MACHINE LEARNING-BASED STOCHASTIC REDUCED MODELING OF GLE AND STATE-DEPENDENT-GLE

By

Zhiyuan She

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computational Mathematics, Science and Engineering—Doctor of Philosophy

2025

ABSTRACT

Predictive modeling of high-dimensional dynamical systems remains a central challenge in numerous scientific fields, including biology, materials science, and fluid mechanics. When clear scale separation is lacking, a reduced model must accurately capture the pronounced memory effects arising from unresolved variables, making non-Markovian modeling essential. In this thesis, we develop and analyze data-driven methods for constructing generalized Langevin equations (GLEs) and extended stochastic differential equations that faithfully encode non-Markovian behaviors.

Building on the Mori–Zwanzig formalism, we first propose an approach to learn a set of non-Markovian features—auxiliary variables that incorporate the history of the resolved coordinates—so that the effective dynamics inherits long-time correlations. By matching evolution of correlation functions in the extended variable space, our method systematically approximates the multi-dimensional GLE without requiring direct estimates of complicated memory kernels. We show that this approach yields stable, high-fidelity reduced models for molecular systems, enabling significantly lower-dimensional simulations that nonetheless reproduce key statistical and dynamical properties of the original system.

We then extend this framework to incorporate state-dependent memory kernels, facilitating enhanced sampling across diverse regions of phase space. We demonstrate that constructing heterogeneous memory kernels—reflecting the local variations in unresolved degrees of freedom—improves the model's accuracy and robustness, especially in systems exhibiting multiple metastable states. Through both numerical experiments and theoretical analysis, we highlight how these data-driven non-Markovian models outperform traditional Markovian or fixed-memory approaches.

To address complex, multi-modal distributions in high-dimensional data, we then modify the latent variable of a KRNet normalizing-flow architecture from a single Gaussian to a mixture-of-Gaussians (MoG). This richer latent representation not only improves the model's expressiveness and training stability but also facilitates discovering collective variables (CVs), as the multi-modal latent space reveals distinct modes corresponding to relevant metastable states or slow degrees of freedom. Through both numerical experiments and theoretical analysis, we show that integrating a

MoG prior into KRNet yields superior density estimation, enhanced sampling of metastable basins, and a more interpretable set of learned CVs.

Altogether, this thesis provides a comprehensive methodology for deriving scalable, memory-embedded reduced dynamics augmented by advanced latent representations. Such models open new possibilities for multi-scale simulations by merging fine-grained molecular fidelity with tractable coarse-grained representations, all while systematically leveraging the benefits of multi-modal latent spaces to identify key low-dimensional features. Our results underscore the practical advantages of incorporating non-Markovian features and a mixture-based flow model in capturing the full complexity of real-world molecular and dynamical systems.

ACKNOWLEDGEMENTS

I extend my heartfelt gratitude to my advisor, Professor Huan Lei, whose unwavering support, insightful guidance, and boundless patience have made this work possible. His passion for research and dedication to academic excellence have continually inspired me to broaden my intellectual horizons and to push myself beyond my comfort zone. I am truly grateful for the countless hours he devoted to reading my drafts, sharing thoughtful feedback, and challenging me to become a better researcher and problem-solver.

I also wish to acknowledge the encouragement and camaraderie of my groupmates, Pei Ge and Liyao Lyu. Their collaborative spirit, readiness to exchange ideas, and commitment to building a supportive research environment greatly enriched my work. Our thought-provoking discussions often revealed new perspectives and solutions that shaped this thesis for the better. I deeply appreciate their help both inside and outside the lab, from meticulous code reviews to the friendly conversations that made my time in graduate school more rewarding.

I am especially thankful to my wife, Xia Li, whose unwavering love and understanding have been an anchor throughout this demanding journey. Her belief in my abilities, willingness to celebrate even the smallest milestones, and compassionate reassurance during times of doubt provided me with the emotional resilience to persevere. I cannot overstate how much her steadfast support has meant to me—her presence has been the guiding light that helped me navigate this path.

To each of you—my advisor, colleagues, and family—your contributions, generosity, and devotion to this endeavor remain deeply appreciated. Your collective influence will continue to shape my academic and personal development for years to come.

TABLE OF CONTENTS

CHAPTER 1

OVERVIEW

The accurate modeling of high-dimensional dynamical systems remains a cornerstone challenge across a range of scientific disciplines, from soft matter and biophysics to fluid mechanics and climate science. Modern simulations of such systems often demand immense computational resources due to the intricate interactions and multi-scale nature of the underlying processes. Although fully resolved models, which include all microscopic variables and fine-grained details, can in principle capture the relevant physics, they frequently prove infeasible in practice because of prohibitive computational cost. As a result, significant efforts in coarse-grained modeling aim to reduce dimensionality and complexity while preserving the essential statistics and long-time behaviors of the original high-dimensional problem.

A central insight in coarse-graining is that purely Markovian models—those assuming instantaneous and memoryless evolution—often fail to reproduce the observed time correlations and transport properties in real systems. These discrepancies become particularly pronounced in situations lacking clear time-scale separations, wherein so-called "fast" or unresolved degrees of freedom exert non-negligible influences over extended time horizons. Consequently, one must explicitly retain memory terms to produce physically accurate reduced dynamics. In theoretical treatments, the Mori–Zwanzig (MZ) formalism provides a foundation for describing the projected dynamics of a lower-dimensional set of variables, augmented with a time-dependent memory kernel and stochastic fluctuation terms. However, direct numerical implementation of these ideas is seldom straightforward because the memory kernel typically lacks a closed-form expression and may require large volumes of data to estimate reliably.

In response, a growing body of literature has turned to data-driven approaches that learn reduced equations or effective models directly from simulation trajectories or experimental data. Such methods circumvent the direct computation of memory kernels by leveraging machine learning techniques to discover the underlying structure of the system. This dissertation focuses on integrating and extending three distinct strategies that address complementary aspects of non-Markovian

modeling for high-dimensional dynamical systems:

1. **Non-Markovian Feature Learning**

   Our first strategy proposes a learning framework that sidesteps the need for explicit memory-kernel estimation by introducing a set of auxiliary variables. These additional variables encapsulate the "history" of the coarse-grained coordinates, effectively transforming what would otherwise be a non-Markovian model into a higher-dimensional extended Markovian system. By matching correlation functions between the full model and the extended reduced model, we ensure that critical temporal dependencies are accurately preserved. This correlation-matching step is instrumental: it encodes the slow relaxations and recurrent configurations that are crucial for capturing long-time dynamics. Numerical experiments on molecular systems demonstrate that non-Markovian feature learning can yield reduced-order simulations with excellent fidelity to the reference trajectories, all while maintaining moderate computational cost.

2. **State-Dependent Memory Kernels**

   While the first approach offers a single global mechanism to embed memory effects, many real systems exhibit state-dependent memory. For instance, macromolecular or fluid systems may have multiple metastable basins, each with distinct relaxation times or energetic barriers. In such cases, it is insufficient to assume a uniform memory kernel throughout the entire state space. Our second strategy addresses this limitation by introducing heterogeneous memory kernels that adapt to the local environment of the resolved variables. Rather than fitting a single kernel function, we allow the memory to vary based on the instantaneous configuration or thermodynamic state. This added flexibility is particularly advantageous for systems with complex free-energy landscapes, as it enables more accurate modeling of basin-to-basin transitions, barrier crossing, and other processes sensitive to local unresolved dynamics. By learning these heterogeneous kernels from data, we capture nuanced variations in the

memory structure, significantly improving sampling efficiency and predictive performance in multi-basin or multi-phase scenarios.

3. **KRNet with a Mixture-of-Gaussians Latent Representation**

   Although robust memory modeling is critical for accurate dynamics, effectively capturing distributional complexity in high-dimensional systems poses an additional challenge. Many normalizing-flow methods, which learn invertible transformations from simple latent spaces to complex data distributions, rely on a single Gaussian prior for the latent variables. Such a unimodal assumption can limit the expressivity of the model, particularly when the target distribution is multi-modal or exhibits heavy tails. Here, we introduce an advanced KRNet architecture that employs a mixture-of-Gaussians (MoG) as the latent prior. By allowing the latent space to be multi-modal, KRNet gains greater flexibility in approximating intricate molecular or continuum distributions. Moreover, analyzing individual mixture components provides insights into physically meaningful collective variables (CVs), such as reaction coordinates or slow degrees of freedom that govern the system's long-time behavior. This MoG-based design not only improves density-estimation accuracy but also enhances interpretability, offering an additional avenue for understanding how different metastable states or configurations map to the underlying latent structure.

Taken as a whole, these three complementary approaches form a cohesive framework for memory-aware, distribution-aware coarse-graining. In the first two, we focus on correctly capturing time correlations and historical dependence—the hallmark of non-Markovian dynamics. In the third, we emphasize handling complex data distributions and uncovering low-dimensional representations. Implemented together, they enable practitioners to develop robust reduced-order models that do not sacrifice crucial multi-scale or multi-modal characteristics of the original system.

Beyond theoretical importance, these techniques offer practical advantages: multi-scale simulations become more feasible as the reduced models require fewer degrees of freedom to achieve a similar predictive capability, potentially reducing wall-clock times while maintaining essential

physical fidelity. Moreover, our approaches naturally expose slow modes and transitions that might otherwise remain obscured in fully detailed simulation data, thereby aiding in physical interpretation. Researchers can identify meaningful collective variables or design specialized sampling protocols targeting critical regions of phase space (e.g., near transition states or interfaces).

Ultimately, the methods presented here reflect a broader trend in computational science: as machine learning and high-performance computing continue to advance, new opportunities emerge for data-driven modeling of complex phenomena. These advances permit us to go beyond naively discarding unresolved scales, instead systematically incorporating their effects through memory terms, latent-variable modeling, or both. The chapters that follow detail each of these methods, their theoretical underpinnings, and the empirical studies that demonstrate their utility. Collectively, they underline the feasibility of incorporating memory effects and multi-modal representations in next-generation coarse-grained simulation frameworks, bridging the gap between brute-force full-resolution models and simpler—but often inaccurate—Markovian approximations.

In summary, the remainder of this dissertation proceeds as follows:

- We begin by examining **non-Markovian feature learning** and explain how auxiliary variables, grounded in correlation-function matching, facilitate extended Markovian representations of intrinsically non-Markovian processes.

- Next, we tackle **state-dependent memory kernels** as a direct way to incorporate local environmental or configurational effects, significantly improving the realism of reduced models in multi-basin systems.

- Finally, we present **KRNet** with a **mixture-of-Gaussians (MoG)** latent space, illustrating how it enhances distribution modeling and offers a pathway to derive physically interpretable collective variables. We highlight its synergy with memory-based approaches, demonstrating an integrated methodology for advanced coarse-grained simulations.

Through these contributions, this dissertation seeks to illustrate the power and flexibility of data-driven, memory-embedded modeling. By capturing temporal dependencies and multi-modal

structures simultaneously, researchers can generate high-fidelity reduced-order simulations capable of exploring complex energy landscapes, long-time dynamical evolution, and rarely visited metastable states. The approaches and results stand to benefit a wide array of fields, ranging from molecular biophysics and materials science to geophysics and fluid mechanics, all of which confront the challenges posed by limited computational budgets and intrinsically non-Markovian dynamics.

CHAPTER 2

DATA-DRIVEN CONSTRUCTION OF STOCHASTIC REDUCED DYNAMICS ENCODED WITH NON-MARKOVIAN FEATURES

2.1 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 spatio-temporal 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 formalism Mori (1965b); Zwanzig (1973) or canonical models such as the generalized Langevin equation (GLE) Zwanzig (2001). Examples include the t-model approximation Chorin et al. (2002), the Galerkin discretization Darve et al. (2009a), regularized integral equation discretization Lange and Grubmüller (2006), the hierarchical construction Chen et al. (2014); Stinis (2015); Zhu and Venturi (2018); Hudson and Li (2020); Price et al. (2021), and so on. Recent studies Ma et al. (2018); Vlachas et al. (2018); Harlim et al. (2020); Wang et al. (2020a) based on the recurrent neural networks Hochreiter and Schmidhuber (1997) 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 certain ansatz, e.g., the fictitious particle Davtyan et al. (2015a), continued fraction Wall (1948); Mori (1965a), rational function Corless and Frazho (2003), such that the memory and the noise terms can be embedded in an extended Markovian dynamics Mori (1965a); Ceriotti et al. (2009); Baczewski and Bond (2013); Davtyan et al. (2015a); Lei et al. (2016a);

Jung et al. (2017a); Lee et al. (2019a); Russo et al. (2019); Ma et al. (2019); Grogan et al. (2020). In addition, non-Markovian models are represented by discrete dynamics with exogenous inputs in form of NARMAX (nonlinear autoregression moving average with exogenous input) Chorin and Lu (2015); Lin and Lu (2021) and SINN (statistics information neural network) Zhu et al. (2022) and parameterized for each specific time step. Recent work by Vroylandt et al. Vroylandt et al. (2022) presents an expectation-maximization method to parameterize the reduced model from the full model trajectories. Refs. Daldrop et al. (2017); Kowalik et al. (2019) present 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 Klippenstein et al. (2021) 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. Lee et al. (2019a). 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 Lei and Li (2021) to construct the non-Markovian reduced dynamics by projecting the full dynamics into a subspace spanned by a set of projection bases in 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 Lei and Li (2021), 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 Kubo (1966), and retains the consistent invariant distribution Español (2004); Noid et al. (2008).

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 Lei et al. (2016a); Lei and Li (2021) 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 towards constructing reliable continuum-level transport model equations Lei et al. (2020); Fang

et al. (2022).

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, 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. Rohrdanz et al. (2011); Pérez-Hernández et al. (2013); Li and Ma (2014); Krivov (2013); Lu and Vanden-Eijnden (2014); Bittracher et al. (2018) on learning resolved variables that retain the Markovianity, Refs. Coifman et al. (2008); Chiavazzo et al. (2017); Crosskey and Maggioni (2017); Ye et al. (2021); Feng et al. (2022); Zieliński and Hesthaven (2022) on learning the slow dynamics on a non-linear manifold, and Refs. Giannakis (2019); Klus et al. (2018); Dibak et al. (2018); Klus et al. (2020) on model reduction of the transfer operator.

2.2 Methods

2.2.1 Problem Setup

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

$$\dot{\boldsymbol{Z}} = \boldsymbol{S}\nabla H(\boldsymbol{Z}), \tag{2.1}$$

where $\boldsymbol{S} = \begin{pmatrix} 0 & \boldsymbol{I} \\ -\boldsymbol{I} & 0 \end{pmatrix}$ is the symplectic matrix, and $H(\boldsymbol{Z})$ is the Hamiltonian function and initial condition is given by $\boldsymbol{Z}(0) = \boldsymbol{Z}_0$. For high-dimensional systems with $N \gg 1$, the numerical simulation of Eq. (2.1) can be computational expensive. It is often desirable to construct a reduced model with respect to a set of low-dimensional resolved variables $\boldsymbol{z}(t) := \boldsymbol{\phi}(\boldsymbol{Z}(t))$ where $\boldsymbol{\phi} : \mathbb{R}^{6N} \to \mathbb{R}^m$ represents the mapping from the full to the coarse-grained state space with $m \ll N$. With the explicit form of $H(\boldsymbol{Z})$ and $\boldsymbol{\phi}(\boldsymbol{Z})$, the evolution of $\boldsymbol{z}(t)$ can be mapped from the initial values via the Koopman operator Koopman (1931), i.e., $\boldsymbol{z}(t) = \mathrm{e}^{t\mathcal{L}}\boldsymbol{z}(0)$, where the Liouville

operator $\mathcal{L}\phi(\boldsymbol{Z}) = -((\nabla H(\boldsymbol{Z}_0))^T \boldsymbol{S}\nabla_{\boldsymbol{Z}_0})\phi(\boldsymbol{Z})$ depends on the full-dimensional phase vector $\boldsymbol{Z}$. Using the Duhamel–Dyson formula, the evolution of $\boldsymbol{z}(t)$ can be further formulated in terms of $\boldsymbol{z}$ based on the Mori-Zwanzig (MZ) projection formalism Mori (1965b); Zwanzig (1973). However, the numerical evaluation of the derived model relies on solving the full-dimensional orthogonal dynamics Chorin et al. (2002), which can be still computational expensive.

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

$$\begin{aligned} \dot{\boldsymbol{q}} &= \boldsymbol{M}^{-1}\boldsymbol{p} \\ \dot{\boldsymbol{p}} &= -\nabla U(\boldsymbol{q}) - \int_0^t \boldsymbol{\theta}(t-\tau)\dot{\boldsymbol{q}}(\tau)\mathrm{d}\tau + \boldsymbol{\mathcal{R}}(t), \end{aligned} \tag{2.2}$$

where $\boldsymbol{q} \in \mathbb{R}^m$ is the so-called collective variables, $\boldsymbol{M}$ is the mass matrix, $U(\boldsymbol{q})$ 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 $\mathcal{R}(t)$ is a stationary colored noise related to $\boldsymbol{\theta}(t)$ through the second fluctuation-dissipation condition Kubo (1966), i.e., $\left\langle \boldsymbol{\mathcal{R}}(t)\boldsymbol{\mathcal{R}}(0)^T \right\rangle = k_B T\boldsymbol{\theta}(t)$. It is worth mentioning that Eq. (2.2) is not exact based on the MZ formalism. In particular, the memory function generally depends on the resolved variables $\boldsymbol{z}$ and the noise term could be non-Gaussian; we refer to Ref. Ayaz et al. (2022) for further discussion. Nevertheless, even for the simplified GLE form (2.2), the accurate construction of the reduced model could remain highly-nontrivial. Specifically, the numerical simulation requires the explicit knowledge of both the free energy $U(\boldsymbol{q})$ and the memory function $\boldsymbol{\theta}(t)$. Several methods based on importance sampling Kumar et al. (1992); Darve and Pohorille (2001); Laio and Parrinello (2002a) and temperature elevation Rosso et al. (2002); Maragliano and Vanden-Eijnden (2006, 2008) 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 $\boldsymbol{q}$, 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. (2.2), 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.,

$$\mathrm{d}\tilde{\boldsymbol{z}} = \boldsymbol{g}\left(\tilde{\boldsymbol{z}}\right)\mathrm{d}t + \boldsymbol{\Sigma}\mathrm{d}\boldsymbol{W}_t, \tag{2.3}$$

where $\tilde{\boldsymbol{z}} := [\boldsymbol{q}; \boldsymbol{p}; \boldsymbol{\zeta}_1; \cdots; \boldsymbol{\zeta}_n]$ represents the extended variables and $\boldsymbol{W}_t$ represents the standard Wiener process. In principle, any such extended system would generally lead to a non-Markovian dynamics for the resolved variables $\boldsymbol{z} = [\boldsymbol{q}; \boldsymbol{p}]$. However, the essential challenge is to determine $\{\boldsymbol{\zeta}_i\}_{i=1}^n$ so that the non-local statistical properties of $\boldsymbol{z}$ can be preserved with sufficient accuracy. Also, the form of $\boldsymbol{g}(\cdot)$ 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. Lei and Li (2021), the features $\{\boldsymbol{\zeta}_i\}_{i=1}^n$ are not the empirically-chosen projection bases; instead, they are learned along with model equation (2.3) for the best approximation of the non-local statistics. This essential difference enables the present method to generate more accurate reduced model and be easy to implement for multi-dimensional resolved variables without empirical treatment for numerical stability.

2.2.2 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 (2.3) for the solute molecule, a main question is how to construct the auxiliary variables $\boldsymbol{\zeta} := [\boldsymbol{\zeta}_1; \boldsymbol{\zeta}_2; \cdots; \boldsymbol{\zeta}_n]$. Intuitively, $\boldsymbol{\zeta}_i$ should depend on the full-dimensional vector $\boldsymbol{Z}$ such that their evolution encodes certain unresolved dynamics orthogonal to the subspace spanned by $\boldsymbol{z}(t)$. A straightforward approach is to represent $\boldsymbol{\zeta}(t)$ as a function of $\boldsymbol{Z}(t)$, i.e., $\boldsymbol{\zeta} = \boldsymbol{h}(\boldsymbol{Z})$. However, the direct construction of the general formulation $\boldsymbol{h}(\boldsymbol{Z})$ 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 $\boldsymbol{Z}$.

To circumvent the above challenges, the key ascribes to formulate $\boldsymbol{\zeta}(t)$ such that it properly encodes the unresolved dynamics of $\boldsymbol{Z}(t)$, and meanwhile, can be easily sampled. One important observation is that the history of $\boldsymbol{p}(t)$ naturally encodes the unresolved dynamics orthogonal to $\boldsymbol{z}(t)$ and the values can be conveniently obtained. To see this, we note that the dynamics follows $\dot{\boldsymbol{p}} = \mathcal{L}\boldsymbol{p}$

where the Liouville operator $\mathcal{L}\phi(\boldsymbol{Z}) = -((\nabla H(\boldsymbol{Z}_0))^T \boldsymbol{S}\nabla_{\boldsymbol{Z}_0})\phi(\boldsymbol{Z})$ depends on the full-dimensional vector $\boldsymbol{Z}$. Therefore, it is possible to construct $\boldsymbol{\zeta}(t)$ by some encoder functions in terms of the time history of $\boldsymbol{p}(t)$, i.e., $\boldsymbol{p}(\tau)$ with $\tau \le t$, such that they retain certain components orthogonal to $\boldsymbol{z}(t)$. This is somewhat similar to the study Lei et al. (2020) 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

$$\begin{aligned} \boldsymbol{\zeta}_i(t) &= \int_0^{+\infty} \boldsymbol{\omega}_i(\tau)\boldsymbol{v}(t-\tau)\mathrm{d}\tau \\ &\approx \sum_{j=1}^{N_w} \boldsymbol{w}_i(j\delta t)\boldsymbol{v}(t-j\delta t) \end{aligned} \tag{2.4}$$

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

$\boldsymbol{\zeta}_i(t)$ can be loosely viewed as a generalized convolution over the history of $\boldsymbol{v}(t)$ whose evolution encodes the orthogonal dynamics of $\boldsymbol{z}(t)$. Therefore, it is possible to learn a set of $\boldsymbol{\zeta}_i(t)$ such that the joint dynamics in terms of both $\boldsymbol{z}(t)$ and $\boldsymbol{\zeta}_i(t)$ can be well approximated by the extended Markovian model (2.3). Moreover, the linear form in terms of $\boldsymbol{v}(t)$ ensures that the invariant density function of $\boldsymbol{\zeta}_i(t)$ retains the Gaussian distribution consistent with $\boldsymbol{v}(t)$. We can further impose a constraint such that $\boldsymbol{v}(t)$ and $\boldsymbol{\zeta}_i(t)$ are uncorrelated. This leads to an additional quadratic term in the energy function of the extended system, i.e., $W(\boldsymbol{q},\boldsymbol{p},\boldsymbol{\zeta}) = U(\boldsymbol{q}) + \frac{1}{2}\boldsymbol{p}^T\boldsymbol{M}^{-1}\boldsymbol{p} + \frac{1}{2}\boldsymbol{\zeta}^T\hat{\boldsymbol{\Lambda}}^{-1}\boldsymbol{\zeta}$, where $\hat{\boldsymbol{\Lambda}}$ represents the covariance matrix of the $\boldsymbol{\zeta}_1, \cdots, \boldsymbol{\zeta}_n$. The reduced dynamics can be written as

$$\mathrm{d}\begin{pmatrix} \boldsymbol{q} \\ \boldsymbol{p} \\ \boldsymbol{\zeta} \end{pmatrix} = \boldsymbol{G}\nabla W(\boldsymbol{q},\boldsymbol{p},\boldsymbol{\zeta})\mathrm{d}t + \boldsymbol{\Sigma}\mathrm{d}\boldsymbol{W}_t, \tag{2.5}$$

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

$$\boldsymbol{G} = \begin{pmatrix} 0 & \boldsymbol{I} & 0 & \cdots & 0 \\ -\boldsymbol{I} & & & & \\ 0 & & & \boldsymbol{J} & \\ \vdots & & & & \\ 0 & & & & \end{pmatrix} \begin{pmatrix} \boldsymbol{I} & 0 & 0 & \cdots & 0 \\ 0 & \multicolumn{3}{c}{\multirow{3}{*}{$\begin{bmatrix} \boldsymbol{I} & & \\ & & \\ & & \hat{\boldsymbol{\Lambda}} \end{bmatrix}$}} & \\ 0 & & & & \\ \vdots & & & & \\ 0 & & & & \end{pmatrix}. \tag{2.6}$$

The matrix $\boldsymbol{J} \in \mathbb{R}^{nm\times nm}$ 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 next subsection. Given the reduced model in form of Eqs. (2.5) and (2.6), the noise covariance matrix can be determined by

$$\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T = -\beta^{-1}(\boldsymbol{J}\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^T\boldsymbol{J}^T), \tag{2.7}$$

where $\beta = 1/k_BT$ and $\boldsymbol{\Lambda} = \mathrm{d}iag(\boldsymbol{I}, \hat{\boldsymbol{\Lambda}})$. The form of $\boldsymbol{\Lambda}$ implies that $\boldsymbol{v}$ and $\boldsymbol{\xi}_i$ are uncorrelated and is consistent with the energy function of the extended system $W(\boldsymbol{q}, \boldsymbol{p}, \boldsymbol{\zeta})$. It also alleviates the non-negative constraint of $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T$ as discussed in Sec. 2.2.3. Furthermore, we can show that model (2.5) strictly satisfies the second-fluctuation dissipation theorem. Specifically, the embedded kernel in Eq. (2.5) takes the form

$$\tilde{\boldsymbol{\theta}}(t) = -\left(\tilde{\boldsymbol{J}}_{11}\delta(t) + \tilde{\boldsymbol{J}}_{12}\mathrm{e}^{\tilde{\boldsymbol{J}}_{22}t}\tilde{\boldsymbol{J}}_{21}\right), \tag{2.8}$$

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

$$\left\langle \tilde{\boldsymbol{\mathcal{R}}}(t)\tilde{\boldsymbol{\mathcal{R}}}(t')^T \right\rangle = -\beta^{-1}\left(\tilde{\boldsymbol{J}}_{12}\mathrm{e}^{\tilde{\boldsymbol{J}}_{22}(t-t')}\tilde{\boldsymbol{J}}_{21} + (\tilde{\boldsymbol{J}}_{11} + \tilde{\boldsymbol{J}}_{11}^T)\delta(t-t')\right) \tag{2.9}$$

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

$$\rho_{\mathrm{eq}} \propto \exp\left[-\beta W(\boldsymbol{q}, \boldsymbol{p}, \boldsymbol{\zeta})\right]. \tag{2.10}$$

We refer to Appendix C and D for details.

We conclude this subsection with two remarks: (I) In principle, the mass matrix $\boldsymbol{M}$ further depends on $\boldsymbol{q}$. Ref. Lee et al. (2019a) reports that the varying mass matrix plays a secondary effect on the reduced dynamics of the molecular system therein; see also Ref. Ayaz et al. (2022) 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 form of Eq. (2.4) can be generalized to retain the state-dependence, e.g., $\boldsymbol{\zeta}_i(t) = \int_0^{+\infty} \boldsymbol{\omega}_i(\tau, \boldsymbol{q}(\tau))\boldsymbol{v}(t-\tau)\mathrm{d}\tau$, 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 (2.5) that approximates the standard GLE (2.2) 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 (2.5) by matching the evolution of the correlation matrices for both the resolved and the extended variables. The learning of reduced models with the heterogeneous memory term will be systematically investigated in the future study.

2.2.3 Joint learning of the reduced dynamics

Construction of the above reduced models relies on the joint learning of the non-Markovian features (2.4) in form of the encoder functions $\{\boldsymbol{\omega}_i(t)\}_{i=1}^n$ and the reduced dynamics (2.5)(2.6) determined by the free energy $U(\boldsymbol{q})$ and the matrix $\boldsymbol{J}$. In this study, we represent the multi-dimensional free energy $U(\boldsymbol{q})$ using a neural network and parameterize it based on the molecular dynamics Frenkel and Smit (2001); we refer to Appendix for details. Furthermore, the covariance of the noise term specified by Eq. (2.7) implies that $\boldsymbol{J}$ and $\boldsymbol{\Lambda}$ (i.e., the encoder functions $\boldsymbol{\omega}_i(t)$) need to satisfy the following constraint

$$\boldsymbol{J}\boldsymbol{\Lambda} + \boldsymbol{\Lambda}\boldsymbol{J}^T \preccurlyeq 0. \tag{2.11}$$

Directly imposing the condition (2.11) becomes cumbersome for the joint learning of $\boldsymbol{J}$ and $\boldsymbol{\omega}_i(t)$. Fortunately, this issue can be avoided by imposing an orthogonal constraint among the

non-Markovian features, i.e.,

$$\begin{aligned}\left[\hat{\boldsymbol{\Lambda}}\right]_{ij} &:= \beta \left\langle \boldsymbol{\zeta}_i, \boldsymbol{\zeta}_j \right\rangle \\ &= \beta \sum_{k,k'} \left\langle \boldsymbol{w}_i(t - k\delta t)\boldsymbol{v}(k\delta t), \boldsymbol{w}_j(t - k'\delta t)\boldsymbol{v}(k'\delta t) \right\rangle \\ &= \delta_{ij}\boldsymbol{I}, \quad 1 \le i, j \le n,\end{aligned} \tag{2.12}$$

where the inner product $\langle \boldsymbol{f}(\boldsymbol{Z}), \boldsymbol{g}(\boldsymbol{Z}) \rangle = \int \boldsymbol{f}(\boldsymbol{Z})\boldsymbol{g}(\boldsymbol{Z})^T \rho(\boldsymbol{Z}) \mathrm{d}\boldsymbol{Z}$ is defined with respect to the equilibrium density function of the full model $\rho(\boldsymbol{Z}) = e^{-\beta H(\boldsymbol{Z})} / \int e^{-\beta H(\boldsymbol{Z})} \mathrm{d}\boldsymbol{Z}$. In addition, we also impose the orthogonal constraints such that $\boldsymbol{\zeta}$ and $\boldsymbol{p}$ are uncorrelated. Therefore, condition (2.11) can be transformed into seeking $\boldsymbol{J}$ s.t. $\boldsymbol{J} + \boldsymbol{J}^T \preccurlyeq 0$, and we represent $\boldsymbol{J}$ by

$$\boldsymbol{J} = -\boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{J}^A, \tag{2.13}$$

where $\boldsymbol{L} \in \mathbb{R}^{(n+1)m \times (n+1)m}$ is the lower-triangle matrix with positive diagonal elements and $\boldsymbol{L}\boldsymbol{L}^T$ represents the Cholesky decomposition of a symmetric positive-definite matrix. $\boldsymbol{J}^A$ represents an anti-symmetric matrix. Unlike the studies Mori (1965a); Ceriotti et al. (2009) based on the direct kernel approximation, we note that $\boldsymbol{J}$ takes a more general form and is not restricted to be diagonal or tri-diagonal.

With the proper form of $\boldsymbol{J}$, we can cast the reduced dynamics into the evolution of the correlation matrices by multiply $\boldsymbol{v}(0)$ to both sides of Eq. (2.5), i.e.,

$$\frac{\mathrm{d}}{\mathrm{d}t} \underbrace{\begin{pmatrix} \langle \boldsymbol{M}\boldsymbol{v}, \boldsymbol{v}(0) \rangle \\ \langle \boldsymbol{\zeta}_1, \boldsymbol{v}(0) \rangle \\ \vdots \\ \langle \boldsymbol{\zeta}_n, \boldsymbol{v}(0) \rangle \end{pmatrix}}_{\boldsymbol{C}_1(t)} + \underbrace{\begin{pmatrix} \langle \nabla U(\boldsymbol{q}), \boldsymbol{v}(0) \rangle \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{\boldsymbol{C}_0(t)} = \boldsymbol{J} \underbrace{\begin{pmatrix} \langle \boldsymbol{v}, \boldsymbol{v}(0) \rangle \\ \langle \boldsymbol{\zeta}_1, \boldsymbol{v}(0) \rangle \\ \vdots \\ \langle \boldsymbol{\zeta}_n, \boldsymbol{v}(0) \rangle \end{pmatrix}}_{\boldsymbol{C}_2(t)}, \tag{2.14}$$

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

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

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

$$\begin{aligned} L_C &= \sum_{j=1}^{N_t} \left\| \boldsymbol{C}_1'(t_j) + \boldsymbol{C}_0(t_j) - \boldsymbol{J}\boldsymbol{C}_2(t_j) \right\|^2 \; L_\Lambda = \|\boldsymbol{\Lambda} - \boldsymbol{I}\|^2 , \\ L &= \lambda_C L_C + \lambda_\Lambda L_\Lambda, \end{aligned} \tag{2.15}$$

where $\boldsymbol{C}_1 = [\langle \boldsymbol{M}\boldsymbol{v}, \boldsymbol{v}(0)\rangle ; \langle \boldsymbol{\zeta}_1, \boldsymbol{v}(0)\rangle ; \cdots ; \langle \boldsymbol{\zeta}_n, \boldsymbol{v}(0)\rangle]$, $\boldsymbol{C}_0$ and $\boldsymbol{C}_2(t)$ are defined similarly in Eq. (2.14). λ_C and λ_Λ are the hyperparameters. 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 $\hat{\boldsymbol{\Lambda}} = \boldsymbol{I}$ such that $\boldsymbol{J}$ in form of Eq. (2.13) strictly satisfies the semi-positive definiteness condition. We emphasize that the non-Markovian encoder weights $\left\{\boldsymbol{w}_i(t_j)\right\}_{j=1}^{N_w}$ do not explicitly appear in the loss function. However, they are involved in the training process along with $\boldsymbol{J}$ since the correlation functions $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ 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 $[\boldsymbol{q}; \boldsymbol{p}; \boldsymbol{\zeta}]$. As shown in Sec. 3.3, 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 leads 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 ADAM optimization algorithm Kingma and Ba (2015) where 1000 points are randomly selected per each training step

We conclude this subsection with the following remarks: (I) Instead of Eq. (2.14), the reduced dynamics can be also cast into the evolution of the correlation matrices by multiplying $\boldsymbol{q}(0)$ to both

sides of Eq. (2.5). 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 $\boldsymbol{M}\dot{\boldsymbol{v}} + \nabla U(\boldsymbol{q})$; 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 following Sec. 3.3, all the reduced models are constructed with this choice. For simple notation, we set it to be the last feature. For example, the fourth-order reduced model is constructed using 4 non-Markovian features. $\boldsymbol{\zeta}_1$, $\boldsymbol{\zeta}_2$ and $\boldsymbol{\zeta}_3$ take the form of Eq. (2.4), and $\boldsymbol{\zeta}_4$ is set to be $\boldsymbol{M}\dot{\boldsymbol{v}} + \nabla U(\boldsymbol{q})$. (III) In principle, for reduced models of Hamiltonian systems, the form of matrix $\boldsymbol{J}$ can be further restricted to

$$\boldsymbol{J} = -\mathrm{diag}(0, \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T) + \boldsymbol{J}^A, \tag{2.16}$$

where $\hat{\boldsymbol{L}} \in \mathbb{R}^{nm\times nm}$ is a lower-triangle matrix. Eq. (2.16) ensures that the embedded kernel in Eq. (2.5) does not contain the Markovian memory term (i.e., $\left(\boldsymbol{J}_{11} + \boldsymbol{J}_{11}^T\right)\delta(t)$). $\tilde{\boldsymbol{\theta}}(t)$ recovers the form of standard GLE and the second fluctuation-dissipation relationship shown in Eq. (2.9) recovers the standard form, i.e., $\left\langle \tilde{\boldsymbol{\mathcal{R}}}(t)\tilde{\boldsymbol{\mathcal{R}}}(t')^T \right\rangle = \beta^{-1}\tilde{\boldsymbol{\theta}}(t-t')$. In this study, both forms yield accurate numerical results; the contribution of the Markovian term constructed by Eq. (2.13) is less than 1%.

2.3 Numerical results

2.3.1 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

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

where $\boldsymbol{Q}_i$ and $\boldsymbol{Q}_j$ are the positions of i-th and j-th 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. Nosé-Hoover thermostat is used with $k_B T = 0.5$ 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 the NVE simulations of the full model for a production stage of 1×10^5 steps using the Velocity-Verlet integrator.

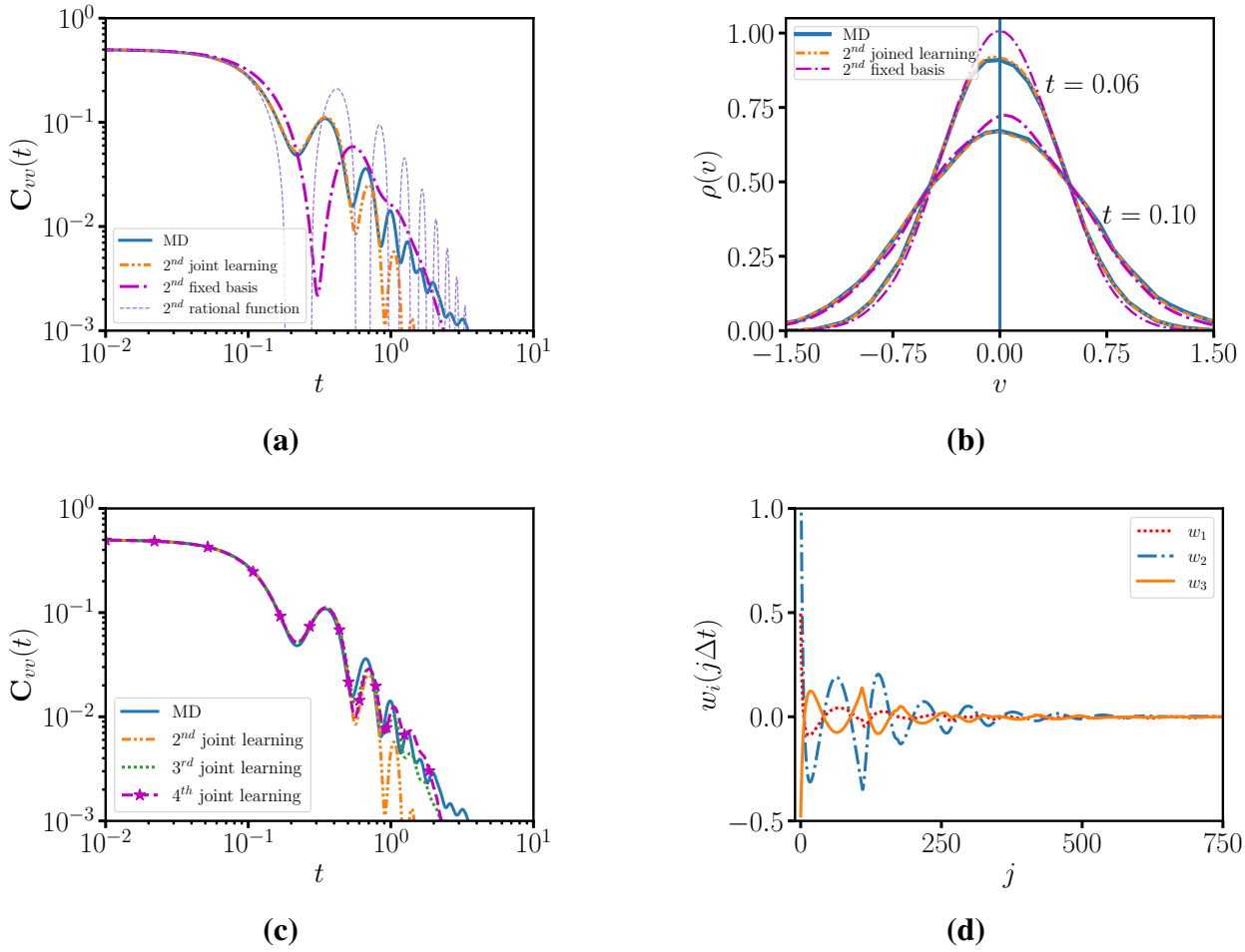


Figure 2.1 Numerical results for a tagged particle in the solvent particle bath. (a) Velocity correlation function $C_{vv}(t)$ obtained from the full MD model and the reduced models constructed by the present method based on the joint learning approximation, the rational function approximation Lei et al. (2016a), and the Petrov–Galerkin projection with fixed bases Lei and Li (2021). (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.

The reduced dynamics in terms of the tagged particle is constructed in form of Eq. (2.5) along with the learning of the non-Markovian features $\{\zeta_i\}_{i=1}^n$. The free energy $U(\boldsymbol{q})$ 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) Lei and Li (2021) and the rational function

approximation Lei et al. (2016a). Fig. 2.1(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 the 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. Fig. 2.1(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. 2.1(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. Fig. 2.1(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. Lei and Li (2021), the present method enables the encoders to be optimized for the best approximation of the non-local statistics, and therefore yields more accurate results.

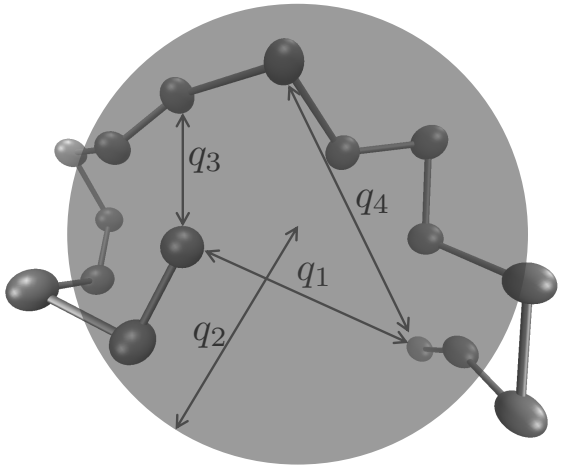


Figure 2.2 A sketch of a chain-molecule represented by united atom model. Reduced models are constructed with respect to a four-dimensional resolved vector $\boldsymbol{q}$, which represents the end-to-end distance (q_1), the radius of gyration (q_2), and the end-to-middle distances (q_3 and q_4), respectively.

2.3.2 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

$$V_{\mathrm{mol}}(\boldsymbol{Q}) = \sum_{i\neq j}^{N} V_{\mathrm{p}}(Q_{ij}) + \sum_{i=1}^{N_b} V_{\mathrm{b}}(l_i) + \sum_{i=1}^{N_a} V_{\mathrm{a}}(\theta_i) + \sum_{i=1}^{N_d} V_{\mathrm{d}}(\phi_i), \tag{2.17}$$

where l_i, θ_i, 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 multiharmonic dihedral interactions whose explicit forms are specified in Appendix A. The atom mass is set to unit, 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 the NVE simulations of the full model for a production stage of 1×10^6 steps using the Velocity-Verlet integrator. Fig. 2.2 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 $q_1 = \|\boldsymbol{Q}_1 - \boldsymbol{Q}_N\|$ as done in the previous work Lei and Li (2021) based on the Petrov-Galerkin method, and compare the numerical results obtained from the two methods. Figure 2.3(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 $t = 0.4$ and $t = 4$. However, the previous method with the same order approximation shows limitation to accurately capture these two patterns.

Figure 2.3(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 $t = 25$. Figure 2.3(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.

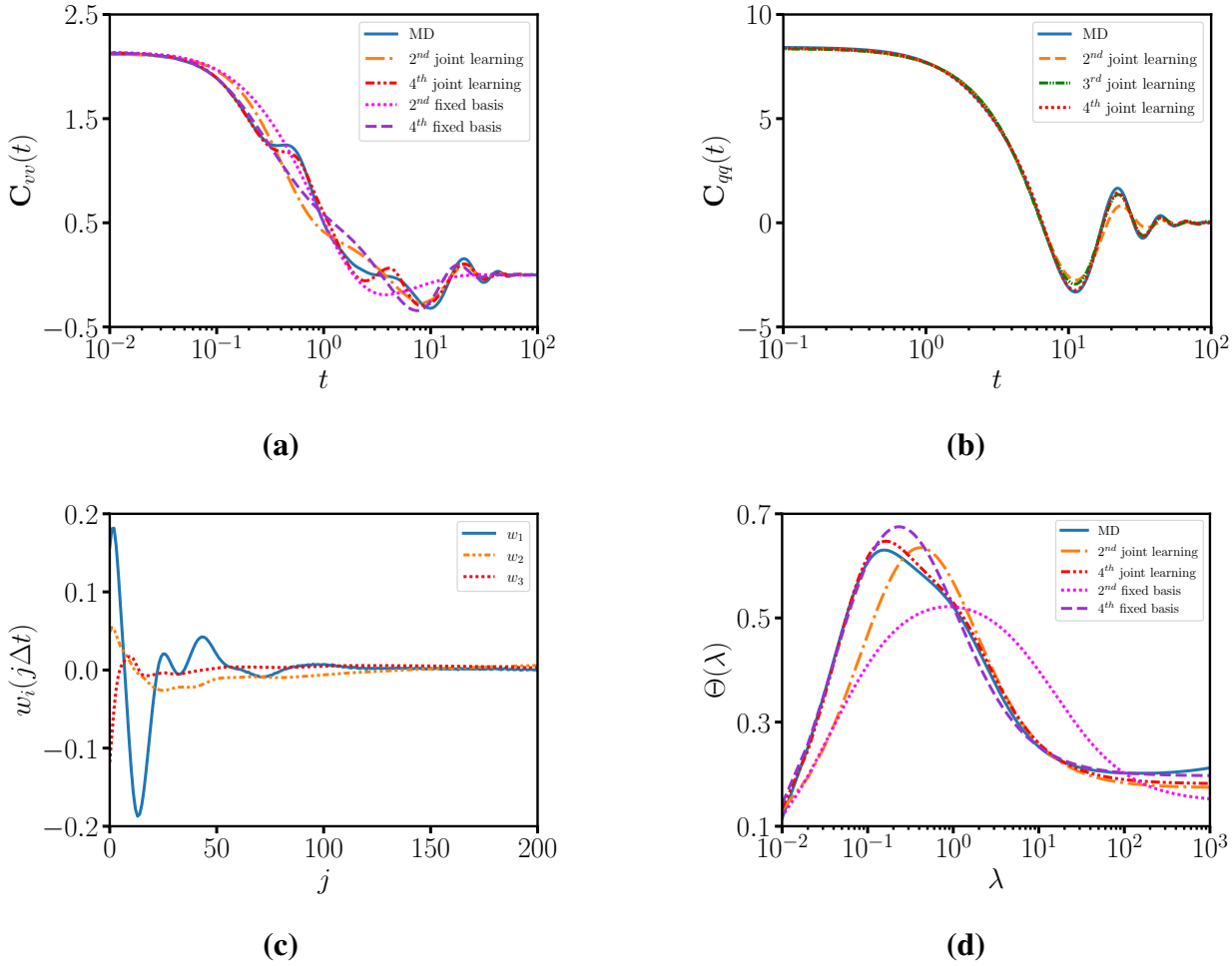


Figure 2.3 Numerical results of a one-dimensional reduced model representing the dynamics of the end–end distance of a polymer molecule system. (a)–(b) Velocity correlation function $C_{vv}(t)$ and the Laplace transform of the memory function $\Theta(\lambda)$ obtained from the full MD simulations and 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.

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 2.3(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 $\boldsymbol{\Theta}(\lambda) = -\boldsymbol{G}(\lambda)\boldsymbol{H}(\lambda)^{-1}$, where $\boldsymbol{G}(\lambda)$ and $\boldsymbol{H}(\lambda)$ are the Laplace transform of the correlation matrices $\boldsymbol{g}(t) = \langle \boldsymbol{M}\dot{\boldsymbol{v}}(t) + \nabla U(\boldsymbol{q}), \boldsymbol{q}(0) \rangle$ and $\boldsymbol{h}(t) = \langle \boldsymbol{v}(t), \boldsymbol{q}(0) \rangle$. 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 the constraint (2.14) 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.

2.3.3 Four-dimensional reduced model of a polymer molecule

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

$$
\begin{aligned}
q_1 &= \|\boldsymbol{Q}_1 - \boldsymbol{Q}_N\|, \\
q_2^2 &= \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{Q}_i - \mathbf{Q}_c\|^2, \quad \mathbf{Q}_c = \frac{1}{N}\sum_{i=1}^{N} \mathbf{Q}_i, \\
q_3 &= \left\| \mathbf{Q}_{\lfloor \frac{N}{2} \rfloor} - \mathbf{Q}_1 \right\|, \\
q_4 &= \left\| \mathbf{Q}_{\lceil \frac{N}{2} \rceil} - \mathbf{Q}_N \right\|,
\end{aligned}
\tag{2.18}
$$

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 $U(\boldsymbol{q})$ 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 (2.5) by minimizing the loss function (2.15).

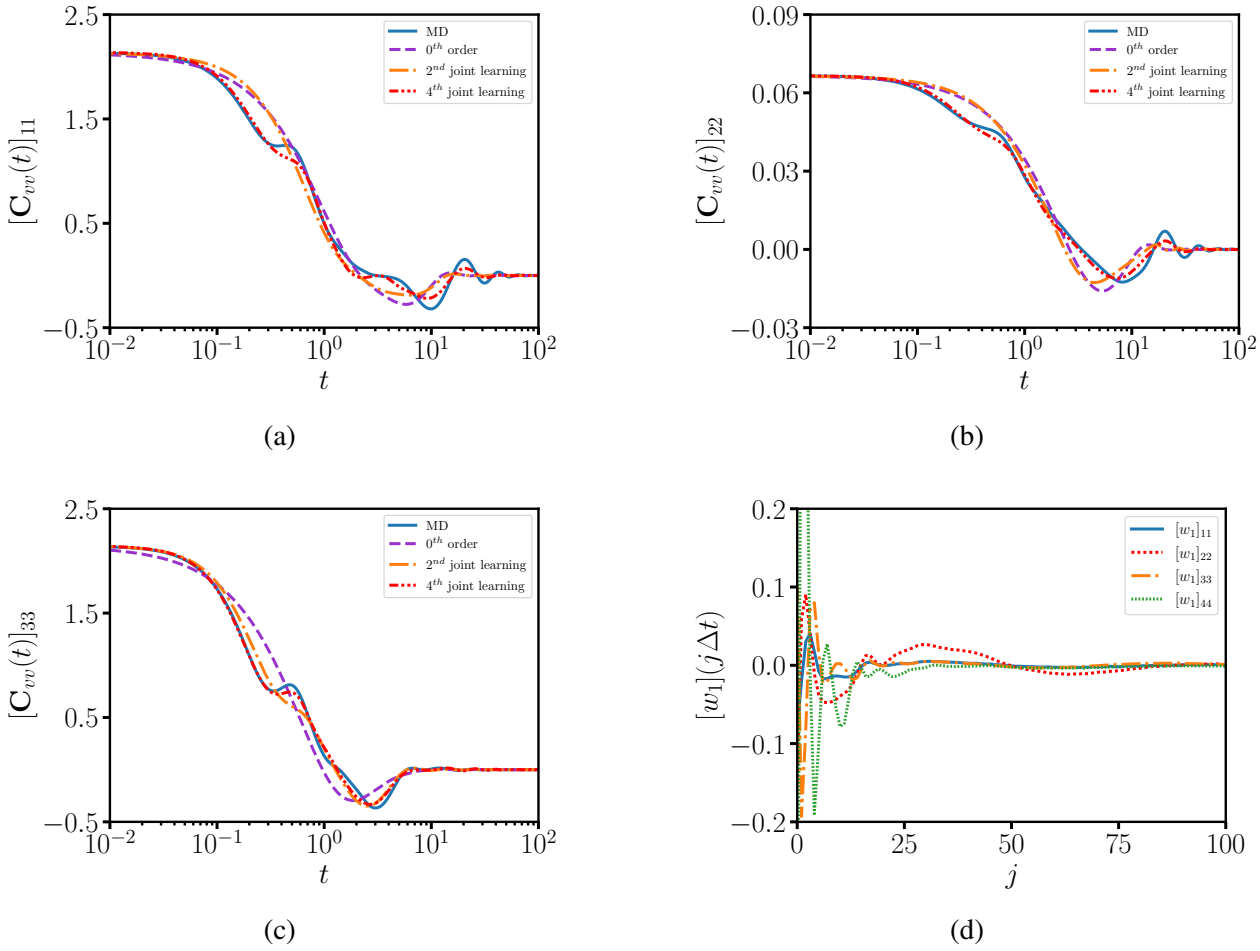


Figure 2.4 Numerical results of a four-dimensional reduced model representing the dynamics of a polymer molecule system, with conformation states characterized by the resolved variables $\boldsymbol{q}$ (see Eq. (2.18)). (a)–(c) Diagonal components of the velocity correlation function $\boldsymbol{C}_{vv}(t) = \langle \boldsymbol{v}(t)\, \boldsymbol{v}(0)^T \rangle$. Note that $[\boldsymbol{C}_{vv}(t)]_{44}$ is omitted because it is similar to $[\boldsymbol{C}_{vv}(t)]_{33}$. (d) Constructed encoder weights of the first non-Markovian feature $\boldsymbol{\zeta}_1$ for the fourth-order reduced model.

Figure 2.4(a-c) show the diagonal components of the velocity correlation matrix $\boldsymbol{C}_{vv}(t) = \left\langle \boldsymbol{v}(t)\boldsymbol{v}(0)^T \right\rangle$ obtained from the full MD and the reduced models using different order approximations. Specifically, the components $[\boldsymbol{C}_{vv}(t)]_{11}$ and $[\boldsymbol{C}_{vv}(t)]_{33}$ show similar values near $t = 0$ since both q_1 and q_3 characterize the distances between the individual particles, e.g., $v_1 = (\boldsymbol{Q}_1 - \boldsymbol{Q}_N) \cdot (\boldsymbol{V}_1 - \boldsymbol{V}_N)/\|\boldsymbol{Q}_1 - \boldsymbol{Q}_N\|$. As the distribution of the velocity difference between the two free-end particles follows $\mathcal{N}(0, 2k_BT\boldsymbol{I})$, the variance of v_1 is $2k_BT$. Similar argument also holds for v_3 and v_4. On the long-time scale, $[\boldsymbol{C}_{vv}(t)]_{11}$ and $[\boldsymbol{C}_{vv}(t)]_{22}$ decay much slower than $[\boldsymbol{C}_{vv}(t)]_{33}$ and $[\boldsymbol{C}_{vv}(t)]_{44}$

and show pronounced oscillations near $t = 10$ and $t = 25$. 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 molecular conformation. Based on the scaling law of the idealized Gaussian chain model de Gennes (1979), the relaxation time of q_1 and q_2 is proportional to N^2. Accordingly, $[\boldsymbol{C}_{vv}(t)]_{11}$ decays four times slower than $[\boldsymbol{C}_{vv}(t)]_{33}$, 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 $[\boldsymbol{C}_{vv}(t)]_{33}$ but less accurate predictions for $[\boldsymbol{C}_{vv}(t)]_{11}$ and $[\boldsymbol{C}_{vv}(t)]_{22}$. The fourth-order model yields good agreement for all the components over the full regime. Fig. 2.4(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.

Fig. 2.5 shows the off-diagonal components of the velocity correlation matrix $\boldsymbol{C}_{vv}(t)$. Similar to the diagonal components, $[\boldsymbol{C}_{vv}(t)]_{12}$ represents the coupling between the dynamics of two global conformation states and therefore exhibits the longest correlation with pronounced oscillations at $t = 10$ and $t = 25$. On the other hand, $[\boldsymbol{C}_{vv}(t)]_{13}$ and $[\boldsymbol{C}_{vv}(t)]_{23}$ represent the coupling between a global state and semi-global state, and therefore exhibit intermediate correlation. In addition, $[\boldsymbol{C}_{vv}(t)]_{34}$ 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 8-th and 9-th atom. The predictions of the second-order reduced model show fairly good agreement with the full MD results for $[\boldsymbol{C}_{vv}(t)]_{13}$ and $[\boldsymbol{C}_{vv}(t)]_{23}$ but less agreement for $[\boldsymbol{C}_{vv}(t)]_{12}$. The fourth-order reduced model yields good agreement for all the components.

Fig. 2.6 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

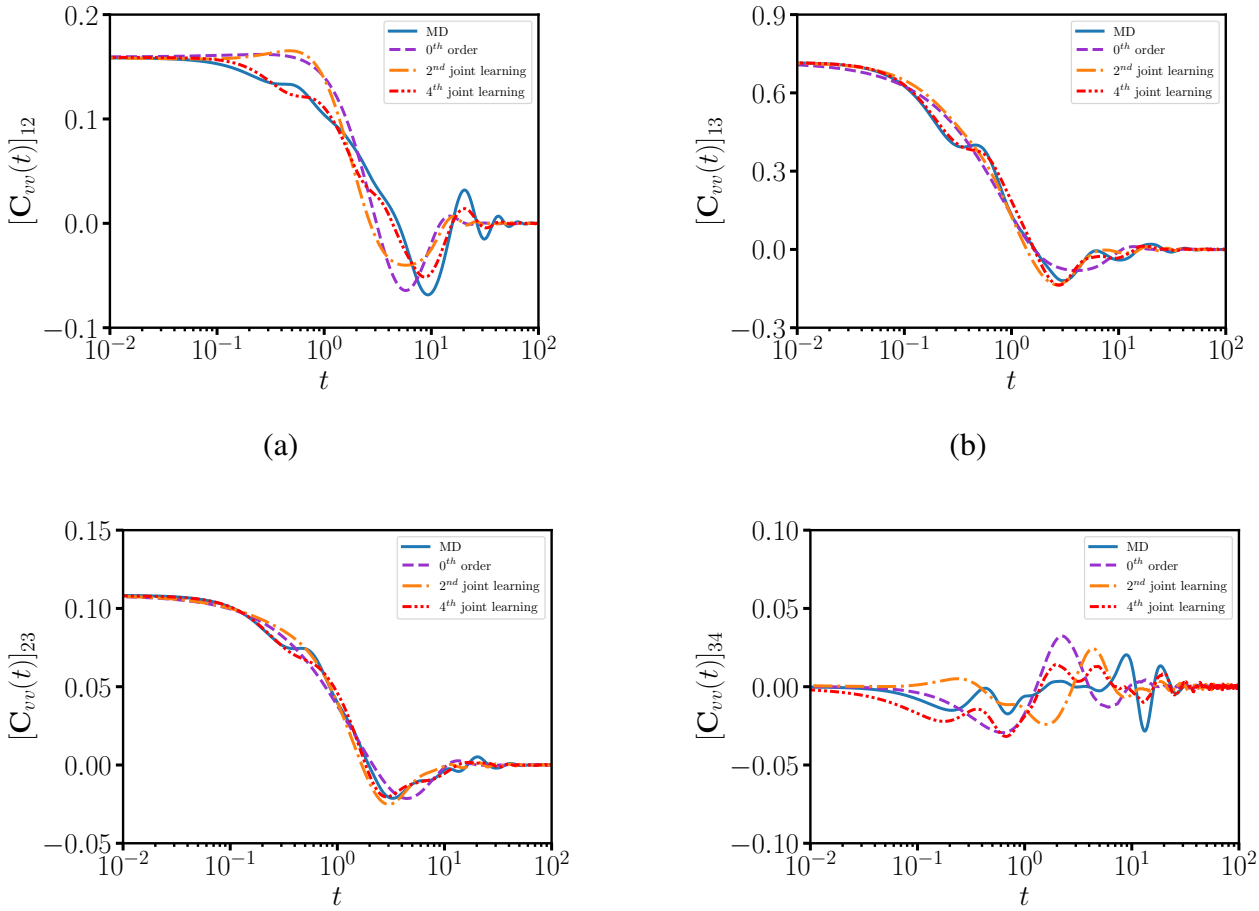


Figure 2.5 (a-d) Off-diagonal components of the velocity correlation function $\boldsymbol{C}_{vv}(t)$ for a polymer molecule system whose conformation states are characterized by a four-dimensional resolved vector $\boldsymbol{q}$ defined by Eq. (2.18).

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 Fig. 2.4 and 2.5 (see also Appendix E 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.

2.4 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 provides an efficient approach to approximate the multi-dimensional generalized Langevin equation.

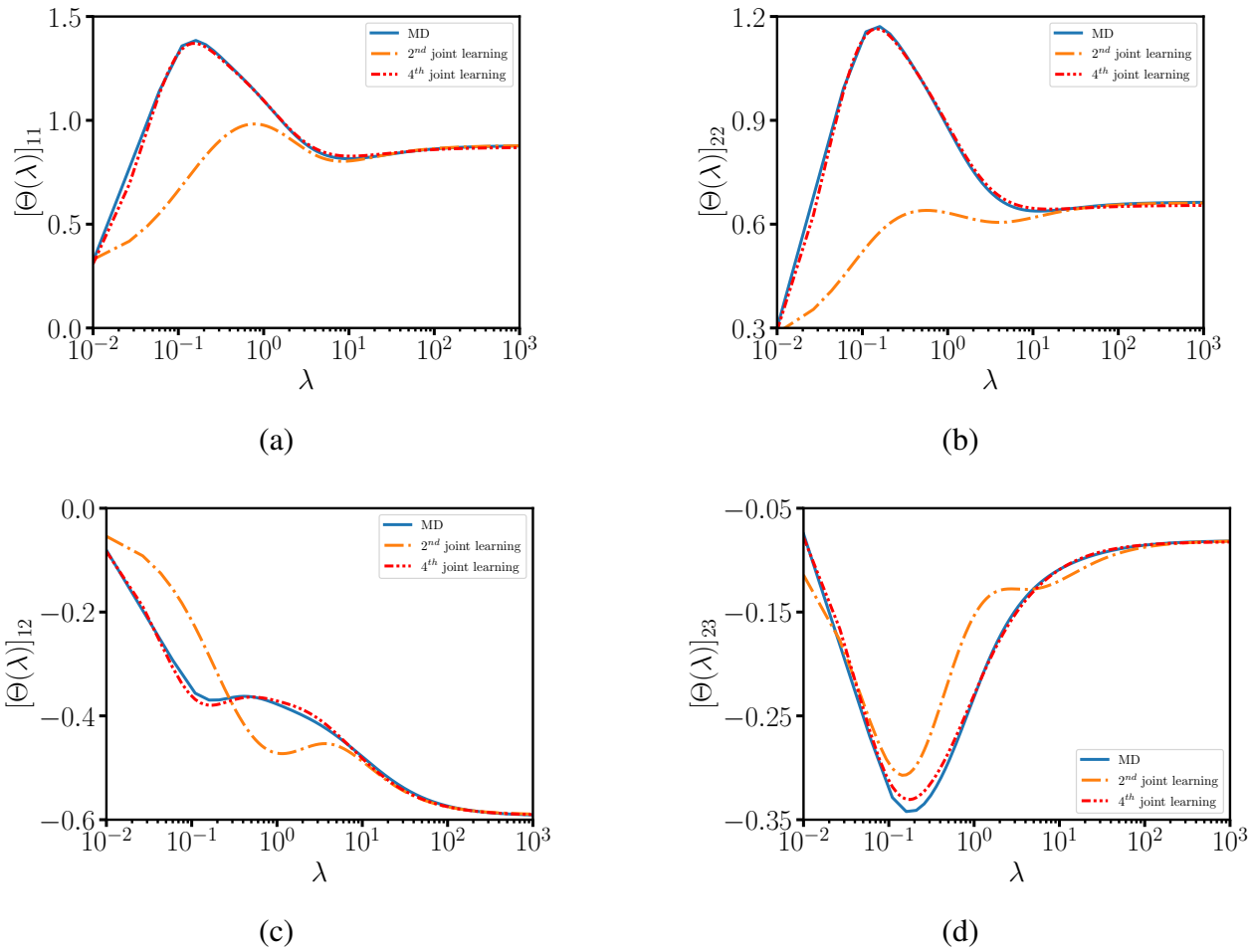


Figure 2.6 (a-d) Components of the embedded matrix-valued kernel $\mathbf{\Theta}(\lambda)$ in the Laplace space obtained from the full MD and a four-dimensional reduced model of a polymer molecule system.

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 Lei et al. (2016a) and the Petrov-Galerkin projection Lei and Li (2021) with the pre-selected fractional derivative bases, the present method enables us to probe the 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.

Building upon the data-driven framework for modeling state-independent memory effects introduced in Chapter 2, Chapter 3 extends this approach to more complex dynamical systems with state-dependent memory. While Chapter 2 demonstrated how a fixed set of convolutional encoders could capture global non-Markovian behavior through auxiliary variables, this assumption becomes limiting in systems where memory varies across configurations — such as when transitions between metastable states occur. To address this, Chapter 3 introduces a heterogeneous encoding architecture that allows the memory kernels to adapt locally to the system's state. This generalization enables a more accurate and flexible representation of reduced dynamics in high-dimensional systems where memory effects are inherently configuration-dependent.

CHAPTER 3

ENHANCED SAMPLING DATA-DRIVEN CONSTRUCTION OF STOCHASTIC REDUCED DYNAMICS ENCODED WITH STATE-DEPENDENT MEMORY

3.1 Introduction

Predictive modeling of multi-scale dynamic systems remains a significant challenge across various fields, including biology, materials science, and fluid physics. A prominent example is coarse-grained molecular dynamics (CGMD), where the goal is to simplify molecular system representations while preserving their essential dynamic behavior. The generalized Langevin equation (GLE) has emerged as a widely used framework for capturing the non-Markovian dynamics inherent in many CGMD processes. A range of approaches has been proposed for parameterizing the memory Lange and Grubmüller (2006); Darve et al. (2009b); Ceriotti et al. (2009); Baczewski and Bond (2013); Davtyan et al. (2015b); Lei et al. (2016b); Russo et al. (2019); Jung et al. (2017b); Lee et al. (2019b); Ma et al. (2019); Wang et al. (2020b,c); Zhu and Venturi (2020); Vroylandt et al. (2022); She et al. (2023); Xie and E (2024) aiming to reconstruct specific dynamic properties accurately. However, recent work Lyu and Lei (2023b); Ge et al. (2024) reveal that recovering isotropic properties alone may be insufficient for accurately reproducing the underlying complex dynamics. These findings underscore the importance of incorporating state-dependent memory effects to achieve precise reconstruction of dynamic behaviors.

The accurate parameterization of a state-dependent memory kernel hinges on effectively capturing the dynamic properties within the phase space. However, practical applications often face challenges due to the inherent complexity of the energy landscape in phase space. This complexity is typically marked by the presence of numerous metastable states, which are separated by significant energy barriers. These barriers hinder transitions between states, making it difficult to comprehensively sample the phase space and accurately reconstruct the memory kernel. Addressing this challenge requires advanced techniques capable of efficiently exploring these landscapes while retaining essential dynamic information. The critical role of sampling in phase space has been widely acknowledged, particularly in the construction of free energy landscapes. To address this, numerous

methodologies have been developed, each offering unique advantages for overcoming sampling challenges. Notable approaches include umbrella sampling Torrie and Valleau (1977), which applies biased potentials to enhance exploration; histogram reweighting Kumar et al. (1992), enabling the integration of data from multiple simulations; metadynamics Laio and Parrinello (2002b); Barducci et al. (2008), which facilitates the escape from metastable states through adaptive biasing; and variational enhanced sampling Valsson and Parrinello (2014); Shaffer et al. (2016); Bonati et al. (2019), a framework that leverages variational principles to optimize bias potentials. These methods collectively underscore the importance of efficient phase space exploration in capturing accurate free energy profiles. Despite their great success and wide application to capture the static properties, the importance of the sampling for the dynamic properties is largely ignored.

In this study, we employ our previously developed consensus-based enhanced sampling technique to simultaneously construct the free energy surface and parameterize the memory kernel. The conservative force is determined through constrained dynamics at selected points in the phase space, while dynamic information at these points is obtained via multiple free dynamics simulations initiated from the same locations.

3.2 Methods

The system under consideration is modeled as a Hamiltonian system with a $6N$-dimensional phase space vector $\mathbf{Z} = [\mathbf{Q}; \mathbf{P}]$, represent the position and momentum vectors, respectively. The dynamics of the system are governed by the equation of motion:

$$\dot{\mathbf{Z}} = \mathbf{S}\nabla H(\mathbf{Z}), \tag{3.1}$$

where $\mathbf{S} = \begin{pmatrix} 0 & \mathbf{I} \\ -\mathbf{I} & 0 \end{pmatrix}$ is the symplectic matrix that preserves the structure of Hamiltonian dynamics, with $\mathbf{I}$ being the identity matrix. and $H(\mathbf{Z})$ denotes the Hamiltonian function. For sufficiently large N, the simulation of Eq. (3.1) becomes computationally prohibitive. However, in many practical scenarios, interest lies in a low-dimensional resolved variable, $\mathbf{z}(t) = \phi(\mathbf{Z}(t))$, where $\phi : \mathbb{R}^{6N} \to \mathbb{R}^m$ serves as a mapping that projects the high-dimensional pace onto a reduced space of interest.

The Mori-Zwanzig (MZ) formalism provides a robust foundation for constructing approximate

dynamics for resolved variables by employing a projection operator. This framework separates the resolved and unresolved components of the system, enabling a reduced description of the dynamics while incorporating memory effects and fluctuating forces to account for the influence of unresolved variables. The projection operator $\mathcal{P}$ maps functions of the full system to functions of the coarse-grained (CG) system, and is defined as:

$$(\mathcal{P}f)(\mathbf{z}) = \frac{\int \delta(\phi(\mathbf{Z}) - \mathbf{z}) f(\mathbf{Z}) \rho(\mathbf{Z}) \mathrm{d}\mathbf{Z}}{\int \delta(\phi(\mathbf{Z}) - \mathbf{z}) \rho(\mathbf{Z}) \mathrm{d}\mathbf{Z}},$$

where $\mathbf{z}$ represents the CG variables, $\Phi(\mathbf{Z})$ is the mapping from the full system to the CG system, and $\rho(\mathbf{Z})$ is the probability density function of the full system. The dynamics of CG variable follows:

$$\frac{\partial}{\partial t}\phi(\mathbf{Z}) = \exp(t\mathcal{L})\mathcal{P}\mathcal{L}\phi(\mathbf{Z}) + \int_0^t \exp\left((t-s)\mathcal{L}\right)\mathcal{P}\mathcal{L}\exp(s\mathcal{Q}\mathcal{L})\mathcal{Q}\mathcal{L}\phi(\mathbf{Z})\mathrm{d}s + \exp(t\mathcal{Q}\mathcal{L})\mathcal{Q}\mathcal{L}\phi(\mathbf{Z}).$$

Here, $\mathcal{L} :=$ is the Liouville operator and $\mathcal{Q} = \mathbf{I} - \mathcal{P}$. Motivate by this, the reduced dynamics can be written as

$$\begin{aligned} \dot{\mathbf{q}} &= \mathbf{M}(\mathbf{q})^{-1}\mathbf{p}, \\ \dot{\mathbf{p}} &= -\mathbf{F}(\mathbf{q}) - \int_0^t \theta(t-\tau)\dot{\mathbf{q}}(\tau)\mathrm{d}\tau + \mathcal{R}(t), \end{aligned} \tag{3.2}$$

Here, $\mathbf{q} = \phi_q(\mathbf{Q})$ denotes a coarse-grained variable, and $\mathbf{p}$ is the corresponding momentum, associated with a mass matrix $\mathbf{M}(\mathbf{q})$. The effective free energy for $\mathbf{q}$ is defined as $U_{\text{eff}}(\mathbf{q}) = -\frac{1}{\beta}\log\int \mathrm{d}\mathbf{Z}\delta(\phi_q(\mathbf{Z}) - \mathbf{q})\rho(\mathbf{Z})$, with $\beta = \frac{1}{K_B T}$ is the inverse temperature. Inspired by previous research She et al. (2023), Equation (3.2) can be reformulated as an extended Markovian process $(\mathbf{q}, \mathbf{p}, \boldsymbol{\xi})$, where $\boldsymbol{\xi}$ represents auxiliary variables that will be defined later. These auxiliary variables serve to capture the memory effects inherent in the original generalized Langevin equation (GLE) and are assumed to follow a Gaussian distribution, $N(0, 1)$. The evolution takes the form

$$\mathrm{d}\begin{pmatrix} \mathbf{q} \\ \mathbf{p} \\ \boldsymbol{\xi} \end{pmatrix} = \begin{pmatrix} 0 & \mathbf{I} & 0 \\ -\mathbf{I} & & \\ 0 & \mathbf{J}(\mathbf{q}) & \end{pmatrix}\begin{pmatrix} \nabla_{\mathbf{q}}\mathcal{F} \\ \nabla_{\mathbf{p}}\mathcal{F} \\ \nabla_{\boldsymbol{\xi}}\mathcal{F} \end{pmatrix}\mathrm{d}t + \begin{pmatrix} 0 & 0 & 0 \\ 0 & & \\ 0 & \boldsymbol{\Sigma}(\mathbf{q}) & \end{pmatrix}\mathrm{d}\mathbf{W}_t, \tag{3.3}$$

where $\mathcal{F}(\mathbf{q}, \mathbf{p}, \boldsymbol{\xi}) = U(\mathbf{q}) + \mathbf{p}^T\mathbf{M}(\mathbf{q})^{-1}\mathbf{p} + \boldsymbol{\xi}^T\boldsymbol{\xi}$ is the free energy for the extended system. The relationship $\boldsymbol{\Sigma}(\mathbf{q})\boldsymbol{\Sigma}(\mathbf{q})^T = -K_B T(\mathbf{J}(\mathbf{q})\boldsymbol{\Lambda} + \boldsymbol{\Lambda}\mathbf{J}(\mathbf{q})^T)$, ensures consistency with the second fluctuation-dissipation theorem, where $\boldsymbol{\Lambda}$ is covariance matrix of $(\mathbf{p}, \boldsymbol{\xi})$. By solving the Fokker-Planck equation,

we have the invariant distribution of the extended system $\rho_e(\mathbf{q},\mathbf{p},\boldsymbol{\xi}) \propto \exp(-\beta\mathcal{F}(\mathbf{q},\mathbf{p},\boldsymbol{\xi}))$. The invariant distribution should be consistent with the free energy for $\mathbf{q}$, with $\int \rho_e(\mathbf{q},\mathbf{p},\boldsymbol{\xi})\mathrm{d}\mathbf{p}\mathrm{d}\boldsymbol{\xi} \propto \exp(-\beta U_{\text{eff}}(\mathbf{q}))$, from which we notice that $U(\mathbf{q}) = U_{\text{eff}}(\mathbf{q}) - \frac{1}{2\beta}\log|\mathbf{M}(\mathbf{q})^{-1}|$. To construct the hidden variables, we define them as a linear combination of past momentum values $\mathbf{p}(t-\delta t),\cdots,\mathbf{p}(t-N_w\delta t)$, i.e

$$\boldsymbol{\xi}_i(t) = \sum_{j=1}^{N_w} \mathbf{w}_{ji}\mathbf{p}(t-i\delta t),$$

where $\mathbf{w}_{ji}$ are the coefficients to be optimized, and N_ξ is the number of momentum terms included in the linear combination. To simplify the process of training and collecting data, we construct our matrix $\mathbf{J}(\mathbf{q})$ as follows

$$\mathbf{J}(\mathbf{q}) = \begin{pmatrix} \mathbf{0} & \mathbf{h}^T(\mathbf{q}) \\ -\mathbf{h}(\mathbf{q})\mathbf{M}^{-1}(\mathbf{q}) & \hat{\mathbf{J}}(\mathbf{q}) \end{pmatrix} \tag{3.4}$$

here $\mathbf{h}(\mathbf{q})$ is a vector represented by a neural network which takes the coarse-grained variable $\mathbf{q}$ as input. Then we use choleschy decomposition to form $\hat{\mathbf{J}}(\mathbf{q}) = -\mathbf{L}(\mathbf{q})\mathbf{L}^T(\mathbf{q}) + \mathbf{A}(\mathbf{q})$. Note that $\mathbf{L}(\mathbf{q})$ is a block-wise lower triangle matrix and $\mathbf{A}(\mathbf{q})$ is a block-wise antisymmetric matrix represented by two different neural networks respectively. Now, the the second fluctuation-dissipation theorem can be simplified as follows

$$\hat{\boldsymbol{\Sigma}}(\mathbf{q})\hat{\boldsymbol{\Sigma}}(\mathbf{q})^T = -K_BT(\hat{\mathbf{J}}(\mathbf{q})\hat{\boldsymbol{\Lambda}} + \hat{\boldsymbol{\Lambda}}\hat{\mathbf{J}}(\mathbf{q}))$$

where

$$\boldsymbol{\Sigma}(\mathbf{q}) = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{\Sigma}}(\mathbf{q}) \end{pmatrix}, \boldsymbol{\Lambda}(\mathbf{q}) = \begin{pmatrix} \mathbf{M}(\mathbf{q}) & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{\Lambda}}(\mathbf{q}) \end{pmatrix} \tag{3.5}$$

With the constructed $\boldsymbol{\xi}$, we can computed the correlation function by multiply $\mathbf{p}(0)$ given $\mathbf{q}(0) = \mathbf{q}$ on both side of Eq. (3.3),

$$\underbrace{\mathrm{d}\begin{pmatrix} \langle \mathbf{p} + \nabla_{\mathbf{q}}\mathcal{F}, \mathbf{p}(0) \,|\mathbf{q}(0) = \mathbf{q}\rangle \\ \langle \boldsymbol{\xi}, \mathbf{p}(0) \,|\mathbf{q}(0) = \mathbf{q}\rangle \end{pmatrix}}_{\mathbf{C}_1(t,\mathbf{q};w)} = \mathbf{J}(\mathbf{q}) \underbrace{\begin{pmatrix} \langle \nabla_{\mathbf{p}}\mathcal{F}, \mathbf{p}(0) \,|\mathbf{q}(0) = \mathbf{q}\rangle \\ \langle \nabla_{\boldsymbol{\xi}}\mathcal{F}, \mathbf{p}(0) \,|\mathbf{q}(0) = \mathbf{q}\rangle \end{pmatrix}}_{\mathbf{C}_2(t,\mathbf{q};w)} \mathrm{d}t. \tag{3.6}$$

By construct $\mathbf{J}(q)$ as $\tilde{\mathbf{J}}(\mathbf{q};\theta_J) = \tilde{\mathbf{\Sigma}}(\mathbf{q};\theta_J)\tilde{\mathbf{\Sigma}}(\mathbf{q};\theta_J)^T$, the loss function can be constructed as

$$\min_{\theta_J,\mathbf{w}} \sum_{i=1}^{N_s} \left\| \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{C}_1(t,\mathbf{q}_i;w) - \tilde{\mathbf{J}}(\mathbf{q}_i;\theta_J)\mathbf{C}_2(t,\mathbf{q}_i;\mathbf{w}) \right\|^2 + \sum_{i=1}^{N_s} \left\| \hat{\mathbf{\Lambda}}(\mathbf{q}_i;\mathbf{w}) - \mathbf{I} \right\|^2, \tag{3.7}$$

where we optimize $\tilde{\mathbf{J}}(\mathbf{q};\theta_J)$ and auxiliary variable $\boldsymbol{\xi}$ depends on $\mathbf{w}$ at the same time on the training set $\{\mathbf{q}_i\}_{i=1}^{N_s}$. The construction of dynamics also depends on an accurate free energy surface $U_{\text{eff}}(\mathbf{q})$ and mass matrix $\mathbf{M}(\mathbf{q})$. It can be computed from restrained dynamics by introducing a harmonic term into full potential, i.e.

$$\mathcal{U}_k(\mathbf{Q},\mathbf{q}) = \mathcal{U}(\mathbf{Q}) + \frac{k}{2}(\phi_q(\mathbf{Q}) - \mathbf{q})^T(\phi_q(\mathbf{Q}) - \mathbf{q}),$$

where k represents the magnitude of the restrained potential and $\mathcal{U}$ is the potential of full system without restraint. The mean force can be computed by $\nabla U_{\text{eff}}(\mathbf{q}) = \lim_{k\to\infty} \mathbf{F}^k(\mathbf{q})$ and $\mathbf{M}(\mathbf{q}) = \lim_{k\to\infty} \mathbf{M}^k(\mathbf{q})$, where

$$\mathbf{F}^k(\mathbf{q}) = \frac{1}{Z_k(\mathbf{q})} \int k(\phi_q(\mathbf{Q}) - \mathbf{q}) \exp(-\beta \mathcal{U}_k(\mathbf{Q},\mathbf{q}))\mathrm{d}\mathbf{Q}$$

and

$$\mathbf{M}^k(\mathbf{q}) = \frac{1}{Z_k(\mathbf{q})} \int \frac{1}{2\beta(\frac{\mathrm{d}\phi_q(\mathbf{Q})}{\mathrm{d}t})^2} \exp(-\beta \mathcal{U}_k(\mathbf{Q},\mathbf{q}))\mathrm{d}\mathbf{Q}.$$

Two neural networks $\tilde{\mathbf{M}}(\mathbf{q};\theta_M)$ and $\tilde{\mathbf{F}}(\mathbf{q};\theta_F)$ is constructed to approximate $\mathbf{M}(\mathbf{q})$ and $\mathbf{F}(\mathbf{q})$ respectively and trained loss function on the same dataset

$$\min_{\theta_M} \sum_{i=1}^{N_s} \|\tilde{\mathbf{M}}(\mathbf{q}_i;\theta_M) - \mathbf{M}(\mathbf{q}_i)\|^2, \min_{\theta_F} \sum_{i=1}^{N_s} \|\tilde{\mathbf{F}}(\mathbf{q}_i;\theta_F) - \mathbf{F}(\mathbf{q}_i)\|^2.$$

The sampling points are adaptively selected by consensus-based enhanced sampling strategy Lyu and Lei (2023a) with a McKean type stochastic differential equation for N_w walker $\mathbf{q}_1, \cdots, \mathbf{q}_{N_w}$

$$\mathrm{d}\mathbf{q}_i = -\frac{1}{\gamma}\nabla_{\mathbf{z}} G(\mathbf{q}_i)\mathrm{d}t + \sqrt{\frac{2}{\kappa_h \gamma}}\mathrm{d}\mathbf{W}_t \tag{3.8}$$

where $G(\mathbf{q}_i) = \frac{1}{2}(\mathbf{q}_i - \mathbf{m})^T\mathbf{V}^{-1}(\mathbf{q}_i - \mathbf{m})$, where $\mathbf{m} = \sum_{i=1}^{N_w} \mathbf{q}_i p(\mathbf{q}_i)$, $\mathbf{V} = (\kappa_l + \kappa_h)\sum_{i=1}^{N_w}(\mathbf{q}_i - \mathbf{m})^T(\mathbf{q}_i - \mathbf{m})p(\mathbf{q}_i)$ and $p(\mathbf{q}_i) = \frac{\exp(-\kappa_l \mathcal{R}(\mathbf{q}_i))}{\sum_j^{N_w} \exp(-\kappa_l \mathcal{R}(\mathbf{q}_j))}$. The $\mathcal{R}(\mathbf{q}_i)$ represents the residual at the point $\mathbf{q}_i$, which we choose to be $\|\tilde{\mathbf{F}}(\mathbf{q}_i;\theta_F) - \mathbf{F}(\mathbf{q}_i)\|^2$ in this project. The first right term in Eq. (3.8) represents the

exploitation term that uses current information to drive the sampler towards the maximum residual region, and the second term is a high temperature exploration term κ_h that explores the unknown region. Notice that the residual also depends on the neural network parameters, then by iterative optimization of our neural network representation and the sampling points, we can get a good training set over the phase space.

3.3 Numerical results

3.3.1 One-dimensional state-dependent reduced model of a polymer molecule

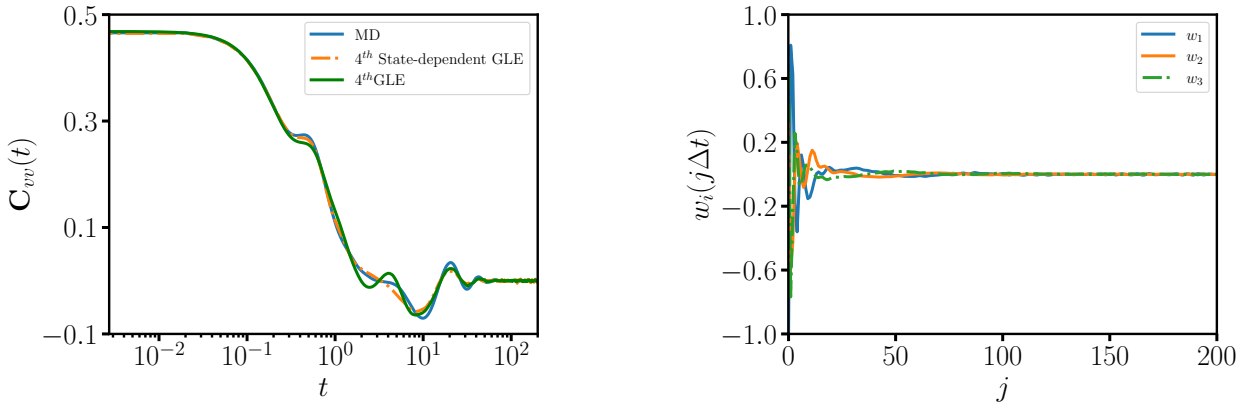
To illustrate the core concept of the present method, we begin with a polymer molecule with $N = 16$ atoms, where the intramolecular potential is defined by

$$V_{\mathrm{mol}}(\mathbf{Q}) = \sum_{i\neq j}^{N} V_{\mathrm{p}}(Q_{ij}) + \sum_{i=1}^{N_b} V_{\mathrm{b}}(l_i) + \sum_{i=1}^{N_a} V_{\mathrm{a}}(\theta_i) + \sum_{i=1}^{N_d} V_{\mathrm{d}}(\phi_i), \tag{3.9}$$

where V_p is the Lennard-Jones intermolecular potential, V_b is the harmonic bonds, V_a and V_d denotes the potential on angle and dihedral angle respectively. The end-to-end distance $q_1 = \|\mathbf{Q}_1 - \mathbf{Q}_N\|$ is used as a collective variable in the 1D reduced dynamics framework to evaluate the effectiveness of our current method. We selected 25 distinct points uniformly from the range of $q_1 \in [2, 18]$ to create our training set.

Four auxiliary variables $\boldsymbol{\xi}_i$ are learned in standard GLE and our state-dependent GLE. The overall velocity correlation function $C_{vv}(t) = \langle v_1(t)v_1(0)\rangle$ in presented in Figure 3.1(a). Both state-dependent GLE and standard GLE align well with the MD result. We also present the encoder weights for non-Markovian features as obtained from the state-dependent GLE approximation in Figure 3.1(b). The encoder functions demonstrate pronounced oscillations at short times, reflecting the dynamic interactions present in the system during that initial period. As time progresses, these oscillations gradually decay to zero, indicating that the influence of these non-Markovian features diminishes at longer time scales. This behavior underscores the transient nature of the non-Markovian dynamics in the system.

However, the fitness of the overall GLE do not represent the good performance of the learned dynamics. We compare the conditional autocorrelation with $\mathbf{q}$ starting from 25 selected points by



(a) **(b)**

Figure 3.1 Numerical results of a one-dimensional reduced model representing the dynamics of the end–end distance of a polymer molecule system. (a) Overall velocity correlation function $C_{vv}(t)$ obtained from MD, 4th order standard GLE, and state-dependent GLE. (b) Encoder weights for the three non-Markovian features of the state-dependent GLE.

standard GLE, state-dependent GLE and MD in Figure 3.2. This comparison will reveal differences of the diffusion behavior at different points on the phase space in MD is captured by the standard state-dependent but not in standard one. Inaccuracy in the diffusion process will in return affect the precision of the transition process.Figure 3.2(a) illustrates the time distribution of q_1 for values greater than 15, comparing results from MD simulations, GLE and state-dependent GLE. The data reveals that the state-dependent GLE provides a closer match to the MD results than the standard GLE method.

However, the fitness of the overall GLE do not represent the good performance of the learned dynamics. We compare the conditional autocorrelation with **q** starting from 25 selected points by standard GLE, state-dependent GLE and MD in Figure 3.2. This comparison will reveal differences of the diffusion behavior at different points on the phase space in MD is captured by the standard state-dependent but not in standard one. Inaccuracy in the diffusion process will in return affect the precision of the transition process.Figure 3.2(a) illustrates the time distribution of q_1 for values greater than 15, comparing results from MD simulations, GLE and state-dependent GLE. The data reveals that the state-dependent GLE provides a closer match to the MD results than the standard GLE method.

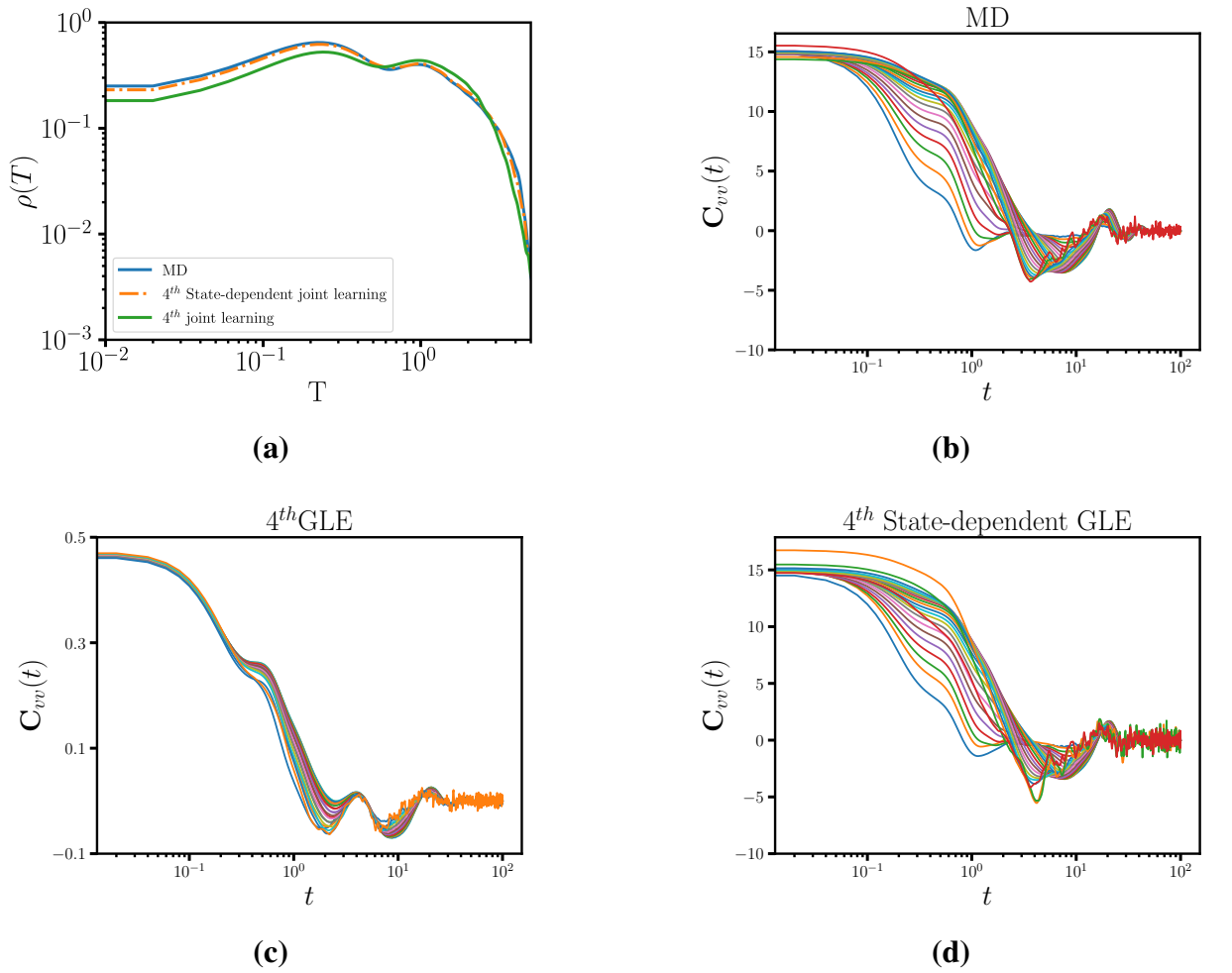


Figure 3.2 Numerical results of a one-dimensional reduced model representing the dynamics of the end–end distance of a polymer molecule system. (a) Distribution of $q_1 > 15$ obtained from the full MD simulations, the 4^{th}-order GLE approximation, and the 4^{th}-order state-dependent GLE approximation. (b–d) Conditional velocity correlation functions obtained from MD, standard GLE, and state-dependent GLE, respectively.

3.3.2 Two-dimensional state-dependent reduced model of an alanine dipeptid

We further demonstrate the effectiveness of our state-dependent reduced modeling framework using the alanine dipeptide molecule (Ace-Ala-Nme), commonly referred to as Ala2. The full-atom molecular dynamics (MD) simulation is performed for a solvated alanine dipeptide immersed in 383 explicit water molecules at 300 K, using the Amber99-SB force field and the TIP3P water model. A time step of 2.5×10^{-4} ps is used for numerical integration.

To reduce dimensionality, we adopt two dihedral angles as collective variables (CVs): the ϕ angle defined by atoms (C, N, C_α, C), and the ψ angle defined by atoms (N, C_α, C, N). These angles

provide a compact representation of the molecular conformations.

A consensus-based sampling strategy selects 1,000 representative configurations from the MD trajectory to train the state-dependent generalized Langevin equation (GLE) model. At each configuration, we compute the conservative force, effective mass, and the velocity autocorrelation function. Four auxiliary variables are introduced to close the non-Markovian system, forming a Markovian embedding that captures state-dependent memory effects.

The model's accuracy is validated by comparing conditional momentum autocorrelation functions from two conformational regions. As shown in Figure 3.3, the state-dependent GLE faithfully reproduces MD results, including subtle oscillatory features that are missed by traditional GLE models with fixed memory kernels.

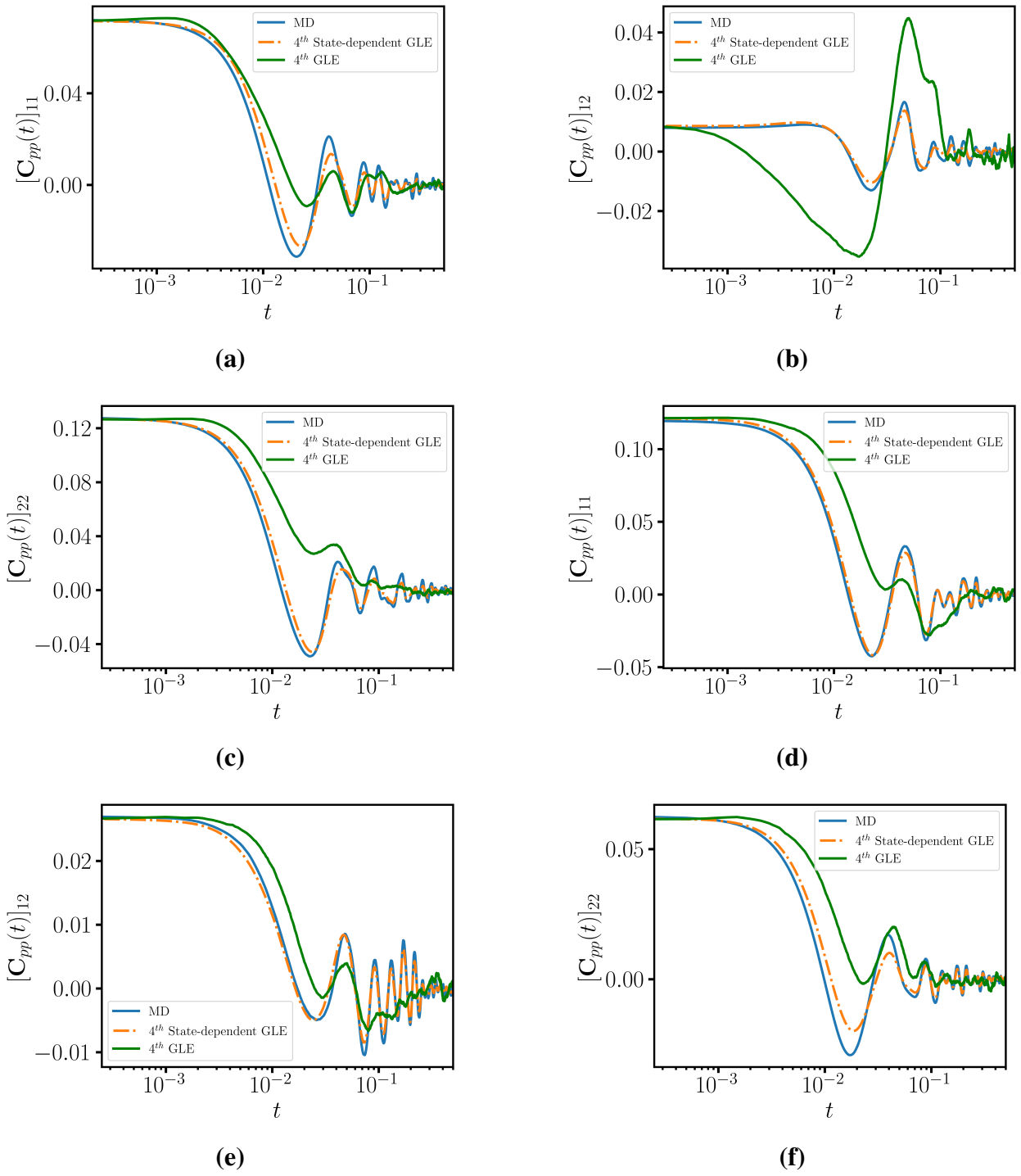
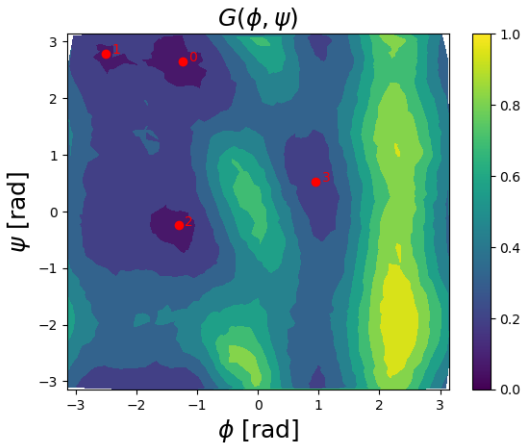


Figure 3.3 Numerical results of the two-dimensional reduced model in terms of the two dihedral angles of the alanine dipeptide system. (a–c) Conditional momentum auto-correlation functions obtained from full MD simulations and the 4^{th}-order GLE approximation at $(\phi, \psi) = (-1.60, 2.78)$. (d–f) Conditional momentum auto-correlation functions at $(\phi, \psi) = (-2.90, -0.16)$.

We also compute the distribution of transition time between different local minima. Four local

minima is selected in Figure 3.4. The distribution of transition time between point 0 and other three points is shown in Figure 3.5



(a)

Figure 3.4 The heatmap of the free energy surface $\mathbf{m}G(\phi, \psi)$. Colored solid circles mark four local minima of the configuration.

Figure 3.6 presents the distribution of time spent at each point, based on results from MD, 4^{th} order of GLE and 4^{th} order of sate-dependent GLE. The results demonstrate that our current method provides improved accuracy for each point compared to GLE method.

3.4 Summary

The state-dependent generalized Langevin equation (GLE) is usefull tool to describe the non-Markovian behavior in many processes in the CGMD problem accurately. In this study, we employ our previously developed consensus-based enhanced sampling strategy to simultaneously construct the heterogeneous memory kernel and the free energy surface. The conservative force is calculated using constrained dynamics at specific points in the phase space, while the dynamic information at these points is gathered through multiple free dynamics initiated from the same locations. We then train our neural network to capture the differences among the various conditional auto-correlation functions. The results demonstrate that our current method provides improved accuracy for each point compared to the GLE method.

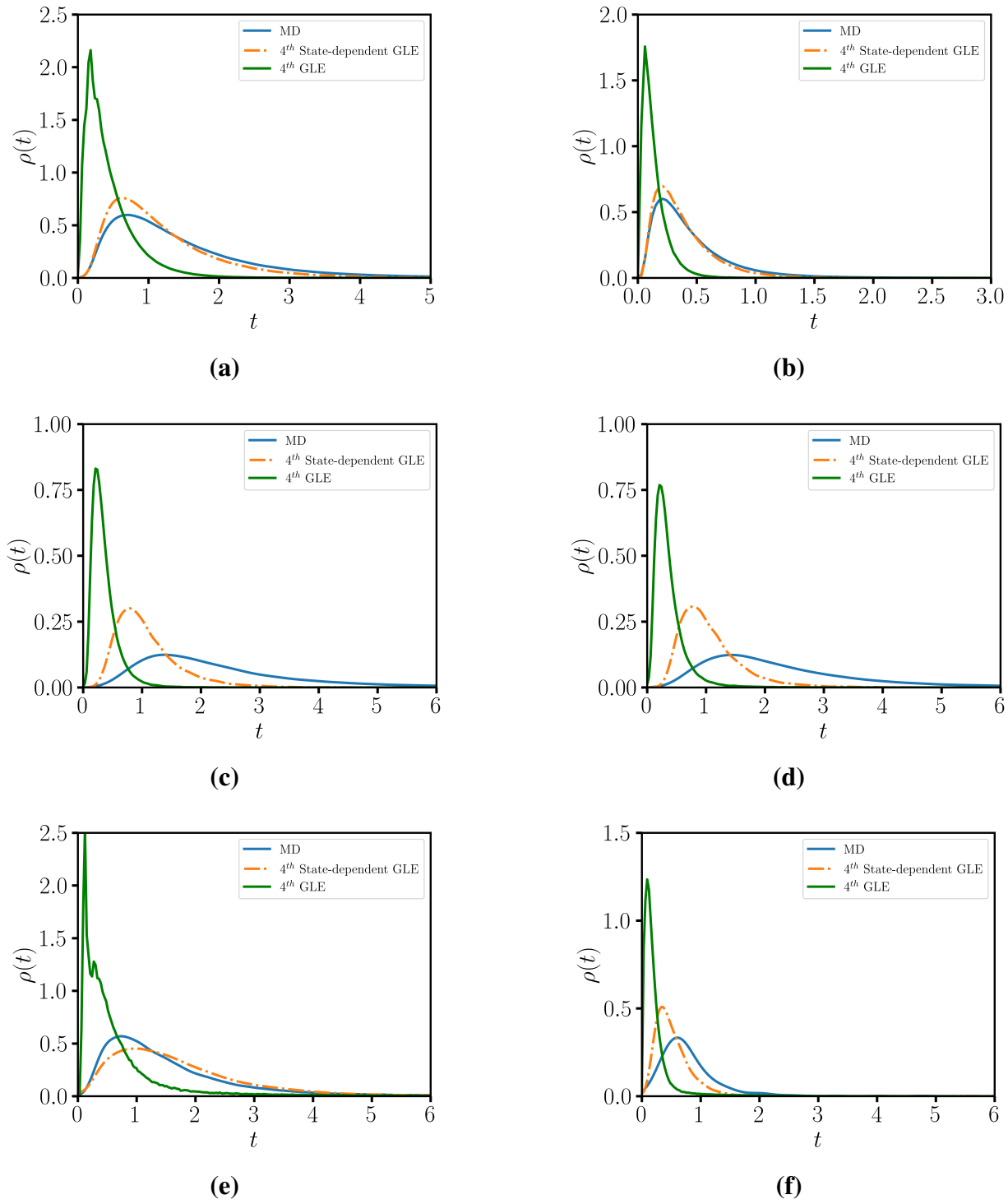


Figure 3.5 Numerical results of a two-dimensional reduced model representing the two dihedral angles of the alanine dipeptide system. (a) Distribution of transition time from position 0 to position 1. (b) From position 1 to 0. (c) From position 0 to 2. (d) From position 2 to 0. (e) From position 0 to 3. (f) From position 3 to 0.

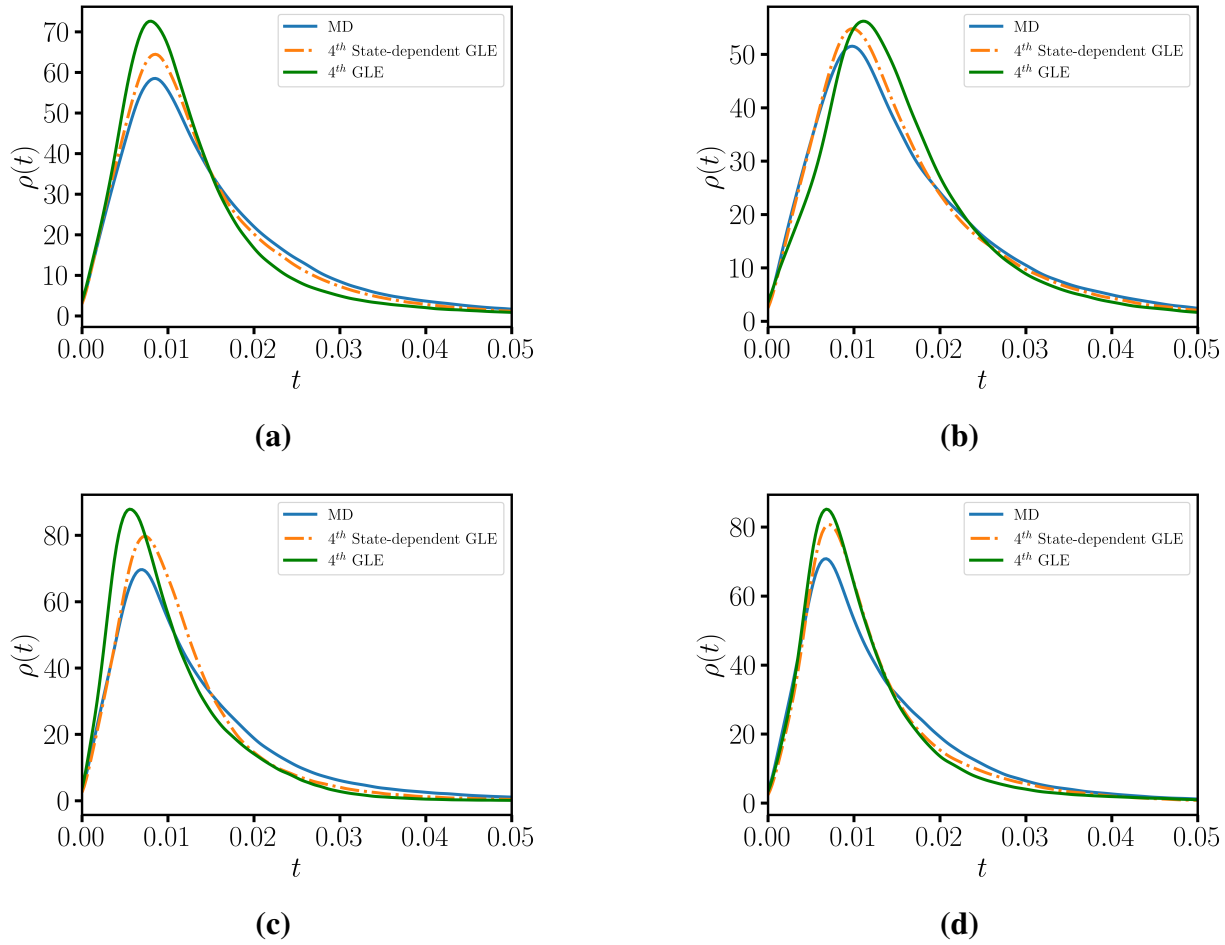


Figure 3.6 Numerical results of a two-dimensional reduced model representing the two dihedral angles of the alanine dipeptide system. (a) Distribution of time periods spent at position 0 before transitioning to position 1. (b–d) Distributions of time spent at positions 1, 2, and 3, respectively.

CHAPTER 4

GENERATIVE MODEL BASED IDENTIFYING METASTABLE STATES IN FULL MOLECULE SPACE

4.1 Introduction

Normalizing flows have gained significant traction in recent years as flexible generative models that provide exact likelihood evaluation and tractable sampling through invertible transformations between data and latent spaces. While highly expressive, their performance critically depends on the structure of the latent prior distribution. In most conventional settings—including in many state-of-the-art normalizing flow architectures—a simple unimodal prior such as a standard multivariate Gaussian is employed. This assumption works well for data distributions that are themselves unimodal or smoothly varying, but it becomes a substantial bottleneck in modeling systems characterized by multimodality, sharp transitions, or complex geometrical features in high-dimensional spaces.

This issue is particularly critical in domains such as molecular dynamics or continuum mechanics, where data often arise from a mixture of metastable states or rare-event transitions. These systems naturally lead to multimodal distributions, with each mode representing a distinct macroscopic configuration or energy basin. While recent developments in normalizing flows—such as NICE, RealNVP, and Glow Dinh et al. (2014, 2017); Kingma and Dhariwal (2018)—have significantly improved the expressivity of the transformation through architectural innovations like coupling layers and conditioning, they still rely on simple unimodal latent priors. As a result, such models are limited in their ability to identify and represent metastable states explicitly, since the prior structure does not reflect the inherent multimodality of the system.

Moreover, during training, these models typically focus on maximizing the overall likelihood and do not incorporate gradient or perturbation-based penalties to enforce key physical constraints—such as requiring the gradient of the log-density to vanish at latent maxima or ensuring that log-density values at these mapped maxima are indeed local maxima in data space. Without these constraints, the learned transformation may distort the latent structure and fail to preserve the correspondence

between latent and data-space modes, ultimately limiting the interpretability and metastable state resolution of the model.

To address these limitations, we propose a generative modeling framework that uses a Mixture-of-Gaussians (MoG) prior to explicitly represent multiple metastable modes in the latent space. The goal is not merely to enhance expressivity, but to enforce a maximum-to-maximum correspondence between the latent space and data space—ensuring that each latent mode is mapped to a high-density metastable state in the observed configuration space. To achieve this, we use an invertible transformation that preserves the structure of the distribution under the change of variables. While our implementation adopts KRNet for this transformation due to its flexibility and scalability, the approach is general and compatible with any expressive normalizing flow architecture. During training, we further impose gradient penalties to enforce vanishing gradients at mapped maxima, and contrastive perturbation penalties to ensure local maximality in data space. This strategy allows the model to capture and preserve the metastable structure inherent in complex physical systems, rather than simply fitting the data distribution in a likelihood sense.

Our design is inspired in part by recent advances in multimodal flow-based generative modeling, such as the bounded KRNet architecture introduced by Peng et al. Peng et al. (2023), which demonstrated that introducing structural constraints in the latent space can significantly improve the accuracy and interpretability of normalizing flows. Building on this line of thinking, our model introduces a Mixture-of-Gaussians (MoG) latent prior not merely for greater flexibility, but to explicitly encode multiple metastable modes that reflect the complex landscape of molecular systems. Each Gaussian component captures a different region of the latent space, which is then mapped—via a KRNet transformation—into a distinct metastable basin in data space. To enforce this mode-to-mode correspondence, we introduce gradient penalties to drive the log-density gradient toward zero at each latent mode, and contrastive perturbation losses to ensure that these mapped points are true maxima in the data space.

Beyond improving generative performance, this design also enhances scientific interpretability. The multimodal latent structure enables soft clustering of generated configurations, where each mode

can be associated with meaningful collective variables (CVs) such as torsion angles, end-to-end distances, or radius of gyration. This provides insight into the system's metastable organization and helps identify the slow reaction coordinates that govern long-time dynamics. In summary, our MoG-based KRNet formulation introduces a novel framework for aligning latent and data-space maxima, enabling both accurate density modeling and interpretable discovery of metastable structure in high-dimensional molecular data.

4.2 Method

4.2.1 Overview of the MoG-KRnet Framework

MoG-KRnet is a bijective generative model that constructs an invertible mapping $f : \mathbb{R}^d \to \mathbb{R}^d$ transforming samples from a hybrid latent distribution p_Z to the data distribution p_X. The key idea is to approximate a transport map that rearranges a tractable base measure into a complex, potentially multimodal data distribution. For a given observation $\mathbf{x} \in \mathbb{R}^d$, the model defines the log-density through the change-of-variable formula:

$$\log p_X(\mathbf{x}) = \log p_Z(f(\mathbf{x})) + \log |\det J_f(\mathbf{x})|,$$

where $J_f(\mathbf{x}) = \nabla_{\mathbf{x}} f(\mathbf{x}) \in \mathbb{R}^{d\times d}$ denotes the Jacobian of f. This formulation permits exact likelihood evaluation and allows optimization via maximum likelihood estimation.

4.2.2 Hybrid Latent Prior

The latent variable $\mathbf{z} \in \mathbb{R}^d$ is decomposed into two independent blocks, denoted $\mathbf{z}_1 \in \mathbb{R}^{d_1}$ and $\mathbf{z}_2 \in \mathbb{R}^{d_2}$, with $d_1 + d_2 = d$. The first block $\mathbf{z}_1$ follows a product of one-dimensional mixture-of-Gaussians:

$$p(\mathbf{z}_1) = \prod_{j=1}^{d_1} \sum_{k=1}^{K_j} \pi_{j,k} \cdot \mathcal{N}(z_j; \mu_{j,k}, \sigma_{j,k}^2),$$

while the second block $\mathbf{z}_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_2})$ is standard Gaussian. This hybrid prior combines multimodal expressiveness with analytical tractability and defines the target measure for the flow map.

4.2.3 KRnet Architecture as Progressive Triangular Transport

Inspired by the Knothe–Rosenblatt rearrangement, MoG-KRnet factorizes the transformation f into a sequence of stage-wise maps:

$$f = f^{(K)} \circ f^{(K-1)} \circ \cdots \circ f^{(1)},$$

where each stage $f^{(k)}$ updates a block of coordinates while conditioning on preceding ones, approximating triangular transport structure. Each $f^{(k)}$ is implemented as a composition of transformations:

$$f^{(k)} = \mathcal{S}^{(k)} \circ \mathcal{A}^{(k)} \circ \mathcal{N}^{(k)} \circ \mathcal{R}^{(k)},$$

where $\mathcal{R}^{(k)}$ is a linear transformation with learnable LU structure, $\mathcal{N}^{(k)}$ is an actnorm layer that ensures zero-mean and unit variance per dimension (with learnable parameters), $\mathcal{A}^{(k)}$ is a stack of affine coupling layers (described in Section 3.4), and $\mathcal{S}^{(k)}$ performs a squeezing operation that reallocates dimension usage over stages.

This layered composition ensures that the full Jacobian J_f remains triangular or block-triangular, allowing for efficient computation of $\log|\det J_f|$ as a sum over the individual layers.

4.2.4 Affine Coupling Transformations

Each affine coupling layer partitions the input $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2]$, and updates $\mathbf{x}_2$ using a scale-and-shift transformation conditioned on $\mathbf{x}_1$:

$$\mathbf{x}_2' = \mathbf{x}_2 \odot (1 + \alpha \cdot \tanh(s(\mathbf{x}_1))) + \gamma \cdot \tanh(t(\mathbf{x}_1)).$$

Here, s and t are neural networks; $\alpha \in (0, 1)$ is a fixed stability parameter (e.g., 0.6); $\gamma \in \mathbb{R}^{d'}$ is a learnable global vector. The inverse transformation is analytically computable, ensuring exact invertibility. The log-determinant of the Jacobian for each coupling layer is efficiently computed as:

$$\log|\det J_{\mathcal{A}}| = \sum_{i=1}^{d'} \log\left(1 + \alpha \cdot \tanh(s_i(\mathbf{x}_1))\right),$$

which contributes additively to the total log-likelihood.

4.2.5 Mode Alignment Regularization

We introduce a geometric regularization mechanism to promote semantic alignment between the latent and data spaces. Let $\mathbf{z}_{\max} \in \mathbb{R}^d$ be the point in latent space corresponding to the global mode of the prior. We define it as the concatenation of the mean of the dominant mixture components in $\mathbf{z}_1$, and the zero vector in $\mathbf{z}_2$:

$$\mathbf{z}_{\max} = [\mu_1^*, \ldots, \mu_{d_1}^*, 0, \ldots, 0],$$

where μ_j^* is the mean of the most probable component for dimension j. We compute the corresponding data-space mode as $\mathbf{x}_{\max} = f^{-1}(\mathbf{z}_{\max})$.

To encourage $\mathbf{x}_{\max}$ to align with a mode of the data distribution, we introduce two regularization terms. The first is a gradient penalty:

$$\mathcal{L}_{\text{grad}} = \left\| \nabla_{\mathbf{x}} \log p_X(\mathbf{x}) \big|_{\mathbf{x}=\mathbf{x}_{\max}} \right\|^2,$$

which encourages stationarity of the log-density at $\mathbf{x}_{\max}$. The second is a local contrastive penalty defined over perturbed neighborhoods:

$$\mathcal{L}_{\text{contrast}} = \frac{1}{M} \sum_{i=1}^{M} \max\left(0, \log p_X(\mathbf{x}_i^{\text{neigh}}) - \log p_X(\mathbf{x}_{\max})\right),$$

where $\mathbf{x}_i^{\text{neigh}} = \mathbf{x}_{\max} + \epsilon_i$, and $\epsilon_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$. This enforces that $\mathbf{x}_{\max}$ is not only a critical point, but a local maximum.

4.2.6 Objective Function

The total loss function used for training is the sum of the negative log-likelihood and the mode-alignment penalties:

$$\mathcal{L} = -\mathbb{E}_{\mathbf{x}\sim\mathcal{D}} \log p_X(\mathbf{x}) + \lambda_{\text{grad}} \cdot \mathcal{L}_{\text{grad}} + \lambda_{\text{contrast}} \cdot \mathcal{L}_{\text{contrast}}.$$

Here, $\lambda_{\text{grad}}, \lambda_{\text{contrast}} \geq 0$ control the strength of the regularization terms.

4.2.7 Sampling and Inference

For density evaluation, an input $\mathbf{x} \in \mathbb{R}^d$ is mapped through the flow to obtain $\mathbf{z} = f(\mathbf{x})$, and the log-likelihood is computed via:

$$\log p_X(\mathbf{x}) = \log p_Z(\mathbf{z}) + \sum_{k=1}^{K} \log |\det J_{f^{(k)}}(\cdot)|.$$

Each sub-map contributes a triangular Jacobian, so the determinant is computed in linear time. The prior log-density is decomposed as:

$$\log p_Z(\mathbf{z}) = \sum_{j=1}^{d_1} \log\left(\sum_{k=1}^{K_j} \pi_{j,k} \cdot \mathcal{N}(z_j; \mu_{j,k}, \sigma_{j,k}^2)\right) - \frac{1}{2}\|\mathbf{z}_2\|^2 - \frac{d_2}{2}\log(2\pi).$$

To generate samples, latent vectors $\mathbf{z} \sim p_Z$ are drawn by sampling each MoG dimension z_j from its categorical mixture and the Gaussian block from standard normal. The resulting $\mathbf{z}$ is passed through the inverse flow $\mathbf{x} = f^{-1}(\mathbf{z})$, which is exact and fully differentiable.

4.3 Numerical Result

4.3.1 Approximation of the Müller-Brown Equilibrium Distribution

We first evaluate the capacity of MoG-KRnet to approximate a complex, multimodal target distribution arising from the well-known Müller-Brown potential—a classical benchmark in molecular simulation that features multiple metastable wells separated by high-energy barriers. The target equilibrium distribution is the Boltzmann-Gibbs measure:

$$p_X(\mathbf{x}) = \frac{1}{Z}\exp\left(-\frac{U(\mathbf{x})}{k_B T}\right),$$

where $U(\mathbf{x})$ denotes the potential energy, $T = 30K$, and $\mathbf{x} \in \mathbb{R}^2$. We simulate overdamped Langevin dynamics,

$$\frac{d\mathbf{x}}{dt} = -\nabla U(\mathbf{x}) + \sqrt{2k_B T} \cdot \boldsymbol{\xi}(t),$$

to generate a reference dataset of 5 million samples. These samples serve as empirical draws from the true equilibrium distribution p_X.

The model is instantiated as a single-stage KRnet with depth $D = 64$, input dimension $d = 2$, and coupling width 256. Each flow stage includes alternating affine coupling layers with learnable LU-based linear transformations and interleaved squeezing operations. The coupling layers implement:

$$\mathbf{z}_2 = \mathbf{z}_2 \odot (1 + \alpha \cdot \tanh(s(\mathbf{z}_1))) + \gamma \cdot \tanh(t(\mathbf{z}_1)),$$

with $\alpha = 0.6$, and s, t being two-layer neural networks with ReLU activations. The latent prior $p_Z(\mathbf{z})$ consists of a product of a one-dimensional mixture-of-Gaussians:

$$p(z_1) = \sum_{k=1}^{2} \pi_k \cdot \mathcal{N}(z_1; \mu_k, \sigma_k^2), \quad \pi = [0.6, 0.4], \quad \mu = [-2, 2], \quad \sigma = [1.0, 0.5],$$

and a standard Gaussian $z_2 \sim \mathcal{N}(0, 1)$.

In addition to maximum likelihood estimation, we introduce geometric regularization to ensure that the high-density region of the latent prior is mapped to a corresponding mode in the target space. This is done via two penalty terms:

1. A local contrastive penalty encourages log-probability at mapped prior mode $\mathbf{x}_{\max}$ to exceed its local neighbors:

$$\mathcal{L}_{\text{contrast}} = \sum_{i=1}^{M} \max\left(0, \log p_X(\mathbf{x}_i^{\text{neigh}}) - \log p_X(\mathbf{x}_{\max})\right),$$

where $\mathbf{x}_i^{\text{neigh}} = \mathbf{x}_{\max} + \epsilon_i$, and $\epsilon_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$.

2. A gradient penalty enforces $\nabla \log p_X(\mathbf{x}_{\max}) \approx 0$:

$$\mathcal{L}_{\text{grad}} = \left\| \nabla_{\mathbf{x}} \log p_X(\mathbf{x}) \big|_{\mathbf{x}=\mathbf{x}_{\max}} \right\|_1.$$

The total objective is:

$$\mathcal{L} = -\mathbb{E}_{\mathbf{x}\sim\mathcal{D}} \left[\log p_X(\mathbf{x})\right] + \lambda_{\text{contrast}} \mathcal{L}_{\text{contrast}} + \lambda_{\text{grad}} \mathcal{L}_{\text{grad}}$$

Training is performed using the Adam optimizer with an initial learning rate of 1×10^{-4}, which is decayed by 10% every 5000 iterations. Penalty coefficients $\lambda_{\text{contrast}}$, λ_{grad}, and λ_{align} are initialized at small values and gradually increased (by 5% every 5000 iterations) to guide the flow toward stable mode alignment while preventing instability in early optimization.

Minibatches of 2D coordinates are drawn from preprocessed subsets $\{\texttt{set}_j\}_{j=0}^{19}$, and the entire training loop runs for 100,000 steps. Gradients are clipped using global norm clipping with a threshold of 0.01. Checkpoints are saved regularly and used for restarting long runs.

Figure 4.1 shows the learned transport map learned by MoG-KRnet. On the left, samples drawn from the latent space exhibit two clearly separated high-density regions corresponding to distinct

components in the mixture-of-Gaussians prior. After training, these two latent modes are transported via the inverse flow f^{-1} into two distinct basins of the Müller-Brown energy landscape, shown on the right.

This demonstrates that the model not only captures the overall multimodal structure of the target distribution but also learns a semantically consistent transport: each latent mode is mapped to a specific metastable state of the physical system. The smoothness and separation of the transformed samples reflect that the triangular KRnet flow—coupled with the hybrid prior and geometric regularization—successfully avoids mode collapse and learns a one-to-one correspondence between latent and physical modes.

Such mode-resolving behavior is difficult to achieve with standard normalizing flows that rely on unimodal priors. In contrast, MoG-KRnet leverages the flexibility of mixture components to assign and map separated probability mass to different energetic regions in a physically meaningful way. This mode alignment improves both sampling fidelity and interpretability, especially in systems where multiple competing basins dominate the dynamics.

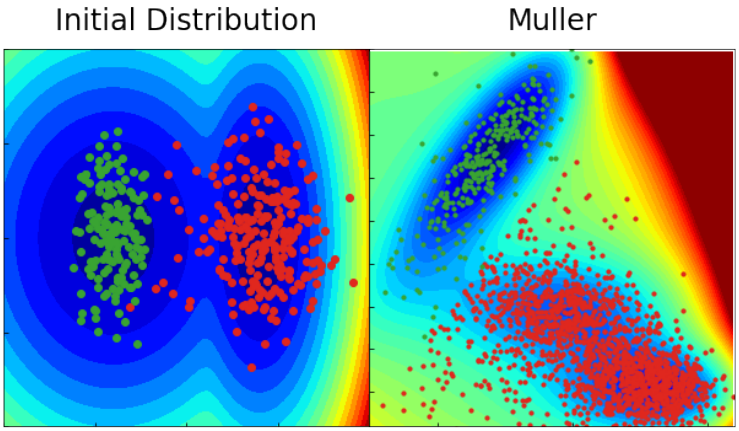


Figure 4.1 Learned transport from latent space to physical space using MoG-KRnet. The left half of the image shows samples drawn from the latent distribution, which has two distinct high-density regions due to the mixture-of-Gaussians prior. The right half shows the transformed samples under the inverse flow f^{-1}, which accurately maps the two latent modes into the two metastable basins of the Müller-Brown potential. This demonstrates the model's ability to perform semantically aligned mode separation, transporting distinct regions of latent mass to physically meaningful targets.

4.3.2 Approximation of the Alanine Dipeptide Equilibrium Distribution

To further assess the scalability and generalization capacity of MoG-KRnet, we apply it to approximate the equilibrium distribution of a higher-dimensional molecular system: alanine dipeptide in implicit solvent. This molecule is a well-known testbed in molecular simulation due to its low dimensionality yet rich conformational landscape, characterized by transitions between metastable states in the Ramachandran (ϕ, ψ)-angle space and the full Cartesian coordinates of selected atoms.

We generate reference samples for alanine dipeptide (Ace-Ala-Nme, commonly referred to as Ala2) via full-atom molecular dynamics (MD) simulation in explicit solvent. The simulation system consists of the alanine dipeptide molecule immersed in 383 TIP3P water molecules. The simulation is carried out at 350 K using the Amber99-SB force field and Langevin dynamics for temperature control. A time step of 2.5×10^{-4} ps is used for numerical integration.

Trajectories are collected from equilibrium simulations and projected onto a reduced coordinate space consisting of Cartesian positions of selected heavy atoms. In total, 5 million configurations are used to construct the dataset $\mathcal{D}_{\text{ala}} \sim p_X$, representing the high-dimensional equilibrium distribution over molecular conformations.

The MoG-KRnet model is constructed to map the equilibrium distribution of alanine dipeptide into a structured latent space. The model input is a 15-dimensional vector representing the Cartesian coordinates of five key atoms—[5, 7, 9, 15, 17]—selected to capture relevant backbone conformational fluctuations while avoiding redundant degrees of freedom. This atom selection encompasses chemically meaningful internal coordinates, including both ϕ and ψ torsions, as well as spatial end-to-end geometry.

The flow transformation $f : \mathbb{R}^{15} \to \mathbb{R}^{15}$ consists of 7 staged mappings inspired by the pseudo-triangular structure of the Knothe–Rosenblatt rearrangement. Each stage contains 24 affine coupling layers with hidden width 128, supplemented by actnorm, LU-based rotations, and squeezing operations. This progressive architecture allows early layers to resolve nonlinear, multimodal features in dominant subspaces, while later layers refine global geometry.

The latent prior p_Z is a hybrid of three independent one-dimensional mixture-of-Gaussians (MoG) and a standard multivariate Gaussian:

$$p_Z(\mathbf{z}) = \prod_{j=1}^{3} \sum_{k=1}^{K_j} \pi_{j,k} \cdot \mathcal{N}(z_j; \mu_{j,k}, \sigma_{j,k}^2) \cdot \mathcal{N}(\mathbf{z}_{4:15}; \mathbf{0}, \mathbf{I}_{12}).$$

This design reflects the assumption that a small number of latent coordinates capture discrete conformational transitions (e.g., basin-hopping), while the remaining degrees of freedom reflect continuous fluctuations in local structure. The independence of latent dimensions facilitates efficient sampling and interpretable mode decomposition. Latent samples are drawn by independently sampling each MoG dimension using categorical selection followed by Gaussian sampling, and appending a standard normal vector for the Gaussian block.

Training is performed on a dataset of 5 million MD samples using a composite loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{NLL}} + \lambda_{\text{grad}}\mathcal{L}_{\text{grad}} + \lambda_{\text{contrast}}\mathcal{L}_{\text{contrast}} + \mathcal{L}_{\text{rep}}.$$

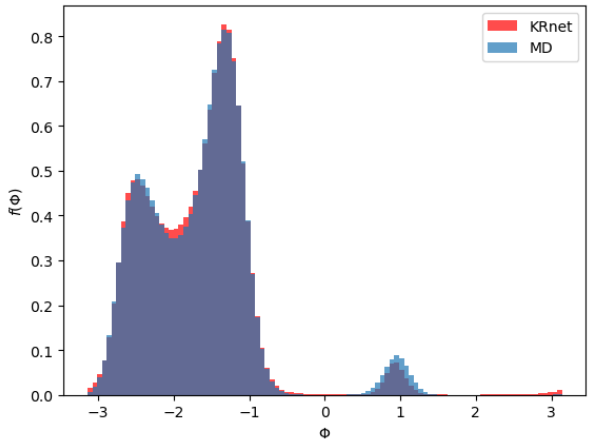
Here, $\mathcal{L}_{\text{NLL}}$ is the negative log-likelihood, $\mathcal{L}_{\text{grad}}$ enforces local maximality at the mapped latent maxima, and $\mathcal{L}_{\text{contrast}}$ ensures neighborhood contrast.

To prevent mode collapse, a pairwise repulsion loss $\mathcal{L}_{\text{rep}}$ is introduced between mapped maxima $\{\mathbf{x}_{\text{max}}^{(i)}\}$ using a soft margin criterion.

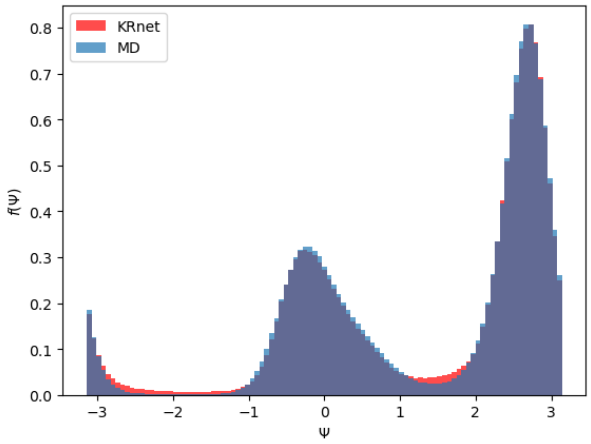
Training runs for 200,000 iterations with the Adam optimizer, an initial learning rate of 10^{-7}, and dynamically scaled penalty weights. During training, KDE diagnostics are periodically used to ensure that generated samples recover the correct distributions in torsional and Euclidean observables.

In Fig. 4.2, we compare the predicted marginal densities of the ϕ and ψ dihedral angles from KRnet to reference MD histograms. MoG-KRnet accurately reproduces all modes in both coordinates and captures the correct relative amplitudes. In particular, the sharp peak near $\psi \approx 3.0$ is learned precisely, and the multimodality in ϕ is preserved without mode collapse.

To assess the learned joint dependencies, we visualize the 2D density over (ϕ, ψ) in Fig. 4.3, along with the eight mode points mapped from latent maxima. The conformational basins of Ala2 are clearly recovered, and the mode points are well-separated, each landing within distinct high-density



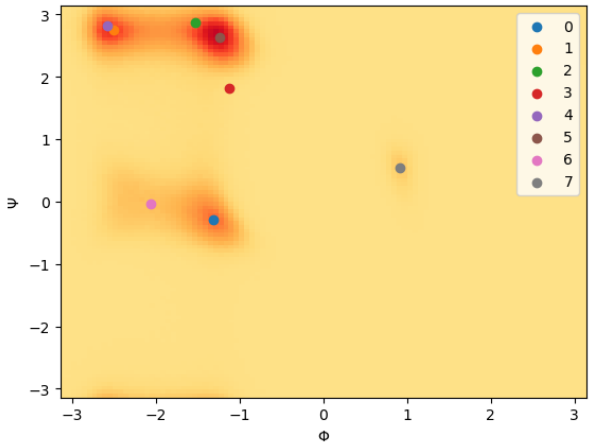
(a) Marginal distribution of ϕ



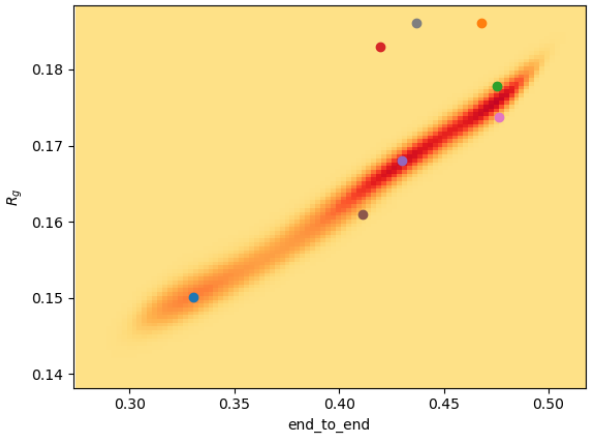
(b) Marginal distribution of ψ

Figure 4.2 Comparison of 1D dihedral angle marginals between KRnet (red) and MD ground truth (blue).

regions. This confirms that MoG-KRnet not only fits the data globally but also identifies meaningful latent structure.



(a) ϕ–ψ dihedral distribution of Ala2. Warm background indicates predicted density; overlaid points mark mapped latent maxima.



(b) End-to-end distance vs. radius of gyration (R_g) from sampled configurations. Latent maxima are projected onto this plane.

Figure 4.3 Predicted equilibrium features of alanine dipeptide from the trained MoG-KRnet model.

An important feature of our MoG-KRnet framework is the explicit mapping of latent density modes to high-probability basins in configuration space. By construction, each mode of the hybrid prior $\mathbf{z}_{\max}^{(i)}$ is mapped through the inverse flow f^{-1} to a corresponding point $\mathbf{x}_{\max}^{(i)} \in \mathbb{R}^d$. These mapped maxima are designed to coincide with peaks in the learned data distribution $p_X(\mathbf{x})$, enforced

during training via gradient and contrastive penalties.

As shown in Figure 4.3, this alignment holds in both the dihedral angle space (ϕ, ψ) and in structural coordinates. In the left panel, each mapped latent mode falls within a distinct conformational basin in the (ϕ, ψ) landscape, indicating that MoG-KRnet captures metastability through a structured latent space. In the right panel, the same latent maxima are distributed across the manifold of end-to-end distance and radius of gyration, further supporting the physical consistency and geometric expressiveness of the learned model. These results confirm that our framework not only generates accurate samples but also provides a semantically meaningful latent representation that aligns with physically interpretable features of the molecular system.

4.4 Summary.

In this work, we proposed MoG-KRnet, a novel flow-based generative framework designed for approximating high-dimensional equilibrium distributions in complex molecular systems. Our approach builds upon the theory of invertible transformations and exploits a staged Knothe–Rosenblatt-inspired architecture to progressively map structured latent representations to physical configuration space. A distinguishing feature of MoG-KRnet is its use of a hybrid latent prior that combines independent one-dimensional mixture-of-Gaussians (MoG) components with standard Gaussian variables. This formulation enables the model to flexibly represent multi-modal distributions while maintaining computational tractability and efficient sampling.

To ensure meaningful correspondence between latent and physical modes, we introduced a mode-alignment strategy during training. This involves constructing the prior such that each latent mode $\mathbf{z}_{\max}^{(i)}$ encodes a distinct peak in the latent density, and then enforcing through loss penalties that each of these modes is mapped to a high-probability region $\mathbf{x}_{\max}^{(i)} = f^{-1}(\mathbf{z}_{\max}^{(i)})$ in the observed space. This is achieved via gradient-based penalties to minimize $\|\nabla_{\mathbf{x}} \log p_X(\mathbf{x}_{\max})\|$, and contrastive penalties to ensure that $\mathbf{x}_{\max}$ is indeed a local maximum compared to its neighbors. This alignment strategy imbues the model with semantic coherence and supports downstream interpretability.

The efficacy of MoG-KRnet was demonstrated on two benchmark systems. For the Müller-Brown potential, we showed that the model accurately captured the bimodal equilibrium distribution and

learned to associate each latent mode with a distinct physical basin. For the more complex alanine dipeptide molecule in explicit solvent, the model was trained on 5 million full-atom MD snapshots and succeeded in learning the joint equilibrium distribution over both angular variables (dihedral angles ϕ, ψ) and structural observables (end-to-end distance and radius of gyration). In all cases, the mapped latent maxima landed squarely in dominant high-density regions of the data space, confirming the success of the mode-to-basin alignment. Furthermore, generated samples from the model reproduced the marginal and joint distributions of key physical features with high fidelity, closely matching empirical histograms derived from MD data.

Together, these contributions underscore the dual strengths of MoG-KRnet: the capacity to approximate complex, multi-modal densities in high dimensions, and the ability to structure the latent space in a physically meaningful and interpretable manner. Our results demonstrate that MoG-KRnet provides not only a powerful generative model but also a principled tool for reduced representation of molecular systems, where the mapping from latent to physical coordinates respects the underlying metastable structure of the dynamics. This makes it particularly well-suited for tasks in coarse-grained modeling, statistical reweighting, and uncertainty-aware exploration of equilibrium configurations.

CHAPTER 5

CONCLUSION

This thesis presents a unified, data-driven framework for constructing reduced-order models of high-dimensional, non-Markovian dynamical systems. By integrating advances in memory-aware modeling and normalizing flow-based latent representations, we address two central challenges in coarse-grained modeling: accurately capturing long-time correlations and resolving complex, multi-modal equilibrium distributions.

We began by developing a novel learning-based approach to non-Markovian stochastic reduced modeling. By augmenting the resolved dynamics with a set of learned auxiliary variables—interpretable as non-Markovian features—we showed that the complex memory effects embedded in full-atom molecular simulations can be faithfully captured without directly estimating memory kernels. This framework builds on the Mori–Zwanzig formalism and circumvents conventional kernel fitting by matching correlation functions in an extended variable space. Numerical results on tagged particles and polymer chains demonstrated excellent agreement with full molecular dynamics (MD) simulations, validating the expressiveness and robustness of the proposed models.

We extended this methodology to incorporate state-dependent memory kernels, thereby enabling more realistic dynamics in systems with heterogeneous free energy landscapes. Our framework captures local variations in unresolved degrees of freedom and accommodates basin-specific relaxation times and noise structures. Through simulations on polymer systems, we observed significant improvements in predictive accuracy and sampling fidelity compared to global or fixed-kernel models.

To handle the challenge of modeling complex equilibrium distributions, we introduced a new Mixture-of-Gaussians (MoG) KRNet architecture—termed KRnet-MoG-GLE—as a probabilistic generative model for full molecular dynamics. By replacing the unimodal latent prior with a flexible MoG prior, our model gains the capacity to represent multiple metastable basins and generate samples that better match empirical distributions. Importantly, we designed the training to enforce mode-alignment, ensuring that the maxima of the latent variables map to high-probability regions in

the data space. This was demonstrated convincingly in the Müller-Brown and alanine dipeptide systems, where our model successfully resolved distinct basins and accurately approximated target observables such as end-to-end distance, radius of gyration, and dihedral angles.

Taken together, the contributions of this work represent a significant step forward in the design of interpretable, memory-embedded, and generative reduced-order models. By marrying the strengths of stochastic modeling with expressive latent-variable architectures, our approach enables efficient exploration and inference in systems that are otherwise intractable due to their high dimensionality and long memory effects.

Ultimately, this thesis lays the foundation for data-driven, physically-consistent reduced models that can serve as scalable surrogates for multi-scale simulations, with broad applicability across molecular biophysics, soft materials, and beyond.

BIBLIOGRAPHY

Ayaz, C., Dalton, B. A., and Netz, R. R. (2022). Generalized langevin equation with a non-linear potential of mean force and non-linear memory friction from a hybrid projection scheme. Physical Review E, 105:054138.

Baczewski, A. D. and Bond, S. D. (2013). Numerical integration of the extended variable generalized Langevin equation with a positive Prony representable memory kernel. The Journal of chemical physics, 139(4):044107.

Barducci, A., Bussi, G., and Parrinello, M. (2008). Well-tempered metadynamics: a smoothly converging and tunable free-energy method. Physical Review Letters, 100(2):020603.

Bittracher, A., Koltai, P., Klus, S., Banisch, R., Dellnitz, M., and Schütte, C. (2018). Transition manifolds of complex metastable systems. Journal of Nonlinear Science, 28(2):471–512.

Bonati, L., Zhang, Y.-Y., and Parrinello, M. (2019). Neural networks-based variationally enhanced sampling. Proceedings of the National Academy of Sciences, 116(36):17641–17647.

Ceriotti, M., Bussi, G., and Parrinello, M. (2009). Langevin equation with colored noise for constant-temperature molecular dynamics simulations. Physical review letters, 102(2):020601.

Chen, M., Li, X., and Liu, C. (2014). Computation of the memory functions in the generalized Langevin models for collective dynamics of macromolecules. J. Chem. Phys., 141:064112.

Chiavazzo, E., Covino, R., Coifman, R. R., Gear, C. W., Georgiou, A. S., Hummer, G., and Kevrekidis, I. G. (2017). Intrinsic map dynamics exploration for uncharted effective free-energy landscapes. Proceedings of the National Academy of Sciences, 114(28):E5494–E5503.

Chorin, A. J., Hald, O. H., and Kupferman, R. (2002). Optimal prediction with memory. Phys. D, 166:239–257.

Chorin, A. J. and Lu, F. (2015). Discrete approach to stochastic parametrization and dimension reduction in nonlinear dynamics. Proceedings of the National Academy of Sciences, 112(32):9804–9809.

Coifman, R. R., Kevrekidis, I. G., Lafon, S., Maggioni, M., and Nadler, B. (2008). Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems. Multiscale Modeling & Simulation, 7(2):842–864.

Corless, M. and Frazho, A. (2003). Linear Systems and Control: An Operator Perspective. Chapman & Hall/CRC Pure and Applied Mathematics. Taylor & Francis.

Crosskey, M. and Maggioni, M. (2017). Atlas: A geometric approach to learning high-dimensional stochastic systems near manifolds. Multiscale Modeling & Simulation, 15(1):110–156.

Daldrop, J. O., Kowalik, B. G., and Netz, R. R. (2017). External potential modifies friction of molecular solutes in water. Physical Review X, 7(4):041065.

Darve, E. and Pohorille, A. (2001). Calculating free energies using average force. The Journal of Chemical Physics, 115(20):9169–9183.

Darve, E., Solomon, J., and Kia, A. (2009a). Computing generalized Langevin equations and generalized fokker-planck equations. Proc. Natl. Acad. Sci., 106(27):10884–10889.

Darve, E., Solomon, J., and Kia, A. (2009b). Computing generalized Langevin equations and generalized Fokker-Planck equations. Proceedings of the National Academy of Sciences, 106(27):10884–10889.

Davtyan, A., Dama, J. F., Voth, G. A., and Andersen, H. C. (2015a). Dynamic force matching: A method for constructing dynamical coarse-grained models with realistic time dependence. J. Chem. Phys., 142.

Davtyan, A., Dama, J. F., Voth, G. A., and Andersen, H. C. (2015b). Dynamic force matching: A method for constructing dynamical coarse-grained models with realistic time dependence. J. Chem. Phys., 142.

de Gennes, P. (1979). Scaling concepts in polymer physics. Cornell Univ. Pr.

Dibak, M., del Razo, M. J., De Sancho, D., Schütte, C., and Noé, F. (2018). Msm/rd: Coupling markov state models of molecular kinetics with reaction-diffusion simulations. The Journal of Chemical Physics, 148(21):214107.

Dinh, L., Krueger, D., and Bengio, Y. (2014). Nice: Non-linear independent components estimation. arXiv preprint arXiv:1410.8516.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2017). Density estimation using real nvp. In International Conference on Learning Representations (ICLR).

Español, P. (2004). Statistical mechanics of coarse-graining. In Novel Methods in Soft Matter Simulations, pages 69–115. Springer.

Fang, L., Ge, P., Zhang, L., E, W., and Lei, H. (2022). DeePN^2: A deep learning-based non-Newtonian hydrodynamic model. Journal of Machine Learning, 1:114–140.

Feng, L., Gao, T., Dai, M., and Duan, J. (2022). Auto-sde: Learning effective reduced dynamics from data-driven stochastic dynamical systems. arXiv preprint:arXiv:2205.04151.

Frenkel, D. and Smit, B. (2001). Understanding molecular simulation: from algorithms to applications, volume 1. Elsevier.

Ge, P., Zhang, Z., and Lei, H. (2024). Data-driven learning of the generalized langevin equation with state-dependent memory. Physical Review Letters, 133(7):077301.

Giannakis, D. (2019). Data-driven spectral decomposition and forecasting of ergodic dynamical systems. Applied and Computational Harmonic Analysis, 47(2):338–396.

Grogan, F., Lei, H., Li, X., and Baker, N. A. (2020). Data-driven molecular modeling with the generalized langevin equation. J. Comput. Phys., 418:109633–109641.

Harlim, J., Jiang, S. W., Liang, S., and Yang, H. (2020). Machine learning for prediction with missing dynamics. Journal of Computational Physics, page 109922.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8):1735–1780.

Hudson, T. and Li, X. H. (2020). Coarse-Graining of Overdamped Langevin Dynamics via the Mori–Zwanzig Formalism. Multiscale Modeling & Simulation, 18(2):1113–1135.

Jung, G., Hanke, M., and Schmid, F. (2017a). Iterative reconstruction of memory kernels. Journal of Chemical Theory and Computation, 13(6):2481–2488.

Jung, G., Hanke, M., and Schmid, F. (2017b). Iterative reconstruction of memory kernels. Journal of Chemical Theory and Computation, 13(6):2481–2488.

Kingma, D. and Ba, J. (2015). Adam: A method for stochastic optimization. International Conference on Learning Representations (ICLR).

Kingma, D. P. and Dhariwal, P. (2018). Glow: Generative flow with invertible 1x1 convolutions. In Advances in neural information processing systems (NeurIPS).

Klippenstein, V., Tripathy, M., Jung, G., Schmid, F., and van der Vegt, N. F. (2021). Introducing memory in coarse-grained molecular simulations. The Journal of Physical Chemistry B, 125(19):4931–4954.

Klus, S., Nüske, F., Koltai, P., Wu, H., Kevrekidis, I., Schütte, C., and Noé, F. (2018). Data-driven model reduction and transfer operator approximation. Journal of Nonlinear Science, 28(3):985–1010.

Klus, S., Nüske, F., Peitz, S., Niemann, J.-H., Clementi, C., and Schütte, C. (2020). Data-driven approximation of the koopman generator: Model reduction, system identification, and control. Physica D: Nonlinear Phenomena, 406:132416.

Koopman, B. O. (1931). Hamiltonian systems and transformation in Hilbert space. Proceedings of the National Academy of Sciences, 17(5):315–318.

Kowalik, B., Daldrop, J. O., Kappler, J., Schulz, J. C., Schlaich, A., and Netz, R. R. (2019). Memory-kernel extraction for different molecular solutes in solvents of varying viscosity in confinement. Physical Review E, 100(1):012126.

Krivov, S. V. (2013). On reaction coordinate optimality. Journal of Chemical Theory and Computation, 9(1):135–146.

Kubo, R. (1966). The fluctuation-dissipation theorem. Reports on Progress in Physics, 29(1):255–284.

Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H., and Kollman, P. A. (1992). THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. Journal of Computational Chemistry, 13(8):1011–1021.

Laio, A. and Parrinello, M. (2002a). Escaping free-energy minima. Proceedings of the National Academy of Sciences, 99(20):12562–12566.

Laio, A. and Parrinello, M. (2002b). Escaping free-energy minima. Proceedings of the National Academy of Sciences, 99(20):12562–12566.

Lange, O. F. and Grubmüller, H. (2006). Collective Langevin dynamics of conformational motions in proteins. The Journal of Chemical Physics, 124(21):214903.

Lange, O. F. and Grubmüller, H. (2006). Collective Langevin dynamics of conformational motions in proteins. J. Chem. Phys., 124:214903.

Lee, H. S., Ahn, S.-H., and Darve, E. F. (2019a). The multi-dimensional generalized Langevin equation for conformational motion of proteins. The Journal of Chemical Physics, 150(17):174113.

Lee, H. S., Ahn, S.-H., and Darve, E. F. (2019b). The multi-dimensional generalized Langevin equation for conformational motion of proteins. The Journal of Chemical Physics, 150(17):174113.

Lei, H., Baker, N. A., and Li, X. (2016a). Data-driven parameterization of the generalized Langevin equation. Proceedings of the National Academy of Sciences, 113(50):14183–14188.

Lei, H., Baker, N. A., and Li, X. (2016b). Data-driven parameterization of the generalized Langevin equation. Proceedings of the National Academy of Sciences, 113(50):14183–14188.

Lei, H. and Li, X. (2021). Petrov–Galerkin methods for the construction of non-Markovian dynamics preserving nonlocal statistics. The Journal of Chemical Physics, 154(18):184108.

Lei, H., Wu, L., and E, W. (2020). Machine learning based non-Newtonian fluid model with molecular fidelity. Physical Review E, 102:043309.

Li, W. and Ma, A. (2014). Recent developments in methods for identifying reaction coordinates.

Molecular Simulation, 40(10-11):784–793.

Lin, K. K. and Lu, F. (2021). Data-driven model reduction, Wiener projections, and the Koopman-Mori-Zwanzig formalism. Journal of Computational Physics, 424:109864.

Lu, J. and Vanden-Eijnden, E. (2014). Exact dynamical coarse-graining without time-scale separation. The Journal of Chemical Physics, 141(4):044109.

Lyu, L. and Lei, H. (2023a). Consensus-based construction of high-dimensional free energy surface. arXiv preprint arXiv:2311.05009.

Lyu, L. and Lei, H. (2023b). Construction of coarse-grained molecular dynamics with many-body non-markovian memory. Physical Review Letters, 131(17):177301.

Ma, C., Wang, J., and E, W. (2018). Model reduction with memory and the machine learning of dynamical systems. Communications in Computational Physics, 25(4):947–962.

Ma, L., Li, X., and Liu, C. (2019). Coarse-graining langevin dynamics using reduced-order techniques. Journal of Computational Physics, 380:170–190.

Maragliano, L. and Vanden-Eijnden, E. (2006). A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations. Chemical Physics Letters, 426(1):168 – 175.

Maragliano, L. and Vanden-Eijnden, E. (2008). Single-sweep methods for free energy calculations. The Journal of Chemical Physics, 128(18):184110.

Mori, H. (1965a). A continued-fraction representation of the time-correlation functions. Prog. Theor. Phys., 34:399–416.

Mori, H. (1965b). Transport, collective motion, and Brownian motion. Progress of Theoretical Physics, 33(3):423–455.

Noid, W. G., Chu, J.-W., Ayton, G. S., Krishna, V., Izvekov, S., Voth, G. A., Das, A., and Andersen, H. C. (2008). The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. J. Chem. Phys., 128(24):244114.

Peng, J., Wu, H., Zhou, Z., and Nie, Q. (2023). Bounded krnet: Density estimation for bounded data with normalizing flows. arXiv preprint arXiv:2305.09063.

Pérez-Hernández, G., Paul, F., Giorgino, T., De Fabritiis, G., and Noé, F. (2013). Identification of slow molecular order parameters for markov model construction. The Journal of Chemical Physics, 139(1):015102.

Price, J., Meuris, B., Shapiro, M., and Stinis, P. (2021). Optimal renormalization of multiscale

systems. Proceedings of the National Academy of Sciences, 118(37):e2102266118.

Rohrdanz, M. A., Zheng, W., Maggioni, M., and Clementi, C. (2011). Determination of reaction coordinates via locally scaled diffusion map. The Journal of Chemical Physics, 134(12):124116.

Rosso, L., Minàry, P., Zhu, Z., and Tuckerman, M. E. (2002). On the use of the adiabatic molecular dynamics technique in the calculation of free energy profiles. The Journal of Chemical Physics, 116(11):4389–4402.

Russo, A., Durán-Olivencia, M. A., Kevrekidis, I. G., and Kalliadasis, S. (2019). Deep learning as closure for irreversible processes: A data-driven generalized Langevin equation. arXiv preprint arXiv:1903.09562.

Shaffer, P., Valsson, O., and Parrinello, M. (2016). Enhanced, targeted sampling of high-dimensional free-energy landscapes using variationally enhanced sampling, with an application to chignolin. Proceedings of the National Academy of Sciences, 113(5):1150–1155.

She, Z., Ge, P., and Lei, H. (2023). Data-driven construction of stochastic reduced dynamics encoded with non-markovian features. Journal of Chemical Physics, 158(3):034102.

Stinis, P. (2015). Renormalized Mori-Zwanzig-reduced models for systems without scale separation. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 471(2176):20140446.

Torrie, G. M. and Valleau, J. P. (1977). Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. Journal of Computational Physics, 23(2):187–199.

Valsson, O. and Parrinello, M. (2014). Variational approach to enhanced sampling and free energy calculations. Phys. Rev. Lett., 113:090601.

Vlachas, P. R., Byeon, W., Wan, Z. Y., Sapsis, T. P., and Koumoutsakos, P. (2018). Data-driven forecasting of high-dimensional chaotic systems with long short-term memory networks. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 474(2213):20170844.

Vroylandt, H., Goudenège, L., Monmarché, P., Pietrucci, F., and Rotenberg, B. (2022). Likelihood-based non-markovian models from molecular dynamics. Proceedings of the National Academy of Sciences, 119(13):e2117586119.

Wall, H. (1948). Analytic Theory of Continued Fractions. Analytic Theory of Continued Fractions. D. Van Nostrand Company.

Wang, Q., Ripamonti, N., and Hesthaven, J. S. (2020a). Recurrent neural network closure of parametric POD-Galerkin reduced-order models based on the Mori-Zwanzig formalism. Journal of Computational Physics, 410:109402.

Wang, Q., Ripamonti, N., and Hesthaven, J. S. (2020b). Recurrent neural network closure of parametric POD-Galerkin reduced-order models based on the Mori-Zwanzig formalism. Journal of Computational Physics, 410:109402.

Wang, S., Ma, Z., and Pan, W. (2020c). Data-driven coarse-grained modeling of polymers in solution with structural and dynamic properties conserved. Soft Matter, 16(36):8330–8344.

Xie, P. and E, W. (2024). Coarse-graining conformational dynamics with multidimensional generalized langevin equation: How, when, and why. Journal of Chemical Theory and Computation.

Ye, F. X. F., Yang, S., and Maggioni, M. (2021). Nonlinear model reduction for slow-fast stochastic systems near manifolds. arXiv preprint:arXiv:2104.02120.

Zhu, Y., Tang, Y.-H., and Kim, C. (2022). Learning stochastic dynamics with statistics-informed neural network. arXiv preprint: arXiv:2202.12278.

Zhu, Y. and Venturi, D. (2018). Faber approximation of the Mori-Zwanzig equation. Journal of Computational Physics, 372:694–718.

Zhu, Y. and Venturi, D. (2020). Generalized langevin equations for systems with local interactions. Journal of Statistical Physics, 178:1217.

Zieliński, P. and Hesthaven, J. S. (2022). Discovery of slow variables in a class of multiscale stochastic systems via neural networks. Journal of Nonlinear Science, 32(4):51.

Zwanzig, R. (1973). Nonlinear generalized Langevin equations. Journal of Statistical Physics, 9(3):215–220.

Zwanzig, R. (2001). Nonequilibrium Statistical Mechanics. Oxford University Press.

APPENDIX A

MICROSCALE MODEL OF THE POLYMER MOLECULE

The polymer molecule is modeled as a bead-spring chain consisting of 4 sub-units. Each sub-unit consists of 4 atoms. The full potential is given by

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

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

$$V_{\mathrm{p}}(Q) = \begin{cases} 4\varepsilon\left[\left(\frac{\sigma}{Q}\right)^{12} - \left(\frac{\sigma}{Q}\right)^{6}\right] - 4\varepsilon\left[\left(\frac{\sigma}{Q_c}\right)^{12} - \left(\frac{\sigma}{Q_c}\right)^{6}\right], & Q < Q_c \\ 0, & Q \geq Q_c \end{cases} \tag{A.2}$$

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

$$V_{\mathrm{b}}(l) = -\frac{k_s}{2} l_0^2 \log\left[1 - \frac{l^2}{l_0^2}\right], \tag{A.3}$$

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

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

Table A.1 Parameters of the FENE bond interactions.

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

$$V_{\mathrm{a}}(\theta) = \frac{k_a}{2}\left(\theta - \theta_0\right)^2, \tag{A.4}$$

Type	k_a	θ_0
1	1.2	114.0
2	1.5	119.7

Table A.2 Parameters of the harmonic angle interaction.

where 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, 4-5-6) are imposed by type-2 potential. The detailed parameter set is given by Tab. A.2.

The dihedral potential V_{d} is modeled by the multiharmonic dihedral potential

$$V_{\mathrm{d}}(\phi) = \sum_{i=1}^{6} A_n \cos^{(n-1)}(\phi), \tag{A.5}$$

where two different types. Type-1 dihedral potential is imposed to dihedral angles formed by 2-3-4-5, 4-5-6-7, $\cdots$. Type-2 dihedral potential is imposed to dihedral angles formed by 3-4-5-6, 7-8-9-10, $\cdots$. The detailed parameter set is given by Tab. A.3.

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

Table A.3 Parameters of the multiharmonic dihedral interaction.

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 $U(\mathbf{m}q)$ for the four-dimensional resolved variables $\mathbf{m}q$ defined by (2.18), we conduct the restraint molecular dynamics simulation to sample the average force. Specifically, for each target configuration $\mathbf{m}q^*$, we impose a biased quadratic potential $U_{\mathrm{b}ias}(\mathbf{m}q, \mathbf{m}q^*)$ by

$$U_{\mathrm{b}ias}(\mathbf{m}q, \mathbf{m}q^*) = \frac{1}{2}\sum_{i=1}^{4} k_i \left(q_i - q_i^*\right)^2, \tag{B.1}$$

where $k_1, \cdots, k_4$ 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 $\left\{\mathbf{m}Q_j\right\}_{j=1}^{N}$ is given by

$$\mathbf{m}F_{\mathrm{b}ias}(\mathbf{m}q, \mathbf{m}q^*) = -\sum_{i=1}^{4} k_i \left(q_i - q_i^*\right) \nabla_{\mathbf{m}Q_j} q_i, \tag{B.2}$$

where the gradient terms are given by

$$\begin{aligned}
\nabla_{\mathbf{m}Q_j} q_1 &= \frac{\mathbf{m}Q_1 - \mathbf{m}Q_N}{q_1}\delta_{j,1} + \frac{\mathbf{m}Q_N - \mathbf{m}Q_1}{q_1}\delta_{j,N}, \\
\nabla_{\mathbf{m}Q_j} q_2 &= \frac{2\left(\mathbf{m}Q_j - \mathbf{m}Q_c\right)}{Nq_2}, \\
\nabla_{\mathbf{m}Q_j} q_3 &= \frac{\mathbf{m}Q_1 - \mathbf{m}Q_{\lfloor \frac{N}{2} \rfloor}}{q_3}\delta_{j,1} + \frac{\mathbf{m}Q_{\lfloor \frac{N}{2} \rfloor} - \mathbf{m}Q_1}{q_3}\delta_{j,\lfloor \frac{N}{2} \rfloor}, \\
\nabla_{\mathbf{m}Q_j} q_4 &= \frac{\mathbf{m}Q_N - \mathbf{m}Q_{\lceil \frac{N}{2} \rceil}}{q_4}\delta_{j,N} + \frac{\mathbf{m}Q_{\lceil \frac{N}{2} \rceil} - \mathbf{m}Q_N}{q_4}\delta_{j,\lceil \frac{N}{2} \rceil},
\end{aligned} \tag{B.3}$$

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

The free energy $U(\mathbf{m}q)$ is approximated by a 4-layer fully connected neural network $\tilde{U}(\mathbf{m}q)$. Each hidden layer has 160 neurons; hyperbolic tangent function is used as the activation function. $\tilde{U}(\mathbf{m}q)$ is trained by minimizing the empirical loss

$$L = \sum_{k=1}^{N_s} \left\| -\nabla_{\mathbf{m}q^{(k)}} \tilde{U}(\mathbf{m}q) - \mathbf{m}F_{\mathrm{b}ias}(\mathbf{m}q, \mathbf{m}q^{(k)}) \right\|^2, \tag{B.4}$$

where $\mathbf{m}q^{(k)}$ represents a sampled configuration. In this work, we construct $\tilde{U}(\mathbf{m}q)$ using $N_s = 400000$ sample points collected from a simulation with a production stage of 1×10^7 steps.

For each configuration, the number of step 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 $\tilde{U}(\mathbf{m}q)$, we numerically evaluate the integration

$$k_B T \mathbf{m} I \equiv \int \mathbf{m}q \otimes \nabla U(\mathbf{m}q) \mathrm{e}^{-U(\mathbf{m}q)/k_B T} \mathrm{d}\mathbf{m}q \Big/ \int \mathrm{e}^{-U(\mathbf{m}q)/k_B T} \mathrm{d}\mathbf{m}q \approx \frac{1}{N_s} \sum_{k=1}^{N_s} \mathbf{m}q^{(k)} \otimes \nabla \tilde{U}(\mathbf{m}q^{(k)}). \tag{B.5}$$

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

$$\frac{1}{N_s} \sum_{k=1}^{N_s} \mathbf{m}q^{(k)} \otimes \nabla \tilde{U}(\mathbf{m}q^{(k)}) = \begin{bmatrix} 1.0362 & -0.0011 & 0.0087 & 0.0062 \\ 0.0094 & 0.9814 & 0.0021 & 0.0018 \\ 0.0096 & 0.0068 & 0.9913 & -0.0020 \\ 0.0076 & 0.0098 & 0.0008 & 0.9913 \end{bmatrix}, \tag{B.6}$$

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

APPENDIX C

FLUCTUATION-DISSIPATION THEOREM OF THE EXTENDED DYNAMICS

For the extended dynamics in form of Eqs. (2.5)(2.6), we can show that the embedded memory kernel $\tilde{\mathbf{m}}\theta(t)$ and fluctuation term $\tilde{\mathbf{m}}\mathcal{R}(t)$ satisfy the second-fluctuation dissipation theorem. Without loss of generality, we set the covariance of the non-Markovian features to be $k_B T\mathbf{m}I$ following the learning method presented in Sec. 2.2.3, i.e., $\mathbf{m}\Lambda = \mathbf{m}I$, $\tilde{\mathbf{m}}J = \mathbf{m}J$.

Proposition C.0.1. *The embedded memory kernel of the extended dynamics* (2.5)(2.6) *takes the form* $\tilde{\mathbf{m}}\theta(t) = -\left(\mathbf{m}J_{11}\delta(t) + \mathbf{m}J_{12}\mathrm{e}^{\mathbf{m}J_{22}t}\mathbf{m}J_{21}\right)$. *Furthermore, by choosing the initial condition of* $\mathbf{m}\zeta$ *and the white noise term* $\mathbf{m}\xi(t) = \mathbf{m}\Sigma\mathbf{m}\dot{W}_t$ *satisfying*

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

the embedded kernel $\tilde{\mathbf{m}}\theta(t)$ *and* $\mathbf{m}\mathcal{R}(t)$ *satisfies the second fluctuation-dissipation theorem, i.e.,*

$$\left\langle \tilde{\mathbf{m}}\mathcal{R}(t)\tilde{\mathbf{m}}\mathcal{R}(t')^T \right\rangle = -\beta^{-1}\left(\tilde{\mathbf{m}}J_{12}\mathrm{e}^{\tilde{\mathbf{m}}J_{22}(t-t')}\tilde{\mathbf{m}}J_{21} + (\tilde{\mathbf{m}}J_{11} + \tilde{\mathbf{m}}J_{11}^T)\delta(t-t')\right). \tag{C.2}$$

Proof. With $\mathbf{m}\Lambda = \mathbf{m}I$ and $\tilde{\mathbf{m}}J = \mathbf{m}J$, we can take the integration of $\mathbf{m}\zeta(t)$ in Eq. (2.5), yielding

$$\mathbf{m}\zeta(t) = \int_0^t \mathrm{e}^{\mathbf{m}J_{22}(t-s)}\mathbf{m}J_{21}\mathbf{m}v(s)\mathrm{d}s + \int_0^t \mathrm{e}^{\mathbf{m}J_{22}(t-s)}\mathbf{m}\xi_2(s)\mathrm{d}s + \mathrm{e}^{\mathbf{m}J_{22}t}\mathbf{m}\zeta(0). \tag{C.3}$$

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

$$\begin{aligned}\mathbf{m}M\dot{\mathbf{m}}v =& -\nabla U(\mathbf{m}q) + \mathbf{m}J_{11}\mathbf{m}v + \int_0^t \mathbf{m}J_{12}\mathrm{e}^{\mathbf{m}J_{22}(t-s)}\mathbf{m}J_{21}\mathbf{m}v(s)\mathrm{d}t \\ &+ \underbrace{\mathbf{m}\xi_1(t)}_{\tilde{\mathcal{R}}_1(t)} + \underbrace{\int_0^t \mathbf{m}J_{12}\mathrm{e}^{\mathbf{m}J_{22}(t-s)}\mathbf{m}\xi_2(s)\mathrm{d}s}_{\tilde{\mathcal{R}}_2(t)} + \underbrace{\mathbf{m}J_{12}\mathrm{e}^{\mathbf{m}J_{22}t}\mathbf{m}\zeta(0)}_{\tilde{\mathcal{R}}_3(t)}.\end{aligned} \tag{C.4}$$

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

$$
\begin{aligned}
\left\langle \tilde{\mathcal{R}}_1(t)\tilde{\mathcal{R}}_1(t')^T \right\rangle &= -\beta^{-1}(\mathbf{m}J_{11} + \mathbf{m}J_{11}^T)\delta(t-t'), \\
\left\langle \tilde{\mathcal{R}}_2(t)\tilde{\mathcal{R}}_2(t')^T \right\rangle &= \int_0^t \int_0^{t'} \mathbf{m}J_{12} \mathrm{e}^{\mathbf{m}J_{22}(t-s)} \left\langle \mathbf{m}\xi_2(s)\mathbf{m}\xi_2(s')^T \right\rangle \mathrm{e}^{\mathbf{m}J_{22}^T(t'-s')} \mathbf{m}J_{12}^T \mathrm{d}s\mathrm{d}s' \\
&= -\beta^{-1} \int_0^t \int_0^{t'} \mathbf{m}J_{12} \mathrm{e}^{\mathbf{m}J_{22}(t-s)} (\mathbf{m}J_{22} + \mathbf{m}J_{22}^T)\delta(s-s')\mathbf{m}J_{21}^T \mathrm{e}^{\mathbf{m}J_{22}^T(t'-s')} \mathbf{m}J_{12}^T \mathrm{d}s\mathrm{d}s' \\
&= -\beta^{-1} \int_0^{t'} \mathbf{m}J_{12} \mathrm{e}^{\mathbf{m}J_{22}(t-s')} (\mathbf{m}J_{22} + \mathbf{m}J_{22}^T)\mathbf{m} \mathrm{e}^{\mathbf{m}J_{22}^T(t'-s')} \mathbf{m}J_{12}^T \mathrm{d}s', \\
&= -\beta^{-1}\mathbf{m}J_{12} \mathrm{e}^{\mathbf{m}J_{22}t + \mathbf{m}J_{22}^T t'} \mathbf{m}J_{12}^T + \beta^{-1}\mathbf{m}J_{12} \mathrm{e}^{\mathbf{m}J_{22}(t-t')} \mathbf{m}J_{12}^T, \quad \forall t' \le t \\
\left\langle \tilde{\mathcal{R}}_3(t)\tilde{\mathcal{R}}_3(t')^T \right\rangle &= \mathbf{m}J_{12} \mathrm{e}^{\mathbf{m}J_{22}t} \left\langle \mathbf{m}\zeta(0)\mathbf{m}\zeta(0)^T \right\rangle \mathrm{e}^{\mathbf{m}J_{22}^T t'} \mathbf{m}J_{12}^T \\
&= \beta^{-1}\mathbf{m}J_{12} \mathrm{e}^{\mathbf{m}J_{22}t} \mathrm{e}^{\mathbf{m}J_{22}^T t'} \mathbf{m}J_{12}^T.
\end{aligned}
\tag{C.5}
$$

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

$$
\begin{aligned}
\left\langle \tilde{\mathcal{R}}_2(t)\tilde{\mathcal{R}}_1(t')^T \right\rangle &= \int_0^t \mathbf{m}J_{12} \mathrm{e}^{\mathbf{m}J_{22}(t-s)} \left\langle \mathbf{m}\xi_2(s)\mathbf{m}\xi_1(t') \right\rangle \mathrm{d}s \\
&= -\beta^{-1} \int_0^t \mathbf{m}J_{12} \mathrm{e}^{\mathbf{m}J_{22}(t-s)} (\mathbf{m}J_{21} + \mathbf{m}J_{12}^T)\delta(t'-s)\mathrm{d}s \\
&= -\beta^{-1}\mathbf{m}J_{12} \mathrm{e}^{\mathbf{m}J_{22}(t-t')} (\mathbf{m}J_{21} + \mathbf{m}J_{12}^T).
\end{aligned}
\tag{C.6}
$$

Combining Eq. (C.5) and Eq. (C.6), we have

$$
\begin{aligned}
\left\langle \tilde{\mathcal{R}}(t)\tilde{\mathcal{R}}(t')^T \right\rangle &= \beta^{-1}\mathbf{m}J_{12} \mathrm{e}^{\mathbf{m}J_{22}(t-t')} \mathbf{m}J_{12}^T - \beta^{-1}\mathbf{m}J_{12} \mathrm{e}^{\mathbf{m}J_{22}(t-t')} (\mathbf{m}J_{21} + \mathbf{m}J_{12}^T) \\
&\quad - \beta^{-1}(\mathbf{m}J_{11} + \mathbf{m}J_{11}^T)\delta(t-t') \\
&= -\beta^{-1} \left(\mathbf{m}J_{12} \mathrm{e}^{\mathbf{m}J_{22}(t-t')} \mathbf{m}J_{21} + (\mathbf{m}J_{11} + \mathbf{m}J_{11}^T)\delta(t-t') \right).
\end{aligned}
\tag{C.7}
$$

□

As a special case, by imposing the restraint specified by Eq. (2.16) such that $\mathbf{m}J_{11} + \mathbf{m}J_{11}^T = 0$ and $\mathbf{m}J_{12} = -\mathbf{m}J_{21}^T$, the memory kernel $\tilde{\mathbf{m}\theta}(t)$ recovers $-\mathbf{m}J_{12} \mathrm{e}^{\mathbf{m}J_{22}t} \mathbf{m}J_{12}^T$ without the Markovian part, and the second fluctuation-dissipation theorem recovers the standard form, i.e.,

$$
\left\langle \tilde{\mathbf{m}\mathcal{R}}(t)\tilde{\mathbf{m}\mathcal{R}}(0)^T \right\rangle = \beta^{-1}\tilde{\mathbf{m}\theta}(t). \tag{C.8}
$$

APPENDIX D

INVARIANT PROBABILITY DENSITY FUNCTION

Proposition D.0.1. *By choosing the white noise following Eq.* (C.1)*, the reduced model* (2.5)(2.6) *retains the invariant density function*

$$\rho_{\mathrm{eq}}(\mathbf{m}q, \mathbf{m}p, \mathbf{m}\zeta) = \exp\left[-\beta W(\mathbf{m}q, \mathbf{m}p, \mathbf{m}\zeta)\right] / \int \exp\left[-\beta W(\mathbf{m}q, \mathbf{m}p, \mathbf{m}\zeta)\right] \mathrm{d}\mathbf{m}q \mathrm{d}\mathbf{m}p \mathrm{d}\mathbf{m}\zeta. \tag{D.1}$$

Proof. By Eq. (C.1), the covariance of the white noise of the full extended system is given by $\mathbf{m}G + \mathbf{m}G^T = \mathrm{d}iag(0, \mathbf{m}\Sigma\mathbf{m}\Sigma^T)$. Accordingly, the Fokker-Plank equation follows

$$\frac{\partial \rho(\mathbf{m}z, t)}{\partial t} = \nabla \cdot \left(-\mathbf{m}G \nabla W(\mathbf{m}z) \rho(\mathbf{m}z, t) - \frac{1}{2}\beta^{-1}(\mathbf{m}G + \mathbf{m}G^T)\nabla \rho(\mathbf{m}z, t) \right), \tag{D.2}$$

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

$$\begin{aligned} \nabla \cdot \left(\beta^{-1}\mathbf{m}G \nabla \rho_{\mathrm{eq}}(\mathbf{m}z, t) - \frac{1}{2}\beta^{-1}(\mathbf{m}G + \mathbf{m}G^T)\nabla \rho_{\mathrm{eq}}(\mathbf{m}z, t) \right) &= \beta^{-1} \nabla \cdot \left(\mathbf{m}G^A \nabla \rho_{\mathrm{eq}}(\mathbf{m}z, t) \right) \\ &\equiv 0, \end{aligned} \tag{D.3}$$

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

APPENDIX E

MEMORY KERNEL OF THE POLYMER MOLECULE SYSTEMS

Fig. E.1 shows the embedded matrix-valued kernels $\mathbf{m}\theta(t)$ of the full MD and the 4D reduced models of the polymer molecule system. Similar to the kernel in the Laplace space $\mathbf{m}\Theta(\lambda)$ shown in Fig. 2.6, the good agreement between the full MD and the reduced models verifies that the reduced model can accurately retain the non-Markovian dynamics of the resolved variables.

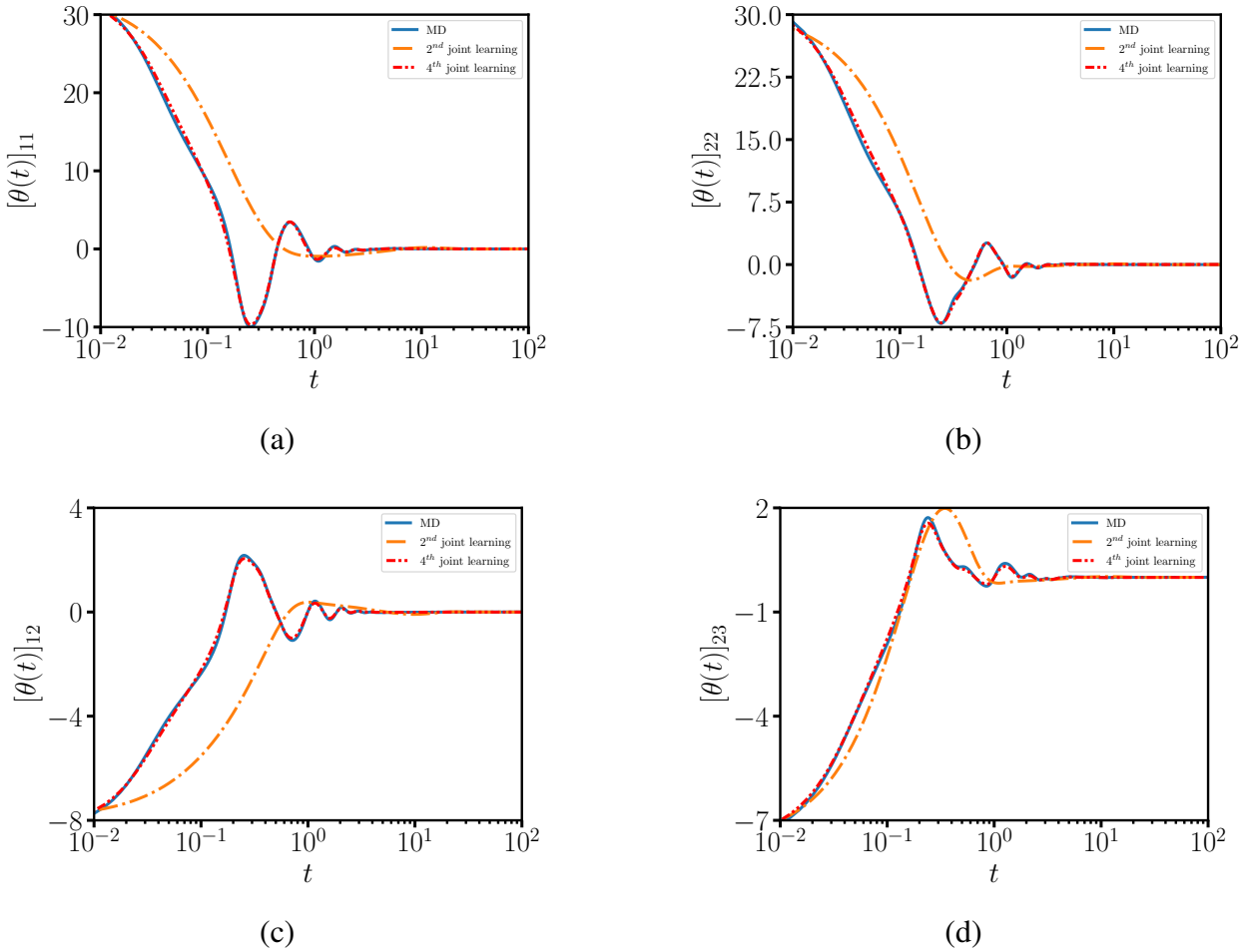


Figure E.1 (a–d) Components of the embedded matrix-valued kernel $\mathbf{m}\theta(t)$ obtained from the full MD and the four-dimensional reduced model of a polymer molecule system She et al. (2023).