Thesis for The Degree of Licentiate of Philosophy

Unsupervised Learning of Biomolecular Dynamics with Multi-Modal Data

Christopher Kolloff

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg
Gothenburg, Sweden, 2024

Unsupervised Learning of Biomolecular Dynamics with Multi-Modal Data

Christopher Kolloff

ISSN 1652-876X

Department of Computer Science and Engineering
Division of Data Science and AI
Chalmers University of Technology | University of Gothenburg
SE-412 96 Göteborg,
Sweden
Phone: +46(0)31 772 1000

Printed by Chalmers Digitaltryck,
Gothenburg, Sweden 2024.

Till min familj.

Unsupervised Learning of Biomolecular Dynamics with Multi-Modal Data

Christopher Kolloff

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg

Abstract

The functioning of cells critically depends on the dynamics of biomolecular systems, such as proteins or nucleic acids. Biophysical experiments as well as Molecular Dynamics (MD) simulations are the primary techniques to model and understand the kinetics and thermodynamics of biomolecules. Despite their shared focus on molecular dynamics, their results often yield differing conclusions due to computational or observational limitations. Combining the two approaches in multi-modal models leads to a more accurate kinetic and thermodynamic understanding of the systems by compensating for their respective weaknesses. However, this integration presents its own set of challenges due to the differences in resolution and timescales between experimental data and MD simulations. In this thesis, we explore the reconciliation of simulation data with experimental evidence as well as the potential of machine learning (ML) to alleviate some of MD's fundamental problems. By incorporating experimental constraints, we demonstrate how integrative kinetic models are more accurate with respect to the "true" ensemble while retaining atomic-level detail. Additionally, we discuss ML's evolving role in the analysis of MD simulation and its potential as an independent method for sampling molecular conformations. The work concludes by highlighting current limitations and future directions for these integrative approaches and proposes potential remedies for ML models to achieve enhanced accuracy and generalizability across different chemical spaces, physical conditions, and timescales. These approaches offer the potential to provide deeper insights into the complex dynamics of biomolecules, which has profound implications for drug design and our understanding biological processes.

Keywords

biomolecular dynamics, data-driven modeling, integrative structural biology, statistical mechanics, generative modeling, augmented Markov models, multi-modal machine learning

List of Publications

Appended publications

This thesis is based on the following publications:

[**Paper I**] **C. Kolloff**, S. Olsson. "Rescuing off-equilibrium simulation data through dynamic experimental data with dynAMMo", *Machine Learning: Science and Technology, Vol. 4, No. 4 (December 2023) doi.org/10.1088/2632-2153/ad10ce.*

[**Paper II**] **C. Kolloff**, S. Olsson. "Machine Learning in Molecular Dynamics Simulations of Biomolecular Systems", *Comprehensive Computational Chemistry, edited by Russell J. Boyd and Manuel Yanez, Elsevier, October 19, 2023, p. 475-492, 978-0-12-823256-9 (ISBN)*

.

Other publications

The following publications were published during my PhD studies, or are to be submitted/currently under revision. However, they are not appended to this thesis, due to contents overlapping that of appended publications or contents not related to the thesis.

[**a**] **C. Kolloff**, A. Mazur, J. K. Marzinek, P. J. Bond, S. Olsson, S. Hiller. "Motional clustering in supra-τ_c conformational exchange influences NOE cross-relaxation rate", *Journal of Magnetic Resonance Volume 338, 2022, 107196, doi.org/10.1016/j.jmr.2022.107196.*

[**b**] M. Bostock, **C. Kolloff**, E. Ilyukhina, S. Asami, S. Olsson, A. Skerra, M. Sattler. "Dynamics in the Protein Dynamics in the Binding of Anticalin D6.2 to the Plant Toxin Colchicine", *to be submitted*
.

Acknowledgment

First and foremost, I would like to thank my supervisor and mentor, Simon Olsson. Through his support and leadership, I was able to do the things I did during this first part of my PhD. I could not have asked for a more inspiring and supportive supervisor. Second of all, I would like to thank my committee, consisting of Devdatt Dubhashi and Alexander Schliep, who have been supporting me from the beginning in everything I wanted to do. I would also like to thank Max Bonomi for agreeing to lead the discussion of the defense.

Furthermore, I want to acknowledge the administration in supporting me in whatever capacity I needed. Special thanks to Fatima Hersi, Kolbjörn Tunström Clara Oders and Jenny Lind, and Samuel Marawgeh. I would not be at the point I am without all of my colleagues and friends from the Data Science and AI division, especially Lena, Juan, Télio, Riccardo, Mathias, Alec and Filip, and Lovisa. Of course, I also want to thank all members (current and previous) from the CSE PhD council, whom I've had the pleasure of working together with. I would also like to thank Graham Kemp, Yinan Yu, Richard Johansson, Fredrik Johansson, and Ashkan Panahi in their capacity as course responsibles during my time as a teaching assistant. I want to thank Rocío Mercado and Adel Daoud for helping me prepare for my Fulbright fellowship interview, Liliane Järmann from the Fulbright office at swissuniversities as well as the staff from the US Embassy in Bern for their support during the application process. Next, I want to acknowledge my great collaborators, Sebastian Hiller, Adam Mazur, Peter Bond, Jan Marzinek, Mark Bostock, and Michael Sattler, with whom I've had the pleasure of having interesting scientific discussions. My PhD was generously funded by the Wallenberg AI, Autonomous Systems and Software Program (WASP), which I am immensely grateful for because it gives me the opportunity to do the science that I love.

Lastly, I extend my heartfelt gratitude to my friends and family in Switzerland for their unwavering support during my transition to Sweden, and for always being there for me. Equally, I am deeply thankful to my friends in Sweden for their companionship and encouragement. Last but certainly not least, I wish to express my profound appreciation to my family here: my partner Jonathan and our dog Kikki.

The love and support I have received from each of you have been indispensable in making this journey possible.

Contents

Part I

Introductory Chapters

Chapter 1

Introduction

In his eloquent reflection on the natural world, Carl Sagan once remarked, "The beauty of a living thing is not the atoms that go into it, but the way those atoms are put together" [1]. This statement rings especially true in the context of large biomolecules, like proteins or nucleic acids, which are comprised of thousands of atoms. Expanding upon Sagan's thought, I would add that the beauty of life is not just about the mere assembly of these atoms. It is the dynamic interactions they engage in that constitutes the very essence of life. Thanks to the dynamics of biomolecules, proteins are able to break down food into its fundamental components; cells can signal when to divide and when not to; and our bodies are able to fight off pathogens in order not to get sick.

The field of biomolecular dynamics is inherently interdisciplinary in its attempt to break down, model, and understand the mechanisms that determine the behavior of macromolecules. This interdisciplinarity is also needed to fully capture the complexity of these systems. In this thesis, we will focus on a small part in this big endeavor. We will look at how to model biomolecular dynamics from a theoretical perspective in a way that allows us to understand the fundamental problems we need to address in order to solve them. We will also look at computational approaches that have been developed to tackle some of these problems, those include Molecular Dynamics (MD) simulation in conjunction with and orthogonal to Machine Learning (ML). Finally, we will consider experimental approaches and especially how the knowledge we gain from experiments can be integrated into computational models.

MD simulations have been developed starting in the early 1950s but the earliest models were severely limited by the computational capacities [2]. It was only in the 1970s that Andrew McCammon, Bruce Gelin, and Martin Karplus developed to study biomolecules [3–5]. Thanks to increasingly better software and hardware, simulations could attain longer timescales, but were still limited to picoseconds [3]. With the rise of specialized software and Force Fields (FF), the method gained a lot of traction to study systems at even longer timescales and bigger systems. With the advent of High-Performance Clusters (HPC) and supercomputers specifically designed for MD [6], several simulations with a trajectory length of tens to hundreds of microseconds are

attainable for standard biomolecules [6].

In the past 15 years, machine learning has emerged as a powerful tool to mitigate some of the problems we face in MD (Section 2.3). Through generative modeling, we have begun to develop independent methods that structurally do not depend on MD simulations [7, 8], although they are still trained using MD data.

Nevertheless, even with the integration of ML into MD and the development of generative models in machine learning, there is a discrepancy between the timescales we are typically interested in for biomolecular systems and the timescales attainable by simulations [9]. Also, as we will see in Section 2.1, modeling the energy function $U(\mathbf{x})$ for a system in a configuration $\mathbf{x}$ turns out to be a hard problem, leading to further discrepancies between experimental observations and the predictions from simulations.

Naturally, it makes sense to combine the two sources of information to improve our models. However, this is also not straight-forward and requires developing many new techniques based on the existing ones. Hence, we discuss and propose ways how experimental observables can be integrated into models derived from simulation data (Section 2.4).

After a summary of my contributions in the field, that is, the development of dynamic Augmented Markov Models – a tool to integrate dynamic experimental data into kinetic models estimated using simulation statistics [10] – as well as a book chapter on "Machine Learning in Molecular Dynamics Simulations of Biomolecular Systems" [11], we will then draw some conclusions and propose future directions (Section 4), where I discuss the most urgent questions in the field at the moment, propose remedies and solutions, and give an outline on what I will be focusing on for the next phase in my PhD.

Chapter 2

Background

2.1 MD Simulations and Experiments: Different Perspectives on the Same Problem

Molecular dynamics simulations and certain biophysical experiments share a common objective: to unravel the dynamics of proteins or nucleic acids. Dynamics encompass the multitude of ways in which these physical systems move, interact, and transition between different states. In order to illustrate what we mean when we talk about dynamics, we shall have a look at Figure 2.1. It shows a 1D energy potential as a function of the state space. There are several energy minima separated by high-energy barriers. The height of these barriers determines the probability/rate at which transitions between the energy minima occur (described by *kinetics*). The depth of the minima reflects the relative stability of the different conformations. A lower minimum means that it is more stable and therefore more populated at equilibrium (described by *thermodynamics*). Naturally, the energy landscape of biomolecules is much more complicated and much more high dimensional. However, the concept of dynamics stays the same. There are two main approaches with which we can investigate the dynamics of biomolecular systems. The first approach is to simulate the system's dynamics computationally by calculating and propagating the forces acting on each atom over a period of time [12]. The second one entails conducting physical experiments to investigate the interactions between different parts of the molecule, gathering data that reflect the dynamics of the "true" molecular ensemble. While both approaches study the same dynamical system, the conclusions we draw from the analysis of the experiments and simulations often differ substantially [13]. In this chapter, I will discuss how these approaches offer insights at varying levels of detail, timescales, and accuracy. I aim to demonstrate that both simulations and experiments, while providing different perspectives, are equally crucial and complementary in enhancing our understanding of molecular dynamics.

2.1.1 MD Simulations

Molecular Dynamics simulations are the principal method for studying the behavior of molecules at atomic resolution [4, 14–16]. The movement of molecules is governed by physical laws, typically approximated using Newton's equations of motion in MD simulations. MD integrators (e.g., Verlet or Langevin) numerically approximate these equations over discrete time steps to simulate the trajectory of molecules. These integration steps are on the femtosecond (i.e., 10^{-15} s) timescale. A key aspect of simulating the (long-timescale) behavior of molecules is modeling its energy landscape. In MD, the potential energy of the system $U(\mathbf{x})$ is calculated for every atom in the molecule using empirically parameterized force fields $\mathbf{F}(\mathbf{x})$. These force fields move the atoms towards the negative gradient of the energy, effectively guiding them to thermodynamically stable configurations. The simulation is set up in a way that it samples from the Boltzmann distribution as τ goes to infinity:

$$p(\mathbf{x}) \propto \exp\left(-\frac{U(\mathbf{x})}{k_{\mathrm{B}}T}\right). \tag{2.1}$$

To understand why, we first have to consider the dynamics of a single particle. Its behavior is modeled using Langevin dynamics that are based on Newton's laws of motion in classical simulations. The Langevin equation [17] describes the deterministic and stochastic forces acting on the particle:

$$m\frac{\partial^2\mathbf{x}}{\partial t^2} = \underbrace{-\nabla U(\mathbf{x})}_{\text{deterministic term}} + \text{friction term} + \text{stochastic term} \tag{2.2}$$

$$= \mathbf{F}(\mathbf{x}), \tag{2.3}$$

where m is the molecule's mass and $\mathbf{x}$ its configuration. The term $m\frac{\partial^2\mathbf{x}}{\partial t^2}$ represents the momentum of the molecule, reflecting the change in velocity of the molecule over time. The Langevin equation can be recast to describe the temporal evolution of the probability density function for the position of the molecule, denoted as $p_\tau(\mathbf{x}_{t+\tau}|\mathbf{x}_t)$. This reformulation particularly important when modeling an ensemble of infinitely many non-interacting copies of the same system. The so-called Fokker-Planck equation [18, 19] has the form:

$$\partial_t p_\tau(\mathbf{x}_{t+\tau}|\mathbf{x}_t) = \nabla\cdot[-p_\tau(\mathbf{x}_{t+\tau}|\mathbf{x}_t)\beta\nabla U(\mathbf{x}) + \nabla p_\tau(\mathbf{x}_{t+\tau}|\mathbf{x}_t)] \tag{2.4}$$

$$= \nabla\cdot[p_\tau(\mathbf{x}_{t+\tau}|\mathbf{x}_t)\beta\mathbf{F}(\mathbf{x}) + \nabla p_\tau(\mathbf{x}_{t+\tau}|\mathbf{x}_t)], \tag{2.5}$$

where $\beta = (k_B T)^{-1}$ is the Boltzmann factor. In this formulation, the first term represents the drift (deterministic part), and the second term represents the diffusion (stochastic part). As we will examine in Section 2.3.2, if a dynamical system is governed by the Fokker-Planck equation, the evolution of the probability density can be described using the transfer operator formalism. At equilibrium, when the time derivative ∂_t approaches zero, the Fokker-Planck equation yields the stationary solution, which corresponds to the Boltzmann distribution:

$$\lim_{\tau \to \infty} p_\tau(\mathbf{x}_{t+\tau}|\mathbf{x}_t) = \pi(\mathbf{x}) \propto \exp\left(-\frac{U(\mathbf{x})}{k_B T}\right), \tag{2.6}$$

This implies that, theoretically, running the simulation for an infinite amount of time would result in sampling the stationary distribution $\pi(\mathbf{x})$ of the system. The probability of observing the system in a particular configuration $\mathbf{x}$ is proportional to the exponential of the negative energy $U(\mathbf{x})$ at a given temperature T. This reflects the thermodynamic propensity of molecular systems to occupy lower energy states.

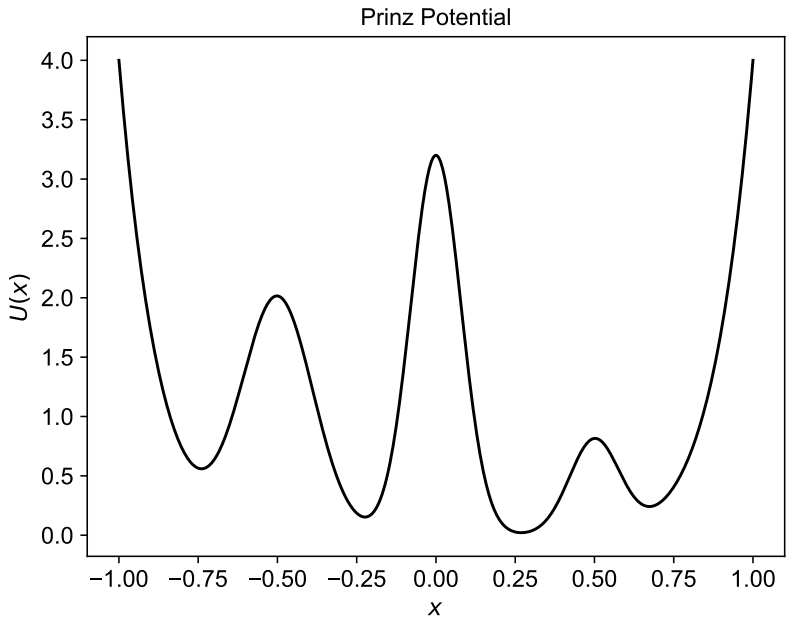

Figure 2.1: Example of a one-dimensional energy landscape using the Prinz potential [20]. The plot illustrates a series of energy minima and high-energy barriers separating them.

In practice, running a simulation until the system reaches equilibrium is computationally infeasible to do so for even small systems – much less for most proteins or nucleic acids, even for state-of-the-art special-purpose high-performance computing clusters [6]. Classical force fields, such as CHARMM [21] and AMBER [22] typically include intra- and intermolecular interaction terms to model the forces acting upon the atoms. Those terms include, among others, long-range electrostatic forces such as Coulomb energy and van der Waals forces, and harmonic elements like bonds, angles, and torsions. In order to access the timescales required to study most biological processes, these force fields are so-called classical approximations, as they do not rely on Density Functional Theory (DFT) calculations [23]. Finite simulation lengths

and the approximations of empirically parameterized force fields have several implications for studying the long-timescale behavior of macromolecules:

1. Many, if not most, biological processes of interest occur on the micro- to millisecond ($10^{-6}-10^{-3}$ s) timescale, some – like drug–target dissociations – take seconds, minutes or hours ($10^{0}-10^{4}$ s) [24], while the integration steps between MD frames are on the femtosecond (10^{-15} s) timescale. Due to the inherent computational limitations, simulating such processes may take years to simulate.

2. The number of pairwise (non-bonded) interactions scales quadratically with the number of atoms, thus rendering the study of large molecules or ensembles computationally demanding.

3. The conformational landscape of biomolecules can be frustrated, making the full exploration of state space nearly impossible. If the simulations are not long enough or if we are simply unlucky (due to the stochastic part of the MD integrator), we may not sample all relevant states. This is illustrated in Figure 2.1 using the example of a 1D energy potential. The deeper the energy minima and the higher the barriers between them, the longer the simulations have to be in order to fully characterize the energy landscape.

4. Even if we sample all relevant states, we often need to observe the transitions between them several times in order to estimate reliable statistical models of kinetic exchange.

5. Modeling the potential energy of the system $U(\mathbf{x})$ is critical as it represents the energy landscape in which the molecule moves. In MD, we approximate U using empirically parameterized force fields that drive the transition density generator, i.e., the MD integrator, to equilibrium (Eq. 2.3, 2.5, and 2.6). Due to the aforementioned FF approximations, predictions derived from simulations often do not match experimental observations.

Despite these limitations, MD simulations provide valuable insights into the atomic details of molecular dynamics [25–30] and many efforts have been made in order to mitigate these problems [31, 32]. To further enhance our understanding of biomolecular dynamics, it is essential to complement these computational insights with experimental investigations. Laboratory experiments offer empirical data, adding an additional layers of validity to our models as well as information that simulations alone may not fully capture.

2.1.2 Experiments

Biophysical techniques, such as Nuclear Magnetic Resonance (NMR) or fluorescence spectroscopy, cryo-Electron Microscopy (cryo-EM) or Small-Angle X-ray/Neutron Scattering (SAXS/SANS) have been developed to study various aspects of conformational dynamics [24, 33–38]. In particular, NMR has the

unique ability to provide atomic-level resolution and detailed dynamic information about molecules in solution, offering insights into their structure, dynamics, and interactions over a wide range of timescales [24, 39]. However, there a several limitations that make the study of biomolecules using experiments difficult:

1. Many biophysical techniques, like NMR and cryo-EM, have inherent limitations in their temporal resolution. This makes capturing rapid molecular dynamics and transient states, which often occur on short timescales, quite challenging. Similarly, SAXS tends to yield data at a lower spatial resolution, typically providing averaged structural information, rather than precise atomic positions. This can limit the level of detail available from experimental approaches.

2. It is often hard to control experimental conditions, such that the experiment can be repeated with high statistical confidence. Also, artifacts, like crystal packing contacts during crystallization or high concentrations that are needed for NMR experiments may not reflect physiological conditions, thus tainting our conclusions.

3. Many experimental techniques provide ensemble-averaged data, which can mask the heterogeneity and the presence of rare or transient states in biomolecular systems that may play an important role in the molecule's function.

Nevertheless, experiments offer a unique access to the "true" molecular ensemble that can yield important insights into the thermodynamics and kinetics of molecular systems. By looking at the advantages and disadvantages of biophysical simulations and experiments, respectively, it is clear that one complements the other: the wide range of timescales accessible through experiments can offset the computational constraints inherent in simulations. Conversely, the atomic details available through MD can complement the lack thereof in experiments. This synergistic relationship between simulations and experiments is key to a comprehensive understanding of biomolecular dynamics [40–48]. Now, we will look at how we can compare kinetic/thermodynamic quantities predicted from simulations with their experimental counterparts.

2.2 Stationary and Dynamic Observables

In statistical mechanics, distinguishing between stationary (or static) and dynamic observables is fundamental in studying the macroscopic behavior of dynamical systems. Observables are physical quantities that can be either measured in experiments or calculated from simulations, thus providing a link between the two approaches. Stationary observables are quantities of a system that stay constant with respect to time, meaning that they report on the equilibrium properties of the system. Those include but are not limited to (melting) temperature, secondary-structure content or molecular weight. For some function $f : X \to \mathbb{C}^n$ or $\mathbb{R}^n$, where $X = \{\mathbf{x} \in \mathbb{R}^{3N} \mid \mathbf{x} \in \Omega\}$ with $\mathbf{x}$ being

the Cartesian coordinates of the configuration, the corresponding stationary observable o_f^{stat} is

$$o_f^{\text{stat}} = \int_\Omega f(\mathbf{x})\pi(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{2.7}$$

Here, Ω is the state space of the system. Eq. 2.7 can be understood as follows: f maps every configuration $\mathbf{x} \in \Omega$ to some specific value. For most observables, such as average helicity or melting temperature, o_f is a scalar, thus, $n = 1$. These observables can be measured through averaging over a long period of time (ensemble averaging), thus giving us access to the thermodynamic properties of the system.

Dynamic observables, on the other hand, are a way of measuring how the system evolves over time and are thus time dependent. For example, a common dynamic observable in biomolecular experiments is to compare the position of a molecule with its position at a later time point (we call this time lag). This so-called autocorrelation function contains the information of how quickly the molecules switches from one state to another or how long it stays in a particular state. We often do not measure the time correlations directly but through some forward model (see Section 2.2.1 for details). The observable calculates the weighted correlation between some functions f and g over time, integrated over all possible states:

$$o_{fg}^{\text{dyn}}(k\tau) = \int_\Omega \int_\Omega f(\mathbf{x}_t) g(\mathbf{x}_{t+k\tau}) p_\tau(\mathbf{x}_{t+k\tau}|\mathbf{x}_t)\boldsymbol{\pi}(\mathbf{x})\,\mathrm{d}\mathbf{x}_t\,\mathrm{d}\mathbf{x}_{t+k\tau}, \tag{2.8}$$

where τ refers to some set time lag and $p_\tau(\mathbf{x}_{t+k\tau}|\mathbf{x}_t)$ is the transition density. k is an integer multiple of τ. However, for simplicity and clarity, k is omitted in the remainder of the thesis. o_{fg} is called and auto-correlation function if $f = g$ and a cross-correlation function otherwise. Dynamic observables are particularly useful when studying relaxation processes or the spectral properties of systems (e.g., the timescales of exchange). Examples of techniques that belong to the class of dynamic observables include T_1 and T_2 relaxation [49] as well as relaxation dispersion techniques (such as $R_{1\rho}$ [50, 51] or Carl-Purcell-Meiboom-Gill (CPMG) [52, 53]) and (single-molecule) Fluorescence Resonance Energy Transfer (smFRET) [54, 55]. None of these techniques are measuring the time correlation directly. $R_{1\rho}$/CPMG relaxation dispersion, for example, measure the Fourier transform of the autocorrelation function [50, 52, 53, 56]. We will now look at how to compare quantities, like relaxation dispersion, obtained from experiments with those predicted from computational models.

2.2.1 Link between Experiments and Simulation

Correlation functions are crucial in quantitatively describing how physical properties evolve over time (see Eq. 2.8). A key aspect of modeling the dynamics of a molecule is to analyze its relaxation behavior, i.e., transitions between molecular conformations. One fundamental assumption we make about systems in equilibrium is the Markov property, which posits that the future state of a system only depends on its present state and not on its past history.

Under this assumption, we can decompose the overall relaxation behavior into the sum of individual relaxation processes, whereby each contribution can be characterized by its own timescale t_i^{ex} and amplitude c_i. Mathematically, this can be represented as:

$$C(\tau) = \sum_i c_i e^{-\tau/t_i^{\text{ex}}}. \tag{2.9}$$

Note that while functionally different, Eq. 2.8 and 2.9 are the same. Eq. 2.9 provides the general description of a correlation function, whereas Eq. 2.8 views the correlation through the lens of the two observable functions $f(\cdot)$ and $g(\cdot)$.

To illustrate the relationship between MD-derived predictions and experimental observations, consider this example of an NMR experiment that characterizes the conformational exchange of molecules on the micro- to millisecond timescale, called $R_{1\rho}$ relaxation dispersion [47, 51, 57]:

$$R_{1\rho}(\nu_1) = (2\pi\nu_0)^2 \int_0^\infty C(\tau)\cos(2\pi\nu_1\tau)\,\mathrm{d}\tau, \tag{2.10}$$

where the factor $(2\pi\nu_0)^2$ is related to the Larmor frequency of nuclei under investigation and ν_1 is the probing frequency. Importantly, Eq. 2.10 gives us a framework to

1. compute the relaxation rates directly from the MD simulations, and
2. compare the predictions from MD directly with the observations obtained from the experiments.

It is important to stress the complementary nature of the two approaches: From the experiments, we obtain the convolution of all exchange processes of the "true" ensemble but lack the atomistic details to be able to explain them. From simulations, we have full observability and are able to perfectly explain all the different exchange processes. However, due to the aforementioned issues with sampling and force field inaccuracies, we often lack the ability to confidently assert that the MD simulations accurately represent the true dynamics of the biomolecular system in its natural state.

This underlines the need for and importance of combining experimental data with simulations. The goal therefore is to obtain detailed mechanistic insights into the different exchange contributions from MD, all while ensuring that the models we estimate are in fact consistent with the experimental data available. Having truly integrative statistical models thus has the potential for a more comprehensive understanding of (bio-)molecular dynamics.

2.3 Application of Machine Learning to Address MD Limitations

So far, we have established the main limitations of MD simulations as well as the usefulness of including experimental observables in obtaining more accurate

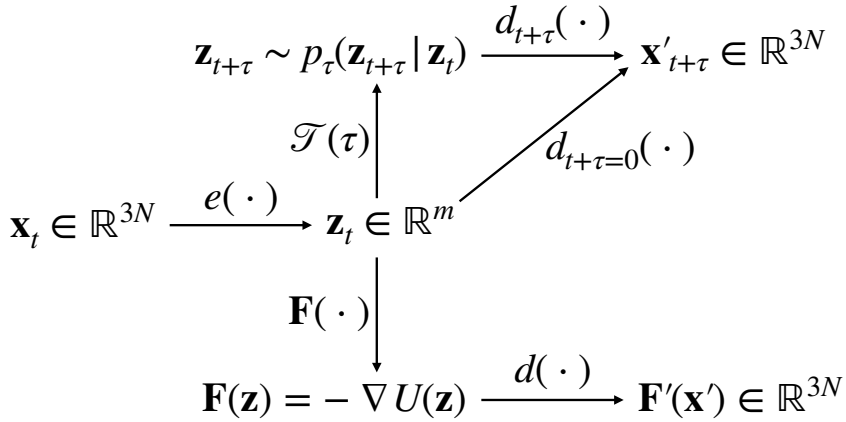

Figure 2.2: Schematic depicting machine learning applications in molecular dynamics. The configuration of a system in Cartesian coordinates is denoted by $\mathbf{x}$, but a more common approach involves working with a lower-dimensional representation $\mathbf{z}$ (where $N > m$), obtained through an encoding operation $e(\cdot)$. The key role of ML in MD includes approximating the transfer operator $\mathcal{T}(\tau)$ for modeling transition density $p_\tau(\mathbf{z}_{t+\tau}|\mathbf{z}_t)$ as well as in predicting the force field $\mathbf{F}(\mathbf{z})$ as the negative gradient of energy $-\nabla U(\mathbf{z})$. Note that $\mathbf{z}$ may be a lower-dimensional representation (e.g., coarse-graining) but it can also include all-atom force field modeling if $m = N$. Finally, ML is used in the decoding process $d_{t+\tau}(\cdot)$, translating the latent representation back to $\mathbf{x}'_{t+\tau}$ in configuration space. When $\tau \neq 0$, this involves reconstructing the time-lagged latent sample. In the same way, ML aids in mapping the force of the coarse-grained representation back to its all-atom equivalent $\mathbf{F}'(\mathbf{x}')$.

mechanistic models of the system's thermodynamic as well as kinetic properties. Now, we will look at how machine learning can be useful in addressing the drawbacks that we face in MD simulations and finally, how experiments can be integrated into ML techniques.

2.3.1 Encoding of Molecular Configurations

Molecular configurations typically live in a very high-dimensional space, and there are several ways of representing them [58, 59]. In classical MD simulations, molecules are represented using their Cartesian coordinates, i.e., $\mathbf{x} \in \mathbb{R}^{3N}$, where N is the number of atoms. Using the same representation for machine learning turns out to be unnecessarily hard and impractical as the machine learning model might incorrectly interpret rotations or translations of the molecule as meaningful changes in the system, whereas they are not. Some quantities, such as the energy or the molecular weight of a molecule, do not change whether the molecule is transformed, that is rotated or translated. We

say that this quantity is *invariant* under some transformation. Mathematically, we want to learn a representation such that the output of some function $f(\cdot)$, e.g., energy, applied to the configuration $\mathbf{x}$ is the same when we apply some transformation $g(\cdot)$: $f(g(\mathbf{x})) = f(\mathbf{x})$. Similarly, other physical quantities, like force, change with the transformation of the molecule in a predictable way. This is referred to as *equivariance*, and we can express it mathematically as $g(f(\mathbf{x})) = f(g(\mathbf{x}))$. The field in machine learning that is concerned with incorporating symmetry information into neural network architectures is called geometric deep learning [58]. ML can therefore help us to find a potentially lower-dimensional *encoding* $\mathbf{z}_t = e(\mathbf{x}_t)$ of our configuration that inherently respects physical symmetries (Figure 2.2). Learning the essential features of a molecule reduces the computational complexity and therefore increases the learning efficiency [60–67]. Another, orthogonal approach of identifying a lower-dimensional encoding is often referred to as *coarse graining* in the MD community [68, 69], and it enables us to study larger molecular systems at longer timescales [70]. The idea is that not all atoms are important in defining the energy landscape of the molecule; and by pooling groups of atoms into "beads", it is possible to run longer simulations for the same amount of computational power. In particular, graph neural networks have played an important role in learning a coarse-grained representation of biomolecules [69].

We will now shift our attention to understanding the framework that lies behind models we can estimate from some encoded latent representation $\mathbf{z}$. These models serve, both, for our understanding of molecular kinetics (Section 2.3.3) as well as for sampling time-lagged conformations (Section 2.3.4).

2.3.2 Transfer Operator Formalism

Recall that one primary objective of molecular dynamics simulations is to learn a model of the probability density $p_\tau(\mathbf{z}_{t+\tau}|\mathbf{z}_t)$. In Section 2.1.1, we introduced the Fokker-Planck equation (Eq. 2.5), a differential equation describing the evolution of the probability density of a molecule's position over time. However, we can also take a different approach of modeling the evolution of that distribution. Instead of using a differential equation, we can describe how a system evolves over time using an operator-based framework. That is, we can construct a so-called transfer operator [71, 72] such that it models the same underlying physical process as described by the Fokker-Planck equation. In this section, we will look at how to approximate this transfer operator $\mathcal{T}_\Omega$ using machine learning. This can be helpful for estimating kinetic models or sampling from the probability density. $\mathcal{T}_\Omega$ is a mathematical construct that describes how the probability density p_τ evolves in the state space Ω over time τ. Mathematically, the relationship between the transfer operator and the probability density can be described as [71, 72]:

$$p_\tau(\mathbf{z}_{t+\tau}|\mathbf{z}_t) = \int_\Omega p(\mathbf{z}_t) p_\tau(\mathbf{z}_{t+\tau}|\mathbf{z}_t)\,\mathrm{d}\mathbf{z} \tag{2.11}$$

$$= \mathcal{T}_\Omega(\tau) p(\mathbf{z}_t) \tag{2.12}$$

$$= \sum_{i=1}^{\infty} \lambda_i(\tau)\,|\boldsymbol{\phi}_i\rangle\,\langle\boldsymbol{\psi}_i|\,. \tag{2.13}$$

The last equation is particularly useful as it allows us to model $\mathcal{T}$ using its *spectral decomposition*. It can be interpreted as follows: A dynamical system can be described by a superposition of i processes that are associated with a pair of eigenvalues λ_i and eigenfunctions $\boldsymbol{\psi}_i$. The eigenvalue decays exponentially at a rate that is governed by the characteristic timescale t_i^{ex} of the process:

$$\lambda_i(\tau) = \exp\left(-\frac{\tau}{t_i^{\mathrm{ex}}}\right). \tag{2.14}$$

Since we are only considering systems in equilibrium, we can assume $0 \leq |\lambda_i| \leq 1$ (conservation of probability), $\lambda_i \in \mathbb{R}$ (reversibility[1]), and $\lambda_1 = 1$, $|\lambda_{i>1}| < 1$ (ergodicity). The corresponding eigenfunction $\boldsymbol{\psi}_i$ encodes a particular mode or pattern of behavior in the system. That is, it describes the probability flow between different states. The stationary density–weighted eigenfunction $\boldsymbol{\phi}_i$ determines the amplitude of the process and is defined as

$$\boldsymbol{\phi}_i = \boldsymbol{\psi}_i \boldsymbol{\pi}. \tag{2.15}$$

This gives us a powerful framework to approximate the probability density $p_\tau(\mathbf{z}_{t+\tau}|\mathbf{z}_t)$ since we can use the spectral components to find the optimal parameters that explain the data. Furthermore, we do not need to characterize infinitely many processes but can resort to a low-rank approximation where small eigenvalues, i.e., $|\lambda_i| \ll 1$, are discarded since their corresponding timescales are fast and thus do not contribute much to the overall signal.

2.3.3 Markov State Models (MSMs)

Markov state models have shown to be very expressive in modeling biomolecular dynamics and their application to MD data have lead to many important insights [73–87]. By using a discretization of Ω into n states, we are able to approximate $\mathcal{T}$ by estimating the number of transitions between states in a count matrix $\mathbf{C} \in \mathbb{N}^{n\times n}$. At the core of the Markov state model approach is the transition matrix $\mathbf{T} \in \mathbb{R}^{n\times n}$. By enforcing physically meaningful constraints for dynamical systems in equilibrium, such as reversibility, ergodicity, and row stochasticity, $\mathbf{T}_{ij}$ will contain the transition probability of the configuration $\mathbf{z}_t^{(i)}$ at time t being in state i to its time-lagged equivalent $\mathbf{z}_{t+\tau}^{(j)}$ in state j. As we will see in Section 2.4, we can use the transfer operator formalism, and in particular Markov state modeling, to predict, both, stationary and dynamic

[1] In order for the dynamical system to be reversible, it is also necessary that $\pi(\mathbf{z}_{t+\tau}) p_\tau(\mathbf{z}_{t+\tau}|\mathbf{z}_t) = \pi(\mathbf{z}_t) p_\tau(\mathbf{z}_t|\mathbf{z}_{t+\tau})$

observables, thus allowing us to compare the predictions based on MD data with experimental data.

2.3.4 Sampling from the Boltzmann Distribution

With the transfer operator formalism as the foundation, we are now able to understand how some models apply this theory to learn a model of the probability density $p_\tau(\mathbf{z}_{t+\tau}|\mathbf{z}_t)$ which allows $\mathbf{z}_{t+\tau} \sim p_\tau(\mathbf{z}_{t+\tau}|\mathbf{z}_t)$ (Figure 2.2). This is an exceptionally difficult task, given that the conformational space is very high-dimensional and even in latent space, the energy landscape is complicated and rugged. Also, due to the exponential relationship between energy and probability (Eq. 2.1), small changes in $\mathbf{z}$ can have a big effect on the probability density. Nevertheless, machine learning offers unique ways to address the sampling time-lagged states of the system given its current state ($\tau \ll \infty$) and from the Boltzmann distribution ($\tau \to \infty$). Biomolecules often have many different low-energy regions with high energy barriers between them. These are often referred to as *metastable states* – and generating statistically independent samples from these energy minima is an important goal in the characterization and modeling of molecular dynamics [88]. The way ML can help in sampling from distributions that are difficult to sample from directly, like the Boltzmann distribution in high-dimensional spaces, is to learn how to map low-energy configurations (sampled from molecular simulations or experimental data) to a simpler latent space and back. ML aids in sampling from complicated distributions, such as the Boltzmann distribution in high-dimensional spaces, by mapping low-energy configurations (derived from molecular simulations or experimental data) to a simpler latent space and back. The two primary ML that currently are being used to achieve this are Normalizing Flows (NFs) [7, 89–93] and Denoising Diffusion Probabilistic Models (DDPMs) [8, 94–97]. Normalizing flows achieve this through a series of invertible transformations, while diffusion models do so by gradually denoising a simple distribution to the target distribution. A conceptually different approach is to learn a function that biases the energy potential to sample rare events (akin to classical *enhanced sampling* in MD simulations) [98–101]. While much effort has been put into developing these models, many open questions remain for developing models that are able to sample valid configurations across state space in a time-lagged fashion. A field where this generalization also is a key aspect is not just in sampling but in deriving a model of the forces that govern the potential energy landscape. This is the field of machine-learned force fields, which we will look at next.

2.3.5 Machine-Learned Force Fields

Traditional force fields are based on empirical data and quantum chemical calculations, thereby limiting the accuracy and scalability when applied to large, complex biomolecular systems. Machine learning offers a promising alternative by estimating force fields that can harness the vast amount of MD data to learn more accurate and scalable models [102]. The force $\mathbf{F}(\mathbf{z})$ exerted

on the atoms is defined as the negative gradient of the potential energy function $\mathbf{F}(\mathbf{z}) = -\nabla U(\mathbf{z})$. Machine learning offers a significant advantage compared to the classical, error-prone force fields (Figure 2.2). ML algorithms can learn complex patterns and dependencies from large datasets. This ability makes ML particularly well-suited for approximating the potential energy surface $U(\mathbf{z})$ of biomolecules. Once trained, ML models can efficiently compute the forces in MD simulations, thereby speeding up the simulation process while maintaining, or even improving, the accuracy of the results. It is worth mentioning that ML FF are being developed for, both, all-atom systems [102–112] as well as coarse-grained ones [69, 113, 114].

Working with the coarse-grained representation is often fine in order to understand the mechanisms-of-action. However, it can be of interest to map the coarse-grained molecule back to its all-atom equivalent, and hence, also the forces. How machine learning can help in decoding the latent representation is what concerns us next.

2.3.6 Decoding the Latent Representation back to Configuration Space

In order for our models to be interpretable and useful, we often need to decode $d(\cdot)$ the compressed, lower-dimensional latent space representation $\mathbf{z}$ into its reconstructed, high-dimensional Cartesian coordinate representation $\mathbf{x}'$. However, we need to distinguish between two tasks. The first is to reconstruct the latent sample $\mathbf{z}$ without time lag ($\tau = 0$) and the other, is to decode the time-propagated sample $\mathbf{z}_{t+\tau}$ back to configuration space. Examples of the former include neural network architectures, such as Variational Autoencoders (VAEs) that generate samples from a latent distribution, approximating the underlying data distribution [61, 66, 99]. However, the much harder case is the latter, which tries to find a way to time-propagate the latent representation and to decode it back to configuration space $\mathbf{x}'_{t+\tau} = d_{t+\tau}(\mathbf{z}_t)$ (see Section 2.3.2). While there have been approaches to learn such a decoder [115, 116], they do not seem suited to be applied to long-timescale MD data since they work well for small τ. And while techniques, such as deep generative MSMs [117], show promising results, the sampled conformations are often physically invalid. Recently, Schreiner et al. proposed a promising approach, called the Implicit Transfer Operator (ITO) to sample (multiple) time-lagged conformations with a large τ from an equivariant latent presentation [8]. In its current implementation, the method is, however, limited in size as well as its ability to generalize across chemical and thermodynamic space.

In chapter 2.3.5, we looked at how machine learning can help in estimating the force field of coarse-grained representations. Mapping this lower-dimensional representation back to the all-atom equivalent is not a trivial task since we have a many-to-one mapping: multiple all-atom configurations might correspond to the same coarse-grained state. Here, graph neural networks play an important role in learning this mapping [118–120], in particular also learning the forces that have been estimated for the coarse-grained model.

Now we have covered the different aspects, ML can be used parallel to MD

and even independent of it to study the dynamics of biomolecules. However, a big question remains: How do we ensure that our machine learning models are thermodynamically and kinetically consistent with experimental observations?

2.4 Using Experimental Constraints in the Modeling of Molecular Conformations

Building on the complementary aspects of experiments and simulations discussed in Section (Section 2.1), we have shown that through the calculation and acquisition of stationary and dynamic observables, we are able to compare MD simulations and experimental data (Section 2.2). We have also looked at the various ways, ML is able to address some of the most fundamental problems in MD (Section 2.3). We also examined the usefulness of the transfer operator formalism as a tool for modeling dynamical systems (Section 2.3.2). Now, we have the background and tools to look at how experimental constraints that report on the kinetics and thermodynamics of biomolecular systems can be incorporated into machine learning techniques.

2.4.1 Computation of stationary/dynamic observables from MSMs

Upon examining Eq. 2.9 and Eq. 2.14, we observe a fundamental connection between the eigenvalues of the transfer operator and the relaxation contributions in the time correlation functions. It highlights how the decay rates in correlation functions are intrinsically linked to the timescales defined by the eigenvalues of the transfer operator. The amplitudes in the correlation function, represented as c_i, can be efficiently computed using the spectral decomposition of the transfer operator $\mathcal{T}$, in particular its approximation, the MSM (Section 2.3.3). The amplitudes are mathematically expressed as:

$$c_i = (\mathbf{f} \cdot \boldsymbol{\phi}_i)^2, \tag{2.16}$$

where $\mathbf{f} \in \mathbb{R}^n$ is a vector representing the average value of the observable function $f(\mathbf{z})$ across all configurations of $\mathbf{z}$ within each state $i \in n$ [47, 73, 121]. ψ_i is the left eigenvector of the ith process. This gives us a way to express dynamic observables using MSMs:

$$o_{ff}^{\text{dyn, MSM}}(\tau) = c_1 + \sum_{i=2}^{n} c_i \exp\left(-\frac{\tau}{t_i^{\text{ex}}}\right), \tag{2.17}$$

where c_1 corresponds to the amplitude of the stationary process with $\lambda = 1$ and thus the stationary distribution π:

$$o_f^{\text{stat, MSM}} = \mathbf{f} \cdot \boldsymbol{\pi}. \tag{2.18}$$

2.4.2 Machine Learning with Kinetic and Thermodynamic Constraints

Obtaining a machine learning model that is consistent with, both, simulation and experimental data requires designing an objective function that is physically consistent by penalizing deviations for predictions made by the model and the experimental data available. In general, such an objective function $\mathcal{L}$ can be expressed as

$$\mathcal{L} = \mathcal{L}^{\text{data}} + \alpha\mathcal{L}^{\text{stat}} + \beta\mathcal{L}^{\text{dyn}}, \tag{2.19}$$

where α and β are both scaling factors. The specific forms of the stationary and dynamic observable loss terms, vary depending on the model's structure and the nature of the data. However, we can define a more generalized expression for these loss terms. For instance, considering stationary observables $\mathbf{o}^{\text{stat}}$, a common approach is to devise a loss function that minimizes the discrepancy between experimental observations and model predictions. Specifically, for m experimental observations of a stationary observable, the thermodynamic loss can be formulated as:

$$\mathcal{L}^{\text{stat}} = \sum_{i=1}^{m} \nu_i \| o_i^{\text{exp}} - o_i^{\text{pred}} \|^2. \tag{2.20}$$

The notation $\| \cdot \|^2$ denotes the squared Euclidean norm (L2 norm), which quantifies the squared difference between the experimental and predicted values. ν_i can be some form of Lagrange multiplier or weight that needs to be learned.

The general form for the dynamic loss term is a bit more involved since we are not just dealing with a single observation per data point but with time correlations (Eq. 2.9). Thus, for m experimental observations with k measurements per observation, we have

In the case of dynamic observables, where we rely on time correlations (Eq. 2.9), the formulation of the loss term must account for the temporal relationships within the data. For a dataset comprising m experimental observations, each consisting of k time-correlated measurements, the dynamic loss term can be structured as follows:

$$\mathcal{L}^{\text{dyn}} = \sum_{i=1}^{m} \sum_{j=1}^{k} \nu_{ij} \| o_i^{\text{exp}}(j\tau) - o_i^{\text{pred}}(j\tau) \|^2. \tag{2.21}$$

In addition to the data constraints, it is often necessary to include other, physically relevant constraints as well. Those include

- Detailed balance/Reversibility: $\pi_i p_{ij} = \pi_j p_{ji}$
- Ergodicity: $\lambda_1 = 1$, $|\lambda_{i>1}| < 1$
- Row stochasticity: $\sum_j p_{ij} = 1$

- Orthonormality of eigenvectors: $\boldsymbol{\Psi}\boldsymbol{\Psi}^\top = \mathbb{I}$, with
$$\boldsymbol{\Psi} = \begin{bmatrix} | & | & & | \\ \boldsymbol{\psi}_1 & \boldsymbol{\psi}_2 & \cdots & \boldsymbol{\psi}_n \\ | & | & & | \end{bmatrix}$$

for systems in equilibrium.

They can be enforced in several ways [122–124], however the most straightforward way of doing so is by using Lagrange multipliers and using fixed-point iteration algorithms [46] or gradient descent [10] to find the optimal parameters.

2.4.3 Balancing Experimental Priors with MD Data Using Maximum Entropy Principles

When biasing statistical models derived from classical MD data with experimental constraints, it is necessary that the inferred probability distribution is the most uniform given the imposed constraints. This ensures that we avoid unwarranted assumptions about the systems. Maximum entropy principles allow us to incorporate experimental knowledge in an information-theoretically optimal way [125, 126]. Suppose you want to find a probability distribution $p(\mathbf{x})$ over a set of n states. The maximum entropy principle states that among all distributions that satisfy the given constraints, the one that maximizes the entropy is the least biased and therefore the most preferable. The entropy H of the distribution $p(\mathbf{x})$ is given by:

$$H[p(\mathbf{x})] = -\sum_n p(\mathbf{x}) \log p(\mathbf{x}). \tag{2.22}$$

Given a set of m experimental constraints $o_i^{\exp} = \sum_n f_i(\mathbf{x})p(\mathbf{x})$ with $i \in m$ (Eq. 2.7), we want to maximize $H[p(\mathbf{x})]$ subject to these constraints. Using Lagrange multipliers λ, we can write the Lagrangian as:

$$\begin{aligned} \mathcal{L} = &-\sum_n p(\mathbf{x}) \log p(\mathbf{x}) \\ &+ \lambda_0 \left(\sum_n p(\mathbf{x}) - 1 \right) + \sum_{i=1}^{m} \lambda_i \left(\sum_n f_i(\mathbf{x}) p(\mathbf{x}) - o_i^{\exp} \right). \end{aligned} \tag{2.23}$$

The first term is the entropy defined in Eq. 2.22. The second ensures the normalization of $p(\mathbf{x})$, and the last enforces the experimental constraints. Differentiating $\mathcal{L}$ with respect to $p(\mathbf{x})$ and setting it to zero gives the maximum entropy distribution, balancing the probability $p(\mathbf{x})$ estimated from simulation data with the experimental evidence available.

Chapter 3

Summary of Included Papers

3.1 Rescuing off-equilibrium simulation data through dynamic experimental data with dynAMMo

In this paper, we developed a method to construct comprehensive mechanistic models of kinetic exchange processes in proteins. Our method integrates molecular dynamics data with dynamic experimental observations in an augmented Markov model, addressing the long-standing challenge of modeling protein dynamics over extensive timescales.

Problem

Understanding the details of biochemical processes, such as the mode-of-action of enzymes, ligand binding and unbinding as well as regulation of signaling pathways, is imperative in, for example, designing drugs that specifically target the biomolecules in question. In order to gain such understanding, the most popular approaches are either to run molecular dynamics simulations or to record biophysical experiments. Nevertheless, one of the most persistent challenges in the field of molecular dynamics is to integrate the two sources into a single kinetic model. This often results in estimating separate models based on the simulation or experimental data, respectively. Those models either lack the atomistic details necessary to explain the kinetic exchange or often fail to reflect the accurate thermodynamic and kinetic realities of the molecules. A unifying approach that utilizes the mechanistic details available from MD simulations as well as reflects the empirical robustness from dynamic experimental data in the most unbiased way is thus essential to bridge this gap in our understanding of dynamic processes.

Contribution

This paper introduces dynamic Augmented Markov Models (dynAMMo) – a method to reconcile simulation statistics with experimental data that report on the kinetics and thermodynamics of molecules in an information-theoretically sound way. By showcasing the performance of the algorithm using two theoretical model systems, we were able to show that our model accurately predicts dynamic experimental observables in two important scenarios:

1. The case where sufficient simulation data are available to reproduce the experimental evidence, however, the kinetics and thermodynamics are biased, thus resulting in a discrepancy between the predictions and the experimental observables.

2. The case where the simulations samples all thermodynamically relevant states but fails to observe the transitions between them. Prior to this method, it would not have been possible reproduce the experimental observables.

Moreover, we demonstrate the failure of our model in a third important scenario:

3. The case where not all thermodynamically important states were sampled. In this scenario, we demonstrated that the model was not able to accurately predict the experimental evidence. This shows that our method does not introduce artifacts since we have to have all relevant information in order to reproduce the experimental observation.

Those three cases are important scenarios with real-world relevance. The first one addresses the issue that force fields in MD simulations often do not reproduce correct kinetic and/or thermodynamic quantities even when all states and transitions are sampled. The second, and arguably most important, addresses the fact that MD simulations are limited in terms of the timescales of exchange that can be reached. MD simulations often do not sample the transitions between relevant states, hitherto hindering researchers in establishing useful mechanistic models. The third and last scenario is important because in the case that we do not have all relevant information available to us, it is imperative to show that the estimated model is not reliable and we can assume that given the data available, we are not able to reproduce the experimental evidence.

The last important contribution of this paper is to show that the method we have developed also works with real-world data. For this, we used Bovine Pancreatic Trypsin Inhibitor (BPTI) as an example. This is a well-studied protein system [28, 34, 127–130] where, we assume to know all relevant structural states. However, extensive experimental data show that there is a discrepancy between the estimated exchange rates [127] and the ones estimated from the simulation [28]. Using our model, we were able to correct for this bias and estimate a mechanistic model of kinetic exchange with experimental accuracy. Furthermore, by artificially recreating the "disconnected states" scenario, we

were also able to show that were were able to reproduce the kinetics from the "biased" scenario, thus underscoring the relevance of our model in a real-world scenario.

Methodology

dynAMMo is a Markov state model that is estimated from simulation and experimental data. A transition matrix $\mathbf{T}$ is first estimated using simulation statistics only. Through the spectral decomposition of $\mathbf{T}$, we get access to the (right) eigenvectors, the eigenvalues, and the stationary distribution. Through the use of an observable function $f(\cdot)$, we can calculate the stationary and dynamic observables, o^{stat} and o^{dynamic}, respectively. This allows us to match the predicted observables with the experimental ones o^{exp}. By constructing a loss function that combines the maximum likelihood of the transition counts with the mean squared error between the predictions and observed data, we can estimate a dynamic augmented Markov model that takes, both, simulation and experimental data into account. In addition, we use Lagrange multipliers to enforce constraints that adhere to the equilibrium properties of physical systems. That is, we constrain the model to be ergodic, reversible, and preserving the measure of the system's state space. Using gradient descent, the spectral components of the transition matrix as well as the Lagrange multipliers are optimized.

3.2 Machine Learning in molecular dynamics simulations of biomolecular systems

In this book chapter, we explore the significant ways in which machine learning techniques have been utilized and developed for studying molecular systems. These techniques are applied both in conjunction with and independently of molecular dynamics simulations.

Problem

Classical MD simulations are fundamentally limited in many ways that affect the characterization and modeling of biomolecular systems. This includes issues of scalability in terms of size and timescale, thermodynamic and kinetic accuracy due to empirical force field parameterization as well as sampling efficiency. In the past years, machine learning has emerged as a powerful tool to alleviate or improve those aforementioned limitations.

Contribution

This book chapter provides computational chemists with a comprehensive survey of the application of machine learning in molecular dynamics simulations of biomolecular systems. It focuses on the fundamental principles in modeling the kinetics of molecular systems. In particular, it explains why Markov state modeling is such a successful and popular approach in doing so. Furthermore, we highlight different sampling techniques and neural network architectures and how they are been used to sample metastable conformations. It also outlines how coarse-grained force fields are being estimated. Finally, we identify the current challenges and questions in the field and what is needed to further improve the models and thus our understanding of biomolecular dynamics.

Chapter 4

Concluding Remarks and Future Directions

Using experimental observations as prior information to enhance the accuracy of models based on Molecular Dynamics (MD) data has significantly advanced the scientific community's knowledge of biomolecular systems [131–133]. Nevertheless, as the arsenal of techniques from the ML community grows, more and better solutions need to be found to profit from the synergy of combining thermodynamic and kinetic prior information with computational data (both from MD simulations as well as generated samples from ML models). Addressing the limitations and future directions of all methods related to ML in MD is beyond the scope of this thesis. However, I will discuss the most pressing aspects in this regard.

A remaining concern in the fields of machine-learned force fields (Section 2.3.5) as well as generative sampling of molecules (Section 2.3.4) is transferability with respect to chemical space, different physical conditions (temperature, pressure, pH, etc.), and timescales. Transferability across chemical space is the ability of ML models to generalize across different types of molecules and chemical compositions. This feature is essential to avoid having to retrain the models for every new molecule encountered. The thousands of structures available in the Protein Data Bank (PDB) as well as the AlphaFold protein structure database [134] together with the increasingly faster and more accessible high-performance computing, there is a big potential to make ML FF or generative models capable of making useful predictions for previously unseen molecules.

Secondly, The probability density of the system's states is highly sensitive to variations in external parameters. For example, small changes in the temperature can significantly alter the thermodynamics and kinetics of molecular systems. It is therefore necessary to work on developing models that are capable of interpolating between a variety of environmental conditions, such as temperature, pressure, or concentration. A recently introduced framework [135] shows promising potential to address this problem. Albergo et al. proposed stochastic interpolants that can bridge two probability density functions by learning a

vector field (*flow*) between them. By including experimental observables, such as dynamic observables at different temperatures, there is the potential to vastly improve these models.

Furthermore, there are still challenges in finding meaningful and interpretable representation of biomolecules that need to be addressed. First of all, more work needs to be done on finding representations that account for physical symmetries as well asthe time dependence in MD data. It has been demonstrated that it is possible to include time lags in the encoding of MD data through Variational Autoencoders (VAE) [66] by including calculations of autocorrelation functions. It would be worth pursuing to also integrate stationary or dynamic experimental observables in the hope to improve the latent representation of the system. Including experimental constraints when learning (equivariant) representations has several advantages. First, this would ensure that the representation is meaningful in that it aligns with real-world observations, making the model more accurate. Second, the combination of equivariant representations and adherence to experimental data can lead to models that generalize better across different molecular systems – addressing the issue of generalization.

Lastly, there is a big potential for further exploiting the transfer operator framework in sampling time-lagged conformations by expanding the Implicit Transfer Operator (ITO) framework [8]. Instead of implicitly learning the sum of all (slow) exchange contributions, one can learn the slow modes explicitly. This would allow us greater access and improve our understanding of the underlying kinetics of the system. By identifying the slowest eigenfunctions, we gain an understanding what the most significant structural transitions are and how slow they are. However, this means that it is necessary to also learn the eigenvalues of the corresponding eigenfunctions, and hence, also the timescales (Eq. 2.14). Taking inspiration from the dynAMMo approach [10], it is conceivable to augment the model by including dynamic experimental observables to match the predictions made by the model with the experimental expectations.

In summary, while the progress in recent years in terms of modeling biomolecular systems has been substantial, there is still much room for improvement. It is important to acknowledge that using empirical force fields combined with enhanced sampling methods is quite effective in sampling the conformational space of molecules [32, 88, 136]; especially compared to the computational resources required to train large ML models. The field of machine learning for the physical sciences is still in its infancy, and there is much potential to improve in order to consistently outperform more traditional molecular dynamics simulations. Future research should focus on harnessing the full potential of transfer operator frameworks as well as learning representations of biomolecules that are consistent with stationary and dynamic experimental data. The incorporation of experimental constraints also has implications on the generalizability and transferability of the models. As we continue to expand and refine these methods, we get closer to a comprehensive understanding of biomolecular dynamics with wide-ranging implications in fields like drug design, molecular engineering, and beyond.