The Journal of Chemical Physics

RESEARCH ARTICLE | JANUARY 17 2023

Data-driven construction of stochastic reduced dynamics encoded with non-Markovian features

Zhiyuan She ; Pei Ge ; Huan Lei

J. Chem. Phys. 158, 034102 (2023)
https://doi.org/10.1063/5.0130033

AIP Publishing

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

Data-driven construction of stochastic reduced dynamics encoded with non-Markovian features

Cite as: J. Chem. Phys. 158, 034102 (2023); doi: 10.1063/5.0130033
Submitted: 8 October 2022 • Accepted: 28 December 2022 •
Published Online: 17 January 2023

Zhiyuan She,[1] Pei Ge,[1] and Huan Lei[2,a)]

AFFILIATIONS

[1] Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, Michigan 48824, USA
[2] Department of Computational Mathematics, Science and Engineering and Department of Statistics and Probability, Michigan State University, East Lansing, Michigan 48824, USA

[a)] **Author to whom correspondence should be addressed:** leihuan@msu.edu

ABSTRACT

One important problem in constructing the reduced dynamics of molecular systems is the accurate modeling of the non-Markovian behavior arising from the dynamics of unresolved variables. The main complication emerges from the lack of scale separations, where the reduced dynamics generally exhibits pronounced memory and non-white noise terms. We propose a data-driven approach to learn the reduced model of multi-dimensional resolved variables that faithfully retains the non-Markovian dynamics. Different from the common approaches based on the direct construction of the memory function, the present approach seeks a set of non-Markovian features that encode the history of the resolved variables and establishes a joint learning of the extended Markovian dynamics in terms of both the resolved variables and these features. The training is based on matching the evolution of the correlation functions of the extended variables that can be directly obtained from the ones of the resolved variables. The constructed model essentially approximates the multi-dimensional generalized Langevin equation and ensures numerical stability without empirical treatment. We demonstrate the effectiveness of the method by constructing the reduced models of molecular systems in terms of both one-dimensional and four-dimensional resolved variables.

I. INTRODUCTION

Predictive modeling of multi-scale dynamic systems is a long-standing problem in many fields, such as biology, materials science, and fluid physics. One essential challenge arises from the high-dimensionality; numerical simulations of the full models often show limitations in the achievable spatiotemporal scales. Alternatively, reduced models in terms of a set of resolved variables are often used to probe the evolution on the scale of interest. However, the construction of reliable reduced models remains a highly non-trivial problem. In particular, for systems without a clear scale separation, the reduced dynamics often exhibits non-Markovian memory effects, where the analytic form is generally unknown. To close the reduced dynamics, existing methods are primarily based on the following two approaches. The first approach seeks various numerical approximations of the memory term by projecting the full dynamics onto the resolved variables based on frameworks such as the Mori–Zwanzig (MZ) formalism[1,2] or canonical models, such as the generalized Langevin equation (GLE).[3] Examples include the t-model approximation,[4] the Galerkin discretization,[5] regularized integral equation discretization,[6] the hierarchical construction,[7–11] and so on. Recent studies[12–15] based on the recurrent neural networks[16] provide a promising approach to learn the memory term of deterministic dynamics. Yet, for ergodic dynamics, how to impose the coherent noise term compensating for the unresolved variables remains open. The second approach parameterizes the memory term by a certain ansatz, e.g., the fictitious particle,[17] continued fraction,[18,19] and rational function,[20] such that the memory and the noise terms can be embedded in an extended Markovian dynamics.[17,19,21–28] In addition, non-Markovian models are represented by discrete dynamics with exogenous inputs in the form of NARMAX (nonlinear autoregression moving average with exogenous input),[29,30] SINN (statistics information neural network),[31] and Ab Initio GLE[32] that are parameterized for each specific time step. Recent work by Vroylandt *et al.*[32] presents an expectation-maximization method to parameterize the reduced model from the full model trajectories. The authors of Refs. 34 and 35 presented an efficient approach based on analyzing the force correlation function

to extract the memory function for the reduced dynamics of aqueous molecules under quadratic confinement potential; see also recent review[36] for further discussion. Despite the overall success, most studies focus on the cases with a scalar memory function. Notably, the reduced model of a two-dimensional GLE is constructed in Ref. 25. To the best of our knowledge, the systematic construction of stochastic reduced dynamics of multi-dimensional resolved variables remains under-explored.

Ideally, to obtain a reliable reduced model, the construction needs to accurately retain the non-Markovian features, enable certain modeling flexibility (e.g., the dimensionality of the resolved variables) and adaptivity (e.g., the order of approximation), and guarantee the numerical stability and robustness. In a recent study, we developed a Petrov–Galerkin approach[37] to construct the non-Markovian reduced dynamics by projecting the full dynamics into a subspace spanned by a set of projection bases in the form of the fractional derivatives of the resolved variables. The obtained reduced model is parameterized as extended stochastic differential equations by introducing a set of test bases. Different from most existing approaches, the construction does not rely on the direct fitting of the memory function. Non-local statistical properties can be naturally matched by choosing the appropriate bases, and the model accuracy can be systematically improved by introducing more basis functions to expand the projection subspace. Despite these appealing properties, the construction relies on the heuristic choices of the projection and test bases. Given the target number of basis, how to choose the optimal basis functions for the best representation of the non-Markovian dynamics remains an open problem. Furthermore, the numerical stability needs to be treated empirically. These issues limit the applications in complex systems with multi-dimensional resolved variables.

In this work, we aim to address the above issues by developing a new data-driven approach to construct the stochastic reduced dynamics of multi-dimensional resolved variables. The method is based on the joint learning of a set of non-Markovian features and the extended dynamic equation in terms of both the resolved variables and these features. Unlike the empirically chosen projection bases adopted in the previous work,[37] the non-Markovian features take an interpretable form that encodes the history of the resolved variables and are learned along with the extended Markovian dynamic such that they are *optimal* for the reduced model representation. In this sense, they represent the optimal subspace that embodies the non-Markovian nature of the resolved variables. The learning process enables the adaptive choices of the number of features and is easy to implement by matching the evolution of the correlation functions of the extended variables. In particular, the explicit form of the encoder function enables us to obtain the correlation functions of these features directly from the ones of the resolved variables rather than the time-series samples. The constructed model automatically ensures numerical stability, strictly satisfies the second fluctuation–dissipation theorem,[38] and retains the consistent invariant distribution.[39,40]

We demonstrate the method by modeling the dynamics of a tagged particle immersed in solvents and a polymer molecule. With the same number of features (or equivalently, the projection bases), the present method yields more accurate reduced models than the previous methods[23,37] due to the concurrent learning of the non-Markovian features. More importantly, reduced models with respect to multi-dimensional resolved variables can be conveniently constructed without the cumbersome efforts of matrix-valued kernel fitting and stabilization treatment. This is well-suited for model reduction in practical applications, where the constructed reduced models often need to retain the non-local correlations among the resolved variables. It provides a convenient approach to construct meso-scale models encoded with molecular-level fidelity and paves the way toward constructing reliable continuum-level transport model equations.[41,42]

Finally, it is worthwhile to mention that the present study focuses on the model reduction of ergodic dynamic systems where the full or part of the resolved variables are specified as known quantities that either retain a clear physical interpretation (e.g., the tagged particle position) or are experimentally accessible (e.g., the polymer end-to-end distance and the radius of gyration). Another relevant direction focuses on learning the slow or Markovian dynamics from the complex dynamic systems where the resolved variables are unknown *a priori*; we refer to Refs. 43–48 on learning resolved variables that retain the Markovianity, Refs. 49–54 on learning the slow dynamics on a non-linear manifold, and Refs. 55–58 on model reduction of the transfer operator.

II. METHODS

A. Problem setup

Let us consider the full model as a Hamiltonian system represented by a $6N$-dimensional phase vector $\mathbf{Z} = [\mathbf{Q}; \mathbf{P}]$, where $\mathbf{Q}$ and $\mathbf{P}$ are the position and momentum vectors, respectively. The equation of motion follows:

$$\dot{\mathbf{Z}} = \mathbf{S}\nabla H(\mathbf{Z}), \tag{1}$$

where $\mathbf{S} = \begin{pmatrix} 0 & \mathbf{I} \\ -\mathbf{I} & 0 \end{pmatrix}$ is the symplectic matrix and $H(\mathbf{Z})$ is the Hamiltonian function, and the initial condition is given by $\mathbf{Z}(0) = \mathbf{Z}_0$. For high-dimensional systems with $N \gg 1$, the numerical simulation of Eq. (1) can be computational expensive. It is often desirable to construct a reduced model with respect to a set of low-dimensional resolved variables $\mathbf{z}(t) := \phi(\mathbf{Z}(t))$, where $\phi : \mathbb{R}^{6N} \to \mathbb{R}^m$ represents the mapping from the full to the coarse-grained state space with $m \ll N$. With the explicit form of $H(\mathbf{Z})$ and $\phi(\mathbf{Z})$, the evolution of $\mathbf{z}(t)$ can be mapped from the initial values via the Koopman operator,[59] i.e., $\mathbf{z}(t) = \mathrm{e}^{t\mathscr{L}}\mathbf{z}(0)$, where the Liouville operator $\mathscr{L}\phi(\mathbf{Z}) = -((\nabla H(\mathbf{Z}_0))^T\mathbf{S}\nabla_{\mathbf{Z}_0})\phi(\mathbf{Z})$ depends on the full-dimensional phase vector $\mathbf{Z}$. Using the Duhamel–Dyson formula, the evolution of $\mathbf{z}(t)$ can be further formulated in terms of $\mathbf{z}$ based on the Mori–Zwanzig (MZ) projection formalism.[1,2] However, the numerical evaluation of the derived model relies on solving the full-dimensional orthogonal dynamics,[4] which can be still computational expensive.

In practice, the resolved variables are often defined by the position vector $\mathbf{Q}$. The MZ-formed reduced dynamics is often simplified into the GLEs, i.e.,

$$\begin{aligned} \dot{\mathbf{q}} &= \mathbf{M}^{-1}\mathbf{p}, \\ \dot{\mathbf{p}} &= -\nabla U(\mathbf{q}) - \int_0^t \boldsymbol{\theta}(t-\tau)\dot{\mathbf{q}}(\tau)\mathrm{d}\tau + \mathscr{R}(t), \end{aligned} \tag{2}$$

where $\mathbf{q} \in \mathbb{R}^m$ is the so-called collective variables, $\mathbf{M}$ is the mass matrix, $U(\mathbf{q})$ is the free energy function, $\boldsymbol{\theta}(t) : \mathbb{R}^+ \to \mathbb{R}^{m\times m}$ is a matrix-valued function representing the memory kernel, and $\mathscr{R}(t)$ is a stationary colored noise related to $\boldsymbol{\theta}(t)$ through the second fluctuation–dissipation condition,[37] i.e., $\langle \mathscr{R}(t)\mathscr{R}(0)^T \rangle = k_B T \boldsymbol{\theta}(t)$. It is worth mentioning that Eq. (2) is not exact based on the MZ formalism. In particular, the memory function generally depends on the resolved variables $\mathbf{z}$ and the noise term could be non-Gaussian; we refer to Ref. 60 for further discussion. Nevertheless, even for the simplified GLE form (2), the accurate construction of the reduced model could remain highly nontrivial. Specifically, the numerical simulation requires the explicit knowledge of both the free energy $U(\mathbf{q})$ and the memory function $\boldsymbol{\theta}(t)$. Several methods based on importance sampling[61–63] and temperature elevation[64–66] have been developed to construct the multi-dimensional free energy function. In real applications, the main challenge often lies in the treatment of the memory kernel $\boldsymbol{\theta}(t)$. In particular, for multi-dimensional collective variables $\mathbf{q}$, the efficient construction of numerically stable matrix-valued memory function remains under-explored.

In this study, we develop an alternative approach to learn the reduced model. Rather than directly constructing the memory function $\boldsymbol{\theta}(t)$ in Eq. (2), we seek a set of non-Markovian features from the full model, denoted by $\{\boldsymbol{\zeta}_i\}_{i=1}^n$, and establish a joint learning of the reduced Markovian dynamics in terms of both the resolved variables and these features, i.e.,

$$\mathrm{d}\tilde{\mathbf{z}} = \mathbf{g}(\tilde{\mathbf{z}})\mathrm{d}t + \boldsymbol{\Sigma}\mathrm{d}\mathbf{W}_t, \tag{3}$$

where $\tilde{\mathbf{z}} := [\mathbf{q}; \mathbf{p}; \boldsymbol{\zeta}_1; \ldots; \boldsymbol{\zeta}_n]$ represents the extended variables and $\mathbf{W}_t$ represents the standard Wiener process. In principle, any such extended system would generally lead to a non-Markovian dynamics for the resolved variables $\mathbf{z} = [\mathbf{q}; \mathbf{p}]$. However, the essential challenge is to determine $\{\boldsymbol{\zeta}_i\}_{i=1}^n$ so that the non-local statistical properties of $\mathbf{z}$ can be preserved with sufficient accuracy. In addition, the form of $\mathbf{g}(\cdot)$ and $\boldsymbol{\Sigma}$ will need to be properly designed such that the reduced model retains the consistent thermal fluctuations and density distribution. In particular, the introduction of auxiliary variables can be loosely considered as approximating the full-dimensional Koopman operator in a sub-space. However, different from Ref. 37, the features $\{\boldsymbol{\zeta}_i\}_{i=1}^n$ are not the empirically chosen projection bases; instead, they are learned along with model equation (3) for the *best approximation* of the non-local statistics. This essential difference enables the present method to generate more accurate reduced models and be easy to implement for multi-dimensional resolved variables without empirical treatment for numerical stability.

B. Non-Markovian features and the extended dynamics

To illustrate the essential idea, let us consider a solute molecule immersed solvent particles. To construct a reduced model (3) for the solute molecule, a main question is how to construct the auxiliary variables $\boldsymbol{\zeta} := [\boldsymbol{\zeta}_1; \boldsymbol{\zeta}_2; \ldots; \boldsymbol{\zeta}_n]$. Intuitively, $\boldsymbol{\zeta}_i$ should depend on the full-dimensional vector $\mathbf{Z}$ such that their evolution encodes certain unresolved dynamics orthogonal to the subspace spanned by $\mathbf{z}(t)$. A straightforward approach is to represent $\boldsymbol{\zeta}(t)$ as a function of $\mathbf{Z}(t)$, i.e., $\boldsymbol{\zeta} = \mathbf{h}(\mathbf{Z})$. However, the direct construction of the general formulation $\mathbf{h}(\mathbf{Z})$ would become impractical since the learning generally involves sampling the individual solvent particles near the solute molecule; the computational cost could become intractable due to the high-dimensionality of $\mathbf{Z}$.

To circumvent the above challenges, the key ascribes to formulate $\boldsymbol{\zeta}(t)$ such that it properly encodes the unresolved dynamics of $\mathbf{Z}(t)$ and, meanwhile, can be easily sampled. One important observation is that the history of $\mathbf{p}(t)$ naturally encodes the unresolved dynamics orthogonal to $\mathbf{z}(t)$ and the values can be conveniently obtained. To see this, we note that the dynamics follows $\dot{\mathbf{p}} = \mathscr{L}\mathbf{p}$, where the Liouville operator $\mathscr{L}\phi(\mathbf{Z}) = -((\nabla H(\mathbf{Z}_0))^T \mathbf{S} \nabla_{\mathbf{Z}_0})\phi(\mathbf{Z})$ depends on the full-dimensional vector $\mathbf{Z}$. Therefore, it is possible to construct $\boldsymbol{\zeta}(t)$ by some encoder functions in terms of the time history of $\mathbf{p}(t)$, i.e., $\mathbf{p}(\tau)$ with $\tau \le t$, such that they retain certain components orthogonal to $\mathbf{z}(t)$. This is somewhat similar to the study[41] on modeling the non-local constitutive dynamics of non-Newtonian fluids by learning a set of features from the polymer configuration space. The main difference is that the present features $\boldsymbol{\zeta}$ are non-Markovian in the temporal space.

Accordingly, we define a set of non-Markovian features $\{\boldsymbol{\zeta}_i\}_{i=1}^n$ by

$$\begin{aligned}\boldsymbol{\zeta}_i(t) &= \int_0^{+\infty} \boldsymbol{\omega}_i(\tau)\mathbf{v}(t-\tau)\mathrm{d}\tau \\ &\approx \sum_{j=1}^{N_w} \mathbf{w}_i(j\delta t)\mathbf{v}(t-j\delta t),\end{aligned} \tag{4}$$

where $\mathbf{v} := \mathbf{M}^{-1}\mathbf{p}$ represents the generalized velocity and $\boldsymbol{\omega}_i : \mathbb{R}^+ \to \mathbb{R}^{m\times m}$ represents the encoder function represented by N_w discrete weights $\{\mathbf{w}_i(j\delta t)\}_{j=1}^{N_w}$ whose values will be determined later.

$\boldsymbol{\zeta}_i(t)$ can be loosely viewed as a generalized convolution over the history of $\mathbf{v}(t)$ whose evolution encodes the orthogonal dynamics of $\mathbf{z}(t)$. Therefore, it is possible to learn a set of $\boldsymbol{\zeta}_i(t)$ such that the joint dynamics in terms of both $\mathbf{z}(t)$ and $\boldsymbol{\zeta}_i(t)$ can be well approximated by the extended Markovian model (3). Moreover, the linear form in terms of $\mathbf{v}(t)$ ensures that the invariant density function of $\boldsymbol{\zeta}_i(t)$ retains the Gaussian distribution consistent with $\mathbf{v}(t)$. We can further impose a constraint such that $\mathbf{v}(t)$ and $\boldsymbol{\zeta}_i(t)$ are uncorrelated. This leads to an additional quadratic term in the energy function of the extended system, i.e., $W(\mathbf{q}, \mathbf{p}, \boldsymbol{\zeta}) = U(\mathbf{q}) + \frac{1}{2}\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p} + \frac{1}{2}\boldsymbol{\zeta}^T\hat{\boldsymbol{\Lambda}}^{-1}\boldsymbol{\zeta}$, where $\hat{\boldsymbol{\Lambda}}$ represents the covariance matrix of the $\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_n$. The reduced dynamics can be written as

$$\mathrm{d}\begin{pmatrix}\mathbf{q}\\ \mathbf{p}\\ \boldsymbol{\zeta}\end{pmatrix} = \mathbf{G}\nabla W(\mathbf{q}, \mathbf{p}, \boldsymbol{\zeta})\mathrm{d}t + \boldsymbol{\Sigma}\mathrm{d}\mathbf{W}_t, \tag{5}$$

where the matrix $\mathbf{G} \in \mathbb{R}^{(2+n)m\times(2+n)m}$ takes the form

$$\mathbf{G} = \begin{pmatrix} 0 & \mathbf{I} & 0 & \cdots & 0 \\ -\mathbf{I} & & & & \\ 0 & & & & \\ \vdots & & \mathbf{J} & & \\ 0 & & & & \end{pmatrix} \begin{pmatrix} \mathbf{I} & 0 & 0 & \cdots & 0 \\ 0 & & & & \\ 0 & & \begin{bmatrix} \mathbf{I} & \\ & \hat{\boldsymbol{\Lambda}} \end{bmatrix} & & \\ \vdots & & & & \\ 0 & & & & \end{pmatrix}. \tag{6}$$

The matrix $\mathbf{J} \in \mathbb{R}^{nm\times nm}$ further determines the statistical properties of the resolved variables and will be learned along with the non-Markovian features $\{\boldsymbol{\omega}_i(t)\}_{i=1}^n$ from the training samples as discussed in Subsection II C. Given the reduced model in the form of Eqs. (5) and (6), the noise covariance matrix can be determined by

$$\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T = -\beta^{-1}(\mathbf{J}\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^T\mathbf{J}^T), \tag{7}$$

where $\beta = 1/k_BT$ and $\boldsymbol{\Lambda} = \mathrm{diag}(\mathbf{I}, \hat{\boldsymbol{\Lambda}})$. The form of $\boldsymbol{\Lambda}$ implies that $\mathbf{v}$ and $\boldsymbol{\xi}_i$ are uncorrelated and is consistent with the energy function of the extended system $W(\mathbf{q}, \mathbf{p}, \boldsymbol{\zeta})$. It also alleviates the non-negative constraint of $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T$ as discussed in Sec. II C. Furthermore, we can show that model (5) strictly satisfies the second-fluctuation–dissipation theorem. Specifically, the embedded kernel in Eq. (5) takes the form

$$\tilde{\boldsymbol{\theta}}(t) = -\left(\tilde{\mathbf{J}}_{11}\delta(t) + \tilde{\mathbf{J}}_{12}e^{\tilde{\mathbf{J}}_{22}t}\tilde{\mathbf{J}}_{21}\right), \tag{8}$$

where $\tilde{\mathbf{J}}_{11} = [\tilde{\mathbf{J}}]_{1:m,1:m}$, $\tilde{\mathbf{J}}_{12} = [\tilde{\mathbf{J}}]_{1:m,m+1:}$, and $\tilde{\mathbf{J}}_{21} = [\tilde{\mathbf{J}}]_{m+1:,1:m}$ are the sub-blocks of the matrix $\tilde{\mathbf{J}} := \mathbf{J}\boldsymbol{\Lambda}$. The colored noise $\tilde{\mathscr{R}}(t)$ in terms of $\mathbf{p}(t)$ is related to $\tilde{\boldsymbol{\theta}}(t)$ by

$$\left\langle \tilde{\mathscr{R}}(t)\tilde{\mathscr{R}}(t')^T \right\rangle = -\beta^{-1}\left(\tilde{\mathbf{J}}_{12}e^{\tilde{\mathbf{J}}_{22}(t-t')}\tilde{\mathbf{J}}_{21} + (\tilde{\mathbf{J}}_{11} + \tilde{\mathbf{J}}_{11}^T)\delta(t-t')\right) \tag{9}$$

with $t' < t$. Moreover, we can show that the reduce model retains the consistent invariant density function with the full model, i.e.,

$$\rho_{\mathrm{eq}} \propto \exp[-\beta W(\mathbf{q}, \mathbf{p}, \boldsymbol{\zeta})]. \tag{10}$$

We refer to Appendixes C and D for details.

We conclude this subsection with two remarks. (I) In principle, the mass matrix $\mathbf{M}$ further depends on $\mathbf{q}$. The authors of Ref. 25 reported that the varying mass matrix plays a secondary effect on the reduced dynamics of the molecular system therein; see also Ref. 60 for the cases of the nonlinear distance coordinate with constant mass. A constant mass matrix is adopted in the present study; reduced models with the varying mass matrix can be constructed by introducing an additional term in the conservative force and will be considered in the future study. (II) The non-Markovian features $\{\boldsymbol{\zeta}_i\}_{i=1}^n$ in the form of Eq. (4) can be generalized to retain the state-dependence, e.g., $\boldsymbol{\zeta}_i(t) = \int_0^{+\infty}\boldsymbol{\omega}_i(\tau, \mathbf{q}(\tau))\mathbf{v}(t-\tau)\mathrm{d}\tau$, which leads to a reduced model with state-dependent non-Markovian memory. In this study, we demonstrate the proposed learning framework by constructing the reduced model (5) that approximates the standard GLE (2) with state-independent memory function $\boldsymbol{\theta}(t)$. As shown in the numerical examples, although $\boldsymbol{\theta}(t)$ is not explicitly constructed, it is well approximated by the memory kernel embedded in the reduced model (5) by matching the evolution of the correlation matrices for both the resolved and extended variables. The learning of reduced models with the heterogeneous memory term will be systematically investigated in the future study.

C. Joint learning of the reduced dynamics

Construction of the above reduced models relies on the joint learning of the non-Markovian features (4) in the form of the encoder functions $\{\boldsymbol{\omega}_i(t)\}_{i=1}^n$ and the reduced dynamics (5) and (6) determined by the free energy $U(\mathbf{q})$ and the matrix $\mathbf{J}$. In this study, we represent the multi-dimensional free energy $U(\mathbf{q})$ using a neural network and parameterize it based on the restraint molecular dynamics.[67] We refer to Appendix B for details. Furthermore, the covariance of the noise term specified by Eq. (7) implies that $\mathbf{J}$ and $\boldsymbol{\Lambda}$ [i.e., the encoder functions $\boldsymbol{\omega}_i(t)$] need to satisfy the following constraint:

$$\mathbf{J}\boldsymbol{\Lambda} + \boldsymbol{\Lambda}\mathbf{J}^T \preccurlyeq 0. \tag{11}$$

Directly imposing condition (11) becomes cumbersome for the joint learning of $\mathbf{J}$ and $\boldsymbol{\omega}_i(t)$. Fortunately, this issue can be avoided by imposing an orthogonal constraint among the non-Markovian features, i.e.,

$$\begin{aligned}\left[\hat{\boldsymbol{\Lambda}}\right]_{ij} &:= \beta\langle\boldsymbol{\zeta}_i, \boldsymbol{\zeta}_j\rangle \\ &= \beta\sum_{k,k'}\left\langle \mathbf{w}_i(t-k\delta t)\mathbf{v}(k\delta t), \mathbf{w}_j(t-k'\delta t)\mathbf{v}(k'\delta t)\right\rangle \\ &= \delta_{ij}\mathbf{I}, \quad 1 \le i, j \le n,\end{aligned} \tag{12}$$

where the inner product $\langle\mathbf{f}(\mathbf{Z}), \mathbf{g}(\mathbf{Z})\rangle = \int \mathbf{f}(\mathbf{Z})\mathbf{g}(\mathbf{Z})^T\rho(\mathbf{Z})\mathrm{d}\mathbf{Z}$ is defined with respect to the equilibrium density function of the full model $\rho(\mathbf{Z}) = e^{-\beta H(\mathbf{Z})}/\int e^{-\beta H(\mathbf{Z})}\mathrm{d}\mathbf{Z}$. In addition, we also impose the orthogonal constraints such that $\boldsymbol{\zeta}$ and $\mathbf{p}$ are uncorrelated. Therefore, condition (11) can be transformed into seeking $\mathbf{J}$ such that $\mathbf{J} + \mathbf{J}^T \preccurlyeq 0$, and we represent $\mathbf{J}$ by

$$\mathbf{J} = -\mathbf{L}\mathbf{L}^T + \mathbf{J}^A, \tag{13}$$

where $\mathbf{L} \in \mathbb{R}^{(n+1)m\times(n+1)m}$ is the lower-triangle matrix with positive diagonal elements and $\mathbf{L}\mathbf{L}^T$ represents the Cholesky decomposition of a symmetric positive-definite matrix. $\mathbf{J}^A$ represents an anti-symmetric matrix. Unlike the studies[19,21] based on the direct kernel approximation, we note that $\mathbf{J}$ takes a more general form and is not restricted to be diagonal or tri-diagonal.

With the proper form of $\mathbf{J}$, we can cast the reduced dynamics into the evolution of the correlation matrices by multiplying $\mathbf{v}(0)$ to both sides of Eq. (5), i.e.,

$$\underbrace{\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\langle\mathbf{M}\mathbf{v}, \mathbf{v}(0)\rangle \\ \langle\boldsymbol{\zeta}_1, \mathbf{v}(0)\rangle \\ \vdots \\ \langle\boldsymbol{\zeta}_n, \mathbf{v}(0)\rangle\end{pmatrix}}_{\mathbf{C}_1(t)} + \underbrace{\begin{pmatrix}\langle\nabla U(\mathbf{q}), \mathbf{v}(0)\rangle \\ 0 \\ \vdots \\ 0\end{pmatrix}}_{\mathbf{C}_0(t)} = \mathbf{J}\underbrace{\begin{pmatrix}\langle\mathbf{v}, \mathbf{v}(0)\rangle \\ \langle\boldsymbol{\zeta}_1, \mathbf{v}(0)\rangle \\ \vdots \\ \langle\boldsymbol{\zeta}_n, \mathbf{v}(0)\rangle\end{pmatrix}}_{\mathbf{C}_2(t)}, \tag{14}$$

where the correlation matrices $\langle\boldsymbol{\zeta}_i(t), \mathbf{v}(0)\rangle$ can be directly obtained from the correlation matrix of the resolved variables $\langle\mathbf{v}(t), \mathbf{v}(0)\rangle$, given the encoder weights, i.e.,

$$\langle\boldsymbol{\zeta}_i(t), \mathbf{v}(0)\rangle = \sum_{j=1}^{N_w}\mathbf{w}_i(t_j)\langle\mathbf{v}(t-t_j), \mathbf{v}(0)\rangle,$$

where $t_j = j\delta t$ and encoder weights $\mathbf{w}_i(t_j)$ will be learned jointly. Therefore, we are able to construct $\mathbf{J}$ from the pre-computed, noise-free correlation matrices instead of the on-the-fly computation from the time-series samples of $\mathbf{q}$ and $\mathbf{p}$. The reduced model can be trained by minimizing the following loss function:

$$L_C = \sum_{j=1}^{N_t} \left\| \mathbf{C}_1'(t_j) + \mathbf{C}_0(t_j) - \mathbf{J}\mathbf{C}_2(t_j) \right\|^2 \quad L_\Lambda = \|\mathbf{\Lambda} - \mathbf{I}\|^2, \qquad L = \lambda_C L_C + \lambda_\Lambda L_\Lambda, \tag{15}$$

where $\mathbf{C}_1 = [\langle \mathbf{M}\mathbf{v}, \mathbf{v}(0)\rangle; \langle \boldsymbol{\zeta}_1, \mathbf{v}(0)\rangle; \ldots; \langle \boldsymbol{\zeta}_n, \mathbf{v}(0)\rangle]$ and $\mathbf{C}_0$ and $\mathbf{C}_2(t)$ are defined similarly in Eq. (14). λ_C and λ_Λ are the hyper-parameters. t_j refers to the discrete time points, and N_t represents the total number of sample points of the correlation matrices obtained from the full model. The loss term L_C ensures that the non-local statistical properties of the resolved variables can be accurately preserved, while the loss term L_Λ ensures the aforementioned orthogonality among the non-Markovian features. To simulate the constructed model, we always set $\hat{\mathbf{\Lambda}} = \mathbf{I}$ such that $\mathbf{J}$ in the form of Eq. (13) strictly satisfies the semi-positive definiteness condition. We emphasize that the non-Markovian encoder weights $\{\mathbf{w}_i(t_j)\}_{j=1}^{N_w}$ do not explicitly appear in the loss function. However, they are involved in the training process along with $\mathbf{J}$ since the correlation functions $\mathbf{C}_1$ and $\mathbf{C}_2$ further depend on the definition of $\boldsymbol{\zeta}_i$, i.e., they are concurrently learned for the best approximation of the extended Markovian dynamics of $[\mathbf{q}; \mathbf{p}; \boldsymbol{\zeta}]$. As shown in Sec. III, this joint learning of both the non-Markovian features and the dynamic equations enables us to probe the optimal representation of the reduced models that lead to more accurate numerical results than the ones constructed by the pre-selected bases and can be easily implemented for models with multi-dimensional resolved variables. In this study, we choose $N_t = 5000$ for all the numerical examples and choose $N_w = 1800$ for the one-dimensional reduced model and 1200 for the four-dimensional reduced model, respectively. The training is conducted by the adaptive moment estimation (ADAM) optimization algorithm[68] where 1000 points are randomly selected per each training step (see Appendix E for the summary of training process).

We conclude this subsection with the following remarks: (I) Instead of Eq. (14), the reduced dynamics can be also cast into the evolution of the correlation matrices by multiplying $\mathbf{q}(0)$ to both sides of Eq. (5). For the present study, we found that both formulations yield accurate reduced models. (II) Rather than learning the full sets of non-Markovian features, we can fix one of them as $\mathbf{M}\dot{\mathbf{v}} + \nabla U(\mathbf{q})$; this ensures that the time-derivatives of correlation functions of the reduced model can accurately match the values of the full model near $t = 0$. In the numerical examples presented in Sec. III, all the reduced models are constructed with this choice. For a simple notation, we set it to be the last feature. For example, the fourth-order reduced model is constructed using four non-Markovian features. $\boldsymbol{\zeta}_1$, $\boldsymbol{\zeta}_2$, and $\boldsymbol{\zeta}_3$ take the form of Eq. (4), and $\boldsymbol{\zeta}_4$ is set to be $\mathbf{M}\dot{\mathbf{v}} + \nabla U(\mathbf{q})$. (III) In principle, for reduced models of Hamiltonian systems, the form of matrix $\mathbf{J}$ can be further restricted to

$$\mathbf{J} = -\operatorname{diag}(0, \hat{\mathbf{L}}\hat{\mathbf{L}}^T) + \mathbf{J}^A, \tag{16}$$

where $\hat{\mathbf{L}} \in \mathbb{R}^{nm\times \mathrm{nm}}$ is a lower-triangle matrix. Equation (16) ensures that the embedded kernel in Eq. (5) does not contain the Markovian memory term [i.e., $(\mathbf{J}_{11} + \mathbf{J}_{11}^T)\delta(t)$]. $\tilde{\boldsymbol{\theta}}(t)$ recovers the form of standard GLE, and the second fluctuation–dissipation relationship shown in Eq. (9) recovers the standard form, i.e., $\langle \tilde{\mathscr{R}}(t)\tilde{\mathscr{R}}(t')^T \rangle = \beta^{-1}\tilde{\boldsymbol{\theta}}(t - t')$. In this study, both forms yield accurate numerical results; the contribution of the Markovian term constructed by Eq. (13) is less than 1%.

III. NUMERICAL RESULTS

A. A tagged particle in aqueous environment

We demonstrate our method by modeling a tagged particle immersed in solvent particles. The particle interaction is governed by the pairwise force,

$$\mathbf{F}_{ij}(\mathbf{Q}_{ij}) = \begin{cases} f_0(1 - Q_{ij}/r_c)\mathbf{e}_{ij}, & Q_{ij} \le r_c, \\ 0, & Q_{ij} > r_c, \end{cases}$$

where $\mathbf{Q}_i$ and $\mathbf{Q}_j$ are the positions of ith and jth particles. $\mathbf{Q}_{ij} = \mathbf{Q}_i - \mathbf{Q}_j$, $Q_{ij} = \|\mathbf{Q}_i - \mathbf{Q}_j\|$, and $\mathbf{e}_{ij} = \frac{\mathbf{Q}_{ij}}{Q_{ij}}$, and r_c is the cut-off distance. The full system consists of 4000 particles in a $10 \times 10 \times 10$ cubic box with periodic boundary condition along each direction. We set $f_0 = 50$, $r_c = 1$, and the particle mass to be unit. A Nosé–Hoover thermostat is used with $k_BT = 0.5$ and time step $\delta t = 2 \times 10^{-3}$. 128 samples are collected from a production stage of 6×10^5 steps, which are used as the initial conditions of NVE simulations of the full model for a production stage of 1×10^5 steps using the velocity-Verlet integrator.

The reduced dynamics in terms of the tagged particle is constructed in the form of Eq. (5) along with the learning of the non-Markovian features $\{\boldsymbol{\zeta}_i\}_{i=1}^n$. The free energy $U(\mathbf{q})$ vanishes for this case. For comparison, we also construct the reduced model using the previous approaches based on the Petrov–Galerkin projection (named as fixed-basis)[37] and the rational function approximation.[23] Figure 1(a) shows the velocity correlation function of constructed models using two non-Markovian features or, equivalently, two projection bases, as well as the second-order rational function approximation. The model constructed by the present (named as the joint-learning) method shows the best agreement with the full model based on molecular dynamics (MD) simulations. The model accuracy can be further examined by the evolution of probability density function (PDF) of the particle velocity. Specifically, we fix the velocity to be zero as $t = 0$ and sample the instantaneous PDF thereafter. Figure 1(b) shows the obtained PDF at $t = 0.06$. The present approach yields more accurate result than the Petrov–Galerkin method. As shown in Fig. 1(c), the accuracy of the reduced model shows further improvement as we increase the number of non-Markovian features. In particular, the reduced model with the fourth-order approximation can accurately capture the oscillations over the full regime. Figure 1(d) shows the obtained encoder weights of the fourth-order approximation. All of the three encoder functions show pronounced oscillations near $t = 0$ and decay to 0 for large t. Unlike the empirically chosen fractional-derivative bases in Ref. 37, the present method enables the encoders

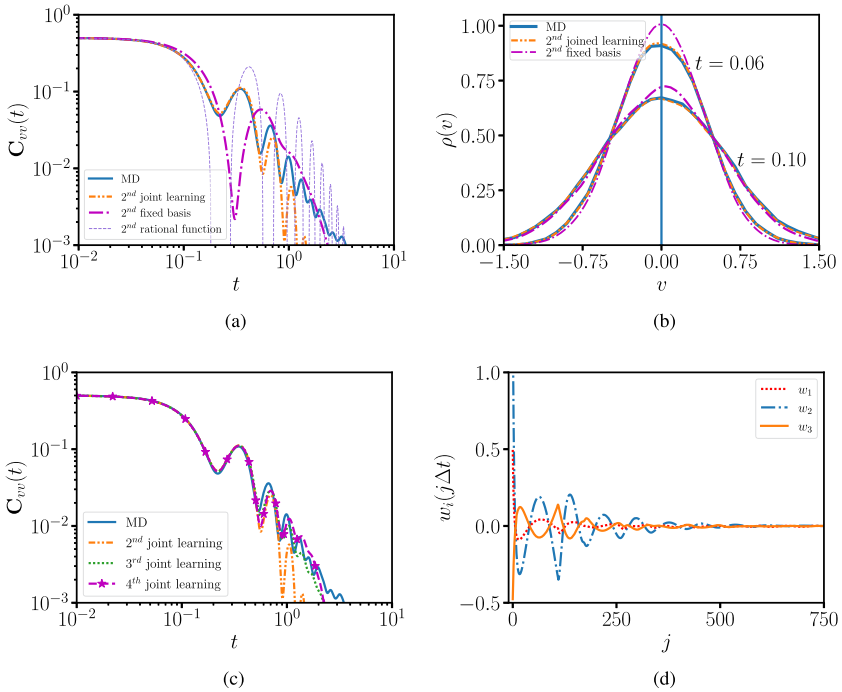

FIG. 1. Numerical results for a tagged particle in the solvent particle bath. (a) Velocity correlation function $C_{vv}(t)$ obtained from the full MD model and the reduced models constructed by the present method based on the joint learning approximation, the rational function approximation,[23] and the Petrov–Galerkin projection with fixed bases.[37] (b) Predicted evolution of the probability density function of the particle velocity obtained from the full MD and the different reduced models with the second-order approximation. The initial velocity v is set to 0 (the vertical line). (c) $C_{vv}(t)$ obtained from the full MD model and different orders of the present joint learning approximation. (d) Encoder weights for the three non-Markovian features obtained from the present joint learning with the fourth-order approximation.

to be optimized for the best approximation of the non-local statistics and therefore yields more accurate results.

B. One-dimensional reduced model of a polymer molecule

We consider the reduced dynamics of a polymer molecule consisting of $N = 16$ atoms. The intramolecular potential is governed by

$$V_{\mathrm{mol}}(\mathbf{Q}) = \sum_{i\neq j}^{N} V_{\mathrm{p}}(Q_{ij}) + \sum_{i=1}^{N_b} V_{\mathrm{b}}(l_i) + \sum_{i=1}^{N_a} V_{\mathrm{a}}(\theta_i) + \sum_{i=1}^{N_d} V_{\mathrm{d}}(\phi_i), \quad (17)$$

where l_i, θ_i, and ϕ_i represent the individual bond length, bond angle, and dihedral angle, respectively. V_{p}, V_{b}, V_{a}, and V_{d} represent the pairwise Lennard-Jones, finite extensible nonlinear elastic bond, harmonic angle, and multi-harmonic dihedral interactions whose explicit forms are specified in Appendix A. The atom mass is set to unit, the thermal temperature k_BT is set to 1.0, and the time step δt is set to be 1×10^{-3}. 512 samples are collected from a production stage of 3×10^6 steps, which are used as the initial conditions of NVE simulations of the full model for a production stage of 1×10^6 steps using the velocity-Verlet integrator. Figure 2 shows a sketch of the polymer molecule.

To examine the effectiveness of the present method, we first construct a 1D reduced dynamics in terms of the end-to-end distance $q_1 = \|\mathbf{Q}_1 - \mathbf{Q}_N\|$ as done in the previous work[37] based on the Petrov–Galerkin method and compare the numerical results obtained from the two methods. Figure 3(a) shows the velocity correlation function $C_{vv}(t) = \langle v_1(t)v_1(0)\rangle$ obtained from the full MD and different orders of fixed-basis and joint-learning approximations. With the same order of approximation, the current method yields better agreement with the MD results. Specifically, the second-order model of the current method can capture the pattern around $t = 4$ and the fourth-order model can capture the patterns around

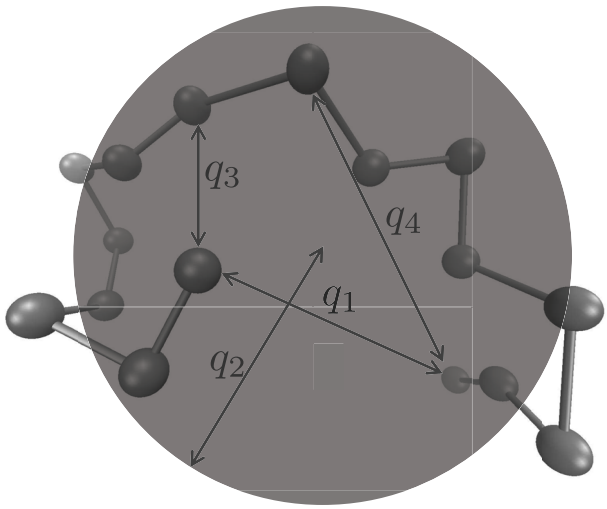

FIG. 2. A sketch of a chain-molecule represented by united atom model. Reduced models are constructed with respect to a four-dimensional resolved vector **q**, which represents the end-to-end distance (q_1), the radius of gyration (q_2), and the end-to-middle distances (q_3 and q_4), respectively.

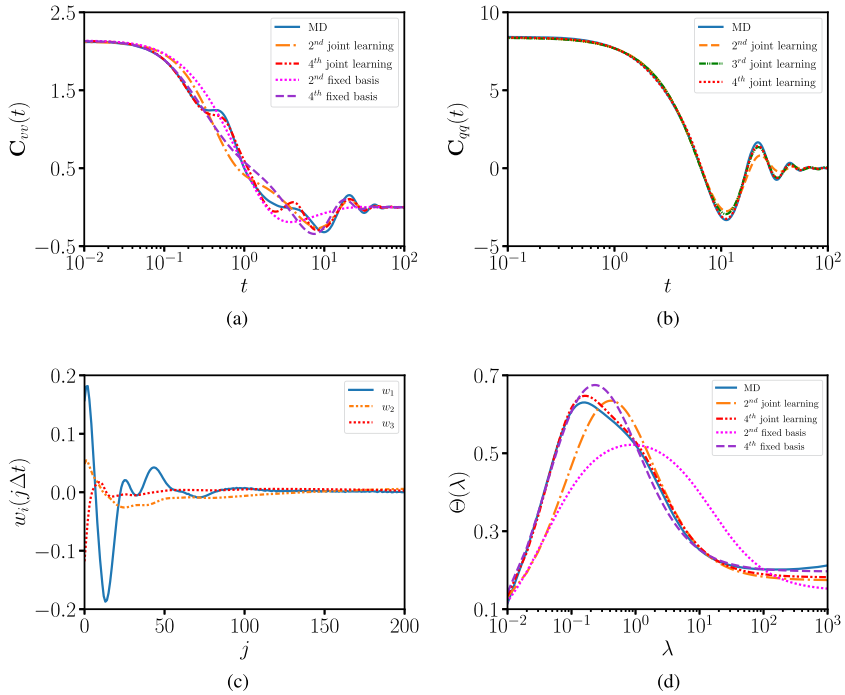

FIG. 3. Numerical results of a one-dimensional reduced model representing the dynamics of the end–end distance of a polymer molecule system. (a) and (b) Velocity correlation function $C_{vv}(t)$ and the Laplace transform of the memory function $\Theta(\lambda)$ obtained from full MD simulations and the different orders of the present joint learning approximation and the Petrov–Galerkin projection with fixed basis approximation. (c) Displacement correlation function $C_{qq}(t)$ obtained from the full MD and different orders of the joint learning approximation. (d) Encoder weights for the three non-Markovian features of the reduced model with the fourth-order approximation.

$t = 0.4$ and 4. However, the previous method with the same order approximation shows limitation to accurately capture these two patterns.

Figure 3(b) shows the displacement auto-correlation function $C_{qq}(t) = \langle q_1(t)q_1(0)\rangle$ obtained from full MD and the reduced models constructed by the present method with different number of non-Markovian features. As we introduce more features, the predicted correlation functions approaches the MD results. In particular, the fourth-order model can capture the oscillations of the MD results at $t = 10$ and 25. Figure 3(c) shows the encoder weights of non-Markovian features for the fourth-order approximation. Similar to the tagged particle system, the encoder functions exhibit pronounced oscillations at the short time and decay to zero at longer time.

The accuracy of the constructed reduced models can be further examined by comparing the embedding memory kernels $\tilde{\boldsymbol{\theta}}(t)$ with the full MD model. Figure 3(d) shows the Laplace transform of the memory kernel of the reduced models $\tilde{\boldsymbol{\Theta}}(\lambda) = \int_0^{+\infty}\tilde{\boldsymbol{\theta}}(t)\exp(-t/\lambda)\mathrm{d}t$. The MD kernel $\boldsymbol{\Theta}(\lambda)$ is obtained by $\boldsymbol{\Theta}(\lambda) = -\mathbf{G}(\lambda)\mathbf{H}(\lambda)^{-1}$, where $\mathbf{G}(\lambda)$ and $\mathbf{H}(\lambda)$ are the Laplace transform of the correlation matrices $\mathbf{g}(t) = \langle \mathbf{M}\dot{\mathbf{v}}(t) + \nabla U(\mathbf{q}), \mathbf{q}(0)\rangle$ and $\mathbf{h}(t) = \langle \mathbf{v}(t), \mathbf{q}(0)\rangle$. Compared with the previous method, the current method yields better agreement with MD results. Specifically, the second- and fourth-order of the joint learning approximation and the fourth-order of the fixed basis approximation show good agreement with the MD result $\boldsymbol{\Theta}(\lambda)$ for λ between 1 and 1000. Furthermore, the fourth-order model of the joint learning approximation can further capture the pronounced peak regime of the MD results near $\lambda = 0.1$. We emphasize that the *memory kernel* $\tilde{\boldsymbol{\theta}}(t)$ *is not explicitly constructed during the learning process*; $\tilde{\boldsymbol{\theta}}(t)$ approaches $\boldsymbol{\theta}(t)$ as we impose constraint (14) such that the correlation matrices of the reduced dynamics match the ones of the full model. This enables us to circumvent the direct fitting of the matrix-valued memory function for multi-dimensional GLEs and efficiently construct the numerically stable reduced model that retains the non-local statistics and coherent noise as shown in the following example.

C. Four-dimensional reduced model of a polymer molecule

Finally, we construct a reduced model in terms of a four-dimensional resolved vector $\mathbf{q} = [q_1, q_2, q_3, q_4]$ defined by

$$\begin{aligned} q_1 &= \|\mathbf{Q}_1 - \mathbf{Q}_N\|, \\ q_2^2 &= \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{Q}_i - \mathbf{Q}_c\|^2, \quad \mathbf{Q}_c = \frac{1}{N}\sum_{i=1}^{N}\mathbf{Q}_i, \\ q_3 &= \left\|\mathbf{Q}_{\lfloor \frac{N}{2}\rfloor} - \mathbf{Q}_1\right\|, \\ q_4 &= \left\|\mathbf{Q}_{\lceil \frac{N}{2}\rceil} - \mathbf{Q}_N\right\|, \end{aligned} \tag{18}$$

where q_1, q_2, q_3, and q_4 represent the end-to-end distance, radius of gyration, and two center-to-end distances, respectively. The four-dimensional free energy function $U(\mathbf{q})$ is constructed by matching the average force sampled from the restraint molecular dynamics and represented by a neural network; we refer to Appendix B for details. Rather than constructing the four-dimensional GLE kernel $\boldsymbol{\theta}(t)$, we directly learn the reduced model (5) by minimizing the loss function (15).

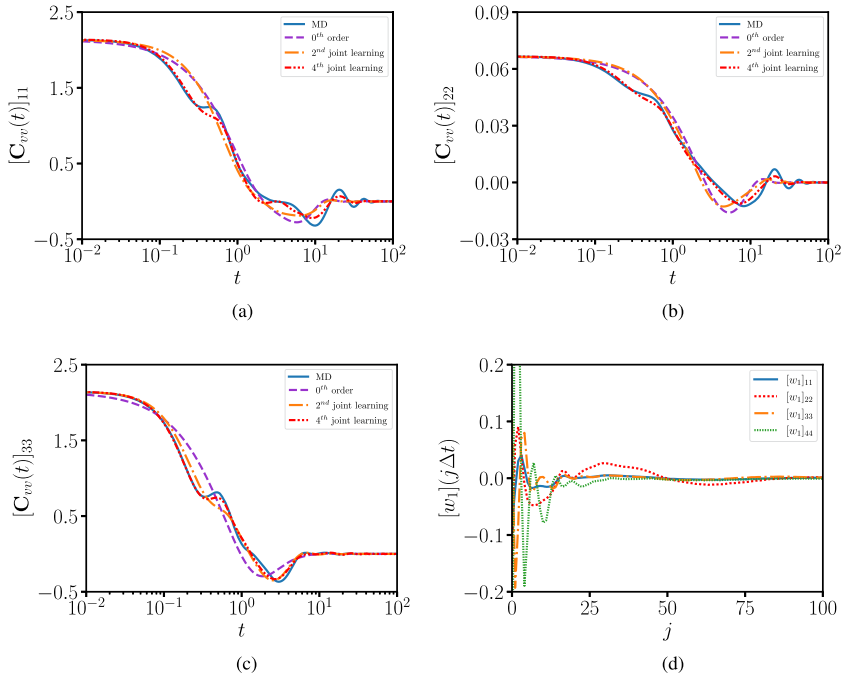

FIG. 4. Numerical results of a four-dimensional reduced model representing the dynamics of a polymer molecule system with conformation states characterized by the resolved variables **q** defined by Eq. (18). (a)–(c) Diagonal components of velocity correlation function $\mathbf{C}_{vv}(t) = \langle \mathbf{v}(t)\mathbf{v}(0)^T \rangle$. $[\mathbf{C}_{vv}(t)]_{44}$ is not shown as it is similar to $[\mathbf{C}_{vv}(t)]_{33}$. (d) Constructed encoder weights of the first non-Markovian features $\boldsymbol{\zeta}_1$ for the reduced model of the fourth-order approximation.

Figures 4(a)–4(c) show the diagonal components of the velocity correlation matrix $\mathbf{C}_{vv}(t) = \langle \mathbf{v}(t)\mathbf{v}(0)^T \rangle$ obtained from the full MD and the reduced models using different order approximations. Specifically, the components $[\mathbf{C}_{vv}(t)]_{11}$ and $[\mathbf{C}_{vv}(t)]_{33}$ show similar values near $t = 0$ since both q_1 and q_3 characterize the distances between the individual particles, e.g., $v_1 = (\mathbf{Q}_1 - \mathbf{Q}_N) \cdot (\mathbf{V}_1 - \mathbf{V}_N)/\|\mathbf{Q}_1 - \mathbf{Q}_N\|$. As the distribution of the velocity difference between the two free-end particles follows $\mathcal{N}(0, 2k_BT\mathbf{I})$, the variance of v_1 is $2k_BT$. Similar argument also holds for v_3 and v_4. On the long-time scale, $[\mathbf{C}_{vv}(t)]_{11}$ and $[\mathbf{C}_{vv}(t)]_{22}$ decay much slower than $[\mathbf{C}_{vv}(t)]_{33}$ and $[\mathbf{C}_{vv}(t)]_{44}$ and show pronounced oscillations near $t = 10$ and 25. The differences can be understood as follow: Compared with the end-to-middle distances q_3 and q_4, the end-to-end distance q_1 and radius of gyration q_2 represent the global states of the

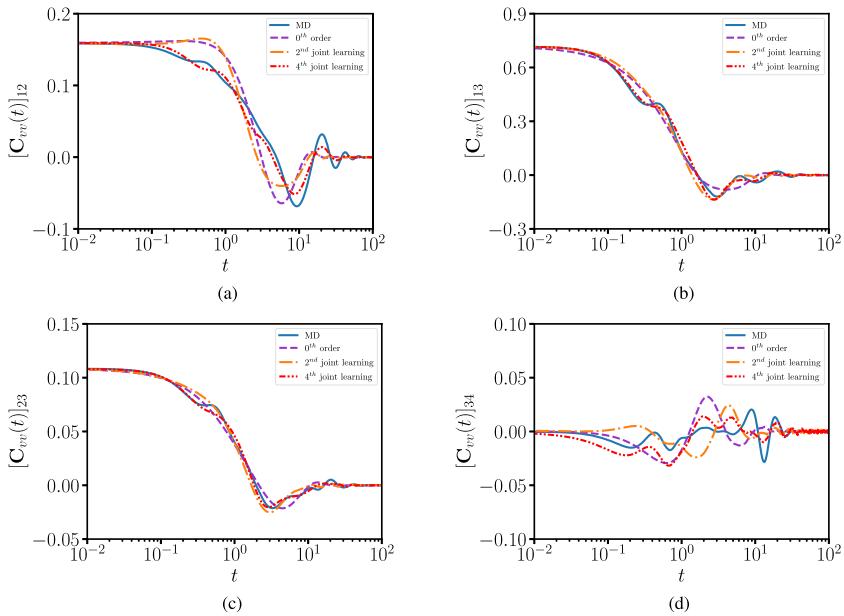

FIG. 5. (a)–(d) Off-diagonal components of the velocity correlation function $\mathbf{C}_{vv}(t)$ for a polymer molecule system whose conformation states are characterized by a four-dimensional resolved vector **q** defined by Eq. (18).

molecular conformation. Based on the scaling law of the idealized Gaussian chain model,[69] the relaxation time of q_1 and q_2 is proportional to N^2. Accordingly, $[\mathbf{C}_{vv}(t)]_{11}$ decays four times slower than $[\mathbf{C}_{vv}(t)]_{33}$, which is qualitatively consistent with the present numerical results.

The transient dynamics of the correlation functions can be accurately captured by the reduced model. As we increase the number of non-Markovian features, the predictions show better agreement with MD results. Specifically, the zeroth-order (i.e., Langevin) model is insufficient to capture the patterns around 0.5 and 5. The second-order model yields an accurate prediction for $[\mathbf{C}_{vv}(t)]_{33}$ but less accurate predictions for $[\mathbf{C}_{vv}(t)]_{11}$ and $[\mathbf{C}_{vv}(t)]_{22}$. The fourth-order model yields good agreement for all the components over the full regime. Figure 4(d) shows the encoder weights of the first non-Markovian feature $\boldsymbol{\zeta}_1$, which naturally encode the non-local statistics among the resolved variables and decay to 0 at large time.

Figure 5 shows the off-diagonal components of the velocity correlation matrix $\mathbf{C}_{vv}(t)$. Similar to the diagonal components, $[\mathbf{C}_{vv}(t)]_{12}$ represents the coupling between the dynamics of two global conformation states and, therefore, exhibits the longest correlation with pronounced oscillations at $t = 10$ and 25. On the other hand, $[\mathbf{C}_{vv}(t)]_{13}$ and $[\mathbf{C}_{vv}(t)]_{23}$ represent the coupling between a global state and a semi-global state and, therefore, exhibit intermediate correlation. In addition, $[\mathbf{C}_{vv}(t)]_{34}$ exhibits weaker correlation compared with the other components since the coupling between the dynamics of q_3 and q_4 is mainly governed by the local bond- and angle-interactions associated with eighth and nineth atom. The predictions of the second-order reduced model show fairly good agreement with the full MD results for $[\mathbf{C}_{vv}(t)]_{13}$ and $[\mathbf{C}_{vv}(t)]_{23}$ but less agreement for $[\mathbf{C}_{vv}(t)]_{12}$. The fourth-order reduced model yields good agreement for all the components.

Figure 6 shows the components of the embedded matrix-valued kernels in the Laplace space obtained from the full MD and the reduced models. In particular, $\tilde{\boldsymbol{\Theta}}(\lambda)$ obtained from the second-order model shows good agreement with $\boldsymbol{\Theta}(\lambda)$ obtained from the full MD within the regime of large λ. The fourth-order model yields good agreement over the full regime, which is consistent with the accurate prediction of the velocity correlation functions shown in Figs. 4 and 5 [see also Appendix F for $\boldsymbol{\theta}(t)$]. While the kernel function $\boldsymbol{\theta}(t)$ is not explicitly constructed in the present method, the accurate recovery of $\boldsymbol{\Theta}(\lambda)$ verifies that the constructed models faithfully retain the non-Markovian dynamics of the resolved variables.

IV. SUMMARY

In this study, we developed a data-driven approach to accurately learn the stochastic reduced dynamics of full Hamiltonian systems with non-Markovian memory. The method essentially provided an efficient approach to approximate the multi-dimensional generalized Langevin equation. Rather than directly fitting the matrix-valued memory kernel, the present method seeks a set of non-Markovian features whose evolution naturally encodes with the orthogonal dynamics of the resolved variables and establishes a joint learning of the extended dynamics in terms of both the resolved variables and the non-Markovian features. Compared with the previous studies based on the rational function approximation[23] and the Petrov–Galerkin projection[37] with the pre-selected fractional derivative bases, the present method enables us to probe the

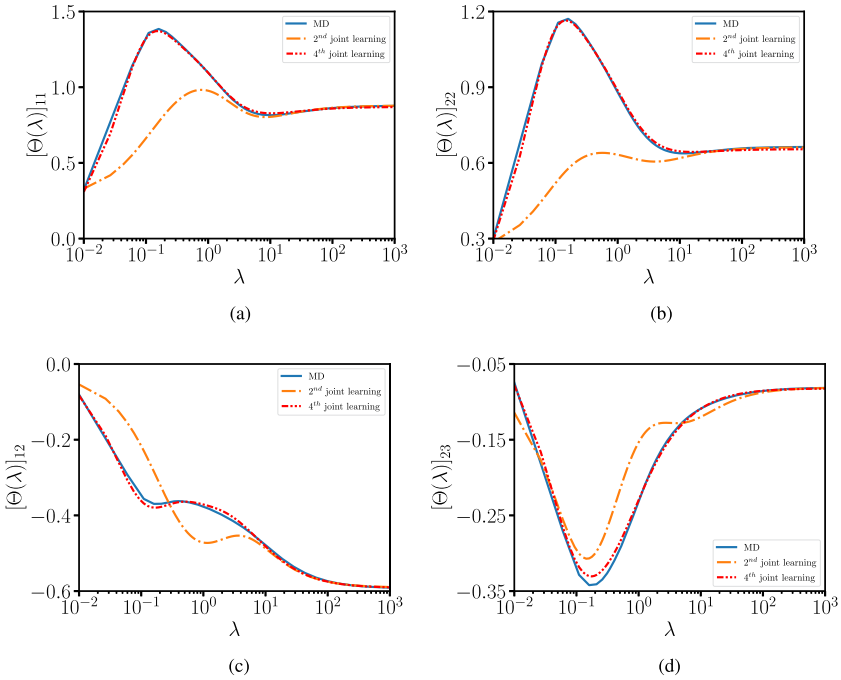

FIG. 6. (a)–(d) Components of the embedded matrix-valued kernel $\boldsymbol{\Theta}(\lambda)$ in the Laplace space obtained from the full MD and a four-dimensional reduced model of a polymer molecule system.

optimal representation of the reduced dynamics through the joint learning of the non-Markovian features. The constructed features retain a clear physical interpretation and can be loosely viewed as the convolution of the velocity history. This enables us to construct the proper learning formulation such that the reduced dynamics strictly preserves the second fluctuation–dissipation theorem and retains the consistent invariant density distribution. Moreover, the learning process does not require the on-the-fly computation of the time correlations of these features from the time-series samples and automatically ensures numerical stability of the constructed model without empirical treatment. This is particularly well-suited for the construction of reduced dynamics of complex systems, such as the conformation dynamics of macromolecular systems, where multi-dimensional resolved variables are often needed to characterize the transition dynamics with non-local cross correlations among the variables.

Despite the overall success, there are several open questions that are worth further exploration. For instance, the encoders are currently formulated as the standard linear convolution operators. Other nonlinear formulations may further facilitate the retaining of the non-Markovian features. In addition, the proper design of the encoder functions that represent state-dependent features may enable us to faithfully construct the reduced dynamics retaining the state-dependent memory term. We leave these issues for future studies.

ACKNOWLEDGMENTS

This work was supported by the Extreme Science and Engineering Discovery Environment (XSEDE) Bridges at the Pittsburgh Supercomputing Center through Allocation No. MTH210005, the Strategic Partnership Grant at Michigan State University, and the National Science Foundation under Grant No. DMS-2110981. We thank the two anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Zhiyuan She: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Visualization (lead); Writing – original draft (equal). **Pei Ge**: Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal). **Huan Lei**: Conceptualization (lead); Data curation (equal); Formal analysis (equal); Funding acquisition (lead); Investigation (equal); Methodology (lead); Project administration (lead); Visualization (equal); Writing – original draft (lead).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

TABLE I. Parameters of the FENE bond interactions.

Type	k_s	l_0
1	0.4	1.8
2	0.64	1.6
3	0.32	1.8

APPENDIX A: MICROSCALE MODEL OF THE POLYMER MOLECULE

The polymer molecule is modeled as a bead-spring chain consisting of four sub-units. Each sub-unit consists of four atoms. The full potential is given by

$$V_{\mathrm{mol}}(\mathbf{Q}) = \sum_{i\neq j}^{N} V_{\mathrm{p}}(Q_{ij}) + \sum_{i=1}^{N_b} V_{\mathrm{b}}(l_i) + \sum_{i=1}^{N_a} V_{\mathrm{a}}(\theta_i) + \sum_{i=1}^{N_d} V_{\mathrm{d}}(\phi_i), \quad \text{(A1)}$$

where V_{p}, V_{b}, V_{a}, and V_{d} represent the pairwise, bond, angle, and dihedral interactions whose detailed forms are specified as below.

The pairwise interaction V_{p} is modeled by the Lennard-Jones potential,

$$V_{\mathrm{p}}(Q) = \begin{cases} 4\varepsilon\left[\left(\frac{\sigma}{Q}\right)^{12} - \left(\frac{\sigma}{Q}\right)^{6}\right] - 4\varepsilon\left[\left(\frac{\sigma}{Q_c}\right)^{12} - \left(\frac{\sigma}{Q_c}\right)^{6}\right], & Q < Q_c, \\ 0, \quad Q \geq Q_c, \end{cases} \quad \text{(A2)}$$

where $\varepsilon = 0.005$, $\sigma = 1.8$, and $Q_c = 10.0$.

The bond potential V_{b} is modeled by the finite extensible nonlinear elastic bond (FENE) potential,

$$V_{\mathrm{b}}(l) = -\frac{k_s}{2} l_0^2 \log\left[1 - \frac{l^2}{l_0^2}\right], \quad \text{(A3)}$$

where there are three different bond types. Within each sub-unit, the atoms 1-2 and 3-4 are connected by the type-1 bond. The 2-3 atoms are connected by the type-2 bond. Finally, the sub-unit groups are connected by the type-3 bond. The detailed parameter set is given in Table I.

The angle potential V_{a} is modeled by the harmonic angle potential,

$$V_{\mathrm{a}}(\theta) = \frac{k_a}{2}(\theta - \theta_0)^2, \quad \text{(A4)}$$

where there are two different types. Within each sub-unit group, the bond angles formed by 1-2-3 and 2-3-4 are imposed by type-1 potential. The bond angles formed by atoms of different sub-unit groups (e.g., 3-4-5 and 4-5-6) are imposed by the type-2 potential. The detailed parameter set is given in Table II.

TABLE II. Parameters of the harmonic angle interaction.

Type	k_a	θ_0
1	1.2	114.0
2	1.5	119.7

TABLE III. Parameters of the multi-harmonic dihedral interaction.

Type	A_1	A_2	A_3	A_4	A_5	A_6
1	0.0673	1.8479	0.0079	−2.2410	−0.0058	0.0051
2	0.1602	−3.9993	0.2483	6.2837	0.0165	−0.0146

The dihedral potential V_d is modeled by the multi-harmonic dihedral potential,

$$V_\mathrm{d}(\phi) = \sum_{i=1}^{6} A_n \cos^{(n-1)}(\phi), \tag{A5}$$

where there are two different types. Type-1 dihedral potential is imposed to dihedral angles formed by 2-3-4-5, 4-5-6-7, ⋯. Type-2 dihedral potential is imposed to dihedral angles formed by 3-4-5-6, 7-8-9-10, ⋯. The detailed parameter set is given in Table III.

APPENDIX B: CONSTRUCTION OF THE FOUR-DIMENSIONAL FREE ENERGY FUNCTION

Accurate construction of the multi-dimensional free energy is a well-known non-trivial problem. To construct the free energy function $U(\mathbf{q})$ for the four-dimensional resolved variables $\mathbf{q}$ defined by (18), we conduct the restraint molecular dynamics simulation to sample the average force. Specifically, for each target configuration $\mathbf{q}^*$, we impose a biased quadratic potential $U_\mathrm{bias}(\mathbf{q}, \mathbf{q}^*)$ by

$$U_\mathrm{bias}(\mathbf{q}, \mathbf{q}^*) = \frac{1}{2}\sum_{i=1}^{4} k_i \left(q_i - q_i^*\right)^2, \tag{B1}$$

where $k_1, \ldots, k_4$ represents the magnitude of the bias potential. We choose the values such that the fluctuations are about 5% of target values. For the polymer molecule considered in the present study, the effective restraint force applied to the full atom $\{\mathbf{Q}_j\}_{j=1}^{N}$ is given by

$$\mathbf{F}_\mathrm{bias}(\mathbf{q}, \mathbf{q}^*) = -\sum_{i=1}^{4} k_i \left(q_i - q_i^*\right) \nabla_{\mathbf{Q}_j} q_i, \tag{B2}$$

where the gradient terms are given by

$$\begin{aligned}
\nabla_{\mathbf{Q}_j} q_1 &= \frac{\mathbf{Q}_1 - \mathbf{Q}_N}{q_1}\delta_{j,1} + \frac{\mathbf{Q}_N - \mathbf{Q}_1}{q_1}\delta_{j,N},\\
\nabla_{\mathbf{Q}_j} q_2 &= \frac{2\left(\mathbf{Q}_j - \mathbf{Q}_c\right)}{N q_2},\\
\nabla_{\mathbf{Q}_j} q_3 &= \frac{\mathbf{Q}_1 - \mathbf{Q}_{\lfloor \frac{N}{2} \rfloor}}{q_3}\delta_{j,1} + \frac{\mathbf{Q}_{\lfloor \frac{N}{2} \rfloor} - \mathbf{Q}_1}{q_3}\delta_{j,\lfloor \frac{N}{2} \rfloor},\\
\nabla_{\mathbf{Q}_j} q_4 &= \frac{\mathbf{Q}_N - \mathbf{Q}_{\lceil \frac{N}{2} \rceil}}{q_4}\delta_{j,N} + \frac{\mathbf{Q}_{\lceil \frac{N}{2} \rceil} - \mathbf{Q}_N}{q_4}\delta_{j,\lceil \frac{N}{2} \rceil},
\end{aligned} \tag{B3}$$

where $\delta_{i,j}$ represents the Kronecker delta function.

The free energy $U(\mathbf{q})$ is approximated by a four-layer fully connected neural network $\tilde{U}(\mathbf{q})$. Each hidden layer has 160 neurons; the hyperbolic tangent function is used as the activation function. $\tilde{U}(\mathbf{q})$ is trained by minimizing the empirical loss,

$$L = \sum_{k=1}^{N_s} \left\| -\nabla_{\mathbf{q}^{(k)}} \tilde{U}(\mathbf{q}) - \mathbf{F}_\mathrm{bias}(\mathbf{q}, \mathbf{q}^{(k)}) \right\|^2, \tag{B4}$$

where $\mathbf{q}^{(k)}$ represents a sampled configuration. In this work, we construct $\tilde{U}(\mathbf{q})$ using $N_s = 400\,000$ sample points collected from a simulation with a production stage of 1×10^7 steps. For each configuration, the number of steps is between 1×10^6 and 6×10^6 such that the empirical sampling error is less than 5% of the mean value.

To verify the accuracy of $\tilde{U}(\mathbf{q})$, we numerically evaluate the integration

$$\begin{aligned}
k_B T \mathbf{I} &\equiv \int \mathbf{q} \otimes \nabla U(\mathbf{q}) \mathrm{e}^{-U(\mathbf{q})/k_B T} \mathrm{d}\mathbf{q} \Big/ \int \mathrm{e}^{-U(\mathbf{q})/k_B T} \mathrm{d}\mathbf{q}\\
&\approx \frac{1}{N_s}\sum_{k=1}^{N_s} \mathbf{q}^{(k)} \otimes \nabla \tilde{U}(\mathbf{q}^{(k)}).
\end{aligned} \tag{B5}$$

Therefore, the difference between the numerical summation and $k_B T\mathbf{I}$ provides a metric. For this case, $k_B T = 1$. The average term yields

$$\frac{1}{N_s}\sum_{k=1}^{N_s} \mathbf{q}^{(k)} \otimes \nabla \tilde{U}(\mathbf{q}^{(k)}) = \begin{bmatrix} 1.0362 & -0.0011 & 0.0087 & 0.0062 \\ 0.0094 & 0.9814 & 0.0021 & 0.0018 \\ 0.0096 & 0.0068 & 0.9913 & -0.0020 \\ 0.0076 & 0.0098 & 0.0008 & 0.9913 \end{bmatrix}, \tag{B6}$$

which verifies that the constructed $\tilde{U}(\mathbf{q})$ is an accurate approximation of $U(\mathbf{q})$.

APPENDIX C: FLUCTUATION–DISSIPATION THEOREM OF THE EXTENDED DYNAMICS

For the extended dynamics in the form of Eqs. (5) and (6), we can show that the embedded memory kernel $\tilde{\boldsymbol{\theta}}(t)$ and fluctuation term $\tilde{\mathcal{R}}(t)$ satisfy the second-fluctuation–dissipation theorem. Without loss of generality, we set the covariance of the non-Markovian features to be $k_B T\mathbf{I}$ following the learning method presented in Sec. II C, i.e., $\boldsymbol{\Lambda} = \mathbf{I}$, $\tilde{\mathbf{J}} = \mathbf{J}$.

Proposition C.1. The embedded memory kernel of the extended dynamics [(5) and (6)] takes the form $\tilde{\boldsymbol{\theta}}(t) = -\left(\mathbf{J}_{11}\delta(t) + \mathbf{J}_{12}\mathrm{e}^{\mathbf{J}_{22}t}\mathbf{J}_{21}\right)$. *Furthermore, by choosing the initial condition of* $\boldsymbol{\zeta}$ *and the white noise term* $\boldsymbol{\xi}(t) = \boldsymbol{\Sigma}\dot{\mathbf{W}}_t$ *satisfying*

$$\begin{aligned}
\left\langle \boldsymbol{\zeta}(0)\boldsymbol{\zeta}(0)^T \right\rangle &= \beta^{-1}\mathbf{I},\\
\left\langle \boldsymbol{\xi}(t)\boldsymbol{\xi}(s)^T \right\rangle &= -\beta^{-1}(\mathbf{J} + \mathbf{J}^T)\delta(t-s),
\end{aligned} \tag{C1}$$

the embedded kernel $\tilde{\boldsymbol{\theta}}(t)$ *and* $\mathscr{R}(t)$ *satisfy the second fluctuation–dissipation theorem,* i.e.,

$$\left\langle \tilde{\mathscr{R}}(t)\tilde{\mathscr{R}}(t')^T \right\rangle = -\beta^{-1}\left(\tilde{\mathbf{J}}_{12}\mathrm{e}^{\tilde{\mathbf{J}}_{22}(t-t')}\tilde{\mathbf{J}}_{21} + (\tilde{\mathbf{J}}_{11} + \tilde{\mathbf{J}}_{11}^T)\delta(t-t')\right). \quad \text{(C2)}$$

Proof. With $\boldsymbol{\Lambda} = \mathbf{I}$ and $\tilde{\mathbf{J}} = \mathbf{J}$, we can take the integration of $\boldsymbol{\zeta}(t)$ in Eq. (5), yielding

$$\boldsymbol{\zeta}(t) = \int_0^t \mathrm{e}^{\mathbf{J}_{22}(t-s)}\mathbf{J}_{21}\mathbf{v}(s)\mathrm{d}s + \int_0^t \mathrm{e}^{\mathbf{J}_{22}(t-s)}\boldsymbol{\xi}_2(s)\mathrm{d}s + \mathrm{e}^{\mathbf{J}_{22}t}\boldsymbol{\zeta}(0). \quad \text{(C3)}$$

Plugging $\boldsymbol{\zeta}(t)$ into the dynamic equation of $\mathbf{v}$ gives

$$\mathbf{M}\dot{\mathbf{v}} = -\nabla U(\mathbf{q}) + \mathbf{J}_{11}\mathbf{v} + \int_0^t \mathbf{J}_{12}\mathrm{e}^{\mathbf{J}_{22}(t-s)}\mathbf{J}_{21}\mathbf{v}(s)\mathrm{d}t + \underbrace{\boldsymbol{\xi}_1(t)}_{\tilde{\mathscr{R}}_1(t)} + \underbrace{\int_0^t \mathbf{J}_{12}\mathrm{e}^{\mathbf{J}_{22}(t-s)}\boldsymbol{\xi}_2(s)\mathrm{d}s}_{\tilde{\mathscr{R}}_2(t)} + \underbrace{\mathbf{J}_{12}\mathrm{e}^{\mathbf{J}_{22}t}\boldsymbol{\zeta}(0)}_{\tilde{\mathscr{R}}_3(t)}. \quad \text{(C4)}$$

We check the covariance matrices of the noise terms, i.e.,

$$\begin{aligned}
\left\langle \tilde{\mathscr{R}}_1(t)\tilde{\mathscr{R}}_1(t')^T \right\rangle &= -\beta^{-1}(\mathbf{J}_{11} + \mathbf{J}_{11}^T)\delta(t-t'),\\
\left\langle \tilde{\mathscr{R}}_2(t)\tilde{\mathscr{R}}_2(t')^T \right\rangle &= \int_0^t\int_0^{t'} \mathbf{J}_{12}\mathrm{e}^{\mathbf{J}_{22}(t-s)}\left\langle \boldsymbol{\xi}_2(s)\boldsymbol{\xi}_2(s')^T \right\rangle \mathrm{e}^{\mathbf{J}_{22}^T(t'-s')}\mathbf{J}_{12}^T\mathrm{d}s\mathrm{d}s' = -\beta^{-1}\int_0^t\int_0^{t'} \mathbf{J}_{12}\mathrm{e}^{\mathbf{J}_{22}(t-s)}(\mathbf{J}_{22} + \mathbf{J}_{22}^T)\delta(s-s')\mathbf{J}_{21}^T\mathrm{e}^{\mathbf{J}_{22}^T(t'-s')}\mathbf{J}_{12}^T\mathrm{d}s\mathrm{d}s'\\
&= -\beta^{-1}\int_0^{t'} \mathbf{J}_{12}\mathrm{e}^{\mathbf{J}_{22}(t-s')}(\mathbf{J}_{22} + \mathbf{J}_{22}^T)e^{\mathbf{J}_{22}^T(t'-s')}\mathbf{J}_{12}^T\mathrm{d}s' = -\beta^{-1}\mathbf{J}_{12}\mathrm{e}^{\mathbf{J}_{22}t + \mathbf{J}_{22}^Tt'}\mathbf{J}_{12}^T + \beta^{-1}\mathbf{J}_{12}\mathrm{e}^{\mathbf{J}_{22}(t-t')}\mathbf{J}_{12}^T, \quad \forall t' \le t,\\
\left\langle \tilde{\mathscr{R}}_3(t)\tilde{\mathscr{R}}_3(t')^T \right\rangle &= \mathbf{J}_{12}\mathrm{e}^{\mathbf{J}_{22}t}\left\langle \boldsymbol{\zeta}(0)\boldsymbol{\zeta}(0)^T \right\rangle \mathrm{e}^{\mathbf{J}_{22}^Tt'}\mathbf{J}_{12}^T = \beta^{-1}\mathbf{J}_{12}\mathrm{e}^{\mathbf{J}_{22}t}\mathrm{e}^{\mathbf{J}_{22}^Tt'}\mathbf{J}_{12}^T.
\end{aligned} \quad \text{(C5)}$$

Moreover, for $t > t'$, all the cross terms vanish except $\left\langle \tilde{\mathscr{R}}_2(t)\tilde{\mathscr{R}}_1(t')^T \right\rangle$, i.e.,

$$\begin{aligned}
\left\langle \tilde{\mathscr{R}}_2(t)\tilde{\mathscr{R}}_1(t')^T \right\rangle &= \int_0^t \mathbf{J}_{12}\mathrm{e}^{\mathbf{J}_{22}(t-s)}\left\langle \boldsymbol{\xi}_2(s)\boldsymbol{\xi}_1(t') \right\rangle \mathrm{d}s\\
&= -\beta^{-1}\int_0^t \mathbf{J}_{12}\mathrm{e}^{\mathbf{J}_{22}(t-s)}(\mathbf{J}_{21} + \mathbf{J}_{12}^T)\delta(t'-s)\mathrm{d}s\\
&= -\beta^{-1}\mathbf{J}_{12}\mathrm{e}^{\mathbf{J}_{22}(t-t')}(\mathbf{J}_{21} + \mathbf{J}_{12}^T).
\end{aligned} \quad \text{(C6)}$$

Combining Eqs. (C5) and (C6), we have

$$\begin{aligned}
\left\langle \tilde{\mathscr{R}}(t)\tilde{\mathscr{R}}(t')^T \right\rangle &= \beta^{-1}\mathbf{J}_{12}\mathrm{e}^{\mathbf{J}_{22}(t-t')}\mathbf{J}_{12}^T - \beta^{-1}\mathbf{J}_{12}\mathrm{e}^{\mathbf{J}_{22}(t-t')}\\
&\quad \times (\mathbf{J}_{21} + \mathbf{J}_{12}^T) - \beta^{-1}(\mathbf{J}_{11} + \mathbf{J}_{11}^T)\delta(t-t')\\
&= -\beta^{-1}\left(\mathbf{J}_{12}\mathrm{e}^{\mathbf{J}_{22}(t-t')}\mathbf{J}_{21} + (\mathbf{J}_{11} + \mathbf{J}_{11}^T)\delta(t-t')\right).
\end{aligned} \quad \text{(C7)}$$

□

As a special case, by imposing the restraint specified by Eq. (16) such that $\mathbf{J}_{11} + \mathbf{J}_{11}^T = 0$ and $\mathbf{J}_{12} = -\mathbf{J}_{21}^T$, the memory kernel $\tilde{\boldsymbol{\theta}}(t)$ recovers $-\mathbf{J}_{12}\mathrm{e}^{\mathbf{J}_{22}t}\mathbf{J}_{12}^T$ without the Markovian part, and the second fluctuation–dissipation theorem recovers the standard form, i.e.,

$$\left\langle \tilde{\mathscr{R}}(t)\tilde{\mathscr{R}}(0)^T \right\rangle = \beta^{-1}\tilde{\boldsymbol{\theta}}(t). \quad \text{(C8)}$$

APPENDIX D: INVARIANT PROBABILITY DENSITY FUNCTION

Proposition D.1. By choosing the white noise following Eq. (C1), the reduced model [(5) and (6)] retains the invariant density function,

$$\rho_{\mathrm{eq}}(\mathbf{q},\mathbf{p},\boldsymbol{\zeta}) = \exp[-\beta W(\mathbf{q},\mathbf{p},\boldsymbol{\zeta})]/\int \exp[-\beta W(\mathbf{q},\mathbf{p},\boldsymbol{\zeta})]\mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p}\mathrm{d}\boldsymbol{\zeta}. \quad \text{(D1)}$$

Proof. By Eq. (C1), the covariance of the white noise of the full extended system is given by $\mathbf{G} + \mathbf{G}^T = \mathrm{diag}(0, \boldsymbol{\Sigma}\boldsymbol{\Sigma}^T)$. Accordingly, the Fokker–Plank equation follows:

$$\frac{\partial \rho(\mathbf{z},t)}{\partial t} = \nabla\cdot\left(-\mathbf{G}\nabla W(\mathbf{z})\rho(\mathbf{z},t) - \frac{1}{2}\beta^{-1}(\mathbf{G} + \mathbf{G}^T)\nabla\rho(\mathbf{z},t)\right), \quad \text{(D2)}$$

where $\rho(\mathbf{z},t)$ represents the probability density function of the extended variables $\mathbf{z} = [\mathbf{q};\mathbf{p};\boldsymbol{\zeta}]$. For $\rho_{\mathrm{eq}}(\mathbf{q},\mathbf{p},\boldsymbol{\zeta}) \propto \exp[-\beta W(\mathbf{q},\mathbf{p},\boldsymbol{\zeta})]$, the RHS follows,

$$\begin{aligned}
&\nabla\cdot\left(\beta^{-1}\mathbf{G}\nabla\rho_{\mathrm{eq}}(\mathbf{z},t) - \frac{1}{2}\beta^{-1}(\mathbf{G} + \mathbf{G}^T)\nabla\rho_{\mathrm{eq}}(\mathbf{z},t)\right)\\
&\quad = \beta^{-1}\nabla\cdot\left(\mathbf{G}^A\nabla\rho_{\mathrm{eq}}(\mathbf{z},t)\right)\\
&\quad \equiv 0,
\end{aligned} \quad \text{(D3)}$$

where the last identity holds because $\mathbf{G}^A$ is anti-symmetric. □

APPENDIX E: SUMMARY OF THE TRAINING PROCESS

We summarize the training process of the present method in Algorithm 1.

APPENDIX F: MEMORY KERNEL OF THE POLYMER MOLECULE SYSTEMS

Figure 7 shows the embedded matrix-valued kernels $\boldsymbol{\theta}(t)$ of the full MD and the 4D reduced models of the polymer molecule

ALGORITHM 1. Training process of the reduced model with non-Markovian features.

Input: learning rate α; Training set $(\mathbf{x}(\mathbf{w}), \mathbf{y}(\mathbf{w})) = \left(\begin{bmatrix} \mathbf{C}_2(1*\delta t;\mathbf{w}) \\ \vdots \\ \mathbf{C}_2(N_t*\delta t;\mathbf{w}) \end{bmatrix}, \begin{bmatrix} \mathbf{C}_1'(1*\delta t;\mathbf{w}) + \mathbf{C}_0(1*\delta t;\mathbf{w}) \\ \vdots \\ \mathbf{C}_1'(N_t*\delta t;\mathbf{w}) + \mathbf{C}_0(N_t*\delta t;\mathbf{w}) \end{bmatrix} \right)$;

Output: $\mathbf{J}^S, \mathbf{J}^A, \mathbf{w}$

$\mathbf{J}^S, \mathbf{J}^A$ are two randomly generated $(n+1)m \times (n+1)m$ matrices.
$\mathbf{A}$ is a $(n+1)m \times (n+1)m$ upper triangular matrix, whose non-zero elements are 1.
$\mathbf{B} = \mathbf{A} - \mathbf{I}$.

function $L(\mathbf{x}(\mathbf{w}), \mathbf{y}(\mathbf{w}), \mathbf{J}^S, \mathbf{J}^A)$
    $\mathbf{L}_S \leftarrow (\mathbf{J}^S \odot \mathbf{A})^T$
    $\mathbf{H}_A \leftarrow (\mathbf{J}^A \odot \mathbf{B})^T$
    $\mathbf{J} \leftarrow -\mathbf{L}_S\mathbf{L}_S^T + \mathbf{H}_A - \mathbf{H}_A^T$
    return $\|\mathbf{J}\mathbf{x}(\mathbf{w}) - \mathbf{y}(\mathbf{w})\|^2 + L_\Lambda\|\mathbf{\Lambda}(\mathbf{w}) - \mathbf{I}\|^2$
end function

for *step* $= 1, 2, \ldots,$ *training_steps* **do**
    $\mathbf{g} \leftarrow \nabla_{\mathbf{w}} L(\mathbf{x}(\mathbf{w}), \mathbf{y}(\mathbf{w}), \mathbf{J}^S, \mathbf{J}^A)$
    $\mathbf{h}_s \leftarrow \nabla_{\mathbf{J}^S} L(\mathbf{x}(\mathbf{w}), \mathbf{y}(\mathbf{w}), \mathbf{J}^S, \mathbf{J}^A)$
    $\mathbf{h}_a \leftarrow \nabla_{\mathbf{J}^A} L(\mathbf{x}(\mathbf{w}), \mathbf{y}(\mathbf{w}), \mathbf{J}^S, \mathbf{J}^A)$
    $\mathbf{w} \leftarrow \mathbf{w} + \mathrm{Adam}(\mathbf{g}, \alpha)$
    $\mathbf{J}_s \leftarrow \mathbf{J}_s + \mathrm{Adam}(\mathbf{h}_s, \alpha)$
    $\mathbf{J}_a \leftarrow \mathbf{J}_a + \mathrm{Adam}(\mathbf{h}_a, \alpha)$
end for

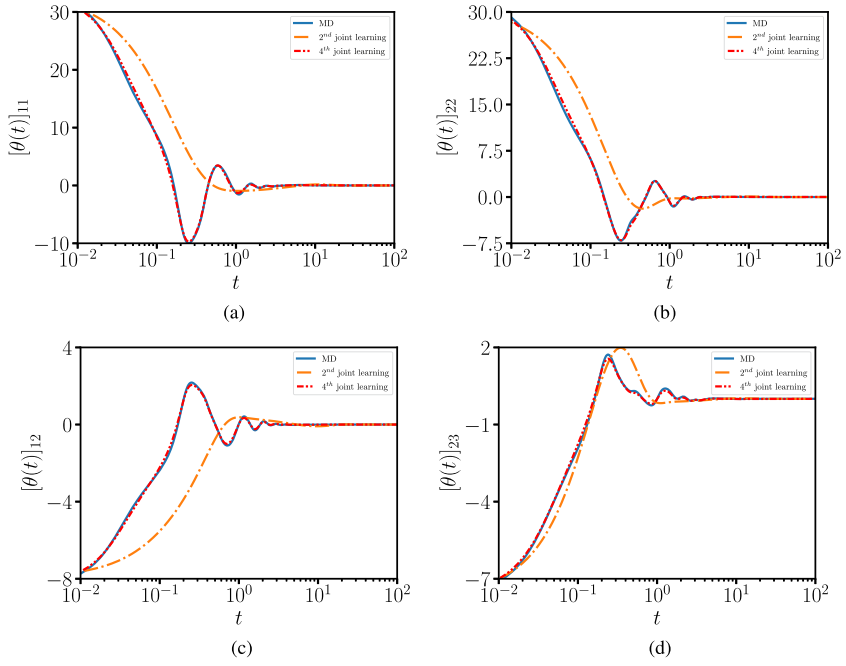

FIG. 7. (a)–(d) Components of the embedded matrix-valued kernel $\boldsymbol{\theta}(t)$ obtained from the full MD and four-dimensional reduced model of a polymer molecule system.

system. Similar to the kernel in the Laplace space $\Theta(\lambda)$ shown in Fig. 6, the good agreement between the full MD and reduced models verifies that the reduced model can accurately retain the non-Markovian dynamics of the resolved variables.

REFERENCES

[1] H. Mori, "Transport, collective motion, and Brownian motion," Prog. Theor. Phys. **33**, 423–455 (1965).
[2] R. Zwanzig, "Nonlinear generalized Langevin equations," J. Stat. Phys. **9**, 215–220 (1973).
[3] R. Zwanzig, *Nonequilibrium Statistical Mechanics* (Oxford University Press, 2001).
[4] A. J. Chorin, O. H. Hald, and R. Kupferman, "Optimal prediction with memory," Physica D **166**, 239–257 (2002).
[5] E. Darve, J. Solomon, and A. Kia, "Computing generalized Langevin equations and generalized Fokker-Planck equations," Proc. Natl. Acad. Sci. U. S. A. **106**, 10884–10889 (2009).
[6] O. F. Lange and H. Grubmüller, "Collective Langevin dynamics of conformational motions in proteins," J. Chem. Phys. **124**, 214903 (2006).
[7] M. Chen, X. Li, and C. Liu, "Computation of the memory functions in the generalized Langevin models for collective dynamics of macromolecules," J. Chem. Phys. **141**, 064112 (2014).
[8] P. Stinis, "Renormalized Mori-Zwanzig-reduced models for systems without scale separation," Proc. R. Soc. A **471**, 20140446 (2015).
[9] Y. Zhu and D. Venturi, "Faber approximation of the Mori-Zwanzig equation," J. Comput. Phys. **372**, 694–718 (2018).
[10] T. Hudson and X. H. Li, "Coarse-graining of overdamped Langevin dynamics via the Mori–Zwanzig formalism," Multiscale Model. Simul. **18**, 1113–1135 (2020).
[11] J. Price, B. Meuris, M. Shapiro, and P. Stinis, "Optimal renormalization of multiscale systems," Proc. Natl. Acad. Sci. U. S. A. **118**, e2102266118 (2021).
[12] C. Ma, J. Wang, and W. E, "Model reduction with memory and the machine learning of dynamical systems," Commun. Comput. Phys. **25**, 947–962 (2018).
[13] P. R. Vlachas, W. Byeon, Z. Y. Wan, T. P. Sapsis, and P. Koumoutsakos, "Data-driven forecasting of high-dimensional chaotic systems with long short-term memory networks," Proc. R. Soc. A **474**, 20170844 (2018).
[14] J. Harlim, S. W. Jiang, S. Liang, and H. Yang, "Machine learning for prediction with missing dynamics," J. Comput. Phys. **428**, 109922 (2020).
[15] Q. Wang, N. Ripamonti, and J. S. Hesthaven, "Recurrent neural network closure of parametric POD-Galerkin reduced-order models based on the Mori-Zwanzig formalism," J. Comput. Phys. **410**, 109402 (2020).
[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput. **9**, 1735–1780 (1997).
[17] A. Davtyan, J. F. Dama, G. A. Voth, and H. C. Andersen, "Dynamic force matching: A method for constructing dynamical coarse-grained models with realistic time dependence," J. Chem. Phys. **142**, 154104 (2015).
[18] H. Wall, *Analytic Theory of Continued Fractions* (D. Van Nostrand Company, 1948).
[19] H. Mori, "A continued-fraction representation of the time-correlation functions," Prog. Theor. Phys. **34**, 399–416 (1965).
[20] M. Corless and A. Frazho, *Linear Systems and Control: An Operator Perspective* (Chapman & Hall/CRC Pure and Applied Mathematics; Taylor & Francis, 2003).
[21] M. Ceriotti, G. Bussi, and M. Parrinello, "Langevin equation with colored noise for constant-temperature molecular dynamics simulations," Phys. Rev. Lett. **102**, 020601 (2009).
[22] A. D. Baczewski and S. D. Bond, "Numerical integration of the extended variable generalized Langevin equation with a positive Prony representable memory kernel," J. Chem. Phys. **139**, 044107 (2013).
[23] H. Lei, N. A. Baker, and X. Li, "Data-driven parameterization of the generalized Langevin equation," Proc. Natl. Acad. Sci. U. S. A. **113**, 14183–14188 (2016).
[24] G. Jung, M. Hanke, and F. Schmid, "Iterative reconstruction of memory kernels," J. Chem. Theory Comput. **13**, 2481–2488 (2017).
[25] H. S. Lee, S.-H. Ahn, and E. F. Darve, "The multi-dimensional generalized Langevin equation for conformational motion of proteins," J. Chem. Phys. **150**, 174113 (2019).
[26] A. Russo, M. A. Durán-Olivencia, I. G. Kevrekidis, and S. Kalliadasis, "Deep learning as closure for irreversible processes: A data-driven generalized Langevin equation," IEEE Trans. Neural Netw. Learn. Syst. 1–13 (2022), see https://ieeexplore.ieee.org/document/9947343.
[27] L. Ma, X. Li, and C. Liu, "Coarse-graining Langevin dynamics using reduced-order techniques," J. Comput. Phys. **380**, 170–190 (2019).
[28] F. Grogan, H. Lei, X. Li, and N. A. Baker, "Data-driven molecular modeling with the generalized Langevin equation," J. Comput. Phys. **418**, 109633–109641 (2020).
[29] A. J. Chorin and F. Lu, "Discrete approach to stochastic parametrization and dimension reduction in nonlinear dynamics," Proc. Natl. Acad. Sci. U. S. A. **112**, 9804–9809 (2015).
[30] K. K. Lin and F. Lu, "Data-driven model reduction, Wiener projections, and the Koopman-Mori-Zwanzig formalism," J. Comput. Phys. **424**, 109864 (2021).
[31] Y. Zhu, Y.-H. Tang, and C. Kim, "Learning stochastic dynamics with statistics-informed neural network," J. Comput. Phys. **474**, 111819 (2023).
[32] P. Xie, R. Car, and W. E, "Ab initio generalized Langevin equations," arXiv:2211.06558 (2022).
[33] H. Vroylandt, L. Goudenège, P. Monmarché, F. Pietrucci, and B. Rotenberg, "Likelihood-based non-Markovian models from molecular dynamics," Proc. Natl. Acad. Sci. U. S. A. **119**, e2117586119 (2022).
[34] J. O. Daldrop, B. G. Kowalik, and R. R. Netz, "External potential modifies friction of molecular solutes in water," Phys. Rev. X **7**, 041065 (2017).
[35] B. Kowalik, J. O. Daldrop, J. Kappler, J. C. Schulz, A. Schlaich, and R. R. Netz, "Memory-kernel extraction for different molecular solutes in solvents of varying viscosity in confinement," Phys. Rev. E **100**, 012126 (2019).
[36] V. Klippenstein, M. Tripathy, G. Jung, F. Schmid, and N. F. A. van der Vegt, "Introducing memory in coarse-grained molecular simulations," J. Phys. Chem. B **125**, 4931–4954 (2021).
[37] H. Lei and X. Li, "Petrov–Galerkin methods for the construction of non-Markovian dynamics preserving nonlocal statistics," J. Chem. Phys. **154**, 184108 (2021).
[38] R. Kubo, "The fluctuation-dissipation theorem," Rep. Prog. Phys. **29**(1), 255–284 (1966).
[39] P. Español, "Statistical mechanics of coarse-graining," in *Novel Methods in Soft Matter Simulations* (Springer, 2004), pp. 69–115.
[40] W. G. Noid, J.-W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, and H. C. Andersen, "The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models," J. Chem. Phys. **128**, 244114 (2008).
[41] H. Lei, L. Wu, and W. E, "Machine learning based non-Newtonian fluid model with molecular fidelity," Phys. Rev. E **102**, 043309 (2020).
[42] L. Fang, P. Ge, L. Zhang, W. E, and H. Lei, "DeePN^2: A deep learning-based non-Newtonian hydrodynamic model," J. Mach. Learn. **1**, 114–140 (2022).
[43] M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi, "Determination of reaction coordinates via locally scaled diffusion map," J. Chem. Phys. **134**, 124116 (2011).
[44] 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).
[45] W. Li and A. Ma, "Recent developments in methods for identifying reaction coordinates," Mol. Simul. **40**, 784–793 (2014).
[46] S. V. Krivov, "On reaction coordinate optimality," J. Chem. Theory Comput. **9**, 135–146 (2013).
[47] J. Lu and E. Vanden-Eijnden, "Exact dynamical coarse-graining without time-scale separation," J. Chem. Phys. **141**, 044109 (2014).

[48]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).
[49]R. R. Coifman, I. G. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler, "Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems," Multiscale Model. Simul. **7**, 842–864 (2008).
[50]E. Chiavazzo, R. Covino, R. R. Coifman, C. W. Gear, A. S. Georgiou, G. Hummer, and I. G. Kevrekidis, "Intrinsic map dynamics exploration for uncharted effective free-energy landscapes," Proc. Natl. Acad. Sci. U. S. A. **114**, E5494–E5503 (2017).
[51]M. Crosskey and M. Maggioni, "ATLAS: A geometric approach to learning high-dimensional stochastic systems near manifolds," Multiscale Model. Simul. **15**, 110–156 (2017).
[52]F. X. F. Ye, S. Yang, and M. Maggioni, "Nonlinear model reduction for slow-fast stochastic systems near manifolds," arXiv:2104.02120 (2021).
[53]L. Feng, T. Gao, M. Dai, and J. Duan, "Auto-SDE: Learning effective reduced dynamics from data-driven stochastic dynamical systems," arXiv:2205.04151 (2022).
[54]P. Zieliński and J. S. Hesthaven, "Discovery of slow variables in a class of multiscale stochastic systems via neural networks," J. Nonlinear Sci. **32**, 51 (2022).
[55]D. Giannakis, "Data-driven spectral decomposition and forecasting of ergodic dynamical systems," Appl. Comput. Harmonic Anal. **47**, 338–396 (2019).
[56]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).
[57]M. Dibak, M. J. del Razo, D. De Sancho, C. Schütte, and F. Noé, "MSM/RD: Coupling Markov state models of molecular kinetics with reaction-diffusion simulations," J. Chem. Phys. **148**, 214107 (2018).
[58]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).
[59]B. O. Koopman, "Hamiltonian systems and transformation in Hilbert space," Proc. Natl. Acad. Sci. U. S. A. **17**, 315–318 (1931).
[60]C. Ayaz, B. A. Dalton, and R. R. Netz, "Generalized Langevin equation with a non-linear potential of mean force and non-linear memory friction from a hybrid projection scheme," Phys. Rev. E **105**, 054138 (2022).
[61]S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, "THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method," J. Comput. Chem. **13**, 1011–1021 (1992).
[62]E. Darve and A. Pohorille, "Calculating free energies using average force," J. Chem. Phys. **115**, 9169–9183 (2001).
[63]A. Laio and M. Parrinello, "Escaping free-energy minima," Proc. Natl. Acad. Sci. U. S. A. **99**, 12562–12566 (2002).
[64]L. Rosso, P. Minàry, Z. Zhu, and M. E. Tuckerman, "On the use of the adiabatic molecular dynamics technique in the calculation of free energy profiles," J. Chem. Phys. **116**, 4389–4402 (2002).
[65]L. Maragliano and E. Vanden-Eijnden, "A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations," Chem. Phys. Lett. **426**, 168–175 (2006).
[66]L. Maragliano and E. Vanden-Eijnden, "Single-sweep methods for free energy calculations," J. Chem. Phys. **128**, 184110 (2008).
[67]D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications* (Elsevier, 2001), Vol. 1.
[68]D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in International Conference on Learning Representations (ICLR), 2015.
[69]P. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, 1979).