal{L} \mathcal{K}^t \phi(x). \quad (6)$$

For SDE dynamics of the form (1), stochastic calculus shows that  $\mathcal{L}$  is a second-order differential operator, and we obtain the Kolmogorov equation,

$$\begin{aligned} \frac{\partial}{\partial t} \phi_t(x) &= \sum_{i=1}^d F(i)(x) \frac{\partial}{\partial x^i} \phi_t(x) + \frac{1}{2} \sum_{i,j=1}^d a(i,j)(x) \frac{\partial^2}{\partial x^i \partial x^j} \phi_t(x) \\ &= F(x) \cdot \nabla \phi_t(x) + \frac{1}{2} a(x) : \nabla^2 \phi_t(x) \\ &= \mathcal{L} \phi_t(x), \end{aligned} \quad (7)$$

where we used the colon notation for the Frobenius inner product between matrices, i.e.,  $A : B = \sum_{i,j} A(i,j)B(i,j)$ . For overdamped Langevin dynamics, the generator is given by

$$\mathcal{L}_{\text{OD}} = -\frac{1}{\gamma} \nabla V \cdot \nabla + \gamma^{-1} \beta^{-1} \Delta. \quad (8)$$

We will summarily refer to the Koopman operators or the generator as *dynamical operators* for the dynamical system  $X_t$ . If the dynamics are reversible with respect to the invariant distribution  $\mu$ , then the generator  $\mathcal{L}$  is symmetric with respect to the inner product,

$$\langle \phi_1, \phi_2 \rangle_\mu = \int_{\mathbb{X}} \phi_1(x) \phi_2(x) d\mu(x). \quad (9)$$

As a consequence, all eigenvalues of the generator  $\mathcal{L}$  are real-valued and non-positive, i.e., for all solutions of the equation,

$$\mathcal{L} \psi_i = \kappa_i \psi_i, \quad (10)$$

we have  $\kappa_i \leq 0$ . An object of prime interest for molecular kinetics are low-lying eigenvalues  $0 \leq -\kappa_i \ll 1$ , as these encode slow motions corresponding to metastable behavior encountered in many examples of molecular systems.<sup>6,27</sup> The connection between close-to-zero eigenvalues and metastability is also apparent in the fact that eigenfunctions of the generator are also eigenfunctions of the Koopman operator at any  $t$ , with eigenvalue  $\lambda_i(t) = e^{\kappa_i t}$ . If  $|\kappa_i| \ll 1$ , then the corresponding eigenvalues  $\lambda_i(t)$  decay slowly, as expected for metastable dynamics.

### C. Finite-dimensional approximation

The basic data-driven approximation algorithm for the Koopman operator is called *extended dynamic mode decomposition* (EDMD),<sup>15</sup> and its adaptation to the generator is then called *gEDMD*.<sup>17</sup> Both methods require choosing a finite *dictionary* of observable functions  $\{\phi_1, \dots, \phi_n\}$ . The projections of the dynamical operators onto the linear span of these functions can be represented by matrices,

$$\mathbf{K}^t = \mathbf{G}^{-1} \mathbf{A}^t, \quad \mathbf{L} = \mathbf{G}^{-1} \mathbf{A}^L \quad (11)$$

with mass and stiffness matrices given by

$$\begin{aligned} \mathbf{G}(i,j) &= \langle \phi_i, \phi_j \rangle_\mu, \quad \mathbf{A}^t(i,j) = \langle \phi_i, \mathcal{K}^t \phi_j \rangle_\mu, \\ \mathbf{A}^L(i,j) &= \langle \phi_i, \mathcal{L} \phi_j \rangle_\mu. \end{aligned} \quad (12)$$

Given a long trajectory  $\{x_1, x_2, \dots, x_{m+1}\}$  of the process  $X_t$ , the data-based estimators for these matrices are given by

$$\begin{aligned} \hat{\mathbf{G}}(i,j) &= \frac{1}{m} \sum_{l=1}^m \phi_i(x_l) \phi_j(x_l), \\ \hat{\mathbf{A}}^t(i,j) &= \frac{1}{m} \sum_{l=1}^m \phi_i(x_l) \phi_j(x_{l+1}), \\ \hat{\mathbf{A}}^L(i,j) &= \frac{1}{m} \sum_{l=1}^m \phi_i(x_l) \mathcal{L} \phi_j(x_l), \end{aligned} \quad (13)$$

where we assumed the time step between samples to equal  $t$  in the formula for  $\hat{\mathbf{A}}^t$ . This is not required, in practice, if a sliding-window estimator is used; see Refs. 12 and 28 for details.

For reversible dynamics, the stiffness matrix  $\mathbf{A}^L$  and its estimator  $\hat{\mathbf{A}}^L$  above can be replaced by the following energy-like expressions:<sup>17</sup>

$$\begin{aligned} \mathbf{A}^L(i,j) &= -\frac{1}{2} \int_{\mathbb{X}} \nabla^\top \phi_i(x) a(x) \nabla \phi_j(x) d\mu(x), \\ \hat{\mathbf{A}}^L(i,j) &= -\frac{1}{2m} \sum_{l=1}^m \nabla^\top \phi_i(x_l) a(x_l) \nabla \phi_j(x_l). \end{aligned} \quad (14)$$

The latter estimator only requires first-order derivatives and retains the symmetry of  $\mathbf{A}^L$  even for finite data.

Eigenvalues of the dynamical operators can be approximated by diagonalizing the data-driven estimators  $\hat{\mathbf{K}}^t = \hat{\mathbf{G}}^{-1} \hat{\mathbf{A}}^t$  and  $\hat{\mathbf{L}} = \hat{\mathbf{G}}^{-1} \hat{\mathbf{A}}^L$ , or equivalently, by solving one of the generalized eigenvalue problems,

$$\hat{\mathbf{A}}^t \mathbf{v}_i = \hat{\lambda}_i(t) \hat{\mathbf{G}} \mathbf{v}_i, \quad \hat{\mathbf{A}}^L \mathbf{v}_i = \hat{\kappa}_i \hat{\mathbf{G}} \mathbf{v}_i. \quad (15)$$

### D. Collective variables

For large molecular systems, the dictionary is typically defined in terms of a set of lower-dimensional descriptors  $z \in \mathbb{R}^N$ ,  $N \leq d$ , often called *collective variables* (CVs), such as intramolecular distances or angles. Formally, CVs are given by a smooth mapping  $\xi : \mathbb{X} \rightarrow \mathbb{R}^N$ . Choosing a dictionary  $\{\phi_i\}_{i=1}^n$  of functions depending only on CV space, that is,  $\phi_i(x) = \phi_i(\xi(x))$ , does not lead to any conceptual changes to the EDMD approach. When approximating the generator, the gEDMD method is straightforwardly adapted by computing all  $x$ -derivatives in the definition of the generator using the chain rule. This leads to a particularly elegant reformulation of the symmetric estimator (14) in terms of an effective diffusion matrix,

$$a^\xi(x) = \nabla_x^\top \xi(x) a(x) \nabla_x \xi(x), \quad (16)$$

where  $\nabla_x \xi$  is the  $d \times N$ -dimensional Jacobian of the mapping  $\xi$ . The symmetric estimator (14) then becomes

$$\hat{\mathbf{A}}^L(i,j) = -\frac{1}{2m} \sum_{l=1}^m \nabla_z^\top \phi_i(z_l) a^\xi(x_l) \nabla_z \phi_j(z_l), \quad (17)$$

as shown in Ref. 17. Note that once the effective diffusion matrices have been computed, only the basis functions and their derivatives in CV-space are required.

### E. Reproducing kernel Hilbert space

A critical modeling decision for the (g)EDMD approach is the choice of the dictionary  $\{\phi_1, \dots, \phi_n\}$ . Recent works<sup>22–25</sup> have shown that accurate results, even for large systems, can be obtained using methods based on *reproducing kernel Hilbert spaces* (RKHS) generated by a kernel function  $k$ . These approaches lead to powerful approximation spaces, yet they only depend on a few model parameters.

A reproducing kernel Hilbert space (RKHS)<sup>21,29</sup> is a Hilbert space  $\mathbb{H}$  with an inner product  $\langle \cdot, \cdot \rangle_{\mathbb{H}}$  of continuous functions on  $\mathbb{X}$ , defined by a symmetric, positive-definite two-argument function  $k(x, y)$ , called the *kernel*. Each such kernel corresponds to a unique RKHS.<sup>29</sup> Given  $k$ , one can define a map from  $\mathbb{X}$  into  $\mathbb{H}$ , called the *feature map*, by

$$\Phi(x)(y) = k(x, y). \quad (18)$$

Functions  $f$  in the RKHS then satisfy the reproducing property,

$$\langle f, \Phi(x) \rangle_{\mathbb{H}} = f(x) \quad \forall f \in \mathbb{H}. \quad (19)$$

*Example II.1.* Arguably the most widely used kernel is the Gaussian radial basis function (RBF) kernel,

$$k_{\sigma}(x, y) = \exp\left(-\frac{1}{2\sigma^2}\|x - y\|^2\right) \quad (20)$$

with bandwidth parameter  $\sigma > 0$ . On a periodic domain  $[-L, L]^d$ , the closely related periodic Gaussian kernel,

$$k_{\sigma}^p(x, y) = \exp\left(-\frac{2}{\sigma^2} \sum_{i=1}^d \sin^2\left(\frac{\Pi}{2L}(x^i - y^i)\right)\right), \quad (21)$$

analogously generates an RKHS of periodic functions.<sup>30</sup>  $\triangle$

For many popular kernels, including the Gaussian RBF kernel, the associated RKHS is infinite-dimensional and densely embedded into the most relevant function spaces. Kernels, therefore, provide powerful approximation spaces, reducing the problem of dictionary selection to the choice of a single function, the kernel, and its hyperparameters, such as the bandwidth for the Gaussian RBF kernel.

### F. RKHS representation of dynamical operators

Given a kernel  $k$  with RKHS  $\mathbb{H}$ , the analogs of the matrices  $\mathbf{G}$ ,  $\mathbf{A}^t$ , and  $\mathbf{A}^L$  in Sec. II C are given by linear operators on  $\mathbb{H}$ ; see Refs. 23 and 24 for a derivation,

$$\begin{aligned} \mathcal{G}f &= \int_{\mathbb{X}} f(x)\Phi(x) d\mu(x), \\ \mathcal{A}^t f &= \int_{\mathbb{X}} \mathcal{H}^t f(x)\Phi(x) d\mu(x), \\ \mathcal{A}^L f &= \int_{\mathbb{X}} \mathcal{L}f(x)\Phi(x) d\mu(x). \end{aligned} \quad (22)$$

Here,  $\Phi$  is the feature map again. The eigenvalue problem (15) then turns into an infinite-dimensional generalized eigenvalue equation for eigenfunctions  $\psi_i \in \mathbb{H}$ ,

$$\mathcal{A}^t \psi_i = \lambda_i(t) \mathcal{G} \psi_i, \quad \mathcal{A}^L \psi_i = \kappa_i \mathcal{G} \psi_i. \quad (23)$$

Natural empirical estimators for the RKHS operators in (22) are given by

$$\begin{aligned} \hat{\mathcal{G}}f &= \frac{1}{m} \sum_{l=1}^m f(x_l)\Phi(x_l), \quad \hat{\mathcal{A}}^t f = \frac{1}{m} \sum_{l=1}^m f(x_{l+1})\Phi(x_l), \\ \hat{\mathcal{A}}^L f &= \frac{1}{m} \sum_{l=1}^m \mathcal{L}f(x_l)\Phi(x_l). \end{aligned} \quad (24)$$

The range of these operators is in the linear span of the feature map at the data sites:  $\mathbb{H}_m = \text{span}\{\Phi(x_l)\}_{l=1}^m$ . Therefore, it makes sense to restrict these operators to the  $m$ -dimensional space  $\mathbb{H}_m$ , which leads to the matrix representations,

$$\begin{aligned} [\mathbf{K}_X](r, s) &= k(x_r, x_s), \\ [\mathbf{K}_X^t](r, s) &= k(x_{r+1}, x_s), \\ [\mathbf{K}_X^L](r, s) &= \frac{1}{2} \sum_{i,j=1}^d a(i, j)(x_r) \frac{\partial^2}{\partial x^i \partial x^j} k(x_r, x_s) \\ &\quad + \sum_{i=1}^d F_i(x_r) \frac{\partial}{\partial x^i} k(x_r, x_s) \end{aligned} \quad (25)$$

and (23) can be replaced by a matrix generalized eigenvalue problem,

$$\mathbf{K}_X^t \mathbf{w}_i = \hat{\lambda}_i(t) \mathbf{K}_X \mathbf{w}_i, \quad \mathbf{K}_X^L \mathbf{w}_i = \hat{\kappa}_i \mathbf{K}_X \mathbf{w}_i. \quad (26)$$

The derivatives in the definition of  $\mathbf{K}_X^L$  are taken with respect to the first argument of the kernel function. It is important to note that, as expected for a kernel method, assembling the matrices (25) requires only kernel evaluations at the data sites, along with kernel derivatives and coefficients of the SDE in the generator setting. For reversible dynamics, there is again an alternative symmetric formulation, which leads to the following Hermitian generalized eigenvalue problem instead of (26),

$$\begin{aligned} \mathbf{K}_X^{\text{rev}} \mathbf{w}_i &= \hat{\kappa}_i \mathbf{K}_X \mathbf{K}_X \mathbf{w}_i, \\ \mathbf{K}_X^{\text{rev}}(r, s) &= -\frac{1}{2m} \sum_{l=1}^m \nabla^T k(x_l, x_r) a(x_l) \nabla k(x_l, x_s). \end{aligned} \quad (27)$$

The derivations of (25)–(27) were shown in Refs. 23 and 24; we repeat them for the reader's convenience in the Appendix.

The linear problems (26) and (27) represent large-scale generalized eigenvalue problems. Solving these problems efficiently for large datasets is one of the central challenges common to most applications of kernel methods. In Sec. III, we present an approach based on a randomized low-rank representation obtained from the Fourier transformation of the kernel function.

### III. LOW-RANK REPRESENTATIONS OF KERNEL EIGENVALUE PROBLEMS

In this section, we will present our main contribution: an approach to solving the generalized eigenvalue problems (26) and (27) using a randomized low-rank decomposition.

### A. Random Fourier features

Random Fourier Features (RFFs) have been introduced in Ref. 26 as a low-rank approximation framework for large kernel matrices. The basis for this approach is Bochner's theorem, stating that a translation-invariant kernel  $k$ , which is normalized to  $k(x, x) = 1$  for all  $x$ , possesses the following stochastic representation:<sup>31</sup>

$$k(x, y) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}^d} e^{-i(x-y)^\top \omega} d\rho(\omega) = \mathbb{E}^{\omega \sim \rho} \left[ e^{-ix^\top \omega} \overline{e^{-iy^\top \omega}} \right]. \quad (28)$$

Here,  $\rho$  is a probability measure in frequency space, called the spectral measure. This representation extends to the derivatives of the kernel function, namely

$$D^\alpha k(x, y) = \mathbb{E}^{\omega \sim \rho} \left[ (-i\omega)^\alpha e^{-ix^\top \omega} \overline{e^{-iy^\top \omega}} \right], \quad (29)$$

where  $\alpha \in \mathbb{N}^d$  is the multi-index and  $\omega^\alpha = (\omega^1)^{\alpha(1)} \dots (\omega^d)^{\alpha(d)}$ . The derivatives in the above expression are again taken with respect to the first argument. For the Gaussian RBF kernel with bandwidth  $\sigma$ , the spectral measure is also Gaussian with bandwidth  $\sigma^{-1}$ . For the periodic Gaussian kernel, the spectral measure is discrete, supported on all wave vectors  $\frac{\pi}{L} \mathbf{k}$  for  $\mathbf{k} \in \mathbb{Z}^d$ , with probabilities,

$$\rho\left(\frac{\pi}{L} \mathbf{k}\right) = \prod_{j=1}^d I_{k(j)}(\sigma^{-2}) e^{-\sigma^{-2}}, \quad (30)$$

where  $I_k$  is the modified Bessel function of the first kind; see Ref. 32. Thus, drawing independent samples from the spectral measure is easily accomplished for many popular kernel functions.

### B. Representation of kernel matrices

We apply the RFF representation to the generalized eigenvalue problems (26) and (27) by substituting it into the kernel matrices  $\mathbf{K}_X, \mathbf{K}_X^t, \mathbf{K}_X^L$  given by (25). We again denote samples in real space (in the form of a long trajectory) by  $\{x_l\}_{l=1}^{m+1}$ . Instead of using one long trajectory, it would also be possible to extract a training dataset  $\{(x_l, y_l)\}_{l=1}^m$  from many short trajectories; see Ref. 28 for details. Additionally, we assume there are  $p$  samples  $\{\omega_u\}_{u=1}^p$  in frequency space, sampled from the spectral measure  $\rho$ . Replacing expectation values with finite-sample averages, we obtain

$$\begin{aligned} \mathbf{K}_X &= [k(x_r, x_s)]_{r,s} = \left[ \mathbb{E}^{\omega \sim \rho} \left[ e^{-ix_r^\top \omega} \overline{e^{-ix_s^\top \omega}} \right] \right]_{r,s} \\ &\approx \frac{1}{p} [\mathbf{M}\mathbf{M}^H]_{r,s}, \\ \mathbf{K}_X^t &= [k(x_{r+1}, x_s)]_{r,s} \approx \frac{1}{p} [\mathbf{M}^t \mathbf{M}^H]_{r,s} \end{aligned} \quad (31)$$

with RFF feature matrices,

$$\mathbf{M} = \left[ e^{-ix_r^\top \omega_u} \right]_{r,u}, \quad \mathbf{M}^t = \left[ e^{-ix_{r+1}^\top \omega_u} \right]_{r,u}. \quad (32)$$

That is, the rows of  $\mathbf{M}$  correspond to the data points, and the columns of  $\mathbf{M}$  correspond to the randomly sampled features.

When approximating the generator, we use the representation of derivatives given in (29), to find

$$\begin{aligned} \mathbf{K}_X^L &= \left[ \frac{1}{2} \sum_{i,j=1}^d a(i, j)(x_r) \frac{\partial^2}{\partial x_i^\top \partial x_j} k(x_r, x_s) \right. \\ &\quad \left. + \sum_{i=1}^d F(i)(x_r) \frac{\partial}{\partial x_i} k(x_r, x_s) \right]_{r,s} \\ &\approx \frac{1}{p} \sum_{u=1}^p \left[ \left( -\frac{1}{2} \sum_{i,j=1}^d a(i, j)(x_r) \omega_u(i) \omega_u(j) \right. \right. \\ &\quad \left. \left. - i \sum_{i=1}^d F(i)(x_r) \omega_u(i) \right) e^{-ix_r^\top \omega_u} \overline{e^{-ix_s^\top \omega_u}} \right]_{r,s} \\ &= \frac{1}{p} [\mathbf{M}^L \mathbf{M}^H]_{r,s} \end{aligned} \quad (33)$$

with generator feature matrix  $\mathbf{M}^L$ , which can be written compactly using the Frobenius inner product as

$$\mathbf{M}^L = \left[ \left( -\frac{1}{2} a(x_r) : (\omega_u \otimes \omega_u) - iF(x_r) \cdot \omega_u \right) e^{-ix_r^\top \omega_u} \right]_{r,u}. \quad (34)$$

We will address the reversible case further below after analyzing the RFF-based approximation in more detail. Together, (31) and (33) provide low-rank approximations of the kernel matrices, which can be easily assembled by the evaluation of complex exponentials.

### C. Solution of the generalized eigenvalue problems

Let us now use the approximations (31) and (33) to solve the generalized eigenvalue problems (26), which read (note that the normalization  $\frac{1}{p}$  can be omitted)

$$\mathbf{M}^t \mathbf{M}^H \mathbf{w}_i = \hat{\lambda}_i(t) \mathbf{M} \mathbf{M}^H \mathbf{w}_i, \quad \mathbf{M}^L \mathbf{M}^H \mathbf{w}_i = \hat{\kappa}_i \mathbf{M} \mathbf{M}^H \mathbf{w}_i. \quad (35)$$

If  $p \leq m$ , we can employ a standard trick to express this equivalently as a lower-dimensional eigenvalue problem. By defining  $\mathbf{v}_i = \mathbf{M}^H \mathbf{w}_i$ , we can directly verify that

$$\begin{aligned} \mathbf{M}^H \mathbf{M}^t \mathbf{v}_i &= \mathbf{M}^H \mathbf{M}^t \mathbf{M}^H \mathbf{w}_i = \hat{\lambda}_i(t) \mathbf{M}^H \mathbf{M} \mathbf{M}^H \mathbf{w}_i \\ &= \hat{\lambda}_i(t) \mathbf{M}^H \mathbf{M} \mathbf{v}_i. \end{aligned} \quad (36)$$

The last equality shows that all non-zero eigenvalues of (35) can be equivalently computed by the lower-dimensional *dual* problem,

$$\mathbf{M}^H \mathbf{M}^t \mathbf{v}_i = \hat{\lambda}_i(t) \mathbf{M}^H \mathbf{M} \mathbf{v}_i, \quad \mathbf{M}^H \mathbf{M}^L \mathbf{v}_i = \hat{\kappa}_i \mathbf{M}^H \mathbf{M} \mathbf{v}_i. \quad (37)$$

This  $p$ -dimensional generalized eigenvalue equation is the central problem to be solved within the context of RFF-based kernel approximation. If  $p$  is much smaller than  $m$ , (37) already leads to computational savings compared to (26).

In fact, it is not even necessary to assemble the matrices in (37), because the equation can be solved by operating only on the RFF feature matrices  $\mathbf{M}, \mathbf{M}^t, \mathbf{M}^L$ . To this end, we apply a whitening transformation, which is widely used within the context of (g)EDMD. Using a truncated singular value decomposition (SVD) of  $\mathbf{M}$ ,

$$\mathbf{M} \approx \mathbf{U}\Sigma\mathbf{W}^H \quad (38)$$

and retaining  $r \leq p$  components of the SVD, we obtain a linear transformation  $\mathbf{T} = \mathbf{W}\Sigma^{-1}$ , which eliminates the matrix on the right-hand sides of (37),

$$\mathbf{T}^H \mathbf{M}^H \mathbf{M} \mathbf{T} = \Sigma^{-1} \mathbf{W}^H \mathbf{W} \Sigma \mathbf{U}^H \mathbf{U} \Sigma \mathbf{W}^H \mathbf{W} \Sigma^{-1} = \text{Id}_{r \times r}. \quad (39)$$

Applying the same transformation to the left-hand sides of (37) leads to reduced matrices,

$$\begin{aligned} \mathbf{R}^t &= \mathbf{T}^H \mathbf{M}^H \mathbf{M}^t \mathbf{T} = \mathbf{U}^H \mathbf{M}^t \mathbf{W} \Sigma^{-1}, \\ \mathbf{R}^L &= \mathbf{T}^H \mathbf{M}^H \mathbf{M}^L \mathbf{T} = \mathbf{U}^H \mathbf{M}^L \mathbf{W} \Sigma^{-1}. \end{aligned} \quad (40)$$

It is then sufficient to diagonalize these reduced matrices in order to solve (37). If the truncation rank of  $\mathbf{M}$  is significantly smaller than  $p$ , additional computational savings can be harnessed this way. Algorithm 1 summarizes our solution method in compact form.

#### D. Interpretation as (g)EDMD problem

The transformation of the dual problem (37) to its reduced form is completely analogous to what is called the whitening transformation for (g)EDMD.<sup>28</sup> This is not a coincidence. Consider a finite dictionary defined by

$$\phi_{\text{RFF}}(x) := \begin{bmatrix} \phi_1(x) \\ \vdots \\ \phi_p(x) \end{bmatrix} = \begin{bmatrix} e^{ix^\top \omega_1} \\ \vdots \\ e^{ix^\top \omega_p} \end{bmatrix}. \quad (41)$$

Using the notation of Sec. II C, the empirical mass and stiffness matrices for this basis set are given by

$$\begin{aligned} \hat{\mathbf{G}}(u, v) &= \frac{1}{m} \sum_{l=1}^m \phi_u(x_l) \overline{\phi_v(x_l)} \\ &= \frac{1}{m} \sum_{l=1}^m e^{ix_l^\top \omega_u} e^{-ix_l^\top \omega_v} = \frac{1}{m} [\mathbf{M}^H \mathbf{M}](u, v), \\ \hat{\mathbf{A}}^t(u, v) &= \frac{1}{m} [\mathbf{M}^H \mathbf{M}^t](u, v), \\ \hat{\mathbf{A}}^L(u, v) &= \frac{1}{m} [\mathbf{M}^H \mathbf{M}^L](u, v), \end{aligned} \quad (42)$$

which are precisely the constituent matrices in (37). Note that we need to consider complex-valued observables at this point, leading to complex conjugation of the second argument in inner products. We conclude that the solution to the dual problem is equivalent to the eigenvalue estimation based on (g)EDMD using a random dictionary of complex plane waves. With this interpretation at hand, we can also derive a convenient symmetric formulation for reversible dynamics by replacing the estimator for  $\hat{\mathbf{A}}^L$  with its symmetric version (14),

$$\begin{aligned} \hat{\mathbf{A}}^L(u, v) &= -\frac{1}{2m} \sum_{l=1}^m \nabla^\top \phi_u(x_l) a(x_l) \nabla \phi_v(x_l) \\ &= -\frac{1}{2m} \sum_{l=1}^m \omega_u^\top a(x_l) \omega_v e^{ix_l^\top \omega_u} e^{-ix_l^\top \omega_v}. \end{aligned} \quad (43)$$

To compute the reduced matrix, we just have to apply the transformation  $\mathbf{T}$  from Sec. III C from the left and from the right, i.e.,

$$\mathbf{R}^L = \mathbf{T}^H \hat{\mathbf{A}}^L \mathbf{T} = \Sigma^{-1} \mathbf{W}^H \hat{\mathbf{A}}^L \mathbf{W} \Sigma^{-1}, \quad (44)$$

resulting again in a symmetric estimator.

#### E. Computational effort

Let us break down the computational effort required for Algorithm 1: The SVD of  $\mathbf{M}$  requires  $\mathcal{O}(mp^2)$  operations. Assembling the reduced matrix costs  $\mathcal{O}(rpm)$  operations for the Koopman operator. In the case of generator approximation, we obtain an additional dependence on the dimension due to the required contraction of derivatives of the kernel function. We can upper-bound the resulting effort as  $\mathcal{O}(mp(d^2 + r))$  in the non-reversible case and by  $\mathcal{O}(m(rpd + r^2d^2))$  in the reversible case. Note that this bound assumes a dense and state-dependent diffusion field  $a$ ; it reduces to linear dependence on  $d$  for diagonal diffusions, such as overdamped Langevin dynamics. Finally, the diagonalization of the reduced matrix amounts to  $\mathcal{O}(r^3)$  floating point operations. Comparing this to the full kernel method, we see that the most drastic computational savings will result from being able to choose  $p$  small compared to  $m$ , as it reduces a factor  $\mathcal{O}(m^3)$  to  $\mathcal{O}(mp^2)$ .

#### IV. METHODS

In this section, we will discuss aspects of the practical application of our spectral RFF method and also introduce the systems studied in the results section below. For the numerical examples, we will use the Gaussian RBF kernel or its periodic counterpart for periodic domains.

##### A. Model selection

Before we can apply Algorithm 1, we have to choose the number of features  $p$ , the truncation threshold for singular values, and as many hyper-parameters as required by the kernel function (e.g., the bandwidth for Gaussian RFF kernels). We select these parameters by cross-validating the VAMP score metric on randomly selected subsets of the data, as explained below. The only exception is the truncation threshold for singular values of  $\mathbf{M}$ . Here, we adopt the rule that all singular values smaller than  $10^{-4}$  times the largest singular value are discarded. This choice is based on prior experience and works well in all considered examples.

The VAMP score metric is based on the *Rayleigh variational principle*, which states that for reversible systems, the exact leading  $K$  eigenfunctions are optimizers of the Rayleigh trace,

$$\begin{aligned} \sum_{i=0}^{K-1} \lambda_i(t) &= \max_{\phi_0, \dots, \phi_{K-1}} \sum_{i=0}^{K-1} \langle \phi_i, \mathcal{K}^t \phi_i \rangle_\mu \\ &= \max_{\phi_0, \dots, \phi_{K-1}} \text{VS}(\phi_0, \dots, \phi_{K-1}), \\ \text{s.t. } \langle \phi_i, \phi_j \rangle_\mu &= \delta_{ij}. \end{aligned} \quad (45)$$

As the sum to be maximized on the right-hand side is a special case of the slightly more general *VAMP score* introduced for Koopman operators in Refs. 7 and 9 and for the Koopman generator in Ref. 33, we label it VS in what follows. For the generator, the same variational principle holds, with  $\lambda_i(t)$  replaced by  $\kappa_i$  and  $\mathcal{K}^t$  replaced by

**ALGORITHM 1.** RFF-based spectral approximation of the operator or generator.

**Input:** data matrix  $\mathbf{X} = [x_1, \dots, x_{m+1}] \in \mathbb{R}^{d \times (m+1)}$ ,  
kernel function  $k$  with spectral measure  $\rho$ ,  
number of features  $p$ , truncation rule for singular values.  
**Output:** Approximate eigenpairs  $(\hat{\lambda}_i(t), \hat{\psi}_i)$  or  $(\hat{\kappa}_i, \hat{\psi}_i)$  of the dynamical operator.

- 1: Draw  $p$  samples  $\{\omega_u\}_{u=1}^p$  from the spectral measure  $\rho$ .
- 2: Form matrix  $\mathbf{M}$  given by (32).
- 3: Non-reversible case: form matrix  $\mathbf{M}^t$  or  $\mathbf{M}^L$  as in (32) or (33).
- 4: Compute the SVD of  $\mathbf{M}$  and choose rank  $r$  according to the truncation rule:  $\mathbf{M} \approx \mathbf{U}\Sigma\mathbf{W}^H$ .
- 5: Form the reduced matrix  $\mathbf{R}^t$  or  $\mathbf{R}^L$  according to (40) or (44).
- 6: Compute eigenpairs of the reduced problem  $\mathbf{R}^t \mathbf{u}_i = \hat{\lambda}_i(t) \mathbf{u}_i$  or  $\mathbf{R}^L \mathbf{u}_i = \hat{\kappa}_i \mathbf{u}_i$ .
- 7: Transform to original RFF basis:  $\mathbf{v}_i = \mathbf{T} \mathbf{u}_i$ ,  $\hat{\psi}_i(x) = \mathbf{v}_i^H \phi_{\text{RFF}}(x)$ .

$\mathcal{L}$ . The variational principle also generalizes to non-reversible systems by considering singular functions instead of eigenfunctions.<sup>9</sup> However, we only analyze reversible systems in what follows. The VAMP score can be easily estimated upon replacing inner products by their empirical estimators, as shown in Sec. II C. The maximal VAMP score within a finite-dimensional model class can also be straightforwardly computed; see again Ref. 7. In this way, we can simply compare the maximal VAMP score obtained for each choice of hyper-parameters, e.g., choice of kernel bandwidth  $\sigma$  and number of Fourier features  $p$ , and choose the best.

To account for the estimation error stemming from the use of finite data and also to avoid over-fitting, we employ a standard cross-validation approach and first compute the optimal functions  $\phi_0, \dots, \phi_{K-1}$  for each model class on a random subsample of the data (training data). We then recompute the VAMP score for the  $K$ -dimensional space spanned by  $\phi_0, \dots, \phi_{K-1}$  on the remaining data (test data). This process is repeated  $n_{\text{test}}$  times, where  $n_{\text{test}}$  is either 10 or 20 below, and we choose the model class that optimizes the test score on average. The ratio of training data to test data is always chosen as a 75%/25% split.

## B. Spectral analysis and clustering

The solution of the eigenvalue problem for the reduced matrix  $\mathbf{R}$  provides estimates of the leading eigenvalues  $\hat{\lambda}_i(t)$  or  $\hat{\kappa}_i$ , respectively, and their associated eigenfunctions  $\hat{\psi}_i$ . The first point to note is that (37) is a complex eigenvalue problem, so the eigenvectors can be scaled by arbitrary complex numbers. Before analyzing the eigenvectors, we simply apply a grid search over complex phase factors to determine the one that minimizes the imaginary part of all eigenvector entries and use the rescaled eigenvectors for further analysis.

Second, we transform the eigenvalues of the Koopman operator into *implied timescales* defined as

$$t_i = -\frac{t}{\log(\hat{\lambda}_i(t))}, \quad (46)$$

which have units of time and can be interpreted as physical relaxation timescales for the  $i$ -th eigenmode of the Koopman operator. Since the timescales are independent of the lag time in the absence of projection and estimation errors,<sup>34</sup> we monitor the con-

vergence of implied timescales with increasing lag time  $t$  in order to select an optimal lag time.<sup>11</sup> We also compute the timescales when approximating the generator; in this case, they are simply given by

$$t_i = -\frac{1}{\hat{\kappa}_i}. \quad (47)$$

In order to analyze metastable sets, we evaluate the top eigenfunctions  $\hat{\psi}_i$  at all data points, and apply the spectral clustering method PCCA (Perron-Cluster Cluster Analysis)<sup>35</sup> to assign each data point  $x_i$  to each metastable set  $q$  with a degree of membership  $\chi_q(x_i)$ , which is a number between zero and one. We then identify data point  $x_i$  as part of metastable set  $q$  if  $\chi_q(x_i) \geq 0.6$ , and the remaining data points are treated as transition states.

## C. Systems and simulation setups

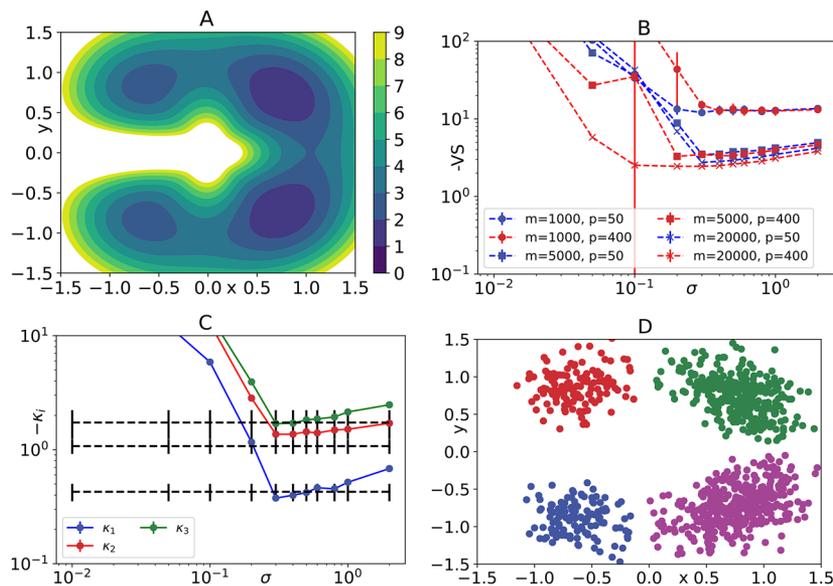
We study four different examples in order to illustrate different aspects of the performance of the proposed method. The first is overdamped Langevin dynamics (4) in a two-dimensional model potential, which is a minor modification of the *Lemon-Slice potential*,<sup>36</sup> given in polar coordinates as

$$V(r, \varphi) = \cos(4\varphi) + 10(r-1)^2 + 1 + \frac{1}{r} + \frac{1}{\cos \varphi}, \quad (48)$$

see Fig. 1 for a visualization. We can easily generate large amounts of simulation data by applying the standard Euler–Maruyama integrator. The elementary simulation time step is set to  $\Delta_t = 10^{-3}$ . Friction  $\gamma$  and inverse temperature  $\beta$  are both set to one. Reference implied timescales are obtained from a Markov state model based on a  $k$ -means discretization of the two-dimensional state space, using 50 discrete states and  $m = 10^5$  data points.

The second system is molecular dynamics simulation data of the alanine dipeptide in explicit water at a temperature of  $T = 300$  K, employing the Langevin thermostat. We generated a total of  $1 \times 10^6$  data points at 1 ps time spacing. Reference implied timescales were computed by a Markov state model in the well-known two-dimensional space of backbone dihedral angles  $\phi$  and  $\psi$ , using a  $30 \times 30$  box discretization and all  $1 \times 10^6$  data points.

The third example is molecular dynamics simulation data of the deca alanine peptide in explicit water at a temperature of  $T = 300$  K;



**FIG. 1.** Results for the Lemon-Slice potential. (a) Contour of the potential (48). (b) Negative VAMP score as a function of the kernel bandwidth for different data sizes  $m$  and feature numbers  $p$ . (c) Leading non-trivial eigenvalues of the negative generator for  $m = 5000$  and  $p = 50$ , as a function of the bandwidth. Black lines indicate the Markov model reference values. (d) Decomposition into four metastable states based on eigenvectors for  $\sigma = 0.4$ ,  $p = 50$ , and  $m = 5000$ . All error bars are based on 20 independent simulations.

see Ref. 37 for a description of the simulation setup. The data comprises a total of  $4 \times 10^6$  time steps at 1 ps time spacing. As reference values, we refer to the MSM analysis presented in Ref. 38, which is built on a 500-state  $k$ -means discretization after applying time-lagged independent component analysis (TICA)<sup>10</sup> in the space of the peptide's 16 backbone dihedral angles.

Finally, we re-analyze the 39-residue protein NTL9, which was simulated on the Anton supercomputer by D. E. Shaw Research; see Ref. 39 for details. The data comprises around  $1.2 \times 10^6$  frames, corresponding to a time spacing of  $\Delta_t = 2$  ns. As a reference, we apply linear TICA to the set of 666 minimal heavy atom inter-residue distances. We also rank these distances according to the fraction of simulation time where a contact is formed, that is, the distance does not exceed 0.35 nm; see Ref. 38, and calculate TICA models using only the first [20, 100, 200, 300, 400, 500, 666] of these distances.

## V. RESULTS

### A. Lemon-Slice potential

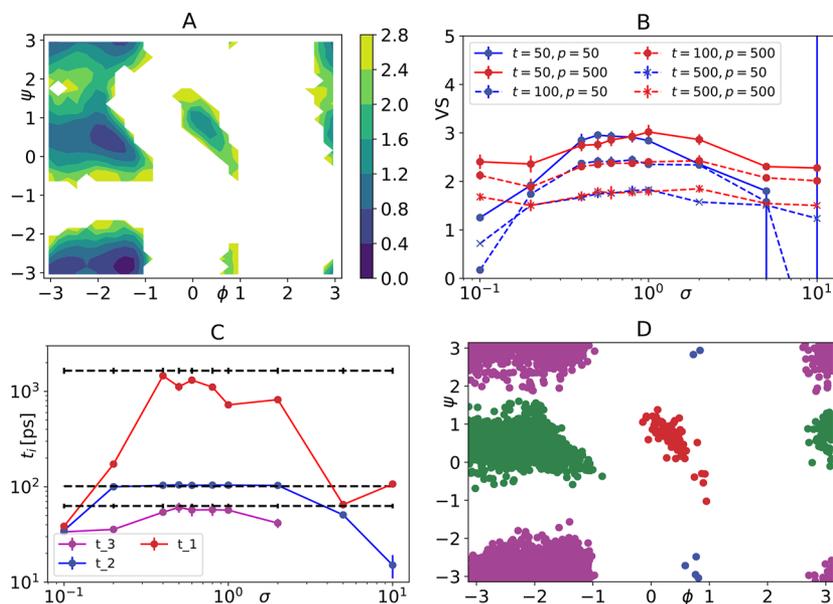
We start by analyzing the Lemon-Slice potential (48), shown in Fig. 1(a), which features four metastable states corresponding to the wells of the potential energy. We use this system as a first test case for approximating the generator  $\mathcal{L}$  due to the availability of all relevant quantities in analytical form.

Figure 1(b) shows the VAMP score as a function of the bandwidth for a selection of data sizes  $m$  and feature numbers  $p$ . We see that while  $m = 1000$  data points are not sufficient, for  $m = 5000$  data points, the score stabilizes for a range of bandwidths  $\sigma \in [0.3, 0.8]$  and already attains optimal values for very small  $p$ . We confirm this finding by plotting the first three non-trivial eigenvalues of

the generator, for  $m = 5000$  and  $p = 50$ , as a function of the bandwidth in Fig. 1(c). Indeed, accurate estimates for eigenvalues can be obtained in the suggested regime of bandwidths. After calculating the eigenvectors of the reduced matrix for  $\sigma = 0.4$  and applying the PCCA analysis, we see that the metastable structure is perfectly recovered; see Fig. 1(d). Thus, the VAMP score reliably guides the hyper-parameter search toward a very efficient approximation of the leading spectral components of this system.

### B. Alanine dipeptide

For alanine dipeptide, we approximate the Koopman operator with lag time  $t$ , using the periodic Gaussian kernel on the well-known two-dimensional reaction coordinate space given by the dihedral angles  $\phi$  and  $\psi$ . Figure 2(a) shows the familiar free energy landscape of the system in the reaction coordinate space. For all RFF-based approximations, we downsample the dataset to  $m = 20\,000$  points, corresponding to a time spacing of 50 ps. We apply the same model selection protocol as before and find that a range of kernel bandwidths  $\sigma \in [0.4, 1.0]$  can be stably identified as optimal in terms of the VAMP score. We observe in Fig. 2(b) that again, a small number of Fourier features,  $p = 50$ , is sufficient to arrive at an optimal VAMP score. We also confirm that the optimal parameter regime is stable across different lag times. Note that the VAMP score necessarily decreases with increasing lag time, as all eigenvalues but  $\lambda_0(t)$  decay exponentially with  $t$ . As shown in Fig. 2(c), the optimal range of bandwidths indeed allows for accurate estimation of the three leading implied timescales of the system. Further, we also verify that the metastable decomposition of the dihedral space into four states is correctly recovered using the RFF model; see Fig. 2(d). The



**FIG. 2.** Results for alanine dipeptide. (a) Free Energy (in kJ/mol) in two-dimensional dihedral space. (b) VAMP Score for selected feature sizes  $p$  and lag times  $t$  as a function of the kernel bandwidth. (c) Implied timescales for  $t = 100$  ps and  $p = 50$  as a function of the bandwidth. (d) Metastable decomposition obtained for  $\sigma = 0.6$ ,  $p = 50$ , and  $t = 100$  ps. The error bars for MSM timescales were generated using Bayesian transition matrix sampling.<sup>40</sup>

dramatic rank reduction for the kernel matrix achieved by using a small number of Fourier features allows us to easily compute eigenvalues and test scores for a broad range of parameters using  $m = 20\,000$  data points in real space. Solving the same problem by means of standard methods would be significantly more costly if the same data size was used.

### C. Deca alanine

We use the deca alanine example to demonstrate that, by means of low-rank kernel methods, generator models can be efficiently learned on a reaction coordinate space that is more than just one- or two-dimensional and that dynamical hypotheses can be tested this way. We project the system onto the space of its 16 interior backbone torsion angles. The generator of full molecular dynamics is not readily available in closed form. It is well known, however, that if observed only in a subset of position space, many systems driven by thermostatically controlled molecular dynamics behave like reversible overdamped Langevin dynamics at long timescales.<sup>41</sup> This can be rigorously shown for underdamped Langevin dynamics by using the inverse friction constant as a re-scaling of time, leading to overdamped dynamics in position space in the large friction limit.

For that reason, we set the full space diffusion matrix to  $\frac{1}{\beta}$  times the identity and employ our reversible estimator. The effective diffusion (16) then turns into

$$a^\xi(z) = \frac{1}{\beta} \nabla \xi^\top(x) \nabla \xi(x), \quad (49)$$

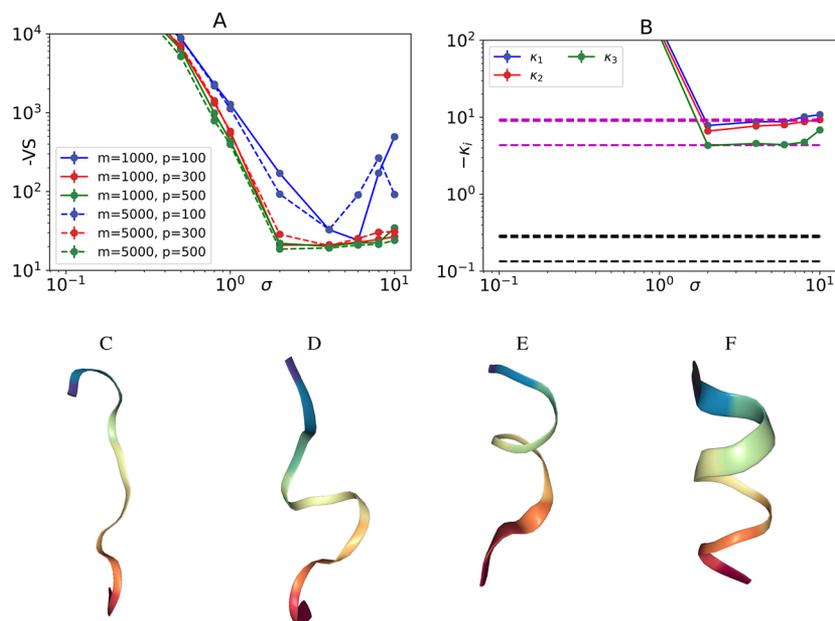
where  $\nabla \xi$  is the Jacobian of the mapping from Euclidean coordinates to the peptide's backbone dihedral angles, which can be evaluated analytically. Note that this approach can be seen as a dynamical

hypothesis: we choose to model the position space dynamics as an overdamped Langevin process and test if its projection onto the dihedral angle space can recover all relevant features of the full MD simulation.

We find in Fig. 3(a) that a kernel bandwidth of  $\sigma \in [2.0, 8.0]$  leads to an optimal VAMP score for the projected generator. Using about  $m = 1000$  data points in real space and  $p = 300$  Fourier features is sufficient to obtain stable results, rendering the calculation highly efficient. The optimality of this regime of bandwidths is confirmed by plotting the first three non-trivial eigenvalues of the generator, as shown in Fig. 3(b). We note that these eigenvalues differ from those of the reference MSM by a constant factor. This is due to the re-scaling of time inherent in assuming overdamped Langevin dynamics in position space, rendering the effective dynamics too fast. However, we also see in Fig. 3 that all slow eigenvalues are indeed re-scaled by the same factor. Moreover, the PCCA analysis of the slowest eigenvectors obtained for the optimal RFF model also recovers the metastable states of the deca alanine system, corresponding to the folded and unfolded states as well as two intermediates. The effective overdamped Langevin model learned by the kernel method thus retains all major features of the original molecular dynamics, up to the re-scaling of time.

### D. NTL9

We use the NTL9 dataset to demonstrate once again the ability of the proposed method to efficiently and robustly analyze large datasets. We downsample the available trajectory data to around  $m = 15\,000$  data points at 200 ns time spacing. As mentioned earlier,  $m = 15\,000$  is still a significant data size for kernel methods.

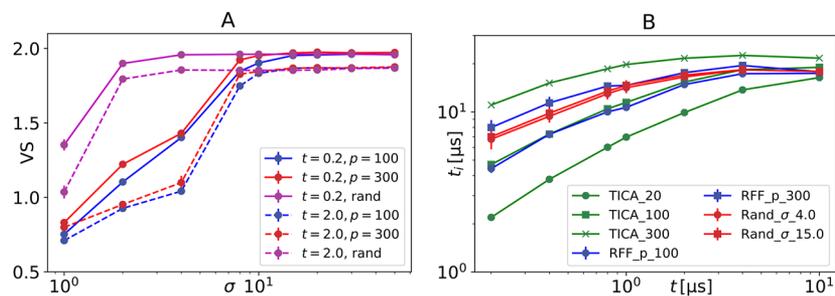


**FIG. 3.** Results for the Koopman generator on the backbone dihedral angle space of the deca alanine peptide. (a) Negative VAMP score for selected data sizes  $m$  and feature sizes  $p$  as a function of the kernel bandwidth. (b) Eigenvalues of the negative generator for  $m = 1000$  and  $p = 300$  as a function of the bandwidth. The reference MSM results are shown in black. Re-scaling MSM eigenvalues by the average ratio between optimal RFF and MSM eigenvalues leads to the magenta lines. (c)–(f) Representative structures for each of the four PCCA states based on the RFF model at  $m = 1000$ ,  $p = 300$ , and  $\sigma = 4.0$ . The error bars for MSM eigenvalues (barely visible) were generated using Bayesian transition matrix sampling.<sup>40</sup>

We project the MD simulations onto the 666-dimensional space of minimal heavy atom inter-residue distances. It is known from previous publications that these coordinates capture the folding process of NTL9 very well.<sup>38,42</sup> Indeed, just using linear TICA on the top-ranked 300 distances allows for an accurate estimation of the slowest implied timescale,  $t_1$ , associated with the folding process; see the top green line in Fig. 4(b).

Figure 4(a) shows that using Gaussian bandwidth parameters  $\sigma \in [10, 50]$  can be identified as optimal in terms of the VAMP score,

again across different lag times. In Fig. 4(b), we first compare the slowest timescale  $t_1$  obtained for  $\sigma = 15$  as a function of the lag time for different feature numbers  $p$  (blue lines). We see that already for  $p \geq 300$ , satisfactory convergence of  $t_1$  with  $t$  can be obtained, reducing the computational cost by about a factor of two thousand compared to the full kernel eigenvalue problem. The folded and unfolded states of the protein can also be identified correctly using these parameters, see Fig. 5. Interestingly, timescale convergence improves with  $p$  in about the same way that the performance



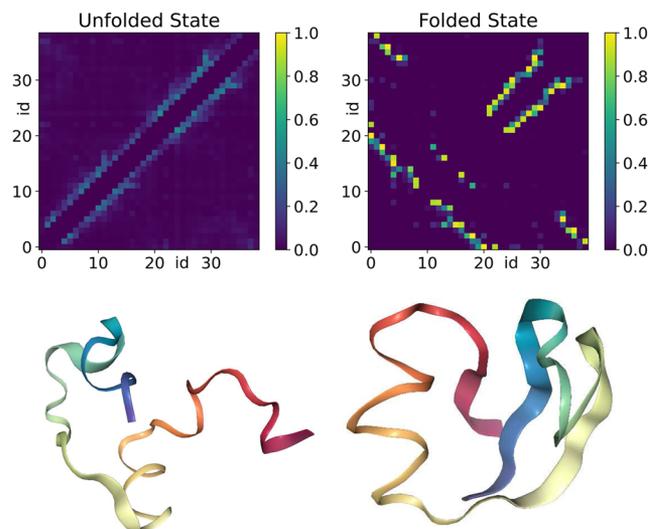
**FIG. 4.** Results for the NTL9 protein. (a) VAMP score for Gaussian RFF approximation as a function of the kernel bandwidth  $\sigma$ . The red and blue lines show the results using all 666 distances for different lag times and different numbers of Fourier features  $p$ . The magenta lines show the average values over ten random selections of only 50 distances, with  $p = 300$  fixed. (b) Slowest implied timescale  $t_1$  as a function of the lag time  $t$ . The green lines show estimates based on linear TICA using the top-ranked 20, 100, and 300 distances. The blue lines show RFF-based estimates on all distances using different values of  $p$ , using  $\sigma = 15.0$ . The red lines are averages over the above-mentioned random subsamples of the distance coordinates, where  $p = 300$  is fixed.

of TICA improves as more and more of the ranked distances are added. However, we show that the kernel method is more robust with respect to the selection of input coordinates by randomly selecting  $d = 50$  out of the 666 distances and recomputing both the VAMP score and implied timescales for different bandwidths, keeping the number of Fourier features fixed at  $p = 300$ . We see in Figs. 4(a) and 4(b) that both the optimal value of the VAMP score as well as the timescale estimates remain stable across random selections of the input distances [magenta lines in panel (a), red lines in panel (b)]. The convergence of timescales is, therefore, mainly dependent on the number of Fourier features  $p$ , which can be automatically tuned using the VAMP score, and not on the selection of input coordinates. Moreover, we also find that under random distance selection, the optimal regime of kernel bandwidths changes as the dimension of the input coordinate space becomes significantly smaller. This is reliably reflected by the behavior of the VAMP score; see again Fig. 4(a).

## VI. CONCLUSIONS

We have introduced a new approach to extracting the dominant eigenvalues of dynamical operators associated with stochastic processes, corresponding to the slow timescale dynamics in molecular kinetics. Building on kernel representations of the Koopman operator and generator, our main novelty is a dual low-rank representation of the kernel matrices using random Fourier features. We have derived the corresponding reduced eigenvalue problem and provided an interpretation of the method in terms of a random dictionary of plane waves. Using four examples of increasing complexity, we have shown that hyper-parameters of the method can be very effectively tuned using the VAMP score metric. Moreover, we have shown that only a few hundred Fourier features were sufficient to obtain accurate and robust results for all systems considered. This finding allowed us to rapidly scan kernel hyper-parameters for data sizes in the ten thousand on coordinate spaces of dimensions up to several hundred. The computational savings compared to the full eigenvalue problem, measured in floating point operations, were on the order of three to four orders of magnitude.

The main promise of our method is the ability to analyze large datasets using a very generic and, in fact, randomly generated basis set. The success of the method depends on the number of required Fourier features. Future research will need to test if this number indeed remains small even for larger systems, which would make the method scalable beyond the benchmarking examples studied here. Combining our approach with re-weighting techniques would allow the application of the method to enhanced sampling simulation data. It will also be interesting to see if kernel functions with more specific properties can be used, for instance, kernels incorporating more detailed symmetries than only translational invariance, kernels acting on mixed types of domain (periodic and non-periodic), or kernels associated with Sobolev spaces, which have rich analytical properties. In addition, concerning the theoretical analysis, it is known that the timescale separation between slow and fast dynamics expected for metastable systems induces an approximate low-rank structure of the Koopman operator. Our results suggest that this structure can be captured by a small number of random fea-



**FIG. 5.** Top Row: Representations of the two PCCA states obtained for  $\sigma = 15$ ,  $p = 300$ , and  $t = 2 \mu\text{s}$ . For each PCCA state, we show the fraction of simulation time during which each residue–residue pair forms a contact; see also Ref. 38. Bottom Row: Representative protein structures for both PCCA states.

tures. Providing a more detailed analysis of this heuristic argument presents another future research direction.

## ACKNOWLEDGMENTS

The authors thank D. E. Shaw Research for providing the molecular dynamics simulation data for NTL9.

## AUTHOR DECLARATIONS

### Conflict of Interest

The authors have no conflicts to disclose.

### Author Contributions

**Felix Nüske:** Conceptualization (lead); Data curation (lead); Formal analysis (equal); Investigation (equal); Methodology (lead); Software (lead); Validation (lead); Visualization (lead); Writing – original draft (lead); Writing – review & editing (lead). **Stefan Klus:** Conceptualization (supporting); Formal analysis (equal); Investigation (equal); Methodology (supporting); Validation (supporting); Visualization (supporting); Writing – original draft (supporting); Writing – review & editing (supporting).

## DATA AVAILABILITY

The methods presented in this paper are available as part of the KoopmanLib library at <https://github.com/fnueske/KoopmanLib>. The complete data and code to reproduce the examples shown above are available from a public repository at <https://doi.org/10.5281/zenodo.8036525>, with the exception of the raw molecular dynamics simulation data, which can be obtained

directly from the authors or, in the case of NTL9, from the copyright owners. PCCA decompositions and all Markov state model results were computed using the deeptime library.<sup>43</sup>

## APPENDIX: DERIVATION OF KERNEL EIGENVALUE PROBLEMS

Our derivation of the kernel eigenvalue problems in Sec. II F follows.<sup>23,24</sup> We compute the matrix elements of the operators  $\hat{\mathcal{E}}$ ,  $\hat{\mathcal{A}}^t$ ,  $\hat{\mathcal{A}}^L$  on the finite-dimensional space  $\mathbb{H}_m$  with respect to its canonical basis  $\{\Phi(x_i)\}_{i=1}^m$ ,

$$\begin{aligned} \langle \Phi(x_r), \hat{\mathcal{E}}\Phi(x_s) \rangle_{\mathbb{H}} &= \frac{1}{m} \sum_{l=1}^m \Phi(x_s)(x_l) \langle \Phi(x_r), \Phi(x_l) \rangle_{\mathbb{H}} \\ &= \frac{1}{m} \sum_{l=1}^m k(x_s, x_l) k(x_l, x_r) \\ &= \frac{1}{m} [\mathbf{K}_X \mathbf{K}_X](r, s), \end{aligned} \quad (\text{A1})$$

$$\begin{aligned} \langle \Phi(x_r), \hat{\mathcal{A}}\Phi(x_s) \rangle_{\mathbb{H}} &= \frac{1}{m} \sum_{l=1}^m \Phi(x_s)(x_{l+1}) \langle \Phi(x_r), \Phi(x_l) \rangle_{\mathbb{H}} \\ &= \frac{1}{m} \sum_{l=1}^m k(x_s, x_{l+1}) k(x_l, x_r) \\ &= \frac{1}{m} [\mathbf{K}_X \mathbf{K}_Y](r, s), \end{aligned} \quad (\text{A2})$$

$$\begin{aligned} \langle \Phi(x_r), \hat{\mathcal{A}}^L\Phi(x_s) \rangle_{\mathbb{H}} &= \frac{1}{m} \sum_{l=1}^m (\mathcal{L}\Phi(x_s))(x_l) \langle \Phi(x_r), \Phi(x_l) \rangle_{\mathbb{H}} \\ &= \frac{1}{m} \sum_{l=1}^m \left[ \nabla F(x_l) \cdot \nabla \Phi(x_s)(x_l) \right. \\ &\quad \left. + \frac{1}{2} a(x_l) : \nabla^2 \Phi(x_s)(x_l) \right] k(x_l, x_r) \\ &= \frac{1}{m} \sum_{l=1}^m \left[ \nabla F(x_l) \cdot \langle \nabla_1 k(x_l, \cdot), \Phi(x_s) \rangle_{\mathbb{H}} \right. \\ &\quad \left. + \frac{1}{2} a(x_l) : \langle \nabla_1^2 k(x_l, \cdot), \Phi(x_s) \rangle_{\mathbb{H}} \right] k(x_l, x_r) \\ &= \frac{1}{m} \sum_{l=1}^m \left[ \nabla F(x_l) \cdot \nabla_1 k(x_l, x_s) \right. \\ &\quad \left. + \frac{1}{2} a(x_l) : \nabla_1^2 k(x_l, x_s) \right] k(x_l, x_r) \\ &= \frac{1}{m} [\mathbf{K}_X \mathbf{K}_X^L](r, s). \end{aligned} \quad (\text{A3})$$

Here, we used the subscript  $\nabla_1$  for the nabla-operator to indicate differentiation with respect to the first variable. To get from the second to the third line, we used the derivative reproducing property,

$$D^\alpha f(x) = \langle D_1^\alpha k(x, \cdot), f \rangle_{\mathbb{H}}, \quad (\text{A4})$$

which holds for any multi-index  $\alpha$  and all RKHS functions  $f$  as soon as the kernel is at least  $2|\alpha|$ -times continuously differentiable in both arguments, see Ref. 20.

We see that in all cases, the full-rank matrix  $\mathbf{K}_X$  can be canceled out, which leaves us with (26). In the reversible case, we replace the operator  $\hat{\mathcal{A}}^L$  with a symmetric bi-linear form

$$\hat{\mathcal{A}}^{\text{rev}}(f, g) = -\frac{1}{2m} \sum_{l=1}^m \nabla^\top f(x_l) a(x_l) \nabla g(x_l). \quad (\text{A5})$$

Computing its matrix elements on  $\mathbb{H}_m$ , we find

$$\begin{aligned} \hat{\mathcal{A}}^{\text{rev}}(\Phi(x_r), \Phi(x_s)) &= -\frac{1}{2m} \sum_{l=1}^m \nabla^\top \Phi(x_r)(x_l) a(x_l) \nabla \Phi(x_s)(x_l) \\ &= -\frac{1}{2m} \sum_{l=1}^m \nabla_1^\top k(x_l, x_r) a(x_l) \nabla_1 k(x_l, x_s) \\ &= \mathbf{K}_X^{\text{rev}}(r, s). \end{aligned} \quad (\text{A6})$$

In this case, we cannot cancel a factor  $\mathbf{K}_X$ , which leads to (27).

## REFERENCES

- P. L. Freddolino, C. B. Harrison, Y. Liu, and K. Schulten, "Challenges in protein-folding simulations," *Nat. Phys.* **6**, 751–758 (2010).
- N. Plattner, S. Doerr, G. De Fabritiis, and F. Noé, "Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling," *Nat. Chem.* **9**, 1005–1011 (2017).
- F. Paul, C. Wehmeyer, E. T. Abualrous, H. Wu, M. D. Crabtree, J. Schöneberg, J. Clarke, C. Freund, T. R. Weikl, and F. Noé, "Protein-peptide association kinetics beyond the seconds timescale from atomistic simulations," *Nat. Commun.* **8**, 1095 (2017).
- S. M. Ulam, *A Collection of Mathematical Problems* (Interscience, 1960).
- M. Dellnitz and O. Junge, "On the approximation of complicated dynamical behavior," *SIAM J. Numer. Anal.* **36**, 491–515 (1999).
- C. Schütte, A. Fischer, W. Huisinga, and P. Deuffhard, "A direct approach to conformational dynamics based on hybrid Monte Carlo," *J. Comput. Phys.* **151**, 146–168 (1999).
- F. Noé and F. Nüske, "A variational approach to modeling slow processes in stochastic dynamical systems," *Multiscale Model. Simul.* **11**, 635–655 (2013).
- F. Nüske, B. G. Keller, G. Pérez-Hernández, A. S. J. S. Mey, and F. Noé, "Variational approach to molecular kinetics," *J. Chem. Theory Comput.* **10**, 1739–1752 (2014).
- H. Wu and F. Noé, "Variational approach for learning Markov processes from time series data," *J. Nonlinear Sci.* **30**, 23–66 (2020).
- G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, "Identification of slow molecular order parameters for Markov model construction," *J. Chem. Phys.* **139**, 015102 (2013).
- J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, "Markov models of molecular kinetics: Generation and validation," *J. Chem. Phys.* **134**, 174105 (2011).
- An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation, Advances in Experimental Medicine and Biology Vol. 797*, edited by G. R. Bowman, V. S. Pande, and F. Noé (Springer Heidelberg, 2014).
- B. O. Koopman, "Hamiltonian systems and transformation in Hilbert space," *Proc. Natl. Acad. Sci. U. S. A.* **17**, 315 (1931).
- I. Mezić, "Spectral properties of dynamical systems, model reduction and decompositions," *Nonlinear Dyn.* **41**, 309–325 (2005).
- M. O. Williams, I. G. Kevrekidis, and C. W. Rowley, "A data-driven approximation of the Koopman operator: Extending dynamic mode decomposition," *J. Nonlinear Sci.* **25**, 1307–1346 (2015).
- C. Schütte, P. Koltai, and S. Klus, "On the numerical approximation of the Perron-Frobenius and Koopman operator," *J. Comput. Dyn.* **3**, 51–79 (2016).

- <sup>17</sup>S. Klus, F. Nüske, S. Peitz, J.-H. Niemann, C. Clementi, and C. Schütte, “Data-driven approximation of the Koopman generator: Model reduction, system identification, and control,” *Physica D* **406**, 132416 (2020).
- <sup>18</sup>A. Mardt, L. Pasquali, H. Wu, and F. Noé, “VAMPnets for deep learning of molecular kinetics,” *Nat. Commun.* **9**, 5 (2018).
- <sup>19</sup>W. Chen, H. Sidky, and A. L. Ferguson, “Nonlinear discovery of slow molecular modes using state-free reversible VAMPnets,” *J. Chem. Phys.* **150**, 214114 (2019).
- <sup>20</sup>H. Wendland, *Scattered Data Approximation* (Cambridge University Press, 2004).
- <sup>21</sup>I. Steinwart and A. Christmann, *Support Vector Machines*, 1st ed. (Springer, New York, 2008).
- <sup>22</sup>M. O. Williams, C. W. Rowley, and I. G. Kevrekidis, “A kernel-based method for data-driven Koopman spectral analysis,” *J. Comput. Dyn.* **2**, 247–265 (2015).
- <sup>23</sup>S. Klus, I. Schuster, and K. Muandet, “Eigendecompositions of transfer operators in reproducing kernel Hilbert spaces,” *J. Nonlinear Sci.* **30**, 283–315 (2020).
- <sup>24</sup>S. Klus, F. Nüske, and B. Hamzi, “Kernel-based approximation of the Koopman generator and Schrödinger operator,” *Entropy* **22**, 722 (2020).
- <sup>25</sup>S. Klus, A. Bittracher, I. Schuster, and C. Schütte, “A kernel-based approach to molecular conformation analysis,” *J. Chem. Phys.* **149**, 244109 (2018).
- <sup>26</sup>A. Rahimi and B. Recht, “Random features for large-scale kernel machines,” in *Advances in Neural Information Processing Systems 20* (Curran Associates, Inc., 2007).
- <sup>27</sup>E. B. Davies, “Metastable states of symmetric Markov semigroups II,” *J. London Math. Soc.* **s2-26**, 541–556 (1982).
- <sup>28</sup>S. Klus, F. Nüske, P. Koltai, H. Wu, I. Kevrekidis, C. Schütte, and F. Noé, “Data-driven model reduction and transfer operator approximation,” *J. Nonlinear Sci.* **28**, 985–1010 (2018).
- <sup>29</sup>N. Aronszajn, “Theory of reproducing kernels,” *Trans. Am. Math. Soc.* **68**, 337–404 (1950).
- <sup>30</sup>D. MacKay, *Introduction to Gaussian Processes*, NATO ASI Series F Computer and Systems Sciences Vol. 168 (Springer, 1998), pp. 133–166.
- <sup>31</sup>S. Bochner, “Monotone funktionen, Stieltjessche integrale und harmonische analyse,” *Math. Ann.* **108**, 378–410 (1933).
- <sup>32</sup>A. Tompkins and F. Ramos, “Fourier feature approximations for periodic kernels in time-series modelling,” in *Proceedings of the AAAI Conference on Artificial Intelligence* (AAAI Publications, 2018), Vol. 32.
- <sup>33</sup>W. Zhang, T. Li, and C. Schütte, “Solving eigenvalue PDEs of metastable diffusion processes using artificial neural networks,” *J. Comput. Phys.* **465**, 111377 (2022).
- <sup>34</sup>M. Sarich, F. Noé, and C. Schütte, “On the approximation quality of Markov state models,” *Multiscale Model. Simul.* **8**, 1154–1177 (2010).
- <sup>35</sup>P. Deuffhard and M. Weber, “Robust Perron cluster analysis in conformation dynamics,” *Linear Algebra Appl.* **398**, 161–184 (2005).
- <sup>36</sup>A. Bittracher, P. Koltai, S. Klus, R. Banisch, M. Dellnitz, and C. Schütte, “Transition manifolds of complex metastable systems,” *J. Nonlinear Sci.* **28**, 471–512 (2018).
- <sup>37</sup>F. Nüske, R. Schneider, F. Vitalini, and F. Noé, “Variational tensor approach for approximating the rare-event kinetics of macromolecular systems,” *J. Chem. Phys.* **144**, 054105 (2016).
- <sup>38</sup>F. Nüske, P. Gelß, S. Klus, and C. Clementi, “Tensor-based computation of metastable and coherent sets,” *Physica D* **427**, 133018 (2021).
- <sup>39</sup>K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, “How fast-folding proteins fold,” *Science* **334**, 517–520 (2011).
- <sup>40</sup>B. Trendelkamp-Schroer, H. Wu, F. Paul, and F. Noé, “Estimation and uncertainty of reversible Markov models,” *J. Chem. Phys.* **143**, 174101 (2015).
- <sup>41</sup>T. Lelièvre, M. Rousset, and G. Stoltz, *Free Energy Computations: A Mathematical Perspective* (World Scientific, 2010).
- <sup>42</sup>L. Boninsegna, F. Nüske, and C. Clementi, “Sparse learning of stochastic dynamical equations,” *J. Chem. Phys.* **148**, 241723 (2018).
- <sup>43</sup>M. Hoffmann, M. Scherer, T. Hempel, A. Mardt, B. de Silva, B. E. Husic, S. Klus, H. Wu, N. Kutz, S. L. Brunton, and F. Noé, “Deeptime: A Python library for machine learning dynamical models from time series data,” *Mach. Learn.: Sci. Technol.* **3**, 015009 (2021).