MODELING DYNAMICS OF MULTI-BODY SYSTEMS VIA MACHINE LEARNING AND NON-MARKOVIAN APPROACHES

by

Yunrui Qiu

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Chemistry)

at the

UNIVERSITY OF WISCONSIN–MADISON

2024

Date of final oral examination: 09/09/2024

The dissertation is approved by the following members of the Final Oral Committee:

Xuhui Huang, Professor, Chemistry
Arun Yethiraj, Professor, Chemistry
Yang Yang, Assistant Professor, Chemistry
Reid Van Lehn, Associate Professor, Chemical and Biological Engineering

Acknowledgment

"Nearly everything is really interesting if you go into it deeply enough. Work as hard and as much as you want to on the things you like to do the best. Don't think about what you want to be, but what you want to do." *-Richard P. Feynman*

Pursuing a Ph.D. is a long and challenging journey. I am deeply thankful for the incredible people I've encountered over the past four years. Their guidance and assistance have made my time in graduate school rich and rewarding.

First and foremost, I must thank my advisor Prof. Xuhui Huang for his constant support. Since I first joined the group during the pandemic period, Prof. Huang has been incredibly patient, assisting me with difficulties in both my research and personal life. Thanks to his help and support, I have been able to continue my Ph.D. journey and transfer to the chemistry program at UW-Madison with a positive outlook. Prof. Huang is not only an outstanding research advisor but also an excellent teacher. His sharp intellect and extensive expertise were essential in guiding me through the challenges I encountered in my research. Under his guidance, I have acquired a lot of knowledge of computational chemistry, statistical mechanics, machine learning and biochemistry. This dissertation could not have been completed without Prof. Huang's invaluable guidance. He has also taught me essential skills in academic presentation, scientific writing, logical reasoning, and time management. These lessons will greatly benefit my entire academic career. Additionally, I am deeply grateful to him for providing me with opportunities to discuss my research with other brilliant scientists at various conferences and to intern at a pharmaceutical company. When I decided to pursue a future career in academia, Prof. Huang provided immense support and strong recommendations that solidified my confidence in this path. In short, I feel incredibly fortunate to have Prof. Huang as my advisor.

Then I would also like to express my deep gratitude to Prof. Arun Yethiraj. He has not only served as a member of my committee but has also taught me a great deal. The two years I spent as his teaching assistant for undergraduate thermodynamics were among my most memorable times

at UW-Madison. His humor, extensive knowledge, and dedication to teaching have always been a source of inspiration and encouragement for me. Our casual conversations often sparked new research ideas, leading to a project on supercooled liquids that opened a new window for my research interests. I sincerely appreciate his contributions to this project, including his encouraging attitude, insightful ideas, and invaluable guidance on writing. Additionally, I have always enjoyed our conversations, which ranged widely from philosophy and family relationships to music. I hope we have more opportunities to talk and collaborate in the future.

Next, I would like to thank Prof. Mark Ediger, another member of my supervision committee. His responsible, rigorous, and positive attitude toward teaching and research has always been a model for me to learn from and strive to emulate. I am deeply impressed by his extensive knowledge and critical thinking across various disciplines. I am so grateful for his tremendous help with my supercooled liquid research, including his valuable suggestions, his introduction to the pioneer Prof. Ludovic Berthier, and the opportunity to present my work at his supercool symposium. I will always remember his valuable and constructive suggestions, as well as his encouragement for my future plans and career. I want to extend my best wishes for his retirement life.

I would like to extend my special thanks to Prof. Yang Yang and Prof. Reid Van Lehn for serving on my thesis examination committee. I look forward to the possibility of collaborating with them more in the future.

Working in a multidisciplinary research area has provided me with the great opportunity to collaborate with researchers across the nation. I am particularly grateful to my collaborators: Prof Weiping Tang and Inhyuk Jang from UW-Madison, Prof. Bin Zhang, Dr. Shuming Liu and Dr. Xingcheng Lin from MIT, as well as Prof. Pratyush Tiwary, Dr. Dedi Wang, and Dr. Eric Beyerle from UMD, College Park. Their hard work has been essential to the success of our collaborative projects. The discussions with them have been invaluable and unforgettable, consistently offering insights from different perspectives that have greatly enhanced my work. Additionally, I appreciate Prof. Pratyush Tiwary for awarding me the IPST fellowship for my postdoctoral position.

I am sincerely grateful for the opportunity to do internship with the outstanding colleagues and friends at Roivant Science Inc. in New York City. They demonstrated how industry can advance

scientific research to significantly impact human lives. I am especially grateful to Dr. Huafeng Xu, Dr. Jesus Izaguirre, and Dr. Asghar Razavi, who mentored me during my internship. They were really supportive and patient, attentively listening to each of my updates and offering encouraging and valuable feedback. I will never forget the insightful discussions with Dr. Xu, which ranged from philosophy to science. Additionally, I want to particularly thank Dr. Woody Sherman and Dr. Rafal Wiewiora. Without their significant efforts and contributions, I would not have been able to complete the PROTAC paper.

I also thank to the colleagues in the Huang group, including Dr. Siqin Cao, Dr. Kirill Konovalov, Dr. Ilona Unarta, Dr. Yeqing Yu, Dr. Xiaowei Wang, Dr. Chongmin Yuan, Dr. Yuqing Wang, Mingyi Xue, Andrew Yik, Eshani Goonetilleke, Bojun Liu, Michael O'Connor, Michael Kalin, Yue Wu, Yichong Lao, Longbang Liu, Andres Lira, Bruce Huang, Peter C. Swanson, and Jenny Yan.

I feel very fortunate to have great friends who enrich my life outside of work. Most importantly, I want to thank Mr. Haomin Li, Dr. Banlan Chi and Mr. Shiwei Lu, who were my classmates and are among my closest friends. I am deeply grateful for their visits to Madison and our phone conversations. Their support has been a great comfort during many challenging times in my Ph.D. journey. I also appreciate Mr. Jiahao Fan and Dr. Pincheng Xie, who are not only friends but also sources of stimulating conversations and inspiration for science and research.

Last but not least, I must thank my family for their unwavering support. I am deeply grateful to my parents, Feng Qiu and Yafang Li, and my grandparents for encouraging and supporting me in every decision I have made. I am so lucky to be their son and grandson. The daily companionship, love, and support from my wife, Yiwen Wang, have given me the strongest motivation to pursue my Ph.D. She makes me the happiest husband in the world and helps me overcome difficulties whenever I face challenges.

Abstract

Multi-body systems are widely prevalent in chemistry, biology, and material sciences. Their complex energy landscapes, heterogeneous dynamics across different time-scales, and numerous pathways pose significant challenges in modeling long-term dynamics and understanding the underlying molecular mechanisms with high spatial and temporal resolution using current experimental and computational techniques. In this thesis, we developed machine learning algorithms and non-Markovian dynamics modeling approaches to tackle these challenges and explore the dynamics of multi-body systems, ranging from biomolecules, such as protein-protein encounter complexes and chromatin, to materials like supercooled liquids. In particular, to bridge the time gap between simulations and the heterogeneous dynamics of interest, and to provide better interpretation for the underlying mechanisms, we developed a non-Markovian dynamic modeling approach called the Integrated Generalized Master Equation (IGME) model. Unlike conventional Markov State Models (MSMs), the IGME model encodes non-Markovian dynamics into time-integration of memory kernel functions and offers more accurate predictions for long-time dynamics based on shorter simulations. Additionally, to categorize diverse pathways with comparable fluxes, we designed the Latent-space Path Clustering (LPC) algorithm, which applies variational autoencoder network to effectively classify multiple pathways into a small set of metastable path channels according to their kinetic similarities and path typologies. Moreover, we have developed an information bottleneck approach for MSM constructions, providing an end-to-end pipeline that achieves state-of-the-art performance. With these effective machine learning and dynamic modeling tools, we studied a protein-protein encounter complex system, where our IGME model successfully predicted multiple non-canonical metastable protein-protein interfaces, supporting the rational design of PROTACs, a promising next-generation cancer treatment drug. Meanwhile, using our LPC algorithm and IGME model, we also explored chromatin folding dynamics and mechanisms, examining the effects of phase separation of nucleosome condensation and DNA linker length, providing insights into the discrepancies between *in vivo* and *in vitro* studies. In addition, we developed an

unsupervised time-lagged approach to efficiently uncover the structural origins of dynamical heterogeneities in multi-body supercooled liquids, addressing a key open question in the field in a much more data-efficient manner.

Published Work and Work in Preparation

[1] Yunrui Qiu, Rafal P. Wiewiora, Jesus A. Izaguirre, Huafeng Xu, Woody Sherman, Weiping Tang and Xuhui Huang, Non-Markovian Dynamic Models Identify Non-Canonical KRAS-VHL Encounter Complex Conformations for Novel PROTAC Design, *JACS Au* **2024**, jacsau.4c00503.

[2] Yunrui Qiu, Inhyuk Jang, Xuhui Huang and Arun Yethiraj, Unsupervised machine learning for supercooled liquids, *arXiv* **2024**, 2404.04473v2. *to be submitted*

[3] Dedi Wang♯, Yunrui Qiu♯, Eric R. Beyerle♯, Xuhui Huang and Pratyush Tiwary, An Information Bottleneck Approach for Markov Model Construction, *J. Chem. Theory Comput.* **2024**, 20(12), 5352–5367. (♯ **contribute equally)**

[4] Yunrui Qiu, Michael S. O'Connor, Mingyi Xue, Bojun Liu and Xuhui Huang, An Efficient Path Classification Algorithm Based on Variational Autoencoder to Identify Metastable Path Channels for Complex Conformational Changes, *J. Chem. Theory Comput.* **2023**, 19(14), 4728-4742.

[5] Yue Wu, Siqin Cao, Yunrui Qiu, and Xuhui Huang, Tutorial on how to build non-Markovian dynamic models from molecular dynamics simulations for studying protein conformational changes, *J. Chem. Phys.* **2023**, 160, 121501.

[6] Siqin Cao, Yunrui Qiu, Michael L. Kalin and Xuhui Huang, Integrative generalized master equation: A method to study long-timescale biomolecular dynamics via the integrals of memory kernels, *J. Chem. Phys.* **2023**, 159, 134106.

[7] Bojun Liu, Mingyi Xue, Yunrui Qiu, Kirill A. Konovalov, Michael S. O'Connor and Xuhui Huang, GraphVAMPnets for Uncovering Slow Collective Variables of Self-Assembly Dynamics, *J. Chem. Phys.* **2023**, 159, 094901.

[8] Andrew Kai-Hei Yik, Yunrui Qiu, Ilona C. Unarta, Siqin Cao, and Xuhui Huang, A step-by-step guide on how to construct quasi-Markov state models to study functional conformational changes of biological macromolecules, *In A Practical Guide to Recent Advances in Multiscale Modeling and Simulation of Biomolecules* **2023**, Book Chapter 10.

[9] Kirill A. Konovalov, Cheng-Guo Wu, Yunrui Qiu, Vijaya Kumar Balakrishnan, Pankaj Singh Parihar, Michael S. O'Connor, Yongna Xing and Xuhui Huang, Disease mutations and phosphorylation alter the allosteric pathways involved in autoinhibition of protein phosphatase 2A, *J. Chem. Phys.* **2023**, 158, 215101.

[10] Bojun Liu♯, Yunrui Qiu♯, Eshani C. Goonetilleke and Xuhui Huang, Kinetic network models to study molecular self-assembly in the wake of machine learning, *MRS Bulletin* **2022**, 47(9), 958-966. (♯ **contribute equally)**

[11] Siqin Cao, Yunrui Qiu, Ilona C. Unarta, Eshani C. Goonetilleke and Xuhui Huang, The Ion-Dipole Correction of the 3DRISM Solvation Model to Accurately Compute Water Distributions around Negatively Charged Biomolecules, *J. Phys. Chem. B* **2022**, 126, 8632-8645.

[12] Yunrui Qiu♯, Shuming Liu♯, Xingcheng Lin, Ilona C. Unarta, Xuhui Huang and Bin Zhang, Nucleosome condensate and linker DNA alter chromatin folding pathways and rates, In prep. **(♯ contribute equally)**

TABLE OF CONTENTS

Page

Page

List of Tables

Table Page

List of Figures

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Chapter 1

Introduction

1.1 Conformational Dynamics of Multi-Body Systems

Conformational dynamics, i.e., the transitions between pairs of conformational states, are crucial in a wide range of multi-body systems across chemistry, biology, and materials science. For instance, protein-protein interactions association and dissociation are essential to numerous vital biological processes[1–4]. Accurate recognition between drug molecules and protein receptors is fundamental for effective drug design[5–8]. The aggregation and separation of biomolecules can significantly influence various biological functions[9, 10]. The supercooling of liquid is vital for synthesizing important glass materials used in both the electronic and pharmaceutical industries[11, 12]. And self-assembly provides a powerful technique for fabricating advanced materials using a bottom-up approach in nanotechnology[13–17]. Investigating the dynamics of conformational changes in multi-body systems at the molecular level is essential for understanding the microscopic mechanisms of many fundamental chemical and biological processes and advancing the rational design of materials. However, studying the dynamics of multi-body systems is highly challenging due to their complexity, and the fact that they often occur on the millisecond time scale or longer at the detailed automatic level[15, 16, 18, 19]. In particular, the involvement of diverse inter-molecular and intra-molecular interaction patterns in the multi-body systems result in complex free energy landscapes, multiple heterogeneous dynamical modes across different timescales and numerous parallel pathways with comparable reactive fluxes. To investigate the kinetics, molecular mechanisms and pathways of conformational changes in multi-body systems,

considerable efforts have been made in both computational and experimental methods[1, 5, 13, 16, 18, 19].

Various experimental tools, including single-molecule fluorescence resonance energy transfer (FRET)[20], dynamic nuclear magnetic resonance (NMR)[21], and laser-induced temperature jump[22], are commonly employed to study the dynamics of conformational changes in multi-body systems. However, their relatively low spatial and temporal resolutions significantly limit their ability to fully elucidate the microscopic details of these conformational changes.

Additionally, computational tools such as all-atom molecular dynamics (MD) simulations have proven to be an effective complementary tool to experimental techniques for studying conformational dynamics and related kinetic pathways in multi-body systems[13, 18]. MD simulations can model the conformational evolution on a femtosecond timescale and at an atomic spatial scale, offering high-resolution dynamical descriptions into conformational changes. However, the time window that all-atom MD simulation can sample is typically much shorter than the timescale of conformational changes in multi-body systems. This limitation arises from the minimum integration steps required for reliable MD simulations and the constrained computational resources available. Additionally, the trajectories obtained from MD simulations are high-dimensional time sequences, making it challenging for people to interpret the kinetics and molecular mechanisms underlying conformational dynamics.

To bridge the gap between the timescales achievable by MD simulations with limited computational resources and the long-term conformational changes in multi-body systems, various techniques have been developed. Enhanced sampling algorithms, such as Metadynamics[23] and Hamiltonian Replica Exchange[24], and adaptive sampling algorithms[25], like weighted ensemble[26], have been integrated with MD simulations and have become popular in the past decade for effectively sampling the energy landscape of complex multi-body systems. Additionally, statistical dynamics models, such as Markov State Models (MSMs)[27–29], non-Markovian dynamics models[30–32] and Langevin equation model [33], have emerged as powerful approaches for modeling the long-time dynamics of conformational changes based on ensembles of short MD trajectories. Meanwhile, it is worth noting that these dynamic models can offer a more refined

coarse-grained representation of MD trajectories, making them much easier to interpret. Recently, with the rise of machine learning techniques, many sampling and dynamic modeling methods have been integrated with various deep learning architectures, demonstrating significant performance improvements[34–37].

1.2 Studying Dynamics Using Dynamic Models and Machine Learning Techniques

Dynamic models based on statistical mechanics, particularly MSMs, have been extensively developed and applied over the past twenty years as a critical method for bridging the timescale gap between unbiased MD simulations and rare events. Robust theories have been proposed and developed from various perspectives, resulting in the development of numerous algorithms in several well-known software packages[38, 39]. The entire pipeline for MSM construction involves multiple steps: partitioning the conformational space into metastable states according to their dynamic metastability and coarse-graining time into discrete units called lag times. With an appropriate selection of the lag time, the continuous dynamics in MD simulations can be modeled as Markovian transitions among different conformational states. Consequently, MSMs can propagate long-time dynamics based on short simulations and coarse-grain MD conformations into a few comprehensive states, facilitating the prediction of their associated thermodynamic and kinetic properties. Additionally, applying Transition Path Theory (TPT)[40, 41] to an MSM offers significant potential for deriving the complete ensemble of kinetic pathways from MD simulations, thereby aiding in the understanding of the underlying mechanisms. This integrated approach has been widely applied to study conformational changes of multi-body systems [13, 18].

However, for MSMs to achieve high predictive accuracy for long-term dynamics, they must be constructed with either a large number of states or a sufficiently long lag time to ensure that intra-state dynamics are relaxed and inter-state transitions are Markovian. This requirement presents significant challenges, as a large number of states complicates the model's interpretability, and the lag time is constrained by the limited duration of short MD simulations. The increasing number of states will also lead to a proliferation of kinetic pathways, particularly in multi-body systems,

making it difficult to understand the mechanisms of conformational changes. Extending lag times will not only require longer simulations but also reduce time resolution and introduce additional statistical errors. These unresolved challenges urgently require the development of new theories and algorithms. In the following chapters, we will present our newly developed non-Markovian dynamics models and demonstrate their effectiveness in overcoming the above-mentioned challenges, providing new physical and biological insights into the dynamics of various multi-body systems.

Meanwhile, with the rapid advancement of machine learning techniques, numerous algorithms and deep neural network architectures have been integrated into the construction pipeline of dynamic models. These innovations have greatly enhanced the quality of dimensionality reduction, clustering, and kinetic lumping procedures, streamlining the previously complex multi-step process [34, 35, 37]. The development of novel deep learning approaches for MSM construction has improved model quality by shortening the required Markovian lag time , and it recently has became a rapidly evolving and promising research area. Expanding and implementing these methodologies to various systems—such as multi-body biological systems (e.g., protein-protein interactions and biomolecule aggregations) and multi-body material systems (e.g., self-assembly and supercooled liquids)—as well as developing new algorithms with novel machine learning architectures to address the challenges in the studies of multi-body systems, holds significant potential for future advancements.

Current model construction algorithms are primarily unsupervised, relying on patterns and structures in input data without predefined targets or labels. In contrast, supervised machine learning methods have been majorly applied to study supercooled liquids, an important type of multi-body system, to automatically detect relevant structural features and establish structure-dynamics relationships[42–44]. These supervised models require extensive training with large datasets that include known long-time dynamics, leading to high computational demands and limited interpretability. Consequently, integrating high-performance unsupervised approaches into the study of supercooled liquids presents a promising alternative for gaining deeper physical insights[45].

1.3 Overview of the Thesis Work

In this thesis, we focus on developing novel computational tools to gain a deeper understanding of the conformational dynamics of the multi-body systems. This development of methods and algorithms is grounded in statistical mechanics theory and advanced machine learning techniques. In Chapter 2 and Chapter 3 we first provide a concise overview of the fundamental theories and general methodologies underlying the MSMs and Generalized Master Equation (GME) methods[31, 46]. Chapter 2 introduces various theories of MSMs (i.e., the projection operator theory and the Variational Approach for Markov Processes (VAMP) theory[34, 47] etc.) and techniques used in MSM construction, and discusses the advantages and limitations of each algorithms. Chapter 3 approaches the problem from a different angle, introducing our newly developed non-Markovian dynamics modeling approaches based on the GME. These approaches effectively address some of the open questions related to MSMs[30, 31]. Both chapters are designed to facilitate the development of subsequent sections in this thesis, whether focused on method development or practical applications.

The general methods and algorithms discussed in Chapter 2 and 3 may not always be directly applicable to real-world multi-body systems, as new research targets often present unique challenges. In Chapter 4, we will introduce a novel algorithm, the Latent-space Path Clustering (LPC) algorithm[48]. This algorithm employs a variational autoencoder to effectively classify multiple transition pathways into a limited set of metastable path channels based on their kinetic similarities. The LPC algorithm significantly enhances the interpretation of complex dynamic mechanisms and surpasses previous methods in performance. This algorithm will be further utilized to elucidate the folding mechanisms of tetra-nucleosome chromatin systems, as discussed in Chapter 7.

In Chapter 5, we leverage a popular deep learning architecture known as the information bottleneck to develop a novel approach for constructing MSMs[35]. Unlike traditional multi-step protocols, this information bottleneck method provides a streamlined, end-to-end pipeline that achieves state-of-the-art performance with minimal hyperparameter tuning. We demonstrate this approach

could shorten the Markovian lag time compared with the conventional methods and expect it to significantly advance the study of conformational dynamics of multi-body systems in the future.

Chapter 6 focus on the application of the various advanced machine learning and dynamic modeling tools to investigate the conformational changes within a critical multi-body system: protein-protein encounter complex[3]. Specifically, we demonstrate how these techniques enhance the prediction of non-canonical protein-protein interfaces at the atomic level, thereby facilitating the rational design of PROTACs, which are highly promising candidates for next-generation cancer treatments.

Subsequently, Chapter 7 showcases the effectiveness of non-Markovian dynamics model and the LPC algorithm in investigating conformational changes in chromatin multi-body systems. Specifically, we examine how the nucleosome aggregated condensate environment and varying linker DNA lengths influence chromatin folding dynamics and mechanisms, offering insights into the discrepancies observed between *in vivo* and *in vitro* studies. In this chapter, we also demonstrate that non-Markovian dynamics model outperforms MSMs in accurately recover the long-time dynamics from short MD trajectories.

Lastly, in addition to the unsupervised machine learning techniques used in the previous chapters, Chapter 8 demonstrates how unsupervised approaches can be employed to investigate the long-time dynamics of supercooled liquids in a more interpretable and data-efficient manner[45]. Especially, we introduce a new time-lagged analysis scheme that utilizes advanced structural descriptors, time-lagged canonical component analysis, and time-lagged autoencoder to elucidate the structural origins of dynamical heterogeneities in supercooled liquids. Compared to the state-of-the-art supervised machine learning methods, our unsupervised scheme can achieve the same accuracy in predicting long-time dynamics without requiring prior knowledge and intensive training.

Chapter 2

Investigating Dynamics of Complex Conformational Changes with Markov State Models

2.1 Hamiltonian Dynamics and Liouville Equation

In classical mechanics, the motion of a system can be described by Hamiltonian equations:

$$\frac{d\boldsymbol{r}}{dt} = \frac{\partial \mathcal{H}}{\partial \boldsymbol{p}}; \quad \frac{d\boldsymbol{p}}{dt} = -\frac{\partial \mathcal{H}}{\partial \boldsymbol{r}} \tag{2.1}$$

where $\boldsymbol{r} = \{r_1, r_2, \ldots r_{3N}\}$ and $\boldsymbol{p} = \{p_1, p_2, \ldots r_{3N}\}$ represent the generalized coordinate and generalized momenta of the system, respectively. Therefore, the time evolution of the system can be described as the trajectory of a point $\{\boldsymbol{\Gamma}(t)\}_{t=0}^{t=N} = \{\boldsymbol{r}(t), \boldsymbol{p}(t)\}_{t=0}^{t=N}$ in the phase space. From the probability point of view, the density distribution could be introduced to describe the probability of the system to be observed at certain region and certain time: $\rho(\boldsymbol{\Gamma_t}, t)$. The evolution of the density distribution must adhere to the conservation law:

$$\rho(\boldsymbol{\Gamma}_t, t)\delta\boldsymbol{\Gamma}(t) = \rho(\boldsymbol{\Gamma}_{t+\Delta t}, t + \Delta t)\delta\boldsymbol{\Gamma}(t + \Delta t) \tag{2.2}$$

Considering the phase space volume at different times, they could be connected through the Jacobi determinant (assuming r_i and p_i are independent):

$$\delta r_i(t + \Delta t)\delta p_i(t + \Delta t) = \begin{vmatrix} \frac{\partial p_i(t+\Delta t)}{\partial p_i(t)} & \frac{\partial p_i(t+\Delta t)}{\partial r_i(t)} \\ \frac{\partial r_i(t+\Delta t)}{\partial p_i(t)} & \frac{\partial r_i(t+\Delta t)}{\partial r_i(t)} \end{vmatrix} \delta r_i(t)\delta p_i(t) \tag{2.3}$$

If the system is conservative, the determinant could be simplified with Hamiltonian equation 2.1:

$$\delta r_i(t + \Delta t)\delta p_i(t + \Delta t) = (1 - (\frac{\partial^2 \mathcal{H}}{\partial r_i \partial p_i})^2 + \frac{\partial^2 \mathcal{H}}{\partial r_i^2}\frac{\partial^2 \mathcal{H}}{\partial p_i^2})\Delta t^2 \delta r_i(t)\delta p_i(t) \tag{2.4}$$

Then the volume element in phase space could be proved as conserved:

$$\lim_{\Delta t \to 0} (\delta\boldsymbol{\Gamma}(t + \Delta t) - \delta\boldsymbol{\Gamma}(t))/\Delta t = 0 \tag{2.5}$$

And the density distribution function should be unchanged along the evolution of the system:

$$\rho(\mathbf{\Gamma}_t, t) = \rho(\mathbf{\Gamma}_{t+\Delta t}, t + \Delta t) \tag{2.6}$$

Therefore, the total time derivative of density distribution should equal to zero:

$$\frac{d\rho(\mathbf{\Gamma_t}, t)}{dt} = \frac{\partial \rho}{\partial t} + \frac{\partial \rho}{\partial \mathbf{\Gamma}} \cdot \frac{\partial \mathbf{\Gamma}}{\partial t} = 0 \tag{2.7}$$

By integrating with Hamiltonian equation 2.1, we could further simplify the second term:

$$\frac{\partial \rho}{\partial \mathbf{\Gamma}} \cdot \frac{\partial \mathbf{\Gamma}}{\partial t} = \sum_{i=1}^{3N} (\frac{\partial \rho}{\partial r_i}\frac{dr_i}{dt} + \frac{\partial \rho}{\partial p_i}\frac{dp_i}{dt}) = \sum_{i=1}^{3N} (\frac{\partial \rho}{\partial r_i}\frac{\partial \mathcal{H}}{\partial p_i} - \frac{\partial \rho}{\partial p_i}\frac{\partial \mathcal{H}}{\partial r_i}) \tag{2.8}$$

Plug back to equation 2.7, we could obtain:

$$\frac{\partial \rho}{\partial t} = -\sum_{i=1}^{3N} (\frac{\partial \rho}{\partial r_i}\frac{\partial \mathcal{H}}{\partial p_i} - \frac{\partial \rho}{\partial p_i}\frac{\partial \mathcal{H}}{\partial r_i}) = -\sum_{i=1}^{3N} (\frac{\partial \mathcal{H}}{\partial p_i}\frac{\partial}{\partial r_i} - \frac{\partial \mathcal{H}}{\partial r_i}\frac{\partial}{\partial p_i})\rho \tag{2.9}$$

The Poisson bracket is further utilized to define the Liouville operator, allowing us to express the Liouville equation as follows:

$$\frac{\partial \rho(\mathbf{\Gamma}_t, t)}{\partial t} = \mathcal{L}\rho(\mathbf{\Gamma}_t, t) \tag{2.10}$$

In equilibrium, Liouville operator satisfies the principle of detailed balance:

$$\mathcal{L}\rho_{eq}(\mathbf{\Gamma}) = 0 \tag{2.11}$$

and Liouville operator is a self-adjoint operator with respect to the equilibrium distribution:

$$\int f_i(\mathbf{\Gamma}) e^{\mathcal{L}\Delta t} f_j(\mathbf{\Gamma}) \rho_{eq}^{-1}(\mathbf{\Gamma}) d\mathbf{\Gamma} = \int f_j(\mathbf{\Gamma}) e^{\mathcal{L}\Delta t} f_i(\mathbf{\Gamma}) \rho_{eq}^{-1}(\mathbf{\Gamma}) d\mathbf{\Gamma} \tag{2.12}$$

2.2 Molecular Dynamics

MD is a computational technique developed to investigate the dynamical behaviors of the molecular systems over time. Based on the Hamiltonian mechanics, MD provides insights into the interactions and dynamics of atoms and molecules with high time resolutions, enabling the exploration of conformational changes within multi-body systems. Given the complex interaction

patterns in multi-body systems, deriving an analytical expression for the Hamiltonian or Liouville operator is infeasible. Instead, classical force fields are usually employed to model inter- and intra-molecular interactions. Classical all-atom force fields typically model potential energy using two main components: bonded and nonbonded interactions, each with various terms for different interaction types. Coarse-graining force fields, on the other hand, streamline complex molecular systems by consolidating groups of atoms or molecules into single interaction sites, enabling more efficient simulations of large-scale systems and extended time scales while preserving key physical phenomena. In this thesis, we will utilize both types of force fields to investigate the conformational changes in multi-body systems at varying levels of resolution.

Since force fields provide an implicit way to express the Hamiltonian, the dynamical propagation scheme can be subsequently established using Hamiltonian mechanics. Various integrators are designed to propagate the dynamics, typically using very short time steps to accurately capture the fastest fundamental motions and ensure the robustness of the calculation. For example, the Velocity-Verlet algorithm and Leaf-Frog algorithm are commonly used. Meanwhile, in MD simulations, thermostats are typically used to regulate and maintain the system's temperature or pressure, ensuring it adheres to the desired ensemble conditions (e.g., *NPT* or *NVT*). The implementation of popular Langevin or Langevin middle integrators can automatically control the temperature based on the fluctuation-dissipation theorem.

As previously mentioned, the time step in MD simulations limits the length of a single trajectory, making it challenging to cover the timescale of conformational changes within multi-body systems. To address this, MSMs are developed using a "divide-and-conquer" approach, where multiple short trajectories could be integrated to infer long-term properties[27–29]. The construction of MSMs can be understood through various theoretical frameworks, including the projection operator approach [46] and the Variational Approach for Markov Processes theory[34, 47].

2.3 Projection Operator and Markov State Models

Phase space is inherently high-dimensional for multi-body systems, and while the evolution of dynamics within it is Markovian, it tends to be exceedingly complex. The Liouville operator,

likewise, is typically complicated, encompassing both slow and fast-evolving kinetics. This complexity is largely depended upon the inherent characteristics of the systems themselves. In the case of dynamics of conformational changes, an approximation is usually made that a few underlying manifolds can effectively describe the dynamics. In general, the spectrum of Liouville operator can be written in order[27, 49]:

$$\lambda_1 = 0 < \lambda_2 \leq \lambda_3 \leq \cdots \leq \lambda_k \ll \lambda_{k+1} \leq \ldots \tag{2.13}$$

and we can always express the evolution of the density distribution in term of the eigenvectors of Liouville operator:

$$|\rho(t+\Delta t)\rangle = |\psi_1\rangle\langle\phi_1|\rho(t)\rangle + \sum_{j=2} |\psi_j\rangle e^{\lambda_j \Delta t}\langle\phi_j|\rho(t)\rangle \tag{2.14}$$

Set the the truncation based on the separation or gap in the spectrum of Liouville operator:

$$|\rho(t+\Delta t)\rangle = \sum_{j=1}^{k} |\psi_j\rangle e^{\lambda_j \Delta t}\langle\phi_j|\rho(t)\rangle + \sum_{j=k+1} |\psi_j\rangle e^{\lambda_j \Delta t}\langle\phi_j|\rho(t)\rangle \tag{2.15}$$

In this case, we exclusively focus on the leading eigenvalues and their associated eigenvectors, which characterize the slowest dynamical modes. This choice is justified by the fact that, as time progresses, the smaller eigenvalues rapidly converge to zero. The eigenvectors corresponding to the slow leading eigenvalues hold significant promise as candidate collective variables or metastable states, because the slowest processes often involve overcoming high free-energy barriers.

To represent the dynamics in a low-dimensional space (e.g. coase-grained metastable states) which is spanned by finite basis set: $\{\chi_1(\mathbf{\Gamma}), \chi_2(\mathbf{\Gamma}), \cdots, \chi_n(\mathbf{\Gamma})\}$, we need to define the projection operator. The general form of projection operator could be written as $\mathbb{P} = |\mathbf{R}\rangle\mathbf{S}^{-1}\langle\mathbf{L}|$ where $|\mathbf{R}\rangle = \rho_{eq}[\chi_1(\mathbf{\Gamma})\,\chi_2(\mathbf{\Gamma}), \cdots, \chi_n(\mathbf{\Gamma})]$, $\langle\mathbf{L}| = [\chi_1(\mathbf{\Gamma})\,\chi_2(\mathbf{\Gamma}), \cdots, \chi_n(\mathbf{\Gamma})]^T$ and $\mathbf{S} = \langle\mathbf{L}|\mathbf{R}\rangle$ is the overlap matrix. Apply the projection operator on the Liouville equation:

$$\begin{aligned} \mathbb{P}\rho(t+\Delta t) &= \mathbb{P}e^{\mathcal{L}\Delta t}\mathbb{P}\rho(t) + \mathbb{P}e^{\mathcal{L}\Delta t}\mathbb{Q}\rho(t) \\ \mathbb{Q}\rho(t+\Delta t) &= \mathbb{Q}e^{\mathcal{L}\Delta t}\mathbb{P}\rho(t) + \mathbb{Q}e^{\mathcal{L}\Delta t}\mathbb{Q}\rho(t) \end{aligned} \tag{2.16}$$

where $\mathbb{P} + \mathbb{Q} = \mathbb{I}$, $\mathbb{Q}$ is the complementary operator. The projected dynamics can be decomposited into Markovian term (first term) and non-Markovian term (high-order term):

$$\mathbb{P}\rho(t+\Delta t) = \mathbb{P}e^{\mathcal{L}\Delta t}\mathbb{P}\rho(t) + \mathbb{P}e^{\mathcal{L}\Delta t}\mathbb{Q}e^{\mathcal{L}\Delta t}\mathbb{P}\rho(t-\Delta t) + \cdots\cdots \tag{2.17}$$

One property of projection operator is: $\langle \mathbf{L}|\mathbb{P}\rangle = \langle \mathbf{L}|\mathbf{R}\rangle \mathbf{S}^{-1}\langle \mathbf{L}| = \langle \mathbf{L}|$, we can apply left projector on the Markovian part of the equation and get:

$$\langle \mathbf{L}|\mathbb{P}\rho(t+\Delta t)\rangle = \langle \mathbf{L}|\mathbb{P}e^{\mathcal{L}\Delta t}\mathbb{P}\rho(t)\rangle \longrightarrow \langle \mathbf{L}|\rho(t+\Delta t)\rangle = \langle \mathbf{L}|e^{\mathcal{L}\Delta t}|\mathbf{R}\rangle \mathbf{S}^{-1}\langle \mathbf{L}|\rho(t)\rangle \tag{2.18}$$

In the matrix form:

$$\begin{pmatrix} \langle \chi_1|\rho(t+\Delta t)\rangle \\ \vdots \\ \langle \chi_n|\rho(t+\Delta t)\rangle \end{pmatrix} = \begin{pmatrix} \langle \chi_1|e^{L\Delta t}|\rho_{eq}\chi_1\rangle & \cdots & \langle \chi_1|e^{L\Delta t}|\rho_{eq}\chi_n\rangle \\ \vdots & \ddots & \vdots \\ \langle \chi_n|e^{L\Delta t}|\rho_{eq}\chi_1\rangle & \cdots & \langle \chi_n|e^{L\Delta t}|\rho_{eq}\chi_n\rangle \end{pmatrix} \tag{2.19}$$

$$\cdot \begin{pmatrix} \langle \chi_1|\rho_{eq}\chi_1\rangle & \cdots & \langle \chi_1|\rho_{eq}\chi_n\rangle \\ \vdots & \ddots & \vdots \\ \langle \chi_n|\rho_{eq}\chi_1\rangle & \cdots & \langle \chi_n|\rho_{eq}\chi_n\rangle \end{pmatrix}^{-1} \begin{pmatrix} \langle \chi_1|\rho(t)\rangle \\ \vdots \\ \langle \chi_n|\rho(t)\rangle \end{pmatrix} \tag{2.20}$$

$$\rho_p(t+\Delta t) = \mathbf{C}(\Delta t)\mathbf{S}^{-1}\rho_p(t) \tag{2.21}$$

where $\mathbf{C}(\Delta t)$ is the time-lagged correlation matrix. $\mathbf{C}(\Delta t)\mathbf{S}^{-1}$ is the normalized time-lagged correlation matrix. Since we typically cannot obtain the analytical formulas for the Liouville operator or Hamiltonian, the correlation matrix can only be estimated from MD simulations or observed experimental data (ensemble average can be estimated from trajectory average)[27]:

$$\begin{aligned} \mathbf{C}_{ij}(\Delta t) &= \frac{1}{N_T - \Delta t}\sum_{i=1}^{N_T-\Delta t} \chi_j(\mathbf{\Gamma}, t)\chi_i(\mathbf{\Gamma}, t+\Delta t) \\ \mathbf{S}_{ij}(\Delta t) &= \frac{1}{N_T - \Delta t}\sum_{i=1}^{N_T-\Delta t} \chi_j(\mathbf{\Gamma}, t)\chi_i(\mathbf{\Gamma}, t) \end{aligned} \tag{2.22}$$

It is worth noting that other methods like Maximum Likelihood Estimate[27] or Beyesian estimation[50] can also be adopted to construct the correlation matrix. Once we estimate the $\mathbf{C}(\Delta t)\mathbf{S}^{-1}$ and assume the Markovian term is the major contributor for the projected dynamics, we can easily propagator our dynamics by using $\mathbf{C}(\Delta t)\mathbf{S}^{-1}$, resulting in a first order master equation.

In the specific case of MSMs, where the low-dimensional space is constructed using indicator functions to define state locations and boundaries within the phase space (i.e., $\chi_i(\boldsymbol{x}) = 1$ (or 0) if

state variable of configuration $\boldsymbol{x}$ belongs (or not) to state i), the projection operator can be further expressed as[46]:

$$\mathbb{P} = \sum_{k=1}^{N} |\rho_{eq}(\mathbf{\Gamma})\chi_k(\mathbf{\Gamma})\rangle \cdot \pi_k^{-1}\langle\chi_k(\mathbf{\Gamma})| \tag{2.23}$$

where π_k stands for the equilibrium population of state k. And the Markovian part of the dynamical equation 2.21 could be written as:

$$p(t + n\Delta t) = \boldsymbol{T}^n(\Delta t)p(t) \tag{2.24}$$

where $p_j(t) = \langle\chi_j(\boldsymbol{x})|\rho(\mathbf{\Gamma}, t)\rangle$ represents the probability of the system is observed in state j at time t, and $\boldsymbol{T}_{ij}(\Delta t) = \langle\chi_i(\boldsymbol{x})|e^{\mathcal{L}\Delta t}|\rho_{eq}(\mathbf{\Gamma})\chi_j(\boldsymbol{x})\rangle\pi_j^{-1}$ is the transition probability matrix (TPM) of lag time Δt.

2.4 Variational Approach for Conformational Dynamic (VAC)

If one system is well sampled and its ergodic coordinates and momentum are recorded, it is possible to construct the evolution propagator in phase space, effectively forming the Liouville operator. However, this is generally not advisable because phase space typically has high dimensionality, and we often do not need to account for the complete translational and rotational movements of the molecules, nor the full degrees of freedom of the solvent. By embedding molecular conformations obtained from MD simulations with translationally and rotationally invariant internal 'features' and excluding all solvent degrees of freedom, we can define a subspace within the phase space. These 'features' are typically nonlinear combinations of the original Cartesian coordinates. In the construction of MSMs, we decompose the high-dimensional phase space into a few metastable states, which can also be regarded as a nonlinear transformation to a low-dimensional state space. As previously demonstrated, the projected dynamics are not necessarily linear and Markovian unless there is a clear temporal separation and we apply a linear transformation to the coordinates. How can we evaluate different definitions of low-dimensional representations or metastable states? The variational approach for conformational dynamics (VAC) shows that the optimal representation or state definition is achieved through the best approximations of the leading slowest dynamical modes of the evolution operator or dynamical propagator[47, 51, 52].

Since the explicit form of the evolution operator is unknown, obtaining its eigenvectors is challenging. In line with quantum mechanics, a straightforward approach is to use a linear combination of basis sets and optimize the combination coefficients. To guide this optimization, certain criteria are needed. It has been demonstrated that, for a linear propagator-driven reversible Markovian process, the eigenvalues derived from the approximated eigenvectors (orthogonal to the top eigenvectors) are consistently smaller than the true eigenvalues. If we consider the Liouville operator as the evolution operator, any trial function can be expressed within the space spanned by the eigenfunctions:

$$\langle f| = \sum_{\alpha} c_{\alpha} \langle \phi_{true,\alpha}| \quad (\sum_{\alpha} c_{\alpha}^2 = 1) \tag{2.25}$$

The variational score could be used to measure how well the trial function approximate the slowest dynamic modes:

$$\langle f|e^{\mathcal{L}\Delta t}|f\rangle_{\rho_{eq}} = \sum_{\alpha}\sum_{\beta} c_{\alpha}c_{\beta}\langle \phi_{true,\alpha}|e^{\mathcal{L}\Delta t}|\phi_{true,\beta}\rangle_{\rho_{eq}} \tag{2.26}$$

$$= \sum_{\alpha}\sum_{\beta} c_{\alpha}c_{\beta}\mu_{true,\beta}\langle \phi_{true,\alpha}|\phi_{true,\beta}\rangle_{\rho_{eq}} \tag{2.27}$$

$$= \sum_{\alpha}\sum_{\beta} c_{\alpha}c_{\beta}\mu_{true,\beta}\delta_{\alpha\beta} = \sum_{\alpha} c_{\alpha}^2\mu_{true,\alpha} \leq \mu_{true,1} \tag{2.28}$$

The similar derivations can be performed iteratively, ensuring that the lower-order trial functions remain orthogonal to the higher-order ones throughout the process. Therefore, the sum of the Rayleigh quotients, which are the top eigenvalues of the TPM or whiten time-lagged correlation matrix, is used as a score to assess the quality of the MSM or low-dimensional collective variables, since it is upper-bounded.

2.5 Variational Approach for Markov Processes (VAMP)

Variational approach for Markov processes (VAMP) theory has been developed based on Koopman operator theory, enabling the description of general Markovian dynamical processes, whatever the process is reversible or nonreversible, stationary or non-stationary [34, 47, 53]. Considering an arbitrary Markovian dynamical system, the system's evolution can be expressed through time

series of state variables, denoted as $\{\boldsymbol{x}_t\}_{t=0}^{T} \in \mathbb{R}^{3N}$. Koopman operator theory ensures that, even if the dynamics are nonlinear, the system's evolution can still be described by a linear Koopman operator, defined as:

$$\mathbb{E}[\boldsymbol{\zeta_1}(\boldsymbol{x}_{t+\tau})] = \mathbf{K}_{\tau}^{T}\mathbb{E}[\boldsymbol{\zeta_0}(\boldsymbol{x}_t)] \tag{2.29}$$

where $\boldsymbol{\zeta_0}(\boldsymbol{x}) = [\zeta_{01}(\boldsymbol{x}), \zeta_{02}(\boldsymbol{x}), \cdots]^T$ and $\boldsymbol{\zeta_1}(\boldsymbol{x}) = [\zeta_{11}(\boldsymbol{x}), \zeta_{12}(\boldsymbol{x}), \cdots]^T$ are feature transformation functions or observable transformation functions with infinite rank that map the state variable $\boldsymbol{x}$ to the feature or observable spaces $\mathcal{L}^2_{\rho_0} = \{\zeta_0 | \langle \zeta_0, \zeta_0 \rangle_{\rho_0} < \infty\}$ and $\mathcal{L}^2_{\rho_1} = \{\zeta_1 | \langle \zeta_1, \zeta_1 \rangle_{\rho_1} < \infty\}$. Following the theory, subsequent questions arise: Given our consistent pursuit of simple representations (e.g., collective variables or metastable states) for expressing dynamics, how can we construct a deduced Koopman operator with a substantially lower rank while still being able to capture long-term dynamical information? And What are the most optimal feature transformation functions for the deduced operator model?

To address these questions, the generalized Eckart-Young Theorem[54] suggests that the error introduced in the approximation of finite and deduced Koopman operators can be minimized by setting $\boldsymbol{\zeta_0}$ and $\boldsymbol{\zeta_1}$ as the top left and right singular functions of the true Koopman operator. To identify the most optimal feature transformation functions, serving as the left and right singular functions of the true Koopman operator, the VAMP theory could be employed. The VAMP theory introduces the VAMP-r score, a summation of the singular values to the power of r of the approximated Koopman operator, which facilitate s the measurement of similarity between the estimated singular functions and the ground truth. A higher VAMP-r score indicates a more accurate approximation of the singular functions. Specifically, the VAMP-2 score is the most commonly used and well-defined, and it is also employed in the implementation of lots of algorithms. Then the question of optimization of low-rank representations could be expressed as a maximization problem with constraints:

$$\arg\max \sum_{i=1}^{k} \sigma_i^2 = \arg\max_{\boldsymbol{\zeta_0}, \boldsymbol{\zeta_1}} \sum_{i=1}^{k} \langle \zeta_{0i}, \mathbf{K}_{\tau} \zeta_{1i} \rangle_{\rho_0}^2 \tag{2.30}$$

$$s.t. \quad \langle \zeta_{0i}, \zeta_{0j} \rangle_{\rho_0} = \delta_{ij} \quad \langle \zeta_{1i}, \zeta_{1j} \rangle_{\rho_1} = \delta_{ij} \tag{2.31}$$

In practice, to solve this problem, we can employ a large amount of trial functions (e.g. linear combinations of the internal coordinates of molecules) to approximate the singular functions. By utilizing the VAMP theory, we can identify the most optimal approximation. However, obtaining the Koopman operator in prior is typically not feasible in real implementations. For instance, in MD simulations, this underlying dynamical propagator is implicitly dependent on the initial setup, including the force field, periodic boundary conditions, etc. If the time sequences of a set of basis functions are observable or can be simulated, the Koopman operator under these normalized basis functions can be estimated using the time-correlation matrix[34]:

$$\overline{\mathbf{K}}_\tau = \mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{C}_{01} \mathbf{C}_{11}^{-\frac{1}{2}} \tag{2.32}$$

Here, the correlation matrices are estimated from the observable of basis functions:

$$\mathbf{C}_{00} = \mathbb{E}_t[\boldsymbol{\zeta_0}(\boldsymbol{x}_t)\boldsymbol{\zeta_0}(\boldsymbol{x}_t)^T] \tag{2.33}$$

$$\mathbf{C}_{01} = \mathbb{E}_t[\boldsymbol{\zeta_0}(\boldsymbol{x}_t)\boldsymbol{\zeta_1}(\boldsymbol{x}_{t+\tau})^T] \tag{2.34}$$

$$\mathbf{C}_{11} = \mathbb{E}_t[\boldsymbol{\zeta_1}(\boldsymbol{x}_{t+\tau})\boldsymbol{\zeta_1}(\boldsymbol{x}_{t+\tau})^T] \tag{2.35}$$

And VAMP-2 score could be thereby calculated through:

$$\hat{R}_2 = ||\mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{C}_{01} \mathbf{C}_{11}^{-\frac{1}{2}}||_F^2 = ||\overline{\mathbf{C}}_{00}^{-\frac{1}{2}} \overline{\mathbf{C}}_{01} \overline{\mathbf{C}}_{11}^{-\frac{1}{2}}||_F^2 + 1 \tag{2.36}$$

where the matrices with bars represent the remove-mean covariance matrices. There are various algorithmic approaches to accomplish this process. The most common method involves pre-selecting a set of basis functions and utilizing the VAMP-2 score to optimize the linear combination coefficients of the functions for approximating the singular functions. This is referred to as time-lagged canonical correlation analysis algorithm[55]. If we additionally enforce the detailed balance constraint for the dynamical process, we can employ a self-adjoint operator (i.e., transfer operator) to propagate the dynamics, leading to the derivation of the time-lagged independent component analysis algorithm[47].

In the case of VAMPnet[34], a neural network that maps high-dimensional continuous state variables directly to metastable state assignments, a more general approach is used, where the basis

functions can be parameterized. This was accomplished through the utilization of two parallel deep neural networks that facilitate nonlinear transformations of the input simple basis functions. By minimizing the loss function, which maximizes the VAMP-2 score, the trained neural network can then function as the optimal basis functions to approximate the true singular functions of the Koopman operator. In the implementation of VAMPnet, encoder neural networks are utilized for non-linear transformations, where the number of neurons decreases gradually layer by layer. Softmax functions are connected to the output layer, allowing for continuous non-linear functions to approximate the singular functions. This setup facilitates the establishment of a direct mapping between conformations and the probabilities of belonging to specific states.

Additionally, by removing the Softmax activation function from the last layer and incorporating the detailed balance condition for the VAMP-2 score, state-free reversible VAMPnets (SRVnets)[37] can be implemented. This network has the capability to identify continuous collective variables for reversible dynamical systems by utilizing non-linear transformed input features.

2.6 Protocol and Cutting-Edge Algorithms for MSM Construction

As mentioned in the previous sections, the projection operator theory and VAMP theory provide solid mathematical foundations for construction of MSMs and the identification of collective variables. Over the past decade, numerous methods and algorithms have been developed and implemented based on these theories and other physical insights[27–29]. This has led to the creation of well-known and user-friendly packages such as MSMBuilder[38] and PyEMMA[39]. In this section, we summarize a general workflow that integrates many recently developed techniques to investigate conformational changes in multi-body systems.

The complete workflow schematic is illustrated in Figure 2.1. First, unbiased MD simulations need to be conducted. Since MSMs can integrate an ensemble of parallel short trajectories to infer long-term dynamical information, performing ultra-long simulations is unnecessary. Instead, multiple simulations initiated from diverse configurations that subsequently sample the local equilibrium distribution reversibly are ideal for constructing MSMs. Next, representative internal

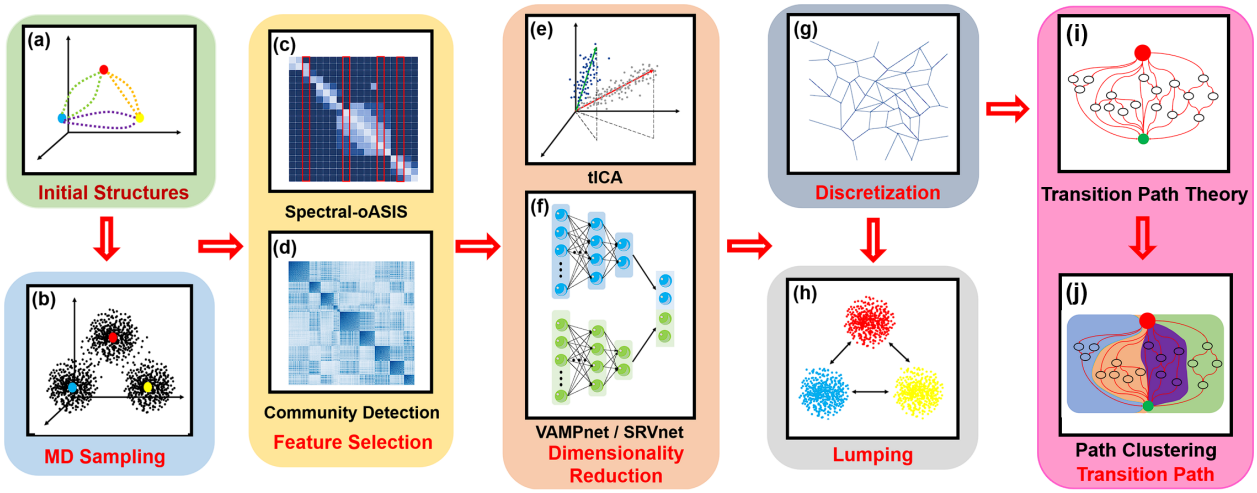

Figure 2.1: The general workflow for MSM construction to investigate conformational changes of multi-body systems. (a-b) First the MD simulations are shooted from multiple diverse initial structures. (c-d) And then the MD configurations are embeded by representative features. (e-f) The feature-embedded configurations are further projected on few collective variables, where the geometric distance represents the kinetic distance. (g) The projected configurations are clustered into microstates and microstate-MSM is constructed and validated. (h) The microstates are lumped into interpretable macrostates to comprehend the underlying mechanisims and dynamics. (i-j) The kinetic transition pathways can be identified based on the microstate-MSM and further classified into explainable metastable path channels.

features are typically selected to embed each MD configuration. By identifying the collective variables from these features, each MD configuration can be projected onto a low-dimensional space, where the geometric distance corresponds to the kinetic difference. The projected configurations are then coarse-grained into microstates, and the MD trajectories can be converted into transitions between these microstates. The microstate-MSM can subsequently be constructed and validated. To interpret the mechanisms and kinetic rates underlying the conformational changes, microstates can be lumped into a few metastable macrostates, and transition path analysis can be performed with the validated microstate-MSM. For conformational changes in multi-body systems, it is typically observed that multiple parallel kinetic pathways exist, each exhibiting comparable reactive fluxes [18, 48, 56]. Therefore, further classification of pathways into a few metastable path channels may enhance the understanding of the underlying mechanisms.

2.6.1 Explore Diverse Initial Configurations

Preparing diverse initial configurations for unbiased MD simulations is crucial for extensively sampling the wider energy landscape. This step can typically be guided by both experimental structures and the integration of various simulation techniques. Experimental methods, such as X-ray crystallography[57], cryo-electron microscopy[58], and nuclear magnetic resonance spectroscopy[21], can provide single or multiple structures that serve as initial seeds for simulations. However, running unbiased MD simulations directly from a limited number of structures may lead to many simulations being trapped in local equilibria, missing the broader scope of the energy landscape. Therefore, various simulation tools have been incorporated with experimental structures to provide a more comprehensive range of structures. For example, enhanced sampling methods such as Metadynamics[23], Hamiltonian Replica Exchange Molecular Dynamics[24], and Temperature Accelerated Molecular Dynamics[59] have proven highly effective in sampling configurations on a global scale. Additionally, coarse-graining MD simulations[60] can be employed to unbiasedly and extensively explore the energy landscape at much greater speed. However, accurately backmapping from coarse-grain structures to all-atom structures remains an ongoing challenge[61]. Traditional methods, such as performing restrained MD simulations based on coarse-graining

structures, along with newly developed deep learning approaches, are being considered for this purpose.

Targeting different aims, other purely computational approaches can also provide reasonable and insightful initial structures. For example, to study protein-protein or protein-ligand complexes, rigid body docking holds great potential for generating diverse initial structural ensembles[62]. And it is easy to incorporate human intuition or experimental observations as constraints during this process[3]. Moreover, with the rapid expansion of deep learning, increasingly powerful models have been developed to generate diverse 3D all-atom structures, such as the series of AlphaFold models[63, 64]. Recently, some generative models have even been shown to produce structures consistent with the Boltzmann distribution, for both biomolecular systems and glassy systems[65–67].

However, global exploration of the energy landscape can sometimes yield unwanted structures, such as the unfolding or dissociation of the entire system, which are far removed from the conformational changes of interest. In such cases, elucidating localized initial pathways between initial, target, or intermediate conformations becomes important[28, 29]. These initial pathways can be generated using various computational algorithms, including the Climber algorithm[68] and the string method[69]. By further optimizing conformations along the initial pathway, the minimum free energy path and important intermediate states can be elucidated. Conducting unbiased MD simulations from different points along this optimal pathway can effectively sample the conformational changes of interest and generate an appropriate input dataset for subsequent MSM construction[28, 29].

2.6.2 Embed Conformations into Representative Features

After obtaining sufficient MD samples of the process we are interested in, the next step is typically to extract appropriate internal features or coordinates to embed the MD configurations. A common strategy for studying simple conformational changes in single-body systems, such as protein folding, is to choose internal pairwise distances, dihedral angles, or the Root Mean Square Distance (RMSD), which are invariant to overall translation and rotation. However, unlike

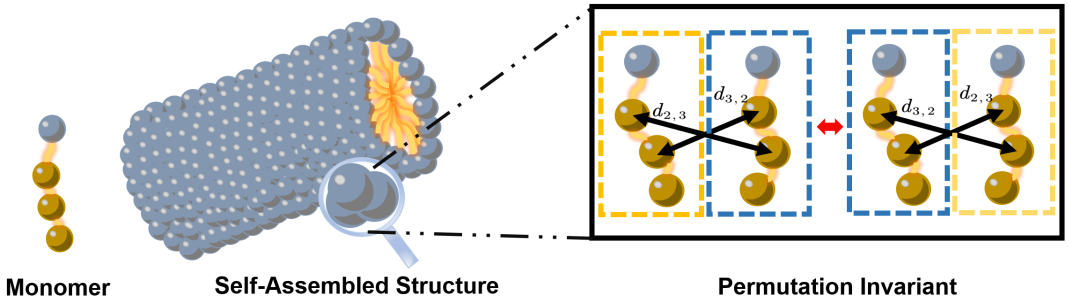

Figure 2.2: A schematic model to illustrate the challenge associated with permutation in multi-body system. The monomer is composed of four particles (the gray sphere represents a hydrophilic particle, and the orange spheres represent hydrophobic particles). When two identical monomers exchange positions, the inter-atom distances $d_{2,3}$ and $d_{3,2}$ change correspondingly, while the overall configuration remains identical. This figure is reproduced from Liu. *et al.*[13]

single-body systems, this step is more difficult and crucial for studying conformational changes in multi-body systems, as there are many additional factors that need to be considered[13, 18]. For instance, a major issue in many multi-body systems (e.g. bio-condensate, supercooled liquid and self-assembly) is that the features should be invariant not only to the translation and rotation of the entire system but also to the permutations of monomers[13, 18]. Additionally, feature construction often requires more question-driven intuition. Sometimes inter-molecular features are more important, while other times, intra-molecular features take precedence. Meanwhile the conformational changes of multi-body systems may involve various collective motions and multi-body interactions[18, 19], so utilizing different kinds of features and considering longer-range interactions can be crucial. Next, various challenges and approaches for feature embedding in multi-body systems are discussed.

Identify permutation-invariant features to describe multi-body structures. In many multi-body systems, monomers are permutable, i.e., exchanging two monomers results in an indistinguishable overall structure. Therefore permutation-invariant features should be employed to study the conformational changes of the entire system[13]. As illustrated in Figure 2.2, each monomer in the multi-body system consists of a chain of four particles. If two monomers are exchanged, the dimer

configurations remain unchanged, but the pairwise distances between the monomers vary with the permutation, making these distances unsuitable as features. To construct permutation-invariant features, three different approaches can be utilized.

One approach is to choose physical structural order parameters or structural descriptors that remain invariant under permutations. Numerous order parameters have been developed in the fields of nucleation and supercooled liquids to encode local structures of monomer, independent of the permutation of surrounding monomers[42, 44, 45, 70]. For example, the coordination numbers and associated Gaussian expanded moments can precisely capture a monomer's local radial density distribution. Steinhardt bond order parameters, using spherical harmonics of different orders, effectively distinguish between ordered and amorphous structures and capture orientational structure differences. Pair entropy order parameters can approximate radial and orientational entropies from radial distribution functions. Additionally, averaging these order parameters for a single monomer with its surrounding monomers incorporates longer-range information. After obtaining these order parameters for each monomer, arranging them in a specific order or taking the pooling average can achieve permutation-invariant structural embedding for the entire system.

An alternative approach is to use non-physical coordinates that represent the overall structure of the system. For instance, the number of monomers in each aggregate and morphology parameter have been demonstrated to serve as explainable and effective coordinates for studying the conformational changes of self-assembly[17]. Additionally, solvent-accessible surface area and modified RMSD have been employed to investigate the kinetics of amphiphilic lipid aggregation[71]. The modified RMSD is invariant to the permutations of indistinguishable particles and can be constructed using the permutation matrix $\mathbf{P}$: $RMSD(A,B) = \frac{1}{\sqrt{N}} min_{\mathbf{P},\mathbf{D},\mathbf{L}} |\mathbf{R}^A - \mathbf{P}(\mathbf{R}^B - \mathbf{D} - \mathbf{L})|$, where $\mathbf{D}$ and $\mathbf{L}$ denote translational and periodic movement, respectively. Moreover, in the graph-based method, aggregated structures are represented by undirected graphs with sub-units as nodes and strong interactions between pairs of sub-units as edges. The number of nodes and edges could be directly used as features for subsequent analysis.

Another recently developed approach involves adopting deep neural networks with specific architectural settings. For example, graph neural networks (GNN)[72] and equivariant neural networks (ENN)[73] have shown great capability in embedding permutation-invariant features. In the implementation of GNN, identical features are initially encoded on permutable nodes. This ensures that after graph convolutional operations and pooling of the entire graph, the output features remain invariant to permutations. ENN, which are designed to maintain the symmetries present in the data, can also effectively embed permutation-invariant features by ensuring that the output remains consistent under permutations of the input nodes. Currently, the development of various approaches to identify permutation-invariant features for ordered and amorphous microscopic structures is still an active and ongoing field.

Classify features to elucidate key collective motions. Due to the complexity of conformational changes in multi-body systems, various collective motions typically exist. Therefore, it is crucial to classify a large set of features and identify those representing the targeted or relevant dynamical motions and filter out noise features that do not show any correlation with others. The MoSAIC algorithm[74] has been developed to identify feature communities where features exhibit high correlation, reflecting collective motions within a multi-body system. This algorithm helps in filtering out features that either remain constant or vary randomly during dynamic processes. Meanwhile, the identified communities offer clear interpretations of collective dynamics, enabling the incorporation of biological intuition into feature selection. Implementing the MoSAIC algorithm involves two main steps: first, measuring the correlation between features and averaging this over the ensemble of trajectories; and second, clustering the features based on these correlations. Pearson correlation is used in the first step, while the Leiden community detection algorithm is employed in the second. The Leiden algorithm views features as nodes and their correlations as edges in a graph, performing clustering to optimize an objective function. The Constant Potts Model (CPM) is used as the objective function, defined as follows:

$$\Phi = \sum_{c} \left[e_c - \gamma \binom{n_c}{2} \right] \tag{2.37}$$

Here, e_c and n_c represent the total edge weights and the number of nodes within cluster c, respectively, while the binomial term $\binom{n_c}{2} = \frac{n_c^2 - n_c}{2}$ indicates the number of possible edges within c. The parameter γ is a hyperparameter that determines the minimum average correlation within clusters. A higher γ value requires a larger average correlation within communities, resulting in more, smaller communities, whereas a lower γ value produces larger but fewer communities.

Select features to capture the slowest dynamics. One alternative criteria used to extract important features is to identify the most effective features for accurately capturing the slowest transition modes in the dynamical multi-body system. The Spectral-oASIS algorithm[75] was designed based on this idea. The Spectral-oASIS algorithm operates through an iterative process: (1) First, k features are randomly chosen, and the sub-correlation matrix is computed from these selected features, $\mathbf{C}_k = \mathbf{C}[:, S] \in \mathbb{R}^{n \times k}$. The full correlation matrix is then reconstructed using: $\tilde{\mathbf{C}} = \mathbf{C}_k \mathbf{W}_k^{-1} \mathbf{C}_k^T$, where $\mathbf{W} = \mathbf{C}_k[S, :]$ is an invertible matrix. (2) Next, the selection criteria, defined by the diagonal reconstruction errors of the matrix weighted by the corresponding eigenvectors, are used to choose features for the next iteration with the aim of minimizing these reconstruction errors. (3) By iteratively applying this process, features are identified that not only best approximate the eigenvalues of the correlation matrix but also effectively distinguish different metastable states within the dynamical system. As the Nyström algorithm is applicable only to symmetric positive semi-definite matrices, and the correlation matrix is considered an approximation for the time-lagged correlation matrix, Spectral-oASIS can also be comprehended through the VAC[47, 51, 52].

2.6.3 Identify Low-dimensional Collective Variables

Typically, the aforementioned feature construction approaches will generate a large number of features, making comprehension challenging. The next step is usually to identify a much lower-dimensional set of collective variables based on these features. Over the last decade, various unsupervised algorithms with different objective functions have been developed. For example,

principal component analysis (PCA) has been developed to construct principal components by linearly combining input features, such that the projected data exhibits the largest variance. These components are considered the most informative collective variables, minimizing the error in reconstructing the original high-dimensional features from them. Similarly, the autoencoder neural network has been designed to achieve this goal through non-linear transformation of input features using trainable neural networks. However, a drawback of PCA and autoencoders is that they do not utilize any time-sequence information from the unbiased simulations, which can sometimes cause the resulting collective variables to overlook significant slow transitions.

To elucidate collective variables that represent the slowest dynamics, many algorithms have been developed within the framework of VAC[47, 51, 52]. One of the most classical and widely used methods is time-lagged independent component analysis (tICA)[47, 76]. Considering the system's evolution as a time series of state variables $\{\boldsymbol{x}_t\}_{t=0}^{T} \in \mathbb{R}^{3N}$, the feature embedding can be expressed using a set of non-linear functions for each time point, resulting in mean-free features $\{f_i(\boldsymbol{x}_t)\}_{i=1}^{K}$. To perform tICA algorithm, the time-lagged correlation matrix $\mathbf{S}(\tau)$ and the covariance matrix $\mathbf{\Sigma}$ need to be estimated based on the set of features:

$$\mathbf{S}(\tau)_{ij} = \frac{1}{N_T - N_\tau} \sum_{l=1}^{N_T - N_\tau} f_i(\boldsymbol{x}(l)) f_j(\boldsymbol{x}(l + N_\tau)) \tag{2.38}$$

$$\mathbf{\Sigma}_{ij} = \frac{1}{N_T} \sum_{n=1}^{N_T} f_i(\boldsymbol{x}(n)) f_j(\boldsymbol{x}(n)) \tag{2.39}$$

where $\tau = N_\tau \Delta t$ (Δt denotes the time interval of the trajectory) and N_T represents the length of the trajectory. The VAC framework suggests that the leading eigenvectors obtained from the eigen-decomposition problem $\mathbf{S}(\tau)\psi_i = \mathbf{\Sigma}\psi_i\lambda_i$ should provide the coefficients for the linear combinations of features that optimally approximate the slowest dynamical modes. In practice, to enforce the detailed balance condition, the time-lagged correlation matrix is usually symmetrized by averaging it with its transpose. By further combining tICA with diffusion map theory, the collective variables can be re-scaled, allowing the Euclidean distance on these collective variables to approximate kinetic distances[76]. In the line with tICA, SRVnets use two trainable encoders to first perform a non-linear transformation of the features and then linearly combine the outputs to

derive collective variables[37]. The SRVnet training objective function is designed to maximize the time-lagged correlations of the collective variables using the VAMP-2 score, while also enforcing the detailed balance constraint. It is worth noting that in the case of conformational changes in multi-body systems, processes sometimes may prefer to be unidirectional; for instance, dissociation in self-assembly are often much slower than associations. This can cause the MD simulation data to be biased or 'out-of-equilibrium' to sample the less favorable processes. To address this challenge, the above-mentioned VAMP theory can be used to construct VAMPnet[34], which is good at performing dimensional reduction for non-reversible Markovian processes.

In addition to the aforementioned methods, several other algorithms based on the VAC framework are available for dimensionality reduction, including kernel-tICA[77], deep-tICA[78], time-lagged autoencodersc[55], and variational dynamics encoders[79]. Moreover, various other algorithms have been developed based on different theories or criteria, such as the past-future information bottleneck[80], state predictive information bottleneck[81], transition manifold methods[82], reaction coordinate flow[83], and relaxation mode analysis[84], among others. In this thesis, we will showcase the effectiveness and advantages of various algorithms across different systems.

2.6.4 Construct and Validate Microstate-MSMs

After projecting the MD configurations onto collective variables, we can cluster them into hundreds or thousands of microstates and model the transition dynamics between these states. As mentioned above, the Euclidean distance between projected conformations in the collective variable space can be considered a good approximation of the kinetic differences between them[76]. Therefore, coarse-graining the projected MD conformations using distance metrics or density distribution in the collective variable space should be effective. Commonly used centroid-based algorithms include the K-Centers algorithm, K-Means algorithm, and K-Medoids algorithm[28, 85]. These unsupervised algorithms define different objective functions to optimize, identify the landmarks (i.e., center conformations or geometric means) for various clusters, and then assign the conformations to the cluster with the nearest landmark. Specifically, the K-Centers algorithm aims to minimize the maximum inter-cluster distance, resulting in more uniform clusters with similar

sizes, while the K-Means algorithm seeks to minimize the total within-cluster variance, leading to more clusters in the dense regions. The K-Medoids algorithm is in line with the K-Means algorithm but designates actual data points as cluster centers instead of the geometric means. However, these algorithms have certain shortcomings. For instance, the number of clusters must be specified as an input parameter beforehand, and the algorithms perform poorly when the metastable regions in the collective variable space are not convex.

In addition to centroid-based algorithms, other clustering methods are available, such as Automatic State Partitioning for Multibody Systems (APM)[19], Density-Based Spatial Clustering of Applications with Noise (DBSCAN)[86], and Adaptive Partitioning by Local Density-Peaks (APLoD)[87]. In particular, the APM algorithm iteratively splits and lumps clusters to determine the appropriate number based on the lifetimes of clusters. The APLoD algorithm automatically identifies the optimal number of clusters by analyzing local density variations within the data, making it effective at identifying and characterizing metastable regions in high-dimensional spaces, even when these regions are non-convex. These algorithms are often more suitable for studying the heterogeneous dynamics within multibody systems.

After clustering the projected MD conformations into microstates, we can construct MSMs based on these states. The crucial step in building MSMs is estimating the transition probabilities with a lag time τ from the MD trajectories:

$$\mathbf{T}_{ij} = p(\boldsymbol{x}(t+\tau) \in j | \boldsymbol{x}(t) \in i) = \frac{\boldsymbol{C}_{ij}(\tau)}{\sum_j \boldsymbol{C}_{ij}(\tau)} \tag{2.40}$$

As indicated in the equation above, the most straightforward method for estimating transition probabilities is to normalize the transition count matrix (TCM) $\mathbf{C}(\tau)$. Each element of the TCM, $\mathbf{C}_{ij}(\tau)$, is constructed by counting the number of transitions from state i to state j across all trajectories. To enforce the detailed balance constraint, the TCM is typically averaged with its transpose using a brute-force approach:

$$\mathbf{C}^{symmetry}(\tau) = \frac{\mathbf{C}(\tau) + \mathbf{C}^T(\tau)}{2} \tag{2.41}$$

Meanwhile, there are many alternative ways to construct the transition probability. For example, the maximum-likelihood estimator (MLE)[27] for a reversible MSM assumes that the observed

transitions within the trajectories are independent and originate from the underlying equilibrium distribution, thereby enabling the use of detailed balance condition as a constraint. In particular, the likelihood function for detecting transition pairs within the simulations in the MLE framework can be expressed as:

$$p(\mathbf{T}|\mathbf{C}^{obs}) \propto \prod_{i,j=1}^{n} \mathbf{T}_{ij}^{\mathbf{C}_{ij}^{obs}} \tag{2.42}$$

By further incorporating the constraint $\pi_i \mathbf{T}_{ij} = \pi_j \mathbf{T}_{ji}$, the Lagrange approach can be used to derive a self-consistent expression for the state equilibrium population and transition probabilities:

$$\pi_i = \sum_j \frac{\mathbf{C}_{ij} + \mathbf{C}_{ji}}{\frac{N_j}{\pi_j} + \frac{N_i}{\pi_i}} \tag{2.43}$$

$$\mathbf{T}_{ij} = \frac{\mathbf{C}_{ij} + \mathbf{C}_{ji}\pi_j}{N_j\pi_i + N_i\pi_j} \tag{2.44}$$

where π_i denotes the equilibrium population for state i and $N_i = \sum_j \mathbf{C}_{ij}$. Additionally, reversible Koopman's method[88] and various estimators designed to construct MSMs from biased simulations (e.g., transition-based reweighting analysis methods[89]) could also be very useful to study the conformational changes of the multi-body systems.

One of the most crucial hyperparameters in constructing an MSM is the lag time. The lag time must be sufficiently long to allow intra-state dynamics to relax to equilibrium and for inter-state transitions to become Markovian, thereby ensuring that the MSM can provide accurate long-term predictions[27]. Typically, the shortest lag time that ensures inter-state transitions are Markovian is chosen as the Markovian lag time. Two common approaches are used to qualitatively evaluate the Markovianity of the model. The first is implied time scale (ITS) analysis, which is calculated using the eigenvalue of the TPM for each dynamical mode:

$$t_i(\tau) = -\frac{\tau}{ln\lambda_i(\tau)} \tag{2.45}$$

where $\lambda_i(\tau)$ is the i^{th} eigenvalue of TPM with lag time of $m\tau$. If the model is Markovian and satisfies the first-order master equation $\lambda_i(m\tau_M) = \lambda_i^m(\tau_M)$, the ITS will converge to value $-\frac{\tau_M}{ln\lambda_i(\tau_M)}$ and will be independent of the lag time. Therefore, the lag time at which the largest ITS reaches a plateau can be considered the shortest Markovian lag time. Additionally, the Chapman-Kolmogorov (CK) test could also be used to assess the Markovianity of the model. This test checks

the consistency between the long-term transition probabilities predicted by the MSM, $\mathbf{T}^m(\tau_M)$, and those directly estimated from the raw MD data with the same lag time, $\mathbf{T}(m\tau_M)$. When performing the CK test, it is important to account for statistical errors introduced by MD sampling, which can be estimated using bootstrapping of the trajectories.

Moreover, it is important to carefully choose various other hyperparameters during the construction of MSMs, such as the number of collective variables and the number of states etc. The VAC[47, 51, 52] and VAMP[34, 47, 53] theories introduced earlier provide useful scoring functions to evaluate the quality of MSMs. The Generalized Matrix Rayleigh Quotient (GMRQ) score[90] (based on VAC) and the VAMP-2 score (based on VAMP) are effective criteria for assessing the model's ability to approximate the leading eigenvectors or singular functions of the dynamical propagator and capture the slowest timescales. Additionally, these scores can be used within a cross-validation framework to balance the model's systematic and statistical errors. For cross-validation with GMRQ, we divide the dataset, which consists of multiple parallel trajectories, into k folds. We use $k-1$ folds as the training set to construct MSMs and the remaining fold as the test data. The GMRQ scores for both the training and testing datasets are defined as follows:

$$\mathbf{GMRQ}_{train} = k^{-1} \sum_{i=1}^{k} \mathbf{Tr}\left(\mathbf{A}^{(-i)T}\mathbf{S}^{(-i)}\mathbf{A}^{(-i)}\left(\mathbf{A}^{(-i)T}\mathbf{\Sigma}^{(-i)}\mathbf{A}^{(-i)}\right)^{-1}\right) \tag{2.46}$$

$$\mathbf{GMRQ}_{test} = k^{-1} \sum_{i=1}^{k} \mathbf{Tr}\left(\mathbf{A}^{(-i)T}\mathbf{S}^{(i)}\mathbf{A}^{(-i)}\left(\mathbf{A}^{(-i)T}\mathbf{\Sigma}^{(i)}\mathbf{A}^{(-i)}\right)^{-1}\right) \tag{2.47}$$

where $\mathbf{S}^{(-i)}$ and $\mathbf{S}^{(i)}$ denote the time-lagged correlation matrices computed from the training and validation dataset (made symmetric to adhere to the detailed balance constraint), respectively. Similarly, $\mathbf{\Sigma}^{(-i)}$ and $\mathbf{\Sigma}^{(i)}$ is the autocorrelation matrices estimated from the training and validation datasets. A^{-i} are the leading eigenvectors of the trained model: $\mathbf{S}^{(-i)}\mathbf{\Sigma}^{(-i)-1}$. Additionally, other scoring functions derived from different perspectives should also be used to evaluate the quality of MSMs and perform cross-validation. These may include measures of metastability[91] and Shannon entropy of the metastable states[92], among others. We will demonstrate the untilization details of more scores in the following chapters.

2.6.5 Lump Microstates to Macrostates to Comprehend the Mechanisms

Microstate-MSMs can be constructed with relatively short lag times, making its construction feasible for most short MD simulations. However, having hundreds or even thousands of microstates can make it challenging to understand the underlying mechanisms. To create a more interpretable and explainable kinetic model, microstates are typically grouped into a smaller number of metastable macrostates based on their kinetic connectivity[28, 29]. Many algorithms have been developed for this step, each with its own pros and cons.

The Perron Cluster Cluster Analysis (PCCA) and Robust Perron Cluster Analysis (PCCA+) algorithms are the most widely used[93, 94]. The PCCA algorithm uses the sign patterns of the leading eigenvectors of the microstate-MSM to define a crisp assignment of microstates to metastable macrostates, optimizing the approximation of the slowest dynamical modes. PCCA+ improves upon this with a fuzzy approach, simultaneously considering multiple leading eigenvectors and optimizing the membership vectors of microstates for each macrostate to be as close as possible to binary values (0 and 1). This provides better performance for dealing with microstates around transition regions. Additionally, many other algorithms have been developed for kinetic lumping and have demonstrated good performance under various criteria. These include the Bayesian Agglomerative Clustering Engine (BACE)[95], Gibbs Lumping algorithm[96], Most Probable Path (MPP) algorithm][97], and Hierarchical Nystrom Expansion Graph (HNEG) algorithm[98]. In this thesis, we will also showcase the state-of-the-art performance of our recently developed information bottleneck approach by comparing it with the PCCA+ and MPP methods.

After assigning the microstates to metastable macrostates, we can construct a macrostate-MSM with similar steps as the microstate-MSM. However, with fewer states, longer lag times are required to allow intra-state dynamics to relax and ensure that inter-state transitions are Markovian, which is often not feasible with relatively short trajectory lengths. To investigate these non-Markovian transition dynamics with short lag times, one could apply our recently developed non-Markovian dynamics models[30–32] (will be introduced in next chapter). Unlike conventional

MSMs, which rely on the first-order master equation, the IGME model uses the generalized master equation, offering a more accurate approach by incorporating time-dependent memory kernel functions to describe the dynamics.

2.6.6 Analyze Transition Pathways

Transition Path Theory (TPT)[40, 41] applied to an MSM offers significant potential for deriving the complete ensemble of kinetic pathways from MD simulations. After constructing and validating the MSM, and identifying the initial source states I and final sink states F for the transition pathways, the committor probabilities can be determined from the TPM of the MSM. The forward committor probability q_i^+ is defined as the probability that, when the system is in state i, it will next reach a state in F rather than a state in I. Consequently, $q_i^+ = 0$ for all $i \in I$ and $q_i^+ = 1$ for all $i \in F$. With these definitions, the committor probabilities q_i^+ can be solved using a set of linear equations:

$$q_i^+ = \sum_j \mathbf{T}_{ij} q_j^+ \tag{2.48}$$

Furthermore, the backward committor probability q_i^- is the likelihood that the system was previously in set I rather than F when being at state i. Under the detailed balance condition, the backward committor probability is given by $q^- = 1 - q^+$. Notably, the forward committor probabilities from equation 2.48 can also be used to identify significant transition states, where the forward committor probability equals 0.5.

The flux f_{ij} along the edge i, j, contributing to the transition from I to F, can then be computed as:

$$f_{ij} = \pi_i q_i^- \mathbf{T}_{ij} q_j^+ \tag{2.49}$$

where π_i represents the stationary population of state i. To eliminate unnecessary detours and recrossings, the net flux matrix can be further defined as:

$$f_{ij,net} = max\{f_{ij} - f_{ji}, 0\} \tag{2.50}$$

Finally, using the net flux matrix, the Dijkstra algorithm[99] can be employed to reveal the most probable pathways from I to F. The detailed steps are as follows:

(i). Mark the initial states in set I as visited and all other states as unvisited.

(ii). Visit the state s_1 with the highest new flux from the initial states in I and designate I as the "parent" state of s_1.

(iii). Among the net flux values from any visited state to its unvisited neighbors, select the state s_n with the highest net flux f_{mn}, where s_m is a visited state and s_n is an unvisited state. Visit s_n and label s_m as the "parent" state.

(iv). Repeat step 3 until a final state in F is reached.

(v). Identify the pathway $P = \{p_1 \in I, p_2, \cdots, p_{l-1}, p_l \in F\}$ with the greatest minimal flux f using the backtracking method: trace the parent states consecutively from F to I and reverse the pathway found.

(vi). Remove the pathway and update the net flux matrix by subtracting the net flux of the pathway from every state in the pathway: $f^+_{p_j,p_{j+1}} \leftarrow f^+_{p_j,p_{j+1}} - f(j \in [1,(l-1)], j \in N)$.

(vii). Repeat steps 1-6 with the modified net flux matrix.

For multi-body systems, such as those involving heterogeneous aggregation, self-assembly, protein-ligand binding, and protein-protein interactions, conformational changes often lead to the identification of numerous parallel pathways with similar probabilities or fluxes[13, 48, 56]. To simplify the interpretation of the intrinsic mechanisms, it is essential to consolidate these kinetic pathways into a few metastable channels. Several path clustering algorithms have been developed to categorize multiple pathways. For instance, the path lumping algorithm [56] uses inter-path flux as the distance metric to group pathways. Additionally, data-driven methods utilizing dynamic time warping (DTW) [100] have been created to classify transition pathways, such as those involved in ligand-receptor binding. In the following chapters, we introduce an algorithm called Latent-space Path Clustering (LPC)[48]. This algorithm uses a variational autoencoder to learn low-dimensional representations of the high-dimensional kinetic distribution of pathways, which are then used for clustering. In particular, we highlight its effectiveness in studying various multi-body systems, including chromatin folding.

Chapter 3

Developing Non-Markovian Dynamic Models to Study Long-time Dynamics

This chapter is reproduced in part with permission from following publications:

[1]. Cao, S.; Qiu, Y; Kalin, M.; & Huang, X. Integrative generalized master equation: A method to study long-timescale biomolecular dynamics via the integrals of memory kernels. *The Journal of Chemical Physics* **2023**, 159, 134106.

[2]. Yik, A. K. H.; Qiu, Y; Unarta, I. C.; Cao, S.; & Huang, X. A step-by-step guide on how to construct quasi-Markov state models to study functional conformational changes of biological macromolecules. *In A Practical Guide to Recent Advances in Multiscale Modeling and Simulation of Biomolecules* (pp. 10-1). Melville, New York: AIP Publishing LLC.

Qiu, Y contributed to the methodology developments, authored the codes, and polished the above papers for both the quasi-Markov State Model and the Integrative Generalized Master Equation method. Qiu, Y also rewrote and summarized the content in this chapter.

3.1 Nakajima-Zwanzig Equation

As mentioned in Chapter 2, the dynamical evolution of a conservative system can be described by the Liouville equation in the phase space. When the system is observed through a few degrees of freedom, the dynamics are projected onto a low-dimensional subspace. In this section, the general analytical equation describing the low-dimensional projected dynamics is derived. The projection of dynamics can be expressed by applying the general projection operator $\mathbb{P}$ on the Liouville equation:

$$\mathbb{P}\frac{\partial}{\partial t}\rho(t) = \mathbb{P}\mathcal{L}(\mathbb{P}+\mathbb{Q})\rho(t) \tag{3.1}$$

where $\mathbb{Q} = \mathbb{I} - \mathbb{P}$ is the complementary operator of $\mathbb{P}$. Meanwhile, one could also operate $\mathbb{Q}$ on the Liouville operator as:

$$\mathbb{Q}\frac{\partial}{\partial t}\rho(t) = \mathbb{Q}\mathcal{L}(\mathbb{P} + \mathbb{Q})\rho(t) \tag{3.2}$$

By performing the Laplace transformation on equation 3.2, we can solve the dynamics projected onto the $\mathbb{Q}$ subspace, which can be expressed as:

$$\mathbb{Q}\rho(t) = e^{\mathbb{Q}\mathcal{L}t}\mathbb{Q}\rho(0) + \int_0^t e^{\mathbb{Q}\mathcal{L}(t-s)}\mathbb{Q}\mathcal{L}\mathbb{P}\rho(s)ds \tag{3.3}$$

Plugging equation 3.3 back to equation 3.1 results in the expression for dynamics projected onto the $\mathbb{P}$ subspace:

$$\frac{\partial}{\partial t}\mathbb{P}\rho(t) = \mathbb{P}\mathcal{L}\mathbb{P}\rho(t) + \mathbb{P}\mathcal{L}e^{\mathbb{Q}\mathcal{L}t}\mathbb{Q}\rho(0) + \int_0^t \mathbb{P}\mathcal{L}e^{\mathbb{Q}\mathcal{L}(t-s)}\mathbb{Q}\mathcal{L}\mathbb{P}\rho(s)ds \tag{3.4}$$

which is named Nakajima-Zwanzig equation[49, 101]. Clearly, from equation 3.3, we can see that the probability density projected onto the $\mathbb{Q}$ space at time t arises from two terms: (1). $e^{\mathbb{Q}\mathcal{L}t}\mathbb{Q}\rho(0)$, the initial projection onto the $\mathbb{Q}$ space, which further evolves to time t, and (2). $\int_0^t e^{\mathbb{Q}\mathcal{L}(t-s)}\mathbb{Q}\mathcal{L}\mathbb{P}\rho(s)ds$, the density initially projected onto the $\mathbb{P}$ space, which transitions back into the $\mathbb{Q}$ space during the interval from 0 to t. Additionally, the right hand side of equation 3.4 is consisting of three terms, showing different resource of influences on dynamics in $\mathbb{P}$ space: (1). $\mathbb{P}\mathcal{L}\mathbb{P}\rho(t)$, the Markov term, which represents the density information that always remains in the $\mathbb{P}$ space. (2). $\mathbb{P}\mathcal{L}e^{\mathbb{Q}\mathcal{L}t}\mathbb{Q}\rho(0)$, the inhomogenous term, represents the effect of the density initially projected onto the $\mathbb{Q}$ space, evolve to time t and projected back to the $\mathbb{P}$ space. (3). $\int_0^t \mathbb{P}\mathcal{L}e^{\mathbb{Q}\mathcal{L}(t-s)}\mathbb{Q}\mathcal{L}\mathbb{P}\rho(s)ds$, the non-Markovian memory term, represents the effect of the density transitioning to the $\mathbb{Q}$ space at time s, evolving within $\mathbb{Q}$ space, and then transitioning back to $\mathbb{P}$ space. Nakajima-Zwanzig equation provides a general analytical formula to describe the projected dynamics.

3.2 Generalized Master Equation (GME)

As mentioned in Chapter 2, constructing an MSM can be viewed as projecting the original dynamics in phase space onto a set of indicator functions $\{\chi_1(\mathbf{\Gamma}), \chi_2(\mathbf{\Gamma}), \cdots, \chi_n(\mathbf{\Gamma})\}$, which are

used to define the state decomposition:

$$\chi_i(\boldsymbol{x}) = \begin{cases} 1 & if \boldsymbol{x} \in \text{state} \quad i \\ 0 & if \boldsymbol{x} \notin \text{state} \quad i \end{cases} \tag{3.5}$$

Based on this definition for state decomposition, the projection from phase space to state space can be expressed using the generalized Hummer-Szabo projection operator[46]:

$$\mathbb{P} = \sum_{k=1}^{N} |\rho_{eq}(\mathbf{\Gamma})\chi_k(\mathbf{\Gamma})\rangle \cdot \pi_k^{-1}\langle\chi_k(\mathbf{\Gamma})| \tag{3.6}$$

where $\rho_{eq}(\mathbf{\Gamma})$ stands for the equilibrium distribution and π_k is the equilibrium population of state k. Applying this operator to the Nakajima-Zwanzig equation 3.4 and assuming the initial density distribution is at equilibrium (ignoring the inhomogenous term), one could obtain the generalized master equation (GME):

$$\dot{\mathbf{T}}(t) = \mathbf{T}(t)\dot{\mathbf{T}}(0) - \int_0^t \mathbf{T}(t-\tau)\mathbf{K}(\tau)d\tau \tag{3.7}$$

where $\mathbf{T}_{ij}(t) = \langle\chi_i(\boldsymbol{x})|e^{\mathcal{L}t}|\chi_j(\boldsymbol{x})\rho_{eq}\rangle\pi_j^{-1}$ is the transition probability matrix (TPM) of lag time t and the $\dot{\mathbf{T}}_{ij}(0) = -\langle\chi_i(\boldsymbol{x})|\mathcal{L}|\chi_j(\boldsymbol{x})\rho_{eq}\rangle\pi_j^{-1}$ is the time-derivative of the TPM at $t = 0$. The $\mathbf{K}_{ij}(t) = \langle\chi_i(\boldsymbol{x})|\mathcal{L}e^{\mathbb{Q}\mathcal{L}t}\mathbb{Q}\mathcal{L}|\chi_j(\boldsymbol{x})\rho_{eq}\rangle\pi_j^{-1}$ represents the time-dependent memory kernel matrix. For dynamical systems with a clear separation of timescales, the relaxation time of memory kernel τ_k (i.e., $\mathbf{K}(t \geq \tau_k)$=0) is typically short. Therefore, the shortest upper bound for the convolution integration term in the GME can be limited to a small time value. In this way, the GME could be rewritten as[30, 31]:

$$\dot{\mathbf{T}}(t) = \mathbf{T}(t)\dot{\mathbf{T}}(0) - \int_0^{min\{t,\tau_k\}} \mathbf{T}(t-\tau)\mathbf{K}(\tau)d\tau \tag{3.8}$$

3.3 quasi-Markov State Models (qMSMs)

The GME serves as an approach that can accurately describe the projected density evolution in the low-dimensional state space. In the line with the construction of conventional MSM, by conducting MD simulations, we can assign MD conformations into different pre-defined metastable

states and estimate the transition probabilities between these states at different lag times from the simulation trajectories, even without knowing the explicit analytical formula of the Hamiltonian and dynamical propagator. Employing GME as the equation to describe the dynamics (MSM utilizes first-order master equation), we can then compute the memory kernel matrices at different times and use it to accurately propagate long-term transition probabilities. Because the relaxation time of the memory kernel is typically much shorter than the Markovian lag time, the required length of MD segment data to construct GME models is significantly shorter than that needed to build MSMs with predictive power.

The quasi-Markov State Model (qMSM)[30] has been developed to provide an iterative scheme for solving the memory kernel matrices and propagating long-term dynamics based on GME. Specifically, qMSM expresses the the derivatives in the equation 3.8 through discretized differences and Riemann sum:

$$\dot{\mathbf{T}}(n\Delta t) = \mathbf{T}(n\Delta t)\dot{\mathbf{T}}(0) - \Delta t \sum_{m=1}^{min[n,\frac{\tau_k}{\Delta t}]} \mathbf{T}((n-m)\Delta t)\mathbf{K}(m\Delta t) \tag{3.9}$$

where Δt is the time step of the simulation. Rearranging the equation gives the expression for the time-dependent memory kernel matrices:

$$\mathbf{K}(n\Delta t) = \frac{\dot{\mathbf{T}}(0)\mathbf{T}(n\Delta t) - \dot{\mathbf{T}}(n\Delta t)}{\Delta t} + \sum_{m=1}^{n} \mathbf{K}(m\Delta t)\mathbf{T}((n-m)\Delta t) \tag{3.10}$$

Consequently, by employing the short-lag-time TPMs $\{\mathbf{T}^{MD}(m\Delta t)\}_{m=0}^{n}$ and their derivatives, i.e., $\dot{\mathbf{T}}^{MD}(m\Delta t) \approx [\mathbf{T}^{MD}((m+1)\Delta t) - \mathbf{T}^{MD}(m\Delta t)]/\Delta t$, which could be directly estimated from MD trajectories, one could calculate the short-lived memory kernel $\{\mathbf{K}(m\Delta t)\}_{m=0}^{\tau_k/\Delta t}$ with greedy algorithm in an iterative way. To characterize the memory relaxation time, the mean integral memory kernel (MIK) is defined as:

$$\mathcal{MIK}(t) = \frac{1}{N}\sqrt{\sum_{i,j=1}^{N}(\int_0^t \mathbf{K}_{ij}(\tau)d\tau)^2} \tag{3.11}$$

which could be employed to assess the cumulative contribution of the memory kernel by different time. Consequently, the time at which the MIK converges will be identified as the memory kernel

relaxation time. Using the calculated memory kernel, the long-term dynamics can be propagated with Equation 3.9. The qMSM has demonstrated great effectiveness in accurately predicting transition rates in studies of conformational changes in large biomolecular systems. However, in the implementation of qMSM, it is observed that the numerical derivations of TPMs lead to significant numerical fluctuations in the calculation of the memory kernel[30, 32]. To improve the model's robustness, we recently developed the integrative generalized master equation (IGME) method[31].

3.4 Integrative Generalized Master Equation (IGME) Models

Due to the complex derivatives and convolutional integration terms in the GME, obtaining an analytical solution is very difficult. The brute-force numerical derivative approach have been found to introduce considerable numerical fluctuations in calculation of memory kernel matrices. To construct a more robust model, our recently developed IGME method adopts the integration of memory kernels for the propagation of dynamics. In particular, considering the dynamics for $t \geq \tau_k$, $\mathbf{T}(t-\tau)$ in the convolutional integration can be expressed using Taylor expansion:

$$\int_0^t \mathbf{T}(t-s)\mathbf{K}(s)ds = \int_0^t \left[\mathbf{T}(t) + \sum_{n=1}^{\infty} \frac{1}{n!}\frac{d^n\mathbf{T}(t)}{dt^n}(-s)^n\right]\mathbf{K}(s)ds \tag{3.12}$$

$$= \mathbf{T}(t)\mathbf{M}_0(t) + \sum_{n=1}^{\infty} \frac{(-1)^n}{n!}\frac{d^n\mathbf{T}(t)}{dt^n}\mathbf{M}_n(t) \tag{3.13}$$

where $\mathbf{M}_n = \int_0^{\tau_k} \mathbf{K}(s)s^n ds$ is the integration of memory kernel for different orders. We note that since the relaxation timescale for the memory kernel is typically much shorter than the slowest timescale of conformational changes in multi-body systems, $\mathbf{M}_n$ would converge to a constant over a very short period. Under this circumstance, i.e., $t \geq \tau_k$, the GME could be expressed as an ordinary differential equation:

$$\mathbf{T}^{-1}(t)\frac{d}{dt}\mathbf{T}(t) = \dot{\mathbf{T}}(0) - \mathbf{M}_0 - \sum_{n=1}^{\infty}\left[\frac{(-1)^n}{n!}\mathbf{T}^{-1}(t)\frac{d^n}{dt^n}\mathbf{T}(t)\right]\mathbf{M_n} \tag{3.14}$$

By solving the equation in a self-consistent manner, an analytical solution can be obtained:

$$\mathbf{T}(t \geq \tau_k) = \mathbf{A}\hat{\mathbf{T}}^{\mathbf{t}} \tag{3.15}$$

$$\ln \hat{\mathbf{T}} = \dot{\mathbf{T}}(0) - \mathbf{M}_0 - \sum_{n=1} \frac{(-1)^n}{n!} (\ln \hat{\mathbf{T}})^n \mathbf{M}_n \tag{3.16}$$

Where $\mathbf{A}$ and $\hat{\mathbf{T}}$ are two integration constant matrices that can be estimated from the TPMs at short lag times, and the equation can be used to propagate dynamics at any time much longer than τ_k. The $\hat{\mathbf{T}}$ matrix represents the infinite-time dynamical behavior and can be used to calculate thermodynamic properties, such as stationary populations, as well as long-time kinetic properties, like implied timescales and mean first passage time. Additionally, the solution provides the second approach to calculate the MIK: the first-order time-integrated memory kernel matrix can be approximately obtained through $\mathbf{M}_0 \approx \dot{\mathbf{T}}(0) - \ln \hat{\mathbf{T}}$.

Equation 3.15 serves as the equation to evolve dynamics with the IGME model. The constant matrix $\mathbf{A}$ and $\hat{\mathbf{T}}$ are the most important parameters to construct the IGME model. In practice, we could estimated these two matrices from multiple short-lag-time TPMs $\{\mathbf{T}^{MD}(\tau_k + m\Delta t)\}_{m=0}^{L_{fit}}$ using the least square fitting approach. Specifically, we can firstly compute the logarithms of the TPMs estimated from the MD simulations $\{\ln \mathbf{T}^{MD}(\tau_k + m\Delta t)\}_{m=0}^{L_{fit}}$ using the spectrum-based method, which involves multiplying the original eigenvectors of the matrices by the logarithms of their eigenvalues. Then, we could perform the least squares fitting for each element of the logarithms of the TPMs, i.e., $[\ln \mathbf{T}^{MD}(t)]_{ij} = b_{ij} + k_{ij}t$, to generate the initial guess for matrix $\mathbf{A}^{init} = exp(\mathbf{b})$ and $\hat{\mathbf{T}}^{init} = exp(\mathbf{k})$. Furthermore, the steepest descent method can be employed to optimize these initial guesses, ensuring that the propagated TPMs have positive eigenvalues and satisfy the detailed balance condition. In particular, we employ the loss function for the optimization of $\mathbf{A}$ and $\hat{\mathbf{T}}$ as:

$$\mathcal{L}_{loss} = \sum_{t=\tau_k} ||\mathbf{T}^{MD}(t) - \mathbf{A}\hat{\mathbf{T}}^t||_F^2 + \alpha \sum_i \left(\sum_j \left[\mathbf{T}^{IGME}(t)\right]_{ij} - 1 \right) \tag{3.17}$$

$$+\beta ||\pi \mathbf{T}^{IGME}(t) - \mathbf{T}^{IGME}(t)^T \pi^T||_F^2 \tag{3.18}$$

where the subscript F denotes the Frobenius norm. And the second term of loss function gurantee the row-normalization requirement of TPMs, while the third term corresponds to the detailed balance requirement. One alternative approach to estimate the $\mathbf{A}$ and $\hat{\mathbf{T}}$ matrices with constraints is the utilization of Lagrange multipliers[102]:

$$\mathcal{L} = \frac{1}{2}\sum_{t} |\ln \mathbf{T}^{MD}(t) - \ln \mathbf{A} - t \ln \hat{\mathbf{T}}|_F^2 + \sum_i \left(\gamma_i \sum_j [\ln \hat{\mathbf{T}}]_{ij} \right) \tag{3.19}$$

where γ_i is the Lagrange multiplier which constrain the row-sum rule of TPMs. With the Lagrangian, we could obtain the estimation of $\mathbf{A}$ and $\hat{\mathbf{T}}$ matrices by solving a linear equation.

While τ_k serves as the lower bound for available time range of IGME, the size of TPMs (i.e., hyper-parameter L_{fit}) used to estimate the $\mathbf{A}$ and $\hat{\mathbf{T}}$ matrices could be optimized. To quantify the errors of IGME models in reproducing the simulation data, we have designed a criterion called the time-averaged root mean squared error (RMSE):

$$\mathcal{RMSE} = \sqrt{\frac{\sum_{n=1}^{L_x}\sum_{i,j=1}^{N}[\pi_i \mathbf{T}_{ij}^{MD} - \pi_i \mathbf{T}_{ij}^{IGME}]^2 dt}{N^2 L_x}} \tag{3.20}$$

Here, π_i represents the stationary population for state i, N is the number of macrostates, and L_x denotes the defined time range for computing RMSE. With this evaluation score, we can systematically and extensively scan multiple different τ_k and L_{fit} values to determine the optimal combination. Additionally, evaluating the RMSE for IGME models with varying numbers of states can help identify the optimal number of states, balancing state resolution and modeling accuracy. While increasing the number of states can enhance resolution, it also enlarges the dimensions of matrices $\mathbf{A}$ and $\hat{\mathbf{T}}$, necessitating the fitting of more elements with limited data and thereby introducing greater statistical errors.

3.5 Investigating the Dynamics of Complex Biomolecular Systems Using Non-Markovian Dynamic Models

Next, we demonstrate the effectiveness of non-Markovian dynamic models, specifically the IGME model, in elucidating the dynamics of a complex biomolecular system—the gate opening mechanism of Taq RNA polymerase (RNAP), as shown in 3.1(a). We conducted MD simulations

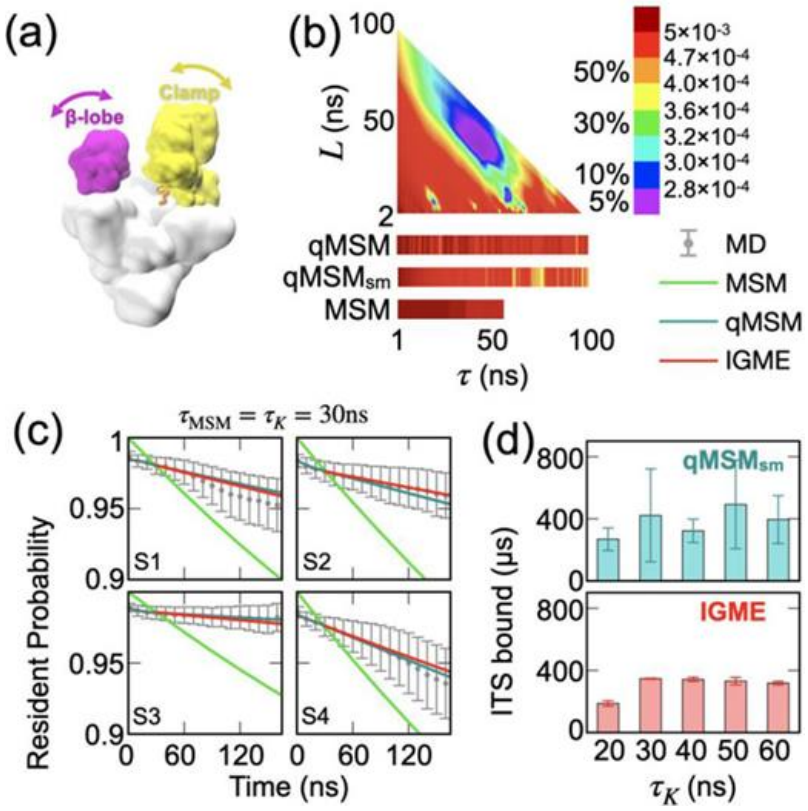

Figure 3.1: The IGME model of the gate opening dynamics of Taq RNAP. (a) The cartoon of the bacterial holoenzyme and its domains (yellow), clamp (magenta), β-lobe (yellow) and Switch two region (orange). (b) The RMSE heatmap of the IGME, qMSM, qMSMsm and MSM. The triangular panel is the result of the IGME at different τ_k and L. The purple color shows the regions of the top 5% most accurate IGME models. (c) The Chapman–Kolmogorov test of the MSM (green), qMSMsm (cyan) and IGME (red) compared against MD simulations (grey). In this panel, $\tau_{MSM} = \tau_k = 30ns$. (d) The bar graph of ITS bounds predicted by the qMSMsm (cyan) and IGME (red). This figure is reproduced from Cao. *et al.*[31]

to capture the conformational changes of the RNAP system [103]. The dataset comprises 306 trajectories, each spanning 200 ns with snapshots saved at 100 ps intervals. To identify the metastable states of the RNAP, we firstly performed the tICA to identify three collective variables from 1,770 number of features (involving alpha Carbon atoms of the clamp, β-lobe, β-protrusion, Switch regions, and active site). Subsequently, four metastable states are identified using a splitting-and-lumping approach. A 100-microstate MSM is initially constructed with the K-Centers algorithm in the collective variable space, and then four states are further identified using the PCCA+ algorithm. We estimate the transition probabilities between pairs of the four metastable states using the maximum likelihood estimator.

For this system, the qMSM fails to converge at certain values of τ_k, and the converged qMSMs exhibit significant fluctuations. To address this, we introduced a smoothing technique, qMSMsm,[103] to reduce the fluctuations in the TPMs used to construct the qMSMs. This smoothing technique significantly improved the convergence of the qMSM from 51% (original qMSM at $\tau_k \in [0, 90]$ ns) to 92% (qMSMsm). For the converged data, as shown in Figure 3.1 (b), the smoothing technique further reduced the error in predicted TPMs from 0.16% (average RMSE of the qMSM within $\tau_k \in [30, 90]$ ns) to 0.07% (qMSM$_{sm}$). However, for complex systems such as Taq RNAP, qMSMs with the smoothing scheme still exhibit significant fluctuations, limiting the accuracy of long-term dynamics predictions. The uncertainty in the variational bound of the slowest implied timescales for qMSM$_{sm}$ ranges from 23% at $\tau_k \sim 40ns$ to 108% at $\tau_k \sim 10ns$, as indicated by the blue bars in Figure. 3.1(d).

In constrast, the IGME produces much more accurate results with smaller numerical fluctuations. As shown in Figure. 3.1(b), the best qMSM$_{sm}$ has an RMSE of 0.04%, which is higher than the RMSE of the green, cyan, blue, and magenta regions representing the top 30% of IGME models from the exhaustive scan. The purple region consisting of the top 5% IGME models could reach an even smaller RMSE ($0.023\% - 0.028\%$) than the qMSM. The enhancement in numerical stability is illustrated in Figure. 3.1(d): when $\tau_k \geq 20ns$, the uncertainty of the ITS bound calculated by IGME ranges from 1% at $\tau_k \sim 30ns$ to 7% at $\tau_k \sim 50ns$. More examples of implementing non-Markovian dynamic models for biomolecular systems could be found in Ref. [31].

Chapter 4

An Efficient Path Classification Algorithm Based on VariationalAutoencoder to Identify Metastable Path Channels for ComplexConformational Changes

This chapter is reproduced in part with permission from Qiu, Y; O'Connor, M. S., Xue, M., Liu, B., & Huang, X. An Efficient Path Classification Algorithm Based on Variational Autoencoder to Identify Metastable Path Channels for Complex Conformational Changes. *Journal of Chemical Theory and Computation* **2023**, 19(14), 4728-4742.

4.1 Introduction

As introduced in Chapter 1, conformational changes, or dynamic transitions between pairs of conformational states, are crucial to many multibody systems, such as protein-protein interaction[3, 5], protein-ligand recognition[7, 8], and the self-assembly of soft materials[13–17]. Investigating the kinetics of these conformational changes is essential for understanding the molecular mechanisms underlying various biological processes and for advancing the rational design of self-assembled materials[104]. Since conformational changes often occur on millisecond timescales and at the molecular level, existing experimental techniques struggle to achieve the necessary spatial and time resolution to directly observe their dominant kinetic pathways. All-atom MD simulations have proven effective in studying these changes and their kinetic pathways, complementing experimental techniques [6, 103]. However, identifying the ensemble of kinetic pathways from MD simulations, which involve the time evolution of all-atom conformations, is challenging[28, 29, 40, 56, 105].

In the chapter 2, we have introduced that TPT applied to a MSM is highly effective for identifying the full ensemble of kinetic pathways from MD simulations. In the framework of MSM, by partitioning the conformational space into metastable states and coarse-graining time into discrete

lag times, continuous dynamics can be modeled as Markovian transitions. TPT then identifies and characterizes kinetic pathways between states, including dominant pathways and fluxes. This approach has been widely used to study conformational changes in chemical and biological processes.

Applying TPT to conformational changes of multi-body systems, especially in heterogeneous aggregation and self-assembly, often results in numerous parallel pathways with comparable probabilities or fluxes. For example, in the aggregation of two hydrophobic molecules with 500 states in explicit solvent[56], 900 pathways are needed to account for 50% of the total flux. This multitude of pathways complicates the understanding of molecular mechanisms. While some protein folding studies demonstrate that a few pathways can capture a substantial fraction of the flux[106], many conformational changes involve a large number of kinetic pathways[107–110]. For instance, the top 200 pathways identified in a MSM for c-Src kinase activation with 1798 states only account for about 18% of the total flux[109]. This extensive number of pathways hinders our understanding of the underlying mechanisms, necessitating the simplification of models by grouping numerous states into fewer, more interpretable states. However, a MSM with fewer states may need longer lag times to achieve Markovianity, which is limited by the length of MD trajectories [30–32]. To construct a high-resolution Markovian model with shorter lag times, the number of states (i.e., microstates) must be increased, resulting in a significant rise in kinetic pathways. Therefore, to gain a deeper understanding of the mechanisms of conformational changes in multi-body systems, it is necessary to develop algorithms that can lump parallel kinetic pathways into representative path channels.

To effectively group kinetic pathways into path channels, it's essential to define the distance or similarity between pairs of pathways and then cluster them accordingly. Our previous path-lumping algorithm [56] employed inter-path flux to define pathway distance, but this single-value metric can sometimes lead to inaccurate clustering by overlooking pathway topology and spatial distribution. This can consequently result in pathways with distinct transition mechanisms being falsely grouped together due to large inter-path flux near initial source or final sink states. To

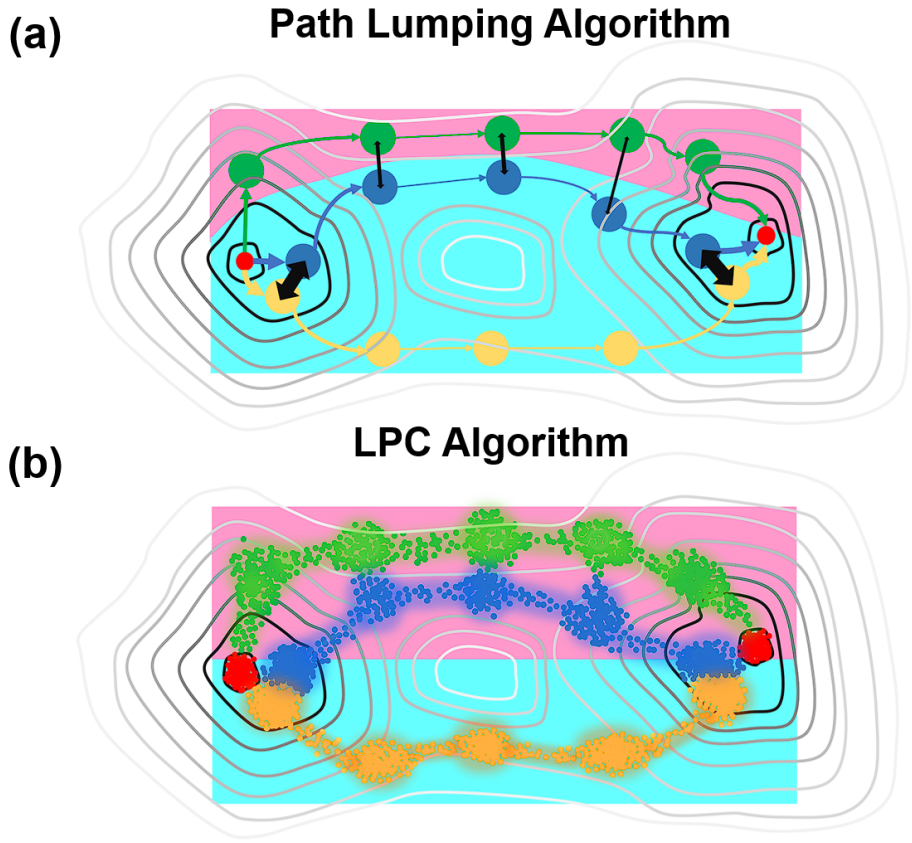

Figure 4.1: Schematic potential illustrating the differences between the path-lumping algorithm and latent-space path clustering (LPC) algorithm. (a) Three distinct pathways (green, blue, and orange) are represented by the connected nodes. Each of the three pathways originates from the energy basin on the left and ends at the one on the right (indicated by the red dots). The thicknesses of the black arrows denote the relative quantities of effective fluxes between nodes crossing different pathways. Since the blue and orange pathways share large inter-path flux through the nodes in the left and right potential basins, the path-lumping algorithm groups the three pathways into two metastable path channels: channel 1 (in pink, containing the green pathway) and channel 2 (in cyan, containing the blue and orange pathways). (b) The LPC algorithm treats each pathway as a continuous flow and considers the topologies and boundaries of the pathways. Since the blue and green pathways have more similar spatial distributions, LPC can correctly identify two metastable path channels: channel 1 (in pink, containing the green and blue pathways) and channel 2 (in cyan, containing the orange pathway). This figure is reproduced from Qiu. *et al.*[48]

achieve more accurate classification, pathways should be considered as continuous flows, incorporating their spatial distributions into the clustering process. However, considering the spatial distributions of pathways is challenging since these kinetic pathways go through the high-dimensional phase space. In the construction of MSMs, MD conformations are typically projected into a lower-dimensional subspace composed of collective variables that capture the slowest transitions before clustered into states [47, 55, 76, 111]. In this collective variable subspace, Euclidean distances between conformations can serve as a reliable indicator of kinetic distances. By leveraging spatial distributions of pathways in the continuous collective variable space and employing deep learning techniques, path classification can be performed with greater accuracy.

In this Chapter, we introduce the Latent-space Path Clustering (LPC) algorithm, designed to group numerous parallel kinetic pathways into a condensed set of metastable path channels based on their spatial distributions within a continuous subspace defined by collective variables. This approach leverages variational autoencoder (VAE) neural network[112], a deep learning framework adept at extracting meaningful features from high-dimensional data and extensively used in studies of dynamics of conformational changes [79, 81]. For instance, time-lagged VAE has been employed to simplify high-dimensional MD conformations into more manageable manifolds through time covariance analysis [79]. Similarly, the state predictive information bottleneck method has utilized VAE to directly map conformations in continuous space to discrete states based on the information bottleneck principle[81]. By harnessing VAE's capabilities, the LPC algorithm embeds the spatial distributions of kinetic pathways into a latent space, facilitating more effective classification.

4.2 Latent-Space Path Clustering (LPC) algorithm

In this section, we introduce the Latent-space Path Clustering (LPC) algorithm, designed to group various kinetic pathways into a set of metastable path channels. The core concept of LPC is to consider the spatial distribution of each kinetic pathway within the continuous collective variable space during path clustering. We begin by using the kinetic-mapping tICA algorithm to construct a low-dimensional collective variable space, where the Euclidean distances between pairs

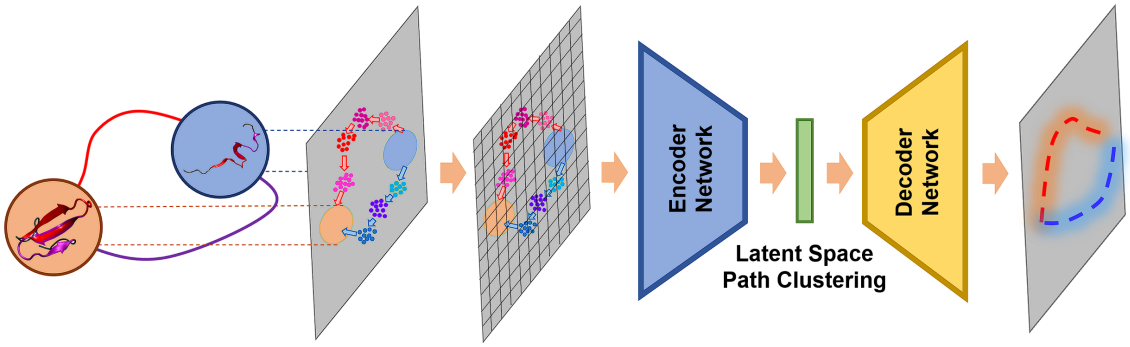

Figure 4.2: Schematic for the LPC algorithm. The initial step involves identifying the transition pathways using MSM and TPT. Subsequently, the conformations associated with each pathway are projected onto the collective variable space, generating the path distributions. The path distributions are then utilized to train the VAE network. Finally, path clustering is performed in the VAE latent space using the K-means algorithm. This figure is reproduced from Qiu. *et al.*[48]

of MD conformations correlate well with their kinetic connectivity. As introduced in Chapter 2, the kinetic-mapping tICA algorithm [76] achieves this by rescaling independent components from tICA using the eigenvalues of the time-correlation matrix based on diffusion map theory. Once the MD conformations are projected onto this collective variable space, they are grouped into states to construct a MSM. Given initial (I) and final (F) states, TPT is applied to identify an ensemble of kinetic pathways, each consisting of a sequence of states visited when transitioning from I to F. More details about MSM construction and TPT are presented in Chapter 2.

To classify kinetic pathways into a set of metastable path channels, it is crucial to define the distance or similarity between pairs of pathways. In the previous path-lumping algorithm[56], the distance between two pathways was determined by inter-path fluxes, which were calculated as the normalized sum of effective flux between every pair of states belonging to the respective pathways. However, relying solely on a single inter-path flux value may not always suffice for accurate path clustering, as it only accounts for the total fluxes connecting two pathways without considering the continuous flow and topology of each kinetic pathway. This can lead to inaccuracies, especially

when effective fluxes between states near I and F dominate the inter-path flux. For example, Figure 4.1a shows two energy basins (left and right) connected by three pathways. The large effective flux (black arrows) between nodes in the basins of the blue and orange pathways causes these pathways to be grouped together by the path-lumping algorithm (highlighted in cyan in Figure 4.1a), even though the blue pathway is kinetically similar to the green pathway (highlighted in pink in Figure 4.1b).

To address the challenge faced by the path-lumping algorithm, we developed the LPC algorithm, which considers each pathway as a flow in the continuous collective variable space and uses the spatial distributions of all pathways for classification rather than inter-path flux. This decouples the pathway identification from the clustering process, allowing kinetic pathways identified by any algorithm (TPT in this study) to be used as input for the LPC algorithm. Since TPT characterizes a kinetic pathway as a sequence of state indices, it is necessary to embed the pathways as flows in the continuous collective variable space. We project MD conformations belonging to the states traversed by a kinetic pathway onto the collective variable space, resulting in a spatial distribution that accurately represents the pathway's shape and flow. Directly using these projections may lead to biased clustering due to high-density regions dominating the distributions (see Figure 4.1 for an example). To address this, we reweight each MD conformation using the inverse of the local conformational density (i.e., $1/N_{local}$) to generate pathways with homogeneous density distributions. This approach focuses on the shapes and spatial distributions of pathways during clustering. Each pathway is then converted into a $1D$ vector with N^k dimensions (where k is the number of collective variables and N is the number of bins along each collective variable).

After generating the 1D vectors for the individual pathways, we input them into a deep neural network called VAE[112], which learns the spatial distributions of pathways. VAE is a popular generative model in the deep learning field with the capability to learn, approximate, and compress high-dimensional distributions[113]. It consists of two parts: the encoder network, which embeds high-dimensional input data into points in a low-dimensional representation called the latent space, and the decoder network, which reconstructs the latent representation back to the original dimensionality. The loss function used to update and train the network comprises two terms: the

reconstruction loss and the Kullback-Leibler divergence. The reconstruction loss measures the difference between the original data and the reconstructions, while the Kullback-Leibler divergence regularizes the encoder framework by applying Gaussian noise to the latent space. Variational Bayesian inference with a Gaussian prior is incorporated into the loss function to maximize the lower bound on the log-likelihood of the observed data[112]. The formula for the loss function is as follows:

$$\mathcal{L}_{VAE} = \mathcal{L}_R + \mathcal{L}_{KL} = \mathbb{E}[||x^{(i)} - x'^{(i)}||] + \mathcal{KL}[q(z)||p(z)] \tag{4.1}$$

where x^i and $x'^{(i)}$ represent the input and output data for the VAE, $q(z)$ is the latent space distribution learned by the encoder network, and $p(z)$ is the prior normal distribution. The VAE generates a representative latent space that captures the spatial distributions of kinetic pathways. When pathways are mapped to points within this latent space, those with high similarity correspond to similar flows in the continuous collective variable space.

The detailed procedures to perform the LPC algorithm is summarized as follow:

(i). Construct an MSM by first projecting the MD simulation dataset onto a low-dimensional collective variable space using kinetic-mapping tICA. Then, use clustering methods to group the conformations into discrete states. Validate the MSM by inspecting the ITS and performing the CK test [27–29].

(ii). With the given initial (I) and final (F) states, employ the TPT to obtain an ensemble of kinetic pathways.

(iii). Embed each pathway as a flow in the collective variable space and convert it to a $1D$ vector with a dimension of N^k, where k is the number of collective variables and N is the number of bins along each collective variable. See Figure 4.2 for an example with $k = 2$.

(iv). Train a VAE using the 1D vectors from the previous step. The encoder will embed each pathway as a single point in the latent space (see Figure 4.2).

(v). Perform K-Means clustering in the VAE latent space to group kinetic pathways into a set of metastable path channels. The distance between a pair of pathways is defined as the Euclidean distance between two points in the latent space.

4.3 Evaluating the LPC Algorithm: Comparison with Path Lumping

To showcase the performance of the LPC algorithm, we applied it to a 2D potential system with numerous parallel transition pathways and four distinct path channels separated by potential energy barriers (see Figure 4.3a). The 2D potential energy for each point $(x, y) \in (0, 60) \times (0, 60)$ on the landscape is given by the following expression:

$$\frac{E}{k_B T} = \frac{1}{2}\left\{1 - 2\left(\frac{x-30}{20}\right)^2 + \left(\frac{x-30}{20}\right)^4\right\}\left\{3 + cos\left(\frac{2\pi y}{30}\right) + 2cos\left(\frac{4\pi y}{30}\right)\right\} \tag{4.2}$$

Among these four channels, the two exhibiting spatial symmetry should have identical fluxes and probabilities. MD simulations of a single particle with mass $m = 1$ were carried out on the 2D potential landscape. The velocity-Verlet integrator was employed with a time step of 0.001, and the Andersen thermostat maintained a temperature of $T = 1/k_B$. At each integration step, the particle's velocity was randomly assigned according to the Boltzmann distribution with $k_B T = 1$ at a probability of 0.001. Reflective boundary conditions prevented the particle from diffusing outside the 2D space. To ensure ergodic sampling, an NVT simulation with 10^8 steps and a saving interval of 100 steps was conducted. The resulting trajectories were then clustered into 500 states using the K-Means algorithm, with the cluster centers serving as initial positions for the production NVT simulations. During production, 500 parallel trajectories were simulated for 10^7 steps each, with a saving interval of 100 steps. From these simulations, 10^5 points per trajectory were sampled, totaling 5×10^7 points for subsequent analysis. To perform the TPT analysis, we selected the states from the 300-state MSM closest to $(10, 30)$ and $(50, 30)$ as the initial and final states of the transition pathways (Figure 4.3b). Finally,the top 500 pathways, contributing 94.8% of the total flux,were grouped into path channels by the LPC algorithm (Figure 4.3c&d).

To apply the LPC algorithm, we initially embedded these 500 pathways as continuous flows in the (x, y) space and then converted them into 500 1D-vectors, each with 900 dimensions (30×30 bins), to form a training dataset for the VAE. After 100 training epochs, the VAE network converged (see Figure 4.4a). Using K-Means clustering in the latent space, we classified these pathways into 4 metastable channels (see Figure 4.4b). Notably, the LPC algorithm's classification results aligned with the shape and symmetry of the potential energy landscape, assigning similar fluxes to the

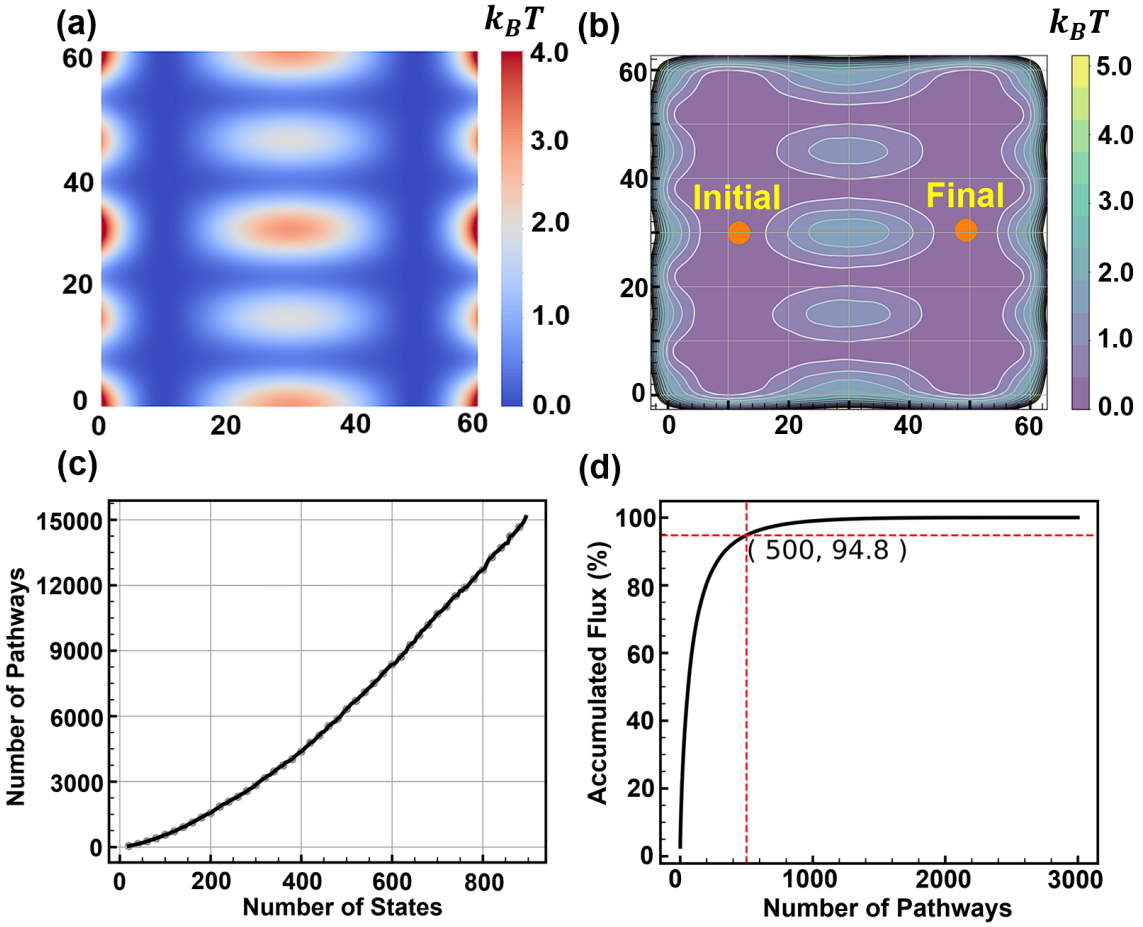

Figure 4.3: $2D$-potential with 4 path channels connecting the two energy basins. (a) The analytic $2D$-potential (see Eq.8 for details). (b) Free energy landscape of the $2D$-potential sampled by MD simulations at $T = 1/k_B$. The distributions of MD conformations are estimated by the Gaussian kernel function. (c) The number of identified pathways from TPT as a function of MSMs containing different numbers of states. (d) The accumulated flux as a function of the number of transition pathways obtained by TPT from a 300-state MSM. This figure is reproduced from Qiu. *et al.*[48]

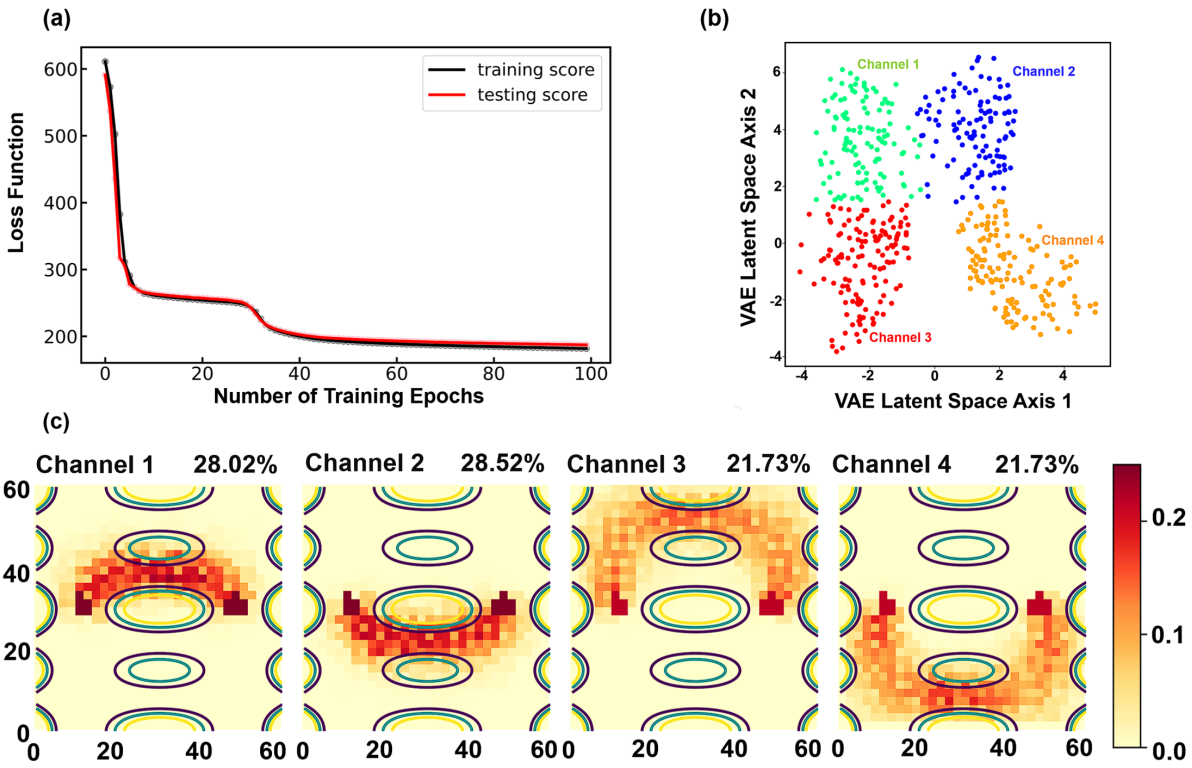

Figure 4.4: Latent-space path clustering (LPC) algorithm classifies 500 pathways into 4 path channels. (a) Loss function as a function of the training epochs for the training and testing of VAE. (b) Distributions of kinetic pathways (each is represented as a point) in the latent space with channel labels. (c) Visualization of identified path channels with each corresponding flux labeled. This figure is reproduced from Qiu. *et al.*[48]

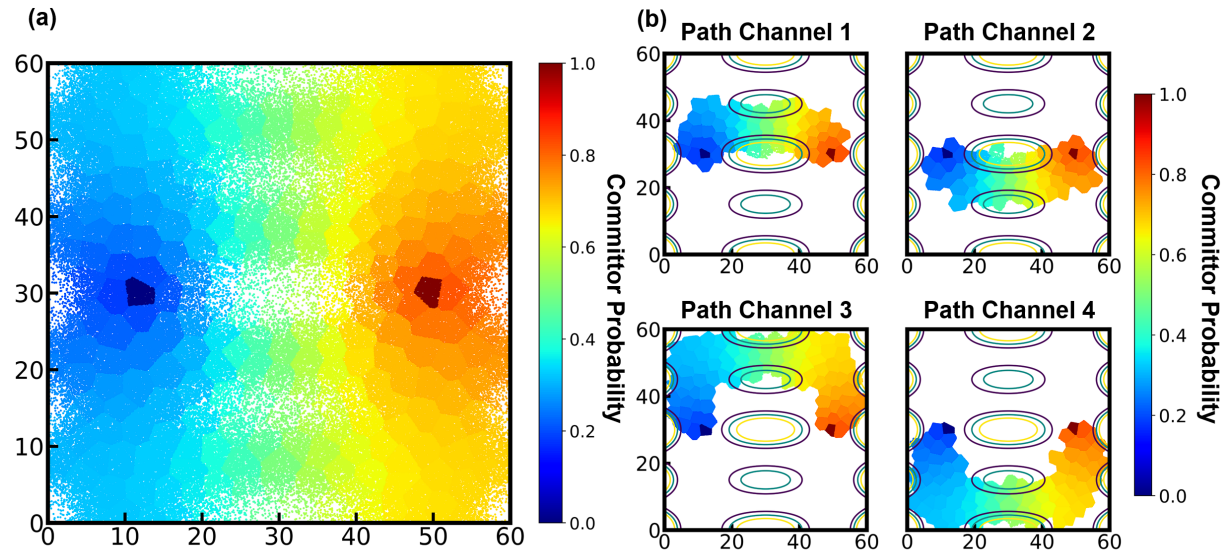

Figure 4.5: Forward committor probabilities for the 2D potential system. (a). Projections of forward committor probabilities of 300 states onto the 2D map. Each state is visualized by overlaying its conformations with the color corresponding to forward committor probabilities. (b). Forward committor probabilities for each of the four metastable path channels. This figure is reproduced from Qiu. *et al.*[48]

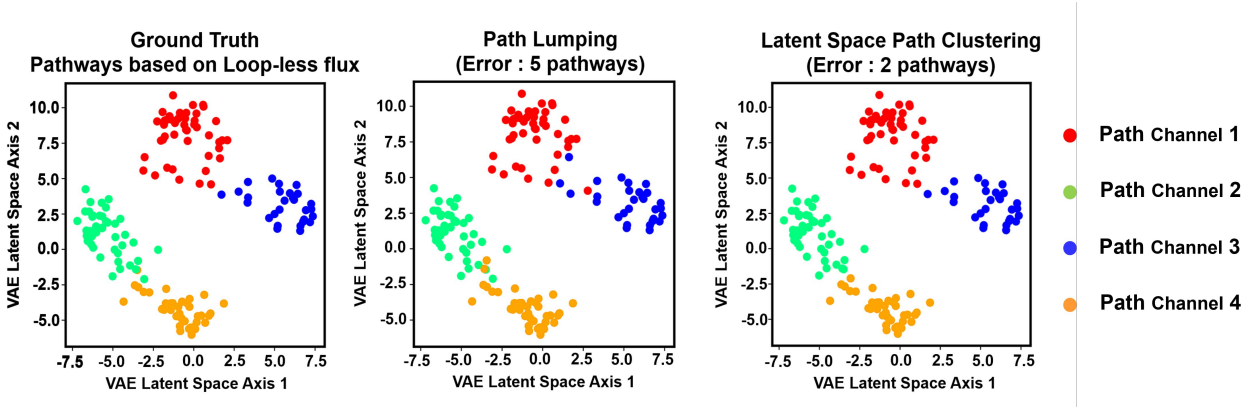

Figure 4.6: Comparison of accuracy between path-lumping algorithm and LPC algorithm. All 154 kinetic pathways (each is represented as a point) are visualized on the same two-dimensional latent space generated from the VAE in LPC and colored according to the classification labels. This figure is reproduced from Qiu. *et al.*[48]

spatially symmetric path channels (see Figure 4.4c). Additionally, to visualize the transition states, we plotted the forward committor probabilities on the 2D space (see Figure 4.5). The transition states for the four path channels are precisely located on their respective energy barriers, clearly distinguishing the left energy basin from the right.

The 2D-potential used in this study clearly distinguishes kinetic pathways into four metastable channels, making it an ideal testbed for evaluating our LPC algorithm against the previous path lumping method. The earlier path lumping algorithm relies on a single metric-namely, the accumulated effective fluxes between states across different pathways-to assess pathway similarity. In contrast, the LPC algorithm leverages the spatial distributions of pathways in continuous collective variable space for clustering. To compare these two approaches, we analyzed an additional set of 154 pathways, identified using the modified TPT protocol described in the path lumping study. This allowed us to evaluate both methods with a clear ”ground truth” established by directly plotting individual pathways on the 2D-potential and accurately assigning them to specific path channels (see Figure 4.6a). As shown in Figures 4.6b and c, the LPC algorithm significantly outperforms the path-lumping approach, providing a more precise clustering result, especially near the channel boundaries (with only 2 mis-assigned pathways by LPC compared to 5 by the path-lumping

method). The path-lumping method's reliance on a single inter-path flux value proved insufficient for accurate clustering. In contrast, the LPC algorithm's use of continuous flow representations and spatial distributions enhances its accuracy in distinguishing pathways.

4.4 Elucidating Path Channels for Hydrophobic Particles Aggregation

As mentioned in the Introduction, heterogeneous self-assembly or aggregation processes often involve numerous parallel kinetic pathways. We now evaluate the effectiveness of our LPC algorithm in analyzing the aggregation dynamics of two hydrophobic 9-(diphenylmethylene)-9H-fluorene (9D9F) molecules in explicit solvent (Figure 4.7a). The simulation data, sourced from a prior study, includes 40 MD trajectories of $100ns$ each[56]. For constructing the MSMs, the final 80 ns of each trajectory were utilized. Additional information regarding the force field and MD setup can be found in the referenced study. To construct the MSM, we used all pairwise distances between heavy atoms of the two 9D9F molecules, resulting in 676 features. These features were then combined and reduced to three collective variables using kinetic-mapping tICA. The conformations were assigned to 500 states using the K-Centers clustering algorithm, and an MSM was built with a Markovian lag time of $30ps$. States were categorized based on their solvent accessible surface area (SASA), with 25 states identified as aggregated, 80 as separated, and the rest as intermediate (Figure 4.7b). TPT identified over 15,000 pathways, and our LPC algorithm grouped the top 10,000 pathways—accounting for 98.2% of the total flux—into four distinct path channels (Figure 4.7c&d).

As illustrated in Figure 4.7, this system reveals a higher number of pathways compared to both the 2D potential and the normal protein folding systems, even though all MSMs involved contain only a few hundred states. This indicates that the dynamics of this hydrophobic aggregation are exceptionally heterogeneous. Additionally, the majority of these kinetic pathways have similar weights, with the most prominent pathway accounting for just 0.22% of the total flux. Consequently, understanding the mechanisms of hydrophobic aggregation is challenging due to the large number of parallel pathways, necessitating the use of path clustering.

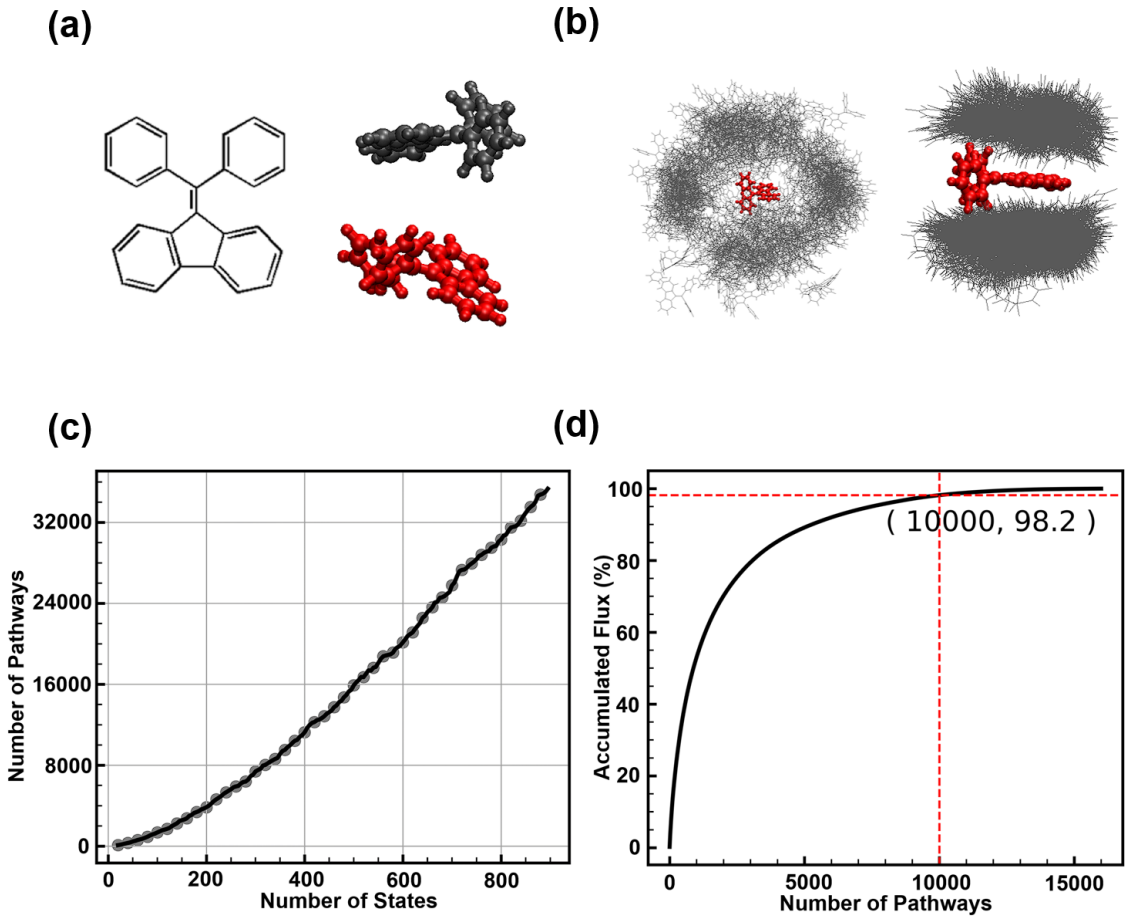

Figure 4.7: The dynamics for the aggregation of two hydrophobic molecules contain a large number of parallel pathways. (a) The chemical structure of 9D9F. (b) Separated and aggregated conformations for the two hydrophobic 9D9F molecules. (c) Number of identified kinetic pathways from TPT as a function of the number of states in MSMs. (d) The accumulated flux as a function of the number of transition pathways. The results are obtained using a 500-state MSM). This figure is reproduced from Qiu. *et al.*[48]

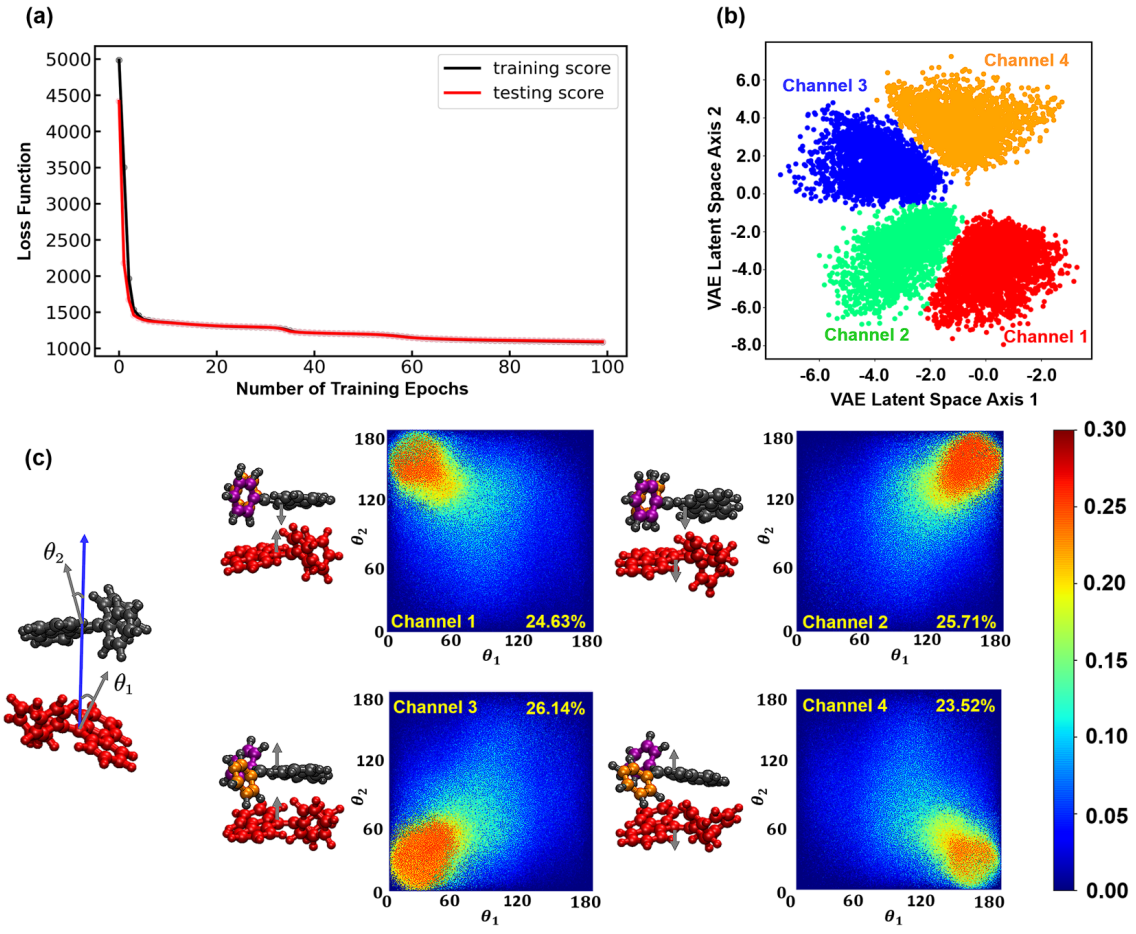

Figure 4.8: LPC clustering algorithm classifies $10,000$ pathways into 4 path channels for the hydrophobic aggregation. (a) Loss function as a function of the training epochs for the training and testing processes of VAE. (b) Distributions of kinetic pathways (each is represented as a point) in the latent space with channel labels. (c) Visualization of 4 path channels by overlaying all kinetic pathways (weighted by their fluxes) belonging to each path channel. To visualize a kinetic pathway, we projected all the MD conformations (re-weighted by the inverse of local density) belonging to this pathway onto (θ_1, θ_2). θ_1 and θ_2 are defined by the normal vector of each 9D9F molecule (grey arrows in the left panel) and the center of mass displacement vector between the two 9D9F molecules (blue arrow in the left panel). This figure is reproduced from Qiu. *et al.*[48]

To implement the LPC algorithm, we initially embedded the kinetic pathways as continuous flows in a three-dimensional collective variable space. We then needed to convert these pathways into 1D-vectors for use as training data for the VAE. However, directly transforming each pathway into a 1D-vector resulted in an impractically high dimensionality for this system (i.e., 50^3). To address this, we projected each kinetic pathway onto three distinct two-dimensional collective variable spaces, each defined by a pair of the three collective variables. This approach yielded 1D-vectors with a significantly reduced dimensionality (i.e., $50^2 \times 3$). After inputting these 1D-vectors into the VAE and training for 100 epochs, both the training and testing scores converged (see Figure 4.8a). The latent space obtained, as shown in Figure 4.8b, exhibited a well-distributed arrangement of data points that formed four distinct clusters, indicating the successful identification of four kinetic metastable path channels.

Finally, we used K-Means algorithm to group $10,000$ pathways into four metastable path channels in the latent space of the trained VAE. As shown in Figure 4.8c, these four channels correspond to four distinct relative orientations of how the two particles approach and aggregate. This suggests the presence of free energy barriers between different aggregation modes, aligning with chemical intuition. Our findings also support the observation that one particle must undergo a mirror inversion to change the aggregation mode, which involves breaking and reforming contacts and overcoming significant energy barriers. Due to the system's spatial symmetry, the four path channels are expected to share an equal amount of flux, which is consistent with our clustering results. Moreover, the clustering results from the LPC algorithm align with previous path lumping studies, confirming the algorithm's effectiveness in analyzing self-assembly dynamics. To illustrate the significant transition states within these four metastable channels, we projected the committor probabilities onto the (θ_1, θ_2) plane and displayed the corresponding transition state structures (defined as states with committor probabilities closest to 0.5). The structures from transition state represent encounter complex conformations specific to each path channel, highlighting distinct aggregation pathways.

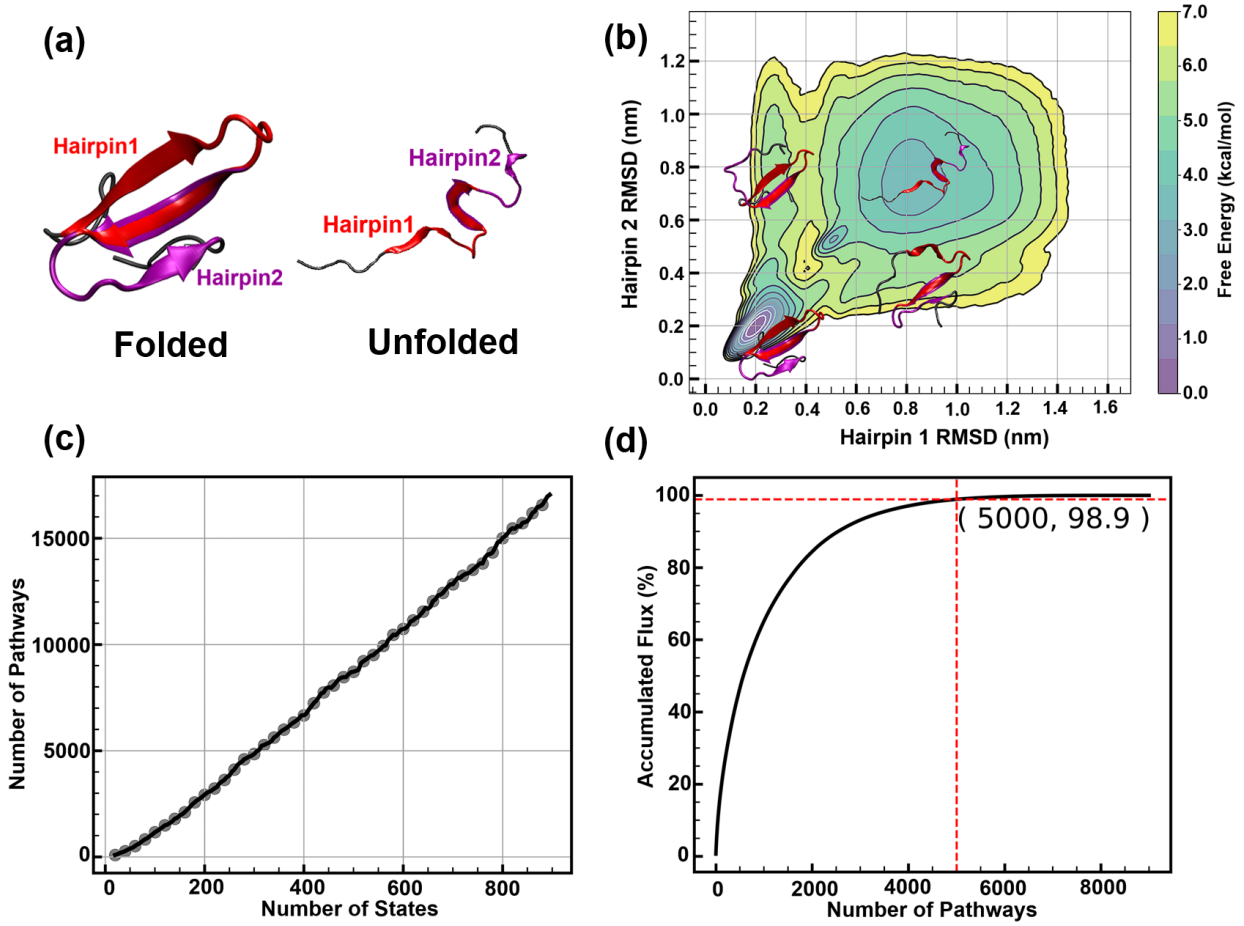

Figure 4.9: The mechanism of Fip35 WW-domain folding. (a) The native and unfolded structures of Fip35 WW-domain. (b) Projection of free energy landscape onto two physical coordinates: RMSDs of the Hairpin 1 (X-axis) and Hairpin 2 (Y-axis) with respect to the native structure. We projected all the MD conformations onto the RMSD space to generate this figure. (c) The relationship between the number of identified pathways from TPT and the number of MSM states. (d) The accumulated flux as a function of the number of transition pathways (based on the 500-state MSM). This figure is reproduced from Qiu. *et al.*[48]

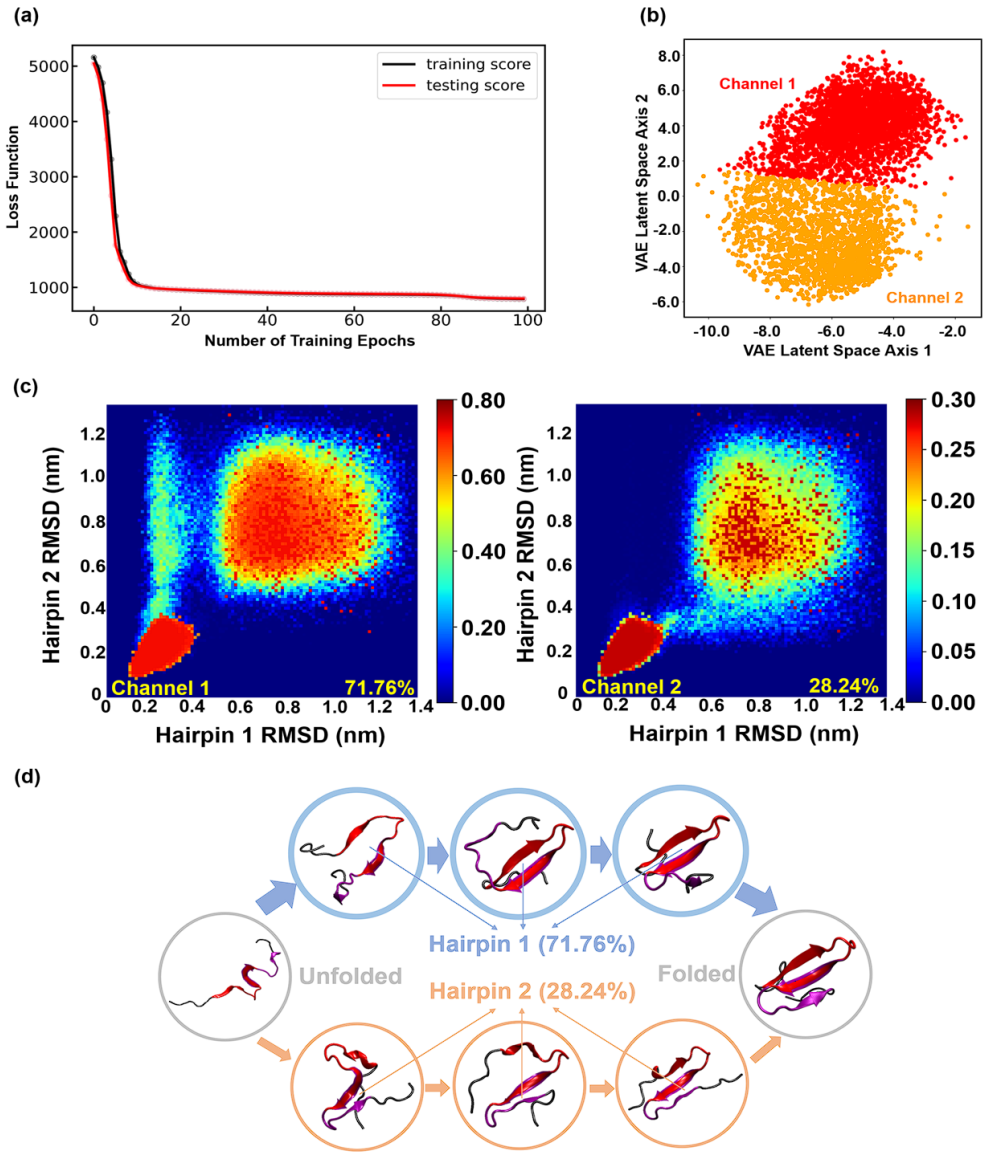

Figure 4.10: The LPC algorithm classifies 5000 pathways into two path channels for the Fip35 WW domain folding. (a) Loss function as a function of the number of training epochs for the training and testing processes of VAE. (b) Distributions of kinetic pathways (each is represented as a point) in the latent space with channel labels. (c) Visualization of the identified path channels using two physical coordinates: the RMSDs of hairpin 1 and hairpin 2 with respect to the folded structure. To visualize each path channel, we overlaid all the kinetic pathways (weighted by their fluxes) belonging to each path channel, and for each pathway, we projected all the MD conformations (reweighted by the inverse of local density)belonging to this pathway onto two RMSDs. (d) Representation of protein conformations along the two path channels. This figure is reproduced from Qiu. *et al.*[48]

4.5 Understanding Protein Folding Mechanisms with the LPC Algorithm

In this section, we demonstrate the effectiveness of the LPC algorithm in studying the folding of a 35-residue Fip35 WW domain. This protein consists of three-stranded β-sheets, with residues $8-23$ forming hairpin 1 and residues $17-30$ forming hairpin 2 (see Figure 4.9a). In recent years, significant efforts have been devoted to investigating the folding mechanism of the Fip35 WW domain[40, 114–118]. Two distinct folding pathways, differing in the order of hairpin 1 and hairpin 2 folding, have been identified. The folding transition can occur through the formation of hairpin 1 followed by hairpin 2 (hairpin 1-hairpin 2 pathway) or in the reverse order (hairpin 2-hairpin 1 pathway)[40, 116, 118, 119]. In 2005, Klimov and Thirumalai employed coarse-graining simulations to examine the impact of secondary structure elements on protein folding [119]. They found that, for the asymmetrical Fip35 WW domain, approximately 70% of the folding pathways involve the initial formation of hairpin 1, followed by the folding of hairpin 2. In 2009, No$\acute{e}$ et al. performed a TPT analysis based on a 50-state MSM and showed similar results, with 70% of the folding transitions occurring through the hairpin 1-hairpin 2 pathway and 30% through the hairpin 2-hairpin 1 pathway[40]. However, their analysis was based on a modest dataset containing 180 MD trajectories of $115ns$ each (approximately $20.7\mu s$ in total). More recently, D. E. Shaw Research published a much more extensive dataset containing two independent ultralong trajectories of 651 and $486\mu s$[114]. Here, we applied our LPC method to analyze this extended dataset and quantify the probabilities of these two metastable channels for the WW domain folding.

Specifically, we performed K-Centers clustering on the three-dimensional collective variables space to generate a 500-state MSM. TPT identified approximately $9,000$ pathways, with the top $5,000$ pathways serving as input for the LPC algorithm (Figure 4.9c&d). As with the hydrophobic aggregation system in the previous section, we created three two-dimensional subspaces by combining each pair of collective variables and further embedded the pathways as continuous flows in each subspace. We then generated a distribution 1D-vector with a length of $50^2 \times 3$ for each of the $5,000$ pathways and inputted them into the VAE. After 100 epochs of training, the network converged, producing a two-dimensional latent space, as shown in Figure 4.10a. Using the K-Means

clustering algorithm, we classified the pathways in the VAE latent space into two metastable path channels (Figure 4.10b). When visualizing the classified path channels in the RMSDs with respect to the folded structure, we observed that each channel corresponds to a specific folding pathway. We found that the hairpin 1-hairpin 2 pathway accounts for 71.76% of the total folding flux, while the hairpin 2-hairpin 1 pathway contributes 28.24% of the flux (Figure 4.10c and d). Our results, based on the D.E. Shaw simulation dataset ($1137\mu s$ in total), are consistent with those obtained by Noé et al[40]. via TPT analysis of a 50-state MSM based on a much smaller dataset ($\sim 20.7\mu s$ in total). The findings by Thirumalai et al.[119] also support our results. Additionally, a previous study demonstrated that a global fitting of experimental ϕ-value data indicates the relative probability of the hairpin 1 - hairpin 2 folding pathway is approximately $67 \pm 5\%$[120], further corroborating our analysis.

To identify the metastable transition states for the Fip35 WW domain, we visualized the committor probabilities for the two path channels in the RMSD space (RMSD to Hairpin 1 *vs.* Hairpin 2, see Figure 4.11). Interestingly, the transition state of path channel 1 shows a fully folded Hairpin 1 but a partially folded Hairpin 2. In contrast, the transition state of path channel 2 has a partially folded Hairpin 1 but a fully folded Hairpin 2. These observations align with the different sequential orders of hairpin folding, effectively distinguishing channel 1 from channel 2 (see Figure 4.10). Based on this transition state analysis, critical residues can be identified, and mutations designed in the 3-stranded beta sheet. These mutations have the potential to disrupt or stabilize the transition state of a specific path channel, thereby allowing modulation of the relative folding flux among path channels.

In summary, the LPC algorithm is particularly effective in elucidating metastable path channels for self-assembly and other heterogeneous dynamic processes where multiple pathways with similar probabilities coexist. Additionally, it can be applied to protein folding and other complex biomolecular conformational changes, where constructing a Markovian model with many states results in numerous kinetic pathways. In our current implementation, we used pathways from the TPT analysis as input to the LPC algorithm. However, it is important to note that the VAE in the LPC algorithm can be trained using pathways from other algorithms, such as Markov Chain

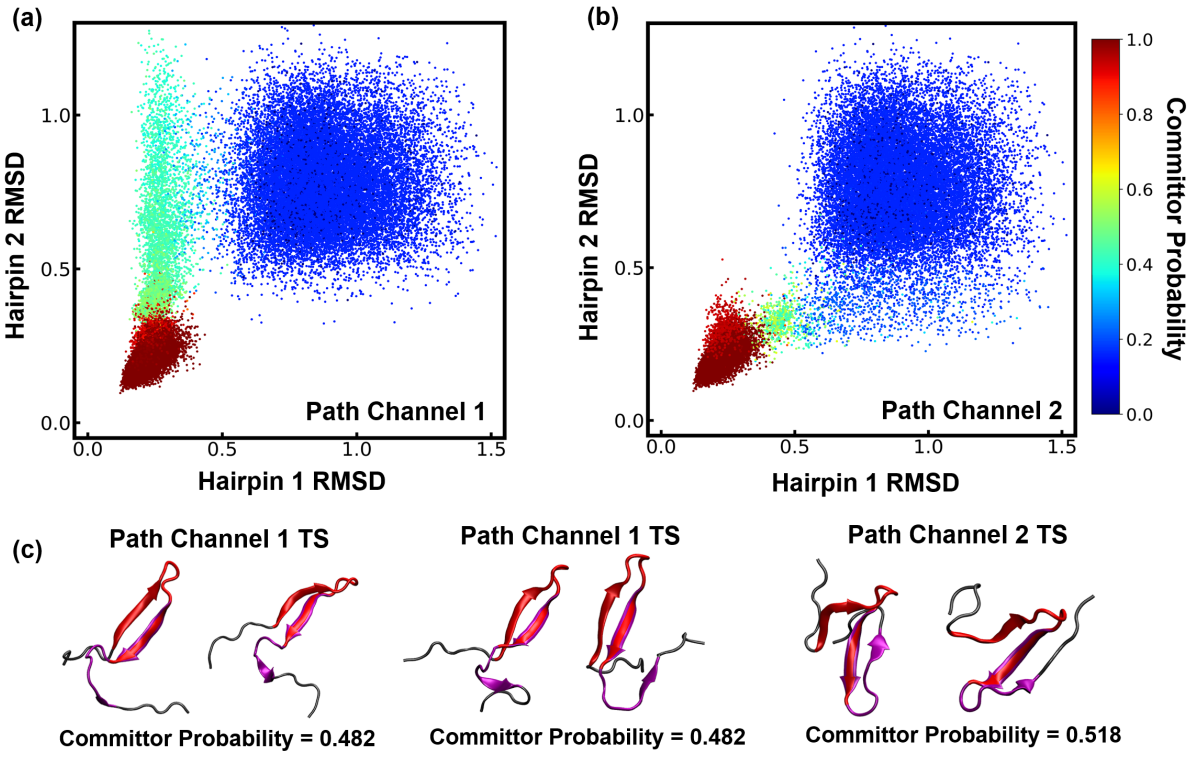

Figure 4.11: Forward committor probabilities for the Fip35 WW-domain folding system. (a-b). The forward committor probabilities are shown separately for each metastable path channel on two physical coordinates: the RMSDs of Hairpin 1 and Hairpin 2 with respect to the folded structure. Two representative configurations are chosen from two states that have committor probabilities closest to 0.5 which correspond to two different folding orders. This figure is reproduced from Qiu. *et al.*[48]

Monte Carlo [121, 122] or MSMPathfinder[123]. Furthermore, our LPC algorithm trains a generative VAE model, with the decoder network capable of generating new continuous pathways in the collective variable space. We anticipate this capability to be valuable for future applications in sampling transition pathways.

However, the LPC algorithm may have certain limitations. Its performance is dependent on the adequacy of conformational sampling; insufficient sampling can introduce bias in MSM and TPT analysis, affecting the accuracy of the input data for LPC algorithm. To address this potential issue, we recommend employing adaptive sampling techniques during MSM construction to improve convergence[124]. Additionally, enhancing the quality of initial conformations for atomistic MD simulations can be achieved by applying enhanced sampling algorithms or coarse-graining simulations before initiating extensive unbiased MD simulations[28, 29].

We also acknowledge that TPT analysis and embedding of transition pathways as continuous flows require careful selection of collective variables in advance. Currently, we use tICA-based kinetic distances, which are linear combinations of input features. However, if this linear approach fails to capture the underlying slow dynamics, alternative non-linear algorithms such as kernel-tICA[77], SRVnets[37], Variational Dynamics Encoders[79], and diffusion maps[125] should be considered. Additionally, the quality of collective variables can be improved by optimizing input features using automatic methods like Spectral-oASIS[75] or through a deeper physical understanding of the system. In the future, we plan to further investigate collective variable selection to enhance the performance of the LPC algorithm.

Finally, it is crucial to validate the theoretical predictions of our LPC algorithm against experimental data. Recent progress in experimental methods has made these comparisons feasible. For instance, Haran and colleagues have used single-molecule fluorescence resonance energy transfer (smFRET) combined with Hidden Markov Model analysis to successfully elucidate the primary transition fluxes between folded and unfolded states in adenylate kinase[126, 127].

4.6 Conclusion

In this work, we have developed the LPC algorithm, which groups an ensemble of kinetic pathways into metastable path channels based on their spatial distributions in the collective variable space. In our algorithm, MD conformations are initially projected onto a low-dimensional space containing a small set of collective variables. Then, MSM and TPT are constructed to obtain the ensemble of pathways, which are subsequently embedded as continuous flows in the collective variable space and used as the training dataset for the VAE. The VAE networks are trained to learn the spatial distributions of kinetic pathways in the continuous collective variable space. In the latent space of trained VAE model, each pathway is represented as a point. Pathways with similar spatial distributions form clusters in the latent space and can be further classified into metastable path channels. We have demonstrated the effectiveness of the LPC algorithm for a 2D potential, the aggregation of two hydrophobic molecules, and the folding of the Fip35 WW domain. In all three systems, the LPC algorithm identifies metastable path channels consistent with our physical intuitions. Using the 2D potential, we further show that the LPC algorithm outperforms previously developed path lumping algorithms. We anticipate that our LPC algorithm can be widely applied to elucidate the metastable path channels underlying complex dynamic processes in various systems.

Chapter 5

Information Bottleneck Approach for Markov Model Construction

This chapter is reproduced in part with permission from Wang, D.; Qiu, Y.; Beyerle, E.; Huang, X.; Tiwary, P. Information Bottleneck Approach for Markov Model Construction, *Journal of Chemical Theory and Computation* **2024**, 20, 12, 5352–5367.

Qiu, Y equally contributed to the development of methodologies, authored the codes, and co-wrote the paper with Wang, D and Beyerle, E. Qiu, Y also rewrote and summarized the content in this chapter.

5.1 Introduction

As introduced in Chapter 2, with advancements in various algorithms for MSMs, it is important to recognize that the specific construction of MSMs and the choice of hyperparameters can significantly affect the quality of the final kinetic model. To quantitatively evaluate the performance of different MSMs and aid in their construction, several metrics have been established. These metrics address various aspects of the model, including improving the metastability of macrostates in the kinetic model[19, 91], optimizing approximations of the principal eigenmodes of dynamics[128], and enhancing the model's ability to capture the leading slowest dynamics[34, 47]. The third approach, known as the Variational Approach for Markov Processes (VAMP), has become the most widely used. It provides a variational score to quantify the discrepancy between the eigenmodes approximated by the model and those of the true dynamical propagator, serving as an objective function for model optimization.

Several approaches utilizing artificial neural networks have proven effective in variational optimization workflows. Recently, VAMPnets have been introduced, combining VAMP and Koopman theory to create a unified, end-to-end data-driven model that directly maps molecular coordinates

to coarse-grained macrostates[34]. Subsequently, State-Free Reversible VAMPnets (SRVs) were introduced to learn nonlinear approximations of the leading slow eigenfunctions [37]. Time-lagged autoencoders use auto-associative neural networks to reconstruct signals with a time lag, with embedded variables from the bottleneck layer serving as collective variables[55]. Another method, variational dynamics encoders (VDEs), integrates time-lagged reconstruction loss with autocorrelation maximization within a variational autoencoder to approximate the dynamical propagator[79].

In this work, we demonstrate a robust protocol for MSMs construction through the use of an information bottleneck approach called State Predictive Information Bottleneck (SPIB)[81]. Previous studies have shown that SPIB could effectively learn low-dimensional collective variables for enhanced sampling in molecular simulations, accelerating various processes. These include permeation[129] and dissociation of medically relevant ligands[130], conformational changes in proteins[131, 132], and nucleation of crystal polymorphs[133]. Here, we focus on demonstrating SPIB as a state-of-the-art approach for automatic construction of multi-resolution MSMs. Unlike existing methods, such as VAMP-based approaches that maximize the Rayleigh coefficient or VAMP-score, SPIB combines the information bottleneck framework with a straightforward heuristic of state metastability at a predefined lag time to unify feature extraction and state division. By utilizing this lag time parameter, SPIB allows for adaptive adjustment of the number of metastable states in the final kinetic model, enabling automatic generation of MSMs at varying resolutions. As a result, SPIB facilitates the automatic clustering and projection of the MD simulation data into a few macrostates and learns directly from MD trajectories a low dimensional latent space where these states are cleanly separated into metastable states, providing a single end-to-end framework that integrates dimension reduction, clustering, and lumping tasks.

5.2 State Predictive Information Bottleneck (SPIB)

Current VAMP-based methods tend to focus either on direct state partitioning, such as VAMPnets[34], or on identifying low-dimensional continuous collective variables from input data, like

SRVnets[37]. In contrast, our approach, the SPIB[81], employs an information bottleneck strategy for MSM construction, providing a unified framework that seamlessly integrates both state partitioning and dimension reduction.

SPIB employs a heuristic to determine both the number and location of potential metastable states for constructing MSMs. The core idea is that if a configuration is in state i at a given time, it should have the highest probability of remaining in state i after a short lag time Δt, since the escape time from a metastable state i should be much longer than Δt. Using this principle, prior study [81] introduced an iterative scheme to dynamically learn the number and location of states on-the-fly.

We initiate with an arbitrary randomly generated state labels for the MD conformations in the trajectory $\{\mathbf{y}_t\}_{t=0}^{T}$, where both the number and definition of labels are some initial guess. The probability that the system starting from $\mathbf{X}$ will be found in state $\mathbf{y}_{\Delta t} = i$ after a lag time Δt, assuming a stationary distribution, can be estimated by the following:

$$\begin{aligned} p(\mathbf{y}_{\Delta t} = i|\mathbf{X}) = \frac{1}{\rho(\mathbf{X})} \lim_{T\to+\infty} \frac{1}{T} \int_0^T \mathbf{1}_{\mathbf{y}_{t+\Delta t}=i} \delta(\mathbf{X} - \mathbf{X}_t) dt \\ \text{where } \rho(\mathbf{X}) = \lim_{T\to+\infty} \frac{1}{T} \int_0^T \delta(\mathbf{X} - \mathbf{X}_t) dt. \end{aligned} \tag{5.1}$$

Here $\rho(\mathbf{X})$ represents the equilibrium density of $\mathbf{X}$ and $\mathbf{1}_{\mathbf{y}_{t+\Delta t}=i}$ is the indicator function defined for state i, which is equal to 1 if the MD trajectory is within state i at time $t + \Delta t$ and equal to 0 otherwise. Since it depends on the input configuration $\mathbf{X}$ and denotes a state-transition probability, we refer to the function $p(\mathbf{y}_{\Delta t}|\mathbf{X})$ as the state-transition density. If the system starting from a high-dimensional configuration $\mathbf{X}$ is most likely to be found in state i after a lag time Δt, then the label of configuration $\mathbf{X}$ will be updated to state i. Consequently, a set of new state labels can be generated according to:

$$\hat{\mathbf{y}}_t = \underset{i}{\text{argmax}}\ p(\mathbf{y}_{t+\Delta t} = i|\mathbf{X}_t). \tag{5.2}$$

Based on the new refined state labels, the state-transition density $p(\mathbf{y}_{\Delta t}|\mathbf{X})$ is re-estimated, and the process is repeated until the state labels converge. This iterative refinement can result in some initial labels being nullified, thereby reducing the total number of states. In this manner, SPIB can

dynamically determine both the number and positions of potential metastable states in the system, starting from any initial states.

Directly estimating the state-transition density using Eq. 5.1 suffers from the curse of dimensionality as the input feature $\mathbf{X}$ is typically high dimensional for complex systems. To alleviate this problem, SPIB seeks to identify a low-dimensional manifold onto which the system's dynamics can be projected, enabling a more robust estimation of the state-transition density $p(\mathbf{y}_{\Delta t}|\mathbf{X})$. This provides a unified pipeline for both dimension reduction and state decomposition. Such a low dimensional representation $\mathbf{z}_t$ is designed to leverage minimal information from the past signal $\mathbf{X}_t$ while accurately predicting its future state label $\mathbf{y}_{t+\Delta t}$. This learning process is enabled by the deep variational information bottleneck framework[134]. Various enhancements have been made to improve the algorithm's robustness since its initial introduction[80, 81], as summarized below.

The network architecture for SPIB is illustrated in Figure 5.1. Given a trajectory $\{\mathbf{X}^1, \cdots, \mathbf{X}^M\}$ and its corresponding state labels $\{\mathbf{y}^1, \cdots, \mathbf{y}^M\}$, where the length M is sufficiently large, the objective function of SPIB is formulated as follows:

$$\underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(\theta) = \quad \frac{1}{M-s}\sum_{n=1}^{M-s}\Big[\log q_\theta(\mathbf{y}^{n+s}|\mathbf{z}^n) - \beta \log \frac{p_\theta(\mathbf{z}^n|\mathbf{X}^n)}{r_\theta(\mathbf{z}^n)}\Big] \tag{5.3}$$

where the encoder $p_\theta(\mathbf{z}|\mathbf{X})$, the decoder $q_\theta(\mathbf{y}|\mathbf{z})$, and the prior $r_\theta(\mathbf{z})$ are probability distributions parameterized by deep neural networks θ. The variable $\mathbf{z}^n$ is sampled from $p_\theta(\mathbf{z}|\mathbf{X}^n)$, where the time interval between $\mathbf{X}^n$ and $\mathbf{X}^{n+s}$ represents the lag time Δt, indicating how long into the future SPIB is set to predict. The first term $\log q_\theta(\mathbf{y}^{n+s}|\mathbf{z}^n)$ evaluates how well our representation predicts the target, while the second term $\log \frac{p_\theta(\mathbf{z}^n|\mathbf{X}^n)}{r_\theta(\mathbf{z}^n)}$ serves as a complexity penalty acting as a regularizer. This regularization term encourages the latent space $\mathbf{z}$ to retain minimal information from the input $\mathbf{X}$, thus fostering a more compact representation. The balance between prediction accuracy and model complexity is governed by the hyperparameter $\beta \in [0, \infty)$.

As usual, we opt to employ a Gaussian encoder with a constant variance:

$$\log p_\theta(\mathbf{z}^n|\mathbf{X}^n) = \log \mathcal{N}(\mathbf{z}^n; \mu, \sigma I) \tag{5.4}$$

where only the mean $\mu = \mu_\theta(\mathbf{X}^n)$ is the output of a neural network whose input is $\mathbf{X}^n$, while the variance σ^2 is a trainable parameter independent of the input $\mathbf{X}$. I denotes the identity matrix. By

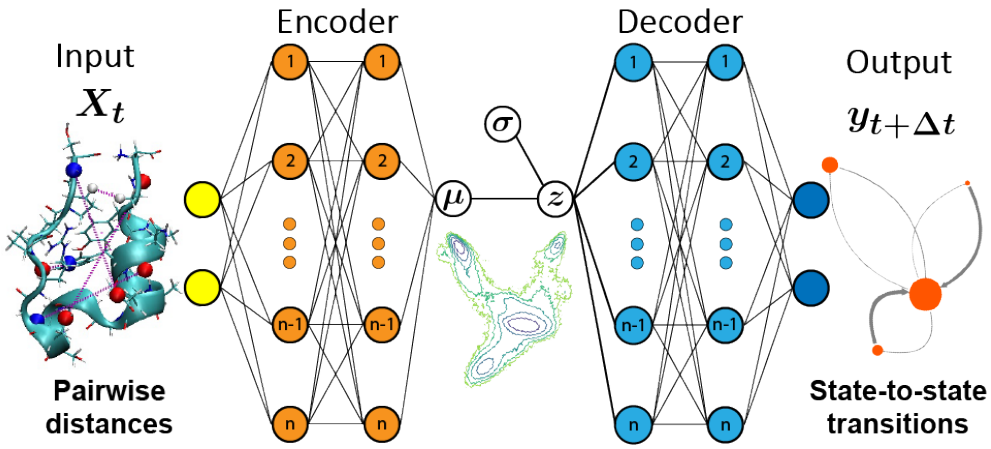

Figure 5.1: Network architecture employed for SPIB consists of both the encoder and decoder as nonlinear neural networks with two hidden layers. SPIB is designed to take features such as pairwise distances, denoted as input $\mathbf{X}_t$, enabling the learning of a low-dimensional latent representation $\mathbf{z}$ for predicting its future state $\mathbf{y}_{t+\Delta t}$ after a lag time Δt. In this modified architecture, the encoder only outputs the mean μ, from which the latent representation $\mathbf{z}$ is then sampled utilizing a position-independent trainable standard deviation σ. For visualization, the left panel illustrates some minimal residue-residue distances of the Trp-cage system. In the middle, an example of the free energy surface of the learned latent space is displayed. The right panel presents a network plot of the output Markov state model. This figure is reproduced from Wang. *et al.*[35]

maintaining a constant variance for $p\theta(\mathbf{z}^n|\mathbf{X}^n)$ across all $\mathbf{X}$, we facilitate the learning of a more homogeneous latent space. Meanwhile, a deep feed forward neural network with softmax outputs is employed in the decoder $q_\theta(\mathbf{y}|\mathbf{z})$:

$$\log q_\theta(\mathbf{y}^{n+s}|\mathbf{z}^n) = \sum_{i=1}^{S} y_i^{n+s} \log \mathcal{D}_i(\mathbf{z}^n;\theta) \tag{5.5}$$

here the state label $\mathbf{y}$ is represented as a one-hot vector with S dimensions, and the decoder function $\mathcal{D}$ produces an S-dimensional softmax output from a neural network. The use of the softmax output in the decoder enables fuzzy assignments to the states $\mathbf{y}$ predicted by SPIB.

Since the latent representation $\mathbf{z}$ is expected to distinguish between different metastable states, a multi-modal distribution for the prior $r_\theta(\mathbf{z})$ is appropriate. We adapt the variational mixture of posteriors prior algorithm to create this multi-modal prior distribution. In this approach, the approximate prior $r_\theta(\mathbf{z})$ is a weighted mixture of various posteriors $p_\theta(\mathbf{z}|\mathbf{X})$, using representative inputs $\{\mathbf{X}_{\text{rep}}^k\}_{k=1}^K$ instead of $\mathbf{X}$:

$$r_\theta(\mathbf{z}) = \sum_{k=1}^{K} \omega_k \; p_\theta(\mathbf{z}|\mathbf{X}_{\text{rep}}^k), \tag{5.6}$$

where K is the number of representative-inputs, and ω_k represents the weight of $p_\theta(\mathbf{z}|\mathbf{X}_{\text{rep}}^k)$ with the constraint $\sum_k \omega_k = 1$. The algorithm for selecting the representative inputs $\{\mathbf{X}_{\text{rep}}^k\}_{k=1}^K$ works as follows: Initially, one sample is randomly chosen from each initial state to create the initial set of representative inputs, with K representing the total number of initial states. After each iteration of model training and state label refinement, all input samples are projected into the learned latent space. We then calculate the center of each newly refined, non-empty metastable state and determine the nearest sample to each center by measuring Euclidean distance in this latent space. These chosen samples become the updated set of representative inputs $\{\mathbf{X}_{\text{rep}}^k\}_{k=1}^K$ for the next iteration. As a result, the algorithm dynamically adjusts the representative inputs $\{\mathbf{X}_{\text{rep}}^k\}_{k=1}^K$, ensuring that the number of representative inputs K always matches the number of states in $\mathbf{y}$. Moreover, by incorporating the mixture of Gaussians prior to Equation 5.3, the regularization term could encourage spatial separation among the state centers within the latent space.

With the optimal parameters θ^* determined by maximizing the SPIB objective function, we can revisit Equation 5.2. By using the deterministic output of SPIB, $\hat{p}(\mathbf{y}_{t+\Delta t}|\mathbf{X}_t) \equiv \mathcal{D}(\mu(\mathbf{X}); \Delta t, \theta^*)$, we can approximate the state-transition density $p(\mathbf{y}_{t+\Delta t}|\mathbf{X}_t)$ effectively. The updated rule for state label assignment can be expressed as:

$$\hat{\mathbf{y}}_t = \underset{i}{\text{argmax}}\ \mathcal{D}_i(\mu(\mathbf{X}_t); \Delta t, \theta^*). \tag{5.7}$$

The SPIB workflow is summarized as follows: To start the training process, SPIB requires both trajectory data, represented by input features, and an initial set of state labels. These initial labels are generated as the first step and serve as prior information, guiding the learning process within SPIB. One straightforward approach for generating initial state labels involves discretizing certain input order parameters based on expert intuition, a technique frequently used in previous research. For more complex systems, particularly those without clear intuitive guidance-such as the protein folding systems and multi-body systems-the initial state assignment can be approached similarly to traditional MSM construction methods. This entails applying dimensionality reduction techniques like PCA or tICA to select a subset of optimal linear combinations from a large set of input features, and then using clustering algorithms such as K-Means or K-Centers to define the discrete states. After generating the initial state labels, both the trajectory data $\mathbf{X}$ and the state labels $\mathbf{y}$ are fed into SPIB. The goal is to determine the optimal latent representation that best captures the key features of the past configuration $\mathbf{X}_t$ to accurately predict the future state $\mathbf{y}_{t+\Delta t}$. Once the model has learned this representation, the state labels are refined according to Equation 5.7, and the updated labels are reintroduced into SPIB. This iterative process continues until the latent representation and state labels converge, allowing for further analysis.

5.3 Evaluating MSM Performance with Quantitative Metrics

SPIB's ability to independently identify metastable states indicates its potential as a valuable tool for constructing MSMs. To assess the quality of the MSM generated by SPIB, we employed a range of quantitative metrics and systematically compared it with MSMs constructed through alternative methodologies. We adhered to the conventional sequence of dimensionality reduction,

clustering, and lumping, but applied different algorithms at each stage. For dimensionality reduction, we employed tICA[47, 76, 111] and PCA; for clustering, we used the k-means algorithm; and for lumping, we applied PCCA+[93] and MPP[92, 97]. This approach yielded four distinct pipeline combinations for constructing MSMs. To provide a comparison with other deep neural network-based methods, we also implemented the widely used VAMPnets[34] as a reference. For both SPIB and VAMPnet, we crisply assign state labels to MD conformations according to the highest probability output from the neural networks.

For the quantitative metrics, in line with benchmark studies on the same HP35 trajectory detailed in References[92, 135], and based on traditional scoring functions introduced in Chapter 2, we selected several metrics: the generalized matrix Rayleigh quotient (GMRQ)[90], metastability score[91], Shannon entropy[135], Davies-Bouldin index (DBI)[135], and implied timescales (ITS)[27]. These metrics are defined as follows:

1. GMRQ score: defined as the sum of the top n eigenvalues, λ_i, of the TPM:

$$\text{GMRQ} = \sum_{i=1}^{n} \lambda_i, \tag{5.8}$$

where n is the total number of eigenvalues scored. According to VAC theory[51] mentioned in Chapter 2, when the dynamics under study are reversible and detailed-balanced, the sum of the eigenvalues of the approximated propagator (i.e., GMRQ score) can serve as a variational score, providing a lower bound to the ground truth. Thus, maximizing VAMP-based scores typically results in larger eigenvalues for reversible propagators and, consequently, higher GMRQ scores[90]. Notably, the GMRQ score is also a widely utilized criterion in cross-validation for selecting optimal model hyperparameters, helping to prevent overfitting[90].

2. Metastability Q: defined as the average trace of the TPM, which assesses the likelihood that the system will stay in the same state after a lag time τ. High metastability generally signifies successful separation of slow inter-state dynamics from faster intra-state dynamics[91]. Explicitly, we define metastability Q as:

$$Q = \frac{1}{S}\text{tr}\left(\mathbf{T}(\tau)\right), \tag{5.9}$$

where S is, as previously defined, the number of metastable states spanned by $\mathbf{T}(\tau)$.

3. Shannon entropy H of the learned metastable states: defined in the usual information theoretic manner as

$$\mathrm{H} = -\sum_i \pi_i \log(\pi_i), \tag{5.10}$$

where π_i denotes the stationary or marginal probability of occupying metastable state i. Specifically, $\left(\pi^T \mathbf{T}(\tau)\right)_i = \pi_i$ (TPM is row-normalized). Thus, a higher value of H suggests that a substantial fraction of states is well-populated, which is preferred over partitions with a few densely populated states and many sparsely populated ones.

4. Furthermore, the DBI measures the ratio between the average distance of data points within a state and the distance between the centers of mass of different states. This metric indicates how effectively the metastable states are separated:

$$\mathrm{DBI} = \frac{1}{N} \sum_i \max_j \frac{s_i + s_j}{r_{ij}}, \tag{5.11}$$

where s_i represents he average distance of points in state i from the centroid of metastable state i, and r_{ij} is the distance between the centroids of states i and j. A low DBI value signifies well-separated, structurally distinct states.

5. Implied timescales (ITS): which monitors the timescales, t_i, of eigenmode i across different lag time, were calculated for TPMs with different lag time[27]:

$$t_i(\tau) = -\frac{\tau}{\ln |\lambda_i|}. \tag{5.12}$$

Here, π represents the lag time used to estimate the TPM, and λ_i denotes the i-th eigenvalue of the TPM. Typically, if the ITS converge and are independent of the lag time τ, it implies that the dynamics of the model satisfy the first-order master equation: $\lambda(n\tau) = \lambda_i(\tau)^n$. This characteristic can help identify the shortest Markovian lag time. Additionally, the Chapman-Kolmogorov (CK) test can be used as a supplementary validation tool to assess the Markovian properties of the model.

From the ITS analysis, two key factors can be used to evaluate the quality of MSMs: the Markovian lag time and the values of the converged timescales. The Markovian lag time is the shortest lag time at which all ITS converge, reflecting the time resolution of the MSMs. A shorter Markovian lag time indicates a more effective separation of slow inter-state dynamics from fast

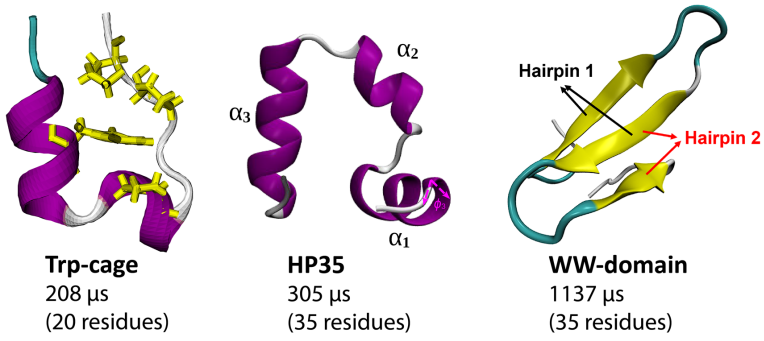

Figure 5.2: Protein systems investigated in this study. Data for all simulations is obtained from the DESRES protein folding trajectories. The duration of the MD simulation and the number of residues are specified for each case. This figure is reproduced from Wang. *et al.*[35]

intra-state dynamics. Moreover, since the lag time is limited by the length of the trajectory, a shorter Markovian lag time can reduce the required simulation length. Additionally, according to VAC theory, a model with larger converged timescales is better at capturing the leading slowest dynamics.

All analyses in this study are performed on the long equilibrium molecular dynamics trajectories of three mini-proteins from the DESRES group[114], namely Trp-cage (PDB:2JOF)[114], HP35 (PDB:2F4K)[136], and WW domain (PDB:2F21)[114] (Figure 5.2). All three datasets are characterized using the minimal residue-residue distances, calculated as the nearest distance between the heavy atoms of two residues separated by at least two neighboring residues in sequence, yielding 153 features for Trp-cage and 528 features each for HP35 and the WW-domain. To maintain computational feasibility, we analyze the HP35 and WW-domain datasets at a resolution of $1ns$ per frame. However, for the smaller Trp-cage system, we retain the original trajectory resolution of $0.2ns$ per frame.

To facilitate the comparison in the next section, we identified an optimal set of hyperparameters through ten-fold cross-validation using the GMRQ score as the scoring metric for all the algorithms employed in this Chapter. A more complex model generally has a higher capacity to capture significant slow dynamics, reflected in a larger GMRQ score. However, limited data necessitates

a balance between model complexity and generalization ability. The VAMP-2 score serves as an alternative criterion, akin to the GMRQ score, particularly when the dynamics are reversible.

To prepare the data, we initially divide the long equilibrium DESRES trajectory into 100 segments of equal length, treating these segments as independent trajectories to simulate the typical MSM construction process using multiple short trajectories. For 10-fold cross-validation, the 100 segments are shuffled and randomly subsampled as part of the train-test split procedure for each fold. All parameters are then evaluated using only the top 3 eigenvalues at a $100ns$ MSM lag time for all three systems. The $100ns$ lag time is confirmed as the Markovian lag time for all systems based on the ITS analysis and CK test. All GRMQ scores and metastability values reported in this chapter are calculated using MSMs with $100ns$ lag time.

5.4 Comparison of SPIB-MSM and Other State-Of-The-Art Methods

In this section, we will first examine the impact of lag time Δt on SPIB performance, followed by an evaluation of MSMs constructed by various methods, including SPIB, using different metrics. SPIB can be seen as a 'fast mode filter,' where the hyperparameter Δt-indicating how far into the future the model should predict-filters out short lived states and controls the degree of dynamic coarse-graining. Choosing the appropriate lag time Δt is essential for simplifying the learning process. When $\Delta t = 0$, SPIB ignores dynamics entirely and focuses only on clustering the input configurations into distinct states. In contrast, for $\Delta t > 0$, SPIB acts as a filter, excluding dynamics that occur faster than Δt. This filtering ability allows SPIB to disregard extraneous details in the dynamical processes, yielding a coarse-grained, dynamics-based understanding. Furthermore, as illustrated in Fig. 5.3, the increase in lag time Δt leads directly to a decrease in the number of metastable states one expects to find after a delay of Δt. This underscores a key advantage of SPIB: it automatically adjusts the number of metastable states based on the chosen Δt. As Δt increases towards infinity, the number of states will progressively decrease to one, representing the system's most stable state.

As shown in Figure 5.3, varying Δt in SPIB results in different numbers of converged metastable states. The lack of a clear plateau at shorter timescales suggests a complex free-energy landscape

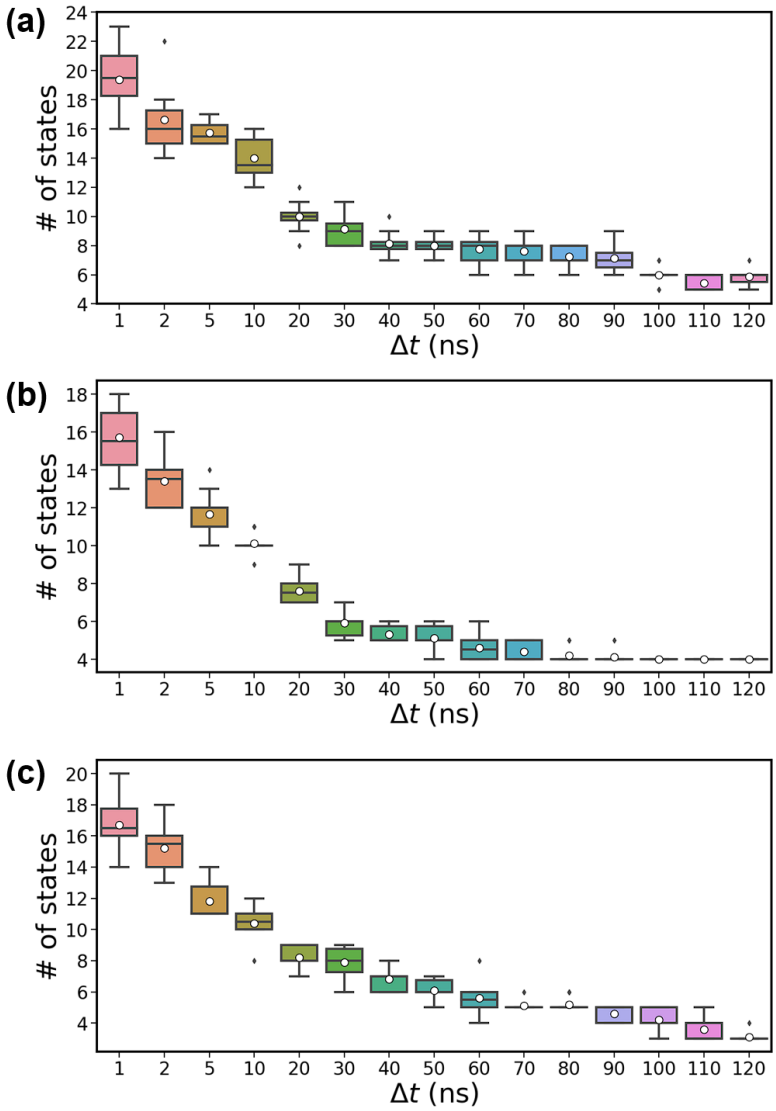

Figure 5.3: Impact of different lag time Δt choices on the number of converged SPIB states in 10-fold cross-validation for all three systems. This figure is reproduced from Wang. *et al.*[35]

in protein folding. Nonetheless, each choice of Δt and the corresponding states effectively capture the system's relevant dynamics at the chosen temporal resolution. This capability, as explored in subsequent subsections, highlights SPIB's strength in providing detailed insights into the hierarchical energy landscapes of simple proteins. To evaluate SPIB's performance both qualitatively and quantitatively, we select two distinct values of Δt: one large and one moderate. This choice results in two sets of MSMs: one with 4 to 5 states for the large Δt and another with approximately 10 states for the moderate Δt. We avoid using smaller Δt values to maintain fewer states for better interpretability.

Specifically, for Trp-cage, we select $\Delta t = 100$ ns which yielded 4 metastable states using SPIB. For HP35, we use $\Delta t = 110$ ns, which generates models with an average of 5 metastable states. Finally, for the WW-domain, we use $\Delta t = 70$ ns which led to 5 metastable states with SPIB. To ensure a fair and meaningful comparison, we align the number of states produced by all baseline methods with those generated by SPIB. This is accomplished by adjusting the relevant hyperparameters: the metastability threshold Q_{min} for MPP, the number of clusters m for PCCA+, and the number of output states for VAMPnets. Additionally, other hyperparameters are optimized through cross-validation using the GMRQ score, ensuring that comparisons are made between equivalently optimized MSMs from different methods.

Analysis reveals that VAMPnet consistently achieves the highest GMRQ scores and metastability across all three systems. This is because maximizing the VAMP-2 score for reversible dynamics inherently maximizes the eigenvalues of the TPM, leading to larger GMRQ scores and generally higher metastability. However, when not using the VAMP-based score as the objective function, SPIB shows comparable performance to both tICA-PCCA+ and VAMPnet in terms of GMRQ score and metastability on the validation data. Beyond demonstrating similar effectiveness to these leading methods in capturing slow dynamics, SPIB also reliably produces well-populated and structurally distinct states. As a result, it achieves comparable entropy and DBI values across all three systems when compared to tICA-PCCA+ and VAMPnet. Notably, other methods sometimes achieve higher entropy, as seen with PCA-PCCA+ for Trp-cage and PCA-MPP for WW-domain, or lower DBI, such as PCA-PCCA+ for WW-domain. However, these improvements in

Table 5.1: Quantitative comparison of different methods for large Δt in terms of GMRQ (scoring based on the top 3 eigenvalues), metastability Q, Shannon entropy, and DBI across three systems. The arrows indicate whether larger or smaller values are better for each metric. The reported values represent the mean along with the standard error of the mean derived from 10-fold cross-validation results. This figure is reproduced from Wang. *et al.*[35]

Model	Trp-cage Train				Trp-cage Validation			
	GMRQ ↑	Q ↑	H ↑	DBI ↓	GMRQ ↑	Q ↑	H ↑	DBI ↓
PCA-PCCA+	3.00 ± 0.02	0.749 ± 0.004	$\mathbf{0.895 \pm 0.006}$	1.807 ± 0.006	2.72 ± 0.08	0.68 ± 0.02	$\mathbf{0.86 \pm 0.04}$	1.85 ± 0.03
tICA-PCCA+	3.42 ± 0.01	0.855 ± 0.002	0.837 ± 0.005	$\mathbf{1.748 \pm 0.003}$	2.9 ± 0.2	$\mathbf{0.76 \pm 0.03}$	0.79 ± 0.05	1.81 ± 0.07
PCA-MPP	2.46 ± 0.03	0.56 ± 0.02	0.71 ± 0.02	2.0 ± 0.1	2.17 ± 0.09	0.50 ± 0.03	0.68 ± 0.05	2.1 ± 0.1
tICA-MPP	2.79 ± 0.08	0.67 ± 0.02	0.51 ± 0.01	2.14 ± 0.05	2.06 ± 0.09	0.50 ± 0.04	0.47 ± 0.05	2.1 ± 0.2
VAMPnet	$\mathbf{3.55 \pm 0.01}$	$\mathbf{0.888 \pm 0.002}$	0.807 ± 0.006	1.822 ± 0.003	$\mathbf{3.0 \pm 0.1}$	$\mathbf{0.76 \pm 0.03}$	0.76 ± 0.05	1.82 ± 0.04
SPIB	3.51 ± 0.01	0.878 ± 0.002	0.797 ± 0.006	1.810 ± 0.003	$\mathbf{3.0 \pm 0.1}$	0.75 ± 0.03	0.75 ± 0.05	$\mathbf{1.76 \pm 0.02}$
Model	HP35 Train				HP35 Validation			
	GMRQ ↑	Q ↑	H ↑	DBI ↓	GMRQ ↑	Q ↑	H ↑	DBI ↓
PCA-PCCA+	3.00 ± 0.02	0.73 ± 0.01	0.65 ± 0.08	3.5 ± 0.2	2.3 ± 0.2	0.62 ± 0.08	0.8 ± 0.1	3.5 ± 0.2
tICA-PCCA+	3.57 ± 0.02	0.80 ± 0.02	0.92 ± 0.05	5.3 ± 0.3	3.0 ± 0.1	$\mathbf{0.69 \pm 0.04}$	0.87 ± 0.06	4.6 ± 0.6
PCA-MPP	3.12 ± 0.02	0.64 ± 0.03	0.91 ± 0.08	$\mathbf{2.5 \pm 0.1}$	2.5 ± 0.2	0.28 ± 0.02	0.89 ± 0.08	$\mathbf{2.69 \pm 0.08}$
tICA-MPP	3.52 ± 0.02	0.76 ± 0.02	1.21 ± 0.05	3.6 ± 0.3	$\mathbf{3.1 \pm 0.1}$	$\mathbf{0.69 \pm 0.03}$	1.18 ± 0.04	3.1 ± 0.3
VAMPnet	$\mathbf{3.65 \pm 0.04}$	$\mathbf{0.83 \pm 0.03}$	0.8 ± 0.2	3.5 ± 0.2	2.9 ± 0.1	0.65 ± 0.02	0.77 ± 0.07	3.1 ± 0.2
SPIB	3.51 ± 0.02	$\mathbf{0.83 \pm 0.01}$	$\mathbf{1.26 \pm 0.01}$	3.74 ± 0.04	3.0 ± 0.1	0.67 ± 0.03	$\mathbf{1.19 \pm 0.03}$	3.4 ± 0.1
Model	WW-domain Train				WW-domain Validation			
	GMRQ ↑	Q ↑	H ↑	DBI ↓	GMRQ ↑	Q ↑	H ↑	DBI ↓
PCA-PCCA+	3.02 ± 0.01	0.652 ± 0.006	0.587 ± 0.006	$\mathbf{1.83 \pm 0.03}$	2.4 ± 0.1	0.49 ± 0.03	0.57 ± 0.04	$\mathbf{1.76 \pm 0.05}$
tICA-PCCA+	3.48 ± 0.01	0.788 ± 0.006	0.591 ± 0.006	2.27 ± 0.02	2.7 ± 0.2	0.58 ± 0.04	0.56 ± 0.04	2.19 ± 0.09
PCA-MPP	2.35 ± 0.06	0.46 ± 0.03	$\mathbf{0.9 \pm 0.1}$	8.5 ± 0.8	2.05 ± 0.07	0.41 ± 0.03	$\mathbf{0.91 \pm 0.09}$	7.9 ± 0.6
tICA-MPP	3.39 ± 0.05	0.79 ± 0.02	0.592 ± 0.006	1.99 ± 0.05	2.6 ± 0.2	0.60 ± 0.05	0.56 ± 0.03	1.8 ± 0.1
VAMPnet	$\mathbf{3.64 \pm 0.01}$	$\mathbf{0.841 \pm 0.003}$	0.592 ± 0.003	2.22 ± 0.02	$\mathbf{2.8 \pm 0.2}$	0.58 ± 0.04	0.56 ± 0.04	2.0 ± 0.1
SPIB	3.59 ± 0.01	0.823 ± 0.003	0.627 ± 0.003	2.29 ± 0.07	$\mathbf{2.8 \pm 0.2}$	$\mathbf{0.67 \pm 0.03}$	0.61 ± 0.05	2.28 ± 0.09

Table 5.2: Quantitative comparison of different methods for moderate Δt in terms of GMRQ (scoring based on the top 5 eigenvalues), metastability Q, Shannon entropy, and DBI across three systems. The reported values represent the mean along with the standard error of the mean derived from 10-fold cross-validation results. This figure is reproduced from Wang. *et al.*[35]

Model	Trp-cage Train				Trp-cage Validation			
	GMRQ ↑	Q ↑	H ↑	DBI ↓	GMRQ ↑	Q ↑	H ↑	DBI ↓
PCA-PCCA+	3.42 ± 0.02	0.401 ± 0.002	1.33 ± 0.02	$\mathbf{2.10 \pm 0.05}$	2.9 ± 0.1	0.31 ± 0.01	$\mathbf{1.27 \pm 0.04}$	$\mathbf{2.17 \pm 0.06}$
tICA-PCCA+	4.22 ± 0.02	0.506 ± 0.005	1.00 ± 0.03	2.55 ± 0.09	3.1 ± 0.1	0.32 ± 0.02	0.92 ± 0.05	3.0 ± 0.1
PCA-MPP	2.96 ± 0.03	0.323 ± 0.008	1.16 ± 0.04	2.3 ± 0.1	2.39 ± 0.08	0.25 ± 0.01	1.12 ± 0.07	2.3 ± 0.1
tICA-MPP	3.52 ± 0.07	0.39 ± 0.01	0.66 ± 0.05	4.3 ± 0.4	2.3 ± 0.1	0.26 ± 0.01	0.61 ± 0.07	4.5 ± 0.3
VAMPnet	$\mathbf{5.05 \pm 0.02}$	$\mathbf{0.735 \pm 0.006}$	1.00 ± 0.01	3.0 ± 0.3	3.0 ± 0.1	0.37 ± 0.05	0.86 ± 0.06	2.5 ± 0.2
SPIB	4.48 ± 0.03	0.538 ± 0.009	$\mathbf{1.35 \pm 0.05}$	3.03 ± 0.08	$\mathbf{3.5 \pm 0.1}$	$\mathbf{0.40 \pm 0.02}$	1.25 ± 0.08	2.8 ± 0.1
Model	HP35 Train				HP35 Validation			
	GMRQ ↑	Q ↑	H ↑	DBI ↓	GMRQ ↑	Q ↑	H ↑	DBI ↓
PCA-PCCA+	3.27 ± 0.08	0.38 ± 0.01	$\mathbf{1.53 \pm 0.02}$	$\mathbf{2.8 \pm 0.1}$	2.6 ± 0.2	0.29 ± 0.02	$\mathbf{1.50 \pm 0.05}$	$\mathbf{2.91 \pm 0.08}$
tICA-PCCA+	4.816 ± 0.008	0.619 ± 0.004	1.45 ± 0.02	5.2 ± 0.2	4.0 ± 0.1	$\mathbf{0.49 \pm 0.02}$	1.40 ± 0.05	4.6 ± 0.2
PCA-MPP	3.83 ± 0.02	0.34 ± 0.02	1.2 ± 0.1	3.4 ± 0.1	2.8 ± 0.2	0.22 ± 0.02	1.1 ± 0.1	3.5 ± 0.1
tICA-MPP	4.47 ± 0.06	0.50 ± 0.03	1.27 ± 0.03	5.8 ± 0.8	$\mathbf{4.1 \pm 0.8}$	0.47 ± 0.03	1.24 ± 0.04	3.9 ± 0.4
VAMPnet	$\mathbf{5.21 \pm 0.08}$	0.61 ± 0.02	1.41 ± 0.04	4.1 ± 0.3	3.4 ± 0.3	0.37 ± 0.02	1.37 ± 0.05	4.3 ± 0.2
SPIB	5.02 ± 0.04	$\mathbf{0.65 \pm 0.02}$	1.511 ± 0.009	4.7 ± 0.1	3.8 ± 0.2	0.46 ± 0.03	1.44 ± 0.05	3.8 ± 0.1
Model	WW-domain Train				WW-domain Validation			
	GMRQ ↑	Q ↑	H ↑	DBI ↓	GMRQ ↑	Q ↑	H ↑	DBI ↓
PCA-PCCA+	3.72 ± 0.01	0.44 ± 0.02	0.67 ± 0.01	$\mathbf{2.13 \pm 0.04}$	2.9 ± 0.1	0.32 ± 0.02	0.65 ± 0.04	$\mathbf{2.11 \pm 0.06}$
tICA-PCCA+	4.65 ± 0.01	0.622 ± 0.004	0.773 ± 0.005	4.72 ± 0.03	$\mathbf{3.7 \pm 0.2}$	0.43 ± 0.02	0.75 ± 0.06	4.30 ± 0.08
PCA-MPP	2.37 ± 0.05	0.221 ± 0.007	$\mathbf{1.2 \pm 0.1}$	7.9 ± 0.5	2.06 ± 0.07	0.19 ± 0.01	$\mathbf{1.1 \pm 0.1}$	7.7 ± 0.5
tICA-MPP	3.51 ± 0.01	0.54 ± 0.01	0.67 ± 0.01	5.0 ± 0.4	2.9 ± 0.2	0.39 ± 0.02	0.65 ± 0.04	4.2 ± 0.4
VAMPnet	$\mathbf{4.91 \pm 0.02}$	$\mathbf{0.669 \pm 0.006}$	0.75 ± 0.01	4.18 ± 0.07	3.4 ± 0.2	$\mathbf{0.44 \pm 0.03}$	0.72 ± 0.05	3.7 ± 0.1
SPIB	4.75 ± 0.02	0.57 ± 0.01	0.95 ± 0.02	6.5 ± 0.4	3.6 ± 0.1	0.42 ± 0.02	0.92 ± 0.05	5.4 ± 0.3

entropy or reductions in DBI often come at the expense of dynamical accuracy, leading to significantly poorer GMRQ scores and metastability compared to SPIB. Thus, the quantitative analysis in Table 5.1 concludes that SPIB delivers state-of-the-art performance across four diverse metrics for assessing the quality of kinetic models in all three protein folding systems.

To gain a more detailed understanding of the underlying dynamics with improved temporal and spatial resolution, we select a moderate Δt to identify additional metastable states using SPIB. Specifically, we set $\Delta t = 10$ ns for Trp-cage and WW-domain, and $\Delta t = 20$ ns for HP35, aiming to capture approximately 10 metastable states in each case. Table 5.2 consolidates all quantitative comparison results for the moderate Δt. As with the large Δt scenario, SPIB shows competitive or slightly superior performance in validation GMRQ scores and metastability compared to tICA-PCCA+ and VAMPnet across all three systems. Notably, SPIB excels at identifying a larger number of well-populated states, resulting in higher entropy scores on both training and validation sets for all three proteins. In contrast, tICA-PCCA+ and VAMPnets often produce lower entropy scores, suggesting they capture many sparsely populated states due to an overemphasis on slow dynamics. Additionally, PCA-PCCA+ consistently shows the lowest DBI, indicating the presence of the most structurally distinct states.

To thoroughly compare the Markovian properties of MSMs constructed using PCA-PCCA+, tICA-PCCA+, VAMPnets, and SPIB, we present a detailed analysis in Figure 5.4. This figure offers an in-depth view of ITS convergence as a function of lag time. The ITS are visualized using the mean value obtained from 10 rounds of bootstrapping, where data is randomly sampled with replacement. SPIB consistently achieves short Markovian lag times and large converged timescales, particularly for the slowest processes, making its performance competitive with VAMPnet. However, VAMPnet outperforms SPIB in Trp-cage and HP35 for other slow processes. This advantage might be due to potential overfitting in VAMPnets, as SPIB maintains comparable GMRQ scores in the validation set (Table. 5.1 and 5.2). This suggests a potential advantage of SPIB: VAMP-based methods are prone to overfitting, especially with a large number of states, due to the statistical uncertainty in estimating singular functions. In contrast, SPIB is trained in a self-consistent manner, which generally makes it more robust and stable.

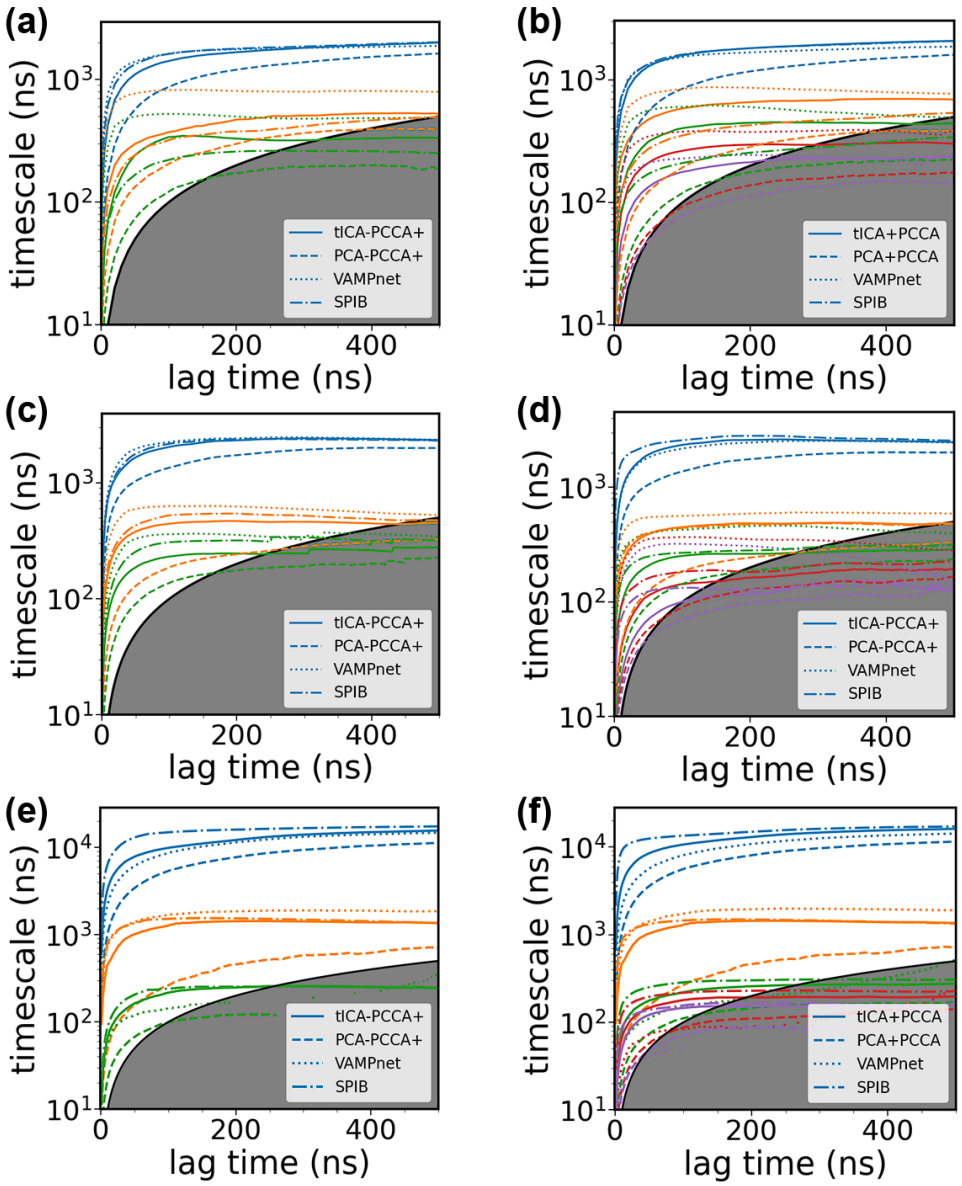

Figure 5.4: Implied timescales as a function of lag time for the MSMs of all systems. The left panels illustrate the results for 4-state MSMs in Trp-cage and 5-state MSMs in HP35 and WW-domain, while the right panels showcase the outcomes for 10-state MSMs. For clarity in presentation, only the mean values from 10 bootstrapping samples are plotted. The shaded gray area represents the region where timescales become equal to or smaller than the lag time and can no longer be resolved. This figure is reproduced from Wang. *et al.*[35]

Across all three systems, SPIB outperforms tICA-PCCA+ and PCA-PCCA+ methods in the ITS, as the latter two fail to converge at the same lag times, and the converged timescales are slightly smaller. This observed trend is attributed to the capacity of SPIB to learn nonlinear transformations of input coordinates, providing enhanced resolution of slower processes, and the use of a continuous basis set for MSM macrostates construction. These enhanced capabilities yield a reduction of discretization errors compared to the Galerkin method when approximating the dynamical propagator. This results in higher converged timescales and the generation of state models with clearer time separations, leading to shorter relaxation times within states and, consequently, a reduced Markovian lag time.

5.5 Elucidating Biophysical Mechanisms with SPIB-MSM

In this section, we will demonstrate SPIB-MSM's ability to elucidate the key biophysical mechanisms involved in the folding of three proteins (i.e., Trp-cage, HP35 and WW-domain). We believe that the SPIB-MSM approach has great potential for future studies of conformational changes in multi-body systems.

5.5.1 Trp-cage

The Trp-cage protein, consisting of 20 residues, is a well-known example of a small folding protein. It folds into a native state with an N-terminal α-helix, a short 3_{10}-helix, a C-terminal polyproline II region, and a hydrophobic core stabilized by interactions between the Trp6 side chain and Pro12, Pro18, and Pro19, as shown in Figure 5.2. Figure 5.5 demonstrates how SPIB effectively identifies metastable states at varying levels of coarse-graining. With a large lag time Δt, SPIB discerns and represents the system with 4 states, as shown in Figure 5.5(a, c). Additionally, using a moderate lag time, SPIB captures more detail by elucidating 10 states, as depicted in Figure 5.5(b, d). The connection between the 4-state and 10-state models is clearly depicted in the Sankey plot shown in Figure 5.5(e). This plot visually maps one set of states to the other, effectively revealing the hierarchical arrangement of metastable states[92].

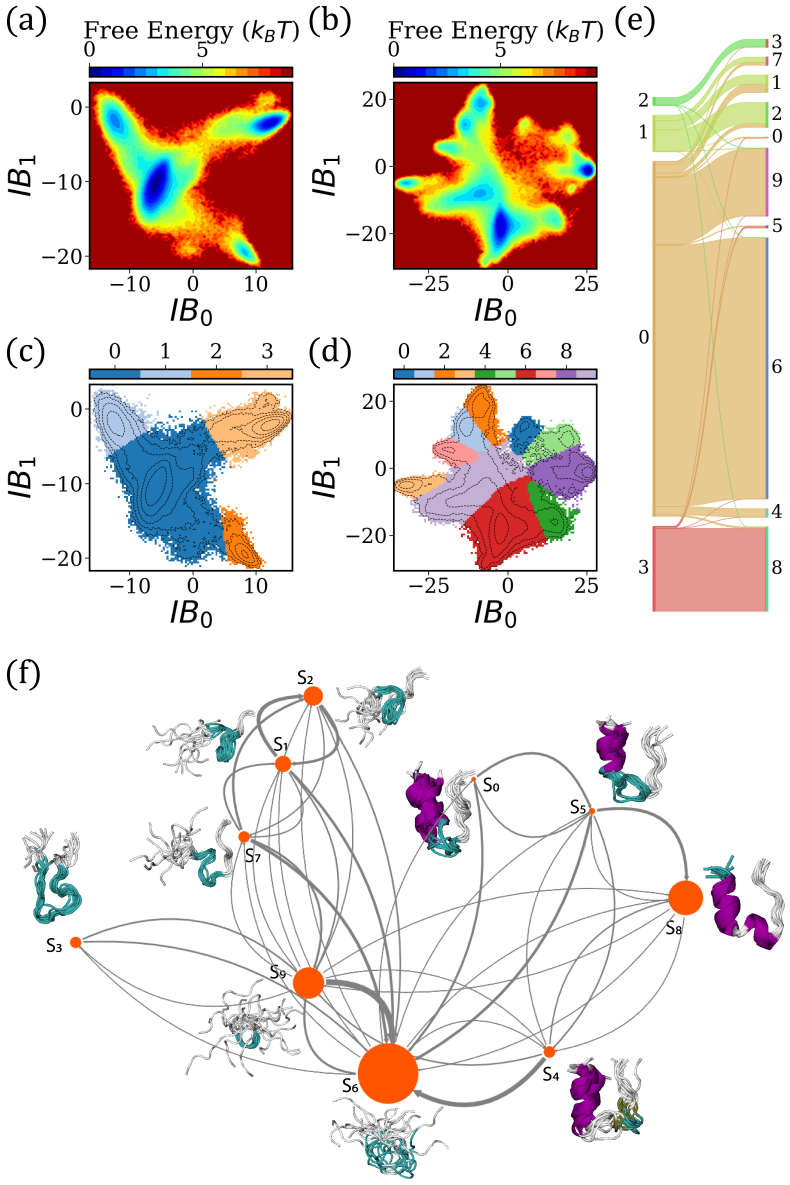

Figure 5.5: Qualitative description of the MSM analysis for Trp-cage protein. (a) and (b) give the free-energy surfaces in the two-dimensional SPIB latent space, denoted by IB_0 and IB_1, for large and moderate Δt, respectively. (c) and (d) give the metastable states learned by SPIB in the case of large and moderate Δt, respectively. (e) The Sankey plot illustrates the corresponding relations between states learned by SPIB using large (left) and moderate (right) Δt. (f) The MSM constructed based on states identified by SPIB, trained with a moderate Δt, is visualized using a flux network. The node size is proportional to the stationary population of the states, and the arrow width is scaled according to jump probabilities. Additionally, ten randomly selected conformations from each state are overlapped and displayed adjacent to the corresponding node. This figure is reproduced from Wang. *et al.*[35]

With a large Δt, SPIB produces a minimal MSM for analysis. Key features of the landscape include a folded state represented by state 3, which is intricately linked to the molten globule state (state 0) through a narrow bottleneck. State 0 connects to the unfolded state (state 1), characterized by multiple turns in the structure, via a small energy barrier, and to a hairpin state (state 2) through a significant energy barrier.

When the lag time Δt is reduced, the initial 4-state model undergoes further refinement, evolving into a more detailed 10-state model, as shown in Figure 5.5(e). The metastable conformational ensembles and associated transitions for the 10 macrostates are visually depicted in Figure 5.5(f). Notably, S_8 corresponds to the folded state, while S_0 and S_5 represent intermediate states that bridge the folded and unfolded states, featuring a partially folded N-terminus. Additionally, S_9 embodies a crossed conformation with a minor central hairpin, and S_6 comprises a mix of molten globule structures and an extended conformation. S_4 indicates a partially unfolded state. Additionally, S_1 and S_2 adopt a braided hairpin-like structure, S_7 appears as a partially compact configuration with multiple turns, and S_3 displays a distinct hairpin conformation. These identified states are consistent with those reported in the literature, confirming their relevance and accuracy[137]. A clear correspondence emerges between the SPIB-learned latent space, as depicted in Figure 5.5(b,d), and the constructed MSM network shown in Figure 5.5(f). This observation suggests that SPIB actually learns a continuous embedding of the MD conformations, serving as an information bottleneck that maximally preserves information about state-to-state transitions.

Additionally, we also comprehensively exam the consequences of untilizing alternative methods for macrostate construction. For the 4-state model, SPIB aligns closely with VAMPnets, whereas tICA-PCCA+ and PCA-PCCA+ identify states but have difficulty with precise boundaries. Conversely, PCA-MPP and tICA-MPP often fail to accurately identify or distinguish the ensemble of unfolded states, frequently missing one or two significant unfolded states. For the 10-state model, SPIB excels in capturing a more refined model with numerous well-populated macrostates. In contrast, other methods struggle to effectively subdivide highly populated states, resulting in the identification of many sparsely populated states. Although using MPP and PCCA+

on PCA helps mitigate this issue, their overall performance in dynamic metrics remains generally suboptimal.

5.5.2 HP35

Figure 5.6 shows the analogous qualitative results for HP35. Figure 5.6(a,b) show the free-energy surfaces in the SPIB latent space for a large (100 ns) and moderate (20 ns) values of Δt, indicating the kinetic model at a coarse and fine resolution, respectively. Figure 5.6(c,d) illustrates the metastable partitions identified by SPIB on these free-energy surfaces. As anticipated, using a smaller lag time enables the identification of a greater number of metastable states. Figure 5.6(e) presents a Sankey plot that maps the topological relationships between the states learned with a large Δt and those learned with a moderate Δt. The topological changes primarily involve a more detailed partitioning of the unfolded state (state 1 in the 5-state model) into a series of substates, while the folded states remain largely unchanged.

Figure 5.6(f) illustrates the state-to-state transition network for the discrete model learned by SPIB at moderate Δt. This network reveals that the two most populated states, S_5 and S_8, are distinguished by changes in the ϕ_3 dihedral angle (Figure. 5.2, middle). Previously, this dihedral angle has been shown to effectively differentiate between folded states in HP35[92]. In our 10-state analysis, S_5 corresponds to $\phi_3 > 0$ rad and S_8 to $\phi_3 < 0$ rad, while in the 5-state analysis, they correspond to states 2 and 4, respectively. The remaining eight states in the 10-state model represent varying degrees of unfolding or misfolding. The unfolded state 1 in the 5-state model is further decomposed into states S_1, S_2, S_3, S_4, and S_7 in the 10-state model. Although all these states exhibit some level of secondary structure, they differ in the proportion and positioning of this structure. Specifically, state S_1 features an unfolded α_2 and misfolded α_3, with α_1 folded. State S_2 has α_1 unfolded and both α_2 and α_3 misfolded. States S_3 and S_4 both show α_1 unfolded while α_2 and α_3 are folded. State S_7 represents a fully unfolded conformation.

State S_3 features α_1 unfolded and both α_2 and α_3 folded, acting as a central node in the folding pathway and indicating that α_1 is the last to fold in the 10-state SPIB model. Transition path theory analysis, tracking the journey from unfolded state S_2 to the folded state S_8, suggests a

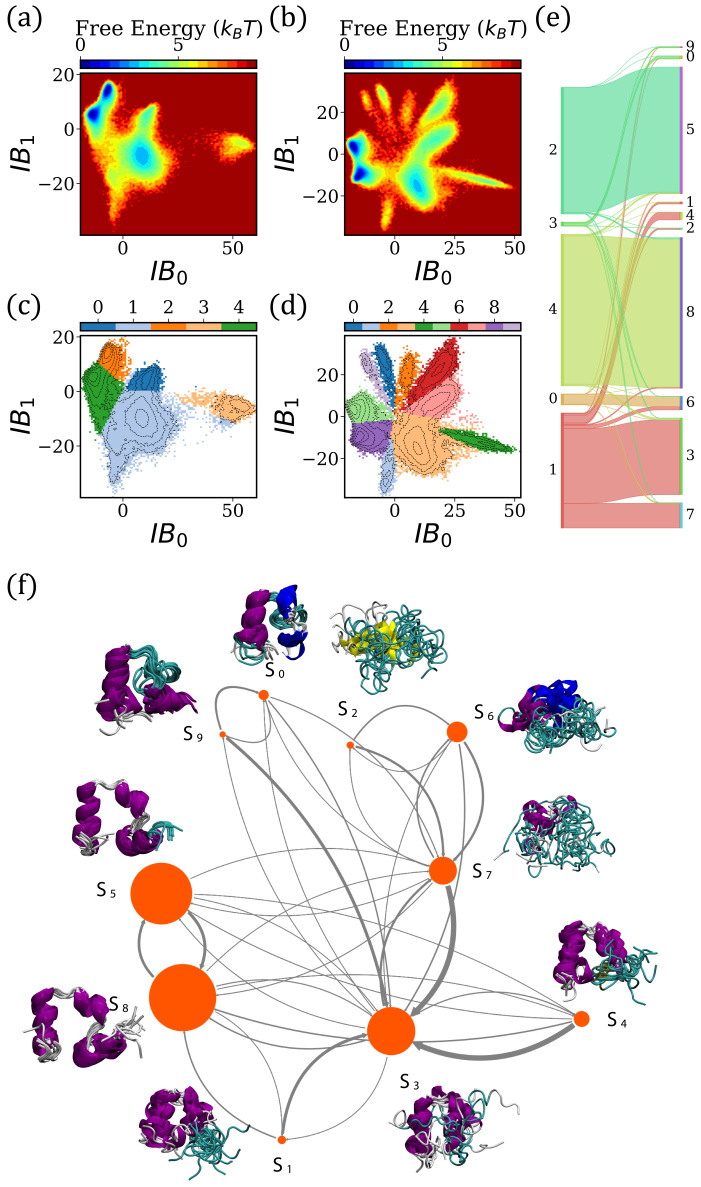

Figure 5.6: Qualitative description of the MSM analysis for HP35 protein. (a) and (b) give the free-energy surfaces in the two-dimensional SPIB latent space for large (100 ns) and moderate ($20ns$) Δt, respectively. (c) and (d) give the metastable states learned by SPIB in the case of large and moderate Δt, respectively. (e) The Sankey plot illustrates the corresponding relations between states learned by SPIB using large (left) and moderate (right) Δt. (f) The MSM constructed based on states identified by SPIB, trained with a moderate Δt, is visualized using a flux network. The node size is proportional to the stationary population of the states, and the arrow width is scaled according to jump probabilities. Additionally, ten randomly selected conformations from each state are overlapped and displayed adjacent to the corresponding node, with the secondary structure for each frame templated on a single, randomly selected frame from all ten. This figure is reproduced from Wang. *et al.*[35]

folding sequence where α_3 folds first, followed by α_2, and then α_1. States S_0 and S_9, where α_1 folds before α_2, are less involved in the main folding trajectory and are identified as misfolded trap states in our model. Overall, S_7 functions as a sink for both the misfolded state S_1 and other unfolded states. However, the qualitative analysis of the 10-state model suggests that there is no single dominant folding-unfolding pathway for the HP35 protein. Instead, the folding process appears to be complex and multifaceted, lacking a clear, predominant route even for this relatively simple protein.

We further make a comparison of the SPIB metastable states at both the 5 and 10 state level of resolution and those discovered by the PCCA+, MPP, and VAMPnet approaches. We find that only SPIB and tICA-MPP can clearly and completely differentiate between the two folded states distinguished by the ϕ_3 dihedral angle. All other methods tested merge these two states into a single folded metastable state at the coarser 5-state level. At the finer 10-state level, all methods are capable of distinguishing the two folded states based on the ϕ_3 angle. However, VAMPnet resolves the two folded states correctly in only 40% of the models, which aligns with its occasional difficulty in predicting the third-slowest timescale in alanine dipeptide systems[34].

5.5.3 WW-Domain

The WW-domain protein consists of 35 residues which could form a three-stranded beta-sheet, with residues 8-23 forming hairpin 1 and residues 17-30 forming hairpin 2 (as shown in Figure 5.2). Recent investigations, employing both experimental techniques and MD simulations, have explored the folding mechanism of the WW-domain. Two distinct folding mechanisms have been elucidated, differing in the folding order of hairpin 1 and hairpin 2. Approximately 70% of the WW-domain protein folding process involves the sequential folding of hairpin 1 followed by hairpin 2, while the remaining 30% undergoes folding in the opposite order[40, 48, 118, 120]. In this Chapter, adopting SPIB for MSMs analysis, we obtained results qualitatively consistent with previous studies.

Constructing MSMs for the WW-domain protein folding system using SPIB, Figure 5.7 showcases the results from two SPIB models trained at $\Delta t = 70$ ns and 10 ns, respectively. Employing

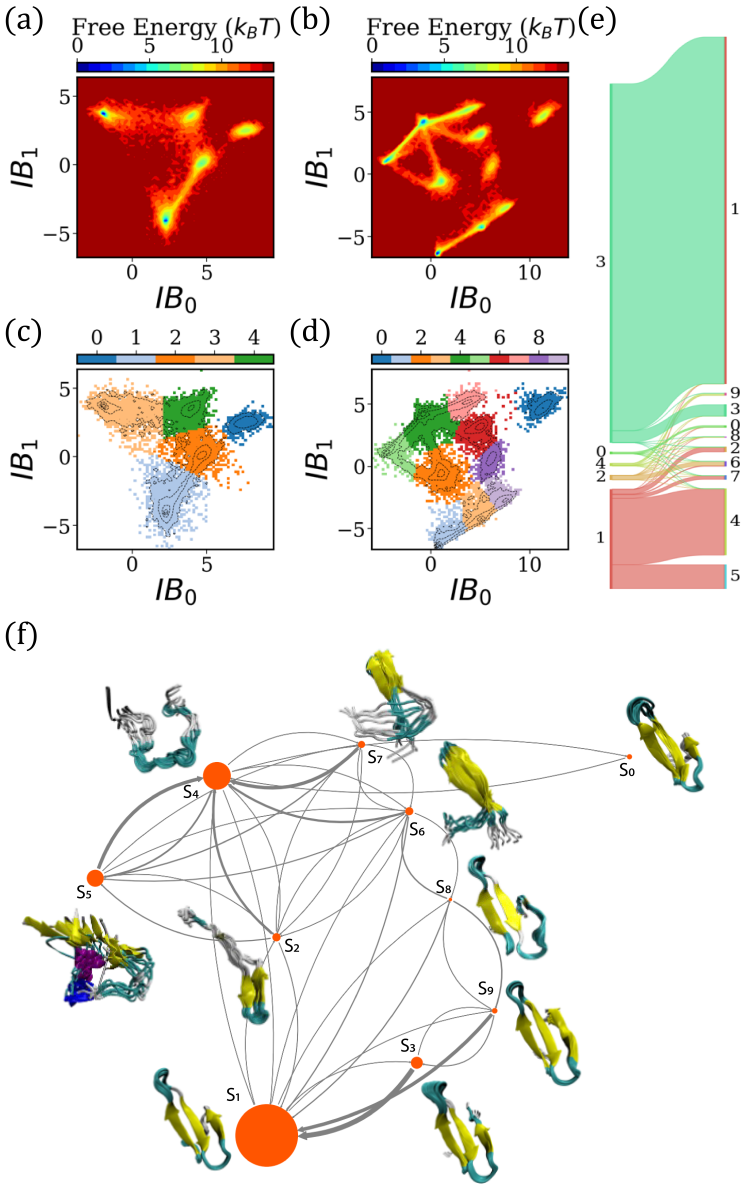

Figure 5.7: Qualitative description of the MSM analysis for WW-domain protein. (a) and (b) give the free-energy surfaces in the two-dimensional SPIB latent space for large and moderate Δt, respectively. (c) and (d) give the metastable states learned by SPIB in the case of large and moderate Δt, respectively. (e) The Sankey plot illustrates the corresponding relations between states learned by SPIB using large (left) and moderate (right) Δt. (f) The MSM constructed based on states identified by SPIB, trained with a moderate Δt, is visualized using a flux network. The node size is proportional to the stationary population of the states, and the arrow width is scaled according to jump probabilities. Additionally, ten randomly selected conformations from each state are overlapped and displayed adjacent to the corresponding node. This figure is reproduced from Wang. *et al.*[35]

a large Δt yielded a highly coarse-grained model with 5 states. The latent space of the trained SPIB model and the distributions of different states are visualized in Figure 5.7(a) and (c). Using a much smaller Δt of 10 ns, a higher-resolution model with 10 states is obtained, as shown in Figure 5.7(b) and (d). The representative conformations for the 10 states and their transition relationships are elucidated in the network flux plot in Figure 5.7(e). The corresponding relationships between the 5-state and 10-state models are illustrated in Figure 5.7(f).

Clearly, both the 5-states and 10-states MSMs introduced by SPIB successfully identified the folded, unfolded, misfolded, and partially folded states. In the 5-state MSM, state 0 corresponds to the misfolded state, while states 1 and 3 correspond to the unfolded and folded states, respectively. State 2 represents the partially folded hairpin 1 state, and state 4 includes both partially folded hairpin 1 and hairpin 2 conformations. This model effectively links the unfolded and folded states through partially folded intermediates and separates the misfolded state. However, due to the limited resolution of the model, distinctions between different folding mechanisms are not very evident.

In the higher resolution 10-state MSM, the two folding mechanisms are clearly identified. States S_4 and S_5 represent the unfolded states, states S_1 and S_3 correspond to the folded states, and S_0 is identified as the misfolded state. Additionally, two distinct pathways between unfolded and folded states emerge: states 7-6-8-9 represent the hairpin 1 to hairpin 2 folding sequence, while state 2 shows the hairpin 2 to hairpin 1 sequence. The widths of the flux arrows support the dominance of the hairpin 1 to hairpin 2 pathway. Interestingly, the folded and unfolded states are further subdivided based on different interactions between the terminal regions of the protein. This model significantly enhances our understanding of the WW-domain protein folding process.

Overall, when comparing with other methods, we find that for the 5-state model, SPIB results align with those produced by tICA-PCCA+, tICA-MPP, and VAMPnets. Given the presence of multiple states with relatively low populations (i.e., $< 1\%$) for the WW-domain, distinguishing the partially folded states proves challenging for PCA-based methods. In the 10-state model, while PCA-based methods still struggle to separate kinetically stable states, VAMPnets face difficulties distinguishing variations within partially folded and unfolded conformations. Results from the

tICA-PCCA+ method are mostly consistent with those from SPIB. Overall, SPIB demonstrates robust performance and effectively distinguishes various meaningful metastable states, highlighting its significant potential for constructing MSMs.

5.6 Discussion and Conclusion

In this work, we utilize the State Predictive Information Bottleneck (SPIB) to construct multi-resolution MSMs from MD simulation trajectories of protein conformational changes. This framework integrates the variational information bottleneck principle with a straightforward heuristic for state metastability, employing a flexible neural network to achieve both feature extraction and state partitioning in a unified approach. Using various quantitative and qualitative metrics across three distinct mini-proteins (Trp-cage, villin headpiece [HP35], and WW-domain), our study highlights the distinct advantages of the SPIB approach over competing methods, with minimal hyperparameter tuning. These advantages include its ability to automatically determine the number of metastable states based on the specified minimum time resolution, achieving an optimal balance between capturing slow dynamics and identifying significantly populated states, and providing a direct visualization of the underlying dynamics.

Without explicitly optimizing the VAMP-based score, SPIB consistently achieves state-of-the-art performance in capturing the slowest dynamical processes, performing comparably or slightly better in validation GMRQ and metastability compared with other advanced methods. Moreover, the top few slowest ITS of MSMs constructed by SPIB rapidly converge to their timescales even with shorter lag times, indicating an accurate Markovian kinetic model.

Beyond its proficiency in capturing slow dynamics, SPIB offers a distinct advantage in constructing a more nuanced MSM with 10 states. While VAMP-based methods optimize overall kinetic performance, they often struggle to subdivide highly populated states further. In contrast, SPIB excels at learning numerous well-populated macrostates, thanks to its optimization of likelihood through the information bottleneck principle, which prioritizes slow transitions with significant probabilities. This characteristic distinguishes SPIB from VAMP-based methods, allowing

it to excel in capturing intricate structural details. This effectiveness enables SPIB to differentiate among various metastable states, particularly in discerning subtle differences within folded or unfolded conformations in protein folding studies.

Our results also indicate that SPIB learns a low-dimensional, continuous embedding of MD conformations that maximally preserves information about state-to-state transitions. This capability facilitates direct visualization and offers a more insightful interpretation of the folding and unfolding pathways in mini-proteins within a 2D space. This contrasts with many existing dimension reduction methods that directly approximate eigenfunctions.

Analysis of all three mini-proteins reveals that SPIB uncovers a hierarchical structure in the free energy landscape governing their folding processes. This organizational structure segments both the native and unfolded basins into several well-populated metastable states, while also identifying a few less populated intermediate or misfolded states. Even in these seemingly simple proteins, folding processes are intricate, involving multiple pathways of folding and unfolding. We believe our algorithm presents a novel, practical, and robust method for constructing MSMs. It is especially promising for future implementation in multi-body systems, with potential applications spanning molecular simulations and the analysis of complex dynamical systems. We expect it to be highly valuable across diverse scientific disciplines.

Chapter 6

Non-Markovian Dynamic Models Identify Non-Canonical KRAS-VHL Encounter Complex Conformations for Novel PROTAC Design

This chapter is reproduced in part with permission from Qiu, Y; Wiewiora, R.P.; Izaguirre, J.A.; Xu, H.; Sherman, W.; Tang, W; & Huang, X; Non-Markovian Dynamic Models Identify Non-Canonical KRAS-VHL Encounter Complex Conformations for Novel PROTAC Design, *JACS Au* **2024**, jacsau.4c00503.

6.1 Introduction

As clearly stated in Chapter 1, protein-protein interactions (PPIs) and protein-ligand complexes are typical multi-body systems encountered in drug discovery. A deep understanding of their thermodynamic and kinetic properties will greatly enhance the rational design of new drug compounds. In this chapter, we will demonstrate how our non-Markovian dynamics models and unsupervised machine learning algorithms can help elucidate non-canonical PPIs between two proteins in complexes with three ligands and predict corresponding properties, thereby facilitating the future design of drug degraders namaed proteolysis targeting chimeras (PROTACs).

Small molecule heterobifunctional degraders, exemplified by PROTACs, have the potential to transform drug discovery and therapeutic interventions by *degrading* proteins instead of *inhibiting* them [138–141]. Unlike the traditional small-molecule inhibitors that block the protein function through occupying the active or allosteric site of the protein of interest (POI), PROTACs can employ functional or non-functional binders to target the POI, inducing its degradation through a catalytic mechanism [142, 143]. This approach provides opportunities to target many undruggable POIs that lack well-defined small molecule binding sites for functional blockade. A PROTAC comprises three distinct components: warhead, linker, and E3 ligand. With the warhead binding to

the POI and E3 ligand binding to the E3 ligase, PROTAC facilitates the proximity between the POI and the E3 ubiquitin ligase, leading to the formation of a ternary complex. This complex could then trigger the ubiquitination of the POI, marking it for degradation by the cellular proteasome machinery. Over the past two decades, significant effort has been dedicated to investigating and designing PROTACs. However, the development of most PROTACs remains highly empirical, involving the time-consuming synthesis and screening of libraries with various linkers between the warhead and the E3 ligase ligand. This process aims to induce favorable non-native PPIs between the POI and E3 ligase.

Throughout the PROTAC-induced targeted protein degradation (TPD), establishing specific PPI between the POI and E3 ligase is critical [5, 144–146]. Many degraders function by leveraging the stabilization of pre-existing but weak PPIs between POIs and E3 ligases [5, 146, 147]. Additionally, both experiments and computational simulations reveal that PPIs of highly productive ternary complexes exhibit noticeable dynamical conformational heterogeneities, distinct from the static contacts found in crystal structures [5, 148–151]. Previous biophysical and structural studies have also demonstrated that different PROTACs, even with the same warhead and E3 ligand but different linkers, can induce distinct PPIs in ternary complexes, leading to significant differences in degradation efficiency [144, 150–158]. Therefore, investigating the complex and dynamic non-native PPIs between the POIs and E3 ligases is critical for understanding TPD mechanisms and guiding the rational design of novel PROTACs. An approach with great potential to explore all possible inherent PPIs between the POI and E3 ligase is to study the POI-E3 ligase encounter complex without the linker [145, 159–161]. The subsequent introduction of the linker to this encounter complex is akin to adding an additional geometric constraint.

All-atom MD simulation offers a promising approach to reveal metastable and dynamical PPIs between the POI and E3 ligase [5, 145]. It has been combined with enhanced sampling techniques to elucidate both the kinetic and thermodynamic properties of a PROTAC system [162]. However, simulating the formation of PPIs presents significant challenges due to the various ways in which the POI and E3 ligase can approach each other, as well as the conformational changes induced upon the formation of the encounter complex. The formation and conformational changes of encounter

complex PPI interfaces often occur on milliseconds timescales, which exceed accessible length of the straightforward MD simulations for a system at the size of approximately 150,000 atoms. Adding to the complexity, there is a lack of dominant PPI, and all PPIs may potentially serve as functional ones for PROTAC design [145]. Therefore, obtaining a comprehensive understanding of the conformational space of the encounter complex and identifying representative PPI interfaces, along with their equilibrium populations and transition rates between them, are challenging.

MSMs built from extensive MD simulations offer a potentially useful technique to address these challenges [13, 27, 29–32, 34, 48, 111, 163]. MSMs model dynamics through a series of Markovian jumps among conformational states at discrete lag times. MSMs also provide a rigorous pipeline to coarse-grain MD conformations into a few comprehensible states according to their dynamic metastability, facilitating the prediction of thermodynamic and kinetic properties associated with them. However, for MSMs to have predictive power, they must be constructed with a sufficiently long lag time to ensure that inter-state transitions become Markovian, posing a major challenge as the lag time is constrained by the length of short MD simulations [30–32]. To address this challenge, we recently developed an approach based on the Generalized Master Equation (GME), called the Integrative Generalized Master Equation (IGME) method[31]. IGME captures non-Markovian dynamics by incorporating time-integrations of memory kernel functions, offering a promising approach to study PPIs in encounter complexes based on relatively limited MD simulation data.

In this chapter, we constructed an IGME model from 2,492 MD trajectories, with an average length of 605 ns ($\sim$1.51 milliseconds in total), to elucidate potential non-native PPIs between the oncogene homologue (KRAS) protein[164–167] and the von Hippel-Lindau (VHL) E3 ligase[168, 169]. KRAS is the oncogene most frequently mutated in cancer[164], and PROTAC-induced TPD is considered as a promising approach for treating KRAS-induced cancer[165, 167]. We here simulated the formations and conformational changes of the encounter complex in the absence of the linker, but with KRAS bound to two different warheads and VHL bound to one ligand (Figure 6.1 a-d). Using our simulation and dynamic modeling protocol, we revealed six metastable states

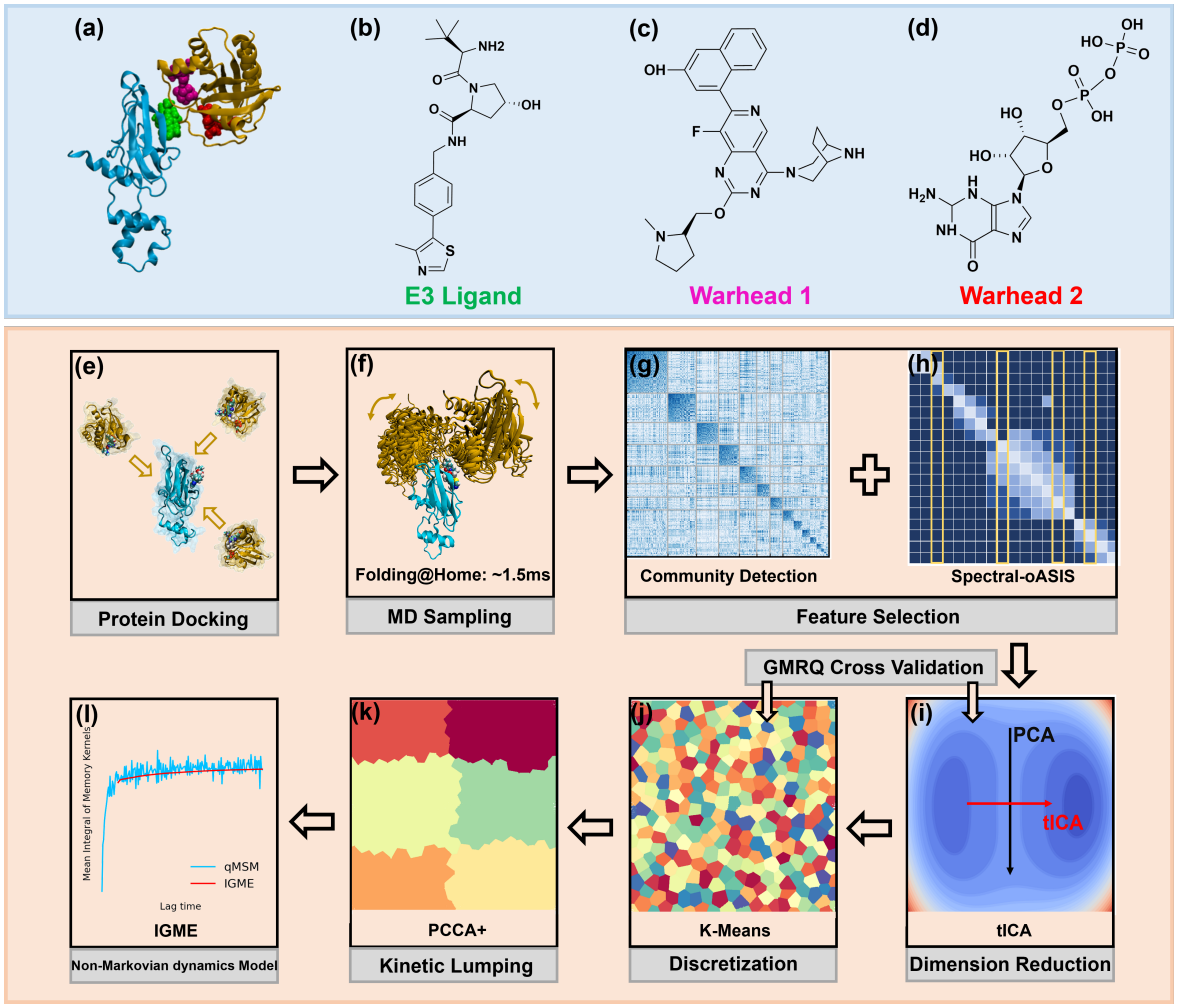

Figure 6.1: The KRAS-VHL encounter complex system (a-d) and the workflow of the construction of the non-Markovian IGME model (e-l). (a). The structure of the encounter complex from rigid protein docking, involving VHL (cyan) and KRAS (orange), along with the E3 ligand and two warheads. (b-d) Chemical structures of E3 ligand (green), warhead 1 (magenta) and warhead 2 (red). (e) Generate initial conformations for the encounter complex through rigid protein docking. (f) Perform extensive MD simulations using Folding@Home to explore the PPI interfaces of the encounter complex. (g-h) Utilize MoSAIC community detection and spectral-oASIS algorithms to extract essential pairwise distance features for representations of the PPI interfaces. (i) Identify the collective variables by tICA. (j) Cluster the projected MD conformations to microstates by K-Means algorithm. The hyperparameters for (i) and (j) are tuned through cross-validation based on the GMRQ score. (k) Lump the microstates to metastable macrostates by PCCA+ algorithm. (l) Model the transition dynamics between macrostates with IGME method. This figure is reproduced from Qiu. *et al.*[3]

characterized by distinct conformations of PPI interfaces and provided the corresponding thermodynamic and kinetic properties for each state. Based on the IGME model, we further evaluated additional structural properties of conformations within each state, such as the spatial proximity of the warhead and E3 ligand and the solvent-exposed sites of both. Consequently, we identified three metastable states that exhibit promising PPI interfaces for future linker design. Conformations from one of our predicted metastable states agree well with a recent ternary crystal structure[166] (with an average interface-RMSD of $5.42 \pm 3.67\mathring{A}$) involving a degrader of promising degradation efficiency.

6.2 Elucidating the Dynamics of KRAS-VHL Encounter Complex Formation: IGME Outperforms MSM

We construct our IGME model from MD trajectories totaling ~ 1.51 milliseconds for studying the conformational changes of the KRAS-VHL encounter complex (see Figure 6.1 panels e-l for our protocol). Specifically, to systematically explore the full ensemble of PPI interface conformations, we employ rigid protein docking to globally search the preferable PPIs from various approaching orientations, and then initiate unbiased MD simulations from these docking poses (see Figure 6.1 e-f and *Methods* for more details). To build the 100-microstate MSM, we first characterize the conformations of the encounter complex using all 25,330 internal pairwise distances between KRAS and VHL residues, and then utilize the Molecular Systems Automated Identification of Cooperativity (MoSAIC) algorithm[74] and Spectral-oASIS algorithm [75] to identify 1,500 important distances as features for subsequent analysis (see Figure 6.1 g-h). The implementation of these two algorithms makes sure that the selected distance features adequately represent various important collective motions around the PPI interfaces, while also effectively capturing the slowest dynamics (see *Methods* for details). Subsequently, we apply the time-lagged independent component analysis (tICA)[47] with kinetic mapping[76] to project the encounter complex conformations onto five collective variables (CVs) (see Figure 6.1 i) and then cluster them into 100 microstates via K-Means algorithm (see Figure 6.1 j). The tICA-related hyperparameters and the number of microstates are optimized using cross-validation with the Generalized Matrix Rayleigh

Quotient (GMRQ) score[90]. More details about the construction and validation of microstate MSM could be found in *Methods*.

To identify the inherent metastable PPIs of the KRAS-VHL encounter complex, we lump 100 microstates into six metastable macrostates using PCCA+[93, 94] and build a 6-macrostate IGME model (see Figure 6.1 k-l). Unlike MSMs which model the dynamics as Markovian processes, IGME utilizes the GME to evolve dynamics, considering the non-Markovian dynamics through time integration of memory kernel functions. Considering that the relaxation time of memory kernel functions is much shorter than the Markovian lag time for MSMs, IGME is able to model dynamics between a handful of metastable states with shorter segments of MD simulations compared to MSMs. As shown in Figure 6.2 a, the integrations of memory kernels reach plateaus at around 50 ns, therefore accurate IGME models could be constructed with the memory kernel relaxation time $\tau_k > 50$ ns. An example of such an IGME model, built from MD simulation segments, each with the length of 150 ns ($\tau_k = 70$ns and an additional segment of $L_{fit} = 80$ ns for fitting, see *Methods* for details), is shown in Figure 6.3(c). In sharp contrast, one MSM constructed with a much longer lag time of $\tau = 250$ ns still predicts significantly faster state-relaxation dynamics compared to the original MD simulations (Figure 6.3 c). Additionally, the time-averaged root mean squared error (RMSE) of the MSMs' predicted dynamics is over an order of magnitude larger than that of the IGME models at different lag times (see Figure 6.2 b and the *Methods* for the details of the RMSE computations). While IGME models consistently capture the slowest timescale and the mean first passage time (MFPT) across a wide range of lag times, MSMs always underestimate these values (see Figure 6.2 c-d). We believe that to achieve comparable performance with IGME models, MSMs would require a considerably longer lag time, which is even beyond the length of our MD simulations. As shown in Figure 6.3 a, IGME models built with a wide range of hyperparameters: τ_k and L_{fit} robustly exhibit small RMSEs, i.e., below 0.1%. For the other results reported in this section, we choose the optimal IGME model as the one with the smallest RMSE value (constructed with $\tau_k = 70$ns and $L_{fit} = 80$ ns) to report the thermodynamic and kinetic properties of the PPI interfaces (see Figure 6.3 a-b).

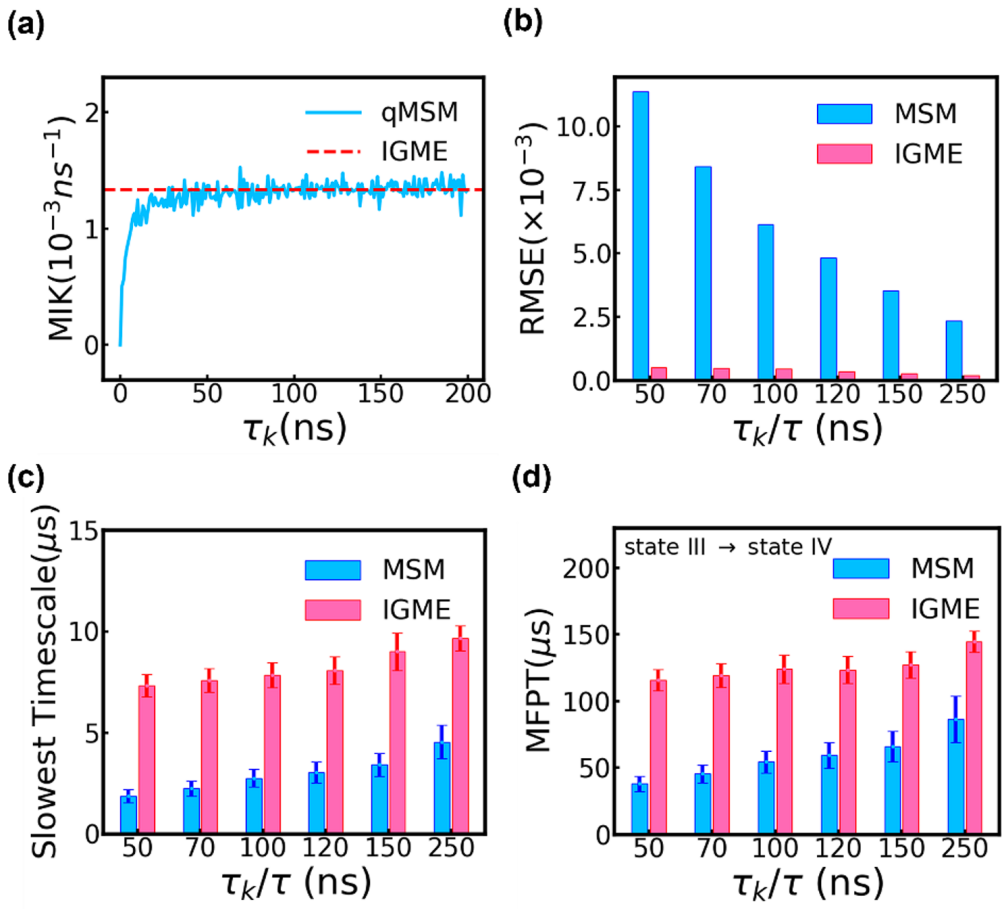

Figure 6.2: Non-Markovian IGME models outperform MSMs in elucidating the dynamics of the KRAS-VHL encounter complex formation. (a) Mean Integral of memory kernels (MIK) with different τ_k for six-states model calculated from qMSM and IGME. (b) Root mean squared error (RMSE) of predicted transition probability matrices with respective to MD simulations, (c) Slowest implied timescale and (d) mean first passage time (MFPT) from State III to State IV, calculated from IGME models and MSMs constructed with various lag times. The error bars represent standard deviations estimated from fifty bootstraps of the data with replacement. This figure is reproduced from Qiu. *et al.*[3]

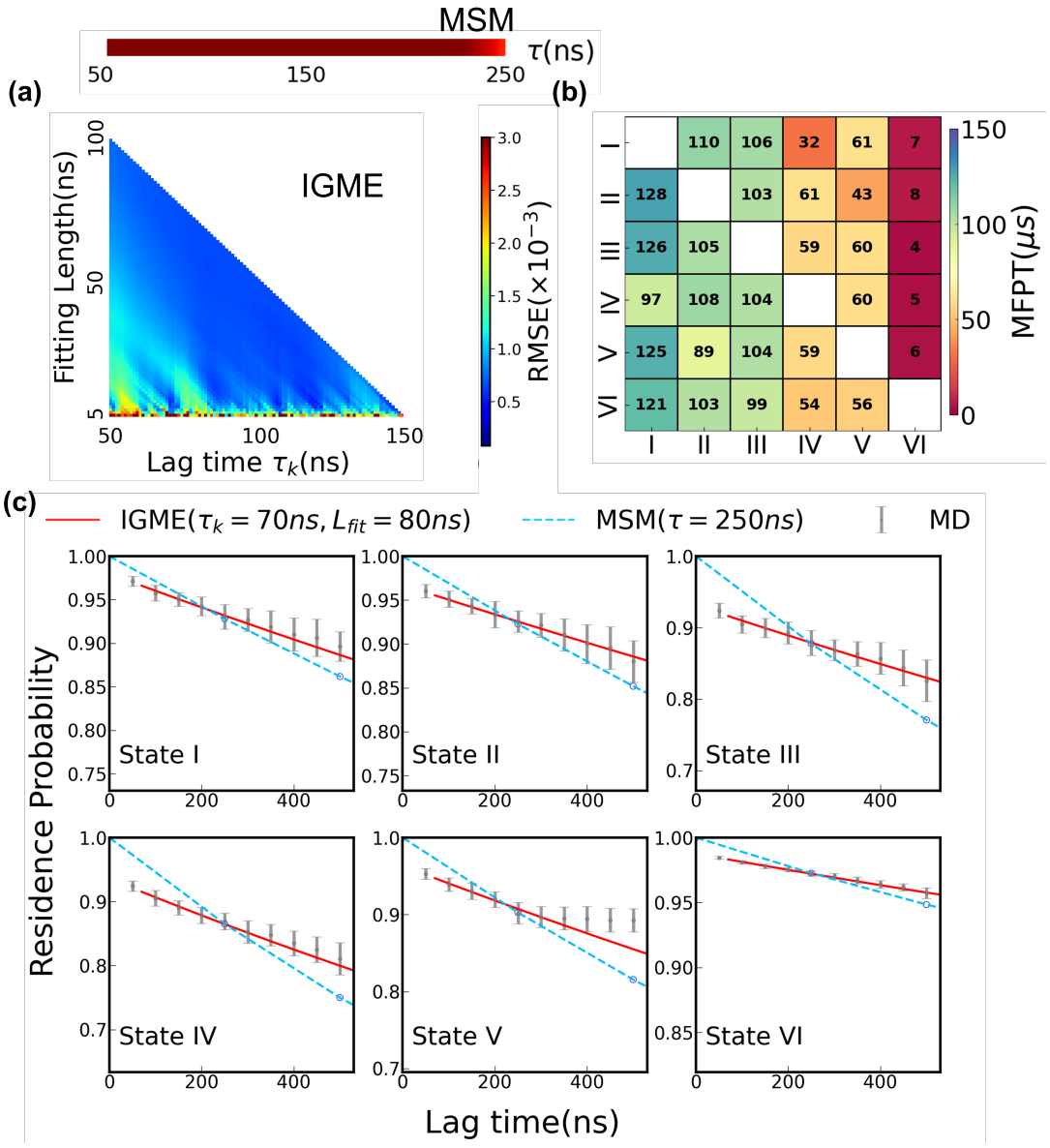

Figure 6.3: Construction and validation of the non-Markovian dynamics model using IGME method. (a) The Root Mean Squared Error (RMSE) of IGME and MSM constructed with varying lag time ranges. (b) Mean First Passage Time (MFPT) between macrostates predicted by the IGME model with the smallest RMSE. (c) The Chapman–Kolmogorov test of the MSM (blue, τ=250ns) and the IGME (red, τ_k=70ns, L_{fit}=80ns) compared against MD simulations (grey). This figure is reproduced from Qiu. *et al.*[3]

6.3 Dynamic Heterogeneity of the Encounter Complex Associated with Diverse Metastable PPI Formation

Based on the optimal IGME model, we observe the metastable PPI interfaces of encounter complex consist of diverse non-native interaction patterns and exhibit significant dynamical heterogeneities. As illustrated in Figure 6.4 a, the free energy landscape of PPI interfaces, projected onto the top two CVs constructed by tICA, reveal various free energy basins. Each basin is associated with distinct metastable macrostate, indicating the intrinsic flexibility and diversity for the formation of the PPI interfaces between KRAS and VHL. Our optimal IGME model also elucidates that State VI is highly populated (72.71%), while the equilibrium populations of the other five states are all below 10% (Figure 6.4 b). Strikingly, we observe significant different PPI interfaces formed and stabilized by diverse chemical interactions between different domains of KRAS and VHL in these six metastable states (see representative structures in Figure 6.4 e). The transition rates between these metastable states are also accurately predicted by our IGME model (Figure 6.3 b). To characterize different PPIs, we first illustrate the PPI patterns using the residue contact maps and their standard deviations between KRAS and VHL. We find that various PPIs display substantially different residue contact maps. Additional analysis of the contact frequency for each residue across PPI interfaces also indicates the heterogeneity of these PPIs. To further examine if there exists preference of specific non-bonded chemical interactions to stabilize these PPIs, we quantify the preferences of amino acid type and interactions for PPIs formed in different macrostates. As a result, we observe that salt bridges and dipolar interactions are present in all PPIs, through the interactions between charged residues (e.g., Glu and Arg) and polar residues (e.g., Gln and His). Interestingly, PPIs in States I and IV also exhibit additional hydrophobic interactions (e.g., via Leu and Val). These observations indicate that KRAS and VHL can form different non-native PPIs via diverse non-bonded interactions. We anticipate that these metastable non-native PPIs could open new opportunities for future PROTAC linker design.

Both previous experimental and computational studies have demonstrated that it is inadequate to uniquely rely on the crystal structure of induced ternary complex to assess the PROTAC performance [150–158]. Instead, the dynamic behaviors of the ternary complex may conduct a more

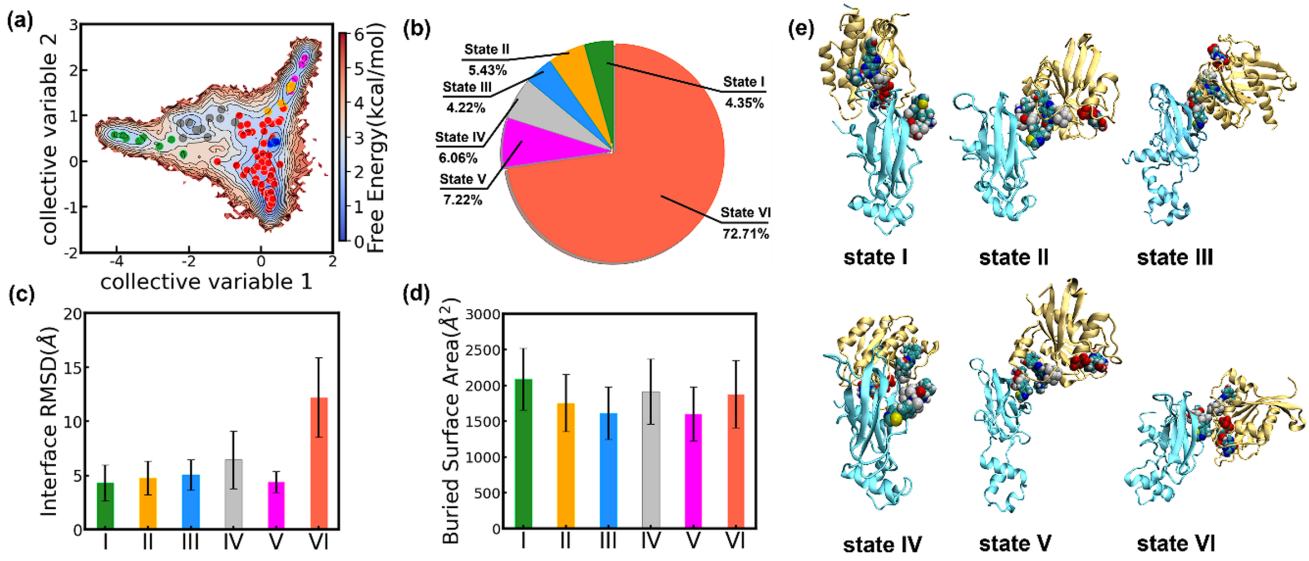

Figure 6.4: Interpretation of non-Markovian dynamics model. (a) The free energy landscape and distribution of states visualized on the top two tICA components. The free energy is estimated from the ultralong trajectory generated by running kinetic Monte Carlo with the microstate-MSM. Each point represents the center of a microstate, and its color corresponds to the macrostate label. (b) Stationary populations for macrostates predicted from the optimal IGME model. (c) The heterogeneity of each macrostate is visualized by calculating the interface-RMSD relative to the state center for all conformations within the state. (d) The buried area of PPI surfaces within each macrostate. (e) The representative conformations for each macrostate (selected from the microstates with the highest population). This figure is reproduced from Qiu. *et al.*[3]

influential role on degradation efficiency. As our simulations of the encounter complex do not include the degrader linker, the encounter complex should generally exhibit much greater heterogeneity among different protein domains. We next examine the structural heterogeneities within each metastable macrostate. The visualizations of multiple encounter complex conformations for each macrostate illustrate the high consistency of PPI interfaces in State I-V and significant flexibilities of interfaces in state VI. By further using the MD conformation located at the geometric center of each macrostate as the reference structure, we compute the interface-RMSD among all MD conformations within each of the six macrostates (Figure 6.4 c). The interface-RMSD is computed using the formula $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{x}_i^{ref})^2}$, where $\boldsymbol{x}_i$ denotes the Cartesian coordinates of atoms in the interface residues (i.e., residues with an average minimal distance to the other protein less than $10\mathring{A}$) after the optimal alignment, $\boldsymbol{x}_i^{ref}$ represents the reference coordinates and N is the number of conformations. Except for the highest populated State VI, all other 5 macrostates exhibit moderate interface-RMSD values which are consistent to those observed in the prior dynamical simulations of other PROTAC-induced ternary complexes[5]. This finding suggests that even in the absence of the degrader linker, KRAS and VHL can display dynamic cooperativity during the formation of the encounter complex, resulting in various well-defined PPIs suitable as a baseline for linker design. Furthermore, we evaluate the stability of the PPIs for each macrostates. Previous studies have elucidated a correlation between buried surface areas (BSA) and experimentally measured binding affinity of the PPIs. We quantify the BSAs of interfaces from different macrostates by subtracting the solvent-accessible surface area (SASA) of the two single proteins from the encounter complex. As shown in Figure 6.4 d, our analysis demonstrates that there are no noticeable differences in the BSA values of the six macrostates in our IGME model. This result is consistent with the equilibrium populations predicted by IGME, where States I-V exhibit comparable populations. Conversely, although State VI has significantly larger populations, it is composed of many short-lived microstates that rapidly inter-convert with each other. Given the substantial structural heterogeneity within State VI, we conclude State VI as a high-entropy state, we conclude State VI as a high-entropy state, where the diverse PPI bonding modes quickly exchange and exhibit low kinetic stability. Consequently, PPI interfaces from States I-V may serve as better

candidates for further PROTAC design. In the current study, we did not observe any metastable state with an incredibly low equilibrium population. However, we note that population predicted from the IGME model could be a very useful criterion in future studies. Specifically, states that exhibit large time separations from other states but have very low populations should be considered inappropriate for the linker design.

Compared to the simple rigid protein docking, our simulation and dynamic modeling protocol demonstrates great power to refine PPI interface patterns and identify the most metastable interfaces. Although the interfaces obtained from docking exhibit great diversity, we found that many of them do not fall into the free energy basins when projected onto the top two CVs. Further characterization of structural differences between docking interfaces and those classified from metastable states revealed generally non-negligible variations in contact map patterns. Additionally, assigning docking interfaces to metastable states showed that while these interfaces span all six states, the majority are predominantly associated with state VI, which is unsuitable for subsequent linker design. This underscores the necessity of dynamic modeling for proper interface classification.

6.4 Shortlisting Predicted PPIs Meeting Linker Constraints

The rational design of PROTAC linkers has been extremely challenged due to the difficulties associated with predicting the pre-existing PPIs between the POI and the E3 ligase. As our simulation and modeling protocol has identified and characterized the stabilities and kinetics of various metastable PPI interfaces, we subsequently consider the geometries of interfaces and ligands within each macrostate to assess their potential for linker design. Throughout the MD simulations, we observe that the warheads and E3 ligand tightly bound to the protein pockets, with only $\sim 3.2\%$ of trajectories showing them diffusing out from the binding site. We further remove out these conformations from the post-analysis. Since the encounter complex exhibits varying extents of conformational changes during the formation of different PPIs, warheads and E3 ligand expose different atoms and take different relative orientations accordingly. To identify the exposed functional groups in the warheads and E3 ligand that could potentially be connected via a linker, we calculate the SASA for each of their atoms (Figure 6.5 a-b). We identify exposed heavy atoms,

defined as those with the top 50% SASA among all atoms, as having the linkage potential. Furthermore, we detect the average pairwise distances between the exposed heavy atom pairs of the E3 ligand and the warheads to further evaluate the feasibility of linker design, as shown in Figure 6.5 c.

Figure 6.5 a-b illustrate that the E3 ligand consistently exhibits larger SASA compared to the warheads, primarily due to the shallow pocket of VHL [170, 171]. The E3 ligand conformations from States I and IV are significantly more exposed than those from other states. This suggests that conformations from these two states may have multiple potential sites to be connected. However, with further examining the interactive profiles between KRAS and VHL in these two states (Figure 6.4 e), we notice that developing any linker based on the PPI interfaces from these two states is impractical, as the warheads and E3 ligand are too far away from each other (between $25\mathring{A}$ and $30\mathring{A}$, see Figure 6.5 c). Prior studies have demonstrated that linker length is one of the most crucial factors determining the effectiveness of PROTACs, and excessively long linkers often result in a reduction in potency. Generally, PROTACs with linkers containing more than 20 atoms have comparatively low potency [152, 159, 172, 173]. Therefore, designing linkers based on the PPI interfaces from States I and IV poses significant challenges. Moreover, after evaluating the conformations from all states, we observe that warhead 2 (see its chemical structure in Fig. 6.1 d) consistently stays distant from the E3 ligase ligand, indicating difficulties in developing a degrader using it.

Consequently, by excluding State VI due to its kinetically unstable PPI, and eliminating State I, State IV, and warhead 2 due to inappropriate distances between warheads and E3 ligand, the PPI interfaces from the remaining three states (II, III, and V) are thought to have the potential for further linker development between E3 ligand and warhead 1. We further visualize the PPI conformations and the relative orientations of ligands for these three states. As shown in Figure 6.5 d-o, the conformations within these three states maintain homogeneous interfaces while also exhibiting slight heterogeneities. Furthermore, the warhead and E3 ligand approach each other at appropriate distances for adding the linker. As shown in Fig. 6.5 f, j, & n, we highlight the E3 ligand and warhead atoms with the top 50% largest SASA using dashed boxes in their chemical

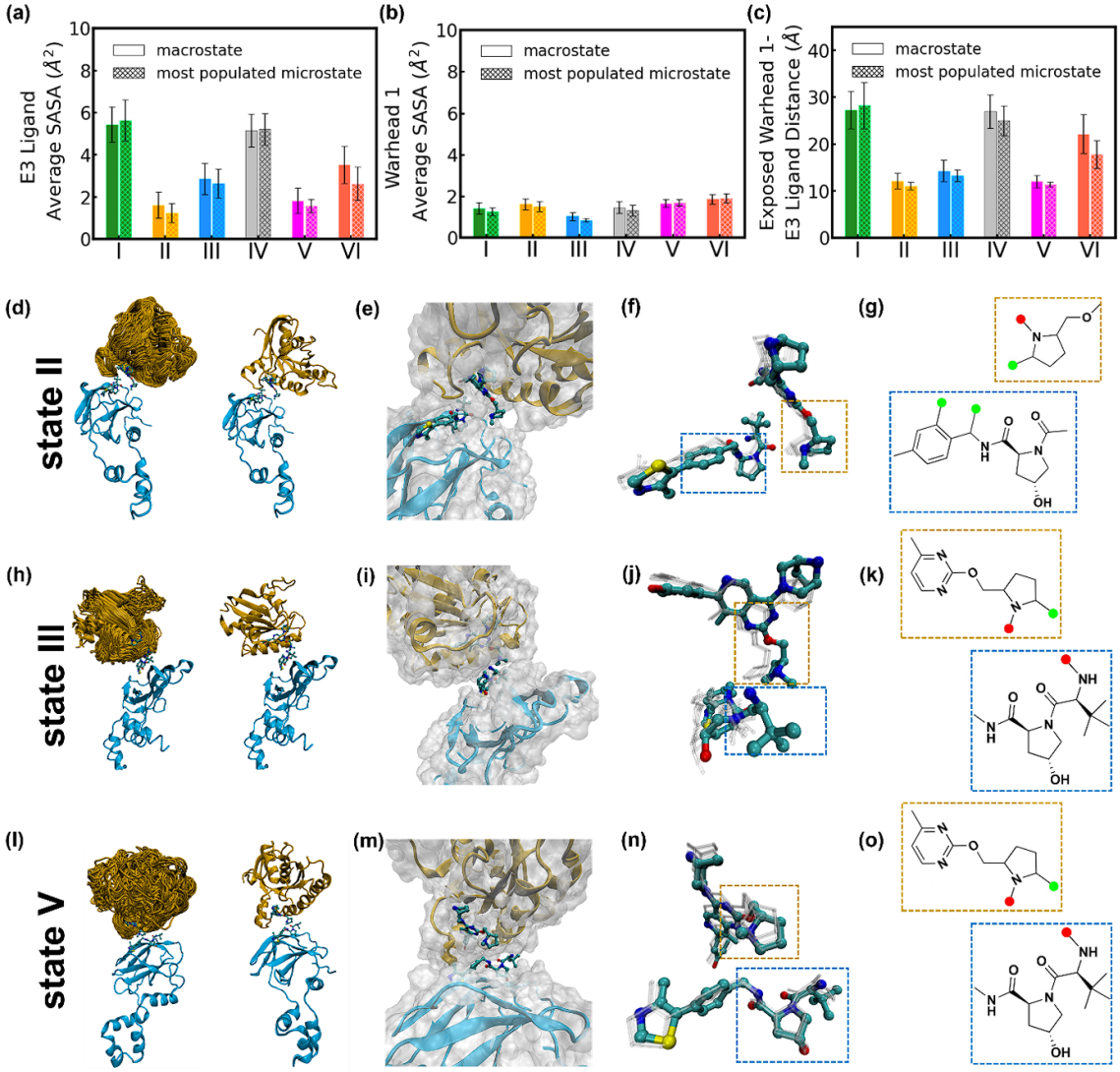

Figure 6.5: PPI interfaces selected for linker design. (a-b). Average Solvent Accessible Surface Area (SASA) depicted for (a) E3 ligand and (b) warhead 1 molecules across all conformations within six macrostates and their respective most populated microstate. (c) The average pairwise distances between the exposed heavy E3 ligand atoms and warhead 1 atoms (top 50% SASA) are calculated across all conformations within six macrostates and their respective most populated microstate. Error bars represent standard deviations. Fifty randomly selected overlapping conformations and one representative single conformation of the PPI interface are visualized for State II (d-e), State III (h-i), and State V (i-m). The relative positions of the E3 ligand-warhead 1 and their partial chemical structures are displayed for State II (f-g), State III (j-k), and State V (n-o). This figure is reproduced from Qiu. *et al.*[3]

structures. In these chemical structures, we have identified potential linkage sites, labeling them with colored dots based on their synthetic ease and frequency of use reported in the literature. In particular, the red dots correspond to atomic sites that are most commonly employed in literature for attaching the linkers, while the green dots indicate sites that are less frequently used for this purpose [174–179]. However, it is worth noting that less frequently used linkage sites for VHL ligands could also lead to highly effective degraders. We believe that these selected conformations may aid in designing linkers that could further stabilize the naturally favorable PPIs.

Through our systematic analysis, we notice there are several key factors to consider when evaluating and selecting appropriate PPI interfaces for linker design. After identifying metastable states through our simulation and modeling protocol, quantifying their structural heterogeneity helps in selecting states with long-lived PPI interfaces and consistent binding modes. States with interface-RMSD values larger than $10\mathring{A}$ should be carefully considered, given that the typical structural heterogeneity of the POI-PROTAC-E3 ligase ternary complex is moderate[5]. Subsequently, the equilibrium populations of states and the BSA of configurations within states can serve as references for PPI binding affinity. States with very low populations (e.g., $< 1\%$) and very small BSAs (e.g., $< 500\mathring{A}$)[147] should be used cautiously for linker design. Then we can employ SASA to identify solvent-exposed heavy atoms on the warhead or ligand as potential linkage sites. Targeting these exposed sites with a linker can largely retain the inherent binding modes of the warhead and ligand. The identification process can be highly system-dependent, as variations in pocket shapes and PPI binding modes significantly influence the conformations of the warhead and ligand. Meanwhile, it is also essential to integrate chemical synthesis knowledge during this step. Additionally, the distance between selected linkage sites offers guidance for determining the appropriate linker length. PPI interfaces with linking sites that are too far apart should be discarded, as typical linkers span 5 to 15 carbon atoms in length[159]. A schematic workflow for evaluating metastable PPI interfaces for linker design is shown in Figure 6.6.

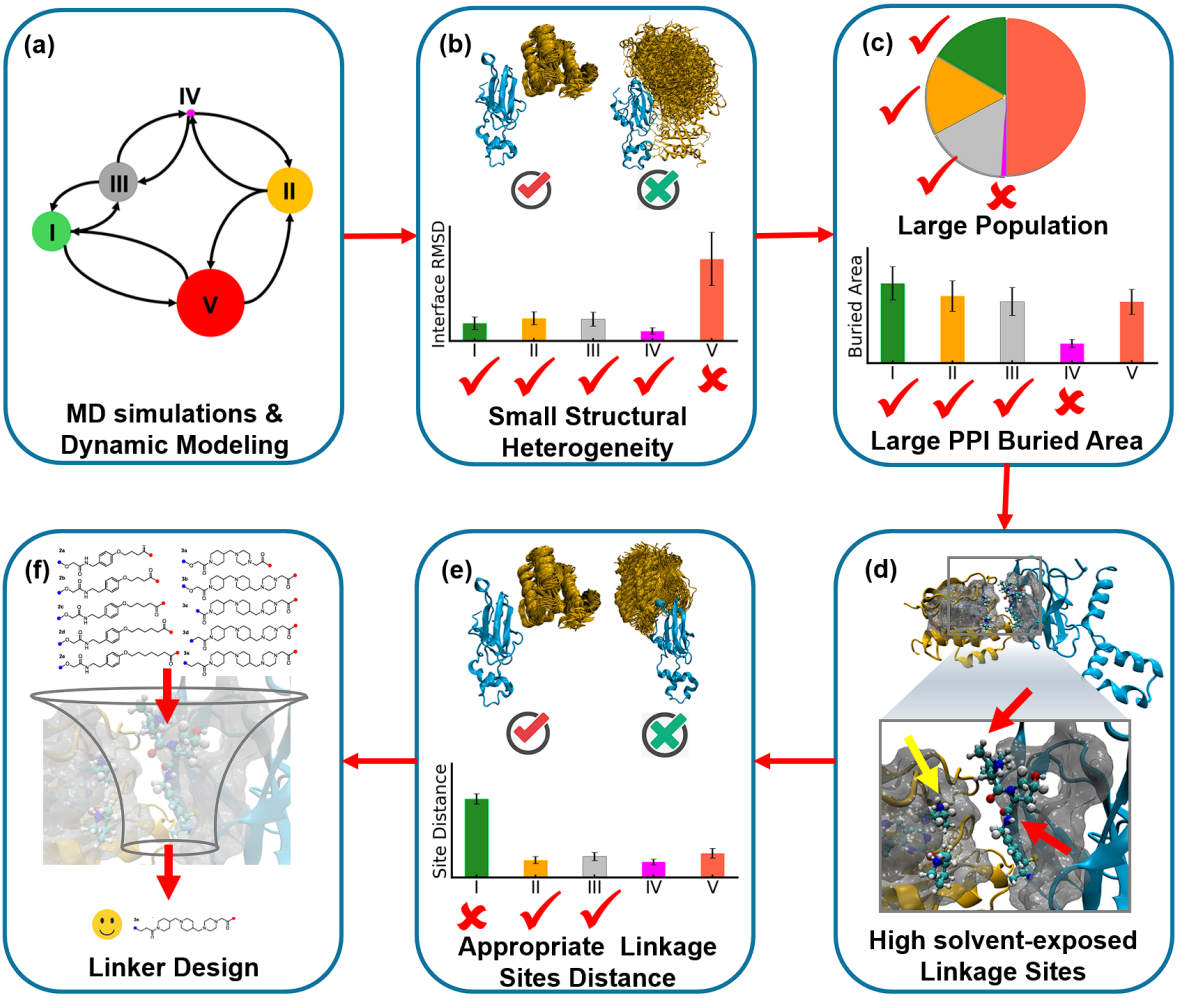

Figure 6.6: The workflow for evaluating metastable PPI interfaces for PROTAC linker design. (a) Perform MD simulations and dynamic modeling of linker-less encounter complex. (b) Quantify structural heterogeneity for metastable states and select states with long-lived consistent PPI binding modes. (c) Use equilibrium populations predicted by the IGME model and buried surface areas to select states with high interface binding affinity. (d) Analyze the high solvent-exposed regions to identify potential linker attachment sites. (e) Compute the distances between attachment sites and filter out PPI interfaces with inappropriate distances. (f) Select linkers that can stabilize the selected metastable PPI interfaces. This figure is reproduced from Qiu. *et al.*[3]

6.5 Predicted PPI Interface Agrees with the Structure Induced by an Experimentally Designed PROTAC

A recent experimental study by Johannes et al. successfully designed and completed the preclinical validation of a single small molecule degrader, targeting KRAS and related mutant cancer proteins with VHL E3 ligase [166]. In this study, the authors elucidated a co-crystal structure of the degrader in complex with KRAS and VHL at a resolution of $2.2\mathring{A}$ (PDB: 8QVU), as visualized in Figure 6.7 a. We observe that the PPI interface in this ternary co-crystal structure is structurally similar to the most populated microstate from our State III (Figure 6.7 a-b). Upon further examination of the residue pairwise distances of the crystal PPI interface, we find a high degree of consistency with the distance map of the ensemble of interfaces within macrostate III (Figure 6.7 c-d), where salt bridges and dipolar interactions stabilize the PPI. The interface-RMSDs between the crystal structure and the ensemble of interfaces from macrostate III and its corresponding most populated microstate are as small as $5.42\pm3.67\mathring{A}$ and $3.76\pm2.37\mathring{A}$, respectively. The interface with the smallest interface-RMSD, visualized in Figure 6.7 a, has only a value of $0.68\mathring{A}$. Additionally, the BSA of crystal structure is $1,556\mathring{A}^2$, which is also consistent with State III ($1,612 \pm 367\mathring{A}^2$) and its most populated microstate ($1,587 \pm 206\mathring{A}^2$). This consistency between the PPI interfaces in State III with the experimental crystal structure provides compelling validation of the predictions from our IGME model. In contrast, among all interfaces generated from protein docking, we found only one that could be assigned to state III and exhibited the smallest interface-RMSD with the crystal structure. However, even the smallest interface-RMSD value is $8.28\mathring{A}$.

Detecting the dynamical interactions and elucidating non-native PPIs between protein pairs without degrader are significant challenges for various experimental methods. The weak binding affinity of the protein encounter complex makes it difficult for the structure biology approaches like X-ray crystallography[180–182]. While NMR spectroscopy or hydrogen-deuterium exchange mass spectrometry can detect the interactions, their time-resolution is rather very limited [5, 182]. Recently, it has also been shown that data-driven machine-learning approaches such as AlphaFold and AlphaFold-Multimer face challenges in accurately predicting non-native PPIs, particularly

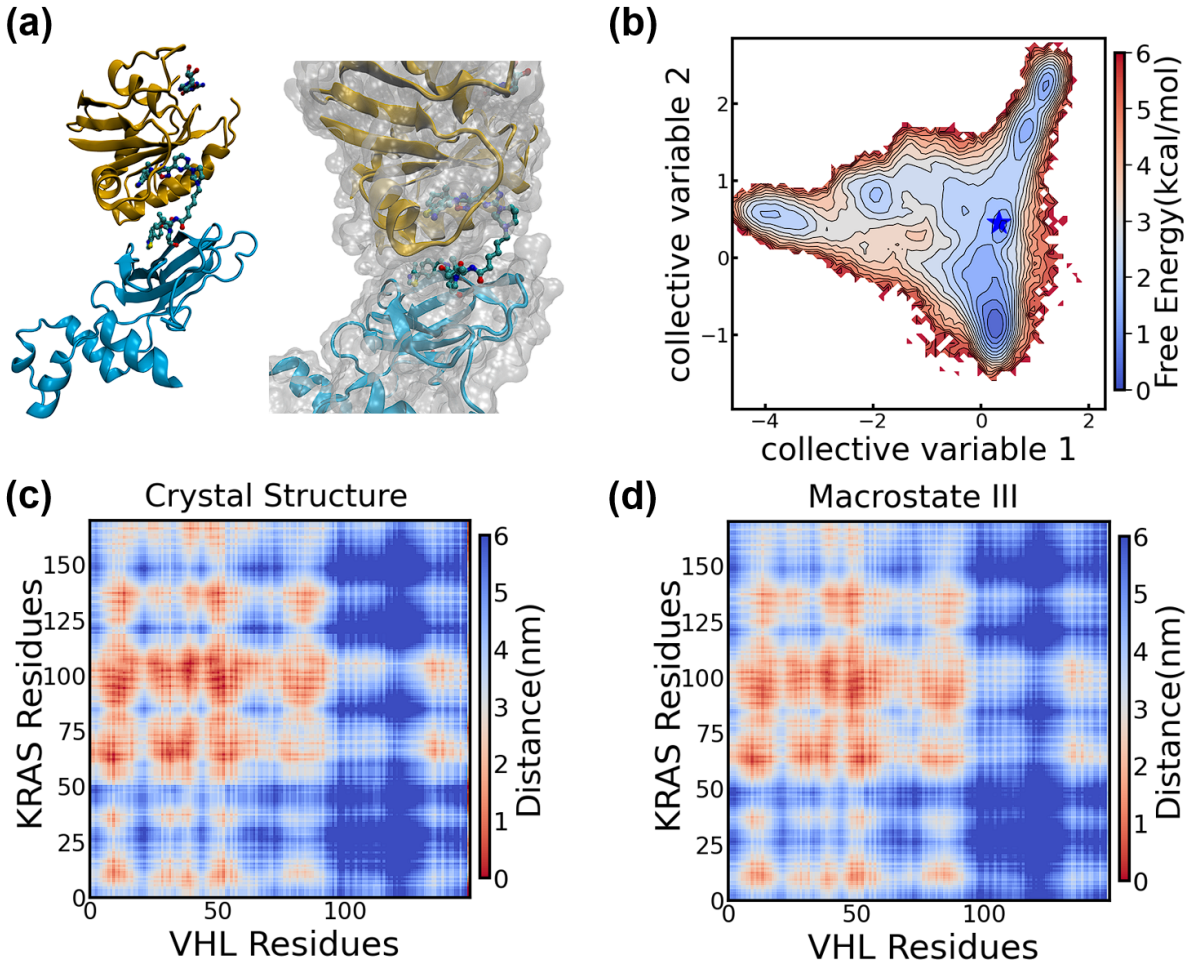

Figure 6.7: Comparison between computationally predicted PPI interfaces and the interface induced by the experimentally designed PROTAC. (a) Structural alignment between the crystal structure (magenta, PDB ID: 8QVU) and one PPI interface from most populated microstate in State III (orange). The interface with the smallest interface-RMSD (0.68$\mathring{A}$) is selected for visualization, and the alignment is based on the VHL protein. (b) Projection (blue star) of the crystal PPI interface of the ternary complex onto the top two CVs. (c) Pairwise distances between KRAS residues and VHL residues in the crystal structure of the ternary complex (PDB: 8QVU). (d) Averaged pairwise distances between KRAS residues and VHL residues across all conformations within macrostate III. This figure is reproduced from Qiu. *et al.*[3]

when the interface area is limited [183, 184]. Our results demonstrate that non-native PPIs in encounter complex could be systematically and accurately predicted in atomistic detail by integrating parallel short unbiased MD simulations with non-Markovian dynamics modeling (i.e., IGME). Our pipeline provides advantages of revealing both non-native PPIs and their dynamic heterogeneities simultaneously, thereby offering ensembles of metastable PPI interfaces for later high-throughput linker design. Moreover, in the future, the transition timescales predicted by our IGME model could be valuable for designing flexible linkers to stabilize two distinct metastable PPI interfaces that interconvert rapidly. Therefore, we anticipate that IGME holds significant potential for generalization in future PROTAC discovery.

6.6 Conclusion

PROTAC-induced TPD is regarded as one of the most promising approaches for small molecule-based drug discovery. However, the rational design of PROTACs remains challenging due to factors such as the large size of the multi-protein system and the complex, dynamic protein interactions. In this study, we present a physics-based approach to identify the complete ensemble of intrinsic and dynamic PPI interfaces between KRAS and VHL proteins by studying the linker-less encounter complex. Specifically, we show that our IGME model, a non-Markovian dynamics model, constructed from extensive MD simulations (~ 1.5 milliseconds), is able to elucidate the inherent metastable states of PPI interfaces and accurately predict their associated thermodynamic and kinetic properties. We demonstrate that IGME models significantly outperform MSMs in predicting slow dynamics associated with the encounter complex formation between KRAS and VHL. The six metastable states revealed in our IGME model represent distinct PPI interfaces of the encounter complex. Upon evaluating the stabilities and geometries of the PPI interfaces in each state, we narrowed down to three states (State II, III, and V) with promising PPI interfaces for future PROTAC linker design. The interfaces from the selected metastable states are primarily maintained by electrostatic interactions and display local dynamic heterogeneity, serving as a good basis for linker docking. We validate our theoretical predictions by showing that one of our selected PPI interfaces (State III) is highly consistent with a recent co-crystal structure containing

the PPI induced by an experimentally validated PROTAC for the KRAS-VHL system. We anticipate that our predicted PPI interfaces for the KRAS-VHL system will provide valuable insights for future linker design. We believe that the rigorous foundations of this strategy, grounded in physical simulations and statistical thermodynamics, will lead to broad applicability across diverse systems, facilitating more efficient designs of efficacious PROTACs.

6.7 Methods

6.7.1 All-Atom MD Simulation Setup for KRAS-VHL Encounter Complex

The structures of VHL protein (PDB ID: 1VCB)[185] and KRAS protein (PDB ID: 7RPZ)[186] are respectively derived from crystal structures identified through X-ray diffraction. We first employ the PyRosetta docking package[62] to generate sets of initial KRAS-VHL encounter complex structures. The rigid-body docking is performed while restraining the distance between the linker attachment atoms of warhead 1 and the E3 ligand as $20\mathring{A}$, ensuring the formation of interfaces appropriate for subsequent design. Subsequently, we ultize the K-Means algorithm to categorize the obtained docking structures into fifty clusters based on their root-mean-square deviations. The structure nearest to each cluster center is chosen as the initial structure for the subsequent MD simulations. We then protonate and solvate the initial poses in cubic boxes with explicit TIP3P water[187] and add counter ions to maintain the neutrality of the system. Next, we employ the OpenMM package[188] to conduct all-atom simulations with power of Folding@Home[189], with the in-house parameterized force field for the small molecules (i.e., ligand and warheads) and the AMBER ff14SB force field[190] for the proteins. The final obtained dataset used for post-analysis consists of 2,492 trajectories, totaling ~ 1.51 milliseconds of aggregate simulation time, with an average trajectory length of 605 nanoseconds.

6.7.2 Construction and Validation for Microstate-MSM

Following our proposed pipeline in Figure 6.1, we build the microstate-MSM to study the inherent PPIs between KRAS and VHL protein. The detailed procedures are present below:

(1). Classify the collective motions at PPI interfaces via MoSAIC: we initially embed representations for PPI interfaces by utilizing the internal pairwise distances between KRAS residues (170 residues) and VHL residues (149 residues), resulting in 25,330 pairwise distance features. Subsequently, we apply the MoSAIC algorithm to cluster these features into 27 communities, with approximately 10% of the features filtered out as unimportant noise. Through visualization of the features within each community and the integration of our biological intuitions, we further exclude 11 communities associated with collective conformational changes unrelated to PPI interfaces, resulting in 16 communities encompassing a total of 14,402 features.

(2). Extract the features capturing slow dynamics by Spectral-oASIS: We employ the spectral-oASIS algorithm for the second round of feature selection, through which 1,500 features are automatically identified. These features are supposed to effectively capture the leading slowest dynamic modes.

(3). Reduce dimensionality by tICA: We employ tICA with kinetic mapping to linearly construct five collective variables (CVs) from 1,500 features. The MD conformations are then projected on these CVs and further clustered into 100 microstates using the K-Means algorithm. The optimal hyper-parameters (i.e., number of CVs, tICA relaxation time and the number of microstates) are determined by cross-validations with the generalized matrix Rayleigh quotient (GMRQ) score. Additionally, we validate that the microstates are well-connected, with multiple reversible transitions observed between them.

(4). Validate the microstate MSM: Based on the 100 microstates model, we further produce Implied Time Scale (ITS) analysis and Chapman-Kolmogorov (CK) test and validate the Markovian lag time for microstates-MSM is 200ns.

6.7.3 IGME Modeling of Encounter Complex

To identify metastable states of the PPI interface, facilitate the interpretation, and acquire the associated thermodynamic and kinetic properties, we employ our recently developed IGME method to construct a model comprising only six representative states. We first utilize the PCCA+

algorithm to lump the 100 microstates into 6 macrostates, given the largest time scale gap is between the 5th and 6th transition modes. We crisply assign each microstate to the macrostate with the highest membership value. Subsequently, we evaluate the connectivity among the six macrostates and demonstrate that each macrostate exhibits reversible transitions with at least four other macrostates. Then the IGME is employed to precisely model the transition dynamics between macrostates, encoding the non-Markovian dynamics through the time-integration of memory kernel functions. Specifically, IGME accurately describes the evolution of the transition probabilities matrices (TPMs) with the lag time longer than memory relaxation time τ_k by $\mathbf{T}(t \geq \tau_k) = \mathbf{A}\hat{\mathbf{T}}^t$, where matrices $\mathbf{A}$ and $\hat{\mathbf{T}}$ are estimated from simulations.

To decide τ_k, we adopt two approaches: one employs our previous quasi-MSM technique, which calculates memory kernel matrices at various times using the greedy algorithm with discretized GME, the other approach involves using IGME to approximate the time-integrated memory kernel. The mean integral memory kernel, defined as $\mathcal{MIK} = \frac{1}{N}\sqrt{\sum_{i,j=1}^{N}(\int_0^t \mathbf{K}_{ij}(\tau)d\tau)^2}$, computed from two approaches are well consistent and the memory relaxation time τ_k is decided as 50ns (Figure 6.2 a).

To construct the optimal non-Markovian dynamics model, we employ multiple sets of TPMs with different lag time range $\{\mathbf{T}^{MD}(\tau_k + n\Delta t)\}_{n=0}^{L_{fit}}$ to estimate the matrices $\mathbf{A}$ and $\hat{\mathbf{T}}$ using least-squares fitting with a Lagrangian approach[102]. The optimal range, parameterized by τ_k and L_{fit} are decided by time-averaged root mean squared error (RMSE) with respect to MD simulations. After a systematic scan, we ultimately identify the optimal fitting range: $\tau_k = 70$ns and $L_{fit} = 80$ns (see Figure 6.3 a).

Chapter 7

Dynamic Modeling Reveals Nucleosome Condensates and Linker DNA Influence Chromatin Folding Pathways and Rates

This chapter is reproduced in part from an unpublished collaborative work by Qiu, Y; Liu, S; Lin, X; Unarta, I; Huang, X; & Zhang, B; with permission. Qiu, Y equally contributed to the development of methodologies, data analysis, performing simulations and manuscript writing with Liu, S. Qiu, Y rewrote and summarized the content in this chapter.

7.1 Introduction

In this chapter, to investigate the dynamical behavior of chromatin systems composed of multi-body nucleosomes and to understand the relationships between structures observed with different experimental techniques, we integrate coarse-grained (CG) models, the previously introduced non-Markovian dynamic modeling approach, and various machine learning techniques, including the LPC algorithm, to systematically study the origins of chromatin flexibility. In particular, our implementation of various novel techniques significantly helps bridge the timescale gap between simulations and rate events and enhance our understanding of the biological mechanisms. And our exploration of how multi-body nucleosome condensates affect chromatin folding dynamics provides new biophysical insights into how phase separation influences genetic functions.

Chromatin organization, essential for packaging DNA in eukaryotic genomes, plays a crucial role in numerous genetic functions. [191–195]. Where available, atomistic structures of chromatin have been invaluable for constructing mechanistic models of gene regulation and other processes [196, 197]. However, the detailed organization at the atomic level has been a topic of contentious debate. The existence of 30-nm fibers, characterized by nucleosomes arranged in a zig-zag pattern and stacking to form a twisted two-column fibril, has been documented for a long

time [196–200]. However, several *in vivo* experimental techniques, including cryo-electron microscopy [196], Micro-C [201], and ChromEMT [202], have indicated a lack of ordered fibril-like structures. Instead, these techniques suggest the presence of 10-nm disordered arrays with dominant local oligomer motifs, such as trimers, α-tetrahedron, and β-rhombus tetramers [201–206]. These conflicting observations have left the principles of chromatin folding, particularly at the short length scale of tens of nucleosomes, unresolved.

A more dynamic perspective on chromatin organization has proven to be insightful. Previous work [207] characterized the folding landscape of a tetra-nucleosome using a residue-level CG model, suggesting that the unfolding of fibril structures leads to the irregular conformations observed in the nucleus, thereby bridging *in vitro* and *in vivo* configurations. Another work [208] further demonstrated that unfolding can generate clutches, consistent with observations from super-resolution imaging of the cell nucleus [209]. A recent cryo-electron tomography (cryo-ET) study further supports the presence of similar folding principles for chromatin both *in vitro* and *in situ* [202]. However, the computational studies employed biased simulations to accelerate chromatin folding and unfolding. While these simulations allow for thermodynamic predictions, extracting accurate kinetic information from them is challenging. Therefore, it is crucial to directly determine whether the irregular *in vivo* chromatin configurations represent intermediate folding stages leading to the zigzag fibril structure.

MSMs provide an efficient method for extracting interpretable kinetic information, enabling the prediction of transition rates and the identification of reaction pathways from MD data [27–29, 40, 41, 48, 210, 211]. Importantly, the construction of such models and prediction of long-timescale dynamics require only distributed short MD trajectories as inputs, which can be efficiently generated in parallel using supercomputing facilities. MSMs have demonstrated success in investigating the dynamics of conformational changes in various chemical and biological systems, including protein folding [52, 106], protein-ligand recognition [6, 105], and the self-assembly of soft materials [13, 16]. Moreover, recent advancements in non-Markovian dynamics models, which employ the generalized master equation to account for memory effects, have led to significant improvements in modeling accuracy and a substantial reduction in the required input MD trajectory length [30–

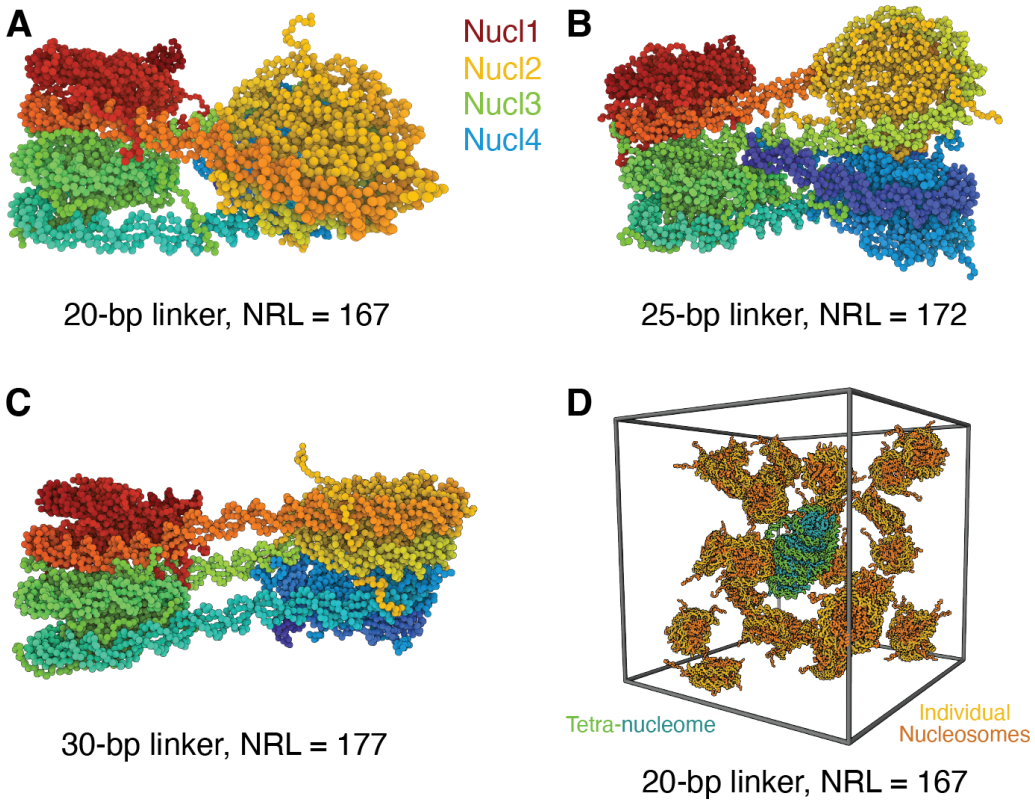

Figure 7.1: Representative configurations for the four tetra-nucleosome systems studied. The three isolated systems feature tetra-nucleosomes of 20-bp (A), 25-bp (B), and 30-bp (C) DNA linker. The corresponding NRL is 167, 172, and 177 bp, respectively. In the fourth system, the tetra-nucleosome with NRL=167 is embedded into a nucleosome condensate.

32, 212]. By applying these dynamics modeling approaches to the study of chromatin folding, we can gain new insights without the potential biases introduced by advanced sampling approaches.

Here, we we perform extensive residue-level CG MD simulations to study chromatin folding. In particular, we investigate the folding of four different systems: three isolated tetra-nucleosomes with nucleosome repeat lengths (NRL) of 167, 172, and 177 base pairs (bp), and a tetra-nucleosome with an NRL of 167 bp embedded within a nucleosome condensate. We construct MSMs and easily interpretable non-Markovian dynamics models for each system to comprehensively and respectively characterize the folding pathways and kinetics. In each system, transition path analysis consistently reveals numerous parallel folding pathways with comparable fluxes. This finding contrasts with typical one-body protein folding and more closely resembles multi-body self-assembly processes. Notably, intermediate states along these pathways resemble *in vivo* chromatin organization. Furthermore, while the tetra-nucleosome with NRL = 167 bp folds stably into the zigzag fibril structure, both the nucleosome condensate environment and a 5 bp longer DNA linker destabilize the fibril conformation, promoting the formation of folding intermediates through different molecular mechanisms. Extending the linker DNA length by 10 or 5 bp has markedly different effects on the chromatin folding landscape: while the 10 bp extension favors the zigzag fibril configuration, the 5 bp extension does not form a specific structure, instead favoring a dynamic ensemble of conformations. These results reinforce the idea that the absence of regular fibril configurations inside cell nuclei could result from chromatin unfolding driven by various factors such as crowding environments and linker length variation.

7.2 Construction of Dynamics Models for Chromatin Folding from Extensive MD Simulations

Our study focused on tetra-nucleosomes, the fundamental units of the zigzag fibril configuration [196, 197], to explore chromatin organization in various biological contexts. Specifically, we configured a series of tetra-nucleosome structures with 20, 25, and 30 bp linkers, corresponding to NRL of 167, 172, and 177 bp (Figure 7.1A-C), respectively. The nucleosomal DNA length is defined as 147 bp, and we refer to the DNA connecting two adjacent nucleosomes as the linker

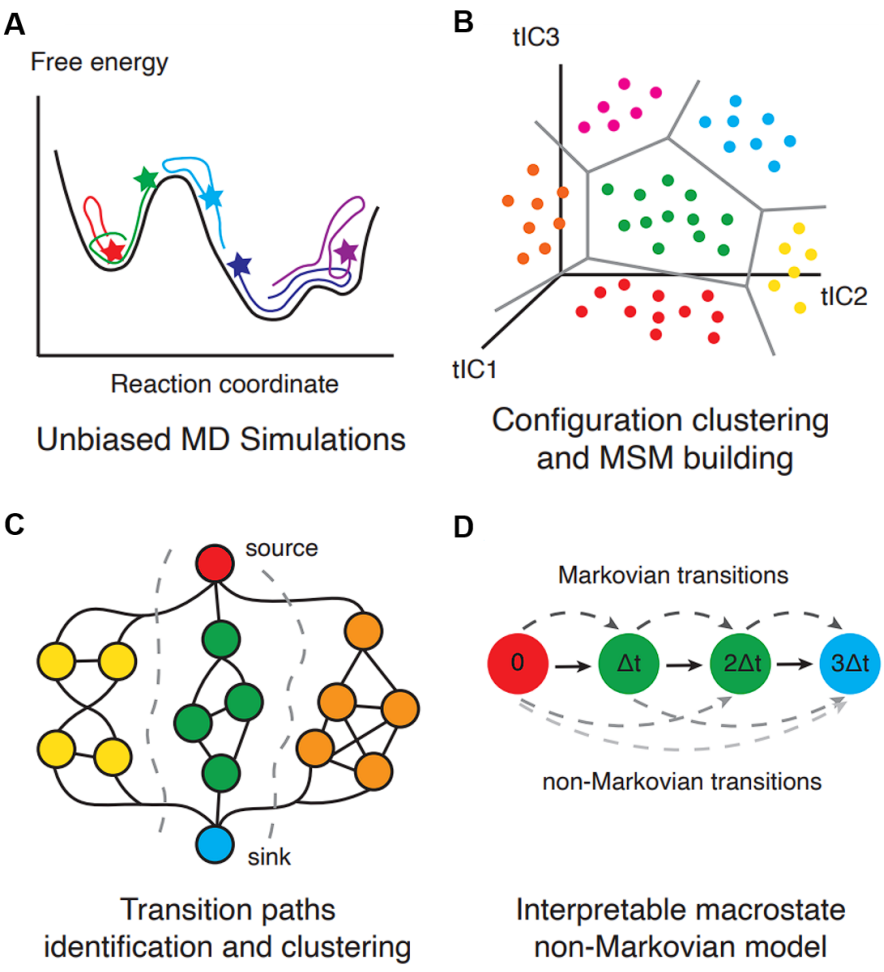

Figure 7.2: Overview of the computational pipeline to elucidate chromatin folding kinetics and pathways. (A) The workflow begins with extensive unbiased MD simulations, initiated from a variety of configurations. (B) Configurations collected from these simulations are then projected onto collective variables constructed by tICA, followed by clustering into microstates to build up MSMs. (C) Subsequently, chromatin folding pathways are identified and reaction channels are lumped using transition path theory and the Latent-space path clustering algorithm. (D) Finally, the microstates are grouped into a few interpretable macrostates, and the transition dynamics between these macrostates are modeled using the generalized master equation that incorporates time-dependent memory kernels (D).

DNA. To simulate a biologically relevant scenario, we further embedded the tetra-nucleosome with a 20 bp linker in a solution of single nucleosomes to account for nuclear crowding effects. Our setup results in an overall nucleosomal concentration of 0.3 mM (Figure 7.1D), consistent with concentrations estimated from *in vitro* nucleosome array condensates [213, 214].

Given the complexity of nucleosome structures and the slow timescale of tetra-nucleosome folding and unfolding, we employed a residue-level CG model to perform MD simulations, making extensive sampling computationally feasible for investigating chromatin conformational dynamics [215–219]. In particular, we utilized a one-bead-per-amino-acid model [220] and a three-bead-per-nucleotide model [221] to maintain sufficient resolution for an accurate description of specific protein-protein and protein-DNA interaction with physical chemistry potentials. These models have been effectively employed to study various protein-DNA systems, accurately reproducing experimental results and offering mechanistic insights [207, 208, 222–227]. Further details on the CG force field are provided in the *Methods* section.

For each of these four systems, we conducted multiple independent, unbiased MD simulations to extensively explore various regions of the chromatin configurational space. As illustrated in Figure 7.2A, these simulations were initiated from distinct conformations, capturing various degrees of chromatin compaction and folding, as identified by the enhanced sampling techniques reported in a previous study [207]. For isolated tetra-nucleosomes with varying linker lengths, we executed 4,643 simulations for each system. In the case of the condensate system, due to its much longer equilibration time, we reduced the number of trajectories to 530 and extended each simulation length. All simulations are sufficiently long to allow for the relaxation of chromatin conformations and the attainment of local equilibrium.

Subsequently, we constructed dynamics models for each system to study chromatin folding over timescales much longer than individual MD simulation trajectories. We projected tetra-nucleosome configurations explored along unbiased trajectories onto collective variables using time-lagged independent component analysis (tICA) [76, 111, 228, 229], and clustered them into microstates via the K-Means algorithm (Figure 7.2B). Microstate-MSMs were then constructed and validated to model chromatin folding dynamics through the implied timescale analysis and

Chapman-Kolmogorov test. Further empolying the transition path theory (TPT) [40, 41, 230], we identified the complete ensemble of kinetic pathways for the each tetra-nucleosome folding. (Figure 7.2C).

To aid in interpreting the underlying folding dynamics, we constructed non-Markovian dynamics models with six metastable macrostates, derived by lumping microstates using the PCCA+ algorithm, and modeled the transition dynamics using the Integrated Generalized Master Equation (IGME) method. [31] (Figure 7.2D). Different from microstate-MSMs, which typically involve hundreds of states to describe slow dynamics, making it challenging to interpret their biological significance, IGME models could accurately capture the slowest dynamics based on only a few representative states. This advantage is due to IGME's consideration of non-Markovian dynamics via time-integrations of memory kernels. We demonstrated that, compared to MSMs constructed with comparable or longer MD segments, IGME models exhibit significantly higher accuracy in predicting long-term dynamics, as referenced by raw MD simulations. More details regarding system setup, simulations, and constructions of MSMs and IGME models are available in the *Methods* section.

7.3 Multiple Reaction Channels of Tetra-nucleosome Folding

We began our investigation with the isolated NRL = 167 tetra-nucleosome system. Inter-nucleosomal distances, d_{13} and d_{24}, are considered effective and interpretable coordinates for studying tetra-nucleosome folding, as previously demonstrated [207]. Therefore, we used the equilibrium populations obtained from the corresponding microstate-MSM to estimate the two-dimensional free energy landscape along these coordinates. (Figure 7.3A) The resulting landscape, featuring an approximate 10 kcal/mol difference between unfolded and folded regions, indicates the stability of the native tetra-nucleosome structures and is consistent with previous estimations from neural network fitting of mean forces[207].

The free energy landscape clearly indicates the potential existence of multiple folding pathways, as suggested in a previous study. Indeed, transition path analysis based on the microstate-MSM revealed thousands of kinetic pathways with comparable fluxes. This demonstrates that

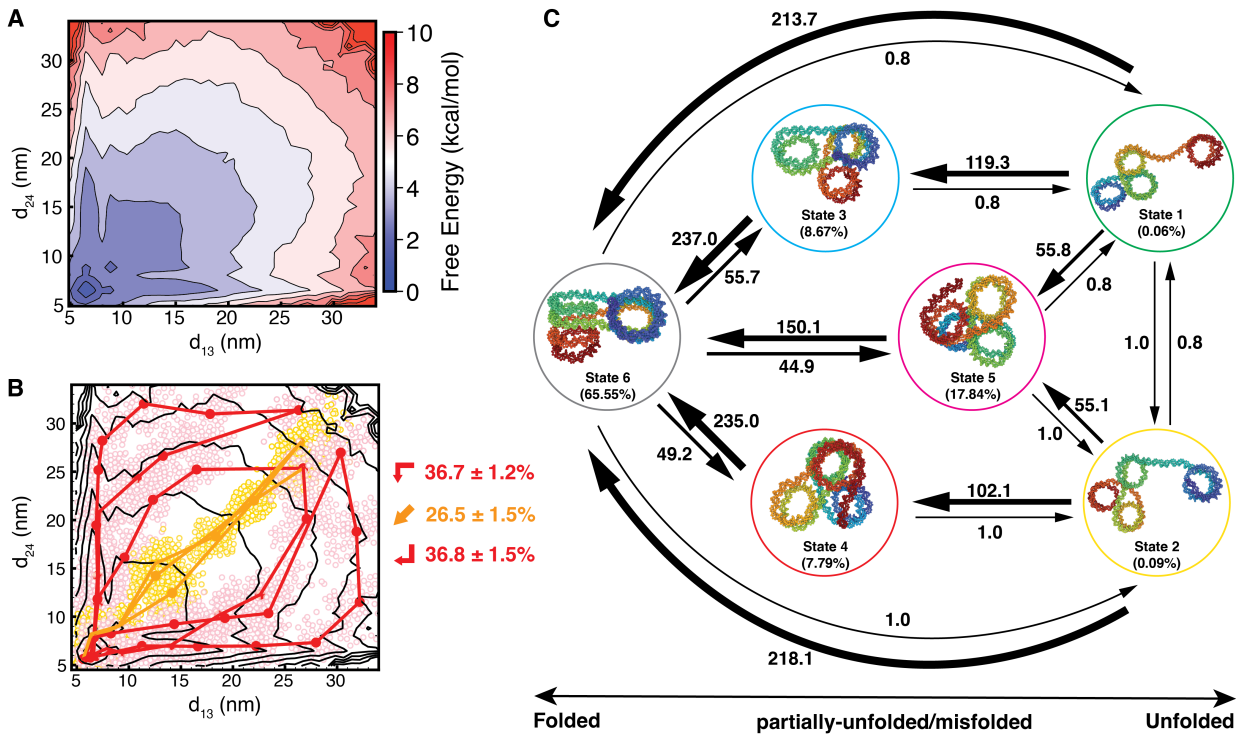

Figure 7.3: Folding pathways and kinetics for the NRL = 167 tetra-nucleosome.(A) The free energy profile along the center-of-geometry distance between 1-3 (d_{13}) and 2-4 (d_{24}) nucleosomes. (B) The three reaction channels for tetra-nucleosome folding. Top three transition pathways with most reactive flux from each one reaction channel are drawn as lines. The filled and open circles represent the centers of, and the MD configurations belonging to, the microstates along the pathways, respectively. The total reactive flux of each reaction channel is provided on the side. (C) Diagram of the non-Markovian dynamics model with six macrostates. The numbers represent rates estimated as inverse MFPT labeled in unit of $(10^9 \text{ steps})^{-1}$ for the corresponding transitions. Histone proteins are not shown in the representative configurations of each state.

tetra-nucleosome folding differs significantly from typical protein folding, where a few leading kinetic pathways contribute the majority of fluxes. Instead, it resembles multi-body systems such as self-assembly, where pathways need to be further grouped to facilitate the understanding of biological mechanisms. Applying the developed Latent-space Path Clustering (LPC) algorithm, three reaction channels for chromatin folding were identified: the sequential channels correspond to processes where one pair of nucleosomes stacks before the other, while the concerted channel represents processes where both pairs of nucleosomes stack simultaneously (Figure 7.3B). The low dimension projection with the two collective variables, however, leave the structural details of chromatin folding unresolved.

Further insight into chromatin folding can be obtained from analyzing an IGME model with six metastable macrostates. The representative conformations, equilibrium populations of these macrostates, their transition network, and transition rates (estimated as the inverse mean first passage times, MFPTs) are illustrated in Figure 7.3C. The sparsely populated states 1 and 2 correspond to unfolded structures. Their high transition rates to the folded state, state 6, and low rates for backward transitions, are consistent with the downhill free energy landscape.

States 3 and 4 exhibit partially unfolded structures in which one nucleosome extrudes away from the remaining three. These configurations resemble those seen in sequential chromatin folding pathways. A cryo-EM study has directly observed tri-nucleosome configurations [231], underscoring the IGME model's capability to identify metastable states. The IGME model also identifies a misfolded state (state 5). This state is characterized by non-canonical stacking between the first and fourth nucleosomes (i.e., i & $i+3$ contacts). To achieve the zigzag conformation of state 6, these non-canonical interactions must be disrupted, leading to slow transition rates. We classify this state as misfolded because transition path analysis indicates that none of the top 3,000 most reactive pathways pass through it. Future cryo-EM studies may determine whether this state can be resolved.

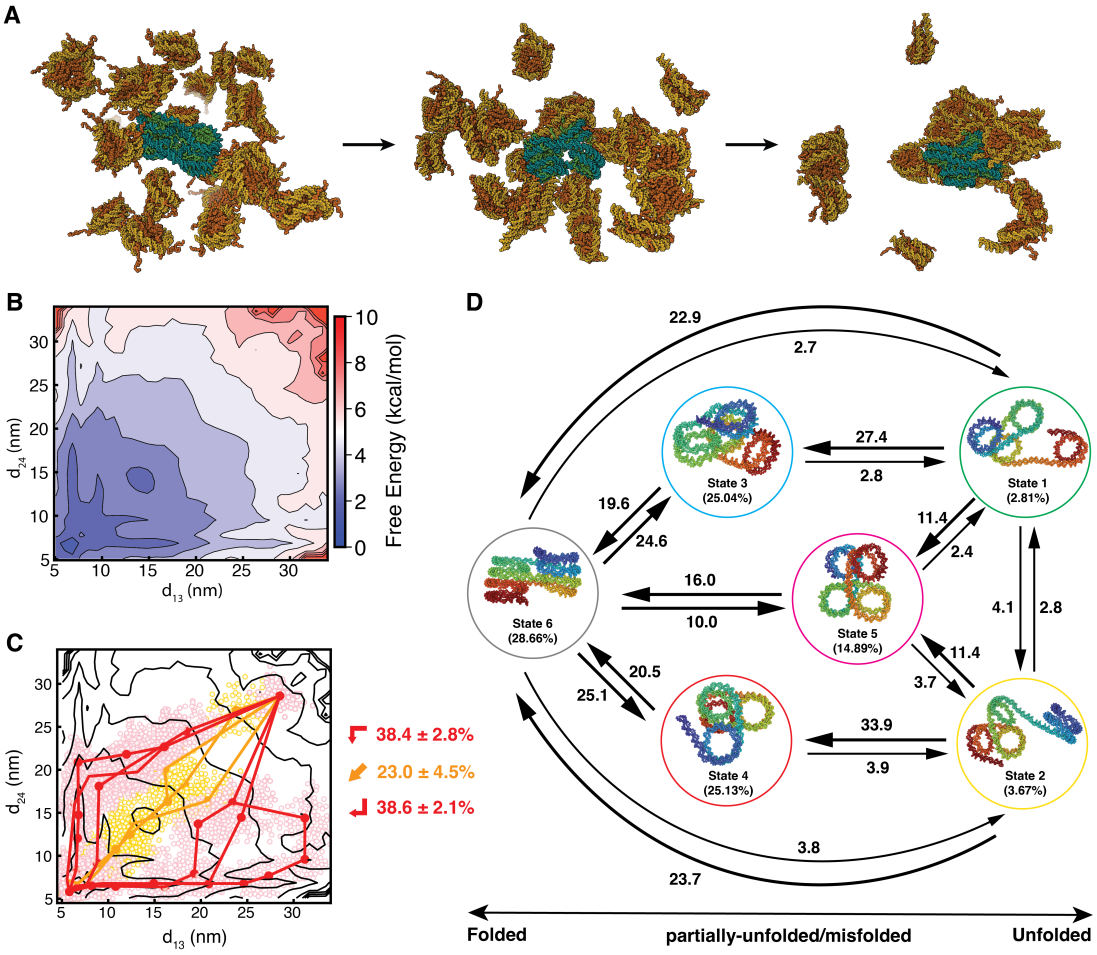

Figure 7.4: Folding pathways and kinetics for the NRL = 167 tetra-nucleosome embedded in nucleosome condensate. (A) Illustration of the starting, middle, and end configurations of the condensate system along a 70 million step long simulation trajectory. The tetra-nucleosome is shown in cyan and green, and individual nucleosomes are shown in yellow and orange. (B) The free energy profile along the center-of-geometry distance between 1-3 (d_{13}) and 2-4 (d_{24}) nucleosomes from the tetra-nucleosome. (C) The three reaction channels for tetra-nucleosome folding. Top three transition pathways with most reactive flux from each one reaction channel are drawn as lines. The filled and open circles represent the centers of, and all the MD configurations belonging to, the microstates along the pathways, respectively. The total reactive flux of each reaction channel is provided on the side. (D) Diagram of the non-Markovian dynamics model with six macrostates. The numbers represent rates estimated as inverse MFPT labeled in unit of $(10^9 \text{ steps})^{-1}$ for the corresponding transitions. Histone proteins are not shown in the representative configurations of each state.

7.4 Nucleosome Condensate Promotes Chromatin Unfolding

Our study on the isolated tetra-nucleosome does not consider the multi-body nuclear environment. The presence of additional nucleosomes may not only act as a crowding agent but also affect the stability of chromatin conformations through direct interactions. Inter-nucleosomal interactions could destabilize the chromatin fiber by forming interdigitated configurations [208, 232]. In this study, we examine the effects of immersing the tetra-nucleosome in the multi-body nucleosome condensate environment on chromatin folding.

In simulations of this condensate system, individual free nucleosomes were initially uniformly distributed throughout the cubic simulation box with the tetra-nucleosome at the center (Figure 7.4A). Over time, they rapidly and spontaneously aggregated around the tetra-nucleosome. This aggregation resulted in contacts between free nucleosomes and those within the tetra-nucleosome. Such contacts can be clearly observed in the end configuration of a typical simulation trajectory provided in Figure 7.4A. They are also evident in the distributions of free nucleosomes around the tetra-nucleosome. We observed that the radial distributions show significant populations at distances as small as 5 nm, a value achievable only with stacked nucleosomes.

We further integrated individual unbiased simulations to construct microstate-MSM for the tetra-nucleosome in the condensate. By using the equilibrium populations of microstates, we estimated the two-dimensional free energy landscape of the tetra-nucleosome. As shown in Figure 7.4B, the overall shape of the free energy landscape remains similar to that of the isolated system, but the free energy values of the folding intermediate structures have decreased. This reduction is attributed to interactions between the tetra-nucleosome and free nucleosomes, which stabilize partially folded structures. However, the overall change in the free energy landscape is relatively small, on the order of 1-2 kcal/mol.

To more directly and accurately examine chromatin folding kinetics, we constructed a six-macrostate IGME model (Figure 7.4D). We observed a significant increase in the populations of partially unfolded states 3 and 4, which are comparable to the population of the folded state 6. This finding underscores the changes in the free energy observed in Figure 7.4B and the impact of

nucleosome condensate on chromatin stability. Notably, the unfolding rates from state 6 to these intermediate states are faster than the folding rates. Therefore, the nucleosome condensate does not fundamentally alter the folding modes of the tetra-nucleosome (i.e., the distribution of fluxes among different reaction channels remains unchanged), but it quantitatively modulates the free energy landscape, favoring unfolding dynamics driven by inter-nucleosomal contacts.

7.5 The Role of DNA Linker Length on Chromatin Folding

Our study thus far has focused on a tetra-nucleosome with an NRL of 167 bp (i.e., 20-bp linkers). However, linker length is known to significantly influence chromatin organization [213, 233–237]. A DNA linker of $10n$ bp facilitates well-aligned stacking between i and $i+2$ nucleosomes to form the zigzag fibril structure [196, 197, 238, 239]. Extending the linker by 5 bp induces a half-turn twist in the DNA, disrupting the stacking between i and $i+2$ nucleosomes and destabilizing the zigzag conformation [240]. These $10n+5$ linkers are significant and have been observed in certain mammalian cells [241]. However, structural characterization of chromatin with $10n+5$ linker DNA is scarce due to the irregularity and instability of their organization. We further investigate how increasing the linker by 5 and 10 bp impacts chromatin stability and folding dynamics using the computational pipelines mentioned earlier.

The folding of the tetra-nucleosome with an NRL of 177 resembles that of an NRL of 167 but exhibits quantitative differences in reactive fluxes of folding. Its free energy landscape also shows a global minimum at small values of d_{13} and d_{24} (Figure 7.5 A), indicating stacked two-column structures. However, the stability of this folded state is reduced, as evidenced by its lower equilibrium populations in the six-macrostate IGME model (Figure 7.5 B). The folded state brings linker DNA into close proximity, increasing electrostatic repulsion for longer linkers, which decreases stability. Intermediate states along the sequential pathway also feature stacked nucleosomes with closely positioned linker DNA, resulting in similar electrostatic penalties. As a result, unlike the NRL = 167 tetra-nucleosome, the NRL = 177 one favors the concerted folding pathway over the sequential ones. Notably, our findings align with those of previous studies[234], who observed an expansion of chromatin with longer $10n$-bp linkers.

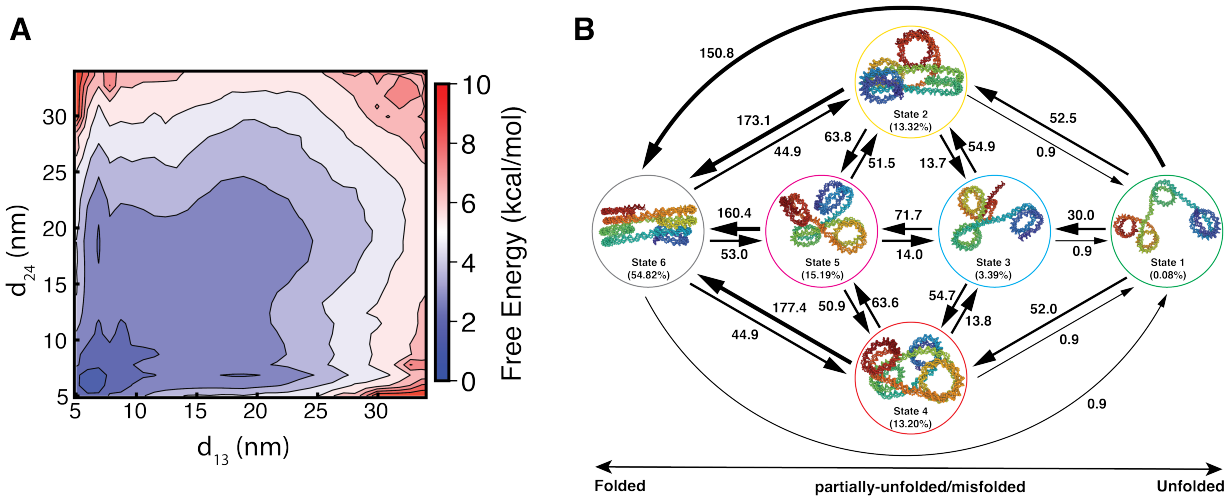

Figure 7.5: Folding pathways and kinetics for the NRL = 177 tetra-nucleosome. (A) The free energy profile along the center-of-geometry distance between 1-3 (d_{13}) and 2-4 (d_{24}) nucleosomes. (B) Diagram of the non-Markovian dynamics model with six macrostates. The numbers represent rates estimated as inverse MFPT labeled in unit of $(10^9 \text{ steps})^{-1}$ for the corresponding transitions. Histone proteins are not shown in the representative configurations of each state.

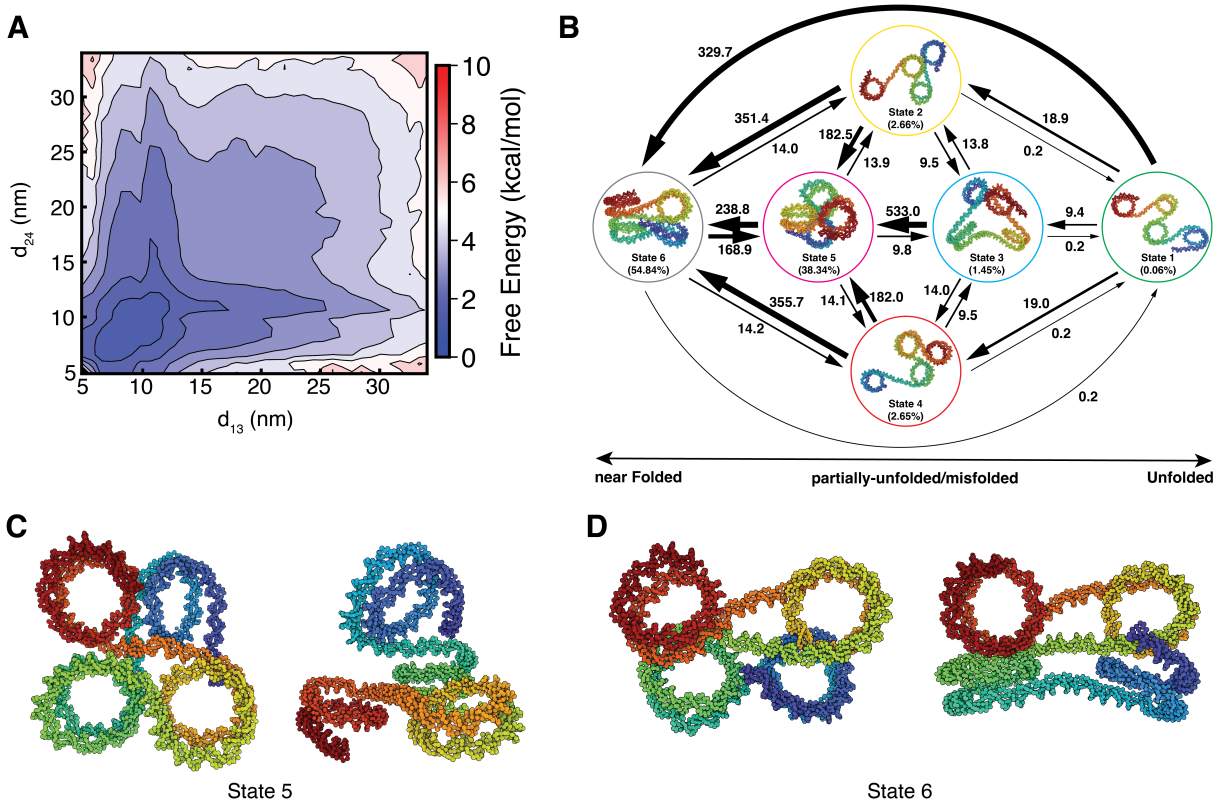

Figure 7.6: Folding pathways and kinetics for the NRL = 172 tetra-nucleosome (A) The free energy profile along the center-of-geometry distance between 1-3 (d_{13}) and 2-4 (d_{24}) nucleosomes from the tetra-nucleosome. (B) Diagram of the non-Markovian dynamics model with six macrostates. The numbers represent rates estimated as inverse MFPT labeled in unit of $(10^9 \text{ steps})^{-1}$ for the corresponding transitions. Histone proteins are not shown in the representative configurations of each state. (C,D) Addition representative structures the compact stable states 5 and 6.

Unlike systems with 10n-bp linker DNA, the NRL = 172 tetra-nucleosome system explores distinct chromatin conformations over a different folding landscape. As shown in Figure 7.6A, the free energy profile has a broader global minimum and a smaller free energy gap between the folded and unfolded states. This broad basin is due to the absence of a single stable conformation. Instead, an ensemble of irregular structures are equally probable. The six-macrostate IGME model (Figure 7.6B) reveals that two states, 5 and 6, emerge as "folded" states with comparable populations. Unlike the folded states in 10n bp linker systems, these states do not feature a dominant stable configuration but rather an ensemble of conformations (Figures 7.6C and 7.6D). While the structures in states 5 and 6 are compact and promote protein-DNA interactions, they do not exhibit perfect stacking between i and $i+2$ nucleosomes due to the topological constraints imposed by the additional 5 bp. Instead, they resemble intermediate states observed along the folding pathways for 10n bp linker systems. The high transition rates between states 5 and 6 further emphasize the dynamic nature of the tetra-nucleosome with 10n+5 bp linkers.

Meanwhile, the unfolded states, 1, 2, and 4, which have more extended structures in the NRL = 172 tetra-nucleosome, also reveal distinct topology. Unlike the extended states seen in 10n bp linkers, where consecutive nucleosomes reside on the same side of the linker DNA, the half-turn linker DNA in NRL = 172 tetra-nucleosome introduces additional twist, causing consecutive nucleosomes to reside on opposite sides of the linker.

7.6 Conclusions and Discussion

We integrated CG model based MD simulations, Markov state modeling, and non-Markovian dynamic modeling techniques to comprehensively characterize chromatin stability and folding mechanisms. These advanced techniques enable us to explore the long-timescale folding dynamics beyond the reach of direct simulations. Our study of a single tetra-nucleosome reveals multiple folding pathways, with intermediate states resembling *in vivo* chromatin organization [205, 242, 243]. For instance, intermediates along the concerted pathway often adopt a coplanar geometry, with all four nucleosomes in the same plane, resembling the β-rhombus configuration [205]. In contrast, intermediates along the sequential pathway have at least one nucleosome out of the plane,

similar to the α-tetrahedron structure [205]. These findings suggest that the absence of regular fibril configurations inside cell nuclei may result from chromatin unfolding driven by factors such as the condensate environment and linker length.

Our examination of the tetra-nucleosome within the nucleosome condensate supports the hypothesis that a crowded environment, typical inside cell nuclei, can drive chromatin unfolding [208]. The stacking interactions that stabilize the fibril configuration in a single chromatin chain may be replaced by contacts between nucleosomes from different chains. The replacement leads to the formation of irregular conformations, even for DNA linker length that favor fibiril structures.

Additionally, we found that $10n$+5 bp linkers destabilize the zigzag fibril configurations, favoring structures that resemble *in vivo* configurations observed by cryo-electron tomography [242, 243]. Chromatin with these DNA linkers tends to adopt compact conformations without perfect stacking between i and $i+2$ nucleosomes. These conformations are irregular and form a dynamic ensemble, similar to the behavior of intrinsically disordered proteins.

Our study demonstrates that residue-level CG models allow predictive modeling of chromatin organization at near-atomistic resolution under conditions where experimental techniques face challenges. While our focus was on the impact of nucleosome condensate on chromatin organization, this approach can be generalized to study other protein condensates, such as HP1α for constitutive heterochromatin. Recent advancements in force fields and software highlight the potential for exciting future research directions [244–246].

7.7 Methods

7.7.1 Coarse-grained molecular dynamics simulations of chromatin

Following previous studies [207, 208], we utilized residue-level CG models to examine chromatin conformational dynamics. Specifically, we combined the 3SPN.2C model [221, 247, 248], which represents DNA with three sites per nucleotide, and the SMOG model [249], which represents histone proteins with one bead per amino acid, to model chromatin. Protein-DNA interactions included electrostatic interactions and non-specific Lennard-Jones potentials. Electrostatic interactions were calculated at 300 K and 150 mM ionic strength using the Debye-Hückel potential.

This setup has been shown to replicate the energetics of DNA unwinding in single nucleosomes [250] and the force-extension curves of chromatin fibers [208]. To reduce computational cost and prevent nucleosome sliding, we rigidified the histone core and the inner 73 bp layer of core DNA during all simulations. Piror study [207] has demonstrated that such rigidification has negligible effects on the chromatin folding landscape.

We studied three isolated tetra-nucleosome systems with NRL of 167, 172, and 177 bp, respectively. For each system, we performed 4,643 independent, unbiased MD simulations starting from distinct initial configurations. Initial configurations for these simulations were prepared using 10,000 tetra-nucleosome structures from a previous study that applied enhanced sampling methods to comprehensively cover the configuration space [207]. We computed the inter-nucleosome distances for the 10,000 structures and selected those with $d_{13} \geq d_{24}$ (4,643 in total) to serve as reference values for further MD simulations. From these reference inter-nucleosome distances, we generated the initial configurations using 0.2 million-step-long restrained MD simulations to bias the d_{13} and d_{24} toward reference values. From these initial configurations, we conducted unbiased NVT simulations for production. For the NRL = 167 system, each trajectory lasted at least 0.8 million steps, while for the NRL = 172 and NRL = 177 systems, the trajectory length was consistently set to 1 million steps. All simulations were performed at 300 K using LAMMPS on CPUs [251], and the pairwise distances between each pair of nucleosome geometric centers, i.e., $(d_{12}, d_{13}, d_{14}, d_{23}, d_{24}, d_{34})$, were recorded every 500 steps.

Additionally, we conducted simulations for a system where the NRL = 167 tetra-nucleosome is embedded in a nucleosome solution. Due to the higher computational cost of this system, we limited the number of independent simulations to 530. Initial configurations of the tetra-nucleosome for these simulations were prepared by selecting structures from the centers of the 530 microstates constructed for the isolated NRL = 167 tetra-nucleosome. Then we placed each of the tetra-nucleosomes in the center of a cubic box with a 55 nm edge, and randomly placed 26 additional single nucleosomes to achieve a total nucleosome concentration of 0.3 mM. We relaxed each of these initial configurations with a 0.2 million-step NVT simulation, during which the tetra-nucleosome was fixed as a rigid body while single nucleosomes were free to move. From

the relaxed configurations, we conducted production NVT simulations of at least 7 million steps, allowing all nucleosomes to evolve freely. Simulations were performed at 300 K with OpenMM and OpenABC packages on GPUs [188, 245]. The pairwise distances between nucleosomes within the tetra-nucleosome, as well as the positions of all nucleosomes, were recorded every 500 steps for analysis.

7.7.2 Markov state modeling and transition path analysis for chromatin folding

We constructed independent microstate-MSMs and performed transition path analysis for all four systems following a consistent protocol, briefly outlined below.

(a) Select the converged segments of the trajectories and duplicate the converged trajectories according to the reflection symmetry of the nucleosome indices. All the following analyses are applied to the converged and duplicated trajectories.

(b) With six inter-nucleosomal distances d_{ij} as input features, apply the time-lagged independent component analysis (tICA) method with kinetic mapping algorithm to further reduce dimensionality and uncover independent collective variables (CVs) that better represent slow timescale dynamics [76, 111, 228, 229].

(c) Group MD conformations into microstates by the K-Means algorithm according to their kinetic similarities based on the geometric distances in the tICA CV space. The hyperparameters of the tICA and clustering (i.e., tICA relaxation time, number of tICs, and number of microstates) are optimized using the cross-validation with the generalized matrix Rayleigh quotient (GMRQ) score [90].

(d) Construct and validate the microstate-MSMs by the Chapman-Kolmogorov (CK) test and implied time scale (ITS) analysis [27, 28].

(e) Employ the Transition Path Theory (TPT) [40, 41, 230] to elucidate the folding kinetic pathways and their corresponding fluxes.

(f) Lump multiple parallel kinetic pathways into a small set of metastable and representative reaction channels using Latent-space Path Clustering (LPC) algorithm to facilitate the understanding

of folding mechanisms [13, 48, 56].

All analyses were conducted using in-house Python codes and codes based on MSMBuilder version 3.8.1. [38].

7.7.3 Identifying reaction channels from transition pathways

By conducting transition path analysis with the microstate-MSM, we identified over 20,000 pathways connecting the unfolded to folded states. Strikingly, the most dominant pathway only contributes 0.09% of the total reactive flux for the folding. This contrasts with typical protein folding, such as NTL9 folding, where the top 10 pathways can account for 25% of the total flux [106]. In the case of tetra-nucleosome folding, approximately 600 pathways are required to reach the same flux level. The observation of the downhill landscape and numerous parallel pathways with comparable fluxes suggests that tetra-nucleosome folding is more akin to heterogeneous aggregation in self-assembly systems than to typical protein folding [13].

To aid in interpreting the multiple parallel pathways with similar fluxes, we grouped the pathways into a small set of metastable reaction channels using the Latent-space Path Clustering algorithm [48]. This method identified three metastable reaction channels: two sequential channels (up and down sequential channels) and one concerted channel (Figure 7.3B). The overall reactive flux in the sequential channels is slightly higher than in the concerted channel, indicating a preference for reactions. This preference may result from more stable intermediates encountered along pathways with increased internucleosomal contacts.

7.7.4 IGME modeling of chromatin folding

To facilitate the interpretation of chromatin folding, for each simulation system, we modeled its folding dynamics with a six-macrostates IGME model[31] (Figure 7.2D). These six macrostates were constructed by lumping hundreds of microstates using the Robust Perron Cluster Analysis algorithm[93, 94]. Unlike MSMs, which approximate transition dynamics between states by Markovian jumps based on the first-order master equation, the IGME model encodes non-Markovian dynamics through time integrations of memory kernels based on the Generalized Master Equation

[31]. Specifically, IGME model describes the evolution of transition probabilities over time using $\mathbf{T}(t \geq \tau_k) = \mathbf{A}\hat{\mathbf{T}}^t$, where $\mathbf{T}(t)$ is the transition probability matrix (TPM) at lag time t, and τ_k is the relaxation time of the memory kernels. The matrices $\mathbf{A}$ and $\hat{\mathbf{T}}$ are constants estimated from the simulation data.

We determined the relaxation time of the memory kernels τ_k by examining the convergence of the mean integral memory kernels. The $\mathbf{A}$ and $\hat{\mathbf{T}}$ matrices were estimated using a set of TPMs at short lag times through least squares fitting with a Lagrangian approach. The optimal fitting range was selected by a systematic scanning to minimize the deviation between the predictions from IGEM model and raw MD data. The optimal IGME models were further validated and compared to MSMs using the Chapman-Kolmogorov test. All results regarding the equilibrium populations and MFPTs for macrostates were derived from the optimal and validated IGME models.

Chapter 8

Exploring Dynamical Heterogeneities in Supercooled Liquids using Unsupervised Machine Learning

This chapter is reproduced in part with permission from Qiu, Y; Jang, I; Huang, X; & Yethiraj, A; Unsupervised machine learning for supercooled liquids *arXiv* **2024**, 2404.04473.

8.1 Introduction

As introduced in Chapter 1, supercooled liquids represent a unique multi-body system with distinct dynamical properties. Over the past decade, the unique heterogeneous dynamics of supercooled liquids have been observed and extensively documented across various systems. However, the relationship between their microscopic structures and long-time dynamics remains a mystery. With advancements in machine learning, various supervised methods have been developed to establish structure-dynamics connections. Following the unsupervised approach developed for MSMs, we extended time-lagged analysis to study the long-term dynamics of supercooled liquids, demonstrating its effectiveness over state-of-the-art supervised methods.

The glass transition is a fascinating phenomenon in physics. When a liquid is cooled with sufficient rapidity, certain substances can avoid crystallization and instead reach a homogeneous amorphous solid state known as the "supercooled" state. The microscopic static structure, for example, the powder x-ray diffraction pattern, is similar to that of a liquid, but the dynamic properties, for example, the viscosity, are slower by several orders of magnitude [11, 12]. Understanding the mechanisms behind the glass transition is one of the major challenges in the field of liquid state physics [42, 252–254].

One of the key dynamic properties of supercooled liquids is the emergence of strong dynamical heterogeneity: molecules in certain regions actively rearrange, while molecules in other regions

remain nearly frozen over the extremely long duration [255, 256]. Characterizing amorphous structures and understanding the relationship between disordered structures and heterogeneous dynamics at the microscopic level are crucial and active areas of research [42, 44, 252–254, 257–260]. Over the years, various structural descriptors have been established to characterize amorphous structures (e.g. local density [257, 261], soft modes [254, 257], potential energy [257, 262–264], Voronoi tessellation [265], etc.), and ongoing efforts have been made to identify key structures related to local relaxation dynamics [42, 70]. However, most manually crafted descriptors have been found to be highly system-dependent and have limited effectiveness in accurately predicting relaxation dynamics [42, 263].

To quantitatively characterize the connections between the structures and subsequent heterogeneous dynamics for supercooled liquids, an important dynamical property called *propensity* has been employed [254, 266]. Propensity is obtained from iso-configurational ensemble simulations, where multiple trajectories are initialized from the same configuration but with different velocities randomly sampled from the equilibrium Boltzmann distribution [254, 266]. The propensity for each individual particle is defined as either the bond-breaking correlation function [258, 259, 267] or the absolute displacement [257] of the particle within a specified time interval, averaged over all iso-configurational trajectories. Simulations clearly show dynamic heterogeneity, i.e., there is a distribution of propensities, with particles of similar propensities clustered spatially. The simulations do not, however, elucidate the key structural origin of this heterogeneity.

Recently, machine learning (ML) techiques have been extensively used to detect relevant structural features and establish structure-dynamics relationships for supercooled liquids. Specifically, supervised ML models, such as support vector machines (SVM)[268–270], multilayer perceptrons (MLP)[44, 259, 271], and graph neural networks (GNN)[43, 272, 273], which require prior training with large datasets where the propensity is known in prior, have shown significant accuracy in predicting dynamical properties from structural pattern. However, a significant drawback of supervised methods is their high computational intensity: the large number of parameters in supervised models necessitates substantial training data [43, 44], and obtaining propensities for these datasets

requires lengthy iso-configurational ensemble simulations, which become progressively challenging as temperatures decrease [259]. Supervised models are also not generalizable, i.e., for every new system, a new training dataset must be generated. Additionally, the low interpretability of supervised models limits the deep understanding of important structures for dynamics, rendering them high-performing yet opaque black boxes [42].

Meanwhile, few unsupervised ML models that do not require prior training with target propensities have been developed. These models autonomously detect structural heterogeneity from a large set of structural descriptors and attempt to correlate it with dynamic heterogeneity [70, 260, 261, 274]. Although unsupervised models are much more data-efficient and explainable, they always provide very limited prediction accuracy for dynamics compared to supervised models [42, 70]. Improving the prediction power of unsupervised models has emerged as one of the major challenges in the ML for supercooled liquids [42]. Many efforts have been made, including the development of better structural descriptors and the testing of various model architectures, but it remains unclear how to construct the effective unsupervised models [70].

Recent studies have shown that unsupervised models, despite using similar or even identical structural descriptors as supervised models, consistently demonstrated considerably lower performance [42, 70, 260]. This suggests that the criteria currently used by unsupervised models to rank feature importance, which are based on feature variances [70, 260], need to be redefined. In this work, we developed a new unsupervised ML protocol to explore the structure-dynamics connections of supercooled liquids. Departing from previous unsupervised models which focus on identifying high-variance components, our approach aims to identify descriptors that exhibit strong correlations over a short time period, referred to as lag time Δt. Specifically, we employed two methods: time-lagged canonical correlation analysis (TCCA) [275] and time-lagged autoencoder (TAE) [55] to extract low-dimensional order parameters (OPs) with highest time-correlations over Δt from high-dimensional structural descriptors. Since the construction of OPs does not use any information about the target output labels, i.e., propensity, they are therefore completely unsupervised. We demonstrate that OPs obtained with small but non-zero Δt, such as thousandths

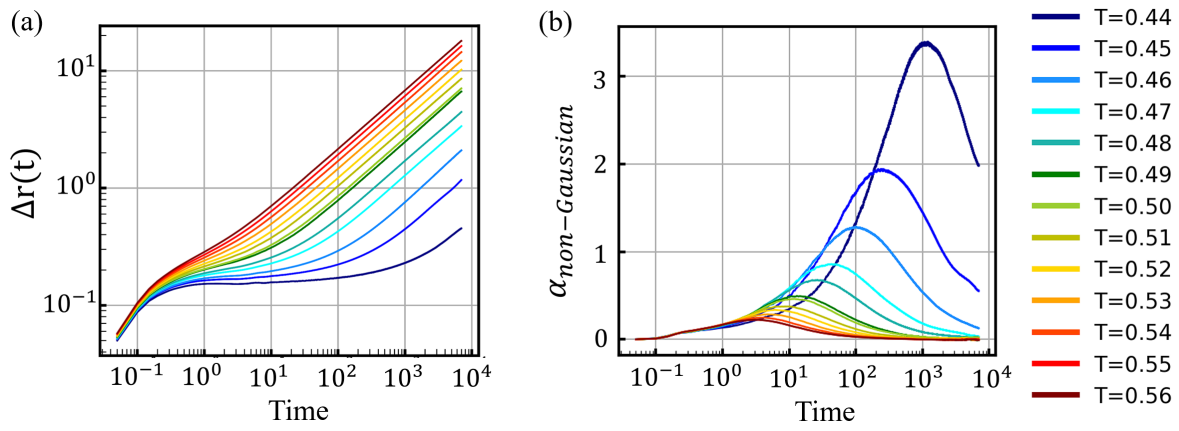

Figure 8.1: Dynamical characteristics of the 3D Kob-Andersen model at different temperatures. (a) Root mean square displacement, (b) Non-Gaussian parameters for different temperatures. This figure is reproduced from Qiu. *et al.*[45]

of the relaxation time of the supercooled liquid, provide valuable insights into dynamic heterogeneity over long timescales. The predictive power of our OPs for long-time propensity is even comparable to that of many supervised methods on the benchmark system. Meanwhile, we show that our unsupervised model, constructed at a single temperature, exhibits robust transferability across different temperatures, highlighting its potential to investigate the dynamical properties of lower-temperature regions. Additionally, our methods highlight the importance of all length-scales, particularly the medium-range density descriptors, in capturing long-time dynamic heterogeneity.

8.2 System, Structural Descriptors and Unsupervised ML Models

8.2.1 Physical system and dynamical observables

We study the benchmark glassy-forming system: the 3D KA 80:20 binary Lennard-Jones mixture [276]. The system consists of 3277 big particles of type A and 819 small particles of type B, which interact via a Lennard–Jones potential: $V_{\alpha\beta}(r) = 4\epsilon_{\alpha\beta}\left[\left(\frac{\sigma_{\alpha\beta}}{r}\right)^{12} - \left(\frac{\sigma_{\alpha\beta}}{r}\right)^{6}\right]$ where $\alpha, \beta \in \{A, B\}$, $\epsilon_{AA} = 1.0$, $\epsilon_{AB} = 1.5$, $\epsilon_{BB} = 0.5$, $\sigma_{AA} = 1.0$, $\sigma_{AB} = 0.8$ and $\sigma_{BB} = 0.88$. The units for distance, time, and temperature are σ_{AA}, $\sigma_{AA}\sqrt{m/\epsilon_{AA}}$, and ϵ_{AA}/k_B, respectively, where k_B is Boltzmann's constant, and m is the mass of the particles. We performed molecular

dynamics simulations using the package LAMMPS [277] to obtain equilibrium configurations at reduced temperatures ranging from 0.44 to 0.56.

Two types of propensities are computed to quantify the dynamic heterogeneity: bond-breaking correlation propensity and absolute displacement propensity. All propensities are calculated by averaging at least 30 independent isoconfigurational simulations. The bond-breaking propensity of particle i after a time interval δt is defined as: $\mathcal{C}_B^i(\delta t) = \langle n_{\delta t}^i / n_0^i \rangle_{iso}$. Here, n_0^i represents the initial number of nearest neighbors for particle i, while $n_{\delta t}^i$ measures the number of those original neighbors retained after a time δt [267]. With the average bond-breaking propensity, the bond-breaking relaxation time τ_α^{BB} could be computed as: $\langle \frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_B^i(t = \tau_\alpha^{BB}) \rangle = 0.5$ [267], where N denotes the number of A-particles. Meanwhile, the displacement-based propensity of particle i after time interval δt is defined as: $\mathcal{D}_i(\delta t) = \langle |\mathbf{r}_i(\delta t) - \mathbf{r}_i(0)| \rangle_{iso}$, where $\mathbf{r}_i(\delta t)$ denotes the position of particle i at time δt [43].

To characterize the dynamic heterogeneity of supercooled liquid, we also compute the non-Gaussian parameter, $\alpha(t) \equiv \frac{3}{5} \frac{\langle (\Delta r)^4 \rangle}{\langle (\Delta r)^2 \rangle^2} - 1$ where Δr is the particle displacement over time t, and $\langle ... \rangle$ denotes the thermodynamic average (see Figure 8.1). The peak value of α is denoted as α_m.

8.2.2 Local environment structural descriptors

We employ the well-established structural descriptors to encode the local structural environment for each individual particle. The local radial density distribution for a given particle i is characterized using multiple Gaussian kernel functions:

$$G_i(r, \delta, s) = \sum_{j \neq i, s_j = s} e^{-\frac{(r_{ij} - r)^2}{2\delta^2}} \tag{8.1}$$

where r_{ij} signifies the distance between particle i and its neighbor j, $s_j = \{A, B\}$ is the species of particle j for which we aim to assess the density. By adjusting the values of r, δ, and s, the G_i capture density distributions for different shells and different types of particles surrounding particle i. G_i with different hyper-parameters constitute a 200-dimensional descriptors vector for each particle.

Additionally, we also construct angular descriptors to represent the angular distribution using spherical harmonics functions. First, for a given particle i, we define the complex quantities as follows:

$$q_i(l, m, r, \delta) = \frac{1}{Z} \sum_{j \neq i} e^{-\frac{(r_{ij}-r)^2}{2\delta^2}} Y_l^m(\mathbf{r}_{ij}) \tag{8.2}$$

where Y_l^m represents the spherical harmonic function of order l, with m being an integer ranging from $-l$ to $+l$. $\mathbf{r}_{ij}$, the distance vector between particle i and j, can be used to determine polar and azimuthal angles. $Z = \sum_{j \neq i} e^{-\frac{(r_{ij}-r)^2}{2\delta^2}}$ is a normalization constant. Subsequently, rotationally invariant angular descriptors are defined by summing over m as:

$$q_i(l, r, \delta) = \sqrt{\frac{4\pi}{2l+1} \sum_{m=-l}^{m=l} |q_i(l, m, r, \delta)|^2} \tag{8.3}$$

The q_i quantifies the l-fold symmetry of the angular density distribution around particle i. By tuning different hyperparameters, q_i forms a 192-dimensional descriptor vector for each particle.

8.2.3 Unsupervised ML models

We use TCCA and TAE models to identify OPs with the strongest time-correlations in the descriptor space. The input data for the models can be represented as $\{\mathbf{X}_i(t), \mathbf{X}_i(t + \Delta t)\}_{i=1}^N$, where $\mathbf{X}_i(t) \in \mathbb{R}^d$ is the structural descriptor for particle i embedded at time t with d dimensions. $\mathbf{X}_i(t+\Delta t)$ is the descriptor for the same particle i, but encoded through the configuration at $t+\Delta t$, which has evolved from the configuration at t. To construct the models, first, we process the input data through mean-centering and covariance matrix whitening operations:

$$\tilde{\mathbf{X}}_i(t) = \mathbf{C}_{00}^{-\frac{1}{2}}(\mathbf{X}_i(t) - \frac{1}{N} \sum_{i=1}^{N} \mathbf{X}_i(t)) \tag{8.4}$$

$$\tilde{\mathbf{X}}_i(t + \Delta t) = \mathbf{C}_{\Delta t \Delta t}^{-\frac{1}{2}}(\mathbf{X}_i(t + \Delta t) - \frac{1}{N} \sum_{i=1}^{N} \mathbf{X}_i(t + \Delta t)) \tag{8.5}$$

where $\mathbf{C}_{00}$ and $\mathbf{C}_{\Delta t \Delta t}$ are self-covariance matrices for $\{\mathbf{X}_i(t)\}_{i=1}^N$ and $\{\mathbf{X}_i(t + \Delta t)\}_{i=1}^N$, respectively. Then both TCCA and TAE aim to identify the optimal solution for transformation $\mathcal{P}$ which

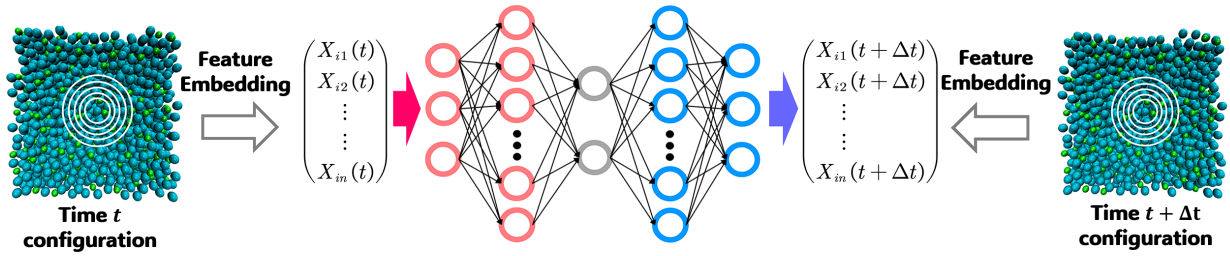

Figure 8.2: Schematic representation of the unsupervised ML model Time-lagged AutoEncoder (TAE). At a given temperature, the input descriptor vector $\mathbf{X}_i(t)$ consists of the local structural descriptors of the i-th particle at time t, while the output descriptor vector $\mathbf{X}_i(t+\Delta t)$ consists of the descriptors for the same particle, at time $t+\Delta t$. After training of TAE, the OPs are the values of the normalized bottleneck space variables (grey). This figure is reproduced from Qiu. *et al.*[45]

could minimize the loss function:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} ||\tilde{\mathbf{X}}_i(t+\Delta t) - \mathcal{P}\tilde{\mathbf{X}}_i(t)||^2 \tag{8.6}$$

TCCA assumes the transformation $\mathcal{P}$ is linear, allowing for an analytical solution to be obtained from the regression [275]:

$$\mathcal{P}_r = \mathbf{C}_{00}^{-\frac{1}{2}}\mathbf{C}_{0\Delta t}\mathbf{C}_{\Delta t\Delta t}^{-\frac{1}{2}} \tag{8.7}$$

where $\mathbf{C}_{0\Delta t}$ is the covariance matrix between $\{\mathbf{X}_i(t)\}_{i=1}^N$ and $\{\mathbf{X}_i(t+\Delta t)\}_{i=1}^N$. TCCA truncates the leading singular vectors of $\mathcal{P}_r$, obtained through SVD, to serve as OPs. Similar to principal component analysis (PCA), which identifies the components exhibiting the largest variances by minimizing the Frobenius norm reconstruction error, the OPs from TCCA show the strongest time-correlations.

The TAE method [55] adopts an autoencoder neural network to non-linearly approximate the $\mathcal{P}$ transformation and minimize the loss function (see Figure 8.2). After optimizing the TAE, it autonomously provides an encoder that maps high-dimensional descriptors to low-dimensional latent variables. We then apply PCA to orthogonalize and reorder the latent variables, thereby obtaining independent OPs.

8.3 Identifying Order Parameters via Unsupervised ML Models

We implement our unsupervised ML models on the 3-dimensional Kob-Andersen (KA) binary mixture system [276]. To obtain equilibrium configurations in different supercooled regimes (from $T = 0.44$ to 0.56), we conduct molecular dynamics simulations at various temperatures with multiple relaxation steps, ensuring the systems are fully relaxed. For each temperature, after generating the independent equilibrium configurations, we further extend each simulation for a short lag time Δt to obtain time-lagged configurations for the construction of our unsupervised models. In line with previous studies, we quantify dynamic heterogeneity using the bond-breaking correlation propensity $\mathcal{C}_B^i(\delta t)$ [258, 259, 267], which measures the fraction of nearest neighbors retained by particle i over the time interval δt, thus representing its mobility. This propensity also offers a way to extract the structural relaxation time, with the bond-breaking relaxation time τ_α^{BB} defined as the time at which the average bond-breaking propensity decreases to 0.5. All results in this section are reported for the large A-particles, but we note that the findings are independent of particle type.

Unlike systems such as the Wahnström model [278] with strong icosahedral order, the structure-dynamics relationship in the KA system is much more complex [42, 70, 260]. To date, no straightforward single structural descriptor has been identified that is highly correlated with the long-time dynamics of the KA system. Both prior supervised and unsupervised methods constructed a large number of fine-grained descriptors to capture subtle local structural differences and then established connections to dynamics. Based on previous studies, our unsupervised protocol can be divided into two steps. First, for each individual particle, we characterize its local structure at a given time t and at a lagged time $t + \Delta t$, respectively. Specifically, we encode the particle's surrounded radial density using 200-dimensional Gaussian kernel functions and its local angular distribution using 192-dimensional spherical harmonic functions. These descriptors have demonstrated great power in predicting dynamic heterogeneity within supervised schemes [44, 268, 271]. Second, TCCA or TAE is employed to identify the optimal transformation for compressing the descriptors at time t into low-dimensional representations, allowing for the most accurate reconstruction of descriptors at time $t + \Delta t$. These low-dimensional representations are considered OPs, along

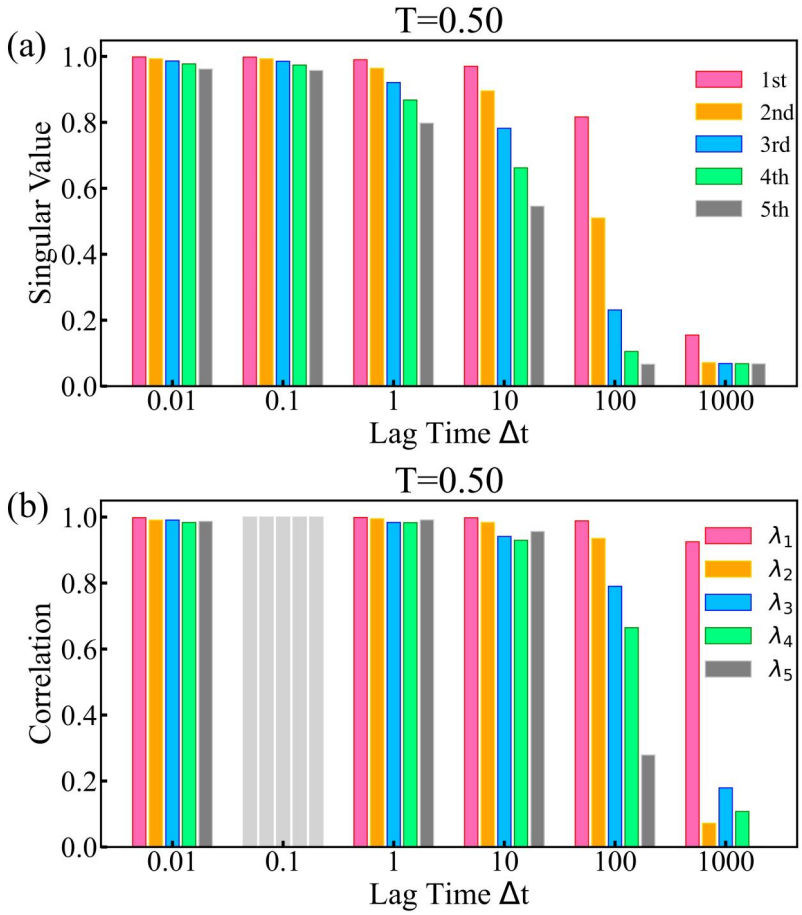

Figure 8.3: Selection of number of OPs and lag time Δt for TCCA models constructed with radial descriptors at T=0.50. (a) Top five singular values obtained from TCCA constructed with different lag times. (b) Pearson correlation between OPs obtained at different lag times (color) and those obtained with lag time of $\Delta t = 0.1$ (light). Each bar represents the correlation calculated between $\lambda_i(\Delta t)$ and $\lambda_i(0.1)$.

which the projections of descriptors lagged by time Δt exhibit the highest correlations. TCCA assumes the transformation is linear and obtains the OPs via singular value decomposition (SVD) of normalized time-correlation matrix. TAE utilizes an autoencoder to learn an optimal non-linear transformation, and its normalized latent space serves as OPs, as shown in Figure 8.2.

TCCA ranks the OPs according to the singular values from SVD, while TAE uses principal component analysis (PCA) to reorder the latent space. The obtained OPs are denoted as $\{\lambda_i(\Delta t)\}_{i=1}^{k}$. There are clearly two critical parameters need to be set for our unsupervised models, i.e., the number of OPs and the lag time Δt. To choose the number of OPs, we analyze how the time correlation of the projected time-lagged descriptors decay along each TCCA OP as Δt increases. Specifically, we examine the radial and angular descriptors separately. As shown in Figure 8.3(a), at temperature $T = 0.50$, the top 5 singular values, which reflect the time correlations along different OPs constructed from radial descriptors, are all close to 1 when Δt is very small but begin to decay quickly beyond the relaxation time ($\tau_\alpha^{BB} \simeq 117$ for $T = 0.50$). Notably, the time correlation along λ_1 decays much more slowly than the other OPs. This indicates that λ_1 constructed from radial descriptors captures the long-term relaxation of structural descriptors. In contrast, we observed that the time correlations along the OPs derived from angular descriptors are generally much weaker and decay more quickly than those from radial descriptors over a broad time range, suggesting that radial descriptors are better suited for capturing long-term dynamics.

Meanwhile, we also quantify the difference between OPs $\{\lambda_i\}_{i=1}^{5}$ constructed with different lag times using the Pearson correlation coefficient. We observe that for both the radial and angular descriptors, leading OPs obtained at very short lag times (e.g. $\Delta t = 0.1$) exhibit strong similarities and high correlations with those obtained at lag times around the relaxation time, especially for λ_1 (see Figure 8.3(b)). This demonstrates that OPs with strong short-time correlations are consistent with those strong long-time correlations, facilitating the inference of long-term dynamics from short-term fluctuations. And we found that all of above observations are independent of temperatures ranging from $T = 0.44$ to $T = 0.56$. Therefore, we will set the lag time to a small value, $\Delta t = 0.1$, and focus on the leading OP built with radial descriptors from TCCA or TAE, λ_1, to study long-term dynamic heterogeneity in the following analysis.

8.4 Explaining Dynamical Heterogeneity with Order Parameters

Using unsupervised TCCA and TAE models, the local radial descriptors of the particle i can be encoded into the OP λ_1, based on its evolution over a very short lag time, $\Delta t = 0.1$. For better visualization, we constructed the TAE with a 2-dimensional bottleneck space. When applying TAE at temperature $T = 0.50$, we observe that λ_1 reveals clear structural heterogeneity, as shown in Figure 8.4(a). Strikingly, we notice that the propensity around $\tau_\alpha^{BB}/2$, where the dynamics are strongly heterogeneous, follows a similar distribution pattern (see Figure 8.4(b)). And the visualization of the TAE bottleneck space further demonstrates the effectiveness of λ_1 in distinguishing between active and passive particles over long timescales (see Figure 8.4(c)). These observations suggest that the structural heterogeneity captured by λ_1 with a short lag time of less than $0.001\tau_\alpha^{BB}$ may provide a compelling explanation for long-term dynamic heterogeneity.

Next, we systematically examine the correlation between the OP λ_1 and the propensity $\mathcal{C}_B(\delta t)$ over a broad time range. Specifically, we use the Pearson correlation coefficient:

$$\rho_{\mathcal{C}_B} = cov(\mathcal{C}_B^i, \lambda_1^i)/\sqrt{var(\mathcal{C}_B^i)var(\lambda_1^i)} \tag{8.8}$$

to quantify the relationship between them. As shown in Figure 8.4(d), we find the λ_1 identified by TAE exhibits a strong correlation ($\rho_{\mathcal{C}_B} > 0.5$) with the propensity from time $\sim \tau_\alpha^{BB}/3$ to $\sim 3\tau_\alpha^{BB}$. The OP λ_1 from TCCA displays very similar behavior but with slightly lower correlations. In contrast, consistent with previous studies[70, 260], the OP constructed by PCA or the Autoencoder using the same descriptors, which represents the highest-variance component in the data, provides very limited insight into dynamic heterogeneity ($\rho_{\mathcal{C}_B} < 0.2$). Our unsupervised models clearly surpass traditional approaches, achieving state-of-the-art performance in accurately correlating structures with heterogeneous dynamics.

Additionally, the four-point susceptibility, $\chi_4(\delta t) = N(\langle \bar{\mathcal{C}_B^2}(\delta t)\rangle - \langle \bar{\mathcal{C}_B}(\delta t)\rangle^2)$, is commonly used to characterize the spatial dynamic heterogeneity of supercooled liquid, where N represents the number of A-type particles and $\bar{\mathcal{C}_B}(\delta t)$ is the averaged propensity [258, 259, 267]. We calculate this time-dependent scalar to quantify how structural factors contribute to the system's dynamic fluctuations over time (see Figure 8.4(d)). At short timescales, since particles predominantly move

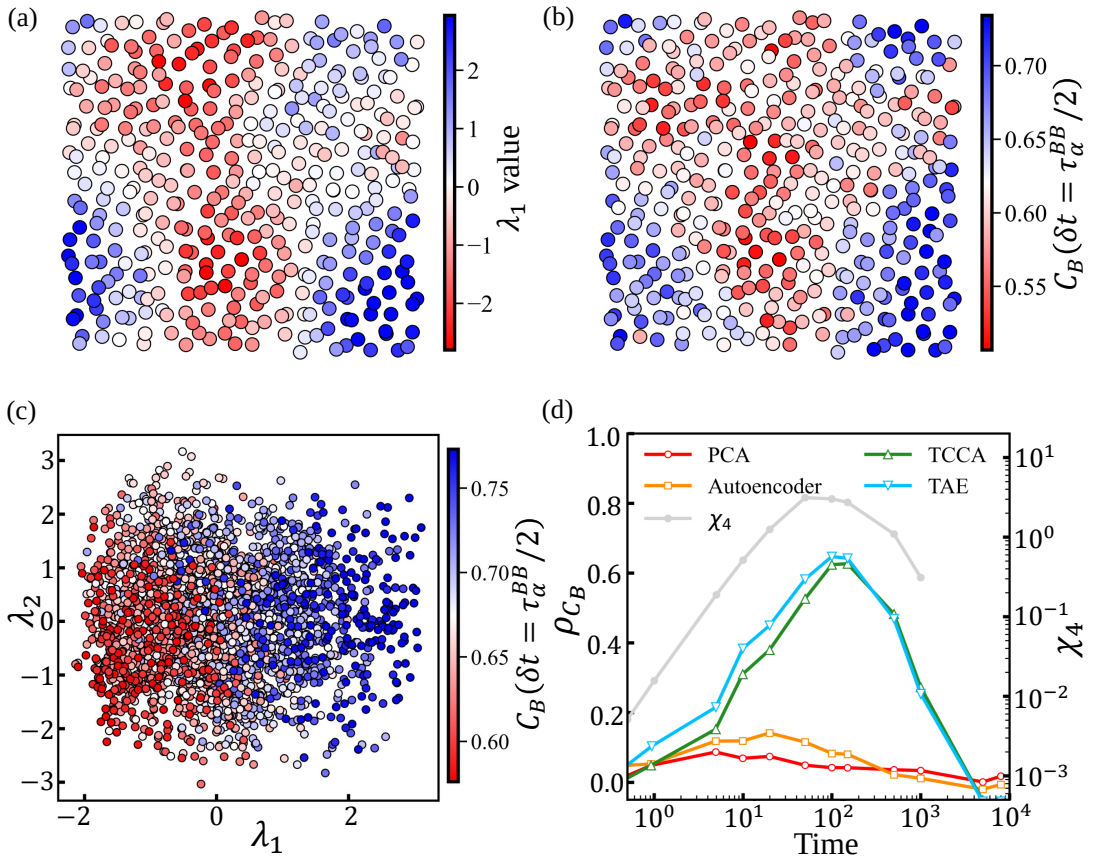

Figure 8.4: Visualization and characterization of dynamical heterogeneity at $T = 0.50$. (a) The 2D slice snapshot of the simulation box is randomly selected and visualized, with each particle colored based on its OP λ_1 value, encoded by the TAE constructed at $T = 0.50$. (b) The same snapshot, but with each particle colored based on its bond-breaking correlation propensity at $\tau_\alpha^{BB}/2$. (c) The 2D latent space of the TAE. Each point represents an individual particle, and the color is assigned based on the $\mathcal{C}_B(\tau_\alpha^{BB}/2)$ propensity. (d) The Pearson correlation coefficient between the leading OP identified by different unsupervised ML models and $\mathcal{C}_B(\delta t)$ over time is displayed. The susceptibility χ_4 over time is shown as the grey curve. This figure is reproduced from Qiu. *et al.*[45]

within their local confined cages, we find they typically display minimal heterogeneity. At longer timescales, however, particles tend to lose memory of their initial structures, leading to the decaying heterogeneity. Notably, we observe that the time dependence of χ_4 closely mirrors the Pearson correlation results of the OP λ_1 from TCCA and TAE, indicating that our unsupervised protocol is highly effective in understanding strongly heterogeneous dynamics.

To further demonstrate the effectiveness of our unsupervised approach, we compared the performance of λ_1 obtained from TAE with conventional physics-based OPs and some supervised models. For consistency with Ref.[43], we chose to use a different propensity defined as the isoconfigurational average of particle absolute displacements, $\mathcal{D}(\delta t)$, rather than $\mathcal{C}_B(\delta t)$. As shown in Figure 8.5(a), TAE demonstrates much better predictive performance for the strong heterogeneous regime compared to the soft modes (SM) [254], which uses the mode participation fraction of each particle for low-frequency soft normal modes, and the Debye-Waller (DW) factor [264], which relies on the ground-truth dynamics up to $3/4$ of the initial value of the intermediate scattering function (i.e.,corresponds to time $\simeq 0.4$ for $T = 0.50$). And TAE's performance in the strong heterogeneity range is comparable to, or even slightly superior to, that of supervised models including graph neural networks (GNN), convolutional neural networks (CNN), and support vector machines (SVM). To our knowledge, this is the first time an unsupervised method has been shown to achieve performance comparable to supervised methods for the benchmark KA system.

Furthermore, as illustrated in Figure 8.5(b), we validate that the TAE OP λ_1, constructed across different temperatures (i.e., from $T = 0.44$ to $T = 0.56$), consistently demonstrates strong performance in correlating with dynamic heterogeneity over a broad time range (i.e., from $\sim \tau_\alpha^{BB}/3$ to $\sim 3\tau_\alpha^{BB}$). As the temperature is decreased, the TAE OP exhibits strong correlation with propensity over longer time scales. Additionally, we examine the quality of other TAE OPs in correlating with dynamic heterogeneity and how the lag time Δt affects their performance. As shown in Figure 8.6, we find that the leading OP, λ_1, consistently shows the strongest correlation with long-term propensity, while the second OP, λ_2, exhibits a slightly higher correlation with short-term dynamics over different temperatures. The other OPs, however, are far less informative. We also observe once lag time is set as $0.1 < \Delta t < \tau_\alpha^{BB}$, the performance of the OPs shows minimal variation across

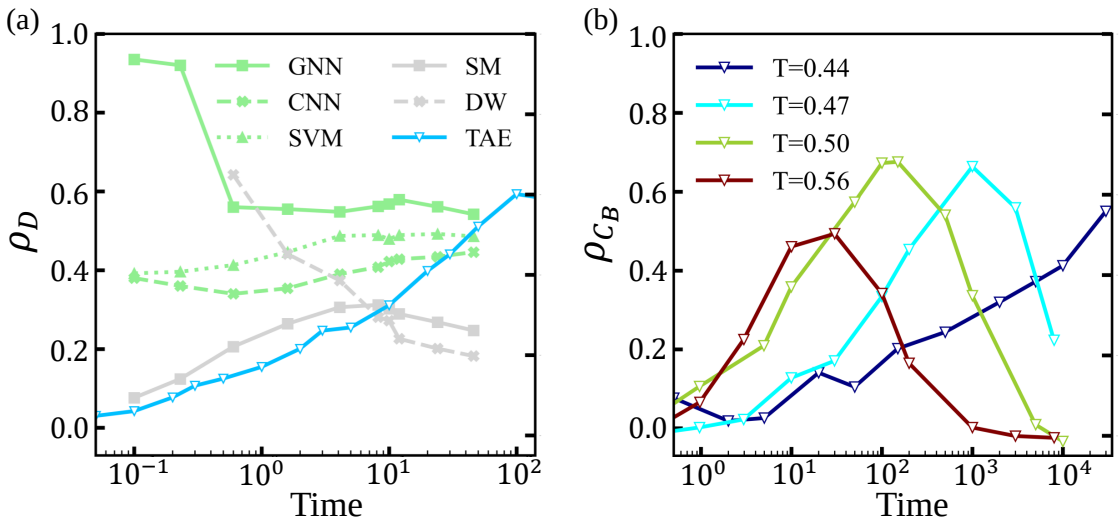

Figure 8.5: Effectiveness of the TAE OP in explaining dynamical heterogeneity. (a) Comparison of supervised ML methods and physics-based OPs from Ref. [43] with the OP λ_1 from TAE in predicting the isoconfigurational average of displacements $\mathcal{D}(\delta t)$ at temperature $T = 0.50$. Pearson correlation coefficients at different time points are shown. (b) Four TAE models are respectively constructed for four different temperatures, and the Pearson correlation coefficients between $\mathcal{C}_B(\delta t)$ and λ_1 are shown for each TAE model. This figure is reproduced from Qiu. *et al.*[45]

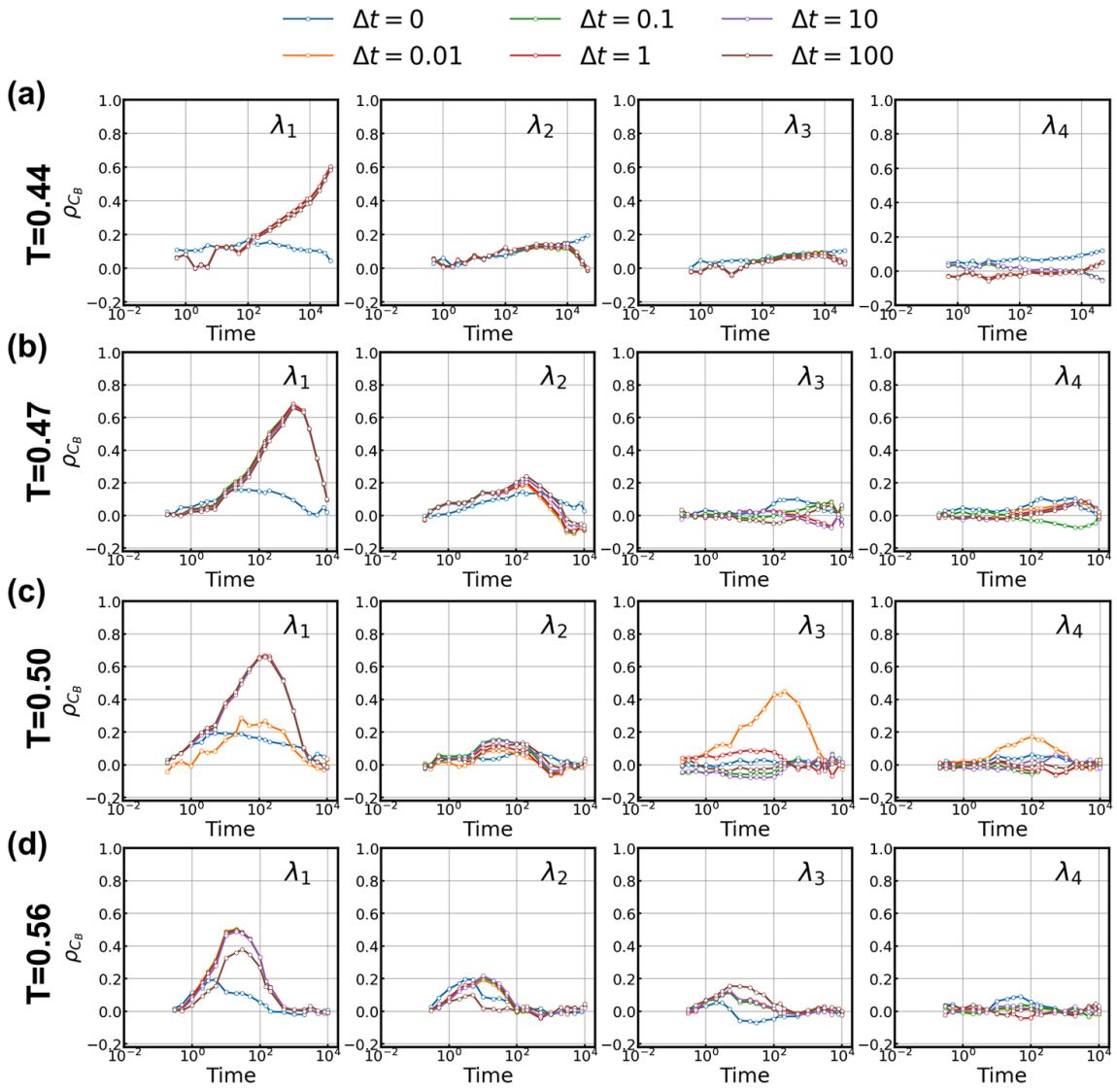

Figure 8.6: Effects of lag time and dimension of latent space on the quantify of OPs. Multiple independent TAE models are constructed with 4-dimensional latent space and various lag times $\Delta t = 0, 0.01, 0.1, 1, 10, 100$ by utilizing radial density descriptors derived from time-lagged configurations at different temperatures: (a) T=0.44, (b) T=0.47, (c) T=0.50 and (d) T=0.56. The correlations between four OPs and bond-breaking propensities are calculated and visualized, respectively. This figure is reproduced from Qiu. *et al.*[45]

different lag times and temperatures. These findings align with the TCCA singular value and OP correlation analysis, where the leading singular values decay much more slowly than the others, and OPs constructed with different lag times show strong correlations with each other.

8.5 Generalizing Predictive Power to Low-Temperatures

The computational cost of simulating supercooled liquids and calculating their propensities around relaxation timescale increases significantly as the temperature decreases. Therefore, it is crucial to investigate the generalization capabilities of the OPs constructed using our unsupervised protocol across different temperatures. We evaluate the transferability of the TAE, constructed at single temperature $T = 0.50$, to other temperatures. Specifically, after optimizing the parameters (i.e., weights and biases) of TAE with 2-dimensional bottleneck space, we then encode the radial descriptors of particles from equilibrium configurations at other temperatures, thereby determining the λ_1 distribution as a function of temperature. Notably, this process does not require additional simulations at different temperatures, only ensembles of configurations from equilibrium simulations.

Encoding multiple A-particles from 13 uniformly spaced state points (from $T = 0.44$ to $T = 0.56$) by transferred TAE, produces 13 distributions for λ_1. As shown in Figure 8.7(a), the mean value of λ_1 gradually increase with the decreasing of the temperatures, indicating the "slow" structural group becomes more dominant. However, the increase in λ_1 is linear and not dramatic and sensitive as in the traditional properties used to quantify the onset of dynamical arrest in supercooled liquids, for example, the peak value of non-Gaussian parameters α_m, shown for comparison. Additionally, the standard deviations of λ_1 are observed to progressively increase as the temperature decreases, demonstrating the structural heterogeneity captured by λ_1 grows with decreasing temperature, consistent with the trend of dynamic heterogeneity.

Furthermore, we quantify the ability of λ_1, encoded from configurations at four different temperatures by the transferred TAE, to predict dynamic propensity at the corresponding temperatures (see Figure 8.7(b)). The predictive performance of the transferred TAE at different temperatures

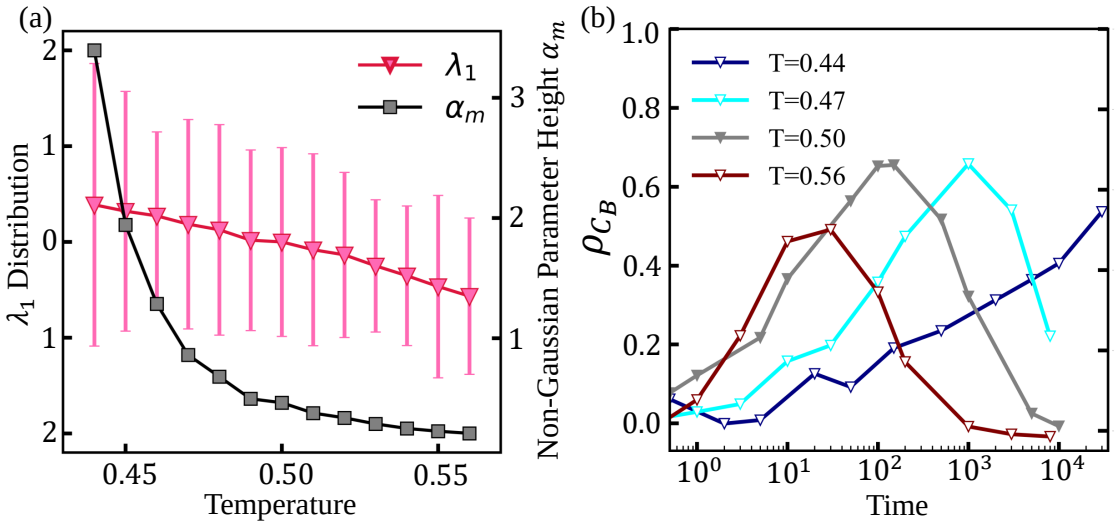

Figure 8.7: Transferability of TAE across different temperatures. The TAE constructed at $T = 0.50$ is applied to encode configurations from other temperatures to obtain λ_1 for those temperatures. (a) The distributions of λ_1 for encoded particles from different temperatures are shown in red bars. The heights α_m of the peaks of the non-Gaussian parameter are shown in black curve for comparison. (b) The correlation between λ_1 obtained from the transferred TAE and propensities is shown for different temperatures. This figure is reproduced from Qiu. *et al.*[45]

is found to be comparable to that of the TAE models specifically constructed for those temperatures (see Figure 8.5(b)). Therefore, the TAE exhibits great predictive capability at temperatures other than where its parameters were optimized, showcasing the excellent transferability of our unsupervised models and underscoring their potential for investigating dynamic heterogeneity at exceptionally low temperatures and over extended timescales.

8.6 Extracting Important Local Structures for Long-Time Dynamics

We have shown that the OP λ_1, constructed using our unsupervised models with radial descriptors, is highly effective in correlating structural differences with long-term dynamic heterogeneity across different temperatures. These unsupervised models are not only far more data-efficient but also significantly more interpretable than supervised models. In particular, TCCA is implemented through a very trivial linear combination of input descriptors, offering an good opportunity to uncover the critical information hidden in the static structure that influences prediction of long-term dynamics.

We characterize the importance of radial densities with different radii by using the TCCA linear combination coefficients of λ_1, scaled by the ensemble-averaged fluctuations of the corresponding densities. As shown in Figure 8.8, we find that radial density across all length scales plays a crucial role in the success of TCCA. Notably, the densities of surrounding B-particles contribute more than those of A-particles, especially at lower temperatures. The medium-range B-particle densities, even beyond the cutoff for potential energy, are particularly crucial. And the importance of radial descriptors remains largely consistent from $T = 0.44$ to $T = 0.56$.

Furthermore, we validate the importance of radial descriptors using TAE by constructing multiple TAEs, each restricted to descriptors from one of three regions: from 0 to the first minimum in the pair correlation function, between the first and second minima, and beyond the second minimum (see Figure S9). We observe that in general, the performance of the TAE OP constructed with a subset of descriptors is weaker than when all descriptors are used (see Figure 8.9). Specifically, when TAE is built using only the densities of surrounding A-particles, the performance of λ_1 is quite limited, whereas using only the densities of surrounding B-particles yields stronger results.

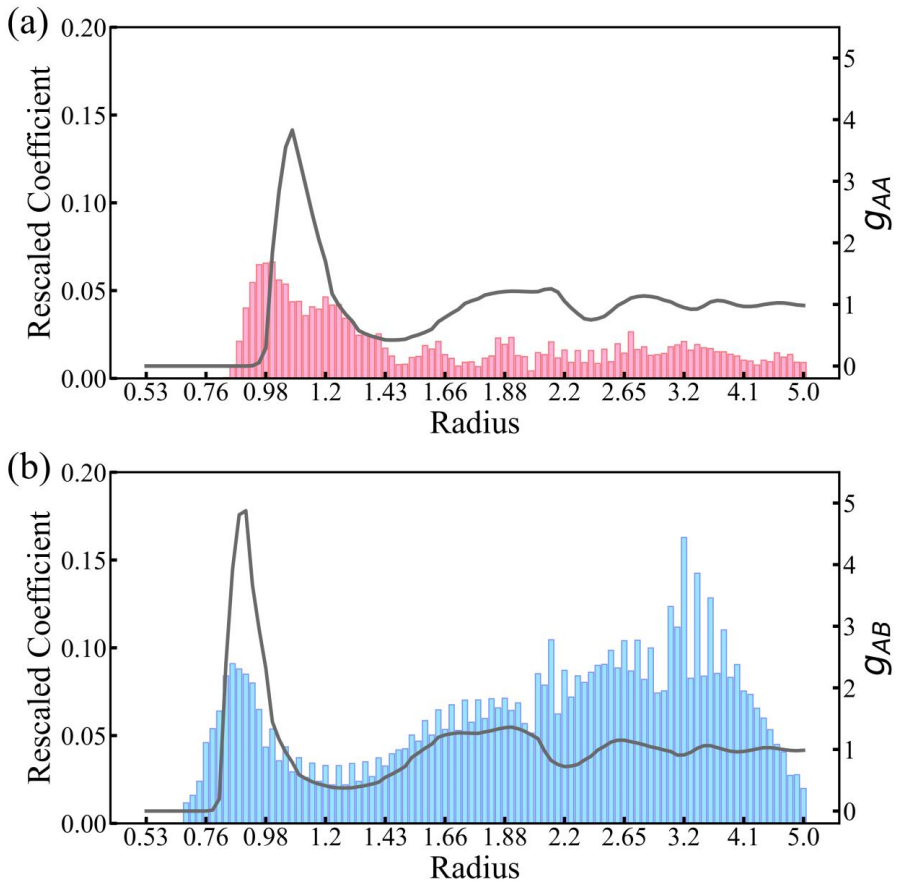

Figure 8.8: Importance of density at varying radii in constructing the OP λ_1. The rescaled linear transformation coefficients for local radial densities at different radii in the construction of λ_1 using TCCA are displayed for (a) A-particles and (b) B-particles, respectively. Each absolute value of coefficient is rescaled by the ensemble-averaged fluctuation of the corresponding density. The gray curves represent the pair correlation functions over different radii. This figure is reproduced from Qiu. *et al.*[45]

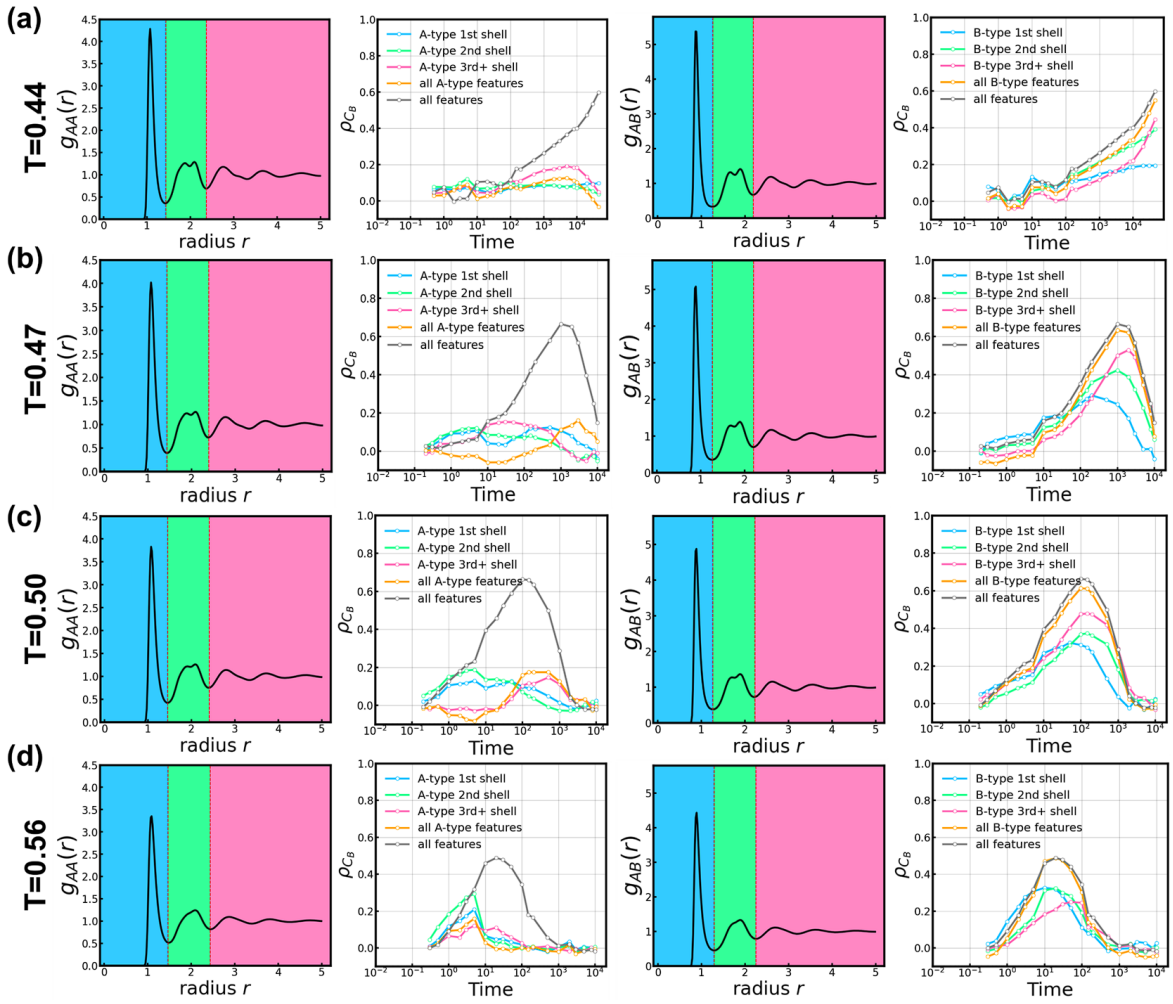

Figure 8.9: Importance of radial descriptors within different shells to capture long-time dynamics heterogeneity. A-particles from equilibrium configurations at temperatures (a) $T = 0.44$, (b) $T = 0.47$, (c) $T = 0.50$, and (d) $T = 0.56$ are embedded using 200-dimensional radial descriptors. These descriptors are subsampled based on reference particle types (left two columns for A-particles, right two columns for B-particles) and shells (regions colored blue, green, and pink). Independent TAE models are constructed with different input descriptors. For each model, the correlations between propensities and the OP λ_1 are calculated and visualized over time. The different curves for correlation are obtained using descriptors from same colored regions in the pair correlation function. This figure is reproduced from Qiu. *et al.*[45]

And it is also noticed that medium-range descriptors (i.e., 3rd+ shell) produce OPs with a stronger correlation to long-term dynamics. Consequently, the outcomes from two unsupervised models are highly consistent with each other. Our findings also align well with previous studies that emphasize the importance of medium-range descriptors, showing that averaging descriptors with neighboring particles can improve model performance [43, 44, 260].

Additionally, based on previous analysis of TCCA singular values, we demonstrate that OPs constructed from angular descriptors show significantly weaker time correlations and may be insufficient for predicting long-term dynamics. We verify this by constructing OPs using both angular descriptors and a combination of radial and angular descriptors with our unsupervised models. Our findings support that, for the KA system, the three-dimensional angles of neighboring particles are not useful for correlating with long-term dynamics. This aligns with a key conclusion from the implementation of GNNs, which is that angles are only useful for short-term predictions, while distances alone are sufficient for studying dynamics in the glassy regime [43].

8.7 Markovian Embedding of Non-Markovian Dynamics

We employ TAE to construct a non-linear propagator that maps the features of a particle at time t to the features of the same particle at time $t + \Delta t$. In this approach, we ignore the historical influence before time t and make the strong assumption that the evolution of dynamics in feature space is Markovian. However, the dynamics of a single particle are expected to have a strong memory effect due to its active interactions with multiple surrounding particles. If we solely utilize the coordinates and velocities of the probe particle to construct the feature space, the dynamics in the feature space should be non-Markovian. This explains why the displacement during a short lag time shows no correlation with long-time propensity at all. The strong memory effect prevents the predictability of long-term dynamics from short-time displacement alone.

In our implementation of TAE, we address this issue by employing the Markovian embedding of non-Markovian dynamics approach.[279, 280] This involves incorporating a substantial amount of additional degrees of freedom, achieved through the consideration of local surrounding structures. The resulting high-dimensional feature space takes into account interactions between the

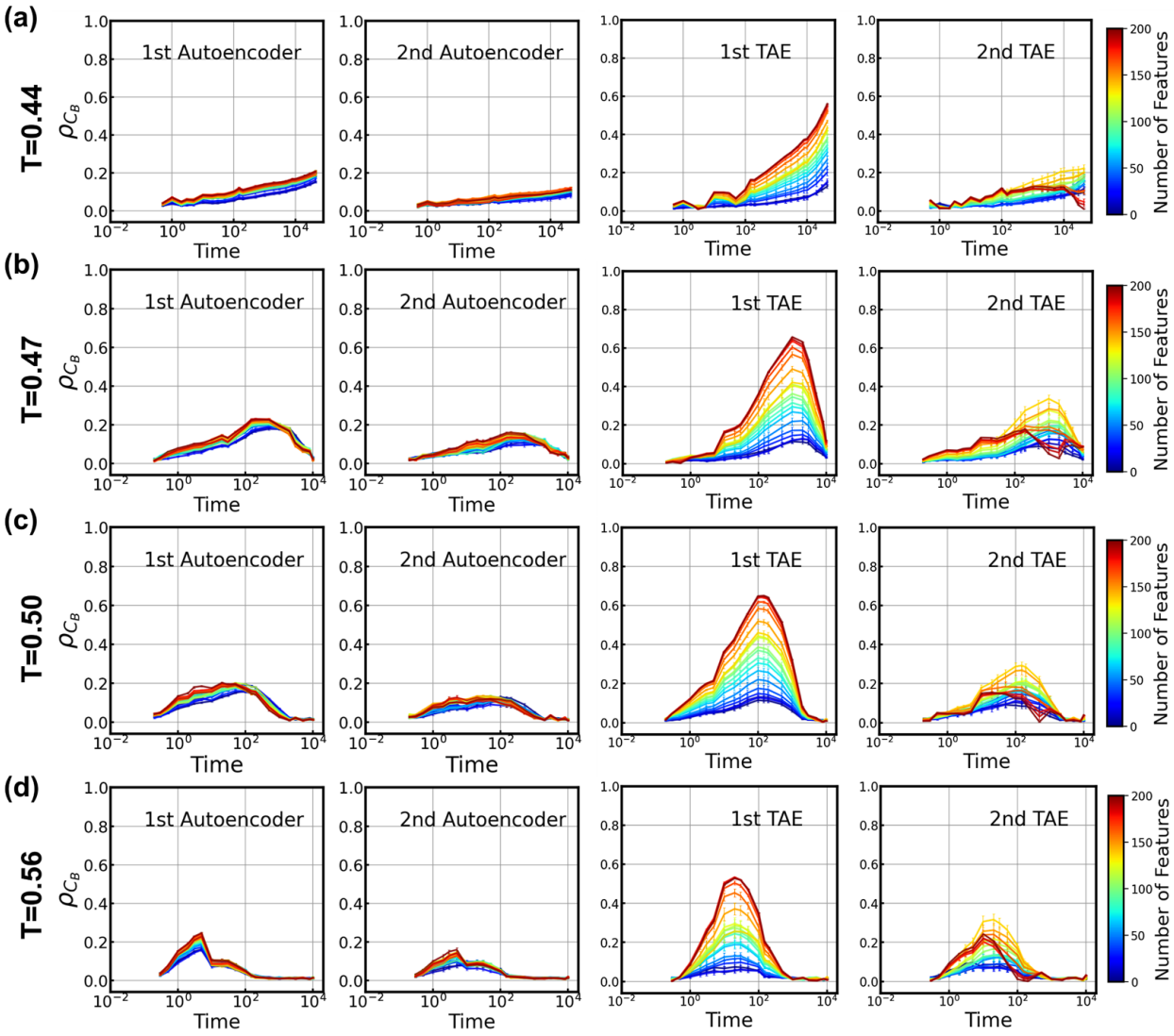

Figure 8.10: Impact of input feature dimension on the quality of order parameters. The 200-dimensional radial density features are randomly subsampled into different dimensions and used to train both AE and TAE models at four respective temperatures: (a) $T = 0.44$, (b) $T = 0.47$, (c) $T = 0.50$, and (d) $T = 0.56$. The number of hidden neurons is consistently set as five times the input feature dimensions for all the models, and the lag time for TAE is fixed at $\Delta t = 0.1$. The subsamplings are repeated twenty times for a specific dimension of input features, and the error bars are estimated accordingly. This figure is reproduced from Qiu. *et al.*[45]

particle and its surroundings, allowing for a better approximation of dynamics as a Markovian process. But since we consider each particle individually and cannot update the evolution of features due to the motions of surrounding particles, the accurate propagation time of dynamics is limited.

To substantiate the above justifications, we evaluate the quality of order parameters obtained from TAE models constructed with varying numbers of features. We fix the TAE lag time at $\Delta t = 0.1$ and randomly sub-sample the 200-dimensional radial features to train TAE models. The number of hidden neurons is consistently set as five times the number of input features. We repeat this process twenty times for a specific number of features. As shown in Figure 8.10, we observe a substantial decrease in the prediction accuracy of TAE order parameters when the input features are truncated. In contrast, the prediction accuracy of order parameters from the $\Delta t = 0$ Autoencoder does not change significantly. Despite the identification of the radial features for B-particle as important, utilizing only a single shell of these features lead to order parameters with significantly reduced predictive power (Figure 8.9). These results demonstrate that expanding the feature space enhances the predictability of long-time dynamics for a single particle using a short lag time, aligning with our concept of Markovian embedding of non-Markovian dynamics.

8.8 Conclusion

Overall, our results from the unsupervised ML models suggest that structural heterogeneity in supercooled liquids could be effectively captured and used to explain long-term dynamic slowdown and heterogeneity in a more data-efficient and interpretable manner. In contrast to supervised models, which are "black boxes" trained with targeted labels from extensive simulations, and to typical unsupervised models that use numerous local structural descriptors to construct OP capturing the most variance, our unsupervised protocol leverages information from extremely short equilibrium simulations to identify OP with the greatest short-time correlation. This one-dimensional OP is shown to be highly effective in representing variations in local structure that correlates with strongly heterogeneous dynamics. An intriguing conclusion thereby drawn is that the essence of long-term dynamics is already encoded in the fluctuating structural behaviors observed at short times.

While substantial efforts have been made to establish clear connections between structure and dynamics—such as developing new ML networks, incorporating additional physical properties, and creating novel structural descriptors—we take a different approach. By applying simple and efficient linear or nonlinear transformations to commonly used descriptors, we achieve remarkable performance across various temperatures, which is even comparable to some supervised models. Note that our models use only trivial descriptors for individual particle as input, without any averaging or coarse-graining. While averaging structural descriptors with neighboring particles can enhance model performance, the hyperparameters for averaging are either empirical [44, 260] or optimized through supervised methods [43]. The straightforward setup of our unsupervised protocol facilitates the implementation and helps identify the most important structures, with a particular emphasis on medium-range structures and particle types. Meanwhile, the robust transferability of our models opens up opportunities for studies of dynamic heterogeneities in lower-temperature regions using data that are more easily obtained at higher temperatures.

It is also important to remember that structural and dynamic heterogeneity can vary from system to system [263]. Our protocol provides a potential approach for exploring structure-dynamics relationships, but further exploration is needed for future. In particular, future work could generalize the approaches to a wider range of glass-forming materials and systems, as well as investigate the relationship between the identified OP and other relevant physical quantities in glassy systems.

Chapter 9

Conclusions and Future Perspectives

Investigating complex dynamics in multi-body systems is of great importance for fields such as chemistry, biology, and materials science. However, due to the unique characteristics of multi-body systems, such as complex energy landscapes with large local fluctuations, heterogeneous dynamics across various timescales, and diverse parallel transition pathways, etc., investigating their microscopic dynamics and mechanisms at high temporal (picoseconds) and spatial (atomic-level) resolution is highly challenging. MD simulation is a powerful tool for revealing the dynamics of multi-body systems with high resolution, but its direct application to such systems remains highly complex. On one hand, the short time steps required for MD simulations make it extremely difficult to sample slow processes that occur on timescales of milliseconds or longer, creating a timescale gap that hinders direct comparison with the timescales from experimental observations. On the other hand, the trajectories generated from MD simulations are high-dimensional sequence data, making them difficult to interpret, particularly when trying to identify transition pathways and understand the mechanisms. To overcome these challenges, the MSMs framework has been developed to bridge the timescale gap by integrating ensembles of short MD trajectories to infer long-term dynamics and to interpret complex dynamical processes through Markovian jumps between few coarse-grained metastable states. However, it should be noted that to construct a MSM with prediction power, a sufficiently long lag time must be employed to ensure that intra-state dynamics are relaxed and inter-state transitions are Markovian. Since the feasible lag time is constrained by the length of the MD trajectory, MSMs with only a few metastable states typically still require prohibitively long simulations. Increasing the number of states can help shorten the lag time, but it also result in numerous parallel pathways and leads to difficulties in understanding the underlying mechanisms. These open questions lead to a dilemma: it is difficult to simultaneously

capture the heterogeneous long-time dynamics in multi-body systems accurately and intuitively understand their molecular mechanisms.

In this thesis, we developed machine learning algorithms and non-Markovian dynamics models to address the aforementioned challenges and dilemmas. Unlike conventional MSMs, which model dynamics as a Markovian process using the first-order master equation, we developed the IGME approach. This approach models dynamics using the generalized master equation, incorporating time-integration of memory kernels to capture non-Markovian behaviors. Since the relaxation time for memory kernels is much shorter than the Markovian lag time, the IGME model allowed us to achieve greater accuracy in modeling long-term dynamics using shorter MD segments. Meanwhile, we developed an effective algorithm called LPC to identify explainable metastable path channels in multi-body systems. This algorithm could group multiple kinetic pathways with comparable fluxes, identified through the integration of transition path theory and a hundred/thousand-state MSM, providing a much clearer interpretation. Additionally, we also designed an information bottleneck approach, which employed the varitional autoencoder deep neural network to construct MSMs. Different from the traditional MSM construction pipeline, this approach can coarse-grain MD conformations into a few metastable states using a set of non-linear basis functions, leading to improved assignments for MD conformations and thereby significantly reducing the Markovian lag time for MSMs. Through the development and implementation of the machine learning techniques and algorithms mentioned above, along with the non-Markovian dynamics models, we expect to better capture long-term heterogeneous dynamics and gain deeper insights into the mechanisms of multi-body systems.

In the future, significant effort will be needed to further improve the workflow for MSM construction and dynamic modeling. For instance, the generalized Langevin equation could replace the generalized master equation to model projected dynamics in a low-dimensional space with higher spatial and temporal resolution. However, additional efforts may be paid to determine the continuous friction kernels and stochastic noise while preserving the fluctuation-dissipation theorem. Meanwhile, high-order or variable-order Markov models could be developed to more accurately capture non-Markovian dynamics and provide better interpretations of the physical insights related

to memory kernels. Additionally, improvements in algorithms for dimension reduction, clustering, and kinetic lumping could optimize the MSM construction protocol, achieving better the time-scale separation between intra-state and inter-state dynamics and thereby shortening the Markovian lag time. In particular, beyond the commonly used VAMP-score-based optimization and metastability maximum optimization schemes, methods that establish metastable state models with reduced memory effects would be highly beneficial. To enhance the interpretation of mechanisms, pathway clustering algorithms are still actively being developed. However, determining the optimal method for quantifying similarities between pathways remains an open question. These path clustering algorithms are crucial and highly relevant for complex multi-body systems, such as protein-ligand complexes.

With the developed techniques, we further applied them to investigate the dynamics in real multi-body systems, including PPIs and chromatin folding. Specifically, in the Chapter 6, we illustrate how the IGME model and advanced feature embedding methods can aid in elucidating the hetergeneous conformations and dynamics of the KRAS-VHL encounter complex. We identified non-canonical PPI interfaces that serve as valuable baselines for future design of PROTAC linkers. Meanwhile, we came up with a pipeline incorporating various metrics to shortlist the most suitable metastable PPI interfaces for PROTAC design, which can be easily generalized to evaluate other systems. In the Chapter 7, we thoroughly examine the impact of DNA linker length and nucleosome condensate environment on the folding dynamics and mechanisms of the tetra-nucleosome. We observed notable differences between tetra-nucleosome folding and typical protein folding, with the former involving many more parallel pathways with similar fluxes. Our LPC algorithm was instrumental in identifying three distinct folding pathways, each corresponding to different mechanisms, and in quantifying their relative dominance. Using the IGME models, we further reveal that both longer DNA linkers and the phase separation of nucleosome condensate environment can destabilize the tetra-nucleosome and promote unfolding rates. Additionally, our dynamic modeling provides insights into the differences between *in vivo* and *in vitro* observations.

For future studies, it remains unclear which computational approach is best for enhancing the heterobiofunctional molecule PROTAC design. we anticipate that the simulation and dynamics

modeling protocol used in our investigation of the KRAS-VHL encounter complex can be extended to other PPI systems, potentially aiding in the design of PROTACs and molecular glues. Additionally, in our study of the tetra-nucleosome within the nucleosome condensate, we did not explore the dynamics of condensate aggregation, which would be a valuable area for future research. In particular, implementing graph neural networks, such as Graph-VAMPnet, could effectively identify metastable intermediate aggregation patterns and elucidate the aggregation mechanisms of nucleosome condensates. Research into the dynamics of bio-condensate multi-body systems is becoming increasingly popular and significant.

Moreover, in Chapter 8, we studied the dynamics of another typical multi-body system: supercooled liquids. Specifically, we present an unsupervised machine learning approach to explore the connection between long-time dynamical heterogeneity in supercooled liquids and their local structures. Different from typical supervised machine learning approaches that rely on training with large datasets where dynamical heterogeneity is pre-modeled from long MD simulations, our unsupervised method identifies order parameters with the highest short-time lagged correlations from high-dimensional structural descriptors. And the order parameters were demonstrated to effectively explain the long-term dynamical heterogeneity of supercooled liquids, achieving prediction accuracy comparable to state-of-the-art supervised machine learning methods. Notably, without additional training, our order parameters exhibited strong generalization ability in low-temperature regions, indicating the success to localize key structural features across varying degrees of supercooling. Further analysis of input feature importance highlights the significance of medium-range features, extending even beyond the potential energy cutoff.

With the advancement of machine learning techniques, predicting dynamical heterogeneities at various timescales from initial structures has become an active and highly pursued research direction in the study of supercooled liquids. In the future, there are four key directions to go. First, improving structural descriptors for disordered amorphous glasses is crucial. Since the significance of different features is difficult to be evaluated in advance and is highly system-dependent, it's important to develop a variety of descriptors that can comprehensively capture structural details from multiple perspectives. Second, applying more advanced supervised machine learning models,

such as different newly developed neural network architectures, remains promising for achieving greater accuracy in dynamics prediction. Third, developing unsupervised methods with higher performance can significantly enhance data efficiency and help understand the importance of different features, while adaptively using these identified features to build supervised models could be promising. Fourth, since dynamical heterogeneities are highly system-dependent, it is crucial to test different algorithms or models across various systems while also creating more benchmark systems and datasets.

Bibliography

(1) Plattner, N.; Doerr, S.; De Fabritiis, G.; Noé, F. Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nature chemistry* **2017**, *9*, 1005–1011.

(2) Scott, D. E.; Bayly, A. R.; Abell, C.; Skidmore, J. Small molecules, big targets: drug discovery faces the protein–protein interaction challenge. *Nature Reviews Drug Discovery* **2016**, *15*, 533–550.

(3) Qiu, Y.; Wiewiora, R. P.; Izaguirre, J. A.; Xu, H.; Sherman, W.; Tang, W.; Huang, X. Non-Markovian Dynamic Models Identify Non-Canonical KRAS-VHL Encounter Complex Conformations for Novel PROTAC Design. *JACS Au* **2024**.

(4) King, N. P.; Sheffler, W.; Sawaya, M. R.; Vollmar, B. S.; Sumida, J. P.; André, I.; Gonen, T.; Yeates, T. O.; Baker, D. Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* **2012**, *336*, 1171–1174.

(5) Dixon, T.; MacPherson, D.; Mostofian, B.; Dauzhenka, T.; Lotz, S.; McGee, D.; Shechter, S.; Shrestha, U. R.; Wiewiora, R.; McDargh, Z. A., et al. Predicting the structural basis of targeted protein degradation by integrating molecular dynamics simulations with structural mass spectrometry. *Nature communications* **2022**, *13*, 5884.

(6) Silva, D.-A.; Bowman, G. R.; Sosa-Peinado, A.; Huang, X. A role for both conformational selection and induced fit in ligand binding by the LAO protein. *PLoS computational biology* **2011**, *7*, e1002054.

(7) Bucher, D.; Grant, B. J.; Markwick, P. R.; McCammon, J. A. Accessing a hidden conformation of the maltose binding protein using accelerated molecular dynamics. *PLoS computational biology* **2011**, *7*, e1002034.

(8) Yu, H. S.; Gao, C.; Lupyan, D.; Wu, Y.; Kimura, T.; Wu, C.; Jacobson, L.; Harder, E.; Abel, R.; Wang, L. Toward atomistic modeling of irreversible covalent inhibitor binding kinetics. *Journal of Chemical Information and Modeling* **2019**, *59*, 3955–3967.

(9) Hyman, A. A.; Weber, C. A.; Jülicher, F. Liquid-liquid phase separation in biology. *Annual review of cell and developmental biology* **2014**, *30*, 39–58.

(10) Alberti, S.; Gladfelter, A.; Mittag, T. Considerations and challenges in studying liquid-liquid phase separation and biomolecular condensates. *Cell* **2019**, *176*, 419–434.

(11) Ediger, M. D.; Angell, C. A.; Nagel, S. R. Supercooled liquids and glasses. *The journal of physical chemistry* **1996**, *100*, 13200–13212.

(12) Debenedetti, P. G.; Stillinger, F. H. Supercooled liquids and the glass transition. *Nature* **2001**, *410*, 259–267.

(13) Liu, B.; Qiu, Y.; Goonetilleke, E. C.; Huang, X. Kinetic network models to study molecular self-assembly in the wake of machine learning. *MRS Bulletin* **2022**, *47*, 958–966.

(14) Xia, C.; Qiu, Y.; Xia, Y.; Zhu, P.; King, G.; Zhang, X.; Wu, Z.; Kim, J. Y.; Cullen, D. A.; Zheng, D., et al. General synthesis of single-atom catalysts with high metal loading using graphene quantum dots. *Nature chemistry* **2021**, *13*, 887–894.

(15) Conti, S.; Cecchini, M. Predicting molecular self-assembly at surfaces: a statistical thermodynamics and modeling approach. *Physical Chemistry Chemical Physics* **2016**, *18*, 31480–31493.

(16) Zeng, X.; Li, B.; Qiao, Q.; Zhu, L.; Lu, Z.-Y.; Huang, X. Elucidating dominant pathways of the nano-particle self-assembly process. *Physical Chemistry Chemical Physics* **2016**, *18*, 23494–23499.

(17) Zheng, X.; Zhu, L.; Zeng, X.; Meng, L.; Zhang, L.; Wang, D.; Huang, X. Kinetics-controlled amphiphile self-assembly processes. *The journal of physical chemistry letters* **2017**, *8*, 1798–1803.

(18) Zhu, L.; Sheong, F. K.; Zeng, X.; Huang, X. Elucidation of the conformational dynamics of multi-body systems by construction of Markov state models. *Physical Chemistry Chemical Physics* **2016**, *18*, 30228–30235.

(19) Sheong, F. K.; Silva, D.-A.; Meng, L.; Zhao, Y.; Huang, X. Automatic state partitioning for multibody systems (APM): an efficient algorithm for constructing Markov state models to elucidate conformational dynamics of multibody systems. *Journal of chemical theory and computation* **2015**, *11*, 17–27.

(20) Ha, T. Single-molecule fluorescence resonance energy transfer. *Methods* **2001**, *25*, 78–86.

(21) Jackman, L., *Dynamic nuclear magnetic resonance spectroscopy*; Elsevier: 2012.

(22) Callender, R.; Dyer, R. B. Probing protein dynamics using temperature jump relaxation spectroscopy. *Current opinion in structural biology* **2002**, *12*, 628–633.

(23) Barducci, A.; Bonomi, M.; Parrinello, M. Metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1*, 826–843.

(24) Fukunishi, H.; Watanabe, O.; Takada, S. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *The Journal of chemical physics* **2002**, *116*, 9058–9067.

(25) Bowman, G. R.; Ensign, D. L.; Pande, V. S. Enhanced modeling via network theory: Adaptive sampling of Markov state models. *Journal of chemical theory and computation* **2010**, *6*, 787–794.

(26) Zuckerman, D. M.; Chong, L. T. Weighted ensemble simulation: review of methodology, applications, and software. *Annual review of biophysics* **2017**, *46*, 43–57.

(27) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. Markov models of molecular kinetics: Generation and validation. *The Journal of chemical physics* **2011**, *134*.

(28) Wang, W.; Cao, S.; Zhu, L.; Huang, X. Constructing Markov State Models to elucidate the functional conformational changes of complex biomolecules. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2018**, *8*, e1343.

(29) Konovalov, K. A.; Unarta, I. C.; Cao, S.; Goonetilleke, E. C.; Huang, X. Markov state models to study the functional dynamics of proteins in the wake of machine learning. *JACS Au* **2021**, *1*, 1330–1341.

(30) Cao, S.; Montoya-Castillo, A.; Wang, W.; Markland, T. E.; Huang, X. On the advantages of exploiting memory in Markov state models for biomolecular dynamics. *The Journal of Chemical Physics* **2020**, *153*.

(31) Cao, S.; Qiu, Y.; Kalin, M. L.; Huang, X. Integrative generalized master equation: A method to study long-timescale biomolecular dynamics via the integrals of memory kernels. *The Journal of Chemical Physics* **2023**, *159*.

(32) Yik, A. K.-h.; Qiu, Y.; Unarta, I. C.; Cao, S.; Huang, X. In *A Practical Guide to Recent Advances in Multiscale Modeling and Simulation of Biomolecules*; AIP Publishing LLC Melville, New York: 2023, pp 10–1.

(33) Hegger, R.; Stock, G. Multidimensional Langevin modeling of biomolecular dynamics. *The Journal of chemical physics* **2009**, *130*.

(34) Mardt, A.; Pasquali, L.; Wu, H.; Noé, F. VAMPnets for deep learning of molecular kinetics. *Nature communications* **2018**, *9*, 5.

(35) Wang, D.; Qiu, Y.; Beyerle, E. R.; Huang, X.; Tiwary, P. Information Bottleneck Approach for Markov Model Construction. *Journal of Chemical Theory and Computation* **2024**.

(36) Mehdi, S.; Smith, Z.; Herron, L.; Zou, Z.; Tiwary, P. Enhanced sampling with machine learning. *Annual Review of Physical Chemistry* **2024**, *75*.

(37) Chen, W.; Sidky, H.; Ferguson, A. L. Nonlinear discovery of slow molecular modes using state-free reversible VAMPnets. *The Journal of chemical physics* **2019**, *150*.

(38) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. MSMBuilder2: modeling conformational dynamics on the picosecond to millisecond scale. *Journal of chemical theory and computation* **2011**, *7*, 3412–3419.

(39) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. PyEMMA 2: A software package for estimation, validation, and analysis of Markov models. *Journal of chemical theory and computation* **2015**, *11*, 5525–5542.

(40) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proceedings of the National Academy of Sciences* **2009**, *106*, 19011–19016.

(41) Vanden-Eijnden, E. et al. Towards a theory of transition paths. *Journal of statistical physics* **2006**, *123*, 503–523.

(42) Jung, G.; Alkemade, R. M.; Bapst, V.; Coslovich, D.; Filion, L.; Landes, F. P.; Liu, A.; Pezzicoli, F. S.; Shiba, H.; Volpe, G., et al. Roadmap on machine learning glassy liquids. *arXiv preprint arXiv:2311.14752* **2023**.

(43) Bapst, V.; Keck, T.; Grabska-Barwińska, A.; Donner, C.; Cubuk, E. D.; Schoenholz, S. S.; Obika, A.; Nelson, A. W.; Back, T.; Hassabis, D., et al. Unveiling the predictive power of static structure in glassy systems. *Nature physics* **2020**, *16*, 448–454.

(44) Boattini, E.; Smallenburg, F.; Filion, L. Averaging local structure to predict the dynamic propensity in supercooled liquids. *Physical Review Letters* **2021**, *127*, 088007.

(45) Qiu, Y.; Jang, I.; Huang, X.; Yethiraj, A. Unsupervised machine learning for supercooled liquids. *arXiv preprint arXiv:2404.04473* **2024**.

(46) Hummer, G.; Szabo, A. Optimal dimensionality reduction of multistate kinetic and Markov-state models. *The Journal of Physical Chemistry B* **2015**, *119*, 9029–9037.

(47) Noé, F.; Nuske, F. A variational approach to modeling slow processes in stochastic dynamical systems. *Multiscale Modeling & Simulation* **2013**, *11*, 635–655.

(48) Qiu, Y.; O'Connor, M. S.; Xue, M.; Liu, B.; Huang, X. An Efficient Path Classification Algorithm Based on Variational Autoencoder to Identify Metastable Path Channels for Complex Conformational Changes. *Journal of Chemical Theory and Computation* **2023**, *19*, 4728–4742.

(49) Zwanzig, R., *Nonequilibrium statistical mechanics*; Oxford university press: 2001.

(50) Trendelkamp-Schroer, B.; Wu, H.; Paul, F.; Noé, F. Estimation and uncertainty of reversible Markov models. *The Journal of chemical physics* **2015**, *143*.

(51) Nuske, F.; Keller, B. G.; Pérez-Hernández, G.; Mey, A. S.; Noé, F. Variational approach to molecular kinetics. *Journal of chemical theory and computation* **2014**, *10*, 1739–1752.

(52) Chodera, J. D.; Noé, F. Markov state models of biomolecular conformational dynamics. *Current opinion in structural biology* **2014**, *25*, 135–144.

(53) Wu, H.; Noé, F. Variational approach for learning Markov processes from time series data. *Journal of Nonlinear Science* **2020**, *30*, 23–66.

(54) Golub, G. H.; Hoffman, A.; Stewart, G. W. A generalization of the Eckart-Young-Mirsky matrix approximation theorem. *Linear Algebra and its applications* **1987**, *88*, 317–327.

(55) Wehmeyer, C.; Noé, F. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *The Journal of chemical physics* **2018**, *148*.

(56) Meng, L.; Sheong, F. K.; Zeng, X.; Zhu, L.; Huang, X. Path lumping: an efficient algorithm to identify metastable path channels for conformational dynamics of multi-body systems. *The Journal of Chemical Physics* **2017**, *147*, 044112.

(57) Woolfson, M. M., *An introduction to X-ray crystallography*; Cambridge University Press: 1997.

(58) Benjin, X.; Ling, L. Developments, applications, and prospects of cryo-electron microscopy. *Protein Science* **2020**, *29*, 872–882.

(59) Perez, D.; Uberuaga, B. P.; Shim, Y.; Amar, J. G.; Voter, A. F. Accelerated molecular dynamics methods: introduction and recent developments. *Annual Reports in computational chemistry* **2009**, *5*, 79–98.

(60) Ingólfsson, H. I.; Lopez, C. A.; Uusitalo, J. J.; de Jong, D. H.; Gopal, S. M.; Periole, X.; Marrink, S. J. The power of coarse graining in biomolecular simulations. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2014**, *4*, 225–248.

(61) Noid, W.; Szukalo, R. J.; Kidder, K. M.; Lesniewski, M. C. Rigorous Progress in Coarse-Graining. *Annual Review of Physical Chemistry* **2024**, *75*, 21–45.

(62) Chaudhury, S.; Lyskov, S.; Gray, J. J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **2010**, *26*, 689–691.

(63) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A., et al. Highly accurate protein structure prediction with AlphaFold. *nature* **2021**, *596*, 583–589.

(64) Evans, R.; O'Neill, M.; Pritzel, A.; Antropova, N.; Senior, A.; Green, T.; Žídek, A.; Bates, R.; Blackwell, S.; Yim, J., et al. Protein complex prediction with AlphaFold-Multimer. *biorxiv* **2021**, 2021–10.

(65) Herron, L.; Mondal, K.; Schneekloth, J. S.; Tiwary, P. Inferring phase transitions and critical exponents from limited observations with Thermodynamic Maps. *arXiv preprint arXiv:2308.14885* **2023**.

(66) Zheng, S.; He, J.; Liu, C.; Shi, Y.; Lu, Z.; Feng, W.; Ju, F.; Wang, J.; Zhu, J.; Min, Y., et al. Predicting equilibrium distributions for molecular systems with deep learning. *Nature Machine Intelligence* **2024**, 1–10.

(67) Jung, G.; Biroli, G.; Berthier, L. Normalizing flows as an enhanced sampling method for atomistic supercooled liquids. *Machine Learning: Science and Technology* **2024**.

(68) Xi, B.; Liu, Z.; Raghavachari, M.; Xia, C. H.; Zhang, L. In *Proceedings of the 13th international conference on World Wide Web*, 2004, pp 287–296.

(69) Weinan, E.; Ren, W.; Vanden-Eijnden, E. String method for the study of rare events. *Physical Review B* **2002**, *66*, 052301.

(70) Coslovich, D.; Jack, R. L.; Paret, J. Dimensionality reduction of local structure in glassy binary mixtures. *The Journal of Chemical Physics* **2022**, *157*.

(71) Weng, J.; Yang, M.; Wang, W.; Xu, X.; Tian, Z. Revealing thermodynamics and kinetics of lipid self-assembly by Markov state model analysis. *Journal of the American Chemical Society* **2020**, *142*, 21344–21352.

(72) Ghorbani, M.; Prasad, S.; Klauda, J. B.; Brooks, B. R. GraphVAMPNet, using graph neural networks and variational approach to Markov processes for dynamical modeling of biomolecules. *The Journal of Chemical Physics* **2022**, *156*.

(73) Pezzicoli, F. S.; Charpiat, G.; Landes, F. P. Rotation-equivariant graph neural networks for learning glassy liquids representations. *SciPost Physics* **2024**, *16*, 136.

(74) Diez, G.; Nagel, D.; Stock, G. Correlation-based feature selection to identify functional dynamics in proteins. *Journal of Chemical Theory and Computation* **2022**, *18*, 5079–5088.

(75) Litzinger, F.; Boninsegna, L.; Wu, H.; Nüske, F.; Patel, R.; Baraniuk, R.; Noé, F.; Clementi, C. Rapid calculation of molecular kinetics using compressed sensing. *Journal of Chemical Theory and Computation* **2018**, *14*, 2771–2783.

(76) Noé, F.; Clementi, C. Kinetic distance and kinetic maps from molecular dynamics simulation. *Journal of chemical theory and computation* **2015**, *11*, 5002–5011.

(77) Schwantes, C. R.; Pande, V. S. Modeling molecular kinetics with tICA and the kernel trick. *Journal of chemical theory and computation* **2015**, *11*, 600–608.

(78) Bonati, L.; Piccini, G.; Parrinello, M. Deep learning the slow modes for rare events sampling. *Proceedings of the National Academy of Sciences* **2021**, *118*, e2113533118.

(79) Hernández, C. X.; Wayment-Steele, H. K.; Sultan, M. M.; Husic, B. E.; Pande, V. S. Variational encoding of complex dynamics. *Physical Review E* **2018**, *97*, 062412.

(80) Wang, Y.; Ribeiro, J. M. L.; Tiwary, P. Past–future information bottleneck for sampling molecular reaction coordinate simultaneously with thermodynamics and kinetics. *Nature communications* **2019**, *10*, 3573.

(81) Wang, D.; Tiwary, P. State predictive information bottleneck. *The Journal of Chemical Physics* **2021**, *154*.

(82) Rohrdanz, M. A.; Zheng, W.; Maggioni, M.; Clementi, C. Determination of reaction coordinates via locally scaled diffusion map. *The Journal of chemical physics* **2011**, *134*.

(83) Wu, H.; Noé, F. Reaction coordinate flows for model reduction of molecular kinetics. *The Journal of Chemical Physics* **2024**, *160*.

(84) Kitao, A.; Hirata, F.; Gō, N. The effects of solvent on the conformation and the collective motions of protein: normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum. *Chemical physics* **1991**, *158*, 447–472.

(85) Bowman, G. R.; Huang, X.; Pande, V. S. Using generalized ensemble simulations and Markov state models to identify conformational states. *Methods* **2009**, *49*, 197–201.

(86) Lemke, O.; Keller, B. G. Density-based cluster algorithms for the identification of core sets. *The Journal of chemical physics* **2016**, *145*.

(87) Liu, S.; Zhu, L.; Sheong, F. K.; Wang, W.; Huang, X. Adaptive partitioning by local density-peaks: An efficient density-based clustering algorithm for analyzing molecular dynamics trajectories. *Journal of Computational Chemistry* **2017**, *38*, 152–160.

(88) Wu, H.; Nüske, F.; Paul, F.; Klus, S.; Koltai, P.; Noé, F. Variational Koopman models: Slow collective variables and molecular kinetics from short off-equilibrium simulations. *The Journal of chemical physics* **2017**, *146*.

(89) Wu, H.; Paul, F.; Wehmeyer, C.; Noé, F. Multiensemble Markov models of molecular thermodynamics and kinetics. *Proceedings of the National Academy of Sciences* **2016**, *113*, E3221–E3230.

(90) McGibbon, R. T.; Pande, V. S. Variational cross-validation of slow dynamical modes in molecular kinetics. *The Journal of chemical physics* **2015**, *142*, 03B621_1.

(91) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *The Journal of chemical physics* **2007**, *126*.

(92) Nagel, D.; Sartore, S.; Stock, G. Selecting features for Markov modeling: a case study on HP35. *Journal of Chemical Theory and Computation* **2023**, *19*, 3391–3405.

(93) Röblitz, S.; Weber, M. Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification. *Advances in Data Analysis and Classification* **2013**, *7*, 147–179.

(94) Deuflhard, P.; Weber, M. Robust Perron cluster analysis in conformation dynamics. *Linear algebra and its applications* **2005**, *398*, 161–184.

(95) Bowman, G. R. Improved coarse-graining of Markov state models via explicit consideration of statistical uncertainty. *The Journal of Chemical Physics* **2012**, *137*.

(96) Wang, W.; Liang, T.; Sheong, F. K.; Fan, X.; Huang, X. An efficient Bayesian kinetic lumping algorithm to identify metastable conformational states via Gibbs sampling. *The Journal of Chemical Physics* **2018**, *149*.

(97) Jain, A.; Stock, G. Identifying metastable states of folding proteins. *Journal of chemical theory and computation* **2012**, *8*, 3810–3819.

(98) Yao, Y.; Cui, R. Z.; Bowman, G. R.; Silva, D.-A.; Sun, J.; Huang, X. Hierarchical Nyström methods for constructing Markov state models for conformational dynamics. *The Journal of chemical physics* **2013**, *138*.

(99) Schulz, F.; Wagner, D.; Weihe, K. Dijkstra's algorithm on-line: An empirical case study from public railroad transport. *Journal of Experimental Algorithmics (JEA)* **2000**, *5*, 12–es.

(100) Ray, D.; Parrinello, M. Data-driven classification of ligand unbinding pathways. *Proceedings of the National Academy of Sciences* **2024**, *121*, e2313542121.

(101) Chandler, D. Introduction to modern statistical. *Mechanics. Oxford University Press, Oxford, UK* **1987**, *5*, 11.

(102) Wu, Y.; Cao, S.; Qiu, Y.; Huang, X. Tutorial on how to build non-Markovian dynamic models from molecular dynamics simulations for studying protein conformational changes. *The Journal of Chemical Physics* **2024**, *160*.

(103) Unarta, I. C.; Cao, S.; Kubo, S.; Wang, W.; Cheung, P. P.-H.; Gao, X.; Takada, S.; Huang, X. Role of bacterial RNA polymerase gate opening dynamics in DNA loading and antibiotics inhibition elucidated by quasi-Markov State Model. *Proceedings of the National Academy of Sciences* **2021**, *118*, e2024324118.

(104) Mai, Y.; Eisenberg, A. Self-assembly of block copolymers. *Chemical Society Reviews* **2012**, *41*, 5969–5985.

(105) Shukla, D.; Meng, Y.; Roux, B.; Pande, V. S. Activation pathway of Src kinase reveals intermediate states as targets for drug design. *Nature communications* **2014**, *5*, 3397.

(106) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. Molecular simulation of ab initio protein folding for a millisecond folder NTL9 (1- 39). *Journal of the American Chemical Society* **2010**, *132*, 1526–1528.

(107) Paul, F.; Noe, F.; Weikl, T. R. Identifying conformational-selection and induced-fit aspects in the binding-induced folding of PMI from Markov state modeling of atomistic simulations. *The Journal of Physical Chemistry B* **2018**, *122*, 5649–5656.

(108) Han, W.; Schulten, K. Characterization of folding mechanisms of Trp-cage and WW-domain by network analysis of simulations with a hybrid-resolution model. *The Journal of Physical Chemistry B* **2013**, *117*, 13367–13377.

(109) Meng, Y.; Shukla, D.; Pande, V. S.; Roux, B. Transition path theory analysis of c-Src kinase activation. *Proceedings of the National Academy of Sciences* **2016**, *113*, 9193–9198.

(110) Zheng, W.; Gallicchio, E.; Deng, N.; Andrec, M.; Levy, R. M. Kinetic network study of the diversity and temperature dependence of trp-cage folding pathways: Combining transition path theory with stochastic simulations. *The Journal of Physical Chemistry B* **2011**, *115*, 1512–1523.

(111) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of slow molecular order parameters for Markov model construction. *The Journal of chemical physics* **2013**, *139*, 07B604_1.

(112) Kingma, D. Auto-Encoding Variational Bayes. *arXiv preprint arXiv:1312.6114* **2013**.

(113) Hinton, G. E.; Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *science* **2006**, *313*, 504–507.

(114) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How fast-folding proteins fold. *Science* **2011**, *334*, 517–520.

(115) Freddolino, P. L.; Liu, F.; Gruebele, M.; Schulten, K. Ten-microsecond molecular dynamics simulation of a fast-folding WW domain. *Biophysical journal* **2008**, *94*, L75–L77.

(116) Ensign, D. L.; Pande, V. S. The Fip35 WW domain folds with structural and mechanistic heterogeneity in molecular dynamics simulations. *Biophysical journal* **2009**, *96*, L53–L55.

(117) Krivov, S. V. The free energy landscape analysis of protein (FIP35) folding dynamics. *The journal of physical chemistry B* **2011**, *115*, 12315–12324.

(118) A Beccara, S.; Škrbić, T.; Covino, R.; Faccioli, P. Dominant folding pathways of a WW domain. *Proceedings of the National Academy of Sciences* **2012**, *109*, 2330–2335.

(119) Klimov, D. K.; Thirumalai, D. Symmetric connectivity of secondary structure elements enhances the diversity of folding pathways. *Journal of molecular biology* **2005**, *353*, 1171–1186.

(120) Weikl, T. R. Transition states in protein folding kinetics: Modeling Φ-values of small β-sheet proteins. *Biophysical journal* **2008**, *94*, 929–937.

(121) Sharpe, D. J.; Wales, D. J. Efficient and exact sampling of transition path ensembles on Markovian networks. *The Journal of Chemical Physics* **2020**, *153*.

(122) Wales, D. J. Calculating rate constants and committor probabilities for transition networks by graph transformation. *The Journal of chemical physics* **2009**, *130*.

(123) Nagel, D.; Weber, A.; Stock, G. MSMPathfinder: Identification of pathways in Markov state models. *Journal of Chemical Theory and Computation* **2020**, *16*, 7874–7882.

(124) Zimmerman, M. I.; Porter, J. R.; Sun, X.; Silva, R. R.; Bowman, G. R. Choice of adaptive sampling strategy impacts state discovery, transition probabilities, and the apparent mechanism of conformational changes. *Journal of chemical theory and computation* **2018**, *14*, 5459–5475.

(125) Ferguson, A. L.; Panagiotopoulos, A. Z.; Debenedetti, P. G.; Kevrekidis, I. G. Systematic determination of order parameters for chain dynamics using diffusion maps. *Proceedings of the National Academy of Sciences* **2010**, *107*, 13597–13602.

(126) Liu, Z.; Thirumalai, D. Cooperativity and folding kinetics in a multidomain protein with interwoven chain topology. *ACS Central Science* **2022**, *8*, 763–774.

(127) Pirchi, M.; Ziv, G.; Riven, I.; Cohen, S. S.; Zohar, N.; Barak, Y.; Haran, G. Single-molecule fluorescence spectroscopy maps the folding landscape of a large protein. *Nature communications* **2011**, *2*, 493.

(128) Gu, H.; Wang, W.; Cao, S.; Unarta, I. C.; Yao, Y.; Sheong, F. K.; Huang, X. RPnet: a reverse-projection-based neural network for coarse-graining metastable conformational states for protein dynamics. *Physical Chemistry Chemical Physics* **2022**, *24*, 1462–1474.

(129) Mehdi, S.; Wang, D.; Pant, S.; Tiwary, P. Accelerating all-atom simulations and gaining mechanistic understanding of biophysical systems through state predictive information bottleneck. *Journal of chemical theory and computation* **2022**, *18*, 3231–3238.

(130) Beyerle, E. R.; Tiwary, P. Thermodynamically optimized machine-learned reaction coordinates for hydrophobic ligand dissociation. *The Journal of Physical Chemistry B* **2024**, *128*, 755–767.

(131) Vani, B. P.; Aranganathan, A.; Wang, D.; Tiwary, P. Alphafold2-rave: From sequence to boltzmann ranking. *Journal of chemical theory and computation* **2023**, *19*, 4351–4354.

(132) Vani, B. P.; Aranganathan, A.; Tiwary, P. Exploring kinase asp-phe-gly (dfg) loop conformational stability with alphafold2-rave. *Journal of chemical information and modeling* **2023**, *64*, 2789–2797.

(133) Zou, Z.; Beyerle, E. R.; Tsai, S.-T.; Tiwary, P. Driving and characterizing nucleation of urea and glycine polymorphs in water. *Proceedings of the National Academy of Sciences* **2023**, *120*, e2216099120.

(134) Tomczak, J.; Welling, M. In *International conference on artificial intelligence and statistics*, 2018, pp 1214–1223.

(135) Nagel, D.; Sartore, S.; Stock, G. Toward a benchmark for Markov state models: The folding of HP35. *The Journal of Physical Chemistry Letters* **2023**, *14*, 6956–6967.

(136) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. Protein folding kinetics and thermodynamics from atomistic simulation. *Proceedings of the National Academy of Sciences* **2012**, *109*, 17845–17850.

(137) Sidky, H.; Chen, W.; Ferguson, A. L. High-resolution Markov state models for the dynamics of Trp-cage miniprotein constructed over slow folding modes identified by state-free reversible VAMPnets. *The Journal of Physical Chemistry B* **2019**, *123*, 7999–8009.

(138) Schneekloth Jr, J. S.; Fonseca, F. N.; Koldobskiy, M.; Mandal, A.; Deshaies, R.; Sakamoto, K.; Crews, C. M. Chemical genetic control of protein levels: selective in vivo targeted degradation. *Journal of the American Chemical Society* **2004**, *126*, 3748–3754.

(139) Zhou, P. Targeted protein degradation. *Current opinion in chemical biology* **2005**, *9*, 51–55.

(140) Raina, K.; Crews, C. M. Chemical inducers of targeted protein degradation. *Journal of Biological Chemistry* **2010**, *285*, 11057–11060.

(141) Békés, M.; Langley, D. R.; Crews, C. M. PROTAC targeted protein degraders: the past is prologue. *Nature Reviews Drug Discovery* **2022**, *21*, 181–200.

(142) Lai, A. C.; Crews, C. M. Induced protein degradation: an emerging drug discovery paradigm. *Nature reviews Drug discovery* **2017**, *16*, 101–114.

(143) Liu, Z.; Hu, M.; Yang, Y.; Du, C.; Zhou, H.; Liu, C.; Chen, Y.; Fan, L.; Ma, H.; Gong, Y., et al. An overview of PROTACs: a promising drug discovery paradigm. *Molecular biomedicine* **2022**, *3*, 46.

(144) Chamberlain, P. P.; Hamann, L. G. Development of targeted protein degradation therapeutics. *Nature chemical biology* **2019**, *15*, 937–944.

(145) Mostofian, B.; Martin, H.-J.; Razavi, A.; Patel, S.; Allen, B.; Sherman, W.; Izaguirre, J. A. Targeted protein degradation: advances, challenges, and prospects for computational methods. *Journal of Chemical Information and Modeling* **2023**, *63*, 5408–5432.

(146) Cowan, A. D.; Ciulli, A. Driving E3 ligase substrate specificity for targeted protein degradation: lessons from nature and the laboratory. *Annual review of biochemistry* **2022**, *91*, 295–319.

(147) Rui, H.; Ashton, K. S.; Min, J.; Wang, C.; Potts, P. R. Protein–protein interfaces in molecular glue-induced ternary complexes: classification, characterization, and prediction. *RSC Chemical Biology* **2023**, *4*, 192–215.

(148) Daurio, N. A.; Zhou, H.; Chen, Y.; Sheth, P. R.; Imbriglio, J. E.; McLaren, D. G.; Tawa, P.; Rachdaoui, N.; Previs, M. J.; Kasumov, T., et al. Examining targeted protein degradation from physiological and analytical perspectives: enabling translation between cells and subjects. *ACS Chemical Biology* **2020**, *15*, 2623–2635.

(149) Li, W.; Zhang, J.; Guo, L.; Wang, Q. Importance of three-body problems and protein–protein interactions in proteolysis-targeting chimera modeling: insights from molecular dynamics simulations. *Journal of Chemical Information and Modeling* **2022**, *62*, 523–532.

(150) Gadd, M. S.; Testa, A.; Lucas, X.; Chan, K.-H.; Chen, W.; Lamont, D. J.; Zengerle, M.; Ciulli, A. Structural basis of PROTAC cooperative recognition for selective protein degradation. *Nature chemical biology* **2017**, *13*, 514–521.

(151) Farnaby, W.; Koegl, M.; Roy, M. J.; Whitworth, C.; Diers, E.; Trainor, N.; Zollman, D.; Steurer, S.; Karolyi-Oezguer, J.; Riedmueller, C., et al. BAF complex vulnerabilities in cancer demonstrated via structure-based PROTAC design. *Nature chemical biology* **2019**, *15*, 672–680.

(152) Cyrus, K.; Wehenkel, M.; Choi, E.-Y.; Han, H.-J.; Lee, H.; Swanson, H.; Kim, K.-B. Impact of linker length on the activity of PROTACs. *Molecular BioSystems* **2011**, *7*, 359–364.

(153) Weerakoon, D.; Carbajo, R. J.; De Maria, L.; Tyrchan, C.; Zhao, H. Impact of PROTAC linker plasticity on the solution conformations and dissociation of the ternary complex. *Journal of Chemical Information and Modeling* **2022**, *62*, 340–349.

(154) Paiva, S.-L.; Crews, C. M. Targeted protein degradation: elements of PROTAC design. *Current opinion in chemical biology* **2019**, *50*, 111–119.

(155) Chan, K.-H.; Zengerle, M.; Testa, A.; Ciulli, A. Impact of target warhead and linkage vector on inducing protein degradation: comparison of bromodomain and extra-terminal (BET) degraders derived from triazolodiazepine (JQ1) and tetrahydroquinoline (I-BET726) BET inhibitor scaffolds. *Journal of medicinal chemistry* **2018**, *61*, 504–513.

(156) Roy, M. J.; Winkler, S.; Hughes, S. J.; Whitworth, C.; Galant, M.; Farnaby, W.; Rumpel, K.; Ciulli, A. SPR-measured dissociation kinetics of PROTAC ternary complexes influence target degradation rate. *ACS chemical biology* **2019**, *14*, 361–368.

(157) Smith, B. E.; Wang, S. L.; Jaime-Figueroa, S.; Harbin, A.; Wang, J.; Hamman, B. D.; Crews, C. M. Differential PROTAC substrate specificity dictated by orientation of recruited E3 ligase. *Nature communications* **2019**, *10*, 131.

(158) Nowak, R. P.; DeAngelo, S. L.; Buckley, D.; He, Z.; Donovan, K. A.; An, J.; Safaee, N.; Jedrychowski, M. P.; Ponthier, C. M.; Ishoey, M., et al. Plasticity in binding confers selectivity in ligand-induced protein degradation. *Nature chemical biology* **2018**, *14*, 706–714.

(159) Dong, Y.; Ma, T.; Xu, T.; Feng, Z.; Li, Y.; Song, L.; Yao, X.; Ashby Jr, C. R.; Hao, G.-F. Characteristic roadmap of linker governs the rational design of PROTACs. *Acta Pharmaceutica Sinica B* **2024**.

(160) Drummond, M. L.; Henry, A.; Li, H.; Williams, C. I. Improved accuracy for modeling PROTAC-mediated ternary complex formation and targeted protein degradation via new in silico methodologies. *Journal of Chemical Information and Modeling* **2020**, *60*, 5234–5254.

(161) Wu, J.; Wang, W.; Leung, C.-H. Computational strategies for PROTAC drug discovery. *Acta Materia Medica* **2023**, *2*, 42–53.

(162) Tang, R.; Wang, Z.; Xiang, S.; Wang, L.; Yu, Y.; Wang, Q.; Deng, Q.; Hou, T.; Sun, H. Uncovering the Kinetic Characteristics and Degradation Preference of PROTAC Systems with Advanced Theoretical Analyses. *JACS Au* **2023**, *3*, 1775–1789.

(163) Wan, H.; Voelz, V. A. Adaptive Markov state model estimation using short reseeding trajectories. *The Journal of chemical physics* **2020**, *152*.

(164) Huang, L.; Guo, Z.; Wang, F.; Fu, L. KRAS mutation: from undruggable to druggable in cancer. *Signal transduction and targeted therapy* **2021**, *6*, 386.

(165) Bond, M. J.; Chu, L.; Nalawansha, D. A.; Li, K.; Crews, C. M. Targeted degradation of oncogenic KRASG12C by VHL-recruiting PROTACs. *ACS central science* **2020**, *6*, 1367–1375.

(166) Popow, J.; Farnaby, W.; Gollner, A.; Kofink, C.; Fischer, G.; Wurm, M.; Zollman, D.; Wijaya, A.; Mischerikow, N.; Hasenoehrl, C., et al. Targeting cancer with small-molecule pan-KRAS degraders. *Science* **2024**, *385*, 1338–1347.

(167) Tolcher, A. W.; Park, W.; Wang, J. S.; Spira, A. I.; Janne, P. A.; Lee, H.-J.; Gill, S.; LoRusso, P.; Herzberg, B.; Goldman, J. W., et al. Trial in progress: A phase 1, first-in-human, open-label, multicenter, dose-escalation and dose-expansion study of ASP3082 in patients with previously treated advanced solid tumors and KRAS G12D mutations. 2023.

(168) Hon, W.-C.; Wilson, M. I.; Harlos, K.; Claridge, T. D.; Schofield, C. J.; Pugh, C. W.; Maxwell, P. H.; Ratcliffe, P. J.; Stuart, D. I.; Jones, E. Y. Structural basis for the recognition of hydroxyproline in HIF-1α by pVHL. *Nature* **2002**, *417*, 975–978.

(169) Rodriguez-Gonzalez, A.; Cyrus, K.; Salcius, M.; Kim, K.; Crews, C.; Deshaies, R.; Sakamoto, K. Targeting steroid hormone receptors for ubiquitination and degradation in breast and prostate cancer. *Oncogene* **2008**, *27*, 7201–7211.

(170) Galdeano, C.; Gadd, M. S.; Soares, P.; Scaffidi, S.; Van Molle, I.; Birced, I.; Hewitt, S.; Dias, D. M.; Ciulli, A. Structure-guided design and optimization of small molecules targeting the protein–protein interaction between the von Hippel–Lindau (VHL) E3 ubiquitin ligase and the hypoxia inducible factor (HIF) alpha subunit with in vitro nanomolar affinities. *Journal of Medicinal Chemistry* **2014**, *57*, 8657–8663.

(171) Diehl, C. J.; Ciulli, A. Discovery of small molecule ligands for the von Hippel-Lindau (VHL) E3 ligase and their use as inhibitors and PROTAC degraders. *Chemical Society Reviews* **2022**, *51*, 8216–8257.

(172) Bemis, T. A.; La Clair, J. J.; Burkart, M. D. Unraveling the role of linker design in proteolysis targeting chimeras: Miniperspective. *Journal of Medicinal Chemistry* **2021**, *64*, 8042–8052.

(173) Troup, R. I.; Fallan, C.; Baud, M. G. Current strategies for the design of PROTAC linkers: a critical review. *Exploration of Targeted Anti-tumor Therapy* **2020**, *1*, 273.

(174) Setia, N.; Almuqdadi, H. T. A.; Abid, M. Journey of von hippel-lindau (VHL) E3 ligase in PROTACs design: From VHL ligands to VHL-based degraders. *European Journal of Medicinal Chemistry* **2023**, 116041.

(175) Zoppi, V.; Hughes, S. J.; Maniaci, C.; Testa, A.; Gmaschitz, T.; Wieshofer, C.; Koegl, M.; Riching, K. M.; Daniels, D. L.; Spallarossa, A., et al. Iterative design and optimization of initially inactive proteolysis targeting chimeras (PROTACs) identify VZ185 as a potent, fast, and selective von Hippel–Lindau (VHL) based dual degrader probe of BRD9 and BRD7. *Journal of medicinal chemistry* **2018**, *62*, 699–726.

(176) Disch, J. S.; Duffy, J. M.; Lee, E. C.; Gikunju, D.; Chan, B.; Levin, B.; Monteiro, M. I.; Talcott, S. A.; Lau, A. C.; Zhou, F., et al. Bispecific estrogen receptor α degraders incorporating novel binders identified using DNA-encoded chemical library screening. *Journal of Medicinal Chemistry* **2021**, *64*, 5049–5066.

(177) Han, X.; Zhao, L.; Xiang, W.; Qin, C.; Miao, B.; Xu, T.; Wang, M.; Yang, C.-Y.; Chinnaswamy, K.; Stuckey, J., et al. Discovery of highly potent and efficient PROTAC degraders of androgen receptor (AR) by employing weak binding affinity VHL E3 ligase ligands. *Journal of medicinal chemistry* **2019**, *62*, 11218–11231.

(178) Shah, R. R.; De Vita, E.; Sathyamurthi, P. S.; Conole, D.; Zhang, X.; Fellows, E.; Dickinson, E. R.; Fleites, C. M.; Queisser, M. A.; Harling, J. D., et al. Structure-guided design and optimization of covalent VHL-targeted sulfonyl fluoride PROTACs. *Journal of Medicinal Chemistry* **2024**, *67*, 4641–4654.

(179) Tovell, H.; Testa, A.; Maniaci, C.; Zhou, H.; Prescott, A. R.; Macartney, T.; Ciulli, A.; Alessi, D. R. Rapid and reversible knockdown of endogenously tagged endosomal proteins via an optimized HaloPROTAC degrader. *ACS chemical biology* **2019**, *14*, 882–892.

(180) Słabicki, M.; Kozicka, Z.; Petzold, G.; Li, Y.-D.; Manojkumar, M.; Bunker, R. D.; Donovan, K. A.; Sievers, Q. L.; Koeppel, J.; Suchyta, D., et al. The CDK inhibitor CR8 acts as a molecular glue degrader that depletes cyclin K. *Nature* **2020**, *585*, 293–297.

(181) Simonetta, K. R.; Taygerly, J.; Boyle, K.; Basham, S. E.; Padovani, C.; Lou, Y.; Cummins, T. J.; Yung, S. L.; von Soly, S. K.; Kayser, F., et al. Prospective discovery of small molecule enhancers of an E3 ligase-substrate interaction. *Nature Communications* **2019**, *10*, 1402.

(182) Vaynberg, J.; Qin, J. Weak protein–protein interactions as probed by NMR spectroscopy. *TRENDS in Biotechnology* **2006**, *24*, 22–27.

(183) Pereira, G. P.; Gouzien, C.; Souza, P. C. T. d.; Martin, J. AlphaFold-Multimer struggles in predicting PROTAC-mediated protein-protein interfaces. *bioRxiv* **2024**, 2024–03.

(184) Yin, R.; Feng, B. Y.; Varshney, A.; Pierce, B. G. Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants. *Protein Science* **2022**, *31*, e4379.

(185) Stebbins, C. E.; Kaelin Jr, W. G.; Pavletich, N. P. Structure of the VHL-ElonginC-ElonginB complex: implications for VHL tumor suppressor function. *Science* **1999**, *284*, 455–461.

(186) Wang, X.; Allen, S.; Blake, J. F.; Bowcut, V.; Briere, D. M.; Calinisan, A.; Dahlke, J. R.; Fell, J. B.; Fischer, J. P.; Gunn, R. J., et al. Identification of MRTX1133, a noncovalent, potent, and selective KRASG12D inhibitor. *Journal of medicinal chemistry* **2021**, *65*, 3123–3133.

(187) Mark, P.; Nilsson, L. Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K. *The Journal of Physical Chemistry A* **2001**, *105*, 9954–9960.

(188) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D., et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology* **2017**, *13*, e1005659.

(189) Voelz, V. A.; Pande, V. S.; Bowman, G. R. Folding@ home: Achievements from over 20 years of citizen science herald the exascale era. *Biophysical journal* **2023**, *122*, 2852–2863.

(190) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of chemical theory and computation* **2015**, *11*, 3696–3713.

(191) Luger, K.; Dechassa, M. L.; Tremethick, D. J. New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? *Nature reviews Molecular cell biology* **2012**, *13*, 436–447.

(192) Rowley, M. J.; Corces, V. G. Organizational principles of 3D genome architecture. *Nature Reviews Genetics* **2018**, *19*, 789–800.

(193) Jerkovic, I.; Cavalli, G. Understanding 3D genome organization by multidisciplinary methods. *Nature Reviews Molecular Cell Biology* **2021**, *22*, 511–528.

(194) Lin, X.; Qi, Y.; Latham, A. P.; Zhang, B. Multiscale modeling of genome organization with maximum entropy optimization. *The Journal of chemical physics* **2021**, *155*.

(195) Liu, S.; Athreya, A.; Lao, Z.; Zhang, B. From Nucleosomes to Compartments: Physicochemical Interactions Underlying Chromatin Organization. *Annual Review of Biophysics* **2024**, *53*.

(196) Song, F.; Chen, P.; Sun, D.; Wang, M.; Dong, L.; Liang, D.; Xu, R.-M.; Zhu, P.; Li, G. Cryo-EM study of the chromatin fiber reveals a double helix twisted by tetranucleosomal units. *Science* **2014**, *344*, 376–380.

(197) Schalch, T.; Duda, S.; Sargent, D. F.; Richmond, T. J. X-ray structure of a tetranucleosome and its implications for the chromatin fibre. *Nature* **2005**, *436*, 138–141.

(198) Williams, S.; Athey, B.; Muglia, L.; Schappe, R.; Gough, A.; Langmore, J. Chromatin fibers are left-handed double helices with diameter and mass per unit length that depend on linker length. *Biophysical journal* **1986**, *49*, 233–248.

(199) Dorigo, B.; Schalch, T.; Kulangara, A.; Duda, S.; Schroeder, R. R.; Richmond, T. J. Nucleosome arrays reveal the two-start organization of the chromatin fiber. *Science* **2004**, *306*, 1571–1573.

(200) Robinson, P. J.; Fairall, L.; Huynh, V. A.; Rhodes, D. EM measurements define the dimensions of the "30-nm" chromatin fiber: evidence for a compact, interdigitated structure. *Proceedings of the National Academy of Sciences* **2006**, *103*, 6506–6511.

(201) Hsieh, T.-H. S.; Weiner, A.; Lajoie, B.; Dekker, J.; Friedman, N.; Rando, O. J. Mapping nucleosome resolution chromosome folding in yeast by micro-C. *Cell* **2015**, *162*, 108–119.

(202) Ou, H. D.; Phan, S.; Deerinck, T. J.; Thor, A.; Ellisman, M. H.; O'shea, C. C. ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science* **2017**, *357*, eaag0025.

(203) Eltsov, M.; MacLellan, K. M.; Maeshima, K.; Frangakis, A. S.; Dubochet, J. Analysis of cryo-electron microscopy images does not support the existence of 30-nm chromatin fibers in mitotic chromosomes in situ. *Proceedings of the National Academy of Sciences* **2008**, *105*, 19732–19737.

(204) Nishino, Y.; Eltsov, M.; Joti, Y.; Ito, K.; Takata, H.; Takahashi, Y.; Hihara, S.; Frangakis, A. S.; Imamoto, N.; Ishikawa, T., et al. Human mitotic chromosomes consist predominantly of irregularly folded nucleosome fibres without a 30-nm chromatin structure. *The EMBO journal* **2012**, *31*, 1644–1653.

(205) Ohno, M.; Ando, T.; Priest, D. G.; Kumar, V.; Yoshida, Y.; Taniguchi, Y. Sub-nucleosomal genome structure reveals distinct nucleosome folding motifs. *Cell* **2019**, *176*, 520–534.

(206) Alvarado, W.; Agrawal, V.; Li, W. S.; Dravid, V. P.; Backman, V.; de Pablo, J. J.; Ferguson, A. L. Denoising Autoencoder Trained on Simulation-Derived Structures for Noise Reduction in Chromatin Scanning Transmission Electron Microscopy. *ACS Central Science* **2023**.

(207) Ding, X.; Lin, X.; Zhang, B. Stability and folding pathways of tetra-nucleosome from six-dimensional free energy surface. *Nature communications* **2021**, *12*, 1091.

(208) Liu, S.; Lin, X.; Zhang, B. Chromatin fiber breaks into clutches under tension and crowding. *Nucleic Acids Research* **2022**, *50*, 9738–9747.

(209) Ricci, M. A.; Manzo, C.; García-Parajo, M. F.; Lakadamyali, M.; Cosma, M. P. Chromatin fibers are formed by heterogeneous groups of nucleosomes in vivo. *Cell* **2015**, *160*, 1145–1158.

(210) Pande, V. S.; Beauchamp, K.; Bowman, G. R. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods* **2010**, *52*, 99–105.

(211) Husic, B. E.; Pande, V. S. Markov state models: From an art to a science. *Journal of the American Chemical Society* **2018**, *140*, 2386–2396.

(212) Dominic III, A. J.; Cao, S.; Montoya-Castillo, A.; Huang, X. Memory unlocks the future of biomolecular dynamics: Transformative tools to uncover physical insights accurately and efficiently. *Journal of the American Chemical Society* **2023**, *145*, 9916–9927.

(213) Gibson, B. A.; Doolittle, L. K.; Schneider, M. W.; Jensen, L. E.; Gamarra, N.; Henry, L.; Gerlich, D. W.; Redding, S.; Rosen, M. K. Organization of chromatin by intrinsic and regulated phase separation. *Cell* **2019**, *179*, 470–484.

(214) Zhang, M.; Díaz-Celis, C.; Onoa, B.; Cañari-Chumpitaz, C.; Requejo, K. I.; Liu, J.; Vien, M.; Nogales, E.; Ren, G.; Bustamante, C. Molecular organization of the early stages of nucleosome phase separation visualized by cryo-electron tomography. *Molecular cell* **2022**, *82*, 3000–3014.

(215) Armeev, G. A.; Kniazeva, A. S.; Komarova, G. A.; Kirpichnikov, M. P.; Shaytan, A. K. Histone dynamics mediate DNA unwrapping and sliding in nucleosomes. *Nature communications* **2021**, *12*, 2387.

(216) Peng, Y.; Li, S.; Onufriev, A., et al. Binding of regulatory proteins to nucleosomes is modulated by dynamic histone tails. Nat Commun 12: 5280, 2021.

(217) Li, S.; Wei, T.; Panchenko, A. Histone variant H2A. Z modulates nucleosome dynamics to promote DNA accessibility. *Biophysical Journal* **2023**, *122*, 218a.

(218) Woods, D. C.; Rodríguez-Ropero, F.; Wereszczynski, J. The Dynamic Influence of Linker Histone Saturation within the Poly-Nucleosome Array. *Journal of Molecular Biology* **2021**, *433*, 166902.

(219) Winogradoff, D.; Aksimentiev, A. Molecular Mechanism of Spontaneous Nucleosome Unraveling. *Journal of Molecular Biology* **2019**, *431*, 323–335.

(220) Clementi, C.; Nymeyer, H.; Onuchic, J. N. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *Journal of molecular biology* **2000**, *298*, 937–953.

(221) Hinckley, D. M.; Freeman, G. S.; Whitmer, J. K.; De Pablo, J. J. An experimentally-informed coarse-grained 3-site-per-nucleotide model of DNA: Structure, thermodynamics, and dynamics of hybridization. *The Journal of chemical physics* **2013**, *139*.

(222) Lequieu, J.; Córdoba, A.; Schwartz, D. C.; de Pablo, J. J. Tension-dependent free energies of nucleosome unwrapping. *ACS central science* **2016**, *2*, 660–666.

(223) Moller, J.; Lequieu, J.; de Pablo, J. J. The free energy landscape of internucleosome interactions and its relation to chromatin fiber structure. *ACS central science* **2019**, *5*, 341–348.

(224) Leicher, R.; Ge, E. J.; Lin, X.; Reynolds, M. J.; Xie, W.; Walz, T.; Zhang, B.; Muir, T. W.; Liu, S. Single-molecule and in silico dissection of the interaction between Polycomb repressive complex 2 and chromatin. *Proceedings of the National Academy of Sciences* **2020**, *117*, 30465–30475.

(225) Latham, A. P.; Zhang, B. On the Stability and Layered Organization of Protein-DNA Condensates. *Biophysical Journal* **2022**, *121*, 1727–1737.

(226) Lin, X.; Leicher, R.; Liu, S.; Zhang, B. Cooperative DNA Looping by PRC2 Complexes. *Nucleic Acids Research* **2021**, *49*, 6238–6248.

(227) Farr, S. E.; Woods, E. J.; Joseph, J. A.; Garaizar, A.; Collepardo-Guevara, R. Nucleosome Plasticity Is a Critical Element of Chromatin Liquid–Liquid Phase Separation and Multivalent Nucleosome Interactions. *Nature Communications* **2021**, *12*, 1–17.

(228) Naritomi, Y.; Fuchigami, S. Slow dynamics of a protein backbone in molecular dynamics simulation revealed by time-structure based independent component analysis. *The Journal of Chemical Physics* **2013**, *139*, 12B605_1.

(229) Schwantes, C. R.; Pande, V. S. Improvements in Markov state model construction reveal many non-native interactions in the folding of NTL9. *Journal of chemical theory and computation* **2013**, *9*, 2000–2009.

(230) Metzner, P.; Schütte, C.; Vanden-Eijnden, E. Transition path theory for Markov jump processes. *Multiscale Modeling & Simulation* **2009**, *7*, 1192–1219.

(231) Dombrowski, M.; Engeholm, M.; Dienemann, C.; Dodonova, S.; Cramer, P. Histone H1 binding to nucleosome arrays depends on linker DNA length and trajectory. *Nature structural & molecular biology* **2022**, *29*, 493–501.

(232) Zhurkin, V. B.; Norouzi, D. Topological polymorphism of nucleosome fibers and folding of chromatin. *Biophysical Journal* **2021**, *120*, 577–585.

(233) Szerlong, H. J.; Hansen, J. C. Nucleosome distribution and linker DNA: connecting nuclear function to dynamic chromatin structure. *Biochemistry and Cell Biology* **2011**, *89*, 24–34.

(234) Correll, S. J.; Schubert, M. H.; Grigoryev, S. A. Short nucleosome repeats impose rotational modulations on chromatin fibre folding. *The EMBO journal* **2012**, *31*, 2416–2426.

(235) Collepardo-Guevara, R.; Schlick, T. Chromatin fiber polymorphism triggered by variations of DNA linker lengths. *Proceedings of the National Academy of Sciences* **2014**, *111*, 8061–8066.

(236) Kenzaki, H.; Takada, S. Linker DNA length is a key to tri-nucleosome folding. *Journal of Molecular Biology* **2021**, *433*, 166792.

(237) Alvarado, W.; Moller, J.; Ferguson, A. L.; de Pablo, J. J. Tetranucleosome interactions drive chromatin folding. *ACS Central Science* **2021**, *7*, 1019–1027.

(238) Garcia-Saez, I.; Menoni, H.; Boopathi, R.; Shukla, M. S.; Soueidan, L.; Noirclerc-Savoye, M.; Le Roy, A.; Skoufias, D. A.; Bednar, J.; Hamiche, A., et al. Structure of an H1-bound 6-nucleosome array reveals an untwisted two-start chromatin fiber conformation. *Molecular cell* **2018**, *72*, 902–915.

(239) Lewis, T. S.; Sokolova, V.; Jung, H.; Ng, H.; Tan, D. Structural basis of chromatin regulation by histone variant H2A. Z. *Nucleic Acids Research* **2021**, *49*, 11379–11391.

(240) Brandani, G. B.; Gopi, S.; Yamauchi, M.; Takada, S. Molecular dynamics simulations for the study of chromatin biology. *Current Opinion in Structural Biology* **2022**, *77*, 102485.

(241) Voong, L. N.; Xi, L.; Sebeson, A. C.; Xiong, B.; Wang, J.-P.; Wang, X. Insights into nucleosome organization in mouse embryonic stem cells through chemical mapping. *Cell* **2016**, *167*, 1555–1570.

(242) Li, Y.; Zhang, H.; Li, X.; Wu, W.; Zhu, P. Cryo-ET study from in vitro to in vivo revealed a general folding mode of chromatin with two-start helical architecture. *Cell Reports* **2023**, *42*.

(243) Hou, Z.; Nightingale, F.; Zhu, Y.; MacGregor-Chatwin, C.; Zhang, P. Structure of Native Chromatin Fibres Revealed by Cryo-ET in Situ. *Nature Communications* **2023**, *14*, 6324.

(244) Latham, A. P.; Zhang, B. Unifying Coarse-Grained Force Fields for Folded and Disordered Proteins. *Current Opinion in Structural Biology* **2022**, *72*, 63–70.

(245) Liu, S.; Wang, C.; Latham, A. P.; Ding, X.; Zhang, B. OpenABC enables flexible, simplified, and efficient GPU accelerated simulations of biomolecular condensates. *PLOS Computational Biology* **2023**, *19*, e1011442.

(246) Lin, X.; Zhang, B. Explicit Ion Modeling Predicts Physicochemical Interactions for Chromatin Organization. *eLife* **2024**, *12*, ed. by Dalal, Y.; Cui, Q., RP90073.

(247) Freeman, G. S.; Hinckley, D. M.; Lequieu, J. P.; Whitmer, J. K.; De Pablo, J. J. Coarse-grained modeling of DNA curvature. *The Journal of chemical physics* **2014**, *141*, 10B615_1.

(248) Lu, W.; Bueno, C.; Schafer, N. P.; Moller, J.; Jin, S.; Chen, X.; Chen, M.; Gu, X.; Davtyan, A.; de Pablo, J. J., et al. OpenAWSEM with Open3SPN2: A fast, flexible, and accessible framework for large-scale coarse-grained biomolecular simulations. *PLoS computational biology* **2021**, *17*, e1008308.

(249) Noel, J. K.; Levi, M.; Raghunathan, M.; Lammert, H.; Hayes, R. L.; Onuchic, J. N.; Whitford, P. C. SMOG 2: a versatile software package for generating structure-based models. *PLoS computational biology* **2016**, *12*, e1004794.

(250) Parsons, T.; Zhang, B. Critical role of histone tail entropy in nucleosome unwinding. *The Journal of chemical physics* **2019**, *150*.

(251) Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; in't Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D., et al. LAMMPS-a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications* **2022**, *271*, 108171.

(252) Tanaka, H.; Tong, H.; Shi, R.; Russo, J. Revealing key structural features hidden in liquids and glasses. *Nature Reviews Physics* **2019**, *1*, 333–348.

(253) Richard, D.; Ozawa, M.; Patinet, S.; Stanifer, E.; Shang, B.; Ridout, S.; Xu, B.; Zhang, G.; Morse, P.; Barrat, J.-L., et al. Predicting plasticity in disordered solids from structural indicators. *Physical Review Materials* **2020**, *4*, 113609.

(254) Widmer-Cooper, A.; Perry, H.; Harrowell, P.; Reichman, D. R. Irreversible reorganization in a supercooled liquid originates from localized soft modes. *Nature Physics* **2008**, *4*, 711–715.

(255) Berthier, L.; Biroli, G.; Bouchaud, J.-P.; Cipelletti, L.; van Saarloos, W., *Dynamical heterogeneities in glasses, colloids, and granular media*; OUP Oxford: 2011; Vol. 150.

(256) Ediger, M. D. Spatially heterogeneous dynamics in supercooled liquids. *Annual review of physical chemistry* **2000**, *51*, 99–128.

(257) Jack, R. L.; Dunleavy, A. J.; Royall, C. P. Information-theoretic measurements of coupling between structure and dynamics in glass formers. *Physical review letters* **2014**, *113*, 095703.

(258) Jung, G.; Biroli, G.; Berthier, L. Predicting dynamic heterogeneity in glass-forming liquids by physics-inspired machine learning. *Physical Review Letters* **2023**, *130*, 238202.

(259) Jung, G.; Biroli, G.; Berthier, L. Dynamic heterogeneity at the experimental glass transition predicted by transferable machine learning. *Physical Review B* **2024**, *109*, 064205.

(260) Boattini, E.; Marín-Aguilar, S.; Mitra, S.; Foffi, G.; Smallenburg, F.; Filion, L. Autonomously revealing hidden local structures in supercooled liquids. *Nature communications* **2020**, *11*, 5479.

(261) Paret, J.; Jack, R. L.; Coslovich, D. Assessing the structural heterogeneity of supercooled liquids through community inference. *The Journal of chemical physics* **2020**, *152*.

(262) Doliwa, B.; Heuer, A. What does the potential energy landscape tell us about the dynamics of supercooled liquids and glasses? *Physical Review Letters* **2003**, *91*.

(263) Hocky, G. M.; Coslovich, D.; Ikeda, A.; Reichman, D. R. Correlation of local order with particle mobility in supercooled liquids is highly system dependent. *Physical review letters* **2014**, *113*, 157801.

(264) Widmer-Cooper, A.; Harrowell, P. Predicting the Long-Time Dynamic Heterogeneity in a Supercooled Liquid on the Basis of Short-Time Heterogeneities. *Physical Review Letters* **2006**, *96*, 185701.

(265) Rieser, J. M.; Goodrich, C. P.; Liu, A. J.; Durian, D. J. Divergence of Voronoi cell anisotropy vector: a threshold-free characterization of local structure in amorphous materials. *Physical review letters* **2016**, *116*, 088001.

(266) Widmer-Cooper, A.; Harrowell, P. On the study of collective dynamics in supercooled liquids through the statistics of the isoconfigurational ensemble. *The Journal of chemical physics* **2007**, *126*.

(267) Guiselin, B.; Scalliet, C.; Berthier, L. Microscopic origin of excess wings in relaxation spectra of supercooled liquids. *Nature Physics* **2022**, *18*, 468–472.

(268) Cubuk, E. D.; Schoenholz, S. S.; Rieser, J. M.; Malone, B. D.; Rottler, J.; Durian, D. J.; Kaxiras, E.; Liu, A. J. Identifying structural flow defects in disordered solids using machine-learning methods. *Physical review letters* **2015**, *114*, 108001.

(269) Schoenholz, S. S.; Cubuk, E. D.; Sussman, D. M.; Kaxiras, E.; Liu, A. J. A structural approach to relaxation in glassy liquids. *Nature Physics* **2016**, *12*, 469–471.

(270) Yang, Z.-Y.; Wei, D.; Zaccone, A.; Wang, Y.-J. Machine-learning integrated glassy defect from an intricate configurational-thermodynamic-dynamic space. *Physical Review B* **2021**, *104*, 064108.

(271) Alkemade, R. M.; Smallenburg, F.; Filion, L. Improving the prediction of glassy dynamics by pinpointing the local cage. *The Journal of Chemical Physics* **2023**, *158*.

(272) Shiba, H.; Hanai, M.; Suzumura, T.; Shimokawabe, T. Botan: Bond targeting network for prediction of slow glassy dynamics by machine learning relative motion. *The Journal of Chemical Physics* **2023**, *158*.

(273) Jiang, X.; Tian, Z.; Li, K.; Hu, W. A geometry-enhanced graph neural network for learning the smoothness of glassy dynamics from static structure. *The Journal of Chemical Physics* **2023**, *159*.

(274) Oyama, N.; Koyama, S.; Kawasaki, T. What do deep neural networks find in disordered structures of glasses? *Frontiers in Physics* **2023**, *10*, 1007861.

(275) Hotelling, H. In *Breakthroughs in statistics: methodology and distribution*; Springer: 1992, pp 162–190.

(276) Kob, W.; Andersen, H. C. Testing mode-coupling theory for a supercooled binary Lennard-Jones mixture I: The van Hove correlation function. *Physical Review E* **1995**, *51*, 4626.

(277) Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *Journal of computational physics* **1995**, *117*, 1–19.

(278) Wahnström, G. Molecular-dynamics study of a supercooled two-component Lennard-Jones system. *Physical Review A* **1991**, *44*, 3752.

(279) Anto-Sztrikacs, N.; Segal, D. Capturing non-Markovian dynamics with the reaction coordinate method. *Physical Review A* **2021**, *104*, 052617.

(280) Ceriotti, M.; Bussi, G.; Parrinello, M. Langevin Equation with Colored Noise for Constant-Temperature Molecular Dynamics Simulations. *Physical review letters* **2009**, *102*, 020601.