Thesis for The Degree of Doctor of Philosophy

Data-Driven Modeling of Biomolecular Dynamics under Physical and Experimental Constraints

Christopher Kara (né Kolloff)

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg
Gothenburg, Sweden, 2026

Data-Driven Modeling of Biomolecular Dynamics under Physical and Experimental Constraints

Christopher Kara

© Christopher Kara, 2026
except where otherwise stated.
All rights reserved.

ISBN 978-91-8103-438-7
Doktorsavhandlingar vid Chalmers tekniska högskola, Ny serie nr 0346-718X.
ISSN 0346-718X
DOI: https://doi.org/10.63959/chalmers.dt/5895

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

Printed by Chalmers Digitaltryck,
Gothenburg, Sweden 2026.

πάντα ῥεῖ. Everything flows.
– *after Heraclitus*

Data-Driven Modeling of Biomolecular Dynamics under Physical and Experimental Constraints

Christopher Kara (né Kolloff)

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

Abstract

Biophysical experiments and biomolecular simulations report on the same processes that govern biological function, yet each provides only a partial view. Experiments offer unbiased but sparse ensemble measurements, while simulations provide atomic detail but suffer from force-field biases and finite sampling. Integrating both into a single model is essential for a complete picture of biomolecular dynamics. This thesis develops the mathematical foundations for this integration through constrained statistical inference using the transfer operator framework and observables as the common language between experiments and simulations, comprising five papers in total. A survey chapter reviews machine learning approaches to molecular dynamics and identifies the absence of experimental consistency mechanisms as a common limitation. An empirical study then predicts dynamics-sensitive experiments from simulation data, revealing significant discrepancies that establish the need for integrative approaches. To address this, dynamic Augmented Markov Models (dynAMMo) are introduced, incorporating experimental measurements as constraints to recover slow exchange processes even from incomplete simulations. The Minimum-Excess-Work framework pushes the integration upstream, offering a method that incorporates stationary observables into probability flow generative models. Finally, applying dynAMMo to an engineered lipocalin demonstrates that these methods yield mechanistic insights that neither simulations nor experiments alone could provide. Together, these contributions progressively shift from comparing experimental and computational data, to correcting kinetic models post hoc, to guiding the generative process itself.

Keywords

biomolecular dynamics, transfer operators, Markov state models, constrained statistical inference, NMR relaxation dispersion, generative modeling, integrative structural biology, AI4Science

List of Publications

Appended publications

This thesis is based on the following publications:

[**Paper I**] **C. Kolloff** and S. Olsson, *Machine Learning in Molecular Dynamics Simulations of Biomolecular Systems*
Comprehensive Computational Chemistry, Elsevier (October 2023), pp. 475–492, ISBN: 978-0-12-823256-9..

[**Paper II**] **C. Kolloff**, A. Mazur, J. K. Marzinek, P. J. Bond, S. Olsson, and S. Hiller, *Motional clustering in supra-τ_c conformational exchange influences NOE cross-relaxation rate*
Journal of Magnetic Resonance, Vol. 338 (May 2022). `https://doi.org/10.1016/j.jmr.2022.107196`

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

[**Paper IV**] **C. Kolloff***, T. Höppe*, E. Angelis*, M. J. Schreiner, S. Bauer, A. Dittadi, and S. Olsson, *Minimum-Excess-Work Guidance: Score-Based Sampling with Experimental Data or Sparse Restraints*
Accepted for publication in Journal of Chemical Theory and Computation. `https://doi.org/10.48550/arXiv.2505.13375`

[**Paper V**] M. J. Bostock*, **C. Kolloff***, E. Jerschke*, S. Asami, A. Skerra, S. Olsson, and M. Sattler, *Conformational Quenching in an Engineered Lipocalin Protein Achieves High Affinity Binding to the Toxin Colchicine*
Angewandte Chemie International Edition, Vol. 64 (October 2025), e202515950. `https://doi.org/10.1002/anie.202515950`

Summary of Contributions

This thesis is based on five appended papers. My contributions to each paper are listed below using the CRediT (Contributor Roles Taxonomy) classification:

[**Paper I**] Investigation (literature review), Writing – original draft, Visualization, Writing – review & editing (reviewer responses).

[**Paper II**] Conceptualization, Methodology (theoretical framework), Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization, Writing – review & editing (reviewer responses).

[**Paper III**] Conceptualization, Methodology (theoretical framework and method design), Software, Investigation, Validation, Writing – original draft, Visualization, Writing – review & editing (reviewer responses).

[**Paper IV**] Conceptualization (Observable Guidance), Methodology (theoretical framework of the minimum-excess-work regularizer, including connections to optimal transport and excess work bounds), Software (Observable Guidance implementation), Investigation, Validation, Visualization, Writing – original draft, Writing – review & editing (reviewer responses). The Path Guidance component was developed by co-authors and is not part of this thesis.

[**Paper V**] Software (computational analysis code), Investigation (molecular dynamics simulations and dynAMMo analysis), Formal analysis (trajectory and dynAMMo results), Visualization (computational figures), Writing – original draft, Writing – review & editing (reviewer responses).

Acknowledgment

My scientific curiosity has always revolved around the same question: how do proteins manage to perform such a vast variety of functions? During my Bachelor's in Molecular Biology, I learned to place them in a cellular context. In my Master's in Biophysics, I was curious to understand what it means to characterize them—experimentally and computationally. For my PhD, I wanted to go further: I wanted to learn how to build mathematical models of their dynamics. I am beyond grateful to my supervisor, Simon, for taking a chance on someone without a classical computer science background, and for giving me the freedom to follow that curiosity into this field. It is surely an understatement that this required a lot of patience and understanding, and I am honored to have been chosen as one of his first PhD students. It is a cliché at this point, but I can honestly say that none of this would have been possible without you, Simon. So thank you for everything. Not for a single second have I regretted moving to Sweden to do my PhD with you.

The five years of this PhD were shaped by a supervision team I could not have wished to be better. Rocío and, before her, Alex served as co-supervisors, and Devdatt as examiner. Together with Simon, they gave me steady support and guidance throughout. I am also grateful to Matteo Degiacomi, who agreed to serve as opponent, and to Mikael Akke, Carme Rovira, Anders Krogh, and Fredrik Johansson for taking the time to read and evaluate this work as members of the grading committee. I look forward to the conversation on June 12. A formative chapter of this PhD was my research exchange at the Research laboratory of Electronics at MIT, made possible by a Fulbright scholarship and a WASP research stint. Working with Tess was phenomenal. Your curiosity, kindness, and scientific expertise were truly inspiring. Much of the work behind Paper IV was forged on the conference circuit. The MEW paper passed through NeurIPS, ICLR, and ICML before finding its final form in JCTC, and the persistence that required was shared above all with Tobias and Andrea. My collaboration with Mark and Michael on the lipocalin study (Paper V) holds a special distinction: it spans the entire arc of this PhD, from my first year to my last. What none of us could have predicted was that dynAMMo (Paper III), which was developed entirely independently, would turn out to be exactly the right tool for the job. The fact that these two projects converged so neatly remains one of the most satisfying surprises of my PhD.

My years at Chalmers, in particular DSAI, would not have been half as good without the colleagues with many of whom I also became friends along the way. I was especially lucky to share an office, organize a study trip to Japan, and form a lasting friendship with Juan. And with Lena, who joined me on that same Japan trip, and who by some wonderful coincidence ended up doing her research exchange at MIT at the same time as me. To the rest of Simon's and Rocío's lab as well as the division as a whole: thank you for making the day-to-day a pleasure. Whether in group meetings, over coffee, or traveling around Sweden together, serving on the PhD Student Council and taking/TA'ing courses with many of you has been one of the unexpected rewards of these five years. Likewise, to my colleagues and friends at MIT: how lucky was I to have spent a phenomenal eight months with you in Cambridge?

Finally – Jonatan, min bage: du har varit där *bokstavligen* från början. Du har följt mig till USA, och nu kommer vi få dela ett liv tillsammans i Schweiz. Jag kan inte vänta! Det är du, jag och tjejen (jö).

Mama, Papa, Nicolas und Lua: Auch ihr wart für mich von Anfang an auf meiner wissenschaftlichen Reise da. Ich freue mich so sehr darauf, nach fünf Jahren endlich wieder (geographisch) näher bei euch zu sein.

And to the rest of my friends: Thank. You.

This work was supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, and by the Fulbright Foreign Student Program, sponsored by the U.S. Department of State and the Swiss–American Fulbright Commission. The contents of this thesis are solely my responsibility and do not necessarily represent the official views of the Fulbright Program, the Government of the United States, or the Swiss–American Fulbright Commission. Computations and data handling were enabled by resources provided by Chalmers e-Commons at Chalmers, by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) on Alvis (projects NAISS 2023/22-1289 and NAISS 2025/22-608), and by the MIT Office of Research Computing and Data. During the preparation of this thesis, large language models were used as light editorial tools for proofreading, prose clarity, and notational consistency. All scientific content, including research questions, methods, analysis, results, and conclusions, is entirely my own.

Gothenburg, May 2026

Contents

Part I

Summary

Chapter 1

Introduction

We now have the resources and methods to predict the three-dimensional structure of virtually any protein given its amino acid sequence and evolutionary context [1]. Many molecular biology questions instead focus on how one specific protein or nucleic acid fits into the context of a cell. That protein might catalyze some reaction [2], signal another molecule to halt gene expression, or interact with other biomolecules [3]. A single structure does not tell us anything about how that protein fits into the bigger picture; how we can connect its structure to its function. Dynamics tells us how the protein changes its structure, how fast the transitions between different conformations occur, and whether these structural changes expose, for example, hidden pockets that are important for its function. In enzymes, for example, conformational exchange resolved by NMR spectroscopy can correlate directly with microscopic rates of substrate turnover [4]. Yet, even for a single well-studied protein system, a reliable quantitative picture continues to be notoriously difficult to obtain [5, 6].

This is not because we lack the tools. Molecular dynamics (MD) simulations track the movements of every single atom [7–10], Nuclear Magnetic Resonance (NMR) spectroscopy tells us how fast the exchange between states occurs [11–13], and Small-Angle X-ray Scattering informs us about the broad structural changes that occur in this exchange [14–16]. The reason that studying biomolecular systems is hard is because each tool captures a different, incomplete view of the same underlying processes [3]. Integrative approaches that combine MD simulations with orthogonal experimental techniques, such as ion-mobility mass spectrometry, demonstrate that joint modeling can recover mechanistic insight inaccessible to either method alone [17, 18]. If they all in some way describe the *same* process, then we should be able to model these different views jointly and obtain a more complete, mechanistic understanding of the true underlying process.

In recent years, advanced statistical models, such as machine learning (ML) techniques, have transformed our ability to make sense of experimental [19–21] and computational data [22–24]. That includes constructing kinetic models from MD trajectories [25, 26], learning slow collective variables [27, 28], and even generating molecular conformers [29–32]. By design, these models learn

from the data and hence also inherit their limitations and biases. Hence, a more complicated and expressive model of biased data is still a biased model. Even highly accurate ML structure predictors benefit from orthogonal experimental constraints when ligand binding or oligomerization are involved [33]. Because these methods probe the same process from different perspectives, we can compensate for their respective weaknesses. This requires a *principled* integration of different data sources to use as reference.

As no single data source is authoritative (simulations carry force-field biases, experiments report on projections of the full dynamics, and each has finite statistical expressiveness), combining them is therefore not simply a matter of fitting a model to more data. It raises the question what to trust, how much, and what to do when two sources disagree.

This thesis does not resolve these epistemological questions – though it would certainly also be interesting if it did. What it does, however, is to formalize the technical side of the problem. In this thesis, I describe how we can build, guide, correct, and gain insights from machine learning and statistical inference models of biomolecular dynamics using physical and experimental constraints.

The background of this thesis is structured in four parts. Chapter 2 poses the question what mathematical language is general enough to describe the dynamics of a biomolecular system regardless of whether we observe it through a simulation or an experiment. Chapter 3 asks how the abstract concepts described in Chapter 2 become measurable, thus allowing us to compare the different views. With the conceptual framework in place, Chapter 4 turns to the data sources themselves and examines what we can expect from simulations and experiments, and where each is reliable or limited. Chapter 5 then asks whether machine learning can overcome these limitations and finds that while ML models are more expressive and flexible, they cannot close the gap between simulations and experiments on their own. This motivates Chapter 6, which develops principled statistical methods for incorporating experimental data into computational models, thus laying the groundwork for the appended papers.

The thesis includes five appended papers. Paper I surveys the landscape of machine learning approaches in molecular dynamics applications and identifies the open problems that motivated the subsequent work. Paper II asks how accurately current simulations can reproduce a particularly demanding class of NMR observables and finds that the answer is sobering, even across state-of-the-art force fields. Paper III and Paper IV each propose different methods that allow for a systematic integration of experimental observables into kinetic (Paper III) and generative models (Paper IV). Paper V moves away from method development to validation of integrative modeling approaches and applies the method developed in Paper III to a real protein system. This validation shows that through the combination of NMR and computational data, we can elucidate a binding mechanism that neither experiments nor simulations alone would have been able to show on their own.

Chapter 2

Stochastic Dynamics and the Transfer Operator

Biomolecules, like all molecules, interact with their environment in a myriad of ways: they are surrounded by water molecules and solvents, bind to membranes [34], other proteins or nucleic acids [35–37], or get modified in various ways [38, 39]. And even though those interactions occur spontaneously and randomly, they create the predictable patterns that constitute the biochemical processes that researchers try to uncover. Most importantly for us, though, the interactions of biomolecules with their surroundings result in changes in their shape, or structure, that we call *conformations*, and the set of all conformations for a given biomolecule is called the *ensemble*.

In order to make sense of how random molecular interactions give rise to predictable dynamical behavior, we need a mathematical framework that captures the essential features of the system while stripping away details that are not relevant. This chapter develops such a framework around the *transfer operator* and its spectral decomposition, which organizes much of what follows in this thesis. Molecular dynamics simulations can be understood as an instantiation that generates samples whose dynamics are governed by the transfer operator; experimental observables probe specific aspects of its spectrum; machine learning methods for dynamics attempt to model it through approximations; and integrative modeling asks how experimental data can constrain our estimate of it. We introduce these ideas here in their general form before making the explicit connections to the methods introduced in this thesis.

2.1 Transfer Operator Framework

At any given time t, a single molecule occupies a configuration $\mathbf{x}_t$ in a high-dimensional configuration space, which in the classical approximation [2, 40, 41] can be represented as $\mathbf{x} \in \Omega \subseteq \mathbb{R}^{3N}$, where N is the number of atoms. Because of thermal noise and the random interactions, the evolution of $\mathbf{x}_t$ is stochastic

rather than deterministic. In the context of this thesis, however, we are not interested in the exact configurational changes of one specific particle but in the macroscopic behavior of the ensemble. The fundamental quantity of interest is therefore not the trajectory itself but the evolution of the probability density $\rho(\mathbf{x}, t)$ of observing configuration $\mathbf{x}_t$ at time t across conformational space Ω. From this viewpoint, we can explore how experiments and computational models of biomolecular dynamics are connected.

We assume the process is Markovian [22, 42, 43], i.e., the probability of a future configuration $\mathbf{x}_{t+\tau}$ depends only on the current configuration $\mathbf{x}_t$, not on any earlier history:

$$p(\mathbf{x}_{t+\tau} \,|\, \mathbf{x}_t, \mathbf{x}_{t-\tau}, ...) = p(\mathbf{x}_{t+\tau} \,|\, \mathbf{x}_t, \tau). \tag{2.1}$$

The lag time $\tau > 0$ denotes a finite time increment over which the stochastic evolution is observed. We further assume the process is time-homogeneous, so that the transition density depends only on τ and not on the absolute time t. Let $\rho(\mathbf{x}, t)$ denote the initial probability density at time t. By integrating over all possible starting configurations, the density $\rho_\tau(\mathbf{x}_{t+\tau}, t+\tau)$ at time $t+\tau$ is given by

$$\rho(\mathbf{x}, t+\tau) = \int_\Omega p(\mathbf{x}_{t+\tau} \,|\, \mathbf{x}_t, \tau)\rho(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}_t. \tag{2.2}$$

This integral defines a linear map on probability densities. We call it the *Perron–Frobenius transfer operator* $\mathcal{T}_\tau$ [43, 44]:

$$(\mathcal{T}_\tau \rho_t)(\mathbf{x}_{t+\tau}) := \int_\Omega p(\mathbf{x}_{t+\tau} \,|\, \mathbf{x}_t, \tau)\rho_t(\mathbf{x}_t, t)\, \mathrm{d}\mathbf{x}_t, \tag{2.3}$$

so that the density at the next time step is simply $\rho(\mathbf{x}, t+\tau) = \mathcal{T}_\tau \rho(\mathbf{x}, t)$.

Spectral Decomposition The transfer operator encodes all dynamical information about the system, but that information is not immediately accessible from Eq. 2.3. To extract it, we decompose $\mathcal{T}_\tau$ into its individual contributions, each corresponding to a single dynamical exchange with a corresponding amplitude and timescale. The result is the *spectral decomposition* of $\mathcal{T}_\tau$, which separates the kinetic and thermodynamic content of the transfer operator and makes both accessible from simulations and experiments. All results in this thesis are applied in a finite-dimensional or discrete setting and do not require the functional-analytic machinery that a fully rigorous treatment would demand.

Applied to an initial density $\rho(\mathbf{x}, t)$, the spectral decomposition of $\mathcal{T}_\tau$ decomposes the time evolution into a series of individual contributions:

$$\rho(\mathbf{x}, t+\tau) = \sum_{i=1}^{\infty} \lambda_i(\tau)\, \boldsymbol{\phi}_i(\mathbf{x}_{t+\tau}) \int_\Omega \boldsymbol{\psi}_i(\mathbf{x}_t)\, \rho(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}_t, \tag{2.4}$$

where each term represents a single dynamical process i, characterized by an eigenvalue $\lambda_i(\tau)$ and a pair of eigenfunctions: a left eigenfunction $\boldsymbol{\psi}_i$ and a

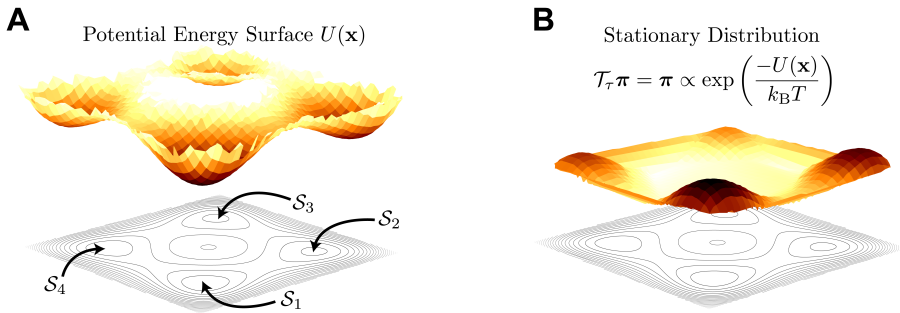

Figure 2.1: **Relationship between the potential energy surface and the equilibrium distribution for a two-dimensional asymmetric quadruple-well potential.** (A) The potential energy surface $U(\mathbf{x})$ (top) and its contour projection (bottom), with energy minima corresponding to four metastable states $\mathcal{S}_1$–$\mathcal{S}_4$. (B) The stationary distribution π obtained via the Boltzmann relation (Eq. 2.6). Regions of low potential energy in (A) correspond to regions of high probability density in (B), and the energy barriers separating the minima become the low-probability regions that govern the rate of conformational exchange between states.

right eigenfunction $\boldsymbol{\phi}_i$. Together, these three components encode the timescale, spatial structure, and amplitude of each process, but their physical meaning becomes clear only once we establish the thermodynamic and kinetic structure of the system. Each eigenvalue $\lambda_i(\tau)$ has an associated exchange timescale t_i^{ex} via:

$$\lambda_i(\tau) = \exp\left(-\frac{\tau}{t_i^{\mathrm{ex}}}\right) \tag{2.5}$$

that characterizes the decay of the process. This relationship separates the thermodynamic and kinetic content of the transfer operator.

2.1.1 Thermodynamics and Kinetics

Equilibrium distribution. For $i = 1$, $\lambda_1 = 1$, which corresponds to $t_1^{\mathrm{ex}} = \infty$ (Eq. 2.5) and designates a process that never decays. The associated eigenfunction is a density that $\mathcal{T}_\tau$ leaves unchanged. We call it the *equilibrium (or stationary) distribution*: $\mathcal{T}_\tau \pi = \pi$. This distribution describes the probability of finding the system in configuration $\mathbf{x}$ after all exchange processes have decayed. In terms of the spectral decomposition (Eq. 2.4), π is the right eigenfunction $\boldsymbol{\phi}_1$ of the first process: the component that persists indefinitely while all transient contributions decay. Thus, all time-independent, i.e., thermodynamic, properties of the system (e.g., free energies or equilibrium populations) are expectations under π, and in the standard equilibrium setting of most molecular applications, it corresponds to the Boltzmann distribution:

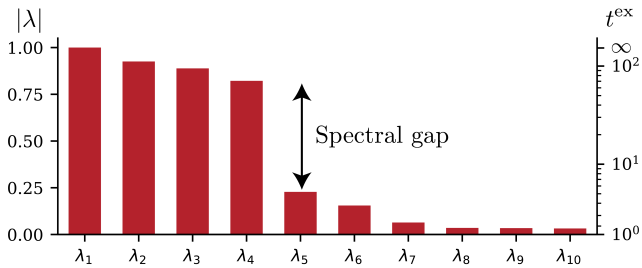

Figure 2.2: **Eigenvalue spectrum of asymmetric quadruple-well potential.** Eigenvalue magnitudes $|\lambda_i|$ (left axis) and corresponding exchange timescales t_i^{ex} (right axis, log scale; Eq. 2.5). The spectral gap between λ_4 and λ_5 separates the slow conformational exchanges from fast local fluctuations.

$$\pi \propto \exp\left(\frac{-U(\mathbf{x})}{k_{\mathrm{B}}T}\right), \tag{2.6}$$

where $U(\mathbf{x})$ corresponds to the energy of the configuration $\mathbf{x}$ and $(k_{\mathrm{B}}T)^{-1} = \beta$ is the inverse temperature that acts as a measure of how the probability of occupying a state changes with temperature.

Fig. 2.1 illustrates this relationship for a two-dimensional asymmetric quadruple-well potential[1]. The potential energy surface (panel A) defines four distinct minima corresponding to metastable states $\mathcal{S}_1$–$\mathcal{S}_4$, separated by energy barriers of varying height. The Boltzmann relation (Eq. 2.6) inverts this landscape so that the low-energy basins become peaks in the stationary distribution (panel B), with deeper wells accumulating more weight in the equilibrium distribution. The barriers, in turn, become the low-probability regions that limit the rate of conformational exchange between states.

Kinetic Processes. The remaining processes $i > 1$, on the other hand, have $0 < |\lambda_i| < 1$ and thus a finite relaxation timescale. They encode the system's kinetics. Notably, the importance of the exchange processes can be weighted by the magnitude of λ_i. Eigenvalues close to 1 indicate *slow processes* whose contributions are relevant for long timescales (see Eq. 2.5). If a process has an eigenvalue close to 0, it is considered fast and thus does not contribute to the long-timescale behavior of the system. Fig. 2.2 shows the eigenvalue spectrum of the quadruple-well potential from Fig. 2.1. The first four eigenvalues have magnitudes close to 1 and correspond to exchange timescales that are orders of magnitude longer than those of the remaining processes. The sharp drop between λ_4 and λ_5 is the spectral gap that separates these slow conformational exchanges from fast intra-well fluctuations. For each of these slow processes, the corresponding eigenfunctions identify where and how probability density is redistributed (Fig. 2.3). The right eigenfunction $\boldsymbol{\phi}_i$ partitions conformational

[1]The asymmetric quadruple-well potential used in this work is given by $V(x_1, x_2) = x_1^4 - \frac{1}{16}x_1^3 - 2x_1^2 + \frac{3}{16}x_1 + x_2^4 - \frac{1}{8}x_2^3 - 2x_2^2 + \frac{3}{8}x_2$, with corresponding overdamped Langevin dynamics defined via the drift $-\nabla V(x_1, x_2)$ and isotropic diffusion coefficient $\sigma = 0.6$.[45]

space into regions that gain and lose density during the exchange, where the positive/negative sign marks the source and sink of probability flow. The left eigenfunction $\boldsymbol{\psi}_i = \boldsymbol{\phi}_i \cdot \pi$ carries the same sign structure but reweights each region by its equilibrium population. This means that the left eigenfunction encodes how much statistical weight each side of the partition carries. A well that participates in the exchange but is sparsely populated contributes less than a different state with a high thermodynamic weight. Thus, the projection $\int_\Omega \boldsymbol{\psi}_i(\mathbf{x}_t)\, \rho(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}_t$ in 3.5 determines how strongly a given initial density $\rho(\mathbf{x}, t)$ excites the process.

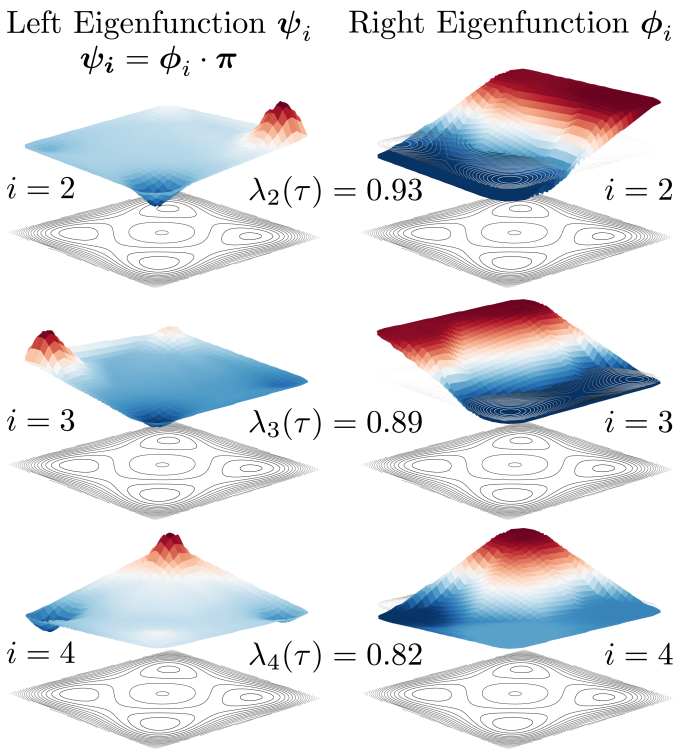

Figure 2.3: **Left (ψ_i) and right (ϕ_i) eigenfunctions of $\mathcal{T}_\tau$ for the three slowest non-trivial processes ($i = 2, 3, 4$) of the asymmetric quadruple-well potential** (contour lines). The sign structure of each right eigenfunction partitions configuration space into source and sink regions of the exchange. The left eigenfunctions $\boldsymbol{\psi}_i = \boldsymbol{\phi}_i \cdot \pi$ carry the same partition, reweighted by equilibrium population. Eigenvalues λ_i carry timescale information (2.5)

To summarize, each spectral component in Eq. 2.4 contributes three pieces of information about a dynamical process:

- **Eigenvalues** $\lambda_i(\tau) = \exp(-\tau/t_i^{\mathrm{ex}})$ set the timescale of the process. λ_i close to 1 correspond to slow conformational exchanges, while λ_i close to 0 indicate fast fluctuations that decay within the lag time (Fig. 2.2).

- **Right eigenfunctions** $\boldsymbol{\phi}_i$ partition conformational space into the regions that exchange probability density during the process. Their sign structure identifies source and sink (positive and negative regions gain and lose density, respectively; Fig. 2.3).

- **Left eigenfunctions** $\boldsymbol{\psi}_i = \boldsymbol{\phi}_i \cdot \pi$ carry the same partition but weighted by equilibrium population. They determine how strongly a given initial density excites the process via the projection $\int_\Omega \boldsymbol{\psi}_i(\mathbf{x}_t)\, \rho(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}_t$ in Eq. 2.4 (Fig. 2.3).

2.1.2 Assumptions for Systems in Thermodynamic Equilibrium

The spectral decomposition introduced in Sec. 2.1 relies on a set of assumptions about dynamical systems in equilibrium [22]:

1. **Conservation of probability.** $\mathcal{T}_\tau$ maps valid probability densities to valid probability densities, i.e. $\int_\Omega \rho(\mathbf{x})\, \mathrm{d}\mathbf{x} = 1$. This bounds all eigenvalues to $|\lambda_i| \leq 1$.[2]

2. **Ergodicity.** Every configuration is reachable from every other in finite time, so no region of Ω is permanently disconnected. This guarantees that the stationary distribution π is unique ($\lambda_1 = 1$ is non-degenerate) and that all other processes eventually decay ($|\lambda_{i>1}| < 1$).

3. **Detailed balance.** At equilibrium the process is microscopically reversible, i.e., the probability of observing a transition from $\mathbf{x}$ to $\mathbf{x}'$ equals that of the reverse: $\pi(\mathbf{x})\, p(\mathbf{x}' \,|\, \mathbf{x}) = \pi(\mathbf{x}')\, p(\mathbf{x} \,|\, \mathbf{x}')$. As a consequence, all eigenvalues are real and non-negative, and the left and right eigenfunctions are related by $\boldsymbol{\psi}_i = \boldsymbol{\phi}_i \cdot \pi$.[3]

Biomolecular systems in equilibrium satisfy these assumptions naturally [22, 43]. Thus, under these assumptions, the eigenspectrum is real, ordered and bounded (Fig. 2.2), where each eigenvalue–eigenfunction pair corresponds to a single interpretable process with a well-defined characteristic timescale and exchange profile.

[2] Repeated application $\mathcal{T}_\tau^n$ propagates each spectral component as λ_i^n, which remains bounded only if $|\lambda_i| \leq 1$

[3] Detailed balance makes $\mathcal{T}_\tau$ self-adjoint with respect to the π-weighted inner product, which additionally guarantees that the eigenfunctions form an orthonormal basis under this inner product.

2.1.3 Low-Rank Approximation of the Transfer Operator

Characterizing biomolecular systems, we are usually interested in their long-timescale behavior since it is the part that is most relevant for biological functions, such as ligand binding, chemical reactions, or signaling [46–48]. From Eq. 2.4, we see that the spectral decomposition expresses the density propagation under $\mathcal{T}_\tau$ as a sum over all dynamical processes. However, in Sec. 2.1.1, we established that the slow processes are the ones whose eigenvalue λ is close to 1 and that the faster contributions ($|\lambda_i| \ll 1$) decay quickly and become negligible at long timescales:

$$\rho(\mathbf{x}, t+\tau) \approx \sum_{i=1}^{k} \lambda_i(\tau)\, \boldsymbol{\phi}_i(\mathbf{x}_{t+\tau}) \int_\Omega \boldsymbol{\psi}_i(\mathbf{x}_t)\, \rho(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}_t \tag{2.7}$$

To characterize the dynamics of biomolecular systems it therefore suffices to capture the k dominant spectral components, since only the few slow processes govern conformational exchange on biologically relevant timescales. The quality of this truncation depends on the *spectral gap* between the retained and discarded eigenvalues, i.e., the separation $\lambda_k - \lambda_{k+1}$ (see Fig. 2.2). A large spectral gap indicates a clear timescale separation between the slow conformational exchange we wish to resolve and the fast local fluctuations we can neglect since they don't contribute to the long-timescale behavior of the system [22].

Regions of conformational space that are separated by a slow process, i.e., connected by an eigenfunction with $|\lambda_i| \approx 1$, are called *metastable states* [49, 50]. These are regions in Ω where the system dwells for long times relative to τ before transitioning to a different state (e.g., the four wells $\mathcal{S}_1$–$\mathcal{S}_4$ in Fig. 2.1, whose pairwise exchanges are resolved by the sign structure of the eigenfunctions in Fig. 2.3).

This finite-dimensional approximation is what lets us estimate, compare, and correct dynamical models using both computational and experimental data. Fig. 2.3 shows these three components for the three slowest non-trivial processes of the quadruple-well potential, illustrating how the eigenfunctions resolve the pairwise exchanges between the four metastable states in Fig. 2.1.

2.2 Molecular Dynamics Simulations

The transfer operator in Sec. 2.1 provides a mathematical framework for analyzing biomolecular dynamics, and molecular dynamics (MD) simulations [40, 51] are the principal computational method for generating them. They produce trajectories at atomic resolution by integrating a stochastic differential equation parameterized by a potential energy function $U(\mathbf{x})$ (see Fig. 2.1). Because $U(\mathbf{x})$ fully determines the transfer operator and its spectrum, any inaccuracies in modeling $U(\mathbf{x})$ lead to distortions of the spectral components of $\mathcal{T}_\tau$.

2.2.1 Langevin Dynamics

In the classical approximation, the motion of a molecule is governed by the Langevin equation [52], which balances deterministic forces, frictional drag, and thermal noise[4]. These three contributions (deterministic forces from the potential energy, frictional drag, and thermal noise) play distinct roles: the first drives the system toward low-energy configurations; the second dissipates kinetic energy through coupling to the solvent; and the third injects thermal energy to maintain the system at temperature T. By the fluctuation–dissipation theorem [53], this balance between dissipation and thermal driving ensures that trajectories asymptotically sample from the Boltzmann distribution (Fig. 2.1B):

$$\lim_{t\to\infty} p(\mathbf{x}_{t+\tau} \,|\, \mathbf{x}_t, \tau) = \pi(\mathbf{x}) \propto \exp(-U(\mathbf{x})/k_{\mathrm{B}}T). \tag{2.8}$$

In many biomolecular applications, inertial effects decay on ps timescales and are much faster than the conformational dynamics of interest. In this regime, the momentum term in the full Langevin equation can be neglected, yielding the overdamped Langevin equation [52]:

$$\gamma \,\mathrm{d}\mathbf{x} = -\nabla_{\mathbf{x}} U(\mathbf{x})\,\mathrm{d}t + \sqrt{2\gamma k_{\mathrm{B}}T}\;\mathrm{d}\mathbf{W}, \tag{2.9}$$

Here, γ is the friction coefficient, $-\nabla U(\mathbf{x})$ is the deterministic force, and $\mathrm{d}\mathbf{W}$ denotes increments of a Wiener process. Rather than tracking individual MD trajectories that this SDE generates, we can equivalently describe how the probability density $\rho(\mathbf{x}, t)$ evolves via the Fokker–Planck equation [54, 55]:

$$\partial_t \rho(\mathbf{x}, t) = \nabla \cdot \left[\frac{1}{\gamma} \rho(\mathbf{x}, t)\, \nabla_{\mathbf{x}} U(\mathbf{x}) + \frac{k_{\mathrm{B}}T}{\gamma} \nabla \rho(\mathbf{x}, t) \right]. \tag{2.10}$$

The transfer operator $\mathcal{T}_\tau$ is its finite-time propagator that maps the density at time t to the density at time $t+\tau$, so that $\rho(\mathbf{x}, t+\tau) = \mathcal{T}_\tau \rho(\mathbf{x}, t)$ as before [43, 44]. The chain from $U(\mathbf{x})$ to the Langevin equation to the Fokker–Planck equation to $\mathcal{T}_\tau$ thus links force-field parameterization to the spectral properties of the transfer operator.

2.2.2 Molecular Force Fields

In the context of this thesis, $U(\mathbf{x})$ is a classical, empirical potential energy function, a *force field* [56–58], whose functional form is fixed. It is parameterized either through fitting to quantum-mechanical calculations [2, 59], experimental thermodynamic data [2, 9], or both [56–58]. In standard biomolecular force fields [60, 61], the force field consists of a sum of terms that model specific instances of interatomic interactions: harmonic potentials for covalent bond lengths and angles, periodic potentials for torsional rotations, pairwise Lennard-Jones terms as well as Coulomb terms for electrostatic interactions. This makes force fields an interpretable but approximate representation of the true

[4] The full Langevin equation is given by $m\frac{\mathrm{d}^2\mathbf{x}}{\mathrm{d}t^2} = -\nabla U(\mathbf{x}) - \gamma\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} + \sqrt{2\gamma k_B T}\,\boldsymbol{\xi}(t)$, where m is the atomic mass, γ the friction coefficient, $U(\mathbf{x})$ a parameterized potential energy function known as the *force field*, and $\boldsymbol{\xi}(t)$ is Gaussian white noise.

$U(\mathbf{x})$. Because these equations of motion must be integrated numerically at femtosecond timesteps, while the conformational dynamics of interest occur on μs–ms timescales, a single simulation spans nine to twelve orders of magnitude in time [62]. This is a fundamental computational bottleneck of MD-based modeling. Due to the aforementioned modeling choices made in approximating $U(\mathbf{x})$ (e.g., parameterization through fitting to experimental data), force fields are systematically inaccurate. The consequences for the statistical limitations and systematic biases are the subject of Chap. 4.

The transfer operator provides the mathematical language for describing biomolecular dynamics, and MD simulations yield realizations of those dynamics through trajectories. But in order to assess whether a simulation faithfully represents the true dynamics, we need quantities that can both be predicted from a model and measured experimentally.

Chapter 3

Observables

Molecular configurations live in a high-dimensional space that cannot be probed directly. A molecule with N atoms occupies $\mathbb{R}^{3N}$, which neither experiments nor intuition can interpret without projection. Chap. 2 introduced the transfer operator as an object that encodes both the thermodynamics and kinetics of molecular systems. $\mathcal{T}_\tau$ acts on probability densities over that same high-dimensional space, but on its own, we also cannot access it directly. What we do have access to, however, are *observables*, i.e., scalar[1] quantities $o(\mathbf{x}) : \Omega \to \mathbb{R}$ that probe the spectral components of $\mathcal{T}_\tau$ by asking how the expected value of a measurable function $o(\cdot)$ changes over time. The chapter develops the mathematical relationship between observables and the spectral structure of $\mathcal{T}_\tau$, with o as the *lingua franca* of this thesis.

3.1 Equilibrium observables

The stationary distribution π (Sec. 2.1.1) encodes which regions of Ω are populated and with what relative (thermodynamic) weight. The equilibrium observable o^{stat} is the expectation of $o(\mathbf{x})$ under π:

$$o^{\text{stat}} = \mathbb{E}_{\mathbf{x}\sim\pi}[o(\mathbf{x})] = \int_\Omega o(\mathbf{x})\pi(\mathbf{x})\,\mathrm{d}\mathbf{x} \tag{3.1}$$

Examples. Three classes of equilibrium observable appear throughout this thesis:

1. **Interatomic distances and NOE-derived averages.** For atoms i and j, the interatomic distance $o(\mathbf{x}) = \|\mathbf{x}_i - \mathbf{x}_j\|$ and their $\langle r^{-6}\rangle$ ensemble average; important for the Nuclear Overhauser Effect (NOE) [11, 63, 64]. (Paper II, [65])

[1] All observables relevant in this thesis (e.g., folding free energy, scalar couplings, or melting temperature) are scalar valued, but the formalism extends straightforwardly to vector-valued observables.

2. **Scalar coupling constants.** ^3J-couplings report on backbone dihedral angles through the Karplus equation [66]:

$$^3J(\phi) = A\cos^2\phi + B\cos\phi + C, \tag{3.2}$$

where ϕ is the relevant dihedral angle, and A, B, and C are empirical Karplus parameters. (Paper IV, [67])

3. **Folding free energy.** $\Delta G = -k_{\mathrm{B}}T\ln(p_{\mathrm{folded}}/p_{\mathrm{unfolded}})$ captures the relative thermodynamic stability of macrostates [5, 68]. (Paper IV, [67])

Different observable functions have different sensitivity to the equilibrium distribution [16]. Global observables like the radius of gyration R_g are dominated by the most populated basins, whereas local structural observables (NOEs, scalar couplings [11, 66]) resolve residue-level populations. The folding free energy ΔG is logarithmically sensitive to the population balance between macrostates and can detect shifts in rare states that other observables miss. The choice of observable therefore determines which aspects of π are probed [5]. For example, NMR order parameters serve as probes of local fluctuations and as bridges between MD-derived dynamics and conformational entropy contributions to binding [69].

3.2 Dynamic Observables

Stationary observables are thermodynamic measures that describe the different populations in the system but carry no information about its kinetics. Two systems with identical stationary distributions (and hence identical stationary observables) can exhibit vastly different dynamics if the energy barriers between different states differ. To probe these barriers, we need observables that capture how a quantity measured at time t relates to the same or a different quantity measured at $t+\tau$. We formalize this with the time-correlation function, which captures two expectations: one over initial configurations from π and one over stochastic evolution to $t+\tau$:

$$\begin{aligned} o^{\mathrm{dyn}}_{o(\mathbf{x}_t),o'(\mathbf{x}_{t+\tau})} &= \mathbb{E}_{\mathbf{x}_t\sim\pi}[\mathbb{E}_{\mathbf{x}_{t+\tau}\sim p(\mathbf{x}_{t+\tau}\,|\,\mathbf{x}_t,\tau)}][o(\mathbf{x}_t)o'(\mathbf{x}_{t+\tau})] && (3.3)\\ &= \int_\Omega\int_\Omega o(\mathbf{x}_t)o'(\mathbf{x}_{t+\tau})p(\mathbf{x}_{t+\tau}\,|\,\mathbf{x}_t)\pi(\mathbf{x})\,\mathrm{d}\mathbf{x}_t\,\mathrm{d}\mathbf{x}_{t+\tau}. && (3.4)\end{aligned}$$

When $o = o'$, we call the quantity $o^{\mathrm{dyn}}_{o,o'}$ an *autocorrelation function*. Otherwise, it is considered a *cross-correlation function*. Dynamic observables of this kind characterize, for example, relaxation phenomena and capture the slow kinetics of biomolecules [12, 70].

Examples. The dynamic observables central to this thesis are:

1. **NOE cross-relaxation rates.** The NOE measures the dipolar cross-relaxation between proximal protons[2]. The underlying correlation function involves the internuclear distance as r_{ij}^{-6}, making it highly sensitive to local geometry. (Paper II, [65])

2. **CPMG and $R_{1\rho}$ relaxation dispersion.** These experiments probe conformational exchange on the µs–ms timescale through the autocorrelation of the chemical shift [72–76]. Like NOEs, they are residue-specific, but they report on slower exchange processes. (Paper III, [77], Paper V, [78])

3.2.1 Spectral decomposition of time-correlation functions

The correlation function in Eq. 3.4 involves an integral over the full transition density $p(\mathbf{x}_{t+\tau} \,|\, \mathbf{x}_t)$, but spectral decomposition of $\mathcal{T}_\tau$ resolves which dynamical processes contribute. Under the equilibrium assumptions of Sec. 2.1.2, each eigenprocess contributes independently to the correlation function [22, 43]:

$$o^{\mathrm{dyn}}_{o,o'}(\tau) = \sum_{i=1}^{\infty} c_i \, \lambda_i(\tau) = \sum_{i=1}^{\infty} c_i \, \exp\!\left(-\frac{\tau}{t_i^{\mathrm{ex}}}\right). \tag{3.5}$$

where

$$c_i = \int_\Omega o(\mathbf{x}) \, \boldsymbol{\psi}_i(\mathbf{x}) \, \mathrm{d}\mathbf{x} \int_\Omega o'(\mathbf{x}) \, \boldsymbol{\psi}_i(\mathbf{x}) \, \mathrm{d}\mathbf{x} \tag{3.6}$$

are the amplitudes determined by projecting the observable functions onto the eigenfunctions of $\mathcal{T}_\tau$.

Eq. 3.5 contains the central insight of this chapter. The exchange processes, encoded by the eigenvalues and eigenfunctions of $\mathcal{T}_\tau$, are properties of the system itself. They are independent from any measurement, though residue-resolved NMR can reveal that apparent local fluctuations are in fact signatures of broader, correlated structural exchange [79]. The property that allows *different* observables to measure *different* aspects of the dynamics is how sensitive they are to each process by modulating its amplitude c_i. Fig. 3.1 illustrates this for two observables on the quadruple-well potential (Fig. 2.1). Because they overlap differently with the eigenfunctions (Fig. 2.3), they weight the same dynamical processes with different amplitudes. No single observable gives access to the full eigenspectrum of $\mathcal{T}_\tau$; an NOE cross-relaxation rate and a $R_{1\rho}$ relaxation dispersion profile, for instance, each provide a differently filtered view of the same underlying dynamics. Central to this thesis, observables are quantities that can be both predicted from a computational model and measured experimentally. Observables therefore provide a direct basis for comparing simulations with experiments, and the next chapter examines the reasons and consequences of discrepancies between them.

[2] Relaxation occurs between all NMR-active nuclei, but proton–proton relaxation dominates due to the high gyromagnetic ratio; we consider only this case. [71]

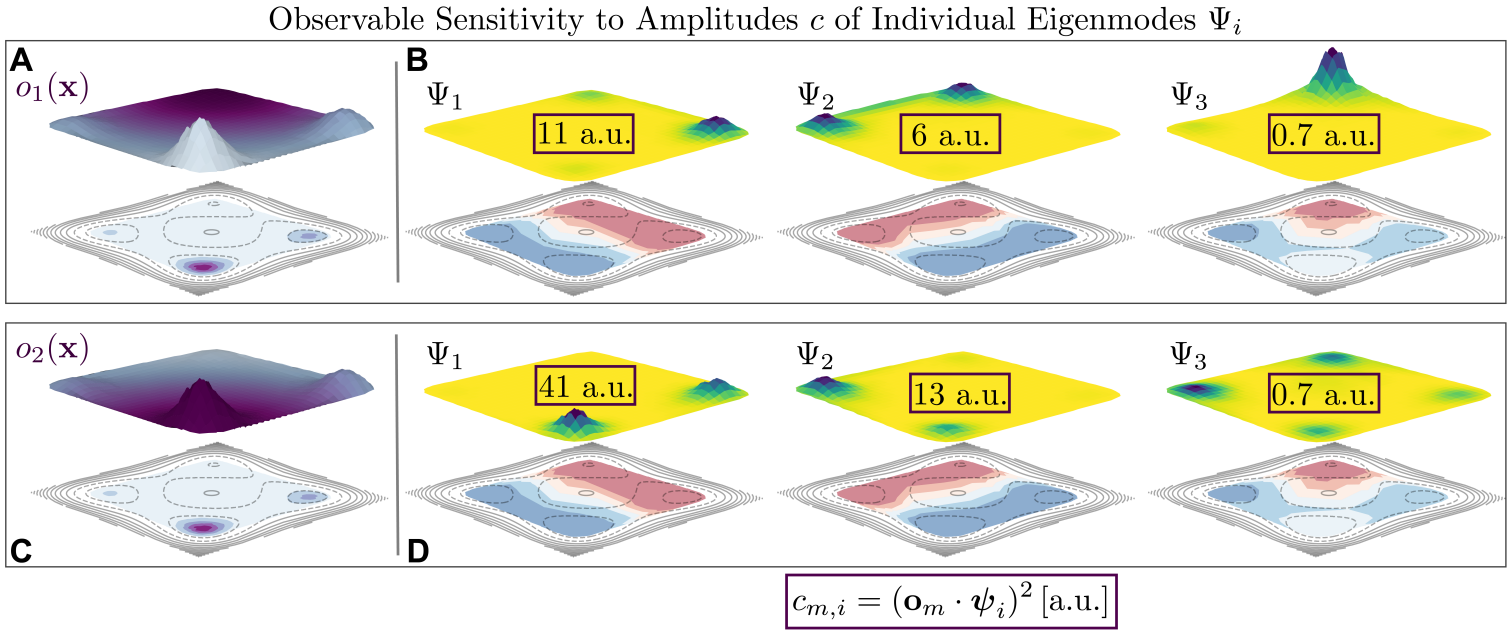

Figure 3.1: **Observable choice determines which dynamical processes are visible.** (A, **C**) Two observable functions $o_1(\mathbf{x})$ and $o_2(\mathbf{x})$ defined on the asymmetric quadruple-well potential (Fig. 2.1). o_1 concentrates its amplitude on the least populated basin, while o_2 concentrates on the most populated one. (B, D) The squared projection coefficients $c_{m,i} = (\mathbf{o}_m \cdot \boldsymbol{\psi}_i)^2$ quantify how strongly each observable couples to each eigenmode Ψ_i. Because o_1 and o_2 overlap differently with the eigenfunctions, they weight the same set of dynamical processes differently. Here, o_2 is roughly four times more sensitive to Ψ_1 and twice as sensitive to Ψ_2 than o_1, while both are equally insensitive to Ψ_3. The same underlying system thus produces distinct correlation functions depending on which observable is measured, and no single observable gives access to the full eigenspectrum.

Chapter 4

The Gap Between Molecular Simulations and Biophysical Experiments

Molecular dynamics can, in principle, recover the complete eigenspectrum of $\mathcal{T}_\tau$, but only from an approximate potential and with finite sampling. This chapter argues that the gap between what simulations predict and what experiments measure is a diagnostic that can guide the integration of both sources of information [6].

4.1 Limitations of Experiments and Simulations

The projection of $\mathcal{T}_\tau$ down to a scalar value is a severely underdetermined inverse problem that does not allow us to reconstruct $\mathcal{T}_\tau$ from experiments alone. In addition, ensemble averaging can mask lowly populated states and dynamic experiments are sensitive only to a narrow timescale range. These constraints motivate the question of what simulations can contribute that experiments alone cannot.

MD trajectories are generated from the force-field-parameterized Langevin dynamics introduced in Sec. 2.2, which in principle allows the full spectral structure of $\mathcal{T}_\tau$ to be recovered. In practice, this recovery is limited by two distinct sources of error.

The first is systematic. Any bias in $U(\mathbf{x})$ propagates into the eigenvalues, eigenfunctions, and stationary distribution of the estimated operator $\hat{\mathcal{T}}_\tau$ (Sec. 2.2.2), and no amount of additional sampling can compensate for this [24, 80, 81].

The second is statistical. The nine-to-twelve-order-of-magnitude gap between integration timestep and biological timescale outlined in Sec. 2.2.2 makes exhaustive sampling computationally infeasible [68] even with special-purpose hardware [82] or distributed platforms [83]. In order to reliably estimate the

large eigenvalues $0 \ll |\lambda_i| < 1$ of $\mathcal{T}_\tau$, the simulation needs to reversibly sample the rare transition events their eigenfunctions encode (Fig. 2.3). The smaller eigenvalues $|\lambda_i| \ll 1$ converge much more quickly, but the fast processes they describe do not necessarily capture the conformational changes we are interested in [22, 25].

4.2 Diagnosing Model Deficiencies Through Observable Discrepancies

The observable functions introduced in Chap. 3 provide a shared quantitative language between simulations and experiments: both make predictions about the same spectral quantities. Equilibrium expectations o^{stat} as well as correlation function amplitudes c_i and timescales t_i^{ex} serve as a direct diagnostic of model quality when comparing mismatches between predicted and measured values. And while they should agree in theory, in practice, the limitations of each approach (Sec. 4.1) prevent them from doing so. The central argument of this section is that such discrepancies are not merely binary indicators *that* something is wrong. They also reveal *what* is wrong and *where* the model fails (though not necessarily *why*), as illustrated schematically in Fig. 4.1.

Interpretability of discrepancies. Using the spectral decomposition of the Koopman operator, we can view the time correlation function as the sum of all processes, where each process has an associated amplitude and timescale (Sec. 3.2.1). Spectral decomposition localizes the mismatch source. If amplitudes are wrong but timescales are correct, the observable projection onto the eigenfunction is incorrect (Fig. 3.1); this can stem from incorrect metastable-state populations (Fig. 2.1) or from missing relevant states. Thermodynamic force-field bias and finite sampling are the main contributors in this case. If the timescales are wrong and the amplitudes correct, this points to a bias in the spectral gap (Fig. 2.2), so timescales are over- or underestimated. This can either be caused by kinetic bias in force-field parameterizations, i.e., by incorrectly modeled energy barriers between metastable states, or, again, by insufficient sampling. Neither diagnostic unambiguously identifies the *cause* of the mismatch; it only indicates whether the discrepancy is thermodynamic or kinetic.

Sensitivity analysis across force fields. Running simulations with different force fields provides a practical, albeit imperfect, way to assess where our predictions are reliable and where experimental integration is needed [84, 85]. If a single simulation disagrees with experiment, the source of error is ambiguous: it could reflect force-field bias, insufficient sampling, or both. Comparing predictions across force fields helps us to disentangle this. If the predicted kinetic parameters for a given region of the protein vary wildly between force fields, the prediction is clearly sensitive to modeling choices, and no individual simulation can be trusted there. This is where integrating experimental data into the model becomes most relevant. Conversely, if all

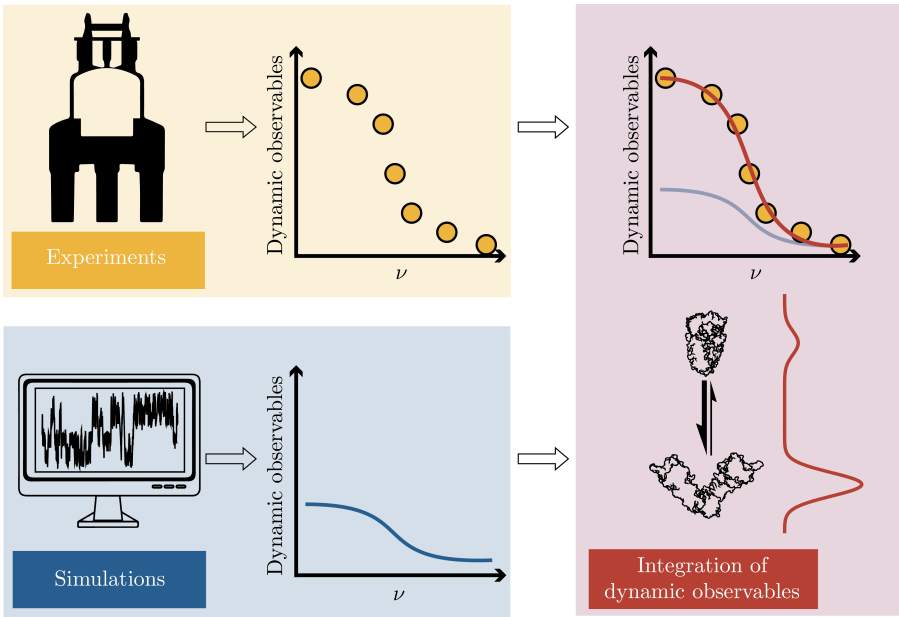

Figure 4.1: **Schematic of integrating dynamic experimental observables with molecular simulations.** Experiments (top left) yield discrete measurements of dynamic observables, while simulations (bottom left) produce continuous predictions from the estimated transfer operator. The amplitude discrepancy between the simulation prediction (blue) and the experimental data (right) indicates that the observable's projection onto the eigenfunctions is incorrect, due to wrong state populations or missing metastable states. The corrected model (red) recovers the experimental amplitudes by constraining the kinetic model to satisfy the measured observables. Adapted from [77].

force fields agree, the prediction is at least robust to parameterization (though not necessarily correct, since force fields share functional forms, training data, and core approximations that could produce shared errors [86, 87]). This case still warrants comparison with their experimental counterparts. Observables like NOE cross-relaxation rates and relaxation dispersion are particularly useful for this kind of analysis because they provide residue-level resolution, which allows us to spatially localize the regions where models diverge and where using experiments in our model as additional information would be most informative [65]. The same principle extends to other structural observables, e.g., chemical crosslinks, which account for conformational heterogeneity rather than relying on static structures, significantly improves the interpretation of experimental data [88].

The diagnostic framework above rests on the premise that principled integration of simulations and experiments is necessary to obtain reliable models of biomolecular dynamics. A practical integration framework must therefore

estimate finite-dimensional approximations of $\mathcal{T}_\tau$ from trajectory data and connect them to experimentally measurable quantities, and machine learning approaches have advanced this substantially in recent years [25, 26, 89–92].

Chapter 5

Machine Learning for Biomolecular Dynamics

5.1 Markov State Models as Finite-Dimensional Approximations of Transfer Operator

The transfer operator introduced in Chap. 2 is an infinitely dimensional object that needs to be approximated, both in space and time, in order for it to be useful in a computational modeling context. Our goal is thus to find a finite-dimensional surrogate estimated from trajectory data; and the Markov State Model (MSM) is the simplest finite-dimensional approximation of the transfer operator [22, 25, 42].

5.1.1 MSM Estimation

MSMs have shown to be simple yet informative models of biomolecular dynamics [46–48, 93–104]. MSMs can be estimated by first featurizing the MD trajectory [25, 105] (e.g., interatomic distances or dihedral angles), projecting these features onto a reduced set of coordinates (e.g., tICA [27, 106] or PCA [107]), and finally discretizing the projected features into n states. The discrete trajectory can then be used to estimate the transition matrix $\mathbf{T}(\tau) \in \mathbb{R}^{n\times n}$ [22], where entry $\mathbf{T}_{ij}(\tau)$ gives the probability of transitioning from state i to j in time τ. This estimation happens via the count matrix $\mathbf{C}(\tau) \in \mathbb{R}^{n\times n}$ that records transitions between state pairs. The simplest maximum-likelihood estimate of $\mathbf{T}(\tau)$ from $\mathbf{C}(\tau)$ is by row-normalizing $\hat{\mathbf{T}}_{ij}(\tau) = \mathbf{C}_{ij}(\tau)/[\sum_j \mathbf{C}_{ij}(\tau)]$, though more statistically robust methods exist [108]. The equilibrium assumptions from Sec. 2.1.2 (stochasticity, ergodicity, and detailed balance) can be imposed as constraints on $\mathbf{T}$, thereby reducing the space of admissible solutions and stabilizing estimation [22, 108].

5.1.2 Observable Prediction from Markov State Models

Diagonalizing $\mathbf{T}$ yields the eigenvalues $\{\lambda_i\}_{i=1}^n$ and the left and right eigenvectors $\{(\mathbf{L}_i, \mathbf{R}_i)\}_{i=1}^n$. They are the discrete-space analogues of the eigenvalues and eigenfunctions introduced in Sec. 2.1 (cf. Fig. 2.2 and Fig. 2.3) and allow us to predict biophysical observables directly from the MSM to their measured counterparts [92, 109].

Given an observable function $o(\mathbf{x})$, the observable vector $\mathbf{o} \in \mathbb{R}^n$ contains the mean value of o within each state. Analogous to Eq. 3.1, the stationary observable can be calculated as the inner product between the observable vector and the discrete stationary distribution: $o^{\text{stat,MSM}} = \mathbf{o} \cdot \boldsymbol{\pi}$.

Dynamic information enters through the time-correlation function, which admits two equivalent forms [109]:

$$C_{\mathbf{o}}(k\tau) = \mathbf{o}^\top \operatorname{diag}(\boldsymbol{\pi})\, \mathbf{T}^k(\tau)\, \mathbf{o} \tag{5.1}$$

$$= (o^{\text{stat,MSM}})^2 + \sum_{i=2}^{n} c_i \exp\left(-\frac{k\tau}{t_i^{\text{ex}}}\right), \tag{5.2}$$

Here, k is a positive integer denoting the number of lag-time steps, $\operatorname{diag}(\boldsymbol{\pi})$ is the diagonal matrix formed from the stationary probabilities, and $\mathbf{T}^k(\tau)$ propagates the MSM by time $k\tau$. The amplitudes $c_i = (\mathbf{o} \cdot \mathbf{L}_i)^2$ (Eq. 3.6) couple the observable to each mode through the density-weighted eigenvectors $\mathbf{L}_i = \mathbf{R}_i \circ \boldsymbol{\pi}$ (see Fig. 3.1 for how different observables weight the same set of modes). The matrix form (Eq. 5.1) can be evaluated directly from $\mathbf{T}$ without diagonalizing it, which is appropriate for observables like NOE cross-relaxation rates that depend on the full correlation function [65]. The spectral form (Eq. 5.2) separates contributions by mode and is used, for example, in relaxation dispersion forward models, where the experimental observable is a function of the individual amplitudes and timescales [77, 92, 110].

Forward models. Biophysical experiments do not measure correlation functions directly, but rather forward functions thereof. The MSM provides the necessary ingredients in each case, which can be either through the spectral components as outlined earlier (amplitudes and timescales) or via the transition matrix $\mathbf{T}$ directly. Below are two examples from NMR for each.

Relaxation dispersion experiments ($R_{1\rho}$ and CPMG) probe conformational exchange on the µs–ms timescale [76, 92, 110]. Here the observable function $o(\mathbf{x})$ is the chemical shift that depends on the local environment of $\mathbf{x}$. It does not have a closed-form expression but it can be calculated using chemical shift predictors [111, 112]. The amplitudes $c_i = (\mathbf{o} \cdot \mathbf{L}_i)^2$ quantify how much of the chemical shift variance is explained by the exchange process i, and the dispersion profiles are functions of these amplitudes and their associated timescales t_i^{ex}:

$$R_{1\rho}(\nu_1) = \kappa_{R_{1\rho}} \sum_{i=2}^{m} c_i \frac{t_i^{\mathrm{ex}}}{1 + (t_i^{\mathrm{ex}} \cdot 2\pi\nu_1)^2} \tag{5.3}$$

$$R^{\mathrm{ex}}_{2,\mathrm{CPMG}}(\nu_{\mathrm{CP}}) = \kappa_{\mathrm{CPMG}} \sum_{i=2}^{m} c_i \, t_i^{\mathrm{ex}} \left(1 - \frac{t_i^{\mathrm{ex}}}{\tau_{\mathrm{CP}}} \tanh \frac{\tau_{\mathrm{CP}}}{t_i^{\mathrm{ex}}}\right), \tag{5.4}$$

where $\kappa_{R_{1\rho}}$ and κ_{CPMG} collect the NMR-specific prefactors[1], and ν_1 as well as ν_{CP} are the experimental running variables whose modulation traces out the dispersion profile.

Predicting **NOE cross-relaxation rates**, in contrast, uses $\mathbf{T}$ directly via Eq. 5.1 and does not require the spectral decomposition. NOEs depend on a spectral density function that is the Fourier transform of the dipolar time-correlation function between spin pairs [113]. The observable vector $\mathbf{o}$ encodes the dipolar coupling for each spin pair (i, j):

$$o(\mathbf{x}) \propto \frac{Y_2^m(\theta_{ij}(\mathbf{x}), \varphi_{ij}(\mathbf{x}))}{r_{ij}(\mathbf{x})^3}, \tag{5.5}$$

Here, Y_2^m is a second-rank spherical harmonic, with $m \in \{-2, \ldots, 2\}$ indexing the angular components of the dipolar interaction. where r_{ij} is the internuclear distance and θ_{ij}, φ_{ij} describe the orientation of the spin pair relative to the external magnetic field. The r_{ij}^{-3} dependence makes $\mathbf{o}$ sensitive to both the distance and the geometry of the spin pair within each state [65].

Beyond MSMs. Using MD-derived MSMs, we can now directly compare MSM-derived predictions of observables with their experimental counterparts. In practice, they often disagree (Sec. 4.2), and a natural question therefore becomes if this is due to the limitations of the MSM itself. After all, MSMs approximate the eigenfunctions of $\mathcal{T}_\tau$ using indicator functions over a hard state-space discretization [22]. Their predictions are therefore bounded by the quality of the molecular representation and the expressiveness of the basis functions. One can therefore wonder if using richer basis functions would reduce these discrepancies. Variational principles guarantee that richer basis functions (linear projections [27] or neural-network parameterizations [26, 89]) can only *improve* the spectral estimate compared to MSM indicator functions [43, 114]. However, all of these approaches optimize the same simulation data. More expressive basis functions therefore reduce the approximation error, but they cannot correct for systematic bias in the force field or recover transitions that were never sampled (Sec. 4.1). Richer representations can encode relevant properties such as local dynamics directly alongside geometric features [115], but no amount of representational flexibility can recover information absent from the training data, which motivates a fundamentally different sampling strategy.

[1] The full expressions with all prefactors are given in [92].

5.2 Generative Models

Generative models learn to draw samples from a target distribution by transforming them from a simple, tractable prior (typically a Gaussian) through a learned map [116]. Deep generative models parameterize this map with neural networks [116], enabling sampling from complex, high-dimensional state distributions [117]. Instead of doing femtosecond timesteps to sample the potential energy landscape like in MD, ML models efficiently sample from a learned Boltzmann distribution directly, without explicitly integrating the equations of motion to probe the equilibrium properties of biomolecular systems [29, 31, 118]. Boltzmann Generators (BGs) [29] or Boltzmann Emulators (BEs) [31] learn invertible maps from a simple latent distribution to approximations of $\pi(\mathbf{x})$ using flow- or diffusion-based models [119, 120]. This extends earlier probabilistic approaches that generated native-like protein conformations by encoding local structural bias [121]. The resulting samples, however, are not time dependent, meaning that we lose kinetic information. More recently, these models have shown impressive capabilities to generalize across chemical space [32].

A different approach that retains the kinetic information is to learn the transition density $p(\mathbf{x}_{t+\tau} \,|\, \mathbf{x}_t, \tau)$ using neural networks [30, 122–124]. Since they encode the time dependence between samples, they allow us to calculate dynamic observables, which equilibrium sampling models cannot. The Implicit Transfer Operator (ITO) approach [30] maps $\mathbf{x}_t$ to samples of $\mathbf{x}_{t+\tau}$ by implicitly encoding $\mathcal{T}_\tau$ without the need to discretize the states and even across different lag times τ. This directly learns the propagator and therefore sidesteps the need for expensive numerical integration that limits MD's ability to reach long timescales.

5.2.1 Score Models and SDEs.

Generative modeling approaches can be understood within a common mathematical framework. They learn to transform a probability distribution from a simple prior to a chosen target distribution [119, 125, 126]. The starting point is a *forward noising process* that gradually destroys the structure in the data. Given samples from a target distribution $q_0(\mathbf{x})$, we can define an SDE that progressively adds noise:

$$\mathrm{d}\mathbf{x}_t = f(\mathbf{x}_t, t)\,\mathrm{d}t + g(t)\,\mathrm{d}\mathbf{W}_t, \quad \mathbf{x}_0 \sim q_0, \tag{5.6}$$

where $f(\mathbf{x}_t, t)$ is a drift coefficient and $g(t)$ a scalar diffusion coefficient that are chosen such that the forward transitions $q_t(\mathbf{x}_t \mid \mathbf{x}_0)$ are Gaussian and available in closed form. Here, the time variable $t \in [0, 1]$ denotes *diffusion time*, not physical time: $t = 0$ corresponds to the data distribution and $t = 1$ to the noise distribution. As the forward process progresses from $t = 0$ to $t = 1$, the marginal q_t gradually loses its dependence on q_0, until all structure in the original data distribution has been replaced by noise.

The key insight that makes generative modeling possible is that this noising process (Eq. 5.6) can be reversed exactly. If the *score* $\nabla_{\mathbf{x}} \log q_t(\mathbf{x}_t)$, i.e., the

gradient of the log-density of the noised distribution at diffusion time t is known, samples from q_0 can be obtained by integrating the reverse-time SDE from $\mathbf{x}_1 \sim q_1$ [126, 127]:

$$\mathrm{d}\mathbf{x}_t = \left[f(\mathbf{x}_t, t) - g(t)^2 \nabla_{\mathbf{x}} \log q_t(\mathbf{x}_t)\right] \mathrm{d}t + g(t)\, \mathrm{d}\mathbf{W}_t. \tag{5.7}$$

The score $\nabla_{\mathbf{x}} \log q_t(\mathbf{x}_t)$ is intractable analytically since it involves the marginal density over all data. We approximate the score with a neural network $s_\theta(\mathbf{x}_t, t)$ trained via *denoising score matching* [126]. Because the forward transitions $q_t(\mathbf{x}_t \mid \mathbf{x}_0)$ are Gaussian by construction, the conditional score $\nabla_{\mathbf{x}} \log q_t(\mathbf{x}_t \mid \mathbf{x}_0)$ is available analytically, and thereby makes it practical to train a score model that generates samples from a learned approximation $p_0(\mathbf{x}_0)$ of the target distribution $q_0(\mathbf{x}_0)$. Replacing the ideal score $\nabla_{\mathbf{x}} \log q_t(\mathbf{x}_t)$ in Eq. 5.7 with $s_\theta(\mathbf{x}_t, t)$ induces a generated family of marginals p_t, where the terminal distribution p_0 approximates q_0.

In addition to the stochastic reverse process, the same family of marginal distributions can be generated deterministically by a *probability flow ODE* [126]:

$$\mathrm{d}\mathbf{x}_t = f(\mathbf{x}_t, t)\, \mathrm{d}t - \frac{1}{2} g(t)^2 \nabla_{\mathbf{x}} \log q_t(\mathbf{x}_t)\, \mathrm{d}t, \tag{5.8}$$

which enables exact likelihood computation via the instantaneous change-of-variables formula [128]. Continuous normalizing flows [128] learn this deterministic transport directly without the intermediate step of defining a forward noising process and learning the score [125].

The reverse-time SDE (Eq. 5.7) is structurally similar to the overdamped Langevin equation (Eq. 2.9): both are SDEs whose drift moves samples along gradients of a density or energy landscape. Langevin dynamics evolve configurations under the force field $\nabla_{\mathbf{x}} U(\mathbf{x})$ in physical time, whereas generative models evolve probability density from noise to data in diffusion time. Crucially, this framework is flexible enough to allow us to *choose* the target distribution. For BGs/BEs, the target is the Boltzmann distribution (Fig. 2.1), so the learned score at $t = 0$ approximates $s_\theta(\mathbf{x}_0, 0) \approx \nabla_{\mathbf{x}_0} \log \pi(\mathbf{x}_0)$; for ITO models, the generated variable is the future configuration conditioned on $\mathbf{x}_t$, and at $t = 0$ the learned score approximates the gradient of the transition density with respect to that future configuration. In both cases, the score function determines which distribution the model generates.

Importantly for this thesis, because the learned score near the end of the reverse process approximates the gradient of a physically meaningful density, such as the Boltzmann distribution or a transition density, modifications to the score correspond to modifications of the physical distribution that the model generates. This opens the possibility of perturbing the score to modify the generated distribution [129–131] and potentially counteract biases inherited from training data. Concretely, we add an auxiliary vector field $h_\vartheta(\mathbf{x}_t, t)$ to $s_\theta(\mathbf{x}_t, t)$. Substituting $s_\theta(\mathbf{x}_t, t) + h_\vartheta(\mathbf{x}_t, t)$ into Eq. 5.7 yields:

$$\mathrm{d}\mathbf{x}_t = \left[f(\mathbf{x}_t, t) - g(t)^2 (s_\theta(\mathbf{x}_t, t) + h_\vartheta(\mathbf{x}_t, t))\right] \mathrm{d}t + g(t)\, \mathrm{d}\bar{\mathbf{W}}_t, \tag{5.9}$$

which transports the same base/noise distribution $p_1 \approx q_1$ to a *different* generated distribution $p'_0(\mathbf{x}_0)$ (Fig. 6.1). Geometrically, this means that the perturbation shifts probability mass toward or away from certain regions in state space. Generative models provide a powerful alternative to MD for overcoming sampling limitations. Although they inherit biases from training data, adding an external vector field $h_\vartheta(\mathbf{x}_t, t)$ offers a principled intervention mechanism that can, in principle, correct those biases. The construction of $h_\vartheta(\mathbf{x}_t, t)$ is developed in Chap. 6.

Chapter 6

Integration of Experimental Observables via Constrained Statistical Inference

The preceding chapters established that simulations and experiments are complementary but incomplete views of the same underlying dynamics. This chapter introduces the framework for treating experimental observables as constraints on statistical dynamical models, thus reframing the relationship between experiments and simulations from one of cross-validation to one of integration. The question is how to combine both sources of information in a way that is consistent with the available data and does not make more assumptions than necessary.

6.1 Experimental and Physical Constraints on Admissible Spectral Components

Optimization using stationary observables is convex. Ideally, we want our predictions to match what we measure experimentally, but as established in Chap. 4, this agreement often fails to hold in practice. Our goal is therefore to *find* a corrected model that satisfies $o_i^{\mathrm{exp}} \stackrel{!}{=} o_i^{\mathrm{stat}}$. To do so, we must first understand what experimental measurements actually impose on the model.

Consider the finite-dimensional MSM setting with n states. The stationary observable is the inner product $o^{\mathrm{stat,MSM}} = \mathbf{o}^\top \boldsymbol{\pi}$ (Eq. 3.1). Crucially, $o^{\mathrm{stat,MSM}}$ is linear in $\boldsymbol{\pi}$. This also means that the experimental constraint $\mathbf{o}^\top \boldsymbol{\pi}' = o^{\mathrm{exp}}$ is a linear equation. For M observables, each constraint remains linear, and the set of all valid probability distributions that satisfy them simultaneously $\mathcal{C}_{\mathrm{stat}}$ is therefore convex, which makes optimization over it tractable.

Optimization using dynamic observables is non-convex. Dynamic observables, on the other hand, do not only constrain the stationary distribution but also the combination of eigenvectors and -values through the spectral decomposition (Eq. 5.2). By coupling the modes (through the amplitudes $c_i = (\mathbf{o} \cdot \mathbf{L}_i)^2$, Eq. 3.6) to the relaxation spectrum (through the timescales $\exp(-\tau/t_i^{\text{ex}})$, Eq. 2.5), a single observable does not just constrain one particular eigenvalue but the specific weighted combination of exponential decays filtered through the observable function. Different observables therefore "see" different linear combinations of the same spectral components (Fig. 3.1). In the MSM case, using Eq. 3.5, we get the spectral form for the correlation function:

$$o^{\text{dyn}}(k\tau) = \underbrace{(\mathbf{o} \cdot \mathbf{L}_1)^2}_{\text{stationary}} + \sum_{i=2}^{n} \underbrace{(\mathbf{o} \cdot \mathbf{L}_i)^2}_{\text{quadratic in } \mathbf{L}_i} \cdot \underbrace{\lambda_i^k}_{\text{power of eigenvalue}} . \tag{6.1}$$

This equation illustrates the nonlinearity of dynamic observables due to the quadratic term for the left eigenvectors and an exponential term for the eigenvalues. Compared to the stationary case, where $\mathbf{o}_i^\top \boldsymbol{\pi}' \stackrel{!}{=} o_i^{\text{exp}}$ is a linear equation, the constraint $o^{\text{dyn}}(k\tau) \stackrel{!}{=} o^{\text{dyn, exp}}(k\tau)$ is a product of unknown eigenvectors with unknown eigenvalues, making the optimization over the set of dynamic constraints $\mathcal{C}_{\text{dyn}}$ non-convex.

Physical Consistency Constraints. In addition to the experimental constraints, the physical assumptions on the transition matrix (stochasticity, detailed balance, and a unique stationary distribution, Sec. 2.1.2) further restrict the space of admissible models, which we denote $\mathcal{C}_{\text{phys}}$.

If we now compare the number of degrees of freedom in configuration space ($\approx 10^4 - 10^6$) with the number of experimental constraints typically at our disposal ($10^0 - 10^2$) [132, 133], the constraint set $\mathcal{C}_{\text{stat}} \cap \mathcal{C}_{\text{dyn}}$ is a large-dimensional manifold of admissible models and not one unique solution. And even if we further constrain the solution space with the physical constraints $\mathcal{C}_{\text{phys}}$ on the transition matrix, the intersection remains a high-dimensional set. What we therefore need is a principled selection criterion to select among the many distinct transition matrices that satisfy all physical and experimental constraints simultaneously.

6.2 Information-Theoretic Principles for Model Selection

The observable and physical constraints $\mathcal{C} = \mathcal{C}_{\text{stat}} \cap \mathcal{C}_{\text{dyn}} \cap \mathcal{C}_{\text{phys}}$, respectively, give us the set of admissible models, but we still need a principled way to rank them. The question then becomes how we do that with the least number of assumptions, and the answer to this question is at the heart of information theory [134, 135].

To make this concrete, consider the stationary case first. Our task is to find a corrected distribution $\boldsymbol{\pi}' \in \mathcal{C}_{\text{stat}}$ (Sec. 6.1). By definition, we know that

this set is consistent with the experimental evidence. If we select a distribution that concentrates probability mass on a few states (see effective sample size discussion in Sec. 6.2.2) such that its expectation value via the observable function $o(\cdot)$ corresponds to the experimental average, the result is a valid distribution, but it also makes strong claims about the relative importance of those states if we lack further evidence that would justify such a choice. If we, instead, choose a distribution $\boldsymbol{\pi}'$ that is closer to the prior distribution $\boldsymbol{\pi}$, the claim is weaker. The principle of maximum entropy of E. T. Jaynes formalizes this intuition [134]. The distribution that makes the fewest assumptions beyond the data we already have (i.e., the experimental data and simulations) is the distribution that maximizes the Shannon entropy: $H[\boldsymbol{\pi}'] = -\sum_j \boldsymbol{\pi}'_j \log \boldsymbol{\pi}'_j$. In our case, our prior knowledge already encodes a great deal of physical knowledge. $\boldsymbol{\pi}$ was estimated from MD trajectory data that was generated using a molecular force field (Eq. 2.9). This information is valuable because, to estimate $\boldsymbol{\pi}'$, we can use $\boldsymbol{\pi}$ as a good first guess that is then *corrected* using the experimental constraints. We can measure how much $\boldsymbol{\pi}'$ deviates from $\boldsymbol{\pi}$ using the Kullback-Leibler (KL) divergence [136]:

$$\mathrm{D}_{\mathrm{KL}}(\boldsymbol{\pi}' \,\|\, \boldsymbol{\pi}) = \sum_j \boldsymbol{\pi}'_j \log \frac{\boldsymbol{\pi}'_j}{\boldsymbol{\pi}_j}. \tag{6.2}$$

The KL divergence is non-negative and equals zero only if $\boldsymbol{\pi}' = \boldsymbol{\pi}$. By minimizing Eq. 6.2, we therefore get the distribution $\boldsymbol{\pi}^*$ that fulfills the experimental and physical constraints *and* is based on the prior knowledge derived from MD simulations in the form of the original stationary distribution [91, 135]:

$$\boldsymbol{\pi}^* = \underset{\boldsymbol{\pi}' \in \mathcal{C}_{\mathrm{stat}}}{\arg\min}\, \mathrm{D}_{\mathrm{KL}}(\boldsymbol{\pi}' \,\|\, \boldsymbol{\pi})\,. \tag{6.3}$$

It is also worth noting that $\mathrm{D}_{\mathrm{KL}}(\boldsymbol{\pi}' \,\|\, \boldsymbol{\pi})$ penalizes shifting probability mass where there is none, or very little, in $\boldsymbol{\pi}$. This avoids "inventing" states the MD simulation hasn't explored and instead shifts mass between the states that it has.

6.2.1 Maximum-Entropy Reweighting

Reweighting the stationary distribution using experimental observables according to maximum-entropy principles is a well-established approach [6, 91, 137–141]. We can solve Eq. 6.3 using the method of Lagrange multipliers for each observable constraint $m \in [M]$. In addition, we introduce a multiplier μ that enforces normalization: $\sum_j \boldsymbol{\pi}'_j = 1$. This gives us the Lagrangian [91, 135]:

$$\mathcal{L} = \sum_j \boldsymbol{\pi}'_j \log \frac{\boldsymbol{\pi}'_j}{\boldsymbol{\pi}_j} + \sum_{m=1}^{M} \alpha_m \left(\underbrace{\sum_j \boldsymbol{\pi}'_j o_{m,j}}_{=o_m^{\mathrm{pred}}} - o_m^{\mathrm{exp}} \right) + \mu \left(\sum_j \boldsymbol{\pi}'_j - 1 \right). \quad (6.4)$$

Solving this constrained minimization via Lagrange multipliers (see, e.g., [91, 135] and Appendix A1 of Paper IV) yields the exponentially tilted distribution

$$\boldsymbol{\pi}^*_j \propto \boldsymbol{\pi}_j \exp\left(- \sum_{m=1}^{M} \alpha_m o_{m,j} \right), \quad (6.5)$$

where the Lagrange multipliers $\{\alpha_m\}_{m=1}^{M}$ are determined by the constraint equation $\sum_j o_m(\pi^*_j) = o_m^{\mathrm{exp}}$. Eq. 6.5 has a clear interpretation: the weight of each state in the prior distribution $\boldsymbol{\pi}$ is scaled by an exponential factor that couples each observable function $o_m(\cdot)$ to its corresponding Lagrange multiplier α_m. If the discrepancy between the prior and the experimental data is minor, α_m will be close to zero. The bigger the discrepancy, the larger α_m will be.

6.2.1.1 Augmented Markov Models

Augmented Markov Models combine classical MSMs (Sec. 5.1) estimated from an MD trajectory with experimental constraints using the maximum-entropy reweighting framework [91]. They apply the exponential tilting of Eq. 6.5 to the discrete stationary distribution $\boldsymbol{\pi}$ of an MSM. The Lagrange multipliers $\{\alpha_m\}$ are fitted to match the experimental values, yielding a corrected stationary distribution $\boldsymbol{\pi}^*$ whose probability mass has been redistributed accordingly. This is a success story for the convex stationary optimization problem, and the corrected thermodynamics of the systems have even shown to improve the prediction of dynamic observables that were not involved in the procedure [91]. Why this is the case warrants a closer inspection.

6.2.2 Limitations of Maximum-Entropy Reweighting

The MaxEnt reweighting approach is widely adopted by the MD community [91, 137, 139–142]. While it yields the correct solution for reweighting the stationary distribution post hoc, it does not take into account the dynamic modes of the system.

Ignorance of kinetics. In the spectral decomposition of the transition matrix (Eq. 2.4), the stationary distribution $\boldsymbol{\pi}$ corresponds to the first eigenvector (i.e., $\lambda_1 = 1$; cf. Fig. 2.2). By definition, the reweighting does not affect the kinetics that are encoded in the remaining $n - 1$ eigenvector/-value pairs. However, despite the fact that AMMs [91] only modify the stationary distribution, they have shown to also improve the prediction of dynamic observables (Sec. 6.2.1.1).

This apparent contradiction is best understood as a special case. When the dominant source of force field error is in the state populations rather than in the barrier heights, correcting $\boldsymbol{\pi}$ can have an effect on the amplitudes of relaxation processes, which indirectly improves the agreement with dynamic observables. But this indirect improvement is contingent upon the kinetics being close to correct in the first place. An example of this is shown in Fig. 4.1. In Chap. 3 we have seen that two systems with identical stationary distributions can exhibit vastly different dynamics if the energy barriers between their states differ (Sec. 3.2). Since dynamic observables couple the amplitudes $c_i = (\mathbf{o} \cdot \mathbf{L}_i)^2$ to the relaxation rates through $\exp(-\tau/t_i^{\mathrm{ex}})$ (Sec. 6.1), they are sensitive to the spectral components that stationary reweighting cannot modify. As seen in Sec. 6.1, constraining the dynamic modes of $\mathbf{T}$ with observables leads to a non-convex optimization problem.

Weight degeneracy. As mentioned earlier, even in the stationary case where MaxEnt reweighting works in principle, the model can only weight the metastable states differently. If the prior is far off from the true distribution, i.e., the KL-divergence between the prior and the posterior is high, the model can compensate by concentrating probability mass on a few states so that the Effective Sample Size (ESS) collapses. ESS is a measure of quality for the reweighting [141], and its formula is given by

$$\mathrm{ESS} = \frac{\left(\sum_{j=1}^{J} w_j\right)^2}{\sum_{j=1}^{J} w_j^2}, \tag{6.6}$$

where $\{w_j\}_{j=1}^{J}$ are the unnormalized importance weights for every state in the distribution. ESS ranges from 1 (all weight concentrated on a single state) to J (uniform weights). When the prior $\boldsymbol{\pi}$ is far from the corrected distribution $\boldsymbol{\pi}^*$, the exponential tilting in Eq. 6.5 produces a few dominant weights and the ESS collapses accordingly [141].

Missing States. Furthermore, maximum-entropy reweighting applies only post hoc and therefore cannot address the underlying force field or sampling biases. If the simulation misses an entire metastable state, no reweighting scheme (stationary or dynamic) can recover it. Solving this problem therefore either requires changing the force field that generates the MD trajectory or, in the ML case, correcting the sampling mechanism of a learned generative model in a way that it shifts probability density towards the missing state.

6.3 Guidance of Generative Models

The maximum-entropy framework corrects a static distribution by finding the minimal KL-divergence perturbation that satisfies experimental constraints (Eq. 6.5). In Sec. 5.2.1, we introduced probability flow models as an alternative to conventional MD as well as the possibility to perturb the probability flow

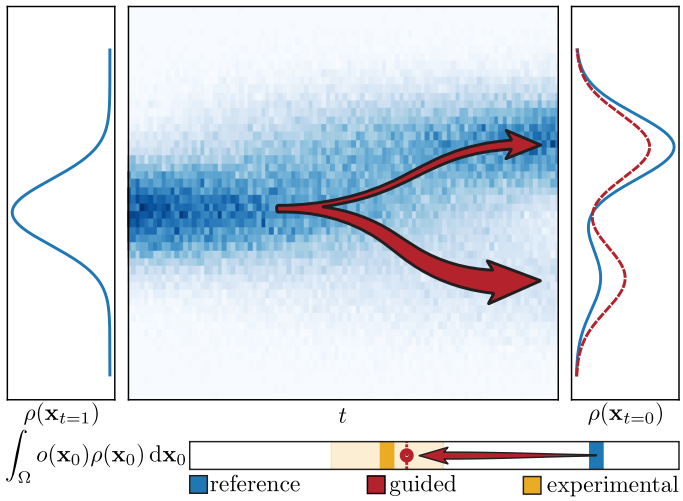

Figure 6.1: **Schematic of observable guidance for probability flow generative models.** The reference model (blue) generates samples by integrating a reverse-time process from $\rho(\mathbf{x}_{t=1})$ to $\rho(\mathbf{x}_{t=0})$. A guidance field $h_\vartheta(\mathbf{x}_t, t)$ perturbs the generative trajectory (red arrows), shifting the output distribution (red, dashed) so that its expectation value $\int_\Omega o(\mathbf{x}_0)\,\rho(\mathbf{x}_0)\,\mathrm{d}\mathbf{x}_0 = \mathbb{E}[o(\mathbf{x}_0)]$ matches the experimental target (yellow), while an information-theoretic regularizer ensures the perturbation introduces no more bias than the data warrant. The bar at the bottom illustrates the constraint. Adapted from [143].

using an external vector field (Eq. 5.9). Although reweighting and guidance have the same goal, i.e., to shift probability density according to external constraints, their approach is different. Guidance targets the generative process itself (Eq. 5.9; Fig. 6.1) instead of just $p(\mathbf{x}_0)$, as is the case for reweighting. Guidance allows the generative model to shift probability density even towards regions where the unperturbed model assigns low or zero density, thereby directly addressing the sampling limitation of MD simulations and hence, also the downstream modeling of the trajectory. It also avoids the ESS collapse in MaxEnt reweighting by construction. Guidance of probability flow models has been an active area of research [131, 144]. However, as we will see, none of the conventional methods apply to the observable constraint setting we are in.

Existing guidance strategies. In the image generation literature, several strategies exist for steering a pre-trained generative model at inference time [129–131, 145]. These methods all assume that supervision is available at the level of individual samples, either through per-sample labels [129], training-time conditioning [130], or per-step gradient signals derived from a single-sample loss [131, 145]. In our setting, however, the supervision is purely distributional, where a single scalar o_m^{exp} constrains the expectation $\mathbb{E}_{\mathbf{x}_0\sim p_0'}[o_m(\mathbf{x}_0)]$ over the entire generated ensemble, and no individual sample can be classified as correct or incorrect in isolation. A suitable framework would need to (i) operate at

inference time on a pre-trained model, (ii) accommodate purely distributional rather than per-sample constraints, and (iii) regularize the perturbation to avoid degenerate solutions. We therefore seek $h_\vartheta(\mathbf{x}_t, t)$ such that the predictions of the guided probability distribution $o(p'(\mathbf{x}_0))$ are in line with the experimental observations. Also, given that there are many possible $p'(\mathbf{x}_0)$ that fulfill that criterion, how do we choose $h_\vartheta(\mathbf{x}_t, t)$ such that it induces the least amount of bias?

6.3.1 Optimal Transport as Physical Regularization

In Sec. 6.2, we established that KL divergence provides a principled selection criterion for correcting static distributions. That is, among all $\boldsymbol{\pi}' \in \mathcal{C}_{\text{stat}}$, choose the one closest to the prior $\boldsymbol{\pi}$ (Eq. 6.3). In a generative setting where the sampling happens through integration of a reverse-SDE, this does not apply directly and we need an analogous notion of "closeness".

Optimal transport provides such a notion. Rather than measuring the distance between two distributions statically as in the KL-divergence case where we compare $\boldsymbol{\pi}'$ and $\boldsymbol{\pi}$, the formulation of Benamou and Brenier [146] recasts the problem for the vector fields $h_\vartheta(\mathbf{x}_t, t)$ that transport a distribution p to a distribution q in unit time. Choosing among all vector fields, it says to find the one that does so with the least total kinetic energy:

$$\inf_h \int_0^1 \int_\Omega \|h_\vartheta(\mathbf{x}_t, t)\|^2 \, p_t(\mathbf{x}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}t, \tag{6.7}$$

subject to mass conservation[1]. The minimum of this functional is the standard metric on the space of probability distributions, namely the squared Wasserstein-2 distance $\mathcal{W}_2^2(p, q)$ [146, 147]. The optimal velocity field is the one that moves mass from p to q with the smallest possible deviation from doing nothing. This intuition builds on the KL-divergence case we used for maximum-entropy reweighting.

The Benamou–Brenier formulation [146] is stated for deterministic flows, but its logic extends naturally to the stochastic setting [148, 149]. Consider a reference generative SDE (Eq. 5.7) that produces a target distribution $p(\mathbf{x}_0)$. Adding a perturbation $h_\vartheta(\mathbf{x}_t, t)$ to the drift yields a modified process with a different terminal distribution $p'(\mathbf{x}_0)$. The kinetic energy of the perturbation

$$\mathcal{E}[h] = \int_0^1 \mathbb{E}_{\mathbf{x}\sim p'_t}\left[\|h_\vartheta(\mathbf{x}_t, t)\|^2\right] \, \mathrm{d}t, \tag{6.8}$$

measures how much the corrected process deviates from the reference dynamics. It upper-bounds both the KL divergence $\mathrm{D}_{\mathrm{KL}}(p'(\mathbf{x}_0) \,\|\, p(\mathbf{x}_0))$ and $\mathcal{W}_2^2(p'(\mathbf{x}_0), p(\mathbf{x}_0))$ between the reference and corrected terminal distributions [148, 149]. Minimizing $\mathcal{E}[h]$ therefore selects the perturbation that changes the generative process as little as possible and represents the probability-flow-space analog of minimizing $\mathrm{D}_{\mathrm{KL}}(\boldsymbol{\pi}' \,\|\, \boldsymbol{\pi})$.

[1] $\partial_t p_t + \nabla \cdot (p_t \, h_\vartheta(\mathbf{x}_t, t)) = 0$ with $p_{t=0} = p$ and $p_{t=1} = q$

6.3.2 Combining Regularization with Observable Constraints

The Benamou–Brenier formulation [146] gives us a principled selection criterion for choosing among the infinitely many drift perturbations $h_\vartheta(\mathbf{x}_t, t)$ that could correct a generative model. We still need to fulfill the observable constraint, namely that $o(p'(\mathbf{x}_0)) \stackrel{!}{=} o^{\text{exp}}$. Combined with the experimental constraints $\mathcal{C}$ (Sec. 6.1), we can therefore use Eq. 6.8 as an optimization objective for the corrected generative model:

$$\min_h \int_0^1 \mathbb{E}_{\mathbf{x}_t \sim p'_t} \left[\|h_\vartheta(\mathbf{x}_t, t)\|^2 \right] \mathrm{d}t \quad \text{subj. to } \mathbb{E}_{\mathbf{x}_0 \sim p'_0}[o_m(\mathbf{x}_0)] = o_m^{\text{exp}} \;\; \forall\, m \in [M], \tag{6.9}$$

where p'_t is the density induced by the perturbed SDE and $p'_0(\mathbf{x}_0)$ is the resulting corrected generated distribution. This is the probability-flow-space analog of the stationary reweighting problem (Eq. 6.3) that minimizes deviations from the prior subject to agreement with experiments. However, Eq. 6.9 exposes a difficulty with no counterpart in the stationary case. The constraints are *expectation values* under the corrected distribution $p'_0(\mathbf{x}_0)$, which is itself an implicit function of h through the entire generative trajectory (Sec. 6.3). Moreover, the M experimental observables are sparse relative to the degrees of freedom of the generative model, so the constraints alone are far too weak to determine h uniquely. The practical challenge is therefore to combine the kinetic energy regularizer with the observable constraints into a tractable objective that allows for gradient-based optimization despite the distributional nature of the constraints.

This chapter developed constrained inference in two stages: first, experimental observables are treated as constraints on the stationary distribution and the maximum-entropy principle selects the least biased correction (Eq. 6.3); second, the same logic is applied to generative models. The stationary problem is solved for thermodynamics [91, 137, 139], but it cannot modify kinetics, and it cannot recover states that the simulation never sampled. Rather than reweighting a fixed ensemble, we perturb the drift of the reverse-time SDE and regularize the perturbation through its kinetic energy, thus selecting the most conservative correction in the sense of optimal transport (Eq. 6.9). The key difficulty is that the observable constraints are expectation values under a distribution that is itself an implicit function of the perturbation, which makes the optimization problem non-standard compared to the classical ML setting [145, 150]. The appended papers develop practical solutions to this problem.

Chapter 7

Summary of Included Papers

Chapters 2 through 6 developed the mathematical and conceptual framework for understanding *why* the limitations of simulations, experiments, and machine learning arise, *where* they enter the modeling pipeline, and *how* experimental data and computational models can be combined through constrained statistical inference. The five appended papers build on this framework by mapping the landscape of current ML approaches in MD, diagnosing the gap between simulations and experiments empirically, proposing methods to close it, and validating the integrative modeling approach on a real protein system.

7.1 Machine Learning in Molecular Dynamics Simulations of Biomolecular Systems

Classical molecular dynamics simulations are fundamentally limited in ways that directly affect how well we can characterize biomolecular systems [62, 151–153] (Chap. 4). In this book chapter [154], we provide a broad, structured survey of how machine learning techniques have been developed and deployed to address these limitations, both in conjunction with and independently of MD simulations. The chapter is organized around three fundamental problems that span the modeling pipeline introduced in the preceding background chapters: estimating and analyzing molecular kinetics, sampling thermodynamically stable states, and parameterizing coarse-grained force fields. The key finding across all three domains is that none of the existing approaches include a principled mechanism for ensuring consistency with experimental observations, a diagnosis that motivates the subsequent papers in this thesis.

For molecular kinetics, the chapter outlines how Markov state models approximate the transfer operator $\mathcal{T}_\tau$ (Chap. 2) through spatial discretization [22, 155, 156], and how variational principles (VAC, VAMP) generalize this by allowing richer basis functions [43, 89, 114]. It also describes how neural network approaches such as VAMPnets learn the dominant eigenfunctions

(Fig. 2.3) end-to-end [26]. The setup presented in the book chapter illustrates that more expressive operator bases reduce approximation error but cannot correct for force-field bias or missing rare events, because all of these methods work on the same underlying simulation data.

For sampling, we survey how generative models learn to map between simple latent distributions and the complex equilibrium distribution of a molecular system [29, 30]. We also cover ML-driven enhanced sampling strategies that learn collective variables or biasing potentials to accelerate barrier crossings [24, 157]. The same diagnosis applies to the generative and enhanced-sampling approaches we survey [24, 29, 30, 157]. Even though the generative paradigm changes *what* we learn, it does not change the *source* we learn it from. For coarse-grained (CG) force fields, we discuss how the free energy of simplified bead representations can be learned from all-atom reference data via force matching, and identify three outstanding challenges: learning the mapping function itself, transferability across thermodynamic and chemical space, and faithful reproduction of kinetics [158–160]. Learning CG force fields also connects to the transfer operator framework, since a CG model that fails to preserve spectral properties will yield incorrect exchange timescales even if the thermodynamics are well reproduced [161].

The book chapter serves two roles in the thesis. First, it expands on the conceptual foundation of ML tools surveyed in Chap. 5 by identifying the key neural network architectures and their physically motivated learning objectives. Second, the diagnosis that few existing approaches ensures experimental consistency frames the research question that the rest of this thesis answers: how can we combine the expressive power of machine learning with the empirical grounding of experimental data to build more accurate models of biomolecular dynamics? The NOE/MSM study (Paper II) demonstrates this gap empirically, dynAMMo (Paper III) addresses it at the level of kinetic model correction, and MEW Guidance (Paper IV) pushes the integration upstream to the generative process itself.

7.2 Motional Clustering in Supra-τ_c Conformational Exchange Influences NOE Cross-Relaxation Rate

Methyl-labeled NMR spectroscopy is one of the primary tools for studying the structure and dynamics of large proteins, where traditional backbone-based approaches become impractical [162, 163]. Methyl–methyl NOE cross-peaks encode distance information between pairs of methyl groups, and their quantitative interpretation requires understanding how conformational dynamics modulate the cross-relaxation rates that give rise to these signals. In this study [65], we set out to predict methyl–methyl NOE cross-relaxation rates from all-atom MD simulations of the 723-residue enzyme Malate Synthase G, with the goal of establishing a computational framework that can be compared against experimental NOESY spectra.

A key challenge is that conformational exchange slower than the rotational correlation time τ_c requires a state-specific treatment of distance averaging rather than a single trajectory average [164]. Estimating a transition matrix yields exchange timescales that classify each methyl–methyl contact as sub- or supra-τ_c, and for slowly exchanging pairs the cross-relaxation rate is computed as a population-weighted combination of state-specific contributions. Since the NOE cross-relaxation rate can be expressed as a correlation function over the transition matrix (Eq. 5.1 and Eq. 5.5), the entire prediction pipeline is grounded in the operator-theoretic framework of the background chapters.

This distinction between the two regimes can be substantial. The paper uses the example of the methyl pair Ile346$^{\delta 1}$–Leu362$^{\delta 2}$, which has two metastable states with effective distances of 3.6 and 6.2 Å that interconvert on the supra-τ_c timescale. Accounting for this exchange properly yields a cross-peak intensity that differs by up to a factor of two from the naive trajectory average. Across the entire protein, roughly 100 out of 1,500 methyl–methyl contacts within the distance cutoff are affected by supra-τ_c exchange.

A more consequential finding appears when the same analysis is repeated across five force fields. Even after the observable prediction is done correctly, the variation between force fields is significantly larger than the correction introduced by accounting for supra-τ_c exchange. The framework enables a rigorous comparison, but it also exposes a problem this study cannot solve. Improved observable calculation cannot address the systematic gap between simulated and true dynamics.

The paper concludes by offering two remedies: either improving force field parameters and re-running the simulations, or correcting the simulation-derived model post hoc using experimental constraints. Paper III (dynAMMo) develops the latter for MSMs. Generative models, meanwhile, largely overcome the sampling limitations of classical MD (Sec. 5.2), and guidance methods such as those developed in Paper IV (MEW Guidance) offer a promising avenue for addressing force-field biases by correcting the generative process itself.

7.3 Rescuing Off-Equilibrium Simulation Data through Dynamic Experimental Data with dynAMMo

dynAMMo [77] corrects simulation-derived kinetic models by formulating MSM estimation as a constrained optimization over both transition counts and dynamic experimental observables (Eq. 5.3 and Eq. 5.4). Where the NOE/MSM study (Paper II) diagnoses the gap between simulated and measured dynamics, dynAMMo addresses this directly. We show that dynAMMo recovers correct exchange timescales from deliberately biased simulations and, in the most demanding test, reconstructs the slowest kinetic process from experimental relaxation data alone when all connecting transitions are absent from the simulation. The optimization acts on the spectral components of $\mathbf{T}$ (eigenvalues, eigenvectors, and stationary distribution; Eq. 2.4, Fig. 4.1) which, together,

determine both the populations and the predicted relaxation profiles. The dynAMMo objective combines the simulation and experimental data in a single loss function [77]:

$$\mathcal{L}\big(\hat{\mathbf{T}} \mid D, C\big) = -\sum_{ij} c_{ij} \log p_{ij} + \sum_{m,k} \big(o_{m,k}^{\mathrm{pred}} - o_{m,k}^{\mathrm{exp}}\big)^2, \tag{7.1}$$

Here, $C = \{c_{ij}\}$ is the transition-count matrix, D denotes the experimental observable data, and p_{ij} is the transition probability from state i to state j under $\hat{\mathbf{T}}$. Lagrange multipliers enforce the physical constraints (detailed balance, ergodicity, and conservation of probability Eq. 2.1.2) and are updated jointly with the spectral components via gradient descent.

In addition, our information-theoretic considerations play an important role in regularizing the estimation. Because the experimental constraints are sparse, many kinetic models are consistent with the data. dynAMMo addresses this ambiguity by seeking the model that deviates minimally from the simulation-derived prior while satisfying the experimental constraints according to maximum entropy principles (Sec. 6.2), extending the minimum-KL principle used for stationary reweighting (Eq. 6.3) to the full spectral decomposition of $\mathbf{T}$. The result is the least-informative correction to the simulation model that is compatible with the available experimental evidence.

We validate dynAMMo on two controlled benchmark systems where the ground truth is known, demonstrating that it recovers correct exchange timescales from simulations with deliberately biased potentials. A particularly informative test case is the disconnected simulation setting: when we artificially remove all transitions connecting the major and minor states, the standard MSM loses the slowest process entirely, but dynAMMo recovers it quantitatively from the experimental relaxation dispersion data alone. We then apply dynAMMo to BPTI [82] using published backbone ^{15}N CPMG data [165] at two magnetic field strengths and obtain a single mechanistic kinetic model that is globally consistent with all experimental dispersion profiles while retaining the atomistic detail of the underlying MD simulations.

dynAMMo thus demonstrates that sparse dynamic experimental data can rescue incomplete or biased simulation data and yield quantitatively predictive mechanistic models. Its fundamental limitation is that it can only operate on states that have already been sampled in the simulation and cannot generate new conformations or fill in unvisited regions of configuration space. This motivates methods that intervene earlier in the modeling pipeline by correcting the generative process itself rather than the kinetic model derived from it.

7.4 Minimum-Excess-Work Guidance

The Minimum-Excess-Work (MEW) framework [67] integrates experimental constraints into pre-trained probability flow generative models by introducing an additive perturbation $h_\vartheta(\mathbf{x}_t, t)$ to the learned score function (Eq. 5.9; Fig. 6.1). The key question is how to choose $h_\vartheta(\mathbf{x}_t, t)$ so that the corrected distribution agrees with experimental expectations while deviating as little as possible from

the prior, and Observable Guidance proposes a solution where the perturbation takes the form

$$h_\vartheta(\mathbf{x}_t, t) = -\eta_t(\vartheta) \sum_m \alpha_m \nabla_{\mathbf{x}_t} O_m(\hat{\mathbf{x}}_t(\mathbf{x}_t)), \tag{7.2}$$

Here, O_m is the differentiable observable function used for guidance, $\hat{\mathbf{x}}_t(\mathbf{x}_t)$ denotes the posterior mean estimate of the clean sample at diffusion time t [166, 167], and $\eta_t(\vartheta)$ is a scalar activation function. The guidance field $h_\vartheta(\mathbf{x}_t, t)$ is the gradient of the observable function with respect to the noisy state $\mathbf{x}_t$, and the Lagrange multipliers α_m are the same object as in the static maximum-entropy solution (Eq. 6.5), where each one couples an observable to its experimental constraint and tilts the distribution accordingly. Observable Guidance thus extends the maximum-entropy principle from static reweighting of a fixed ensemble to the generative modeling realm. In the static setting, the multipliers produce the exponentially tilted distribution that is closest in KL divergence to the prior (Eq. 6.3). In the generative setting, the same multipliers steer the probability flow toward an observable-consistent terminal distribution, enabling the model to place probability mass in regions where the unperturbed model assigns low or zero density (Fig. 6.1).

The theoretical foundation for this extension is the excess work functional (Eq. 6.8), which quantifies the kinetic energy of $h_\vartheta(\mathbf{x}_t, t)$ integrated over the generative trajectory. Through the Benamou–Brenier formulation of optimal transport [146], this functional upper-bounds both the KL divergence and the squared Wasserstein-2 distance between the reference and corrected terminal distributions, providing a principled notion of closeness in probability flow space. Minimizing excess work subject to experimental constraints (Eq. 6.9) therefore selects the perturbation that changes the generative process as little as is needed (Eq. 6.3). The paper also develops Path Guidance for targeted sampling of rare transition states; that work was done in collaboration and is not part of this thesis.

We validate Observable Guidance on two coarse-grained protein Boltzmann Emulators. On chignolin, guidance corrects the folding free energy from -1.27 to -1.82 ± 0.01 kcal/mol against an experimental target of -1.87 kcal/mol, demonstrating that the correction recovers distributional fidelity rather than merely fitting the target observable. On EnHD, MEW Guidance was applied to the state-of-the-art Boltzmann Emulator BioEmu [31] using only ten experimental ${}^3J_{\text{HN-HA}}$ couplings [168], reducing the quality factor [169, 170] from 0.147 to 0.037. These empirical results also show that the effective sample size collapse that plagues post-hoc reweighting (Eq. 6.6) is avoided.

MEW thus moves the integration of experimental constraints from the level of model analysis, where dynAMMo (Paper III) operates, to the level of sample generation. Observable Guidance enables a principled correction of probability flow generative models on an ensemble level. The corrected ensemble is consistent with both the learned prior and the experimental evidence and provides a physics-grounded alternative to standard fine-tuning when experimental constraints are sparse relative to the model's degrees of freedom, as is typical for dynamic NMR observables.

7.5 Conformational Quenching in an Engineered Lipocalin Protein Achieves High Affinity Binding to the Toxin Colchicine

The engineered lipocalin Colchicalin binds the clinically relevant plant toxin colchicine with picomolar affinity ($K_D = 120$ pM), yet the mechanism of ligand recognition remained unresolved. Crystallographic analysis showed that the overall β-barrel fold is preserved upon binding, but loop L3 undergoes a large outward displacement to vacate the binding pocket [171], raising the question of whether recognition proceeds by induced fit or conformational selection. This distinction is increasingly understood to be quantitative rather than absolute. NMR relaxation dispersion studies of galectin-3 have shown that ligand-specific stabilization of a transition state can switch the dominant binding pathway between conformational selection and induced fit [172], and that the energetic structure of the binding pathway can be mapped through linear free-energy relationships [173]. Answering this question requires a kinetic model that resolves the exchange between binding-competent and binding-incompetent conformations, but MD simulations of the apo state never sampled the slow loop transition connecting these states. Without observed transitions, a conventional MSM cannot be estimated, and the experimental CPMG data alone report on exchange timescales without providing a structural model. In this study [78], we use dynAMMo (Paper III) to bridge exactly this gap. We integrate the disconnected simulation ensembles with NMR relaxation dispersion data and obtain the first kinetic model of the Colchicalin conformational exchange, identifying three-state exchange in the apo state and establishing conformational selection as the dominant binding mechanism.

While fast ps–ns backbone dynamics were unchanged upon binding, ^{15}N CPMG relaxation dispersion experiments revealed substantial µs–ms exchange in the apo state, concentrated in the L3 loop region. These exchange contributions were markedly reduced upon colchicine binding, indicating that complex formation quenches the large-amplitude loop motions. The resulting dynAMMo model identified three-state exchange in the apo form. Two binding-incompetent states, comprising ∼97% of the equilibrium population, are characterized by loop L3 occluding the binding pocket in fully closed and partially closed conformations. A third, low-population binding-competent state (∼3%) features a displaced L3 that exposes the ligand pocket, with exchange occurring in the fast millisecond regime. The apo state thus transiently samples a minor open conformation that permits colchicine access, and ligand binding then quenches the slow loop dynamics and rigidifies the complex. In the context of this thesis, the lipocalin study validates the integrative approach in a real protein engineering setting. Using dynAMMo yields a global kinetic model at atomic resolution that mechanistically describes the conformational exchange across the entire protein, providing insights into the binding mechanism and the role of loop dynamics in ligand recognition that may inform rational drug design for lipocalin-based therapeutics.

Chapter 8

Discussion and Future Work

8.1 Discussion

A recurring finding across all papers in this thesis is that the observables that carry the richest information about biomolecular dynamics are precisely the ones that are hardest to predict from simulations [6].

Dynamic constraints recover unsampled kinetics. A key insight throughout all papers is that the integration of and subsequent optimization using dynamic observables (e.g., relaxation dispersion or NOE cross-relaxation) is significantly harder than using stationary observables only. While convex optimization methods are enough for stationary constraints, the nonlinear nature of dynamic constraints (quadratic in eigenvectors and exponential in eigenvalues, Eq. 6.1) requires much more careful consideration. The dynAMMo study (Paper III) and the subsequent lipocalin study (Paper V), on the other hand, show that dynamic constraints recover exchange timescales absent from simulation data. The method leverages the fact that time correlation functions encode information about the eigenspectrum of the transfer operator through the amplitudes and decay rates of the exchange processes (Fig. 3.1). Stationary observables, which by definition do not contain this information, are therefore not sufficient to bridge entirely disconnected regions of state space.

Sparse constraints generalize beyond their targets. The empirical results of the study also suggest that the guided ensemble is consistent across different, cross-validated observables. Likewise, dynAMMo (Paper III) and the lipocalin study (Paper V) demonstrate that a single dynAMMo model fitted to backbone ^{15}N CPMG dispersion data recovers a three-state kinetic model whose predictions are consistent with independent methyl CPMG and NOE experiments that were never used in the optimization. This would not be possible if different observables addressed isolated spectral components. A

common thread between dynAMMo and MEW Guidance is therefore that sparse experimental constraints improve the model well beyond the specific quantities they target, though the mechanism differs between the stationary and the dynamic setting.

In the dynamic case, the spectral structure of correlation functions (Eq. 3.5) explains this behavior. The spectral decomposition of the correlation function (Eq. 3.5) shows that each observable filters the eigenspectrum through its amplitudes c_i (Fig. 3.1). When two observables report on the same processes, i.e., they have a non-negligible amplitude on the same eigenmode, correcting the ensemble to match one observable simultaneously improves the prediction of the other.

For the stationary case in MEW Guidance, eigenmodes do not enter, because the constraints are expectations under the equilibrium distribution (Sec. 3.1) rather than time-correlation functions. Here, the mechanism is that the observable functions $o_i(\mathbf{x})$ are structurally correlated in configuration space. The covariance analysis of the Engrailed Homeodomain (EnHD, [174]) ^3J-couplings (Paper IV, Fig. S7) shows that 10 out of 43 couplings capture most of the conformational variation, because the backbone dihedral angles they depend on are not independent. This structural redundancy explains why correcting a sparse subset of observables improves cross-validated observables that were not part of the guidance. The constrained and unconstrained observables report on the same underlying degrees of freedom. On chignolin, correcting a single thermodynamic observable reduced the KL divergence to the reference distribution from 0.329 to 0.005; on EnHD, ten couplings reduced a cross-validated quality factor from 0.147 to 0.037. In both cases, the practical implication is the same. The effective dimensionality of the observable space appears to be low enough in the studied protein systems that a modest number of well-chosen experiments constrains the model globally.

From static reweighting to dynamic correction. The MEW Guidance study (Paper IV) addresses a structural limitation of the maximum-entropy framework developed in Chap. 6. Classical MaxEnt reweighting provides the information-theoretically best answer to the question: given a fixed set of samples, what is the closest distribution in KL divergence that satisfies the experimental constraints? This is effective when the prior ensemble already contains sufficient support in the relevant regions of configuration space. As discussed in Sec. 6.2.2, when the prior is far from the target, the exponential tilting concentrates probability mass on a few samples and the effective sample size collapses. It exposes the limitations that redistributing weight among existing samples cannot create probability mass where none exists. MEW seeks the minimal perturbation of the generative process that yields samples satisfying the constraints. The regularizer is no longer a KL divergence between distributions but the integrated kinetic energy of the perturbation along the probability flow path (the excess work in the Benamou–Brenier sense). This shifts the integration of experimental constraints from a static reweighting problem to a dynamic optimal transport problem, even for stationary observables. The excess work bounds (Propositions 1 and 2 in Paper IV) guarantee that

the guided distribution stays close to the prior in both Wasserstein distance and KL divergence simultaneously, with the excess work acting as an upper bound on both. The practical consequence is visible in the EnHD experiment: post-hoc reweighting achieves comparable observable agreement but at an effective sample size of 0.255, whereas MEW avoids this collapse entirely because it modifies the generator rather than reweighting its output. The key conceptual contribution of MEW is therefore circumventing the ESS collapse that limits importance-sampling-based reweighting approaches by modifying the generative process rather than its output. It does this by changing the mathematical category of the problem from static inference to dynamic control of the generative trajectory.

NOE cross-relaxation as an end-to-end benchmark. The NOE/MSM study (Paper II) reveals that proper modeling of NOE cross-relaxation rates from simulations involves a series of modeling choices that each introduce errors. The consequence of the sensitivity of NOEs is one of the most stringent tests available of a computational model. Because the cross-relaxation rate depends simultaneously on the force field accuracy (through local geometry), the sampling completeness (through the population of metastable states), and the correct treatment of motional timescales (NOEs integrate motions from the ps to the ms timescale [63]), a model that quantitatively reproduces methyl—methyl NOE build-up curves has necessarily gotten all three right. This makes NOE build-up curves one of the most demanding benchmarks for validating integrative models of biomolecular dynamics.

8.2 Limitations

Error propagation. A limitation shared by dynAMMo (Paper III) and MEW Guidance (Paper IV) is that neither method currently distinguishes between three layers of uncertainty that enter the constrained inference problem. Both methods report errors by running N optimizations from different initializations or random seeds, quantifying sensitivity to the optimization procedure. This can be referred to as *optimization uncertainty*. For dynAMMo, this is due to the non-convexity of the dynamic constraints; for MEW, it is due to the stochasticity of the generative process and the Bayesian optimization of guidance parameters.

The second layer is *experimental uncertainty*. The measurements that define the constraints carry their own errors, which determine how tightly the model needs to match the experimental values. Currently, both methods treat the experimental targets as exact, meaning that the optimization may spend effort fitting within the variance of the noise of the measurement.

The third layer is *forward model uncertainty*. The predicted observable values depend on empirical parameters of the observable function, e.g., the parameterization of the Karplus equation [66] for ^3J-couplings (Paper IV) or the error in the chemical shift predictor used to calculate relaxation dispersion experiments (Paper III). Errors in these parameters affect the target the

optimization is trying to reach.

What makes matters worse is that these three layers are coupled. An error in the experimental measurement or in the forward model changes the feasible set of solutions, which in turn changes what the optimization can and should converge to. Because the methods currently conflate all three sources into a single residual, it is not possible to determine whether a remaining discrepancy between prediction and experiment reflects an inadequate model, noisy data, or an inaccurate forward model. A more thorough error modeling is therefore required to investigate said discrepancy.

Dependence on the quality of the prior. Both dynAMMo (Paper III) and MEW Guidance (Paper IV) apply minimal corrections to an existing model, either through maximum-entropy reweighting of a kinetic model or through minimum-excess-work perturbation of a generative process. The underlying principle in both cases is to change as little as possible while satisfying the experimental constraints. This is appropriate when the prior model is already close to the truth, but it becomes a limitation when the prior is substantially wrong. If the force field carries large systematic errors, as is often the case for intrinsically disordered proteins [175, 176], or if the generative model has failed to learn important regions of configuration space, the minimum-perturbation solution will find the nearest feasible model to a poor prior, which may not be the physically correct one. Neither method provides a diagnostic for whether the prior is adequate in the first place. In practice, cross-validation against observables whose amplitudes weight different regions of the eigenspectrum (Fig. 3.1) offers some protection. But this requires additional experimental data and does not guarantee that unobserved degrees of freedom are well described.

Fixed state decomposition in dynAMMo. dynAMMo operates on a discrete Markov state model whose state decomposition (the number of states and their discretization in configuration space) is fixed before the experimental constraints enter the optimization. Discretization errors (both in the number of states and in the low-dimensional projection used to define them) have been shown to affect the quality of estimated eigenfunctions [22]. If the projection fails to separate conformations that are kinetically distinct, the state boundaries will not align with the true metastable structure, and the subsequent constrained optimization inherits this problem. Variational approaches such as VAMPnets [26], which learn eigenfunctions in continuous space and optimize the projection end-to-end (Sec. 5.1.2), would in principle remove this bottleneck, though integrating experimental constraints into such a framework remains an open problem.

8.3 Future Work

Guidance of time-lagged generative models using dynamic constraints. A natural next step in this line of work is to generalize MEW Guidance to include dynamic observables as well. MEW currently only enforces stationary

constraints; extending it to time correlation functions therefore requires expressive time-lagged generative models that depend on the transition density rather than on the stationary distribution only. There are two possible routes, which differ in interpretability and technical feasibility.

First, it would be advantageous to explicitly learn the slow eigenfunctions and corresponding eigenvalues instead of the transition density implicitly (as in ITO [30]). Current ITO models encode the transition density $p(\mathbf{x}_{t+\tau} \mid \mathbf{x}_t, \tau)$ implicitly in the weights of the network. This enables the generation of time-lag-conditioned samples without having access to the dominant exchange processes that determine the evolution of the probability density. Such a model would allow for the incorporation of dynamic experimental data in a similar fashion to dynAMMo [77] but without the limitations of state-space discretization or post-hoc modeling. However, learning continuous eigenfunctions that satisfy orthonormality under the π-weighted inner product (see Sec. 6.2.2) is significantly more difficult than in the MSM case [26, 77, 177]. In addition, the experimental constraints reintroduce the same problems of optimizing non-convex manifolds. Thus, a natural synthesis of continuous generative modeling and spectral constraints would be a remedy to most of the downstream modeling artifacts we currently face, but its technical implementation would come with open questions that go beyond what either method can solve on its own.

A more feasible alternative in the medium term would be to treat the ITO model as a black box. Rather than learning eigenfunctions explicitly, one would use the model's time-lagged samples to compute correlation functions $C_o(\tau)$ and match them against experimental measurements by perturbing the score function (Fig. 6.1, analogous to the stationary MEW setup). The trade-off is that this approach does not give direct access to the leading eigenmodes. This solution would be a less interpretable one compared to the explicit transfer operator approach but a technically much easier compromise.

Experimental design for integrative modeling. A second direction is to develop strategies for choosing which experiments to run in order to maximally reduce the remaining uncertainty in the model. The discussion above shows that many observables carry overlapping information; the design question is which ones do not. That is, the question is which proposed measurement would maximally reduce the set of models given the current constraints.

The first is how to quantify the information content of a single observable about the spectrum of $\mathcal{T}_\tau$. The amplitudes c_i (Eq. 3.5, Fig. 3.1) determine which eigenmodes an observable reports on, but they depend on the eigenfunctions themselves, which are part of what the model is trying to determine. This creates a circular problem that requires careful consideration.

The second is how to measure complementarity. Two observables may each be highly informative in isolation but redundant together if they share spectral support, so the relevant quantity is the conditional information given the constraints already imposed.

The third question is about what the right design objective is. For stationary observables, the uncertainty is over the population vector. The constraint set is convex and established criteria from Bayesian experimental design apply

directly [178, 179]. For dynamic observables, the uncertainty lives in eigenvalues and eigenvectors jointly on a non-convex manifold, and it is not obvious how to define an analogous objective. A practical approximation would be to run many dynAMMo optimizations from different initializations to characterize the space of feasible solutions, and then ask which proposed experiment would shrink that space the most. This requires a principled information-theoretic formulation and remains an open problem.